Sequential Hypothesis Tests of Multinomial Count Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sequential hypothesis tests of equality and contrasts among arbitrarily many time-inhomogeneous Bernoulli and Poisson counting processes are constructed based on a new sequential test of multinomial point hypotheses. For multinomial, Bernoulli and Poisson counting processes we provide confidence sequences for the probability vector, all contrasts in log-probabilities and all contrasts in log-intensities, respectively. Combined with sequential p-values, these provide an "always-valid" framework for inference, controlling the Type I probability under optional stopping/continuation and continuous monitoring. These methods are demonstrated with three applications relevant to online controlled experiments: sample ratio mismatch testing, conversion rate optimization, and software canary testing.

1 Introduction

In many areas of experimentation, one is frequently interested in the comparison of *counts* between treatment groups of an experiment (arms). Examples abound, such as a web developer wishing to compare the number of newly subscribed users among different versions of a signup page or a DevOps engineer wishing to compare the number of errors logged by different software versions Kuo and Yang [1996]. In these applications and others, controlled experiments can test statistical hypotheses based solely on the counts observed within each arm.

However, performing experiments incurs a cost, which is often linear in the sample size. The cost could stem from orchestrating an expensive experiment or several opportunity costs. There is a reward-associated opportunity cost in assigning an experimental unit to a particular arm when the observable outcome may be better under another. If a unit cannot be assigned to multiple experiments simultaneously, then there is also a learning-associated opportunity cost in using a unit to test a particular hypothesis when the scientific conclusions may be stronger testing another. A cost-minimizing goal of the statistical methodology is then to test hypotheses with the smallest sample size possible while maintaining the same statistical guarantees. Faster conclusion of experiments improves the overall agility of the experimenter, increasing the number of hypotheses tested and the rate of learning. We have included a critique of fixed-n testing, the advantages of sequential testing and corresponding literature review in the appendix.

2 A Sequential Multinomial Test

2.1 Theory

Our development of sequential tests for different kinds of count data begins with a sequential test for multinomial observations. The construction follows the following sequence of steps common in the literature Shafer et al. [2011], Waudby-Smith and Ramdas [2020], Howard et al. [2021]. Step 1:

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Define a relevant Bayes factor. Step 2: Show that the Bayes factor is a nonnegative supermartingale under the null hypothesis. Step 3: Use martingale inequalities to construct a test martingale that controls the frequentist Type I probability below a desired level u. Step 4: Invert the sequential test based on the test martingale to obtain a confidence sequence with a coverage guarantee of at least $1-u$. Extensions to other kinds of count data, including Bernoulli, Binomial and Poisson counting processes, are then obtained by recognizing relationships that exist to the multinomial distribution.

We note that this construction can be identified as an mSPRT-type procedure using mixture martingales. The Bayes factor can be interpreted as the statistic used in the mixture sequential probability ratio test, which is obtained by integrating the likelihood of the alternative with respect to a chosen weighting function. In this case, the likelihood is a multinomial likelihood, and the weight function is the Dirichlet prior distribution that will be used. The purpose of this paper is not to study the theoretical properties of general mSPRT-type procedures or sequential tests developed via mixture martingales (this have been studied previously in Wald and Wolfowitz [1948] and Kaufmann and Koolen [2021]), rather, to develop appropriate mSPRT-type tests for a number of challenging applications in modern online experimentation.

For step 1, consider a sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots$ of independent $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variables with $\boldsymbol{\theta} \in \triangle^d$, the $d-1$ simplex. We use bold typeset to denote vectors. Under the null hypothesis, M_0, it is assumed that

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots | M_0 \overset{\text{i.i.d.}}{\sim} \mathrm{Multinomial}(1, \boldsymbol{\theta}_0). \tag{1}$$

To construct a model over alternatives, M_1, we place a conjugate Dirichlet prior over alternative values of $\boldsymbol{\theta}$

$$\begin{aligned} \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots | \boldsymbol{\theta}, M_1 &\overset{\text{i.i.d.}}{\sim} \mathrm{Multinomial}(1, \boldsymbol{\theta}), \\ \boldsymbol{\theta} | M_1 &\sim \text{Dirichlet}(\boldsymbol{\alpha}_0). \end{aligned} \tag{2}$$

The following expressions are simplest with a uniform prior over the simplex achieved by setting $\alpha_{0,i} = 1$. In many applications, however, we expect departures from the null to be small and encode this information into the Dirichlet prior by concentrating it about $\boldsymbol{\theta}_0$ with the choice $\alpha_{0,i} = k\theta_{0,i}$ for a concentration parameter $k \in \mathbb{R}^+$. Let $S_i^n = \sum_{j=1}^n x_{j,i}$ and $\boldsymbol{S}_n = (S_1^n, \ldots, S_d^n) \in \mathbb{R}^d$. In addition, let $|\boldsymbol{v}| = \sum_i v_i$ denote the element-wise sum of a vector $\boldsymbol{v}$, $\boldsymbol{v}^{\boldsymbol{w}} = \prod_i v_i^{w_i}$ to denote element-wise exponentiation of two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, and Beta to denote the multivariate Beta function $\mathrm{Beta}(\boldsymbol{v}) := (\prod_i \Gamma(v_i))/\Gamma(\sum_i v_i)$. The resulting Bayes factor comparing models M_1 to M_0 is given by

$$BF_{10}(\boldsymbol{x}_{1:n}) = \frac{\mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n)}{\mathrm{Beta}(\boldsymbol{\alpha}_0)} \frac{1}{\boldsymbol{\theta}_0^{\boldsymbol{S}_n}}. \tag{3}$$

which appears as early as Good [1967] (derivation in Appendix A.4). In a Bayesian analysis, the Bayes factor multiplied by the prior odds gives the posterior odds of M_1 over M_0. In this work, we take the prior odds to be unity so that the terms Bayes factor and posterior odds can be used interchangeably.

In sequential applications, it often makes sense to compute Equation (3) recursively. Let $O_n(\boldsymbol{\theta}_0)$ denote the posterior odds at n, then

$$O_n(\boldsymbol{\theta}_0) = \frac{\mathrm{Beta}(\boldsymbol{\alpha}_{n-1} + \boldsymbol{x}_n)}{\mathrm{Beta}(\boldsymbol{\alpha}_{n-1})} \frac{1}{\boldsymbol{\theta}_0^{\boldsymbol{x}_n}} O_{n-1}(\boldsymbol{\theta}_0), \tag{4}$$

where $\boldsymbol{\alpha}_n = \boldsymbol{\alpha}_{n-1} + \boldsymbol{x}_n$ and $O_0(\boldsymbol{\theta}_0) = p(M_1)/p(M_0) = 1$. Details are provided in Appendix A.5. The dependence of $O_n(\boldsymbol{\theta}_0)$ on the observed data $\boldsymbol{x}_{1:n}$ is implicit in this notation, yet the null value $\boldsymbol{\theta}_0$ being tested is made explicit to aid the discussion of confidence sequences in Theorem 2.4.

Step 2 in our construction is to demonstrate that this is a nonnegative supermartingale under the null hypothesis M_0.

Theorem 2.1. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a sequence of independent $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variables, $\mathcal{F}_{n-1} = \sigma(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n-1})$ and consider the sequence of posterior odds $O_n(\boldsymbol{\theta}_0)$ defined in Equation (4) with $O_0(\boldsymbol{\theta}_0) = 1$, then*

$$\mathbb{E}_{M_0}[O_n(\boldsymbol{\theta}_0) | \mathcal{F}_{t-1}] = O_{n-1}(\boldsymbol{\theta}_0) \tag{5}$$

The proof is found in Appendix A.6. Theorem 2.1 states that $O_n(\boldsymbol{\theta}_0)$ is a nonnegative martingale under the null hypothesis with respect to the canonical filtration.

Step 3 is to use the posterior odds to construct a test martingale.

Theorem 2.2. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a sequence of independent* $\text{Multinomial}(1, \boldsymbol{\theta})$ *random variables and consider the sequence of posterior odds $O_n(\boldsymbol{\theta}_0)$ defined in Equation* (4) *with $O_0(\boldsymbol{\theta}_0) = 1$. Then*

$$\mathbb{P}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left(\exists n \in \mathbb{N} : O_n(\boldsymbol{\theta}_0) \geq 1/u\right) \leq u \tag{6}$$

for all $u \in [0, 1]$.

The proof is provided in Appendix A.7. The time-uniform bound presented in Theorem 2.2 controls the deviations of a stochastic process for all t simultaneously and is essential for proving the correctness of sequential tests and verifying the optional stopping and optional continuation properties. It provides a valid stopping rule: reject the null at time $\tau = \inf\{n \in \mathbb{N} : O_n(\boldsymbol{\theta}_0) \geq 1/u\}$. Simply stated, a practitioner who rejects the null hypothesis as soon as the posterior odds become larger than $1/u$ incurs a frequentist Type I error probability of at most u. Shafer et al. [2011], Johari et al. [2021] bring this idea back to more familiar territory by constructing a *sequential p-value* by tracking the running supremum of the posterior odds and taking its inverse, or equivalently

$$\begin{aligned} p_0 &= 1 \text{ and} \\ p_n &= \min(p_{n-1}, 1/O_n(\boldsymbol{\theta}_0)). \end{aligned}$$

It follows from this definition and equation (6) that

$$\mathbb{P}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left(\exists n \in \mathbb{N} : p_n \leq u\right) \leq u, \tag{7}$$

which is an easily digestible generalization of a fixed-n p-value to sequential settings. Instead of holding only at some pre-specified $n \in \mathbb{N}$, this guarantee holds *for all* $n \in \mathbb{N}$. This construction is shown in Figure 1. A simulation empirically demonstrating the control of false positives under continuous monitoring relative to a χ^2 test is shown in Appendix A.9.

Before completing Step 4, it is useful to show that this sequential test is not trivial. For this test to have utility it must possess the ability to control not only Type I errors, as in theorem 2.2, but also Type II errors. This is provided by the following theorem

Theorem 2.3. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a sequence of independent* $\text{Multinomial}(1, \boldsymbol{\theta})$ *random variables and consider the sequence of posterior odds $O_n(\boldsymbol{\theta}_0)$ defined in Equation* (4) *with $O_0(\boldsymbol{\theta}_0) = 1$. If $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then*

$$\frac{1}{n} \log O_n(\boldsymbol{\theta}_0) \to D_{KL}(\boldsymbol{\theta} || \boldsymbol{\theta}_0) \qquad a.s. \tag{8}$$

where $D_{KL}(\boldsymbol{\theta} || \boldsymbol{\theta}_0)$ is the Kullback Leibler divergence of the true Multinomial distribution with true parameter $\boldsymbol{\theta}$ from the Multinomial distribution under the null hypothesis with null parameter $\boldsymbol{\theta}_0$.

The proof is given in Appendix A.8. Theorem 2.3 states that if the null hypothesis is not true, with $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then the Bayes factor will diverge to infinity and exceed the $1/u$ threshold in theorem 2.2 (a.s.), or equivalently that the sequential p-value converges to zero and falls below the u threshold (a.s.). In other words, this test is guaranteed to reject the null almost surely if the null is incorrect, which is considered to be *asymptotically power 1* by Robbins [1970]. This result follows simply from the posterior consistency of Bayes factors. We now state step 4 of the construction.

Theorem 2.4. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a sequence of independent* $\text{Multinomial}(1, \boldsymbol{\theta})$ *random variables and consider the sequence of posterior odds $O_n(\boldsymbol{\theta}_0)$ defined in Equation* (4) *with $O_0(\boldsymbol{\theta}_0) = 1$. Let $C_n(u) = \{\boldsymbol{\theta} \in \triangle^d : O_n(\boldsymbol{\theta}) < 1/u\}$ denote the set of parameter vectors that would not be rejected by the test at the u level, then*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta} \in C_n(u) \text{ for all } n \in \mathbb{N}\right) \geq 1 - u \tag{9}$$

for all $u \in [0, 1]$.

A simple corollary of theorem 2.4 is that $\mathbb{P}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta} \in \bigcap_{n=1}^{\infty} C_n(u)\right) \geq 1 - u$. This result provides a confidence statement for sequentially estimating the true parameter vector $\boldsymbol{\theta}$ as the experiment progresses. The confidence set $C_n(u)$ for $\boldsymbol{\theta}$ is a convex subset of $\triangle^d$, with convexity following from the concavity of the multinomial log-likelihood. Confidence intervals on the individual elements of $\boldsymbol{\theta}$ can be obtained by projecting $C_n(u)$ onto the coordinate axes in the following manner.

Corollary 2.5. *For $C_n(u)$ as in Theorem 2.4, let*

$$j^+_{n,i}(u) = \sup\{\theta_i : \boldsymbol{\theta} \in C_n(u)\},$$
$$j^-_{n,i}(u) = \inf\{\theta_i : \boldsymbol{\theta} \in C_n(u)\},$$

then

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\forall i : \theta_i \in \bigcap_{n=1}^{\infty} [j^-_{n,i}(u), j^+_{n,i}(u)]\right) \geq 1 - u. \tag{10}$$

$j^+_{n,i}(u)$ can be computed by solving the following convex optimization program

$$\begin{aligned} \max \quad & \theta_i \\ \text{s.t.} \quad & c + \log u \leq \sum_i S_i^n \log \theta_i \\ & \sum_i \theta_i = 1 \end{aligned} \tag{11}$$

where $c = \log \mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n) - \log \mathrm{Beta}(\boldsymbol{\alpha}_0)$. The constraints in the optimization program simply define $C_n(u)$. Similarly, $j^-_{n,i}(u)$ is obtained by minimizing θ_i over this set.

2.2 Application: Sample Ratio Mismatch Tests

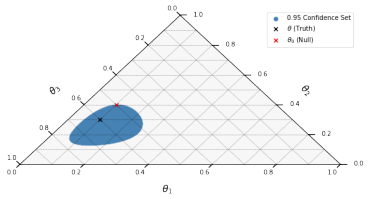
Comparing the counts of experimental units assigned to each arm of a multi-arm experiment can often detect bias and errors in the experiment. Most online controlled experiments follow simply randomized designs, whereby a new unit is randomly assigned to one of d arms according to a vector of pre-specified probabilities $\boldsymbol{\theta}$. The assignment outcome for a new unit is independent of other units (individualistic), unit-level covariates, potential outcomes (unconfounded) and can therefore be summarized as an independent $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variable. Under these assumptions, the assignment mechanism can be considered ignorable when performing inference on causal estimands such as the average treatment effect [Imbens and Rubin, 2015].

Although simple in theory, the systems that perform assignments quickly grow in complexity as the number of concurrent experiments increases [Tang et al., 2010]. This increased complexity increases the risk of introducing bugs that cause departures from the intended assignment mechanism, breaking the assumption of ignorability and rendering causal estimates invalid. Zhao et al. [2016] provides an account of an incorrect hashing algorithm introducing bias into the assignment mechanism.

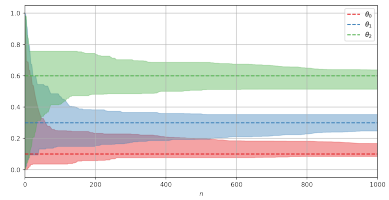
After assignment and measurement, data passes through processing and cleansing pipelines before analysis. If incorrect cleansing logic is applied, there is a risk that specific observations may be selectively removed, introducing a “missing not at random" missing data mechanism, rendering causal estimates invalid Rubin [1976]. Fabijan et al. [2019] describes an experiment in which units from the treatment arm were unintentionally removed, with the probability of their removal depending on their observed outcomes.

An arm with a surprisingly low or high number of units is usually symptomatic of an implementation error in the experiment and is colloquially referred to as a *sample ratio mismatch* (SRM) Fabijan et al. [2019]. These errors can be caught by comparing the counts of experimental units in each arm against the intended assignment probabilities. It is now considered a good practice to validate the experiment setup by performing a χ^2 test, comparing the observed counts against the expected counts under the intended assignment mechanism Chen et al. [2018]. However, the χ^2 test is an example of a *fixed-n* test, providing statistical guarantees when performed *once*. Due to this limitation, it is typically performed after data collection and prior to analysis. To reveal a bug that renders an expensive experiment invalid, only after the experiment is finished, would be less than ideal. Ideally SRMs are detected as early as possible so that the implementation error can be corrected before more units enter the experiment. This necessitates sequential testing. We provide a sequential multinomial test for testing a point null based on the counts of each outcome.

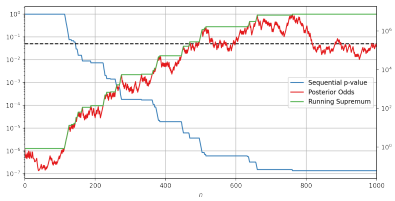
Figure 1 shows the application of the sequential multinomial test to a representative experiment exhibiting an SRM. In this example the intended assignment probabilities are $\boldsymbol{\theta}_0 = (0.1, 0.4, 0.5)$. These propensity scores are assumed to be known by the analyst and will subsequently be used



(a) Joint Confidence Set at $n = 144$



(b) Simultaneous Marginal Confidence Sequences



(c) Sequential p-value and Bayes Factor

Figure 1: (a) The confidence set $C_{144}(0.05)$ at $n = 144$ as defined in Theorem 2.4. The true $\boldsymbol{\theta} = (0.1, 0.3, 0.6)$ is marked with a black cross, and the null hypothesis $\boldsymbol{\theta}_0 = (0.1, 0.4, 0.5)$ is denoted with a red cross. (b) Simultaneous 0.95 confidence sequences that cover the individual elements of $\boldsymbol{\theta} = (0.1, 0.3, 0.6)$ obtained from Corollary 2.5 and computed via the optimization program in Equation (11). (c) (Left axis) Sequential p-value (blue) defined in Equation (7). Critical value $u = 0.05$ (dashed-black). $\boldsymbol{\theta} = (0.1, 0.3, 0.6)$, $\boldsymbol{\theta}_0 = (0.1, 0.4, 0.5)$. $p_n < 0.05$ for all $n \geq 144$. (Right axis) The posterior odds defined in Equation (3) (red), with the running supremum (green).

to perform inference about average treatment effects. Due to a bug, either in randomization or data processing, some units are being preferentially assigned to certain arms based on their covariates, or data is being lost not at random, resulting in observations arriving with probabilities $\boldsymbol{\theta} = (0.1, 0.3, 0.6)$. The sequential p-value is less than 0.05 for all $n \geq 144$. This is the smallest n for which $\boldsymbol{\theta}_0 \notin C_n(0.05)$.

3 Inhomogeneous Bernoulli Processes

Wald [1947] described an experiment to test the null hypothesis that two guns are equally precise. Two guns are fired at a target simultaneously for many attempts, and their success is recorded. Each attempt is modelled as a Bernoulli trial and, under the null hypothesis, the probability of each gun hitting the target on the same attempt is equal. However, the success probability varies across attempts due to gusty wind conditions. As the guns are shot simultaneously, the wind condition affects each gun equally. If the null is rejected, choosing the gun that obtained the greatest number of successes seems reasonable.

3.1 Theory

Suppose a new experimental unit is randomly assigned to one of d experiment arms at time t, according to assignment probabilities $\boldsymbol{\rho} \in \triangle^d$, and a Bernoulli outcome is observed. The Bernoulli probability for arm i at time t is parameterized by $p_i(t) = e^{\mu(t)} e^{\delta_i}$ so that the time-varying effect is multiplicative and common to all arms. The improvement of arm j over arm i at any time is then $p_j(t)/p_i(t) = \exp(\delta_j - \delta_i)$, and the difference on the log-scale is simply $\delta_j - \delta_i$. Suppose Bernoulli failures are ignored and arms are compared only through their counts of Bernoulli successes. The (conditional) probability that the next Bernoulli success comes from arm i is

$$\theta_i = \frac{\rho_i e^{\delta_i}}{\sum_{j=1}^{d} \rho_j e^{\delta_j}}, \tag{12}$$

which is independent of the time-varying effect. The arm from which the next Bernoulli success arrives is, therefore, a $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variable, and the counts of Bernoulli successes for each arm are an ancillary statistic with respect to the time-varying nuisance parameter $\mu(t)$. Framing the problem this way allows the sequential multinomial test to perform inference on $\boldsymbol{\delta}$. Simple hypotheses about $\boldsymbol{\delta}$ can therefore be translated into testing simple hypotheses about $\boldsymbol{\theta}$. Equality among success probabilities can be tested by simply testing the null multinomial hypothesis $\boldsymbol{\theta}_0 = \boldsymbol{\rho}$. The individual components δ_i are not identifiable, as adding a constant to each element results in the same $\boldsymbol{\theta}$, yet contrasts of the form $\sum_i a_i \delta_i$ for $\sum_i a_i = 0$ are identifiable.

Let $\sigma_{\boldsymbol{\rho}} : \mathbb{R}^d \to \triangle^d$ denote a generalization of the softmax function to include $\boldsymbol{\rho}$, with $\sigma_{\boldsymbol{\rho}}(\boldsymbol{\delta})_i$ equal to the right hand side of Equation (12).

Theorem 3.1. *Let $K_n(u) = \sigma_{\boldsymbol{\rho}}^{-1}(C_n(u))$, then*

$$\mathbb{P}[\boldsymbol{\delta} \in K_n(u) \textit{ for all } n \in \mathbb{N}] \geq 1 - u \tag{13}$$

The proof is a direct consequence of Theorem 2.4. The following corollary yields simultaneous confidence sequences for all contrasts.

Corollary 3.2. *Let $K_n(u) = \sigma_{\boldsymbol{\rho}}^{-1}(C_n(u))$ and $\mathcal{A}^d = \{\boldsymbol{a} \in \mathbb{R}^d : \sum_i a_i = 0\}$ denote the set of all d-dimensional contrasts. For all $\boldsymbol{a} \in \mathcal{A}^d$ define*

$$l^+_{n,\boldsymbol{a}}(u) = \sup\{\sum_i a_i \delta_i : \boldsymbol{\delta} \in K_n(u)\},$$

$$l^-_{n,\boldsymbol{a}}(u) = \inf\{\sum_i a_i \delta_i : \boldsymbol{\delta} \in K_n(u)\},$$

then

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\forall \boldsymbol{a} \in \mathcal{A}^d : \sum_i a_i \delta_i \in \bigcap_{n=1}^{\infty} [l^-_{n,a}(u), l^+_{n,a}(u)]\right) \geq 1 - u.$$

The upper bound $l^+_{n,a}(u)$ is the solution to the following convex optimization

$$\begin{aligned} \max \quad & \sum_i a_i \delta_i \\ \text{s.t.} \quad & c \leq \sum_i S_i^n \left(\delta_i + \log \rho_i - \log \sum_j \rho_j e^{\delta_j}\right) \end{aligned} \tag{14}$$

where $c = \log \mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n) - \log \mathrm{Beta}(\boldsymbol{\alpha}_0) + \log u$. Convexity follows from the log-sum-exponential function. The lower bound $l^-_{n,a}(u)$ is the solution to the corresponding minimization problem. This is visualized in Figure 2.

A hypothesis can be rejected at the u level as soon as the set that it defines fails to intersect with the confidence set $K_n(u)$. Note that $K_n(u_1) \subset K_n(u_2)$ for $u_1 > u_2$. To obtain a sequential p-value, we seek the largest u such that the null is not rejected. That is, we seek the smallest set $K_n(u)$ such that there is a non-empty intersection with the subset of $\mathbb{R}^d$ defined by the hypothesis. This too can be achieved by a convex optimization program. One can simply maximize u over the feasible set defined by the intersection of the $K_n(u)$ and the set defined by the hypothesis. Suppose one wishes to test the hypothesis $\delta_0 \geq \delta_1$ and $\delta_0 \geq \delta_2$. The sequential p-value at time n is the inverse of the solution to the following convex program

$$\begin{aligned} \max \quad & q \\ \text{s.t.} \quad & c \leq \log(q) + \sum_i S_i^n \left(\delta_i + \log \rho_i - \log \sum_j \rho_j e^{\delta_j}\right) \\ & \delta_0 \geq \delta_1 \\ & \delta_0 \geq \delta_2 \end{aligned} \tag{15}$$

where $c = \log \mathrm{Beta}(\boldsymbol{\alpha}_0 + \boldsymbol{S}_n) - \log \mathrm{Beta}(\boldsymbol{\alpha}_0)$. A simulation study presenting sequential p-values and confidence sequences on $\boldsymbol{\delta}$ is given in Appendix A.11

(a) (b)

Figure 2: (a) Assuming p_1 and p_0 are stationary and estimating them using multinomial confidence sequences from from Corollary 2.5. (b) Assuming $p_1(t)$ and $p_0(t)$ are dynamic but with $p_1(t)/p_0(t) = \exp(\delta_1 - \delta_0)$ and estimating this quotient using the confidence sequences from equation Corollary 3.2. (Left axis) Confidence sequences are visualized with shaded regions and MLE estimates are visualized with solid lines. (Right axis) sequential p-value.

3.2 Application: Conversion Rate Optimization

In the online experimentation a *conversion* is used to describe a Bernoulli trial such as user signup or purchase. Many modern experiments are designed to increase conversion rate and drive additional revenue. The typical setup is that there are d different *experiences* (such as different versions of a signup page), and each experience i is assumed to have some conversion probability p_i. When a new user visits the experience, they convert or not, and a successful or unsuccessful conversion event is logged. One might think that it would be easy to estimate p_i sequentially using the multinomial confidence sequences of section 2 and the counts of successful and unsuccessful conversions of arm i. However, one soon runs into problems as illustrated below...

The following case study is taken from a signup funnel experiment at a leading internet streaming company, from whom we have obtained consent to use in this paper. It was hypothesized that adding additional methods of payments might increase overall numbers of conversions. Alternatively, it is also be plausible that additional payment methods may simply cannibalize conversions from other payment methods, not increasing overall number of conversions at all. An A/B test was created to address these hypotheses, one arm with the currently offered payment methods, the other with additional payment methods. Figure 2 shows the application of the multinomial confidence sequences to estimating the conversion probabilities for arms 0 and 1. Note that the running intersection of always-valid confidence intervals becomes the empty set, and the MLE exits the confidence sequence. This would a rare event (with probability less than α) if the assumptions of section 2 were true. This is usually a sign that the conversion probabilities are not constant, but time-varying, invalidating a commonly made assumption in conversion rate experimentation.

Many online conversion experiments share similarities with Wald's gun-shooting example. The success probability of a conversion is often time-varying due to external factors such as day of the week, recent product launches, or new promotions. These external factors are likely to affect all arms of the experiment equally, so it is reasonable to expect the time variation to be common to all. Moreover, it is often the case in these experiments that an unsuccessful signup (a 0 Bernoulli outcome) is not directly observed. When a successful signup occurs, and event is typically logged, but if no signup occurs then an unsuccessful signup must be inferred by the absence of a successful signup (typically through an ad-hoc definition such as "an experimental unit failed to convert within T units of time since allocation"). This definition of conversion can make the analysis sensitive to the arbitrary choice of time window, and can present challenges to a sequential analysis if one must wait T units of time before the first Bernoulli outcome is realized.

Instead, we propose using the assumptions of section 3, which only requires the counts of successful signups. This relaxes the assumption of constant conversion probabilities to a constant multiplicative difference between them, which is typically more realistic in practice. The confidence sequence on the multiplicative constant for the payment methods experiment is shown in 2 (b). A winning arm can be declared in approximately 1 week instead of 9, to which all subsequent visitors can be directed, dramatically increasing signups relative to the fixed-n experiment.

4 Inhomogeneous Poisson Counting Processes

4.1 Theory

A counting process is a stochastic process $\{N(t) : t \geq 0\}$ satisfying $N(0) = 0$, $N(t) \in \mathbb{N}_0$ and $N(s) \leq N(t)$ for $s \leq t$. The inhomogeneous Poisson counting process is defined by an *intensity function* $\lambda : \mathbb{R} \to \mathbb{R}^+$ that is locally integrable, $\int_B \lambda(t)dt \leq \infty$ for all bounded Borel measurable sets $B \in \mathbb{R}$, defining a measure $\Lambda(B) = \int_B \lambda(t)dt$ Kingman [1992]. For any collection of disjoint Borel measurable sets $B_1, B_2, \ldots$ the inhomogeneous Poisson counting process has the property that $N(B_i)$ are independent $\mathrm{Poisson}(\Lambda(B_i))$ random variables. The inhomogeneous Poisson counting process can be defined in terms of an inhomogeneous Poisson point process on the real line by simply counting the number of points in a set. For our applications these points correspond to times of events. At any time t the probability density of the time-difference s to the next point is given by $g(s) = \lambda(t+s)\exp(-\int_0^s \lambda(t+s)ds)$. The independent increments property implies a *memoryless* property of the process, that the counts in the next time interval or waiting time until the next point are independent of the process history. This is a necessary property to establish the following theorem.

Theorem 4.1. *Consider d inhomogeneous Poisson point processes with intensity functions $\lambda_i(t) = \rho_i e^{\delta_i}\lambda(t)$ for $i \in \{1, 2, \ldots, d\}$. Let each point produced by process i be marked with the corresponding process index i. At any time t, such as immediately after the previous point, the probability that the next point has mark i is given by*

$$\theta_i = \frac{\rho_i e^{\delta_i}}{\sum_{j=1}^d \rho_j e^{\delta_j}}. \tag{16}$$

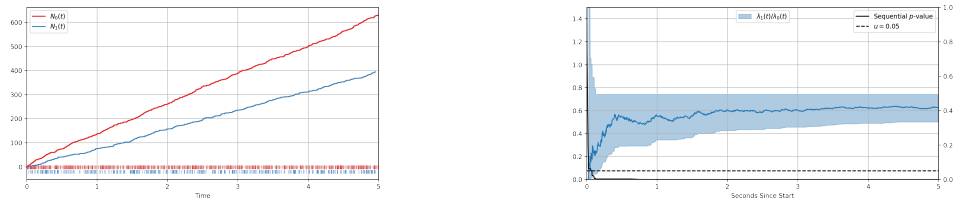
This gives the probability that the next point in time is from process i. The proof is given in the Appendix A.10. Theorem 4.1 states that the sequence of marks can be considered a sequence of $\mathrm{Multinomial}(1, \boldsymbol{\theta})$ random variables, allowing the sequential multinomial test to perform inference on $\boldsymbol{\delta}$. For example, a sequential test of equality among d time-inhomogeneous Poisson point processes ($\lambda_i(t) = \lambda_j(t)$ for all pairs i and j) can be obtained from the sequential multinomial test of the hypothesis $\boldsymbol{\theta}_0 = (1/d, \ldots, 1/d)$. Once again, the total counts in each arm is a statistic that is ancillary to the time-varying nuisance parameter $\lambda(t)$. A simulation study presenting sequential p-values and confidence sequences on $\boldsymbol{\delta}$ is given in Appendix A.12

4.2 Application: Software Canary Testing

When continuously deploying new software to users, DevOps engineers often adopt the practice of *canary testing* Schermann et al. [2018]. A canary test is a controlled experiment in which users are randomly assigned to the current software or a newer release candidate. The experimenter's goal is to study the performance of the release candidate in a production environment before releasing it globally, essentially acting as a quality control gate before full deployment. If the release candidate has significantly worse performance, it is blocked, and developers must resolve the offending issues. This strategy helps to prevent bugs from reaching all users. However, performance regressions are still experienced by those in the experiment. To minimize harm to these users, performance regressions should be detected in real-time, and the canary terminated as soon as possible. This necessitates sequential testing methods.

Performance regressions are measured in terms of the rates of *events*. These events could correspond to errors, failures, or any occurrence of interest that are sent to a central logging service by each instance of the software. The data naturally forms a marked 1-dimensional point process in time, recording the timestamp and type of event. The instantaneous rate is expected to be time-varying due to varying traffic and usage patterns.

The following example is taken from a leading internet streaming company, from whom we have obtained consent to user in this paper, in which a bug that affected approximately 60% of all devices was detected with this methodology in less than 1 second. In this canary experiment, *successful play starts* (SPS) are carefully monitored between the existing software and the release candidate. Whenever a title is requested by the user in the streaming application and the title successfully begins playback, then an SPS event is sent by the application to the central logging system. On the receiving end, SPS events are being received from treatment and control arms of the experiment. If significantly fewer SPS events are being received from the treatment group running the release candidate, then it



(a) (b)

Figure 3: (a) Ruglplot shows the timestamps of SPS events being received, while solid lines show the counting processes for arms 0 and 1. (b) (Left axis) Shaded region shows the confidence sequence for the quotient of inhomogeneous Poisson process intensity functions $\lambda_1(t)/\lambda_0(t)$ obtained from Corollary 3.2. Blue solid line shows the MLE of $\lambda_1(t)/\lambda_0(t) = e^{\delta_1 - \delta_0}$. (Right axis) Black solid line shows sequential p-value.

indicates there is an issue with the new software version that is preventing some streams from starting. Figure 3 shows that the confidence sequence on $\lambda_1(t)/\lambda_0(t)$ falls below 1.0 in less than a second, indicating that the instantaneous rate of SPS events for arm 1 is less than the instantaneous rate of SPS events for arm 0. In this case, the canary experiment was aborted and the offending bug was identified, preventing a serious bug from being released globally to all users.

5 Conclusion

The contributions of this paper provide an "always-valid" approach to inference in a number of important applications dealing with count data. This permits experiments to be continuously monitored and enables optional stopping, which can greatly reduce both the time required for experiments to complete and the potential harm to units in the experiment. We first introduced a sequential test for Multinomial hypotheses using a mixture martingale construction, which has already proven to be an effective solution for rapidly detecting sample ratio mismatches in online controlled experiments. We then used this result to develop a sequential test of equality and contrasts in inhomogeneous Bernoulli processes, which has practically demonstrated dramatic speedups in decision making in conversion experiments. Contrary to many widely used models, our approach does not assume that the conversion probabilities are constant, which is often violated in the real world. Moreover, only successful conversions, such as signups, are observed in practice. This is different from Bernoulli outcome models in which it is assumed that both successes and failures are observed. Our proposed sequential test enjoys the added convenience of only requiring successful Bernoulli outcomes to be observed. Lastly, we used the sequential Multinomial test to develop a sequential test for equality and contrasts in time-inhomogeneous Poisson counting processes. These play an important role in the monitoring of systems, such as data pipelines and software usage. This methodology has been productionized at an internet streaming company for the safe roll-out of new software. It has been able to detect bugs and performance regressions at a sub-second level, preventing these issues from being released to all users.

The confidence sequences provided in this paper allow inference to be made at any time and the ability to use data-dependent stopping rules for ending an experiment. This in itself can dramatically speed up the time to reach conclusions with valid statistical guarantees. An obvious extension to this work is to combine these confidence sequences with a strategy for adapting the assignment probabilities, assigning fewer units to suboptimal arms, and more to the optimal arm. The confidence sequences presented here can be used to construct an adaptive algorithm that identifies the best arm with tunably high probability through the least upper confidence bound (LUCB) algorithm, as has been done successfully in Howard and Ramdas [2022].

References

Lynn Kuo and Tae Young Yang. Bayesian computation for nonhomogeneous poisson processes in software reliability. *Journal of the American Statistical Association*, 91(434):763–773, 1996. ISSN

01621459. URL http://www.jstor.org/stable/2291671.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, bayes factors and p -values. *Statist. Sci.*, 26(1):84–101, 02 2011. doi: 10.1214/10-STS347. URL https://doi.org/10.1214/10-STS347.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20204–20214. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e96c7de8f6390b1e6c71556e4e0a4959-Paper.pdf.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055 – 1080, 2021. doi: 10.1214/20-AOS1991. URL https://doi.org/10.1214/20-AOS1991.

A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948. ISSN 00034851. URL http://www.jstor.org/stable/2235638.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021. URL http://jmlr.org/papers/v22/18-798.html.

I. J. Good. A bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 29(3):399–431, 1967. ISSN 00359246. URL http://www.jstor.org/stable/2984384.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 2021. doi: 10.1287/opre.2021.2135. URL https://doi.org/10.1287/opre.2021.2135.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970. ISSN 00034851. URL http://www.jstor.org/stable/2239848.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings 16th Conference on Knowledge Discovery and Data Mining*, pages 17–26, Washington, DC, 2010.

Zhenyu Zhao, Miao Chen, Don Matheson, and Maria Stone. Online experimentation diagnosis and troubleshooting beyond aa validation. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 498–507, 2016. doi: 10.1109/DSAA.2016.61.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Aleksander Fabijan, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer, and Pavel Dmitriev. Diagnosing sample ratio mismatch in online controlled experiments: A taxonomy and rules of thumb for practitioners. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2156–2164, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330722. URL https://doi.org/10.1145/3292500.3330722.

Nanyu Chen, Min Liu, and Ya Xu. Automatic detection and diagnosis of biased online experiments, 2018.

A. Wald. *Sequential analysis*. J. Wiley & sons, Incorporated, 1947. URL https://books.google.com/books?id=0nREAAAAIAAJ.

J.F.C. Kingman. *Poisson Processes*. Oxford Studies in Probability. Clarendon Press, 1992. ISBN 9780191591242. URL https://books.google.com/books?id=VEiM-OtwDHkC.

Gerald Schermann, Jürgen Cito, Philipp Leitner, Uwe Zdun, and Harald C. Gall. We're doing it live: A multi-method empirical study on continuous experimentation. *Information and Software Technology*, 99:41–57, 2018. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2018.02.010. URL https://www.sciencedirect.com/science/article/pii/S0950584917302136.

Steven R. Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704 – 1728, 2022. doi: 10.3150/21-BEJ1388. URL https://doi.org/10.3150/21-BEJ1388.

P. Armitage. *Interim Analyses in Clinical Trials*. Multiple Comparisons, Selection and Applications in Biometry. CRC Press, 1993. ISBN 9780824788957.

Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1933.0009.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952. doi: bams/1183517370. URL https://doi.org/.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. ISBN 9780534243128. URL https://books.google.com/books?id=0x_vAAAAMAAJ.

F. J. Anscombe. Fixed-sample-size analysis of sequential observations. *Biometrics*, 10(1):89–100, 1954. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3001665.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1517–1525, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3097992. URL https://doi.org/10.1145/3097983.3097992.

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244, 1969. ISSN 00359238. URL http://www.jstor.org/stable/2343787.

Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld. Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, 91(435):1228–1235, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291741.

Jean Ville. *Étude critique de la notion de collectif*. 1939. URL http://eudml.org/doc/192893.

A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 06 1945. doi: 10.1214/aoms/1177731118. URL https://doi.org/10.1214/aoms/1177731118.

Harold Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222, 1935. doi: 10.1017/S030500410001330X.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291091.

Allard Hendriksen, Rianne de Heide, and Peter Grünwald. Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations. *Bayesian Analysis*, 16(3):961 – 989, 2021. doi: 10.1214/20-BA1234. URL https://doi.org/10.1214/20-BA1234.

James O. Berger, Lawrence D. Brown, and Robert L. Wolpert. A unified conditional frequentist and bayesian test for fixed and sequential simple hypothesis testing. *Ann. Statist.*, 22(4):1787–1807, 12 1994. doi: 10.1214/aos/1176325757. URL https://doi.org/10.1214/aos/1176325757.

J. O. Berger, B. Boukai, and Y. Wang. Unified frequentist and bayesian testing of a precise hypothesis. *Statist. Sci.*, 12(3):133–160, 09 1997. doi: 10.1214/ss/1030037904. URL https://doi.org/10.1214/ss/1030037904.

James O. Berger, Benzion Boukai, and Yinping Wang. Simultaneous bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*, 86(1):79–92, 1999. ISSN 00063444. URL http://www.jstor.org/stable/2673538.

Jerome Cornfield. A bayesian test of some classical hypotheses, with applications to sequential clinical trials. *Journal of the American Statistical Association*, 61(315):577–594, 1966. ISSN 01621459. URL http://www.jstor.org/stable/2282772.

D. A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967. ISSN 0027-8424. doi: 10.1073/pnas.58.1.66. URL https://www.pnas.org/content/58/1/66.

Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H. Kennedy, and Aaditya Ramdas. Doubly robust confidence sequences for sequential causal inference, 2021.

C. Jennison and B.W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, 1999. ISBN 9781584888581. URL https://books.google.com/books?id=qBrpTcAYtNQC.

Ivair R Silva and Martin Kulldorff. Continuous versus group sequential analysis for post-market drug and vaccine safety surveillance. *Biometrics*, 71(3):851–858, 2015.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17(none):257 – 317, 2020. doi: 10.1214/18-PS321. URL https://doi.org/10.1214/18-PS321.

P. A. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes with degree-two exponential polynomial rate function. *Operations Research*, 27(5):1026–1040, 1979. doi: 10.1287/opre.27.5.1026. URL https://doi.org/10.1287/opre.27.5.1026.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our work is motivated by three important sequential applications in online experimentation. Testing proper randomization via sample ratio mismatch testing. Testing hypotheses among time-varying probabilities in conversion rate optimization and testing hypotheses among time-varying poisson process intensities in software canary experiments. We have provided viable sequential testing solutions to all of three of these applications via sequential multinomial tests, sequential tests of inhomogeneous Bernoulli processes, and sequential tests of time-inhomogeneous Poisson counting processes.
   (b) Did you describe the limitations of your work? [Yes] We have explained in the introductory section that we do not study the theoretical properties such as optimality our proposed tests. Our tests are specific instances of a generic construction known as the method of mixture martingales and mixture sequential probability ratio tests. The optimality properties of these tests are well established and we have linked to the relevant literature. Our work focuses on providing specific solutions to some important applications using these constructions.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] All proofs include the assumptions, and we have even presented examples of what can go wrong when these assumptions do not hold - see the application of multinomial confidence sequences when the probabilities are time varying in section 3

(b) Did you include complete proofs of all theoretical results? [Yes] All proofs are in the appendix

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have a number of simulations in the appendix, which all include clear instructions on how they were created

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We include two real-world case studies of how this methodology is employed with practical value at a leading internet streaming company. A draft of this paper has been reviewed internally by said company and cleared for submission

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]