Hypothesis Testing in Adaptively Sampled Data: ART to Maximize Power Beyond *iid* Sampling

Dae Woong Ham[1*†] and Jiaze Qiu[1*†]

[1*]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, 02138, MA, U.S.A .

*Corresponding author(s). E-mail(s): daewoongham@g.harvard.edu; jiazeqiu@g.harvard.edu;
†These authors contributed equally to this work.

Abstract

Testing whether a variable of interest affects the outcome is one of the most fundamental problem in statistics and is often the main scientific question of interest. To tackle this problem, the conditional randomization test (CRT) is widely used to test the independence of variable(s) of interest ($\boldsymbol{X}$) with an outcome ($\boldsymbol{Y}$) holding other variable(s) ($\boldsymbol{Z}$) fixed. The CRT uses "Model-X" inference framework that relies solely on the *iid* sampling of $(\boldsymbol{X}, \boldsymbol{Z})$ to produce exact finite-sample $\boldsymbol{p}$-values that are constructed using any test statistic. We propose a new method, the *adaptive randomization test* (ART), that tackles the same independence problem while allowing the data to be adaptively sampled. Like the CRT, the ART relies solely on knowing the (adaptive) sampling distribution of $(\boldsymbol{X}, \boldsymbol{Z})$. Although the ART allows practitioners to flexibly design and analyze adaptive experiments, the method itself does not guarantee a powerful adaptive sampling procedure. For this reason, we show substantial power gains obtained from adaptively sampling compared to the typical *iid* sampling procedure in a multi-arm bandit setting and an application in conjoint analysis. We believe that the proposed adaptive procedure is successful because it takes arms that may initially look like "fake" signals due to random chance and stabilizes them closer to "null" signals and samples more/less from signal/null arms.

Keywords: Conditional Independence Testing, Randomization Inference, Adaptive Sampling, Model-X, Non-parametric Testing

1 Introduction

Independence testing is ubiquitous in statistics and often the main task of interest in variable selection problems. For example, it is used in causal inference for testing the absence of any treatment effect for various applications [1–3]. More specifically, social scientists may wonder if a political candidate's gender may affect voting behavior while controlling for all other gender related stereotypes to isolate the true effect of gender [4–6]. Biologists may also be interested in the effect of a specific gene on a characteristic after holding all other genes constant [7].

In the independence testing problem, the main objective is to test whether a response Y is statistically affected by a variable of interest X while holding other variable(s) Z fixed. Informally speaking, we aim to test $Y \perp\!\!\!\perp X \mid Z$, where Z can be the empty set for an unconditional test. For the aforementioned gender example, Y is voting responses, X is the political candidate's gender, and Z are the candidate's personality, party affiliation, etc. One way to approach this problem is the *model-based* approach that uses parametric or semi-parametric methods such as regression while assuming some knowledge of $Y \mid (X, Z)$. Recently, the *design-based* approach has been increasingly gaining popularity [1, 2, 8] to tackle the independence testing problem. In an influential paper [3], the authors introduce the conditional randomization test (CRT), which uses a *design-based* or the "Model-X" approach to perform randomization based inference. This approach assumes nothing about the $Y \mid (X, Z)$ relationship but shifts the burden on requiring knowledge of the $X \mid Z$ distribution (hence named "Model-X"). In exchange, the CRT has exact type-1 error control while allowing the user to propose any test statistics, including those from complicated machine learning models, to increase power. We remark that if the data was collected from an experiment, then the distribution of the experimental variables (X, Z) is immediately available and the CRT can be classified as a non-parametric test.

The CRT, however, does require that (X, Z) is collected independently and identically (*iid*) from some distribution, which may not be always appropriate or desired. For example, large tech companies, such as Uber or Doordash, have rich experimental data that are sequentially and adaptively collected, i.e., the next treatment is sampled as a function of all of its previous history [9, 10]. Despite this non-*iid* experimental setup, the companies are interested in performing hypothesis tests on whether a certain treatment or features of their products affects the response in any way. Additionally, many practitioners may prefer an adaptive sampling procedure as it can be more effective to detect an effect since obtaining a large number of samples is often difficult and costly.

1.1 Our Contributions and Overview

Given this motivation, a natural direction is to weaken the *iid* assumption in the "Model-X" randomization inference approach and allow testing adaptively collected data. Therefore, the main contribution of our paper is we allow the same "Model-X" randomization inference procedure under adaptively collected data, i.e., we allow the data (X_t, Z_t) to be sequentially collected at time t as a function of the historical values of $X_{1:(t-1)}, Z_{1:(t-1)}, Y_{1:(t-1)}$, where $X_{1:(t-1)}$ denotes the vector of $(X_1, \ldots, X_{t-1})$ and $Z_{1:(t-1)}, Y_{1:(t-1)}$ is defined similarly. To the best of our knowledge, there does not exist a general randomization inference procedure that enjoys all the same benefits as that of the CRT while allowing for adaptively sampled data (see Section 1.2 for related works).

Our contribution is useful in both the experimental stage (the focus of this paper), i.e., allowing experimenters to construct powerful adaptive sampling procedures, and the analysis stage, i.e., after the data was adaptively collected as long as the analyst knows how the data was adaptively sampled. We name our method the ART (Adaptive Randomization Test) and we remark that the validity of the ART, like the CRT, does not require any knowledge of $Y \mid (X, Z)$ and leverages the distribution of (X, Z). Therefore, in an experimental setting, the ART can also be viewed as a non-parametric test.

In Section 2 we formally introduce the proposed method, ART, and prove how the ART leverages the known distribution of (X, Z) to produce exact finite-sample valid p-values for any test statistic. Although this formally allows practitioners to adaptively sample data to potentially increase power, it does not give any guidance on how to choose a reasonable adaptive procedure. Consequently, we first showcase the ART in the normal-means model setting (Section 3), a special case of the "multi-arm" bandit setting, through simulations and a theoretical asymptotic power analysis. Secondly, we also explore the ART's potential in a factorial survey setting in Section 4 through an application to a recent conjoint study concerning the role of gender discrimination in political candidate evaluation [4]. For both examples, we find that the ART can be uniformly more powerful than the CRT with a typical *iid* sampling scheme. We postulate that adaptive procedures leveraging evidence of signals can potentially enhance statistical power, provided that the degree of adaptivity is appropriately kept "in check". As shown in Sections 3-4 and in [11], excessive adaptation, i.e., sampling one or a few arms with very high probability, can lead to favorable regret-minimizing performance but unfavorable inferential properties. We generally contend that adaptive procedures that carefully balance the competing demands of exploration and exploitation are promising for improvements in statistical power and leave further empirical and theoretical validation to future works.

1.2 Related Works

In this section, we put our proposed method in the context of the current literature. The ART methodology is in the intersection of reinforcement learning and "Model-X" randomization inference procedures. As far as we know, our paper is the first to weaken the *iid* assumption and allow adaptive testing in the context of randomization inference when specifically tackling the independence testing problem. We remark that [12] considers *unconditional* randomization testing in sequentially adaptively sampled treatment assignments. However, this work does not cover the more general case of conditional randomization testing and assumes a causal inference framework under the finite-population view, i.e., conditioning on the potential outcomes [13]. Our work differs in that we allow for both super-population and finite-population view and additionally generalize to the *conditional* independence testing problem for general sequentially adaptive procedures (see Section 2.3 for more details). We also acknowledge that [14] (and references within) contain mentions of randomization inference in adaptive settings but serves primarily as a literature summary of randomization inference and provides no formal testing for general adaptive procedures.

As hinted above, many ideas from the reinforcement learning literature can also be useful starting points to construct a sensible adaptive procedure. For example, we find ideas from the multi-arm bandit literature, including the Thompson sampling [15], epsilon-greedy algorithms [16], and the UCB algorithm [17] to be useful when constructing the adaptive sampling procedure. Although ideas from reinforcement learning can be utilized when performing the ART, the objective of independence testing is different than that of a typical reinforcement learning problem. This difference is illustrated and further emphasized in the theoretical analysis of the normal means bandit problem in Section 3.

1.3 The Conditional Randomization Test (CRT)

We begin by introducing the CRT that requires an *iid* sampling procedure. The CRT assumes that the data $(X_t, Z_t, Y_t) \overset{iid}{\sim} f_{XZY}$ for $t = 1, 2, \ldots, n$, where f_{XZY} denotes the joint probability density function (pdf) or probability mass function (pmf) of (X, Z, Y) and n is the total sample size. For brevity, we refer to both probability density function and probability mass function as pdf.[1] The CRT aims to test whether the variable of interest X affects the distribution of Y conditional on Z, i.e., $Y \perp\!\!\!\perp X \mid Z$. If Z is the empty set, the CRT reduces to the (unconditional) randomization test. The CRT tests $Y \perp\!\!\!\perp X \mid Z$ by creating "fake" resamples $\tilde{X}_t^b$ for $t = 1, 2, \ldots, n$ from the conditional distribution $X \mid Z$ induced by f_{XZ}, the joint pdf of (X, Z), for $b = 1, 2, \ldots, B$, where B is the Monte-Carlo parameter of choice. More formally, the fake resamples $\tilde{X}_t^b$ are sampled in the following way,

$$\tilde{X}_t^b \sim \frac{f_{XZ}(\tilde{x}_t^b, z_t)}{\int_x f_{XZ}(x, z_t)\mathrm{d}x} \text{ for } t = 1, 2, \ldots, n, \tag{1}$$

where the right hand side is the pdf of the conditional distribution $X \mid Z$ induced by the joint pdf f_{XZ}, lower case $\tilde{x}_t^b$ represents the realization of random variable $\tilde{X}_t^b$, and each $\tilde{X}_t^b$ is sampled *iid* for $b = 1, 2, \ldots, B$ independently of X and Y. Since each sample X_t only depends on the current Z_t, the right hand side of Equation (1) is a conditional distribution that is a function of only its

[1] Neither the CRT nor our paper needs to assume the existence of the pdf. However, for clarity and ease of exposition, we present the data generating distribution with respect to a pdf.

current Z_t. Under the conditional independence null, $Y \perp\!\!\!\perp X \mid Z$, the authors of [3] show that $(\tilde{\mathbf{X}}^{\mathbf{1}}, \mathbf{Z}, \mathbf{Y})$, $(\tilde{\mathbf{X}}^{\mathbf{2}}, \mathbf{Z}, \mathbf{Y})$, ..., $(\tilde{\mathbf{X}}^{\mathbf{B}}, \mathbf{Z}, \mathbf{Y})$, and $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ are exchangeable, where $\mathbf{X}$ denotes the complete collection of $(X_1, X_2, \ldots, X_n)$. $\tilde{\mathbf{X}}^{\mathbf{b}}$, $\mathbf{Z}$, and $\mathbf{Y}$ are defined similarly. This implies that any test statistic $T(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is also exchangeable with $T(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y})$ under the null. This key exchangeability property allows practitioners to use any test statistic T when calculating the final p-value. More formally, the CRT proposes to obtain a p-value in the following way,

$$p_{\text{CRT}} = \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbf{1}_{\{T(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y}) \geq T(\mathbf{X}, \mathbf{Z}, \mathbf{Y})\}}\right], \tag{2}$$

where the addition of 1 is included so that the null p-values are stochastically dominated by the uniform distribution. Due to the exchangeability of the test statistics, the p-value in Equation (2) is guaranteed to have exact type-1 error control, i.e., $\mathbb{P}(p_{\text{CRT}} \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$ (under the null) despite the choice of T and any $Y \mid (X, Z)$ relationship. This also allows the practitioner to ideally choose a test statistic to powerfully distinguish the observed test statistic with the resampled fake test statistic such as the sum of the absolute value of the main effects of X from a penalized Lasso regression [18].

2 Methodology

2.1 Sequential Adaptive Sampling Procedure

The ART, like the CRT, is tied to a specific sampling procedure. Although it generalizes the *iid* sampling procedure, it still relies on a specific sequentially adaptive sampling procedure. Therefore, we also refer to the sequentially adaptive sampling procedure as the ART sampling procedure and now formally present the definition of this procedure.

Definition 2.1 (Sequential Adaptive Sampling Procedure - The ART sampling procedure)**.** We say the sample $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ follows a sequential adaptive sampling procedure A if the sample obeys the following sequential data generating process.

$$\begin{aligned} (X_1, Z_1) &\sim f_1^A(x_1, z_1), \ \ Y_1 \sim f_Q(x_1, z_1) \\ &\cdots \\ (X_t, Z_t) &\sim f_t^A(x_t, z_t \mid x_1, z_1, y_1, \ldots, x_{t-1}, z_{t-1}, y_{t-1}), \ \ Y_t \sim f_Q(x_t, z_t), \end{aligned}$$

where lower case (x_t, z_t, y_t) denotes the realization of the random variables (X_t, Z_t, Y_t) at time t, respectively, f_t^A denotes the joint pdf of (X_t, Z_t) given the past realizations, and f_N denotes the pdf of the response Y_t as a function of only the current (X_t, Z_t).

Definition 2.1 captures a general sequential adaptive experimental setting, where an experimenter adaptively samples the next values of (X_t, Z_t) according to an adaptive sampling procedure f_t^A that may be dependent on all the history (including the outcome) while "nature" f_N determines the next outcome. We emphasize that f_N is generally unknown and in most cases hard to model exactly. We also remark that practitioners need not implement a fully adaptive scheme, e.g., f_t^A can remain identical and even independent of the history for many t if the researcher wishes to only adapt at some time points (see Section 3 for an adaptive sampling scheme that only adapts once).

The left panel of Figure 1 visually summarizes the sequential adaptive procedure, where we allow the next sample to depend on all the history (including the response). Although Definition 2.1 makes no assumption about the adaptive procedure f_t^A (even allowing the adaptive procedure to change across time), it does implicitly assume that the response Y has no carryover effects, i.e., f_N is only a function of its current realizations (x_t, z_t) as there are no arrows in Figure 1 from previous (X_{t-1}, Z_{t-1}) into current Y_t. It also assumes that f_N is stationary and does not change across time. Both of these assumptions are typically invoked in the sequential reinforcement learning literature [19**?**].

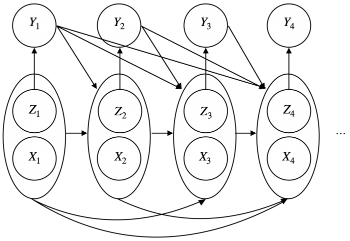

Fig. 1 Schematic diagram of the ART sampling procedure in Definition 2.1. The directed arrows denote the order in how the random variable(s) may affect the corresponding random variable(s).

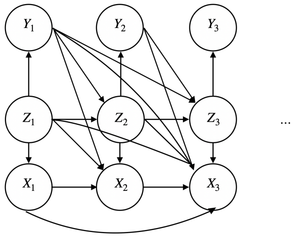

Fig. 2 Schematic diagram of the convenient adaptive sampling procedure that satisfies Assumption 1. As before, the directed arrows denote the order in how the random variable(s) may affect the corresponding random variable(s).

2.2 Hypothesis Test

Given the sampling procedure defined in Definition 2.1, the main objective is to determine whether the variable of interest X affects Y after controlling for Z. Because the sampling scheme is no longer *iid*, testing $Y \perp\!\!\!\perp X \mid Z$ requires further notation and formalization. In the CRT, the null hypothesis of interest is formally $Y_t \perp\!\!\!\perp X_t \mid Z_t$ for all $t = 1, 2, \ldots, n$. Since the data is sampled *iid*, $Y_t \perp\!\!\!\perp X_t \mid Z_t$ reduces to testing $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$ using the whole data since the subscript t is irrelevant. However, for an adaptive collected data, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$ is trivially false for any non-degenerate adaptive procedure A because $\mathbf{X}$ depends on $\mathbf{Y}$ through f_t^A. Just like the CRT, the practitioners are interested in whether X affects Y for each sample t. We now formalize this by testing the following null hypothesis H_0 against H_1,

$$\begin{aligned} H_0 &: f_N(x,z) = f_N(x',z) \text{ for all } x, x' \in \mathcal{X}, z \in \mathcal{Z} \\ H_1 &: f_N(x,z) \neq f_N(x',z) \text{ for some } x, x' \in \mathcal{X}, z \in \mathcal{Z}, \end{aligned} \tag{3}$$

where $\mathcal{X}$ denotes the entire domain of X that captures all possible values of X regardless of the distribution of X induced by the adaptive procedure. For example, if X is a univariate discrete variable that can take any integer values, then $\mathcal{X} = \mathbb{Z}$ even if the adaptive procedure A only has a finite support with positive probability only on values $(-1,0,1)$. In such a case, testing H_0 using the aforementioned adaptive procedure A will only be powerful up to the restricted support induced by A. $\mathcal{Z}$ is defined similarly as the entire domain for Z.

We finish this subsection by connecting H_0 to the causal inference literature. First, H_0 captures the same notion as the CRT null of $Y \perp\!\!\!\perp X \mid Z$ because if X makes any distributional impact on Y given Z, then H_0 is false. On the other hand, if H_0 is false, then the CRT null is trivially false. Recently, the authors of [2] show that the CRT null is equivalent to testing the following causal hypothesis

$$H_0^{\text{Causal}} : Y_t(x,z) \overset{d}{=} Y_t(x',z) \text{ for all } x, x' \in \mathcal{X}, z \in \mathcal{Z},$$

where $Y_t(x,z)$ is the potential outcome for individual t at values $X = x, Z = z$ and we have implicitly assumed the Stable Unit Treatment Value Assumption (SUTVA) assumption [13], where the potential outcomes of each individual t is a function of only its own values (X_t, Z_t). The proposed H_0 already captures the causal hypothesis H_0^{Causal} because $f_Q(x,z)$ characterizes the causal relationship between (X,Z) and Y. To formally establish this in the potential outcome framework, we define $Y_t(x,z) \overset{i.i.d}{\sim} f_N(x,z)$ from a super-population framework, i.e., the potential outcomes are viewed as random variables. Then H_0 is equivalent to the causal hypothesis H_0^{Causal}. Additionally, if the researcher wishes to think in terms of the finite-population framework, i.e., conditioning on the potential outcomes and units in the sample, then only a simple modification of Definition 2.1 is needed. We first replace obtaining the response $Y_t \sim f_N$ in Definition 2.1 from a stochastic f_N to a fixed potential outcome $Y_t = Y_t(x_t, z_t)$ at every time point t, where $Y_t(x_t, z_t)$ is the deterministic (non-random) potential outcome of individual t with values $X_t = x_t$ and $Z_t = z_t$. Then H_0 reduces to the sharp Fisher null that states $Y_t(x,z) = Y_t(x',z)$ for all $x, x' \in \mathcal{X}, z \in \mathcal{Z}$ and all individuals t in our finite population. This finite-population testing framework is the one proposed in [12], where the authors perform the unconditional randomization test in a sequential adaptive setting like ours.

2.3 Adaptive Randomization Test (ART)

Since (X_t, Z_t, Y_t) are no longer sampled *iid* from some joint distribution, the main challenge is to construct $\tilde{\mathbf{X}}^{\mathbf{b}}$ such that $(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y})$ and $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ are still exchangeable to ensure the validity of the p-value in Equation (2). A necessary condition for the joint distributions of $(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y})$ and $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ to be exchangeable is that they are equal in distribution. For our sequential adaptive sampling procedure, X_t depends on all the history including the response and it is unclear how to construct our resamples.

To solve this, we propose a natural resampling procedure that respects our sequential adaptive setting in Definition 2.1. Before formally presenting the resampling procedure, we provide intuition on how to construct valid resamples $\tilde{\mathbf{X}}^{\mathbf{b}}$. Similar to the CRT, the key is to create fake copies of X by replicating the original sampling procedure of X conditional on $\mathbf{Z}, \mathbf{Y}$. For the CRT sampling procedure, this reduces to sampling X_t *iid* from the conditional distribution of $X_t \mid Z_t$ for all $t = 1, 2, \ldots, n$. In our sequential adaptive sampling procedure, this reduces to sampling X_t conditional on the history as done exactly in the original adaptive sampling procedure since X_t does not depend on the future values of Z and Y. We now formalize this in the following definition.

Definition 2.2 (Natural Adaptive Resampling Procedure)**.** Given data $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, $\tilde{\mathbf{X}}^b$ follows the natural adaptive resampling procedure if $\tilde{\mathbf{X}}^b$ satisfies the following data generating process,

$$\tilde{X}_1^b \sim \frac{f_1^A(\tilde{x}_1^b, z_1)}{\int_z f_1^A(\tilde{x}_1^b, z)\mathrm{d}z}, \tilde{X}_2^b \sim \frac{f_2^A(\tilde{x}_2^b, z_2 \mid \tilde{x}_1^b, z_1, y_1)}{\int_x f_2^A(x, z_2 \mid \tilde{x}_1^b, z_1, y_1)\mathrm{d}x}, \ldots,$$

$$\tilde{X}_n^b \sim \frac{f_n^A(\tilde{x}_n^b, z_n \mid \tilde{x}_1^b, z_1, y_1, \ldots, \tilde{x}_{n-1}^b, z_{n-1}, y_{n-1})}{\int_x f_n^A(x, z_n \mid \tilde{x}_1^b, z_1, y_1, \ldots, \tilde{x}_{n-1}^b, z_{n-1}, y_{n-1})\mathrm{d}x},$$

for $b = 1, 2, \ldots, B$ independently conditional on $(\mathbf{Z}, \mathbf{Y})$, where $\tilde{x}_t^b$ are dummy variables representing $\tilde{X}_t^b$.

Similar to Equation (1), Definition 2.2 formalizes how each $\tilde{X}_t$ is sequentially sampled from the conditional distribution of $X_t \mid (X_{1:(t-1)}, Z_{1:t}, Y_{1:(t-1)})$. We call this the natural adaptive resampling procedure (NARP) because at each time t the fake resamples $\tilde{X}_t^b$ are sampled from the original sequential adaptive distribution of X_t conditional on $Z_{1:t}$ and $Y_{1:(t-1)}$ (see Appendix C for further discussions about the NARP). Just like the CRT, Definition 2.2 requires one to sample from a conditional distribution. For this practically important consideration, we propose a more practical alternative where the experimenter, at each time t, samples Z_t first and then samples the variable of interest X_t from $X_t \mid Z_{1:t}, Y_{1:(t-1)}$ at every time step (as opposed to simultaneously sampling (X_t, Z_t) from a joint distribution). This alternative procedure loses very little generality but allows the NARP in Definition 2.2 to directly sample from the already available conditional distribution.

Unfortunately resampling from the NARP does not immediately guarantee a valid p-value. Recall that we require our resampled $\tilde{\mathbf{X}}^b$ to be exchangeable with $\mathbf{X}$ conditional on $(\mathbf{Z}, \mathbf{Y})$. A necessary condition of exchangeability requires the joint distribution of $(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z})$ be the same as that of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. In particular, the following distributional relationship is always true for any t when assuming the NARP,

$$\tilde{X}_{1:(t-1)} \perp\!\!\!\perp Z_t \mid \left(Y_{1:(t-1)}, Z_{1:(t-1)}\right), \tag{4}$$

because $\tilde{X}_{1:(t-1)}$ is a random function of only $\left(Y_{1:(t-1)}, Z_{1:(t-1)}\right)$ and not the future Z_t. Equation (4) directly shows that Z can not depend on previous X because we require $(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z})$ and $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ to be exchangeable. This constraint turns out to be both sufficient and necessary to ensure validity of using the ART with the NARP to test H_0 as formally stated in Theorem 2.1 and Theorem 2.2.

Assumption 1 (Z can not adapt to previous X)**.** For each $t = 1, 2, \ldots, n$ we have by basic rules of probability $f_t^A(x_t, z_t \mid x_{1:(t-1)}, z_{1:(t-1)}, y_{1:(t-1)}) = g_t^A(x_t \mid x_{1:(t-1)}, z_{1:(t-1)}, y_{1:(t-1)}, z_t) h_t^A(z_t \mid x_{1:(t-1)}, z_{1:(t-1)}, y_{1:(t-1)})$, where g_t^A, h_t^A denotes the conditional and marginal density functions induced by the joint pdf of f_t^A, respectively. We say an adaptive procedure A satisfies Assumption 1 if $h_t^A(z_t \mid x_{1:(t-1)}, z_{1:(t-1)}, y_{1:(t-1)})$ does not depend on $x_{1:(t-1)}$, for $t = 2, 3, \ldots, n$.

Assumption 1 states that the sequential adaptive procedure A does not allow Z_t to depend on $X_{1:(t-1)}$. For the gender example above, Assumption 1 does not allow other factors, e.g., party affiliation, candidate personality, etc., to depend on the previous values of gender. However, Assumption 1 still allows the practitioner to sample the next values of gender based on all the historical data, even sampling more of male or female based on a strong interaction with other factors. Although Assumption 1 does restrict our adaptive procedure, it is crucial that each X_t and Z_t are still allowed to adapt by looking at its own previous values and the previous responses.

We visually summarize Assumption 1 and a more convenient, but not necessary, way to conduct a restricted adaptive sampling procedure in Figure 2. Figure 2 shows a set of arrows from Z_t into X_t as opposed to them being simultaneously generated as in Figure 1 to allow the proposed NARP in Definition 2.2 to conveniently sample directly from the already available conditional distribution. Assumption 1 is also satisfied in Figure 2 as there exist no arrows from any $X_{t'}$ into Z_t for $t' < t$. Before stating our main theorem, we summarize the ART procedure in Algorithm 1. We note that although the p-value p_{ART} in Equation (5) is similar to p_{CRT} in Equation (2), the resamples $\tilde{X}^b$ are different in the two procedures. We now state the main theorem that shows the finite-sample validness of using the ART for testing H_0.

```
Algorithm 1 ART p-value
Input: Adaptive procedure $A$, test statistic $T$, total number of resamples $B$
Given an adaptive procedure $A$, obtain $n$ samples of $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ according to the sequential adaptive procedure in Definition 2.1.
for b = 1 to B do : Sample $\tilde{X}^{(b)}$ according to the NARP in Definition 2.2;
Output:
```

$$p_{\text{ART}} := \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbf{1}_{\{T(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y}) \geq T(\mathbf{X}, \mathbf{Z}, \mathbf{Y})\}}\right] \tag{5}$$

Theorem 2.1 (Valid p-values using the ART)**.** Suppose the adaptive procedure A follows the adaptive procedure in Definition 2.1 and satisfies Assumption 1. Further suppose that the resampled $\tilde{X}^b$ follows the NARP in Definition 2.2 for $b = 1, 2, \ldots B$. Then the p-value p_{ART} in Algorithm 1 for testing H_0 is a valid p-value. Equivalently, $\mathbb{P}(p_{\text{ART}} \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$. In addition, p_{ART} is also a valid p-value conditional on $\mathbf{Y}$ and $\mathbf{Z}$.

This theorem allows testing H_0 for sequentially adaptive sampling procedures through randomization inference. Before concluding this section, as alluded before, we state formally in Theorem 2.2 that our assumption is indeed necessary to establish the exchangeability of $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y})$ if we follow the natural adaptive procedure in Definition 2.2. The proofs of Theorem 2.1 and Theorem 2.2 can be found in Appendix D.

Theorem 2.2 (Necessity of Assumption)**.** For an adaptive procedure A, if the resampled $\tilde{\mathbf{X}}^b$ follows the natural adaptive resampling procedure in Definition 2.2 and $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y})$ are exchangeable, then Assumption 1 must hold, i.e., Assumption 1 is necessary.

3 ART in Normal Means Model

In this section, we explore the ART under the well-known normal-means setting [20]. We first introduce the normal-means setting, the sampling procedures we consider, and the test statistic in Section 3.1. We then present two main theorems, Theorem A.1 and Theorem A.2, that characterize the asymptotic power of both the *iid* procedure and a naïve, but still insightful, two stage adaptive sampling procedure under local alternatives of $O(n^{-1/2})$ distance in Section 3.2. Finally, we numerically evaluate these two theorems to illustrate when the adaptive sampling procedure leads to an increase of power in Section 3.3. Lastly, we postulate the main reasons for why an adaptive sampling procedure is more powerful than an *iid* sampling procedure in Section 3.4.

3.1 Normal Means Model

Formally, the normal-means model is characterized by the following model.

$$Y \mid (X = j) \sim \mathcal{N}(\theta_j, 1), \qquad \text{for } j \in \mathcal{X} := \{1, 2, \ldots, p\},$$

where j refers to the p different possible integer values of X.[2] We refer to the different values of X as different arms. For this setting there are no other experimental variables Z. Our task is to characterize power under the alternative, i.e., when at least one arm of X has a different mean than that of the other arms. For simplicity, we consider an alternative where only one arm has a positive non-zero mean while the remaining $p-1$ arms have zero mean. This leads to the following one-sided alternative.

$$\mathrm{H}_1^{\mathrm{NMM}} : \text{there exists only one } j^\star \text{ such that } \theta_{j^\star} = h > 0 \text{ and } \theta_j = 0, \forall j \neq j^\star.$$

As usual, our null assumes that X does not affect Y in any way,

$$H_0^{\mathrm{NMM}} : \theta_j = 0, \forall j \in \{1, 2, \ldots, p\}.$$

Given a budget of n samples, our task is to come up with a reasonable adaptive sampling procedure that leads to a higher power than that of the typical uniform *iid* sampling procedure using CRT. Because we do not use a fully adaptive procedure for this setting but a simplified two step adaptive procedure, we use subscript i instead of t to denote the sample index for this section. We now formally state the general *iid* sampling procedure.

Definition 3.1 (Normal Means Model: *iid* Sampling procedure with Weight Vector q)**.** We call a sampling procedure *iid with weight vector* $q = (q_1, q_2, \cdots, q_p)$ if each sample of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is sampled independently and $\mathbb{P}(X_i = j) = q_j$, for all $i = 1, 2, \ldots, n$ and $j \in \mathcal{X}$.

We note that this definition is more general than the uniform iid sampling procedure that pulls each arm with equal probability, i.e., $q = (1/p, 1/p, \ldots, 1/p)$. We use $\mathbf{X} \sim \mathcal{M}(q)$ to compactly describe this *iid* sampling procedure. With a slight abuse of notation, we also use $X_i \sim \mathcal{M}(q)$ to denote the above distribution of X_i.

Despite the simplicity of the normal-means setting, analyzing the power of a fully adaptive procedure is generally theoretically infeasible. Therefore, we consider a naïve "two stage" adaptive procedure. The first stage is an exploration stage that follows the typical *iid* sampling procedure while the second stage is again another *different iid* sampling procedure that adapts once based on the first stage's data. More specifically, the second stage will adapt by reweighting the probability of pulling each arm by a function of

[2] We remind the reader that p is used to denote the cardinality of $\mathcal{X}$ as opposed to the dimension of $\mathcal{X}$.

the sample mean. Under the alternative, we expect the arm with the true signal will on average have a higher sample mean, thus we can exploit this arm more in the second stage. Furthermore, the adaptive procedure will also detect arms that, by chance, lead to higher sample means. In such a case, we can additionally identify these “fake” signal arms and sample more to “denoise” and reduce the variance from these arms. We note that this two-stage adaptive procedure does not utilize the full potential of an adaptive sampling procedure, but we show that even a simple two stage adaptive procedure can lead to insightful gains and conclusions. We formally summarize the adaptive procedure in Definition 3.2.

Definition 3.2 (Normal Means Model: Two Stage Adaptive Sampling procedure)**.** An adaptive sampling procedure is called a *two stage adaptive sampling procedure* with *exploration parameter* ϵ, *reweighting function* f and *scaling parameter* t if $(\mathbf{X}, \mathbf{Y})$ are sampled by the following procedure. First, for $1 \leq i \leq [n\epsilon]$,

$$X_i \overset{iid}{\sim} \mathcal{M}(q); \quad Y_i \overset{iid}{\sim} f_N(x_i).$$

Second, for each $j \in \mathcal{X}$, we compute the sample mean for each arm using the $[n\epsilon]$ samples from the first stage,

$$\bar{Y}_j^{\mathrm{F}} := \frac{\sum_{i=1}^{[n\epsilon]} Y_i \mathbf{1}_{X_i=j}}{\sum_{i=1}^{[n\epsilon]} \mathbf{1}_{X_i=j}},$$

in which the superscript “F” stands for the first stage. Third, we calculate a reweighting vector $Q \in \mathbb{R}^p$ as a function of $\bar{Y}_i^{\mathrm{F}}$’s that captures the main adaptive step,

$$Q_j = \frac{f(t\sqrt{n} \cdot \bar{Y}_j^{\mathrm{F}})}{\sum_{k=1}^{p} f(t\sqrt{n} \cdot \bar{Y}_k^{\mathrm{F}})}. \tag{6}$$

Finally, we sample the second batch of samples using the new weighting vector, namely, for $[n\epsilon] + 1 \leq i \leq n$

$$X_i \overset{iid}{\sim} \mathcal{M}(Q); \quad Y_i \overset{iid}{\sim} f_N(x_i).$$

We comment that $f(\cdot)$ denotes the adaptive re-weighting function. For example if $f(x) = e^x$, then this reweighs the probability by an exponential function, where t is a hyper-parameter of choice and a larger value of t will lead to a more disproportional sampling of different arms for the second stage. We also scale the reweighting function by $\sqrt{n}$ because the signal decreases with rate $1/\sqrt{n}$ as we describe now in the following section.

3.2 Power Analysis Through Local Asymptotics

Although practically one could simulate the power for both the *iid* sampling procedure and the adaptive sampling procedure, we theoretically characterize the power for deeper insights and exploration across an entire grid of different signal strengths and number of arms of X. To characterize the asymptotic power of both the uniform *iid* sampling procedure and the two stage adaptive sampling procedure, we use key ideas from the classical local asymptotic theory [21]. We remark that for our setting we apply local asymptotic theory to characterize the power of different sampling procedures as opposed to characterizing the distribution of different test statistics of the data from a fixed sampling procedure.

In our asymptotic setting, we keep p fixed and let $n \to \infty$. To avoid the power from approaching one, we scale our signal strength h proportional to the standard parametric rate $n^{-1/2}$, i.e.,

$$h = h_0/\sqrt{n} > 0, \tag{7}$$

where h_0 is a positive constant.

As introduced in Definition 3.1, we first analyze the power under an *iid* sampling procedure with arbitrary weight vector $q = (q_1, q_2, \cdots, q_p)$ such that q_i's are all positive and $\sum_{i=1}^{p} q_i = 1$. Without loss of generality, we assume under H_1^{NMM} the signal is in the first arm, i.e., $j^\star = 1$. Consequently, we have under $\mathrm{H}_1^{\text{NMM}}$,

$$\begin{aligned}
\mathbf{X} \sim & \mathcal{M}(q), \\
Y_i \mid X_i = 1 & \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(\frac{h_0}{\sqrt{n}}, 1\right), \\
Y_i \mid X_i = j & \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1), \text{ for } j \neq 1.
\end{aligned}$$

To compute the p-value as done in Equation (5), we need a reasonable test statistic. We use maximum of all sample means for each arm as the main proposed test statistic,

$$T(\mathbf{X}, \mathbf{Y}) = \max_{j \in 1,2,\ldots,p} \bar{Y}_j := \max_{j \in 1,2,\ldots,p} \frac{\sum_{i=1}^{n} Y_i \mathbf{1}_{X_i = j}}{\sum_{i=1}^{n} \mathbf{1}_{X_i = j}}. \tag{8}$$

We remark that another natural test statistic, $\bar{Y}$ (the sample mean), is degenerate in our testing framework since it does not depend on $\mathbf{X}$ or $\tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}}$ is the fake copy. For the sake of notation simplicity, we define the following resampled test statistic

$$\tilde{T}(\tilde{\mathbf{X}}, \mathbf{Y}) = \max_{j \in 1,2,\ldots,p} \tilde{\bar{Y}}_j := \max_{j \in 1,2,\ldots,p} \frac{\sum_{i=1}^{n} Y_i \mathbf{1}_{\tilde{X}_i = j}}{\sum_{i=1}^{n} \mathbf{1}_{\tilde{X} = j}},$$

in which, formally speaking, $\tilde{\mathbf{X}} = (\tilde{X}_1^1, \ldots, \tilde{X}_n)^1 := \tilde{\mathbf{X}}^1$ and readers should comprehend $\tilde{\mathbf{X}}$ as a generic copy of $\tilde{\mathbf{X}}^b$. Lastly, to deal with the Monte-Carlo parameter B, we show in Appendix E that as $B \to \infty$ the power of testing H_1

against H_0 is equal to

$$\mathbb{P}\left(\mathbb{P}\left[T(\mathbf{X},\mathbf{Y}) > z_{1-\alpha}\left(\tilde{T}(\tilde{\mathbf{X}},\mathbf{Y})\right)\middle|\mathbf{Y}\right]\right), \tag{9}$$

where $\mathbb{P}$ denotes the probability measure induced by all randomness and $z_{1-\alpha}\left(\tilde{T}(\tilde{\mathbf{X}},\mathbf{Y})\right)$ is the $1-\alpha$ quantile of the distribution of $\tilde{T}(\tilde{\mathbf{X}},\mathbf{Y})$ conditioning on $\mathbf{Y}$.

With the above setting, one can explicitly derive the joint asymptotic distributions of $\bar{Y}_j$'s, $\bar{\tilde{Y}}_j$'s and $\bar{Y}$ under the alternative H_1. Consequently, the asymptotic power of both the *iid* sampling procedures (Theorem A.1) and our two-stage adaptive sampling procedures (Theorem A.2) with test statistic T as defined in Equation 8 can be characterized. For brevity and readability, we omit the explicit formulation of the theorems here, but they have been derived and are included in Appendix A. Therefore, for now, we will proceed assuming that these asymptotic results have been established.

Although we characterize the asymptotic power for the *iid* sampling procedures for any general weight vector q, the default choice of weight vector q should be $(1/p, 1/p, \ldots, 1/p)$ since the practitioner typically has no prior information about which arm is more important. We refer to this choice of q as the uniform *iid* sampling procedure. Suppose an oracle knows which arm is the signal. Then a naïve, but natural idea for the oracle would be to sample more from the arm with signal (large value of q_1) to maximize power. As shown in the next section, this is not necessarily the best strategy. In other words, the optimizer $\hat{q}_1 := \arg\max_{q_1} \mathrm{Power}_{\mathrm{iid}}(q)$ is not always larger than $1/p$, illustrating that it is actually better to sometimes sample less from the actual signal arm depending on the signal strength. This hints at the well known bias-variance trade-off between the mean difference of T and $\tilde{T}$ and their variances. Another natural idea is to construct an adaptive procedure that up-weights or down-weights the signal arm according to the oracle weight. However, Section 3.3 shows this naïve strategy is not always recommended as the adaptive procedure can do better than even the oracle *iid* sampling procedure.

Furthermore, we remark that the final expression that characterizes the asymptotic power in both Theorem A.1 and Theorem A.2 are not immediately insightful due to the complicated nature of both the "maximum" test statistic and the adaptive sampling procedure. Though Theorem A.1 and Theorem A.2 are not directly interpretable, the computational cost of evaluating it numerically is less than naïvely simulating the adaptive procedure for a large value of n by a factor of $O(n)$. Moreover, since the asymptotic power does not depend on n, the conclusion is naturally more consistent and unified when compared to the empirical power obtained from simulating with different large sample size. Apart from the computational advantages the theorem provides, it is also of theoretical interest as our work leverages local asymptotic power analysis to characterize the distributions under different sampling strategies as opposed to characterizing the distributions under different test statistics. In addition, these two theorems can also serve as a starting point and motivating example for theoretically analyzing the power of the ART for future works.

3.3 Power Results

Given the asymptotic results alluded in Section 3.2 and formally presented in Appendix A, we now attempt to understand how the ART using an adaptive sampling procedure may be more powerful than the CRT using an *iid* sampling procedure. As alluded in the previous subsection, if a practitioner knows which arm contains the signal, then a naïve but natural adaptive strategy is to up-weight or down-weight the known signal arm according to the oracle. We formally define the oracle in the following way, where we assume, without loss of generality, $j^\star = 1$,

$$q_1^\star := \arg \max_{0 \leq q_1 \leq 1} \text{Power}_{\text{iid}}(q(q_1)),$$

in which $q(q_1) := (q_1, (1-q_1)/(p-1), (1-q_1)/(p-1), \ldots, (1-q_1)/(p-1)) \in \mathbb{R}^p$ denotes the sampling probabilities of all p arms, where the first signal arm has probability q_1 and the remaining arms (that have no signal) equally share the remaining sampling probability. Let $q^\star = q(q_1^\star)$, i.e., the oracle *iid* sampling procedure that samples the known treatment arm in an optimal way. We refer to the *iid* sampling with weight vector $q^\star$ as the "oracle *iid* sampling procedure."[3]

Next, we use numerical evaluations of Theorem A.1 and Theorem A.2 to compare the power of the (two-stage) adaptive sampling procedure, uniform *iid* sampling, and the oracle *iid* sampling procedure across a grid of possible signal strengths h_0 and number of arms p. For the adaptive sampling procedure described in Definition 3.2, we choose the reweighting function f to be the exponential function, i.e., $f(x) = \exp(x)$.

Figure 3 shows how the ART's power with the proposed adaptive sampling procedure is greater than that of both the uniform *iid* sampling procedure and even the oracle *iid* sampling procedure. To produce this figure, we first fix an arbitrary, but reasonable, combination of hyper-parameters for the ART, i.e., we set exploration parameter $\epsilon = 0.5$ and reweighting parameters $t_0 = \log 2$ and $t = t_0/h_0$. As a reminder, exploration parameter $\epsilon = 0.5$ implies the adaptive procedure spends half of the sampling budget on exploration and only adapts once by reweighting (see Definition 3.2) after the first half of the *iid* samples are collected. The choice of $t_0 = \log 2$ allows the first arm (containing the real signal) to get roughly twice more sampling weight than the remaining arms in the second stage in expectation. Appendix F shows additional simulations with different choices for the adaptive parameters (ϵ, t_0),

[3] $q^\star$ is not formally the most optimal *iid* sampling procedure for all possible *iid* sampling procedure since we consider the maximum power when only varying q_1 while imposing the remaining arms to all have equal probabilities. However, we do not imagine any other reasonable *iid* sampling procedure to have a stronger power than $q^\star$ since the remaining $p-1$ arms with no signals are not differentiable in any way, thus we lose no generality by setting them with equal probability.

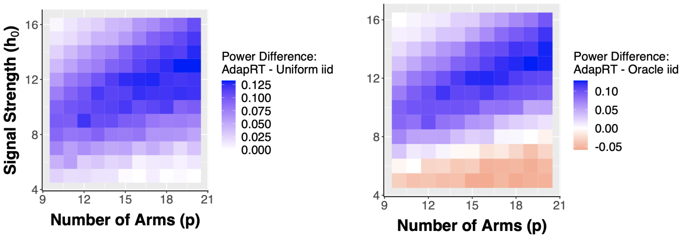

Fig. 3 The figure shows the difference between the local asymptotic power of the ART using the adaptive sampling procedure in Definition 3.2 (with a fixed arbitrary choice of hyper-parameters $\epsilon = 0.5$ and $t = \log 2/h_0$) and the CRT using an *iid* sampling procedure for different values of signal strength h_0 and number of arms p. All tests use the test statistic defined in Equation 8. The left plot shows that the power of the adaptive sampling procedure is almost uniformly higher than that of the default uniform *iid* sampling. The right plot shows that the power of the adaptive sampling procedure is higher than that of even the oracle *iid* sampling procedure when the signal strength is relatively high. We note that values on the top left corners of both heatmaps are close to 0 only because the power of all three sampling procedures is almost degenerately one. The significance level is $\alpha = 0.05$. These heat maps are generated based on Monte Carlo evaluations of Theorem A.1 and Theorem A.2.

demonstrating that the results presented here are not sensitive to the initially chosen parameters.

Figure 3 shows that the power of the ART from the adaptive sampling procedure is uniformly better than that of the CRT using the default uniform *iid* sampling procedure. For example, in areas that have high number of arms and signal, the adaptive sampling procedure can have close to 10 percentage points higher power than the uniform *iid* sampling procedure. The right panel of Figure 3 surprisingly shows that the adaptive sampling procedure can be more powerful than even the oracle *iid* sampling procedure when the signal strength is relatively high. This power difference can be as large as 10 percentage points when the signal and number of arms are high. However, we note that the adaptive sampling procedure's power can be lower than that of the oracle *iid* sampling procedure when the signal is low. We postulate further in Section 3.4 how and why the ART may be helping in power. We note that for both panels in Figure 3, the top left corners of the heatmaps have zero difference between the two sampling procedures because this regime of strong signal and low p results in a degenerate power close to one, allowing no significant differences.

3.4 Understanding why Adapting Helps

In this subsection, we summarize some of the insights we find from the above analysis of the normal means model. Our goal is to characterize key ideas of why adapting is helpful so practitioners can also build their own successful adaptive procedure. We acknowledge that all statements here are respect to the specific normal-means model setting, but we believe that the main ideas

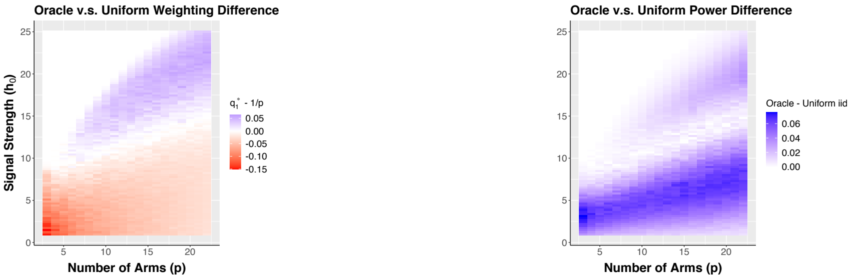

Fig. 4 This figure compares the theoretical power of the CRT from an *iid* oracle sampling procedure with the CRT from an uniform *iid* sampling procedure. The first panel on the left compares whether oracle $q_1^\star$ should down-weight (less than $1/p$) or up-weight (more than $1/p$) the signal arm. The second panel compares the power difference between the oracle and uniform sampling procedures.

should generalize to different applications and scenarios as shown in Section 4 for instance. Unfortunately, it is difficult to theoretically verify many of the presented insights because the power of the ART and the CRT depends on the behavior of also the resampled test statistics. For example, even if we empirically verify that the adaptive procedure is sampling arms with zero signal with lower probability, it does not directly imply the power is greater because the resampled test statistic may exhibit the same behavior. This would make both the *observed* and *resampled* test statistic approximately indistinguishable, leading to an insignificant p-value. Therefore, Figure 3 should serve as the main result that highlights how adapting can indeed help. Nevertheless, we attempt to show some empirical evidence of how adapting is helping.

As pointed out at the beginning of Section 3.3, a natural idea is to try to design adaptive strategies that mimic the oracle *iid* procedure. However, the power gain shown in the left plot of Figure 3 can not be attributed to only mimicking the oracle *iid* sampling procedure because the right plot of Figure 3 shows the adaptive sampling procedure can be more powerful than even the oracle *iid* sampling as long as the signal strength is not too low. Additionally, it is unclear if the oracle sampling procedure always samples the signal arm with higher probability as our adaptive sampling procedure does. Consequently, to understand the oracle sampling procedure's behavior further, we present Figure 4 that compares the oracle sampling procedure's behavior with the *iid* uniform sampling procedure.

The left plot of Figure 4 shows that the oracle up-weights and also down-weights the signal arm depending on h_0 and p. For example, the red regions shows that the oracle actually down-weights the signal arm to spend more sampling budget on other arms. Therefore, if mimicking the oracle sampling procedure is the ideal solution, the adaptive procedure should down-weight the signal arm for the red regions in Figure 4. However, when comparing Figure 3 and the left plot in Figure 4, we see that the up-weighting (since $t > 0$) adaptive procedure can actually beat not only the uniform *iid* sampling procedure but also the oracle *iid* sampling procedure. This shows that the adaptive procedure is doing more than just mimicking the oracle sampling procedure.

Instead, as alluded previously, we believe the main intuition behind the success of the ART is for the following three reasons. As expected, the first reason is that an adaptive sampling procedure can, to some extent, mimic the oracle *iid* procedure and achieve closer-to-oracle sampling proportions on average (at least for the regimes that up-weight the signal arm). Additionally, and most importantly, when the adaptive sampling procedure samples more from the arms that look like signal it is not only sampling from the arms that is truly the real signal but also the arms that are "fake" signals due to random chance. This allows the adaptive procedure to de-noise these "fake" signal arms to a correctly null state. Thirdly, adapting also down-weights arms (with high probability) that contain no signal, allowing our remaining samples to focus on exploring the more relevant arms. In summary, we find that adaptive procedure that sample more from signal arms generally lead to improvements in power. However, as shown in Figures 3-4, this power advantage vanishes when the level of adaptivity becomes "unchecked", i.e., too far from the uniform *iid* sampling procedure. Consequently, we postulate that adaptive procedure that balance both exploration and exploitation are generally promising for improvements in power.

4 Application of ART in Conjoint Studies

In this section, we further demonstrate the power of the ART in a popular factorial design called conjoint analysis. Conjoint analysis, introduced more than half a century ago [22], is a factorial survey-based experiment designed to measure preferences on a multidimensional scale. Recently the authors of [2] also introduced the CRT in the context of conjoint analysis to test whether a variable of interest X matters at all for a response Y given Z.

Unlike the analysis performed in Section 3, we do not theoretically characterize the asymptotic power and in exchange consider a fully adaptive procedure and a more complicated test statistic. We apply our proposed methodology on a recent conjoint study concerning the role of gender discrimination in political candidate evaluation [4, 23]. In this study, the authors conduct an experiment based on a sample of voting-eligible adults in the U.S. collected in March 2016, where each of the 1,583 respondents were given 10 pairs of political candidates with uniformly sampled levels of gender and twelve other factors (see original article for details). The respondents were then forced to choose one of the two pair of candidate profiles to vote into office, which is our main binary response Y. The study consists of a total of $7,915$ responses, where the primary objective was to test whether gender (X) matters in voting behavior (Y) while controlling for other variables such as age, race, etc. (Z).

The authors of [4] were able to find a statistically significant effect of candidate's gender on voting behavior of Presidential candidates. We attempt to answer this important question of whether gender matters in voting behavior had the experimenter ran the same experiment for the first time but with a lower sample budget $n < 7,915$. To run this quasi-experiment, we assume the original data of size $7,915$ is the population and we draw samples (without replacement) from the original dataset according to our experiment. For simplicity, suppose X is gender and Z is only candidate party. Since each sample consists of a *pair* of profiles, one potential sample may be $X_1 = (\text{Male}, \text{Female})$ and $Z_1 = (\text{Democrat}, \text{Democrat})$, indicating the left profile was a Democratic male candidate and the right profile was Democratic female candidate. Given such a sample, we obtain the subsequent response Y from the original study of 7,915 samples from randomly drawing response Y with corresponding pair of profiles with a Democratic male candidate and a Democratic female candidate. Once we draw this response Y, we do not put it back into the population. Since Z in the original study contained twelve other factors, the probability of observing a unique sequence of a particular (X, Z) is close to zero. For this reason, we only run this quasi-experiment for up to one other Z, namely the candidate's party affiliation (Democratic or Republican) because the authors of [2] suggest potential strong interactions with gender.

Following the intuition presented in Section 3, we build an adaptive procedure that samples arms of (X, Z) such that there is more evidence of a signal based on each arm mean. We similarly set an initial exploration ϵ parameter and use a test statistic based on the sum of the estimated coefficients related to X from a Lasso logistic regression of $\mathbf{Y}$ with main effects of $\mathbf{X}$ and $\mathbf{Z}$ and their interactions. In the original experiment, all factor levels were sampled uniformly and independently. We use this as the baseline *iid* sampling procedure for comparison. Appendix G contains further details of the adaptive procedure and test statistic along with simulation results.

	iid sampling procedure - CRT	Adaptive sampling procedure - ART
$n = 500$	0.13	0.14
$n = 1,000$	0.14	0.17
$n = 2,000$	0.24	0.30
$n = 3,000$	0.31	0.40

Table 1 The two columns represent the power of the CRT with the uniform *iid* sampling procedure and the ART for testing H_0 when $\alpha = 0.1$, where X is gender (Male or Female) in the gender political candidate study in [4] and Z is the candidate's party affiliation (Democratic or Republican). Each row represents a different sample size n that aims to replicate the original experiment had the researchers re-ran the experiment with the respective sampling procedures.

Table 1 shows the power results using both the *iid* sampling procedure and the proposed adaptive sampling procedure. In particular, Table 1 shows that the power of detecting gender effects using the adaptive sampling procedure is consistently higher than that from using the *iid* sampling procedure. For example, when $n = 3,000$ (approximately 37% of the original sample size), we observe a power difference of 9 percentage points with the *iid* sampling procedure only having 31% power, approximately a 30% increase of power.

Acknowledgments. We thank Lucas Janson, Iavor Bojinov and Subhabrata Sen for advice and feedback.

Appendix A Asymptotic Results for the Normal Means Model

The two asymptotic power analysis results we omitted for conciseness of the main text in Section 3.2 are stated here. Proofs of results presented in this section are all in Appendix E.

Theorem A.1 (Normal Means Model: Power of RT under *iid* sampling procedures)**.** Upon taking $B \to \infty$, the asymptotic power of the *iid* sampling procedure with probability weight vector $q = (q_1, q_2, \cdots, q_p)$, as defined in Definition 3.1, with respect to the RT with the "maximum" test statistic, is equal to

$$\text{Power}_{\text{iid}}(q) = \mathbb{P}\left(T_{\text{iid}} \geq z_{1-\alpha}\left(\tilde{T}_{\text{iid}}\right)\right),$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the distribution of $\tilde{T}_{\text{iid}}$. T_{iid} and $\tilde{T}_{\text{iid}}$ are defined/generated as a function of $G := (G_1, G_2, \ldots, G_{p-1})$ and $H := (H_1, H_2, \ldots, H_{p-1})$, both of which are independent and follow the same $(p-1)$ dimensional multivariate Gaussian distribution $\mathcal{N}(0, \Sigma(q))$. T_{iid} and $\tilde{T}_{\text{iid}}$ are then defined as

$$T_{\text{iid}} = T_{\text{iid}}(q, G, H) := \max\left(\{H_1 + h_0\} \cap \{H_j, j = 2, \ldots, p-1\} \cap \left\{-\frac{1}{q_p}\sum_{i=1}^{p-1} q_j H_j\right\}\right);$$

$$\tilde{T}_{\text{iid}} = \tilde{T}_{\text{iid}}(q, G, H) := h_0 q_1 + \max\left(\{G_j, j = 1, \ldots, p-1\} \cap \left\{-\frac{1}{q_p}\sum_{j=1}^{p-1} q_j G_j\right\}\right).$$

Finally, $\Sigma(q)$ is specified by

$$\Sigma(q) := D(q)^{-1}\Sigma_0(q)D(q)^{-1}. \tag{A1}$$

where matrices Σ_0 and D are defined as

$$\Sigma_0(q) := \begin{bmatrix} v(q_1) & -q_1q_2 & \cdots & -q_1q_{p-1} \\ -q_1q_2 & v(q_2) & \cdots & -q_2q_{p-1} \\ \cdots & \cdots & \cdots & \cdots \\ -q_1q_{p-1} & -q_2q_{p-1} & \cdots & v(q_{p-1}) \end{bmatrix} \in \mathbb{R}^{(p-1)\times(p-1)},$$

with $v(x) = x(1-x)$, and $D(q) := \text{diag}(q_1, q_2, \ldots, q_{p-1}) \in \mathbb{R}^{(p-1)\times(p-1)}$.

We note that if we assume p to be "large" (in a generic sense) and our sampling probabilities $q_j = O(1/p)$ for all j, then the diagonal elements of $\Sigma(q)$ will be generally much larger than the off-diagonal elements. Consequently, G and H in Theorem A.1 will have approximately independent coordinates, thus both T_{ind} and $\tilde{T}_{\text{iid}}$ are characterized by nearly independent Gaussian distributions.

By an argument similar to proof for Theorem A.1, we can also derive the asymptotic power for our two-stage adaptive sampling procedures.

Theorem A.2 (Normal Means Model: Power of the ART under two-stage adaptive sampling procedures)**.** Upon taking $B \to \infty$, the asymptotic power of a two-stage adaptive sampling procedures with *exploration parameter* ϵ, *reweighting function* f, *scaling parameter* t and test statistic T as defined in Definition 3.2, with respect to the ART with the "maximum" test statistic, is equal to

$$\mathrm{Power}_{\mathrm{adap}}\left(\epsilon, t, f\right) := \mathbb{P}_{R^{\mathrm{F}}, G^{\mathrm{F}}, R^{\mathrm{S}}, H^{\mathrm{F}}}\Bigg(\mathbb{P}\Bigg(T_{\mathrm{adap}} \geq z_{1-\alpha}(\tilde{T}_{\mathrm{adap}} \mid R^{\mathrm{F}}, R^{\mathrm{S}}, H^{\mathrm{F}}, H^{\mathrm{S}}) \\ \Bigg| R^{\mathrm{F}}, R^{\mathrm{S}}, H^{\mathrm{F}}, H^{\mathrm{S}}\Bigg)\Bigg) \tag{A2}$$

where $z_{1-\alpha}(\tilde{T}_{\mathrm{adap},j} \mid R^{\mathrm{F}}, R^{\mathrm{S}}, H^{\mathrm{F}}, H^{\mathrm{S}})$ denotes the $1-\alpha$ quantile of the conditional distribution of $\tilde{T}_{\mathrm{adap}}$ given R^{F}, R^{S}, G^{F} and G^{S}. Further more,

$$T_{\mathrm{adap}} = \max_{j \in \{1,2,\ldots,p\}} T_{\mathrm{adap},j}; \qquad \tilde{T}_{\mathrm{adap}} = \max_{j \in \{1,2,\ldots,p\}} \tilde{T}_{\mathrm{adap},j};$$

$$\begin{aligned} T_{\mathrm{adap},j} &= \frac{q_j\sqrt{\epsilon}W_j + Q_j\sqrt{(1-\epsilon)}\left[H_j^{\mathrm{S}} + R^S + \mathbf{1}_{j=1}\sqrt{1-\epsilon}h_0\right]}{\epsilon q_j + (1-\epsilon)Q_j}; \\ \tilde{T}_{\mathrm{adap},j} &= \frac{q_j\sqrt{\epsilon}\tilde{W}_j + \tilde{Q}_j\sqrt{(1-\epsilon)}\left(G_j^{\mathrm{S}} + R^{\mathrm{S}} + \sqrt{1-\epsilon}h_0 Q_1\right)}{\epsilon q_j + (1-\epsilon)\tilde{Q}_j}, \end{aligned}$$

where R^{F}, R^{S}, G^{F}, G^{S}, H^{F}, H^{S}, Q, $\tilde{Q}$, W and $\tilde{W}$ are random quantities generated from the following procedure. First, generate $R^{\mathrm{F}} \sim \mathcal{N}(0,1)$, $G^{\mathrm{F}} \sim \mathcal{N}\left(0, \Sigma(q)\right)$, and $H^{\mathrm{F}} \sim \mathcal{N}\left(0, \Sigma(q)\right)$ independently, where $\Sigma(\cdot)$ is defined in Equation A1. Second, compute

$$\begin{aligned} W_j &= H_j^{\mathrm{F}} + R^{\mathrm{F}} + \mathbf{1}_{j=1}\sqrt{\epsilon}h_0, \text{ for } j \in \{1,2,\ldots,p-1\}, \\ \tilde{W}_j &= G_j^{\mathrm{F}} + R^{\mathrm{F}} + \sqrt{\epsilon}h_0 q_1, \text{ for } j \in \{1,2,\ldots,p-1\}, \\ W_p &= -\frac{1}{q_p}\sum_{j=1}^{p-1} q_j H_j^{\mathrm{F}} + R^{\mathrm{F}} + \sqrt{\epsilon}h_0 q_1(1-q_1), \\ \tilde{W}_p &= -\frac{1}{q_p}\sum_{j=1}^{p-1} q_j G_j^{\mathrm{F}} + R^{\mathrm{F}} + \sqrt{\epsilon}h_0 q_1. \end{aligned}$$

Third, compute

$$Q_j = \frac{f(W_j/\sqrt{\epsilon})}{\sum_{j=1}^{p} f(W_j/\sqrt{\epsilon})} \quad \text{and} \quad \tilde{Q}_j \quad = \frac{f(\tilde{W}_j/\sqrt{\epsilon})}{\sum_{j=1}^{p} f(\tilde{W}_j/\sqrt{\epsilon})}.$$

We note that with a slight abuse of notation, the Q defined here is the asymptotic distributional characterization of Equation 6. Lastly, generate $R^{\mathrm{S}} \sim \mathcal{N}(0,1)$, $H^{\mathrm{S}} \sim \mathcal{N}\left(0, \Sigma(Q)\right)$ and $G^{\mathrm{S}} \sim \mathcal{N}\left(0, \Sigma\left(\tilde{Q}\right)\right)$ independently.

Appendix B Multiple Testing

Our proposed method tests H_0 for a single variable of interest X conditional on other experimental variables Z. However, the practitioner may be interested in testing multiple H_0 for multiple variables of interest (including variables from Z).

To formalize this, denote $X = (X^1, X^2, \ldots, X^p)$ to contain p variables of interest, each of which can also be multidimensional. Informally speaking, our objective is to perform p tests of $Y \perp\!\!\!\perp X^j \mid X^{-j}$ for $j = 1, 2, \ldots p$, where X^{-j} denotes all variables in X except X^j. Given a fixed j, our proposed methodology in Section 2.1-2.3 can be used to test any single one of these hypothesis. The main issue with directly extending our proposed methodology for testing all $j = 1, 2, \ldots p$ variables is that Assumption 1 does not allow X^{-j} to depend on previous X^j but X^j may depend on previous X^{-j} when testing a single hypothesis $Y \perp\!\!\!\perp X^j \mid X^{-j}$. This asymmetry may cause this assumption to hold when testing for X^j but simultaneously not hold when testing for $X^{j'}$ for $j \neq j'$. Thus, in order to satisfy Assumption 1 for all variables of interest simultaneously, we modify our procedure such that each X_t^j is independent of $X_{t'}^{j'}$ for all j, j' and $t' \leq t$. In other words, we force each X_t^j to be sampled according to its *own* history $X^j_{1:(t-1)}$ and the history of the response but not the history and current values of $X^{j'}$ for $j \neq j'$ and for every j. We formalize this in following assumption.

Assumption 2 (Each X^j does not adapt to other $X^{j'}$)**.** For each $t = 1, 2, \ldots, n$ suppose each $X_t = (X_t^1, X_t^2, \ldots, X_t^p)$ are sampled according to a sequential adaptive sampling procedure A: $X_t \sim f_t^A(x_t^1, x_t^2, \ldots, x_t^p \mid x^{-j}_{1:(t-1)}, x^j_{1:(t-1)}, y_{1:(t-1)})$. We say an adaptive procedure A satisfies Assumption 2 if f_t^A can be written into following factorized form, for $t = 2, 3, \ldots, n$,

$$f_t^A(x_t^1, x_t^2, \ldots, x_t^p \mid x^{-j}_{1:(t-1)}, x^j_{1:(t-1)}, y_{1:(t-1)}) = \prod_{j=1}^{p} f^A_{t,j}(x_t^j \mid x^j_{1:(t-1)}, y^j_{1:(t-1)})$$

with every $f^A_{t,j}(\cdot \mid x^t_{1:(t-1)}, y^t_{1:(t-1)})$ being a valid probability measure for all possible values of $(x^t_{1:(t-1)}, y^t_{1:(t-1)})$.

Assumption 2 states that X^j can not adapt based on the history of any other $X^{j'}$ for all $j \neq j'$. This assumption is sufficient to satisfy Assumption 1 when testing H_0 for any X^j for any $j = 1, 2, \ldots, p$, thus leading to a valid p-value for every X^j simultaneously when using the proposed ART procedure in Algorithm 1. Although our framework gives valid p-values for each of the multiple tests, we need to further account for multiple testing issues. For example, one naïve way to control the false discovery rate is to use the Benjamini Hochberg procedure [24], but this is not the focus of our paper.

Appendix C Discussion of the Natural Adaptive Resampling Procedure

Keen readers may argue the NARP is merely a practical choice but an unnecessary one, thus no longer requiring Assumption 1. Exchangeability requires $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $(\tilde{\mathbf{X}}^b, \mathbf{Z}, \mathbf{Y})$ to be equal in distribution. Consequently, if one could sample the entire data vector $\tilde{\mathbf{X}}$ from the conditional distribution of $\mathbf{X} \mid (\mathbf{Z}, \mathbf{Y})$, then this construction of $\tilde{\mathbf{X}}$ would satisfy the required distributional equality. In general, however, it is well known that it is difficult to sample from a complicated graphical model [25]. To illustrate this, we show how constructing valid resamples $\tilde{\mathbf{X}}^b$ for even two time periods may be difficult without Assumption 1 with the following equations.

$$\begin{aligned}
&P(X_1 = x_1, X_2 = x_2 \mid Z_1 = z_1, Z_2 = z_2, Y_1 = y_1, Y_2 = y_2)\\
&= \frac{P(X_2 = x_2 \mid X_1 = x_1, Z_1 = z_1, Z_2 = z_2, Y_1 = y_1)}{\int_x P(Z_2 = z_2 \mid X_1 = x, Y_1 = y_1, Z_1 = z_1) \mathrm{d}P(X_1 = x \mid Z_1 = z_1)}\\
&\quad \cdot P(Z_2 = z_2 \mid X_1 = x_1, Y_1 = y_1, Z_1 = z_1) P(X_1 = x_1 \mid Z_1 = z_1)\\
&\propto P(X_2 = x_2 \mid X_1 = x_1, Z_1 = z_1, Z_2 = z_2, Y_1 = y_1)\\
&\quad \cdot \left[P(Z_2 = z_2 \mid X_1 = x_1, Y_1 = y_1, Z_1 = z_1) P(X_1 = x_1 \mid Z_1 = z_1)\right].
\end{aligned}$$

This follows directly from elementary probability calculations. Since any valid construction of $\tilde{\mathbf{X}}^b$ must have that $P(\tilde{X}_1 = x_1, \tilde{X}_2 = x_2 \mid Z_1 = z_1, Z_2 = z_2, Y_1 = y_1, Y_2 = y_2) = P(X_1 = x_1, X_2 = x_2 \mid Z_1 = z_1, Z_2 = z_2, Y_1 = y_1, Y_2 = y_2)$, the above equation shows that it is generally hard to construct valid resamples due to the normalizing constant in the denominator of the second line. We further note that Assumption 1 bypasses this problem because $P(Z_2 = z_2 \mid X_1 = x_1, Y_1 = y_1, Z_1 = z_1)$ is now independent of the condition $X_1 = x_1$. Therefore, the denominator in the second line is always $P(Z_2 = z_2 \mid X_1 = x_1, Y_1 = y_1, Z_1 = z_1)$, cancelling out with the numerator.

Although sampling from a distribution that is known up to a proportional constant has been extensively studied in the Markov Chain Monte Carlo (MCMC) literature [26], many MCMC methods introduce extra computational burden to an already computationally expensive algorithm that requires $B + 1$ resamples and computation of test statistic T. Moreover, it is unclear how “approximate” draws from the desired distribution in a MCMC algorithm may impact the exact validness of the p-values. This problem may be exacerbated when the sample size n is large because the errors for each resamples could exponentially accumulate across time. Therefore, we choose to use the

NARP along with Assumption 1 as the proposed method because it avoids these complications.

Appendix D Proof of Main Results Presented in Section 2

Proof of Theorem 2.1 By definition of our resampling procedure, under H_0,

$$\tilde{X}_1 \mid (Y_1, Z_1) \overset{\mathrm{d}}{=} \tilde{X}_1 \mid Z_1 \overset{\mathrm{d}}{=} X_1 \mid Z_1 \overset{\mathrm{d}}{=} X_1 \mid (Y_1, Z_1)$$

where the last "$\overset{\mathrm{d}}{=}$" is by the null hypothesis of conditional independence, namely $X_1 \perp\!\!\!\perp Y_1 \mid Z_1$. Moreover, it also suggests

$$(\tilde{X}_1, Y_1, Z_1) \overset{\mathrm{d}}{=} (X_1, Y_1, Z_1).$$

Then we will prove the following statement holds for any $k \in \{1, 2, \ldots, n\}$ by induction,

$$(\tilde{X}_{1:k}, Y_{1:k}, Z_{1:k}) \overset{\mathrm{d}}{=} (X_{1:k}, Y_{1:k}, Z_{1:k}). \tag{D3}$$

Assuming Equation D3 holds for $k-1$, we now prove it also holds for k. For simplicity, in the rest of this proof, we will use $P(\cdot)$ as a generic notation for *pdf* or *pmf*, though the proof holds for more general distributions without a *pdf* or *pmf*. First,

$$\begin{aligned}
&P\left[(\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right] \\
\overset{\text{(i)}}{=}&P\left[Z_k \mid (\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
&\cdot P\left[(\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
\overset{\text{(ii)}}{=}&P\left[Z_k \mid (Y_{1:(k-1)}, Z_{1:(k-1)}) = (y_{1:(k-1)}, z_{1:(k-1)})\right] \\
&\cdot P\left[(\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
\overset{\text{(iii)}}{=}&P\left[Z_k \mid (Y_{1:(k-1)}, Z_{1:(k-1)}) = (y_{1:(k-1)}, z_{1:(k-1)})\right] \\
&\cdot P\left[(X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
\overset{\text{(iv)}}{=}&P\left[Z_k \mid (X_{1:(k-1)} Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
&\cdot P\left[(X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
=&P\left[(X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right],
\end{aligned} \tag{D4}$$

where (i) is simply by Bayes rule; (ii) is because $Z_k \perp\!\!\!\perp \tilde{X}_{1:k-1} \mid (Y_{1:(k-1)}, Z_{1:(k-1)})$ since $\tilde{X}_{1:k-1}$ is a random function of only $Y_{1:(k-1)}$ and $Z_{1:(k-1)}$; and lastly, (iii) is by induction assumption; (iv) is by Assumption 1. Moreover,

$$\begin{aligned}
&P\left[(\tilde{X}_{1:k}, Y_{1:k}, Z_{1:k}) = (x_{1:k}, y_{1:k}, z_{1:k})\right]\\
\overset{\text{(i)}}{=}&P\left[Y_k = y_k \mid (\tilde{X}_{1:k}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:k}, y_{1:(k-1)}, z_{1:k})\right]\\
&\cdot P\left[(\tilde{X}_{1:k}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:k}, y_{1:(k-1)}, z_{1:k})\right]\\
\overset{\text{(ii)}}{=}&P\left[Y_k = y_k \mid Z_k = z_k\right] \cdot P\left[(\tilde{X}_{1:k}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:k}, y_{1:(k-1)}, z_{1:k})\right]\\
\overset{\text{(iii)}}{=}&P\left[Y_k = y_k \mid Z_k = z_k\right] \cdot P\left[\tilde{X}_k = x_k \mid (\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right]\\
&\cdot P\left[(\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right]\\
\overset{\text{(iv)}}{=}&P\left[Y_k = y_k \mid Z_k = z_k\right] \cdot P\left[X_k = x_k \mid (X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right]\\
&\cdot P\left[(\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right]\\
\overset{\text{(v)}}{=}&P\left[Y_k = y_k \mid Z_k = z_k\right] \cdot P\left[X_k = x_k \mid (X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right]\\
&\cdot P\left[(X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right]\\
=&P\left[(X_{1:k}, Y_{1:k}, Z_{1:k}) = (x_{1:k}, y_{1:k}, z_{1:k})\right],
\end{aligned}$$

where (i) is again simply by Bayes rule; (ii) is because Y_k is a random function of only Z_k (up to time k) under the null H_0 and thus is independent of anything with index smaller or equal to k conditioning on Z_k; (iii) is again by Bayes rule; (iv) is by Definition 2.2; and finally (v) is by the previous equation above. Equation D3 is thus established by induction, as a corollary of which, we also get for any $k \le n$,

$$\tilde{X}_{1:n} \mid (Y_{1:n}, Z_{1:n}) \overset{\text{d}}{=} X_{1:n} \mid (Y_{1:n}, Z_{1:n})$$

Finally, note that $\tilde{X} \perp\!\!\!\perp X \mid (Y, Z)$. So, conditioning on (Y, Z), $\tilde{X}$ and X are exchangeable, which means the p-value defined in Equation 5 is conditionally valid, conditioning on (Y, Z). Since $\mathbb{P}\left(p < \alpha \mid Y, Z\right) \le \alpha$ holds conditionally, it also holds marginally. □

Proof of Theorem 2.2 Note that Assumption 1 was only utilized once in the proof of Theorem 2.1, namely (iv) of Equation D4. So upon assuming $(\tilde{X}_{1:k}, Y_{1:k}, Z_{1:k}) \overset{\text{d}}{=} (X_{1:k}, Y_{1:k}, Z_{1:k})$, we know immediately from Equation D4 that

$$\begin{aligned}
&P\left[Z_k = z_k \mid (Y_{1:(k-1)}, Z_{1:(k-1)}) = (y_{1:(k-1)}, z_{1:(k-1)})\right]\\
=&P\left[Z_k = z_k \mid (X_{1:(k-1)} Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right]
\end{aligned}$$

which is exactly Assumption 1. □

Appendix E Proof of Results Presented in Appendix A

Before proving the main power results, we first state a self-explanatory lemma concerning the effect of taking B to go to infinity, which justifies assuming

B to be large enough and ignoring the effect of discrete p-values like the one defined in Equation 5. Similar proof arguments are made in [27], thus we omit the proof of this lemma. The lemma states that as $B \to \infty$, conditioning on any given values of $(X, \mathbf{Y}, \mathbf{Z})$,

$$\begin{aligned} p\text{-value} &:= \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbf{1}_{\{T(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y}) \geq T(\mathbf{X}, \mathbf{Z}, \mathbf{Y})\}}\right] \\ &\xrightarrow{\text{a.s.}} \mathbb{P}\left(T(\tilde{\mathbf{X}}^{\mathbf{b}}, \mathbf{Z}, \mathbf{Y}) \geq T(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \mid \mathbf{Y}, \mathbf{Z}\right). \end{aligned}$$

Lemma E.1 (Power of ART under $B \to \infty$)**.** For any adaptive sapling procedure A satisfies Definition 2.1 and any test statistic T, as we take $B \to \infty$, the asymptotic conditional power of ART (with CRT being an degenerate special case) condition on (Y, Z) is equal to

$$\mathbb{P}\left(T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \geq z_{1-\alpha}(T(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z})) \mid \mathbf{Y}, \mathbf{Z}\right),$$

while the unconditional (marginal) power is equal to

$$\mathbb{P}_{\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}}\left(\mathbb{P}\left(T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \geq z_{1-\alpha}(T(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z})) \mid \mathbf{Y}, \mathbf{Z}\right)\right).$$

Note that the joint distribution of $(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z})$ is implicitly specified by the sampling procedure A.

Lemma E.2 (Normal Means Model with *iid* sampling procedures: Joint Asymptotic Distributions of $\bar{Y}_j$'s, $\bar{\tilde{Y}}_j$'s and $\bar{Y}$ Under the Alternative H_1)**.** Define

$$T_{\text{all}} = \left(\bar{\tilde{Y}}_1, \bar{\tilde{Y}}_2, \ldots, \bar{\tilde{Y}}_{p-1}, \bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_{p-1}, \bar{Y}\right)^T \in \mathbb{R}^{2p-1}.$$

Upon assuming the normal means model introduced in Section 3, under the alternative H_1 with $h = h_0/\sqrt{n}$, as $n \to \infty$,

$$\sqrt{n} \cdot T_{\text{all}} \xrightarrow{\text{d}} T_{\text{all}}^{\infty},$$

with

$$T_{\text{all}}^{\infty} = \begin{pmatrix} G_1 + R + h_0 q_1 \\ G_2 + R + h_0 q_1 \\ \cdots \\ G_{p-1} + R + h_0 q_1 \\ H_1 + R + h_0 \\ H_2 + R \\ \cdots \\ H_{p-1} + R \\ R \end{pmatrix} \in \mathbb{R}^{2p-1},$$

where $G := (G_1, G_2, \ldots, G_{p-1})$ and $H := (H_1, H_2, \ldots, H_{p-1})$ both follow the same $(p-1)$ dimensional multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ and R is a standard normal random variable. Note that Σ was defined in the statement of Theorem A.1. Moreover, G, H and R are independent.

Remark 1. Roughly speaking, after removing means, R captures the randomness of $\mathbf{Y}$ being sampled from its marginal distribution; H captures the randomness of sampling $\mathbf{X}$ conditioning on $\mathbf{Y}$; lastly, G captures the randomness of resampling $\tilde{\mathbf{X}}$ given $\mathbf{Y}$.

Remark 2. We also note that we do not include characterizing the distribution of $\tilde{\bar{Y}}_p$ or $\bar{Y}_p$ to avoid stating the convergence in terms of a degenerate multivariate Gaussian distribution since $\bar{Y}_p$ is a deterministic function given $\bar{Y}$ and the remaining $p-1$ means of the other arms.

Proof of Lemma E.2 We first characterize the conditional distribution of $\tilde{\bar{Y}}_j$. For any $j \in \{1, 2, \ldots, p\}$,

$$\begin{aligned}\tilde{\bar{Y}}_j &:= \frac{\sum_{i=1}^n Y_i \mathbf{1}_{\tilde{X}_i = j}}{\sum_{i=1}^n \mathbf{1}_{\tilde{X}_i = j}} \\ &= \frac{1}{\sqrt{n}} \left[\frac{1}{q_j} \frac{\sum_{i=1}^n Y_i \left(\mathbf{1}_{\tilde{X}_i = j} - q_j \right)}{\sqrt{n}} + \frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \right] \frac{q_j n}{\sum_{i=1}^n \mathbf{1}_{\tilde{X}_i = j}}.\end{aligned}$$

By Central Limit Theorem, since $\mathrm{Var}\left(Y_i(\mathbf{1}_{\tilde{X}_i = j} - q_j) \right) \to q_j(1 - q_j)$ as $n \to \infty$,

$$\frac{\sum_{i=1}^n Y_i \left(\mathbf{1}_{\tilde{X}_i = j} - q_j \right)}{\sqrt{q_j(1 - q_j)n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1),$$

which together with Slutsky's Theorem and the fact that $q_j n / \sum_{i=1}^n \mathbf{1}_{\tilde{X}_i = j} \to 1$ almost surely gives,

$$J_{j,n} := \sqrt{n} \tilde{\bar{Y}}_j - \frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}\left(0, \frac{v(q_j)}{q_j^2} \right),$$

where $v(q_j) = \mathrm{Var}(\mathrm{Bern(q_j)}) = \mathrm{Var}(\mathbf{1}_{\tilde{X}_j = 1}) = q_j(1 - q_j)$. Additional to these one dimensional asymptotic results, we can also derive their joint asymptotic distribution. Before moving forward, we define a few useful notations,

$$\mathbf{J}_{-p,n} := (J_{1,n}, J_{2,n}, \ldots, J_{p-1,n}) \in \mathbb{R}^{p-1},$$

$$V_i := \left(Y_i(\mathbf{1}_{\tilde{X}_i = 1} - q_1), Y_i(\mathbf{1}_{\tilde{X}_i = 2} - q_2), \cdots, Y_i(\mathbf{1}_{\tilde{X}_i = p-1} - q_{p-1}) \right) \in \mathbb{R}^{p-1},$$

$$\bar{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n \mathrm{Var}(V_i),$$

and

$$\Sigma_0 := \mathrm{Var}\left(\left(\mathbf{1}_{\tilde{X}_i = 1}, \mathbf{1}_{\tilde{X}_i = 2}, \cdots, \mathbf{1}_{\tilde{X}_i = p-1} \right) \right) = \begin{bmatrix} v(q_1) & -q_1 q_2 & -q_1 q_3 & \cdots & -q_1 q_{p-1} \\ -q_1 q_2 & v(q_2) & -q_2 q_3 & \cdots & -q_2 q_{p-1} \\ -q_1 q_3 & -q_2 q_3 & v(q_3) & \cdots & -q_3 q_{p-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -q_1 q_{p-1} & -q_2 q_{p-1} & -q_3 q_{p-1} & \cdots & v(q_{p-1}) \end{bmatrix}. \tag{E5}$$

By Multivariate Lindeberg-Feller CLT (see for instance [28]),

$$\sqrt{n}\bar{\Sigma}_n^{-1/2}\left(\bar{V}-\mathbb{E}\bar{V}\right)\xrightarrow{\mathrm{d}}\mathcal{N}\left(0,I_{p-1}\right). \tag{E6}$$

which further gives

$$\sqrt{n}\left(\bar{V}-\mathbb{E}\bar{V}\right)\xrightarrow{\mathrm{d}}\mathcal{N}\left(0,\Sigma_0\right)$$

because of

$$\lim_{n\to\infty}\bar{\Sigma}_n=\Sigma_0.$$

Therefore we have

$$\mathbf{J}_{-p,n}\xrightarrow{\mathrm{d}}\mathcal{N}\left(0,\Sigma\right), \tag{E7}$$

where

$$\Sigma=D^{-1}\Sigma_0D^{-1}$$

with

$$D=\operatorname{diag}(q_1,q_2,\cdots,q_{p-1})\in\mathbb{R}^{(p-1)\times(p-1)}. \tag{E8}$$

Roughly speaking, this suggests that after removing the shared randomness induced by $\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$, all the $\sqrt{n}\bar{\tilde{Y}}_j$'s are asymptotically independent and Gaussian distributed.

Next, we turn to $\bar{Y}_j$. Note that in this part we will view X_i as generated from $F_{X|Y}$ after the generation of Y_i according to its marginal distribution. The only difference in the observed test statistic and the above is that we have

$$X_i \mid Y_i\sim\mathcal{M}(q_i^\star)$$

with $q_i^\star=(q_{i,1}^\star,q_{i,2}^\star,\cdots,q_{i,p}^\star)$ and

$$q_{i,j}^\star=\frac{q_j\mathcal{N}\left(Y_i;\ \frac{h_0}{\sqrt{n}}\mathbf{1}_{j=1},1\right)}{\sum_{k=1}^p q_k\mathcal{N}\left(Y_i;\ \frac{h_0}{\sqrt{n}}\mathbf{1}_{k=1},1\right)}=\frac{\exp\left[-\frac{1}{2}\left(Y_i-\frac{h_0}{\sqrt{n}}\mathbf{1}_{j=1}\right)^2\right]}{\sum_{k=1}^p q_k\exp\left[-\frac{1}{2}\left(Y_i-\frac{h_0}{\sqrt{n}}\mathbf{1}_{k=1}\right)^2\right]}$$

instead. Again, Multivariate Lindeberg-Feller CLT gives,

$$\sqrt{n}(\bar{\Sigma}_n^\star)^{-1/2}\left(\bar{V}^\star-\mathbb{E}\bar{V}^\star\right)\xrightarrow{\mathrm{d}}\mathcal{N}\left(0,I_{p-1}\right), \tag{E9}$$

with

$$V_i^\star:=\left(Y_i(\mathbf{1}_{X_i=1}-q_{i,1}^\star),Y_i(\mathbf{1}_{X_i=2}-q_{i,2}^\star),\cdots,Y_i(\mathbf{1}_{X_i=p-1}-q_{i,p-1}^\star)\right)\in\mathbb{R}^{p-1},$$

$$\bar{\Sigma}_n^\star=\frac{1}{n}\sum_{i=1}^n\operatorname{Var}\left(V_i^\star\right).$$

Note that, since

$$\lim_{n\to\infty}\operatorname{Var}\left(Y_i(\mathbf{1}_{X_i=j}-q_{i,j}^\star)\right)=q_j(1-q_j)$$

and

$$\lim_{n\to\infty}\operatorname{Cov}\left(Y_i(\mathbf{1}_{X_i=j_1}-q_{i,j_1}^\star),Y_i(\mathbf{1}_{X_i=j_2}-q_{i,j_2}^\star)\right)=-q_{j_1}q_{j_2},$$

we have

$$\lim_{n\to\infty}\bar{\Sigma}_n^\star=\Sigma_0,$$

which further gives

$$\sqrt{n}\left(\bar{V}^\star-\mathbb{E}\bar{V}^\star\right)\xrightarrow{\mathrm{d}}\mathcal{N}\left(0,\Sigma_0\right). \tag{E10}$$

Similar to $\mathbf{J}$'s, we define $\mathbf{J}^\star$'s as well,

$$J_{j,n}^\star:=\sqrt{n}\bar{Y}_j-\frac{\sum_{i=1}^n q_{i,j}^\star Y_i}{q_j\sqrt{n}}=\frac{\sum_{i=1}^n Y_i\mathbf{1}_{X_i=j}}{q_j\sqrt{n}}-\frac{\sum_{i=1}^n q_{i,j}^\star Y_i}{q_j\sqrt{n}}+o_p(1)=\frac{\sqrt{n}\left(\bar{V}^\star\right)_j}{q_j}+o_p(1).$$

and

$$\mathbf{J}^{\star}_{-p,n} := (J^{\star}_{1,n}, J^{\star}_{2,n}, \ldots, J^{\star}_{p-1,n}) \in \mathbb{R}^{p-1},$$

which together with Equation E10 gives

$$\mathbf{J}^{\star}_{-p,n} \xrightarrow{\text{d}} \mathcal{N}(0, \Sigma)\,. \tag{E11}$$

Note that though Equation E7 and Equation E11 are almost exactly the same, it does not suggest $\bar{Y}_j$'s and $\tilde{\bar{Y}}_j$'s have the same asymptotic distribution, since the "mean" parts that have been removed actually behave differently, namely $\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$ and $\frac{\sum_{i=1}^n q^{\star}_{i,j} Y_i}{q_j\sqrt{n}}$, as demonstrated in Lemma E.3, Lemma E.4, Lemma E.5 and Lemma E.6. Roughly speaking, under this $\sqrt{n}$ scaling, the randomness that leads to the Gaussian noise part in CLT is the same across them as demonstrated in Equation E7 and Equation E11, but the Gaussian distribution they are converging to have different means.

Finally, following exactly the same logic, we can further derive the following joint asymptotic distribution of $\mathbf{J}_{-p,n}$, $\mathbf{J}^{\star}_{-p,n}$ and $\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$. Letting

$$\mathbf{J}_{\text{ALL}} := \left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}, \mathbf{J}_{-p,n}, \mathbf{J}^{\star}_{-p,n}\right) \in \mathbb{R}^{2p-1},$$

we have

$$\mathbf{J}_{\text{ALL}} \xrightarrow{\text{d}} \mathcal{N}(0, \Sigma_{\text{ALL}}) := \mathcal{N}\left(0, \begin{bmatrix} 1 & 0 & 0 \\ 0 & \Sigma & 0 \\ 0 & 0 & \Sigma \end{bmatrix}\right).$$

□

Lemma E.3. As $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^n Y_i^2 \xrightarrow{\text{a.s.}} 1 \quad \text{and} \quad \frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(h_0 q_1, 1).$$

Proof By defining $E_i := S_i W_i + (1 - S_i) G_i \sim \mathcal{N}(0, 1)$, we have

$$Y_i = E_i + \frac{S_i h_0}{\sqrt{n}}$$

Note that E_i and S_i are not independent. Thus,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n Y_i^2 &= \frac{1}{n}\sum_{i=1}^n \left(E_i + \frac{S_i h_0}{\sqrt{n}}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n E_i^2 + \frac{1}{n^2}\sum_{i=1}^n S_i h_0 + \frac{1}{n^{3/2}}\sum_{i=1}^n 2h_0 E_i S_i \\ &\xrightarrow{\text{a.s.}} 1, \end{aligned}$$

since by Law of Large Numbers the last two terms will vanish asymptotically and the first term will converge to $\mathbb{E}(E_i^2) = 1$. Moreover,

$$\begin{aligned} \frac{\sum_{i=1}^n Y_i}{\sqrt{n}} &= \frac{\sum_{i=1}^n E_i}{\sqrt{n}} + h_0 \frac{\sum_{i=1}^n S_i}{n} \\ &\xrightarrow{\text{d}} \mathcal{N}(h_0 q_1, 1)\,, \end{aligned}$$

where the last line is obtained by applying CLT to the first term and LLN to the second term. □

Lemma E.4. As $n \to \infty$,

$$\frac{\sum_{i=1}^{n} q_{i,1}^{\star} Y_i}{q_1\sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}\left(q_1 h_0, 1\right).$$

Proof We first show

$$\lim_{n\to\infty} \mathbb{E}\left(\sqrt{n} q_{i,1}^{\star} Y_i\right) = h_0. \tag{E12}$$

Recall that Y_i can be seen as a mixture of two normal distributions $\mathcal{N}(0,1)$ and $\mathcal{N}\left(\frac{h_0}{\sqrt{n}}, 1\right)$ with weights $1-q_1$ and q_1. Thus $\mathbb{E}\left(\sqrt{n} q_{i,1}^{\star} Y_i\right)$ is equal to

$$\sqrt{n}\int_{\mathbb{R}} \frac{y q_1 e^{-(y-h_0/\sqrt{n})^2/2}}{q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1)e^{-y^2/2}} \left[(1-q_1)\frac{1}{\sqrt{2\pi}}e^{-y^2/2} + q_1\frac{1}{\sqrt{2\pi}}e^{-(y-h_0/\sqrt{n})^2/2}\right] \mathrm{d}y$$
$$:= A_0 + A_1.$$

Note that with a change of variable $h = h_0/\sqrt{n}$,

$$\begin{aligned}
\lim_{n\to\infty} A_1 &= \frac{q_1^2\sqrt{n}}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{y e^{-(y-h_0/\sqrt{n})^2/2}}{q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h_0/\sqrt{n})^2/2}\mathrm{d}y \\
&= \lim_{h\to 0} \frac{q_1^2 h_0}{\sqrt{2\pi}} \left[\frac{1}{h}\int_{\mathbb{R}} \frac{y e^{-(y-h)^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2}\mathrm{d}y\right] \\
&= \frac{q_1^2 h_0}{\sqrt{2\pi}} \left.\frac{\mathrm{d}\left[\int_{\mathbb{R}} \frac{y e^{-(y-h)^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2}\mathrm{d}y\right]}{\mathrm{d}h}\right|_{h=0} \\
&= \frac{q_1^2 h_0}{\sqrt{2\pi}} \int_{\mathbb{R}} \left.\frac{\mathrm{d}\left[\frac{y e^{-(y-h)^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2}\right]}{\mathrm{d}h}\right|_{h=0} \mathrm{d}y \\
&= \frac{q_1^2 h_0}{\sqrt{2\pi}} \int_{\mathbb{R}} (2-q_1) y^2 e^{-y^2/2}\mathrm{d}y \\
&= h_0 q_1^2 (2-q_1).
\end{aligned}$$

Similarly,

$$\lim_{n\to\infty} A_0 = h_0 q_1 (1-q_1)^2.$$

Equation E12 is thereby established. Then we compute $\lim_{n\to\infty} \mathrm{Var}(q^\star_{i,1} Y_i)$ using the same strategy.

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{Var}(q^\star_{i,1} Y_i) \\
&= \lim_{n\to\infty} \left\{ \mathbb{E}\left[(q^\star_{i,1} Y_i)^2 \right] - \left[\mathbb{E}(q^\star_{i,1} Y_i) \right]^2 \right\} \\
&= \lim_{n\to\infty} \mathbb{E}\left[(q^\star_{i,1} Y_i)^2 \right] \\
&= \lim_{n\to\infty} \int_{\mathbb{R}} y^2 \left[\frac{q_1 e^{-(y-h_0/\sqrt{n})^2/2}}{q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1) e^{-y^2/2}} \right]^2 \\
&\quad \cdot \left[(1-q_1) \frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h_0/\sqrt{n})^2/2} \right] \mathrm{d}y \qquad \text{(E13)} \\
&= \int_{\mathbb{R}} \lim_{h\to 0} \left\{ y^2 \left[\frac{q_1 e^{-(y-h)^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1) e^{-y^2/2}} \right]^2 \right. \\
&\quad \left. \cdot \left[(1-q_1) \frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h)^2/2} \right] \right\} \mathrm{d}y \\
&= \int_{\mathbb{R}} q_1^2 \frac{e^{-y^2/2}}{\sqrt{2\pi}} \mathrm{d}y \\
&= q_1^2.
\end{aligned}$$

Combining Equation E12 and Equation E13, the lemma is thus established by Central Limit Theorem. □

Following exactly the same logic, we have the following parallel lemma for $j \neq 1$.

Lemma E.5. For $j \neq 1$, as $n \to \infty$,

$$\frac{\sum_{i=1}^n q^\star_{i,j} Y_i}{q_j \sqrt{n}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

Proof We first show

$$\lim_{n\to\infty} \mathbb{E}\left(\sqrt{n} q^\star_{i,j} Y_i \right) = 0.$$

Again, recall that Y_i can be seen as a mixture of two normal distributions $\mathcal{N}(0,1)$ and $\mathcal{N}\left(\frac{h_0}{\sqrt{n}}, 1\right)$ with weights $1-q_1$ and q_1. Thus $\mathbb{E}\left(\sqrt{n} q^\star_{i,j} Y_i\right)$ is equal to

$$\sqrt{n} \int_{\mathbb{R}} \frac{y q_j e^{-y^2/2}}{q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1) e^{-y^2/2}} \left[(1-q_1) \frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h_0/\sqrt{n})^2/2} \right] \mathrm{d}y$$
$$:= B_0 + B_1.$$

With a change of variable $h = h_0/\sqrt{n}$, we have

$$
\begin{aligned}
\lim_{n\to\infty} B_1 &= \frac{q_1 q_j \sqrt{n}}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{y e^{-y^2/2}}{q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h_0/\sqrt{n})^2/2} \mathrm{d}y \\
&= \lim_{h\to 0} \frac{q_1 q_j h_0}{\sqrt{2\pi}} \left[\frac{1}{h} \int_{\mathbb{R}} \frac{y e^{-y^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2} \mathrm{d}y \right] \\
&= \frac{q_1 q_j h_0}{\sqrt{2\pi}} \left. \frac{\mathrm{d}\left[\int_{\mathbb{R}} \frac{y e^{-y^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2} \mathrm{d}y \right]}{\mathrm{d}h} \right|_{h=0} \\
&= \frac{q_1 q_j h_0}{\sqrt{2\pi}} \int_{\mathbb{R}} \left. \frac{\mathrm{d}\left[\frac{y e^{-y^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2} \right]}{\mathrm{d}h} \right|_{h=0} \mathrm{d}y \\
&= \frac{q_1 q_j h_0}{\sqrt{2\pi}} \int_{\mathbb{R}} (1-q_1) y^2 e^{-y^2/2} \mathrm{d}y \\
&= h_0 q_j q_1 (1-q_1).
\end{aligned}
$$

Similarly,

$$\lim_{n\to\infty} B_0 = -h_0 q_j q_1 (1-q_1).$$

Finally, we have $\lim_{n\to\infty} \mathrm{Var}(q^{\star}_{i,1} Y_i) = q_j^2$ as well, which by CLT finishes the proof. □

We can further write down their asymptotic joint distribution. We note that $q^{\star}_{i,j} = \frac{q_j}{q_2} q^{\star}_{i,2}$ deterministically for $j > 2$, thus it suffices to only include $j = 1, 2$ in the joint asymptotic distribution.

Lemma E.6. As $n \to \infty$,

$$\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}, \frac{\sum_{i=1}^n q^{\star}_{i,1} Y_i}{q_1 \sqrt{n}}, \frac{\sum_{i=1}^n q^{\star}_{i,2} Y_i}{q_2 \sqrt{n}} \right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mu_3, \Sigma_3\right),$$

where

$$\mu_3 = \left(h_0 q_1, h_0 q_1 (2-q_1), h_0 q_1 (1-q_1)\right)^T \in \mathbb{R}^3,$$

and $\Sigma_3 \in \mathbb{R}^{3\times 3}$ is equal to

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

In other words, asymptotically these three random variables are completely linearly correlated.

Proof By Lemma E.4, it suffices to show

$$\begin{aligned}\lim_{n\to\infty}\operatorname{Cor}\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}},\frac{\sum_{i=1}^n q_{i,1}^\star Y_i}{q_1\sqrt{n}}\right)&=\lim_{n\to\infty}\operatorname{Cor}\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}},\frac{\sum_{i=1}^n q_{i,2}^\star Y_i}{q_2\sqrt{n}}\right)\\&=\lim_{n\to\infty}\operatorname{Cor}\left(\frac{\sum_{i=1}^n q_{i,1}^\star Y_i}{q_1\sqrt{n}},\frac{\sum_{i=1}^n q_{i,2}^\star Y_i}{q_2\sqrt{n}}\right)=1,\end{aligned}$$

which can be established by the following three displays. First,

$$\begin{aligned}&\lim_{n\to\infty}\operatorname{Cov}\left(Y_i,q_{i,1}^\star Y_i\right)\\&=\lim_{n\to\infty}\mathbb{E}\left(Y_i\cdot q_{i,1}^\star Y_i\right)\\&=\lim_{n\to\infty}\int_{\mathbb{R}}\left\{\frac{y^2q_1e^{-(y-h_0/\sqrt{n})^2/2}}{q_1e^{-(y-h_0/\sqrt{n})^2/2}+(1-q_1)e^{-y^2/2}}\right.\\&\quad\left.\left[(1-q_1)\frac{1}{\sqrt{2\pi}}e^{-y^2/2}+q_1\frac{1}{\sqrt{2\pi}}e^{-(y-h_0/\sqrt{n})^2/2}\right]\right\}\mathrm{d}y\\&=\lim_{h\to0}\int_{\mathbb{R}}\frac{y^2q_1e^{-(y-h)^2/2}}{q_1e^{-(y-h)^2/2}+(1-q_1)e^{-y^2/2}}\left[(1-q_1)\frac{1}{\sqrt{2\pi}}e^{-y^2/2}+q_1\frac{1}{\sqrt{2\pi}}e^{-(y-h)^2/2}\right]\mathrm{d}y\\&=\int_{\mathbb{R}}\lim_{h\to0}\left\{\frac{y^2q_1e^{-(y-h)^2/2}}{q_1e^{-(y-h)^2/2}+(1-q_1)e^{-y^2/2}}\left[(1-q_1)\frac{1}{\sqrt{2\pi}}e^{-y^2/2}+q_1\frac{1}{\sqrt{2\pi}}e^{-(y-h)^2/2}\right]\right\}\mathrm{d}y\\&=q_1\int_{\mathbb{R}}y^2\frac{e^{-y^2/2}}{\sqrt{2\pi}}\mathrm{d}y\\&=q_1;\end{aligned}$$

secondly,

$$\begin{aligned}&\lim_{n\to\infty}\operatorname{Cov}\left(Y_i,q_{i,2}^\star Y_i\right)\\&=\lim_{n\to\infty}\mathbb{E}\left(Y_i\cdot q_{i,2}^\star Y_i\right)\\&=\lim_{n\to\infty}\int_{\mathbb{R}}\left\{\frac{y^2q_2e^{-y^2/2}}{q_1e^{-(y-h_0/\sqrt{n})^2/2}+(1-q_1)e^{-y^2/2}}\right.\\&\quad\left.\left[(1-q_1)\frac{1}{\sqrt{2\pi}}e^{-y^2/2}+q_1\frac{1}{\sqrt{2\pi}}e^{-(y-h_0/\sqrt{n})^2/2}\right]\right\}\mathrm{d}y\\&=\lim_{h\to0}\int_{\mathbb{R}}\frac{y^2q_2e^{-y^2/2}}{q_1e^{-(y-h)^2/2}+(1-q_1)e^{-y^2/2}}\left[(1-q_1)\frac{1}{\sqrt{2\pi}}e^{-y^2/2}+q_1\frac{1}{\sqrt{2\pi}}e^{-(y-h)^2/2}\right]\mathrm{d}y\\&=\int_{\mathbb{R}}\lim_{h\to0}\left\{\frac{y^2q_2e^{-y^2/2}}{q_1e^{-(y-h)^2/2}+(1-q_1)e^{-y^2/2}}\left[(1-q_1)\frac{1}{\sqrt{2\pi}}e^{-y^2/2}+q_1\frac{1}{\sqrt{2\pi}}e^{-(y-h)^2/2}\right]\right\}\mathrm{d}y\\&=q_2\int_{\mathbb{R}}y^2\frac{e^{-y^2/2}}{\sqrt{2\pi}}\mathrm{d}y\\&=q_2;\end{aligned}$$

and finally

$$
\begin{aligned}
&\lim_{n\to\infty} \operatorname{Cov}\left(q^{\star}_{i,1} Y_i, q^{\star}_{i,2} Y_i\right) \\
&= \lim_{n\to\infty} \mathbb{E}\left(q^{\star}_{i,1} q^{\star}_{i,2} Y_i^2\right) \\
&= \lim_{n\to\infty} \int_{\mathbb{R}} \Bigg\{ \frac{y^2 q_1 q_2 e^{-(y-h_0/\sqrt{n})^2/2} e^{-y^2/2}}{\left[q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1) e^{-y^2/2}\right]^2} \\
&\quad \left[(1-q_1)\frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h_0/\sqrt{n})^2/2}\right] \Bigg\} \mathrm{d}y \\
&= \lim_{h\to 0} \int_{\mathbb{R}} \frac{y^2 q_1 q_2 e^{-(y-h)^2/2} e^{-y^2/2}}{\left[q_1 e^{-(y-h)^2/2} + (1-q_1) e^{-y^2/2}\right]^2} \left[(1-q_1)\frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h)^2/2}\right] \mathrm{d}y \\
&= \int_{\mathbb{R}} \lim_{h\to 0} \left\{ \frac{y^2 q_1 q_2 e^{-(y-h)^2/2} e^{-y^2/2}}{\left[q_1 e^{-(y-h)^2/2} + (1-q_1) e^{-y^2/2}\right]^2} \left[(1-q_1)\frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h)^2/2}\right] \right\} \mathrm{d}y \\
&= q_1 q_2 \int_{\mathbb{R}} y^2 \frac{e^{-y^2/2}}{\sqrt{2\pi}} \mathrm{d}y \\
&= q_1 q_2.
\end{aligned}
$$

□

Appendix F Additional Simulations in Normal-means Model

To show that our results presented in Section 3.3 are not sensitive to the initially chosen adaptive parameters and to also further optimize for multiple adaptive procedures A as shown in Algorithm 1, we create Figure F1. Figure F1 shows the power of the ART using different combinations of the adaptive parameters, ϵ and reweighting value t_0, in three different scenarios of p and h_0.

Figure F1 shows that an adaptive procedure with exploration parameter $\epsilon = 0.7$ seems to be a favorable choice across different signal strengths. Additionally, we find that the optimal reweighting parameter t can be different across different scenarios but does not seem to matter largely across the different scenarios. We find that our initially chosen parameter of $\epsilon = 0.5$ in Section 3.3 was not necessarily the most optimal choice, demonstrating the robustness of the results presented in Section 3.3

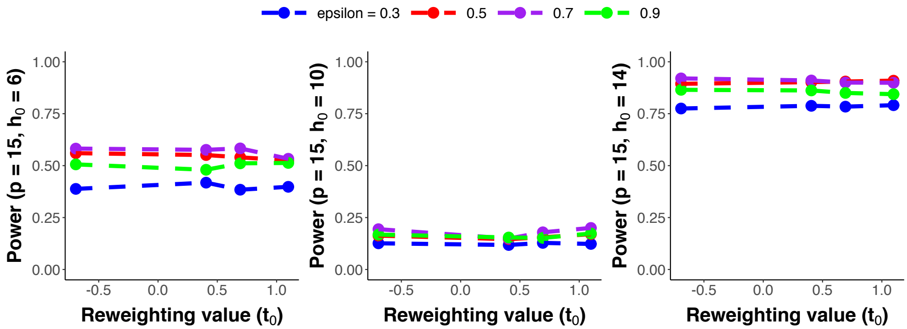

Fig. F1 Each panel showcases the power for different exploration parameter ϵ across different reweighting parameter t_0, where $t = t_0/h_0$. The panels differ by different signal strengths $h_0 = 6, 10, 14$ while the number of arms are fixed at $p = 15$.

Appendix G Details of ART in Conjoint Analysis

In this section we first give further details of the ART used in Section 4 along with some additional simulation results.

G.1 Simulation Setup

For our simulation setup, (X, Z) each contain one factor with four levels, i.e., $X_t^L, X_t^R, Z_t^L, Z_t^R$ take values $1, 2, 3, 4$. The response model follows a logistic regression model with main effects and interactions on only one specific combination,

$$\begin{aligned}\Pr(Y_t = 1 \mid X_t, Z_t) = \text{logit}^{-1}\Big[&\beta_X \mathbf{1}\{X_t^L = 1, X_t^R \neq 1\} - \beta_X \mathbf{1}\{X_t^L \neq 1, X_t^R = 1\}\\ &+ \beta_Z \mathbf{1}\{Z_t^L = 1, Z_t^R \neq 1\} - \beta_Z \mathbf{1}\{Z_t^L \neq 1, Z_t^R = 1\}\\ &+ \beta_{XZ} \mathbf{1}\{X_t^L = 1, Z_t^L = 2, X_t^R \neq 1, Z_t^R \neq 2\}\\ &- \beta_{XZ} \mathbf{1}\{X_t^L \neq 1, Z_t^L \neq 2, X_t^R = 1, Z_t^R = 2\}\Big],\end{aligned}$$

where the first four indicators force main effects β_X, β_Z of X and Z, respectively, on the first levels of each factor and the last two indicators force an interaction effect β_{XZ} between the first and second level of factors X and Z. For example, $\mathbf{1}\{X_t^L = 1, Z_t^L = 2, X_t^R \neq 1, Z_t^R \neq 2\}$ is one if the left profile values of (X, Z) are $(1, 2)$, respectively, but the right profile values of (X, Z) are not $(1, 2)$ simultaneously. We note that the interaction indicator is still one if $(X_t^L, Z_t^L) = (1, 2)$ and $(X_t^R, Z_t^R) = (1, 3)$ as long as both (X_t^L, Z_t^L) and (X_t^R, Z_t^R) are not $(1, 2)$ simultaneously. For the left plot of Figure G2, $\beta_X = \beta_Z = 0.6$ while $\beta_{XZ} = 0.9$ while we vary the sample size in the x-axis. For the right plot of Figure G2, the interaction $\beta_{XZ} = 0$ while we vary

$\beta_X = \beta_Z = (0, 0.3, 0.6, 0.9, 1.2)$ in the x-axis with a fixed sample size of $n = 1,000$. Lastly, our response model assumes "no profile order effect" since all main and interaction effects are repeated symmetrically for the right and left profile (except we shift the sign because $Y = 1$ refers to the left profile being selected).

G.2 Adaptive Procedure and Test Statistic

We first give a detailed description of our adaptive procedure then formally define the test statistic used in Section 4.

We define $X_t \sim \text{Multinomial}(p^X_{t,1}, p^X_{t,2}, \ldots, p^X_{t,K^2})$, where $p^X_{t,j}$ represents the probability of sampling the jth arm (arm refers to each unique combination of left and right factor levels) out of K^2 possible arms and K is the total levels of X. For example, in our simulation setup $K = 4$ and there are 16 possible arms, $(1, 1), (1, 2)$, etc., and $p^Z_{t,j}$ is defined similarly. The uniform *iid* sampling procedure pulls each arm with equal probability, i.e., $p^X_{t,j} = \frac{1}{K^2}, p^Z_{t,j} = \frac{1}{L^2}$ for every j and L is the total number of factor levels for factor Z.[4] Although we present our adaptive procedure when Z contains only one other factor (typical conjoint analysis have 8-10 other factors), our adaptive procedure loses no generality in higher dimensions of Z.

We now propose the following adaptive procedure that adapts the sampling weights of $p^X_{t,j}, p^Z_{t,j}$ at each time step t in the following way,

$$p^X_{t,j} \propto |\bar{Y}^X_{j,t} - 0.5| + |N(0, 0.01^2)|, \qquad p^Z_{t,j} \propto |\bar{Y}^Z_{j,t} - 0.5| + |N(0, 0.01^2)|, \quad \text{(G14)}$$

where $\bar{Y}^X_{j,t}$ denotes the sample mean of $Y_1, Y_2, \ldots, Y_{t-1}$ for arm j in variable X, $\bar{Y}^Z_{j,t}$ is defined similarly, and $N(0, 0.01^2)$ denotes a Gaussian random variable with mean zero and variance 0.01^2 (the two Gaussians in Equation (G14) are drawn independently). Equation (G14) samples more from arms that look like signal (further away from 0.5). We add a slight perturbation in case $\bar{Y}^X_{j,t}$ is exactly equal to 0.5 at any time point t to discourage an arm from having zero probability to be sampled.

With this reweighting procedure, we build our adaptive procedure. Just like Definition 3.2, we also have an ϵ adaptive parameter that denotes the beginning $[n\epsilon]$ samples that are used for "exploration" by using the typical uniform *iid* sampling procedure. In the remaining samples, we adapt by changing the weights according to Equation (G14). This adaptive sampling procedure immediately satisfies Assumption 1 and also Assumption 2 since each variable only looks at its own history and previous responses. Algorithm 2 summarizes the adaptive procedure.

We now give the test statistic under consideration. Although the authors of [2] consider a complex Hierarchical Lasso model to capture all second-order interactions, we consider a simple cross-validated Lasso logistic test statistic

[4] We also note that conjoint applications do indeed default to the uniform *iid* sampling procedure (or a very minor variant from it) [4, 29].

Algorithm 2 Adaptive Procedure for Conjoint Studies

```
Given adaptive parameter $\epsilon$
for $t = 1$ to $[n\epsilon]$ do
    Sample $X_t \sim$ Multinomial$(p^X_{t,1}, p^X_{t,2}, \dots, p^X_{t,K^2})$, where $p^X_{t,j} = \frac{1}{K}$ for all $j = 1, 2, \dots, K^2$
    Sample $Z_t \sim$ Multinomial$(p^Z_{t,1}, p^Z_{t,2}, \dots, p^Z_{t,L^2})$, where $p^Z_{t,j} = \frac{1}{L}$ for all $j = 1, 2, \dots, L^2$
for $t = [n\epsilon] + 1$ to $n$ do
    Sample $X_t \sim$ Multinomial$(p^X_{t,1}, p^X_{t,2}, \dots, p^X_{t,K^2})$, where $p^X_{t,j}$ is given in Equation (G14)
    Sample $Z_t \sim$ Multinomial$(p^Z_{t,1}, p^Z_{t,2}, \dots, p^Z_{t,L^2})$, where $p^Z_{t,j}$ is given in Equation (G14)
```

that fits a Lasso logistic regression of $\mathbf{Y}$ with main effects of $\mathbf{X}$ and $\mathbf{Z}$ and their interactions due to the simplicity of this simulation setting. This leads to the following test statistic

$$T^{\text{lasso}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) = \sum_{k=1}^{K-1} |\hat{\beta}_k| + \sum_{k=1}^{K-1} \sum_{l=1}^{L-1} |\hat{\gamma}_{kl}|, \tag{G15}$$

where $\hat{\beta}_k$ denotes the estimated main effects for level k out of K levels of X (one is held as baseline) and $\hat{\gamma}_{kl}$ denotes the estimated interaction effects for level k of X with level l of L total levels of Z.

This test statistic also imposes the "no profile order effect" constraints, i.e., we do not separately estimate coefficients for the left and right profiles to increase power. When fitting a Lasso logistical regression of $\mathbf{Y}$ with main effects and interaction of $(\mathbf{X}, \mathbf{Z})$, we obtain a separate effect for both the left and right effects. Since the "no profile order effect" constraints the left and right effects to be similar, we formally impose the following constraints

$$\hat{\beta}_k = \hat{\beta}^L_k = -\hat{\beta}^R_k, \quad \hat{\gamma}_{kl} = \hat{\gamma}^L_{kl} = -\hat{\gamma}^R_{kl}, \tag{G16}$$

where the superscripts L and R denote the left and right profile effects, respectively. To incorporate this symmetry constraint, we split our original $\mathbb{R}^{n\times(4+1)}$ data matrix $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ into a new data matrix with dimension $\mathbb{R}^{2n\times(2+1)}$, where the first n rows contain the values for the left profile (and the corresponding Y) and the next n rows contain the values for the right profile with new response $1 - Y$, [2] shows that this formally imposes the constraints in Equation (G16) by destroying any profile order information in the new data matrix.

G.3 Simulation Results

We first compare the power of our adaptive procedure stated in Algorithm 2 with the *iid* setting where each arm for X and Z are drawn uniformly at random under the simulation setting described in Appendix G.1. We empirically

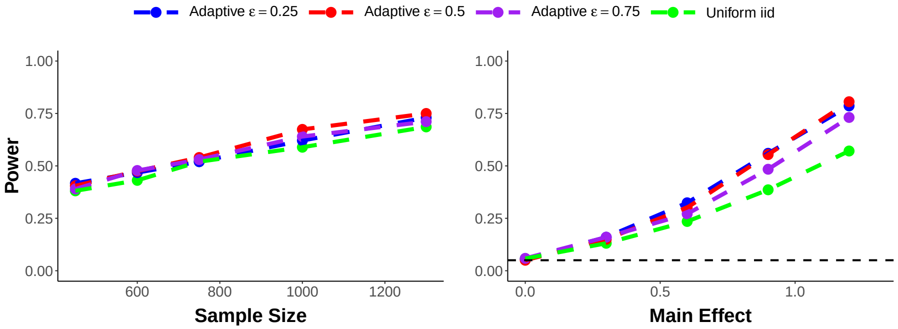

Fig. G2 The figure shows how the power of the ART (based on adaptive sampling procedure in Algorithm 2) and the CRT (based on an *iid* sampling procedure) varies as the sample size increases (left plot) or the main effect increases (right plot). All power curves are calculated from 1,000 Monte-Carlo calculated p-values using Equation (5) with $B = 300$ and test statistic given in Equation (G15) with their respective resampling procedures. The blue, red, and purple power curves denote the power of the ART using the adaptive procedure described in Algorithm 2 and $\epsilon = 0.25, 0.50, 0.75$, respectively. The green power curve denotes the power of the uniform *iid* sampling procedure. The black dotted line in the right panel shows the $\alpha = 0.05$ line. Finally, the standard errors are negligible with a maximum value of 0.016.

compute the power as the proportion of $1,000$ Monte-Carlo p-values less than $\alpha = 0.05$.

For the left panel of Figure G2, we increase sample size when there exist both main effects and interaction effects of X. More specifically, we vary our sample size $n = (450, 600, 750, 1,000, 1,300)$ while fixing the main effects of X and Z at 0.6 and a stronger interaction effect at 0.9 (these refer to the coefficients of the logistic response model defined in Appendix G). For the right panel of Figure G2, we increase the main effects of X and Z with no interaction effect and a fixed sample size at $n = 1,000$. We also vary the exploration parameter ϵ in Algorithm 2 to $\epsilon = 0.25, 0.5, 0.75$.

Both panels of Figure G2 show that the power of the ART with the proposed adaptive sampling procedure is uniformly greater than that of the CRT with a typical uniform *iid* sampling procedure (green). For example when $n = 1,000$ in the left panel, there is a difference in 8.5 percentage points (59% versus 67.5%) between the *iid* sampling procedure and the adaptive sampling procedure with $\epsilon = 0.5$ (red). When the main effect is as strong as 1.2 in the right panel, there is a difference in 24 percentage points (57% versus 81%) between the *iid* sampling procedure and the adaptive sampling procedure with $\epsilon = 0.5$. Additionally, when the main effect is 0 in the right panel, thus under H_0, the power of all methods, as expected, has type-1 error control as the power for all methods are near $\alpha = 0.05$ (dotted black horizontal line).

References

[1] Bates, S., Sesia, M., Sabatti, C., Candès, E.: Causal inference

in genetic trio studies. Proceedings of the National Academy of Sciences **117**(39), 24117–24126 (2020) https://arxiv.org/abs/https://www.pnas.org/content/117/39/24117.full.pdf. https://doi.org/10.1073/pnas.2007743117

[2] Ham, D.W., Imai, K., Janson, L.: Using machine learning to test causal hypotheses in conjoint analysis (2022). https://doi.org/10.48550/arXiv.2201.08343

[3] Candès, E., Fan, Y., Janson, L., Lv, J.: Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B **80**(3), 551–577 (2018)

[4] Ono, Y., Burden, B.C.: The contingent effects of candidate sex on voter choice. Political Behavior, 1–25 (2018)

[5] Arrow, K.J.: What has economics to say about racial discrimination? Journal of Economic Perspectives **12**(2), 91–100 (1998). https://doi.org/10.1257/jep.12.2.91

[6] Lupia, A., Mccubbins, M.: The democratic dilemma: Can citizens learn what they need to know? The American Political Science Review **94** (2000). https://doi.org/10.2307/2586046

[7] Skarnes, W., Rosen, B., West, A., Koutsourakis, M., Roake, W., Iyer, V., Mujica, A., Thomas, M., Harrow, J., Cox, T., Jackson, D., Severin, J., Biggs, P., Fu, J., Nefedov, M., de Jong, P., Stewart, A., Bradley, A.: A conditional knockout resource for the genome-wide study of mouse gene function. Nature **474**, 337–42 (2011). https://doi.org/10.1038/nature10163

[8] Berrett, T., Wang, Y., Barber, R., Samworth, R.: The conditional permutation test for independence while controlling for confounders. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82** (2019). https://doi.org/10.1111/rssb.12340

[9] Farronato, C., MacCormack, A., Mehta, S.: Innovation at uber: The launch of express pool. Harvard Business School Case) **82** (2018)

[10] Glynn, P., Johari, R., Rasouli, M.: Adaptive Experimental Design with Temporal Interference: A Maximum Likelihood Approach. arXiv (2020). https://doi.org/10.48550/ARXIV.2006.05591. https://arxiv.org/abs/2006.05591

[11] Offer-Westort, M., Coppock, A., Green, D.P.: Adaptive experimental design: Prospects and applications in political science. American Journal of Political Science **65**(4), 826–844 (2021) https://arxiv.org/abs/https:

//onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/pdf/10.1111/ajps.12597. https://doi.org/10.1111/ajps.12597

[12] Bojinov, I., Shephard, N.: Time series experiments and causal estimands: Exact randomization tests and trading. Journal of the American Statistical Association, (2019)

[13] Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, USA (2015)

[14] Rosenberger, W.F., Uschner, D., Wang, Y.: Randomization: The forgotten component of the randomized clinical trial. Statistics in Medicine **38**(1), 1–12 (2019) https://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7901. https://doi.org/10.1002/sim.7901

[15] Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**(3-4), 285–294 (1933) https://arxiv.org/abs/https://academic.oup.com/biomet/article-pdf/25/3-4/285/513725/25-3-4-285.pdf. https://doi.org/10.1093/biomet/25.3-4.285

[16] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA (2018)

[17] Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics **6**(1), 4–22 (1985). https://doi.org/10.1016/0196-8858(85)90002-8

[18] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) **58**(1), 267–288 (1996)

[19] Shi, C., Xiaoyu, W., Luo, S., Zhu, H., Ye, J., Song, R.: Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. Journal of the American Statistical Association, 1–29 (2022). https://doi.org/10.1080/01621459.2022.2027776

[20] James, W., Stein, C.: Estimation with quadratic loss Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Berkeley. CA USA: Univ. of California Press (1961)

[21] Le Cam, L.: On the asymptotic theory of estimation and testing hypotheses. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, pp. 129–156 (1956). University of California Press

[22] Luce, R.D., Tukey, J.W.: Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of Mathematical Psychology **1**(1), 1–27 (1964). https://doi.org/10.1016/0022-2496(64)90015-X

[23] Ono, Y.: Replication Data for: The Contingent Effects of Candidate Sex on Voter Choice (2018). https://doi.org/10.7910/DVN/IZKZET

[24] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B **57**(1), 289–300 (1995)

[25] Wainwright, M.J., Jordan, M.I., *et al.*: Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning **1**(1–2), 1–305 (2008)

[26] Liu, J.S.: Monte Carlo Strategies in Scientific Computing, p. 344. Springer, New York, Berlin, Heidelberg (2008)

[27] Wu, J., Ding, P.: Randomization tests for weak null hypotheses in randomized experiments. Journal of the American Statistical Association **116**(536), 1898–1913 (2021) https://arxiv.org/abs/https://doi.org/10.1080/01621459.2020.1750415. https://doi.org/10.1080/01621459.2020.1750415

[28] Ash, R.B., Doleans-Dade, C.A.: Probability and Measure Theory, 2nd edn. Harcourt/Academic Press, Burlington, MA, USA (1999)

[29] Hainmueller, J., Hopkins, D.J.: The hidden american immigration consensus: A conjoint analysis of attitudes toward immigrants. American Journal of Political Science (2015). https://doi.org/10.1111/ajps.12138