Hypothesis Testing in Sequentially Sampled Data: AdapRT to Maximize Power Beyond *iid* Sampling*

Dae Woong Ham[†1] and Jiaze Qiu[‡1]

[1]*Department of Statistics, Harvard University.*

March 31, 2022

Abstract

Testing whether a variable of interest affects the outcome is one of the most fundamental problem in statistics. It is often the main scientific question of interest and also ubiquitously used in variable selection. To tackle this issue, the conditional randomization test (CRT) is widely used to test the independence of a variable of interest (X) with an outcome (Y) holding some controls (Z) fixed. The CRT uses randomization based inference that relies solely on the random *iid* sampling of (X, Z) to produce exact finite-sample p-values that are constructed using any test statistic. We propose a new method the *adaptive randomization test* (AdapRT) that similarly tests whether X affects Y given Z but does not require the data (X, Z) be sampled *iid* from a distribution. AdapRT enjoys the same benefits as the CRT but allows the experimenter to sequentially sample (X, Z) at each time step as a function of all previous values of (X, Z) and the outcome Y. AdapRT also allows an analyst to perform randomization inference on any given data that has known sequential dependencies. We also provide a pipeline for practitioners to diagnosis the efficiency of their proposed adaptive procedures. We then showcase the AdapRT and the proposed pipeline in a particular "multi-arm" bandit problem known as the normal-mean model. Under this setting, we theoretically characterize the power of the *iid* sampling scheme and the AdapRT and empirically find that AdapRT can uniformly outperform the typical uniform *iid* sampling scheme that pulls all arms with equal probability. We also surprisingly find that the AdapRT can be more powerful than even the oracle *iid* sampling scheme when the signal is relatively strong. We discover that the proposed adaptive procedure is successful because it up-weights arms with high signal, down-weights arms with no signal, and most importantly stabilizes arms that may initially look like "fake" signal. We then apply the AdapRT to a popular factorial survey design setting known as conjoint analysis. We similarly find the AdapRT to be more power than the *iid* sampling scheme both through simulations and an application to a recent conjoint study on political candidate evaluation .

Keywords: Independence Testing, Conditional Independence Testing, Reinforcement Learning, Randomization Inference, Design Based Inference, Sequential Sampling, Adaptive Sampling, Dynamic Sampling, Non-parametric Testing, Model-X

*Both authors contributed equally and are listed in alphabetical order.

[†]Email: daewoongham@g.harvard.edu

[‡]Email: jiazeqiu@g.harvard.edu

1 Introduction

Independence testing is ubiquitous in statistics. It is often the main task of interest in variable selection problems. For example, it is an important tool in causal inference for different applications (Bates et al., 2020; Ham, Imai and Janson, 2022; Candès et al., 2018). 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 (Ono and Burden, 2018; Arrow, 1998; Lupia and Mccubbins, 2000). Biologists may be interested in the effect of a specific gene on a characteristic after holding all other genes constant (Skarnes et al., 2011).

More specifically, the objective is to test whether a response Y is statistically affected by a variable of interest X while controlling for other variables Z. Informally speaking, the main objective is 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 can be 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, *design-based* approach has been increasingly gaining popularity (Ham, Imai and Janson, 2022; Bates et al., 2020; Bojinov and Shephard, 2019a) to tackle the independence testing problem. In an influential paper (Candès et al., 2018), the authors introduce the conditional randomization test (CRT), which uses the "Model-X" approach to perform randomization based inference. We refer to this testing approach as the "Model-X" randomization inference approach. This approach assumes nothing about the $Y \mid (X, Z)$ relationship but shifts the burden on requiring knowledge of $X \mid Z$. 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 (X, Z) is immediately available and the CRT can be classified as a non-parametric approach to testing.

Although the "Model-X" randomization inference approach proposed by the CRT is close to "assumption-free" in an experimental setting (or when the distribution of (X, Z) is exactly known), it does require that the sampling scheme of (X, Z) is collected *iid*. On the other hand, there is a growing literature on sequentially and adaptively collected data in reinforcement learning. In this setting, an experimenter is faced with the task of sampling the next values of (X, Z) as a function of all the previous values of (X, Z) and Y to maximize an objective (Sutton and Barto, 2018a). For example, the commonly known "multi-arm bandit" problem (Slivkins, 2019) aims to sample the next value of X, often referred to as sampling the next "arms" of X, that would give a higher value of outcome Y. In this setting, there are a finite but various choices of arms the experimenter can pull, in which the primary objective is to sequentially optimize which arm to pull to produce the maximum reward. There exist many popular algorithms such as the Thompson sampling (Thompson, 1933) and epsilon greedy algorithms (Sutton and Barto, 2018b) that use sequentially adaptive sampling schemes to provide guarantees of reaching a desirable objective.

Given these two settings, a very natural question is to incorporate the ideas of a sequentially adaptive procedure in the context of independence testing through the "Model-X" randomization inference approach. Although the reinforcement learning literature does not concern itself with the independence testing problem, many key ideas in reinforcement learning boils down to identifying arms of X that have high signal. Therefore, one could expect that incorporating a sequentially adaptive sampling scheme, such as those that are used in reinforcement learning, can be helpful in answering the independence testing problem. To provide a simple naive example, consider the aforementioned multi-arm bandit problem with a typical *iid* sampling framework that pulls each arm with equal probability for all n samples. If there exist only one arm that has a weak signal, then the power, even using the most powerful machine learning algorithms, should be relatively low. On the other hand, if the experimenter was allowed to sequentially sample, then he/she would be able to sample more from the arm with signal, where the data would clearly show more evidence of a signal, thus perhaps leading to a greater power.

Given this intuition, it may be desirable to tackle the independence testing problem using adaptively sampled data. In general, regardless of using randomization based inference, parametric inference, non-parametric inference, etc., it is difficult to get a valid test for the independence testing problem when X is adaptively sampled as a function of Y. For example, even if one knew that the true function of $Y \mid X$ was linear, a parametric linear regression approach would lead to many false discoveries due to the possible "fake" dependencies induced by the adaptive procedure. In order to construct a valid test, one would need to decouple the adaptive procedure from the "true" relationship between Y and X, which is a generally difficult task. Consequently, many existing approaches, like the CRT, requires the *iid* assumption and have no guarantees for a valid test under a sequentially sampled scheme. We note that even if the data did not necessarily come from an experiment but was collected through some sequential process such as a time series, the existing methods will mostly fail. Although recent work by (Bojinov and Shephard, 2019a) extends randomization inference in the context of time series with specific carryover assumptions, there is still a lack of a general procedure that allows independence testing for sequentially collected data. Our paper successfully incorporates the sequentially adaptive sampling scheme in the context of the "Model-X" randomization framework while enjoying all the same benefits as those enjoyed by the CRT. We now summarize our contributions.

1.1 Our Contributions

The main contribution of our paper is we allow the same "Model-X" randomization inference procedure under sequentially collected data that depends on all the previous history. More specifically, we allow the data (X_t, Z_t) to be sequentially collected at time t as a function of the historical values of $X_{1:(t-1)}, Z_{1:(t-1)}, Y_{1:(t-1)}$. In this non *iid* sampling scheme, the CRT theory is insufficient to guarantee a valid finite-sample p-value. Our contribution now allows experimenters to flexibly use any sequentially adaptive procedure they desire to potentially increase power compared to an *iid* sampling scheme. Our contribution also allows the analyst to run the "Model-X" randomization inference approach for any sequentially collected data as long as the analyst knows how the data was sequentially sampled. We call our approach the AdapRT (Adaptive Randomization Test) and we remark that AdapRT, like the CRT, does not require any knowledge of $Y \mid (X, Z)$. Therefore, in a pure experimental framework, AdapRT can also be viewed as a non-parametric test.

In Section 2 we formally introduce the proposed method AdapRT and show that the "Model-X" randomization inference approach can be used in sequentially sampled data while providing exact finite-sample valid p-values for any test statistic. Although this formally allows practitioners to adaptively sample data, it does not give any guidance on how to choose a reasonable adaptive scheme. Therefore, we provide a pipeline practitioners can follow in Section 2.5. We then apply this pipeline and show how AdapRT can be more powerful than the *iid* sampling scheme given the same test statistic and sample size in two common scenarios. We first explore in Section 3 the AdapRT in the normal-means model setting, which is special case of the "multi-arm" bandit setting. In this section, we not only apply our proposed pipeline but theoretically characterize the power to give further understanding on how adaptive schemes may help increase power under different scenarios. We find that adapting can significantly increase power compared to the typical uniform *iid* sampling scheme that pulls all arms with equal probability due to three main reasons. 1) AdapRT allows the experimenter to sample more of the signal arms. 2) It also allows the experimenter to sample more from arms that may look like potential signals but is actually not. This is useful because such a sampling scheme then stabilizes the initially "fake" signal looking arm. 3) Finally, AdapRT also down-weights sampling from arms that "clearly" (with high confidence) contain no signal. This allows our remaining samples to focus on and explore the other arms. We also surprisingly find a stronger conclusion, namely that AdapRT can be more powerful than the oracle *iid* sampling scheme when the signal is relatively strong (see Section 3.3 for details). We then explore in Section 4 the AdapRT in a factorial survey setting - often known as conjoint analysis (Hainmueller, Hopkins and Yamamoto, 2014; Ham, Imai and Janson, 2022). We empirically apply

the proposed pipeline in this different setting and to a recent conjoint study that studies whether the gender of political candidates matter given other factors (Ono and Burden, 2018). We find similar results to those presented in Section 3, where AdapRT is significantly more power than the *iid* sampling scheme. Section 5 concludes with a discussion and remarks about future work.

1.2 Related Works

In this section, we put our proposed method in the context of the current literature. The AdapRT 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 introduce adaptive testing in the context of randomization inference when specifically tackling the independence testing problem. However, just like any adaptive procedure, we need to choose a sensible procedure that will increase power. Consequently, many ideas from the reinforcement learning literature can 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 (Thompson, 1933) and epsilon greedy algorithms (Sutton and Barto, 2018b), to be useful when constructing the adaptive sampling scheme. Although ideas from reinforcement learning can be utilized when performing AdapRT, we emphasize that the objective of independence testing is different than that of a typical reinforcement learning problem. For example, in the multi-arm bandit problem with a binary response Y, the researcher is interested in pulling the arms that maximize the number of $Y = 1$ (reward). For the inference problem, the AdapRT may be more interested in not only pulling arms with strong reward but also the arms that give high probability to sample $Y = 0$ since they both inform that X is not independent of Y. Furthermore, in the independence testing problem, we also care about sampling the "useless" arms to use as a baseline comparison. Therefore, the "oracle" sampling policy would not only want to sample the arm with a strong signal but also sample the other "useless" arms with no signal. However, in the bandit problem, the oracle would always sample the arm with the highest reward. This difference is illustrated and further emphasized in the theoretical analysis of the normal means bandit problem in Section 3.3.

There also exist literature dealing with the appropriateness of "Model-X" and building powerful test statistics for the application of interest (Bates et al., 2020; Ham, Imai and Janson, 2022; Candès et al., 2018; Bojinov and Shephard, 2019a) when using the CRT. Although all this is crucial when performing randomization inference, this is not the main focus of our paper. We assume we are either in an experimental framework where the experimenter is deciding on how to efficiently design the experiment to powerfully answer the question of interest or the analyst is already given a dataset with knowledge of how (X, Z) was (adaptively) collected. Furthermore, we do not explore which test statistics may be powerful when using the AdapRT. Although this is an interesting avenue for future research and also very important to increase power, we approach this problem from the design stage. In other words, we fix a reasonable test statistic that is commonly used in a typical *iid* sampling scheme and vary the possible sampling scheme namely the *iid* sampling scheme and a sequential sampling scheme. Since our objective is to show that the AdapRT can be useful and more powerful than the CRT, it is sufficient to show that it can beat the power of the *iid* sampling scheme with a fixed a reasonable test statistic.

1.3 The Conditional Randomization Test (CRT)

Before proposing our method we briefly introduce the CRT in this section. The CRT assumes that $(X_t, Z_t) \overset{i.i.d}{\sim} f_{XZ}$ for $t = 1, 2, \ldots, n$, where f_{XZ} denotes the joint probability density function (pdf) or probability mass function (pmf) of (X, Z) 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 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} 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,

$$\tilde{X}_t^b \sim \frac{f_{XZ}(\tilde{x}_t^b, Z_t)}{\int_z f_{XZ}(\tilde{x}_t^b, z) dz} \text{ for } t = 1, 2, \ldots, n, \tag{1}$$

where the right hand side is the pdf of the conditional distribution $X \mid Z$ induced by the joint pdf f_{XZ} 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 current Z_t. Under the conditional independence null, $Y \perp\!\!\!\perp X \mid Z$, Candès et al. show that $(\tilde{\mathbf{X}}^{\mathbf{1}}, \mathbf{Z}, \mathbf{Y})$, $(\tilde{\mathbf{X}}^{\mathbf{2}}, \mathbf{Z}, \mathbf{Y})$, $\ldots$, $(\tilde{\mathbf{X}}^{\mathbf{B}}, \mathbf{Z}, \mathbf{Y})$, 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,

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

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 5 is guaranteed to have exact type I error control, i.e., $\mathbb{P}(p_{\text{CRT}} \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$ under the null $Y \perp\!\!\!\perp X \mid Z$ despite the choice of T and any relationship of $Y \mid (X, Z)$. Therefore, under this framework, practitioners would ideally choose a test statistic that is powerful to detect a signal between Y and X such as the sum of the absolute value of the main effects of X from a penalized Lasso regression (Tibshirani, 1996).

2 Methodology

With the goal of tackling the independence testing problem for sequentially generated data using a "Model-X" randomization inference approach, we now introduce our proposed method - the Adaptive Randomization Test (AdapRT).

2.1 Sequential Adaptive Sampling

First, we formally present our definition of sequentially adaptive procedures.

Definition 2.1 (Sequential adaptive procedure)**.** We say the sample $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ follows an adaptive procedure A if the sample obeys the following sequential data generating process.

$$\begin{aligned}
&(X_1, Z_1) \sim f_1^A(x_1, z_1), \quad Y_1 \sim f_Q(x_1, z_1) \\
&(X_2, Z_2) \sim f_2^A(x_2, z_2 \mid x_1, z_1, y_1), \quad Y_2 \sim f_Q(x_2, z_2) \\
&\qquad \vdots \\
&(X_t, Z_t) \sim f_2^A(x_t, z_t \mid x_1, z_1, y_1, \ldots, x_{t-1}, z_{t-1}, y_{t-1}), \quad Y_t \sim f_Q(x_t, z_t),
\end{aligned}$$

[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

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 probability density function of (X_t, Z_t) given the past realizations, and f_Q denotes the probability density function 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 a procedure f_t^A that may be dependent on all the history (including the outcome) while "nature" f_Q determines the next outcome. We emphasize that f_Q is generally unknown and in most cases hard to model exactly. 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_Q 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_Q is stationary and does not change across time. Both of these assumptions are typically invoked in the sequential reinforcement learning literature (Shi et al., 2022; Sutton and Barto, 2018a; Bojinov and Shephard, 2019b). Our methodology naturally extends to non-stationary models, where f_Q also depends on t. However, for presentational clarity, we present our method in the common stationary scenario. Moreover, we add that our methodology and main theoretical results should naturally extend when there are simple structural carryover effects, but we leave this for future work.

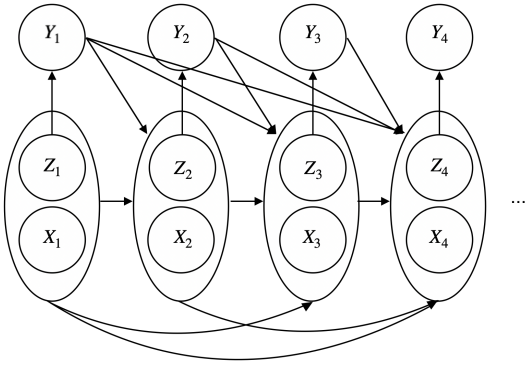

Figure 1: Schematic diagram of the Sequential Adaptive Sampling Scheme in Definition 2.1. The directed arrows denote the order in how the random variable(s) may affect the corresponding random variable(s).

Given the adaptive procedure defined in Definition 2.1, the main objective is to determine whether the variable of interest X affects Y after controlling for Z. Unlike the CRT setting, the data is no longer *iid* and formalizing the main objective of testing $Y \perp\!\!\!\perp X \mid Z$ requires further notation. For example, in the CRT procedure, 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 *iid*, the null hypothesis can ignore the subscript t and test the null using the whole data. However, for the adaptive case, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$ is trivially false for any non-degenerate adaptive procedure A since $\mathbf{X}$ 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,

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

where $\mathcal{X}$ denotes the entire possible 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 our 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.

We finish this subsection with a discussion of H_0. 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, Ham, Imai and Janson show that the CRT null is equivalent to testing the following causal hypothesis,

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

where $Y_t(x, z)$ is the potential outcome for individual t at values $X = x, Z = z$. The proposed H_0 implicitly captures this causal hypothesis because $f_q(x, z)$, by definition, characterizes the causal relationship between (X, Z) and Y. To formally establish this in the potential outcome framework, we define $Y_t(x, z) \overset{i.i.d}{\sim} f_Q(x, z)$ from a super-population framework (Imbens and Rubin, 2015). Then H_0 is indeed also testing H_0^{Causal}, i.e., whether X causally impacts Y after accounting for Z.

2.2 Adaptive Randomization Test (AdapRT)

The main contribution of this paper is that we allow the "Model-X" randomization inference approach when testing H_0 under a sequential adaptive setting in Definition 2.1. Since (X_t, Z_t, Y_t) are no longer sampled *iid* from some joint distribution, it is not obvious how 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 5. 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. To achieve this, Candès et al. construct the fake resamples $\tilde{\mathbf{X}}^{\mathbf{b}}$ from the conditional distribution of $X \mid Z$ as done in Equation 1, which directly satisfies the exchangeability criteria due to the *iid* setting of X and Z. However, in the sequential adaptive case, X_t depends on all the history including the response and it is unclear how to construct our resamples because we can not assume knowledge of f_Q.

Although there is actually not only one way to obtain valid resamples of X (further discussed in Section 2.4), we propose the most natural resampling procedure that respects our sequential adaptive setting in Definition 2.1. Before formally presenting the resampling procedure, we build intuition on how we construct valid resamples $\tilde{\mathbf{X}}^{\mathbf{b}}$. Similar to the CRT, the key is to create the fake copies of X by sampling $\tilde{X}$ as if it were the real X conditional on Z, Y. For the *iid* CRT sampling scheme, this reduces to sampling X_t *iid* from the conditional distribution of $X \mid Z$ for all $t = 1, 2, \ldots, n$. In our sequential adaptive sampling scheme, we similarly sequentially sample $\tilde{X}_t$ as if it were the original X_t but conditional on the history of (Z, 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,

$$\tilde{X}_1^b \sim \frac{f_1^A(\tilde{x}_1^b, z_1)}{\int_z f_1^A(\tilde{x}_1^b, z) dz}, \tilde{X}_2^b \sim \frac{f_2^A(\tilde{x}_2^b, z_2 \mid \tilde{x}_1^b, z_1, y_1)}{\int_z f_1^A(\tilde{x}_1^b, z \mid \tilde{x}_1^b, z_1, y_1) dz}, \ldots, \tilde{X}_n^b \sim \frac{f_t^A(\tilde{x}_n^b, z_n \mid \tilde{x}_1^b, z_1, y_1, \ldots, \tilde{x}_{n-1}^b, z_{n-1}, y_{n-1})}{\int_z f_1^A(\tilde{x}_n^b, z \mid \tilde{x}_1^b, z_1, y_1, \ldots, \tilde{x}_{n-1}^b, z_{n-1}, y_{n-1}) dz},$$

for $b = 1, 2, \ldots, B$ independently condition 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 $Y_{1:(t-1)}$. Just like the CRT, Definition 2.2 requires one to sample from a conditional distribution. For this practically important consideration, we propose another 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 since it is the same original sequential sampling scheme. We refer to this as the convenient adaptive sampling scheme.

Unfortunately resampling from the NARP does not come for free. Recall that we require our resampled $\tilde{\mathbf{X}}^b$ to be exchangeable with $\mathbf{X}$ conditional on $(\mathbf{Z}, \mathbf{Y})$, which implies the distributional dependency among $(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z})$ should be the same as that among $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. In particular, note that the following relationship is always true assuming NARP, for any t,

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

because $\tilde{X}_{1:(t-1)}$ is a random function of only $\left(Y_{1:(t-1)}, Z_{1:(t-1)}\right)$. Equation 4 shows that Z_t is independent of previous fake resamples of X. To satisfy the exchangeability criteria, we also need the same distributional relationship in Equation 4 to be satisfied by the original sampling scheme of $\mathbf{X}$. This leads to the following natural assumption where Z can not depend on previous X, which turns out to be both sufficient and necessary to ensure validity of using AdapRT 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 probability density function 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 $x_{1:(t-1)}$.

Assumption 1 states that the sequential adaptive procedure A does not allow Z_t to depend on the history of X. Although this does restrict our adaptive procedure, it is crucial that each X_t and Z_t are allowed to adapt by looking at its own previous values and the previous responses.

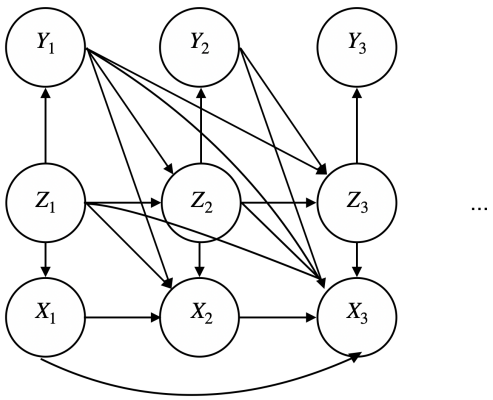

Figure 2: Schematic diagram of the convenient Adaptive Sampling Scheme that satisfies Assumption 1. As before, the directed arrows denote the order in how the random variable(s) may affect the corresponding random variable(s).

We visually summarize Assumption 1 and a more convenient, but not necessary, way to conduct a restricted adaptive sampling scheme 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. We emphasize that this is a mere practical convenience and not necessary for the general AdapRT procedure to work. Assumption 1 is also

Algorithm 1: AdapRT p-value

Input: Adaptive procedure A, test statistic T, total number of re-samples 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 $b = 1, 2, \dots, B$ **do**
  Sample $\tilde{X}^{(b)}$ according to the NARP in Definition 2.2;
Output:

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

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 our AdapRT procedure in Algorithm 1

We note that although the p-value calculation of p_{AdapRT} in Equation 5 is similar to p_{CRT}, the resamples $\tilde{X}^b$ are different in the two procedures. We now state the main theorem that gives the validity of using AdapRT for testing H_0.

Theorem 2.1 (Validness of p-values under AdapRT)**.** 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 $b = 1, 2, \dots B$ conditionally independent of $(\mathbf{Z}, \mathbf{Y})$. Then the p-value p_{AdapRT} in Algorithm 1 for testing H_0 is a valid p-value. Equivalently, $\mathbb{P}(p_{\text{AdapRT}} \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$.

Remark 1. We want to comment that p_{AdapRT} is also a valid p-value condition on $\mathbf{Y}$ and $\mathbf{Z}$.

The proof of Theorem 2.1 is in Appendix 6.1. This theorem is the main result of this paper, which allows non-parametric testing of H_0 for sequentially collected data. Moreover, since this result is valid for any sequential/adaptive sampling scheme, it opens new avenues of research to analyze which adaaptive procedures may lead to better power. 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.

Theorem 2.2 (Necessity of Assumptions)**.** For any 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 be satisfied.

The proof is in Appendix 6.1.

2.3 Multiple Testing

So far we have introduced our proposed method to test H_0 for a single variable of interest X conditional on Z. However, the practitioner may be interested in testing multiple H_0 for different variables of interest.

To formalize this, denote $X = (X^1, X^2, \dots, X^p)$ 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 $j = 1, 2, \dots p$, where X^{-j} denotes all variables in X except X^j. Given a fixed j, our proposed methodology in Section 2.1-2.2 can be used to test any single one of these hypothesis. The main issue with directly extending our proposed methodology for testing for all $j = 1, 2, \dots p$ is that Assumption 1 does not allow X^{-j} to depend on previous X^j but X^j may depend on previous X^{-j}. This asymmetry may cause this assumption

to hold when testing for X^j but not when testing for $X^{j'}$ for $j \neq j'$. For example suppose we are testing $Y \perp\!\!\!\perp X^1 \mid X^{-1}$ and allow X^1 to depend on the history of X^2. This would satisfy Assumption 1, but Assumption 1 would be violated when testing for $Y \perp\!\!\!\perp X^2 \mid X^{-2}$. In order to satisfy Assumption 1 for all variables of interest simultaneously, we further modify our sampling procedure such that Assumption 1 holds for all $j = 1, 2, \dots p$. A simple and sufficient way to modify our procedure is to impose each X_t^j to be 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 $X_{1:(t-1)}^j$ and the history of the response but not the history and current values of $X^{j\prime}$ for $j \neq j'$. We formalize this in following assumption.

Assumption 2. For each $t = 1, 2, \dots, n$ suppose each $X_t = (X_t^1, X_t^2, \dots, X_t^p)$ are sampled according to a sequential adaptive sampling procedure A: $X_t \sim f_t^A(x_t^1, x_t^2, \dots, x_t^p \mid x_{1:(t-1)}^{-j}, x_{1:(t-1)}^j, y_{1:(t-1)})$. We say an adaptive procedure A satisfies Assumption 2 if f_t^A can be written into following factorized form,

$$f_t^A(x_t^1, x_t^2, \dots, x_t^p \mid x_{1:(t-1)}^{-j}, x_{1:(t-1)}^j, y_{1:(t-1)}) = \prod_{j=1}^{p} f_{t,j}^A(x_t^j \mid x_{1:(t-1)}^j, y_{1:(t-1)}^j)$$

with every $f_{t,j}^A(\cdot | x_{1:(t-1)}^t, y_{1:(t-1)}^t)$ being a valid probability measure for all possible values of $(x_{1:(t-1)}^t, y_{1:(t-1)}^t)$.

Assumption 2 states that X^j cannot adapt and be dependent on any of the other $X^{j'}$. This assumption is trivially sufficient to satisfy Assumption 1 when testing H_0 for each X^j, thus leading to a valid p-value for every X^j simultaneously, using the proposed AdapRT scheme in Algorithm 1. Although our framework gives valid p-values for each of the multiple tests, we still need to account for multiple testing issues. For example, one naive way to control the false discovery rate is to use the Benjamini Hochberg procedure (Benjamini and Hochberg, 1995). Since this is not the focus of our paper, we leave the multiple testing issue to future research.

2.4 Discussion of the Natural Adaptive Resampling Procedure

Keen readers may argue that we do not have to follow the NARP, thus no longer needing Assumption 1. Since we only need $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $(\tilde{\mathbf{X}}^b, \mathbf{Z}, \mathbf{Y})$ to be exchangeable under the null, one could try to find a way to resample $\tilde{\mathbf{X}}^b$ to allow this exchangeability. A necessary condition for exchangeability to hold is that $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $(\tilde{\mathbf{X}}^b, \mathbf{Z}, \mathbf{Y})$ should be equal in distribution. If one could sample the entire data vector $\tilde{\mathbf{X}}$ from the conditional distribution of $\mathbf{X} \mid (\mathbf{Z}, \mathbf{Y})$, this construction of the resamples 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 Wainwright, Jordan et al. (2008) and may require knowledge of f_Q (even under the null) while AdapRT requires no knowledge about f_Q. There may exist another resampling procedure that leverages the sequential nature of the original data generating process that not only does not require Assumption 1 but also does not require knowledge of f_Q. Although such a resampling procedure may exist, we believe to actually sample from the required distribution may be computationally infeasible given the complicated dependencies induced by the adaptive procedure. The NARP avoids this issue because it resamples $\tilde{\mathbf{X}}^b$ from the already known original adaptive procedure. Therefore, we propose the AdapRT with the natural resampling procedure that satisfies Assumption 1 as the main proposed method and leave the possibility for different resampling procedures for future research.

2.5 Pipeline for Practitioners: Choosing a Powerful Adaptive Procedure

We conclude this section with a general guideline to practitioners on how they may determine if an adaptive scheme is reasonable. Although AdapRT described in Algorithm 1 allows a practitioner to test H_0 with a

sequential adaptive procedure, the method itself does not give any hints on how a practitioner can create a reasonable adaptive procedure A to more powerfully reject H_0. Unfortunately, there is no general guideline that will allow a class of adaptive schemes to always be more powerful than a standard *iid* data collection because the power depends on the alternative hypothesis, i.e., the true model of f_Q. Although we propose some adaptive procedures that work well under certain scenarios in Section 3 and 4, we leave the exploration of optimal adaptive schemes for specific applications for future research. Instead, we present a general pipeline practitioners can follow to determine whether a certain adaptive procedure can be helpful.

Since both the *iid* sampling procedure and the AdapRT procedure have exactly valid p-values, we can measure the "goodness" of an adaptive sampling scheme with respect to the statistical power of the test. The power of both the *iid* and the adaptive procedure will depend on a given test statistic T and the true model f_Q. The power of the AdapRT will additionally be dependent on the proposed adaptive procedure. Our goal is to present a pipeline to determine whether the power of a proposed adaptive procedure will be higher than that of a standard *iid* sampling scheme given a fixed sample size n and the same test statistic T. Although we provide theory on how to determine the power of a specific adaptive procedure under a normal-means model in Section 3, it is in general difficult to theoretically characterize power, especially for a complicated adaptive scenario. Therefore, we propose the practitioner run a power analysis given a guess of f_Q based on domain expertise. For concreteness, consider the motivating bandit example where a practitioner wants to determine if a treatment with multiple levels has any effect on the response. The practitioner can assume a simple linear model for f_Q with a weak main effect on a few levels of X. Suppose further there is only an experimental budget of $n = 50$. Then, the practitioner can sample each treatment level *iid* uniform and obtain the corresponding response Y from f_Q for a sample size of 50. Given this sample, the practitioner can obtain one CRT based p-value p_{CRT} given a reasonable test statistic such as the sum squared of the coefficients of X from a linear regression. The practitioner can repeat this whole process say 1000 times to empirically obtain a power estimate of this typical uniform *iid* sampling scheme. Similarly, if the practitioner has multiple adaptive procedures he/she is interested in potentially implementing but unsure if the power will be higher or possibly even lower, then the practitioner can repeat the same power analysis for the adaptive procedure(s).

This pipeline will allow the practitioners to empirically diagnosis whether the proposed adaptive procedures are helpful in increasing power given a guess of f_Q. Unfortunately, this pipeline leads to an accurate diagnosis only if f_Q is close to the truth. To mitigate the reliance on specifying the correct f_Q, the practitioner can repeat this procedure for several different specifications and parameter values of f_Q. For example, f_Q can vary between both linear and non-linear models with weak, medium, and strong main effects. If the results show a specific adaptive procedure A uniformly dominating an *iid* sampling scheme across all (or most) the different specifications, then the practitioner should be more comfortable with employing the adaptive sampling scheme in the real experiment. Lastly, we note that this proposed pipeline may be computationally expensive especially if the practitioner wants to try many specifications of f_Q and many different adaptive procedures.

We summarize our proposed guideline in Algorithm 2. We denote m to be the total number of proposed adaptive schemes, which can also differ by only by a single adaptive parameter for the same class of adaptive schemes. For example, in the ϵ-greedy adaptive scheme (Sutton and Barto, 2018b), the practitioner can vary different values of ϵ since it is not clear which ϵ may be the best suitable one for a particular scenario. We also emphasize that although Algorithm 2 does require the practitioner to have a guess of f_Q, a practitioner can run algorithm 2 for a class of different f_Q's to see if a specific adaptive procedure is successful across all (or most) hypothesized f_Q.

A reasonable value of P would be 500. For the remainder of the paper, we demonstrate Algorithm 2 under various common settings of interest. In Section 3, we theoretically characterize a reasonable adaptive sampling scheme's local asymptotic power in the normal-means model and show through simulations how it can lead to a greater power than a typical *iid* sampling scheme. We then propose in Section 4 a more com-

Algorithm 2: Pipeline to Practitioners

Input: f_Q, test statistic T, Adaptive procedures $A_1, A_2, \ldots, A_m$, standard *iid* sampling scheme, total number of p-values P, significance level α ;

for $i = 1, 2, \ldots, P$ **do**

Sample data $(\mathbf{X}, \mathbf{Z})$ from a standard *iid* sampling scheme and obtain $\mathbf{Y}$ from f_Q

Obtain p-value p_{CRT} for the *iid* sampling scheme with the CRT procedure and test statistic $T(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$

for $j = 1, 2, \ldots, m$ **do**

Sample data $(\mathbf{X^j}, \mathbf{Z^j}, \mathbf{Y^j})$ from a sequential adaptive sampling scheme A_j and f_Q

Obtain m p-values p_{AdapRT} for each adaptive sampling scheme, where the jth p-value corresponding to adaptive sampling scheme A_j is computed using Equation 5 and test statistic $T(\mathbf{X^j}, \mathbf{Z^j}, \mathbf{Y^j})$ and the fake samples obtained from the NARP in Definition 2.2.

Output: Empirical power for each sampling scheme, calculated as the proportion of P p-values less than α.

plicated and powerful adaptive sampling scheme in the setting of factorial experiments, commonly known as conjoint analysis. We show through simulations that the power can significantly increase when using the proposed adaptive scheme and apply the same adaptive scheme to a recent conjoint study that studies whether a political candidate's gender matters in voting preference.

3 AdapRT in Normal Means Model

In this section, we explore the AdapRT under the well-known normal-means setting James and Stein (1961). Using local asymptotic power analysis, we theoretically characterize the power of the typical (uniform) *iid* sampling scheme and a naive, but still insightful, two stage adaptive sampling scheme.

We first introduce the normal-means setting, the sampling schemes we consider, and the test statistic in Section 3.1. We then present two main theorems, Theorem 3.1 and Theorem 3.2, that characterize the asymptotic power of both the *iid* and adaptive sampling schemes respectively under local alternatives of $O(n^{-1/2})$ distance in Section 3.2. Finally, we numerically evaluate Theorem 3.1 and Theorem 3.2 to illustrate when the adaptive sampling scheme leads to an increase of power. Simultaneously, we also use show in Section 3.3 to concretely showcase the proposed pipeline in Section 2.5 that illustrates both when and how we should adapt in this setting.

3.1 Normal Means Model

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

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

where j refers to the p different possible integer values of X. We often refer to different values of X as different arms. 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.

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

As usual, our null assumes that X does not affect Y in any way, i.e., all arms have the same mean.

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

Given a budget of n samples, our task is to come up with a reasonable adaptive sampling scheme that leads to a higher power than that of the typical uniform *iid* sampling scheme. For consistency of notations, in this section, we use i (instead of t) to denote the sample index. We now introduce the typical uniform *iid* sampling scheme.

Definition 3.1 (Normal Means Model: *iid* Sampling Scheme with Weight Vector q)**.** We call a sampling scheme *iid with weight vector* $q = (q_1, q_2, \cdots, q_p)$ if each sample of $\mathbf{X} = \left(X_1, X_2, \dots, X_n\right)$ is sampled independently and

$$\mathbb{P}(X_i = j) = q_j, \text{ for any } i \in 1, 2, \dots, n \text{ and } j \in \mathcal{X}. \tag{6}$$

We note that this definition is more general than the classical uniform iid sampling scheme, in which $q = (1/p, 1/p, \dots, 1/p)$. We further denote $\mathbf{X} \sim \mathcal{M}(q)$ to compactly describe the *iid* sampling scheme for $\mathbf{X}$. 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, analysing the power of a fully adaptive procedure is generally theoretically infeasible. For example, if every sample is dependent on the previous history, no central limit theorems will allow any distributional statements due to this heavy dependence. To make the problem theoretically tractable, we consider a naive "two stage" adaptive procedure. The first stage will be an exploration stage that follows the typical *iid* sampling scheme while the second stage will adapt only once given the first stage data. The second stage will adapt by re-weighting the probability of pulling each arm by a function of the sample mean. The main intuition is that under the alternative we consider, the arm that has the signal on average will have a higher sample means. Furthermore, the adaptive procedure will also detect arms that by chance lead to a higher sample mean. In such a case, we can identify these "noisy" arms also and sample more to "de-noise" these arms. We note that this two-stage adaptive scheme does not utilize the full potential of the ability to be adaptive, but we show that even a simple adaptive scheme such as this can lead to insightful gains and conclusions. We formally summarize the two stage adaptive sampling scheme in Definition 3.2.

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

$$X_i \overset{iid}{\sim} \mathcal{M}(q), \text{ for } 1 \le i[n\epsilon];$$
$$Y_i \overset{iid}{\sim} f_Q(x_i).$$

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

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

in which the superscript "F" stands for the first stage. Third, we calculate a re-weighting vector $Q \in \mathbb{R}^p$ as a function of $\bar{Y}_i^{\text{F}}$'s,

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

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

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

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 X for the second stage. For example if $f(x) = exp(x)$, then this reweighs the probability by an exponential function. We also scale the re-weighting function by $\sqrt{n}$ because the signal arm will contain signal that decreases as a function of $\sqrt{n}$ as we describe now in the following section.

3.2 Theoretical Power Analysis Through Local Asymptotics

To theoretically characterize the power of the two sampling schemes, we use key ideas from the classical local asymptotic theory Le Cam (1956). We remark that for our setting we apply the classical local asymptotic theory to characterize the power of different sampling schemes as opposed to characterizing the distribution of different test statistics of the data from a fixed sampling scheme. In our asymptotic setting, we keep p fixed and let $n \to \infty$. To avoid the power degenerately approaching one, we scale our "signal strength" h proportionally to the standard parametric rate $n^{-1/2}$, formally

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

where h_0 is a positive constant.

As introduced in Definition 3.1, we first analyze the power under an *iid* sampling scheme 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 H_1^{NMM},

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

Equivalently,

$$Y_i = S_i\left(W_i + \frac{h_0}{\sqrt{n}}\right) + (1 - S_i)G_i,$$

where W_i and G_i are standard normal random variables, $S_i \coloneqq \mathbb{1}_{X_i=1} \sim \text{Bernoulli}(\text{q}_1)$ and all W_i's, G_i's and S_i's are independent.

Following the CRT procedure in Section 1.3, since there is no Z to condition on, the fake resample copies, $\{\tilde{\mathbf{X}}^b\}_{b=1}^B$, are generated independently from exactly the same distribution as $\mathbf{X}$,

$$\tilde{X}_i^b \overset{\text{i.i.d}}{\sim} \mathcal{M}(q).$$

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

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

Note that another natural testing statistic $\bar{Y}$ is degenerate in our RT framework, since it does not even depend on $\mathbf{X}$ or $\tilde{\mathbf{X}}$. For the sake of notation simplicity, we define the following *re-sampled test statistic*

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

in which, formally speaking, $\tilde{\mathbf{X}} = (\tilde{X}_1^1, \ldots, \tilde{X}_n^1) \coloneqq \tilde{\mathbf{X}}^{\mathbf{1}}$ and readers should comprehend $\tilde{\mathbf{X}}$ as a generic copy of $\tilde{\mathbf{X}}^b$. We show in the Appendix that as $B \to \infty$ the power of testing H_1 against H_0 is equal to

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

As formally shown in the Appendix, one can explicitly derive the joint asymptotic distributions of $\bar{Y}_j$'s, $\tilde{\bar{Y}}_j$'s and $\bar{Y}$ under the alternative H_1. Consequently, we state the first main theorem of this section which characterize asymptotic power of *iid* sampling schemes under RT with test statistic T, as defined in Equation 8.

Theorem 3.1 (Normal Means Model: Power of RT under *iid* sampling schemes)**.** Upon taking $B \to \infty$, the asymptotic power of the *iid* sampling scheme 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

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

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

$$\begin{aligned} T_{\mathrm{ind}} =& T_{\mathrm{ind}}(q, G, H) \\ \coloneqq& \max\left(\left\{H_1 + h_0\right\} \cap \left\{H_j, j = 2, \ldots, p-1\right\} \cap \left\{-\frac{1}{q_p}\sum_{i=1}^{p-1} q_j H_j\right\}\right) \end{aligned} \tag{10}$$

and

$$\begin{aligned} \tilde{T}_{\mathrm{iid}} =& \tilde{T}_{\mathrm{iid}}(q, G, H) \\ \coloneqq& h_0 q_1 + \max\left(\left\{G_j, j = 1, \ldots, p-1\right\} \cap \left\{-\frac{1}{q_p}\sum_{j=1}^{p-1} q_j G_j\right\}\right). \end{aligned} \tag{11}$$

Matrices Σ_0 and D are defined as

$$\Sigma_0(q) \coloneqq \begin{bmatrix} v(q_1) & -q_1 q_2 & -q_1 q_3 & \cdots & -q_1 q_{p-1} \\ -q_1 q_2 & v(q_2) & -q_2 q_3 & \cdots & -q_2 q_{p-1} \\ -q_1 q_3 & -q_2 q_3 & v(q_3) & \cdots & -q_3 q_{p-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -q_1 q_{p-1} & -q_2 q_{p-1} & -q_3 q_{p-1} & \cdots & v(q_{p-1}) \end{bmatrix} \in \mathbb{R}^{(p-1)\times(p-1)},$$

and

$$D(q) \coloneqq \mathrm{diag}(q_1, q_2, \ldots, q_{p-1}) \in \mathbb{R}^{(p-1)\times(p-1)}.$$

Finally,

$$\Sigma(q) \coloneqq D(q)^{-1} \Sigma_0(q) D(q)^{-1}. \tag{12}$$

Remark 2. The default choice of weight vector q should be $(1/p, 1/p, \ldots, 1/p)$ since typically the statistician has no prior information about which arm the signal might come from. This choice of q corresponds to uniform *iid* sampling.

Remark 3. Unlike the previous remark, suppose we are an oracle and know which arm the signal may come from under the alternative. Then a naive, but natural idea, would be to choose large values of q_1 to get more "signal" or richer "information", hoping to maximize testing power. As we will further demonstrate in the next subsection, it is true that choosing different values of q_1 might increase power. However, the optimizer $\hat{q}_1 := \arg\max_{q_1} \text{Power}_{\text{iid}}(q)$ is not necessarily larger than $1/p$, which means sometimes it is actually better to 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 (between T and $\tilde{T}$) and their variances, which can also be observed through comparing Equation 10 and Equation 11.

Remark 4. Following the previous remark, it is again naive but natural to propose an adaptive procedure that should up-weight or down-weight the signal arm according to the oracle q_1. However, Section 3.3 shows this intuition is not always true as the adaptive procedure can do much better than even the oracle *iid* sampling procedure.

Remark 5. 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 Σ will be generally much larger than the off-diagonal elements. Consequently G and H will have approximately independent coordinates and both $T_{\text{ind}}, \tilde{T}_{\text{iid}}$ are characterized by nearly independent Gaussian distribution.

By an argument similar to proof for Theorem 3.1, we can also derive the asymptotic power for our two-stage adaptive sampling schemes.

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

$$\text{Power}_{\text{adap}}(\epsilon, t, f) := \mathbb{P}_{R^{\text{F}}, G^{\text{F}}, R^{\text{S}}, H^{\text{F}}} \left(\mathbb{P} \left(T_{\text{adap}} \geq z_{1-\alpha}(\tilde{T}_{\text{adap}} \mid R^{\text{F}}, R^{\text{S}}, H^{\text{F}}, H^{\text{S}}) \mid R^{\text{F}}, R^{\text{S}}, H^{\text{F}}, H^{\text{S}} \right) \right) \tag{13}$$

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

$$T_{\text{adap}} = \max_{j \in \{1,2,\ldots,p\}} T_{\text{adap},j}$$

$$\tilde{T}_{\text{adap}} = \max_{j \in \{1,2,\ldots,p\}} \tilde{T}_{\text{adap},j}$$

$$T_{\text{adap},j} = \frac{q_j \sqrt{\epsilon} W_j + Q_j \sqrt{(1-\epsilon)} \left[H_j^{\text{S}} + R^{S} + \mathbb{1}_{j=1} \sqrt{1-\epsilon} h_0 \right]}{\epsilon q_j + (1-\epsilon) Q_j}$$

$$\tilde{T}_{\text{adap},j} = \frac{q_j \sqrt{\epsilon} \tilde{W}_j + \tilde{Q}_j \sqrt{(1-\epsilon)} \left(G_j^{\text{S}} + R^{\text{S}} + \sqrt{1-\epsilon} h_0 Q_1 \right)}{\epsilon q_j + (1-\epsilon) \tilde{Q}_j}$$

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. Recall the definition of $\Sigma(\cdot)$ in Equation 12. First, generate $R^{\text{F}} \sim \mathcal{N}(0,1)$, $G^{\text{F}} \sim \mathcal{N}(0, \Sigma(q))$,

and $H^{\mathrm{F}} \sim \mathcal{N}(0, \Sigma(q))$ independently. Second, compute

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

Third, compute

$$
\begin{aligned}
Q_j &= \frac{f(W_j/\sqrt{\epsilon})}{\sum_{j=1}^{p} f(W_j/\sqrt{\epsilon})}, \\
\tilde{Q}_j &= \frac{f(\tilde{W}_j/\sqrt{\epsilon})}{\sum_{j=1}^{p} f(\tilde{W}_j/\sqrt{\epsilon})}.
\end{aligned}
$$

Lastly, generate $R^{\mathrm{F}} \sim \mathcal{N}(0,1)$, $H^{\mathrm{S}} \sim \mathcal{N}(0, \Sigma(Q))$ and $G^{\mathrm{S}} \sim \mathcal{N}\left(0, \Sigma(\tilde{Q})\right)$ independently.

Remark 6. Unfortunately, unlike Theorem 3.1, though we are able to carry out asymptotic analysis for two-stage adaptive procedures, due to the complicated nature of both our "maximum" test statistic and the adaptive way of sampling, the final formula for asymptotic power, namely Equation 13, is not immediately as insightful as one would hope for.

Remark 7. Though Theorem 3.2 is not directly interpretable, the computational cost of evaluating it numerically is less than simulating the adaptive procedure for a large value of n by a factor of $O(n)$. Moreover, since the asymptotic power characterized in Theorem 3.1 and Theorem 3.2 does not depend on n, the conclusion is naturally more consistent and unified comparing to simulating with different large n's.

Remark 8. Apart from the computational advantages it provides, Theorem 3.2 is also of theoretical interest by itself, since local asymptotic power analysis is performed over different testing statistics while we are using it to analyse and compare different sampling strategies. In addition, it can also serve as a starting point and motivating example for analyzing more general adaptive procedures and potentially inference on sequentially/adaptively sampled data in general.

3.3 Power Advantage of AdapRT

Given the theory presente in the previous section, we now attempt to understand how AdapRT may be more powerful than the *iid* sampling scheme. A natural starting point for coming up with a reasonable adaptive strategy is to consider the "oracle" *iid* sampling scheme The oracle *iid* sampling scheme can also serve as an important benchmark. As alluded in Remark 3, if a statistician knows which arm the signal might come from, then he or she would naturally think of the possibility of increasing power by sampling more (or less) from that arm, essentially treating that arm differently. We formally define the oracle in the following way, where we assume, without loss of generality, $j^\star = 1$,

$$
q_1^\star \coloneqq \arg\max_{0 \le q_1 \le 1} \mathrm{Power}_{\mathrm{iid}}(q(q_1)),
$$

in which $q(q_1) \coloneqq \left(q_1, (1-q_1)/(p-1), (1-q_1)/(p-1), \ldots, (1-q_1)/(p-1)\right) \in \mathbb{R}^p$ denotes the sampling probabilities of all p arms, where the first signal arm has probability q_1 and the remaining arms (that all have

no signal) equally share the remaining sampling probability. Let $q^\star = q(q_1^\star)$. From now on, we will use "oracle *iid* sampling scheme" to refer to *iid* sampling with weight vector $q^\star$. Although $q^\star$ is not technically the optimal *iid* sampling scheme for any possible *iid* sampling scheme since we consider the maximum power when only vaarying q_1 and the remaining arms to all equally share the remaining probability, 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 signal can not be differentiated in any way.

Next, we use numerical evaluations of Theorem 3.1 and Theorem 3.2 to compare the testing power of (two-stage) AdapRT, uniform *iid* sampling and the oracle *iid* sampling scheme $q^\star$ across a wide range of possible signal strengths h_0 and number of arms p. For the AdapRT procedure described in Definition 3.2, we choose the re-weighting function f to be the exponential function, i.e., $f(x) = exp(x)$.

To produce Figure 3, we first fix an arbitrary, but reasonable, combination of hyper-parameters for the AdapRT, i.e., we set $\epsilon = 0.5, t_0 = \log 2$ and $t = t_0/h_0$. We do this to demonstrate AdapRT's (empirical) robustness as we present the preliminary power results without necessarily changing the AdapRT adaptive parameters according to different values of h_0 and p. As a reminder, $\epsilon = 0.5$ means we spend half of our sampling budget on exploration and only adapt once by re-weighting (see Definition 3.2) after the first half of samples are collected. The choice of $t_0 = \log 2$ allows the first arm (containing the real signal) will get roughly twice more sampling weight than the remaining arms in the second stage. The left panel of Figure 3 shows that AdapRT is almost uniformly better than the classical/default uniform *iid* sampling. For example, in areas that have high number of arms and signal, the AdapRT can have close to 10% more power than the uniform *iid* sampling scheme. The right panel of Figure 3 surprisingly shows that the AdapRT can beat the even oracle *iid* sampling scheme when signal strength is relatively high. We see this power difference between the oracle *iid* sampling scheme can be as big as 10% when the signal and number of arms are high. However, we do see the AdapRT losing to the oracle *iid* sampling scheme when the signal is low. We explain further in Section 3.4 how and why the AapRT is helping in power. We note that for both panels in Figure 3, the top left corners of the heatmaps have zero difference btween the two sampling schemes because this strong signal low p regime have close to a degenerate power of one, allowing no significant differences to show.

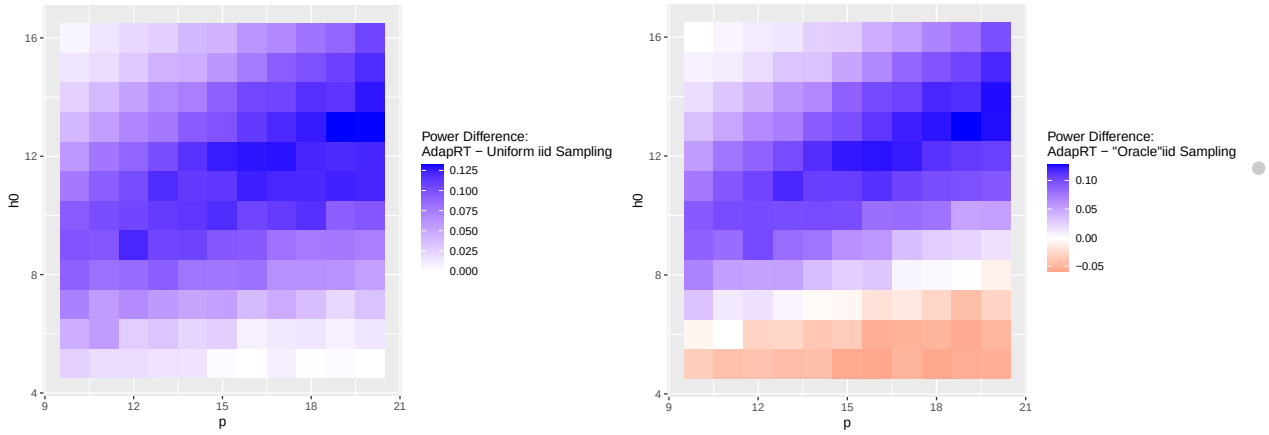

Figure 3: The figure shows difference between local asymptotic power of the AdapRT (with a rather arbitrary choice of hyper-parameters $\epsilon = 0.5$ and $t = \log 2/h_0$) and *iid* sampling for different values of signal strength h_0 and number of arms p. The test statistic is the same "maximum of means" statistic as defined in Equation 8. The left plot showcases the fact that AdapRT is almost uniformly better than classical/default uniform *iid* sampling. The right plot shows that AdapRT can beat the oracle *iid* sampling procedure when signal strength is relatively high. Note that values on the top left corners of both heatmaps are close to 0 only because power of all three sampling schemes is almost 1 in this region and have no significant difference. We chose significance level $\alpha = 0.05$.

Figure 3 already shows how the AdapRT for an arbitrary choice of adaptive parameters can be uniformly more powerful than a typical *iid* sampling scheme. We note that Figure 3 does indeed demonstrate the pipeline propoosed in Algorithm 2. Since Figure 3 tests one specific adaptive procedure against the *iid* sampling scheme for a variety of f_Q, the practitioner can now be comfortable to use the proposed AdapRT with adaptive parameters chosen to $\epsilon = 0.5$ and $t = \log 2/h_0$ to run their experiment. However, to further optimize for multiple adaptive procedures A as shown in Algorithm 1, we additionally create Figure 7 to explore the different adaptive parameters that may lead to a more optimal adaptive procedure. Furthermore, since the practitioner will most likely know the number of arms he/she has for the respective problem, we fix p and vary the value of the signal.

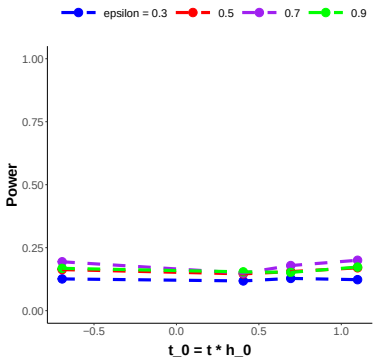

Figure 4: $p = 15, h = 6$

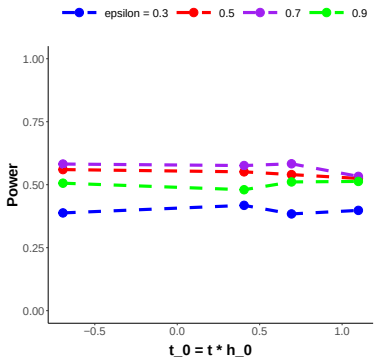

Figure 5: $p = 15, h = 10$

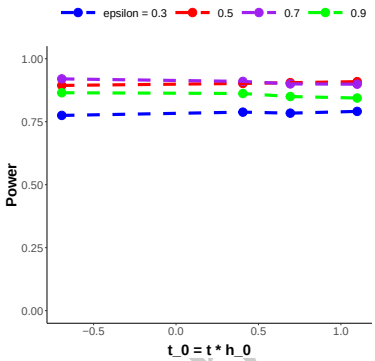

Figure 6: $p = 15, h = 14$

Figure 7: The figure above showcases the pipeline (see Section 2.5) of how to use pre-study simulation to choose among different possible adaptive schemes, or in this particular example, different combinations of hyper-parameters.

This is still an incomplete draft, more details will be added to this subsection.

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 why adapting could help so practitioners can understand the main intuition needed to build their own successful adaptive scheme. We note that all statements here are respect to the specific normal-means model setting, but we believe that the main intuition should carry to many different applications and scenarios. We further demonstrate how this intuition may carry to different scenarios in Section 4

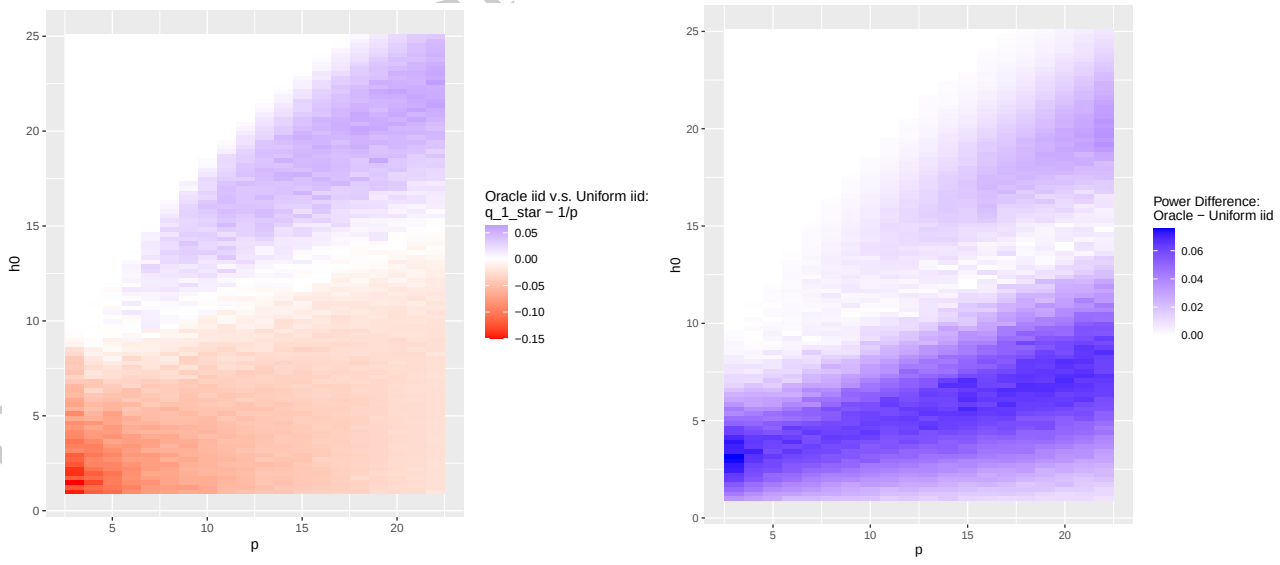

Figure 8: This figure shows whether the oracle $q_1^\star$ should down-weight (less than $\frac{1}{p}$) or up-weight (more than $\frac{1}{p}$) the signal arm.

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, which indeed is a useful starting point. However, the power gain shown in Figure 3 can not be attributed to only mimicing the oracle *iid* sampling scheme. Because, as shown in Figure 3, AdapRT can actually beat oracle *iid* sampling as long as the signal strength is not too low. Moreover, as demonstrated in Figure 8, whether $q_1^\star$ is smaller or larger than $1/p$ actually depends on both h_0 and p, which means if the only piece of intuition is to mimic the oracle, sometimes (roughly the red region plot on the left in Figure 8) one should choose a adaptive procedure that down-weights the arms with large means. However, by comparing Figure 3 and plot on the left in Figure 8, we see that our up-weighting (since $t > 0$) adaptive procedure can actually beat not only uniform *iid* but also oracle *iid* in parts of the region that we are supposed to down-weight.

Instead, we believe the main intuition behind the success of AdapRT is for the following three reasons. As expected, the first two reasons is that the AdapRT is sampling more from the arms with signal and less from the arms that clearly do not look like signal. However, when the AdapRT is sampling from the arms that look like it has signal it is not only sampling from the arms that is truly the real signal but also the arms that are "fake" signals but may look like real signal due to chance. Therefore, the third and most crucial reason is that AdapRT will also sample more from these noisy "fake" signal arms and stabilize these arms to show that it is indeed not a signal.

This is still an incomplete draft, more details will be added to this subsection.

4 AdapRT in Conjoint Studies

For this section, we turn to a completely different setting and demonstrate how AdapRT can help in a popular factorial design - conjoint analysis. Conjoint analysis, introduced more than half a century ago (Luce and Tukey, 1964), is a factorial survey-based experiment designed to measure preferences on a multidimensional scale. Conjoint analysis has been extensively used by marketing firms to determine desirable product characteristics (e.g., Bodog and Florian, 2012; Green, Krieger and Wind, 2001). Recently, it has gained popularity among social scientists (Hainmueller, Hopkins and Yamamoto, 2014; Raghavarao, Wiley and Chitturi, 2010) who are interested in studying individual preferences concerning elections (e.g., Ono and Burden, 2018), immigration (e.g., Hainmueller and Hopkins, 2015), employment (e.g., Popovic, Kuzmanovic and Martic, 2012), and other issues. Recently Ham, Imai and Janson 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.

Following the guideline proposed in Algorithm 2, we first perform our power analysis through simulations in Section 4.1 to show how the proposed adaptive procedure can be helpful in a conjoint setting. Unlike the adaptive procedure proposed in Section 3, we do not theoretically characterize the power but consider a more realistic fully adaptive procedure and complicated test statistic. We then apply our proposed methodology on a recent conjoint study that studies whether a political candidate gender's matter in voting behavior given the candidate's age, experience, etc (Ono and Burden, 2018). We show how the proposed adaptive procedure in Section 4.1 leads to a greater power than that of an *iid* sampling scheme by "boostrapping" from the original sample.

4.1 Conjoint Analysis

In a typical conjoint design, respondents are forced to choose between two profiles presented to them - often known as a forced-choice conjoint design (Ham, Imai and Janson, 2022; Hainmueller, Hopkins and Yamamoto, 2014; Ono and Burden, 2018). We refer to the two profiles as the "left" (L) and "right" (R)

profiles[2]. In this forced-choice design, the response Y is a binary variable that takes value 1 if the respondent chooses the left profile and zero otherwise. For simplicity, we consider the case when there are only two factors that are randomized. The first factor, X, is our factors of interest (for example the candidate's gender) and the second factor, Z, is a control factor (for example the candidate's political party). Although both X and Z are single factors, both of them two-dimensional because each X and Z consists of the left and right profile values. More formally, each respondent t observes $X_t = (X_t^L, X_t^R)$ and $Z_t = (Z_t^L, Z_t^R)$, where the superscripts L and R denote the left and right profiles respectively. In our simulation setup, we allow both factors of X and Z to have up to four levels, i.e., all $X_t^L, X_t^R, Z_t^L, Z_t^R$ take values $1, 2, 3, 4$. The response Y is then generated from the following logistic regression model with main effects and interactions on only one specific level,

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

where the first four indicators force main effects β_X, β_Z of X and Z, respectively, to exist in only the first levels of each factor and the last indicator force an interaction effect β_{XZ} in the first and second level of factor X, Z respectively. We purposefully choose the interaction to act on a different level of Z as to not "stack" up the interaction with the main effects. Lastly, our response model f_Q assumes "no profile order effect", which is commonly assumed in conjoint analysis (Hainmueller, Hopkins and Yamamoto, 2014; Ham, Imai and Janson, 2022). The "no profile order effect" assumption states that whether the left profile was on the right or vice-versa does not matter. We see this is explicitly written in the above response model f_Q as we repeat all main and interaction effects symmetrically for the right and left profile (except we shift the sign because $Y = 1$ refers to the left profile being selected). We later incorporate this information into the test statistic.

4.2 The Adaptive Procedure

To give intuition on why adapting may help, we first consider the typical uniform *iid* sampling scheme, where all levels for each factor are sampled with equal probability. If our sample size n is not sufficiently large enough, we may end up with insufficient samples for levels of X with signal and may by chance have levels that look like "fake" signal due to noise. Given exactly the same sample size n, adapting can mitigate such issues by allowing the sampling scheme to "screen out" levels that definitely do not look like signal, thus allocating the remaining samples to explore the more noisy levels that may not be true signals. Additionally, the adaptive procedure should ideally also identify levels of X that do contain signal and sample more from it to clearly separate out a strong signal with noise. Therefore, the main idea for why an adaptive procedure can be helpful here is very similar to Section 3.

To capture this intuition, we sample $X_t \sim \text{Multinomial}(p_{t,1}^X, p_{t,2}^X, \ldots, p_{t,K^2}^X)$, where $p_{t,j}^X$ represents the probability of sampling the jth arm out of K^2 possible arms and K is the total factor levels of X. For example, in our simulation setup $K = 4$ and we have 16 possible arms, $(1,1), (1,2)$, etc., and $p_{t,j}^Z$ is defined similarly. Our goal is to come up with an adaptive procedure that assigns weights $p_{t,j}^X$ and $p_{t,j}^Z$ that may help increase power compared to the power from a uniform *iid* sampling scheme, where $p_{t,j}^X = \frac{1}{K^2}, p_{t,j}^Z = \frac{1}{L^2}$ for every j and L is the total number of factor levels for factor Z. We also note that real conjoint applications

[2]The profiles are not necessarily always presented side by side.

Algorithm 3: Adaptive Procedure for Conjoint Studies

Given adaaptive parameter ϵ **for** $t = 1, 2, \ldots, [n\epsilon]$ **do**

Sample $X_t \sim \text{Multinomial}(p^X_{t,1}, p^X_{t,2}, \ldots, p^X_{t,K^2})$, where $p^X_{t,j} = \frac{1}{K}$ for all $j = 1, 2, \ldots, K^2$

Sample $Z_t \sim \text{Multinomial}(p^Z_{t,1}, p^Z_{t,2}, \ldots, p^Z_{t,L^2})$, where $p^Z_{t,j} = \frac{1}{L}$ for all $j = 1, 2, \ldots, LL^2$

for $t = [n\epsilon] + 1, \ldots, n$ **do**

Sample $X_t \sim \text{Multinomial}(p^X_{t,1}, p^X_{t,2}, \ldots, p^X_{t,K^2})$, where $p^X_{t,j}$ is given in Equation 14

Sample $Z_t \sim \text{Multinomial}(p^Z_{t,1}, p^Z_{t,2}, \ldots, p^Z_{t,L^2})$, where $p^Z_{t,j}$ is given in Equation 14

do indeed use the uniform *iid* sampling scheme (or a very minor variant from it) (Hainmueller and Hopkins, 2015; Ono and Burden, 2018). Although we present our adaptive procedure when the dimension of Z is only one (typical conjoint analysis have 8-10 othehr 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,

$$p^X_{t,j} \propto |\bar{Y}^X_{j,t} - 0.5| + |N(0, 0.01^2)| \qquad p^Z_{t,j} \propto |\bar{Y}^Z_{j,t} - 0.5| + |N(0, 0.01^2)|, \tag{14}$$

where $\bar{Y}^X_{j,t}$ denotes the sample mean of $Y_1, Y_2, \ldots, Y_{t-1}$ for arm j, $\bar{Y}^Z_{j,t}$ 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 14 are drawn independently). To give intuition why Equation 14 may be a reasonable adaptive procedure, consider what we expect to observe when jth arm is $X = (1, 1)$. In this case, there is no signal so all previous samples that have both left and right profiles with value $X = 1$ should have roughly $\bar{Y}^X_{j,t} = 0.5$. Therefore, in expectation we do not expect to sample more from this arm in the future. We add a slight perturbation in case $\bar{Y}^X_{j,t}$ is exactly equal to 0.5 to discourage an arm from having zero probability. As a numerical example consider the following four arms of X: (1,1), (1, 2), (1,3), (2,4). Suppose we observe at time t the corresponding sample means for the four arms as: $0.52, 0.60, 0.61, 0.59$ respectively. Under our simulation setup, X has a main effect in the first level so having an average of $0.60, 0.61$ for the second and third arm is reasonable. Additionally the first and last arm are "useless" combinations but by chance the last arm looks like a signal. After normalizing, our new adaptive probabilities will be roughly $0.063, 0.312, 0.343, 0.281$. We can see that the first arm, which we are quite certain is not a signal, is less likely to be sampled from, allowing the sampling scheme to use more budget on sampling the signal arms and the fourth "noisier" arm. This matches our intuition as we gain precision in the true signal while also allowing the noisy "fake" signals to stabilize. We finally note that the adaptive weights chosen in Equation 14 is only one of many different ways to adapt. We merely choose one naive and reasonable procedure for exposition and leave the theoretical characterization for optimal adaptive procedures under the conjoint setting to future research.

With this main re-weighting procedure, we build our adaptive procedure. Just like Definition 3.2, we also have an ϵ adaptive procedure parameter that denotes the beginning $[n\epsilon]$ samples that are used for "exploration". In this exploration stage, we sample from the typical uniform *iid* sampling scheme, i.e., $p^X_{t,j} = p^Z_{t,j}$ for all j and $t = 1, 2, \ldots, [n\epsilon]$. This is necessary because if we adapt right away we may end up with very high noise estimates of $\bar{Y}^X_{j,t}$, leading to unreasonable adaptive schemes. In the remaining samples, we adapt by changing the weights according to Equation 14. We note that this adaptive sampling scheme immediately satisfies Assumption 1 and also Assumption 2 since each variable only looks at its own history and previous responses (not the other variables' history). We summarize our adaptive procedure in Algorithm 3.

Lastly, in order for us to compute the p-value in Equation 5, we need a reasonable test statistic T. We emphasize that any test statistic leads to a valid finite-sample p-value. Although Ham, Imai and Janson consider a complicated Hierarchical Lasso model to capture all interactions, due to the simplicity of this

setting, we consider a simple cross-validated Lasso logistic test statistic that fits a Lasso logistic regression of $\mathbf{Y}$ with main effects of $\mathbf{X}$ and $\mathbf{Z}$ and their interaction. Further, to increase power we incorporate the common "no profile order effect" as done in (Hainmueller, Hopkins and Yamamoto, 2014; Ham, Imai and Janson, 2022), which states that the order of the pair of profiles do no matter, i.e., whether the left profile comes first or later. In other words, we do not expect the main effects corresponding to the left profile of X to be any different than the main effects corresponding to the right profile. To incorporate this symmetry constraint, we split our original $\mathbb{R}^{n\times 2\times 2+1}$ data matrix $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ into a new data matrix with dimension $\mathbb{R}^{n\times 2+1}$, 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 $1 - Y$ (see (Hainmueller, Hopkins and Yamamoto, 2014; Ham, Imai and Janson, 2022) for more details). This leads to the following test statistic

$$T^{\text{lasso}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) = \sum_{k=1}^{K-1} |\hat{\beta}_k| + \sum_{k=1}^{K-1} \sum_{l=1}^{L-1} |\hat{\gamma}_{kl}|, \tag{15}$$

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 levels of Z. Given this test statistic, we are now ready to calculate the p-value in Equation 5 for different settings of the simulation.

4.3 Simulation Results

We now compare the power of our adaptive procedure stated in Algorithm 3 with the *iid* setting where each arm for X and Z are drawn uniformly at random, i.e., $p^X_{t,j} = p^Z_{t,j} = 1/16$ for all j under the simulation setting described in Section 4.1. We empirically compute the power as the proportion of 1000 Monte-Carlo p-values less than $\alpha = 0.05$.

In accordance with the guideline presented in Algorithm 2, we simulate the power under different f_Q and different parameterizations of our proposed adaptive procedure. We first vary f_Q by increasing sample size when there exist both main effects and interaction effects of X in the left panel of Figure 9. For this simulation setting, we vary our sample size $n = (450, 750, 1000, 1200, 1500)$ and fix $\beta_X = \beta_Z = 0.6$ and $\beta_{XZ} = 1.1$. For the next setting, we consider only main effects of X and Z (no interaction effects) in the right panel of Figure 9. In this panel, we increase β_X and β_Z to $(0, 0.2, 0.4, 0.6, 0.8)$ and explore the power of the two procedures. To also test for various adaptive procedures, we vary the one natural adaptive sampling parameter ϵ in Algorithm 2 to $\epsilon = 0.25, 0.5, 0.75$.

Both panels of Figure 9 show that the adaptive power is uniformly dominating the uniform *iid* power (green) despite using the same test statistic and under the same simulation setting. For example when $n = 1000$ in the left panel, there is a difference in 10 percentage points (74% versus 84%) between the *iid* sampling scheme and the adaptive sampling scheme with $\epsilon = 0.5$ (red). When the main effect is as strong as 0.5 in the right panel, there is a difference in 26 percentage points (36% versus 62%) between the *iid* sampling scheme and the adaptive sampling scheme with $\epsilon = 0.5$. Additionally, when the main effect is 0 in the right panel (in this case we are under H_0), Figure 9 shows that the power of all methods, as expected, has type-1 error control as both powers are near $\alpha = 0.05$ (the black horizontal line shows the $\alpha = 0.05$ line). Lastly, we note that both $\epsilon = 0.25, 0.5$ are suitable choices for the adaptive procedure as the power when $\epsilon = 0.75$ (purple) seems to be slightly lower than that of the other adaptive power curves (although still higher than the power curve of the *iid* sampling scheme).

4.4 Application: Role of Gender in Political Candidate Evaluation

We now apply our proposed method to a recent conjoint study which examines whether voters prefer candidates of one gender over the other after controlling for other candidate characteristics (Ono and Burden,

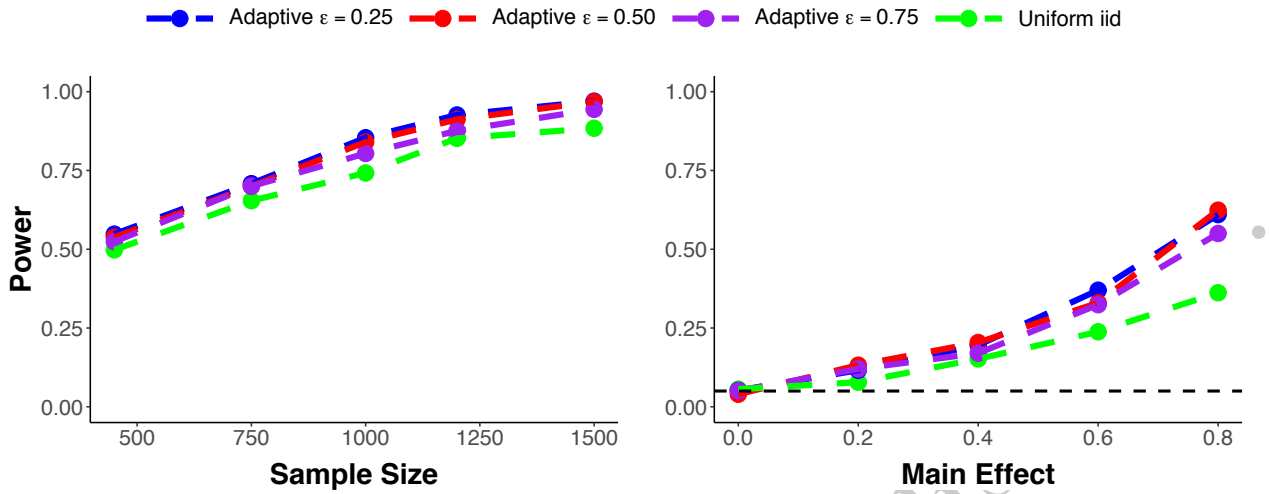

Figure 9: The figure shows how the power of the Adaptive and i.i.d based CRT tests varies as the sample size increases (left plot) or main effect increases (right plot). All power curves are calculated from 1000 Monte-Carlo calculated p-values using Equation 5 with $B = 300$ and test statistic given in Equation 15. The blue, red, and purple power curves denote the power of the AdapCRT using adaptive procedure described in Algorithm 3 and $\epsilon = 0.25, 0.50, 0.75$ respectively. The green power curve denotes the power of the uniform *iid* sampling scheme. 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.

2018). 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 independently randomized levels of: gender, age, race, family, experience in public office, salient personal characteristics, party affiliation, policy area of expertise, position on national security, position on immigrants, position on abortion, position on government deficit, and favorability among the public (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 15,830 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).

Ono and Burden were able to find a weak statistically significant effect of gender (p-value of around 0.04). Furthermore the authors were unable to find a statistically significant effect of gender when considering only Congressional candidates (as opposed to Presidential candidates), which accounted for half the data. Recent results by Ham, Imai and Janson show that gender still matters even for Congressional candidate. We attempt to answer this important question of whether gender matters in voting behavior (for all candidates) had the experimenter ran the experiment for the first time but with a lower sample size or budget $n < 15,830$.

To run this quasi-experiment, we pretend the original data of size $15,830$ is the population and we draw samples (with replacement) from the original dataset according to our experiment. Due to the similarity with the popular boostrap procedure (Efron, 1979), we refer to this quasi-experiment scheme as "bootstrapping" the data. Our bootstrapping approach aims to recreate an experimental setting had Ono and Burden ran this experiment again with the same objective to test H_0 with X being gender but now allowed to use an adaptive scheme to sample (X, Z) instead of a uniform *iid* scheme. As reference, the original experimental design was independently and identically uniformly sampled factors for all factors in the experiment. For example, the left and right profiles' gender was either "Male" or "Female" with equal probability.

The "bootstrapping" procedure is as follows. 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 $X_1 = (\text{Male}, \text{Female})$

	iid CRT	AdapCRT
$n = 2,000$	0.14	0.15
$n = 5,000$	0.25	0.33
$n = 7,500$	0.36	0.43
$n = 10,000$	0.43	0.51
$n = 12,500$	0.51	0.62

Table 1: The two columns represent the power of the *iid* CRT and the Adaptive CRT respectively for testing H_0, where X is gender in the gender political candidate study in (Ono and Burden, 2018) and Z is the the candidate's position on abortion. Each row represents a different bootstrapped sample size n. The power is calculated from the proportion of 1000 p-values less than $\alpha = 0.1$. Each p-value is calculated using Equation 5 using the appropriate resamples for the corresponding procedure with test statistic in Equation 15. The Adaptive CRT uses adaptive procedure in Algorithm 3.

and $Z_1 = (\text{Democrat}, \text{Democrat})$, indicating the left profile was a Democratic male candidate and the right profile was Democratic female candidate. If we draw such a sample, we can then obtain the response drawing a response Y from the original study of 15,830 samples that also had a pair of profiles with a Democratic male candidate and a Democratic female candidate. Since Z in the original study contained 12 other factors, the probability of observing a unique sequence of a particular (X, Z) is close to zero due to the curse of dimensionality. For example, if Z contained only two more factors such as candidate age and experience in public office, then there may exist no samples in the original study that contain a specific profile that is a Democratic male with 20 years of experience in public office and 50 years of age. For this reason, we only "bootstrap" the available data up to one other Z, namely the candidate's position on abortion. We choose this variable because Ham, Imai and Janson suggests possible strong interactions of gender with the candidate's position on abortion. Since our aim is to show that adapting can help achieve a greater power than that of the *iid* procedure, it is sensible to try to use other factors Z that may help in power as long as both experimental procedures use the same test statistic and variables.

Given a budget constraint $n < 15,830$, we compute the power of the adaptive sampling scheme and the *iid* sampling scheme by computing 1000 p-values using n bootstrapped sample of the original 15,830 sample, where the p-value is computed using Equation 5 using the appropriate resamples for the corresponding procedure and 1000 p-values are computed from 1000 possible bootstrapped samples of the original data. We then empirically compute the proportion of the 1000 p-values less than $\alpha = 0.1$. Since the applied setting is exactly the same as that of the simulation setting in Section 4.1, we use the same adaptive procedure in Algorithm 3 with $\epsilon = 0.5$ as suggested by Section 4.3 and the same test statistic in Equation 15. We also similarly impose "no profile order effect" in the test statistic as described in Section 4.2.

Table 1 shows the power results using both the *iid* CRT procedure described in Section 1.3 and our proposed AdapRT. As a reminder, the original experimental design in (Ono and Burden, 2018) performed an *iid* sampling scheme where each factor levels for every factor had equal weights independent of all other factors. Table 1 shows that the power of the AdapRT is consistently and non-trivially higher than that of the *iid* sampling scheme. For example, when $n = 5000$, we observe a power difference of 8 percentage points with the *iid* sampling scheme only having 25% power, leading to approximately a 30% increase of power. Since the adaptive procedure intuition is similar to the one proposed in the normal-means model, the same reasoning can be explained to describe the success of this particular adaptive procedure. In other words, the adaptive procedure was able to focus more on the arm of both X and Z that had potential signal and less on the uninformative arms such as the arm that compares the same gender for both profiles. More crucially, the adaptive procedure was able to stabilize and "de-noise" the more noisy arms that could have contained "fake" signals by chance.

5 Concluding Remarks

In this paper, we introduce the Adaptive randomization test (AdapRT) that allows the "Model-X" randomization inference approach for sequentially collected data. The AdapRT, like the CRT, tackles the fundamental independence testing problem in statistics. We show through various simulations and empirical examples how a sequential sampling scheme can lead to a more power test compared to the typical *iid* sampling scheme. In particular, we demonstrate the AdapRT's advantages in the normal-means model and conjoint settings. We find that adaptively sampling can help for three main reasons. The first two reasons relate to how the AdapRT allows the experimenter to sample more from arms with potential signals. This allows the AdapRT to not only sample arms that contain the true signal but also arms that may initially look like signal but do not actually have any signal. This is still useful because sampling more from these "fake" signals stabilizes the signal. Thirdly, AdapRT also down-weights arms that (with high probability) contain no signal, allowing our remaining samples to focus on exploring the more relevant arms.

Our work, however, is not comprehensive. While our work analyzes two common settings where the AdapRT is clearly helpful, there exist many future research that can further explore how to build efficient adaptive procedures with theoretical and empirical guarantees under many different scenarios for the respective application. Secondly, as briefly discussed in Section 2.3, AdapRT can tackle the problem of rendering multiple p-values at the same time, but it is not clear if one could make theoretical or empirical guarantees about its properties in the context of multiple testing and variable selection, like FDR control. Thirdly, with the goal of extending our methodology beyond independence testing, it would be interesting to explore possible ways to combine adaptive sampling with other ideas from "Model-X" framework. For instance, Zhang and Janson recently proposed a method called Floodgate that goes beyond independence testing by additionally characterizing the strength of the dependency. It would be interesting to incorporate our adaptive framework in this Floodgate setting. Lastly, our AdapRT is crucially reliant on the natural adaptive resampling procedure (NARP) for the validity of the p-values in p_{AdapCRT}. As mentioned in Section 2.4, it may be possible to also have a feasible resampling procedure that does not require Assumption 1 but enjoys the same benefits of AdapCRT.

References

Arrow, Kenneth J. 1998. "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives* 12:91–100.

Bates, Stephen, Matteo Sesia, Chiara Sabatti and Emmanuel Candès. 2020. "Causal inference in genetic trio studies." *Proceedings of the National Academy of Sciences* 117:24117–24126.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57:289–300.

Bodog, Simona and G.L. Florian. 2012. "Conjoint Analysis in Marketing Research." *Journal of Electrical and Electronics Engineering* 5:19–22.

Bojinov, Iavor and Neil Shephard. 2019a. "Time Series Experiments and Causal Estimands: Exact Randomization Tests and Trading." *Journal of the American Statistical Association*.

Bojinov, Iavor and Neil Shephard. 2019b. "Time Series Experiments and Causal Estimands: Exact Randomization Tests and Trading." *Journal of the American Statistical Association* 114:1665–1682.
URL: *https://doi.org/10.1080/01621459.2018.1527225*

Candès, Emmanuel, Yingying Fan, Lucas Janson and Jinchi Lv. 2018. "Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection." *Journal of the Royal Statistical Society: Series B* 80:551–577.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7:1 – 26.
URL: *https://doi.org/10.1214/aos/1176344552*

Green, Paul, Abba Krieger and Yoram Wind. 2001. "Thirty Years of Conjoint Analysis: Reflections and Prospects." *Interfaces* 31:S56–S73.

Hainmueller, Jens and Daniel J. Hopkins. 2015. "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants." *American Journal of Political Science*.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22:1–30.

Ham, Dae Woong, Kosuke Imai and Lucas Janson. 2022. "Using Machine Learning to Test Causal Hypotheses in Conjoint Analysis.".

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

James, W and C Stein. 1961. "Estimation with quadratic loss Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Berkeley.".

Le Cam, Lucien. 1956. 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*. University of California Press pp. 129–156.

Luce, R.Duncan and John W. Tukey. 1964. "Simultaneous conjoint measurement: A new type of fundamental measurement." *Journal of Mathematical Psychology* 1:1 – 27.

Lupia, Arthur and Mathew Mccubbins. 2000. "The Democratic Dilemma: Can Citizens Learn What They Need to Know?" *The American Political Science Review* 94.

Ono, Yoshikuni and Barry C. Burden. 2018. "The Contingent Effects of Candidate Sex on Voter Choice." *Political Behavior*.

Popovic, Milena, Marija Kuzmanovic and Milan Martic. 2012. "Using Conjoint Analysis To Elicit Employers' Preferences Toward Key Competencies For A Business Manager Position." *Management - Journal for theory and practice of management* 17:17–26.

Raghavarao, D., J.B. Wiley and P. Chitturi. 2010. *Choice-based conjoint analysis: Models and Designs*. Chapman and Hall/CRC.

Shi, Chengchun, Wang Xiaoyu, Shikai Luo, Hongtu Zhu, Jieping Ye and Rui Song. 2022. "Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework." *Journal of the American Statistical Association*.

Skarnes, William, Barry Rosen, Anthony West, Manousos Koutsourakis, Wendy Roake, Vivek Iyer, Alejandro Mujica, Mark Thomas, Jennifer Harrow, Tony Cox, David Jackson, Jessica Severin, Patrick Biggs, Jun Fu, Michael Nefedov, Pieter de Jong, Adrian Stewart and Allan Bradley. 2011. "A conditional knockout resource for the genome-wide study of mouse gene function." *Nature* 474:337–42.

Slivkins, Aleksandrs. 2019. "Introduction to Multi-Armed Bandits." *Foundations and Trends® in Machine Learning* 12:1–286.
URL: *http://dx.doi.org/10.1561/2200000068*

Sutton, Richard and Andrew Barto. 2018a. *Reinforcement learning: an introduction. Adaptive Computation and Machine Learning*. MIT Press.

Sutton, Richard S. and Andrew G. Barto. 2018b. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book.

Thompson, William R. 1933. "ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES." *Biometrika* 25:285–294.
URL: *https://doi.org/10.1093/biomet/25.3-4.285*

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58:267–288.
URL: *http://www.jstor.org/stable/2346178*

Wainwright, Martin J, Michael I Jordan et al. 2008. "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning* 1:1–305.

Zhang, Lu and Lucas Janson. 2020. "Floodgate: inference for model-free variable importance." *arXiv preprint arXiv:2007.01283*.

6 Appendix

6.1 Proof of Main Results Presented in Section 2

Proof of Theorem 2.1. First of all, 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,

$$(\tilde{X}_{1:k}, Y_{1:k}, Z_{1:k}) \overset{\mathrm{d}}{=} (X_{1:k}, Y_{1:k}, Z_{1:k}). \tag{16}$$

Assuming Equation 16 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,

$$\begin{aligned}
&P\left[(\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right] \\
\overset{\text{(i)}}{=}&P\left[Z_k \mid (\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
&\cdot P\left[(\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
\overset{\text{(ii)}}{=}&P\left[Z_k \mid (Y_{1:(k-1)}, Z_{1:(k-1)}) = (y_{1:(k-1)}, z_{1:(k-1)})\right] \\
&\cdot P\left[(\tilde{X}_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
\overset{\text{(iii)}}{=}&P\left[Z_k \mid (Y_{1:(k-1)}, Z_{1:(k-1)}) = (y_{1:(k-1)}, z_{1:(k-1)})\right] \\
&\cdot P\left[(X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
\overset{\text{(iv)}}{=}&P\left[Z_k \mid (X_{1:(k-1)}Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
&\cdot P\left[(X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:(k-1)}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:(k-1)})\right] \\
=&P\left[(X_{1:(k-1)}, Y_{1:(k-1)}, Z_{1:k}) = (x_{1:(k-1)}, y_{1:(k-1)}, z_{1:k})\right],
\end{aligned} \tag{17}$$

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) 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 16 is thus established by induction, as a corollary of which, we also get for any $k \leq n$,

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

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 $\mathbb{P}(p < \alpha \mid Y, Z) \leq \alpha$ 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 17, which naturally ensures its necessity. □

Finally, we 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. That is equivalent to say as $B \to \infty$, conditioning on any given values of $(X, \mathbf{Y}, \mathbf{Z})$,

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

Lemma 6.1 (Power of AdapRT under $B \to \infty$)**.** For any adaptive sapling scheme A satisfies Definition 2.1 and any test statistic T, as we take $B \to \infty$, the asymptotic conditional power of AdapRT (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 inexplicitly specified by the sampling procedure A.

6.2 Proof of Results Presented in Section 3

Lemma 6.2 (Normal Means Model: Joint Asymptotic Distributions of $\bar{Y}_j$'s, $\tilde{\bar{Y}}_j$'s and $\bar{Y}$ Under the Alternative H_1)**.** Define

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

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 $G \coloneqq (G_1, G_2, \ldots, G_{p-1})$ and $H \coloneqq (H_1, H_2, \ldots, H_{p-1})$ both follow the same $(p-1)$ dimensional multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ and R is a standard normal random variable. Moreover, G, H and R are independent.

Remark 9. 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 re-sampling $\tilde{\mathbf{X}}$ given $\mathbf{Y}$.

Remark 10. Note that index of p is not included to avoid stating the convergence in terms of a degenerate multivariate Gaussian distribution.

Proof of Lemma 6.2. First of all, we try to characterize the conditional distribution of $\tilde{\bar{Y}}_j$. For any $j \in \{1, 2, \ldots, p\}$,

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

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

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

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

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

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

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

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

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

and

$$\Sigma_0 \coloneqq \mathrm{Var}\left(\left(\mathbb{1}_{\tilde{X}_i=1}, \mathbb{1}_{\tilde{X}_i=2}, \cdots, \mathbb{1}_{\tilde{X}_i=p-1}\right)\right) = \begin{bmatrix} v(q_1) & -q_1q_2 & -q_1q_3 & \cdots & -q_1q_{p-1} \\ -q_1q_2 & v(q_2) & -q_2q_3 & \cdots & -q_2q_{p-1} \\ -q_1q_3 & -q_2q_3 & v(q_3) & \cdots & -q_3q_{p-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -q_1q_{p-1} & -q_2q_{p-1} & -q_3q_{p-1} & \cdots & v(q_{p-1}) \end{bmatrix}. \tag{18}$$

By Multivariate Lindeberg-Feller CLT (reference needed here),

$$\sqrt{n}\bar{\Sigma}_n^{-1/2}\left(\bar{V} - \mathbb{E}\bar{V}\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, I_{p-1}\right). \tag{19}$$

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

$$\mathbf{J}_{-p,n} \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \Sigma\right), \tag{20}$$

where

$$\Sigma = D^{-1}\Sigma_0 D^{-1}$$

with

$$D = \mathrm{diag}(q_1, q_2, \cdots, q_{p-1}) \in \mathbb{R}^{(p-1)\times(p-1)}. \tag{21}$$

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

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

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

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

instead. Again, Multivariate Lindeberg-Feller CLT gives,

$$\sqrt{n}(\bar{\Sigma}_n^\star)^{-1/2}\left(\bar{V}^\star - \mathbb{E}\bar{V}^\star\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, I_{p-1}\right), \tag{22}$$

with

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

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

Note that, since $\lim_{n\to\infty} \mathrm{Var}\left(Y_i(\mathbb{1}_{X_i=j} - q_{i,j}^\star)\right) = q_j(1-q_j)$ and $\lim_{n\to\infty} \mathrm{Cov}\left(Y_i(\mathbb{1}_{X_i=j_1} - q_{i,j_1}^\star), Y_i(\mathbb{1}_{X_i=j_2} - q_{i,j_2}^\star)\right) = -q_{j_1}q_{j_2}$,

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

which further gives

$$\sqrt{n}\left(\bar{V}^\star - \mathbb{E}\bar{V}^\star\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \Sigma_0\right). \tag{23}$$

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

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

and

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

which together with Equation 23 gives

$$\mathbf{J}_{-p,n}^\star \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \Sigma\right). \tag{24}$$

Note that though Equation 20 and Equation 24 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 $\frac{\sum_{i=1}^n q_{i,j}^\star Y_i}{q_j\sqrt{n}}$, as demonstrated in Lemma 6.3, Lemma 6.4, Lemma 6.5 and Lemma 6.6. Roughly speaking, under this $\sqrt{n}$ scaling, the randomness that leads to the Gaussian noise part in CLT is the same across them, 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 $\mathbf{J}_{-p,n}$, $\mathbf{J}_{-p,n}^\star$ and $\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$. Letting

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

we have

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

□

Lemma 6.3. As $n \to \infty$,

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

Proof. By defining $E_i \coloneqq S_i W_i + (1 - S_i)G_i \sim \mathcal{N}(0,1)$, 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,

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

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

Lemma 6.4. As $n \to \infty$,

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

Proof. We first show

$$\lim_{n\to\infty} \mathbb{E}\left(\sqrt{n} q_{i,1}^{\star} Y_i\right) = h_0. \tag{25}$$

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

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

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

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{Var}(q_{i,1}^\star Y_i) &= \lim_{n\to\infty} \left\{ \mathbb{E}\left[(q_{i,1}^\star Y_i)^2 \right] - \left[\mathbb{E}(q_{i,1}^\star Y_i) \right]^2 \right\} \\
&= \lim_{n\to\infty} \mathbb{E}\left[(q_{i,1}^\star Y_i)^2 \right] \\
&= \lim_{n\to\infty} \int_{\mathbb{R}} y^2 \left[\frac{q_1 e^{-(y-h_0/\sqrt{n})^2/2}}{q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1)e^{-y^2/2}} \right]^2 \left[(1-q_1)\frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h_0/\sqrt{n})^2/2} \right] \mathrm{d}y \\
&= \int_{\mathbb{R}} \lim_{h\to 0} \left\{ y^2 \left[\frac{q_1 e^{-(y-h)^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} \right]^2 \left[(1-q_1)\frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h)^2/2} \right] \right\} \mathrm{d}y \\
&= \int_{\mathbb{R}} q_1^2 \frac{e^{-y^2/2}}{\sqrt{2\pi}} \mathrm{d}y \\
&= q_1^2.
\end{aligned} \tag{26}$$

Combining Equation 25 and Equation 26, 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 6.5. For $j \neq 1$, as $n \to \infty$,

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

Proof. We first show

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

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 $\mathbb{E}\left(\sqrt{n} q_{i,j}^{\star} Y_i\right)$ is equal to

$$\sqrt{n}\int_{\mathbb{R}} \frac{y q_j e^{-y^2/2}}{q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1)e^{-y^2/2}} \left[(1-q_1)\frac{1}{\sqrt{2\pi}} e^{-y^2/2} + q_1 \frac{1}{\sqrt{2\pi}} e^{-(y-h_0/\sqrt{n})^2/2}\right] \mathrm{d}y \coloneqq \boldsymbol{B}_0 + \boldsymbol{B}_1.$$

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

$$\begin{aligned}
\lim_{n\to\infty} \boldsymbol{B}_1 &= \frac{q_1 q_j \sqrt{n}}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{y e^{-y^2/2}}{q_1 e^{-(y-h_0/\sqrt{n})^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h_0/\sqrt{n})^2/2} \mathrm{d}y \\
&= \lim_{h\to 0} \frac{q_1 q_j h_0}{\sqrt{2\pi}} \left[\frac{1}{h}\int_{\mathbb{R}} \frac{y e^{-y^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2} \mathrm{d}y\right] \\
&= \frac{q_1 q_j h_0}{\sqrt{2\pi}} \left.\frac{\mathrm{d}\left[\int_{\mathbb{R}} \frac{y e^{-y^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2} \mathrm{d}y\right]}{\mathrm{d}h}\right|_{h=0} \\
&= \frac{q_1 q_j h_0}{\sqrt{2\pi}} \int_{\mathbb{R}} \left.\frac{\mathrm{d}\left[\frac{y e^{-y^2/2}}{q_1 e^{-(y-h)^2/2} + (1-q_1)e^{-y^2/2}} e^{-(y-h)^2/2}\right]}{\mathrm{d}h}\right|_{h=0} \mathrm{d}y \\
&= \frac{q_1 q_j h_0}{\sqrt{2\pi}} \int_{\mathbb{R}} (1-q_1) y^2 e^{-y^2/2} \mathrm{d}y \\
&= h_0 q_j q_1 (1-q_1).
\end{aligned}$$

Similarly,

$$\lim_{n\to\infty} \boldsymbol{B}_0 = -h_0 q_j q_1 (1-q_1).$$

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

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

Lemma 6.6. As $n \to \infty$,

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

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 6.4, it suffices to show

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

which can be established by the following three computations,

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

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

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

□