University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2010 *Paper* 258

Targeting The Optimal Design In Randomized Clinical Trials With Binary Outcomes And No Covariate

Antoine Chambaz* Mark J. van der Laan†

*Laboratoire MAP5, Université Paris Descartes and CNRS, achambaz@u-paris10.fr

†University of California - Berkeley, laan@berkeley.edu

Targeting The Optimal Design In Randomized Clinical Trials With Binary Outcomes And No Covariate

Antoine Chambaz and Mark J. van der Laan

Abstract

This article is devoted to the asymptotic study of adaptive group sequential designs in the case of randomized clinical trials with binary treatment, binary outcome and no covariate. By adaptive design, we mean in this setting a clinical trial design that allows the investigator to dynamically modify its course through data-driven adjustment of the randomization probability based on data accrued so far, without negatively impacting on the statistical integrity of the trial. By adaptive group sequential design, we refer to the fact that group sequential testing methods can be equally well applied on top of adaptive designs. Prior to collection of the data, the trial protocol specifies the parameter of scientific interest. In the estimation framework, the trial protocol also a priori specifies the confidence level to be used in constructing frequentist confidence intervals for the latter parameter and the related inferential method, which will be based on the maximum likelihood principle. In the testing framework, the trial protocol also a priori specifies the null and alternative hypotheses regarding the latter parameter, the wished type I and type II errors, the rule for determining the maximal statistical information to be accrued, and the frequentist testing procedure, including conditions for early stopping. Furthermore, we assume that the protocol specifies a user-supplied optimal unknown choice of randomization scheme, and we will focus on that randomization scheme which minimizes the asymptotic variance of the maximum likelihood estimator of the parameter of interest.

We obtain that, theoretically, the adaptive design converges almost surely to the targeted unknown randomization scheme. In the estimation framework, we obtain that our maximum likelihood estimator of the parameter of interest is a strongly

consistent estimator, and it satisfies a central limit theorem. We can estimate its asymptotic variance, which is the same as that it would feature had we known in advance the targeted randomization scheme and independently sampled from it. Consequently, inference can be carried out as if we had resorted to independent and identically distributed (iid) sampling. In the testing framework, we obtain that the multidimensional t-statistics that we would use under iid sampling still converges to the same canonical distribution under adaptive sampling. Consequently, the same group sequential testing can be carried out as if we had resorted to iid sampling. Furthermore, a comprehensive simulation study that we undertake validates the theory. It notably shows in the estimation framework that the confidence intervals we obtain achieve the desired coverage even for moderate sample sizes. In addition, it shows in the testing framework that type I error control at the prescribed level is guaranteed, and that all sampling procedures only suffer from a very slight increase of the type II error.

A three-sentence take-home message is: "Adaptive designs do learn the targeted optimal design and inference and testing can be carried out under adaptive sampling as they would under the targeted optimal randomization probability iid sampling. In particular, adaptive designs achieve the same efficiency as the fixed oracle design. This is confirmed by a simulation study, at least for moderate or large sample sizes, across a large collection of targeted randomization probabilities."

1 Introduction

This article is devoted to the asymptotic study of *adaptive group sequential designs* in the case of randomized clinical trials with binary treatment, binary outcome and no covariate. Thus, the experimental unit writes as $O = (A, Y)$ where the treatment A and the outcome Y are dependent Bernoulli random variables. Typical parameters of scientific interest are $\Psi_+ = E(Y|A = 1) - E(Y|A = 0)$ (additive scale) or $\Psi_\times = \log E(Y|A = 1) - \log E(Y|A = 0)$ (multiplicative scale, which we will consider hereafter). One can interpret causally such parameters whenever one is willing to postulate the existence of a full data structure $X = (X(0), X(1))$ containing the two counterfactual outcomes under the two possible treatments and such that $Y = X(A)$ and A independent of X. If so indeed, $\Psi_+ = E(X(1)) - E(X(0))$ and $\Psi_\times = \log E(X(1)) - \log E(X(0))$. Let us know explain what we mean by adaptive group sequential design.

1.1 The notion of adaptive group sequential designs.

By *adaptive design*, we mean in this setting a clinical trial design that allows the investigator to dynamically modify its course through data-driven adjustment of the randomization probability based on data accrued so far, without negatively impacting on the statistical integrity of the trial. This definition is slightly adapted from [10], the introductory article to the proceedings (to which many articles cited below belong) of a workshop entitled "Adaptive clinical trial designs: Ready for prime time?" held in October 2004, and jointly sponsored by the FDA and Harvard-MIT Division of Health Science and Technology. Using the definition of prespecified sampling plans given in [8], let us emphasize that we assume that, prior to collection of the data, the trial protocol specifies: the parameter of scientific interest, and

- estimation framework: the confidence level to be used in constructing frequentist confidence intervals for the latter parameter, the related inferential method;
- testing framework: the null and alternative hypotheses regarding the latter parameter, the wished type I and type II errors, the rule for determining the maximal statistical information to be accrued, the frequentist testing procedure (including conditions for early stopping).

Furthermore, we assume that the protocol specifies a user-supplied optimal unknown choice of randomization scheme: our adaptive design does not belong to the class of prespecified sampling schemes in that it targets the latter optimal unknown choice of randomization scheme, learning it based on accrued data. We will focus in this article on maximum likelihood estimation and testing. The considered user-supplied optimal unknown choice of randomization scheme will be that which minimizes the asymptotic variance of our maximum likelihood estimator of the parameter of interest. Even though other choices may be interesting (and dealt with along the same lines), we feel strongly that minimizing the asymptotic variance of our estimator is a particularly sensible choice, as it guarantees narrower confidence intervals and earlier decision to reject the null for its alternative or not. Quoting James Hung of the FDA (about design adaptation in general — see [14]), our adaptive design meets clearly stated objectives, it is certainly a "more careful planning, not sloppy planning".

By adaptive *group sequential* design, we refer to the fact that group sequential testing methods can be equally well applied on top of adaptive designs. According to Hu and Rosenberg (quoting from [13]), "The basic statistical formulation of a sequential testing procedure requires determining the joint distribution of the sequentially computed test statistics. Under response-adaptive randomization, this is a difficult task. There has been little theoretical work done to this point, nor has there been any evaluation of sequential monitoring in the context of sequential estimation procedures [*i.e.* targeted adaptive designs] such as the double adaptive biased coin design." These authors end their final chapter with a quote from [22]: "Surprisingly, the link between response-adaptive randomization and sequential analysis has been tenuous at best, and this is perhaps the logical place to search for open research topics." Indeed, we will determine the limit joint distribution of the sequentially computed test statistics based on our results, and provide a theoretical background to rely upon for adaptive design group sequential testing procedures.

1.2 Bibliography.

The literature on adaptive designs is vast and we apologize for not including all of it. Quite misleadingly, the expression "adaptive design" has also been used in the literature for sequential testing and, in general, for designs that allow data-adaptive stopping times for the whole study (or for certain treatment arms) which achieve the wished type I and type II errors requirements when testing a null hypothesis against its alternative.

In the literature dedicated to what corresponds to our definition of adaptive design, such adaptive designs are referred to as "response-adaptive randomization" designs (see the quote from [13] above). Of course, data-adaptive randomization schemes have a long history, that goes back to the 1930s, and we refer to Section 1.2 in [13] and Section 17.4 in [17] to provide the reader with a comprehensive historical perspective.

The organization of the Section 1.1 illustrates the fact that we have decided to tackle separately the group sequential testing problem from the data-adaptive determination of the randomization probability in response to data collected so far — a choice justified by the fact that separating the characterization of a group sequential testing procedure from the adaptation of the randomization probability makes perfect sense from a methodological point of view. Resorting to the same organization here is more delicate, because response-adaptive treatment allocation is indebted to early studies in the context of sequential statistical analysis, such as [1, 6, 9] among many others. However, the authors of [13] manage to trace back the idea of incorporating randomization in the context of adaptive treatment allocation designs to [31].

On the one hand, regarding the adaptation of the randomization probability, data-adaptive randomization schemes belong to either the "urn model" or "double adaptive biased coin" families (see the quote from [13] above), depending on whether the approach is nonparametric or not. Adaptive designs based on urn models (so called because the randomization scheme can be modeled after different ways of pulling various colored balls from an urn) notably include the seminal "randomized play-the-winner rule" from the aforementioned article [31] or the more recent "drop-the-loser rule" [15]; they do not target a specific user-supplied optimal unknown choice of randomization scheme. The theory of this type of adaptive design is presented in detail in chapter 4 of [13], with a comprehensive bibliography. On the contrary, targeting a specific user-supplied optimal unknown choice of randomization scheme is the core of adaptive designs based on flipping a (data-adaptively) biased coin. More precisely, the latter targeted randomization scheme is expressed as a function $f(\theta)$ of an unknown parameter θ of the response model, and the adaptive design is characterized by the sequence $f(\theta_n)$ which is based on updated estimates θ_n of θ as the data accrue. For instance, the targeted randomization scheme we will consider (namely, the minimizer of the asymptotic variance of our maximum likelihood estimator of the parameter of interest in clinical trials with binary treatment, binary outcome and no covariate) is a function of the two marginal probabilities of success. Again, the authors of [13] manage to trace back this kind of procedure to [7]. A series of articles including [23, 12] address the theoretical study of such adaptive designs, or investigate their properties based on simulations [11]. Overall, the most relevant reference for our present article is certainly [13] (already cited many times), which concerns asymptotic theory for likelihood-based estimation (not testing) based on data-adapted randomization schemes in clinical trials.

On the other hand, regarding the group sequential testing problem, let us emphasize that we consider the case where one starts with a large up-front commitment sample size and uses group sequential testing to allow early stopping rather than starting out with a small commitment of sample size and extending it if necessary — the latter distinction is taken from [20]. Therefore, negative results obtained *e.g.* in [18, 25] for such procedures (inconveniently referred to as adaptive designs methods in [20, 25]) that start out with a small commitment do not apply at all to our procedure. On the contrary, we can build upon the thorough understanding of group sequential methods as exposed in [17, 21], and more recently explored in [19].

Furthermore, there is also a rich literature on the Bayesian approach to adaptive designs. The reader is referred to [4, 16, 3, 2] for further details.

Finally, this article builds upon the seminal technical report [26] which paves the way to robust

and efficient estimation in randomized clinical trials thanks to adaptation of the design in a variety of settings.

1.3 Forthcoming results in words.

We will state, prove and verify by simulation various properties of adaptive designs in the framework of clinical trials with binary treatment, binary outcome and no covariate. Following the same presentation as in Section 1.1, we obtain that the adaptive design converges almost surely to the targeted unknown randomization scheme (Theorem 1), and that

- estimation framework:
  - our maximum likelihood estimator of the parameter of interest is a strongly consistent estimator (Theorem 1), it satisfies a central limit theorem (Theorem 2); we can estimate its asymptotic variance, which is the same as that it would feature had we known in advance the targeted randomization scheme and independently sampled from it (Theorem 2); consequently, inference can be carried out as if we had resorted to independent and identically distributed (iid) sampling;
  - furthermore, the comprehensive simulation study that we undertake validates the theory, notably showing that the confidence intervals we obtain achieve the desired coverage even for moderate sample sizes;
- testing framework:
  - the multidimensional t-statistics that we would use under iid sampling still converges to the same canonical distribution under adaptive sampling (Theorem 3); consequently, the same group sequential testing can be carried out as if we had resorted to iid sampling;
  - furthermore, the comprehensive simulation study that we undertake validates the theory, notably showing that type I error control at the prescribed level is guaranteed, and that all sampling procedures only suffer from a very slight increase of the type II error.

A three-sentence take-home message is "Adaptive designs do learn the targeted optimal design and inference and testing can be carried out under adaptive sampling as they would under the targeted optimal randomization probability iid sampling. In particular, adaptive designs achieve the same efficiency as the fixed oracle design. This is confirmed by a simulation study, at least for moderate or large sample sizes, across a large collection of targeted randomization probabilities."

In essence, everything works as predicted by theory. However, theory also warns us that gains cannot be dramatic in the particular setting of clinical trials with binary treatment, binary outcome and no covariate. Nonetheless, this article is important: it provides a theoretical template and tools for asymptotic analysis of robust adaptive designs in less constrained settings, which we will consider in future work. This notably includes the setting of clinical trials *with covariate*, binary treatment, and *discrete or continuous outcome*, or the setting of clinical trials *with covariate*, binary treatment, and *possibly censored time-to-event* among others. Resorting to targeted maximum likelihood estimation [27] along with adaptation of the design provides substantial gains in efficiency.

Finally, we want to emphasize that the whole adaptive design methodology that we develop here is only relevant for clinical trials in which a substantial number of observations are available before all patients are randomized. From now on, we assume that the clinical trial's time scale permits the application of the adaptive design methodology.

1.4 Organization of the article.

The article is organized as follows. We define the targeted optimal design in Section 2, and describe how to adapt to it in Section 3. The asymptotic study of the maximum likelihood estimator of the parameter of interest under adaptive design is addressed in Section 4, focusing on strong

consistency in Section 4.1 and on asymptotic normality in Section 4.2. In Section 5 we show how a group sequential testing procedure can be applied on top of the adaptive design methodology. We present the results of a simulation study in Sections 6 and 7. Section 6 is dedicated to the investigation of moderate and large sample size properties of the adaptive design methodology with respect to estimation and assessment of uncertainty. Section 7 is dedicated to the performances of the adaptive design group sequential testing methodology. In both sections, our data-adaptive methodology is applied to a large collection of problems. How well the method performs is determined across the collection of problems, which requires coming up with tailored tests that we present in Appendix. In Appendix A.1 we present an important building block for consistency results. It consists of a uniform Kolmogorov strong law of large numbers for martingales sums that essentially relies on a maximal inequality for martingales. Another important building block for central limit theorems is presented in Appendix A.2, where we derive a central limit theorem for discrete martingales. The two tailored tests that we repeatedly use in Sections 6 and 7 in order to evaluate the simulations are carefully described in Appendix A.3 and A.4. The latter tests provide single p-values for multiple pairwise comparisons in the context of our simulations, notably dealing with the multiplicity of elementary tests carried out. We conclude in Appendix A.5 with a series of tables summarizing the results of the simulation study presented and interpreted in Sections 6 and 7.

Finally, in order to ease the reading, we highlight throughout the text the most important results. We notably point out how to construct confidence intervals and how to apply a group sequential testing procedure while targeting the optimal design and thus accruing observations data-adaptively. We also stress in which terms the theoretical study and the simulations validate the latter methods. Moreover, we compare the performances of the targeted optimal design sampling scheme with those of the oracle iid sampling scheme (*i.e.* the targeted scheme). Seven highlight are scattered in the article, respectively entitled

1. *pointwise estimation and confidence interval* (Section 4.2),
2. *targeted optimal design adaptive group sequential testing* (Section 5.1),
3. *empirical validation of central limit theorem* (Section 6.2),
4. *empirical coverage of the confidence intervals* (Section 6.3),
5. *empirical widths of confidence intervals* (Section 6.4),
6. *empirical type I and type II errors* (Section 7.2), and
7. *empirical sample sizes at decision* (Section 7.3).

2 Balanced versus optimal treatment mechanisms

2.1 The observed data structure and related likelihood.

We consider the simplest example of randomized trials, where an observation writes as $O = (A, Y)$, A being a binary treatment of interest and Y a binary outcome of interest. We postulate the existence of a full data structure $X = (X(0), X(1))$ containing the two counterfactual (or potential) outcomes under the two possible treatments. The observed data structure $O = (A, X(A)) = (A, Y)$ only contains the outcome corresponding to the treatment the experimental unit did receive. Therefore O is a missing data structure on X with missingness variable A.

We denote the conditional probability distributions of treatment A by

$$g(a|x) = P(A = a|X = x).$$

We assume that the *randomization* (or *coarsening at random*, abbreviated to *CAR*) assumption holds: for all $a \in \mathcal{A} = \{0, 1\}, x \in \mathcal{X} = \{0, 1\}^2$,

$$g(a|x) = P(A = a|X(a) = x(a)), \tag{1}$$

We denote by $\mathcal{G}$ the set of such CAR conditional distributions of A given X, referred to as the set of fixed designs. In the framework of this article, (1) is equivalent to

$$g(a|x) = g(a)$$

for all $a \in \mathcal{A}, x \in \mathcal{X}$: $g \in \mathcal{G}$ if and only if the random variables A and X are independent. We only consider such treatment mechanisms in the rest of Section 2.

The distribution P_X of the full data structure X has two marginal Bernoulli distributions characterized by $\theta = (\theta_0, \theta_1) \in]0,1[^2$ with $\theta_0 = E_{P_X} X(0)$ and $\theta_1 = E_{P_X} X(1)$ (the only identifiable part of P_X). Therefore, introducing $\mathcal{X}(O) = \{x \in \mathcal{X} : x(A) = Y\}$ (the set of full data structure realizations compatible with O), the likelihood of O writes as

$$\begin{aligned} L(O) &= \sum_{x \in \mathcal{X}(O)} P(O, X = x) = \sum_{x \in \mathcal{X}(O)} P(O|X = x) P(X = x) \\ &= \sum_{x \in \mathcal{X}(O)} g(A|x) P(X = x) = g(A) P(X \in \mathcal{X}(O)) \\ &= \theta_A^Y (1 - \theta_A)^{1-Y} g(A) = \theta(O) g(A), \end{aligned}$$

using the convenient shorthand notation $\theta(O) = \theta_A^Y (1 - \theta_A)^{1-Y}$. Because of the form of the likelihood, we can say that the observed data structure O is obtained under (θ, g).

2.2 Efficient influence curve and efficient asymptotic variance for the log-relative risk.

Say that the parameter of scientific interest is

$$\Psi(\theta) = \log \frac{\theta_1}{\theta_0},$$

the log-relative risk; of course, the sequel straightforwardly applies to other choices, such as the excess risk.

In a classical randomized trial, we would determine a (deterministic) treatment mechanism g (therefore complying with CAR) and sample as many iid copies of O as necessary.

The theory of semiparametric statistics teaches us that the efficient influence curve for parameter θ_0 under (θ, g) is

$$D_0^\star(\theta, g)(O) = (Y - \theta_0) \frac{\mathbb{1}\{A = 0\}}{g(0)} \tag{2}$$

and that for parameter θ_1 under (θ, g) is

$$D_1^\star(\theta, g)(O) = (Y - \theta_1) \frac{\mathbb{1}\{A = 1\}}{g(1)}. \tag{3}$$

Then the delta-method (and page 386 in [28]) implies that the efficient influence curve for parameter $\Psi(\theta)$ under (θ, g) writes as

$$\mathrm{IC}(\theta, g)(O) = -\frac{\mathbb{1}\{A = 0\}}{\theta_0 g(0)} (Y - \theta_0) + \frac{\mathbb{1}\{A = 1\}}{\theta_1 g(1)} (Y - \theta_1) \tag{4}$$

so that the efficient asymptotic variance under (θ, g) is

$$\begin{aligned} E_{\theta,g} \mathrm{IC}(\theta, g)^2(O) &= E_{\theta,g} \left\{ (Y - E_\theta(Y|A))^2 \left(\frac{\mathbb{1}\{A = 0\}}{\theta_0^2 g(0)^2} + \frac{\mathbb{1}\{A = 1\}}{\theta_1^2 g(1)^2} \right) \right\} \\ &= \frac{\sigma^2(\theta)(0)}{\theta_0^2 g(0)} + \frac{\sigma^2(\theta)(1)}{\theta_1^2 g(1)} \\ &= \frac{1 - \theta_0}{\theta_0 g(0)} + \frac{1 - \theta_1}{\theta_1 g(1)} \end{aligned}$$

with notation $\sigma^2(\theta)(a) = \mathrm{Var}_\theta(Y|A = a) = \theta_a(1 - \theta_a)$, the conditional variance of Y given $A = a$.

2.3 A relative efficiency criterion.

Defining $\mathrm{OR}(\theta) = \frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}$, the efficient asymptotic variance as a function of the treatment mechanism g is minimized at the *optimal treatment mechanism* characterized by

$$g^\star(\theta)(1) = \frac{1}{1+\sqrt{\mathrm{OR}(\theta)}} = \frac{\sqrt{\theta_0(1-\theta_1)}}{\sqrt{\theta_0(1-\theta_1)}+\sqrt{\theta_1(1-\theta_0)}}, \tag{5}$$

known as the Neyman allocation (see [13] page 13). Interestingly, $g^\star(\theta)(1) \leq \frac{1}{2}$ whenever $\theta_0 \leq \theta_1$, meaning that the Neyman allocation $g^\star(\theta)$ favors the inferior treatment. The corresponding optimal efficient asymptotic variance then writes as

$$v^\star(\theta) = \left(\sqrt{\frac{1-\theta_0}{\theta_0}} + \sqrt{\frac{1-\theta_1}{\theta_1}}\right)^2.$$

In contrast, the standard balanced treatment characterized by $g^b(1) = \frac{1}{2}$ features an efficient asymptotic variance

$$v^b(\theta) = 2\left(\frac{1-\theta_0}{\theta_0} + \frac{1-\theta_1}{\theta_1}\right),$$

hence the relative efficiency criterion

$$R(\theta) = \frac{v^\star(\theta)}{v^b(\theta)} = \frac{1}{2} + \frac{\sqrt{\mathrm{OR}(\theta)}}{1+\mathrm{OR}(\theta)} \in (\tfrac{1}{2}, 1]. \tag{6}$$

The definition of our relative efficiency criterion illustrates the fact that we decide to consider the balanced treatment mechanism as a benchmark. We emphasize that *any* fixed design could be chosen as benchmark treatment mechanism, with minor impact on the study we expose below.

It is worth noting that $v^b(\theta) = v^\star(\theta)$, or in other words that the so-called balanced treatment mechanism is actually optimal, if and only if $\theta_0 = \theta_1$. In particular, there is no gain to expect from adapting the treatment mechanism in terms of type I error control when testing the null "$\Psi(\theta) = 0$" against its negation. In addition, the following bound involving the relative efficiency criterion on one side and the log-relative risk on the other holds:

$$R(\theta) \leq \frac{1}{2} + \frac{\sqrt{e^{\Psi(\theta)}}}{1+e^{\Psi(\theta)}} \in (\tfrac{1}{2}, 1].$$

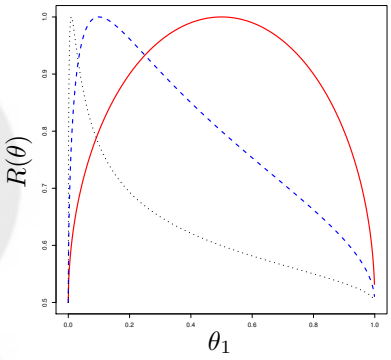

Figure 1: Plot of the relative efficiency $R(\theta)$ as a function of θ_1 for different values of θ_0 ($\theta = (\theta_0, \theta_1)$). The solid, dashed and dotted curves respectively correspond to $\theta_0 = \frac{1}{2}, \frac{1}{10}, \frac{1}{100}$.

We present in Figure 1 three curves $\theta_1 \mapsto R(\theta)$ for three different values of θ_0. It notably illustrates that when θ_0 is small, $R(\theta)$ can be significantly lower than 1 for values of $\Psi(\theta)$ which

are not very large. For instance, $\theta = (\frac{1}{100}, \frac{5}{100})$ yields $\Psi(\theta) \simeq 1.609$, $R(\theta) \simeq 0.868$ and optimal treatment mechanism characterized by $g^\star(\theta)(1) \simeq 0.305$. Were we given the optimal treatment mechanism in advance, we would obtain confidence intervals (based on the central limit theorem and Slutsky's lemma) whose widths are approximately $\sqrt{R(\theta)} \simeq 0.931$ times those of the corresponding confidence intervals we would have got using the balanced treatment mechanism.

However the gain could actually be more dramatic than the previous example let think.

Let us consider again the testing setting: we want to test the null "$\Psi(\theta) = 0$" against the alternative "$\Psi(\theta) > 0$" with type I error α and power $(1 - \beta)$ at some user-defined alternative $\psi > 0$.

Thanks to the delta-method, we know that the maximum likelihood estimator of $\Psi(\theta)$

$$\Psi_n = \log \frac{\sum_{i=1}^n Y_i \mathbb{1}\{A_i = 1\}}{\sum_{i=1}^n \mathbb{1}\{A_i = 1\}} - \log \frac{\sum_{i=1}^n Y_i \mathbb{1}\{A_i = 0\}}{\sum_{i=1}^n \mathbb{1}\{A_i = 0\}}$$

based on n iid copies $O_i = (A_i, Y_i)$ of O is asymptotically efficient. It is furthermore natural to refer to

$$I_n = \frac{n}{s_n^2},$$

the inverse of the estimated variance of Ψ_n at time n, as the *statistical information* available at that time. Under ψ, the central limit theorem applies and teaches us that $\sqrt{I_n}(\Psi_n - \psi)$ converges in distribution, as n grows to infinity, to the standard normal distribution.

Deciding to reject the null if $\sqrt{I_n}\Psi_n \geq \xi_{1-\alpha}$ yields a test with asymptotic type I error α. In order to ensure that its asymptotic power at alternative ψ is $(1 - \beta)$, it suffices to choose n such as

$$n = \inf \left\{ t \geq 1 : I_t \geq \left(\frac{\xi_{1-\alpha} - \xi_\beta}{\psi} \right)^2 = I_{\max} \right\}, \tag{7}$$

$I_{\max}$ being the so-called *maximum committed information*.

For n large enough, $I_n \simeq n/v^b(\theta)$ if we use the balanced treatment mechanism, while I_n would have been approximately equal to $n/v^\star(\theta)$, had we used the optimal treatment mechanism. Substituting bluntly $n/v^b(\theta)$ or $n/v^\star(\theta)$ to I_n in (7), we see that the ratio of the testing times n^b (using the balanced treatment mechanism) and $n^\star$ (using the optimal one) satisfies

$$\frac{n^\star}{n^b} \simeq \frac{v^\star(\theta)}{v^b(\theta)} = R(\theta),$$

the relative efficiency criterion. In other words, were we given the optimal treatment mechanism in advance, we would in average need to sample $R(\theta) \in (\frac{1}{2}, 1)$ times the number of observations required when using the balanced treatment mechanism. In the previous example where $\theta = (\frac{1}{100}, \frac{5}{100})$, setting $\alpha = 0.05$, $\beta = 0.1$ and the alternative parameter $\psi = \Psi(\theta) \simeq 1.609 > 0$, the maximal committed information is $I_{\max} \simeq 3.306$, $n^\star \simeq 676.901$ and $n^b \simeq 780.248$.

In summary, resorting to the balanced treatment mechanism may be a very poor (inefficient) choice. Since the optimal treatment mechanism $g^\star(\theta)$ can be learned from the data, why not use it? Of course, targeting the optimal treatment mechanism on the fly implies losing independence between successive observations, making the study of the design much more involved. However, we present and study in the sequel such a methodology. It is built on the seminal technical report [26].

3 Targeting the optimal design

3.1 Adaptive coarsening at random assumption.

We denote by A_i, $X_i = (X_i(0), X_i(1))$, $Y_i = X_i(A_i)$ and $O_i = (A_i, Y_i)$ the treatment assignment, full data structure, outcome, and observation for experimental unit i. Whereas $X_1, \ldots, X_n$ are

assumed iid, the random variables $A_1, \ldots, A_n$ are not independent anymore since we want to adapt the treatment mechanism based on past observations. Defining

$$\begin{aligned} \mathbf{A}_n &= (A_1, \ldots, A_n), \\ \mathbf{X}_n &= (X_1, \ldots, X_n), \\ \mathbf{O}_n &= (O_1, \ldots, O_n), \end{aligned}$$

and for every $i = 0, \ldots, n$

$$\begin{aligned} \mathbf{A}_n(i) &= (A_1, \ldots, A_i), \\ \mathbf{X}_n(i) &= (X_1, \ldots, X_i), \\ \mathbf{O}_n(i) &= (O_1, \ldots, O_i) \end{aligned}$$

(with convention $\mathbf{A}_n(0) = \mathbf{X}_n(0) = \mathbf{O}_n(0) = \emptyset$), let $\mathbf{g}_n(\cdot|\mathbf{X}_n)$ denote the conditional distribution of the design settings $\mathbf{A}_n$ given the full data $\mathbf{X}_n$: by the chain rule,

$$\mathbf{g}_n(\mathbf{A}_n|\mathbf{X}_n) = \prod_{i=1}^{n} P(A_i|\mathbf{A}_n(i-1), \mathbf{X}_n), \tag{8}$$

hence the additional notation

$$g_i(a_i|\mathbf{a}(i-1), \mathbf{x}) = P(A_i = a_i|\mathbf{A}_n(i-1) = \mathbf{a}(i-1), \mathbf{X}_n = \mathbf{x})$$

for all $1 \leq i \leq n$, $\mathbf{a} = (a_1, \ldots, a_n) \in \mathcal{A}^n$, $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$.

In this new setting, we state the following adaptive counterpart of the CAR assumption (1): for all $1 \leq i \leq n$, $\mathbf{a} \in \mathcal{A}^n$, $\mathbf{x} \in \mathcal{X}^n$, letting $o_i = (a_i, x_i(a_i))$ be the corresponding realization of observation O_i,

$$\begin{aligned} g_i(a_i|\mathbf{a}(i-1), \mathbf{x}) &= P(A_i = a_i|X_i = x_i, \mathbf{O}_n(i-1) = \mathbf{o}(i-1)) \\ &= P(A_i = a_i|X_i(a_i) = x_i(a_i), \mathbf{O}_n(i-1) = \mathbf{o}(i-1)). \end{aligned}$$

With obvious convention, the new *adaptive randomization* (or *adaptive CAR*) assumption also writes as

$$g_i(a_i|\mathbf{a}(i-1), \mathbf{x}) = g_i(a_i|x_i, \mathbf{o}(i-1)) = g_i(a_i|x_i(a_i), \mathbf{o}(i-1)) \tag{9}$$

for all $1 \leq i \leq n$, $\mathbf{a} \in \mathcal{A}^n$, $\mathbf{x} \in \mathcal{X}^n$; it states that for each i A_i is conditionally independent of the full data $\mathbf{X}_n$ given the observed data $\mathbf{O}_n(i-1)$ for the first $(i-1)$ experimental units and the full data X_i for the ith experimental unit, and in addition that the conditional probability of $A_i = a_i$ given X_i and $\mathbf{O}_n(i-1)$ actually only depends on the observed part $X_i(a_i)$ and $\mathbf{O}_n(i-1)$. In particular, (8) reduces to

$$\mathbf{g}_n(\mathbf{A}_n|\mathbf{X}_n) = \prod_{i=1}^{n} g_i(A_i|X_i(A_i), \mathbf{O}_n(i-1)), \tag{10}$$

which justifies the notation $\mathbf{g}_n = (g_1, \ldots, g_n)$. Note that in the framework of this article, a treatment mechanism $\mathbf{g}_n$ complies with the adaptive CAR assumption (10) if and only if, for all $1 \leq i \leq n, \mathbf{a} \in \mathcal{A}^n, \mathbf{x} \in \mathcal{X}^n$,

$$g_i(a_i|\mathbf{a}(i-1), \mathbf{x}) = g_i(a_i|\mathbf{o}(i-1)).$$

Note finally that we find useful to consider g_i satisfying (9) as random element (through $\mathbf{O}_n(i-1)$) of the set $\mathcal{G}$ of CAR designs.

3.2 Data generating mechanism for adaptive design and likelihood.

Given the available data $\mathbf{O}_n(i-1) = (O_1, \ldots, O_{i-1})$ at step i, one draws X_i from P_X independently of $\mathbf{X}_n(i-1)$, then one calculates the conditional distribution $g_i(\cdot|X_i, \mathbf{O}_n(i-1))$ and samples A_i given X_i from it, the next observation finally being $O_i = (A_i, X_i(A_i))$. Regarding the likelihood of $\mathbf{O}_n$, if

$$\begin{aligned}\mathcal{X}(\mathbf{O}_n) &= \{\mathbf{x} \in \mathcal{X}^n : x_i(A_i) = Y_i, i \leq n\} \\ &= \otimes_{i=1}^n \{x \in \mathcal{X} : x(A_i) = Y_i\} \\ &= \otimes_{i=1}^n \mathcal{X}(O_i)\end{aligned}$$

(the set of those realizations $\mathbf{x}$ of $\mathbf{X}_n$ compatible with $\mathbf{O}_n$), then

$$\begin{aligned}L(\mathbf{O}_n) &= \sum_{\mathbf{x} \in \mathcal{X}(\mathbf{O}_n)} P(\mathbf{O}_n, \mathbf{X}_n = \mathbf{x}) \\ &= \sum_{\mathbf{x} \in \mathcal{X}(\mathbf{O}_n)} \mathbf{g}_n(\mathbf{A}_n|\mathbf{x}) P(\mathbf{X}_n = \mathbf{x}) \\ &= \prod_{i=1}^n g_i(A_i|Y_i, \mathbf{O}_n(i-1)) P(\mathbf{X}_n \in \mathcal{X}(\mathbf{O}_n)) \\ &= \prod_{i=1}^n g_i(A_i|Y_i, \mathbf{O}_n(i-1)) \prod_{i=1}^n \theta(O_i), \end{aligned} \tag{11}$$

the third and fourth equalities being derived from the adaptive CAR equality (10) and from independence of $X_1, \ldots, X_n$ respectively. Thus, the likelihood remarkably factorizes into the product of a θ-factor and a $\mathbf{g}_n$-factor. Thanks to the form of the likelihood, we can say that $\mathbf{O}_n$ is obtained under $(\theta, \mathbf{g}_n)$. For convenience, we will also write sometimes that $\mathbf{O}_n$ is *obtained under* $\mathbf{g}_n$*-adaptive sampling scheme* without specifying the parameter θ. Likewise, we will later refer to data *obtained under iid* g^b*-balanced* or *under iid* $g^\star$*-optimal sampling schemes*.

3.3 Strategy.

Let $\theta_i = (\theta_{i,0}, \theta_{i,1})$ denote for each $i \leq n$ the maximum likelihood estimator of $\theta = (\theta_0, \theta_1) \in]0,1[^2$ based on $\mathbf{O}_n(i)$ (with convention $\theta_{i,a} = \frac{1}{2}$ as long as no relevant observation is available). Thanks to the form of the log-likelihood exhibited in (11) and as soon as $\sum_{j=1}^i \mathbb{1}\{A_j = a\} > 1$, $\theta_{i,a}$ is the empirical mean

$$\theta_{i,a} = \frac{\sum_{j=1}^i Y_j \mathbb{1}\{A_j = a\}}{\sum_{j=1}^i \mathbb{1}\{A_j = a\}},$$

as if we used a deterministic treatment mechanism (and observations were iid).

These empirical means yield plug-in estimates of $\sigma^2(\theta)(a)$:

$$\sigma_i^2(a) = \theta_{i,a}(1 - \theta_{i,a}),$$

as well as plug-in estimates of the optimal treatment mechanism $g^\star(\theta)$ introduced in (5), characterized by $g_1^s(1|\mathbf{O}_n(0)) = \frac{1}{2}$ and for $i \geq 1$,

$$g_{i+1}^s(1|\mathbf{O}_n(i)) = \frac{\sqrt{\theta_{i,0}(1-\theta_{i,1})}}{\sqrt{\theta_{i,0}(1-\theta_{i,1})} + \sqrt{\theta_{i,1}(1-\theta_{i,0})}} \tag{12}$$

(sometimes abbreviated to $g_{i+1}^s(1)$), hence a first adaptive CAR treatment mechanism $\mathbf{g}_n^s = (g_1^s, \ldots, g_n^s)$. Another interesting choice is also considered here, which we characterize iteratively by $g_1(1|\mathbf{O}_n(0)) = \frac{1}{2}$ and for $i \geq 1$,

$$g_{i+1}(1|\mathbf{O}_n(i)) = \underset{\gamma \in (0,1)}{\arg\min} \left\{ \frac{1}{i+1} \left(\sum_{j=1}^i g_j(1|\mathbf{O}_n(j-1)) + \gamma \right) - g_i^s(1) \right\}^2. \tag{13}$$

This alternative choice aims at obtaining a balance between the two treatments which, at experiment i, closely approximates $g^\star(\theta)$, in the sense that $\frac{1}{i}\sum_{j=1}^{i} \mathbb{1}\{A_j = 1\} \simeq g^s_{i-1}(1)$, the current best guess. This second definition is more *aggressive* in the pursuit of the optimal treatment mechanism, as it tries to compensate on the fly for early sub-optimal sampling.

A technical condition was actually left aside in the definition of $\mathbf{g}^s_n$. Because we want to exclude the possibility that the adaptive design stops a treatment arm with probability tending to 1, we impose that $g_{i+1}(1|\mathbf{O}_n(i)) \in [\delta; 1-\delta]$ for a small $\delta > 0$ (such that $\delta < \min_{a\in\mathcal{A}} g^\star(\theta)(a)$ and $1-\delta > \max_{a\in\mathcal{A}} g^\star(\theta)(a)$) by letting $g^\star_1(1|\mathbf{O}_n(0)) = \frac{1}{2}$ and for $i \geq 1$,

$$g^\star_{i+1}(1|\mathbf{O}_n(i)) = \min\left\{1-\delta, \max\left\{\delta, \frac{\sqrt{\theta_{i,0}(1-\theta_{i,1})}}{\sqrt{\theta_{i,0}(1-\theta_{i,1})}+\sqrt{\theta_{i,1}(1-\theta_{i,0})}}\right\}\right\} \tag{14}$$

(sometimes abbreviated to $g^\star_{i+1}(1)$), thus characterizing the adaptive CAR treatment mechanism $\mathbf{g}^\star_n = (g^\star_1, \ldots, g^\star_n)$. Similarly, we substitute $g^\star_i(1)$ to $g^s_i(1)$ and allow γ to vary in $[\delta, 1-\delta]$ only in (13), yielding another adaptive CAR treatment mechanism $\mathbf{g}^a_n$ (where the superscript a stands for *aggressive*).

In the rest of this article, we investigate theoretically and by intensive simulations the properties of the data-adaptive designs based on $\mathbf{g}^\star_n$ ($\mathbf{g}^a_n$ is only considered through the simulation study).

4 Asymptotic study

We address in this section the asymptotic statistical study of the method presented in the previous section. We derive strong consistency results in Section 4.1 and a central limit theorem in Section 4.2.

4.1 Strong consistency.

The following consistency result holds, which teaches us that the method does learn what is the optimal design.

Theorem 1. *Let θ_n be the maximum likelihood estimator of $\theta \in]0,1[^2$ based on $\mathbf{O}_n$ sampled from $(\theta, \mathbf{g}^\star_n)$, for the adaptive CAR treatment mechanism $\mathbf{g}^\star_n$ characterized by (14). Then θ_n converges almost surely to θ. Consequently, Ψ_n is a strongly consistent estimate of $\Psi(\theta)$ and $\mathbf{g}^\star_n$ converges to the optimal design $g^\star(\theta)$ in the sense that $g^\star_n(1)$ and $\frac{1}{n}\sum_{i=1}^{n} g^\star_i(1)$ both converge almost surely to $g^\star(\theta)(1)$.*

Proof. Let us first introduce the following estimating function

$$\begin{aligned} D(\vartheta)(O) &= (D_0(\vartheta)(O), D_1(\vartheta)(O)) \\ &= ((Y-\vartheta_0)\mathbb{1}\{A=0\}, (Y-\vartheta_1)\mathbb{1}\{A=1\}). \end{aligned} \tag{15}$$

We also denote by $P_{\theta,g^\star_i}$ the conditional distribution of O_i given $\mathbf{O}_n(i-1)$, so that $P_{\theta,g^\star_i}D(\vartheta) = E[D(\vartheta)(O_i)|\mathbf{O}_n(i-1)]$.

For each $i \leq n$

$$0 = P_{\theta,g^\star_i}D(\theta) \tag{16}$$

(according to the terminology introduced in [26], $\vartheta \mapsto D(\vartheta)$ is a martingale estimating function for θ) while

$$0 = \frac{1}{n}\sum_{i=1}^{n} D(\theta_n)(O_i), \tag{17}$$

because θ_n is the maximum likelihood estimator of θ. Equation (17) can be rearranged into

$$0 = \frac{1}{n}\sum_{i=1}^{n}\left[D(\theta_n)(O_i) - P_{\theta,g^\star_i}D(\theta_n)\right] + \frac{1}{n}\sum_{i=1}^{n} P_{\theta,g^\star_i}D(\theta_n)$$

$$= \quad M_n(D(\theta_n)) + \frac{1}{n}\sum_{i=1}^{n} P_{\theta,g_i^\star} D(\theta_n),$$

which makes clear that if $M_n(D(\theta_n))$ converges to 0 almost surely, then $\frac{1}{n}\sum_{i=1}^{n} P_{\theta,g_i^\star} D(\theta_n)$ does too. Furthermore,

$$P_{\theta,g_i^\star} D(\theta_n) = ((\theta_0 - \theta_{n,0}) g_i^\star(0|\mathbf{O}_n(i-1)), (\theta_1 - \theta_{n,1}) g_i^\star(1|\mathbf{O}_n(i-1)))$$

whence $\frac{1}{n}\sum_{i=1}^{n} P_{\theta,g_i^\star} D(\theta_n)$ converges to 0 almost surely if and only if θ_n does to θ, since (14) guarantees that $\frac{1}{n}\sum_{i=1}^{n} g_i^\star(a|\mathbf{O}_n(i-1)) \in [\delta, 1-\delta]$ for both $a \in \mathcal{A}$.

Now, $|M_n(D(\theta_n))|$ is itself upper-bounded by

$$\sup_{\vartheta \in [0,1]^2} |M_n(D(\vartheta))|$$

which converges to 0 almost surely by application of Theorem 8, justified by the fact that $\sup_{\vartheta\in[0,1]} \|D(\vartheta)\|_\infty < \infty$ and because the standard entropy of $\mathcal{F} = \{D(\vartheta) : \vartheta \in [0,1]\}$ for the supremum norm satisfies $H(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = \mathcal{O}(\log 1/\varepsilon)$ (see [28], example 19.7).

Finally, the strong consistency of θ_n straightforwardly yields that $g_n^\star(1)$ converges almost surely to $g^\star(\theta)(1)$, since the mapping $\theta_n \mapsto g_{n+1}^\star(1)$ is continuous. This in turn implies that the Cesàro mean $\frac{1}{n}\sum_{i=1}^{n} g_i^\star(1)$ also converges almost surely to the same limit. □

4.2 Asymptotic normality.

A central limit theorem.

We have established strong consistency of θ_n in the previous subsection. In order to provide statistical inference, we need now to establish that $\sqrt{n}(\theta_n - \theta)$ converges in distribution so that one can construct confidence intervals for θ and come up with valid testing procedures. The following central limit theorem actually holds.

Theorem 2. *Let θ_n be the maximum likelihood estimator of $\theta \in]0,1[^2$ based on $\mathbf{O}_n$ sampled from $(\theta, \mathbf{g}_n^\star)$, for the adaptive CAR treatment mechanism $\mathbf{g}_n^\star$ defined in (14). Let $D_0^\star(\theta, g^\star(\theta))$, $D_1^\star(\theta, g^\star(\theta))$ and $\mathrm{IC}(\theta, g^\star(\theta))$ be as defined in (2), (3) and (4) with $g = g^\star(\theta)$. Let $D^\star(\theta, g^\star(\theta)) = (D_0^\star(\theta, g^\star(\theta)), D_1^\star(\theta, g^\star(\theta)))$. Then under $(\theta, \mathbf{g}_n^\star)$*

$$\sqrt{n}(\theta_n - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D^\star(\theta, g^\star(\theta))(O_i)(1 + o_P(1)). \tag{18}$$

Furthermore, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D^\star(\theta, g^\star(\theta))(O_i)$ is a normalized discrete martingale which converges under $(\theta, \mathbf{g}_n^\star)$ to a centered Gaussian distribution with covariance matrix

$$\begin{aligned} \Sigma^\star &= P_{\theta,g^\star(\theta)} D^\star(\theta, g^\star(\theta))^\top D^\star(\theta, g^\star(\theta)) \\ &= \mathrm{diag}\left(\frac{\theta_0(1-\theta_0)}{g^\star(\theta)(0)}, \frac{\theta_1(1-\theta_1)}{g^\star(\theta)(1)}\right). \end{aligned} \tag{19}$$

The latter is consistently estimated with its empirical counterpart

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} D^\star(\theta_n, g_n^\star)^\top D^\star(\theta_n, g_n^\star)(O_i) \\ &\qquad = \frac{1}{n}\sum_{i=1}^{n} \mathrm{diag}\left((Y_i - \theta_{n,0})^2 \frac{\mathbb{1}\{A_i = 0\}}{g_n^\star(0)^2}, (Y_i - \theta_{n,1})^2 \frac{\mathbb{1}\{A_i = 1\}}{g_n^\star(1)^2}\right) \end{aligned} \tag{20}$$

as if the sampling was iid.

Thus under $(\theta, \mathbf{g}_n^\star)$, we also have

$$\sqrt{n}(\Psi_n - \Psi(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IC}(\theta, g^\star(\theta))(O_i) + o_P(1), \tag{21}$$

and convergence in distribution of the normalized discrete martingale $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IC}(\theta, g^\star(\theta))(O_i)$ to a centered Gaussian distribution with variance $v^\star(\theta)$, the optimal *efficient asymptotic variance. The latter is finally consistently estimated with either $v^\star(\theta_n)$ or*

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{IC}(\theta_n, g_n^\star)(O_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \left((Y_i - \theta_{n,0})^2 \frac{\mathbb{1}\{A_i = 0\}}{\theta_{n,0}^2 g_n^\star(0)^2} + (Y_i - \theta_{n,1})^2 \frac{\mathbb{1}\{A_i = 1\}}{\theta_{n,1}^2 g_n^\star(1)^2} \right)$$

as if sampling was iid.

Construction of confidence intervals.

We wish to construct a confidence interval for $\Psi(\theta)$ based on $\mathbf{O}_n$ sampled under $(\theta, \mathbf{g}_n^\star)$. Let us denote by s_n^2 either consistent estimates of $v^\star(\theta)$ based on $\mathbf{O}_n$ as introduced in Theorem 2, and let $\xi_{1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of the standard normal distribution. Thanks to the latter theorem,

Highlight 1 (pointwise estimation and confidence interval)**.** *In view of Theorems 1 and 2, the estimator Ψ_n of $\Psi(\theta)$ obtained under $(\theta, \mathbf{g}_n^\star)$ sampling scheme is strongly consistent, the estimated probability of being treated $g_n^\star(1)$ also converging almost surely to the optimal probability of being treated $g^\star(\theta)(1)$. In addition, the confidence interval*

$$\left[\Psi_n \pm \frac{s_n}{\sqrt{n}} \xi_{1-\alpha/2} \right]$$

obtained under $(\theta, \mathbf{g}_n^\star)$ sampling scheme has asymptotic coverage $(1-\alpha)$.

This theoretical result is validated with simulations in Section 6.3.

Proof of Theorem 2.

Proof. As already seen in the proof of Theorem 1, $0 = \frac{1}{n} \sum_{i=1}^{n} D(\theta_n)(O_i) = P_{\theta, g_i^\star} D(\theta)$ for every $i \leq n$, therefore yielding the following equality:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(D(\theta_n)(O_i) - D(\theta)(O_i) \right) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(D(\theta)(O_i) - P_{\theta, g_i^\star} D(\theta) \right). \tag{22}$$

Defining $\bar{g}_n(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{A_i = a\}$ for all $a \in \mathcal{A}$ straightforwardly implies that the left-hand side quantity in (22) also writes as $\sqrt{n}(\theta - \theta_n)\Delta_n$, with $\Delta_n = \mathrm{diag}(\bar{g}_n(0), \bar{g}_n(1))$. Since $g_n^\star(0)$ and $g_n^\star(1)$ are positive, it is almost sure that for n large enough, $\bar{g}_n(0)$ and $\bar{g}_n(1)$ are positive. We consider such a sample size in the sequel. So the diagonal matrix Δ_n is invertible, and equation (22) is equivalent to

$$\sqrt{n}(\theta_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(D(\theta)(O_i) - P_{\theta, g_i^\star} D(\theta) \right) \Delta_n^{-1}, \tag{23}$$

where obviously $\Delta_n^{-1} = \mathrm{diag}(1/\bar{g}_n(0), 1/\bar{g}_n(1))$. It remains to prove that Δ_n^{-1} converges in probability to a deterministic matrix in order to derive (18)

To this end, note that for both $a \in \mathcal{A}$, $\bar{g}_n(a) = \frac{1}{n} S_n(a) + \frac{1}{n} \sum_{i=1}^{n} g_i^\star(a)$ where $S_n(a) = \sum_{i=1}^{n} (\mathbb{1}\{A_i = a\} - g_i^\star(a))$ is a discrete martingale sum. Since its increments are uniformly bounded, $S_n(a)$ converges almost surely and in mean square, hence $\frac{1}{n} S_n(a)$ converges in probability to 0. Because Theorem 1 ensures that $\frac{1}{n} \sum_{i=1}^{n} g_i^\star(a)$ converges to $g^\star(\theta)(a)$ in probability, we conclude

that Δ_n^{-1} converges in probability to the matrix $\Delta_\infty^{-1} = \mathrm{diag}(1/g^\star(\theta)(0), 1/g^\star(\theta)(1))$. At this stage, we obtain that (18) holds true.

Define $W_n = \sum_{i=1}^n P_{\theta,g_i^\star} D^\star(\theta, g^\star(\theta))^\top D^\star(\theta, g^\star(\theta))$ and $\Sigma_n = \frac{1}{n} E W_n$. One has

$$\frac{1}{n} W_n = \mathrm{diag}\left(\theta_0(1-\theta_0) \frac{\frac{1}{n}\sum_{i=1}^n g_i^\star(0)}{g^\star(\theta)(0)^2}, \theta_1(1-\theta_1) \frac{\frac{1}{n}\sum_{i=1}^n g_i^\star(1)}{g^\star(\theta)(1)^2} \right).$$

Theorem 1 guarantees that $\frac{1}{n}\sum_{i=1}^n g_i^\star(a)$ converges almost surely to $g^\star(\theta)(a)$ for both $a \in \mathcal{A}$ (and consequently in L^1 norm since $\frac{1}{n}\sum_{i=1}^n g_i^\star(a) \in [0,1]$ for all $n \geq 1$, by virtue of the dominated convergence theorem), hence $\frac{1}{n} W_n$ converges in probability to $\Sigma^\star$ as given by (19), and Σ_n converges to $\Sigma^\star$ too. Therefore Theorem 10 applies and yields that $\frac{1}{\sqrt{n}}\sum_{i=1}^n D^\star(\theta, g^\star(\theta))(O_i)$ is a normalized discrete martingale which converges under $(\theta, \mathbf{g}_n^\star)$ to a centered Gaussian distribution with covariance matrix $\Sigma^\star$, as stated. Consequently, (18) can be rewritten as

$$\sqrt{n}(\theta_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n D^\star(\theta, g^\star(\theta))(O_i) + o_P(1).$$

In addition, Theorem 10 teaches us that $\frac{1}{n}\sum_{i=1}^n D^\star(\theta_n, g_n^\star)^\top D^\star(\theta_n, g_n^\star)(O_i)$ consistently estimates $\Sigma^\star$, and it is readily seen that (20) holds. We complete the proof by a straightforward application of the delta-method. □

5 Targeted optimal design group sequential testing

Obviously, the sequence of estimators $(\Psi_n)_{n\geq 1}$ can be used to carry out the test of the null "$\Psi(\theta) = \psi_0$" against its unilateral alternative "$\Psi(\theta) > \psi_0$" for some $\psi_0 \in \mathbb{R}$. We build in this section a *group sequential testing procedure*, that is a testing procedure which repeatedly tries to make a decision at intervals rather than once all data are collected, or than after every new observation is obtained (such a testing procedure would be said fully sequential). We refer to [17, 21] for a general presentation of group sequential testing procedures.

5.1 The targeted optimal design group sequential testing procedure.

Formal description of the targeted optimal design group sequential testing procedure.

We wish to test the null "$\Psi(\theta) = \psi_0$" against "$\Psi(\theta) > \psi_0$" with asymptotic type I error α and asymptotic type II error β at some $\psi_1 > \psi_0$. We intend to proceed group sequentially with $K \geq 2$ steps, and we wish to rely on a multidimensional t-statistic of the form

$$(\tilde{T}_1, \ldots, \tilde{T}_K) = \left(\frac{\sqrt{N_k}(\Psi_{N_k} - \psi_0)}{s_{N_k}} \right)_{k \leq K}, \tag{24}$$

where each N_k is a carefully chosen (random) sample size and where s_n^2 estimates the asymptotic variance of $\sqrt{n}(\Psi_n - \Psi(\theta))$ under $(\theta, \mathbf{g}_n^\star)$ sampling (see Theorem 2).

To this end, let $0 < p_1 < \ldots < p_K = 1$ be increasingly ordered proportions. Consider the α-spending and β-spending strategies $(\alpha_1, \ldots, \alpha_K)$ and $(\beta_1, \ldots, \beta_K)$, *i.e.* K-tuples of positive numbers such that $\sum_{k=1}^K \alpha_k = \alpha$ and $\sum_{k=1}^K \beta_k = \beta$. One could for instance choose α-spending and β-spending functions f_α, f_β, that are increasing functions from $[0,1]$ to $[0,1]$ such that $f_\alpha(0) = f_\beta(0) = 0$ and $f_\alpha(1) = f_\beta(1) = 1$, and set $\sum_{l=1}^k \alpha_l = f_\alpha(p_k)\alpha$, $\sum_{l=1}^k \beta_l = f_\beta(p_k)\beta$ for all $k \leq K$.

Now, let $(Z_1, \ldots, Z_K)$ follow the centered Gaussian distribution with covariance matrix $\mathcal{C} = (\sqrt{p_{k\wedge l}/p_{k\vee l}})_{k,l\leq K}$ and let us assume that there exists a unique value $I > 0$, the so-called *maximum committed information* from now on denoted by $I_{\max}$, such that there exist a rejection boundary $(a_1, \ldots, a_K)$ and a futility boundary $(b_1, \ldots, b_K)$ satisfying $a_K = b_K$, $P(Z_1 \geq a_1) = \alpha_1$, $P(Z_1 + \psi_1\sqrt{p_1 I} \leq b_1) = \beta_1$, and for every $1 \leq k < K$,

$$P(\forall j \leq k, b_j < Z_j < a_j \text{ and } Z_{k+1} \geq a_{k+1}) \quad = \quad \alpha_{k+1},$$

$$P(\forall j \leq k, b_j < Z_j + \psi_1 \sqrt{p_j I} < a_j \text{ and } Z_{k+1} + \psi_1 \sqrt{p_{k+1} I} \leq b_{k+1}) \quad = \quad \beta_{k+1}.$$

Note that the closer ψ_1 is to ψ_0, the larger is $I_{\max}$ (actually, $\psi_1 \mapsto \psi_1 \sqrt{I_{\max}}$ is both upper bounded and bounded away from zero). Heuristically, the closer ψ_1 is to ψ_0, the more difficult it is to decide between the null and its alternative while preserving the required type II error at ψ_1, the more information is needed to proceed. In this setting, it is natural to refer to the inverse of the variance of Ψ_n as an amount of statistical information collected so far. The latter information writes as $\frac{n}{s_n^2}$, notably making clear that the product $s_n^2 I_{\max}$ and number of observations n are on the same scale.

In this spirit, let us finally define for each $k \leq K$

$$N_k = \inf \left\{ n \geq 1 : \frac{n}{s_n^2} \geq p_k I_{\max} \right\}.$$

Under $(\theta, \mathbf{g}_n^\star)$, if $v^\star(\theta) I_{\max}$ is large, then N_k tend to be large too.

The targeted optimal design group sequential testing rule finally writes as follows:
starting from $k = 1$,

> if $\tilde{T}_k \geq a_k$ then reject the null and stop accruing data,
> if $\tilde{T}_k \leq b_k$ then fail rejecting the null and stop accruing data,
> if $b_k < \tilde{T}_k < a_k$ then set $k \leftarrow k + 1$ and repeat.

If $(\tilde{T}_1, \ldots, \tilde{T_K})$ had the same distribution as $(Z_1, \ldots, Z_K)$, then the latter rule would yield a testing procedure with the required type I error and type II error at the specified alternative parameter.

The merit is clear of targeting the fixed design $g^\star(\theta)$ that makes the estimate Ψ_n have the optimal (*i.e.* smallest) asymptotic variance, the random variables N_k (*i.e.* successive number of observations required for testing) then being, at least informally, stochastically smaller than they would have been had another fixed design been used (or targeted).

Carrying out the targeted optimal design group sequential testing procedure.

We wish to test the null "$\Psi(\theta) = \psi_0$" against "$\Psi(\theta) > \psi_0$" with asymptotic type I error α and asymptotic type II error β at some $\psi_1 > \psi_0$. We intend to proceed group sequentially with $K \geq 2$ steps.

Highlight 2 (targeted optimal design adaptive group sequential testing)**.** *To do so:*

1. *Choose increasingly ordered proportions* $0 < p_1 < \ldots < p_K = 1$.
2. *Compute numerically the maximum committed information* $I_{\max}$*, rejection and futility boundaries* $(a_1, \ldots, a_K)$ *and* $(b_1, \ldots, b_K)$.
3. *Starting from* $k = 1$*,*
   (a) *keep sampling under* $(\theta, \mathbf{g}_n^\star)$*; as soon as* N_k *data are collected,*
   (b) *compute* $\tilde{T}_k$*,*
   (c) *apply the following rule:*
   if $\tilde{T}_k \geq a_k$ *then reject the null and stop accruing data,*
   if $\tilde{T}_k \leq b_k$ *then fail rejecting the null and stop accruing data,*
   if $b_k < \tilde{T}_k < a_k$ *then set* $k \leftarrow k + 1$ *and repeat.*

We investigate the behavior of this group sequential testing procedure from a theoretical perspective in Section 5.2, and by simulations in Section 7, focusing on empirical type I and type II errors in Section 7.2 and on empirical sample sizes at decision in Section 7.3.

5.2 Asymptotic study of the targeted optimal design group sequential testing procedure powered at local alternatives.

In order to tackle the asymptotic study of the targeted optimal design group sequential testing procedure, we resort to contiguity arguments. According to Chapter 6 in [28], "contiguity arguments are a technique to obtain the limit distribution of a sequence of statistics under underlying laws Q_n from a limiting distribution under laws P_n." Here the laws P_n describe a null distribution under investigation (the distribution of the test statistic under $(\theta, \mathbf{g}_n^\star)$), and the laws Q_n correspond to an alternative hypothesis.

Proving the validity of the targeted optimal design group sequential testing procedure (as defined in Section 5.1) powered at local alternatives is outside the scope of this article. We rather consider a slightly simpler version where the random N_k are replaced by deterministic n_k. We conjecture that the theorem we prove in the simpler deterministic sample sizes context can be extended to the random sample sizes context. The simulation study that we undertake in Section 7 confirms the conjecture.

Once again, let $0 < p_1 < \ldots < p_K = 1$ be increasingly ordered proportions for some integer $K \geq 2$, and define $n_k = \lceil np_k \rceil$ (the smallest integer not smaller than np_k) for each $k \leq K$. We wish to the test the null against its alternative based on the multidimensional t-statistic

$$(T_1, \ldots, T_K) = \left(\frac{\sqrt{n_k}(\Psi_{n_k} - \psi_0)}{s_{n_k}} \right)_{k \leq K} \tag{25}$$

(s_n^2 estimates the asymptotic variance of $\sqrt{n}(\Psi_n - \Psi(\theta))$ under $(\theta, \mathbf{g}_n^\star)$ — see Theorem 2). Before going any further, we state a crucial theorem which describes how the test statistic converges towards the so-called *canonical distribution*.

Theorem 3. *Consider $h = (h_0, h_1) \in \mathbb{R}^2$ satisfying both $h_1 > h_0$ and $\gamma h_1 + \gamma^{-1} h_0 \neq 0$ where $\gamma = \mathrm{OR}(\theta) g^\star(\theta)(1)/g^\star(\theta)(0)$. Define $\theta_{h/\sqrt{n}} = (\theta_0(1 + h_0/\sqrt{n}), \theta_1(1 + h_1/\sqrt{n}))$ for all $n \geq n_0$ large enough to ensure $\theta_{h/\sqrt{n}} \in]0,1[^2$. The sequence $(\theta_{h/\sqrt{n}})_{n \geq n_0}$ defines a sequence $(\psi_n)_{n \geq n_0}$ of contiguous parameters ("from direction h"), with $\psi_n = \Psi(\theta_{h/\sqrt{n}}) > \Psi(\theta)$.*

Introduce the mean vector $\mu(h) = (h_1 - h_0)(\sqrt{p_1}, \ldots, \sqrt{p_K})/\sqrt{v^\star(\theta)}$ and the covariance matrix $\mathcal{C} = (\sqrt{p_{k \wedge l}/p_{k \vee l}})_{k,l \leq K}$. Then:

(i) under $(\theta, \mathbf{g}_n^\star)$, $(T_1, \ldots, T_K)$ converges in distribution, as n tends to infinity, to the centered Gaussian distribution with covariance matrix $\mathcal{C}$;

(ii) under $(\theta_{h/\sqrt{n}}, \mathbf{g}_n^\star)$, $(T_1, \ldots, T_K)$ converges in distribution, as n tends to infinity, to the Gaussian distribution with mean $\mu(h)$ and covariance matrix $\mathcal{C}$.

Say we want to perform a test such with asymptotic type I error α and asymptotic power $(1 - \beta)$ at the limit of the sequence of contiguous parameters $(\psi_n)_{n \geq n_0}$, *i.e.* such that *(a)* the probability of rejecting the null for its alternative under $(\theta, \mathbf{g}_n^\star)$, and *(b)* the probability of failing to reject the null for its alternative under $(\theta_{h/\sqrt{n}}, \mathbf{g}_n^\star)$ converge (as n tends to infinity) towards α and β, respectively.

Consider the α-spending and β-spending strategies $(\alpha_1, \ldots, \alpha_K)$ and $(\beta_1, \ldots, \beta_K)$. It is usually assumed that the next lemma holds:

Lemma 4. *In the framework of Theorem 3, let $(Z_1, \ldots, Z_K)$ follow the centered Gaussian distribution with covariance matrix $\mathcal{C}$ (as defined in Theorem 3). Assume that $\alpha + \beta < 1$. There exits a unique $\varepsilon > 0$, a unique rejection boundary $(a_1, \ldots, a_K)$, a unique futility boundary $(b_1, \ldots, b_K)$ such that $a_K = b_K$, $P(Z_1 \geq a_1) = \alpha_1$, $P(Z_1 + \mu(\varepsilon h)_1 \leq b_1) = \beta_1$, and for every $1 \leq k < K$,*

$$\begin{aligned} P(\forall j \leq k, b_j < Z_j < a_j \text{ and } Z_{k+1} \geq a_{k+1}) &= \alpha_{k+1}, \\ P(\forall j \leq k, b_j < Z_j + \mu(\varepsilon h)_j < a_j \text{ and } Z_{k+1} + \mu(\varepsilon h)_{k+1} \leq b_{k+1}) &= \beta_{k+1}. \end{aligned}$$

Given such rejection and futility boundaries, we proceed as follows: starting from $k = 1$,

if $T_k \geq a_k$ then reject the null and stop accruing data,
if $T_k \leq b_k$ then fail rejecting the null and stop accruing data,
if $b_k < T_k < a_k$ then set $k \leftarrow k+1$ and repeat

Theorem 3 and Lemma 4 teach us that the group sequential testing procedure described above satisfies the requirements on stepwise type I and type II error control, once h is replaced with εh (which actually corresponds to a shift in n in the definition of the sequence of contiguous alternatives $(\theta_{h/\sqrt{n}})_{n \geq n_0}$).

5.3 Proof of Theorem 3.

Theorem 3 is a corollary of the following

Lemma 5. *Denote by Λ_n the* log*-likelihood ratio of the $(\theta_{h/\sqrt{n}}, \mathbf{g}_n^\star)$ experiment with respect to the $(\theta, \mathbf{g}_n^\star)$ experiment, as defined in Theorem 3. There exists a constant $\tau^2 > 0$ such that, under $(\theta, \mathbf{g}_n^\star)$, the vector $(\sqrt{n_1}(\Psi_{n_1} - \Psi(\theta)), \ldots, \sqrt{n_K}(\Psi_{n_K} - \Psi(\theta)), \Lambda_n)$ converges in distribution, as n tends to infinity, to the Gaussian distribution with mean $(0, \ldots, 0, -\frac{1}{2}\tau^2)$ and covariance matrix*

$$\begin{pmatrix} \left(\sqrt{\frac{p_{k\wedge l}}{p_{k\vee l}}} v^\star(\theta)\right)_{k,l \leq K} & \sqrt{v^\star(\theta)}\mu(h)^\top \\ \sqrt{v^\star(\theta)}\mu(h) & \tau^2 \end{pmatrix}.$$

In particular, the $(\theta_{h/\sqrt{n}}, \mathbf{g}_n^\star)$ and $(\theta, \mathbf{g}_n^\star)$ experiments are mutually contiguous.

It is easy to obtain the limiting distribution of $(T_1, \ldots, T_K)$ under $(\theta, \mathbf{g}_n^\star)$ from Lemma 5. Le Cam's third lemma solves the problem of obtaining the limiting distribution of $(T_1, \ldots, T_K)$ under $(\theta_{h/\sqrt{n}}, \mathbf{g}_n^\star)$ from the convergence under $(\theta, \mathbf{g}_n^\star)$ exhibited in Lemma 5. The second limiting distribution is still Gaussian, has the same asymptotic covariance matrix as under $(\theta, \mathbf{g}_n^\star)$, but differs by its asymptotic mean which is no longer 0.

Proof of Theorem 3. By the continuous mapping theorem, we obtain from Lemma 5 the convergence in distribution under $(\theta, \mathbf{g}_n^\star)$ of $(\sqrt{n_1}(\Psi_{n_1} - \Psi(\theta)), \ldots, \sqrt{n_K}(\Psi_{n_K} - \Psi(\theta)))$ to the centered Gaussian distribution with covariance matrix $(v^\star(\theta)\sqrt{p_{k\wedge l}/p_{k\vee l}})_{k,l \leq K}$. Then Slutsy's lemma straightforwardly yield the first convergence *(i)*.

Regarding *(ii)*, we first invoke Lemma 5 and Le Cam's third lemma (see Example 6.7 in [28]) in order to obtain that, under $(\theta_{h/\sqrt{n}}, \mathbf{g}_n^\star)$, the vector $(\sqrt{n_1}(\Psi_{n_1} - \Psi(\theta)), \ldots, \sqrt{n_K}(\Psi_{n_K} - \Psi(\theta)))$ converges in distribution to the Gaussian distribution with mean $\sqrt{v^\star(\theta)}\mu(h)$ and covariance matrix $(v^\star(\theta)\sqrt{p_{k\wedge l}/p_{k\vee l}})_{k,l \leq K}$. In addition, the $(\theta_{h/\sqrt{n}}, \mathbf{g}_n^\star)$ and $(\theta, \mathbf{g}_n^\star)$ experiments are mutually contiguous, implying that if s_n^2 estimates $v^\star(\theta)$ under $(\theta, \mathbf{g}_n^\star)$, then it also estimates $v^\star(\theta)$ under $(\theta_{h\sqrt{n}}, \mathbf{g}_n^\star)$. We apply again Slutsky's lemma in order to obtain the second convergence *(ii)*. □

Proof of Lemma 5. Let us consider first the log-likelihood ratio of the $(\theta_{h/\sqrt{n}}, \mathbf{g}_n^\star)$ experiment with respect to the $(\theta, \mathbf{g}_n^\star)$ experiment. The shorthand notations $\theta(O) = \theta_A^Y(1-\theta_A)^{1-Y}$ and $\theta_{h/\sqrt{n}}(O) = [\theta_A(1+h_A/\sqrt{n})]^Y[1-\theta_A(1+h_A/\sqrt{n})]^{1-Y}$ and (11) readily yield

$$\begin{aligned}
\Lambda_n &= \log \prod_{i=1}^n \frac{g_i^\star(A_i|Y_i, \mathbf{O}_n(i-1))}{g_i^\star(A_i|Y_i, \mathbf{O}_n(i-1))} \prod_{i=1}^n \frac{\theta_{h/\sqrt{n}}(O_i)}{\theta(O_i)} \\
&= \sum_{i=1}^n \mathbb{1}\{A_i = 0\} \left[Y_i \log\left(1 + \frac{h_0}{\sqrt{n}}\right) + (1-Y_i)\log\left(1 - \frac{\theta_0 h_0}{(1-\theta_0)\sqrt{n}}\right)\right] \\
&\quad + \mathbb{1}\{A_i = 1\} \left[Y_i \log\left(1 + \frac{h_1}{\sqrt{n}}\right) + (1-Y_i)\log\left(1 - \frac{\theta_1 h_1}{(1-\theta_1)\sqrt{n}}\right)\right] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n L_1(O_i) - \frac{1}{2n} \sum_{i=1}^n L_2(O_i) + o_P(1),
\end{aligned}$$

with

$$\begin{aligned} L_1(O_i) &= \mathbb{1}\{A_i=0\}\frac{Y_i-\theta_0}{1-\theta_0}h_0 + \mathbb{1}\{A_i=1\}\frac{Y_i-\theta_1}{1-\theta_1}h_1, \\ L_2(O_i) &= \mathbb{1}\{A_i=0\}\left[Y_i+(1-Y_i)\left(\frac{\theta_0}{1-\theta_0}\right)^2\right]h_0^2 \\ &\quad +\mathbb{1}\{A_i=1\}\left[Y_i+(1-Y_i)\left(\frac{\theta_1}{1-\theta_1}\right)^2\right]h_1^2. \end{aligned}$$

First, we observe that, since L_2 is bounded (and measurable), $\frac{1}{n}\sum_{i=1}^n L_2(O_i) = \frac{1}{n}\sum_{i=1}^n P_{\theta,g_i^\star}L_2 + \frac{1}{n}\sum_{i=1}^n [L_2(O_i) - P_{\theta,g_i^\star}L_2] = \frac{1}{n}\sum_{i=1}^n P_{\theta,g_i^\star}L_2 + o_P(1)$ by virtue of the Kolmogorov law of large numbers. Now, $P_{\theta,g_i^\star}L_2 = g_i^\star(0)\frac{\theta_0 h_0^2}{1-\theta_0} + g_i^\star(1)\frac{\theta_1 h_1^2}{1-\theta_1}$, hence $\frac{1}{n}\sum_{i=1}^n P_{\theta,g_i^\star}L_2 = \frac{1}{n}\sum_{i=1}^n g_i^\star(0)\frac{\theta_0 h_0^2}{1-\theta_0} + \frac{1}{n}\sum_{i=1}^n g_i^\star(1)\frac{\theta_1 h_1^2}{1-\theta_1} = g^\star(\theta)(0)\frac{\theta_0 h_0^2}{1-\theta_0} + g^\star(\theta)(1)\frac{\theta_1 h_1^2}{1-\theta_1} + o_P(1)$ by virtue of Theorem 1. In summary, we obtain that

$$\Lambda_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n L_1(O_i) - \frac{1}{2}\tau^2 + o_P(1) \tag{26}$$

for $\tau^2 = g^\star(\theta)(0)\frac{\theta_0 h_0^2}{1-\theta_0} + g^\star(\theta)(1)\frac{\theta_1 h_1^2}{1-\theta_1}$.

Second, we define $Z_i = (\mathbb{1}\{i \le n_1\}, \ldots, \mathbb{1}\{i \le n_K\})$ and introduce the bounded (and measurable) function f such that $f(O_i, Z_i) = (Z_i \mathrm{IC}(\theta, g^\star(\theta))(O_i), L_1(O_i))$. Let us show that $M_n = \frac{1}{n}\sum_{i=1}^n [f(O_i,Z_i) - P_{\theta,g_i^\star}f] = \frac{1}{n}\sum_{i=1}^n f(O_i,Z_i)$ (this equality holds because for all $i \le n$, one has $P_{\theta,g_i^\star}\mathrm{IC}(\theta, g^\star(\theta)) = P_{\theta,g_i^\star}L_1 = 0$) satisfies a central limit theorem. In view of Theorem 10, let $W_n = \sum_{i=1}^n P_{\theta,g_i^\star} f^\top f$ and $\Sigma_n = \frac{1}{n}EW_n$. The entries of matrix W_n write either $\sum_{i=1}^{n_k} P_{\theta,g_i^\star}L_1\mathrm{IC}(\theta,g^\star(\theta))$, or $\sum_{i=1}^{n_k\wedge n_l} P_{\theta,g_i^\star}\mathrm{IC}(\theta,g^\star(\theta))^2$, or $\sum_{i=1}^n P_{\theta,g_i^\star}L_1^2$. Now,

- $A_n = \sum_{i=1}^{n_k} P_{\theta,g_i^\star}L_1\mathrm{IC}(\theta,g^\star(\theta)) = \frac{h_1}{g^\star(1)}\sum_{i=1}^{n_k} g_i^\star(1) - \frac{h_0}{g^\star(0)}\sum_{i=1}^{n_k} g_i^\star(0)$, so that $\frac{1}{n}EA_n = p_k(h_1-h_0) + o(1)$ and $\frac{1}{n}A_n - \frac{1}{n}EA_n = o_P(1)$ since the almost sure convergence of the bounded sequence $\frac{1}{n}\sum_{i=1}^n g_i^\star(a)$ towards $g^\star(\theta)(a)$ (see Theorem 1) implies its convergence in L^1 norm to the same limit;
- $B_n = \sum_{i=1}^{n_k\wedge n_l} P_{\theta,g_i^\star}\mathrm{IC}(\theta,g^\star(\theta))^2 = \frac{1-\theta_1}{\theta_1 g^\star(\theta)(1)^2}\sum_{i=1}^{n_k\wedge n_l} g_i^\star(1) + \frac{1-\theta_0}{\theta_0 g^\star(\theta)(0)^2}\sum_{i=1}^{n_k\wedge n_l} g_i^\star(0)$, hence $\frac{1}{n}EB_n = p_{k\wedge l}\left(\frac{1-\theta_1}{\theta_1 g^\star(\theta)(1)} + \frac{1-\theta_0}{\theta_0 g^\star(\theta)(0)}\right) + o(1) = p_{k\wedge l}v^\star(\theta) + o(1)$ and $\frac{1}{n}B_n - \frac{1}{n}EB_n = o_P(1)$ for the same reasons as above;
- $C_n = \sum_{i=1}^n P_{\theta,g_i^\star}L_1^2 = \frac{\theta_1 h_1^2}{1-\theta_1}\sum_{i=1}^n g_i^\star(1) + \frac{\theta_0 h_0^2}{1-\theta_0}\sum_{i=1}^n g_i^\star(0)$, hence $\frac{1}{n}EC_n = \tau^2 + o(1)$ and $\frac{1}{n}C_n - \frac{1}{n}EC_n = o_P(1)$ for the same reasons as above.

Those calculations notably teach us that, setting $m = (h_1-h_0)(p_1,\ldots,p_K)$ and $\Sigma_0 = (p_{k\wedge l})_{k,l\le K}$, Σ_n converges to

$$\Sigma = \begin{pmatrix} v^\star(\theta)\Sigma_0 & m^\top \\ m & \tau^2 \end{pmatrix}.$$

Is Σ a positive definite covariance matrix? Well, Σ_0 is a positive definite covariance matrix (that of the vector $(B_{p_1},\ldots,B_{p_K})$ where $(B_t)_{t\ge 0}$ is a standard Brownian motion), hence the symmetric matrix Σ is a positive definite covariance matrix if and only if its determinant $\det(\Sigma) > 0$. Subtracting $(h_1-h_0)/v^\star(\theta)$ times the Kth row of Σ to its last row, we get that $\det(\Sigma) = v^\star(\theta)^K \det(\Sigma_0) \times (\tau^2 - (h_1-h_0)^2/v^\star(\theta))$. Now, using $v^\star(\theta) = \frac{1-\theta_0}{\theta_0 g^\star(\theta)(0)} + \frac{1-\theta_1}{\theta_1 g^\star(\theta)(1)}$ and $\gamma h_1 + \gamma^{-1}h_0 \neq 0$ (required in Theorem 3) yields

$$\begin{aligned} &v^\star(\theta)\tau^2 - (h_1-h_0)^2 \\ &= h_1^2\left(\frac{\theta_1 v^\star(\theta)}{1-\theta_1}g^\star(\theta)(1) - 1\right) + h_0^2\left(\frac{\theta_0 v^\star(\theta)}{1-\theta_0}g^\star(\theta)(0) - 1\right) + 2h_0h_1 \end{aligned}$$

$$= (\gamma h_1 + \gamma^{-1} h_0)^2 > 0.$$

In summary, Σ is a positive definite covariance matrix, the conditions of Theorem 10 are met, and therefore $\sqrt{n}M_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(O_i, Z_i)$ converges in distribution to the centered Gaussian distribution with covariance matrix Σ.

Let $\Delta_n = \text{diag}(\sqrt{n/n_1}, \ldots, \sqrt{n/n_K}, 1)$, $\Delta = \text{diag}(1/\sqrt{p_1}, \ldots, 1/\sqrt{p_K}, 1)$; obviously, $\Delta_n = \Delta + o(1)$ and $\sqrt{n}M_n\Delta_n = \sqrt{n}M_n\Delta + o_P(1)$. Invoking (21) in Theorem 2 and (26), it holds that, under $(\theta, \mathbf{g}_n^\star)$,

$$\begin{aligned}(\sqrt{n_1}(\Psi_{n_1} - \Psi(\theta)), \ldots, \sqrt{n_K}(\Psi_{n_K} - \Psi(\theta)), \Lambda_n) \\ = (0, \ldots, 0, -\tfrac{1}{2}\tau^2) + \sqrt{n}M_n\Delta_n + o_P(1) \\ = (0, \ldots, 0, -\tfrac{1}{2}\tau^2) + \sqrt{n}M_n\Delta + o_P(1).\end{aligned}$$

This entails the convergence in distribution, under $(\theta, \mathbf{g}_n^\star)$, of $(\sqrt{n_1}(\Psi_{n_1} - \Psi(\theta)), \ldots, \sqrt{n_K}(\Psi_{n_K} - \Psi(\theta)), \Lambda_n)$ to the Gaussian distribution with mean $(0, \ldots, 0, -\frac{1}{2}\tau^2)$ and covariance matrix $\Delta\Sigma\Delta$. Simple calculations finally reveal that $\Delta\Sigma\Delta$ equals the positive definite covariance matrix given in the lemma. □

6 Simulation study of the performances of targeted optimal design adaptive estimation

In this section, we carry out a simulation study of the performances of targeted optimal design adaptive procedures in terms of estimation and uncertainty assessment. The two main questions at stake are "Do the confidence intervals obtained under the targeted optimal design adaptive sampling scheme guarantee the desired coverage?" and "How well do they compare with the intervals we would obtain under the targeted optimal iid sampling scheme?"

We carefully present the simulation scheme in Section 6.1. We validate with simulations the central limit Theorem 2 that we derived theoretically in Section 4.2. The section culminates in Section 6.3 with the investigation of the covering properties of the confidence intervals based on the data-driven sampling schemes. Then, we consider the performances in terms of widths of the confidence intervals in Section 6.4, Section 6.5 finally containing an illustration of the procedure.

6.1 The simulation scheme.

Define $\varepsilon = 0.1$ and the ε-net $\Theta_0 = \{(i\varepsilon, j\varepsilon) : 1 \leq i \leq j \leq 9\}$ over the set $\{(\theta_0, \theta_1) : \varepsilon \leq \theta_0 \leq \theta_1 \leq 1 - \varepsilon\}$. It has cardinality $\#\Theta_0 = 45$. The log-relative risk function Ψ maps Θ_0 onto the set $\Psi(\Theta_0) \subset [0; 2.1973]$, see Table 1, which is well described by its cumulative distribution function (cdf) plotted in Figure 2. The set $R(\Theta_0) \subset [0.6097; 1]$ is presented in Table 2. It is also interesting to look in Figure 3 at the left-hand plot of $\{(\Psi(\theta), R(\theta)) : \theta \in \Theta_0\}$. All $\theta \in \Theta_0$ which are on the diagonal are associated with a log-relative risk $\Psi(\theta) = 0$ and a relative efficiency $R(\theta) = 1$ and are therefore represented by the single point $(0, 1)$. It is also seen in the left-hand plot of Figure 3 that the relative efficiency $R(\theta)$ can be significantly lower than 1 even when $\Psi(\theta)$ is not large.

Table 3 and the two right-hand plots in Figure 3 are even more interesting, because our search of efficiency relies for each $\theta \in \Theta_0$ on targeting its optimal treatment mechanism $g^\star(\theta)$. In Table 3 we report the various optimal proportions of treated $g^\star(\theta)(1)$. In the two right-hand plots in Figure 3, we represent the optimal proportion of treated $g^\star(\theta)(1)$ against the log-relative risk $\Psi(\theta)$ (middle plot) and against the relative efficiency $R(\theta)$ (rightmost plot). Table 3 and the rightmost plot in Figure 3 both illustrate the closed form equality

$$g^\star(\theta)(1) = \frac{1}{2}\left(1 - \sqrt{\frac{1}{R(\theta)} - 1}\right)$$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0	0.693	1.099	1.386	1.609	1.792	1.946	2.079	2.197
.2	-	0	0.405	0.693	0.916	1.099	1.253	1.386	1.504
.3	-	-	0	0.288	0.511	0.693	0.847	0.981	1.099
.4	-	-	-	0	0.223	0.405	0.560	0.693	0.811
.5	-	-	-	-	0	0.182	0.336	0.470	0.588
.6	-	-	-	-	-	0	0.154	0.288	0.405
.7	-	-	-	-	-	-	0	0.134	0.251
.8	-	-	-	-	-	-	-	0	0.118
.9	-	-	-	-	-	-	-	-	0

Table 1: Values of $\Psi(\theta)$ for $\theta \in \Theta_0$ (with precision 10^{-3}).

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1	0.962	0.904	0.850	0.800	0.753	0.708	0.662	0.610
.2	-	1	0.982	0.945	0.900	0.850	0.796	0.735	0.662
.3	-	-	1	0.988	0.958	0.916	0.862	0.796	0.708
.4	-	-	-	1	0.990	0.962	0.916	0.850	0.753
.5	-	-	-	-	1	0.990	0.958	0.9	0.800
.6	-	-	-	-	-	1	0.988	0.945	0.850
.7	-	-	-	-	-	-	1	0.982	0.904
.8	-	-	-	-	-	-	-	1	0.962
.9	-	-	-	-	-	-	-	-	1

Table 2: Values of $R(\theta)$ for $\theta \in \Theta_0$ (with precision 10^{-3}).

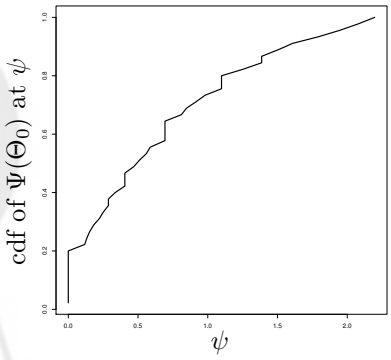

Figure 2: Cumulative distribution function $\psi \mapsto \frac{1}{\#\Theta_0} \sum_{\theta \in \Theta_0} \mathbb{1}\{\Psi(\theta) \leq \psi\}$ of $\Psi(\Theta_0)$.

which can be easily derived from (5) and (6) (using that $g^\star(\theta)(1) \leq \frac{1}{2}$ because $\theta_0 \leq \theta_1$). The above equality, related table and figure teach us that more significant gains in terms of relative efficiency $R(\theta)$ correspond to smaller optimal proportions of treated $g^\star(\theta)(1)$.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0.500	0.400	0.337	0.290	0.250	0.214	0.179	0.143	0.100
.2	-	0.500	0.433	0.380	0.333	0.290	0.247	0.200	0.143
.3	-	-	0.500	0.445	0.396	0.348	0.300	0.247	0.179
.4	-	-	-	0.500	0.449	0.400	0.348	0.290	0.214
.5	-	-	-	-	0.500	0.449	0.396	0.333	0.250
.6	-	-	-	-	-	0.500	0.445	0.380	0.290
.7	-	-	-	-	-	-	0.500	0.433	0.337
.8	-	-	-	-	-	-	-	0.500	0.400
.9	-	-	-	-	-	-	-	-	0.500

Table 3: Values of $g^\star(\theta)(1)$ for $\theta \in \Theta_0$ (with precision 10^{-3}).

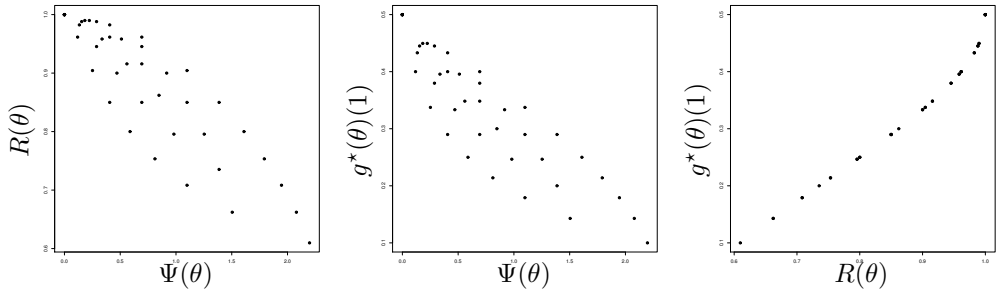

Figure 3: Plots of $\{(\Psi(\theta), R(\theta)) : \theta \in \Theta_0\}$ (left), $\{(\Psi(\theta), g^\star(\theta)(1)) : \theta \in \Theta_0\}$ (middle) and $\{(R(\theta), g^\star(\theta)(1)) : \theta \in \Theta_0\}$ (right).

Let $n = (100, 250, 500, 750, 1000, 2500, 5000)$ be a sequence of sample sizes. For every $\theta \in \Theta_0$, we estimate $M = 1000$ times the log-relative risk $\Psi(\theta)$ based on $\mathbf{O}^m_{n_7}(n_i)$, $m = 1, \ldots, M, i = 1, \ldots, 7$, under

- iid (θ, g^b)-balanced sampling,
- iid $(\theta, g^\star(\theta))$-optimal sampling,
- $(\theta, \mathbf{g}^\star_{n_7})$-adaptive sampling,
- $(\theta, \mathbf{g}^a_{n_7})$-adaptive sampling.

We choose $\delta = 0.01$ in (14).

6.2 Empirical distribution of maximum likelihood estimates.

In Theorem 2 we proved that a central limit result holds for Ψ_n when targeting the optimal design, as it is obviously the case under iid sampling. In order to check by simulations that remarkable property and to determine how quickly the limit is reached, we propose the following procedure.

Testing the empirical distribution of maximum likelihood estimates.

For every $\theta \in \Theta_0$, all types of sampling, and each sample size n_i, we compare the empirical distribution of the (centered and rescaled) estimators of $\Psi(\theta)$

$$Z(\theta)_{n_i,m} = \frac{\sqrt{n_i}\left(\Psi_{n_i}(\mathbf{O}^m_{n_7}(n_i)) - \Psi(\theta)\right)}{\sqrt{v(\theta)}}, \quad m = 1, \ldots, M$$

(where $v(\theta) = v^b(\theta)$ under balanced iid sampling and $v(\theta) = v^\star(\theta)$ otherwise) with its standard normal theoretical limit distribution in terms of two-sided Kolmogorov-Smirnov goodness-of-fit test. This results in a collection of independent p-values $\{P(\theta)^{clt}_{n_i} : \theta \in \Theta_0, i = 1, \ldots, 7\}$ which are uniformly distributed under the null hypothesis stating that all $Z(\theta)_{n_i,m}$ follow the standard normal distribution.

Under the null, $\{P(\theta)^{clt}_{n_i} : \theta \in \Theta_0\}$ contains iid copies of the Uniform distribution over $[0; 1]$ for every $i = 1, \ldots, 7$. This statement can be tested in terms of one-sided Kolmogorov-Smirnov goodness-of-fit procedure, the alternative stating that these iid random variables are stochastically smaller than a uniform random variable, hence 7 final p-values for each sampling scheme as reported in Table 4.

	sample size						
sampling scheme	n_1	n_2	n_3	n_4	n_5	n_6	n_7
iid g^b-balanced	$p < 0.001$	0.001	0.001	0.089	0.124	0.369	0.886
iid $g^\star$-optimal	$p < 0.001$	0.001	0.612	0.094	0.381	0.764	0.947
$\mathbf{g}^\star_n$-adaptive	$p < 0.001$	$p < 0.001$	0.027	0.060	0.042	0.156	0.123
$\mathbf{g}^a_n$-adaptive	$p < 0.001$	$p < 0.001$	$p < 0.001$	0.003	0.006	0.136	0.460

Table 4: Checking the central limit theorem validity by simulation. We test if the independent random variables $\{P(\theta)^{clt}_{n_i} : \theta \in \Theta_0\}$ are uniformly distributed over $[0; 1]$ according to the one-sided Kolmogorov-Smirnov goodness-of-fit test, the alternative stating that they are stochastically smaller than a uniform random variable: we report p-values for each sample size n_i, $i = 1, \ldots, 7$ and each sampling scheme.

Empirical validation of the central limit theorem.

It is not surprising that p-values are very small for smaller sample sizes n_1, n_2, n_3. Considering each sampling scheme (*i.e.* each row of Table 4) separately, we conclude that

the central limit theorem is not rejected under

- iid g^b-balanced sampling for any sample size $n_i \geq n_4 = 750$,
- iid $g^\star$-optimal sampling for any sample size $n_i \geq n_3 = 500$,
- $\mathbf{g}^\star_n$-adaptive sampling for any sample size $n_i \geq n_3 = 500$,
- $\mathbf{g}^a_n$-adaptive sampling for any sample size $n_i \geq n_6 = 2500$,

adjusting for multiple testing in terms of the Benjamini and Yekutieli procedure for controlling the False Discovery Rate at level 5%.

Less formally, the Gaussian limit theoretically guaranteed by the central limit theorem is reached under iid $g^\star$-optimal and $\mathbf{g}^\star_n$-adaptive sampling schemes as soon as 500 observations are accrued. The limit is reached as soon as 750 observations are collected when considering the iid g^b-balanced sampling scheme. This is a very satisfying result for the $\mathbf{g}^\star_n$-adaptive sampling scheme. On the contrary, the limit is reached for a surprisingly large minimal sample size under $\mathbf{g}^a_n$-adaptive sampling scheme. This said, we are less interested in the minimal sample size required to reach the Gaussian limit than in the minimal sample size required to guarantee the desired coverage

properties to our confidence intervals. The coverage properties of our confidence intervals are investigated in Section 6.3.

In conclusion,

Highlight 3 (empirical validation of central limit theorem)**.** *In view of Theorem 2, the convergence of* $\sqrt{n}(\Psi_n - \Psi(\theta))$ *to its limit Gaussian distribution under adaptive* $(\theta, \mathbf{g}_n^\star)$ *sampling scheme is empirically reached as soon as 500 observations are accrued. This is as good as what we get under the iid* $(\theta, g^\star)$ *optimal sampling scheme.*

Illustrating the convergence.

To give a sense of how well the standard normal limit distribution is reached, it is interesting to consider, for each adaptive sampling scheme and for the corresponding first sample size for which the central limit theorem is not rejected, that empirical cdf which is the farthest to the standard normal limit cdf. How far an empirical cdf is from the standard normal cdf is measured in terms of p-value of the two-sided Kolmogorov-Smirnov goodness-of-fit test. For a sample size $n_3 = 500$ (the first sample size for which the central limit theorem is not rejected under the $\mathbf{g}_n^\star$-adaptive sampling scheme; that first sample size is $n_6 = 2500$ for the $\mathbf{g}_n^a$-adaptive sampling scheme), it is also interesting to compare the worse empirical cdf obtained under $\mathbf{g}_n^\star$-adaptive sampling scheme to the worse empirical cdf obtained under $\mathbf{g}_n^a$-adaptive sampling scheme.

Thus, we represent in Figure 4 (left) the empirical cdf of the sequence $(Z(\theta^-)_{n_3,m})_{m\leq M}$ with $\theta^- = \arg\min_{\theta\in\Theta_0} P(\theta)_{n_3}^{clt}$ under adaptive $(\theta^-, \mathbf{g}_{n_3}^\star)$ sampling. We obtain $\theta^- = (0.3, 0.9)$ (for which $\Psi(\theta^-) = 1.0986$). Even though $P(\theta^-)_{n_3}^{clt} \simeq 0.0017$, the empirical cdf and its limit are almost superposable.

Similarly, we represent in Figure 4 (middle) the empirical cdf of the sequence $(Z(\theta'^-)_{n_6,m})_{m\leq M}$ associated with $\theta'^- = \arg\min_{\theta\in\Theta_0} P(\theta)_{n_6}^{clt}$ under adaptive $(\theta'^-, \mathbf{g}_{n_6}^a)$ sampling. We obtain $\theta'^- = (0.1, 0.9)$ (for which $\Psi(\theta'^-) = 2.1972$). Again, the empirical cdf and its limit are almost superposable.

Finally, we also represent in Figure 4 (right) the empirical cdf of the sequence $(Z(\theta^-)_{n_3,m})_{m\leq M}$, that of the sequence $(Z(\theta''^-)_{n_3,m})_{m\leq M}$ associated with $\theta''^- = \arg\min_{\theta\in\Theta_0} P(\theta)_{n_3}^{clt}$ under adaptive $(\theta''^-, \mathbf{g}_{n_3}^a)$ sampling, that is before the asymptotic distribution is reached for that design, and their common limit. We obtain $\theta''^- = \theta'^- = (0.1, 0.9)$. A logarithmic scale is used on the y-axis in order to enhance the differences occurring at the left tail. The $Z(\theta''^-)_{n_3,m}$'s are visibly stochastically (empirically) larger than the $Z(\theta'^-)_{n_3,m}$'s, themselves slightly stochastically (empirically) larger than a standard normal random variable.

6.3 Empirical coverage of the confidence intervals.

We invoke the central limit theorem (Theorem 2) in order to construct confidence intervals for the log-relative risk. The empirical validation of the theorem presented in Section 6.2 also provides us with an indirect validation of the coverage properties of those confidence intervals. However it is interesting to test directly if the coverage requirements are satisfied. Obviously, Section 6.3 is the most important subsection of Section 6.

Testing the empirical coverage of the confidence intervals.

Set $\alpha = 5\%$. For every $\theta \in \Theta_0$, all types of sampling, every iteration m and each sample size n_i, we estimate the asymptotic variance of the maximum likelihood estimator $\Psi_{n_i}(\mathbf{O}_{n_7}^m(n_i))$ with $s(\theta)_{n_i,m}^2$ and build the confidence interval

$$\mathcal{I}(\theta)_{n_i,m} = \left[\Psi_{n_i}(\mathbf{O}_{n_7}^m(n_i)) \pm \frac{s(\theta)_{n_i,m}}{\sqrt{n_i}}\xi_{1-\alpha/2}\right]$$

where $\xi_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile if the standard normal distribution. We are interested in the empirical coverage guaranteed by $\mathcal{I}(\theta)_{n_i,m}$ (its width will be considered in Section 6.4).

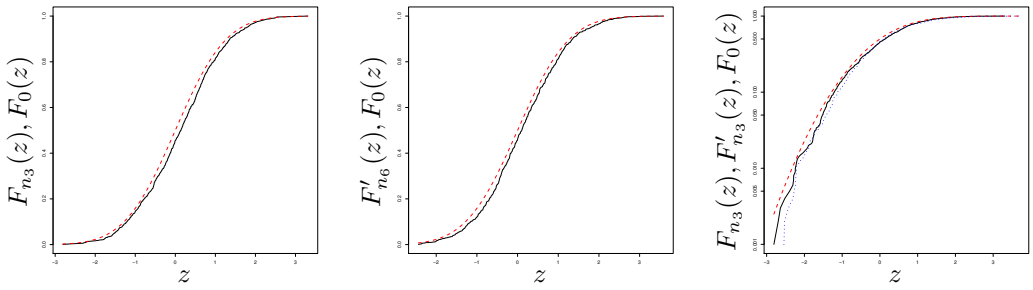

Figure 4: Giving a sense of how well the standard normal limit distribution is reached under each adaptive sampling scheme. Left: *Under* $\mathbf{g}_n^\star$*-adaptive sampling scheme and for sample size* $n_3 = 500$, empirical cdf F_{n_3} (solid line) of the sequence $(Z(\theta^-)_{n_3,m})_{m\leq M}$ whose empirical distribution is the further from its limit standard normal distribution. The reference limit cdf F_0 is also plotted (dashed). Middle: *Under* $\mathbf{g}_n^a$*-adaptive sampling scheme and for sample size* $n_6 = 2500$, empirical cdf F'_{n_6} (solid line) of the sequence $(Z(\theta'^-)_{n_6,m})_{m\leq M}$ whose empirical distribution is the further from its limit standard normal distribution. The reference limit cdf F_0 is also plotted (dashed). Right: Empirical cdf F_{n_3} (solid line; it is the same as that plotted in the leftmost graph), empirical cdf F''_{n_3} (dotted line) of the sequence $(Z(\theta''^-)_{n_3,m})_{m\leq M}$ obtained under $\mathbf{g}_n^a$-adaptive sampling scheme whose empirical distribution is the further from its limit standard normal distribution, and their *common* limit cdf F_0 (dashed). In this last graph only, we use a logarithmic scale on the y-axis in order to enhance the differences at the left tail.

Empirical coverage of intervals $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$, that is proportions

$$c(\theta)_{n_i} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\{\Psi(\theta) \in \mathcal{I}(\theta)_{n_i,m}\}, \quad \theta \in \Theta_0, i = 1, \ldots, 7,$$

are reported in Tables 12 and 13 (for $i = 1, 2, 3, 4$ and $i = 5, 6, 7$ respectively) for iid g^b-balanced sampling, in Tables 14 and 15 (for $i = 1, 2, 3, 4$ and $i = 5, 6, 7$ respectively) for iid $g^\star$-optimal sampling, in Tables 16 and 17 (for $i = 1, 2, 3, 4$ and $i = 5, 6, 7$ respectively) for $\mathbf{g}_n^\star$-adaptive sampling, and in Tables 18 and 19 (for $i = 1, 2, 3, 4$ and $i = 5, 6, 7$ respectively) for $\mathbf{g}_n^a$-adaptive sampling.

Because those tables are very dense, we invite the reader to skim through them and rather comment on Figure 5 before testing if the empirical coverage behaves as it should.

In Figure 5, the leftmost boxplot at each sample size (associated to iid g^b-balanced sampling scheme) serves as a benchmark. There is no striking difference between them and the corresponding boxplots associated to iid $g^\star$-optimal sampling. Surprisingly, a rather good coverage is guaranteed at sample sizes $n_1 = 100, n_2 = 250$, *i.e.* even before the central limit theorem is empirically validated (see Section 6.2). In contrast, the boxplots associated to the adaptive designs reveal a very poor empirical coverage at the smallest sample sizes $n_1 = 100$ and $n_2 = 250$. When the sample size is larger than or equal to n_3, the boxplots associated to the adaptive designs illustrate an empirical coverage that compares equally to that of the independent designs. This is in agreement with the empirical validation of the central limit theorem for $\mathbf{g}_n^\star$-adaptive design, but not for $\mathbf{g}_n^a$-adaptive design.

More rigorously now, the independent rescaled empirical coverage proportions $\{Mc(\theta)_{n_i} : \theta \in \Theta_0\}$ should be distributed according to the Binomial distribution with parameter $(M, 1-a)$ with $a = \alpha$ for every $i = 1, \ldots, 7$. This property can be tested in terms of our tailored test (see Section A.3), the alternative stating that $a > \alpha$. This results in a collection of 7 p-values for each sampling scheme, as reported in Table 5.

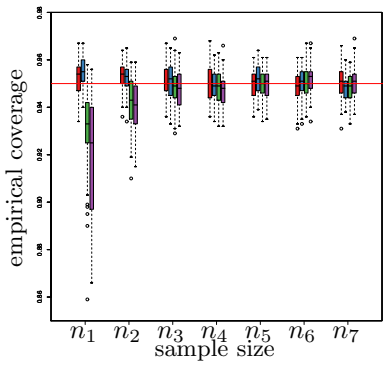

Figure 5: Boxplots representing the empirical coverage proportions $\{c(\theta)_{n_i} : \theta \in \Theta_0\}$ for $i = 1, \ldots, 7$ (each sample size) and each sampling scheme: from left to right at each sample size, iid g^b-balanced, iid $g^\star$-optimal, $\mathbf{g}_n^\star$-adaptive and $\mathbf{g}_n^a-$adaptive sampling schemes. Every box features a solid horizontal line showing the mean value, its bottom and top limits corresponding to the first and third quartiles. Its whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range. An horizontal line indicating the aimed level 95% is added.

	sample size						
sampling scheme	n_1	n_2	n_3	n_4	n_5	n_6	n_7
iid g^b-balanced	0.498	0.966	0.923	0.247	0.369	0.045	0.925
iid $g^\star$-optimal	0.995	0.981	0.769	0.533	0.958	0.586	0.007
$\mathbf{g}_n^\star$-adaptive	$p < 0.001$	$p < 0.001$	0.218	0.160	0.722	0.645	0.179
$\mathbf{g}_n^a$-adaptive	$p < 0.001$	$p < 0.001$	0.028	0.009	0.425	0.898	0.717

Table 5: Checking the adequateness of the coverage guaranteed by our simulated confidence intervals. We test if the independent rescaled empirical coverage Binomial random variables $\{Mc(\theta)_{n_i} : \theta \in \Theta_0\}$ have parameter $(M, 1-\alpha)$, the alternative stating that they have parameter $(M, 1-a)$ with $a > \alpha$: we report p-values for each sample size n_i, $i = 1, \ldots, 7$ and each sampling scheme. The tailored test used here is presented in Section A.3.

Empirical validation of the coverage of the confidence intervals.

Considering each sampling scheme (*i.e.* each row of Table 5) separately, we conclude that

the $(1-\alpha)$-coverage cannot be declared defective under

- iid g^b-balanced sampling for any sample size n_i,
- iid $g^\star$-optimal sampling for any sample size n_i,
- $\mathbf{g}_n^\star$-adaptive sampling for any sample size $n_i \geq n_3 = 500$,
- $\mathbf{g}_n^a$-adaptive sampling for any sample size $n_i \geq n_4 = 750$,

adjusting for multiple testing in terms of the Benjamini and Yekutieli procedure for controlling the False Discovery Rate at level 5%. Note that the coverage validity under iid optimal sampling for the largest sample size $n_7 = 5000$ is barely obtained.

Less formally, the confidence intervals obtained under both iid sampling schemes achieve the desired coverage for any sample size (that is as soon as 100 observations are collected). Satisfactorily, the confidence intervals obtained under $\mathbf{g}_n^\star$-adaptive sampling scheme achieve the desired coverage when the sample size exceeds $n_3 = 500$ (indeed, most numbers in Tables 16 and 17 are very close to 0.95 for $i = 3, 4, 5, 6, 7$). Regarding the $\mathbf{g}_n^a$-adaptive sampling scheme, $n_4 = 750$ accrued data at least are required to guarantee the desired coverage of the confidence intervals. This is much better than the minimal sample size of $n_6 = 2500$ necessary to reach the Gaussian limit in the central limit theorem (see Section 6.2).

In conclusion,

Highlight 4 (empirical coverage of the confidence intervals)**.** *In view of Theorem 2 and its implications in terms of construction of confidence intervals, the confidence intervals that we obtain under adaptive $(\theta, \mathbf{g}_n^\star)$ sampling scheme achieve the desired coverage as soon as 500 observations are accrued. In contrast, the confidence intervals we get under the iid $(\theta, g^\star)$ optimal sampling scheme feature the desired coverage as soon as 100 observations are collected.*

6.4 Empirical widths of the confidence intervals.

Now we know that, for moderate and large sample sizes, the confidence intervals we obtain under both adaptive sampling schemes meet the coverage requirements. In this subsection, we investigate the empirical widths of the confidence intervals. We expect to show that the confidence intervals obtained under adaptive sampling schemes are narrower than those obtained under iid g^b-balanced sampling scheme, and also that they are not significantly wider than the confidence intervals obtained under the iid $g^\star$-optimal sampling scheme.

So we focus here on the empirical widths of intervals $\mathcal{I}(\theta)_{n_i,m}$. A preliminary inspection teaches us that the empirical distributions of the widths $\{|\mathcal{I}(\theta)_{n_i,m}| : m = 1, \ldots, M\}$ are unimodal and roughly symmetric at the mode (this is not a surprise, at least under iid sampling: the squared width $|\mathcal{I}(\theta)_{n_i,m}|^2$ is proportional to $s(\theta)^2_{n_i,m}/n_i$ and $\sqrt{n}(s(\theta,g)^2 - v(\theta,g))$ is asymptotically normal). It is therefore meaningful to report only means and standard deviations. So we introduce the quantities

$$r(\theta)_{n_i,m} = \frac{s(\theta)_{n_i,m}}{\sqrt{v^\star(\theta)}} - 1$$

($s(\theta)_{n_i,m}/\sqrt{v^\star(\theta)}$ is the ratio of the width of $\mathcal{I}(\theta)_{n_i,m}$ over its optimal width), and report the empirical mean and standard deviation of $\{r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ in Tables 20 and 21 (for $i = 1, 2, 3, 4$ and $i = 5, 6, 7$ respectively) for iid g^b-balanced sampling, in Tables 22 and 23 (for $i = 1, 2, 3, 4$ and $i = 5, 6, 7$ respectively) for iid $g^\star$-optimal sampling, in Tables 24 and 25 (for $i = 1, 2, 3, 4$ and $i = 5, 6, 7$ respectively) for $\mathbf{g}_n^\star$-adaptive sampling, and in Tables 26 and 27 (for $i = 1, 2, 3, 4$ and $i = 5, 6, 7$ respectively) for $\mathbf{g}_n^a$-adaptive sampling.

We start with qualitative comments. As expected, empirical means on the diagonal of every table quickly decrease to 0 when the sample size increases (for such θ's, the iid g^b-balanced sampling is optimal). We also remark that for every sampling scheme and $\theta \in \Theta_0$, the corresponding empirical means converge towards $\sqrt{v^b(\theta)/v^\star(\theta)}-1$ (for iid g^b-balanced sampling) or 0 (otherwise) while the corresponding standard deviations decrease as the sample size increases: this is due to the convergence of $s(\theta)^2_{n_i,m}$ towards $v^b(\theta)$ (for iid g^b-balanced sampling) or $v^\star(\theta)$ (otherwise). So this simulation study seems to confirm that it is possible indeed, as theoretically proven in Section 4.2, to get confidence intervals of asymptotic level $(1-\alpha)$ as narrow as the optimal ones that we would obtain, had we known in advance the corresponding optimal treatment mechanism characterized by (5).

Testing the empirical widths of confidence intervals.

Now, the latter qualitative comments are backed by quantitative results that we obtain in a testing framework. On one hand indeed, the widths $\{|\mathcal{I}(\theta)_{n_i,m}| : m = 1,\ldots,M\}$ of the M confidence intervals obtained under iid $g^\star$-optimal sampling provide us with an empirical counterpart of a benchmark distribution of optimal width for sample size n_i. On the other hand the distributions of the widths of the confidence intervals at sample size n_i obtained under both adaptive sampling schemes are the empirical counterparts of two distributions which may be close to the empirical benchmark distribution (at least, the theory teaches us that the empirical distributions under iid $g^\star$-optimal sampling and $\mathbf{g}^\star_n$-adaptive sampling schemes converge, as the sample size increases, to the the same Dirac probability distribution). Similarly, the rescaled widths $\{\sqrt{R(\theta)}|\mathcal{I}(\theta)_{n_i,m}| : m = 1,\ldots,M\}$ of the M confidence intervals obtained under iid g^b-balanced sampling give rise to the empirical counterpart of a distribution which should be close to the empirical benchmark distribution (at least again, the theory teaches us that the empirical distributions under iid optimal and balanced sampling schemes converge, as the sample size increases, to the the same Dirac probability distribution).

Therefore we can test at each sample size and across Θ_0, in terms of our tailored test for comparison of widths (see Section A.4), if for each intermediate sample size the two distributions of widths under both adaptive sampling schemes coincide with the benchmark distribution (null), rather than being stochastically *larger* (alternative hypothesis). This yields 14 p-values all almost equal to one, see Table 6. In other words, no matter the sample size, we cannot conclude that the widths of the confidence intervals obtained under either adaptive sampling scheme are larger than their counterparts obtained under iid $g^\star$-optimal sampling.

Regarding the comparison of the iid g^b-balanced and $g^\star$-optimal sampling schemes, we can test at each sample size and across Θ_0, in terms of our tailored test for comparison of widths (see Section A.4), if for each intermediate sample size the distribution of rescaled widths under iid g^b-balanced sampling scheme coincides with the benchmark distribution (null), rather than being stochastically *smaller* (alternative hypothesis). This yields 7 p-values all smaller than 10^{-6}. In other words, no matter the sample size, we can conclude that the widths of the confidence intervals obtained under iid $g^\star$-optimal sampling scheme are stochastically larger than their rescaled (by the corresponding factor $\sqrt{R(\theta)}$) counterparts obtained under iid g^b-balanced sampling for some $\theta \in \Theta_0$. This is not very surprising: rescaling is meant here to adjust the means, the variances being for instance possibly still different for some $\theta \in \Theta_0$.

However we can slightly adapt the procedure we just presented, rescaling more modestly by a sub-optimal factor. We compare now, in the same terms, the empirical benchmark distributions of optimal width with the empirical distributions of $\{\sqrt{R(\theta)}^\rho|\mathcal{I}(\theta)_{n_i,m}| : m = 1,\ldots,M\}$ under iid g^b-balanced sampling for the arbitrarily chosen $\rho = 0.9$. We obtain the 7 p-values reported in Table 6.

Empirical validation of the widths of confidence intervals.

Considering each sampling scheme (*i.e.* each row of Table 6) separately, we conclude that

	sample size						
sampling scheme	n_1	n_2	n_3	n_4	n_5	n_6	n_7
iid g^b-balanced	$p < 0.001$	0.018	0.025	0.149	0.588	1.000	0.997
$\mathbf{g}_n^\star$-adaptive	1.000	1.000	1.000	1.000	1.000	1.000	0.977
$\mathbf{g}_n^a$-adaptive	1.000	1.000	1.000	0.999	0.995	1.000	1.000

Table 6: Comparing the widths of our confidence intervals. First row: We report p-values derived at each sample n_i, $i = 1, \ldots, 7$ when comparing, across Θ_0, the empirical distributions of rescaled widths (by a factor $\sqrt{R(\theta)}^\rho$ with $\rho = 0.9$) under iid g^b-balanced sampling to the empirical distributions of widths obtained under iid $g^\star$-optimal sampling, in terms of our tailored test for comparison of widths (see Section A.4), the alternative hypothesis stating that the latter are stochastically larger than the former. Second and third rows: We report p-values derived at each sample n_i, $i = 1, \ldots, 7$ when comparing, across Θ_0, the empirical distributions of widths obtained under $\mathbf{g}_n^\star$-adaptive sampling (second row), or under $\mathbf{g}_n^a$-adaptive sampling (third row), to the empirical distributions of widths obtained under iid $g^\star$-optimal sampling, in terms of our tailored test for comparison of widths (see Section A.4), the alternative hypothesis stating in both cases that the latter are stochastically smaller than the former distributions.

- the confidence intervals produced under iid g^b-balanced sampling and rescaled by the corresponding factor of the form $\sqrt{R(\theta)}^\rho$ ($\rho = 0.9$) are not stochastically narrower than those produced under iid $g^\star$-optimal sampling for any sample size $n_i \geq n_2 = 250$,
- the confidence intervals produced under $\mathbf{g}_n^\star$-adaptive sampling are not stochastically wider than those produced under iid $g^\star$-optimal sampling for any sample size n_i,
- the confidence intervals produced under $\mathbf{g}_n^a$-adaptive sampling are not stochastically wider than those produced under iid $g^\star$-optimal sampling for any sample size n_i,

adjusting for multiple testing in terms of the Benjamini and Yekutieli procedure for controlling the False Discovery Rate at level 5%.

In conclusion,

Highlight 5 (empirical widths of confidence intervals)**.** *In view of Theorem 2 and for any sample size, the widths of the confidence intervals obtained under adaptive $(\theta, \mathbf{g}_n^\star)$ sampling scheme are not significantly greater than the widths of the confidence intervals that we obtain under iid $(\theta, g^\star)$ optimal sampling scheme.*

6.5 Illustrating example.

So far, we have been concerned with results averaged across randomly sampled trajectories and θ's ranging over Θ_0. Here we present as an illustrating example four trajectories produced by the iid (θ, g^b) and $(\theta, g^\star)$ sampling schemes and the adaptive $(\theta, \mathbf{g}_n^\star)$ and $(\theta, \mathbf{g}_n^a)$ sampling schemes for $\theta = (0.2, 0.6) \in \Theta_0$.

For each of them, we report the point estimates $\Psi_{n_i}(\mathbf{O}_{n_7}^1(n_i))$ of $\Psi(\theta) = 1.099$ at every sample size n_i, as well as the estimated standard deviations $s(\theta)_{n_i,1}^2$, confidence intervals $\mathcal{I}(\theta)_{n_i,1}$, and estimates $g_{n_i}^\star(1)$ and $g_{n_i}^a(1)$ of the optimal proportion of treated $g^\star(\theta)(1) = 0.290$ for the two adaptive procedures — see Table 7.

In addition, we exhibit in Figure 6 several plots illustrating (from left to right) how the sequences θ_n, Ψ_n, $\mathbf{g}_n$, and $s(\theta)_n^2$ evolve as the sample size increases when applying the two adaptive sampling schemes. The most striking feature in the figure, which is representative of all the trajectories we have observed, concerns the adaptive treatment mechanism sequence. Estimating (or targeting) the optimal treatment mechanism is the driving force of our new adaptive estimation procedure. It is proven in Theorem 1 that $g_n^\star(1)$ and, therefore, the cumulated mean $\frac{1}{n}\sum_{i=1}^n g_i^\star(1)$, converge to the optimal proportion of treated $g^\star(\theta)(1)$ when the sampling scheme is characterized

by (14). We see here that it is the cumulated mean of $g_n^a(1)$ only that converges to $g^\star(\theta)(1)$ when considering the $(\theta, \mathbf{g}_n^a)$ sampling scheme.

sampling scheme	sample size n_1	n_2	n_3	n_4	n_5	n_6	n_7	
(θ, g^b)	1.053	1.285	1.259	1.167	1.118	1.158	1.117	$\Psi_{n_i}(\mathbf{O}^1_{n_7}(n_i))$
	2.611	2.999	2.993	2.964	2.954	3.164	3.046	$s(\theta)_{n_i,1}$
	$[0.541, 1.565]$	$[0.913, 1.656]$	$[0.997, 1.522]$	$[0.955, 1.379]$	$[0.935, 1.301]$	$[1.034, 1.282]$	$[1.033, 1.202]$	$\mathcal{I}(\theta)_{n_i,1}$
	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$g^b(1)$
$(\theta, g^\star)$	0.783	1.131	1.032	1.068	1.097	1.080	1.101	$\Psi_{n_i}(\mathbf{O}^1_{n_7}(n_i))$
	2.873	2.836	2.798	2.823	2.863	2.831	2.809	$s(\theta)_{n_i,1}$
	$[0.220, 1.346]$	$[0.779, 1.482]$	$[0.787, 1.278]$	$[0.865, 1.270]$	$[0.919, 1.274]$	$[0.969, 1.191]$	$[1.023, 1.179]$	$\mathcal{I}(\theta)_{n_i,1}$
	0.290	0.290	0.290	0.290	0.290	0.290	0.290	$g^\star(\theta)(1)$
$(\theta, \mathbf{g}^\star_n)$	1.329	1.212	1.272	1.219	1.238	1.090	1.111	$\Psi_{n_i}(\mathbf{O}^1_{n_7}(n_i))$
	3.560	2.984	2.928	2.939	2.936	2.804	2.829	$s(\theta)_{n_i,1}$
	$[0.631, 2.027]$	$[0.842, 1.582]$	$[1.015, 1.528]$	$[1.008, 1.429]$	$[1.056, 1.420]$	$[0.980, 1.200]$	$[1.033, 1.190]$	$\mathcal{I}(\theta)_{n_i,1}$
	0.281	0.277	0.262	0.273	0.269	0.291	0.288	g_{n_i}
$(\theta, \mathbf{g}^a_n)$	1.216	1.178	1.247	1.369	1.278	1.122	1.136	$\Psi_{n_i}(\mathbf{O}^1_{n_7}(n_i))$
	2.544	2.709	2.770	3.094	2.990	2.853	2.842	$s(\theta)_{n_i,1}$
	$[0.718, 1.715]$	$[0.842, 1.514]$	$[1.004, 1.490]$	$[1.147, 1.590]$	$[1.092, 1.463]$	$[1.010, 1.233]$	$[1.058, 1.215]$	$\mathcal{I}(\theta)_{n_i,1}$
	0.317	0.060	0.367	0.213	0.630	0.750	0.309	g_{n_i}

Table 7: Illustrating example. We report here the point estimates $\Psi_{n_i}(\mathbf{O}^1_{n_7}(n_i))$ of $\Psi(\theta) = 1.099$ ($\theta = (0.2, 0.6)$) at every sample size n_i, the estimated standard deviations $s(\theta)_{n_i,1}$, confidence intervals $\mathcal{I}(\theta)_{n_i,1}$, and estimates $g^\star_{n_i}(1)$ and $g^a_{n_i}(1)$ of the optimal proportion of treated $g^\star(\theta)(1) = 0.290$ for the two iid and two adaptive $(\theta, \mathbf{g}^\star_n)$ and $(\theta, \mathbf{g}^a_n)$ sampling schemes. See also Figure 6.

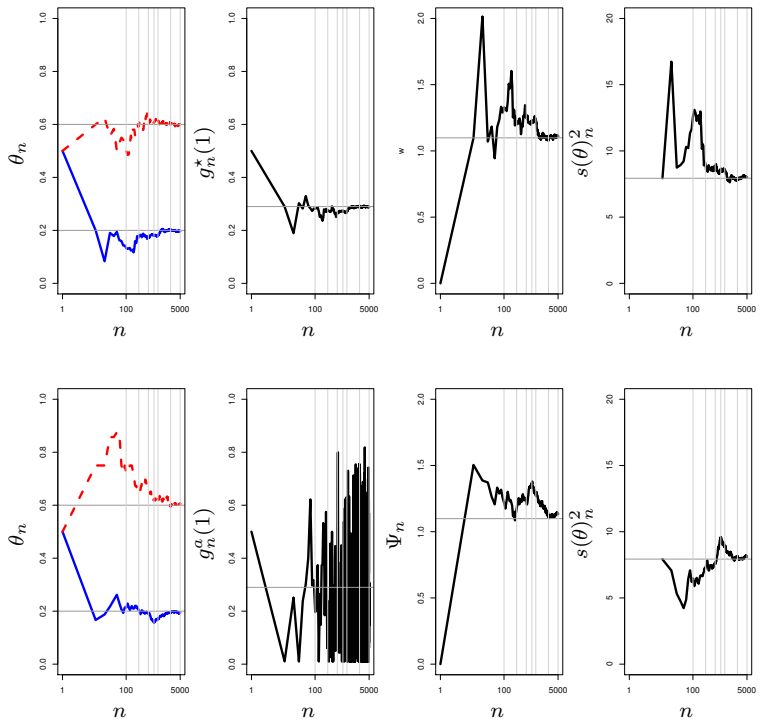

Figure 6: Illustrating how the two adaptive $(\theta, \mathbf{g}_n^\star)$, top, and $(\theta, \mathbf{g}_n^a)$, bottom, sampling schemes behave as the sample size increases (on x-axis, logarithmic scale; the vertical grey lines indicate sample sizes n_i, $i = 1, \ldots, 7$). From left to right we represent at the same scale over columns the sequences θ_n, $\Psi_n = \Psi(\theta_n)$, $g_n^\star(1)$ or $g_n^a(1)$ and $s(\theta)_n^2$. Horizontal grey lines indicate the theoretical limits of the plotted sequences. By convention, θ_n is initiated at $(\frac{1}{2}, \frac{1}{2})$ while $g_n^\star(1)$ and $g_n^a(1)$ are initiated at $\frac{1}{2}$. Another convention requires that at least 10 observations are collected before computing $s(\theta)_n^2$ for the first time, explaining why the corresponding plots start at $n = 10$ rather than $n = 1$. The most striking feature is how smoothly $g_n^\star(1)$ converges to $g^\star(\theta)(1)$ when $\mathbf{g}_n = \mathbf{g}_n^\star$, top, as opposed to when $\mathbf{g}_n = \mathbf{g}_n^a$, bottom.

7 Simulation study of the performances of targeted optimal group sequential testing procedure powered at local alternatives

In this section, we carry out a simulation study of the performances of targeted optimal design adaptive procedures in terms of group sequential testing. The three main questions at stake are "Does the group sequential testing procedure under the targeted optimal design adaptive sampling scheme guarantee the desired type I error?", then "Does it guarantee the desired power?", then lastly "How well does it compare with the group sequential testing procedure under the targeted optimal iid sampling scheme?"

We carefully present the simulation scheme in Section 7.1. The section culminates in Section 7.2 with the investigation of the properties of the adaptive group sequential testing procedure in terms of type I and type II errors. We conclude in Section 7.3 with the simulation study of its performances in terms of sample sizes at decision.

7.1 The simulation scheme (continued).

For every $\theta = (\theta_0, \theta_1) \in \Theta = \Theta_0 \setminus \{(.1,.7),(.1,.8),(.1,.9),(.2,.8),(.2,.9),(.3,.9)\}$[1], we test $M = 1000$ times the null "$\psi = \Psi(\theta)$" against the alternative "$\psi > \Psi(\theta)$" with asymptotic type I error $\alpha = 5\%$ and type II error $\beta = 10\%$ at $\psi = \Psi(\theta) + \Delta(\theta)$, where $\Delta(\theta) = \Psi(\theta + (0,\eta)) - \Psi(\theta) = \log(1 + \eta/\theta_1)$ (with $\eta = 0.05$) is a small increment. Depending on whether we want to investigate the empirical behaviors of the different testing procedures with respect to type I (i) or type II (ii) errors, we resort for $\theta \in \Theta$ to

(i) Empirical type I error study:

- iid (θ, g^b) balanced sampling,
- iid $(\theta, g^\star)$ optimal sampling,
- $(\theta, \mathbf{g}_n^\star)$ adaptive sampling,
- $(\theta, \mathbf{g}_n^a)$ adaptive sampling,

(ii) Empirical type II error study:

- iid $(\theta + (0,\eta), g^b)$ balanced sampling,
- iid $(\theta + (0,\eta), g^\star)$ optimal sampling,
- $(\theta + (0,\eta), \mathbf{g}_n^\star)$ adaptive sampling,
- $(\theta + (0,\eta), \mathbf{g}_n^a)$ adaptive sampling.

We apply a one-sided group sequential testing procedure (as described in Section 5.1) based on proportions $(p_1, p_2, p_3, p_4) = (0.25, 0.50, 0.75, 1)$ and α- and β-spending functions both equal to $t \mapsto t^2$. The designs (values of $\Delta(\theta)$, maximum committed information $I_{\max}$, rejection and futility regions bounds) are reported in Table 8.

θ_1	$\log(1+\eta/\theta_1)$	$I_{\max}(\theta_1)$	rejection/futility boundaries
.1	0.405	56.561	(2.734, 2.302, 2.006, 1.716)
			(-0.973, 0.136, 0.965, 1.716)
.2	0.223	186.747	(2.734, 2.301, 2.013, 1.720)
			(-0.973, 0.138, 0.963, 1.720)
.3	0.154	391.32	(2.734, 2.307, 2.008, 1.716)
			(-0.973, 0.135, 0.962, 1.716)
.4	0.118	670.281	(2.734, 2.303, 2.010, 1.718)
			(-0.973, 0.145, 0.962, 1.718)
.5	0.095	1023.632	(2.734, 2.298, 2.008, 1.715)
			(-0.973, 0.140, 0.961, 1.715)
.6	0.080	1451.372	(2.734, 2.305, 2.006, 1.716)
			(-0.973, 0.135, 0.962, 1.716)
.7	0.069	1947.12	(2.734, 2.300, 2.006, 1.715)
			(-0.976, 0.133, 0.957, 1.715)
.8	0.061	2530.022	(2.734, 2.306, 2.006, 1.715)
			(-0.973, 0.139, 0.961, 1.715)
.9	0.054	3180.931	(2.734, 2.299, 2.006, 1.715)
			(-0.973, 0.138, 0.963, 1.715)

Table 8: Description of one-sided sequential testing designs with proportions $(p_1, p_2, p_3, p_4) = (0.25, 0.50, 0.75, 1)$, α- and β-spending functions both equal to $t \mapsto t^2$, and asymptotic type I error $\alpha = 5\%$ and type II error $\beta = 10\%$. Every $\theta = (\theta_0, \theta_1) \in \Theta$ is associated with the single entry corresponding with θ_1. For each entry, we provide (with precision 10^{-3}) the value of the increment $\Delta(\theta) = \log(1 + \eta/\theta_1)$ that yields the parameter $\psi_1 = \Psi(\theta) + \Delta(\theta)$ at which the test of "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" is powered, the maximum committed information $I_{\max}(\theta_1)$, the rejection region bounds (above) and the futility region bounds (below); note that, of course, the final rejection and futility bounds coincide.

[1] Those six values are left aside because it would be computationally demanding to consider them too: for $\theta = (\theta_0, \theta_1) \in \Theta_0 \setminus \Theta$, the key quantities $v^\star(\theta) I_{\max}(\theta_1)$ and $v^b(\theta) I_{\max}(\theta_1)$ (which can be interpreted as average maximal sample sizes at decision) are very large.

7.2 Empirical type I and type II errors.

Let us consider here the empirical type I and type II errors. We wish to answer the questions "Does the group sequential testing procedure under the targeted optimal design adaptive sampling scheme guarantee the desired type I error?" and "Does it guarantee the desired power?"

Testing the empirical type I and type II errors.

Empirical type I errors, that is proportions $\{a(\theta) : \theta \in \Theta\}\}$ such that $Ma(\theta)$ is the number of times the null was falsely rejected for its alternative by the testing procedure of "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" powered at $\psi = \Psi(\theta + (0, \eta))$, are reported in Table 28.

Empirical type II errors, that is proportions $\{b(\theta) : \theta \in \Theta\}$ such that $Mb(\theta)$ is the number of times the null was falsely *not* rejected for its alternative by the testing procedure "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" powered at $\psi = \Psi(\theta + (0, \eta))$, are reported in Table 29.

In both tables, the numbers are strikingly close to the wished values (0.05 for Table 28, and 0.9 for Table 29).

Here again we rely on testing to assess rigorously if the requirements on type I and II errors are met. To this end, we use that the independent rescaled empirical proportions $\{Ma(\theta) : \theta \in \Theta\}$ should be distributed according to the Binomial distribution with parameter (M, a) with $a = \alpha$. This property can be tested in terms of our tailored test, the alternative stating that $a > \alpha$ (see Section A.3). This results in 4 p-values, as reported in Table 9. Similarly, the independent rescaled empirical proportions $\{Mb(\theta) : \theta \in \Theta\}$ should be distributed according to the Binomial distribution with parameter $(M, b) = (M, 10\%)$. This property can also be tested in terms of our tailored test, the alternative stating that $b > \beta = 10\%$ (see Section A.3). This results in 4 p-values, and we also report them in Table 9. The latter p-values teach us that the study is under-powered. It remains to assert whether the study is slightly or strongly under-powered: to this end we now test the null stating that the independent rescaled empirical proportions $\{Mb(\theta) : \theta \in \Theta\}$ are distributed according to the Binomial distribution with parameter $(M, b) = (M, 11\%)$ against the alternative stating that $b > 11\%$. The corresponding 4 p-values are also reported in Table 9.

sampling scheme	type I error	type II error (10%)	type II error (11%)
iid g^b-balanced	0.974	$p < 0.001$	0.107
iid $g^\star$-optimal	1.000	0.293	0.180
$\mathbf{g}_n^\star$-adaptive	0.552	$p < 0.001$	0.185
$\mathbf{g}_n^a$-adaptive	0.511	$p < 0.001$	0.185

Table 9: Checking the adequateness of the type I errors and powers of our simulated targeted optimal group sequential testing procedures. We test if the rescaled empirical type I errors Binomial random variables $\{Ma(\theta) : \theta \in \Theta\}$ have parameter (M, α), the alternative stating that they have parameter (M, a) with $a > \alpha$ and report (in the second column) the obtained p-values for each sampling scheme. We also test if the rescaled empirical type II errors Binomial random variables $\{Mb(\theta) : \theta \in \Theta\}$ have parameter $(M, \beta) = (M, 10\%)$ (third column) or $(M, 11\%)$, the alternative stating that they have parameter (M, b) with $b > 10\%$ and $b > 11\%$ respectively, and report the obtained p-values for each sampling scheme. The tailored test used here is presented in Section A.3.

Empirical validation of type I and type II errors.

Considering each sampling scheme (*i.e.* each row of Table 9) separately, we conclude that

- the type I error control cannot be declared defective for any sampling procedure or, in less formal terms, that the type I error control is guaranteed for both iid g^b-balanced and $g^\star$-optimal sampling schemes as well as for both $\mathbf{g}_n^\star$-adaptive and $\mathbf{g}_n^a$-adaptive sampling schemes;
- the group sequential testing procedures are all slightly under-powered, in the sense that:

– the type II error control is declared defective for all sampling schemes (for each of them, there exists at least one $\theta \in \Theta$ for which the type II error is likely larger than $\beta = 10\%$);
– however, the type II error control cannot be declared defective for any sampling procedure when substituting $\beta' = 11\%$ to $\beta = 10\%$ or, in less formal terms, a 11% (rather than 10%) control of the type II error is guaranteed for both iid g^b-balanced and $g^\star$-optimal sampling schemes as well as for both $\mathbf{g}_n^\star$-adaptive and $\mathbf{g}_n^a$-adaptive sampling schemes.

This summary notably confirms the conjecture that the Theorem 3 we prove in Section 5.2 for group sequential testing procedures at deterministic sample sizes still holds for "real-life" group sequential testing procedures at random sample sizes, as described in Section 5.1.

In conclusion,

Highlight 6 (empirical type I and type II errors)**.** *In view of Section 5.1 and Theorem 3, the* $(\theta, \mathbf{g}_n^\star)$ *adaptive group sequential testing procedure achieves the desired type I error when testing against local alternatives. It is slightly underpowered in the sense that the type II error control is guaranteed at 89% instead of 90% – but the same holds for the iid* $(\theta, g^\star)$ *optimal group sequential testing procedure.*

7.3 Empirical distributions of sample size at decision.

Let us now consider the empirical distributions of sample size at decision, and answer the question "How well does it compare with the group sequential testing procedure under the targeted optimal iid sampling scheme?"

Testing the empirical sample size at decision.

We report in Table 30 the mean sample sizes at decision for each $\theta \in \Theta$ when checking the adequateness of type I error control of our group sequential testing procedures of "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" powered at $\psi = \Psi(\theta + (0, \eta))$. We also report in Table 31 the mean sample sizes at decision for each $\theta \in \Theta$ when checking the adequateness of type II error control of our group sequential testing procedures of "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" powered at $\psi = \Psi(\theta + (0, \eta))$. Inspecting Tables 30 and 31 tells us, at least in terms of mean sample sizes at decision and regarding either empirical type I or type II errors, first that the two adaptive group sequential testing procedures perform as well as the iid $g^\star$-optimal group sequential testing procedure, and second that the three latter procedures perform (sometimes, much) better than the iid g^b-balanced group sequential testing procedure when balanced and optimal iid procedures differ. As a summary, we provide in Table 10 a comparison of mean sample sizes at decision when resorting to iid g^b-balanced group sequential testing procedure with respect to $\mathbf{g}_n^\star$-adaptive group sequential testing procedure. Naturally, the further the percentage is away from the diagonal, the larger is the gain. Sometimes, the gain is dramatic.

Again, we push further the comparison between empirical distributions of sample size at decision under each group sequential testing procedure (in the same spirit as the comparison of widths in Section 6.4). On the one hand, the sample sizes at decision $\{S(\theta, g^\star)_m : m = 1, \ldots, M\}$ of the M independent copies of the iid $g^\star$-optimal group sequential testing procedure provides us with an empirical counterpart of a benchmark distribution of optimal sample size at decision. On the other hand, we also have at hand the empirical distributions of sample sizes at decision obtained under iid g^b-balanced and both $\mathbf{g}_n^\star$-adaptive and $\mathbf{g}_n^a$-adaptive group sequential testing procedures which we see as empirical counterparts of distributions that we would like to compare to the aforementioned benchmark distribution.

Regarding the comparison of the iid group sequential testing procedures, we propose to test across Θ, in terms of our tailored test for comparison of sample sizes at decision (see Section A.4), if the distribution of sample size at decision under iid g^b-balanced group sequential testing procedure rescaled by a factor $R(\theta)$ coincides with the benchmark distribution (null), rather than being stochastically smaller (alternative hypothesis). This yields a p-value smaller than 10^{-6}. In other

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1%	6%	10%	14%	26%	31%	-	-	-
.2	-	2%	1%	5%	12%	18%	24%	-	-
.3	-	-	2%	0%	6%	9%	16%	24%	-
.4	-	-	-	1%	2%	2%	11%	17%	33%
.5	-	-	-	-	1%	-1%	6%	11%	26%
.6	-	-	-	-	-	2%	1%	3%	13%
.7	-	-	-	-	-	-	-2%	3%	12%
.8	-	-	-	-	-	-	-	3%	4%
.9	-	-	-	-	-	-	-	-	6%

gains when evaluating empirical type I error

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	2%	6%	19%	21%	32%	36%	-	-	-
.2	-	0%	4%	8%	14%	21%	30%	-	-
.3	-	-	-4%	2%	9%	13%	22%	33%	-
.4	-	-	-	0%	2%	5%	13%	27%	53%
.5	-	-	-	-	0	2%	10%	19%	41%
.6	-	-	-	-	-	1%	3%	11%	29%
.7	-	-	-	-	-	-	3%	4%	30%
.8	-	-	-	-	-	-	-	3%	21%
.9	-	-	-	-	-	-	-	-	10%

gains when evaluating empirical type II error

Table 10: Comparing mean sample sizes at decision when resorting to iid g^b-balanced group sequential testing procedure with respect to $\mathbf{g}_n^\star$-adaptive group sequential testing procedure. The top table corresponds to Table 30 and evaluation of empirical type I error, while the bottom table to Table 31 and evaluation of empirical type II error. Entries are of the form $(\bar{S}(\theta, g^b) - \bar{S}(\theta, \mathbf{g}_n^\star))/\bar{S}(\theta, \mathbf{g}_n^\star)$, where $\bar{S}(\theta, g^b)$ (respectively, $\bar{S}(\theta, g^b)$) denotes the empirical mean sample size at decision under iid g^b-balanced sampling (respectively, $\mathbf{g}_n^\star$-adaptive sampling).

words, we can conclude that there exists some $\theta \in \Theta$ for which the sample size at decision under iid $g^\star$-optimal group sequential testing is stochastically larger than the corresponding $R(\theta)$-rescaled sample size at decision under iid g^b-balanced group sequential testing. However, we can slightly adapt the procedure we just presented, rescaling more modestly by a sub-optimal factor. We compare now, in the same terms and with the same benchmark distribution, the distribution of sample size at decision under iid g^b-balanced group sequential testing procedure rescaled by a factor $R(\theta)^\rho$ for the arbitrarily chosen $\rho = 0.45$. The two p-values thus obtained are reported in Table 11. Regarding the comparison of the $\mathbf{g}_n^\star$-adaptive and $\mathbf{g}_n^a$-adaptive group sequential testing procedures to the iid $g^\star$-optimal group sequential testing procedure, we propose to test across Θ, in terms of our tailored test for comparison of sample sizes at decision (see Section A.4), if the distributions of sample size at decision under either adaptive group sequential testing procedures coincides with the benchmark distribution (null), rather than being stochastically larger (alternative hypothesis). This yields 4 p-values (two when investigating the behaviors with respect to type I error, two with respect to type II error) that we report in Table 11.

Empirical validation of sample sizes at decision.

Considering each sample scheme (*i.e.* each row of the table) separately, we conclude that

- the sample sizes at decision obtained under iid (θ, g^b) balanced group sequential testing and rescaled by the corresponding factor $R(\theta)^\rho$ ($\rho = 0.45$) are not stochastically smaller than the sample sizes at decision obtained under iid $(\theta, g^\star)$ optimal group sequential testing;
- the sample sizes at decision obtained under iid $(\theta + (0, \eta), g^b)$ balanced group sequential testing procedure and rescaled by the corresponding factor $R(\theta + (0, \eta))^\rho$ ($\rho = 0.45$) are not stochastically smaller than the sample sizes at decision obtained under iid $(\theta + (0, \eta), g^\star)$ optimal group sequential testing procedure;
- the sample sizes at decision obtained under both $(\theta, \mathbf{g}_n^\star)$ and $(\theta, \mathbf{g}_n^a)$ adaptive group sequential

sampling scheme	type I error	type II error
iid g^b-balanced	0.625	0.727
$\mathbf{g}_n^\star$-adaptive	0.994	1.000
$\mathbf{g}_n^a$-adaptive	0.898	0.949

Table 11: Comparing across Θ the empirical distributions of sample sizes at decision. First row: We report p-values derived when comparing, across Θ, the empirical distributions of rescaled sample sizes at decision (by a factor $R(\theta)^\rho$ with $\rho = 0.45$) under iid g^b-balanced sampling to the empirical counterpart of the benchmark distributions of sample sizes at decision obtained under iid $g^\star$-optimal sampling, in terms of our tailored test for comparison of sample sizes at decision (see Section A.4), the alternative hypothesis stating that the latter are stochastically larger than the former. Second and third rows: We report p-values derived when comparing, across Θ, the empirical distributions of sample sizes at decision obtained under $\mathbf{g}_n^\star$-adaptive sampling (second row), or under $\mathbf{g}_n^a$-adaptive sampling (third row), to the empirical counterpart of the benchmark distributions of sample sizes at decision obtained under iid $g^\star$-optimal sampling, in terms of our tailored test for comparison of sample sizes at decision (see Section A.4), the alternative hypothesis stating in both case that the latter are stochastically smaller than the former distributions. The second column corresponds to data gathered when investigating the behaviors of the group sequential testing procedures in terms of type I error, the third column corresponding to the same investigation but in terms of type II error.

testing procedures are not stochastically larger than the sample sizes at decision obtained under iid $(\theta, g^\star)$ optimal group sequential testing procedure;

- the sample sizes at decision obtained under both $(\theta + (0, \eta), \mathbf{g}_n^\star)$ and $(\theta + (0, \eta), \mathbf{g}_n^a)$ adaptive group sequential testing procedures are not stochastically larger than the sample sizes at decision obtained under iid $(\theta + (0, \eta), g^\star)$ optimal group sequential testing procedure.

Overall, the main message stated in less formal terms is that both adaptive group sequential testing procedures perform as well as the optimal iid group sequential testing procedure with respect to sample size at decision, either under the null or under the alternative.

Highlight 7 (empirical sample sizes at decision)**.** *In view of Section 5.1 and Theorem 3, the* $(\theta, \mathbf{g}_n^\star)$ *adaptive group sequential testing procedure behaves as the iid* $(\theta, g^\star)$ *optimal group sequential testing procedure in terms of sample sizes at decision, both under the null and under local alternatives.*

8 Discussion

We have studied in this article the properties of a new adaptive group sequential design methodology for randomized clinical trials with binary treatment, binary outcome and no covariate (the experimental unit writes as $O = (A, Y) \in \{0, 1\}^2$, A being the assigned treatment and Y the corresponding outcome).

Prior to accruing data, the trial protocol must specify Ψ, the parameter of interest. Regarding the estimation of Ψ, the trial protocol must specify the confidence level to be used in constructing the confidence interval. Regarding the testing of Ψ, the trial protocol must specify the null and alternative hypotheses, the wished type I error, the alternative parameter at which the test is to be powered and the related wished type II error. If the investigator wants to resort to a group sequential testing procedure, then the trial protocol must also specify the number of intermediate tests, the related proportions, the α- and β-spending strategies (then the maximum committed information, rejection and futility boundaries are fully determined). Finally, the trial protocol must specify the (fixed) targeted design. We decided to focus in this article on the log-relative risk $\Psi = \log E(Y|A = 1) - \log E(Y|A = 0)$ and on that design $g^\star$ which minimizes the asymptotic variance of the maximum likelihood estimator of Ψ. Other choices can be treated likewise.

The methodology is adaptive in the sense that the estimator of $g^\star$, which appears to be strongly consistent (see Highlight 1), is alternatively used in the process of accruing new data, then updated and so on. The resulting maximum likelihood estimator of Ψ, Ψ_n, is strongly consistent (see Highlight 1). It satisfies a central limit theorem, and performs as well (in terms of asymptotic variance) as its counterpart under iid sampling using $g^\star$ itself (see Highlight 1). Therefore, one easily constructs confidence intervals which are as narrow as the intervals one would get, has one known in advance $g^\star$ and used it to sample independently data (see Highlight 1). Those theoretical results are validated with simulations. Notably, a test across a large collection of data-generating distributions indexed by Θ shows that the limiting Gaussian law is empirically reached by the sequence of laws of $\sqrt{n}(\Psi_n - \Psi)$ as soon as 500 observations are collected. This is as good as what one would get, has one known in advance $g^\star$ and used it to sample independently data (see Highlight 3). Most importantly, another test across Θ reveals that the wished coverage is achieved as soon as 500 observations are collected. In contrast, a sample size of 100 observations would suffice, has one known in advance $g^\star$ and used it to sample independently data (see Highlight 4). This is the price to pay for adapting. In conclusion, yet another test across Θ shows that, whenever the sample size exceeds 100, the widths of confidence intervals obtained under adaptive sampling schemes are not significantly greater than the widths of the intervals one would get, has one known in advance $g^\star$ and used it to sample independently data (see Highlight 5).

Furthermore, we explain how a group sequential testing procedure can be equally well applied on top of the adaptive sampling methodology (see Highlight 2). An accompanying theoretical result validates the adaptive group sequential testing procedure in the context of contiguous null and alternative hypotheses. It is supported by simulations. Most importantly, a test across a large collection of pairs of null and local alternative hypotheses indexed by Θ' demonstrates that the adaptive group sequential testing procedure achieves the desired type I error (see Highlight 6). Moreover, a complementary test across Θ' reveals that the adaptive group sequential testing procedure is very slightly under-powered. Interestingly, has one known in advance $g^\star$ and used it to sample independently data, the resulting group sequential testing procedure would suffer from the same minor lack of power (see Highlight 6). Finally, a last test across Θ' shows that the laws of sample sizes at decision under adaptive group sequential testing procedure do not significantly differ from the laws of sample sizes at decision that one would get, has one known in advance $g^\star$ and used it to sample independently data and apply the iid group sequential testing procedure (see Highlight 7).

As stated in the abstract, a three-sentence take-home message is "Adaptive designs do learn the targeted optimal design and inference and testing can be carried out under adaptive sampling as they would under the targeted optimal randomization probability iid sampling. In particular, adaptive designs achieve the same efficiency as the fixed oracle design. This is confirmed by a simulation study, at least for moderate or large sample sizes, across a large collection of targeted randomization probabilities." In essence, everything works as predicted by theory. However, theory also warns us that gains cannot be dramatic in the particular setting of clinical trials with binary treatment, binary outcome and no covariate. Nonetheless, this article is important: it provides a theoretical template and tools for asymptotic analysis of robust adaptive designs in less constrained settings, which we will consider in future work. This notably includes the setting of clinical trials *with covariate*, binary treatment, and *discrete or continuous outcome*, or the setting of clinical trials *with covariate*, binary treatment, and *possibly censored time-to-event* among others. Resorting to targeted maximum likelihood estimation [27] along with adaptation of the design provides substantial gains in efficiency.

References

[1] P. Armitage. *Sequential medical trials*. Wiley, New-York, 1975.

[2] A. Banerjee and A. A. Tsiatis. Adaptive two-stage designs in phase II clinical trials. *Statistics in Medicine*, 25:3382–3395, 2006.

[3] D. A. Berry. Bayesian clinical trials. *Nature Reviews*, 5, 2006. Drug discovery.

[4] D. A. Berry and D. K. Stangl. *Bayesian biostatistics*. Marcel Dekker, New-York, 1996.

[5] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, 1999.

[6] H. Chernoff and S. N. Roy. A Bayes sequential sampling inspection plan. *Annals of Mathematical Statistics*, 36:1387–1407, 1965.

[7] J. R. Eisele. The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference*, 38:249–261, 1994.

[8] S. S. Emerson. Issues in the use of adaptive clinical trial designs. *Statistics in Medicine*, 25:3270–3296, 2006.

[9] B. J. Flehinger and T. A. Louis. Sequential treatment allocation in clinical trials. *Biometrika*, 58:419—426, 1971.

[10] H. L. Golub. The need for mor efficient trial designs. *Statistics in Medicine*, 25:3231–3235, 2006.

[11] F. Hu and W. F. Rosenberg. Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98:671–678, 2003.

[12] F. Hu and L-X. Zhang. Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *Annals of Statistics*, 32:268–301, 2004.

[13] F. H. Hu and W. F. Rosenberg. *The theory of response-adaptive randomization in clinical trials*. Wiley, 2006.

[14] H-M. J. Hung. Discussion. *Statistics in Medicine*, 25:3313–3314, 2006.

[15] A. Ivanova. A play-the-winner type urn design with reduced variability. *Metrika*, 58:1–13, 2003.

[16] D. J., K. R. Abrams, and J. P. Myles. *Bayesian approaches to clinical trials and health care evaluation*. Wiley, Chichester, 2004.

[17] C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton, FL, 2000.

[18] C. Jennison and B. W. Turnbull. Mi-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22:971–993, 2003.

[19] Y. Lokhnygina and A. A. Tsiatis. Optimal two-stage group-sequential designs. *J. Statist. Plann. Inference*, 138(2):489–499, 2008.

[20] C. R. Mehta and N. R. Patel. Adaptive, group sequential and decision theoretic approaches to sample size determination. *Statistics in Medicine*, 25:3250–3269, 2006.

[21] M. A. Proschan, G. K. K. Lan, and J. T. Wittes. *Statistical Monitoring of Clinical Trials: A Unified Approach*. Statistics for biology and health. Springer, New-York, 2006.

[22] W. F. Rosenberg. Randomized urn models and sequential design. *Sequential Analysis*, 21:1–41, 2002. (with discussion).

[23] W. F. Rosenberg, N. Stallard, A. Ivanova, C. N. Harper, and M. L. Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57:909–913, 2001.

[24] P. K. Sen and J. M. Singer. *Large sample methods in statistics*. Chapman & Hall, New York, 1993. An introduction with applications.

[25] A. A. Tsiatis and C. R. Mehta. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90:367–368, 2003.

[26] M. J. van der Laan. The construction and analysis of adaptive group sequential designs. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2008. Paper 232.

[27] M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *Int. J. Biostat.*, 2:Art. 11, 40pp, 2006.

[28] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[29] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[30] R. van Handel. On the minimal penalty for markov order estimation, 2009.

[31] L. J. Wei and S. D. Durham. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73:840–843, 1978.

A Appendix

Normalized martingale sums of the form $M_n(f) = \frac{1}{n}\sum_{i=1}^n [f(O_i, Z_i) - P_{\theta,g_i} f]$ with $Z_i = Z_i(\mathbf{O}_n(i-1))$ play a central role in this study. The Kolmogorov strong law of large numbers (see *e.g.* Theorem 2.4.2 in [24]) guarantees that $M_n(f)$ converges in probability to 0 almost surely for any uniformly bounded function f. However, in order to get consistency results, we need a uniform convergence result for $\sup_{f\in\mathcal{F}} |M_n(f)|$ for a certain class $\mathcal{F}$. This issue is addressed in A.1. Similarly, in order to get a central limit theorem, we need the convergence of $\sqrt{n}M_n(f)$ to a Gaussian random variable. We derive this result in Section A.2 from a standard central limit theorem for discrete martingales (see *e.g.* Theorem 3.3.7 in [24]).

Sections A.3 and A.4 are dedicated to the description of the tailored tests used throughout the simulation study of Sections 6 and 7. The latter tests provide single p-values for multiple pairwise comparisons in the context of our simulations, notably dealing with the multiplicity of elementary tests carried out.

Finally additional tables are gathered in Section A.5.

A.1 Building block for consistency results.

Let $\mathbf{O}_n$ be a sequence of successive observations obtained as described in Section 3. We denote by $Z_i = Z_i(\mathbf{O}_n(i-1)) \in \mathcal{Z} \subset \mathbb{R}^d$ a summary measure of $\mathbf{O}_n(i-1)$ of fixed dimension d (for instance $Z_i = \theta_{i-1} \in \mathbb{R}^2$, the current maximum likelihood estimator of θ at step i). Let $\mathcal{F}$ be a class of bounded (and measurable) functions of $(o, z) = (a, y, z)$ such that $\sup_{f\in\mathcal{F}} \|f\|_\infty = U < \infty$ (for instance, $f(O_i, Z_i) = D(\vartheta)(O_i) - P_{\theta,g_i^\star} D(\vartheta)$ for some $\vartheta \in [0,1]^2$, where the dependency wrt $Z_i = \theta_{i-1}$ is conveyed through $g_i^\star$). Defining

$$M_n(f) = \frac{1}{n}\sum_{i=1}^n [f(O_i, Z_i) - P_{\theta,g_i} f]$$

for all $f \in \mathcal{F}$ and $n \geq 1$, we note that $nM_n(f)$ is a discrete martingale sum.

Our uniform convergence result, Theorem 8, essentially relies on a maximal inequality for martingales taken from [30] (Proposition A.2) which we now present.

Let ϕ be the function characterized over $\mathbb{R}$ by $\phi(x) = e^x - x - 1$. We define the generalized entropy with bracketing as follows. Let $n \geq 1$, $K > 0$ and $\varepsilon > 0$ be given. A finite collection

$\{(\Lambda_i^j, \Upsilon_i^j)_{i\le n}\}_{j\le N}$ of random variables is called a $(n, \mathcal{F}, K, \varepsilon)$-bracketing set if Λ_i^j and Υ_i^j are (measurable) functions of $\mathbf{O}_n(i)$ for all $i \le n, j \le N$ and if for every $f \in \mathcal{F}$, there exists $j \le N$ (the map $f \mapsto j$ is non-random) such that P-almost surely, for all $i \le n$,

$$\Lambda_i^j \le f(O_i, Z_i) \le \Upsilon_i^j$$

and such that, for all $j \le N$,

$$\frac{2K^2}{n}\sum_{i=1}^{n} E\left[\phi\left(\frac{|\Upsilon_i^j - \Lambda_i^j|}{K}\right)\middle|\mathbf{O}_n(i-1)\right] \le \varepsilon^2.$$

We denote by $\mathcal{N}(n, \mathcal{F}, K, \varepsilon)$ the cardinality N of the smallest $(n, \mathcal{F}, K, \varepsilon)$-bracketing set, and call $\mathcal{H}(n, \mathcal{F}, K, \varepsilon) = \log \mathcal{N}(n, \mathcal{F}, K, \varepsilon)$ its generalized entropy with bracketing. Then,

Theorem 6 (Proposition A.2 in [30]). *Fix $K > 0$ and define*

$$R_{n,K}(f) = \frac{2K^2}{n}\sum_{i=1}^{n} P_{\theta,g_i}\phi\left(\frac{|f|}{K}\right)$$

for all $f \in \mathcal{F}$ and $n \ge 1$. There exists a universal constant $C > 0$ (the choice $C = 100$ works) such that, for any $n \ge 1$, $R > 0$,

$$P\left(\sup_{f\in\mathcal{F}} \mathbb{1}\{R_{n,K}(f) \le R\} \max_{i\le n} \frac{i}{n} M_i(f) \ge \alpha\right) \le 2\exp\left\{-\frac{n\alpha^2}{C^2(c_1+1)R}\right\}$$

for any $\alpha, c_0, c_1 > 0$ satisfying $c_0^2 \ge C^2(c_1+1)$ and

$$\frac{c_0}{\sqrt{n}}\int_0^{\sqrt{R}} \sqrt{\mathcal{H}(n, \mathcal{F}, K, x)}dx \le \alpha \le \frac{c_1 R}{K}. \tag{27}$$

Note that the uncountable supremum is interpreted as an essential supremum under P, in order to avoid measurability issues.

It appears that it is important to understand the behavior of the random variables $R_{n,K}(f)$ for $f \in \mathcal{F}$. Furthermore, condition (27) may be hard to check because of the relatively intricate definition of the generalized entropy with bracketing. The following lemma provides answers to both issues. Notably, $\mathcal{H}(n, \mathcal{F}, K, \varepsilon)$ is here easily related to the standard entropy for the supremum norm $H(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = \log N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$. Recall that $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$ is the cardinality of the smallest finite collection $(\ell^j, u^j)_{j\le N}$, where $\ell^j \le u^j$ are (measurable) functions of $(o, z) = (a, y, z)$ satisfying $\|u^j - \ell^j\|_\infty \le \varepsilon$ for all $j \le N$, such that for every $f \in \mathcal{F}$, there exists $j \le N$ for which $\ell^j \le f \le u^j$.

Lemma 7. *Recall that $\sup_{f\in\mathcal{F}} \|f\|_\infty = U < \infty$. It holds that:*

(i) For all $n \ge 1$ and $f \in \mathcal{F}$, $R_{n,4U}(f) \le \frac{4}{3}U^2$.

(ii) For all $n \ge 1$ and $\varepsilon > 0$, $\mathcal{H}(n, \mathcal{F}, 4U, \sqrt{2}\varepsilon) \le H(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$.

Proof. First, arbitrarily choose $f \in \mathcal{F}$. For each $i \le n$ and all $m \ge 2$, since $\|f\|_\infty \le U$, one has $P_{\theta,g_i}|f|^m \le U^m \le \frac{m!}{2}U^m$, hence

$$2(4U)^2 P_{\theta,g_i}\phi\left(\frac{|f|}{2U}\right) = 32U^2\sum_{m\ge 2}\frac{P_{\theta,g_i}|f|^m}{m!(4U)^m} \le 16U^2\sum_{m\ge 2} 4^{-m} = \frac{4}{3}U^2.$$

This straightforwardly entails (i).

Second, fix $\varepsilon > 0$, define $N = \exp\{H(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)\}$ and let the collection $(\ell^j, u^j)_{j\le N}$ satisfying $\|u^j - \ell^j\|_\infty \le \varepsilon$ for all $j \le N$, be such that for every $f \in \mathcal{F}$, there exists $j \le N$ for which

$\ell^j \leq f \leq u^j$. Introduce $\Lambda_i^j = \ell^j(O_i, Z_i) \wedge (-U)$ and $\Upsilon_i^j = u^j(O_i, Z_i) \vee U$ for every $i \leq n, j \leq N$. For each $f \in \mathcal{F}$, there exists $j \leq N$ (the map $f \mapsto j$ is non-random) such that $-U \leq \Lambda_i^j \leq f(O_i, Z_i) \leq \Upsilon_i^j \leq U$ for all $i \leq n$.

Set $j \leq N$. For each $i \leq n$ and all $m \geq 2$, one has $E[|\Upsilon_i^j - \Lambda_i^j|^m | \mathbf{O}_n(i-1)] \leq (2U)^{m-2}\varepsilon^2 \leq \frac{m!}{2}(2U)^{m-2}\varepsilon^2$, hence

$$2(4U)^2 E\left[\phi\left(\frac{|\Upsilon_i^j - \Lambda_i^j|}{4U}\right)\middle| \mathbf{O}_n(i-1)\right] = 32U^2 \sum_{m\geq 2} \frac{E[|\Upsilon_i^j - \Lambda_i^j|^m | \mathbf{O}_n(i-1)]}{m!(4U)^m} \leq 2\varepsilon^2.$$

Property (ii) immediately follows. □

Now we can state and prove our building block for consistency results:

Theorem 8. *Recall that* $\sup_{f\in\mathcal{F}} \|f\|_\infty = U < \infty$. *If* $\int_0^{\sqrt{2/3}U} \sqrt{H(\mathcal{F}, \|\cdot\|_\infty, x)}dx < \infty$ *then for all* $\alpha > 0$ *there exists* $c > 0$ *such that, for* n *large enough,*

$$P\left(\sup_{f\in\mathcal{F}} M_n(f) \geq \alpha\right) \leq 2e^{-nc}.$$

Consequently, $\sup_{f\in\mathcal{F}} |M_n(f)|$ *converges to 0 almost surely.*

Proof. Fix $\alpha > 0$ and choose $K = 4U$, $R = \frac{4}{3}U^2$, $c_1 = \frac{\alpha K}{R}$, $c_0 = C\sqrt{c_1 + 1}$ (C is the universal constant introduced in Theorem 6) and let n_1 be the smallest integer such that $\sqrt{n} \geq \frac{c_0}{\alpha}\int_0^{\sqrt{R/2}} \sqrt{H(\mathcal{F}, \|\cdot\|_\infty, x)}dx$. Note that (ii) in Lemma 7 guarantees that condition (27) from Theorem 8 is met, while condition (i) of the same lemma implies that $\mathbb{1}\{R_{n,K}(f) \leq R\} = 1$. Therefore, Theorem 8 ensures that, for all $n \geq n_1$,

$$P\left(\sup_{f\in\mathcal{F}} M_n(f) \geq \alpha\right) \leq P\left(\sup_{f\in\mathcal{F}} \max_{i\leq n} \frac{i}{n} M_i(f) \geq \alpha\right) \leq 2e^{-nc},$$

where $c = \alpha^2/c_0^2 R$.

Now one can assume without loss of generality that for each $f \in \mathcal{F}$, $-f \in \mathcal{F}$ too (otherwise, define $\mathcal{F}' = \mathcal{F} \cup \{-f : f \in \mathcal{F}\}$ and note that $\sup_{f\in\mathcal{F}'} \|f\|_\infty = U$ and $N(\mathcal{F}', \|\cdot\|_\infty, \varepsilon) \leq 2N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$ for all $\varepsilon > 0$). Obviously then, $\sup_{f\in\mathcal{F}} |M_n(f)| = \max\{\sup_{f\in\mathcal{F}} M_n(f), \sup_{f\in\mathcal{F}} M_n(-f)\} = \sup_{f\in\mathcal{F}} M_n(f)$. We conclude by virtue of the Borel-Cantelli lemma. □

A.2 Building block for central limit theorems.

First, we obtain a central limit theorem for univariate discrete martingale sums as a by-product of a classical theorem (see *e.g.* Theorem 3.3.7 in [24]). Second, we rely on it and invoke the Cramér-Wold device (see *e.g.* Theorem 3.2.4 in [24]) in order to extend the result to the case of multivariate discrete martingale sums.

Univariate case.

We use the same framework and notation as those exposed at the beginning of Section A.1. For a given *real-valued* $f \in \mathcal{F}$, let us introduce

$$\begin{aligned} w_n(f)^2 &= \sum_{i=1}^n P_{\theta,g_i} f^2, \\ s_n(f)^2 &= E w_n(f)^2 = \sum_{i=1}^n E f(O_i, Z_i)^2, \quad \text{and} \\ \sigma_n(f)^2 &= \frac{s_n(f)^2}{n}. \end{aligned}$$

The following univariate central limit theorem holds:

Theorem 9. *Assume that for all $i = 1, \ldots, n$, $g_i(A_i|\mathbf{A}_n(i-1), \mathbf{X}_n) = g_i(A_i|X_i(A_i), \mathbf{O}_n(i-1))$ (by virtue of the adaptive CAR assumption (9)) only depends on $\mathbf{O}_n(i-1)$ through Z_i. If $\liminf \sigma_n(f)^2 > 0$ and $\frac{1}{n}w_n(f)^2 - \frac{1}{n}Ew_n(f)^2 = \frac{1}{n}w_n(f)^2 - \sigma_n(f)^2$ converges in probability to 0, then $w_n(f)^2/s_n(f)^2$ converges to 1 in probability and $\sqrt{n}\frac{M_n(f)}{\sigma_n(f)}$ converges in distribution to the standard normal distribution.*

Furthermore, the empirical mean $\hat{\sigma}_n(f)^2 = \frac{1}{n}\sum_{i=1}^n f(O_i, Z_i)^2$ mimics $\sigma_n(f)^2$ in the sense that $\hat{\sigma}_n(f)^2 - \sigma_n(f)^2$ converges to 0 in probability. Consequently, $\sqrt{n}\frac{M_n(f)}{\hat{\sigma}_n(f)}$ also converges in distribution to the standard normal distribution.

This result notably teaches us that we can estimate the asymptotic variance of $\sqrt{n}M_n(f)$ by considering $O_1, \ldots, O_n$ as independent draws from P_{θ,g_i}, treating each g_i as a given deterministic fixed design in $\mathcal{G}$. Therefore, the parametric or non-parametric bootstrap ignoring the dependence structure of $\mathbf{O}_n$ consistently estimate the limiting variance.

Proof. Regarding the behavior of $\hat{\sigma}_n(f)^2$, note that

$$\hat{\sigma}_n(f)^2 = \sigma_n(f)^2 + \left(\frac{1}{n}w_n(f)^2 - \sigma_n(f)^2\right) + \frac{1}{n}M_n' = \sigma_n(f)^2 + \frac{1}{n}M_n' + o_P(1)$$

where $M_n' = \sum_{i=1}^n (f(O_i, Z_i)^2 - P_{\theta,g_i}f^2)$ is a discrete martingale sum with uniformly bounded terms. The Kolmogorov strong law of large numbers (see *e.g.* Theorem 2.4.2 in [24]) guarantees that $\frac{1}{n}M_n'$ converges almost surely, hence in probability, to 0. So $\hat{\sigma}_n(f)^2 - \sigma_n(f)^2 = o_P(1)$ and, by Slutsky's lemma, $\sqrt{n}M_n(f)/\hat{\sigma}_n(f)$ converges to a standard normal distribution when $\sqrt{n}M_n(f)/\sigma_n(f)$ does.

To prove this, we invoke the central limit theorem for discrete martingales (see *e.g.* Theorem 3.3.7 in [24]). First, we must check that $w_n(f)^2/s_n(f)^2$ converges to 1 in probability. Now, since $\liminf \sigma_n(f)^2 > 0$, there exist $n_1 \geq 1, C > 0$ such that $n \geq n_1$ yields $\sigma_n(f)^2 \geq C$ and also

$$\left|\frac{w_n(f)^2}{s_n(f)^2} - 1\right| = \frac{\left|\frac{1}{n}w_n(f)^2 - \sigma_n(f)^2\right|}{\sigma_n(f)^2} \leq \frac{1}{C}\left|\frac{1}{n}w_n(f)^2 - \sigma_n(f)^2\right| = o_P(1)$$

by assumption. Second, we must check that for every $\varepsilon > 0$,

$$\sum_{i=1}^n E\left(f(O_i, Z_i)^2 \mathbb{1}\{f(O_i, Z_i)^2 \geq \varepsilon^2 s_n(f)^2\}\right) = o(s_n(f)^2). \tag{28}$$

Fix $\varepsilon > 0$ and $n > \max\{n_1, \|f\|_\infty^2/\varepsilon^2 C\}$: $\mathbb{1}\{f(O_i, Z_i)^2 \geq \varepsilon^2 s_n(f)^2\} \leq \mathbb{1}\{\|f\|_\infty^2 \geq n\varepsilon^2 C\} = 0$, so that the left-hand side expression in (28) is bounded while $s_n(f)^2$ goes to infinity. Therefore, the central limit theorem for discrete martingales applies, and implies that $\sqrt{n}\frac{M_n(f)}{\sigma_n(f)}$ converges in distribution to the standard normal distribution. This concludes the proof. □

Multivariate case.

Let us state the multivariate version of Theorem 9. It involves the following multidimensional counterparts of $w_n(f)^2$, $s_n(f)^2$ and $\sigma_n(f)^2$ in the case that $f \in \mathcal{F}$ takes values in $\mathbb{R}^r$ (expectations are taken componentwise):

$$\begin{aligned} W_n(f) &= \sum_{i=1}^n P_{\theta,g_i} f f^\top, \\ S_n(f) &= EW_n(f) = \sum_{i=1}^n Ef(O_i, Z_i) f(O_i, Z_i)^\top, \quad \text{and} \\ \Sigma_n(f) &= \frac{S_n(f)}{n}. \end{aligned}$$

For every positive definite symmetric matrix Σ, we denote by $\Sigma^{-1/2}$ the positive definite symmetric matrix such that $(\Sigma^{-1/2})^2$ is the inverse of Σ.

Theorem 10. *Assume that for all $i = 1, \ldots, n$, $g_i(A_i|\mathbf{A}_n(i-1), \mathbf{X}_n) = g_i(A_i|X_i(A_i), \mathbf{O}_n(i-1))$ (by virtue of the adaptive CAR assumption (9)) only depends on $\mathbf{O}_n(i-1)$ through Z_i. If $\Sigma_n(f)$ converges to a positive definite covariance matrix $\Sigma(f)$ and $\frac{1}{n}W_n(f) - \frac{1}{n}EW_n(f) = \frac{1}{n}W_n(f) - \Sigma_n(f)$ converges componentwise in probability to 0, then $\sqrt{n}M_n(f)$ converges in distribution to the centered Gaussian law over $\mathbb{R}^r$ with covariance matrix $\Sigma(f)$.*

Furthermore, the empirical mean $\hat{\Sigma}_n(f) = \frac{1}{n}\sum_{i=1}^n f(O_i, Z_i)f(O_i, Z_i)^\top$ is such that $\hat{\Sigma}_n(f)$ converges to $\Sigma(f)$ in probability. Consequently, $\hat{\Sigma}_n(f)$ is invertible with probability tending to 1 as n tends to infinity, and $\sqrt{n}\hat{\Sigma}_n(f)^{-1/2}M_n(f)$ converges in distribution to the standard normal distribution over $\mathbb{R}^r$.

Proof. The Cramér-Wold device (see *e.g.* Theorem 3.2.4 in [24]) teaches us that $\sqrt{n}M_n(f)$ converges in distribution to the centered Gaussian law over $\mathbb{R}^r$ with covariance matrix $\Sigma(f)$ if and only if $\sqrt{n}\lambda^\top M_n(f)$ converges to the centered Gaussian distribution over $\mathbb{R}$ with variance $\lambda^\top\Sigma(f)\lambda$ for all $\lambda \in \mathbb{R}^r$. Arbitrarily choose $\lambda \in \mathbb{R}^r$ and consider $\lambda^\top M_n(f) = M_n(\lambda^\top f)$. Does Theorem 9 apply?

Note first that $\sigma_n(\lambda^\top f)^2 = \lambda^\top\Sigma_n(f)\lambda$ and also that $\frac{1}{n}w_n(\lambda^\top f)^2 - \sigma_n(\lambda^\top f)^2 = \lambda^\top(\frac{1}{n}W_n(f) - \Sigma_n(f))\lambda$. Therefore $\liminf \sigma_n(\lambda^\top f)^2 = \lim \sigma_n(\lambda^\top f)^2 = \lambda^\top\Sigma(f)\lambda > 0$, $\frac{1}{n}w_n(\lambda^\top f)^2 - \sigma_n(\lambda^\top f)^2$ converges to 0 in probability, Theorem 9 applies and yields that $\sqrt{n}\lambda^\top M_n(f)$ converges to the desired Gaussian distribution, hence the stated convergence of $\sqrt{n}M_n(f)$.

Following the same lines as in the proof of Theorem 9, we then remark that

$$\hat{\Sigma}_n(f) = \Sigma_n(f) + \left(\frac{1}{n}W_n(f) - \Sigma_n(f)\right) + \frac{1}{n}M'_n = \Sigma(f) + \frac{1}{n}M'_n + o_P(1),$$

where $M'_n = \sum_{i=1}^n (ff^\top(O_i, Z_i) - P_{\theta,g_i}ff^\top)$ is a discrete (multivariate) martingale sum with uniformly bounded terms. The Kolmogorov-Smirnov strong law of large numbers (see eg Theorem 2.4.2 in [24]) guarantees that $\frac{1}{n}M'_n$ converges almost surely, hence in probability, to 0. The rest follows because the set of invertible symmetric matrices is open in the set of symmetric matrices, and thanks to Slutsky's lemma. □

A.3 A tailored test of empirical coverage, type I error and power.

Many times in this article we wish to test if the requirements on confidence intervals coverage, type I error and power of tests are met across several data generating distributions (characterized by Θ_0 or a subset Θ of it). Those three issues can be addressed in a common simple framework.

Single p-value for multiple pairwise comparisons.

In each case, the decision must be made based on iid Binomial random variables $\{B(\theta) : \theta \in \Theta\}$ with parameters (M, p), the null stating "$p = \pi$" and its alternative "$p < \pi$" (when testing the empirical coverage and power) or "$p > \pi$" (when testing the empirical type I error).

Let us denote by F_p the Binomial cdf with parameters (M, p). Every $B(\theta)$ is associated with a p-value $P(\theta)$ which is either $F_\pi(B(\theta))$ (when testing the empirical coverage and power) or $(1 - F_\pi(B(\theta)))$ (when testing the empirical type I error). Rather than using the latter p-values directly, we consider the randomly perturbed $P'(\theta) = P(\theta) + \varepsilon U(\theta)$, for iid Uniform random variables $\{U(\theta) : \theta \in \Theta\}$ over $[0, 1]$ and a small real number $\varepsilon > 0$. The substitution of $P'(\theta)$ to $P(\theta)$ is advantageous because the distribution $G_{p,\varepsilon}$ of $P'(\theta)$ is continuous, whereas that of $P(\theta)$ is not (and we do observe several ties in every situation). In addition, $P'(\theta)$ is stochastically smaller under the alternative than under the null: $G_{\pi,\varepsilon} \leq G_{p,\varepsilon}$, small values of $P'(\theta)$ therefore being more likely under the alternative than under the null.

The final step consists of comparing the empirical distribution of $\{P'(\theta) : \theta \in \Theta\}$ with $G_{\pi,\varepsilon}$. We decide to do so in terms of one-sided Kolmogorov-Smirnov goodness-of-fit, the alternative stating that the common cdf of the $P'(\theta)$'s lies above $G_{\pi,\varepsilon}$. We estimate $G_{\pi,\varepsilon}$ by its empirical counterpart based on 10^6 simulated random variables drawn from $G_{\pi,\varepsilon}$. This procedure yields a single p-value $\Pi_{p,\varepsilon}$ whose distribution under the null "$p = \pi$" is uniform over $[0, 1]$.

Simulation study.

The tailored procedure we just described inherits a great sensibility to departures from the null from that of the Kolmogorov-Smirnov goodness-of-fit test. This is the obvious conclusion drawn from the study of Figure 7. We reproduce by simulation the behavior of the p-value $\Pi_{p,\varepsilon}$ of our procedure in the context of empirical coverage, type I error and power quality assessment. By symmetry, it is equivalent to evaluate the performance of our procedure for testing 95%-coverage and 5%-type I error quality assessment. We simulate 1000 independent copies of a collection of 45 (the cardinality of Θ_0) iid Binomial random variables with parameter (M, p), p ranging

- from $\pi = 0.050$ (correct type I error) to 0.055 (inflated type I error),
- from $\pi = 0.900$ (correct power at alternative) to 0.895 (deflated power at alternative)

by steps of length 10^{-3}. The constant ε is arbitrarily set to 10^{-6}. Based on those simulated datasets, we can estimate accurately the cdf $u \mapsto P_p(\Pi_{p,\varepsilon} \leq u)$ of $\Pi_{p,\varepsilon}$ in each configuration. The fit to a Uniform distribution over $[0, 1]$ under the two nulls is perfect. In addition, the procedure features large power at local alternatives.

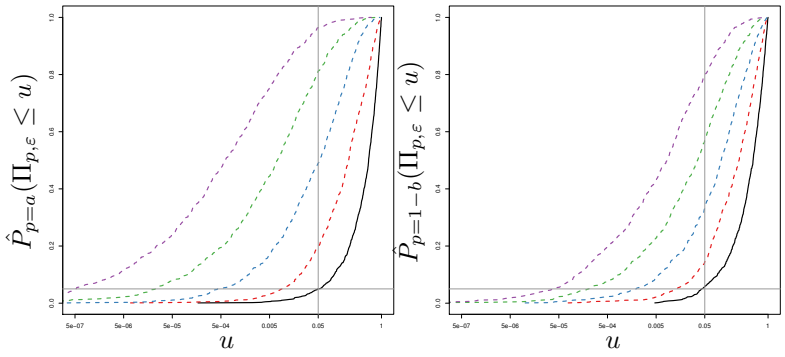

Figure 7: Illustrating the sensibility of our tailored test of empirical coverage, type I error and power. Left: estimated cdf $u \mapsto \hat{P}_{p=a}(\Pi_{p,\varepsilon} \leq u)$ for $a = 0.050$ (rightmost curve) to $a = 0.055$ (leftmost curve). Right: estimated cdf $u \mapsto \hat{P}_{p=1-b}(\Pi_{p,\varepsilon} \leq u)$ for $b = 0.100$ (rightmost curve) to $c = 0.105$ (leftmost curve). The two plots share the same range on x-axis (in logarithmic scale) and y-axis. A vertical and an horizontal lines at the reference level $\alpha = 5\%$ are drawn on each plot.

A.4 A tailored test for multiple pairwise comparisons of empirical distributions of confidence interval widths or sample sizes at decision.

In the same spirit as in the previous section, we want to come up with a common framework to compare two empirical distributions of confidence interval widths or sample sizes at decision across several data generating distributions (characterized again by Θ_0 or a subset of it). Some extra care is needed because the distribution of sample size at decision has atoms.

Single p-value for pairwise comparison.

Denote by P_F and $\mathrm{P}_{F'}$ two probability distributions over the real line, with respective cdf's F and F' that may have jumps. We want to test the null "$F = F'$" against the alternative "$\exists\, t \in \mathbb{R} : F'(t) < F(t)$". The test is based on two independent n-tuples of iid random variables $O_1, \ldots, O_n \sim \mathrm{P}_F$ and $O'_1, \ldots, O'_n \sim \mathrm{P}_{F'}$. We decide to rely on the two-sample one-sided Kolmogorov-Smirnov statistic T_n: letting $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{O_i}$ and $\mathbb{P}'_n = n^{-1} \sum_{i=1}^n \delta_{O'_i}$ be the empirical distributions and

$\mathcal{F} = \{\mathbb{1}(-\infty, t] : t \in \mathbb{R}\}$, $T_n = \sqrt{n/2}\sup_{f\in\mathcal{F}}[(\mathbb{P}_n - \mathbb{P}'_n)f]$, the rescaled maximum gap between the empirical counterparts of F and F' over the set of those points where the former dominates the latter. The following lemma is a well-known result (see Chapter 19 of [28] and pages 85 and 142 of [5]):

Lemma 11. *Under the null, $T_n \rightsquigarrow \sup_{f\in\mathcal{F}} \mathbb{G}_F f$ where the P_F-Brownian bridge $\mathbb{G}_F$ is the zero mean Gaussian process over $\mathcal{F}$ characterized by the covariance structure $E\mathbb{G}_F f\mathbb{G}_F g = \mathrm{P}_F fg - \mathrm{P}_F f\mathrm{P}_F g$ for every $f, g \in \mathcal{F}$. Under the alternative, T_n tends to infinity almost surely. In addition, if F is continuous, then $\mathrm{P}(\sup_{f\in\mathcal{F}} \mathbb{G}_F f \geq t) = e^{-2t^2}$ for all $t \geq 0$.*

This result suggests to reject the null when the observed value of the test statistic T_n is larger than the upper α-quantile of its limit distribution under the null. The asymptotic p-value $P_n = \mathrm{P}(\sup_{f\in\mathcal{F}} \mathbb{G}_F f \geq t)|_{t=T_n}$ is known in closed form when F is continuous (for instance when comparing distributions of confidence intervals widths): $P_n = e^{-2T_n^2}$, but not otherwise (for instance when comparing distributions of sample size at decision). However, it is possible to resort to the bootstrap to estimate P_n when F is not continuous.

Define $\mathbb{F}_n = \frac{1}{2}(\mathbb{P}_n + \mathbb{P}'_n)$, the pooled empirical measure. Let $O_1^*, \ldots, O_n^*$ and $O_1'^*, \ldots, O_n'^*$ be two independent iid samples from $\mathbb{F}_n$. They give rise to the bootstrapped empirical distributions $\hat{\mathbb{P}}_n = n^{-1}\sum_{i=1}^n \delta_{O_i^*}$ and $\hat{\mathbb{P}}'_n = n^{-1}\sum_{i=1}^n \delta_{O_i'^*}$, and to the bootstrapped empirical process $\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \hat{\mathbb{P}}'_n)$, hence finally to $\hat{T}_n = \sqrt{n/2}\sup_{f\in\mathcal{F}} \hat{\mathbb{G}}_n f = \sqrt{n/2}\sup_{f\in\mathcal{F}}[(\hat{\mathbb{P}}_n - \hat{\mathbb{P}}'_n)f]$. The next lemma teaches us that the conditional distribution of $\hat{T}_n$ consistently estimates the limit distribution of T_n under the null (see Chapter 2.9 of [29] for the definition of conditional convergence in distribution; the lemma is a straightforward application of Theorem 3.6.2 in [29]):

Lemma 12. *$\hat{T}_n \rightsquigarrow \sup_{f\in\mathcal{F}} \mathbb{G}_F f$ given almost every sequence $O_1, O_2, \ldots, O'_1, O'_2, \ldots$.*

Consequently, it is possible to estimate P_n when F is not continuous by generating a large number B of independent copies $\hat{T}_n^b$ of $\hat{T}_n$ and using $P_n \simeq B^{-1}\sum_{b=1}^B \mathbb{1}\{\hat{T}_n^b \geq T_n\}$.

Single p-value for multiple pairwise comparisons.

Say that each $\theta \in \Theta$ comes with the empirical counterparts of two distributions $\mathrm{P}_{F(\theta)}, \mathrm{P}_{F'(\theta)}$ and that we wish to test the *global null* "$\forall\ \theta \in \Theta, F(\theta) = F'(\theta)$" against the alternative "$\exists\ \theta \in \Theta, \exists\ t \in \mathbb{R} : F'(\theta)(t) < F(\theta)(t)$". Here $\mathrm{P}_{F(\theta)}$ may for instance be the distribution of the width of a confidence interval for the log-relative risk obtained under iid optimal $(\theta, g^\star)$ sampling and $\mathrm{P}_{F'(\theta)}$ that of a confidence interval for the same parameter obtained under adaptive $(\theta, \mathbf{g}_n^\star)$ sampling for a given sample size n_i, see Section 6. Or $\mathrm{P}_{F(\theta)}$ may the distribution of sample size at decision for sequential testing of a certain couple of hypotheses under iid optimal $(\theta, g^\star)$ sampling and $\mathrm{P}_{F'(\theta)}$ that of sample size at decision for sequential testing of the same hypotheses under adaptive $(\theta, \mathbf{g}_n^\star)$ sampling, see Section 7.

We just showed how to associate a p-value $P(\theta)$ with each θ while testing "$F(\theta) = F'(\theta)$" against "$\exists\ t \in \mathbb{R} : F'(\theta)(t) < F(\theta)(t)$". Under the so-called global null hypothesis, the independent p-values $\{P(\theta) : \theta \in \Theta\}$ are uniformly distributed over $[0, 1]$. Under its alternative, at least one of them is more likely to take smaller values than a Uniform random variable. Proceeding as in Section A.3, assuming in particular that the p-values are identically distributed (even when the global null does not hold, a reasonable assumption in our context), we finally test the global null against its alternative in terms of one-sided Kolmogorov-Smirnov goodness-of-fit test: we compare the empirical distribution of $\{P(\theta) + \varepsilon U(\theta) : \theta \in \Theta_0\}$ for arbitrarily fixed small real number $\varepsilon > 0$ and iid Uniform random variables $\{U(\theta) : \theta \in \Theta\}$ with the Uniform distribution over $[0, 1]$ (we neglect the impact of the ε times Uniform random variable terms), the alternative stating that the common cdf of the $P(\theta)$'s lies above that of the Uniform distribution. Finally, we decide to reject the global null for its alternative if the latter Kolmogorov-Smirnov test yields a rejection of its null hypothesis.

Simulation study.

In order to illustrate the tailored procedure for multiple pairwise comparisons of empirical distributions of sample sizes at decision that we just described, we propose to carry out the following simulation study.

In the first place, let us present objective of the simulation study. Say that we first retrieve the $\#\Theta$ independent samples $\{S(\theta, g^\star)_m : m = 1, \ldots, M\}$ (for each $\theta \in \Theta$) of sample sizes at decision obtained while testing "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" in the framework of Section 7.2 and the simulation study of type I error. Let us denote by $\mathbb{P}_M(\theta)$ the empirical distribution of $S(\theta, g^\star)_1, \ldots, S(\theta, g^\star)_M$. For each $\theta \in \Theta$, we draw two independent random samples of length 1000 from $\mathbb{P}_M(\theta)$: the first one is considered as a reference sample, while the second one is (deterministically) perturbed in order to yield a second sample whose distribution is either the same as the distribution of the reference sample or dominated by the distribution of the reference sample (*i.e.* values from the second sample tend to be larger than values from the first sample). Therefore we obtain two independent collections of $\#\Theta$ independent random samples of same lengths. We then resort to our tailored procedure for multiple pairwise comparisons of empirical distributions of sample sizes in order to compare them, yielding a p-value $\Pi_{p,\varepsilon}$ (where p indicates how we perturb the second sample and ε is the arbitrarily fixed small real number used in the testing procedure). The objective of the simulation study that we are on the verge of describing is to investigate the distribution of $\Pi_{p,\varepsilon}$.

Now, let us present the simulation scheme. We repeat $M = 1000$ times the following steps for each $p \in \{0, 10^{-4}, 10^{-3}, 2.5 \times 10^{-3}, 5 \times 10^{-3}\}$: at the mth iteration,

- for every $\theta \in \Theta$, draw under $\mathbb{P}_M(\theta)$ two independent n-tuples with $n = 1000$ that we denote by $(S_1(\theta), \ldots, S_n(\theta))$ and $(S'_1(\theta), \ldots, S'_n(\theta))$;
- for every $\theta \in \Theta$, perturb the second random sample by introducing, for all $i \leq n$, $S''_i(\theta) = \lceil (1+p) S'_i(\theta) \rceil$;
- in order to compare the two independent collections $\{(S_1(\theta), \ldots, S_n(\theta)) : \theta \in \Theta_0\}$ and $\{(S''_1(\theta), \ldots, S''_n(\theta)) : \theta \in \Theta_0\}$, apply the multiple pairwise comparisons of empirical distributions of sample sizes at decision test procedure presented in the previous subsection (with $B = 1000$ and $\varepsilon = 10^{-6}$), therefore yielding a single p-value $\Pi^m_{p,\varepsilon}$.

Based on this simulated dataset, we can estimate accurately the cdf $u \mapsto P_p(\Pi_{p,\varepsilon} \leq u)$ of $\Pi_{p,\varepsilon}$ in each configuration. We finally represent in Figure 8 those estimated cdfs. The fit to a Uniform distribution over $[0, 1]$ under the null (*i.e.* when $p = 0$) is excellent. In addition, the procedure exhibits good performances in terms of power when $p \geq 2.5 \times 10^{-3}$. This is a very good result, given the mean sample sizes at decision reported in Table 30 (they approximately range between 800 and 13000).

A.5 A summary of the results of the simulation studies carried out for this article.

In this final section, we report summaries of the results of the simulation studies carried out for this article.

- Tables 12 to 19 provide empirical coverage of the confidence intervals obtained under iid g^b-balanced sampling scheme (Tables 12 and 13), iid $g^\star$-optimal sampling scheme (Tables 14 and 15), $\mathbf{g}^\star_n$-adaptive sampling scheme (Tables 16 and 17), and $\mathbf{g}^a_n$-adaptive sampling scheme (Tables 18 and 19). See Section 6.3 for more details.
- Tables 20 to 27 provide the empirical means and standard deviations of a criterion comparing the optimal widths of confidence intervals with the widths of confidence intervals obtained under iid g^b-balanced sampling scheme (Tables 20 and 21), iid $g^\star$-optimal sampling scheme (Tables 22 and 23), $\mathbf{g}^\star_n$-adaptive sampling scheme (Tables 24 and 25), and $\mathbf{g}^a_n$-adaptive sampling scheme (Tables 26 and 27). See Section 6.4 for more details.

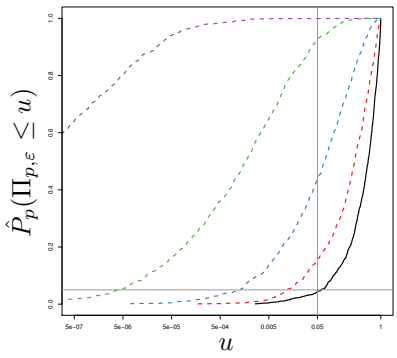

Figure 8: Illustrating the sensibility of our tailored test for multiple pairwise comparisons of empirical distributions of sample sizes at decision. We represent the estimated cdf $u \mapsto \hat{P}_p(\Pi_{p,\varepsilon} \leq u)$ for $p = 0$ (rightmost curve) to $p = 5 \times 10^{-3}$ (leftmost curve). We use a logarithmic scale on the x-axis. A vertical and an horizontal lines at the reference level $\alpha = 5\%$ are drawn.

- Tables 28 and 29 respectively provide empirical type I and type II errors obtained under iid g^b-balanced, iid $g^\star$-optimal, $\mathbf{g}_n^\star$-adaptive, and $\mathbf{g}_n^a$-adaptive sampling schemes. See Section 7.2 for more details.
- Tables 30 and 31 provide mean sample sizes at decision obtained when investigating the adequateness of type I and type II errors, respectively, under iid g^b-balanced, iid $g^\star$-optimal, $\mathbf{g}_n^\star$-adaptive, and $\mathbf{g}_n^a$-adaptive sampling schemes. See Section 7.3 for more details.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.963	.966	.957	.952	.966	.954	.955	.953	.955
.2	-	.965	.956	.954	.954	.959	.947	.948	.943
.3	-	-	.962	.952	.963	.954	.942	.937	.941
.4	-	-	-	.954	.958	.964	.952	.948	.945
.5	-	-	-	-	.956	.957	.952	.952	.957
.6	-	-	-	-	-	.948	**.967**	.936	**.934**
.7	-	-	-	-	-	-	.941	.940	.938
.8	-	-	-	-	-	-	-	.958	.940
.9	-	-	-	-	-	-	-	-	.953

sample size $n_1 = 100$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.952	.950	.961	.959	.960	.957	.954	.952	.951
.2	-	.951	.956	.954	.950	.943	.951	.949	.943
.3	-	-	.962	**.964**	.948	.963	.956	.954	**.936**
.4	-	-	-	.940	.958	.956	.956	.953	.952
.5	-	-	-	-	.955	.957	.960	.946	.953
.6	-	-	-	-	-	.958	.955	.943	.957
.7	-	-	-	-	-	-	.952	.946	.957
.8	-	-	-	-	-	-	-	.959	.954
.9	-	-	-	-	-	-	-	-	.941

sample size $n_2 = 250$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.947	.951	**.967**	.952	.937	.950	.939	.947	.943
.2	-	.948	.944	.954	.957	.948	.944	.959	.962
.3	-	-	.963	.958	.957	.961	.954	**.936**	.944
.4	-	-	-	.952	.950	.956	.956	.950	.948
.5	-	-	-	-	.958	.943	.948	.958	.956
.6	-	-	-	-	-	.956	.938	.943	.946
.7	-	-	-	-	-	-	.954	.952	.949
.8	-	-	-	-	-	-	-	.953	.950
.9	-	-	-	-	-	-	-	-	.948

sample size $n_3 = 500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.943	.948	.961	.941	.957	.951	.941	.957	.937
.2	-	.950	.941	.941	.950	.946	.947	.964	.954
.3	-	-	**.968**	.953	.955	.953	.957	.939	.957
.4	-	-	-	.950	.950	.954	.957	.938	.944
.5	-	-	-	-	.946	.949	.957	.963	.954
.6	-	-	-	-	-	.947	**.936**	.940	.944
.7	-	-	-	-	-	-	.944	.938	.953
.8	-	-	-	-	-	-	-	.957	.952
.9	-	-	-	-	-	-	-	-	.956

sample size $n_4 = 750$

Table 12: Empirical coverage of $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$ for $i = 1, 2, 3, 4$ under iid balanced (θ, g^b) sampling. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.952	.941	.955	.956	.959	.945	.941	.956	.944
.2	-	.956	.947	.947	.953	.955	.951	.954	.953
.3	-	-	.953	.956	.950	**.961**	.947	.943	.957
.4	-	-	-	.948	.943	.951	.950	**.936**	.951
.5	-	-	-	-	.953	.941	.951	.952	.943
.6	-	-	-	-	-	.944	.941	.954	.945
.7	-	-	-	-	-	-	.937	.950	.949
.8	-	-	-	-	-	-	-	.959	.954
.9	-	-	-	-	-	-	-	-	.954

sample size $n_5 = 1000$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.947	.958	.952	**.961**	.956	.949	**.931**	.948	.942
.2	-	.954	.947	.933	.944	.950	.950	.959	.949
.3	-	-	.948	.950	.937	.944	.949	.950	.945
.4	-	-	-	.953	.949	.948	.945	.942	.950
.5	-	-	-	-	.945	.949	.943	.944	.951
.6	-	-	-	-	-	.958	.953	.950	.943
.7	-	-	-	-	-	-	.956	.953	.944
.8	-	-	-	-	-	-	-	.957	.953
.9	-	-	-	-	-	-	-	-	.948

sample size $n_6 = 2500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.953	.949	.942	.960	.963	**.966**	.941	.958	.946
.2	-	.951	.937	**.931**	.943	.959	.955	.944	.944
.3	-	-	.956	.948	.960	.958	.948	.955	.959
.4	-	-	-	.947	.944	.952	.954	.950	.948
.5	-	-	-	-	.950	.954	.945	.952	.946
.6	-	-	-	-	-	.954	.952	.945	.956
.7	-	-	-	-	-	-	.948	.951	.956
.8	-	-	-	-	-	-	-	.944	.947
.9	-	-	-	-	-	-	-	-	.954

sample size $n_7 = 5000$

Table 13: Empirical coverage of $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$ for $i = 5, 6, 7$ under iid balanced (θ, g^b) sampling. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.962	.960	.957	.951	.962	.954	**.967**	.958	.965
.2	-	.956	.958	.952	**.940**	.961	.961	.948	.951
.3	-	-	.952	.944	.943	.961	.966	.948	.958
.4	-	-	-	.947	.942	.947	.957	.958	.946
.5	-	-	-	-	.955	.953	.947	.951	.952
.6	-	-	-	-	-	.962	.954	.960	.954
.7	-	-	-	-	-	-	.948	.959	.953
.8	-	-	-	-	-	-	-	.958	.959
.9	-	-	-	-	-	-	-	-	.961

sample size $n_1 = 100$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	**.965**	.963	.949	.952	.949	.943	.960	.960	.951
.2	-	.956	.946	.953	.950	.955	.959	.953	.953
.3	-	-	.951	.953	.955	**.934**	.953	.959	.951
.4	-	-	-	.956	.955	.957	.947	.963	.951
.5	-	-	-	-	.946	.953	.947	.956	.952
.6	-	-	-	-	-	.946	.956	.940	.948
.7	-	-	-	-	-	-	.940	.958	.950
.8	-	-	-	-	-	-	-	.951	.959
.9	-	-	-	-	-	-	-	-	.946

sample size $n_2 = 250$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.944	.948	.952	.963	.954	.951	.961	.953	.962
.2	-	.952	.945	.954	.944	.959	.952	.943	.960
.3	-	-	.946	**.933**	.960	.944	.960	.957	.949
.4	-	-	-	.944	.952	.951	.958	.950	.939
.5	-	-	-	-	.949	.951	.934	.959	.953
.6	-	-	-	-	-	.950	.937	.952	.952
.7	-	-	-	-	-	-	.941	.944	.959
.8	-	-	-	-	-	-	-	**.965**	.951
.9	-	-	-	-	-	-	-	-	.954

sample size $n_3 = 500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.941	.955	.959	.955	.942	.954	.955	.953	.949
.2	-	.951	.952	.962	.942	.948	.941	.952	.948
.3	-	-	.952	**.934**	**.962**	.948	.956	.957	.955
.4	-	-	-	.943	.950	.943	.953	.947	.953
.5	-	-	-	-	**.934**	.947	.949	.954	.944
.6	-	-	-	-	-	.947	.945	.945	.950
.7	-	-	-	-	-	-	.945	.946	.949
.8	-	-	-	-	-	-	-	.957	.957
.9	-	-	-	-	-	-	-	-	.949

sample size $n_4 = 750$

Table 14: Empirical coverage of $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$ for $i = 1, 2, 3, 4$ under iid optimal $(\theta, g^\star)$ sampling. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.948	.955	.959	.962	.962	.950	.961	.950	.955
.2	-	.953	.954	.959	.946	.940	.949	.944	.946
.3	-	-	.960	.941	.951	.945	.952	.952	.953
.4	-	-	-	.945	.947	.949	.945	.960	**.964**
.5	-	-	-	-	.955	.954	**.939**	.951	.946
.6	-	-	-	-	-	.957	.952	.949	.957
.7	-	-	-	-	-	-	.961	.959	.943
.8	-	-	-	-	-	-	-	.955	.954
.9	-	-	-	-	-	-	-	-	.954

sample size $n_5 = 1000$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.942	.945	.950	.952	.952	.951	.950	.950	.956
.2	-	.960	.935	.958	.955	**.933**	.947	.944	.952
.3	-	-	**.961**	.953	.951	.952	.948	.956	.942
.4	-	-	-	.959	.941	.956	.955	.951	.937
.5	-	-	-	-	.949	.955	.955	.948	.956
.6	-	-	-	-	-	.951	.938	.958	.946
.7	-	-	-	-	-	-	.955	.935	.949
.8	-	-	-	-	-	-	-	.951	.948
.9	-	-	-	-	-	-	-	-	.953

sample size $n_6 = 2500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.948	.957	.951	.947	.940	.956	.938	.949	.956
.2	-	.946	.954	.943	.943	.950	.949	.957	.949
.3	-	-	.955	.950	.949	.946	.956	.950	**.960**
.4	-	-	-	.949	.946	.940	.951	.941	**.939**
.5	-	-	-	-	.951	.944	.945	.957	.951
.6	-	-	-	-	-	.940	.945	.949	.947
.7	-	-	-	-	-	-	**.939**	.946	.951
.8	-	-	-	-	-	-	-	**.939**	.948
.9	-	-	-	-	-	-	-	-	.940

sample size $n_7 = 5000$

Table 15: Empirical coverage of $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$ for $i = 5, 6, 7$ under iid optimal $(\theta, g^\star)$ sampling. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	**.958**	**.958**	.950	.935	.930	.942	.941	.946	.947
.2	-	.937	.951	.938	.941	.933	.919	**.958**	.948
.3	-	-	.942	.935	.950	.928	.925	.930	.946
.4	-	-	-	.934	.937	.936	.923	.927	.945
.5	-	-	-	-	.932	.914	.928	.899	.929
.6	-	-	-	-	-	.924	.927	.905	.933
.7	-	-	-	-	-	-	.890	.895	.927
.8	-	-	-	-	-	-	-	.898	.903
.9	-	-	-	-	-	-	-	-	**.859**

sample size $n_1 = 100$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.942	**.959**	.951	.943	.941	.945	.954	.933	.943
.2	-	.936	.955	.954	.954	.950	.948	.955	.945
.3	-	-	.935	.953	.953	.951	.945	.944	.922
.4	-	-	-	.943	.958	.948	.941	.932	.932
.5	-	-	-	-	.945	.938	.940	.928	.951
.6	-	-	-	-	-	.942	.955	.926	.919
.7	-	-	-	-	-	-	.928	.942	.927
.8	-	-	-	-	-	-	-	.941	.935
.9	-	-	-	-	-	-	-	-	**.910**

sample size $n_2 = 250$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.960	.956	.948	.941	.944	.944	.946	.939	.931
.2	-	.952	.950	**.969**	.953	.958	.953	.954	.959
.3	-	-	.939	.944	.952	.955	.948	.947	.950
.4	-	-	-	.946	.958	.964	.953	.932	.936
.5	-	-	-	-	.961	.950	.944	.948	.949
.6	-	-	-	-	-	.949	.958	.951	**.929**
.7	-	-	-	-	-	-	.949	.942	.944
.8	-	-	-	-	-	-	-	.946	.943
.9	-	-	-	-	-	-	-	-	.934

sample size $n_3 = 750$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.959	.938	.951	.944	.952	.945	.934	**.932**	.954
.2	-	.947	.957	.962	.950	.954	.959	.955	.942
.3	-	-	.940	.946	.955	.958	.958	.938	.938
.4	-	-	-	.950	.954	**.963**	.947	.936	.953
.5	-	-	-	-	.958	.949	.937	.946	.948
.6	-	-	-	-	-	.949	.962	.948	.938
.7	-	-	-	-	-	-	.943	.950	.945
.8	-	-	-	-	-	-	-	.939	.946
.9	-	-	-	-	-	-	-	-	.950

sample size $n_4 = 750$

Table 16: Empirical coverage of $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$ for $i = 1, 2, 3, 4$ under adaptive $(\theta, \mathbf{g}_n^\star)$ sampling. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	**.961**	.952	.946	.946	.950	.945	.935	.937	.953
.2	-	.956	.948	.954	.949	.958	.961	.952	.952
.3	-	-	.945	.955	**.961**	.945	.952	.949	.946
.4	-	-	-	.956	**.961**	.953	.946	.946	.950
.5	-	-	-	-	.958	.952	.941	.946	.950
.6	-	-	-	-	-	.945	.954	.956	.946
.7	-	-	-	-	-	-	.954	.958	.954
.8	-	-	-	-	-	-	-	.935	.949
.9	-	-	-	-	-	-	-	-	**.934**

sample size $n_5 = 1000$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.950	.943	.937	**.967**	.946	.959	**.936**	.943	.952
.2	-	.955	.948	.949	.961	.950	.955	.956	.947
.3	-	-	.951	.949	.949	.955	.952	.950	.946
.4	-	-	-	.945	.951	.963	.941	.945	.947
.5	-	-	-	-	.963	.955	.952	.948	.951
.6	-	-	-	-	-	.965	.938	.945	.947
.7	-	-	-	-	-	-	.938	.950	.954
.8	-	-	-	-	-	-	-	.944	.953
.9	-	-	-	-	-	-	-	-	.955

sample size $n_6 = 2500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.952	.947	.945	.954	.948	.952	.943	.936	.937
.2	-	.951	.951	.947	**.959**	.947	.955	.945	.956
.3	-	-	.937	.944	.949	.955	.938	.952	.956
.4	-	-	-	.953	.940	.959	.946	.953	.941
.5	-	-	-	-	.957	.946	.943	**.933**	.944
.6	-	-	-	-	-	.954	.946	.941	.951
.7	-	-	-	-	-	-	.950	.949	.944
.8	-	-	-	-	-	-	-	.950	.949
.9	-	-	-	-	-	-	-	-	.954

sample size $n_7 = 5000$

Table 17: Empirical coverage of $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$ for $i = 5, 6, 7$ under adaptive $(\theta, \mathbf{g}_n^\star)$ sampling. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.938	.952	.950	.939	.931	.924	.953	.946	.949
.2	-	.941	.947	.940	.927	.920	.897	.930	**.956**
.3	-	-	.954	.925	.933	.900	.889	.912	.948
.4	-	-	-	.944	.932	.922	.885	.880	.929
.5	-	-	-	-	.926	.929	.887	.889	.910
.6	-	-	-	-	-	.921	.897	.873	.919
.7	-	-	-	-	-	-	.907	.897	.913
.8	-	-	-	-	-	-	-	.866	.868
.9	-	-	-	-	-	-	-	-	**.811**

sample size $n_1 = 100$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.943	.951	.949	**.959**	.931	.933	.930	**.915**	.955
.2	-	.957	.952	.943	.941	.949	.918	.922	.934
.3	-	-	.958	.939	**.959**	.951	.940	.935	.924
.4	-	-	-	.949	.956	.955	.944	.934	.926
.5	-	-	-	-	.949	.949	.941	.934	.915
.6	-	-	-	-	-	.940	.945	.937	.923
.7	-	-	-	-	-	-	.951	.943	.925
.8	-	-	-	-	-	-	-	.942	.937
.9	-	-	-	-	-	-	-	-	.918

sample size $n_2 = 250$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	**.963**	.957	.951	.954	.952	.953	.944	.942	.948
.2	-	.955	.951	.954	.959	.941	.956	.939	.941
.3	-	-	.953	.940	.953	.946	.948	.934	.933
.4	-	-	-	.959	.954	.946	.951	.945	.946
.5	-	-	-	-	.958	.959	.938	.954	**.932**
.6	-	-	-	-	-	.952	.942	.936	.941
.7	-	-	-	-	-	-	.956	.942	.941
.8	-	-	-	-	-	-	-	.953	.944
.9	-	-	-	-	-	-	-	-	.935

sample size $n_3 = 750$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.955	.950	.941	.951	.949	.949	.933	.945	.948
.2	-	**.966**	.958	.960	.942	.944	.955	.938	.950
.3	-	-	.947	.958	.941	.950	.941	.936	.947
.4	-	-	-	.949	.948	.953	.949	.948	.947
.5	-	-	-	-	.948	.957	.952	.945	**.932**
.6	-	-	-	-	-	.955	.947	.935	.938
.7	-	-	-	-	-	-	.955	.944	.936
.8	-	-	-	-	-	-	-	.943	.946
.9	-	-	-	-	-	-	-	-	.937

sample size $n_4 = 750$

Table 18: Empirical coverage of $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$ for $i = 1, 2, 3, 4$ under adaptive $(\theta, \mathbf{g}_n^a)$ sampling. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.957	.952	.940	.959	.954	.954	.946	.941	.940
.2	-	.952	**.961**	.955	.947	.948	.956	.945	.950
.3	-	-	.956	.953	.949	.943	.939	.941	.947
.4	-	-	-	.960	.954	.947	.945	.951	.946
.5	-	-	-	-	.953	.955	.951	.951	.949
.6	-	-	-	-	-	.956	.957	.951	.955
.7	-	-	-	-	-	-	.954	.939	.938
.8	-	-	-	-	-	-	-	**.935**	.948
.9	-	-	-	-	-	-	-	-	.940

sample size $n_5 = 1000$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.953	.948	.950	.950	.955	.955	.950	.954	.942
.2	-	.953	.956	.948	.953	.944	.955	.944	**.967**
.3	-	-	.951	.959	.949	.962	.949	.957	.955
.4	-	-	-	.953	.948	.953	.965	.959	.961
.5	-	-	-	-	.948	.949	.953	.954	.952
.6	-	-	-	-	-	**.934**	.953	.946	.954
.7	-	-	-	-	-	-	.959	.940	.948
.8	-	-	-	-	-	-	-	.947	.948
.9	-	-	-	-	-	-	-	-	.941

sample size $n_6 = 2500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.962	.955	.947	.950	.949	.953	**.937**	.943	.956
.2	-	.953	.960	.947	.952	.952	.954	.942	.965
.3	-	-	.947	**.969**	.937	.951	.942	.943	.956
.4	-	-	-	.964	.954	.941	.950	.953	.952
.5	-	-	-	-	.949	.949	.944	.959	.946
.6	-	-	-	-	-	.952	.946	.945	.954
.7	-	-	-	-	-	-	.955	.949	.940
.8	-	-	-	-	-	-	-	.953	.949
.9	-	-	-	-	-	-	-	-	.956

sample size $n_7 = 5000$

Table 19: Empirical coverage of $\mathcal{I}(\theta)_{n_i,m}, m = 1, \ldots, M$ for $i = 5, 6, 7$ under adaptive $(\theta, \mathbf{g}_n^a)$ sampling. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	11±25	10±25	13±28	16±32	21±32	26±36	27±38	32±39	42±45
.2	-	5±15	5±16	7±16	9±18	12±19	17±24	21±24	27±26
.3	-	-	3±13	3±12	5±13	7±15	10±17	14±19	21±20
.4	-	-	-	3±12	2±11	3±12	6±13	10±15	16±17
.5	-	-	-	-	2±11	2±11	4±12	7±13	13±15
.6	-	-	-	-	-	1±11	1±11	3±12	9±16
.7	-	-	-	-	-	-	1±11	1±12	6±15
.8	-	-	-	-	-	-	-	1±13	2±15
.9	-	-	-	-	-	-	-	-	-1±17

sample size $n_1 = 100$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	4±14	5±13	9±16	11±15	15±17	19±19	22±20	26±21	33±24
.2	-	2±9	3±9	4±9	7±11	10±11	14±13	19±13	25±15
.3	-	-	1±7	1±7	3±8	6±9	9±10	14±11	20±12
.4	-	-	-	1±7	1±7	3±7	5±8	9±9	15±10
.5	-	-	-	-	1±6	1±7	2±7	6±8	13±10
.6	-	-	-	-	-	0±7	1±7	3±8	9±9
.7	-	-	-	-	-	-	0±7	1±8	5±9
.8	-	-	-	-	-	-	-	0±8	2±9
.9	-	-	-	-	-	-	-	-	0±11

sample size $n_2 = 250$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	2±8	3±8	7±10	9±10	13±11	17±13	20±14	24±14	31±15
.2	-	1±6	2±6	3±6	6±7	9±8	13±9	18±9	24±10
.3	-	-	0±5	1±5	3±5	5±6	8±7	13±8	20±9
.4	-	-	-	1±5	1±5	2±5	5±6	9±7	15±8
.5	-	-	-	-	0±4	1±5	2±5	6±6	13±7
.6	-	-	-	-	-	0±5	1±5	3±5	9±6
.7	-	-	-	-	-	-	0±5	1±6	5±6
.8	-	-	-	-	-	-	-	0±6	2±7
.9	-	-	-	-	-	-	-	-	0±8

sample size $n_3 = 750$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1±6	3±7	7±8	9±9	12±9	17±10	20±11	24±11	30±12
.2	-	1±5	1±5	3±5	6±6	9±6	13±7	17±7	23±8
.3	-	-	0±4	1±4	2±4	5±5	8±5	13±6	19±7
.4	-	-	-	0±4	1±4	2±4	5±4	9±5	15±6
.5	-	-	-	-	0±4	1±4	2±4	6±5	12±6
.6	-	-	-	-	-	0±4	1±4	3±4	9±5
.7	-	-	-	-	-	-	0±4	1±5	5±5
.8	-	-	-	-	-	-	-	0±5	2±5
.9	-	-	-	-	-	-	-	-	0±6

sample size $n_4 = 750$

Table 20: Empirical means and standard deviations (rounded to the nearest integers) of $\{100 \times r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ for $i = 1, 2, 3, 4$ under iid balanced (θ, g^b) sampling.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1±6	3±6	6±7	9±7	12±7	16±8	19±9	24±9	29±10
.2	-	0±4	1±4	3±4	6±5	9±6	13±6	17±6	23±7
.3	-	-	0±3	1±4	2±4	5±4	8±5	13±5	19±6
.4	-	-	-	0±3	1±3	2±4	5±4	9±5	15±5
.5	-	-	-	-	0±3	1±3	2±4	6±4	12±5
.6	-	-	-	-	-	0±3	1±4	3±4	8±4
.7	-	-	-	-	-	-	0±3	1±4	5±4
.8	-	-	-	-	-	-	-	0±4	2±5
.9	-	-	-	-	-	-	-	-	0±5

sample size $n_5 = 1000$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±3	2±4	6±4	9±4	12±5	15±5	19±6	23±6	28±6
.2	-	0±2	1±3	3±3	6±3	9±3	12±4	17±4	23±4
.3	-	-	0±2	1±2	2±2	5±3	8±3	13±3	19±4
.4	-	-	-	0±2	1±2	2±2	5±3	9±3	15±3
.5	-	-	-	-	0±2	0±2	2±2	5±3	12±3
.6	-	-	-	-	-	0±2	1±2	3±2	8±3
.7	-	-	-	-	-	-	0±2	1±2	5±3
.8	-	-	-	-	-	-	-	0±3	2±3
.9	-	-	-	-	-	-	-	-	0±3

sample size $n_6 = 2500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±2	2±3	5±3	9±3	12±3	15±4	19±4	23±4	28±4
.2	-	0±2	1±2	3±2	6±2	9±2	12±3	17±3	23±3
.3	-	-	0±1	1±2	2±2	5±2	8±2	12±2	19±3
.4	-	-	-	0±1	1±1	2±2	5±2	8±2	15±2
.5	-	-	-	-	0±1	1±1	2±2	5±2	12±2
.6	-	-	-	-	-	0±1	1±1	3±2	8±2
.7	-	-	-	-	-	-	0±2	1±2	5±2
.8	-	-	-	-	-	-	-	0±2	2±2
.9	-	-	-	-	-	-	-	-	0±2

sample size $n_7 = 5000$

Table 21: Empirical means and standard deviations (rounded to the nearest integers) of $\{100 \times r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ for $i = 5, 6, 7$ under iid balanced (θ, g^b) sampling.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	9±25	7±22	7±22	7±21	6±20	7±23	5±20	4±20	5±21
.2	-	5±15	4±14	4±14	3±14	2±13	3±13	3±14	2±14
.3	-	-	3±13	3±11	2±12	2±11	2±12	2±12	3±14
.4	-	-	-	2±11	2±10	2±11	1±10	2±11	1±13
.5	-	-	-	-	1±10	1±10	1±11	1±11	1±13
.6	-	-	-	-	-	1±10	1±11	1±11	1±13
.7	-	-	-	-	-	-	1±12	1±12	1±13
.8	-	-	-	-	-	-	-	1±13	0±15
.9	-	-	-	-	-	-	-	-	-1±17

sample size $n_1 = 100$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	4±13	2±11	3±11	2±11	2±11	2±11	1±10	2±11	2±12
.2	-	2±8	2±8	2±8	1±8	1±8	1±8	1±8	1±9
.3	-	-	1±7	1±7	1±7	1±7	1±7	1±7	1±8
.4	-	-	-	1±7	1±6	1±7	0±7	1±7	1±8
.5	-	-	-	-	0±6	0±6	0±7	0±7	0±8
.6	-	-	-	-	-	1±6	0±7	0±7	0±8
.7	-	-	-	-	-	-	0±7	0±7	0±8
.8	-	-	-	-	-	-	-	0±8	0±9
.9	-	-	-	-	-	-	-	-	0±10

sample size $n_2 = 250$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	2±8	2±7	2±7	1±7	1±7	1±7	0±7	1±8	1±8
.2	-	1±6	1±5	1±5	1±5	1±6	1±6	1±6	1±6
.3	-	-	1±5	0±5	1±5	0±5	0±5	0±5	0±6
.4	-	-	-	0±5	0±4	0±5	0±5	0±5	0±6
.5	-	-	-	-	0±5	0±5	0±5	0±5	0±6
.6	-	-	-	-	-	0±4	0±5	0±5	0±6
.7	-	-	-	-	-	-	0±5	0±5	0±6
.8	-	-	-	-	-	-	-	0±6	0±6
.9	-	-	-	-	-	-	-	-	0±7

sample size $n_3 = 750$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1±6	1±6	1±6	1±6	0±6	1±6	0±6	1±6	1±6
.2	-	1±5	0±4	1±4	0±4	0±4	0±5	0±5	0±5
.3	-	-	1±4	0±4	0±4	0±4	0±4	0±4	0±5
.4	-	-	-	0±4	0±4	0±4	0±4	0±4	0±4
.5	-	-	-	-	0±4	0±4	0±4	0±4	0±5
.6	-	-	-	-	-	0±4	0±4	0±4	0±5
.7	-	-	-	-	-	-	0±4	0±4	0±5
.8	-	-	-	-	-	-	-	0±5	0±5
.9	-	-	-	-	-	-	-	-	0±6

sample size $n_4 = 750$

Table 22: Empirical means and standard deviations (rounded to the nearest integers) of $\{100 \times r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ for $i = 1, 2, 3, 4$ under iid optimal $(\theta, g^\star)$ sampling.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1±5	1±5	1±5	0±5	0±5	1±5	0±5	0±5	1±5
.2	-	0±4	0±4	0±4	0±4	0±4	0±4	0±4	0±4
.3	-	-	0±3	0±3	0±3	0±4	0±3	0±4	0±4
.4	-	-	-	0±3	0±3	0±3	0±3	0±3	0±4
.5	-	-	-	-	0±3	0±3	0±3	0±3	0±4
.6	-	-	-	-	-	0±3	0±3	0±3	0±4
.7	-	-	-	-	-	-	0±3	0±4	0±4
.8	-	-	-	-	-	-	-	0±4	0±4
.9	-	-	-	-	-	-	-	-	0±5

sample size $n_5 = 1000$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±4	0±3	0±3	0±3	0±3	0±3	0±3	0±3	0±3
.2	-	0±2	0±2	0±2	0±2	0±2	0±2	0±3	0±3
.3	-	-	0±2	0±2	0±2	0±2	0±2	0±2	0±2
.4	-	-	-	0±2	0±2	0±2	0±2	0±2	0±2
.5	-	-	-	-	0±2	0±2	0±2	0±2	0±2
.6	-	-	-	-	-	0±2	0±2	0±2	0±3
.7	-	-	-	-	-	-	0±2	0±2	0±3
.8	-	-	-	-	-	-	-	0±2	0±3
.9	-	-	-	-	-	-	-	-	0±3

sample size $n_6 = 2500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2
.2	-	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2
.3	-	-	0±2	0±2	0±1	0±2	0±2	0±2	0±2
.4	-	-	-	0±1	0±1	0±1	0±1	0±2	0±2
.5	-	-	-	-	0±1	0±1	0±1	0±2	0±2
.6	-	-	-	-	-	0±1	0±1	0±2	0±2
.7	-	-	-	-	-	-	0±2	0±2	0±2
.8	-	-	-	-	-	-	-	0±2	0±2
.9	-	-	-	-	-	-	-	-	0±2

sample size $n_7 = 5000$

Table 23: Empirical means and standard deviations (rounded to the nearest integers) of $\{100 \times r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ for $i = 5, 6, 7$ under iid optimal $(\theta, g^\star)$ sampling.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	6±23	4±18	3±18	2±17	2±18	3±20	1±19	1±20	1±20
.2	-	2±14	2±13	1±12	1±14	0±14	-1±14	-1±14	-2±15
.3	-	-	1±11	1±12	0±12	0±12	-2±12	-3±13	-4±13
.4	-	-	-	1±11	0±11	0±11	-1±12	-2±13	-4±14
.5	-	-	-	-	0±11	-1±11	-1±12	-3±13	-4±14
.6	-	-	-	-	-	-1±11	-2±12	-2±13	-4±15
.7	-	-	-	-	-	-	-2±13	-3±14	-5±16
.8	-	-	-	-	-	-	-	-3±15	-5±17
.9	-	-	-	-	-	-	-	-	-8±20

sample size $n_1 = 100$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	2±12	1±10	1±10	1±10	1±10	1±10	0±10	0±11	-1±11
.2	-	0±8	1±8	0±7	0±8	0±8	0±8	0±9	-2±9
.3	-	-	0±7	0±7	1±7	0±7	-1±7	-1±8	-2±9
.4	-	-	-	0±7	0±7	0±6	-1±7	-1±8	-2±9
.5	-	-	-	-	-1±7	0±6	0±7	-1±8	-2±9
.6	-	-	-	-	-	-1±7	-1±7	-1±7	-2±9
.7	-	-	-	-	-	-	-1±7	0±7	-2±9
.8	-	-	-	-	-	-	-	-1±8	-2±10
.9	-	-	-	-	-	-	-	-	-3±11

sample size $n_2 = 250$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1±8	1±7	1±7	1±7	0±7	0±7	0±7	0±8	-1±8
.2	-	0±6	0±6	0±5	0±5	0±5	0±6	0±6	-1±6
.3	-	-	0±5	0±5	0±5	0±5	0±5	0±5	-1±6
.4	-	-	-	0±5	0±5	0±4	0±5	0±5	-1±6
.5	-	-	-	-	0±5	0±4	0±5	-1±5	-1±6
.6	-	-	-	-	-	0±5	0±5	0±5	-1±6
.7	-	-	-	-	-	-	0±5	0±5	-1±6
.8	-	-	-	-	-	-	-	-1±6	0±7
.9	-	-	-	-	-	-	-	-	-1±8

sample size $n_3 = 750$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1±6	1±6	1±6	0±6	0±6	0±6	0±6	0±6	-1±6
.2	-	0±5	0±5	0±4	0±4	0±4	0±4	0±5	-1±5
.3	-	-	0±4	0±4	0±4	0±4	0±4	0±4	-1±5
.4	-	-	-	0±4	0±4	0±4	0±4	0±4	0±5
.5	-	-	-	-	0±4	0±4	0±4	-1±4	0±5
.6	-	-	-	-	-	0±4	0±4	0±4	0±5
.7	-	-	-	-	-	-	0±4	0±4	0±5
.8	-	-	-	-	-	-	-	0±5	0±5
.9	-	-	-	-	-	-	-	-	-1±6

sample size $n_4 = 750$

Table 24: Empirical means and standard deviations (rounded to the nearest integers) of $\{100 \times r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ for $i = 1, 2, 3, 4$ under adaptive $(\theta, \mathbf{g}_n^\star)$ sampling.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±5	0±5	0±5	0±5	0±5	0±5	0±5	0±5	0±5
.2	-	0±4	0±4	0±4	0±4	0±4	0±4	0±4	-1±4
.3	-	-	0±3	0±3	0±3	0±3	0±3	0±4	0±4
.4	-	-	-	0±3	0±3	0±3	0±3	0±4	0±4
.5	-	-	-	-	0±3	0±3	0±3	0±4	0±4
.6	-	-	-	-	-	0±3	0±3	0±3	0±4
.7	-	-	-	-	-	-	0±3	0±4	0±4
.8	-	-	-	-	-	-	-	0±4	0±5
.9	-	-	-	-	-	-	-	-	-1±5

sample size $n_5 = 1000$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±3	0±3	0±3	0±3	0±3	0±3	0±3	0±3	0±3
.2	-	0±3	0±2	0±2	0±2	0±2	0±2	0±2	0±3
.3	-	-	0±2	0±2	0±2	0±2	0±2	0±2	0±3
.4	-	-	-	0±2	0±2	0±2	0±2	0±2	0±2
.5	-	-	-	-	0±2	0±2	0±2	0±2	0±2
.6	-	-	-	-	-	0±2	0±2	0±2	0±2
.7	-	-	-	-	-	-	0±2	0±2	0±3
.8	-	-	-	-	-	-	-	0±2	0±3
.9	-	-	-	-	-	-	-	-	0±3

sample size $n_6 = 2500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2
.2	-	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2
.3	-	-	0±2	0±2	0±1	0±1	0±2	0±2	0±2
.4	-	-	-	0±1	0±1	0±1	0±1	0±2	0±2
.5	-	-	-	-	0±1	0±1	0±1	0±2	0±2
.6	-	-	-	-	-	0±1	0±1	0±2	0±2
.7	-	-	-	-	-	-	0±2	0±2	0±2
.8	-	-	-	-	-	-	-	0±2	0±2
.9	-	-	-	-	-	-	-	-	0±2

sample size $n_7 = 5000$

Table 25: Empirical means and standard deviations (rounded to the nearest integers) of $\{100 \times r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ for $i = 5, 6, 7$ under adaptive $(\theta, \mathbf{g}_n^\star)$ sampling.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	6±22	4±19	4±18	2±17	3±18	2±18	0±17	-1±20	0±21
.2	-	2±14	2±13	1±13	0±14	0±13	-1±15	-3±15	-4±15
.3	-	-	1±12	1±12	0±12	-1±13	-2±14	-4±14	-6±13
.4	-	-	-	0±11	-1±11	-1±12	-2±13	-4±14	-6±14
.5	-	-	-	-	0±11	-1±11	-2±13	-4±14	-7±15
.6	-	-	-	-	-	-1±11	-2±12	-3±14	-7±15
.7	-	-	-	-	-	-	-3±13	-3±14	-7±17
.8	-	-	-	-	-	-	-	-5±16	-7±18
.9	-	-	-	-	-	-	-	-	-9±21

sample size $n_1 = 100$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	2±11	1±10	1±10	1±10	1±10	0±11	0±11	-1±11	-2±12
.2	-	1±8	1±7	0±7	0±7	0±8	0±8	-1±9	-3±10
.3	-	-	1±7	0±7	0±7	0±7	0±7	0±8	-3±9
.4	-	-	-	0±6	0±6	0±6	0±7	-1±8	-3±9
.5	-	-	-	-	0±6	0±7	0±7	0±7	-2±9
.6	-	-	-	-	-	0±7	-1±7	-1±8	-2±9
.7	-	-	-	-	-	-	-1±7	-1±8	-2±10
.8	-	-	-	-	-	-	-	-1±8	-2±10
.9	-	-	-	-	-	-	-	-	-2±12

sample size $n_2 = 250$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1±8	1±7	0±7	0±7	1±7	1±7	0±7	0±8	-1±8
.2	-	1±6	0±5	0±5	0±5	0±5	0±5	0±6	-1±7
.3	-	-	0±5	0±5	1±5	0±5	0±5	0±5	-1±6
.4	-	-	-	0±4	0±4	0±5	0±5	0±5	-1±6
.5	-	-	-	-	0±5	0±5	0±5	0±5	-1±6
.6	-	-	-	-	-	0±5	0±5	0±5	-1±6
.7	-	-	-	-	-	-	0±5	0±5	-1±6
.8	-	-	-	-	-	-	-	0±6	-1±7
.9	-	-	-	-	-	-	-	-	-1±8

sample size $n_3 = 750$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1±6	0±6	0±6	0±6	1±6	0±6	0±6	0±6	0±7
.2	-	0±5	0±4	0±4	0±4	0±4	0±4	0±5	-1±5
.3	-	-	0±4	0±4	0±4	0±4	0±4	0±4	-1±5
.4	-	-	-	0±4	0±4	0±4	0±4	0±4	0±4
.5	-	-	-	-	0±4	0±4	0±4	0±4	-1±5
.6	-	-	-	-	-	0±4	0±4	0±4	0±5
.7	-	-	-	-	-	-	0±4	0±4	0±5
.8	-	-	-	-	-	-	-	0±5	-1±5
.9	-	-	-	-	-	-	-	-	0±6

sample size $n_4 = 750$

Table 26: Empirical means and standard deviations (rounded to the nearest integers) of $\{100 \times r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ for $i = 1, 2, 3, 4$ under adaptive $(\theta, \mathbf{g}_n^a)$ sampling.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±5	0±5	0±5	0±5	1±5	0±5	0±5	0±5	0±6
.2	-	0±4	0±4	0±4	0±4	0±4	0±4	0±4	0±4
.3	-	-	0±3	0±3	0±3	0±3	0±4	0±4	-1±4
.4	-	-	-	0±3	0±3	0±3	0±3	0±3	0±4
.5	-	-	-	-	0±3	0±3	0±3	0±4	0±4
.6	-	-	-	-	-	0±3	0±3	0±4	0±4
.7	-	-	-	-	-	-	0±4	0±4	0±4
.8	-	-	-	-	-	-	-	0±4	0±5
.9	-	-	-	-	-	-	-	-	0±5

sample size $n_5 = 1000$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±3	0±3	0±3	0±3	0±3	0±3	0±3	0±3	0±3
.2	-	0±2	0±2	0±2	0±2	0±2	0±2	0±3	0±3
.3	-	-	0±2	0±2	0±2	0±2	0±2	0±2	0±2
.4	-	-	-	0±2	0±2	0±2	0±2	0±2	0±2
.5	-	-	-	-	0±2	0±2	0±2	0±2	0±2
.6	-	-	-	-	-	0±2	0±2	0±2	0±2
.7	-	-	-	-	-	-	0±2	0±2	0±3
.8	-	-	-	-	-	-	-	0±3	0±3
.9	-	-	-	-	-	-	-	-	0±3

sample size $n_6 = 2500$

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2
.2	-	0±2	0±2	0±2	0±2	0±2	0±2	0±2	0±2
.3	-	-	0±2	0±1	0±1	0±2	0±2	0±2	0±2
.4	-	-	-	0±1	0±1	0±1	0±1	0±2	0±2
.5	-	-	-	-	0±1	0±1	0±1	0±2	0±2
.6	-	-	-	-	-	0±1	0±1	0±2	0±2
.7	-	-	-	-	-	-	0±2	0±2	0±2
.8	-	-	-	-	-	-	-	0±2	0±2
.9	-	-	-	-	-	-	-	-	0±2

sample size $n_7 = 5000$

Table 27: Empirical means and standard deviations (rounded to the nearest integers) of $\{100 \times r(\theta)_{n_i,m} : m = 1, \ldots, M\}$ for $i = 5, 6, 7$ under adaptive $(\theta, \mathbf{g}_n^a)$ sampling.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.033	.060	.050	.043	**.028**	.042	-	-	-
.2	-	.054	.061	.042	.045	.059	.039	-	-
.3	-	-	.057	.051	.045	.039	.049	**.061**	-
.4	-	-	-	.046	.046	.051	.037	.055	.059
.5	-	-	-	-	.044	.047	.054	.038	.043
.6	-	-	-	-	-	.052	.052	.056	.029
.7	-	-	-	-	-	-	.046	.048	.058
.8	-	-	-	-	-	-	-	.050	.043
.9	-	-	-	-	-	-	-	-	.042

iid balanced sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.056	.046	.042	.062	.054	.037	-	-	-
.2	-	.057	.058	.043	.046	.049	.053	-	-
.3	-	-	.060	.058	.058	.057	.037	.047	-
.4	-	-	-	.048	.050	.060	.048	.052	**.036**
.5	-	-	-	-	.052	.048	.061	.063	.047
.6	-	-	-	-	-	.050	**.064**	.054	.045
.7	-	-	-	-	-	-	.053	.045	.048
.8	-	-	-	-	-	-	-	.047	.055
.9	-	-	-	-	-	-	-	-	.048

$(\theta, \mathbf{g}_n^\star)$ adaptive sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.048	.056	.048	.045	.049	.040	-	-	-
.2	-	.050	.044	.055	.049	.052	.051	-	-
.3	-	-	.044	**.037**	.050	.050	.052	.043	-
.4	-	-	-	.053	.040	.051	.057	.055	.055
.5	-	-	-	-	.046	.049	**.062**	**.062**	.038
.6	-	-	-	-	-	.052	.049	.053	**.037**
.7	-	-	-	-	-	-	.051	.048	.046
.8	-	-	-	-	-	-	-	.043	.046
.9	-	-	-	-	-	-	-	-	.046

iid optimal sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.045	.041	.041	.050	.049	.047	-	-	-
.2	-	.045	.045	.065	.044	.055	.054	-	-
.3	-	-	.046	**.034**	.047	.056	.045	.044	-
.4	-	-	-	.043	.044	.055	.050	.058	.040
.5	-	-	-	-	.053	.062	.062	.045	.062
.6	-	-	-	-	-	.046	.057	.049	.069
.7	-	-	-	-	-	-	.048	.055	.055
.8	-	-	-	-	-	-	-	.053	.042
.9	-	-	-	-	-	-	-	-	**.078**

$(\theta, \mathbf{g}_n^a)$ adaptive sampling

Table 28: Empirical type I errors $a(\theta)$ under (top left) iid balanced (θ, g^b) sampling, (top right) iid optimal $(\theta, g^\star)$ sampling, (bottom left) adaptive $(\theta, \mathbf{g}_n^\star)$ sampling, (bottom right) adaptive $(\theta, \mathbf{g}_n^a)$ sampling for every $\theta = (\theta_0, \theta_1) \in \Theta = \Theta_0 \setminus \{(.1, .7), (.1, .8), (.1, .9), (.2, .8), (.2, .9), (.3, .9)\}$, the sequential testing procedure of "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" being powered at $\Psi(\theta + (0, \eta))$ with $\eta = 0.05$ and for asymptotic type I error $\alpha = 5\%$ and type II error $\beta = 10\%$. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.895	.896	.885	.895	**.862**	.892	-	-	-
.2	-	.905	.888	.895	.896	.907	.892	-	-
.3	-	-	.912	.902	.885	.899	.887	.895	-
.4	-	-	-	.903	.902	.897	.876	.880	.892
.5	-	-	-	-	.892	.900	.886	.888	.880
.6	-	-	-	-	-	**.913**	.885	.900	.891
.7	-	-	-	-	-	-	.908	.905	.883
.8	-	-	-	-	-	-	-	.895	.883
.9	-	-	-	-	-	-	-	-	.870

iid balanced sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.912	.897	.894	.890	.898	.896	-	-	-
.2	-	.904	.900	.893	.902	**.870**	.886	-	-
.3	-	-	.896	.911	.890	.901	**.914**	.875	-
.4	-	-	-	.905	.903	.876	.909	.901	.905
.5	-	-	-	-	.897	.908	.896	.899	.909
.6	-	-	-	-	-	.892	**.914**	.888	.908
.7	-	-	-	-	-	-	.900	.891	.893
.8	-	-	-	-	-	-	-	.899	.906
.9	-	-	-	-	-	-	-	-	.901

iid optimal sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.890	.885	.895	.883	.902	.879	-	-	-
.2	-	.895	.896	.895	.895	.893	.894	-	-
.3	-	-	.890	**.878**	.890	.904	.892	.896	-
.4	-	-	-	.903	.900	.897	.905	.898	**.916**
.5	-	-	-	-	.900	.889	.893	.900	.889
.6	-	-	-	-	-	.898	.887	.910	.905
.7	-	-	-	-	-	-	.902	.896	.898
.8	-	-	-	-	-	-	-	**.878**	.895
.9	-	-	-	-	-	-	-	-	.887

$(\theta, \mathbf{g}_n^\star)$ adaptive sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	.898	.911	.889	.898	.912	.889	-	-	-
.2	-	.911	**.913**	.882	.907	.893	**.913**	-	-
.3	-	-	.902	.895	.906	.900	.905	.895	-
.4	-	-	-	.882	.910	.886	.904	.894	.900
.5	-	-	-	-	.894	.891	**.881**	.887	.901
.6	-	-	-	-	-	**.881**	.882	.900	.894
.7	-	-	-	-	-	-	.889	.893	.911
.8	-	-	-	-	-	-	-	.893	.896
.9	-	-	-	-	-	-	-	-	.882

$(\theta, \mathbf{g}_n^a)$ adaptive sampling

Table 29: Empirical type II errors $b(\theta)$ under (top left) iid balanced $(\theta + (0,\eta), g^b)$ sampling, (top right) iid optimal $(\theta + (0,\eta), g^\star)$ sampling, (bottom left) adaptive $(\theta + (0,\eta), \mathbf{g}_n^\star)$ sampling, (bottom right) adaptive $(\theta + (0,\eta), \mathbf{g}_n^a)$ sampling for every $\theta \in \Theta = \Theta_0 \setminus \{(.1,.7),(.1,.8),(.1,.9),(.2,.8),(.2,.9),(.3,.9)\}$, the sequential testing procedure of "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" being powered at $\Psi(\theta + (0,\eta))$ with $\eta = 0.05$ and for asymptotic type I error $\alpha = 5\%$ and type II error $\beta = 10\%$. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1.240	2.950	5.284	8.270	12.341	**16.747**	-	-	-
.2	-	1.809	2.961	4.303	6.061	8.153	10.026	-	-
.3	-	-	2.208	3.108	4.093	5.104	6.360	7.644	-
.4	-	-	-	2.406	3.022	3.705	4.512	5.318	6.200
.5	-	-	-	-	2.453	2.834	3.360	3.755	4.215
.6	-	-	-	-	-	2.285	2.543	2.731	2.870
.7	-	-	-	-	-	-	1.975	2.047	2.022
.8	-	-	-	-	-	-	-	1.520	1.342
.9	-	-	-	-	-	-	-	-	**0.851**

iid balanced sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1.243	2.780	4.805	7.173	9.684	**12.535**	-	-	-
.2	-	1.758	2.852	4.219	5.444	6.759	8.224	-	-
.3	-	-	2.202	3.027	3.952	4.798	5.558	6.209	-
.4	-	-	-	2.412	3.026	3.665	4.140	4.491	4.562
.5	-	-	-	-	2.459	2.864	3.197	3.359	3.341
.6	-	-	-	-	-	2.319	2.511	2.620	2.440
.7	-	-	-	-	-	-	1.999	2.040	1.824
.8	-	-	-	-	-	-	-	1.496	1.295
.9	-	-	-	-	-	-	-	-	**0.842**

iid optimal sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1.229	2.788	4.785	7.274	9.765	**12.818**	-	-	-
.2	-	1.780	2.934	4.114	5.390	6.905	8.087	-	-
.3	-	-	2.168	3.100	3.853	4.704	5.472	6.146	-
.4	-	-	-	2.379	2.963	3.640	4.075	4.543	4.648
.5	-	-	-	-	2.439	2.853	3.172	3.379	3.341
.6	-	-	-	-	-	2.246	2.522	2.651	2.534
.7	-	-	-	-	-	-	2.006	1.987	1.808
.8	-	-	-	-	-	-	-	1.482	1.293
.9	-	-	-	-	-	-	-	-	**0.804**

$(\theta, \mathbf{g}_n^\star)$ adaptive sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1.240	2.786	4.684	7.065	9.397	**12.685**	-	-	-
.2	-	1.752	2.851	4.118	5.441	6.903	8.291	-	-
.3	-	-	2.116	3.059	3.890	4.711	5.601	6.223	-
.4	-	-	-	2.393	3.000	3.590	4.180	4.435	4.573
.5	-	-	-	-	2.481	2.876	3.194	3.375	3.371
.6	-	-	-	-	-	2.312	2.447	2.597	2.546
.7	-	-	-	-	-	-	2.009	2.024	1.860
.8	-	-	-	-	-	-	-	1.498	1.305
.9	-	-	-	-	-	-	-	-	**0.815**

$(\theta, \mathbf{g}_n^a)$ adaptive sampling

Table 30: Mean sample sizes at decision (times 10^{-3}) when checking the adequateness of the type I errors under (top left) iid balanced (θ, g^b) sampling, (top right) iid optimal $(\theta, g^\star)$ sampling, (bottom left) adaptive $(\theta, \mathbf{g}_n^\star)$ sampling, (bottom right) adaptive $(\theta, \mathbf{g}_n^a)$ sampling for every $\theta = (\theta_0, \theta_1) \in \Theta = \Theta_0 \setminus \{(.1, .7), (.1, .8), (.1, .9), (.2, .8), (.2, .9), (.3, .9)\}$, the sequential testing procedure of "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" being powered at $\Psi(\theta + (0, \eta))$ with $\eta = 0.05$ and for asymptotic type I error $\alpha = 5\%$ and type II error $\beta = 10\%$. The lowest and highest values are emphasized.

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1.102	2.944	5.678	8.871	13.234	**18.193**	-	-	-
.2	-	1.697	3.025	4.586	6.399	8.546	11.064	-	-
.3	-	-	2.065	3.110	4.226	5.466	6.780	8.392	-
.4	-	-	-	2.349	3.058	3.795	4.661	5.576	6.468
.5	-	-	-	-	2.412	2.913	3.383	3.917	4.324
.6	-	-	-	-	-	2.277	2.524	2.770	2.912
.7	-	-	-	-	-	-	1.940	1.977	1.998
.8	-	-	-	-	-	-	-	1.408	1.265
.9	-	-	-	-	-	-	-	-	**0.674**

iid balanced sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1.075	2.780	4.886	7.367	10.281	**13.220**	-	-	-
.2	-	1.679	2.882	4.220	5.651	7.231	8.553	-	-
.3	-	-	2.113	3.010	4.024	4.766	5.543	6.307	-
.4	-	-	-	2.358	3.014	3.649	4.176	4.425	4.316
.5	-	-	-	-	2.424	2.786	3.204	3.292	3.146
.6	-	-	-	-	-	2.279	2.477	2.528	2.231
.7	-	-	-	-	-	-	1.920	1.913	1.629
.8	-	-	-	-	-	-	-	1.397	1.086
.9	-	-	-	-	-	-	-	-	**0.637**

iid optimal sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1.077	2.767	4.771	7.314	10.042	**13.357**	-	-	-
.2	-	1.698	2.908	4.227	5.606	7.091	8.539	-	-
.3	-	-	2.147	3.052	3.884	4.828	5.574	6.294	-
.4	-	-	-	2.338	3.004	3.606	4.123	4.381	4.235
.5	-	-	-	-	2.406	2.856	3.076	3.289	3.075
.6	-	-	-	-	-	2.245	2.456	2.490	2.250
.7	-	-	-	-	-	-	1.882	1.896	1.541
.8	-	-	-	-	-	-	-	1.374	1.045
.9	-	-	-	-	-	-	-	-	**0.614**

$(\theta, \mathbf{g}_n^\star)$ adaptive sampling

θ	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	1.062	2.707	4.857	7.487	10.230	**13.277**	-	-	-
.2	-	1.721	2.842	4.151	5.634	7.195	8.462	-	-
.3	-	-	2.112	3.040	3.922	4.810	5.560	6.252	-
.4	-	-	-	2.350	3.063	3.596	4.049	4.483	4.288
.5	-	-	-	-	2.459	2.868	3.102	3.294	2.984
.6	-	-	-	-	-	2.217	2.461	2.478	2.194
.7	-	-	-	-	-	-	1.916	1.826	1.536
.8	-	-	-	-	-	-	-	1.387	1.050
.9	-	-	-	-	-	-	-	-	**0.611**

$(\theta, \mathbf{g}_n^a)$ adaptive sampling

Table 31: Mean sample sizes at decision (times 10^{-3}) when checking the adequateness of the type II errors under (top left) iid balanced $(\theta+(0,\eta), g^b)$ sampling, (top right) iid optimal $(\theta+(0,\eta), g^\star)$ sampling, (bottom left) adaptive $(\theta+(0,\eta), \mathbf{g}_n\star)$ sampling, (bottom right) adaptive $(\theta+(0,\eta), \mathbf{g}_n^a)$ sampling for every $\theta \in \Theta = \Theta_0 \setminus \{(.1,.7),(.1,.8),(.1,.9),(.2,.8),(.2,.9),(.3,.9)\}$, the sequential testing procedure of "$\psi = \Psi(\theta)$" against "$\psi > \Psi(\theta)$" being powered at $\Psi(\theta+(0,\eta))$ with $\eta = 0.05$ and for asymptotic type I error $\alpha = 5\%$ and type II error $\beta = 10\%$. The lowest and highest values are emphasized.