EFFICIENCY REQUIRES ADAPTATION

BY ADAM LANE,

University of Cincinnati

The majority of historical designs are \textit{a priori} in nature, where \textit{a priori} indicates a design can be specified in advance of the experiment. The conventional wisdom is that the set of \textit{a priori} designs is sufficient to produce efficient experiments. This work challenges this convention and finds that efficiency requires data dependent strategies. Specifically, in the context of a sequential experiment, where observations are accrued in a series of runs, an adaptive design is proposed that is guaranteed to be more efficient than any corresponding \textit{a priori} design.

1. Introduction The Cramér Rao lower bound (CRLB) provides an explicit bound on the achievable precision of any parameter estimate obtained following an experiment. The CRLB($\xi$) is a function of the design, $\xi = [x(1), \ldots, x(n)]^T$, where $x(j)$ represents the support point of observation $j (= 1, \ldots, n)$. A critical challenge for an experimenter is to select a design that is efficient with respect to the CRLB.

Different experimental settings require different design strategies. Factorial designs are commonly used in experiments with many treatment factors. Optimal designs minimize the CRLB with respect to a specific convex optimality criterion. Minimax designs are robust against worst case scenarios. While, each of these designs are valuable in the appropriate context, they also, often, satisfy a broader definition of efficiency. Specifically, a design $\xi_1$ is \textit{efficient} if there does not exist a design, $\xi_2$, such that CRLB($\xi_1$) \textgreater CRLB($\xi_2$) in Loewner order. For two symmetric matrices $A$ and $B$, $A > B$ in Loewner order if $A - B$ is positive definite.

While the CRLB is the most popular lower bound there exist tighter alternatives [Bhattacharyya (1948); van Trees (1968); Bobrovsky, Mayer-Wolf and Zakai (1987)]. A design is more robustly efficient if it satisfies an analogous definition with respect to a tighter lower bound. In Section 2 such a lower bound is introduced. Efficiency, as considered in this work, is formally defined with respect to this alternative lower bound [Section 3].

Historically, the design literature has primarily focused on \textit{a priori} designs - designs that can be specified in advance of the experiment. The set of \textit{a priori} designs includes deterministic designs [eg. factorial, optimal and minimax designs] and designs selected via data independent random processes [e.g. Wu (1981); Li (1983); Waite and Woods (2021)].

Conventional wisdom suggests that the set of \textit{a priori} designs is sufficient to produce efficient experiments. The main result of this work finds this convention is incorrect; efficiency requires data dependent strategies. This apparent paradox is proven by developing an adaptive design that when applied to any \textit{a priori} design reduces the lower bound [proposed in Section 2] in Loewner order. Adaptive designs can be implemented in experiments that are observed in a series of sequential runs and are characterized by the use of the data accrued from preceding runs to determine the design of the current run.

1.1. Robust Adaptive Designs Adaptive designs have been employed in experiments where the CRLB depends on the unknown parameters. In such cases \textit{a priori} designs are
only efficient locally at the true value of the unknown parameters. As an illustration, consider an experiment with two treatments, denoted 1 and 2, with means $\eta_1$ and $\eta_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, respectively. Neyman allocation states that an a priori design that allocates $\rho = \sigma_1/(\sigma_1 + \sigma_2)$ proportion of the observations to treatment 1, and the remaining observations to treatment 2, is efficient for estimating the difference in the means. If $\rho$ is unknown then one could guess its value and use a design that is efficient with respect to this guess. This strategy can be inefficient if the guess is far from reality. An adaptive design that assigns observation $(j + 1)$ to treatment 1 with probability $\hat{\rho}(j)$ is a robust alternative to guessing, where $\hat{\rho}(j)$ is an estimate of $\rho$ based on the first $j$ observations [Hu and Rosenberger (2003)]. Designs that adaptively estimate the parameters to produce robust experiments have received significant attention [Box and Hunter (1965); Lai and Robbins (1979); Fedorov and Leonov (2013); Lane, Yao and Flounoy (2014); Kim and Flourny (2014)].

While the described adaptive design is more robust than guessing, it is not more efficient than the a priori design that allocates $\rho$ observations to treatment 1. Even though this ideal design is unknown, and cannot be used in practice, it is included in the set of a priori designs. In contrast to robust adaptive designs, the adaptive design framework developed in this work incorporates ancillary statistics into the design and is more efficient than any a priori design. Ancillary statistics, by definition, contain no information about the model parameters; however, certain ancillary statistics impact the quality of the parameter estimates [Fisher (1934); Cox (1958); Efron and Hinkley (1978)]. Fisher (1934), Basu (1969) and Ghosh, Reid and Fraser (2010) refer to such ancillary statistics as a reference set or relevant subset - a subset of the sample space on which inference should be restricted. Post-data conditional inference on relevant subsets has been thoroughly investigated [Fisher (1934); Cox (1958); Basu (1969); McCullagh (1992); Efron and Hinkley (1978); Ghosh, Reid and Fraser (2010)]. An illustrative example is provided in Section 2.

1.2. Summary of the Contributions Recently, Lane (2019, 2020, 2021) proposed designs that incorporate the relevant subset, referred to as relevant subset designs (RSDs), into an optimal design framework. The current work significantly extends the literature on RSD, and more broadly experimental design, in several key ways.

In Section 2 a lower bound is derived from the relevant subset. This alternative bound is tighter than the CRLB and provides a metric to evaluate the efficiency of RSDs. In Section 3 a novel RSD is developed with the expressed intent of reducing the proposed lower bound. This approach places limited restrictions on the experimental design. The setting could correspond to a factorial, optimal, minimax, randomized design, etc. Additionally, there is limited restriction on the method used to estimate $\theta$. This represents a significant extension of the existing literature on RSDs which has currently only considered the maximum likelihood estimate (MLE) in the context of optimal design.

In Section 4 the theoretical benefits of the proposed RSD are presented. The primary result is that given any a priori design the proposed RSD increases the efficiency in Loewner order. This is a significant generalization of the theoretical benefits associated with existing RSDs which have only been shown to be efficient, primarily through heuristics and simulations, with respect to a single convex optimality criterion. It should be noted that this is all done in the context of a general nonlinear model framework. In Section 6 a simulation study is conducted and it is found that the proposed RSD is more efficient than pertinent alternatives in every case considered.

2. Lower Bounds on Variance One of the most common experimental objectives is to characterize the relationship between a sequence of univariate responses, $y = [y(1), \ldots, y(n)]^T$, and a set of corresponding explanatory variables $x(1), \ldots, x(n)$. In this
work the set of explanatory variables, denoted \( \xi = [x(1), \ldots, x(n)] \), is assigned by the experimenter and is referred to as the design where \( x(j) \) is a \( s \times 1 \) dimensional vector for \( j = 1, \ldots, n \). This relationship is often assessed using an additive error regression model of the form

\[
y(j) = \eta(x(j), \theta) + \varepsilon(j),
\]

where \( \eta(x, \theta) \) is a, potentially nonlinear, function that relates \( x \) to \( y \) through the \( p \)-dimensional parameter of interest \( \theta \) subject to an error term \( \varepsilon \).

Let \( X = (x_1, \ldots, x_d)^T \) denote the unique levels of the explanatory variables used in an experiment and \( w = (w_1, \ldots, w_d)^T \) represents the corresponding allocation weights, where \( w_i = n_i/n, n_i = \sum_j t_i(j), t_i(j) \) is an indicator variable such that \( t_i(j) = 1 \) if \( x(j) = x_i \) and 0 otherwise and \( n = \sum_i n_i \).

**Definition 1.** For a given design space, \( \mathcal{X} \), a design \( \xi \) is an \( n \)-point a priori deterministic design if (1) \( x_i \in \mathcal{X}, \) where \( \mathcal{X} \) is a compact subset of \( \mathbb{R}^s \); (2) \( 0 \leq w_i \leq 1, \sum w_i = 1 \) and \( nw_i \) are integers for all \( i = 1, \ldots, d < \infty \); and (3) \( n \) is a pre-specified constant.

Let the set \( \Xi_n \) represent the collection of all \( n \)-point a priori deterministic designs. As outlined in the introduction the set of a priori designs includes both deterministic and data independent random strategies. For the latter, the design used in the experiment, \( \xi \), is an observed realization from a probability measure \( \pi \) with corresponding probability space \( (\mathcal{X}, \Xi_n, \pi) \). The probability measure \( \pi \) is an a priori random design if \( \pi(\xi|\mathcal{Y}) = \pi(\xi) \), where \( \mathcal{Y} = [\mathcal{Y}(1), \ldots, \mathcal{Y}(n)]^T \) is the random response vector. Calligraphic capital letters indicate random variables; whereas, roman letters correspond to observed realizations. This is a deviation from the standard practice of using capital roman letters to indicate random variables. This notation was implemented to distinguish random variables from observed matrices, which are indicated with bold capital roman letters, e.g. \( X \). The term a priori references the fact that, due to their independence of the responses, they can be specified in advance of the experiment. The set of deterministic designs is a subset of the randomized designs; i.e. any deterministic design \( \xi \in \Xi_n \) can be obtained by setting \( \pi \) equal to \( \delta_{\xi} \), where \( \delta_{\xi} \) is a measure that assigns unit mass to \( \xi \).

2.1. Cramér-Rao Lower Bound A critical question for an experimenter is to determine a design that produces efficient estimates of \( \theta \). Due to its connection to the CRLB the Fisher information in the sample is an important quantity in evaluating this objective.

In the context of model (1) the Fisher information in the sample on \( \theta \) can be constructed from the elemental information. Atkinson et al. (2014) define the elemental information as the information from a single observation associated with a distribution in its standard form, as opposed to with respect to the parameterization induced by the inclusion of explanatory variables. Let \( L(\eta|y) = f(y|\eta), l_y = \log L(\eta|y), \hat{I}_y = (\partial/\partial \eta) \log L(\eta|y), \dot{I}_y = (\partial^2/\partial \eta^2) \log L(\eta|y) \) and \( i_y = -\frac{1}{\dot{I}_y}, \) where \( f(y|\eta) \) is the probability density function (PDF) of \( y \). Consider the following condition.

**Condition 1.** (1) \( \varepsilon = [\varepsilon(1), \ldots, \varepsilon(n)]^T \) is a sequence of independent and identically distributed random variables; (2) \( \Theta \in \Theta \), where \( \Theta \) is an open set in \( \mathbb{R}^p \); (3) \( (\partial/\partial \theta)\eta(x, \theta) \) and \( (\partial^2/\partial \theta^2)\eta(x, \theta) \) exist for all \( x \in \mathcal{X} \); (4) \( (\partial/\partial \eta)f(y|\eta) \) and \( (\partial^2/\partial \eta^2)f(y|\eta) \) exist; and (5) \( (\partial/\partial \eta) \) can be passed twice under the integral sign in \( \int f(y|\eta)dy \) and once in \( \int \theta f(y|\eta)dy \).

Under Condition 1 the elemental information \( \mu = \mathbb{E}[i_y] \) is non-negative. For clarity of exposition \( \mu \) is a shared term for all observations. Cases where this is violated can often
be transformed into the current setting. For example, suppose the elemental information is \( \mu s(x) \), where \( s(x) \) is a known function of \( x \); multiplying \( y \) and \( \eta(x, \theta) \) by \( s^{1/2}(x) \) reduces to the current setting.

Let \( \eta_i = \eta(x_i, \theta) \), \( L(\eta_i|y, t_i) = \prod_{j=1}^n L[\eta_i|y(j)]t_i(j), l_{y, t_i} = \sum_{j=1}^n t_i(j)l_{y(j)} \) and \( l_{y, t_i} = \sum_{j=1}^n t_i(j)d_i(j) \). The observed information on \( \eta_i \) is \( i_{y, t_i} = \sum_{j=1}^n t_i(j)d_i(j) \), where \( t_i = [t_i(1), \ldots, t_i(n)]^T \). Note, observations with \( t_i(j) = 0 \) do not contribute to \( i_{y, t_i} \). The Fisher information in the sample, for a given the design \( \xi \), on \( \eta_i \) is \( \mathbb{E}[\mathcal{I}_{\mathcal{Y}, \mathcal{T}}|\xi] = n\mu_i \). The preceding expectation implicitly assumes \( \xi \) is an a priori deterministic design; the extension of the results in this work to random design strategies is considered in Sections 3.2 and 4.1.

Switching to the Fisher information in the sample on \( \theta \), denote the likelihood of \( \theta \) as \( L(\theta|y, T) = \prod_{i=1}^d L(\eta_i|y, t_i) \). The observed information on \( \theta \) is

\[
I_{y, T}(\theta) = -\frac{\partial^2}{\partial \theta^2} \log L(\theta|y, T) = \sum_{j=1}^n \left[ i_{y, t_i} \frac{\partial \eta(x_i, \theta)}{\partial \theta} \frac{\partial \eta(x_i, \theta)}{\partial \theta^T} - i_{y, t_i} \frac{\partial^2 \eta(x_i, \theta)}{\partial \theta^2} \right],
\]

where \( T = (t_1, \ldots, t_d)^T \) is observed allocation matrix. Under Condition 1 the Fisher information in the sample on \( \theta \)

\[
F_\xi(\theta) = \mathbb{E}[I_{\mathcal{Y}, \mathcal{T}}(\theta)|\xi] = n\mu \sum_{i=1}^d w_i \frac{\partial \eta(x_i, \theta)}{\partial \theta} \frac{\partial \eta(x_i, \theta)}{\partial \theta^T},
\]

where, in the present notation \( \mathcal{Y} \) and \( \mathcal{T} \) are the random response vector and allocation matrix, respectively. The dependence of \( F_\xi \) on \( \theta \) will be omitted when the meaning is clear. Note that the design, \( \xi \), completely determines \( \mathcal{T} \) and that it is assumed that the design \( \xi \) results in a positive definite \( F_\xi \).

The most commonly used lower bound on the variance of an estimate, say \( \hat{\theta} \), with finite expectation \( b(\theta) = \mathbb{E}[\hat{\theta}|\xi] \) is the Cramér-Rao Lower Bound (CRLB) which states, under Condition 1, that

\[
\text{Var}[\hat{\theta}|\xi] \geq \frac{\partial b(\theta)}{\partial \theta} F_\xi^{-1} \frac{\partial b(\theta)}{\partial \theta^T}.
\]

In many settings asymptotically unbiased estimators are of interest. For example, the MLE is unbiased to the order \( O(n^{-1}) \) which implies that \( (\partial/\partial \theta)b(\theta) = I_p(1 + O(n^{-1})) \), where \( I_p \) is the \( p \times p \) identity matrix. Bias correction methods can be implemented to further reduce the bias if desired [Hall and Martin (1988); Firth (1993); Pronzato and Pazman (1994); Kosmidis and Firth (2009); Cordeiro and Cribari-Neto (2014)]. Since, the bias has a negligible impact, as \( n \) gets large, a common approach in the design of experiments is to consider the unbiased version of the CRLB which states that if \( \mathbb{E}[\hat{\theta}|\xi] = \theta \) then

\[
\text{Var}[\hat{\theta}|\xi] \geq F_\xi^{-1}.
\]

Unless explicitly stated otherwise all future references to the CRLB references (2). The CRLB has been a significant motivator for designing experiments that are efficient with respect to \( F_\xi^{-1} \). The conventional wisdom being that reducing the lower bound improves estimation and inference following an experiment.

### 2.2. A Relevant Subset Lower Bound

This section develops an alternative lower bound based on the relevant subset. The relevant subset is embedded in the definition of sufficiency. Suppose \( (\theta, A) \) is a sufficient statistic, where \( \theta \) is an estimate of \( \theta \) and \( A \) is an ancillary random variable. In such cases \( \theta \) is not sufficient and inference based on \( \theta \) alone results in a loss of information. Fisher (1934) argued that this information is recovered by conditioning
on the relevant subset, the subset of the sample space where $A = A$, and by extension that the Fisher information in the relevant subset, derived below, is more meaningful than the Fisher information in the sample. Recall, in the current notation $A$ is a random ancillary variable and $A$ is its observed realization.

Analogous to the derivation of $F_{\xi}$, the Fisher information in the relevant subset on $\theta$ is constructed from the Fisher information in the relevant subset on $\eta_i$. Let $f(y_i|a_i, \eta_i)$ be the conditional probability density of $y_i = \{y(j): t_i(j) = 1, j = 1, \ldots, n\}$, where $a_i$ is an ancillary random variable constructed from the data corresponding to the support point $x_i$. For this section it is required to replace Condition 1 with the more restrictive condition stated below.

**CONDITION 2.** Condition 1 (1-4); (5') $(\partial/\partial \eta_i)$ can be passed twice under the integral sign in $\int f(y_i|a_i, \eta_i)dy_i/a_i$ and once in $\int \phi(y_i|a_i, \eta_i)dy_i/a_i$ for all $a_i \in \mathcal{A}_i$, where $\mathcal{A}_i$ is the support of $a_i$, $i = 1, \ldots, d$; (6) $b_A(\theta) = E[\theta|A = A, \xi] = \theta$ for all $A \in \mathcal{A}$, where $\mathcal{A}$ is the support of $A = (a_1, \ldots, a_d)^T$; and (7) $F_{\xi}$ is positive definite.

The Fisher information in the relevant subset on $\eta_i$ is $h_{\mathcal{A}_i} = E[\mathcal{I}_Y, T|A_i, \xi]$ and the Fisher information in the relevant subset on $\theta$ is

$$H_A(\theta) = E[I_{Y,T}(\theta)|A, \xi] = \sum_{i=1}^{d} h_{A_i} \left( \frac{\partial\eta(x_i, \theta)}{\partial \theta} \frac{\partial\eta(x_i, \theta)}{\partial \theta^T} \right),$$

where $A = (A_1, \ldots, A_d)^T$ is the random ancillary matrix with observed realization $A = (a_1, \ldots, a_d)^T$. To derive (3) the fact that $E[I_{Y,T}|A_i, \xi] = 0$ was used; see the proof of Proposition 1 for details [supplemental materials]. The dependence of $H_A$ on $\theta$ is omitted when the meaning is clear. See Fisher (1934), Basu (1969), Ghosh, Reid and Fraser (2010) and Lane (2021) for additional details regarding the Fisher information in the sample versus the Fisher information in the relevant subset. Illustrative examples are provided later in this section to help differentiate the measures of information.

An argument in favor of $F_{\xi}$ is that its inverse forms a lower bound on the variance of estimates of $\theta$. A similar argument for $H_A$ is the following conditional CRLB.

**PROPOSITION 1.** Under Condition 2 $\text{Var}[\hat{\theta}|A, \xi] \geq H_A^{-1}$ for all $\xi \in \Xi_n$.

See the supplemental materials for proof. Noting that $\text{Var}[\hat{\theta}|A, \xi] \geq F_{\xi}^{-1}$ is not guaranteed reinforces the value of $H_A$ from a conditional perspective. Historically, the literature regarding conditioning on the relevant subset has primarily focused on post-data conditional inference [Fisher (1934); Cox (1958); Basu (1969); McCullagh (1992); Efron and Hinkley (1978); Ghosh, Reid and Fraser (2010)]. Before an experiment is conducted the relevant subset is unknown and conditional properties are an insufficient design motivator. It is more important to assess a designs unconditional properties; e.g. does a design improve the unconditional variance of the estimates of $\theta$. To this end Proposition 1 is used to derive the following tighter lower bound on the unconditional variance of an estimate of $\theta$.

**PROPOSITION 2.** Under Condition 2

$$\text{Var}[\hat{\theta}|\xi] \geq E[H_A^{-1}|\xi] \geq F_{\xi}^{-1}$$

for all $\xi \in \Xi_n$. 
See the supplemental materials for proof. The term\(E[H_{\mathbf{A}}^{-1} | \xi]|\) will be referred to as the relevant subset lower bound (RSLB). Proposition 2 shows that the RSLB is tighter, in Loewner order, than the CRLB. A notable case where equality is obtained for both signs in (4) is the normal linear model with constant variance.

**Remark 1.** The condition that \(b_{\mathbf{A}}(\theta) = \theta\) for all \(\mathbf{A}\) is stronger than the condition that \(v(\theta) = \theta\) required for the CRLB. As was the case for the traditional CRLB it is reasonable to expect that most estimates of interest have negligible conditional bias. This discussion suggests that \(E[H_{\mathbf{A}}^{-1} | \xi]\) is generally the dominant term in the RSLB just as \(F_{L}^{-1}\) is the dominant term in the CRLB.

**Remark 2.** There exist tighter alternatives to the CRLB [Bhattacharyya (1948); van Trees (1968); Bobrovsky, Mayer-Wolf and Zakai (1987), and others]. A search for the bound (4) in the existing literature did not yield any similar expressions. Due to the simplicity of its derivation it is possible, perhaps even likely, that its existence has been noted. One potential reason (4) may not be well known that a closed form expression for the distribution of \(H_{\mathbf{A}}^{-1}\) may not exist which makes computing its expectation a challenge. This is examined in detail in Section 2.4. Conversely, the CRLB is relatively straightforward to compute.

In light of Remark 2 it is important to note that the RSLB is not presented as a post-data alternative to the CRLB. Instead, it is meant to demonstrate the connection between efficiency and the relevant subset and, more importantly for the current work, that the RSLB represents a theoretical benchmark for assessing the quality of candidate designs based on the conventional wisdom that improving a tight lower bound on the variance improves inference. The value of the RSLB as a benchmark is more precisely described in Section 3.

The remainder of this section derives the relevant subset, the information measures and the lower bounds in two distinct settings. The first setting is an illustrative example where the MLE attains the RSLB but not the CRLB. This first example provides intuition for the ancillary error variance \(A(j)\). Specifically, \(\mathcal{Y}(j) \mid A(j) = a(j) \sim N[\eta(j), 1/a(j)]\), where \(\eta(j) = f^{T}(x(j))\theta\) is restricted to be linear with respect to \(\theta\) for \(j = 1, \ldots, n\). The log-likelihood of \(\eta_{i}\) is

\[
(5) \quad l(\eta_{i} | y, t_{i}) \propto \sum_{j=1}^{n} t_{i}(j)a(j)[y(j) - \eta_{i}]^{2}.
\]

From its definition the Fisher information in relevant subset on \(\eta_{i}\) is \(h_{a_{i}} = \sum_{j} t_{i}(j)a(j)\), where \(a_{i} = \{a(j) : t_{i}(j) = 1, j = 1, \ldots, n\}\). The MLE, \(\hat{\eta}_{i}\), is the value of \(\eta_{i}\) that minimizes (5) and can be obtained via a weighted least squares approach as \(\hat{\eta}_{i} = h_{a_{i}}^{-1} \sum_{j} t_{i}(j)a(j)y(j)\). The MLE is a linear combination of conditionally independent normal random variables which implies that \(\hat{\eta}_{i}|a_{i}\) is normally distributed with mean \(\eta_{i}\) and variance \(h_{a_{i}}^{-1}\).

To extend this to \(\theta\), note that \(\log\ L(\theta | y, T) \propto \sum_{i} h_{a_{i}}[\eta_{i} - \theta^{T}f(x_{i})]^{2}\) and the MLE of \(\theta\) is also the solution to a weighted least squares problem; i.e. \(\hat{\theta} = H_{\mathbf{A}}^{-1}F^{T}K_{\mathbf{A}}\hat{\eta}\), where \(\hat{\eta} = (\hat{\eta}_{1}, \ldots, \hat{\eta}_{d})^{T}\), \(\mathbf{A} = (a_{1}, \ldots, a_{d})^{T}\), \(K_{\mathbf{A}} = \text{diag}(h_{a_{1}}, \ldots, h_{a_{d}})\), \(F\) is a \(d \times p\) matrix with \(i\)th row \(f^{T}(x_{i})\) and \(H_{\mathbf{A}} = F^{T}K_{\mathbf{A}}F\). Since \(\hat{\theta}\) is a linear combination of conditionally normal random variables it follows that \(\hat{\theta}|\mathbf{A}\) is normally distributed with mean \(\theta\) and variance \(H_{\mathbf{A}}^{-1}\). This directly implies that \(\text{Var}[\hat{\theta}|\xi] = E[H_{\mathbf{A}}^{-1} | \xi] \geq F_{\xi}^{-1}\) which confirms that the MLE attains the RSLB but not the CRLB.
2.3.1. Example: Normal Inverse-Gamma  
Return to the two treatment example in Section 1 and let 
\( A(j) \sim \text{Gamma}(\alpha, \beta) \), where \( \alpha, \beta > 0 \). In the notation of model (1) \( f^T(x_i) = x_i, \eta_i = x_i^T \theta \) and the design space \( \mathcal{X} \) contains two points \( x_1 = (1,1)^T \) and \( x_2 = (1,0)^T \) corresponding to treatments 1 and 2, respectively. Under the present scenario, \( h_{\alpha} \) is the sum of gamma distributed random variables which implies that \( h_{\alpha} \sim \text{Gamma}(nw;\alpha, \beta) \) and \( h_{\alpha}^{-1} \sim \text{Inverse-Gamma}(nw;\alpha, \beta) \). From the definition of the normal inverse-gamma (NIG) distribution it can be shown that \((\bar{\eta}_i, h_{\alpha}^{-1}) \sim \text{NIG}(\bar{\eta}_i, nw;\alpha, \beta) \) [see Gelman et al. (2014) ch. 2]. The MLE of \( \Delta = \eta_1 - \eta_2 \), denoted \( \hat{\Delta} = \bar{\eta}_1 - \bar{\eta}_2 \), has mean \( \mathbb{E}[\hat{\Delta}|\xi] = \Delta \) and variance \( \text{Var}[\hat{\Delta}|\xi] = \beta[1/(\alpha nw_1 - 1) + 1/(\alpha nw_2 - 1)] \).

Recall from Section 1 that \( \rho \) represents the efficient allocation proportion. In the current setting the treatment variances are equal an \( \rho = 1/2 \); i.e. the design \( \xi_\rho \) that assigns an equal number of observations to each treatment (i.e. \( w_1 = w_2 = 1/2 \)) is efficient. In this illustration \( \rho \) is known and there is no need to estimate it using the adaptive procedure described in the introduction. For the balanced design \( \text{Var}[\hat{\Delta}|\xi_\rho] = 2 \beta/(n\alpha/2 - 1) \). Experiments require \( n > 2/\alpha \) observations in order to ensure a finite variance. As shown above, the variance of \( \hat{\Delta} \) and the RSLB coincide, i.e. \( \text{Var}[\hat{\Delta}|\xi_\rho] = \mathbb{E}[H_{\bar{\Delta}}^{-1}|\xi_\rho] \). One interesting feature is that \( F_{\xi_\rho}^{-1} = \{\mathbb{E}[H_{\bar{\Delta}}|\xi_\rho]\}^{-1} = \beta/(n\alpha) \) is always finite, regardless of sample size.

2.4. General Setting  
One common condition placed on additive error regression models, stated in (1), is that the errors are from the location family. The location family is characterized by the relation \( f(y|\eta) = f(\varepsilon|0) \) and it is well established that there exists and exact ancillary, \( A = (a_1, \ldots, a_d)^T \), where \( a_i = [y_{i(2)} - y_{i(1)}, \ldots, y_{i(n)} - y_{i(n-1)}]^T \) and \( y_{i(j)} \) is the \( j \)th order statistic of \( y_i \). This section outlines the advantages of the location condition for the current work.

At the beginning of this section the Fisher information in the relevant subset on \( \theta, H_A \), was constructed from the Fisher information in the relevant subset on \( \eta_i, h_{\alpha} \). Instead of having to work with the \( p \)-dimensional parameter, \( \theta \), each of the one-dimensional parameters, \( \eta_1, \ldots, \eta_d \), can be examined separately. Let \( l(\eta_i|\hat{\eta}_i) = l(\eta_i|y,t_i) = l(\hat{\eta}_i|y,t_i); \) Fisher (1934) famously demonstrated that

\[
p(\hat{\eta}_i|a_i, \eta_i) = c_{\alpha_i} e^{l(\eta_i, \hat{\eta}_i)},
\]

where \( c_{\alpha_i} \) is a normalizing constant that ensures the right hand side integrates to one. It is important to point out that it is not necessary to compute the ancillary statistic for the methods proposed in this work. What is required is that \( h_{\alpha_i} = \mathbb{E}[i_{y,t}|a_i, \xi] \) can be computed using analytical or numerical methods; examples are provided below to illustrate a useful approach to obtain this quantity. Once \( h_{\alpha_i} \) has been computed \( H_A \) can be computed using (3).

A second advantage of the location setting is that it allows for an invariant form of information to be considered. Denote the invariant observed information as \( i_{y,t}^{(\phi)} = i_{y,t}/\mu \), where the superscript \( (\phi) \) is used to recognized that it is invariant to any one-to-one transformation, \( \phi(\eta_i) \). Noting that (6) is invariant with respect to one-to-one transformations [Barndorff-Nielsen and Cox (1994) ch 6] yields

\[
g_{\alpha_i} = \mathbb{E}[i_{y,t}^{(\phi)}|a_i, \xi] = \mu h_{\alpha_i},
\]

where \( g_{\alpha_i} \) is the invariant information in the relevant subset. Two practical examples are provided below. The design proposed in the next section depends only on \( g_{\alpha_1}, \ldots, g_{\alpha_d} \) and is thus invariant to one-to-one transformations of \( \eta_i \).

**Remark 3.** The challenge with computing \( \mathbb{E}[H_A^{-1}|\xi] \), alluded to in Remark 2 can now be made more explicit. A closed form expression for the distribution of the inverse of \( H_A \)
likely does not exist, which makes obtaining its expectation via analytical means impossible. Monte Carlo methods can be used in place of exact solutions, but such methods are computationally expensive. As previously stated, this discussion indicates that the RSLB may not be a suitable replacement to the CRLB as a post-data variance lower bound. One critical feature of the adaptive design proposed in Section 3 is that it does not require $E[H\xi^{-1}\xi]$ to be computed.

2.4.1. Example: Generalized Normal The generalized normal distribution (GND) has PDF $f(y|\eta) \propto e^{-\zeta|y-\eta|^2}$ with $y, \eta \in \mathbb{R}$ and $\zeta, \tau > 0$. Special cases include the double exponential distribution for $\zeta = 1$, the normal distribution for $\zeta = 2$ and the uniform distribution as $\zeta \to \infty$. The log-likelihood of $\eta$ is $l(\eta|y,t_i) = -\zeta^{-1}\sum_{j=1}^n t_i(j)[|y(j) - \eta_j|/(\zeta \tau)]^\zeta$ and the observed information from subject $j$ is $i_{y(j)}(\eta(j)) = (\zeta - 1)[|y(j) - \eta(j)|/\tau]^\zeta - 2/\tau^2$ which has expectation $\mu = \zeta^2(1-1/\zeta)\Gamma(2 - 1/\zeta)/[2\tau^2\Gamma(1/\zeta)]$, where $\eta(j) = \eta(x(j), \theta)$.

To compute $h_{\alpha_i}$ use the change of variable, $v = \eta_i - \eta_i$ to obtain
\[ h_{\alpha_i} = \int_{y,t_i, (\eta_i - v)e^{\hat{\eta}_i - v, \hat{\eta}_i}} dv / \int_{y,t_i} e^{\hat{\eta}_i - v, \hat{\eta}_i} dv \]

which can be computed using numeric integration. The dependence of $i_{y,t_i}(\eta_i)$ on $\eta_i$ has been made explicit for clarity. The invariant measure is recovered by dividing by $\mu$; i.e., $g_{\alpha_i} = h_{\alpha_i}/\mu$. This assumes that $\mu$ is known. In the event that $\mu$ is unknown it can be replaced with its MLE $\hat{\mu}$.

2.4.2. Example: Cauchy The Cauchy distribution has PDF $f(y|\eta) \propto 1/[1 + (y - \eta)/\tau^2]$ for $y, \eta \in \mathbb{R}$ and $\tau > 0$ which implies that $l(\eta|y,t_i) = -\sum_{j=1}^n t_i(j)\log[1 + u^2(j)/\tau^2], i_{y(j)} = 2[\tau^2 - u^2(j)/\tau + u^2(j)]^2$ and $\mu = 1/(2\tau^2)$, where $u(j) = y(j) - \eta(j)$. The steps from the preceding example can be repeated to obtain $g_{\alpha_i}$.

3. Relevant Subset Design As stated previously, the RSLB is difficult to compute and may not be widely accepted as a post-data variance bound replacement for the CRLB. In this section it is shown that the RSLB has significant value from a design perspective. The following additional condition is required for the adaptive design proposed in this work.

CONDITION 3. (1) The error distribution is a member of the location family; (2) $\gamma^2 = \text{Var}[\hat{\theta}_j] < \infty$, where $\gamma = [(\nu_{20}/\nu_{20} - \nu_{11})/\nu_{20}]^{1/2}$ and $\nu_{kl} = E[l_k l_j + E[l_k^2]/|\eta_i - \eta_i|^2]$ (3) $E[(\hat{\eta}_i - \eta_i)^2] < \infty$ (4) $E[(\hat{\theta}_j - \eta_j)^2] < \infty$ (5) $(\partial^k / \partial \eta^k) \log f(y|\eta)$ is bounded for $k = 1, \ldots, 4$; and $4$ $(\partial^3 / \partial w_i w_j w_k) M_{\xi}^{-1}$ is bounded for all $i,j,k = 1, \ldots, d$, where $M_{\xi} = F_{\xi}/(\mu\nu)$.

Historically, designs that minimize the CRLB, i.e. “minimize” $F_{\xi}^{-1}$, have been considered efficient; note, $F_{\xi}^{-1}$ is a matrix and cannot be minimized in general. As described in Section 2 the CRLB is not necessarily tight. In order to describe the efficiency guaranteed by designs that are efficient with respect to the CRLB consider the following Proposition.

PROPOSITION 3. Under Condition 2 and 3
\[ \lim_{n \to \infty} n^2 e^{T} (\text{Var}[H\xi^{-1}\xi] - F_{\xi}^{-1}) e \mu \lim_{n \to \infty} n^{-1} \text{tr} \{ D (\text{Var}[u_{\xi}] + \text{Var}[v_{\xi}] ) \} \]

for all $\xi \in \Xi$, where $u_{\xi} = (u_{a_1}, \ldots, u_{a_d})^T, u_a = g_a - w_i \sum_{i=1}^d g_a, v_{\xi} = (v_{a_1}, \ldots, v_{a_d})^T, v_a = w_i (\sum_{i=1}^d g_a - n), D$ is a $d \times d$ symmetric matrix with ijth element equal to $r_i(c)r_j(c) s_{ij}, r_i(c) = \hat{\eta}^T(x_i) M_{\xi}^{-1} c, s_{ij} = \hat{\eta}^T(x_i) M_{\xi}^{-1} \hat{\eta}(x_j)$ and $\hat{\eta}(x_i) = (\partial / \partial \theta) \eta(x_i, \theta)$.
In the proof of Proposition 3, provided in the supplemental materials, it is shown that \( n^{-1}\text{Var}[\mathbf{v}_A|\xi] \rightarrow \gamma^2 \mathbf{w} \mathbf{w}^T \) and \( n^{-1}\text{Var}[\mathbf{u}_A|\xi] \rightarrow \gamma^2 \mathbf{W}(\mathbf{I}_d - \mathbf{w} \mathbf{w}^T) \) as \( n \rightarrow \infty \), where \( \mathbf{W} = \text{diag}(\mathbf{w}) \). This implies that the right hand side of (7) converges to a constant as \( n \rightarrow \infty \).

Propositions 2 and 3 indicate that a design that is efficient with respect to \( \mathbf{F} \) is also provided a motivation for an adaptive design. The difference between \( \mathbf{F} \) gives a rigorous and adaptive approach is derived that eliminates the variability associated with \( \mathbf{F} \). This definition is used to confirm the main result of this work, that a design \( \xi_1 \in \Xi_n \) is first order efficient if there does not exist a design \( \xi_2 \in \Xi_n \) such that

\[
nc^T[\mathbf{F}^{-1}_{\xi_1} - \mathbf{F}^{-1}_{\xi_2}]c > 0
\]

for all nonzero \( c \). Note \( c^T(\mathbf{A} - \mathbf{B})c > 0 \) for all nonzero \( c \) implies \( \mathbf{A} \succ \mathbf{B} \) in Loewner order. Not all designs satisfying this definition of efficiency have practical value; however, if efficiency is an objective then the design used in the experiment should.

Intuitively, a designs efficiency is improved if it minimises the RSLB, i.e. “minimizes” \( E[\mathbf{H}^{-1}_A|\xi] \). This in combination with Proposition 3 motivates a higher order definition of efficiency; specifically, a design \( \xi_1 \in \Xi_n \) is not second order efficient if there exists a design \( \xi_2 \in \Xi_n \) such that

\[
\lim_{n \rightarrow \infty} n^2 c^T(E[\mathbf{H}^{-1}_A|\xi_2] - E[\mathbf{H}^{-1}_A|\xi_1])c > 0
\]

for all nonzero \( c \). This is a reversal of the previous definition. Equation (8) describes a method to determine if a design is first order efficient; conversely, (9) describes a method to determine if a design is not second order efficient. Despite its limitation (9) is a useful tool to eliminate inefficient designs. This definition is used to confirm the main result of this work, that no a priori design is second order efficient.

Proposition 3 also provides a motivation for an adaptive design. The difference between the RSLB and CRLB, is dominated by the variances of \( \mathbf{u}_A \) and \( \mathbf{v}_A \). In the next section an adaptive approach is derived that eliminates the variability associated with \( \mathbf{u}_A \); thus reducing the gap between the RSLB and CRLB and increasing efficiency. Section 4 gives a rigorous treatment of this result. It is important to note that \( \mathbf{u}_A \) and \( \mathbf{v}_A \) are functions of the invariant information in the relevant subset \( g_{a_1}, \ldots, g_{a_d} \).

3.1. Randomized Relevant Subset Design A sequential experiment consists of a series of runs, \( r = 1, \ldots, R_{\text{max}} \), where each run has a potentially random number, \( \mathcal{N}\{r\} \), of independent observations. While the total number of runs is random the total sample size, \( n \), is fixed. The notation \( \{r\} \) indicates a variable corresponds to run \( r \); whereas \( \{1; r\} \) corresponds to a variable that incorporates the data from all runs up to and including run \( r \). The curly bracket is used to distinguish the \( r \)th run from the \( j \)th observation, which is denoted with a parentheses. For example, \( y(j) \) is the response of observation \( j \); \( y\{r\} = [y(\mathcal{N}\{1; r - 1\} + 1), \ldots, y(\mathcal{N}\{1; r\})]^T \) are the responses of the observations in run \( r \); and \( y\{1; r\} = [y^T\{1\}, \ldots, y^T\{r\}]^T \) are the responses from all observations from the first \( r \) runs, where \( \mathcal{N}\{1; r\} = \sum_{s=1}^{r} \mathcal{N}\{s\} \) is the number of observations in the first \( r \) runs. As a second example, \( i_{y(j)}^{(\phi)} \) is the invariant observed information for the \( j \)th observation:

\[
i_{y(r), t.\{r\}}^{(\phi)} = \sum_{j=\mathcal{N}\{1; r - 1\}+1}^{\mathcal{N}\{1; r\}} t_{i(j)}^{(\phi)} i_{y(j)}^{(\phi)}
\]

is the invariant observed information from run \( r \); and \( i_{y\{1; r\}, t.\{1; r\}}^{(\phi)} = \sum_{s=1}^{r} i_{y(r), t.\{r\}}^{(\phi)} \) is the total invariant observed information from all runs up to and including run \( r \).

Adaptive designs are characterized by the use of the data from the preceding \( r \) runs to determine the design of run \( r + 1 \). In this work a method is proposed that uses the relevant
subset from the first \(r\) runs to determine the probability of assigning each observation in run \(r + 1\) to the \(i\)th support point, denoted \(\mathcal{P}_i\{r + 1\} = P\{\mathcal{T}_i(j) = 1|\mathcal{A}\{1;r\}\}\), and the sample size of run \(r + 1, \mathcal{N}\{r + 1\}\). As described \(\mathcal{P}_i\{r + 1\}\) and \(\mathcal{N}\{r + 1\}\) are \(\mathcal{A}\{1;r\}\) measurable. To develop the proposed design, consider the first two runs of experiment where \(\mathcal{N}\{1\} = n\{1\}\) and \(\mathcal{N}_i\{1\} = w_i n\{1\}\) are pre-specified constants and it is desired to determine the sample size of the second run, \(\mathcal{N}\{2\}\), and the probability of assigning each observation in the second run to the \(i\)th support point, \(\mathcal{P}_i\{2\}\).

The most important quantity in the proposed design is \(u_{\mathcal{A}}\), defined as \(u_{a_i} = g_{a_i} - w_i \sum_{j=1}^{d} g_{a_j}\) in Proposition 3, see Section 2.4 for details on its computation. For the derivation of the adaptive design proposed in this section the following alternative expression is informative

\[
u_{\mathcal{A},\{1;r\}} = E[S_i\{1;r\} | \mathcal{A}_i\{1;r\}, \xi],
\]

where \(S_i\{1;r\} = \sum_{s=1}^{r} S_i\{s\}, S_i\{r\} = \sum_{j=\mathcal{N}\{1;r\} - 1}^{\mathcal{N}\{1;r\}} s_{y(j)}\) and \(s_{y(j)} = \{T_i(j) - w_i\}^{(g)}\).

Consider the expectation of \(u_{\mathcal{A},\{1;2\}}\) conditional on \(\mathcal{A}\{1\}\)

\[
E[u_{\mathcal{A},\{1;2\}}|\mathcal{A}\{1\}, \xi] = E[S_i\{1\} + S_i\{2\} | \mathcal{A}\{1\}, \xi] = u_{\mathcal{A},\{1\}} + \sum_{j=n\{1\} + 1}^{\mathcal{N}\{1;2\}} E[s_{y(j)} | \mathcal{A}\{1\}, \xi]
\]

(10)

\[
= u_{\mathcal{A},\{1\}} + \mathcal{N}\{2\}(\mathcal{P}_i\{2\} - w_i)
\]

The first line is from the definition of \(u_{\mathcal{A},\{1;2\}}\) and the fact that \(\mathcal{A}_i\{1\} \subseteq \mathcal{A}_i\{1;2\}\) [see the proof of Lemma 1 for proof]; the second is obtained from the definitions of \(u_{\mathcal{A},\{1\}}\) and \(S_i\{2\}\) and by recognizing that \(\mathcal{N}\{1;2\}\) is \(\mathcal{A}_i\{1\}\) measurable; the third line follows from

\[
E[s_{y(j)} | \mathcal{A}\{1\}, \xi] = E[i^{(g)}_{y(j)} | \mathcal{T}_i(j)](P\{\mathcal{T}_i(j) = 1|\mathcal{A}\{1\}\} - w_i) = \mathcal{P}_i\{2\} - w_i
\]

for \(j = n\{1\} + 1, \ldots, \mathcal{N}\{2\}\).

Consider a careful selection of \(\mathcal{N}\{2\}\) and \(\mathcal{P}_i\{2\}\). Set the right hand side of (10) equal to 0 and solve for \(\mathcal{P}_i\{2\}\) to obtain

\[
\mathcal{P}_i\{2\} = w_i - u_{\mathcal{A},\{1\}}/\mathcal{N}\{2\}.
\]

As defined, \(\mathcal{P}_i\{2\}\) is a probability and it is implicitly required that \(\mathcal{P}_i\{2\} \geq 0\). This is only guaranteed if \(\mathcal{N}\{2\}\) is sufficiently "large". This is ensured by setting the right hand side of (11) to zero, solving to obtain \(\mathcal{N}_i\{2\} = u_{\mathcal{A},\{1\}}/w_i \) and defining

\[
\mathcal{N}\{2\} = \lceil \max\{\mathcal{N}_1\{2\}, \ldots, \mathcal{N}_d\{2\}\} \rceil,
\]

where \(\lceil \cdot \rceil\) is the ceiling operator. The ceiling operation ensures \(\mathcal{N}\{2\}\) is an integer and the max ensures that all \(\mathcal{P}_i\{2\} \geq 0\). Plugging this selection of \(\mathcal{N}\{2\}\) and \(\mathcal{P}_i\{2\}\) into (10) yields \(E[u_{\mathcal{A},\{1;2\}} | \mathcal{A}\{1\}] = 0\). To appreciate the significance of this accomplishment consider an application of law of total variance which states that

\[
Var[u_{\mathcal{A},\{1;2\}} | \xi] = Var[E[u_{\mathcal{A},\{1;2\}} | \mathcal{A}\{1\}, \xi] | \xi] + E[Var[u_{\mathcal{A},\{1;2\}} | \mathcal{A}\{1\}, \xi] | \xi] = E[Var[u_{\mathcal{A},\{1;2\}} | \mathcal{A}\{1\}, \xi] | \xi].
\]

(13)

The above states that using the proposed adaptive procedure forces the expectation of \(u_{\mathcal{A},\{1;2\}}\), conditional on \(\mathcal{A}\{1\}\), to equal zero which in turn reduces its variance. In fact as shown in the below lemma variance of \(u_{\mathcal{A},\{1;2\}}\) is eliminated for large \(n\{1\}\).
LEMMA 1. Under Condition 2 and $N(2)$ and $P_1(2)$ as defined above, if $n\{1\} \to \infty$ then
\[
\lim_{n\{1\} \to \infty} (n\{1\})^{-1} \text{Var}[u_{A(1;2)} | \xi] = 0.
\]

See section 8 for proof. Lemma 1 represents a significant improvement compared to a deterministic design where it was found that $(n\{1;2\})^{-1} \text{Var}[u_{A(1;2)} | \xi] = \gamma^2 W (I_d - \mathbf{w} \mathbf{w}^T)$. Recalling Proposition 3 this implies that the adaptive procedure considerably reduces the RSLB as compared to a a priori deterministic design.

The two-run demonstration describes the first step of a more general approach where this procedure is repeated until the total sample size, $n$, is exhausted. The general approach, described below, is referred to as the Randomized Relevant Subset Design (RRSD).

ALGORITHM 1. Randomized Relevant Subset Design (RRSD)

1. Determine an a priori design, $\xi \in \Xi_n$ with support $X = (x_1, \ldots, x_d)^T$ and allocation weights $\mathbf{w} = (w_1, \ldots, w_d)^T$.
2. Allocate the first $n\{1\}$, observations according to design $\xi$. If $n\{1\}w_i$ are integers for allocate exactly $n\{1\}w_i$ to treatment $i$; otherwise allocate with probability $P_i(1) = w_i$ to treatment $i$.
3. Compute $N_{\text{temp}}\{r + 1\} = \lfloor \max\{N_i\{r + 1\}, \ldots, N_d\{r + 1\}\} \rfloor$, where $N_i\{r + 1\} = u_{A_i(1;r)} / w_i$.
4. If $N_{\text{temp}}\{r + 1\} \leq n - N\{1; r - 1\}$ let $N\{r + 1\} = N_{\text{temp}}\{r + 1\}$ and $P_i\{r + 1\} = w_i - u_{A_i(1;r)} / N\{r + 1\}$; else, let $P_i\{r + 1\} = w_i$ and $N\{r + 1\} = n - N\{1; r - 1\}$.
5. Allocate observations $j = N\{1;r\} + 1, \ldots, N\{1;r + 1\}$ to the $i$th support point with probability $P_i\{r + 1\}$.
6. Repeat steps 3-6 until $N\{1;r + 1\} = n$.

The if-else condition in step 4 is required to ensure the adaptive design does not result in an experiment with a sample size greater than $n$. Let $\hat{\xi} = [\hat{x}(1), \ldots, \hat{x}(n)]^T$ represent the observed design following the RRSD with initial design $\xi$, where $\hat{x}(j)$ is the support point of observation $j$ assigned by algorithm. Additionally, the dependence on runs $\{1; R_{\text{max}}\}$ is often omitted for variables corresponding to all $n$ observations in the experiment, e.g. $A = A\{1; R_{\text{max}}\}$. An ancillary random variable is marked with a check, $\check{A}_\xi$, indicates that it corresponds to a RRSD, where the subscript $\xi$ is used to denote that fact that the corresponding RRSD was initialized by design $\xi$. Ancillary random variables without a check, $A_\xi$, corresponds to an experiment with deterministic design $\xi$. Estimates of $\theta$ are functions of the observed design, a superscript will be used to indicate this dependence; e.g. $\hat{\theta}_\xi$ and $\check{\theta}_\xi$ represent estimates following a deterministic design and the RRSD, respectively.

A note on the deterministic design, $\xi$, that initializes Algorithm 1. The selection of this design can be based on any existing deterministic design strategy, factorial, optimal, minimax, etc. The only requirement is that it satisfies Definition 1. The selection of $\xi$ is critical to the overall performance of the RRSD. It is shown in the Section 4 that implementing the RRSD always increases the efficiency compared to the design used to initialize it. If $\xi$ is inefficient then the RRSD is likely to be inefficient compared to an alternative RRSD initialized with a more efficient deterministic design. It is critical to recognize that the proposed approach does not negate the importance of a priori deterministic design strategies.
3.2. Random A Priori Designs  
Up to this point only deterministic a priori designs have been considered. As stated in Section 1, the algorithm and main results of this work extend to random a priori designs. The conceptual change from a deterministic to a random strategy is that in the latter the design $\xi$ is an observed realization from a probability measure $\pi$, which places positive probability on one or more elements of $\Xi_n$. The only alteration to the RRSD algorithm required is that the experimenter must first randomized by generating an observed design $\xi$ from $\pi$ and then adapt by initializing Algorithm 1 with this design. The importance of this randomize then adapt strategy is outlined in Section 4.1.

3.2.1. Minimax Random Design  
As an illustration consider a slightly generalized version of the Waite and Woods (2021) Example 4.2.1 where $x \in \mathcal{X} = [-1, 1]$, $\eta(x, \theta) = f^T(x)\theta$ and $f(x) = (1, x, x^2)^T$. The stated objective is to minimize the maximum prediction variance over $\mathcal{X}$. The minimax deterministic design, denoted $\xi_G$, minimizes the $\max_{x \in \mathcal{X}} \text{Var}[f^T(x)\theta]$ and is efficient with respect to this objective. If $n$ is divisible by 3 then $\xi_G$ places $n/3$ observations on each point -1, 0 and 1. If $n \mod 3 = 1$ then $\xi_G$ places $(n - 1)/3$ observations on each support point and arbitrarily assigns the remaining observation to one of -1, 0 or 1. Waite and Woods (2021) show that a strategy where $(n - 1)/3$ observations are placed on each support point and the remaining observation is randomly assigned, with probability 1/3, to -1, 0 or 1 outperforms the deterministic approach. Let $\pi_G$ represent the minimax randomized design.

The RRSD is only guaranteed to improve the efficiency of the design used to initialize it; a RRSD initialized with $\xi_G$ is not necessarily more efficient than the random design $\pi_G$ and visa versa. If the objective is to improve $\pi_G$ one should first randomize by generating an observed design, $\xi$, from $\pi_G$ and then adapt by initializing the RRSD with this observed realization. This randomize then adapt strategy is guaranteed to be more efficient than $\pi_G$ [Section 4.1]. This example is considered further in the simulation study of Section 6.

Under the previously established notation, $\bar{\pi}_G$, $\bar{A}_{x_G}$ and $\bar{\theta}_{\bar{G}}$ are the observed design, an ancillary variable and an estimate of $\theta$ following the RRSD with initial design $\pi_G$.

4. Main Result  
The adaptive algorithm was motivated by the RSLB, shown in Section 2 to hold for any design $\xi \in \Xi_n$. This is appropriate to motivate the proposed design; however, it is insufficient to reach the desired conclusion, that adaptation improves efficiency. For this it is required to extend Proposition 2 to the RRSD.

**Lemma 2.** Under the conditions of Proposition 2 and Condition 3  
\[
\text{Var}[\theta_\xi | \xi] \geq E[H^{-1}_{A_\xi} | \xi] \geq F^{-1}_{\xi}
\]

for all $\xi \in \Xi_n$.

See section 8 for proof. Note that the expectation and variance in the lemma are conditional on the initial design $\xi$ and not the observed design $\bar{\xi}$.

The second inequality in Lemma 2 shows that the RRSD has the same CRLB as the deterministic design used to initialize it. Now that it has been shown that the same lower bounds hold for the RRSD its efficiency, relative to $\xi$, can be assessed. The main result of this work is obtained in the following theorem.

**Theorem 2.** Under Conditions 2 and 3  
\[
\lim_{n \to \infty} n^2 c^T (E[H^{-1}_{A_\xi} | \xi] - E[H^{-1}_{A_{\bar{G}} | \xi}]) c \geq 0
\]

for any initial design $\xi \in \Xi_n$ and all non-zero $c \in \mathbb{R}^p$ with equality if and only if $d = 1$ or $\gamma = 0$. 
See section 8 for proof. Recall the definition of second order efficiency in (9); Theorem 2 states that no a priori deterministic design is second order efficient, i.e. efficiency requires adaptation.

As previously stated, the RSLB and the CRLB coincide in normal linear models with constant variance, which corresponds to case where \( \gamma = 0 \) and is excluded. A note on why \( d = 1 \) is excluded. The RRSD algorithm increases the efficiency of an experiment, relative to the deterministic design \( \xi \), by adaptively altering the pre-specified allocation weights, \( w = (w_1, \ldots, w_d)^T \), to the pre-specified support points, \( X = (x_1, \ldots, x_d)^T \). If \( d = 1 \) then \( \xi \) has only a single support point, with weight \( w_1 = 1 \), and \( \mathcal{P}_r \{ r \} = 1 \) for all \( r \) in Step 4 of Algorithm 1. When \( d = 1 \) the RRSD results in a design that is equivalent to \( \xi \).

**Remark 4.** Many existing adaptive designs strategies are known to produce biased estimates [Coad and Ivanova (2001); Brannath, Knig and Bauer (2006); Flournoy and Oron (2019); Marschner (2021); Lane (2022)]. This is not expected to be the case for the RRSD. The main intuition for this remark is that the RRSD is ancillary by construction. In other words no information about the parameter \( \theta \) is used in the design and is not expected to induce the additional bias observed in existing adaptive designs.

4.1. *Extension to Randomized Strategies* Theorem 2 restricts the initial design to be from the set of a priori deterministic designs. The extension of Algorithm 1 to a priori random designs is provided in Section 3.2. The important aspect of the extension is to first, randomly generate a design \( \xi \) from \( \pi \); then adapt by initializing the RRSD with this observed realization. For the randomize then adapt strategy it can be directly observed that Theorem 2 holds for any given \( \xi \) by taking the expectation with respect to the probability measure \( \pi \). The following corollary formalizes this statement and extends Theorem 2 to randomized design strategies.

**Corollary 1.** Under the conditions of Theorem 2 for any probability measure such that \( \pi(\xi|\mathcal{Y}) = \pi(\xi) \) with probability space \( (\mathcal{X}, \Xi_n, \pi) \)

\[
\lim_{n \to \infty} n^2 c^T \left( \mathbb{E}[H_{\mathcal{A}_r}^{-1}] - \mathbb{E}[H_{\mathcal{A}_r}^{-1}] \right) c \geq 0
\]

for all non-zero \( c \in \mathbb{R}^p \) with equality if and only if \( d = 1 \) or \( \gamma^2 = 0 \).

As stated previously, the RRSD is only guaranteed to be better than the a priori design used to initialize it. For example, a RRSD initialized by a randomized strategy is only guaranteed to be more efficient than the corresponding a priori randomized strategy. It is not guaranteed to be more efficient than a similar a priori deterministic design. Section 6.2.1 provides an illustrative example of this discussion.

5. Deterministic Relevant Subset Design The RRSD is, in a sense, a proof of concept design. It was constructed to provide a theoretically tractable framework to prove that an adaptive design always exists that is more efficient than any a priori design. While the RRSD represents a convenient solution; it is likely inefficient compared to adaptive deterministic alternatives. An adaptive deterministic design is distinguished from an a priori deterministic design in that, for the former, given the data from the first \( r \) runs the design of run \( r + 1 \) is completely determined; whereas, in the latter the design is completely determined in advance of the experiment. Randomization has many known benefits; but, efficiency is not necessarily one of them. Even in cases where random strategies outperform deterministic ones, e.g. the example in Section 3.2.1, the randomization should be implemented before the
adaptive design is initialized. It is likely that deterministic alternatives to the RRSD will lead to greater efficiency.

Below a one-step ahead deterministic relevant subset design is proposed. In a one-step ahead setting each run, $r$, corresponds to a single observation (i.e. $r = j$) and the number of runs is equal to the sample size.

**ALGORITHM 2. Deterministic Relevant Subset Design (DRSD)**

1. Repeat steps 1-3 of Algorithm 1.
2. For $r = n\{1\} + 1, \ldots, n$ let $\mathcal{P}_1\{r + 1\} = w_i - u_{\mathcal{A},\{1; r\}}/\mathcal{N}\{r + 1\}$.
3. Find $i\{r + 1\} = \arg \max_{i = 1, \ldots, d} \mathcal{P}_1\{r + 1\}$.
4. Allocate the $\{r + 1\}$th observation to the support point to $x_i\{r + 1\}$.

Let $\hat{\xi} = [\hat{x}(1), \ldots, \hat{x}(n)]^T$ represent the design observed following the DRSD with initial design $\xi$, where $\hat{x}(j)$ is the support point for observation $j$ assigned by the DRSD algorithm. Additionally, $\hat{\mathcal{A}}_\xi$ and $\hat{\theta}_\xi$ are ancillary random variable and estimate of $\theta$ following the DRSD initialized with $\xi$.

Intuitively, selecting design points one at a time will further increase efficiency. It is not trivial to extend the main results to the DRSD and such results are not provided in the current work. Instead the justification for the DRSD will be demonstrated in a series of simulation studies in the next section.

**6. Simulation Study** In this section the efficiency of the RRSD and DRSD are compared to a priori alternatives in a variety of settings.

**6.1. Difference in the Means** Return to the difference in the means example considered in Sections 1 and 2.4.1 and recall that the balanced design, $\xi_\rho$, is the most efficient fixed design. To demonstrate the RRSD and DRSD algorithms, let $\alpha = \beta = 1/4$, $\eta_1 = \eta_2 = 1$ and $n\{1\} = 4$. Following Step 1 of Algorithm 1 the adaptive procedures are initialized by placing 2 observations on each treatment. Suppose $a_i\{1\} = \{0.8, 1.1\}$ and $a_2\{1\} = \{1.8, 2.2\}$ are observed. Based on these values $u_{\mathcal{A},\{1\}} = -1.05$, $u_{\mathcal{A},\{1\}} = 1.05$, $\mathcal{N}\{2\} = 3$, $\mathcal{P}_1\{2\} = 0.85$, $\mathcal{P}_2\{2\} = 0.15$ and run 2 of the RRSD contains three observation which are assigned to treatment 1 with probability 0.85 and to treatment 2 with probability 0.15. Recall, for the DRSD $r = j$ and it can be seen that $i\{5\} = \arg \max_i \mathcal{P}_i\{5\} = 1$, i.e. the 5th observation is assigned to treatment 1.

Recall, $\xi_\rho$ and $\hat{\xi}_\rho$ denote the RRSD and DRSD initialized by $\xi_\rho$. Additionally $\mathcal{A}_{\xi_\rho}$, $\hat{\mathcal{A}}_{\xi_\rho}$, and $\hat{\mathcal{A}}_{\xi_\rho}$ represent an ancillary random variable from $\xi_\rho$, $\hat{\xi}_\rho$ and the $\xi_\rho$, respectively.

To compare the performance of the a priori deterministic design, $\xi_\rho$; the RRSD, $\hat{\xi}_\rho$; and the DRSD, $\hat{\xi}_\rho$, their efficiencies, defined as

\[
\text{LB-EFF-} \xi_\rho = \text{E}[H^{-1}_{\mathcal{A}_\rho} | \xi_\rho]/F^{-1}_{\xi_\rho}, \quad \text{LB-EFF-} \hat{\xi}_\rho = \text{E}[H^{-1}_{\hat{\mathcal{A}}_\rho} | \xi_\rho]/F^{-1}_{\xi_\rho}, \quad \text{LB-EFF-} \hat{\xi}_\rho = \text{E}[H^{-1}_{\hat{\mathcal{A}}_\rho} | \xi_\rho]/F^{-1}_{\xi_\rho},
\]

are reported. Efficiency is defined relative to the CRLB and is bounded above by 1; the larger the value the more efficient the design.

Figure 1 plots LB-EFF-$\xi_\rho$ (solid), LB-EFF-$\hat{\xi}_\rho$ (dashed) and LB-EFF-$\hat{\xi}_\rho$ (dotted) on the $y$-axis for $\alpha = \beta = 1/4$ (left) and $\alpha = \beta = 1/8$ (right) as a function of the sample size, $n$, on the $x$-axis. The thin horizontal line at one represents the CRLB. LB-EFF-$\xi_\rho$ is based on the exact results of Section 2.4.1; whereas, LB-EFF-$\hat{\xi}_\rho$ and LB-EFF-$\hat{\xi}_\rho$ are obtained from a simulation with 2000 iterations. As predicted by Theorem 2 the solid curve corresponding to the
Efficiency requires adaptation

RRSD is uniformly greater than the dotted curve corresponding to the a priori deterministic design. Additionally, as predicted in Section 5 the dashed line corresponding to the DRSD is uniformly more efficient than the RRSD. The gain in efficiency can be quite significant, for example when \( \alpha = \beta = 1/8 \) and \( n = 36 \) the efficiency of the DRSD, the RRSD and the a priori deterministic design are 0.76, 0.71 and 0.56, respectively.

LB-EFF measures the efficiency with respect to the RSLB; intuitively this implies that for any estimate of the difference in the means the adaptive designs are expected to be more efficient than the corresponding a priori alternative. One convenient feature of the current setting is that the variance of the MLE and the RSLB are equivalent. This immediately implies that RRSD and DRSD reduce the variance of the MLE in the exact same manner that the RSLB was reduced.

6.2. Optimal Design

A primary objective in optimal design is to minimize \( F_\xi \) with respect to a convex optimality criterion, denoted \( \Psi(\cdot) \). Formally, an a priori deterministic design \( \xi \) is \( \Psi \)-optimal if

\[
\xi_\Psi = \arg\min_{\xi \in \Xi_n} \Psi(F_\xi),
\]

The optimal design has long been considered a benchmark of efficiency, see Kiefer (1959); Kiefer and Wolfowitz (1959, 1960) for early foundational works; see Pukelsheim (2006) and Atkinson, Donev and Tobias (2007) for reference texts. The \( D \)-criterion, \( \Psi(F_\xi) = |F_\xi|^{-1/p} \), and the \( A \)-criterion defined, for non-singular \( F_\xi \) as \( \Psi(F_\xi) = \text{Tr}(F_\xi^{-1}) \) are two commonly used criteria. The definition of efficiency in (14) is narrower than considered in the current work, where efficiency is in Loewner order.

The significance of Theorem 2 is that implementing a RRSD improves any a priori design in Loewner order. This implies that the RRSD is more efficient with respect to most of the so called alphabetic optimally criteria \( A, D, c, E, G, L \), etc., see Atkinson, Donev and Tobias (2007) for a description of commonly used criteria. For example, suppose a RRSD is initialized with the \( D \)-optimal design, \( \xi_D \), not only is the RRSD more efficient with respect to the \( D \) criterion, it is also more efficient with respect to all of the other aforementioned alphabetic criteria.

6.2.1. G-Optimal (Minimax) Design

A nice illustration of the improvement in efficiency offered by the adaptive approaches is the minimax example outlined in Section 3.2.1. The minimax a priori deterministic design corresponds to the \( G \)-optimal design with \( \Psi(F_\xi) = \max_{x \in \mathcal{X}} f^T(x)F_\xi^{-1}f(x) \), \( \mathcal{X} = [-1, 1] \) and \( f(x) = (1, x, x^2)^T \). For this simulation study, it is assumed the errors follow a generalized normal distribution (GND), with \( \theta = (1, 1, 1)^T \).
\( \zeta = 10 \) and \( \tau = 1 \), as described in Section 2.4.1. Recall, that if \( n \) is divisible by 3 then \( \xi_G = \pi_G \) place equal weight on the points \(-1, 0\) and \( 1 \). For this reason only sample sizes such that \( n \mod 3 = 1 \) are examined. The adaptive designs are initialized with \( n \{1\} = 6 \) with 2 observations placed on each point \((-1, 0\) and \( 1 \)) in the first run.

This simulation study considers four designs; the two \( a \) priori designs, \( \xi_G \) and \( \pi_G \); and the two corresponding DRSDs initialized by \( \xi_G \) and \( \pi_G \), denoted \( \hat{\xi}_G \) and \( \hat{\pi}_G \), respectively. As was the case in the difference in the means example, the RRSD was slightly less efficient than the DRSD; for clarity of exposition the RRSD results are not included.

As was the case in the difference in the means example, the RRSD was slightly less efficient than the DRSD; for clarity of exposition the RRSD results are not included.

In the difference in the means example the \( \text{Var}\[\hat{\theta}\|\xi\] = E[1_{\mathcal{A}(\xi)}\|\xi\]} \). For GND errors this is not the case; in this section the efficiency of the RSLB and the variance of the MLE are examined with respect to certain convex optimality criteria. Specifically, the \( \Psi \)-efficiency of the RSLB and variance of the MLE for the \( a \) priori deterministic design \( \xi_G \) defined as

\[
\text{LB-EFF}_{\Psi;\xi_G} = \frac{\Psi\{E[H_{\mathcal{A}(\xi)}^{-1}\|\xi_G]\}}{\Psi\{F_{\hat{\xi}^{-1}}\}} \quad \text{and} \quad \text{Var-\text{EFF}}_{\Psi;\xi_G} = \frac{\Psi\{\text{Var}[\hat{\theta}\|\xi_G]\}}{\Psi\{F_{\hat{\xi}^{-1}}\}},
\]

respectively, are reported; recall that \( \hat{\theta}\|\xi \) is the MLE following an experiment with design \( \xi \). Similarly, \( \text{LB-EFF}_{\Psi;\pi_G} \) and \( \text{Var-\text{EFF}}_{\Psi;\pi_G} \) is the \( \Psi \)-efficiency of the \( a \) priori random minimax design; \( \text{LB-EFF}_{\Psi;\xi_G} \) and \( \text{Var-\text{EFF}}_{\Psi;\xi_G} \) is the \( \Psi \)-efficiency of a DRSD initialized with \( \xi_G \); and \( \text{LB-EFF}_{\Psi;\pi_G} \) and \( \text{Var-\text{EFF}}_{\Psi;\pi_G} \) is the \( \Psi \)-efficiency of a DRSD initialized with \( \pi_G \). All of the above measures are defined with respect to the CRLB corresponding to the \( \Psi \)-optimal design, \( \xi_G \), and are bounded above by 1.

The top left panel of Figure 2 plots \( \text{LB-EFF}_{\Psi;\xi_G} \) (dotted), \( \text{LB-EFF}_{\Psi;\pi_G} \) (dashed), \( \text{LB-EFF}_{\Psi;\xi_G} \) (dot-dashed) and \( \text{LB-EFF}_{\Psi;\pi_G} \) (solid), obtained from 2000 iterations, on the \( y \) axis for various sample sizes \( n \) such that \( n \mod 3 = 1 \) on the \( x \) axis. The thin horizontal line at one corresponds to the CRLB. As expected from Theorem 2, the DRSDs, \( \hat{\pi}_G \) and \( \hat{\pi}_G \), are more efficient than their respective \( a \) priori initial designs, \( \xi_G \) and \( \pi_G \). An additional feature is that the RSLB for the two \( a \) priori designs coincide; whereas, \( \hat{\pi}_G \) is nearly uniformly more efficient than \( \hat{\pi}_G \). Interestingly, this indicates that a randomized strategy does not in itself improve the RSLB; however, initializing the DRSD with \( \pi_G \), in place of \( \xi_G \), can improve the RSLB. Perhaps random strategies are more useful in an adaptive setting than in an \( a \) priori setting.

The top right panel of Figure 2 plots \( \text{Var-\text{EFF}}_{\Psi;\xi_G} \) (dotted), \( \text{Var-\text{EFF}}_{\Psi;\pi_G} \) (dashed), \( \text{Var-\text{EFF}}_{\Psi;\xi_G} \) (dot-dashed) and \( \text{Var-\text{EFF}}_{\Psi;\pi_G} \) (solid) on the \( y \) axis for the same series of simulations presented in the top left panel. The solid curve, which corresponds the DRSD\( \pi_G \), is uniformly greater than the dashed curve, which corresponds to \( \pi_G \). Similarly, the dot-dashed curve is uniformly greater than the dotted curve which confirms that the \( \xi_G \) is uniformly more efficient than \( \xi_G \). These statements support the claim that the adaptive procedures are more efficient than their initial design. For the sample size \( n = 8 \), the efficiency of the \( a \) priori random design \( \pi_G \) is 0.31 is greater than the \( \hat{\pi}_G \) efficiency of 0.25. This illustrates the point that the proposed adaptive strategies are only guaranteed to improve the initial design. For comparison the efficiency of \( \xi_G \) and \( \hat{\pi}_G \) are 0.24 and 0.35, respectively. Note, for all other sample sizes examined \( \hat{\pi}_G \) was more efficient than \( \hat{\pi}_G \), which indicates that the benefit of the \( a \) priori random strategy disappears quickly as \( n \) increases.

The preceding discussion concerns the efficiency of the various designs with respect to the \( G \)-criterion. One important aspect of Theorem 2 is that the improvement is in Loewner order. To illustrate, the bottom row of Figure 2 plots the same series of simulations and efficiency measures as the top row for \( \Psi \) equal to the \( D \)-criterion. For the \( D \)-criterion the efficiency of the RSLB (bottom left) and the variance of the MLE (bottom right) for \( \xi_G \) and \( \pi_G \) are in perfect correspondence. This equality occurs because the \( G \)-optimal, \( \pi_G \), offers no guarantee.
for the $D_\tau$ or any other, optimality criteria. The solid and dot-dashed curves representing the $\xi_G$ and DRSD$\pi G$ offer a nice illustration of the global efficiency of proposed adaptive designs. The two adaptive procedures nearly perfectly coincide and are both uniformly more efficient than the a priori designs $\xi_G$ and $\pi_G$ with respect to the $D$-criterion. Repeating the analysis for any of the commonly used alphabetic criteria produces a similar result. This is a direct result of the improvement in Loewner order.

6.3. Factorial Design  Consider an experiment with two treatment factors $x = (x_1, x_2)^T$, design space $\mathcal{X} = [0,1]^2$, $f^T(x) = (1, x_1, x_2, x_1 x_2)^T$ and $\theta = (1, 1, 1, 1)^T$. The factorial design, $\xi_F = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, places equal weight on each point and allows for the efficient estimation of the main and interaction effects. For $n\{1\} = 8$ the first run places 2 observations on each point in $\xi_F$. It is assumed that the errors are Cauchy distributed, with $\tau = 1$, as described in Section 2.4.2.

The top row of Table 1 reports the RSLB, $n E[H_{\xi_F}^{-1}]$, for $\mathcal{A}$ equal to $\mathcal{A}_{\xi_F}$ (left), $\tilde{\mathcal{A}}_{\xi_F}$ (middle) and $\bar{\mathcal{A}}_{\xi_F}$ (right) from 2000 iterations for the experiment described above with sample size $n = 60$. The second row of Table 1 reports the variance of the MLE, $n \text{Var}[\tilde{\theta}_{\xi} | \xi_F]$, for $\xi$ equal to $\xi_F$ (left), $\bar{\xi}_F$ (middle) and $\tilde{\xi}_F$ (right) for the same scenario as in the top row. The bottom row of Table 1 reports the CRLB, $n F_{\xi_F}^{-1}$, for reference.

Recall the discussion following Proposition 3 which stated that the objective of the adaptive procedures was to reduce the difference between the RSLB and CRLB. The top row of Table 1 confirms that is objective is accomplished; the RSLB corresponding to $\xi_F$ (column 2) and $\xi_F$ (column 3) are markedly closer to $F_{\xi_F}^{-1}$ than the RSLB corresponding to $\xi_F$ (column 1). The RSLB is a lower bound for all estimates of $\theta$; as such it is expected that the adaptive procedures will reduce the variance of estimates of $\theta$. For the MLE this benefit is confirmed...
The check notation is used to indicate an object corresponds to where it can be seen that the variance of the MLE obtained following the RRSD, e.g. following Algorithm \( \xi \), of the MLE from an experiment with design RRSD and DRSD were uniformly more efficient than the corresponding design. Both the adaptive designs are initialize by an \( a \) priori design. The adaptive designs are only guaranteed to be more efficient than the initial \( a \) priori design. It is critical for the overall performance that the RRSD and DRSD are initialized using efficient \( a \) priori designs. Efficient \( a \) priori design strategies are critical to the success of the RRSD and DRSD.

7. Discussion In this work two adaptive designs were developed, the randomized relevant subset design (RRSD) and the deterministic relevant subset design (DRSD), that are more efficient than any corresponding \( a \) priori design, with respect to Loewner order. An \( a \) priori design is any design that is independent of the observed data and can be determined completely in advance of the experiment. This improvement was accomplished by first, deriving a relevant subset lower bound (RSLB) that represents a tighter alternative to the Cramér-Rao lower bound (CRLB). And second, by developing an adaptive algorithm that incorporate the relevant subset into the design of the experiment with the expressed intent of reducing the RSLB.

The main theoretical results in this work are restricted to asymptotic efficiency. A simulation study demonstrated the finite sample benefits of the proposed adaptive designs. Both the RRSD and DRSD were uniformly more efficient than the corresponding \( a \) priori alternatives for a variety of examples.

The proposed adaptive designs are initialize by an \( a \) priori design. The adaptive designs are only guaranteed to be more efficient than the initial \( a \) priori design. It is critical for the overall performance that the RRSD and DRSD are initialized using efficient \( a \) priori designs. Efficient \( a \) priori design strategies are critical to the success of the RRSD and DRSD.

8. Technical Details In this section the supporting arguments for the main theoretical results are presented. Technical details for supporting lemmas are provided in the supplemental materials.

8.1. Proof of Lemma 1 The check notation is used to indicate an object corresponds to the RRSD, e.g. \( \check{Y} \), \( \check{A} \) and \( \check{S}_i \). This notation is introduced for the RRSD in the paragraph following Algorithm 1. The following lemma is required for the validity of equation (10).

| Table 1 |
|---|
| The RSLBs (top row), \( nE[H^{-1}_{\check{A}_F} | \xi_F] \) (left), \( nE[H^{-1}_{\check{A}_F} | \xi_F] \) (middle) and \( nE[H^{-1}_{\check{A}_F} | \xi_F] \) (right); and the variance of the MLEs (second row), \( nVar[\hat{\theta}_{\xi_F} | \xi_F] \) (left), \( nVar[\hat{\theta}_{\xi_F} | \xi_F] \) (middle) and \( nVar[\hat{\theta}_{\xi_F} | \xi_F] \) (right) for a factorial experiment from 2000 iterations with sample size \( n = 60 \). The CRLB (bottom row), \( nF_{\xi_F}^{-1} \), is provided for reference. |
| \( nE[H^{-1}_{\check{A}_F} | \xi_F] \) | \( nE[H^{-1}_{\check{A}_F} | \xi_F] \) | \( nE[H^{-1}_{\check{A}_F} | \xi_F] \) |
| (9.2 -9.2 -9.2 9.2) | (8.8 -8.8 -8.8 8.8) | (8.3 -8.3 -8.3 8.3) |
| \( nVar[\hat{\theta}_{\xi_F} | \xi_F] \) | \( nVar[\hat{\theta}_{\xi_F} | \xi_F] \) | \( nVar[\hat{\theta}_{\xi_F} | \xi_F] \) |
| (10.1 -10.2 -10.2 10.6) | (9.4 -9.3 -9.4 9.6) | (8.5 -8.4 -8.8 8.9) |
| \( nF_{\xi_F}^{-1} \) |
| (8.0 -8.0 -8.0 8.0) |
| (8.0 16.0 8.0 -16.0) |
| (8.0 8.0 16.0 -16.0) |
| (8.0 -16.0 -16.0 32.0) |
Lemma 3. \( \mathcal{A}\{1\} \) is \( \mathcal{A}\{1; 2\} \) measurable.

The second term in the right hand side of (13) can be written as

\[
E[\text{Var}[u_{\mathcal{A}\{1; 2\}} | \mathcal{A}\{1\}]] = E[\text{Var}[E[\check{S}\{1; 2\} | \mathcal{A}\{1; 2\}]] | \mathcal{A}\{1\}]
\]

by the definition of \( u_{\mathcal{A}\{1\}} \).

Lemma 4. Under the current conditions

\[
\text{Var}[E[\check{S}\{1; 2\} | \mathcal{A}\{1; 2\}]] | \mathcal{A}\{1\}] = \text{Var}[E[\check{S}\{2\} | \mathcal{A}\{1; 2\}]] | \mathcal{A}\{1\}] + O(1)
\]

Lemma 4 implies that the right hand side of (15) is

\[
E[\text{Var}[E[\check{S}\{1; 2\} | \mathcal{A}\{1; 2\}]] | \mathcal{A}\{1\}] = E[\text{Var}[E[\check{S}\{2\} | \mathcal{A}\{1; 2\}]] | \mathcal{A}\{1\}] + O(1).
\]

The bulk of the remainder of the proof derives a bound for the right hand side of the above ignoring the \( O(1) \) term.

\[
E[\text{Var}[E[\check{S}\{2\} | \mathcal{A}\{1; 2\}]] | \mathcal{A}\{1\}] \\
\leq E[\text{Var}[\check{S}\{2\} | \mathcal{A}\{1\}] \\
= E \left( \sum_{j=n(1)+1}^{N(1;2)} \text{Var}[(\check{T}_i(j) - w_i)s_{\check{j}(j)}] | \mathcal{A}\{1\} \right) \\
= E \left( \sum_{j=n(1)+1}^{N(1;2)} \text{Var}[(\check{T}_i(j) - w_i)E[s_{\check{j}(j)} | \check{T}_i(j)] | \mathcal{A}\{1\}] \\
+ E[(\check{T}_i(j) - w_i)\text{Var}[s_{\check{j}(j)} | \check{T}_i(j)] | \mathcal{A}\{1\}] \right) \\
\leq E[N\{2\}] \\
= \left\{ E \left[ \max_i \left\{ u_{\mathcal{A}\{1\}}(1)/w_i \right\} \right] + 1 \right\}.
\]

The first line is due to the law of total variance; the second is from the definition of \( \check{S}\{2\} \), the fact that \( (\check{T}_i(j) - w_i)s_{\check{j}(j)} \) \( j = n(1) + 1, \ldots, N\{1; 2\} \) given \( \mathcal{A}\{1\} \) is a sequence of independent and identically distributed random variables and the fact that \( N\{1; 2\} \) is \( \mathcal{A}\{1\} \) measurable; the third equality is due to the law of total variance; the fourth is obtained by recognising that given \( \check{T}_i(j) \) them mean and variance of \( s_{\check{j}(j)} \) is \( \mu \) and 1, respectively, see the proof of Proposition 3 in the supplemental materials; the final line is from the definition of \( N\{2\} \) in (12), where the +1 is required to account for the ceiling operator.

Next consider the term

\[
E \left[ \max_i \left\{ u_{\mathcal{A}\{1\}}(1)/w_i \right\} \right] \leq E \left[ \max_i \left\{ u_{\mathcal{A}\{1\}}(1)/w_i \right\}^2 \right]^{1/2} \leq E \left[ \sum_i (u_{\mathcal{A}\{1\}}(1)/w_i)^2 \right]^{1/2} \\
\leq w_i^{-1} \left\{ \sum_i \text{Var} \left[ u_{\mathcal{A}\{1\}}(1) \right] \right\}^{1/2} = O([n\{1\}]^{1/2}).
\]

The first line is due to Hölder’s inequality, the second is due to the fact that a sum of list of squares is greater than the maximum of a list squared, the third is from the a priori nature
of the first run design, specifically, the independence property of \( u_{A_i}\{1\}, i = 1, \ldots, d \) and the fact that \( \mathbb{E}[u_{A_i}\{1\}] = 0 \); and the final line is again from the property of \textit{a priori} designs where \( \text{Var}[u_{A_i}\{1\}] = O((n\{1\})) \). Putting the preceding together yields

\[
\lim_{n\{1\} \to \infty} \text{Var}[u_{A_{\{1;2\}}} \mid \xi] / n\{1\} = 0
\]
as stated.

8.2. \textit{Proof of Lemma 2}  Lemma 12 [supplemental materials] implies that Proposition 1 extends to the RRSD and taking the expectation implies the first inequality in Propositions 2, provided the necessary conditions which are assumed.

For the second inequality the following lemmas is required.

\textbf{Lemma 5.} \textit{Under the present conditions} \( \mathbb{E}[h_{A_i} \mid \xi] = 0 \) for \( i = 1, \ldots, d \).

Now that the expectation of \( h_{A_i} \) has been established the extension of Jensen’s inequality given in Groves and Rothenberg (1969) implies that \( \mathbb{E}[H^{-1}_{A_i} \mid \xi] \geq F^{-1}_{\xi} \).

8.3. \textit{Proof of Theorem 2}  Proposition 3 and Lemma 2 implies that we can write

(16) \[
\lim_{n \to \infty} n^{2} c^T (\mathbb{E}[H^{-1}_{A_i} \mid \xi] - \mathbb{E}[H^{-1}_{A_i} \mid \xi]) c = \mu \lim_{n \to \infty} n^{-1} \text{Tr}[D(\text{Var}[u_{A_i} \mid \xi] + \text{Var}[v_{A_i} \mid \xi] - \text{Var}[u_{A_i} \mid \xi] - \text{Var}[v_{A_i} \mid \xi])].
\]
The first lemma establishes that the limiting behavior of \( \text{Var}[v_{A_i} \mid \xi] \).

\textbf{Lemma 6.} \textit{Under the present conditions} \( \lim_{n \to \infty} n^{-1} \text{Var}[v_{A_i} \mid \xi] = \gamma^2 w w^T \).

The implications being that the right hand side of (16) reduces to \( \text{Tr}[D(\text{Var}[u_{A_i} \mid \xi] - \text{Var}[u_{A_i} \mid \xi]) \). The following lemma implies that the term \( n^{-1} \text{Var}[u_{A_i} \mid \xi] \) disappears for large \( n \).

\textbf{Lemma 7.} \textit{Under the present conditions} \( \text{Var}[u_{A_i} \mid \xi] = O(n^{1/2}) \).

From Lemmas 6 and 7 we can now state that

\[
\lim_{n \to \infty} n^{2} c^T (\mathbb{E}[H^{-1}_{A_i} \mid \xi] - \mathbb{E}[H^{-1}_{A_i} \mid \xi]) c = \gamma^2 \text{Tr}[D(W - w w^T)] \geq 0.
\]
See the proof of Proposition 3 for additional details. The if only if claim is verified by the following lemma.

\textbf{Lemma 8.} \textit{For all} \( \xi \in \Xi_n \), \( \text{Tr}[D(W - w w^T)] \geq 0 \) \textit{with equality if and only if} \( d = 1 \).

This concludes the proof of the theorem.

\textbf{SUPPLEMENTARY MATERIAL}

\textbf{Supplemental Materials for Efficiency Requires Adaptation}

\textbf{9. Supplemental Introduction}  All expectations and variances are conditional on the design \( \xi \) and thus this is omitted.
10. Proof of Proposition 1 First, note that

\[(S17)\quad \frac{\partial}{\partial \theta} \log f(y|\theta) = \sum_{i=1}^{d} \frac{\partial \eta(x_i, \theta)}{\partial \eta_i} \frac{\partial}{\partial \eta_i} \log f(y_i|\eta_i),\]

where \(y_i = \{y(j) : t_i(j) = 1, j = 1, \ldots, n\}\). From the ancillarity of \(a_i\) its distribution is not a function of \(\eta_i\); therefore,

\[(S18)\quad \frac{\partial}{\partial \eta_i} \log f(y_i|\eta_i) = \frac{\partial}{\partial \eta_i} [\log f(y_i|a_i, \eta_i) + \log f(a_i)] = \frac{\partial}{\partial \eta_i} \log f(y_i|a_i, \eta_i).\]

Continuing, Condition 2 (5') ensures that

\[E \left[ \frac{\partial}{\partial \eta_i} \log f(y_i|a_i, \eta_i) | A_i \right] = 0 \implies E \left[ \frac{\partial}{\partial \theta} \log f(y|\theta) | A \right] = 0.\]

Using the alternate notation of Section 2 the above implies that \(E[\dot{Y}, T|A_i] = 0\). Equations (S17) and (S18) imply that \(\partial \log f(y|\theta)/(\partial \theta) = \partial \log f(y|A, \theta)/(\partial \theta)\); therefore,

\[\text{Cov} \left[ \dot{\theta}, \frac{\partial}{\partial \theta} \log f(y|\theta) | A \right] = E \left[ \dot{\theta} \frac{\partial}{\partial \theta} \log f(y|A, \theta) | A \right] = \frac{\partial}{\partial \theta} E[\dot{\theta} | A] = 0\]

from Condition 2 (5') and the conditional unbiased assumption. From a final appeal to Condition 2 (5') it is straightforward to show that

\[(S19)\quad \text{Var} \left[ \frac{\partial}{\partial \eta_i} \log f(y_i|A_i, \eta_i) | A_i \right] = E \left[ \frac{\partial^2}{\partial \eta_i^2} \log f(y_i|A_i, \eta_i) | A_i \right] = h_{A_i}.\]

From (S17), (S18), (S19) and an appeal to the independence of the responses

\[\text{Var} \left[ \frac{\partial}{\partial \theta} \log f(y|\theta) | A \right] = H_A(\theta).\]

Finally, an appeal to the Cauchy-Schwarz inequality yields \(\text{Var}[\dot{\theta}|A] \geq H^{-1}_A\) as stated.

11. Proof of Proposition 2 The stated result follows from the below

\[\text{Var}[\dot{\theta}] = E[\text{Var}[\dot{\theta}|A]] + \text{Var}[E[\dot{\theta}|A]]\]

\[= E[\text{Var}[\dot{\theta}|A]] \geq E[H^{-1}_A] \geq F^{-1}_\xi.\]

where the first line is from the law of total variance; the second is from the unbiased assumption; the third is due to an application of Proposition 1; and the final line is obtained from an application of Proposition 1 and the extension of Jensen’s inequality for inverse matrices proven in Groves and Rothenberg (1969).

12. Proof of Proposition 3 Let \(Z = (Z_1, \ldots, Z_d)^T\), where \(Z_i = n^{-1/2}(h_{A_i} - nw_i \mu)\). Under the assumed conditions a Taylor expansion for \(Z\) around its expectation \(0\) yields

\[(S20)\quad n^2 c^T(E[H^{-1}_A] - F^{-1}_\xi) c = E[Z^T D Z] + o(1).\]
Let $Z_i = n^{-1/2}(h_i, -nw_i, \mu)$ and note that per the assumed conditions $Z_i - Z_i = O(n^{-1/2})$. Therefore, $Z_i$ can be replaced with $Z_i$ in the right hand side of (S20) without impacting the order of approximation. Noting that $Z_i = \mu n^{1/2}(u_i, v_i)$

$$E[Z^T D Z] = \text{Tr}(D \text{Var}[Z]) = n^{-1} \mu \text{Tr}(D(\text{Var}[u_i] + E[u_i v_i^T] + E[v_i u_i^T] + \text{Var}[v_i])),$$

First consider the expectations

$$E[u_i] = E \left[ \sum_{j=1}^n \{t_i(j) - w_i\}^2 \right] = \sum_{j=1}^n \{t_i(j) - w_i\} = 0,$$

$$E[v_i] = w_i E \left[ \left\{ \sum_{j=1}^n (i_{Y(j)} - 1) \right\} \right] = 0.$$

Let $q_i = \sum_{j=1}^n t_i(j)g_j(y_j)$, where $q_j(y_j) = \{i_{Y(j)}^2 - 1 - n_{11}i_{Y(j)}\}/\gamma$ has mean 0 and variance 1. Further, let $Q = (q_1, \ldots, q_t)^T$.

**Lemma 9.** Under a deterministic design Under the current conditions $u_{\alpha_i} - u_{\beta_i} = O(1)$.

Proof. Barndorff-Nielsen and Cox (1994) state that if $(\tilde{\eta}_i, \alpha_i)$ and $\alpha_i$ are approximately sufficient and ancillary to the order $O(n^{-1/2})$ then $p(\tilde{\eta}_i|\eta_i, \alpha_i) = p(\tilde{\eta}_i|\eta_i, \alpha_i) + O(n^{-1})$, where $p(\tilde{\eta}_i|\eta_i, \alpha_i) = \{i_{Y.T}^2/2\pi\}^{1/2}e^{i_{Y.T}^2/2}$, where $i_{Y.T}$ represents the observed information, $i_{Y.T}$, evaluated at $\eta_i = \tilde{\eta}_i$. The limiting distribution of $q_i$ is normal with mean 0 and variance 1 which implies that it satisfies the $O(n^{-1/2})$ approximate sufficiency and ancillary condition. The importance of the $p^1$-formula is, first, it implies that every $O(n^{-1/2})$ (or higher) approximate ancillary generates the same $p^1$-formula and thus $u_{\alpha_i} - u_{\beta_i} = O(1)$.

Condition 3 ensures that the information in the relevant subset is accurately approximated by the $p^1$-formula; specifically, $a_i - g^a_i = O(1)$, where $g^a_i = E[i_{Y.T}]$. From its definition, in Theorem 3, the preceding implies that $u_{\alpha_i} - u_{\beta_i} = O(1)$ and by extension $u_{\alpha_i} - u_{\beta_i} = O(1)$ as stated.

In words Lemma 9 ensures that all approximate ancillaries yield the same approximate information. The implication begin that it only needs to be shown that $E[u_{\alpha_i} v_{\alpha_i}^T] = 0$ and then an appeal to Lemma 9 implies the stated result. This outlines the remainder of the proof.

Condition 3 ensures that $E[i_{Y.T}^2|Q_i] = 0$; therefore,

$$g_{Q_i} = E[i_{Y.T}^2|Q_i] = \gamma (E[Q_i|Q_i] + nw_i) = \gamma (Q_i + nw_i).$$

From its definition $u_{Q_i} = (u_{Q_1}, \ldots, u_{Q_t})^T$, where $u_{Q_i} = g_{Q_i} - w_i \sum_i g_{Q_i}$ can be alternatively expressed as $u_{Q_i} = \gamma \sum_{j=1}^n [t_i(j) - w_i]g_{Q(j)} + 1$. Similarly, $v_{Q_i} = (v_{Q_1}, \ldots, v_{Q_t})^T$, where $v_{Q_i} = w_i \sum_i g_{Q_i} - n$ can be alternatively expressed as $v_{Q_j} = \gamma \sum_{j=1}^n g_{Y(j)}$.

Consider the variance and covariances terms of $u_{Q_i}$

$$\text{Var}[u_{Q_i}] = \gamma^2 \sum_{j=1}^n [t_i(j) - w_i]^2 \text{Var}[g_{Y(j)}] = \gamma^2 \sum_{j=1}^n [t_i(j) - w_i]^2 \text{Var}[g_{Y(j)}] = \gamma^2 \sum_{j=1}^n [t_i(j) - w_i]^2$$

$$= \gamma^2 w_i (1 - w_i)$$
implies that
\[ E[u_Q, u_{Q_0}] = \gamma^2 \sum_{j=1}^{n} \sum_{l=1}^{n} \{t_i(j) - w_i\}\{t_k(l) - w_k\}E[\{q_{Y(j)} + 1\}\{q_{Y(l)} + 1\}] \]
\[ = \gamma^2 \sum_{j=1}^{n} \{t_i(j) - w_i\}\{t_k(j) - w_k\} \]
\[ = -n\gamma^2 w_i w_k. \]

Note the independence property of \( Y \) was used in the above. Next consider
\[ E[v_Q, v_{Q_0}] = \gamma^2 w_i w_k \left[ \sum_{j=1}^{n} \{q_{Y(j)}\} \right]^2 = w_i w_k \sum_{j=1}^{n} \text{Var}[q_{Y(j)}] = n\gamma^2 w_i w_k \]

Now consider
\[ E[u_Q, v_{Q_0}] = \gamma^2 w_k \sum_{j=1}^{n} \sum_{l=1}^{n} \{t_i(j) - w_i\}\{q_{Y(j)} + 1\} = \gamma^2 w_k \sum_{j=1}^{n} \{t_i(j) - w_i\} = 0, \]
where the independence of responses was again used in the above. The above implies that \( \text{Var}[u_Q] = n\gamma^2 W(I_d - w w^T) \), \( \text{Var}[u_{Q_0}] = n\gamma^2 w w^T \) and \( E[u_Q v_{Q_0}] = 0 \). This and Lemma 9 implies that \( \text{Var}[u_A] = n\gamma^2 W(I_d - w w^T) \), \( \text{Var}[u_{A_0}] = n\gamma^2 w w^T \) and \( E[u_A v_{A_0}] = 0 \) hold at the limit as \( n \to \infty \) for any approximate ancillary \( A \).

13. Proof of Lemma 3

The following lemma is stated for random variables following the entire experiment. While this significantly more than is required these more general results imply the necessary results for the first two runs and will be useful for future results.

**Lemma 10.** \( \tilde{A}\{1; s\} \) is \( \tilde{A}\{1; r\} \) measurable for all \( s \leq r = 1, \ldots, R_{\text{max}} \).

An important property of the location family is that \( \tilde{A}\{1; s\} \) is \( \tilde{A}\{1; r\} \) measurable for all \( r \leq s \) and \( r = 1, \ldots, R_{\text{max}} \). To show this, the following notation is introduced. Note, from Fraser (2004) the \( \tilde{A}\{1; r\} \) can be written as
\[ \tilde{A}\{1; r\} = \{\tilde{A}_i\{1; r; 1\}, \ldots, \tilde{A}_i\{1; r; n\{1\}\}, \ldots, \tilde{A}_i\{1; r; N\{1; r - 1\} + 1\}, \ldots, \tilde{A}_i\{1; r; N_i\{1; r\}\}\}^T, \]
where \( \tilde{A}_i\{1; r; j\} = y_i(j) - \overline{Y}_i\{1; r\} \), \( y_i(j) \) is the \( j \)th observation with support \( x_i \) and \( \overline{Y}_i\{1; r\} = [N\{1; r\}]^{-1} \sum_{i=1}^{n\{1\}} y_i(j) \) and \( \tilde{A}\{1; r\} = [\tilde{a}_1\{1; r\}, \ldots, \tilde{a}_d\{1; r\}]^T \). To demonstrate this note \( \tilde{A}_i\{1; r; j\} \) can be written as
\[ \tilde{A}_i\{1; j\} = \tilde{A}_i\{1; r; j\} - [n\{1\}]^{-1} \sum_{i=1}^{n\{1\}} \tilde{A}_i\{1; r; j\} \]
for \( j = 1, \ldots, n\{1\} \). Noting that the distribution of \( n\{1\} \) is known demonstrates that \( a_i\{1\} \) is \( a_i\{1; r\} \), measurable. Per the definition of the RRSD algorithm it can be seen that \( N_i\{2\} \) is \( a_i\{1\} \) measurable. Repeating as before we obtain
\[ \tilde{A}_i\{1, 2; j\} = \tilde{A}_i\{1; r; j\} - [N_i\{1; 2\}]^{-1} \sum_{i=1}^{N_i\{1; 2\}} \tilde{A}_i\{1; r; j\} \]
for \( j = 1, \ldots, N \{1; 2\} \); implying that \( \tilde{a}_i \{1; 2\} \) is \( \tilde{a}_i \{1; r\} \), measurable. Repeating these steps sequentially, it can be shown that in general

\[
\tilde{A}_i \{1; s; j\} = \tilde{A}_i \{1; r; j\} - \sum_{i=1}^{N_i \{1; s\}} \tilde{A}_i \{1; r; j\}
\]

for \( j = 1, \ldots, N \{1; s\} \); from which it can be concluded that \( \tilde{a}_i \{1; s\} \) is \( \tilde{a}_i \{1; r\} \), measurable for all \( s \leq r \).

**14. Proof of Lemma 4** To begin recall the following lemma from Efron and Hinkley (1978).

**Lemma 11.** Under the present conditions

\[
E[\hat{\eta}_i - \eta_i | \mathbf{A}_i] = -\hat{\eta}_i^{(3)}_{\mathbf{Y}, \mathbf{T}}/(2\hat{\eta}_i^{2}_{\mathbf{Y}, \mathbf{T}}) + O_p(n^{-2})
\]

\[
\text{Var}[\hat{\eta}_i | \mathbf{A}_i] = \hat{\eta}_i^{-1}_{\mathbf{Y}, \mathbf{T}} + O_p(n^{-2})
\]

\[
\hat{\eta}_i^{-1/2}(\hat{\eta}_i - \eta_i)|\mathbf{A}_i \to N(0, 1),
\]

where convergence is in probability as \( n \to \infty \).

The following lemma ensures that the conditional distribution of the MLE is the same as if no adaptation had taken place.

**Lemma 12** (Lane (2019) Theorem 5.1). Let \((\hat{\mathbf{Y}}, \hat{\mathbf{T}}, \hat{\mathbf{A}})\) correspond to the responses, treatment allocation matrix and ancillary complement following an experiment using the RRSD; then

\[
f(\hat{\mathbf{Y}}|\hat{\mathbf{T}}, \hat{\mathbf{A}}) \overset{d}{=} f(\hat{\mathbf{Y}}|\hat{\mathbf{A}})
\]

where \( f(\hat{\mathbf{Y}}|\hat{\mathbf{A}}) \) has the exact same form as an experiment with a deterministic design with \( \mathbf{A} = \hat{\mathbf{A}} \).

Proof. The result of the lemma is verified by the following sequence of a successive conditioning arguments. First,

\[
f(\hat{\mathbf{Y}}, \hat{\mathbf{T}}, \hat{\mathbf{A}}) = f(\hat{\mathbf{T}}|\hat{\mathbf{Y}}, \hat{\mathbf{A}}) f(\hat{\mathbf{Y}}, \hat{\mathbf{A}}) = f(\hat{\mathbf{T}}|\hat{\mathbf{A}}) f(\hat{\mathbf{Y}}, \hat{\mathbf{A}}).
\]

The equality is due to the fact that the RRSD uses only information in \( \hat{\mathbf{A}} \) to determine \( \hat{\mathbf{T}} \). Alternatively, we can also write

\[
f(\hat{\mathbf{Y}}, \hat{\mathbf{T}}, \hat{\mathbf{A}}) = f(\hat{\mathbf{Y}}|\hat{\mathbf{T}}, \hat{\mathbf{A}}) f(\hat{\mathbf{T}}|\hat{\mathbf{A}}) f(\hat{\mathbf{A}}).
\]

Combining the two results and re-arranging we obtain

\[
f(\hat{\mathbf{Y}}|\hat{\mathbf{T}}, \hat{\mathbf{A}}) = f(\hat{\mathbf{Y}}|\hat{\mathbf{A}}).
\]

It can now be concluded that the treatment assignment implemented by the RRSD does not alter the conditional distribution of the responses. Of course the only way the RRSD influences the design is through the selection of the allocations; therefore we can conclude that the conditional distribution of the responses is the same as if \( \hat{\mathbf{T}} \) were pre-specified at its observed value. \( \square \)
Lemmas 11 and 12 ensures the validity of the Taylor series of \( i \) with respect to \( \eta_i \) at \( \hat{\eta}_i \)

\[
i \mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}} = i \mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}} + j^{(3)}_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} (\hat{\eta}_i \{1;2\} - \eta_i) \\
- j^{(4)}_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} (\hat{\eta}_i \{1;2\} - \eta_i)^2 + o(1).
\]

Therefore,

\[
\hat{S}_i \{1\} = \hat{S}_i \{1\} + \mathcal{B}_i \{1;2\} - w_i \sum_{i=1}^{d} \mathcal{B}_i \{1;2\} + o(1),
\]

where \( \hat{S}_i \) denotes \( \hat{S}_i \) evaluated at the MLE of \( \eta_i \) and \( \mathcal{B}_i \{1;2\} = (\hat{\eta}_i \{1;2\} - \eta_i) j^{(3)}_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} + (\hat{\eta}_i \{1;2\} - \eta_i)^2 j^{(4)}_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} / 2 \). Noting that \( j^{(k)}_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} \) is \( \mathcal{A}_i \) measurable for all \( k = 1, 2, \ldots \) an application of Lemma 11 yields

\[
(S21) \quad \text{E}[\hat{S}_i \{1\}|\mathcal{A}_i \{1;2\}] = \hat{S}_i \{1\} + \hat{B}_i - w_i \sum_{i=1}^{d} \hat{B}_i = \hat{S}_i \{1\} + O(1),
\]

where \( \hat{B}_i = [(j^{(3)}_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} / j_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} + j^{(4)}_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} / j_{\mathcal{Y}_{\{1;2\}, \mathcal{T}_i\{1;2\}}} / 2]. The \( \mathcal{A}_i \{1\} \)-measurability of \( S_i \{1\} \) and (S21) implies that

\[
\text{Var}[E[S_i \{1\}|\mathcal{A}_i \{1;2\}|\mathcal{A}_i \{1\}] = \text{Var}[E[S_i \{2\}|\mathcal{A}_i \{1;2\}|\mathcal{A}_i \{1\}] + O(1)
\]

as stated.

15. Proof of Lemma 5 All results are conditional on the initial design \( \xi \). Further, all variables in this section pertain to the RRSD and thus the convention of using a check is not used since there should be no confusion.

Note that \( (h_{\mathcal{A}} - n w_i \mu) / \mu = u_{\mathcal{A}} + v_{\mathcal{A}} \). The following lemma implies Lemma 5. Note this lemma does not require the ancillary to be exact.

**LEMMA 13.** Under the present conditions \( E[v_{\mathcal{A}}] = 0 \) and \( E[u_{\mathcal{A}}] = 0 \).

For \( v_{\mathcal{A}} \), we find that

\[
E[v_{\mathcal{A}}] = w_i \sum_{i=1}^{d} (E[i_{\mathcal{Y}_{\{1;2\}}, \mathcal{T}_i} - n] = w_i \left( E \left[ \sum_{i=1}^{d} \sum_{j=1}^{n} \mathcal{T}_i(j) i_{\mathcal{Y}_{\{n\}}}^{(i)} \right] - n \right)
\]

\[
= w_i \sum_{j=1}^{n} (E[i_{\mathcal{Y}_{\{n\}}}] - n) = 0,
\]

for all \( r = 1, \ldots, R_{\max} \). The first equality is from the definition of \( v_{\mathcal{A}} \); the second is from the definition of \( i_{\mathcal{Y}_{\{n\}}} \); the third is due to the fact that \( \sum_{j=1}^{n} \mathcal{T}_i(j) = 1 \); and the final equality is from \( E[i_{\mathcal{Y}_{\{n\}}}^{(i)}] = 1 \).

For \( u_{\mathcal{A}} \), start by noting that

\[
E\left[i_{\mathcal{Y}_{\{r\}}, \mathcal{T}_r}^{(i)} | \mathcal{A}_i \{1; r - 1\}\right] = E\left[\sum_{j=N^{(i)} \{1; r - 1\} + 1}^{N \{1; r\}} \mathcal{T}_r(j) E[i_{\mathcal{Y}_{\{n\}}}^{(i)} | \mathcal{T}_r(j)] \right] \mathcal{A}_i \{1; r - 1\}
\]
The first equality is from the definition of \( i^{(i)}_j \), \( \mathcal{S}(r) \) is independent of the past with unit expectation and that \( \mathcal{N}_{1 \mid r} \) is \( \mathcal{A}(1 \mid r - 1) \) measurable; the final equality is due to the definition of \( \mathcal{P}_i \). Consider the conditional expectation

\[
E[u_{\mathcal{A}(1 \mid r)} | \mathcal{A}(1 \mid r - 1)]
\]

\[
= E \left[ E[\mathcal{S}(1 \mid r) \mid \mathcal{A}(1 \mid r - 1)] | \mathcal{A}(1 \mid r - 1) \right]
\]

\[
= u_{\mathcal{A}(1 \mid r - 1)} + E \left[ \mathcal{S}(r) \mid \mathcal{A}(1 \mid r - 1) \right]
\]

\[
= u_{\mathcal{A}(1 \mid r - 1)} + E \left[ i^{(i)}_j \mathcal{Y}(r) \mathcal{T}(r) - w_i \sum_{j=1}^d i^{(i)}_j \mathcal{Y}(r) \mathcal{T}(r) \mid \mathcal{A}(1 \mid r - 1) \right]
\]

\[
= u_{\mathcal{A}(1 \mid r - 1)} + \mathcal{N}_{r} \left\{ \mathcal{P}_i \mid \mathcal{A}(1 \mid r - 1) \right\}
\]

The first equality is from the definition of \( u_{\mathcal{A}(1 \mid r)} \); the second follows from the fact that \( u_{\mathcal{A}(1 \mid r - 1)} \) and \( \mathcal{A}(1 \mid r - 1) \) are \( \mathcal{A}(1 \mid r) \) measurable; the third is from the definition \( \mathcal{S}(r) \); the fourth is from (S22); the final line is obtained by recognizing that \( \sum_{j=1}^d \mathcal{P}_i = 1 \).

Let \( \mathcal{G}(r) = I[\mathcal{M} \mid r > n - \mathcal{N}(1 \mid r - 1)] \). From Algorithm 1 the definition of \( \mathcal{P}_i \) differs based on whether \( \mathcal{G}(r) \) is 0 or 1. Note that \( \mathcal{G}(r) \) is \( \mathcal{A}(1 \mid r - 1) \) measurable. First, for the case where \( \mathcal{G}(r) = 0 \) we obtain

\[
E[\mathcal{S}(1 \mid r) \mid \mathcal{G}(r) = 0] = E[u_{\mathcal{A}(1 \mid r)} \mid \mathcal{A}(1 \mid r - 1), \mathcal{G}(r) = 0] = 0
\]

This is obtained directly by recognizing that if \( \mathcal{G}(r) = 0 \) then \( \mathcal{P}_i(r) = w_i - u_{\mathcal{A}(1 \mid r - 1)} / \mathcal{N}(r) \) and plugging this fact into (S22). Second, for the case where \( \mathcal{G}(r) = 1 \) we obtain

\[
E[u_{\mathcal{A}(1 \mid r)} \mid \mathcal{A}(1 \mid r - 2), \mathcal{G}(r) = 1]
\]

\[
= E[\mathcal{S}(1 \mid r) \mid \mathcal{A}(1 \mid r - 1), \mathcal{G}(r) = 1] \mid \mathcal{A}(1 \mid r - 2), \mathcal{G}(r) = 1
\]

\[
= E[\mathcal{S}(1 \mid r - 1) \mid \mathcal{A}(1 \mid r - 2), \mathcal{G}(r) = 1]
\]

\[
= E[u_{\mathcal{A}(1 \mid r - 1)} \mid \mathcal{A}(1 \mid r - 2), \mathcal{G}(r - 1) = 0]
\]

\[
= 0
\]

The first line is due to successive conditioning; the second is obtained by recognizing that if \( \mathcal{G}(r) = 1 \) then \( \mathcal{P}_i(r) = w_i \) and \( E[\mathcal{S}_i \mid \mathcal{G}(r) = 1] = E[\mathcal{S}_i \mid \mathcal{G}(r) = 1] = 0 \); the third line is again from successive conditioning, noting that if \( \mathcal{G}(r) = 1 \) then \( \mathcal{G}(r - 1) = 1 \) and the definition of \( u_{\mathcal{A}(1 \mid r - 1)} \); the final equality is from (S24). Equations (S24) and (S25) imply
that

\[ (S26) \quad E[u_{\tilde{\mathcal{A}}_i\{1;r\}}|G\{r\} = 0] = E[S_i\{1;r\}|G\{r\} = 0] = 0 \quad \text{and} \]

\[ (S27) \quad E[u_{\tilde{\mathcal{A}}_i\{1;r\}}|G\{r\} = 1] = E[S_i\{1;r\}|G\{r\} = 1] = 0. \]

The above verifies that

\[ E[u_{\tilde{\mathcal{A}}_i\{1;r\}}] = 0 \]

for all \( r \geq 2 \). Note, it is very straightforward to verify that \( E[u_{\tilde{\mathcal{A}}_i\{1\}}] = 0 \). This verifies that \( E[u_{\tilde{\mathcal{A}}_i\{1;r\}}] = 0 \) for all \( r \) and implies the stated result.

#### 16. Proof of Lemma 6

To begin note that Lemma implies that the following holds for fixed designs and Lemma 12 ensures that it extends to the RRSD

\[ E[\hat{\eta}_i - n|\tilde{\mathcal{A}}] = E[\hat{\eta}_i - n + \nu_{11}\hat{\eta}_i|\tilde{\mathcal{A}}] = E[\hat{Q}_i|\tilde{\mathcal{A}}] \]

\[ = E[\hat{Q}_i]\hat{\mathcal{A}} + O(1) = \hat{Q}_i + O(1). \]

Next recall that \( \hat{\eta}_i^{(k)}(y) = [\partial^{k}/\partial \eta^{k}] \log f_\eta(y) \) is the \( k \)-th derivative log likelihood for a single observation. The location condition implies that \( \hat{\eta}_i^{(k)}(y) = t_0^{(k)}(\varepsilon) \), where \( t_0^{(k)}(\varepsilon) = (-1)^k[\partial^{k}/\partial \varepsilon^{k}] \log f_0(\varepsilon) \). The importance of this is to say that log-likelihood and its derivatives depend only on the errors and not on the responses. The error distribution is unaffected by the design. The consequence of this is that since \( \sum_{i=1}^{d} \hat{Q}_i = \sum_{j=1}^{n} \hat{q}_{(j)} \) is a function of the first and second derivative of the log-likelihood is actually only a function of only the errors and thus its distribution is exactly the same as if no adaptation had taken place. This implies that

\[ n^{-1} \lim_{n \to \infty} \text{Var}[^{\hat{\mathcal{A}}}] = n^{-1} \lim_{n \to \infty} \text{Var}\left[ \sum_{i=1}^{d} \hat{Q}_i \right] = \gamma^2 \]

as stated.

#### 17. Proof of Lemma 7

To begin we write

\[ \text{Var}[u_{\tilde{\mathcal{A}}_i\{1;r\}}] = \text{Var}[E[u_{\tilde{\mathcal{A}}_i\{1;r\}}|G\{r\}]] + E[\text{Var}[u_{\tilde{\mathcal{A}}_i\{1;r\}}|G\{r\}]] \]

\[ = E[\text{Var}[u_{\tilde{\mathcal{A}}_i\{1;r\}}|G\{r\}]] + \text{Var}[u_{\tilde{\mathcal{A}}_i\{1;r\}}|G\{r\} = 0]P\{G\{r\} = 0\} \]

\[ + \text{Var}[u_{\tilde{\mathcal{A}}_i\{1;r\}}|G\{r\} = 1]P\{G\{r\} = 1\}. \]

(S28)

The first line is from the law of total variance; the second follows directly from (S26) and (S27); and the third is from the definition of variance. Consider the term

\[ \text{Var}[u_{\tilde{\mathcal{A}}_i\{1;r\}}|G\{r\} = 0] = \text{Var}[E[u_{\tilde{\mathcal{A}}_i\{1;r\}}|\mathcal{A}_i\{1;r-1\}, G\{r\} = 0]|G\{r\} = 0] \]

\[ = E[\text{Var}[u_{\tilde{\mathcal{A}}_i\{1;r\}}|\mathcal{A}_i\{1;r-1\}, G\{r\} = 0]|G\{r\} = 0] \]

\[ = E[\text{Var}[u_{\tilde{\mathcal{A}}_i\{1;r\}}|\mathcal{A}_i\{1;r-1\}, G\{r\} = 0]|G\{r\} = 0]. \]

(S29)
The first line is from the law of total variance and the second follows directly from (S26). Similarly, the term
\[
\text{Var}[u_i| \mathcal{A}_i = 1, r] = \text{Var} \left[ E[u_i| \mathcal{A}_i = 1, r-1, \mathcal{G}(r) = 1] | \mathcal{G}(r) = 1 \right]
\]
\[
= E \left[ \text{Var}[u_i| \mathcal{A}_i = 1, r-1, \mathcal{G}(r) = 1] | \mathcal{G}(r) = 1 \right]
\]
(S30)
\[
= E \left[ \text{Var}[u_i| \mathcal{A}_i = 1, r-1, \mathcal{G}(r) = 1] | \mathcal{G}(r) = 1 \right].
\]
The first line is from the law of total variance and the second follows directly from (S27). The following lemma provides a bound on both (S29) and (S30).

**Lemma 14.** Under the current conditions
\[
\text{Var}[u_i| \mathcal{A}_i = 1, r, \mathcal{G}(r) = g(r)]
\]
\[
= \text{Var} \left[ E[\tilde{S}_i(r) + \ldots + \tilde{S}_i| \mathcal{A}_i = 1, r, \mathcal{G}(r) = g(r)] | \mathcal{G}(r) = g(r) \right] + O_p(1)
\]
for \( s = r - 1 \) and \( r - 2 \), where \( g(r) = 0 \) or \( 1 \).

The steps of the proof of Lemma 4 can effectively be repeated with a simple change of notation to incorporate the additional runs. Due to the similarity with what has already been shown the proof of Lemma 14 has been omitted.

Additionally, from an application of the law of total variance
\[
\text{Var} \left[ E[\tilde{S}_i(r) + \ldots + \tilde{S}_i| \mathcal{A}_i = 1, r, \mathcal{G}(r) = g(r)] | \mathcal{G}(r) = g(r) \right]
\]
\[
\leq \text{Var} \left[ \tilde{S}_i(r) + \ldots + \tilde{S}_i| \mathcal{A}_i = 1, s, \mathcal{G}(r) = g(r) \right].
\]
Therefore (S28) can be bounded as
\[
\lim_{n \to \infty} n^{-1/2} \text{Var}[u_i| \mathcal{A}_i = 1, r] \leq \lim_{n \to \infty} n^{-1} E[(1 - \mathcal{G}(r)) \text{Var}[\tilde{S}_i(r)| \mathcal{A}_i = 1, r - 1]]
\]
\[
+ \lim_{n \to \infty} n^{-1/2}E[\mathcal{G}(r) \text{Var}[\tilde{S}_i(r) + \tilde{S}_i - 1| \mathcal{A}_i = 1, r - 2]]
\]
(S31)
\[
= \lim_{n \to \infty} n^{-1/2} E[(1 - \mathcal{G}(r)) \text{Var}[\tilde{S}_i(r)| \mathcal{A}_i = 1, r - 1]]
\]
\[
+ \lim_{n \to \infty} n^{-1/2}E[\mathcal{G}(r) \text{Var}[\tilde{S}_i(r)| \mathcal{A}_i = 1, r - 2]]
\]
(S32)
\[
+ \lim_{n \to \infty} n^{-1/2}E[\mathcal{G}(r) \text{Var}[\tilde{S}_i(r) - 1| \mathcal{A}_i = 1, r - 2]]
\]
(S33)
\[
+ \lim_{n \to \infty} n^{-1/2} E[\mathcal{G}(r) \text{Cov}[\tilde{S}_i(r), \tilde{S}_i(r) - 1| \mathcal{A}_i = 1, r - 2]].
\]

Consider the interior of the expectation in (S32)
\[
\text{Var}[\tilde{S}_i(r)| \mathcal{A}_i = 1, r - 2, \mathcal{G}(r) = 1]
\]
\[
= \text{Var} \left[ \sum_{j=N(1,r-1) + 1}^{N(1,r)} i_{\gamma(j)} \left( T_i(j) - w_i \right) | \mathcal{Q}(1, r - 2), \mathcal{G}(r) = 1 \right]
\]
\[
= \sum_{j=N(1,r-1) + 1}^{N(1,r)} \left( E[(T_i(j) - w_i) \text{Var}[i_{\gamma(j)}| \mathcal{G}(r) = 1] | \mathcal{G}(r) = 1] \right)
\]
\[
= \sum_{j=N(1,r-1) + 1}^{N(1,r)} E[(T_i(j) - w_i)^2 | \mathcal{G}(r) = 1]
\]
(S35)
\[
= a_1 N(r).
\]
where $a_1 = w_i(1 - w_i)$. One important aspect of the preceding derivations is that $\mathcal{G}\{r\} = 1$ implies that $r$ is the final run and that $\mathcal{P}_i\{r\} = w_i$ are fixed. Additionally, this implies that given $\mathcal{G}\{r\} = 1$ that $\tilde{T}_i(j)$ are independent Bernoulli random variables with probability $w_i$. The first equality is obtained from the definition of $\mathcal{S}_i\{r\}$; the second equality is from the preceding comment that given $\mathcal{G}\{r\} = 1$ then $r$ is the final run and that $\mathcal{N}\{r\} = n - \mathcal{N}\{1; r - 1\}$ is $\tilde{A}\{1; r - 2\}$ measurable and that $i_{\mathcal{Y}(j)}^{(h)}$ is a sum $\mathcal{N}\{r\}$ independent random variables; the third equality is obtained by recalling that $\mathbb{E}[i_{\mathcal{Y}(j)}^{(h)}] = 1$, $\text{Var}[i_{\mathcal{Y}(j)}^{(h)}] = \gamma^2$ and as previously described, given $\mathcal{G}\{r\} = 1$ that $\tilde{T}_i(j)$ are independent Bernoulli random variables with probability $w_i$; the final equality is by simplifying terms.

Using (S35) equation (S32) can be written as
\[
\mathbb{E}[\text{Var}[\mathcal{S}_i\{r\}|\tilde{A}\{1; r - 2\}, \mathcal{G}\{r\} = 1]|\mathcal{G}\{r\} = 1]P[\mathcal{G}\{r\} = 1] = a_1\mathbb{E}[\mathcal{G}\{r\}\mathcal{N}\{r\}] .
\]

Consider the interior of the expectation in (S31) and recall that $i_{\mathcal{Y}(j)}^{(h)}[\tilde{T}_i(j) - w_i], j = \mathcal{N}\{1; r - 1\}, \ldots, \mathcal{N}\{1; r\}$ given $\tilde{A}\{1; r - 1\}$ is a sequence of $\mathcal{N}\{r\}$ independent and identically distributed random variables; therefore,
\[
\text{Var}[\mathcal{S}_i\{r\}|\tilde{A}\{1; r - 1\}] = \sum_{j=\mathcal{N}\{1; r - 1\}+1}^{\mathcal{N}\{1; r\}} \text{Var}[i_{\mathcal{Y}(j)}^{(h)}[\tilde{T}_i(j) - w_i]|\tilde{A}\{1; r - 1\}] ,
\]
where
\[
\text{Var}[i_{\mathcal{Y}(j)}^{(h)}[\tilde{T}_i(j) - w_i]|\tilde{A}\{1; r - 1\}]
= \mathbb{E}\left[\text{Var}[i_{\mathcal{Y}(j)}^{(h)}[\tilde{T}_i(j) - w_i]|\tilde{T}_i(j)|\tilde{A}\{1; r - 1\}]ight]
+ \text{Var}\left[\mathbb{E}[i_{\mathcal{Y}(j)}^{(h)}[\tilde{T}_i(j) - w_i]|\tilde{T}_i(j)|\tilde{A}\{1; r - 1\}]ight]
= \gamma^2\mathbb{E}[(\tilde{T}_i(j) - w_i)^2|\tilde{A}\{1; r - 1\}] + \text{Var}[\tilde{T}_i(j)|\tilde{A}\{1; r - 1\}]
= \gamma^2\mathbb{E}[\tilde{T}_i(j)^2 - 2\tilde{T}_i(j)w_i + w_i^2|\tilde{A}\{1; r - 1\}] + \mathcal{P}_i\{r\}[1 - \mathcal{P}_i\{r\}]
= \gamma^2\{\text{Var}[\tilde{T}_i(j)|\tilde{A}\{1; r - 1\}] + \mathbb{E}[\tilde{T}_i(j)|\tilde{A}\{1; r - 1\}]^2
- 2\mathcal{P}_i\{r\}w_i + w_i^2\} + \mathcal{P}_i\{r\}[1 - \mathcal{P}_i\{r\}]
= \gamma^2\{\mathcal{P}_i\{r\}[1 - \mathcal{P}_i\{r\}] + \mathcal{P}_i^2\{r\} - 2\mathcal{P}_i\{r\}w_i + w_i^2\} + \mathcal{P}_i\{r\}[1 - \mathcal{P}_i\{r\}]
\leq a_2 ,
\]
where $a_2 = (\gamma^2 + 1)/4$. Using the above we obtain
\[
\text{Var}[\mathcal{S}_i\{r\}|\tilde{A}\{1; r - 1\}] \leq a_2\mathcal{N}\{r\} .
\]

The above was derived for an arbitrary $r$; therefore, (S33) can be similarly bounded as
\[
\text{Var}[\mathcal{S}_i\{r - 1\}|\tilde{A}\{1; r - 2\}] \leq a_2\mathcal{N}\{r - 1\} .
\]

For the covariance term in (S34) first recognize that
\[
\mathbb{E}[\mathcal{S}_i\{r\}\mathcal{S}_i\{r - 1\}|\tilde{A}\{1; r - 2\}, \mathcal{G}\{r\} = 1]
= \mathbb{E}[\mathcal{S}_i\{r - 1\}\mathbb{E}[\mathcal{S}_i\{r\}|\tilde{A}\{1; r - 1\}, \mathcal{G}\{r\} = 1]|\tilde{A}\{1; r - 2\}, \mathcal{G}\{r\} = 1]
= 0 .
\]
The first line is from successive conditioning and the second is due to the fact that
\( E[\tilde{S}_i \{r\} | \bar{A}_i \{r; r-1\}, G\{r\} = 1] = 0 \). The preceding implies that
\[ \text{Cov}[\tilde{S}_i \{r\}, \tilde{S}_i \{r-1\} | \bar{A}_i \{r; r-2\}, G\{r\} = 1] = 0 \]
which in turn implies that (S34) equals 0.

Finally, (S28) can be bounded as
\[
\lim_{n \to \infty} n^{-1/2} \text{Var}[u_{\bar{A}_i \{1; r\}}] \\
\leq \lim_{n \to \infty} n^{-1/2} \left[ a_2 E[(1 - G\{r\})N\{r\} | r] + a_1 E[G\{r\}N\{r\}] + a_2 E[G\{r\}N\{r-1\}] \right]
\]
(S36) \leq \lim_{n \to \infty} n^{-1/2} \left[ (a_1 + a_2)E[N\{r\}] + a_2 E[N\{r-1\}] \right],
since \( a_1, a_2 > 0 \).

A useful expression for bounding the right hand side of the above is
\[
E[N\{r\}] \leq E \left[ \max_i \left\{ \frac{u_{\bar{A}_i \{1; r-1\}}}{w_i} \right\} + 1 \right]
\leq \left( E \left[ \max_i \left\{ \frac{u_{\bar{A}_i \{1; r-1\}}}{w_i} \right\}^2 \right] \right)^{1/2} + 1
\leq \left( E \left[ \max_i \left\{ \left( \frac{u_{\bar{A}_i \{1; r-1\}}}{w_i} \right)^2 \right\} \right] \right)^{1/2} + 1
\leq \left( E \left[ \sum_{i=1}^d \left( \frac{u_{\bar{A}_i \{1; r-1\}}}{w_i} \right)^2 \right] \right)^{1/2} + 1
\leq \left( \sum_{i=1}^d w_i^{-2} E \left[ u_{\bar{A}_i \{1; r-1\}}^2 \right] \right)^{1/2} + 1
= w_{\min}^{-1} \left( \sum_{i=1}^d \text{Var} u_{\bar{A}_i \{1; r-1\}} \right)^{1/2} + 1
\]
which holds for all \( r = 3, \ldots, R_{\text{max}} \), where \( w_{\min} = \min_i w_i > 0 \). The first line is from the definition of \( N\{r\} \) in Algorithm 1, where the +1 is to account for the ceiling operator; the second line is due to Holder; the third is from a property of maximum of a list; the fourth is from the fact that the maximum of a list of squares is less than the total of the squares; the fifth is obtained by rearranging terms; the sixth is a result of the fact that \( E [u_{\bar{A}_r}] = 0 \).

Consider that
\[
\lim_{n \to \infty} n^{-1/2} \sum_{i=1}^d \text{Var}[u_{\bar{A}_i \{1; r\}}] \\
\leq \lim_{n \to \infty} n^{-1/2} \left[ (a_1 + a_2)w_{\min}^{-1} \left( \sum_{i=1}^d \text{Var} u_{\bar{A}_i \{1; r-1\}} \right) \right]^{1/2}
+ a_2 w_{\min}^{-1} \left( \sum_{i=1}^d \text{Var} u_{\bar{A}_i \{1; r-2\}} \right)^{1/2}
\leq \lim_{n \to \infty} n^{-1/2} d(a_1 + a_2) \left( w_{\min}^{-1} \left( d(a_1 + a_2)E[N\{r-1\}] + da_2 E[N\{r-2\}] \right) \right)^{1/2}
\]
First it is shown that \( W_d > 1 \) for large values of \( N \). Note, it is very likely that \( N \) is due to summing over a form of the approximation bound, \( O(n^{1/2}) \), could be considerably refined; however, the result as stated is sufficient for the current work. This concludes the proof of Lemma 6.

18. Proof of Lemma 8 First it is shown that \( D \) is positive semi-definite. Let \( P = HM_\xi^{-1}H^T \) denote the hat matrix which is well known to be positive semi-definite, where \( H \) is a \( d \times p \) matrix with \( i \)th row equal to \( \hat{\eta}(x_i) \). Therefore, \( D \) can be expressed as 
\[
D = RP_R = RP^{1/2}P^{1/2}R,
\]
where \( R = \text{diag}(r) \). This implies \( D \) is positive semi-definite.

Let \( \overline{W} = (W - \hat{w}w^T) \); if \( d = 1 \) then \( tr(D\overline{W}) = 0 \) and equality is achieved. If \( d > 1 \) then \( \overline{W} \) is positive definite and \( \lambda_{\min}(\overline{W}) > 0 \). Since \( D \) is positive semi-definite, Kleinman and Athans (1968) implies that
\[
tr(D\overline{W}) \geq \lambda_{\min}(\overline{W})tr(D) \geq 0
\]
Therefore, if \( d > 1 \) (S37) can only equal 0 only if \( tr(D) = 0 \). It is useful to use the expanded form of \( tr(D) \) written as
\[
tr(D) = \sum_{i=1}^{d} \sum_{k=1}^{d} c^T M_\xi^{-1} \hat{\eta}(x_i)\hat{\eta}^T(x_i)M_\xi^{-1} \hat{\eta}(x_k)\hat{\eta}^T(x_k)M_\xi^{-1}c
\]
\[
\geq c^T \left[ \sum_{i=1}^{d} w_i M_\xi^{-1} \hat{\eta}(x_i)\hat{\eta}^T(x_i)M_\xi^{-1} \sum_{k=1}^{d} w_k \hat{\eta}(x_k)\hat{\eta}^T(x_k)M_\xi^{-1} \right]c
\]
\[
= c^T M_\xi^{-1} \sum_{i=1}^{d} w_i \hat{\eta}(x_i)\hat{\eta}^T(x_i)M_\xi^{-1}c
\]
\[
= c^T M_\xi^{-1}c
\]
\[
> 0
\]
for all non-zero \( c \) per the condition that inverse of \( M_\xi \) exists. This concludes the proof.

REFERENCES

ATKINSON, A. C., DONEV, A. N. and TOBIAS, R. (2007). Optimum Experimental Designs with SAS. Oxford Statist, Sci. Series, 34, Oxford University Press.

ATKINSON, A., FEDOROV, V., HERZBERG, A. and ZHANG, R. (2014). Elemental information matrices and optimal experimental designs for generalized regression models. J. Statist. Plann. Inference 144 81-91.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). Inference and Asymptotics. Chapman and Hall, London.

BASU, D. (1969). Recovery of ancillary information. Sankhya, A 26 3-16.

BHATTACHARYYA, A. (1948). On some analogues of the amount of information and their use in statistical estimation. Sankhya 8 315-328.

BOBROVSKY, B. Z., MAYER-WOLF, E. and ZAKAI, M. (1987). Some classes of global Cramér-Rao bounds. Ann. Stat. 15 1428-1438.

BOX, G. E. P. and HUNTER, W. G. (1965). Sequential design of experiments for nonlinear models. In Proceedings of the Scientific Computing Symposium: Statistics (J. J. Korth, ed.) 113-137. White Plains: IBM.

BRANNATH, W., KNIG, F. and BAUER, P. (2006). Estimation in flexible two stage designs. Stat. Med. 25 3366-3381.
Coad, D. S. and Ivanova, A. (2001). Bias calculations for adaptive urn designs. *Seq. Anal.* **20** 91-116.

Cordeiro, G. M. and Cribari-Neto, F. (2014). *An Introduction to Bartlett Correction and Bias Reduction.* Springer, New York.

Cox, D. R. (1958). Some problems with statistical inference. *Ann. Math. Stat.* **29** 357-372.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimate: observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–483.

Fedorov, V. V. and Leonov, S. L. (2013). *Optimal Design for Nonlinear Response Models.* CRC Press, Boca Raton.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27-38.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc.* **144** 285-307.

Flournoy, N. and Oron, A. P. (2019). Bias induced by adaptive dose-finding designs. *J. Appl. Stat.* DOI: 10.1080/02664763.2019.1649375.

Fraser, D. A. S. (2004). Ancillaries and conditional inference. *Statist. Sci.* **19** 333-369.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014). *Bayesian Data Analysis,* Third ed. CRC Press, Boca Raton, FL.

Ghosh, M., Reid, N. and Fraser, D. A. S. (2010). Ancillary statistics: a review. *Statist. Sinica* **20** 1309-1332.

Groves, T. and Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika* **3** 690-691.

Hall, P. and Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika* **75** 661-671.

Hu, F. and Rosenberger, W. F. (2003). Optimality, variability, power: evaluating response-adaptive randomization Procedures for Treatment Comparisons. *J. Amer. Statist. Assoc.* **98** 671-678.

Kiefer, J. (1959). Optimal experimental designs (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **21** 272-319.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *Ann. Math. Stat.* **30** 271-294.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.* **14** 363-366.

Kim, S. and Flournoy, N. (2014). Optimal experimental design for systems with bivariate failures under a bivariate weibull function. *J. R. Stat. Soc. Ser. C. App. Stat.* **64** 413-432.

Kleinman, D. L. and Athans, M. (1968). The design of suboptimal linear time-varying systems. *IEEE Trans. Automat. Contr.* **AC-13** 159-159.

Kosmidis, I. and Firth, D. (2009). Bias reduction in the exponential family nonlinear models. *Biometrika* **96** 793-804.

Lai, T. L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* **7** 1196-1212.

Lane, A. (2019). Optimality of observed information adaptive designs. *arXiv:1912.08162*.

Lane, A. (2020). Adaptive designs for optimal observed Fisher information. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **4** 1029-1058.

Lane, A. (2021). Optimal relevant subset designs in nonlinear models. *arXiv:2106.09633*.

Lane, A. (2022). Conditional information and inference in response-adaptive allocation designs. *Stat. in Med.* **41** 390-406.

Lane, A., Yao, P. and Flournoy, N. (2014). Information in a two-stage adaptive optimal design. *J. Statist. Plann. Inference* **144** 173–187.

Li, K. C. (1983). Minimaxity for randomized designs: some general results. *Ann. Stat.* **11** 225-239.

Marschner, I. C. (2021). A general framework for the analysis of adaptive experiments. *Statistical Science* **36** 465-492.

McCullagh, P. (1992). Conditional inference in Cauchy models. *Biometrika* **79** 247-259.

Pronzato, L. and Pazman, A. (1994). Bias correction in nonlinear regression via two-stages least-squares estimation. In *Proc. 10th IFAC/IFORS Symp. on Identification and System Parameter Estimation* (M. Blanke and T. Soderstrom, eds.) 137-142. Danish Automation Society.

Pukelsheim, F. (2006). *Optimal design of experiments.* SIAM, Philadelphia, PA.

Van Trees, H. L. (1968). *Detection, Estimation and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory.* Wiley, New York.

Waite, T. W. and Woods, D. C. (2021). Minimax efficient random experimental design strategies with application to model-robust design for prediction. *J. Amer. Statist. Assoc.* DOI:10.1080/01621459.2020.1863221.

Wu, C. F. (1981). On the robustness and efficiency of some randomized designs. *Ann. Stat.* **9** 1168-1177.