Guarantees for Comprehensive Simulation Assessment of Statistical Methods

James Yang[1,2], T. Ben Thompson[1], and Michael Sklar[1]

[1]Confirm Solutions Inc., Stanford, CA, U.S.A.
[2]Department of Statistics, Stanford University, Stanford, CA, U.S.A.

September 9, 2024

Abstract

Simulation can evaluate a statistical method for properties such as Type I Error, FDR, or bias on a grid of hypothesized parameter values. But what about the gaps between the grid-points? Continuous Simulation Extension (CSE) is a proof-by-simulation framework which can supplement simulations with (1) confidence bands valid over regions of parameter space or (2) calibration of rejection thresholds to provide rigorous proof of strong Type I Error control. CSE extends simulation estimates at grid-points into bounds over nearby space using a model shift bound related to the Renyi divergence, which we analyze for models in exponential family or canonical GLM form. CSE can work with adaptive sampling, nuisance parameters, administrative censoring, multiple arms, multiple testing, Bayesian randomization, Bayesian decision-making, and inference algorithms of arbitrary complexity. As a case study, we calibrate for strong Type I Error control a Phase II/III Bayesian selection design with 4 unknown statistical parameters. Potential applications include calibration of new statistical procedures or streamlining regulatory review of adaptive trial designs. Our open-source software implementation *imprint* is available at https://github.com/Confirm-Solutions/imprint.

1 Introduction

Modern experiment designs can reduce the cost or length of clinical trials, with approaches such as adaptive stopping, multiple testing, and Bayesian decision-making. But, regulated experiments require design properties such as Type I Error and bias to be well-established. Analyzing the properties of an ambitious design may require specialized methods and software, and regulators have historically regarded simulations as a lesser form of evidence of design quality.

This paper will demonstrate how comprehensive guarantees for a wide class of statistical procedures can be built on a simulation backbone. We provide an open-source implementation available at https://github.com/Confirm-Solutions/imprint, which can provide objective verification of the properties of compatible designs. The techniques we develop can also offer finite-sample conservative confidence intervals and performance analysis for methods including logistic modeling and parameterized survival analysis.

To introduce our setting and terminology, it may be helpful to discuss an example. In Section 10.2 we will use our method, Continuous Simulation Extension (CSE), to calibrate a complex Bayesian design. The outcome of each patient in arm i is binary, modeled as independent $Bernoulli(p_i)$, and the space of unknown statistical parameters is therefore $(p_0, p_1, p_2, p_3) \in [0, 1]^4$. This Bayesian design would be very difficult to analyze mathematically, as it has interim decisions for stopping or adaptively dropping arms and repeatedly uses Bayesian posterior calculations within a hierarchical model. Nevertheless we shall seek to analyze the Type I Error behavior of this design across the null hypothesis space, and calibrate the decision rule to achieve an overall guarantee of Type I Error control. Using CSE's Calibration procedure, we run large-scale simulations to ensure strong Type I Error control over the 4-dimensional cube $(p_0, p_1, p_2, p_3) \in [27\%, 76\%]^4$ (see Sections 7 and 10.2 and Appendix B for full details). The guarantee level is set to 2.5%. We give evidence in Section 10.2 that the calibration process likely results in a Type I Error of about 2.3%, or better

if we had run a larger simulation - but the overall 2.5% guarantee accounts for uncertainty in the calibration process itself and is valid-in-finite-samples over the full region $[23\%, 76\%]$[4].

Section 2 discusses related literature and regulatory use of simulations for Type I Error control of clinical trials.

Section 3 describes CSE and its two procedure options, Validation and Calibration.

Section 4 analyzes the Tilt-Bound and compare against alternative inequalities.

Section 5 discusses further setup and terminology for applying CSE Validation and Calibration.

Sections 6 and 7 detail the Validation and Calibration procedures, respectively.

Section 8 discusses how Calibration can offer conservative simulation-based confidence intervals.

Section 9 discusses handling adaptive designs with CSE.

Section 10 performs CSE for two case studies. In 10.1 we use Validation to analyze a Bayesian basket trial design similar to Berry et al (2013) [1]. In 10.2 we Calibrate for strong Type I Error control a complex Phase II-III four-arm Bayesian design with interim decisions based on a hierarchical model, and possible arm-dropping.

Proofs are deferred to the Appendix.

2 Background and Regulatory Context on the use of Simulation in Trial Design

As part of proposing a statistical procedure, general wisdom (and sometimes regulatory guidance [2]) dictate that the Type I Error, power, FDR, and/or bias must be analyzed. But, as stated in an FDA experience review [3], "for less well understood adaptive approaches, the major design properties (eg, type I error rate and power) often cannot be assessed through analytical derivations."

When other methods fail, simulation is a robust fallback; at the very least, it offers analysis under specific assumed scenarios. FDA guidance includes advice on the use of simulations to evaluate Type I Error and other operating characteristics of clinical trial designs [4, 5], and regulatory teams such as the Complex Innovative Design program [6] have validated designs with simulation assessment.

This recent progress comes despite a historical view that simulation could not match the guarantees of mathematical analysis. Grieve 2016 [7], citing Wang and Bretz 2010 [8], summarizes: "The view that control of the type I error is unprovable by simulation was re-iterated in a panel discussion on adaptive designs at the Third Annual FDA/DIA Statistics Forum held in Washington in 2009." Collignon et al. [9] discuss "[the regulatory perspective of the EMA as of 2018] as regards the implementation of adaptive designs in confirmatory clinical trials," and state: "A special need, justified from the context of the trial, would be required to justify a design where type I error rate control can only be ensured by means of statistical simulations, in particular where a design would be available in which type I error rate control can be ensured analytically."

Surprisingly, we show that *analytically valid proofs of Type I Error control can be constructed from a backbone of simulation.* In addition, we provide software to offer rigorous affirmative answers to several questions of regulatory interest for simulation assessment, such as:

- Have enough simulations been completed to establish the performance target?
- Is the gridding of simulation parameters adequately fine?
- Given the possibility of selection of the scenarios by motivated designers, or prior constrained optimization of the design which could move Type I Error to where it is not being checked - are the overall results sufficiently representative?

Similar questions were raised in a public FDA workshop in March 2018 [10]. CBER/CDER 2019 draft guidance on adaptive design [4] says of simulations: "In many cases, it will not be possible to estimate Type I error probability for every set of null assumptions even after taking clinical and mathematical considerations into account. It is common to perform simulations on a grid of plausible values and argue based on the totality of the evidence from the simulations that maximal Type I error probability likely does not exceed a desired level across the range covered by the grid." But what constitutes such a sufficient totality of evidence

is up to interpretation. According to the FDA Complex Innovative Design Program's 2021 progress report [11], additional simulations or scenarios were requested by regulators in 3 out of the 5 case studies considered. As Campbell (2013) states, "there is an extra regulatory burden for a Bayesian or adaptive submission at the design stage. It is simply a lot more work to review and to ask the sponsor to perform simulations for additional scenarios and in some cases the reviewer may need to reproduce the simulation results." [12] We hope that for applicable trials, CSE software can allow for objective verification of standards and streamline this process.

A common goal in stochastic simulation with unknown input parameters is to integrate the operating characteristics over the uncertainty [13], using techniques such as Bayesian analysis [14, 15], bootstrapping [16, 17], or the delta method [18, 19]. However in pivotal medical trials, rather than integrating over the uncertainty of the treatment effect(s), regulators typically request worst-case control of the Type I Error. Justifications include protection against over-optimism in the use of expert judgment or Bayesian priors; regulators' resource and human capital constraints which favor the use of simple and sweeping guarantees; bluff-prevention incentives, because the large ratio between trial costs and profits could incentivize running trials on poor drug candidates unless a minimum frequentist condition is met ([20, 21]); and idiosyncrasies of new treatments which can lead to challenges in defining appropriate prior distributions or uncertainty widths as inputs to such an integration. This is not to say that Bayesian methods are ruled out or discouraged - see [5, 22, 12].

Nonparametric estimation tools including approaches such as kriging and Gaussian Processes are commonly seen across many disciplines for analyzing simulation outputs [23, 24, 25]. In clinical trial simulations, the 'adaptr' package uses Gaussian Process modeling of the Type I Error surface to calibrate the Type I Error of adaptive and Bayesian designs.[26] These nonparametric approaches often build a probabilistic meta-model for the response surface f and can provide effective interpolations. Yet because f typically has a fixed ground truth the modeling assumptions for interval coverage may be false or unverifiable.

For the task of determining the worst-case performance, the Distributionally Robust Optimization (DRO) literature contains several works which involve simulation to assess worst-case performance of a model over confidence sets. [27] considers a climate model with uncertain multivariate Gaussian parameters μ, Σ and seeks the worst-case performance; [28] considers metrics such as Value-at-Risk using simulations and model shifts of Gaussian financial models; and [29] takes a general simulate-and-discretize approach and attempt non-convex optimization, although convergence to a global optimum may be hard to verify. In contrast to these works, our approach yields guarantees in finite samples for general experiments within GLM and canonical exponential family model classes.

Our bounding approach is particularly well-suited for statistical experiments: counter-intuitively, the presence of noise is helpful to ensure the computation is feasible. The Renyi divergence quantities required will typically be finite and and support sufficiently-sharp inequalities; rapid or non-smooth "phase changes" in the target metric will not occur over regions where the target metric depends smoothly on θ, due to the smooth sample likelihood of GLM and exponential families. These favorable conditions may not hold for simulations in other disciplines, such as deterministic physical systems.

Although we introduce new analysis to achieve our task, similar theoretical tools have long existed. For example, Pinsker's inequality can explicitly bound changes in the probabilities of sets at nearby parameter values; but Pinsker's inequality [30] is highly inefficient compared to our Tilt-Bound - see our Section 4.3. Our precursor work [31] used a conservative Taylor Expansion bound, although according to our investigation in Section 4.3 the Tilt-Bound is again superior.

The Tilt-Bound can be written in terms of Renyi divergences, and is similar to bounds in other works such as [32] Corollary 3, and [33] Theorem 9. [34] computes Renyi divergences for exponential families. [35] establishes convexity of the Renyi divergence $D_\alpha(P||Q)$ in P and Q, a result which applies to mixture distributions but not the parametric families we study. Our Theorem 2 establishes a novel quasi-convexity result for the Tilt-Bound which allows for computationally easy worst-case control over volumes of space.

3 Introduction to CSE

Superseding Chapter 5 of Sklar [31], we propose bounds for error quantification of simulation results and develop analysis guarantees covering bounded *regions* of the input parameter space. We use the term

Continuous Simulation Extension (CSE) to refer to the mathematical tools developed in in Section 4 which we use for two concrete procedures, Validation (Section 6) and Calibration (Section 7):

Validation: (Section 6) Given a fixed design, provide a lower or upper-bound estimates, $(\hat{\ell}(\cdot), \hat{u}(\cdot))$, of an operating characteristic (such as Type I Error, false-discovery rate (FDR), or bias of bounded estimator) over a (bounded) parameter space with a pointwise-valid confidence bound with confidence parameter δ. Thus, we have

$$\forall \theta \in \Theta, \mathbb{P}\left(f(\theta) \leq \hat{u}(\theta)\right) \geq 1 - \delta \text{ and } \mathbb{P}\left(f(\theta) \geq \hat{\ell}(\theta)\right) \geq 1 - \delta$$

where Θ is a bounded region of parameter space and $f(\theta)$ is the unknown operating characteristic of interest.

Calibration: (Section 7) Calibrate the critical threshold of a design to achieve a provable bound on the expected Type I Error of the selected threshold on a (bounded) parameter space. That is, we show how to select a (random) critical threshold, denoted $\hat{\lambda}^*$, such that

$$\forall \theta \in \Theta, \mathbb{E}\left[f_{\hat{\lambda}^*}(\theta)\right] \leq \alpha$$

where $f_\lambda(\theta)$ is the Type I Error of the design with critical threshold λ under parameter θ, and Θ is the null hypothesis region of parameter values.

As inputs to these basic tasks, we require (1) a modeling family for outcome generation parametrized by the unknown vector $\theta \in \Theta$, with a likelihood that is exponential family or canonical form GLM with respect to the parameter θ, (2) a protocol plan which specifies the statistical procedure to be performed to analyze the data, (3) repeated simulations of the data and procedure taken from a fine grid of parameter values θ_j for $j \in 1, \ldots, J$.

Where CSE is applicable, the guarantees of calibration and validation are finite-sample valid and conservative. However, many simulations may be required to reduce its conservative slack. Increasing the density and number of simulations typically reduces the width of validation bands and improves the average power of calibration tests.

CSE provides an alternative to mathematical analysis for deriving valid error guarantees and confidence intervals. For example, as we discuss in Section 8, due to the equivalence of statistical tests and confidence intervals, CSE calibration can reduce the problem of deriving a conservative upper confidence interval around an estimator $\hat{\theta}(X)$ to merely "guessing" a data-dependent relative width:

$$\left(-\infty, \hat{\theta}(X) + \lambda u(X)\right)$$

where λ is a tunable scaling factor which will be calibrated to correct the worst-case error in its coverage probability.

We should clarify that the space Θ of unknown parameters must consist of *statistical parameters* which are connected to the simulation outcomes through a likelihood model. We contrast statistical parameters to *design parameters* such as maximum sample sizes for the protocol, which we assume are fixed or conditioned on and thus removed as dimensions from the simulation model. Covariates could be viewed as either statistical parameters (if analyzed as part of the random model) or, preferably, as design parameters (if conditioned-on).

Fortunately, the presence of complications such as interim analyses, highly adaptive stopping, or complex algorithms may affect the number of simulations required only minimally (though a complex algorithm may naturally extend the processing time of each individual simulation). This scaling is favorable for highly adaptive or bandit-like designs, where other "brute force" methods such as dynamic programming may struggle with the large space of possible states. Intuitively, CSE is "brute forcing" the space of input parameters, but not the possible states of the simulation.

4 The Tilt-Bound for CSE

In this section we introduce a key technical tool for *Continuous Simulation Extension* (CSE). The CSE approach, described further in Section 5, uses simulation to estimate the performance over an input grid of

model parameter values, and then bounds the extent to which that performance can vary across nearby space. To perform bounding, CSE uses an inequality we call the *Tilt-Bound*. Given the family of distributions P_θ for the underlying data, a "knowledge point" θ_0, and a "target point" θ, the Tilt-Bound gives a deterministic upper bound on the change of operating characteristics such as probabilities of an arbitrary event (e.g. Type I Error or power of any design), false-discovery rate (FDR), or bias of bounded estimators assuming knowledge of the outcome metric at θ_0.

In Section 4.1, we state the guarantee of the Tilt-Bound and its properties. In Section 4.2, we provide an explicit treatment of the Tilt-Bound on the Normal location family $\{\mathcal{N}(\theta, 1) : \theta \in \Theta\}$ as a concrete example. In Section 4.3, we perform a brief numerical comparison of the quality of the Tilt-Bound against Pinsker's Inequality and the Taylor Expansion approach as in Sklar [31] to demonstrate the tightness of the Tilt-Bound. In Section 5, we outline the general workflow of applying CSE to prepare for the validation and calibration procedures in Sections 6 and 7.

4.1 Tilt-Bound and its Properties

We first state the Tilt-Bound and prove its guarantee in Theorem 1.

Theorem 1 (Tilt-Bound)**.** *Let $\{P_\theta : \theta \in \Theta\}$ denote a family of distributions with density*

$$p_\theta(x) = \exp\left[g_\theta(x) - A(\theta)\right] \tag{1}$$

for some functions $g_\theta(x) := g(\theta, x)$ and $A(\theta)$ such that $\int p_\theta(x)d\mu(x) = 1$ for some base measure μ. Denote

$$\Delta_\theta(v, x) := g_{\theta+v}(x) - g_\theta(x) \tag{2}$$

$$\psi(\theta, v, q) := \log \mathbb{E}_\theta\left[e^{q\Delta_\theta(v,X)}\right] \tag{3}$$

Suppose $F : \mathbb{R} \to [0, 1]$ is a measurable function and $f(\theta) := \mathbb{E}_\theta[F(X)]$ where $X \sim P_\theta$. Fix any "knowledge point" $\theta_0 \in \Theta$ and "target point" $\theta_0 + v \in \Theta$. Then, for any $q \in [1, \infty]$,

$$f(\theta_0 + v) \le f(\theta_0)^{1-\frac{1}{q}} \exp\left[\frac{\psi(\theta_0, v, q)}{q} - \psi(\theta_0, v, 1)\right] \tag{4}$$

The bound (4) is called the Tilt-Bound*, denoted as:*

$$U(\theta_0, v, q, a) := a^{1-\frac{1}{q}} \exp\left[\frac{\psi(\theta_0, v, q)}{q} - \psi(\theta_0, v, 1)\right] \tag{5}$$

Proof. Fix any $1 \le q \le \infty$ and define its Hölder conjugate $p := \left(1 - q^{-1}\right)^{-1}$, with the convention that $p = \infty$ if $q = 1$ and vice versa. Define

$$\ell_\theta(x) := \log p_\theta(x) = g_\theta(x) - A(\theta)$$

Fix any $\theta_0 \in \Theta$. Then, for any $\theta \in \Theta$,

$$\begin{aligned}
f(\theta) = \mathbb{E}_\theta[F(X)] &= \mathbb{E}_{\theta_0}\left[F(X)e^{\ell_\theta(X)-\ell_{\theta_0}(X)}\right] \\
&\le \|F(X)\|_{L^p(P_{\theta_0})} \left\|e^{\ell_\theta(X)-\ell_{\theta_0}(X)}\right\|_{L^q(P_{\theta_0})} \\
&= \|F(X)\|_{L^p(P_{\theta_0})} \left\|e^{g_\theta(X)-g_{\theta_0}(X)}\right\|_{L^q(P_{\theta_0})} e^{-(A(\theta)-A(\theta_0))} \\
&\le f(\theta_0)^{\frac{1}{p}} \left(\mathbb{E}_{\theta_0}\left[e^{q\left(g_\theta(X)-g_{\theta_0}(X)\right)}\right]\right)^{\frac{1}{q}} e^{-(A(\theta)-A(\theta_0))}
\end{aligned}$$

Noting that

$$\psi(\theta_0, v, 1) \equiv A(\theta_0 + v) - A(\theta_0)$$

we have that for any $v \in \Theta - \theta_0$,

$$f(\theta_0 + v) \le f(\theta_0)^{1-\frac{1}{q}} \exp\left[\frac{\psi(\theta_0, v, q)}{q} - \psi(\theta_0, v, 1)\right]$$

□

Our primary use-case of Theorem 1 is when $F(x)$ is the indicator that a design $\mathcal{D}$ rejects the null using data x, so that $f(\theta)$ is the Type I Error whenever θ is in the null-space. Theorem 1 then says that given the knowledge of the Type I Error at a parameter θ_0 and the family of distributions $\{P_\theta : \theta \in \Theta\}$, we can provide a (deterministic) upper bound of the Type I Error at a new point θ.

Remark 1 (Extension to Bounded Functions)**.** *To apply the Tilt-Bound in the case of a function F bounded in $[a, b]$, we may simply standardize the range to $[0, 1]$ with a shift and scale and apply Theorem 1 to the rescaled target $\tilde{F}(x) := \frac{F(x)-a}{b-a}$ to get*

$$\frac{f(\theta_0 + v) - a}{b - a} \leq \left(\frac{f(\theta_0) - a}{b - a}\right)^{1-\frac{1}{q}} \exp\left[\frac{\psi(\theta_0, v, q)}{q} - \psi(\theta_0, v, 1)\right]$$

Rearranging,

$$f(\theta_0 + v) \leq a + (b - a)\left(\frac{f(\theta_0) - a}{b - a}\right)^{1-\frac{1}{q}} \exp\left[\frac{\psi(\theta_0, v, q)}{q} - \psi(\theta_0, v, 1)\right]$$

Remark 2 (Extension to Lower Bound)**.** *Although the Tilt-Bound is stated as an upper bound, one can easily extend the result of Theorem 1 to construct a lower bound. Indeed, given $F : \mathbb{R} \to [0, 1]$, we may consider $\tilde{F} := 1 - F$ and apply Theorem 1 to get that*

$$1 - f(\theta_0 + v) \leq (1 - f(\theta_0))^{1-\frac{1}{q}} \exp\left[\frac{\psi(\theta_0, v, q)}{q} - \psi(\theta_0, v, 1)\right]$$

Rearranging,

$$f(\theta_0 + v) \geq 1 - (1 - f(\theta_0))^{1-\frac{1}{q}} \exp\left[\frac{\psi(\theta_0, v, q)}{q} - \psi(\theta_0, v, 1)\right]$$

Remark 3 (Optimized Tilt-Bound)**.** *The Tilt-Bound is based on a direct application of Hölder's Inequality to develop a* family of bounds *indexed by a parameter $q \in [1, \infty]$. Since the bound holds for every $q \in [1, \infty]$, it can also be minimized over q for sharper results. That is, we have that*

$$f(\theta_0 + v) \leq \inf_{q \in [1,\infty]} U(\theta_0, v, q, f(\theta_0)) \tag{6}$$

Similarly, we also have the optimal worst-case bound over a space $H \subseteq \Theta - \theta_0$

$$\sup_{v \in H} f(\theta_0 + v) \leq \inf_{q \in [1,\infty]} \sup_{v \in H} U(\theta_0, v, q, f(\theta_0)) \tag{7}$$

We refer to results (6) and (7) as the Optimized Tilt-Bound *and the* Tilewise Optimized Tilt-Bound*, respectively. These optimized Tilt-Bounds will be useful in Sections 6 and 7.*

Both (6) and (7) are optimization problems; Theorems 2 and 3 show these will be straightforward to evaluate. Indeed, Theorem 2 shows that the minimization problem in both (6) and (7) are one-dimensional quasi-convex minimization problems in general. With an additional "linearity" condition of the family of distributions, Theorem 3 shows that the Tilt-Bound is quasi-convex as a function of the displacement v as well. Therefore if H is a polytope, the supremum in (7) can be reduced to a (finite) maximum of the Tilt-Bound on the vertices of H. We state Theorems 2 and 3 and leave the proof in Appendices A.1 and A.2.

Theorem 2 (Quasi-convexity in q of the Tilt-Bound)**.** *Let $\{P_\theta : \theta \in \Theta\}$ be as in (1) and U be as in (5). Fix any $\theta_0 \in \Theta \subseteq \mathbb{R}^d$, a set $S \subseteq \mathbb{R}^d$, and $a \geq 0$. Assume that for all $v \in S$, $\Delta_{\theta_0}(v, X)$ in (2) is not constant P_{θ_0}-a.s.. Then, $W(q) := \sup_{v \in S} U(\theta_0, v, q, a)$ is quasi-convex and there exists a global minimizer $q^\star \equiv q^\star(\theta_0, v, a) \in [1, \infty]$. Moreover, if $a > 0$, S is finite, and W is not identically infinite, then $W(q)$ is strictly quasi-convex and the minimizer $q^\star$ is unique.*

Theorem 3 (Quasi-convexity in v of the Tilt-Bound)**.** *Consider the setting of Theorem 1. Fix any $\theta_0 \in \Theta$. Suppose that $\Delta(v, x) \equiv \Delta_{\theta_0}(v, x)$ in (2) is linear in v, i.e. $\Delta(v, x) = W(x)^\top v$ for some vector $W(x) \in \mathbb{R}^d$. Then, the Tilt-Bound (5) is quasi-convex as a function of v.*

Note that Theorem 1 holds generally with minimal assumptions. In particular, Theorem 1 applies to two large classes of models: exponential families, and canonical generalized linear models (when conditioning on covariates). We discuss these in Examples 1 and 2. It is worth emphasizing that Theorems 1 to 3 can apply to truly arbitrary tests. That is, rejections may be based on any parametric, non-parametric, Bayesian, or black-box methods. The Tilt-Bound will be valid in any case; in practice, the key question will be whether the parametric data generation model is sufficiently well-specified that the guarantee is relevant.

Example 1 (Exponential Family)**.** *Suppose* $\{P_\theta : \theta \in \Theta\}$ *is an exponential family with natural parameter* θ *so that the density is of the form*

$$p_\theta(x) = \exp\left[g_\theta(x) - A(\theta)\right]$$

where $g_\theta(x) = T(x)^\top \theta$*. Then, the conditions for Theorem 1 hold.*

Further, the Tilt-Bound simplifies to a simple formula. Indeed, for any $\theta_0 \in \Theta$*,*

$$\Delta(v, x) = g_{\theta_0+v}(x) - g_{\theta_0}(x) = T(x)^\top v$$

Since,

$$\psi(\theta, v, q) = \log \mathbb{E}_\theta \left[e^{q\Delta(v,X)}\right] = \log \mathbb{E}_\theta \left[e^{qT(X)^\top v}\right] = A(\theta + qv) - A(\theta)$$

the Tilt-Bound simplifies to

$$U(\theta_0, v, q, a) = a^{1-\frac{1}{q}} \exp\left[\frac{A(\theta_0 + qv) - A(\theta_0)}{q} - (A(\theta_0 + v) - A(\theta_0))\right] \tag{8}$$

Since $\Delta(v, X)$ *is not constant* P_{θ_0}*-a.s. for any* v*, we may apply Theorem 2. Moreover, since* Δ *is linear in* v*, we are in position to apply Theorem 3 as well.*

In (8), we see that the exponent term is a difference of directional secants. As difference of slopes contains Hessian information, it is intuitive to think about the exponent term as "curvature information" in P_θ*. In fact, under the (natural parameter) exponential family, the Hessian of* $A(\theta)$ *is precisely the negative of the Fisher's Information. The name "Tilt-Bound" comes from the fact that the exponent term mimics exponential-tilts in the theory of exponential family [36, 37].*

Example 2 (Canonical Generalized Linear Model (GLM))**.** *Under the canonial GLM framework, we model each response* y_i *as exponential family where the natural parameter is parameterized as* $x_i^\top \theta$ *(fixed* x_i*'s). Hence, letting* $y \in \mathbb{R}^n$ *be the vector of responses and* $X \in \mathbb{R}^{n\times d}$ *denote the matrix of covariates with each row as* $x_i^\top$*, the density of* y *is of the form*

$$p_\theta(y) = \exp\left[g_\theta(y) - A(\theta)\right]$$

with $g_\theta(y) := y^\top X\theta$*. Then, the conditions for Theorem 1 hold.*

Further, for every fixed $\theta_0 \in \Theta$*,*

$$\Delta(v, y) = g_{\theta_0+v}(y) - g_{\theta_0}(y) = y^\top Xv = T_X(y)^\top v$$

where $T_X(y) := X^\top y$*. Similar to Example 1,*

$$\psi(\theta, v, q) = \log \mathbb{E}_\theta \left[e^{q\Delta(v,X)}\right] = A(\theta + qv) - A(\theta)$$

so, the Tilt-Bound simplifies to the same form as in (8). By the same arguments as in Example 1, we are in position to apply Theorems 2 and 3.

The Tilt-Bound is closely connected to the Rényi divergence, which is defined by

$$D_\alpha(Q\|P) = \frac{1}{\alpha - 1} \log \mathbb{E}_P \left[\left(\frac{dQ}{dP}\right)^\alpha\right]$$

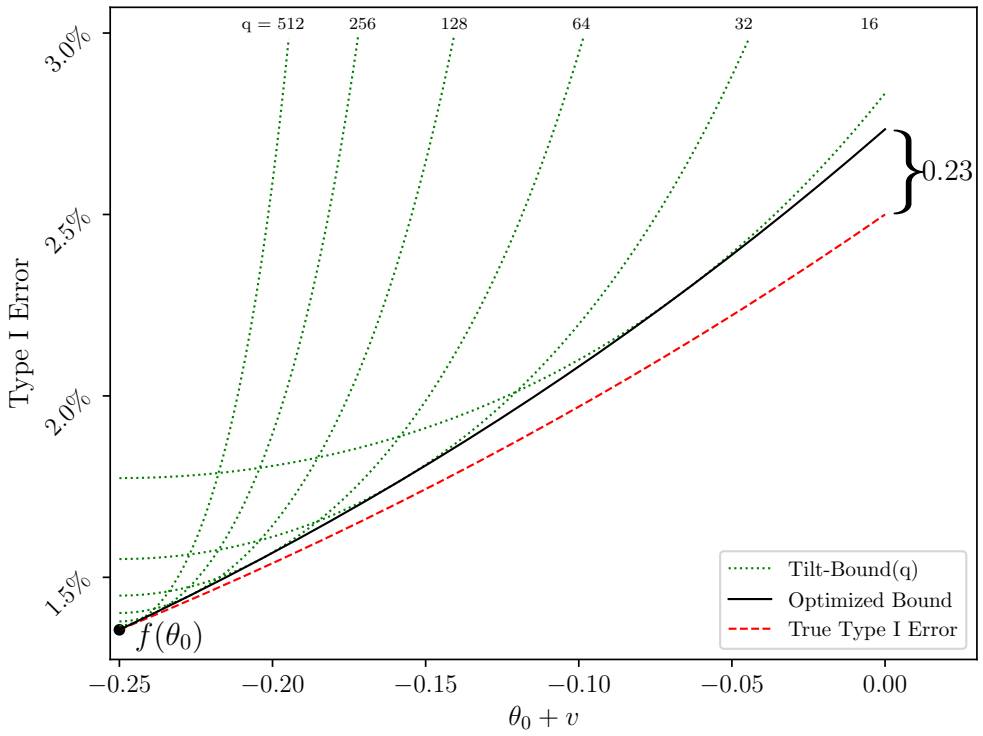


Figure 1: The normal Tilt-Bound (10) for various fixed values of q (**green** dotted lines) are overlaid with the Optimal Tilt-Bound (**black** solid line), which optimizes for q at every θ. It has been assumed that the true Type I Error at the knowledge point $\theta_0 = -0.25$ is known. As a baseline comparison, we plot the true Type I Error (**red** dotted line). In this case, the bound is only mildly conservative, although the slack grows with distance from the knowledge point.

for any $\alpha \in (0,1) \cup (1,\infty)$ [38]. The Rényi divergence can be extended to $\alpha \in \{0, 1, \infty\}$ by taking limits. In the same spirit of Cressie-Read's Power Divergence family of statistics that embeds many important goodness-of-fit statistics [39], Rényi divergence embeds many important divergences such as the Kullback-Leibler (KL) divergence [38, 40, 41]. van Erven and Harremoes [38] shows in Theorem 8 that the Rényi divergence satisfies a change-of-measure inequality that is precisely the claim in Theorem 1. Indeed, there is an explicit relationship between the Tilt-Bound and Rényi divergence:

$$U(\theta_0, v, q, a) = a^{1-\frac{1}{q}} \exp\left\{\left(1 - \frac{1}{q}\right) D_q(P_{\theta_0+v} \| P_{\theta_0})\right\}$$

However, to the best of our knowledge, we are not aware of previous results using quasi-convex analysis of the Rényi divergence as we do to extend tractable statistical guarantees (such as for Type I Error control).

4.2 Tilt-Bound on Normal Location Family

Since our primary demonstrative example, the one-sided z-test, assumes data under a Normal location family, we give an explicit treatment of the Tilt-Bound in this setting. The Normal location family is given by

$$\{\mathcal{N}(\theta, 1) : \theta \in \Theta\} \tag{9}$$

which is an exponential family with the log-partition function $A(\theta) = \frac{\theta^2}{2}$. Unless stated otherwise, the demonstrative example is the one-sided z-test for testing

$$H_0 : \theta \leq 0 \quad H_1 : \theta > 0$$

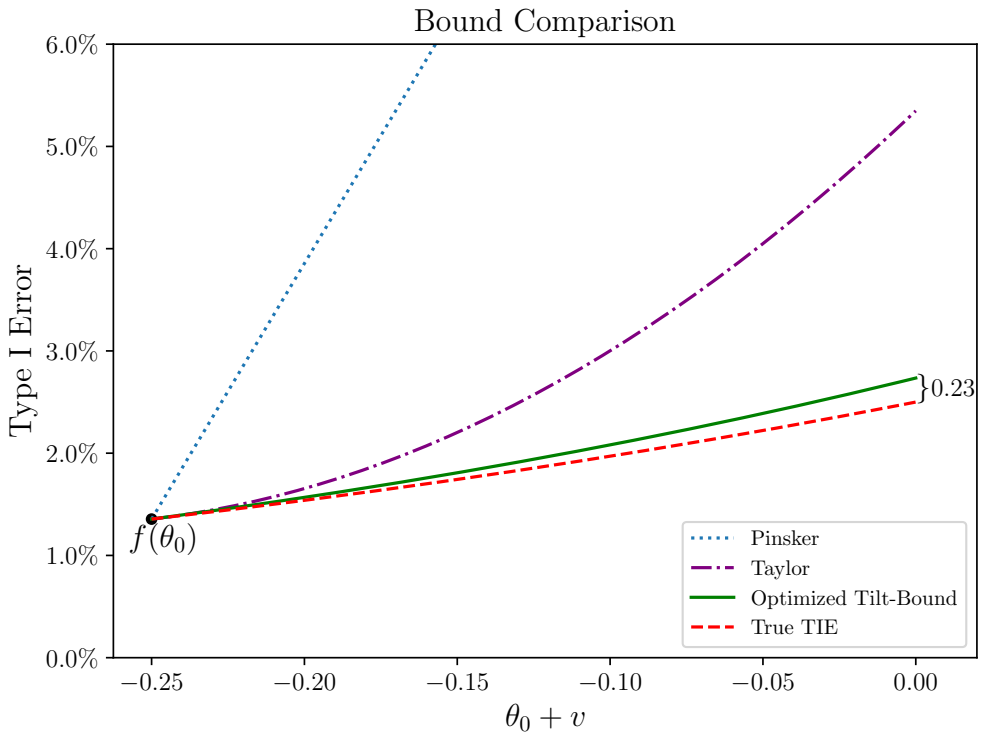


Figure 2: Comparison of methods for constructing an upper bound for the Type I Error. The true (unknown) Type I Error function is the z-test power function, $f(\theta) = \Phi\left(\theta - \Phi^{-1}(1 - 2.5\%)\right)$. To create these upper bounds from data at the origin point $\theta_0 = -0.25$, we assume exact knowledge of $f(\theta_0)$. The Taylor Bound also uses exact knowledge of $f'(\theta_0)$ and the upper bound $f''(\theta) \leq 1$ as in [31].

where $X \sim \mathcal{N}(\theta, 1)$. Recall that the one-sided z-test rejects if $X > z_{1-\alpha}$, the $1 - \alpha$ quantile of the standard normal distribution.

For the Normal location family (9), the Tilt-Bound simplifies to:

$$U(\theta_0, v, q, a) = a^{1-\frac{1}{q}} \exp\left(\frac{(q-1)v^2}{2}\right) \tag{10}$$

In Figure 1, we plot the Tilt-Bound for various fixed values of $q \in \{2^4, 2^5, \ldots, 2^9\}$ and overlay the Optimized Tilt-Bound as discussed in Remark 3. Here, we assume that the knowledge point is $\theta_0 := -0.25$ and the Type I Error of the z-test with $\alpha = 0.025$, $f(\theta_0)$, is given. We see that for any fixed value of q there is some displacement v for which the Tilt-Bound is undesirably loose, but the Optimized Tilt Bound is tight over a wide range of v. Indeed, the largest difference occurs at the furthest target point at $\theta_0 + v = 0$ with a gap of only 0.23%.

4.3 Inequality Comparison

We demonstrate the quality of the Tilt-Bound approximation by comparing with existing tools, namely Pinsker's Inequality and the Taylor approximation given in Sklar [31].

Consider the one-sided z-test as described in Section 4.2 with $\theta_0 = -0.25$ and $\alpha = 0.025$. In Figure 2, we plot the following curves:

- True Type I Error:

$$f(\theta) = \Phi\left(\theta - \Phi^{-1}(1 - \alpha)\right)$$

  where Φ is the standard normal CDF.

- The Optimized Tilt-Bound for the Normal location family (10):

$$f(\theta_0 + v) \leq \inf_{q \geq 1} U(\theta_0, v, q, f(\theta_0)) = \inf_{q \geq 1} \left\{ f(\theta_0)^{1-\frac{1}{q}} \exp\left(\frac{(q-1)v^2}{2}\right) \right\}$$

- The Taylor Expansion Bound used by Sklar [31] assuming perfect estimation of the true value $f'(\theta_0)$ and the upper bound $f''(\theta) \leq 1$, which yields

$$f(\theta_0 + v) \leq f(\theta_0) + f'(\theta_0)v + \frac{v^2}{2}$$

- Pinsker's Inequality:

$$f(\theta_0 + v) \leq f(\theta_0) + \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_{\theta_0} \| P_{\theta_0 + v})} = f(\theta_0) + \frac{|v|}{2}$$

We observe that the Optimized Tilt-Bound is superior by a wide margin and stays close to the true Type I Error. At this distance of 0.25 standard deviation, the Optimized Tilt-Bound yields 2.73% where the true Type I Error is 2.5%. The relative advantage of the Tilt-Bound is consistent across other choices of origin point θ_0 and distance v, except that the Taylor bound can be better for very small distances. However, in practice, the Taylor bound as described in Sklar [31] requires conservative estimation of $f'(\theta_0)$ in addition to $f(\theta_0)$, which can erase this advantage in practical numbers of simulations. Thus, we generally recommend the Optimized Tilt-Bound.

We further remark that the bound performance in Figure 2 will hold for model classes where the likelihood model is asymptotically Gaussian and the test of interest is asymptotically equivalent to a z-test. Indeed, an approximate z-test can then be recovered under an appropriate re-mapping of the parameters to the signal-to-noise scale, in which case, these inequalities are also recoverable with an appropriate scaling (assuming the likelihoods and bounds also converge to those of Gaussians). Even more generally, in a multi-dimensional asymptotically Gaussian parametric model with an asymptotically linear test statistic, an analysis similar to Figure 2 may describe the performance along a one-dimensional line of parameter space corresponding to the direction of the test.

5 CSE Application Workflow

In this section, we give a high-level overview of the workflow of applying CSE.

Before proceeding with CSE, we must begin with a bounded parameter region of interest Θ. In many cases, physical bounds on realistic outcomes and effect sizes will restrict Θ. Alternatively, attention can be restricted to a bounded confidence set with the procedure of Berger and Boos [42] (1994), at a minor cost to the net confidence level. Or, as a supplement to CSE analysis on a bounded region, loose upper bounds on the remainder of the space may be derivable with analytical techniques.

Although the use of CSE is not limited to Type I Error, for simplicity we restrict our discussion in this section to the case of Type I Error and Family-Wise Error Rate (FWER) in the case of multiple testing. However, we note in passing that it is straightforward to extend these results to study FDR and the bias of bounded estimators.

A basic issue will arise as we move to multiple hypothesis testing: the FWER is not necessarily continuous across hypothesis boundaries. Fortunately, with p hypotheses, the FWER function can be decomposed into 2^p smooth parts and thus divide-and-conquered.

We begin with null hypotheses $\mathcal{H}_j \subseteq \mathbb{R}^d$ for $j = 1, \ldots, p$. Define $\Theta := \{\theta : \theta \in \bigcup_{j=1}^{p} \mathcal{H}_j\}$ to be the null hypothesis space, i.e. the space of parameters such that at least one of the null hypotheses is true. The null hypotheses induce a natural partition of Θ into $2^p - 1$ subsets with disjoint interiors where each subset completely resolves whether each null hypothesis is true. Concretely, the partition is defined by $P(b)$ for some *(null) configuration* $b \in \{0,1\}^p \setminus \{\vec{0}\}$ where

$$P(b) := \bigcap_{j:b_j=0} \mathcal{H}_j^c \cap \bigcap_{j:b_j=1} \mathcal{H}_j \tag{11}$$

Following this setup, we define a few terminologies.

Definition 1 (Tile and Platten)**.** *Suppose $\mathcal{H}_j$ are null hypotheses for $j = 1, \ldots, p$. Let Θ be the null hypothesis space induced by $\{\mathcal{H}_j\}_{j=1}^p$. A subset $H \subseteq \Theta$ is a* tile *if the interior of H is a subset of $P(b)$ as in (11) for some configuration $b \in \{0,1\}^p \setminus \{\vec{0}\}$. Consider a collection of tiles $\{H_i\}_{i=1}^I$ with disjoint interiors that partition Θ and a collection of arbitrary points $\{\theta_i\}_{i=1}^I \subseteq \Theta$ associated with each tile at the same index. We say the pair $\{H_i\}_{i=1}^I, \{\theta_i\}_{i=1}^I$ is a* platten.

For every null configuration $b \in \{0,1\}^p \setminus \{\vec{0}\}$, the Type I Error to control FWER at any $\theta \in P(b)$, is given by $f_b(\theta) := \mathbb{E}_\theta[F_b(X)]$ where $X \sim P_\theta$ and $F_b(x)$ is the indicator that a given (arbitrary) test with data x rejects for some j where $b_j = 1$, i.e. rejects at least one of the true null hypotheses in configuration b.

We give a sketch of the CSE application workflow. Consider a platten $\{H_i\}_{i=1}^I$, $\{\theta_i\}_{i=1}^I$. Let $b(i)$ be such that H_i corresponds to the null configuration $b(i)$. Suppose the user gathers some information about the Type I Error at each θ_i. We apply Theorem 1 for each tile to get that

$$\sup_{v \in H_i - \theta_i} f_{b(i)}(\theta_i + v) \leq \inf_{q \geq 1} \sup_{v \in H_i - \theta_i} U(\theta_i, v, q, f_{b(i)}(\theta_i))$$

where U is the Tilt-Bound (5). That is, we use the Tilewise Optimized Tilt-Bound to bound the worst-case Type I Error in each tile. Combining the inference at each θ_i and the Tilewise Optimized Tilt-Bound, we "extend" our inference to each tile. Using a divide-and-conquer strategy to cover many tiles, we again "extend" our inference to *all of* Θ. In Sections 6 and 7, we give concrete methods utilizing this workflow. Specifically, Section 6 constructs valid confidence intervals on all of Θ by extending the confidence intervals at each θ_i; Section 7 constructs a critical threshold for any arbitrary test that achieves (average) level α on all of Θ by combining calibrated thresholds gathered at each θ_i.

6 Validation Procedure

In this section, we describe our *validation procedure* that provides confidence bounds on the Type I Error for any bounded null hypothesis space Θ. Concretely, given an arbitrary design $\mathcal{D}$ and the family of distributions for the data $\{P_\theta : \theta \in \Theta\}$, we wish to construct random functions $\left(\hat{\ell}(\cdot), \hat{u}(\cdot)\right)$ such that for any $\delta > 0$,

$$\forall \theta \in \Theta, \mathbb{P}\left(\hat{\ell}(\theta) \leq f(\theta)\right) \geq 1 - \delta \text{ and } \mathbb{P}\left(\hat{u}(\theta) \geq f(\theta)\right) \geq 1 - \delta$$

where $f(\theta)$ is the Type I Error of the design $\mathcal{D}$ when data comes from P_θ. Note that we give a pointwise valid guarantee for $\hat{u}$ rather than a uniform guarantee. Under the frequentist framework, since we assume the existence of only one true parameter that generates the data, it is sufficient to give a pointwise guarantee. To simplify the discussion, we restrict our attention to the upper bound $\hat{u}(\cdot)$, however, our methodology readily extends to the construction of $\hat{\ell}$ (see Remark 2).

We now discuss the validation procedure in detail. Following the workflow laid out in Section 5, we first discuss in Section 6.1 the simple method of constructing valid confidence bounds for a point-null. In Section 6.2, we extend this guarantee at a point to a tile using CSE. In Section 6.3, we divide-and-conquer to construct our desired upper bound function $\hat{u}(\cdot)$ to give guarantees on all of Θ. To summarize the validation procedure, we demonstrate it on the one-sided z-test in Section 6.4.

6.1 Validation on a Single Point

Validation on a single point is the simplest case. The task at hand is to construct an upper bound $\hat{u}$ given any $\delta > 0$ such that at a point θ_0,

$$\mathbb{P}_{\theta_0}\left(f(\theta_0) \leq \hat{u}\right) \geq 1 - \delta$$

Although there are many choices for $\hat{u}$, we use the Clopper-Pearson upper bound when $f(\theta)$ is the Type I Error function. (In the case of bounded estimates, this can be replaced with other inequalities such as Hoeffding's inequality [43]; but we refrain from exploring this further). Given simulations $i = 1, \ldots, N$, let

B_i be the indicator that simulation i (falsely) rejects. Let $R = \sum_{i=1}^{N} B_i$ be the number of false rejections. Then set the Clopper-Pearson upper bound

$$\hat{u} := \text{Beta}^{-1}(1-\delta, R+1, N-R) \tag{12}$$

where $\text{Beta}^{-1}(q, \alpha, \beta)$ is the qth quantile of $\text{Beta}(\alpha, \beta)$ distribution [44]. An attractive feature of the Clopper-Pearson bound is that $\hat{u}$ is finite-sample valid.

For analysis of FDR or bias of a bounded estimator, the Clopper-pearson interval can be replaced with another finite-sample interval (such as Hoeffding confidence bounds.)

6.2 Validation on a Tile

We now discuss how to "extend" the confidence bound from Section 6.1 to a tile.

Let H be a tile (see Definition 1) with an associated point θ_0. Suppose that $\hat{u}_0$ is the $(1-\delta)$ Clopper-Pearson upper bound (12). Fix any displacement $v \in H - \theta_0$. Using the fact that the Tilt-Bound $a \mapsto U(\theta_0, v, q, a)$ is increasing, which is immediate from the definition, we have that

$$\mathbb{P}_{\theta_0}\left(f(\theta_0 + v) \le U(\theta_0, v, q, \hat{u}_0)\right) \ge \mathbb{P}\left(f(\theta_0) \le \hat{u}_0\right) \ge 1 - \delta \tag{13}$$

This shows that we may extend the upper bound function over H as

$$\hat{u}(\theta_0 + v) := U(\theta_0, v, q, \hat{u}_0) \tag{14}$$

to achieve a pointwise valid confidence guarantee. As a conservative simplification we may consider the worst-case Tilt-Bound estimate, flattening the bound over H:

$$\hat{u}(\cdot) := \sup_{v \in H - \theta_0} U(\theta_0, v, q, \hat{u}_0)$$

Additionally, with the linearity condition of Theorem 3, we can compute this supremum easily for convex polytope H (see Theorem 3). Finally, in either case, we may also minimize over $q \in [1, \infty]$ to get a tighter bound for free (see Theorem 2 regarding computation).

In practice, we typically choose θ_0 close to elements of H (e.g. the center of H). However, the theory does not enforce any particular way of choosing θ_0. Also, as H shrinks in size, we incur less extrapolation cost from the Tilt-Bound, resulting in a tighter upper bound $\hat{u}$.

6.3 Validation on a General Bounded Space

Given a general bounded null hypothesis space Θ, we begin with a platten $\{H_i\}_{i=1}^I$, $\{\theta_i\}_{i=1}^I$ (see Definition 1). To construct an upper bound $\hat{u}(\cdot)$ for all of Θ, we use the method described in Section 6.2 to construct valid upper bounds $\hat{u}_i(\cdot)$ for each tile H_i associated with θ_i. Then, we simply combine this collection into one function, namely

$$\hat{u}(\theta) := \hat{u}_i(\theta) \quad \text{if } \theta \in H_i$$

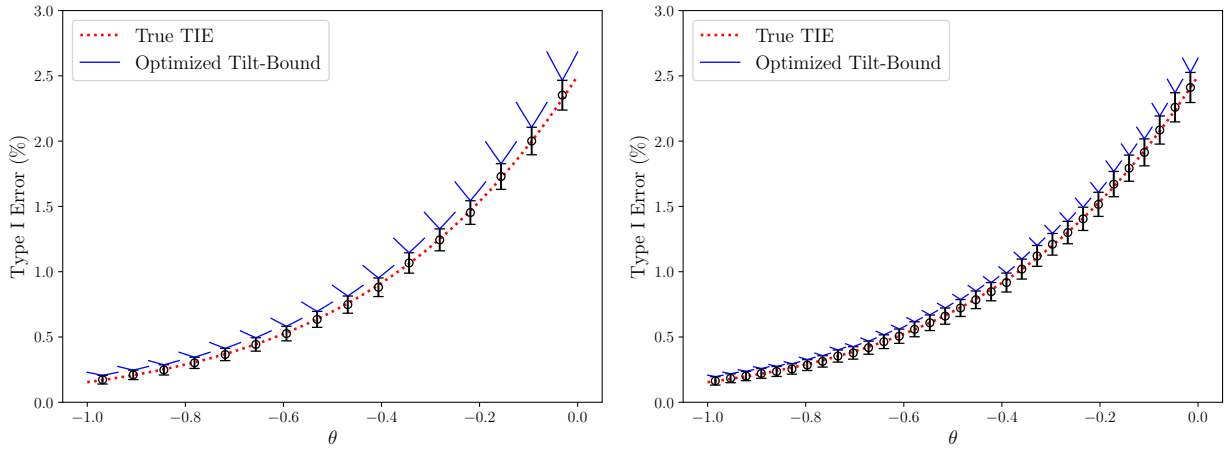
Note that by definition of platten, the interior of H_i are disjoint so that $\hat{u}(\theta)$ is well-defined whenever θ lies in the interior of some H_i. For θ on the boundaries of possibly multiple H_i, one may arbitrarily define $\hat{u}$ to take any one of the corresponding $\hat{u}_i$, so long as the choice is not random. Or, one could also define $\hat{u}(\theta) := \max_{j:\theta \in H_j} \hat{u}_j(\theta)$, the maximum bound across all tiles containing θ. The validation procedure is summarized in Algorithm 1.

It is worth noting that $\hat{u}$ only depends on simulations at finitely many points, namely $\{\theta_i\}_{i=1}^I$, so $\hat{u}$ is computable. However, the implication is indeed that we have a pointwise valid confidence bound on all of Θ, not just at the simulated points $\{\theta_i\}_{i=1}^I$.

In practice, the usefulness of $\hat{u}$ depends on the fineness of the tiles. A rougher partition will require less computation, but will result in a more conservative $\hat{u}(\cdot)$. In any case, the confidence guarantee is valid in finite samples.

Algorithm 1 Validation Procedure

1: Construct a platten $\{H_i\}_{i=1}^I$, $\{\theta_i\}_{i=1}^I$ for the (bounded) null hypothesis space Θ.
2: Construct a $(1-\delta)$ Clopper-Pearson interval at θ_i.
3: Use the Optimized Tilt-Bound to extend the confidence bound at θ_i to all points in H_i.
4: Combine the tilewise confidence bounds into a single bound on Θ.



(a) **(b)**

Figure 3: Plot of the optimized Tilt-Bound (**blue** solid line) for the z-test Clopper-Pearson confidence intervals (**black** vertical bars) overlaid with the true Type I Error (**red** dotted line). Figure 3a uses 16 equally-spaced simulation points and Figure 3b uses 32 points. This shows that with a finer gridding, the cost of the optimized Tilt-Bound quickly becomes negligible.

6.4 Validation for The Z-Test

We now demonstrate the validation procedure for the one-sided z-test (see Section 4.2) for testing $H_0 : \theta \leq 0$ at level $\alpha = 2.5\%$. Traditionally, the null hypothesis space is $(-\infty, 0]$, however, to satisfy the boundedness assumption, we limit our attention to $\Theta := [-1, 0]$.

Following Section 6.3, we choose equally-spaced points in the interval $(-1, 0)$, which naturally define a platten for Θ where θ_i are taken to be the centers of the tiles (intervals) H_i with equal radius. For each simulation point θ_i, we construct $(1-\delta)$ Clopper-Pearson upper bounds, where $\delta \equiv 0.05$. We then compute the upper bounds for each tile as in Section 6.2.

Figure 3 shows the validation results for the one-sided z-test using 16 simulation points (3a) and 32 simulation points (3b). Here, the Optimized Tilt-Bound refers to the upper bound $\hat{u}$ as in (14) optimized over $q \in [1, \infty]$. We overlay the true Type I Error for comparison and include the Clopper-Pearson intervals at each simulation point. Notice that the bounds are all above the true Type I Error curve, lying close to the curve. Depending on the tile, the flatness of the bound adapts as well: the smaller the θ, the flatter the bound. This demonstrates that the Tilt-Bound is using the "curvature information" of the Normal location family quite well and accurately depicts the area of difficult estimation, namely the area close to the null hypothesis boundary. By comparing Figures 3a and 3b, we see that the cost of the Tilt-Bound becomes neglible as the platten becomes finer, to the point where the cost of Clopper-Pearson bounds dominate. Note, however, that the Clopper-Pearson bounds can be tightened by increasing the number of simulations.

7 Calibration Procedure

The *calibration procedure* provides a level-α expected Type I Error control guarantee at every point of the bounded null hypothesis region Θ. In the sense that a confidence interval's guarantee is valid averaged over possible observations of the data, the calibration guarantee is valid averaged over the possible threshold

choices that result from the process. In general, the actual expected Type I Error is smaller than the nominal guarantee, and can be much smaller if the simulation scale is not sufficiently large. But the guarantee is still valid in this case, and the sponsor is incentivized to provide more comprehensive simulations to increase power. The fail-safe properties of calibration may be highly desirable for regulatory applications. These properties should be compared against the *validation* procedure described earlier in Section 6 in which the design's critical threshold is fixed, but randomness is present in the guarantee level so that a target could be randomly overshot. Moreover, the validation confidence bound only holds with probability $1 - \delta$.

Given an arbitrary design $\mathcal{D}$, the family of distribution for the data $\{P_\theta : \theta \in \Theta\}$, and the level α, we wish to construct a (random) threshold $\hat{\lambda}$ such that

$$\forall \theta \in \Theta, \, \mathbb{E}\left[f_{\hat{\lambda}}(\theta)\right] \leq \alpha$$

where $f_\lambda(\theta)$ is the Type I Error of the design $\mathcal{D}$ using the critical threshold λ, when the data comes from P_θ. To be more precise, we assume $f_\lambda(\theta)$ takes on the form

$$f_\lambda(\theta) := \mathbb{P}_\theta\left(S(X) < \lambda\right)$$

where $S(X)$ is the test statistic of $\mathcal{D}$ and $X \sim P_\theta$. As an example, for the one-sided z-test, it is natural to consider $S(X) := -X$ so that $S(X) < \lambda$ corresponds to an upper-tail rejection. Equivalently, we could use the p-value $S(X) := 1 - \Phi(X)$ where Φ is the standard normal CDF.

As in Section 6, we first discuss how to calibrate for a single point in Section 7.1. Section 7.2 extends this guarantee at a point to a tile using CSE. In Section 7.3, we divide-and-conquer to construct a single critical threshold that achieves level α on all of Θ. Finally, we apply the calibration procedure on the one-sided z-test in Section 7.4.

7.1 Calibration on a Single Point

We begin with the calibration procedure on a single null point θ_0.

Suppose N simulations are performed at θ_0 so that we collect $S_1, \ldots, S_N$ i.i.d. samples of the test statistic under P_{θ_0}. Let

$$\hat{\lambda}^\star := S_{(\lfloor (N+1)\alpha \rfloor)} \tag{15}$$

where $S_{(i)}$ is the ith order statistic among $S_1, \ldots, S_N$. Then, by an exchangeability argument, Theorem 4 shows that

$$\mathbb{E}_{\theta_0}\left[f_{\hat{\lambda}^\star}(\theta_0)\right] \leq \frac{\lfloor (N+1)\alpha \rfloor}{N+1} \leq \alpha$$

Hence, this result allows us to *target* a Type I Error upper bound at a given point for any level α by choosing the appropriate quantile of the test statistics.

Theorem 4 tells us more. First, it lower bounds the average Type I Error in (16). Moreover, (17) shows that the variance of the procedure consists of a component that decays at rate $O\left(\frac{1}{N}\right)$ and a second component due to jumps in the distribution. This second component vanishes in the case of a continuous test statistic $S(X)$ and can thus be removed by randomizing the test - in general, this may be achieved by adjoining an independent uniform random variable to the sample space, and re-defining the test statistic as (S, u) and the threshold as (λ, c) to be compared in the lexicographic order. As a result, we also have concentration of $f(\hat{\lambda}^\star)$. Thus, power is well-maintained for sensible tests without heavy discretization.

Theorem 4 (Pointwise calibration)**.** *Let $S_1, \ldots, S_N$ be any i.i.d. random variables. Fix any $\alpha \in \left[\frac{1}{N+1}, \frac{N}{N+1}\right]$. Define the following functions:*

$$\begin{aligned} f_+(\lambda) &:= \mathbb{P}\left(S \leq \lambda\right) \\ f(\lambda) &:= \mathbb{P}\left(S < \lambda\right) \\ \Delta f(\lambda) &:= f_+(\lambda) - f(\lambda) \end{aligned}$$

Finally, let $\hat{\lambda}^\star$ *as in* (15) *and* $\delta_{N,\alpha} := \mathbb{E}\left[\Delta f(\hat{\lambda}^\star)\right]$. *Then,*

$$\frac{\lfloor (N+1)\alpha \rfloor}{N+1} - \delta_{N,\alpha} \leq \mathbb{E}\left[f(\hat{\lambda}^\star)\right] \leq \frac{\lfloor (N+1)\alpha \rfloor}{N+1} \tag{16}$$

Moreover,

$$\operatorname{Var} f(\hat{\lambda}^\star) \leq O\left(\frac{1}{N}\right) + \delta_{N,\alpha}\left(\frac{2\lfloor (N+1)\alpha \rfloor}{N+1} - \delta_{N,\alpha}\right) \tag{17}$$

7.2 Calibration on a Tile

We now discuss calibration on a tile.

Let H be a tile (see Definition 1) with an associated point θ_0. We wish to construct $\hat{\lambda}^\star$ such that

$$\sup_{v \in H - \theta_0} \mathbb{E}\left[f_{\hat{\lambda}^\star}(\theta_0 + v)\right] \leq \alpha$$

Note that for any random rejection rule $\hat{\lambda}$, the Tilt-Bound guarantee in Theorem 1 holds so that

$$\mathbb{E}\left[f_{\hat{\lambda}}(\theta_0 + v)\right] \leq U(\theta_0, v, q, \mathbb{E}\left[f_{\hat{\lambda}}(\theta_0)\right]) \tag{18}$$

for any $q \geq 1$. This is because by Fubini's Theorem, we may view

$$\mathbb{E}\left[f_{\hat{\lambda}}(\theta)\right] = \mathbb{E}_{X \sim P_\theta}\mathbb{E}_{\hat{\lambda}}\left[\mathbb{1}_{S(X) < \hat{\lambda}}\right] = \mathbb{E}_{X \sim P_\theta}\left[G(X)\right]$$

where $G(x) := \mathbb{P}\left(S(x) < \hat{\lambda}\right) \in [0,1]$. Hence, we may apply Theorem 1 with G as the measurable function to integrate.

By back-solving (18) for $\mathbb{E}\left[f_{\hat{\lambda}}(\theta_0)\right]$, we get that

$$\sup_{v \in H - \theta_0} U\left(\theta_0, v, q, \mathbb{E}\left[f_{\hat{\lambda}}(\theta_0)\right]\right) \leq \alpha \iff \mathbb{E}\left[f_{\hat{\lambda}}(\theta_0)\right] \leq \inf_{v \in H} U^{-1}(\theta_0, v, q, \alpha) \tag{19}$$

where U^{-1} is the *Inverted Tilt-Bound* defined by

$$U^{-1}(\theta_0, v, q, \alpha) = \left[\alpha \exp\left(-\frac{\psi(\theta_0, v, q)}{q} + \psi(\theta_0, v, 1)\right)\right]^{\frac{q}{q-1}} \tag{20}$$

Hence, the bound in (19) can be thought of as a target level at θ_0 to ensure the test will be level α on all of H. Note that the bound in (19) can be maximized over $q \in [1, \infty]$ to get the *least* conservative target at θ_0. So, we have reduced the problem to constructing $\hat{\lambda}^\star$ such that

$$\mathbb{E}\left[f_{\hat{\lambda}^\star}(\theta_0)\right] \leq \sup_{q \in [1,\infty]} \inf_{v \in H} U^{-1}(\theta_0, v, q, \alpha) =: \alpha' \tag{21}$$

By a similar argument as in Theorem 3, as soon as H is a convex polytope and the linearity condition is satisfied, the infimum reduces to a finite minimum. And similar to Theorem 2, the supremum is a one-dimensional quasi-concave maximization problem. Together, the bound α' in (21) can be easily computed. To construct $\hat{\lambda}^\star$, we simply use Section 7.1 with "α" as α'. Hence, by relying on the Tilt-Bound to find a target level for θ_0 and simulating only at θ_0, we have successfully extended the Type I Error control to all of H, that is, we have constructed a computable $\hat{\lambda}^\star$ such that

$$\sup_{v \in H - \theta_0} \mathbb{E}\left[f_{\hat{\lambda}^\star}(\theta_0 + v)\right] \leq \alpha$$

Algorithm 2 Calibration Procedure

1: Construct a platten $\{H_i\}_{i=1}^I$, $\{\theta_i\}_{i=1}^I$ for the (bounded) null hypothesis space Θ.
2: For each tile H_i and simulation point θ_i, construct $\hat{\lambda}_i^\star$ (15) that satisfies (21).
3: Select the most conservative threshold $\hat{\lambda}^\star := \min_{i=1,\dots,I} \hat{\lambda}_i^\star$.

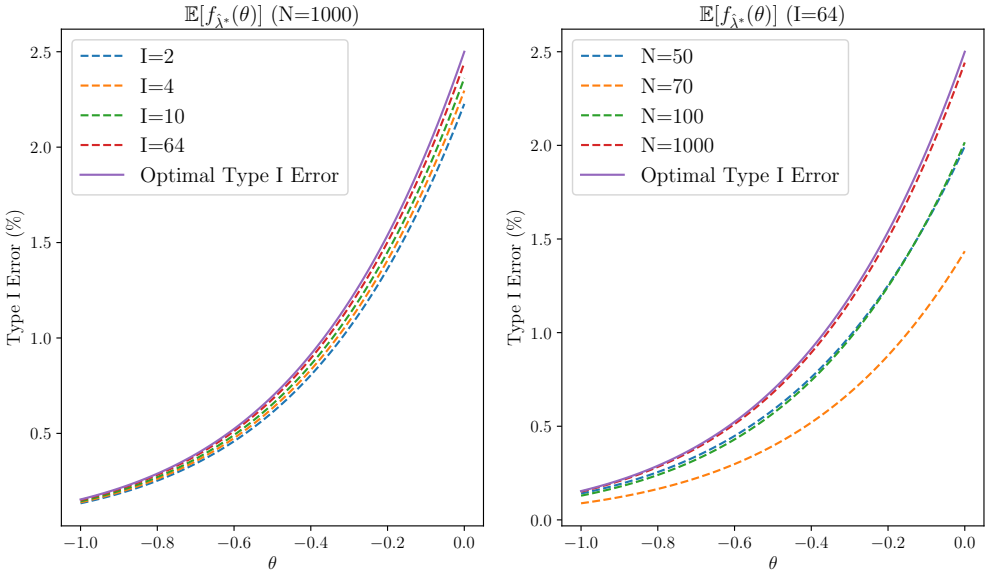


Figure 4: We consider the one-sided z-test on $\Theta \equiv [-1, 0]$. Both plots show a comparison of the average Type I Error $\mathbb{E}\left[f_{\hat{\lambda}^\star}(\theta)\right]$ using the calibrated threshold $\hat{\lambda}^\star$ as in (22). It is assumed that correlated simulations will be used generate the z-score of dataset i, k by translating a shared error $Z_k \sim N(0, 1)$ by θ_i. The curves shown for $\mathbb{E}\left[f_{\hat{\lambda}^\star}(\theta)\right]$ are derived analytically. The left fixes the calibration simulation size $N = 1000$ while varying the number of tiles. Conversely, the right fixes the number of tiles $I = 64$ while varying the number of simulations per-tile. Together they show that with greater computational power, nearly-optimal Type I Error can be achieved.

7.3 Calibration on a General Bounded Space

Given a general bounded null hypothesis space Θ, we begin with a platten $\{H_i\}_{i=1}^I$, $\{\theta_i\}_{i=1}^I$ (see Definition 1). For each tile H_i, we first perform tilewise calibration as described in Section 7.2 to obtain calibrated thresholds $\hat{\lambda}_i^\star$ that achieves level α within each H_i. It is noteworthy that these guarantees permit correlating samples or re-using the simulation RNG across different values of θ_i so long as the simulations corresponding to each individual tile H_i remain i.i.d.. Correlating the simulations offers not only computational savings but also smooths out the estimated Type I Error surface, which reduces the procedure's conservatism. After performing these calibrations, we may select the most conservative threshold over of the tiles,

$$\hat{\lambda}^\star := \min_{i=1,\dots,I} \hat{\lambda}_i^\star \tag{22}$$

Then, since the rejection sets $\{S(X) < \lambda\}$ is monotonic in λ,

$$\sup_{\theta\in\Theta} \mathbb{E}\left[f_{\hat{\lambda}^\star}(\theta)\right] = \max_{i=1,\dots,I} \sup_{\theta\in H_i} \mathbb{E}\left[f_{\hat{\lambda}^\star}(\theta)\right] \leq \max_{i=1,\dots,I} \sup_{\theta\in H_i} \mathbb{E}\left[f_{\hat{\lambda}_i^\star}(\theta)\right] \leq \alpha$$

To summarize, Algorithm 2 describes the full calibration procedure.

7.4 Calibration on The Z-Test

In this section, we apply the calibration procedure on the one-sided z-test.

Figure 4, shows the Type I Error performance of the calibration procedure(see Section 4.2). We plot the average Type I Error $\theta \mapsto \mathbb{E}\left[f_{\hat{\lambda}^\star}(\theta)\right]$ where $\hat{\lambda}^\star$ is constructed as in (22). We also show the optimal Type I Error, which is the usual curve $\theta \mapsto 1 - \Phi(z_{1-\alpha} - \theta)$. In the left plot, the calibration simulation size is fixed to $N = 1000$ and the number of tiles I is varied. Conversely, the right plot fixes the number of tiles $I = 64$ and varies the calibration simulation size. Both curves show close-to-optimal Type I Error performance with large I and N. In the right plot the performance is not strictly increasing in N; this is simply due to discretization error in equation (15) which occurs when α is not a multiple of $\frac{1}{N+1}$. For $\alpha = 2.5\%$, choosing N divisible by 200 is convenient to keep the discretization effect small.

8 From Calibration to Confidence Sets and Intervals

The calibrations discussed earlier in Section 7 can be used to form confidence regions and confidence intervals for parameters and estimands of interest, such as a treatment effect $p_1(\theta) - p_0(\theta)$. Below are the steps to construct a $(1-\alpha)$ confidence upper-bound on an inference target $e(\theta)$. To form a two-sided interval, one may simply intersect an upper and a lower interval, each at level $1 - \alpha/2$.

We describe the algorithm for constructing an upper confidence interval:

1: Select test statistic s_i for the inference target $e(\theta)$. If desired, one may choose a different test statistic for each tile H_i.

2: Compute the tile-specific calibrations of Section 7.2, resulting in critical values $\hat{\lambda}_i(\alpha)$.

3: Form a multi-dimensional confidence region $C_{1-\alpha}$ within the domain Θ, as a union of the non-rejected tiles:

$$C_{1-\alpha} := \bigcup_{i: s_i \geq \hat{\lambda}_i(\alpha)} H_i$$

4: Form the confidence set $C^e_{1-\alpha}$ from the image of $C_{1-\alpha}$ under the function e, defined as

$$C^e_{1-\alpha} := \{e(\theta) : \theta \in C_{1-\alpha}\}$$

This construction trivially attains coverage of at least $1-\alpha$. The resulting confidence sets are conservative in general, but if an asymptotically linear test statistic is used on an asymptotically Gaussian model, the cost may be small. The construction could perhaps be improved by developing higher-order test corrections to cause $C_{1-\alpha}$ to appear like a level set of $e(\theta)$ (thus reducing loss), or through the use of an additional re-calibration. We leave these topics to future work.

9 Handling CSE with Adaptive Procedures

In this section we discuss how our theory developed for exponential families applies to a wide class of designs and outcome mechanisms. In particular, we discuss

(1) **Adaptive Data Collection**: a design where the number of data points to be collected follows a pre-specified plan.

(2) **Administrative Censoring**: the right-censoring of survival data that occurs at time of analysis or study conclusion.

(3) **Latent Variables**: outcome data with multiple states such as an HMM, or multi-step randomness such as a random frailty.

To apply CSE to problems with adaptive sampling or stopping elements, we take the approach of embedding the adaptive model within a larger i.i.d. model. That is, in the case of an adaptive design with filtration $\mathcal{F}_t$ with maximum time (or sample size) T, we may trivially embed $\mathcal{F}_t$ in a finer σ-algebra $\mathcal{F}$ with i.i.d. sampling structure and possibly any other independent information. Then, the simple structure of $\mathcal{F}$

allows application of the Tilt-Bound to $\mathcal{F}$-measurable tests. Because $\mathcal{F}$ is finer, our theory immediately goes through for any tests which are $\mathcal{F}_t$-adapted.

For example, consider analyzing a one-arm trial with exponential outcomes with statistical parameter λ, with possible adaptive stopping and administrative censoring at time t. The filtration is $\mathcal{F}_t = \sigma(X_i \mathbb{1}_{\{X_i \leq s\}}, 1 \leq i \leq n, s \leq t)$ where $X_i \sim \text{Exp}(\lambda)$. We may embed this model in the larger $\mathcal{F} = \sigma(X_i, 1 \leq i \leq n)$, and perform simulations of i.i.d. exponentials. We must, of course, take care that the simulation's sequential rejection functions are correctly implemented and do not use any information past time t. The resulting Tilt-Bound on the Type I Error or other metric over the σ-algebra $\mathcal{F}$ thus directly implies a bound under $\mathcal{F}_\tau$ for any stopping time τ.

Next, consider the case of an adaptive sampling of binomial arms, such as in an analysis of the multi-arm bandit problem with K arms, each with $\text{Bern}(p_k)$ outcomes. In this case, the maximum possible number of samples that could be taken from each arm is N. We may embed this model in a σ-algebra $\mathcal{F}$ that contains a total of NK independent data points with $X_{nk} \sim \text{Bern}(p_k)$. In this case, the Tilt-Bound can be applied to a matrix of N samples from each of K independent Bernoullis.

Note that in (1), $F(X)$ must accord with the decision rule and prospective sampling decisions up to time τ, and not depend on data beyond time τ. In (2), $F(X)$ must further not depend on the specific data values that are larger than the censoring times γ_i, beyond knowledge that $\mathbb{1}(s_i > \gamma_i)$ where s_i is the i'th survival time. For (3), $F(X)$ must not depend on the value of the latent variable (whether the unknown Hidden Markov Model (HMM) state or unknown individual patient hazard) beyond what can be determined by the "visible" outcome data.

9.1 Calibration of Adaptive Designs with Interim Stopping for Efficacy

For calibration of adaptive designs with multiple interim analyses, defining the monotone family of designs indexed by λ may require some care. It is tempting to define λ as a global rejection-threshold across interim analyses, such as a critical threshold for the Bayesian posterior probability of efficacy for each treatment arm. This approach works for single-hypothesis studies, but in adaptive multi-arm designs it sometimes breaks the monotonicity condition required for calibration, because failing to reject one arm might cause a different arm to be rejected in the future. To avoid this issue, the design in Section 10.2 tunes the Bayesian posterior threshold only at the final analysis, rather than at all analyses simultaneously.

In practice, it may be sufficient to index λ as the critical rejection threshold of the final analysis,but in interest of absolute rigor we remark that it is possible for this calibration to fail if the simulations at some θ_i reject with probability greater than 2.5% before the final analysis. This issue does technically have a general, robust solution: the indexed family of designs can be expanded by removing the mass of the rejection set in order, from the largest time to the smallest time. To achieve this formally, one would define λ according to Siegmund's ordering [45], or equivalently, the lexicographic order on (τ, S_τ), where τ is the stopping time and S_τ is a rejection statistic, such that a smaller value of S_t should favor rejection. This extended definition of λ embeds the calibration of the final analysis in a rigorous way and ensures the viability of the overall Type I Error proof.

However in practice, if a calibration eliminates an intended final analysis, this occurrence may indicate poor setup of the design or insufficient simulations. In this case, the practitioner should be empowered to reconsider their design and simulation setup. Loss of formal Type I Error control is preferable to using a poor or poorly-calibrated design.

10 Application of CSE

We discuss two main applications of CSE. Section 10.1 studies a Bayesian basket trial from Berry et al. [1] where we apply the validation procedure to achieve a tight upper bound estimate of the Type I Error surface. Section 10.2 studies a complex Phase II/III selection design where we apply the calibration procedure to search for the critical threshold of the test that achieves level α.

10.1 Validation on a Bayesian Basket Trial

In this section, we examine the Type I Error (Family-Wise Error Rate) of a Bayesian basket trial, modeled after the design of Berry et al. [1]. Using the validation procedure described in Section 6, we obtain a tight upper bound estimate of the Type I Error using the Tilt-Bound. The resulting landscape in Figure 5 is complex and non-monotonic. A total of 7.34 trillion simulations of the trial were performed, taking 4 hours on a single Nvidia V100 GPU.

The model we consider examines the effectiveness of a single treatment applied to different groups of patients distinguished by a biomarker. The model has 4 arms indexed $i = 1, \ldots, 4$ and each patient has a binary outcome indicating treatment success or failure with probability p_i. Every arm has $n_i = 35$ patients. We model the outcome for each arm as:

$$y_i \sim \text{Binom}(n_i, p_i) \tag{23}$$

The hierarchical model allows for data-derived borrowing between the arms and is described in the log-odds space as:

$$\begin{aligned} \theta_i|\mu, \sigma^2 &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \mathcal{N}(-1.34, 100) \\ \sigma^2 &\sim \Gamma^{-1}(0.0005, 0.000005) \end{aligned}$$

where θ_i determines p_i according to:

$$\theta_i = \theta_{0i} + \text{logit}(p_i) \tag{24}$$

where $\text{logit}(p) = \log(p/(1-p))$ and θ_{0i} is a pre-determined constant indicating the effectiveness of the standard of care for arm i. Our use of the inverse-Gamma prior is for historical reasons, and may no longer be recommended (see [46] for suggestions of alternative priors). This hierarchical model borrows between the arms with greater posterior certainty if the four groups perform very similarly. Rejection for arm i is made according to the posterior probability using the following rule:

$$\mathbb{P}\left(p_i > p_{0i}|\mathbf{y}\right) > 85\% \tag{25}$$

where $\mathbf{y} := (y_1, y_2, y_3, y_4)$ is the vector of the binomial responses y_i.

We consider $\theta_i \in [-3.5, 1.0]$ for every i and show four different meaningful portions of the resulting Tilt-Bound surface in Figure 5. The Tilt-Bound is as high as 60% for some parameter combinations, which is not surprising given that the design was not optimized to control the Type I Error. Nevertheless, the Tilt-Bound is well-controlled at the global null point where $\theta_i = \theta_{oi}$ for all arms. This is consistent with the results of Berry et al. [1], which calibrated these Bayesian analysis settings to ensure a low Type I Error at the global null point for a slightly more complicated version of this design.

The Tilt-Bound shows significant multi-modality with high error both near the null hypothesis boundary and at another peak further into the alternative space of θ_1 and θ_2. This multi-modality is due solely to the sharing effects (as the design is not adaptive). The open triangle and open star on Figure 5A indicate the peaks in the Tilt-Bound and correspond, respectively, to the "One Nugget" and "2 Null, 2 Alternative" simulations from Berry et al. [1]. The first peak (indicated by an open triangle in Figure 5A) occurs when the three variables, θ_1, θ_3 and θ_4, are at their respective null hypothesis boundaries and the final variable, θ_2 is deep into its alternative space. The sharing effect from the alternative-space variable *pulls* estimates of the three null-space variables across the rejection threshold, resulting in high Type I Error. Similarly, the second peak (indicated by an open star in Figure 5A) occurs when the un-plotted variables, θ_3 and θ_4, lie on their respective null hypothesis boundaries and the hierarchical model pulls their estimates towards the plotted alternative-space variables θ_1 and θ_2.

10.2 Calibration on a Phase II/III Selection Design

In this section, we calibrate a Phase II/III Bayesian selection design which was suggested as a case study by an FDA official in personal communication. We wish to perform the calibration procedure as in Section 7 to obtain a level α test. Apart from calibration, the design is not otherwise optimized for performance. For general discussion on performing calibration of designs with adaptive stopping, see Section 9.1.

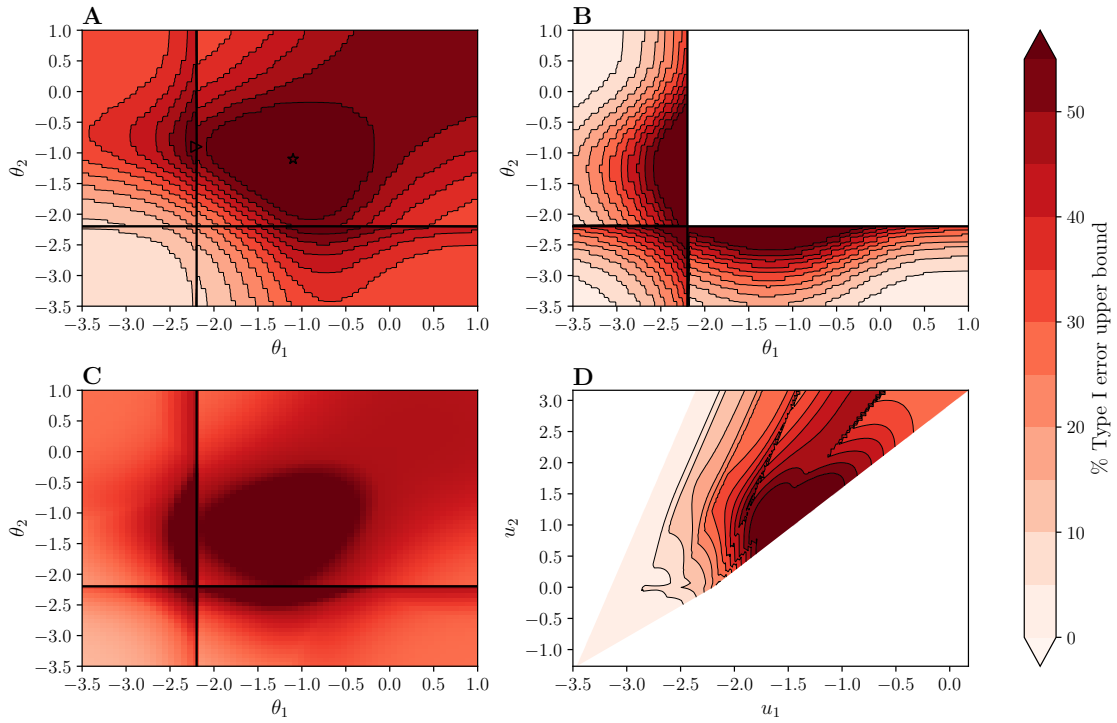


Figure 5: In plots **A**, **B** and **C**, the dark black lines indicate the boundary between the null space and the alternative space for each plotted parameter. **A)** The optimal Tilt-Bound as a function of θ_1 and θ_2 when $\theta_3 = \theta_4 = \theta_{critical}$. The open triangle and open star indicate the two peaks in Type I Error discussed in the text. **B)** The optimal Tilt-Bound as function of θ_1 and θ_2 with θ_3 and θ_4 fixed at the global maximum. Note that the white space in the figure indicates the parameter sets would fall in the alternative space for all arms, and therefore, the Tilt-Bound is not meaningful. **C)** For each value of (θ_1, θ_2), the error surface shown is the worst-case Type I Error over (θ_3, θ_4). **D)** For each $u_1 = (\theta_1 + \theta_2 + \theta_3 + \theta_4)/4$ and $u_2 = \max_i(\theta_i - \theta_{0i})$, the error surface shown is the worst-case Type I Error over the remaining 2 dimensions

This Phase II/III selection design has two stages: the first stage has 3 treatment arms and 1 control arm, and it may select a treatment for continuation against control in the second stage. Outcomes are assumed $\text{Bern}(p_i)$ for $i = 0, \ldots, 3$, with $i = 0$ representing the control arm, and trial decisions are informed by the Bayesian hierarchical model described in Section 10.1 using data from all arms. At each of 3 interim analyses in the first stage, decision options include whether to stop for futility, to drop one or more poor-performing treatments, or to select a treatment for acceleration to the second stage. The second stage has one interim and one final analysis. The total number of patients across all arms and stages is at most 800 with at most 350 in any single arm. A full description is given in Appendix B.

We now restrict our attention to a bounded region of parameter space: the 4-dimensional cube $\Theta = [-1, 1]^4$, where $\theta \in \Theta$ assigns $\text{logit}(p_i)$ for each arm i (so that p_i ranges approx. [27%, 73%]). Each treatment arm for i in $1, 2, 3$ has a null hypothesis denoted H_i, with null region given by

$$\Theta_i = \{\theta \in [-1, 1]^4 : \theta_i \leq \theta_0\}$$

Calibration is only computationally tractable for this problem if we can use larger tiles for some (uninteresting) regions of space and small tiles for others. We adaptively use pilot simulations to select the geometry of tiles and number of simulations to perform. The smallest tiles have half-width 0.000488 whereas the largest tiles have half-width 0.0625. To grid the entire space at the density of the smallest tiles would require 281 trillion tiles. Instead, we used 38.6 million tiles. Tiles in regions of low Type I Error (FWER) use as few as 2048 simulations, while simulations in regions of high Type I Error use up to 524,288 simulations (Figure 7). In total, adaptive simulation reduces the total number of required simulations from 1.5×10^{20} to 960 billion for a total computational savings of approximately 160 million times. As noted above, without adaptivity,

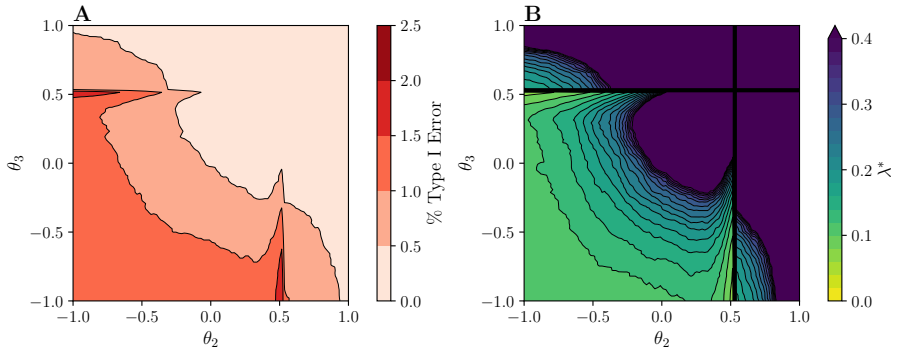


Figure 6: Calibration results of the Phase II/III design. Both plots slice the domain by fixing 2 parameters, $\theta_0 = \theta_1 = 0.533$. Figure **A** shows the Tilt-Bound (Type I Error upper bound) profile for the selected threshold $\lambda = 0.06253$. Figure **B** shows the critical value λ^* separately for each tile such that its Tilt-Bound is 2.5%.

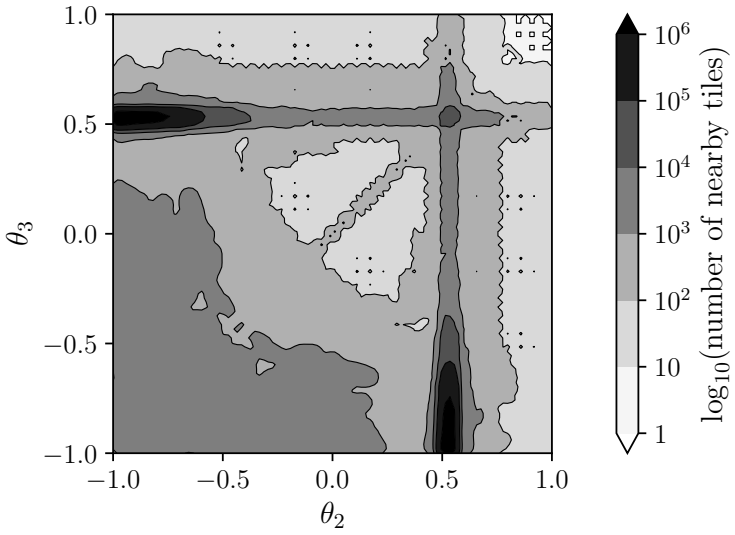


Figure 7: The number of tiles within a ball of radius 0.04 in the θ space. This shows how the adaptive algorithm assigns many more tiles to regions with high Type I Error. We show a slice of the domain where $\theta_0 = \theta_1 = 0.533$.

the solution here would be impossible even with the largest supercomputers available today. Instead, we are able to produce these results in 5 days with a single Nvidia V100 GPU.

We estimate using bootstrapping that at the critical point, the conservative bias due to estimation uncertainty accounts for an expected loss in Type I Error of about 0.16% (out of a total target 2.5%); the amount of Type I Error spent on Tilt-Bound CSE from the simulation point to the edges of tiles is 0.06%. (We note that the latter quantity is not "purely slack" because the Type I Error at the simulation point is generally not the worst-case on its tile.)

The Tilt-Bound profile of the calibrated design is presented in Figure 6. The maximum of the Tilt-Bound occurs at the tile centered at $\theta = (0.4925, 0.4925, 0.4925, -1.0)$. Surprisingly, the worst Type I Error point *does not* occur at the global null (where all treatments perform equally to control), but rather when one treatment performs poorly. It appears this "rogue arm" paradoxically *increases* the chance that other treatments are incorrectly rejected, because the observed heterogeneity causes a reduction in Bayesian borrowing. Unlike the design studied in Section 10.1, borrowing effects from highly successful arms do not cause Type I Error inflation. The Type I Error profile is non-monotonic with respect to θ_i and maximized away from the global null.

We remark that with the standard method of simply performing simulations at each grid point (i.e.

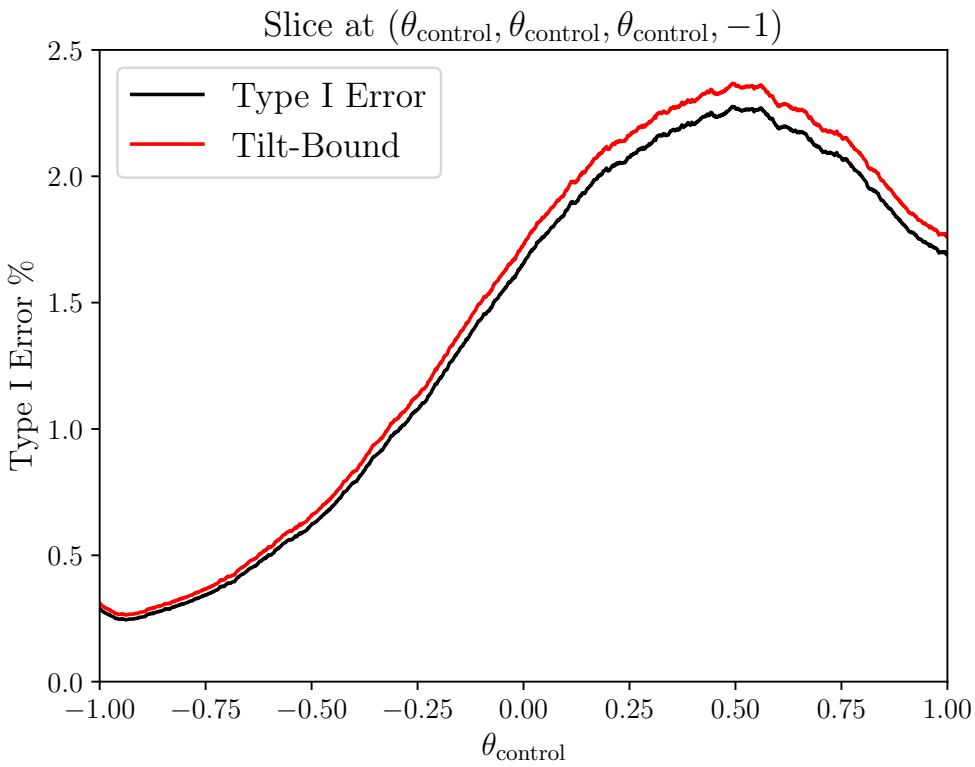


Figure 8: Type I Error estimates, with Tilt-bound from a separate simulation with fixed $\lambda = .06253$, focused on the 1-dimensional set where $\theta_0 = \theta_1 = \theta_2 = \theta_{control}$ and $\theta_3 = -1$. The worst case Type I Error appears when the trio of parameters is set to approximately 0.5.

without CSE), it would be difficult to establish worst-case control. Even after correctly guessing at the $(\theta, \theta, \theta, -1)$ form of the worst-case point, one would still need a 1-dimensional search of Type I Error over θ (such as Figure 8). In contrast, the "proof-by-simulation" approach establishes Type I Error control over the region of interest unambiguously.

However, CSE guarantees do not necessarily extend *outside* of the region of focus. If θ_3 is permitted to drop below -1, as shown in Figure 9 the worst-case Type I Error rises above 2.5% and asymptotes at approximately 3.1%. If a 2.5% guarantee were desired for this area, then the initial domain Θ should have been extended to a wider region at the outset of calibration (although at an additional computational cost). Interestingly, it is technically possible to give a bound which extends Figure 9 to the left all the way to $\theta_3 = -\infty$ ($p_3 = 0$) at the cost of a small additional penalty. See description of Figure 9.

11 Discussion

Proof-by-simulation is a powerful and robust methodology. It places minimal constraints on the structure of the test and sampling decisions, and therefore can successfully validate procedures that defy classical analysis. Continuous Simulation Extension (CSE) enhances simulation-gridding with guarantees, providing an objective standard and removing doubts about potentially missing areas of the null hypothesis space or selective presentation of results.

A practical advantage of CSE is that it analyzes the design as represented in code. Thus, CSE offers designers a "free pass" to use approximate estimators and inference, and defuses theoretical uncertainties such as convergence of MCMC samplers. In fact, even if the inference code were to contain an undiscovered error, CSE guarantees would remain technically valid.

To ensure CSE's conservatism is kept small, the number of simulations required grows with the dimension

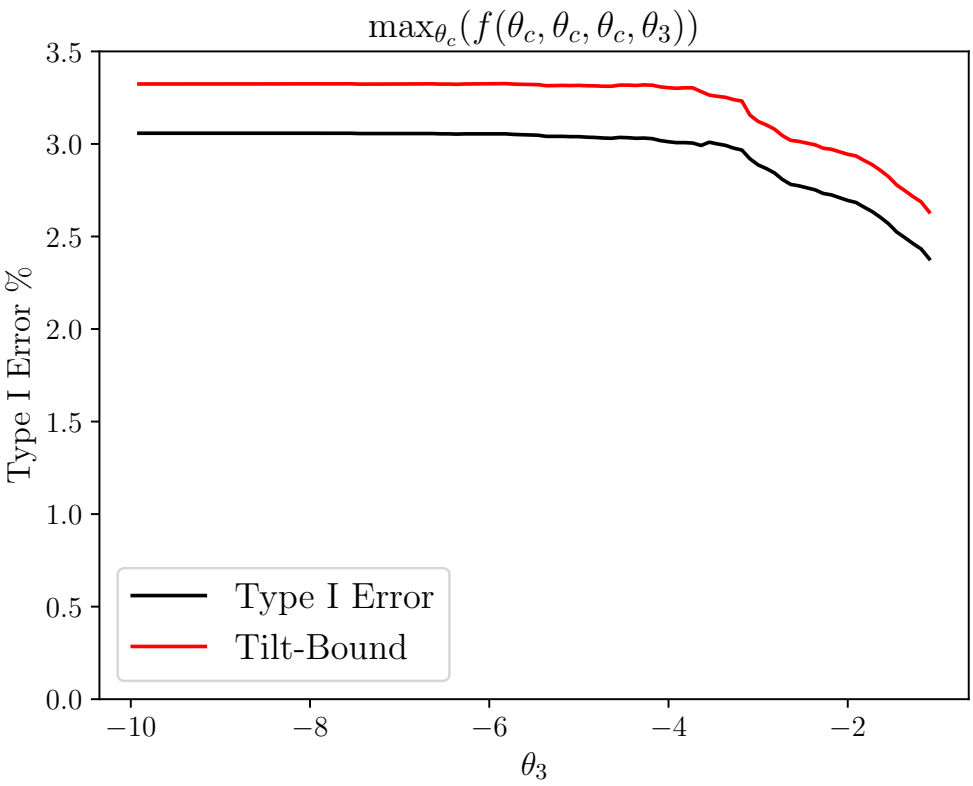


Figure 9: Worst-case Type I Error as a function of θ_3, where the remaining θ_i are constrained equal to each other and then set to the value which maximizes Type I Error given θ_3. Validation has been performed over this domain with λ set to 0.06253; the Tilt-Bound is shown in red. In fact, it is possible to extend this graph from the leftmost point $\theta_3 = -10$ ($p_3 = 0.000045$) all the way to $\theta_3 = -\infty$ ($p_3 = 0$) at the cost of a small additional expansion of the bound by about 0.3% (not shown). This is possible because the sequence of distributions as $\theta_3 \to -\infty$ converges in total variation, and if $\theta_i < -10$, the arm is overwhelmingly likely to dropped before it generates any successes whatsoever. From the left end-point $\theta_3 = -10$, since we can safely assume that this arm would be dropped after observing zero successes out of $n_3 = 50$ patients, the resulting bound would be a penalty of $< 0.3\%$ added to the Type I Error at the leftmost point of the figure.

of the parameter space. Our examples demonstrate that up to 4 unknown parameters can be handled using trillions of simulations; problems with fewer unknown parameters will be easier and require fewer simulations. We anticipate that large-scale computing will only become easier with time. In addition, our framework can be made more efficient with improvements to tile geometry, better adaptive gridding, and extensions to importance sampling. Thus, the set of compatible applications will grow.

In real medical experiments with adaptive elements, recruitment and logistical uncertainties may leave sample sizes and other "design parameters" undetermined at the start, to be revealed as the trial goes on; or the trial may undergo significant unplanned modifications. In this case, a specialized form of CSE can ensure maintenance of statistical control by recalibrating "group-sequentially" at each interim, using the conditional Type I Error principle similar to [47] and alternately conditioning on ancillary parameters and re-evaluating conditional Type I Error budget at each stopping-point. However this process may involve many repeated simulations over which simulation error can accumulate; we leave a full elucidation to future work.

For regulatory applications, a further subjective issue remains: how to justify the modeling family over which CSE will provide its guarantees? The history of previously accepted models is important here, but outside the scope of this paper. Yet it bears mentioning that CSE should be easier for regulators to accept than parametric *regression* methods. To perform inference on a parametrized regression model, such as a confidence interval on the parameters of a logistic regression, requires *estimability* of the model as well as

the *relevance* of the model class for inference. CSE relies only on the relevance; therefore even in cases where a parametric regression model is too uncertain or unstable to use, CSE can still be justified.

We also remark that even if distributional assumptions are unclear, additional evidence of robustness can be given by performing CSE for each of them, or by embedding the models in a larger family. In particular, if a Brownian motion limit applies for the statistics of interest, even simulations taken under the wrong parametric family may offer partial evidence of statistical control. For example: consider applying CSE using an incorrect assumption that data follow $\Gamma(\alpha, \beta)$ outcomes. The test statistics for these simulations may follow an approximate $\mathcal{N}(\mu, \sigma^2)$ distribution. Then, using CSE to cover a region of (α, β) is approximately covering a region of (μ, σ^2) for the test statistic. In this case, using CSE to control Type I Error over a wide range range of (α, β) would imply asymptotic control over a wide class of distributions with similar convergence properties.

Creative innovations in design optimization can be supported by CSE tools. For example, CSE can tighten or level out the Type I Error surface of a design. In design optimization, CSE may be useful as an "inner loop" to ensure that the current version of the design is valid, or to identify and quantify constraint violations. More speculatively, this framework could be used to calibrate "black-box" design protocols, like using a neural network to determine randomization probabilities or stopping decisions as a function of the sufficient statistics.

But our greatest hope is that these proof-by-simulation techniques can speed up innovation cycles. If properly implemented, we believe proof-by-simulation can reduce the human capital cost of parsing and validating new trial designs, improve regulatory consistency by enabling rigorous satisfaction of objective standards, and offer innovators in design a rapid alternative to peer review for establishing formal claims. Yet, much practical work remains to expand our software implementation of CSE for general use. In particular, the case studies in this work and Sklar [31] have been limited to Gaussian and Binomial models. To ensure practical usability, future research should expand these investigations to other model classes, develop introductory materials, continue to improve methodology with techniques such as importance sampling to increase efficiency, and build infrastructure for large-scale simulations and cost estimation to ensure that users find CSE methods cost-effective. We invite statisticians and regulators to explore our open-source code repository at [48] and consider its use for the next generation of statistical methodology.

12 Acknowledgements

We thank Alex Constantino and Gary Mulder for significant volunteer contributions to our software. Thanks also to Daniel Kang, Art Owen, Narasimhan Balasubramanian, and various regulators and biostatisticians for help and advice. This work was supported by ACX Grants.

References

[1] Scott M Berry, Kristine R Broglio, Susan Groshen, and Donald A Berry. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase ii oncology clinical trials. *Clinical Trials*, 10(5): 720–734, 2013.

[2] US Food, Drug Administration, et al. Statistical principles for clinical trials (ich e-9). In *International Conference on Harmonization. US Food and Drug Administration, DHHS*, 1998.

[3] Min Lin, Shiowjen Lee, Boguang Zhen, John Scott, Amelia Horne, Ghideon Solomon, and Estelle Russek-Cohen. Cber's experience with adaptive design clinical trials. *Therapeutic Innovation & Regulatory Science*, 50:195–203, 2016.

[4] FDA. Adaptive design clinical trials for drugs and biologics. guidance for industry. *Federal Register*, 2019.

[5] FDA. Guidance for the use of bayesian statistics in medical device clinical trials. *Federal Register*, 2010.

[6] U.S. Food and Drug Administration. Complex innovative trial designs brochure. URL https://www.fda.gov/media/129256/download.

[7] Andrew P Grieve. Idle thoughts of a 'well-calibrated'bayesian in clinical drug development. *Pharmaceutical statistics*, 15(2):96–108, 2016.

[8] Sue-Jane Wang and Frank Bretz. From adaptive design to modern protocol design for drug development: part i. editorial and summary of adaptive designs session at the third fda/dia statistics forum. *Drug Information Journal*, 44(3):325–331, 2010.

[9] Olivier Collignon, Franz Koenig, Armin Koch, Robert James Hemmings, Frank Pétavy, Agnès Saint-Raymond, Marisa Papaluca-Amati, and Martin Posch. Adaptive designs in clinical trials: from scientific advice to marketing authorisation to the european medicine agency. *Trials*, 19(1):1–14, 2018.

[10] Promoting the Use of Complex Innovative Designs in Clinical Trials, 2018. URL https://www.fda.gov/drugs/news-events-human-drugs/promoting-use-complex-innovative-designs-clinical-trials. Accessed: 2023-03-10.

[11] Dionne Price and John Scott. The us food and drug administration's complex innovative trial design pilot meeting program: progress to date. *Clinical Trials*, 18(6):706–710, 2021.

[12] Gregory Campbell. Similarities and differences of bayesian designs and adaptive designs for medical devices: a regulatory view. *Statistics in Biopharmaceutical Research*, 5(4):356–368, 2013.

[13] Russell R Barton, Henry Lam, and Eunhye Song. Input uncertainty in stochastic simulation. In *The Palgrave Handbook of Operations Research*, pages 573–620. Springer, 2022.

[14] Stephen E Chick. Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research*, 49(5):744–758, 2001.

[15] Faker Zouaoui and James R Wilson. Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions*, 36(11):1135–1151, 2004.

[16] Russell R Barton and Lee W Schruben. Uniform and bootstrap resampling of empirical distributions. In *Proceedings of the 25th conference on Winter simulation*, pages 503–508, 1993.

[17] Russell R Barton, Barry L Nelson, and Wei Xie. Quantifying input uncertainty via simulation confidence intervals. *INFORMS journal on computing*, 26(1):74–87, 2014.

[18] Russell CH Cheng and Wayne Holland. Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation Simulation*, 60(3):183–205, 1998.

[19] Russell CH Cheng and Wayne Holland. Calculation of confidence intervals for simulation output. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 14(4):344–362, 2004.

[20] Aleksey Tetenov. An economic theory of statistical testing. Technical report, cemmap working paper, 2016.

[21] Stephen Bates, Michael I Jordan, Michael Sklar, and Jake A Soloff. Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812*, 2022.

[22] EMA. Complex clinical trials–questions and answers. 2022.

[23] Jack PC Kleijnen. Kriging metamodeling in simulation: A review. *European journal of operational research*, 192(3):707–716, 2009.

[24] Mohamed Iskandarani, Shitao Wang, Ashwanth Srinivasan, W Carlisle Thacker, Justin Winokur, and Omar M Knio. An overview of uncertainty quantification techniques with application to oceanic and oil-spill simulations. *Journal of Geophysical Research: Oceans*, 121(4):2789–2808, 2016.

[25] Roman Schefzik, Thordis L Thorarinsdottir, and Tilmann Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. 2013.

[26] Anders Granholm, Aksel Karl Georg Jensen, Theis Lange, and Benjamin Skov Kaas-Hansen. adaptr: an r package for simulating and comparing adaptive clinical trials. *Journal of Open Source Software*, 7 (72):4284, 2022.

[27] Zhaolin Hu, Jing Cao, and L Jeff Hong. Robust simulation of global warming policies using the dice model. *Management science*, 58(12):2190–2206, 2012.

[28] Paul Glasserman and Xingbo Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14 (1):29–58, 2014.

[29] Soumyadip Ghosh and Henry Lam. Robust analysis in stochastic simulation: Computation and performance guarantees. *Operations Research*, 67(1):232–249, 2019.

[30] Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10), 2009.

[31] Michael Sklar. Adaptive experiments and a rigorous framework for type i error verification and computational experiment design, 2022. URL https://arxiv.org/abs/2205.09369.

[32] Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8):4986–5004, 2021.

[33] Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2016.

[34] Frank Nielsen and Richard Nock. On r\'enyi and tsallis entropies and divergences for exponential families. *arXiv preprint arXiv:1105.3259*, 2011.

[35] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[36] Ronald W. Butler. *Saddlepoint approximations with applications*. Cambridge University Press, 2007.

[37] D. Siegmund. Importance sampling in the monte carlo study of sequential tests. *The Annals of Statistics*, 4(4), 1976. doi: 10.1214/aos/1176343541.

[38] Tim van Erven and Peter Harremoes. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, jul 2014. doi: 10.1109/tit.2014.2320500.

[39] Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464, 1984. ISSN 00359246. URL http://www.jstor.org/stable/2345686.

[40] Amedeo Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, PP:1–1, 05 2021. doi: 10.1109/TIT.2021.3085190.

[41] Luc Bégin, Pascal Germain, Francois Laviolette, and Jean-Francis Roy. Pac-bayesian bounds based on the rényi divergence. 01 2016.

[42] Roger L Berger and Dennis D Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

[43] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[44] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

[45] T. L. Lai and W. Li. Confidence intervals in group sequential trials with random group sizes and applications to survival analysis. *Biometrika*, 93(3):641–654, 2006.

[46] Kristen M Cunanan, Alexia Iasonos, Ronglai Shen, and Mithat Gönen. Variance prior specification for a basket trial design using bayesian hierarchical modeling. *Clinical Trials*, 16(2):142–153, 2019.

[47] Franz Koenig, Werner Brannath, Frank Bretz, and Martin Posch. Adaptive dunnett tests for treatment selection. *Statistics in medicine*, 27(10):1612–1625, 2008.

[48] ConfirmSolutions. Imprint: Proof-by-simulation software for statistical validation, 2022. URL https://github.com/Confirm-Solutions/imprint.

A Proof of Main Results

In this section, we provide proofs of several main theorems for CSE stated in the main text. We first state a few preliminary results.

Lemma 5 (Basic Properties)**.** *Consider the setting of Theorem 1. Fix any $\theta_0 \in \Theta$. For every $q \geq 1$ and $v \in \mathbb{R}^d$, define*

$$\mathcal{D}_q := \{v \in \mathbb{R}^d : \psi(\theta_0, v, q) < \infty\} \tag{26}$$

$$\varphi_q(v) := \frac{\psi(\theta_0, v, q)}{q} - \psi(\theta_0, v, 1) \tag{27}$$

$$\mathcal{I}_v := \{q \geq 1 : \psi(\theta_0, v, q) < \infty\} \tag{28}$$

Let q_1, q_2 be any pair such that $1 \leq q_1 < q_2$. Then, the following statements hold:

(i) $\mathcal{D}_{q_1} \supseteq \mathcal{D}_{q_2}$.

(ii) If $v \mapsto \Delta_{\theta_0}(v, x)$ is linear, then $\mathcal{D}_q$ is convex.

(iii) $\mathcal{D}_q = \{v : 0 \leq \varphi_q(v) < \infty\}$.

(iv) $\mathcal{I}_v$ is convex and $q \mapsto \psi(\theta_0, v, q)$ is convex on $\mathcal{I}_v$. It is strictly convex if $\Delta_{\theta_0}(v, X)$ is not constant P_{θ_0}-a.s.

Proof. If $v \in \mathcal{D}_{q_2}$, then $\psi(\theta_0, v, q_2) < \infty$. By Jensen's Inequality,

$$\begin{aligned} \left\| e^{\Delta_{\theta_0}(v,X)} \right\|_{L^{q_1}(P_{\theta_0})} &\leq \left\| e^{\Delta_{\theta_0}(v,X)} \right\|_{L^{q_2}(P_{\theta_0})} \\ \implies \frac{1}{q_2}\psi(\theta_0, v, q_1) &\leq \frac{1}{q_2}\psi(\theta_0, v, q_2) < \infty \end{aligned} \tag{29}$$

Thus, we have (i).

Without loss of generality, now assume that $\mathcal{D}_q$ is non-empty. For any v_1, v_2 and any $\lambda \in (0, 1)$,

$$\mathbb{E}_{\theta_0}\left[e^{q\Delta_{\theta_0}(\lambda v_1 + (1-\lambda)v_2, X)}\right] \leq \mathbb{E}_{\theta_0}\left[e^{q\Delta_{\theta_0}(v_1, X)}\right]^{\lambda} \mathbb{E}_{\theta_0}\left[e^{q\Delta_{\theta_0}(v_2, X)}\right]^{1-\lambda}$$

by Hölder's Inequality and linearity of $\Delta_{\theta_0}(\cdot, X)$. Taking log on both sides,

$$\psi(\theta_0, \lambda v_1 + (1-\lambda)v_2, q) \leq \lambda\psi(\theta_0, v_1, q) + (1-\lambda)\psi(\theta_0, v_2, q) < \infty$$

Thus, $\lambda v_1 + (1-\lambda)v_2 \in \mathcal{D}_q$. This proves (ii).

Next, consider φ_q. Note that $\psi(\theta_0, v, 1)$ cannot be $-\infty$ because P_{θ_0+v} is a probability measure absolutely continuous to P_{θ_0}. Also, $\mathcal{D}_1 \supseteq \mathcal{D}_q$ by (i). Therefore, for $v \in \mathcal{D}_q$, we must have $\varphi_1(v) = 0$. Further, (29) shows that $q \mapsto \frac{\psi(\theta_0, v, q)}{q}$ is non-decreasing. Thus,

$$\mathcal{D}_q := \{v : |\varphi_q(v)| < \infty\} = \{v : 0 \leq \varphi_q(v) < \infty\}$$

Thus, this proves (iii).

Finally, we show that $\mathcal{I}_v$ is convex. For notational ease, let $Z := \Delta_{\theta_0}(v, X)$. For any $q_1 \neq q_2 \in \mathcal{I}_v$ and $\lambda \in (0,1)$, we apply Hölder's Inequality to get that

$$\mathbb{E}_{\theta_0}\left[e^{(\lambda q_1 + (1-\lambda)q_2)Z}\right] \leq \mathbb{E}_{\theta_0}\left[e^{q_1 Z}\right]^{\lambda} \mathbb{E}_{\theta_0}\left[e^{q_2 Z}\right]^{1-\lambda}$$

Taking log on both sides,

$$\psi(\theta_0, v, \lambda q_1 + (1-\lambda)q_2) \leq \lambda\psi(\theta_0, v, q_1) + (1-\lambda)\psi(\theta_0, v, q_2) < \infty$$

Hence, $\lambda q_1 + (1-\lambda)q_2 \in \mathcal{I}_v$. This proves that $\mathcal{I}_v$ is convex and $q \mapsto \psi(\theta_0, v, q)$ is convex on $\mathcal{I}_v$. Note that Hölder's Inequality is an equality if and only if there exists $\alpha, \beta \geq 0$ not both zero such that

$$\alpha e^{q_1 Z} = \beta e^{q_2 Z}$$

This can only occur if $q_1 = q_2$ or Z is constant P_{θ_0}-a.s. Hence, if Z is additionally not constant P_{θ_0}-a.s., then $q \mapsto \psi(\theta_0, v, q)$ is strictly convex on $\mathcal{I}_v$. This proves (iv). □

Lemma 6 (Monotonicity of $\tilde{\varphi}_q$)**.** *Consider the setting of Theorem 3. Fix any $\theta_0 \in \Theta \subseteq \mathbb{R}^d$, $v_0 \in \mathbb{R}^d$, $u \in \mathbb{R}^d$ such that $u \perp v_0$, and $q \geq 1$. Define*

$$\mathcal{D} := \{t \in \mathbb{R} : v_0 + tu \in \mathcal{D}_q\}$$
$$\tilde{\varphi}_q(h) := \varphi_q(v_0 + hu)$$

where $\mathcal{D}_q$ and φ_q are given by (26) and (27), respectively. Then, for all $h \in \mathcal{D} \cap (0, \infty)$, $\frac{\partial}{\partial h}\tilde{\varphi}_q(h) \geq 0$.

Proof. Without loss of generality, assume $\mathcal{D} \cap (0, \infty)$ is non-empty and consider any $h \in \mathcal{D} \cap (0, \infty)$. We begin with the expression from the chain rule:

$$\frac{\partial}{\partial h}\tilde{\varphi}_q(h) = \nabla\varphi_q(v_0 + hu)^\top u$$
$$\nabla\varphi_q(v) = \frac{\nabla_v\psi(\theta_0, v, q)}{q} - \nabla_v\psi(\theta_0, v, 1)$$

Since $h > 0$, letting $u_h := hu$, we may equivalently show that

$$h\frac{\partial}{\partial h}\tilde{\varphi}_q(h) = \nabla\varphi_q(v_0 + u_h)^\top u_h \geq 0$$

To establish the above, by an integration argument, it suffices to show that

$$\frac{\partial}{\partial t}\left(\frac{\nabla_v\psi(\theta_0, v_0 + u_h, t)^\top u_h}{t}\right) \geq 0$$

for $t \in [1, q]$. Note that all terms are well-defined for any $t \in [1, q]$ since

$$v_0 + u_h \in \mathcal{D}_q = \bigcap_{t \in [1,q]} \mathcal{D}_t$$

by Lemma 5 (i). By integrating, we then have that

$$h\frac{\partial}{\partial h}\tilde{\varphi}_q(h) = \int_1^q \frac{\partial}{\partial t}\left(\frac{\nabla_v\psi(\theta_0, v_0 + u_h, t)^\top u_h}{t}\right) dt \geq 0$$

For any $v \in \mathbb{R}^d$, define a family of distributions

$$\mathcal{Q}_v := \{Q_{t,v} : t \geq 1,\ \psi(\theta_0, v, t) < \infty\}$$

where $Q_{t,v}$ is given by

$$dQ_{t,v}(x) := \exp\left[t\Delta(v,x) - \psi(\theta_0, v, t)\right] dP_{\theta_0}(x)$$

Using the definition of ψ and $Q_{t,v}$, we have that

$$\begin{aligned}\nabla_v \psi(\theta_0, v, t) &= te^{-\psi(\theta_0,v,t)}\mathbb{E}_{\theta_0}\left[\nabla_v \Delta(v,X)e^{t\Delta(v,X)}\right] \\ &= t\mathbb{E}_{Q_{t,v}}\left[\nabla_v \Delta(v,X)\right]\end{aligned}$$

Hence, for any $t \in [1, q]$,

$$\frac{\nabla_v \psi(\theta_0, v_0+u_h, t)^\top u_h}{t} = \mathbb{E}_{Q_{t,v_0+u_h}}\left[\nabla_v \Delta(v_0+u_h, X)^\top u_h\right] \tag{30}$$

Note that $Q_{t,v}$ forms an exponential family with natural parameter t, sufficient statistic $\Delta(v,x)$, and log-partition function $\psi(\theta_0, v, t)$ with base measure P_{θ_0}. By standard results for exponential family, the derivative of (30) with respect to t is

$$\begin{aligned}\frac{\partial}{\partial t}\left(\frac{\nabla_v \psi(\theta_0, v_0+u_h, t)^\top u_h}{t}\right) &= \mathrm{Cov}_{Q_{t,v_0+u_h}}\left(\nabla_v \Delta(v_0+u_h, X)^\top u_h,\ \Delta(v_0+u_h, X)\right) \\ &= u_h^\top\left[\mathrm{Var}_{Q_{t,v_0+u_h}} W(X)\right](v_0+u_h)\end{aligned}$$

Let P_0 denote the projection matrix onto $\mathrm{span}(v_0)^\perp$. Then, since $v_0 \perp u_h$ by hypothesis, we have $P_0(v_0 + u_h) = P_0 u_h = u_h$. Consequently,

$$u_h^\top\left[\mathrm{Var}_{Q_{t,v_0+u_h}} W(X)\right](v_0+u_h) = (v_0+u_h)^\top P_0\left[\mathrm{Var}_{Q_{t,v_0+u_h}} W(X)\right](v_0+u_h) \tag{31}$$

Since P_0 and $\mathrm{Var}_{Q_{t,v_0+u_h}} W(X)$ are both positive semi-definite, their product must have all non-negative eigenvalues. Hence, the right-side of (31) is non-negative. This proves that $\frac{\partial}{\partial h}\tilde{\varphi}_q(h) \geq 0$ on $\mathcal{D} \cap (0, \infty)$. □

Lemma 7. *Let $f(x) = \frac{f_1(x)}{f_2(x)}$ where $f_1 : C \mapsto \mathbb{R}$ and $f_2 : C \mapsto (0, \infty)$ for some convex set C of a vector space. If f_1 is strictly convex and f_2 concave, then f is strictly quasi-convex.*

Proof. Fix any $x \neq y \in C$ and $\lambda \in (0,1)$. Then,

$$\begin{aligned}f(\lambda x + (1-\lambda)y) &< \frac{\lambda f_1(x) + (1-\lambda)f_1(y)}{\lambda f_2(x) + (1-\lambda)f_2(y)} \\ &= \alpha f(x) + (1-\alpha)f(y) \leq \max(f(x), f(y))\end{aligned}$$

where $\alpha = \frac{\lambda f_2(x)}{\lambda f_2(x) + (1-\lambda)f_2(y)}$. □

A.1 Proof of Theorem 2

Proof. Fix $\theta_0 \in \Theta$, a set $S \subseteq \mathbb{R}^d$, and $a \geq 0$. Without loss of generality, assume $a > 0$. If $a = 0$, we may simply set $q^* \equiv \infty$. Note that any $q \geq 1$ achieves the minimum, so the minimizer is not unique.

Let $\mathcal{I}_v$ be as in (28). Define $\mathcal{I} := \bigcap_{v \in S} \mathcal{I}_v$. By Lemma 5 (iv), $\mathcal{I}_v$ is convex for every $v \in S$ so $\mathcal{I}$ is convex as well. It suffices to minimize

$$q \mapsto \sup_{v \in S} U(\theta_0, v, q, a)$$

on $\mathcal{I}$ since it is infinite on $\mathcal{I}^c$. Without loss of generality, assume that $\mathcal{I}$ is non-empty, or equivalently, $q \mapsto \sup_{v \in S} U(\theta_0, v, q, a)$ is not identically infinite. Otherwise, we may set $q^* \equiv \infty$ and the minimizer is not unique.

We show that $q \mapsto \sup_{v \in S} U(\theta_0, v, q, a)$ is quasi-convex on $\mathcal{I}$ and is strict if S is finite. Note that

$$\sup_{v \in S} U(\theta_0, v, q, a) = a \cdot \exp\left[\sup_{v \in S}\left[\frac{\psi(\theta_0, v, q) - \log(a)}{q} - \psi(\theta_0, v, 1)\right]\right]$$

To establish the desired claim, we show that $\tilde{\psi}(q)$ is (strictly, if finite S) quasi-convex on $\mathcal{I}$ where

$$\tilde{\psi}(q) := \sup_{v \in S} \tilde{\psi}_v(q)$$
$$\tilde{\psi}_v(q) := \frac{\psi(\theta_0, v, q) - \log(a)}{q} - \psi(\theta_0, v, 1)$$

It suffices to show that $q \mapsto \tilde{\psi}_v(q)$ is strictly quasi-convex on $\mathcal{I}_v$ for every $v \in S$. Note that $\tilde{\psi}_v(q) = \frac{f_1(q)}{f_2(q)}$ where

$$f_1(q) := \psi(\theta_0, v, q) - \log(a) - q\psi(\theta_0, v, 1)$$
$$f_2(q) := q$$

Lemma 5 shows that $q \mapsto \psi(\theta_0, v, q)$ is strictly convex on $\mathcal{I}_v$, which directly shows the same for $f_1(q)$. Then, Lemma 7 shows that $\tilde{\psi}_v(q)$ is strictly quasi-convex.

Finally, since $\tilde{\psi}$ is (strictly, if finite S) quasi-convex on $\mathcal{I}$, we have the following quasi-convex program:

$$\begin{aligned} &\underset{q}{\text{minimize}} && \tilde{\psi}(q) \\ &\text{subject to} && q \in \mathcal{I} \end{aligned}$$

This proves the existence (and uniqueness for finite S) of a global minimizer $q^* \in \mathcal{I} \subseteq [1, \infty]$. □

A.2 Proof of Theorem 3

Proof. Fix any $\theta_0 \in \Theta$. For every $q \geq 1$, define $\mathcal{D}_q$ as in (26). By Lemma 5 (ii) and (iii), $\mathcal{D}_q \equiv \{v : |\varphi_q(v)| < \infty\}$ is convex, where φ_q is as in (27). Now, fix any $q \geq 1$. To establish quasi-convexity of the Tilt-Bound (5) as a function of v, it therefore suffices to show that $\varphi_q(v)$ is quasi-convex along all one-dimensional line segments contained in $\mathcal{D}_q$. Let a, b be any two points in $\mathcal{D}_q$, and $\overline{ab}$ the line segment joining them. We may consider the directional unit vector $u = \frac{b-a}{\|b-a\|}$, extend the line segment to make a full line $\overleftrightarrow{ab}$, and then drop a perpendicular vector v_0 which measures the distance from the origin to $\overleftrightarrow{ab}$. Hence, without loss of generality, it suffices to show that $h \mapsto \varphi_q(v_0 + hu)$ is quasi-convex for all vectors v_0 and u such that $u \perp v_0$ and $h \in \mathcal{D}_{q,v_0,u} := \{t \in \mathbb{R} : v_0 + tu \in \mathcal{D}_q\}$. Note that $\mathcal{D}_{q,v_0,u}$ is an interval by convexity of $\mathcal{D}_q$.

Without loss of generality, fix any such v_0 and u. Denote $\mathcal{D} \equiv \mathcal{D}_{q,v_0,u}$ for notational ease. Define $\tilde{\varphi}_q(h) := \varphi_q(v_0 + hu)$. We wish to show that $\tilde{\varphi}_q$ is quasi-convex on $\mathcal{D}$. Lemma 6 shows that $\frac{\partial}{\partial h}\tilde{\varphi}_q(h) \geq 0$ for all $h \in \mathcal{D} \cap (0, \infty)$, which shows that $\tilde{\varphi}_q(h)$ is non-decreasing for $h \in \mathcal{D} \cap (0, \infty)$. Note that since $\Delta(\cdot, X)$ is linear, we have by standard exponential family that for every $\tilde{q} \geq 1$, $v \mapsto \psi(\theta_0, v, \tilde{q})$ is continuous on $\mathcal{D}_{\tilde{q}}$. Then, by Lemma 5 (i), $v \mapsto \psi(\theta_0, v, 1)$ is continuous on $\mathcal{D}_q$ so φ_q is continuous on $\mathcal{D}_q$. Hence, $\tilde{\varphi}_q$ is non-decreasing on $\mathcal{D} \cap [0, \infty)$. Therefore, applying the same logic for $-u$, $\tilde{\varphi}_q$ must either be monotone on $\mathcal{D}$ or decreasing from the left end-point of $\mathcal{D}$ to 0 then increasing to the right end-point of $\mathcal{D}$. Thus, $\tilde{\varphi}_q$ is quasi-convex along the segment $\mathcal{D}$, as needed. □

A.3 Proof of Theorem 4

Proof. Define $\tilde{U}_i := f(S_i)$ for every $i = 1, \ldots, N$. Then, by monotonicity of f,

$$\mathbb{E}\left[f(S_{(\lfloor (N+1)\alpha \rfloor)})\right] = \mathbb{E}\left[\tilde{U}_{(\lfloor (N+1)\alpha \rfloor)}\right] \tag{32}$$

We claim that $\tilde{U}_i$ are sub-uniform, that is, they satisfy the following property:

$$\mathbb{P}\left(\tilde{U} \leq x\right) \geq x$$

for every $x \in [0,1]$. To prove this, we first define a pseudo-inverse function

$$f^{-1}(y) := \inf\{\lambda : f(\lambda) \geq y\}$$

for all $y \in [0,1]$. Note that

$$f(f^{-1}(y)) \leq y \leq f_+(f^{-1}(y))$$

by left-continuity of $f(\lambda)$. Hence, for any $x \in [0,1]$,

$$\mathbb{P}\left(\tilde{U} \leq x\right) = \mathbb{P}\left(f(S) \leq x\right) \geq \mathbb{P}\left(S \leq f^{-1}(x)\right) = f_+(f^{-1}(x)) \geq x$$

By considering a possibly enlarged probability space, we may construct a uniform random variable $U_i \sim U(0,1)$ for each $\tilde{U}_i$ such that $\tilde{U}_i \leq U_i$. In particular, we have that $\tilde{U}_{(k)} \leq U_{(k)}$ for any $k = 1, \ldots, N$. Continuing from (32),

$$\mathbb{E}\left[\tilde{U}_{(\lfloor (N+1)\alpha \rfloor)}\right] \leq \mathbb{E}\left[U_{(\lfloor (N+1)\alpha \rfloor)}\right]$$

Note that $U_{(k)} \sim \text{Beta}\left(k, N-k+1\right)$ for every $k = 1, \ldots, N$. Hence,

$$\mathbb{E}\left[U_{(\lfloor (N+1)\alpha \rfloor)}\right] = \frac{\lfloor (N+1)\alpha \rfloor}{N+1}$$

Next, note that

$$\begin{aligned} & f(s) = f_+(s) - \Delta f(s) \\ \implies & \mathbb{E}\left[f(S_{(\lfloor (N+1)\alpha \rfloor)})\right] = \mathbb{E}\left[f_+(S_{(\lfloor (N+1)\alpha \rfloor)})\right] - \mathbb{E}\left[\Delta f(S_{(\lfloor (N+1)\alpha \rfloor)})\right] \end{aligned}$$

Hence, it suffices to show that

$$\mathbb{E}\left[f_+(S_{(\lfloor (N+1)\alpha \rfloor)})\right] \geq \frac{\lfloor (N+1)\alpha \rfloor}{N+1}$$

By a similar argument as before, we construct a pseudo-inverse

$$f_+^{-1}(y) := \inf\{\lambda : f_+(\lambda) > y\}$$

so that

$$f(f_+^{-1}(y)) \leq y \leq f_+(f_+^{-1}(y))$$

Hence, for any $x \in [0,1)$ and $x < y \leq 1$,

$$\mathbb{P}\left(f_+(S) < y\right) \leq \mathbb{P}\left(S < f_+^{-1}(y)\right) = f(f_+^{-1}(y)) \leq y$$

Taking $y \downarrow x$,

$$\mathbb{P}\left(f_+(S) \leq x\right) \leq x$$

Since $f_+ \in [0,1]$, the result trivially holds for $x = 1$. Hence, $f_+(S)$ is super-uniform. This implies that

$$\mathbb{E}\left[f_+(S_{(\lfloor (N+1)\alpha \rfloor)})\right] \geq \mathbb{E}\left[U_{(\lfloor (N+1)\alpha \rfloor)}\right] = \frac{\lfloor (N+1)\alpha \rfloor}{N+1}$$

This concludes the proof of (16).

We now prove (17). For notational ease, let $k := \lfloor (N+1)\alpha \rfloor$. Then,

$$\begin{aligned} \text{Var}\, f(S_{(k)}) &= \mathbb{E}\left[f(S_{(k)})^2\right] - \mathbb{E}\left[f(S_{(k)})\right]^2 \\ &\leq \mathbb{E}\left[U_{(k)}^2\right] - \mathbb{E}\left[f(S_{(k)})\right]^2 \\ &\leq \text{Var}\, U_{(k)} + \left(\frac{k}{N+1}\right)^2 - \left(\frac{k}{N+1} - \mathbb{E}\left[\Delta f(S_{(k)})\right]\right)^2 \\ &= \text{Var}\, U_{(k)} + \mathbb{E}\left[\Delta f(S_{(k)})\right]\left(\frac{2k}{N+1} - \mathbb{E}\left[\Delta f(S_{(k)})\right]\right) \end{aligned}$$

Since $\text{Var}\, U_{(\lfloor (N+1)\alpha \rfloor)} = O(\frac{1}{N})$, we have the desired claim. □

B Phase II/III Design Details

The Phase II component begins with 4 arms, including 1 arm for control (a_0) and 3 treatment arms (a_1, a_2, a_3), and randomizes a maximum of 400 patients. Outcomes are assumed to be immediately observable, and distributed Bern(p_i) for each arm a_i. Randomization in the Phase II portion is blocked $1:1:1:1$ for the first 200 patients. An interim analysis occurs after the first 200 patients and then again after every 100 subsequent patients. In the interim analysis each remaining non-control arm is analyzed according to two metrics, p^i_{best} and p^i_{success}, according to the Bayesian hierarchical model presented in Section 10.1. The metric p^i_{success} is meant to approximate the conditional power if arm a_i were accelerated to Phase III. More precisely, p^i_{success} equals the probability under the current Bayesian posterior model that if 200 patients are added to both of the arms a_i and a_0, i.e. adding a completed Phase III dataset to the analysis, the resulting final posterior will have $\mathbb{P}(p_i > p_0) > 95\%$. This probability is pre-computed by simulation for a grid of possible data values, so that during calibration a rapid approximation can be made by interpolating a lookup table. If $p^i_{\text{success}} > 70\%$ for any treatment, Phase II concludes with a selection and the current best arm is accelerated to the Phase III portion. Otherwise, there is an assessment of which, if any, arms should be dropped. For $i = 1, 2, 3$, if $p^i_{\text{best}} := \mathbb{P}\left(p_i = \max_j p_j\right) < 15\%$, then arm a_i is dropped for futility and future patients in the Phase II will be split evenly among the remaining arms with fractional patients thrown out. If all treatment arms are dropped, the trial ends in futility. If the Phase II portion reaches its third and final analysis (≈ 400 patients randomized), the selection threshold for the best arm to progress to Phase III is lowered to $p^i_{\text{success}} > 60\%$. If no arm achieves this, the trial stops for futility.

If an arm is selected for Phase III, then up to 400 patients will be block-randomized 1:1 between the selected treatment and control. When 200 patients have been thus randomized, an interim analysis is performed. If at this interim $\mathbb{P}(p_i > p_0) > 95\%$ for the selected i, then success is declared immediately. If $p^i_{\text{success}} < 20\%$, where p^i_{success} is the current posterior probability that if 100 patients are added to each a_i and a_0 the resulting posterior will have $\mathbb{P}(p_i > p_0) > 95\%$, the Phase III stops for futility. If neither occurs, we progress to the final analysis with a final block of 200 block-randomized patients. The success criterion at this final analysis will be a threshold λ for the final posterior probability of superiority $\mathbb{P}(p_i > p_0) > 1 - \lambda$. λ is left as a tuning parameter, and is used to calibrate the design for Type I Error control. The selected λ was 6.253%.