Optimality of Observed Information Adaptive Designs in Linear Models

Adam Lane
Cincinnati Children’s Hospital Medical Center
Adam.Lane@cchmc.org

Abstract

This work considers experimental design in linear models with additive errors. A traditional objective in design is to minimize the variance of the estimates of the model parameters. The optimal design, which is found by minimizing a convex function of the expected Fisher information, accomplishes this objective, approximately. The inverse of expected Fisher information is asymptotically equivalent to the variance of the maximum likelihood estimate. It is often remarked that observed Fisher information is a better measure of the variance of the maximum likelihood estimate than the expected Fisher information [Efron and Hinkley (1978)]. However, unlike expected Fisher information, observed Fisher information depends on the observed data and cannot be used to design an experiment in advance of data collection. In a sequential experiment the observed Fisher information from past observations is available to incorporate into the design of the current observation. In this work an adaptive design that incorporates observed Fisher information is proposed. It is shown that this proposed design is optimal, at the limit, with respect to inference and conditional mean square error. In a simulation study the proposed adaptive design performs nearly uniformly better than the optimal design.

Keywords: adaptive design, conditional inference, observed Fisher information, relevant subsets, optimal design.

1 Introduction

An observed information adaptive design is a sequential design that incorporates the observed Fisher information available from past observations into the design of the current and future observations. Lane (2017) introduced observed information adaptive designs and demonstrated, primarily with heuristics and a simulation study, that they can reduce the mean square error (MSE) of the maximum likelihood estimate (MLE). This reduction was demonstrated in comparison to the MSE of the MLE from fixed optimal designs and adaptive optimal designs, described below.

In the current work a linear regression model with additive errors is considered. The properties of the linear model are used to propose an observed information adaptive design. The main result of this work is that the proposed design is optimal with respect to inference and conditional MSE, at the limit.

Throughout, it is assumed the error distribution is a member the location family. A distribution is a member of the location family if its density, \( f_\eta \), has the property that \( f_\eta(y) = f_0(y - \eta) \). The importance of the location family is briefly described here and is elaborated on in Sections 2 and 3. Fisher (1934) described the conditionality resolution for the location family. A conditionality resolution establishes the existence of a pair, \((\hat{\eta}, a)\), that is a one to one function of the minimal sufficient statistic, where \(\hat{\eta} \) is the MLE and \(a\) is an ancillary random variable. A random variable is ancillary if its distribution is independent of the model parameters. Fisher referred to \(a\) as the configuration statistic; in this work, \(a\) is referred to as the ancillary configuration statistic to emphasize its ancillary nature. The conditionality resolution implies that, in general, \(\hat{\eta}\) is not sufficient and inference based on the MLE alone represents a loss of information. Conditional on the ancillary configuration statistic reduces the sample space to a relevant subset and the full information available in the data is recovered in the distribution \(\hat{\eta}|a\). Cox (1958), Fisher (1961), Efron and Hinkley (1978), McCullagh (1992), Sundberg (2003), Fraser (2004) and Ghosh, Reid and Fraser (2010) argue in favor of inference conditional on relevant subsets. If inference is defined as the process of assessing the variability of an estimate, with a standard error, then the conclusion is that \(\text{Var}[\hat{\eta}|a]\) is a more relevant measure of the variability of the MLE than \(\text{Var}[\hat{\eta}]\). Barndorff-Nielsen (1980) generalized this conditionality resolution to transformation models.
In practice, the conditional distribution, \( \hat{\eta} | \mathbf{a} \), is often intractable or unknown. Efron and Hinkley (1978) demonstrated, in the location family, that observed Fisher information can be viewed as a precise approximation to the information in the relevant subset, see (2) and (3) below. Fisher (1934) showed that the joint distribution of responses can be factored as

\[
f(\mathbf{y}) = f(\hat{\eta}, \mathbf{a}) = c(\mathbf{a}) g_a(\hat{\eta} - \eta),
\]

where \( \mathbf{y} \) represents a vector of independent random responses drawn from \( f_\eta \). Assuming derivatives and integrals are exchangeable, the expected and observed Fisher information, with respect to \( \eta \) are

\[
\mathcal{F} = E \left[ \frac{\partial^2 \log g_a(\hat{\eta} - \eta)}{\partial \eta^2} \right] \quad \text{and} \quad i_a = \left[ \frac{\partial^2 \log g_a(\hat{\eta} - \eta))}{\partial \eta^2} \right]_{\eta=\hat{\eta}},
\]

respectively. In a seminal paper Efron and Hinkley (1978) establish that

\[
\begin{align*}
\var[\hat{\eta} | \mathbf{a}] &= i_a^{-1} \{1 + O_p(n^{-1})\} \quad \text{(2)}
\qquad \qquad \qquad \quad \text{Var}[\hat{\eta} | \mathbf{a}] = \mathcal{F}^{-1} \{1 + O_p(n^{-1/2})\}. \quad \text{(3)}
\end{align*}
\]

In the location family, \( i_a \) and \( \mathcal{F} \) do not depend on \( \eta \); and, further, \( i_a \) is itself ancillary. McCullagh (1984) extends (2) and (3) to scale parameters with \( \mathbf{a} \) replaced by an approximate ancillary statistic. The role of approximate ancillary statistics in conditional inference has received significant attention [e.g. Cox (1980), Skovgaard (1986) and Reid and Fraser (2010)].

To summarize, the conditional variance is the appropriate measure of the variability of the MLE and the observed Fisher information is a better approximation of the conditional variance than the expected Fisher information. Therefore, it is more appropriate to use observed Fisher information for statistical inference. Observed Fisher information is relevant outside of the location family and has been considered in many areas of statistical research, see Louis (1982), Firth (1993), Barndorff-Nielsen and Sorensen (1994), Murphy and Vaart (1999), Lystig and Hughes (2002) and Reid (2003). In adaptive design May and Flournoy (2009) and Lin, Flournoy and Rosenberger (2019) consider the asymptotic distribution of parameter estimates normalized with observed Fisher information and Lane (2017) proposed adaptive designs for observed Fisher information.

The above discussion was presented for a one dimensional parameter \( \eta \). In linear regression, described in Section 2, \( \eta = \beta^T f_\sigma(x) \), where \( \beta \) is a \( p \) dimensional vector and \( f_\sigma(x) \) is a function of the design points. As a result, expected and observed Fisher information, with respect to \( \beta \), are \( p \times p \) matrices. The extension of (2) and (3) to linear regression is discussed in Section 2.

The proposed design in this work is rooted in the theory of optimal design, which has received and continues to receive significant attention. Many of the early foundational works are due to Kiefer (1959, 1961) and Kiefer and Wolfowitz (1959, 1960). Throughout, this is referred to as fixed optimal design, since it is fixed in advance of data collection. In fixed optimal design a traditional objective is to minimize the inverse of expected Fisher information. In general, expected Fisher information is a matrix and the minimum of its inverse does not exist. Instead the minimization is done with respect to a convex optimality criterion selected according to the objectives of the experiment. For example, the \( D \)-optimal criterion (the negative log determinant) minimizes the volume of the confidence ellipsoid of the parameter estimates. See Fedorov (1972), Pukelsheim (2006) and Atkinson, Donev and Tobias (2007) for further details regarding optimal design.

Analogous to optimal design, Lane (2017) defined the objective of observed information adaptive designs as minimizing the inverse of observed Fisher information, subject to a convex optimality criterion. Unlike expected Fisher information, which is known \textit{a priori}, observed Fisher information depends on the observed data and is unknown prior to data collection. If the data can be collected sequentially, in a series of runs, then the observed Fisher information from the past runs is available and can be used to determine the design of the current and future runs. Incorporating the observed Fisher information into the design can be viewed as designing the experiment to maximize the information in the relevant subset rather than with respect to the entire sample space.

In linear models the expected Fisher information does not depend on the model parameters, and as a result fixed optimal designs are considered \textit{globally} optimal in the sense that are optimal over the entire
parameter space. The consequence of this is that adaptive optimal designs are not applicable in the context of the current manuscript. Adaptive optimal designs are a class of designs used in the context of locally optimal design. A design is locally optimal if it depends on the model parameters. Box and Hunter (1965) introduced adaptive optimal designs and they have been further investigated in Dragalin and Fedorov (2005), Dragalin, Fedorov and Wu (2007) and others.

Fixed optimal designs, reviewed in Section 3, represent the most relevant comparative method to the design proposed in this paper. In this same section the theory of fixed optimal design is used to develop the observed information adaptive design proposed in this work.

There are two commonly described challenges associated with inference following adaptive designs. As stated in Ford et al. (1985), after the completion of an adaptive procedure the design is not ancillary and thus conditional inference may represent an information loss [Rosenberger and Lachin (2002) Chapter 11]. The design proposed in this paper uses only the ancillary configuration statistics from past observations to select the design for the current observation. Therefore, the design remains ancillary and there is no information lost by conditioning. The second challenge is an induced dependence in the responses. This dependence can negatively impact inference. Specifically, bias and non-approximate normality of the parameter estimates are potential side effects [Lane and Flournoy (2012), Lane, Yao and Flournoy (2014), Lane, Wang and Flournoy (2016), and others]. It is shown in section 4 that these negative side effects do not occur for the design proposed in Section 4. Further, it is shown that inference can be conducted conditionally exactly as if the design were fixed in advance.

The main result of this paper, stated in Section 4, is that the proposed adaptive design is optimal, at the limit, with respect to inference and conditional MSE relative to the corresponding fixed optimal design. The proposed design and the corresponding fixed optimal design are compared in a simulation study in section 5. The simulation study includes different error distributions, mean functions, sample sizes and optimality criteria. The proposed adaptive design performed nearly uniformly better than the fixed optimal design over the entire set of simulated examples.

2 Model and Information

In design it is assumed that independent replicates can be observed at any \( x \) within the design region, denoted \( \mathcal{X} \), at the experimenters discretion. An exact design is comprised of a set of \( d \) design, or support, points, \( x_i \), with corresponding weights, \( w_i = n_i/n \), for \( i = 1, \ldots, d \), and is denoted

\[
\xi_n = \left\{ \begin{array}{c} x_1 \\ w_1 \\ x_2 \\ w_2 \\ \ldots \\ x_d \\ n_d \end{array} \right\},
\]

where the total sample size is \( n = \sum n_i \). In this section it is assumed responses were obtained without adaptation at some predetermined design \( \xi_n \). In section 3 a sequential experiment is introduced. Let \( y_i = (y_{i1}, \ldots, y_{in})^T \) represent the vector of responses observed at the \( i \)th design point and let \( y = (y_1^T, \ldots, y_n^T)^T \) be the complete \( n \times 1 \) response vector. It is assumed the responses were observed from the following linear regression model

\[
y_{ij} = \beta^T f_s(x_i) + \varepsilon_{ij}, \tag{4}
\]

where \( \beta \) is a \( p \times 1 \) vector within the parameter space \( B \); \( x \) is a \( s \times 1 \) vector within the design region, \( \mathcal{X} \); \( \mathcal{X} \) is a compact subset of \( \mathbb{R}^s \) and \( f_s(x) \) is a mapping from \( \mathcal{X} \) to \( \mathbb{R}^p \). The mean function for the \( i \)th design point will, at times, be denoted \( \eta_i = \beta^T f_s(x_i) \). Further, it is assumed that \( \varepsilon = (\varepsilon_1^T, \ldots, \varepsilon_n^T)^T \), where \( \varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in})^T \), is a vector of \( n \) independent and identically distributed (i.i.d.) random variables satisfying the location family constraint \( f_0(\varepsilon) = f_0(y - \eta) = f_0(y) \).

2.1 Information

In the introduction the information with respect to \( \eta \) was described. Atkinson, Fedorov, Herzberg and Zhang (2014) refer to the per-observation expected Fisher information with respect to \( \eta \) as the expected elemental information. Denote the log likelihood as \( l_j(\eta) = \log f_0(y_j - \eta) \). Let \( \hat{l}_j(\eta) = (\partial/\partial \eta) l_j(\eta) \) and \( \tilde{l}_j(\eta) = (\partial^2/\partial \eta^2) l_j(\eta) \), then, assuming that derivatives and integrals can be exchanged, the expected elemental information is

\[
\mu = -E \left[ \tilde{l}_j(\eta) \right].
\]
From the i.i.d. nature of the errors $\mathcal{F} = n\mu$. Lane (2017) defines the observed elemental information as the per-observation observed Fisher information with respect to $\eta$. In the current context the observed elemental information is $i_a/n$. As previously stated, Efron and Hinkley (1978) establish that $i_a^{-1}$ is a better approximation to the conditional variance of the MLE than $\mathcal{F}^{-1}$ and should be used for inference.

The information in the regression model is constructed from the elemental information. Let $a_i$ be the ancillary configuration statistic for the $i$th design point and define the ancillary configuration matrix as $A = (a_1, \ldots, a_d)^T$. Further, let $\tilde{\eta}_i$ be the MLE for the $i$th design point with respect to the parameterization $\eta_i$ and $\tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_d)^T$.

From the factorization given in (1) and assuming derivatives and integrals can be exchanged, the per-subject expected Fisher information for model (4) with design $\xi_n$ is

$$M(\xi_n) = \frac{1}{n} \sum_{i=1}^d \mathbb{E} \left[ \frac{\partial^2 \log g_a[\tilde{\eta}_i - \beta^T f_x(x_i)]}{\partial \beta^2} \right] = \mu F^T WF,$$

where $F$ is a $n \times p$ matrix with $ith$ row $f_x(x_i)^T$ and $W = \text{diag}(w_1, \ldots, w_d)$. The per-subject expected Fisher information is denoted $M = M(\xi_n)$ for shorthand, when the meaning is clear. For a regression model with errors in the location family $M$ does not depend on the model parameters.

The per-subject observed Fisher information at $\beta'$ in linear regression is defined as

$$J_A(\beta', x) = \frac{1}{n} \sum_{i=1}^d \left[ \frac{\partial^2 \log g_a[\hat{\eta}_i - \beta^T f_x(x_i)]}{\partial \beta^2} \right]_{\beta = \beta'},$$

where $i_a$ is the per-observation observed Fisher information with respect to $\eta_i$ and $\tilde{\eta}$ with respect to the parameterization $\eta$ and $\tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_d)^T$.

A traditional representation of observed Fisher information is obtained by letting $\beta' = \hat{\beta}$, where $\hat{\beta}$ is the MLE of $\beta$. This is the definition used in one of the observed information adaptive designs proposed in Lane (2017). Recalling that $\eta_i = \beta^T f_x(x_i)$ it can be seen that by evaluating $\eta_i$ at $\hat{\eta}_i$ one obtains the observed elemental information, in the square brackets of (5), for $i = 1, \ldots, d$. As was shown in the introduction, the observed elemental information does not depend on $\hat{\eta}_i$; therefore,

$$J_A(x) = \frac{1}{n} \sum_{i=1}^d \left[ \frac{\partial^2 \log g_a[\tilde{\eta}_i - \eta_i]}{\partial \eta^2} \right]_{\eta_i = \hat{\eta}_i} f_x(x_i)f_x(x_i)^T = \frac{1}{n} F^T I_A F,$$

is a alternative definition of observed Fisher information, where $I_A = \text{diag}(i_a, \ldots, i_a)$. The dependence of the observed Fisher information on $x$ will be omitted when the meaning is clear. Throughout the remainder of this work any reference to observed Fisher information refers to $J_A$ unless explicitly stated otherwise.

### 2.2 Weighted Least Squares Estimate

The observed Fisher information, as defined in (6), is not a standard one used in linear regression. The more common definition is $J_A(\hat{\beta}, x)$. In this section (6) is justified, approximately, for a weighted least squares estimate (WLSE) of $\beta$. From the independence of responses (2) implies that

$$\text{Var}[\hat{\eta}_i] = I^{-1}_A[1 + O_p(n^{-1})].$$

The above representation suggests that from the conditional perspective a weighted least squares approach represents a straightforward alternative to the MLE. Specifically, the WLSE of $\beta$ proposed is

$$\tilde{\beta} = [F^T I_A F]^{-1} F^T I_A \tilde{\eta}.$$

For the WLSE (2) and (3) extend to the regression setting, i.e.,

$$n\text{Var}[\tilde{\beta}] = (F^T I_A F)^{-1} F^T I_A \text{Var}[\tilde{\eta}] F (F^T I_A F)^{-1} = J_A^{-1}[1 + O_p(n^{-1})].$$

and

$$n\text{Var}[\tilde{\beta}] = M^{-1}[1 + O_p(n^{-1/2})].$$
It is intuitive that the above might also hold for the MLE of $\hat{\beta}$ as well; but it is more involved to show. If the conditional argument, that $\text{Var}[\hat{\beta} | A]$ is a more appropriate measure of the variability of the WLSE than $\text{Var}[\hat{\beta}]$ is accepted, then it follows that the inverse of observed Fisher information, $J^{-1}$, is a more appropriate variance approximation than the inverse of expected Fisher information, $M^{-1}$. To summarize, the primary point of this section has been to show that inference should be based on the entries of the observed Fisher information and not the entries of the expected Fisher information. The discussion in this section has been under the assumption of a non-sequential experiment. In Section 4 it is shown that (7) and (8) extends to the observed information adaptive design proposed in Section 3.

A remark on the WLSE; in this paper the WLSE of $\beta$ is considered due to its convenient form. The WLSE is not expected to be an improvement over the MLE in terms of efficiency. In fact, unconditionally they share the same order of approximation to $M^{-1}$, i.e.,

$$n\text{Var}[\hat{\beta}] = M^{-1}[1 + O(n^{-1})] \quad \text{and} \quad n\text{Var}[\hat{\beta}] = M^{-1}[1 + O(n^{-1})].$$

All the examples presented in Section 5 were repeated for the MLE, see the supplemental materials, and it was found that the MLE and the WLSE performed similarly, with respect to efficiency.

3 Optimal Designs

This section has two parts. The first part reviews important characteristics of fixed optimal design. In the second part the notation for a sequential experiment is introduced and an observed information adaptive design (OAD) for the linear model is proposed.

3.1 Fixed Optimal Design

The fixed optimal design (FOD) minimizes $M$, with respect to a convex criterion function, denoted $\Psi(\cdot)$. The FOD is considered fixed in that it is not adaptive. Specifically, the FOD is defined as

$$\xi^* = \arg \min_{\xi \in \Xi} \Psi \{M(\xi)\},$$

where $\Xi$ represents the set of all permissible designs. The set of permissible designs has two common characterizations, exact and continuous. An optimal design is exact design if (9) is solved for $\Xi = \Xi_n$, where $\Xi_n$ represents the set of all possible exact designs as defined in the preceding section. For continuous designs the integer restriction is relaxed such that $\Xi = \Xi_\Delta$, where $\Xi_\Delta$ contains designs where $w_i$ satisfy $0 \leq w_i \leq 1$ and $\sum_i w_i = 1$. In the current work the focus is on continuous optimal design.

In this work it is assumed $\Psi$ is as a positive-homogeneous convex optimality criterion, see Pukelsheim (2006); in practice, $\Psi$ is selected according to the primary objective(s) of the experiment. For non-singular $M$ the $D$-optimal design, with $\Psi(M) = |M|^{-1/p}$, minimizes the confidence ellipsoid of the parameter estimates. The $A$-optimal design, with $\Psi(M) = \text{Tr}(M^{-1})$, minimizes the average variances of the parameter estimates. A $c$-optimal design minimizes the variance of a linear combination of the parameter estimates, $c^T \hat{\beta}$, where $c$ is a known vector of constants. In $c$-optimal design $\Psi(M) = c^T M^{-1} c$. Atkinson, Donev and Tobias (2007) [ch. 10] provide a review of many commonly used optimality criteria. For shorthand $M^*_s = M(\xi^*_s)$ and $\Psi^* = \Psi(M^*_s)$ are used.

An important characteristic of the linear model is that the FOD does not depend on the model parameters. As a result all FODs are globally optimal, in the sense that they minimize $\Psi$, evaluated at the expected Fisher information, for all $\beta \in B$. However, it will be shown in Section 4 that FOD are less efficient than the OAD proposed in the next section, with respect to inference and conditional MSE.

Designs for linear models are well researched and there exist ubiquitous resources available to find FODs. In special cases, there exist analytic solutions to the optimal design problem [Guest (1958) and Farrell, Kiefer and Walbran (1968)]. Analytic solutions are the exception and in general it is required to use an algorithmic search to find FODs. Algorithms to find continuous designs are ubiquitous, with early examples found in Wynn (1970) and Fedorov (1972). The continuous optimal design problem remains an active area of research and there exist many modern solutions [Yu (2011), Yang, Biedermann and Tang (2013), Harman, Filov and Richtrik (2019) and others]. In practice all experiments must be evaluated at an exact design. To find exact designs there are two common approaches. A continuous optimal design can be rounded to satisfy the integer restriction [Pukelsheim and Rieder (1992)]. Alternatively, search algorithms can be used.
to directly find exact designs; Fedorov (1972), Mitchell (1974) and Atkinson, Donev and Tobias (2007) are classical examples of exact design search algorithms.

Many algorithms to construct continuous FODs are sequential. In the FOD setting sequential means that the design points (and not the responses) for the preceding \( j - 1 \) observations are available to determine the design for the \( j \)th observation. Once the optimal design is found it is not necessary to conduct the experiment sequentially. This is different from the definition of sequential in the adaptive design setting, where both the design points and responses of past observations are known. Adaptive optimal designs have been proposed by extending sequential algorithms [Dragalin and Fedorov (2005), Dragalin, Fedorov and Wu (2007) and others].

In this section a general first order sequential algorithm to find continuous optimal designs is reviewed. This approach will be adapted in the next section to optimize observed Fisher information. In \( D \) optimal design the first order algorithm is often referred to as the Fedorov-Wynn algorithm [Wynn (1970) and Fedorov (1972)]. Fedorov (2010) reviews the first order approach for a general \( \Psi \). The first order approach is based on the general equivalence theorem [Kiefer and Wolfowitz (1960)]. Let \( \xi \) be a design with support, \( x \), and unit allocation. The sensitivity function is defined as the derivative of \( \Psi \) in the direction of \( \xi \), i.e., the sensitivity function is

\[
\phi(x, \xi) = \nabla_{\xi} \Psi \{ M(\xi) \},
\]

where \( \nabla_{\xi} \) is the directional derivative. The general equivalence theorem states that the following are equivalent

1. \( \xi^* \) is the continuous optimal design
2. \( \xi^* = \arg \max_{\xi \in \Xi_\Delta} \min_{x \in X} \phi(x, \xi) \) and
3. \( \min_{x \in X} \phi(x, \xi^*) = 0 \) and the equality occurs only at the support points of the design.

The general equivalence theorem implies that for any arbitrary non-optimal design \( \xi \in \Xi_\Delta \) the minimum of \( \phi(x, \xi) < 0 \). Suppose, the design for the first \( j - 1 \) observations \( \xi(j - 1) \) is a non-optimal design such that \( M[\xi(j - 1)] \) is non-singular. Let

\[
\xi(j) = (1 - \alpha_{j-1})\xi(j - 1) + \alpha_{j-1}\xi^*,
\]

where \( \alpha_{j-1} \) is the step size of the search algorithm. For the design \( \xi(j) \) the expected Fisher information can be written as

\[
M\{\xi(j)\} = (1 - \alpha_{j-1})M\{\xi(j - 1)\} + \alpha_{j-1}M(\xi^*). \tag{10}
\]

For a given step size the objective of a sequential algorithm is to determine the point \( x \) that minimizes \( \Psi\{M[\xi(j)]\} \). Based on the general equivalence theorem the design point for \( j \)th observation in the first order sequential algorithm is

\[
\tau_j = \min_{x \in X} \phi(x, \xi(j)).
\]

Note, \( \tau_j \) is not necessarily an optimal design point. Instead, the sequential algorithm is run until a pattern develops that is suggestive of the optimal design. Then the algorithm is repeated based on the results of the initial search. It is possible to also search for the \( \alpha_{j-1} \) that minimizes \( \Psi\{M[\xi(j)]\} \) [Fedorov (1972)], but this is not desired in the context of an adaptive experiment. Selecting \( \alpha_{j-1} = \frac{1}{1 + (j - 1)^{-1}} = \frac{1}{j} \) mimics a sequential experiment where the \( j \)th observation is placed at the point that minimizes the sensitivity function. For this selection of \( \alpha_{j-1} \) (10) can be written as

\[
M[\xi(j)] = \alpha_{j-1}[(j - 1)M\{\xi(j - 1)\} + M(\xi^*)]. \tag{11}
\]

In this case \( \alpha_{j-1} \) is a scaler and will not influence the selection of the \( j \)th design point. Further, \( M[\xi(j)] \) is the sum of the initial expected Fisher information, weighted by the number of observations used to define the initial design, \( (j - 1) \), and the expected Fisher information from one additional observation with support \( x \).
3.2 Adaptive Design

Lane (2017) proposed two adaptive designs to optimize observed Fisher information in a more general setting where the observed and expected Fisher information can depend on the model parameters. As previously remarked this results in FODs that depend on the model parameters. An OAD will also have a local dependence. Lane (2017) addressed this dependence by proposing two designs, the local observed information adaptive design (LOAD) and the maximum likelihood estimated observed information adaptive design (MOAD). As shown in Section 2, for the linear model, observed and expected Fisher information do not have a local dependence on the model parameters. The design proposed in this section is not equivalent to the LOAD or MOAD procedure; however, it closely resembles the MOAD procedure.

3.2.1 Sequential Experiment

In Section 2 the model was introduced for non-sequential responses. In the sequential setting the sample is an ordered set of experimental runs. In the fully sequentially setting each run is comprised of a single observation. Specifically, the data for observations 1, \ldots, j – 1 are known prior to the design assignment of the jth observation, for j = 2, \ldots, n. The ancillary configuration statistic for the ith design point from the first j runs is denoted \( a_i(j) \) and the corresponding ancillary configuration matrix is denoted \( A(j) = [a_1(j), \ldots, a_d(j)] \). Note \( a_i(n) \) and \( A(n) \) represent these quantities from the full data; however, their dependence on \( n \) is omitted.

Now some quantities relevant to the sequential observed Fisher information are defined. Let \( q_{a,(j)} = i_{a,(j)}/\mu \) and \( Q_{A(j)} = \sum q_{a,(j)} \) and define

\[
\tau_{A(j)} = \begin{pmatrix} x_1 & x_2 & \cdots & x_d \\ \omega_{a_1(j)} & \omega_{a_2(j)} & \cdots & \omega_{a_d(j)} \end{pmatrix},
\]

where \( \omega_{a_i(j)} = q_{a,(j)}/Q_{A(j)} \). As stated in Barndorff-Nielsen and Sorensen (1994), \( i_{a,(j)} \geq 0 \) which implies \( q_{a,(j)} \geq 0 \); however, there could exist positive probability that it equals zero. This is unlikely to occur in practice and this problem can be solved by setting \( q_{a,(j)} \) to be equal to a small positive constant when this occurs. This ensures that \( \tau_{A(j)} \) is a continuous design, i.e. \( \tau_{A(j)} \in \Xi_\Delta \). As shown in Lane (2017), the observed Fisher information from the first j observations can be written

\[
J_{A(j)}(x) = \frac{1}{j} Q_{A(j)} M(\tau_{A(j)}).
\]

The above expression shows that, after run j, \( J_{A(j)} \) is proportional, up to a known constant, to the expected Fisher information evaluated at the design \( \tau_{A(j)} \). Each of the quantities introduced in this section are functions of only the ancillary configuration statistics and are themselves ancillary.

Lane (2017) considers a more general batched design setting where a run can consist of several observations. Here a fully sequential setting is considered to simplify the notation and the procedure. The method and theoretical results presented in this paper can be extended to the batched setting.

3.2.2 Observed Information Adaptive Design for the Linear Model

For the construction of the adaptive design it is assumed that the continuous FOD, with respect to \( \Psi \), denoted \( \xi_\Psi \), is known or can be computed. It is not required that the FOD be found using a first order sequential approach. It can be found using any consistent continuous optimal design approach. The optimal design points and their corresponding optimal allocations are denoted \( x^* = (x_1^*, \ldots, x_d^*)^T \) and \( w^* = (w_1^*, \ldots, w_d^*)^T \), respectively. The objective of the proposed OAD, described in this section, is to sequentially allocated observations to the optimal design points. Non-optimal design points will not be considered.

To incorporate observed Fisher information into the first order sequential algorithm let

\[
\tau(j) = (1 - \alpha_j)\tau_{A(j-1)} + \alpha_j \xi.
\]

Note, \( \tau(j) \in \Xi_\Delta \) since \( \tau_{A(j)} \in \Xi_\Delta \). Further, since only the optimal design points are considered the support points of \( \tau_{A(j)} \) are \( x^* \). For the adaptive algorithm let the step size \( \alpha_{j-1} = [1 + Q_{A(j-1)}]^{-1} \). This can be seen to be analogous the FOD, with \( \alpha_{j-1} = 1/j \), by evaluating the expected Fisher information at \( \tau(j) \)

\[
M\{\tau(j)\} = (1 - \alpha_{j-1})M\{\tau_{A(j-1)}\} + \alpha_{j-1}M(\xi)
= \alpha_{j-1}(j-1)J_{A(j)}(x^*) + M(\xi),
\]

(12)
The parallel between (11) and (12) is clear; \(a_{j-1} \) is a scaler, as it was in (11), and will not influence the selection of the \(j\)th design point; and \(M(\tau(j))\) is now the sum of the observed Fisher information, weighted by the number of past observations, \((j-1)\), and the expected Fisher information from one additional observation with support \(x\).

The search for the design point that minimizes the sensitivity function could be carried out using the same sequential algorithm that was used to find the FOD. However, this is unnecessary; the support points of the FOD are known; therefore, the search for the design point for the \(j\)th observation need only be over \(x^*\). In fact, if \(\omega_{a_j} > w_j^*\) then it indicates that the current observed allocation exceeds the optimal allocation and the search can be improved by excluding such points. Therefore, the adaptive procedure places the \(j\)th observation at

\[
\hat{x}_j = \min_{x \in x^{**}} \phi[x, \tau(j)],
\]

where \(x^{**}\) is the subset of optimal design points such that \(\omega_{a_i} < w_i^*\). The following observed information adaptive design (OAD) is proposed for linear regression models.

**Observed Information Adaptive Design (OAD)**

1. For observations \(j = 1, \ldots, kd\) initiate the design by placing \(k\) observations on each of the optimal design points \(x_1^*, \ldots, x_d^*\).
2. Compute \(t_{a_i}(j-1)\), \(q_{a_i}(j-1)\), \(Q_{A(j-1)}\), \(\omega_{a_i}(j-1)\), and \(\tau_{A(j-1)}\), where \((j-1)\) indicates that it the relevant quantity is computed using the ancillary statistics from the first \(j-1\) observations.
3. The design point for observation \(j = kd + 1, \ldots, n\) is

\[
\hat{x}_j = \min_{x \in x^{**}(j)} \phi[x, \tau(j)],
\]

where \(x^{**}(j)\) is the is the set of design points such that \(\omega_{a_i(j-1)} < w_i^*\). In the next section the properties of the OAD are examined.

A remark on selecting \(k\). For certain distributions if \(k\) is small, say 1, then the observed Fisher information may be volatile. The value \(k\) should be selected to reduce this volatility.

### 4 Inference and Optimality

This section contains the main results of the current work. First, it is shown that inference following an OAD can be conducted as if no adaptation had taken place. Second, the optimality of the OAD with respect to inference and conditional MSE is shown.

#### 4.1 Conditional Inference

There are two classic challenges associated with inference following adaptive designs. As stated in Ford et al. (1985), after the completion of an adaptive procedure the design is not ancillary. It is argued that analyses conditional on the design results in information loss [see Rosenberger and Lachin (2002) Chapter 11]. This is not the case for the OAD. The OAD uses only the ancillary information contained in \(A(j)\) to select the design for observation \(j+1\). Therefore, the design remains ancillary and there is no information lost by conditioning.

A second challenge in inference results from the induced dependence of the response of the \(j\)th observation on the responses from the preceding \(j-1\) observations. This dependence can affect the distribution of estimates following an adaptive procedure. Inference following adaptive designs has received significant attention in the literature [see Lane and Flournoy (2012), Lane, Yao and Flournoy (2014), Lane, Wang and Flournoy (2016), Lin, Flournoy and Rosenberger (2019)]. The theorem and corollary presented in this section demonstrate that the conditional distribution of the responses is not affected by the OAD procedure.

For the remainder of this work random variables associated with an OAD will be represented with a check. For example, \(\hat{y}\) and \(\hat{A}\) represent the responses and the ancillary configuration matrix for the OAD. Variables accented with a star correspond to a FOD, eg. \(y^*\) and \(A^*\). Variables without an accent are arbitrary and can be either an OAD or a FOD. Additionally, let \(\hat{X} = (\hat{x}_1, \ldots, \hat{x}_n)\) represent the sequence of observed designs points for an OAD and let \(X = (x_1, \ldots, x_n)\) represent the sequence for an arbitrary fixed design. Further,
the support points of the FOD, \( x^* \), are the support points for the OAD. The observed Fisher information of the OAD and FOD are denoted \( J_{\hat{A}} \) and \( J_{A^*} \), respectively.

**Theorem 4.1.** If \( X = \bar{X} \) and \( A = \hat{A} \) then

\[
\tilde{Y} | \hat{A} \overset{d}{=} Y | \{X, A\},
\]

where \( \overset{d}{=} \) represents equality in distribution.

This theorem states that conditional on the ancillary configuration matrix, \( \hat{A} \), the distribution of responses following an OAD is the same as a fixed design with \( X = \bar{X} \). The implication is that with respect to conditional inference the adaptive nature of the design can be ignored. The WLSE is a function of the responses following an OAD is the same as a fixed design with \( X = \bar{X} \).

**Corollary 4.1.** Under conditions A.1

\[
E[\hat{\beta} - \beta | A] = O_p(n^{-1})
\]

\[
n \text{Var}[\hat{\beta} | A] = J_{\hat{A}}^{-1} \{1 + O_p(n^{-1})\}
\]

\[
n(\hat{\beta} - \beta) J_{\hat{A}} (\hat{\beta} - \beta) | A = \chi^2_p + O_p(n^{-1})
\]

where \( A \) can be either \( \hat{A} \) or \( A^* \) and \( \chi^2_p \) denotes a \( \chi^2 \)-distribution with \( p \) degrees of freedom.

The \( O_p(n^{-1}) \) terms for the expectation and variance are given explicitly in the appendix. This theorem states that the properties of the observed Fisher information shown in the fixed setting of Section 2 extend to the OAD. Therefore, large sample inference after an OAD can be conducted using the observed Fisher information with the same order of accuracy, \( O_p(n^{-1}) \), present in fixed experiments.

### 4.2 Optimality

In this section inference and conditional MSE following a FOD and an OAD are compared. In Section 2 it was argued that inference for a fixed design should be based on the entries of \( J_A \). Further, Theorem 4.1 established that the OAD can be analyzed exactly as a fixed design without loss of information or concern regarding the conditional distribution. Therefore, the OAD should be analyzed using the entries of \( J_{\hat{A}} \).

Suppose confidence intervals for \( \beta \) are of interest. Corollary 4.1 demonstrated that approximate \((1 - \alpha)\%\) confidence intervals for \( \beta \) can be constructed for either a FOD or an OAD as

\[
\hat{\beta}_i \pm \{\chi^2_p(1-\alpha) | J_{\hat{A}}^{-1} |_{ii}\}^{1/2},
\]

where \( | J_{\hat{A}} |_{ii} \) is the element in the \( i \)th position on the diagonal and \( \chi^2_p(1-\alpha) \) is the \( 1 - \alpha \) quantile of a \( \chi^2 \)-distribution with one degree of freedom. The use of \( J_{\hat{A}} \) for inference implies that inference is optimized by minimizing \( \Psi(J_{\hat{A}}) \). All further references to optimality with respect to inference refers to minimizing \( \Psi(J_{\hat{A}}) \). It should be understood that statements regarding the optimality of the OAD refer to the optimality of the OAD relative to the corresponding FOD.

Statistical curvature describes the asymptotic difference between observed and expected Fisher information, Efron and Hinkley (1978). Efron (1975) defined statistical curvature as

\[
\gamma = \left( \frac{\nu_2 \nu_{20} - \nu_{11}}{\nu_2^3} \right)^{1/2},
\]
\[ \nu_{jk} = E[\dot{\bar{l}}(\eta)(\ddot{\bar{l}}(\eta) + E[\dot{\bar{l}}(\eta)]^2)^k]. \]

In the location family \( \mu = \nu_{20} \). Efron (1975) showed that for the location and scale family \( \gamma \) does not depend on the mean parameter, \( \eta \), and it is invariant under monotonic transformations. The independence of \( \gamma \) and the mean parameter ensures that statistical curvature is constant across the design region. Intuitively, it is expected that the difference in efficiency between the OAD and the FOD is a function of the statistical curvature. This intuition is be proven correct in Theorem 4.2.

A second form of curvature affecting the relative efficiency of the OAD and FOD is the design curvature - the curvature associated with the optimal design. Design curvature is a function of the Hessian matrix of \( \Psi \) evaluated at the optimal design, \( H^* = \nabla^2 \Psi(M) \bigg|_{M=M^*_\Phi} \).

The Hessian matrix of a convex optimality criterion evaluated at its minimum is positive definite and describes how small changes in the allocation weights away from the optimal allocations impact efficiency.

**Theorem 4.2.** Under conditions A.1

\[
    nE[\Psi(J_{A^*}) - \Psi(J_{\hat{A}})] \to \frac{\gamma^2}{2} \text{tr}(H^*V^*) \geq 0
\]

as \( n \to \infty \) with equality if and only if \( \gamma = 0 \) or \( d = 1 \), where \( V^* = \text{diag}[w^*_{d,1}] - w^*_{d,1}w^*_{d,1}^T \) and \( w^*_{d,1} = (w^*_1, \ldots, w^*_{d-1})^T \).

This theorem establishes that if a model has non-zero statistical curvature, \( \gamma > 0 \), and the optimal design \( \xi^*_\Phi \) has positive weight on more than one point, \( d > 1 \) then the OAD is optimal, compared to the corresponding FOD. This will lead to narrower confidence intervals or greater power depending on the nature of the optimality criterion. In the next section this will be more clearly illustrated by an increase in power in the context of \( c \)-optimality.

This theorem also provides insight into when the improvement in the relative efficiency of the OAD can be expected to be significant. As the statistical curvature, \( \gamma \), increases the relative benefit of the OAD will increase. Similarly, designs with large curvature will result in an increase in the relative efficiency of the OAD. As the total number of optimal design points, \( d \), increases the number of terms in the trace increases. However, as \( d \) increases the entries of \( V^* \) decrease. This might indicate that as \( d \) increases the relative benefit of the OAD may not be monotone.

Theorem 4.2 established that large sample inference for the OAD is superior to inference following a FOD. This result can be extended to the conditional mean square error (MSE). Note, MSE and observed Fisher information are inversely related. Therefore, the design that is optimal with respect to the conditional MSE is defined as the design that minimizes \( \Psi\{\text{MSE}[\hat{\beta}|A]^{-1}\} \).

**Theorem 4.3.** Under conditions A.1

\[
    nE\left[ \Psi\{\text{MSE}[\hat{\beta}|A^*]^{-1}\} - \Psi\{\text{MSE}[\hat{\beta}|\hat{A}]^{-1}\} \right] \to \frac{\gamma^2}{2} \text{tr}(H^*V^*) \geq 0
\]

as \( n \to \infty \) with equality if and only if \( \gamma = 0 \) or \( d = 1 \).

The interpretation of the above theorem is similar to that of Theorem 4.2. Specifically, the OAD is optimal with respect to the conditional MSE, at the limit.

4.2.1 Example: \( c \)-Optimality

In this section \( c \)-optimal design is used to illustrate the optimality results of Theorems 4.2 and 4.3. Recall, a \( c \)-optimal design minimizes the approximate variance of \( c^T \hat{\beta} \) with \( \Psi(M) = c^T M^{-1} c \). In this section the relationship with power and optimality is examined and it is shown that the OAD is more powerful, for large \( n \), than the FOD.
Corollary 4.1 established, under the null hypothesis, \( E[c^T \hat{\beta}] = C_0 \),
\[
[c^T J_A^{-1} c]^{-1} (c^T \hat{\beta} - C_0)^2 | A = \chi^2_1 + O_p(n^{-1}).
\]

In the presence of an alternative hypothesis, say \( E[c^T \tilde{\beta}] = C_1 \), the power of a \( \chi^2 \) test is calculated as
\[
\text{Power} = 1 - P \left\{ \chi^2_1(\lambda) \geq \chi^2_{1(1-\alpha)} \right\},
\]
where \( \chi^2_{1}(\lambda) \) is non-central \( \chi^2 \)-distribution with one degree of freedom and non-centrality parameter \( \lambda \). In the current context the non-centrality parameter is
\[
\lambda_A = n\delta^2 [c^T J^{-1}_A c]^{-1},
\]
where \( \delta = C_1 - C_0 \). Power is a non-decreasing function of the non-centrality parameter.

**Corollary 4.2.** Under conditions A.1
\[
E[\lambda_{A^*}] \rightarrow \delta^2 h_n [c^T M^{-1}_c c]^{-1} \left[ n - \gamma^2 \text{tr}(V^* D^*) \right]
\]
\[
E[\lambda_{A}] \rightarrow \delta^2 h_n [c^T M^{-1}_c c]^{-1} \left[ n \right],
\]
where \( \text{tr}(V^* D^*) > 0 \) if \( d > 0 \),
\[
h_n = 1 - \frac{1}{2n\mu} \left( \frac{\rho^2}{\mu} + \mu_4 \right).
\]
and \( D^* \) is a symmetric matrix with the \( i \)th entry equal to \( \{ [f(x_i) + f(x_d)]^T (F^T W^* F)^{-1} [f(x_i) + f(x_d)] \} \).

The interpretation of this corollary is as follows; suppose an experiment is conducted using a FOD with sample size \( n \). It would be expected that the power of this experiment would be equal to an OAD with a sample size of \( n - \gamma^2 \text{tr}(V^* D^*) \) fewer observations than the FOD to have the same power, at least approximately.

To further illustrate this point, consider an experiment with \( p \) treatments, indexed by \( 0, \ldots, p-1 \), where each observation can receive only 1 treatment. This can be written as a linear regression model with \( f(x) = (1, x_1, \ldots, x_{p-1})^T \), where \( x_1 = 1 \) for the treatment \( i \) and 0 otherwise, \( i = 1, \ldots, p-1 \). Suppose, treatment 0 represents a standard treatment and 1, \ldots, \( p-1 \) is a collection experimental treatments. If it is assumed that all experimental treatments have the same directional effect on the response then it could be of interest to test if the total of the experimental treatment effects are nonzero. The null hypothesis of this hypothetical experiment is \( H_0 : \sum_i \beta_i = 0 \). This null hypothesis is rejected if \( [c^T J^{-1}_A c]^{-1} (\sum_i \beta_i)^2 > \chi^2_{1(1-\alpha)} \).

The most common setting for this particular set up is for \( p = 2 \), where the null hypothesis reduces to \( H_0 : \beta_1 = 0 \).

A \( c \)-optimal design, with \( c = (0, 1_{p-1}^T)^T \), maximizes the power of this test. It can be verified by the general equivalence theorem that the \( c \)-optimal design allocates 1/2 of the observations to treatment 0 and divides the remaining 1/2 evenly among the \((p-1)\) experimental treatments. Corollary 4.2 implies, after some basic algebra, that
\[
E[\lambda_{A^*}] \rightarrow \frac{\delta^2 h_n}{2\mu} \left[ n - (p-1)\gamma^2 \right] \quad (13)
\]
\[
E[\lambda_{A}] \rightarrow \frac{\delta^2 h_n}{2\mu} \left[ n \right], \quad (14)
\]
where \( \delta = \sum_i \beta_i \). Therefore, a FOD with sample size \( n \) has approximately the same expected power as an OAD with a sample size of \( n - (p-1)\gamma^2 \).

Equations (13) and (14) present an interesting relationship between power and increases in the number of model parameters, in this case the number of treatments \( p \). For a FOD (13) indicates that as \( p - 1 \) increases the power decreases at a rate of \( p - 1 \). Conversely, as \( p \) increases no decrease in power is expected for the
OAD, at least for large $n$ and finite $p$. This may have interesting implications for large $p$ examples; however, this requires additional consideration since conditions A.1 explicitly exclude large $p$.

Two error distributions will be considered as illustrative examples. First, suppose that $\varepsilon$ has a Student $t$-distribution with parameter $v$. This model will be referred to as a Student $t$-regression (STR). A STR with $v = 1$ corresponds to the Cauchy distribution. The density of the residuals is

$$f(\varepsilon) = \left[\sqrt{v}B(1/2, v/2)\right]^{-1}(1 + \varepsilon^2/v)^{-(1+v)/2},$$

where $B(a, b)$ is the Beta function. The statistical curvature can be computed as

$$\gamma^2 = \frac{6[19 + 3v(6 + v)]}{v(v + 1)(v + 5)(v + 7)},$$

see Efron (1975). In this example $\gamma \to \infty$ as $v \to 0$ and is a decreasing as a function of $v$. For a STR with $v = 1/2$, 1 and 2 $\gamma^2 \approx 5.6, 2.5$ and 1.1, respectively.

A second example is a regression model based on Fisher’s problem of the Nile [Fisher (1974)]. In Fisher’s gamma hyperbola model two independent random variables $(s, t) = (z_1 e^\gamma, z_2 e^{-\gamma})$ are observed simultaneously, where $z_k \sim \text{Gamma}[v, 1]$, $k = 1, 2$. In the supplement it is shown that this can be written as a linear regression model a by setting $y = \log(s/t)/2$. It is also shown that the WLSE has closed form for this error distribution. In gamma hyperbola regression (GHR) $\gamma^2 = (2v)^{-1}$. In this example $\gamma \to \infty$ as $v \to 0$ as before.

A simulation study was conducted for the previously described $c$-optimal experimental design. Figure 1 presents the results: from 10,000 iterations, for $n c^T \text{MSE}[\hat{\beta}|c, c^T J_A^{-1}c$ and power (left to right) for the STR, with $v = 1$, and the GHR, with $v = 1/4$. The first row is a STR with $p = 2$; the second row is a GHR $p = 2$, the third row is a STR $p = 4$ and the fourth row is GHR $p = 4$. Sample sizes ranged from $n = 20$ to 120 and $\beta = 1_p$.

The first column of Figure 1 plots $n c^T \text{MSE}[\hat{\beta}|c$ for the OAD (solid line) and the FOD (dashed line). A design is optimal with respect to MSE if $c^T \text{MSE}[\hat{\beta}|c$ is minimized, the lower the better. For both error distributions, both values of $p$ and all $n$ the OAD resulted in a lower $c^T \text{MSE}[\hat{\beta}|c$ than the FOD in all but a single case ($p = 4, n = 30$). This confirms the result of Theorem 4.3 where the large sample optimality of the OAD with respect to conditional MSE was proven.

The second column of Figure 1 plots $c^T J_A^{-1}c$ again for the OAD (solid line) and FOD (dashed line). As defined, inference is optimized by minimizing $c^T J_A^{-1}c$, again the lower the better. For all cases considered the OAD resulted in a lower $c^T J_A^{-1}c$ than the FOD. This confirms the result of Theorem 4.2 where the large sample optimality, with respect to inference, of the OAD was proven.

It is well known that, under certain conditions, the variance of the estimates from a FOD are subject to the Cramer Rao Lower bound (CRLB), i.e.,

$$\text{Var}[\hat{\beta}] \geq M_c^{-1},$$

where $\geq$ is with respect to Loewner order. In columns 1 and 2 $c^T M_c^{-1}c$ is plotted as the dotted line. The $c$-optimality criterion preserves Loewner ordering, i.e., $n c^T \text{Var}[\hat{\beta}|c \geq c^T M_c^{-1}c$. Since $\text{Var}[\hat{\beta}] \geq \text{MSE}[\hat{\beta]}$ the dotted line is a lower bound for column 1.

The final column of Figure 1 plots the power, for the $\chi^2$-test, for the OAD (solid line) and FOD (dashed line). The greater the power the better. For this example the OAD was uniformly more powerful than the FOD, confirming Corollary 4.2. This sequence of figures provides additional information regarding the magnitude of the difference. For instance, when $p = 2$ the power of the OAD is greater than the FOD; however, the improvement is minimal. If $\gamma$ were increased it is expected that the improvement would also increase. When $p = 4$ the improvement in power is significant. For example, it is often desired to select the minimum sample size that attains a nominal power of 0.8, power is defined as the probability of rejecting the null hypothesis. For a STR with $p = 4$ an OAD with $n = 60$ attains this nominal power. For the same example, the FOD required an additional 10 observations to achieve the same nominal power. This represents a 16.7% increase in the sample size. This is close to $(p - 1)\gamma^2 = 7.5$ (12.5%), the increase predicted by the corollary. As stated, an increase in $\gamma$ will lead to greater discrepancies between the OAD and FOD.

The point of this illustration has been to show that the OAD can increases efficiency and power compared to a FOD. Further, the results of the corollary provide a straightforward method to compute the relative
benefits of the OAD. In the example provided all that was needed is to compute \((p - 1)\gamma^2\) in order to understand the approximate gain in power.

The \(c\)-optimality criterion is not unique with respect to optimality. The \(c\) criterion was used here because it has a straightforward interpretation. Similar optimality results can be obtained for other common optimality criteria. For example, the OAD will be optimal for the \(A\)-optimality criterion with respect to the average confidence interval length. The OAD will be optimal for the \(D\)-optimality criteria for the volume of the confidence ellipsoid. In the next section a simulation is conducted for \(A\)- and \(D\)- optimality.

5 Extended Simulation Study

In this section the previous simulation is extended to \(D\)- and \(A\)-optimal designs. Recall that \(\Psi(M)\) is equal to \(|M|^{-1/p}\) and \(\tr(M^{-1})\) for \(D\)- and \(A\)- optimality, respectively. A quadratic mean function with \(r\) covariates and design region \(X = [-1, 1]^r\) is considered. This corresponds to a linear model with \(f_x(x) = (1, x_1, x_1^2, \ldots, x_r, x_r^2)^T\). The support points of the \(D\)- and \(A\)-optimal designs are the \(3^r\) support points of a full factorial design with levels -1, 0 and 1. The \(D\)-optimal design allocates equal weight to each of the \(3^r\) design points. The optimal allocations for the \(A\)-optimal design do not have a closed form. The \(A\)-optimal
designs were found using the OptimalDesign package in R [Harman and Filov (2016)].

In addition to comparing the OAD and FOD with respect to \( \Psi\{\text{MSE}\hat{\beta}^{-1}\} \) and \( \Psi(J_\beta) \) the relative efficiency of the two methods is considered. The efficiency of the OAD, relative to the FOD, with respect to inference and conditional MSE, are defined as

\[
\text{REff}_{inf} = \frac{\Psi(J_{\beta A})}{\Psi(J_\beta)} \quad \text{and} \quad \text{REff}_{MSE} = \frac{\Psi(\text{MSE}\hat{\beta}_{A^{-1}})}{\Psi(\text{MSE}\hat{\beta}^{-1})},
\]

respectively, where \( \hat{\beta}_A \) is the WLSE given \( A \). A \( \text{REff} > 1 \) indicates the OAD is more efficient than the FOD for each measure. Similar to power, \( \text{REff} \) provides insight into the magnitude of the differences.

### 5.1 D Optimality

Figure 2 presents the results, from 10,000 iterations, for \( |\text{MSE}\hat{\beta}^{1/p}, J_\beta^{-1/p} \) and \( \text{REff} \) (left to right) for both the STR, with \( v = 1 \), and the GHR, with \( v = 1/4 \). The first row is a STR with \( r = 1 \); the second row is a GHR \( r = 1 \), the third row is a STR \( r = 2 \) and the fourth row is GHR \( r = 2 \). Sample sizes ranged from \( n = 30 \) to 120 and \( \beta = 1_p \). In columns 1 and 2 the benchmark, \( |M_{\beta}^1|^{-1/p} \), is plotted as a dotted line.

The first column of Figure 2 plots \( |\text{MSE}\hat{\beta}|^{1/p} \) for the OAD (solid line) and FOD (dashed line). As was the case for \( c \)-optimality the lower the \( |\text{MSE}\hat{\beta}|^{1/p} \) the better. For all cases considered the OAD resulted in a lower \( |\text{MSE}\hat{\beta}|^{1/p} \) than the FOD. This again confirms the result of Theorem 4.3.

The second column of Figure 2 plots \( |J_\beta|^{-1/p} \) for the OAD (solid line) and the FOD (dashed line). As previously remarked, inference is optimized by minimizing \( |J_\beta|^{-1/p} \), again the lower the better. For both error distributions, both values of \( r \) and all \( n \) the OAD resulted in a lower \( |J_\beta|^{-1/p} \) than the FOD. As expected from Theorem 4.2.

The final column of Figure 2 plots the \( \text{REff}_{inf} \) (dashed line) and the \( \text{REff}_{MSE} \) (solid line). Note, in this figure it is not of primary interest to compare the lines to each other. Instead figures in this column summarize the results of from the first two columns. A \( \text{REff} \) greater than 1 indicates the OAD is more efficient than the FOD for each measure. The OAD was uniformly more efficient than the FOD both with respect to inference and conditional MSE. For example, for a STR with \( r = 2 \) the \( \text{REff}_{inf} \) reaches a maximum of approximately 1.14 when \( n = 40 \), and while it slowly decreases from there it is still 1.06 when \( n = 120 \). For the same example the \( \text{REff}_{MSE} \) hovers around 1.08 for \( n = 30 \) to 120.

### 5.2 A Optimality

Figure 3 presents the results for the \( A \)-optimal design under the the same conditions presented in Figure 2. The results for \( A \)-optimality are similar to the results for \( D \)-optimal designs. In almost all cases considered, \( \text{tr}(\text{MSE}\hat{\beta}) \) and \( \text{tr}(J_\beta^{-1}) \) are less for the OAD than for the FOD. This translates to \( \text{REff}_{inf} \) and \( \text{REff}_{MSE} \) greater than one. The conclusion is that the OAD optimizes inference, conditional MSE and efficiency compared to the FOD.

The simulation study considered 3 difference optimality criteria, for each optimality criteria there were 2 different mean functions with 10 sample sizes within each for a total of 120 total different cases. In all cases \( \Psi(J_\beta) \) was less than \( \Psi(J_{\beta A}) \). In other words the OAD was uniformly more efficient than the FOD, with respect to inference. Similarly, in all but 3 cases \( \Psi(\text{MSE}\hat{\beta}_{A^{-1}}) \) was less than \( \Psi(\text{MSE}\hat{\beta}_{A^{-1}}) \). Two exceptions occurred for the \( A \)-optimal design in the STR quadratic model with \( r = 2 \) and \( n = 60, 100 \). The other exception was the OAD was for the STR \( c \)-optimal design with \( p = 4 \) and \( n = 30 \). The conclusion is that the OAD is more efficient than the FOD with respect to conditional MSE.

The WLSE was considered in this work due to its convenient form. It is important that the results of the simulations are not restricted to only this estimate. The entire simulation study was repeated using the MLE of \( \beta, \hat{\beta} \), in place of the WLSE in the supplemental materials. The figures for MLE corresponding to Figures 1, 2 and 3 lead to the same conclusions. This confirms that the OAD is relevant to the MLE as well as the WLSE.

### 6 Discussion

In this work experimental design for linear models was considered. The objective of the proposed adaptive design was to minimize the observed Fisher information subject to a convex optimality criterion, \( \Psi \). The
method was contrasted against the fixed optimal design (FOD), which minimizes \( \Psi \) evaluated at the expected Fisher information. The focus of the comparison was on optimizing inference and conditional MSE. Optimal inference was defined as minimizing \( \Psi \) evaluated at the observed Fisher information. Similarly, a design is optimal with respect to conditional MSE if it minimizes \( \Psi \) evaluated at the inverse of the conditional MSE. In this work it was shown that the proposed adaptive design, denoted OAD, was optimal, at the limit, with respect to both inference and conditional MSE, as compared to the FOD.

In addition to the theoretical results a simulation study was conducted to compare the OAD to the FOD. Three optimality criteria were considered, \( D \), \( A \) and \( c \) optimality. The study included different error distributions, mean functions and various sample sizes. It was found that the OAD was optimal with respect to conditional MSE, inference and power in nearly all cases considered. This suggests that the large sample benefits found in the theoretical results are also present in small to moderate sample sizes. The primary conclusion is that using observed Fisher information in an adaptive design can significantly improve the efficiency of the parameter estimates.

Focusing on a linear model with additive errors made the theoretical results tractable. A more general model was considered in Lane (2017), where it was shown, primarily with heuristics and a simulation study,
that observed information adaptive designs can reduce the MSE of the parameter estimates. Based on the agreement between the simulation study in Section 5 and that of Lane (2017) it might be conjectured that the theoretical results extend to more general models. Extending the theoretical findings in Section 4 is far from trivial.

The motivation for incorporating observed Fisher information into designs is based on the idea of conditional inference on relevant subsets [Cox (1958), Fisher (1961), Efron and Hinkley (1978), McCullagh (1992), Sundberg (2003) and Fraser (2004)]. Ghosh, Reid and Fraser (2010) suggest that there is an emerging consensus in favor of conditional inference on relevant subsets. The objective of maximizing observed Fisher information can be viewed as maximizing the information in the relevant subset, at least approximately. In contrast to the FOD, which can be viewed as maximizing the information with respect to the entire sample space. Form this perspective the results of this work can be viewed as additional evidence of in favor of relevant subsets. The fact that the relevant subset can be used to design an experiment and reduce MSE is itself an indication of importance of the relevant subsets.
A Technical Conditions

Denote the kth derivative of the log likelihood as

$$l^{(k)}(y - \eta) = \frac{\partial^k}{\partial \eta^k} \log f_0(y - \eta).$$

The conditions required for the majority of technical results in this paper are stated below.

**Condition A.1.**

1. **Conditions on the design.**
   (a) The total sample size, \( n \), is fixed.
   (b) \( \Psi(\cdot) \) is a positive-homogeneous convex function with degree \(-1\).
   (c) The continuous optimal design \( \xi^*_\eta \) exists and the number of optimal design points with positive weight, \( d \), is finite.

2. **Conditions on the distribution of responses**
   (a) \( \beta \in B \), where \( B \) is an open subset of \( \mathbb{R}^p \).
   (b) The residual vector, \( \varepsilon \), is a vector of i.i.d random variables satisfying the location family condition.
   (c) \( E[(\eta_i - \eta)^2] < \infty \)
   (d) For \( k = 1, \ldots, 4 \), \( l^{(k)}(y - \eta) \) exists and \( E[|l^{(k)}(y - \eta)|] < \infty \)
   (e) \( l_i > 0 \) and \( E[l_i^2] < \infty \).
   (f) If \( |\eta - \eta'| < \delta \) then \( |l^{(5)}(y - \eta) - l^{(5)}(y - \eta')| < K_\delta(y, \eta), \) where \( \lim_{\delta \to 0} E[K_\delta(y, \eta)] = 0. \)

The conditions are split into 2 parts; conditions on the design (A.1.1) and conditions on the model (A.1.2). The design conditions are mild. It is only required that the sample size is not data dependent and that the optimal design exists and has finite support. The conditions on the responses are nearly the same as those required in Lemma 1 and 2 of Efron and Hinkley (1978). Here the conditions are slightly stronger since the 4th derivative is assumed to exist with finite expectation and the 5th derivative is bounded. In Efron and Hinkley (1978) only the first 3 derivatives were required assumed to exist with finite expectation and the 4th derivative was bounded.

B Proof of Theorem 4.1

We begin by establishing that in the location family the distribution of the residuals are unaffected by the OAD procedure.

**Lemma B.1.** If \( \varepsilon \) is the vector of residuals from the OAD procedure then

$$f(\varepsilon) = f(\varepsilon),$$

where \( \varepsilon \) is the vector of residuals obtained from any arbitrary fixed design.

Proof of lemma. For the OAD procedure the data from the past observations is only used to determine the mean function of the present response. Therefore, conditional on the mean function the responses are independent of the past data. This independence property has been regularly noted in adaptive designs [see Ford et al. (1985)]. In the current setting this implies \( \tilde{y}_{11|\eta_1}, \ldots, \tilde{y}_{1n_1|\eta_1}, \ldots, \tilde{y}_{dn|\eta_d} \) are independent random variables. Therefore \( (\tilde{y}_{11|\eta_1}, \ldots, \tilde{y}_{1n_1|\eta_1}, \ldots, \tilde{y}_{dn|\eta_d}, \ldots, \tilde{y}_{dn_d|\eta_d}) = (\tilde{\varepsilon}_{11|\eta_1}, \ldots, \tilde{\varepsilon}_{1n_1|\eta_1}, \ldots, \tilde{\varepsilon}_{dn|\eta_d}, \ldots, \tilde{\varepsilon}_{dn_d|\eta_d}) \) are also independent random variables. However, \( \tilde{\varepsilon}_{ij} \) are ancillary and thus the distribution of \( \tilde{\varepsilon}_{ij|\eta_i} \) is the same as the distribution of \( \tilde{\varepsilon}_{ij} \) which implies \( \tilde{\varepsilon} \) is a sequence of i.i.d random variables with the same distribution as if no adaptation had taken place. □

Return to the proof of the theorem. After run \( j \) the number of observations with design point \( x_i^* \), denoted \( N_i(j) \), is random. Denote the vector of responses observed at the \( i \)th design point from the first \( j \) runs as \( \tilde{y}_{i}(j) = (\tilde{y}_{i1}, \ldots, \tilde{y}_{iN_{i}(j)})^T \). Further, define \( \tilde{\mu}(j) \) to be the mean of the responses corresponding to these observations. Fisher (1934) showed that the ancillary configuration statistic can be represented as \( \tilde{a}_i(j) = E[y_{1i} - \tilde{\mu}(j), \ldots, y_{in_{i}(j)} - \tilde{\mu}(j)] = [\tilde{\varepsilon}_{1i} - \bar{\varepsilon}(j), \ldots, \tilde{\varepsilon}_{in_{i}(j)} - \bar{\varepsilon}(j)]^T \). Let \( \sigma(\tilde{a}_i(j)) \) be the sub-sigma field generated by \( \tilde{a}_i(j) \). Written in this way it can be shown that \( \sigma(\tilde{a}_i(j)) = \tilde{\sigma}_i(j) \). In other words \( \tilde{\sigma}_i(j - 1) \) is \( \tilde{\sigma}_i(j) \) measurable. This clearly implies \( \tilde{\sigma}(j - 1) \) is \( \sigma\{\tilde{\sigma}(j)\} \) measurable. Now, since the
design point for the $j$th observation, $\tilde{x}_j$ is determined by the observed Fisher information which is a function of $\tilde{A}(j - 1)$ it is $\sigma\{\tilde{A}(j - 1)\}$ measurable. Further this implies that the entire design matrix for an OAD procedure $\tilde{X}$ is $\sigma\{\tilde{A}(n)\}$ measurable, i.e. $\tilde{X}|\tilde{A} = \tilde{X}$. This discussion yields
\[
f[\tilde{y}|\tilde{A}] = f[\tilde{y}|\tilde{X}, \tilde{A}].
\]
Note that $\tilde{y}_{ij}|\tilde{X} = \tilde{y}_{ij}|\eta_i$. From the proof of Lemma B.1, $\tilde{Y}|\tilde{X}$ is a sequence of i.i.d. random variables which implies the stated result.

C Proof of Corollary 4.1

We begin by stating Lemma 1 and 2 from Efron and Hinkley (1978). Let $a_i$ be the ancillary configuration statistic for any arbitrary fixed design.

**Lemma C.1.** [Lemma 1 and 2 Efron and Hinkley (1978)] Under conditions A.1 the following hold
\[
E[\tilde{\eta}_i - \eta_i|a_i] = -\frac{1}{2} l^{(3)}(a_i)^{-2} + o_p(n^{-1})
\]
\[
Var[\eta_i|a_i] = \{l^{(3)}_a(i) \{l^{(3)}(a_i)^{-3/2}\}^2 + \frac{1}{2} l^{(4)}_a(i) \{l^{(3)}(a_i)^{-2} + o_p(n^{-1})\}
\]
\[
i(a_i)(\tilde{\eta}_i - \eta_i)^2|\{A \rightarrow \chi^2_1 + O_p(n^{-1})\}
\]
for $i = 1, \ldots, d$.

From the above lemma we can see for the WLSE that
\[
E[\tilde{\beta}|A] = [F^T I_A F]^{-1} F^T I_A E[\tilde{\eta}|A]
\]
\[
= \beta - \frac{1}{2} [F^T I_A F]^{-1} F^T I_A^{-1} K_A + o_p(n^{-1}),
\]
and
\[
nVar[\tilde{\beta}|A] = n[F^T I_A F]^{-1} F^T I_A Var[\tilde{\eta}|A]
\]
\[
= n[F^T I_A F]^{-1} + n[F^T I_A F]^{-1} F^T I_A \left\{ I_A^{-1} K_A + \frac{1}{2} I_A^{-3} L_A \right\} I_A F [F^T I_A F]^{-1} + o_p(n^{-1})
\]
\[
= J_A^{-1} + \frac{1}{n} [J_A^{-1} F^T \left\{ I_A^{-2} K_A + \frac{1}{2} I_A^{-2} L_A \right\} F J_A^{-1} + o_p(n^{-1}),
\]
where $K_A = (l^{(3)}_a, \ldots, l^{(3)}_a)^T$ and $L_A = (l^{(4)}_a, \ldots, l^{(4)}_a)^T$. Note the term in the curly bracket is $O_p(1)$ Further, we can use the independence of responses and the result from Efron and Hinkley (1978) directly to obtain
\[
n(\tilde{\beta} - \beta)|A \rightarrow \chi^2_1 + O_p(n^{-1}).
\]
The above are obtained for an arbitrary fixed design from the results of Efron and Hinkley (1978). Recognizing that these are all all conditional results immediately implies they hold for the OAD by Theorem 4.1, which completes the proof.

D Proof of Theorem 4.2

In what follows random variables with a check, eg. $\tilde{a}_i$, denote those obtained from an OAD, random variables with a star, eg. $a^*_i$, denote those obtained from an OAD and random variables without a check or star, eg. $a_i$, denote arbitrary random variables that could have been obtained from either a OAD or FOD.

Let $\omega_A = (\omega_a, \ldots, \omega_{ad})^T$, where $\omega_{ad} = 1 - \sum_{i=1}^{d-1} \omega_a$, and note that from the positive homogeneity assumption of $\Psi$ that
\[
\Psi\{J_A\} = \Psi\{Q_A M(\tau_A)/n\} = \frac{n}{Q_A} \Psi\{M(\tau_A)\}.
\]
(15)
A Taylor expansion of $\Psi\{M(\tau_A)\}$ with respect to $\omega_{(d),A}$ around $w^*_{(d)}$, where the subscript $(d)$ indicates the $d$th entry of the vector is removed, yields
\[
\Psi\{M(\tau_A)\} = \Psi\{M(\xi^*)\} + (\omega_{(d),A} - w^*_{(d)})^T \nabla \Psi\{M(\xi^*)\} + \frac{1}{2}(\omega_{(d),A} - w^*_{(d)})^TH^*(\omega_{(d),A} - w^*_{(d)}) \\
+ o_p \left[ (\omega_{(d),A} - w^*_{(d)})^T(\omega_{(d),A} - w^*_{(d)}) \right].
\] (16)

The gradient of $\Psi$ must be found under the constraint $\omega_{a,k} = 1 - \sum_{i=1}^{d-1} \omega_{a,i}$. Under this constraint the total derivative treating $\omega_{a,i}$ as fixed is
\[
\frac{d}{d\omega_{a,i}} \Psi\{M(\tau_A)\} = \frac{\partial}{\partial \omega_{a,i}} \Psi\{M(\tau_A)\} + \frac{\partial}{\partial \omega_{a,d}} \Psi\{M(\tau_A)\} \frac{d\omega_{a,d}}{d\omega_{a,i}}
\]
\[
= -f^T(x^*_i) \frac{\partial \psi}{\partial M} f^T(x^*_i) - f^T(x^*_i) \frac{\partial \psi}{\partial M} f^T(x^*_i) \frac{d\omega_{a,d}}{d\omega_{a,i}}
\]
for $i = 1, \ldots, d-1$. From the general equivalence theorem it is known that $f^T(x^*_i) \frac{\partial \psi}{\partial M} f^T(x^*_i)|_{M=M^*_p} = \text{tr}(M \frac{\partial \psi}{\partial M})|_{M=M^*_p}$ for all optimal design points. Further, taking the derivative of the constraint yields $d\omega_{a,d} = -d\omega_{a,i}$. Therefore,
\[
\left[ \frac{d}{d\omega_{a,i}} \Psi\{M(\tau_A)\} \right]_{M(\tau_A)=M^*_p} = 0
\] (17)

which implies the gradient is a vector of zeros under the constraint.

Using (15), (16) and (17) we can write
\[
n [\Psi\{J_{A^*}\} - \Psi\{J_\bar{A}\}] = n \left[ \frac{n}{Q_{A^*}} \Psi\{M(\tau_{A^*})\} - \frac{n}{Q_{\bar{A}}} \Psi\{M(\tau_{\bar{A}})\} \right]
\]
\[
= n\Psi^* \left( \frac{n}{Q_{A^*}} - \frac{n}{Q_{\bar{A}}} \right)
\]
\[
+ \left( \frac{n}{Q_{A^*}} \right)^2 \frac{n}{2} (\omega_{(d),A^*} - w^*_{(d)})^T H^*(\omega_{(d),A^*} - w^*_{(d)})
\]
\[
- \left( \frac{n}{Q_{\bar{A}}} \right)^2 \frac{n}{2} (\omega_{(d),\bar{A}} - w^*_{(d)})^T H^*(\omega_{(d),\bar{A}} - w^*_{(d)})
\]
\[
+ o_p \left[ (\omega_{(d),A^*} - w^*_{(d)})^T(\omega_{(d),A^*} - w^*_{(d)}) \right]
\]
\[
+ o_p \left[ (\omega_{(d),\bar{A}} - w^*_{(d)})^T(\omega_{(d),\bar{A}} - w^*_{(d)}) \right].
\] (23)

The following sequence of lemmas will be used to describe the asymptotic behavior of terms (19) - (23). Many of these results can be readily shown for the FOD. However, in the OAD the ancillary statistics are not independent and thus more care is needed to show these results extend to this setting. The first lemma can be considered a corollary to Lane (2017) Theorem 1.

**Lemma D.1.** Under conditions A.1

(i) $\omega_{a,i} - w_i^* = O_p(n^{-1/2})$.

(ii) $\omega_{a,i} - w_i^* = O_p(n^{-1})$.

for $i = 1, \ldots, d$.

For part (i) the conditions are more restrictive than the conditions in Lane (2017) Theorem 1 and thus part (i) holds. For part (ii) all conditions are satisfied by Lemma 2 in Lane (2017) trivially except the
condition that \( \hat{x}_j \) occurs at a point where \( w'_{ij} > 0 \), where \( w'_{ij} \) is defined as

\[
w'_{ij} = w^*_i + Q_{A(j-1)}\{w^*_i - \omega_{a_i(j-1)}\} \\
= w^*_i(1 + Q_{A(j-1)}) - Q_{A(j-1)}\omega_{a_i(j-1)} \\
> Q_{A(j-1)}(w^*_i - \omega_{a_i(j-1)}).
\]

Therefore, if \( w^*_i > \omega_{a_i(j-1)} \) then \( w'_{ij} > 0 \). By definition, the OAD only searches for the \( j \)th optimal design point over the set of \( x^* \) such that \( w^*_i > \omega_{a_i(j-1)} \). Therefore, the conditions of Lane (2017) Theorem 1 are satisfied for the OAD and \((ii)\) holds. □

The next lemma describes the difference between the observed Fisher information from an OAD and FOD

**Lemma D.2.** Under conditions A.1

\[
E\left[\sum_{i=1}^{d} (\hat{i}_{a_i} - i_{a_i}^*)\right] = o(1).
\]

Proof. Denote the \( k \)th derivative of the log likelihood of the \( ij \)th observation as

\[
l^{(k)}(y_{ij} - \eta_i) = \frac{\partial^k}{\partial \eta^k} \log f_0(y_{ij} - \eta_i)
\]

and note that \( l^{(k)}(y_{ij} - \eta_i) = l^{(k)}(\varepsilon_{ij}) \). Now let

\[
l^{(k)}_{n_i}(\varepsilon_i) = \sum_{j=1}^{n_i} l^{(k)}(\varepsilon_{ij}) \quad \text{and} \quad l^{(k)}(\varepsilon) = \sum_{i=1}^{d} l^{(k)}_{n_i}(\varepsilon_i)
\]

denote the log likelihoods for the \( i \)th design point and the full data, respectively. Further, denote the \( k \)th derivative of the log likelihood for the \( ij \), design point evaluated the MLE as

\[
l^{(k)}_{a_i} = \left[\frac{\partial^k}{\partial \eta^k} g_{a_i}(\hat{\eta}_i - \eta_i)\right]_{\eta_i=\hat{\eta}_i},
\]

where \( \hat{i}_{a_i} = -l^{(2)}_{a_i} \). It is important to note that

\[
l^{(k)}_{n_i}(\varepsilon_i) = \frac{\partial^k}{\partial \eta^k} \log g_{a_i}(\hat{\eta}_i - \eta_i).
\]

The first five derivatives exist; therefore, a Taylor series expansion yields

\[
n^{-1}l^{(k)}_{n_i}(\varepsilon_i) = n^{-1}l^{(k)}_{a_i}(\varepsilon_i) + \hat{n}_i^{-1}(\hat{\eta}_i - \eta_i){l^{(k+1)}_{a_i}} + o_p(1) \tag{24}
\]

for \( k = 1, \ldots, 4 \). Re-arranging terms we can alternatively write (24), scaled by \( n^{-1} \), as

\[
n^{-1}l^{(k)}_{a_i} = n^{-1}l^{(k)}_{n_i}(\varepsilon_i) - \hat{n}_i^{-1}(\hat{\eta}_i - \eta_i){l^{(k+1)}_{a_i}} + o_p(1) \tag{25}
\]

In the above quantities it should be noted that \( n_i \) is not fixed for the OAD as it is for the FOD. Instead in an OAD the observed \( n_i \) is a random variable and is denoted \( N_i(n) \). The following lemma describes the limiting behavior of the random sample sizes of the OAD on each optimal support point.

**Lemma D.3.** Under conditions A.1

\[
n^{-1}E[N_i(n)] \rightarrow w^*_i
\]

in probability as \( n \rightarrow \infty \) for \( i = 1, \ldots, d \).
Proof. Recognizing that $i_{a_i} = -\hat{l}_i^{(2)}$ and $\hat{\eta}_i - \eta_i = O_p(n^{-1/2})$ then taking the expectation of (25) with $k = 2$ yields
\[
n^{-1}E[i_{a_i}] = -n^{-1}E[\hat{l}_i^{(2)}(\varepsilon_i)] + o(n^{-1/2})
\]
\[
= -n^{-1}E[\hat{l}_i^{(2)}(\varepsilon_i)|N_i(n)] + o(n^{-1/2})
\]
\[
= n^{-1}\mu E[N_i(n)] + o(n^{-1/2}).
\]
(26)

After re-arranging we get
\[
E[i_{a_i}/n] - \mu E[N_i(n)/n] = o(n^{-1/2}).
\]

Next consider the variance
\[
n^{-2}Var[i_{a_i}] = n^{-1}Var[\hat{l}_i^{(2)}(\varepsilon_i)] + o(1)
\]
\[
= n^{-2}E[Var[\hat{l}_i^{(2)}(\varepsilon_i)|N_i(n)]] + n^{-2}Var[E[\hat{l}_i^{(2)}(\varepsilon_i)|N_i(n)]] + o(1)
\]
\[
= Var[N_i(n)/n] + o(1)
\]
(27)

since $Var[\hat{l}_i^{(2)}(\varepsilon_i)|N_i(n)] < \infty$ by conditions A.1. Chebychev’s inequality ensures $N_i(n)/n$ will converge in probability to its expectation if $Var[N_i(n)/n] \to 0$ as $n \to \infty$. From (27) this is equivalent to $n^{-2}Var[i_{a_i}] \to 0$.

Consider the sum over all $i$. Since $n$ is fixed $\sum_i N_i(n) = n$ and the total for (25) with $k = 2$ is
\[
n^{-1} \sum_i \hat{l}_i^{(k)} = n^{-1} \sum_i \hat{l}_i^{(k)}(\varepsilon_i) + o_p(n^{-1/2})
\]
\[
= n^{-1}l_i^{(k)}(\varepsilon) + o_p(n^{-1/2}).
\]

Therefore, since the residuals are i.i.d,
\[
\sum_i E[i_{a_i}/n] - \mu = o(n^{-1/2}).
\]

Further, the variance of the sum using (26) is
\[
n^{-2}Var\left[\sum_i i_{a_i}\right] = n^{-1}Var[l_i^{(2)}(\varepsilon_{ij})] + o(1)
\]
\[
= o(1).
\]

since by assumption $E[(l_i^{(2)}(\varepsilon_{ij}))^2] < \infty$. Therefore the variance of the sum of $\sum_i i_{a_i}$, scaled by $n^{-2}$, goes to 0 and hence $n^{-2}Var[i_{a_i}] \to 0$ for $i = 1, \ldots, d$. This implies, by Chebychev’s inequality, that $n^{-1} \sum_i i_{a_i} \to \mu$ in probability as $n \to \infty$. This in addition to Lemma D.1 can be used to show that
\[
E[i_{a_i}/n] \to \mu w_i^*
\]
as $n \to \infty$. Therefore,
\[
n^{-1}E[N_i(n)] \to w_i^*
\]
as stated. □

The following lemma gives the asymptotic behavior of the higher order derivatives of log likelihood functions evaluated at the MLE.

**Lemma D.4.** Under conditions A.1
\[
n^{-1}l_i^{(k)} \to w_i^* \mu_k
\]
in probability as $n \to \infty$ for $k = 1, \ldots, 4$, where $\mu_k = E[l_i^{(k)}(\varepsilon_{ij})]$.  

21
Proof. Using steps similar to those used to show (26) and (27) it can be shown that

\[ n^{-1} \mathbb{E}[\hat{i}_{a_i}^{(k)}] = n^{-1} \mu_4 \mathbb{E}[N_i(n)] + o(n^{-1/2}) \]
\[ n^{-2} \text{Var}[\hat{i}_{a_i}^{(k)}] = \text{Var}[N_i(n)/n] + o(1). \]

The lemma follows from Lemma D.3 and Chebychev’s inequality. □

Returning to the proof of Lemma D.2. After re-arranging a three term Taylor expansion for \( i_{a_i} \) we get

\[ i_{a_i} = -\hat{I}_{N_i(n)}^{(2)}(\varepsilon_i) + (\hat{\eta}_i - \eta_i)^{l_i^{(3)}} + (\hat{\eta}_i - \eta_i)^2 \frac{1}{2} l_i^{(4)} + (\hat{\eta}_i - \eta_i)^3 \frac{1}{6} l_i^{(5)} + o_p(1). \]  

(28)

In Lemma C.1 it was shown that

\[ n\mathbb{E}[\hat{\eta}_i - \eta_i | a_i] = -\frac{n\eta_i^{l_i^{(3)}}}{\hat{a}_i} + o_p(1) \]
\[ n\mathbb{E}[(\hat{\eta}_i - \eta_i)^2 | a_i] = 1 + o_p(1). \]

Now from Lemma D.4 we can write the above as

\[ n\mathbb{E}[\hat{\eta}_i - \eta_i | a_i] = -\frac{\mu_3}{2\mu^2} + o_p(1) \]
\[ n\mathbb{E}[(\hat{\eta}_i - \eta_i)^2 | a_i] = \mu^{-1} + o_p(1). \]

From conditions A.1 it can be shown that \( l^{(5)}_{a_i} \) is bounded in probability. Therefore, the expectation of (28) can be written

\[ \mathbb{E}[i_{a_i}] = -\mathbb{E}[l_{N_i(n)}^{(2)}(\varepsilon_i)] - w_i^o \left( \frac{\mu_3^3}{2\mu^2} + \frac{\mu_4}{2\mu} \right) + o(1). \]

The above holds for both the OAD and FOD; therefore,

\[ \sum_{i=1}^{d} E \left[ i_{\hat{a}_i} - i_{a_i}^* \right] = E[l_n^{(2)}(\varepsilon^*) - l_n^{(2)}(\hat{\varepsilon})] + o(1). \]

Now by Lemma B.1 the residuals from the OAD and FOD have the same distribution and thus

\[ E[l_n^{(2)}(\varepsilon^*) + l_n^{(2)}(\hat{\varepsilon})] = 0 \]

Therefore,

\[ \sum_{i=1}^{d} E \left[ i_{\hat{a}_i} - i_{a_i}^* \right] = o(1) \]

as stated. □

The next lemma describes the asymptotic distribution of \( \omega_{A^*} \). Note, this result holds only for the FOD.

**Lemma D.5.** Under conditions A.1

\[ \sqrt{n} (\omega_{A^*} - w_{\gamma}^{(2)}) = N[0, \gamma^2 V^*]. \]

Proof. Efron and Hinkley (1978) prove, under the assumed conditions, that

\[ \sqrt{n} (q_{A^*}/n - w_i^*) \to N(0, w_i^* \gamma^2) \]

for \( i = 1, \ldots, d \) in distribution as \( n \to \infty \). The independence assumption implies that \( n^{-1/2} (q_{A^*}/n - w^*) \) is a vector of independent random variables with an approximate normal distribution with mean \( \mathbf{0} \) and variance \( \gamma^2 W_{(d)} \), where \( W_{(d)} = \text{diag}(w_{(d)}' \gamma^2) \).
Let \( q_{A^*} = (q_{a_1^*}, \ldots, q_{a_2^*})^T \) and \( g(q_{A^*}) = q_{(d)A^*}(1_d^T q_{A^*})^{-1} \) then
\[
\sqrt{n}(\omega_{(d),A^*} - w_{(d)}^*) = \sqrt{n}[q_{(d),A^*}(1_d^T q_{A^*})^{-1} - w_{(d)}^*] \\
= \sqrt{n}[g(q_{A^*}) - w_{(d)}^*]^T.
\]

Note \( g(w^*) = w_{(d)}^* \) and \( g(w^*) = I_d - Jw_{(d)}^* \), where \( Jw_{(d)}^* = (w_{1d}^* 1_{d-1}^T, \ldots, w_{(d-1)d}^* 1_{d-1}^T) \). The delta method implies the result of the lemma after some basic algebra. \( \square \)

Now we return to the proof of the theorem. Recall \( Q_A = \mu^{-1} \sum_i i_{a_i} \), thus Lemma D.2 ensures that the expectation of (19) is \( o(1) \). Lemma D.1 ensures the expectation of (21), (22) and (22) are \( o(1) \). Equation (20) is a quadratic form of a random variable \( \omega_{(d),A^*} \). We consider only convex functions \( \Psi \), as a consequence the Hessian matrix, \( H^* \) is positive definite. Thus Lemma D.5 ensures
\[
E \left[ \frac{n}{2}(\omega_{(d),A^*} - w_{(d)}^*)^T H^*(\omega_{(d),A^*} - w_{(d)}^*) \right] \rightarrow \frac{\gamma^2}{2} \text{tr}(H^*V^*)
\]
Combing the results for all terms we can write
\[
nE [\Psi(J_{A^*}) - \Psi(J_A)] \rightarrow \frac{\gamma^2}{2} \text{tr}(H^*V^*)
\]
as \( n \rightarrow \infty \) as stated. From Fang, Loparo and Feng (1994) the lower bound on the trace can be written,
\[
\frac{\gamma^2}{2} \text{tr}(H^*V^*) \geq \frac{\lambda_n(H^*) \text{tr}(V^*)}{2} = \frac{\lambda_n(H^*)}{2}(d - 1).
\]
where \( \lambda_n(H^*) \) is the smallest eigenvalue of \( H^* \). For the if direction, the above is clearly 0 if either \( \gamma = 0 \) or \( d = 1 \). For the only if direction, note that \( \omega_A = 1 \), is a constant. Therefore, the expansion in (16) reduces to
\[
\Psi(M(\tau_A)) = \Psi(M(\xi^*)).
\]
This above is true for both the OAD and FOD and thus (18) is 0 if \( d = 1 \). This concludes the proof of the if and only if statement.

**E Proof of Theorem 4.3**

From Corollary 4.1 it is straightforward to show that
\[
\text{MSE}[\hat{\beta}|A] = J_A^{-1} + \frac{1}{n} R_A + o_p(n^{-1}),
\]
where
\[
R_A = nJ_A^{-1}F^T \left\{ \frac{5}{4} I_A^{-2} K_A^2 + \frac{1}{2} I_A^{-1} L_A \right\} F J_A^{-1}.
\]
First, Lemma D.4 directly implies that \( I_A \rightarrow \mu W^* \), \( K_A \rightarrow \mu_3 W^* \) and \( L_A \rightarrow \mu_4 W^* \) in probability as \( n \rightarrow \infty \). Therefore,
\[
R_A \rightarrow R (w^*)
\]
in probability as \( n \rightarrow \infty \), where
\[
R (w^*) = \frac{1}{2 \mu} \left\{ 5 \mu_3^2 / 2 \mu + \mu_4 \right\} (F^T W^* F)^{-1} F^T F (F^T W^* F)^{-1}.
\]
Therefore,
\[
\psi(\text{MSE}[\hat{\beta}|A]) = \Psi \{ J_A \} + o_p(1).
\]
This holds for both the OAD and the FOD by Theorem 4.1. Therefore, the difference
\[
 n \left[ \Psi\{\text{MSE}[\hat{\beta}|A^*]^{-1}\} - \Psi\{\text{MSE}[\hat{\beta}|\tilde{A}]\} \right] = n \left[ \Psi\{J_{A^*}\} - \Psi\{J_{\tilde{A}}\} \right] + o_p(1).
\]
Taking the expectation of both sides yields, by Theorem 4.2,
\[
nE \left[ \Psi\{\text{MSE}[\hat{\beta}|A^*]^{-1}\} - \Psi\{\text{MSE}[\hat{\beta}|\tilde{A}]\} \right] = nE \left[ \Psi\{J_{A^*}\} - \Psi\{J_{\tilde{A}}\} \right] + o(1)
= \frac{\lambda^2}{2} \text{tr}(H^*V^*) + o(1)
\]
as stated.

**F Proof of Corollary 4.2**
Let \( A \) be the ancillary configuration matrix from a FOD, then the non-centrality parameter is
\[
\lambda_A = n\delta^2(c^T[J_A]^{-1}c)^{-1} = n\delta^2 \frac{Q_A}{n} \{c^T[M(\tau A)]^{-1}c\^{-1}.
\]
Employing a Taylor expansion yields similar to that used in the proof of Theorem 4.2 we get
\[
\{c^T[M(\tau A)]^{-1}c\^{-1} = \{c^T M^*_c^{-1}c\^{-1} + (\omega(d)_A - w^*_d)^T \nabla \{[c^T M^*_c^{-1}c]\^{-1}\}
= \frac{1}{2}(\omega(d)_A - w^*_d)^T \nabla^2 \{[c^T M^*_c^{-1}c]\^{-1}\}(\omega(d)_A - w^*_d) + o_p(n^{-1}).
\]
Let \( \Omega_A = \text{diag}(\omega_A^*) \) then \( M(\tau A) = \mu(F^T\Omega_A F)^{-1}. \) The first partial derivatives are
\[
\frac{\partial}{\partial \omega_{a_i}} \{c^T[M(\tau A)]^{-1}c\}^{-1} = \mu \frac{\partial}{\partial \omega_{a_i}} \{c^T(F^T\Omega_A F)^{-1}c\}^{-1}
= \mu \{c^T(F^T\Omega_A F)^{-1}c\}^{-2} \left\{[f^T(x_a^*)(F^T\Omega_A F)^{-1}c]^2 - [f^T(x_a^*)]^2(F^T\Omega_A F)^{-1}c]^2 \right\}.
\]
The second partial derivative are
\[
\frac{\partial^2}{\partial \omega_{a_i} \partial \omega_{a_i}} \{c^T[M(\tau A)]^{-1}c\}^{-1}
= \mu \frac{\partial^2}{\partial \omega_{a_i} \partial \omega_{a_i}} \{c^T(F^T\Omega_A F)^{-1}c\}^{-1}
= 2\{c^T(F^T\Omega_A F)^{-1}c\}^{-3} \left\{[f^T(x_a^*)(F^T\Omega_A F)^{-1}c]^2 - [f^T(x_a^*)]^2(F^T\Omega_A F)^{-1}c]^2 \right\}^2
+ \left\{c^T[M(\tau A)]^{-1}c\}^{-2} \left\{2f^T(x_a^*)(F^T\Omega_A F)^{-1}c [f^T(x_a^*)]^2(F^T\Omega_A F)^{-1}f(x_a^*)
- f^T(x_a^*)(F^T\Omega_A F)^{-1}c f^T(x_a^*)(F^T\Omega_A F)^{-1}f(x_a^*)
- 2f^T(x_a^*)(F^T\Omega_A F)^{-1}c [f^T(x_a^*)][f^T(x_a^*)]^2f(x_a^*)
- f^T(x_a^*)(F^T\Omega_A F)^{-1}c f^T(x_a^*)(F^T\Omega_A F)^{-1}f(x_a^*) \right\}.
\]
Note from the general equivalence theorem it is known that \( (c^T(F^T\Omega_A F)^{-1}c)^{1/2} = f^T(x_a^*)(F^T\Omega_A F)^{-1}c \), for \( i = 1, \ldots, d - 1 \) and \( (c^T(F^T\Omega_A F)^{-1}c)^{1/2} = f^T(x_a^*)(F^T\Omega_A F)^{-1}c. \) Therefore, after some algebra it can be shown that
\[
\frac{\partial}{\partial \omega_{a_i}} \{c^T[M(\tau A)]^{-1}c\}^{-1} \bigg|_{M(\tau A)} = 0
\]
\[
\frac{\partial^2}{\partial \omega_{a_i} \partial \omega_{a_i}} \{c^T[M(\tau A)]^{-1}c\}^{-1} \bigg|_{M(\tau A)} = -2\{c^T M^*_c^{-1}c\}^{-1} \left\{[f(x_a^*)]^2 + f(x_a^*)]^T(F^T\Omega_A F)^{-1}[f(x_a^*) + f(x_a^*)] \right\}.
\]
Therefore we can rewrite (29) as

\[
\lambda^*_A = n\sigma^2 \frac{Q_A}{n} \left( c^T M_c^{-1} c \right)^{-1} - \frac{Q_A}{n} \left( c^T M_c^{-1} c \right)^{-1} (\omega_{(d),A} - \mu_{(d)})^T D^* (\omega_{(d),A} - \mu_{(d)}) + o_p(n^{-1}),
\]

where \( D^* \) is a positive definite symmetric matrix with the \( ik \)th entry equal to \([f(x_i^*) + f(x_d^*)] (F^T W^* F)^{-1} [f(x_i^*) + f(x_d^*)] \). From equation (28)

\[
n^{-1} E[Q_A] = n^{-1} \mu^{-1} E[-l_n^{(2)}(\varepsilon_i) - \left( \frac{\mu_3^2}{2\mu^2} - \frac{\mu_4}{2\mu} \right)] + o(1)
\]

(30)

where

\[
h_n = 1 - \frac{1}{2n\mu} \left( \frac{\mu_3^2}{\mu} + \mu_4 \right).
\]

Now we can compute the expectations of the non-centrality parameters. First for the FOD, using (30) and Lemma D.5 we get

\[
E[\lambda^*_A] = n\sigma^2 \left( c^T M_c^{-1} c \right)^{-1} \left\{ h_n - \frac{h_n}{n} \gamma^2 \text{tr}(V^* D^*) \right\} + o(1)
\]

\[
= \delta^2 h_n \left( c^T M_c^{-1} c \right)^{-1} \left\{ n - \gamma^2 \text{tr}(V^* D^*) \right\} + o(1),
\]

Using (30) and Lemma D.1 we can find the expected value for the OAD

\[
E[\lambda^*_\tilde{A}] = n\sigma^2 h_n \left( c^T M_c^{-1} c \right)^{-1} + o(1)
\]

which completes the proof.

References
Atkinson, A., Fedorov, V., Herzberg, A., Zhang, R., 2014. Elemental information matrices and optimal experimental designs for generalized regression models. Journal of Statistical Planning and Inference 144, 81–91.
Atkinson, A.C., Donev, A.N., Tobias, R., 2007. Optimum Experimental Designs with SAS. Oxford Statistical Science Series, 34, Oxford University Press.
Barndorff-Nielsen, O.E., 1980. Conditionality resolutions. Biometrika 67, 293–310.
Barndorff-Nielsen, O.E., Sorensen, M., 1994. A review of some aspects of asymptotic likelihood theory for stochastic processes. International Statistical Review 62, 133–165.
Box, G., Hunter, W., 1965. Sequential design of experiments for nonlinear models, in: Korth, J.J. (Ed.), Proceedings of the Scientific Computing Symposium: Statistics, White Plains: IBM, pp. 113–137.
Cox, D.R., 1958. Some problems with statistical inference. Annals of Mathematical Statistics 29, 357–372.
Cox, D.R., 1980. Local ancillarity. Biometrika 67, 279–286.
Dragalin, V., Fedorov, V., 2005. Adaptive designs for dose–finding based on efficacy–toxicity response. Journal of Statistical Planning and Inference 136, 1800–1823.
Dragalin, V., Fedorov, V., Wu, Y., 2007. Adaptive designs for selecting drug combinations based on efficacy–toxicity response. Journal of Statistical Planning and Inference 136, 1800–1823.
Efron, B., 1975. Defining the curvature of a statistical problem (with applications to second order efficiency). Annals of Statistics 3, 1189–1242.
Efron, B., 1978. Assessing the accuracy of the maximum likelihood estimate: observed versus expected Fisher information (with discussion). Biometrika 65, 457–483.
Fang, Y., Loparo, K.A., Feng, X., 1994. Inequalities for the trace of a matrix product. IEEE Transactions on Automatic Control 39, 2489–2490.
Farrell, R.H., Kiefer, J., Walbran, A., 1968. Optimum multivariate designs, in: Proceedings of the 5th Berkeley Symposium, University of California Press, Berkeley, CA. pp. 113 – 138.
Fedorov, V., 1972. Theory of Optimal Experiments. Academic Press: New York.
Fedorov, V., 2010. Optimal experimental design. Wiley Interdisciplinary Reviews: Computational Statistics 2, 581–589.
Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38.
Fisher, R., 1934. Two new properties of mathematical likelihood. Proceedings of Royal Society 144, 285–307.
Fisher, R., 1974. Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh, UK.
Fisher, R.A., 1961. Sampling the reference set. Sankhya: The Indian Journal of Statistics, Series A 23, 3–8.
Ford, I., Titterington, D.M., Wu, C.F.J., 1985. Inference and sequential design. Biometrika 72, 545–551.
Fraser, D.A.S., 2004. Ancillaries and conditional inference. Statistical Science 19, 333–369.
Ghosh, M., Reid, N., Fraser, D., 2010. Ancillary statistics: a review. Statistica Sinica 20, 1309–1332.
Guest, P., 1958. The spacing of observations in polynomial regression. Annals of the Mathematical Statistics 29, 294–299.
Harman, R., Filov, L., 2016. OptimalDesign: Algorithms for D-, A-, and IV-Optimal Designs. URL: https://CRAN.R-project.org/package=OptimalDesign r package version 0.2.
Harman, R., Filov, L., Richtrik, P., 2019. A randomized exchange algorithm for computing optimal approximate designs of experiments. Journal of the American Statistical Association doi:10.1080/01621459.2018.1546588.
Kiefer, J., 1959. Optimal experimental designs (with discussion). Journal of the Royal Statistical Society (Series B) 21, 272–319.
Kiefer, J., 1961. Optimum designs in regression problems, II. The Annals of Mathematical Statistics 32, 298–325.
Kiefer, J., Wolfowitz, J., 1959. Optimum designs in regression problems. The Annals of Mathematical Statistics 30, 271–294.
Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canadian Journal of Mathematics 14, 363–366.
Lane, A., 2017. Adaptive Designs for Optimal Observed Fisher Information. arXiv e-prints arXiv:1712.08499.
Lane, A., Flournoy, N., 2012. Two-stage adaptive optimal design with fixed first-stage sample size. Journal of Probability and Statistics DOI: 10.1155/2012/436239.
Lane, A., Wang, H., Flournoy, N., 2016. MoDa 11 – Advances in Model-Oriented Design and Analysis. Springer (eds. Kunert, J. Muller, C. H. Atkinson A. C.), chapter Conditional Inference in Two-Stage adaptive experiments via the bootstrap. pp. 173–182.
Lane, A., Yao, P., Flournoy, N., 2014. Information in a two-stage adaptive optimal design. Journal of Statistical Planning and Inference 144, 173–187.
Lin, Z., Flournoy, N., Rosenberger, W.F., 2019. Random Norming Aids Analysis of Non-linear Regression Models with Sequential Informative Dose Selection. arXiv e-prints arXiv:1905.09722.
Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society (Series B) 44, 226–233.
Lystig, T.C., Hughes, J.P., 2002. Exact computation of the observed information matrix for hidden markov models. Journal of Computational and Graphical Statistics 11, 678–689.
May, C., Flournoy, N., 2009. Asymptotics in response-adaptive designs generated by a two-color, randomly reinforced urn. Annals of Statistics 37 (2), 1058–1078.
McCullagh, P., 1984. Local sufficiency. Biometrika 71, 233–244.
McCullagh, P., 1992. Conditional inference in Cauchy models. Biometrika 79, 247–259.
Mitchell, T., 1974. An algorithm for the construction of “D-optimum” experimental designs. Technometrics 16, 203–210.
Murphy, S.A., Vaart, A.V.D., 1999. Observed information in semi-parametric models. Bernoulli 5, 381–412.
Pukelsheim, F., 2006. Optimal design of experiments. SIAM, Philadelphia, PA.
Pukelsheim, F., Rieder, S., 1992. Efficient rounding of approximate designs. Biometrika 79, 763–770.
Reid, N., 2003. Asymptotics and the theory of inference. The Annals of Statistics 6, 1695–1731.
Reid, N., Fraser, D.A.S., 2010. Mean loglikelihood and higher-order approximations. Biometrika 97, 159–170.
Rosenberger, W.F., Lachin, J.M., 2002. Randomization in Clinical Trials. Wiley, New York.
Skovgaard, I., 1986. Successful improvement of the order of ancillarity. Biometrika 73, 516–519.
Sundberg, R., 2003. Conditional statistical inference and quantification of relevance. Journal of the Royal Statistical Society (Series B) 65, 299–315.

Wynn, H.P., 1970. The sequential generation of D-optimum experimental designs. The Annals of Mathematical Statistics 41, 1655–1664.

Yang, M., Biedermann, S., Tang, E., 2013. On optimal designs for nonlinear models: A general and efficient algorithm. Journal of the American Statistical Association 108, 1411–1420.

Yu, Y., 2011. Monotonic convergence of a general algorithm for computing optimal designs. The Annals of Statistics 38, 1593–1606.
Supplemental Materials for Optimality of Observed Information Adaptive Designs in Linear Models

1 Gamma Hyperbola Linear Regression

In this section we show the gamma hyperbola model can be written as a linear regression model. The model is derived for a fixed design. We begin by stating the distribution of \((s, t) = (z_1 e^\eta, z_2 e^{-\eta})\), which is the product of two independent gamma random variables, i.e.

\[
f_\eta(s_{ij}, t_{ij}) = e^{-(s_{ij} e^{\eta_i} + t_{ij} e^{-\eta_i})} \frac{1}{\Gamma[\beta]}(s_{ij} t_{ij})^{\beta-1}
\]

where \(\eta_{ij} = \beta^T f(x_i)\). Let \(a_{ij} = \sqrt{s_{ij} t_{ij}}\) and recall \(\frac{1}{2} \log[s_{ij}/t_{ji}]\) then using a change of variable we get

\[
f_\eta(y_{ij}, a_{ij}) = \frac{1}{\Gamma[\beta]} a_{ij}^{\beta-1} e^{-a_{ij}(e^{(y_{ij} - \eta_i) + \eta_i - \eta_i})}
\]

\[
= \frac{1}{\Gamma[\beta]} a_{ij}^{\beta-1} e^{2a_{ij} \cosh(y_{ij} - \eta_i)}.
\]

Using the above expression we can now see that the location family condition that \(f_0(\varepsilon) = f_0(y - \eta) = f_0(y)\) is satisfied.

1.1 Closed Form Expression for the WLSE

In this section the closed form of the WLSE is given. Let \(S_i = \sum_{j=1}^{n_i} s_{ij}\) and \(T_i = \sum_{j=1}^{n_i} t_{ij}\). We begin by stating the distribution of \((S_i, T_i)\)

\[
f_\eta(S_i, T_i|\xi, \theta) = e^{-\sum_{i=1}^{d} (S_i e^{\eta_i} + T_i e^{-\eta_i})} \prod_{i=1}^{d} \frac{1}{\Gamma[n_i \beta]}(S_i T_i)^{n_i \beta-1}
\]

Now let \(a_i = \sqrt{S_i T_i}\) and \(\hat{\eta}_i = \frac{1}{2} \log[S_i/T_i]\) then again by a change of variable we get

\[
f_\eta(\hat{\eta}_i, a_i) = \frac{1}{\Gamma[n_i \beta]} (a_i)^{n_i \beta-1} e^{-a_i(e^{(\hat{\eta}_i - \eta_i) + \hat{\eta}_i - \eta_i})}.
\]

The likelihood equations are

\[
\frac{\partial \log f(\eta, r)}{\partial \eta} = 2r_i \sinh(\hat{\eta}_i - \eta_i) e^{-\sum_{i=1}^{d} 2r_i \cosh(\hat{\eta}_i - \eta_i)} = 0.
\]

Setting \(\hat{\eta}_i = \eta_i\) satisfies these equations and thus \(\hat{\eta}_i\) is the MLE of \(\eta_i\). Therefore, the WLSE is

\[
\hat{\beta} = [F^T I_A F]^{-1} F^T I_A \hat{\eta}
\]

which has a closed form since \(\hat{\eta}\) has a closed form.

Additionally, the conditional distribution can be written as

\[
f(\hat{\eta}|r) = \frac{e^{-2 \sum_{i=1}^{d} r_i \cosh(\hat{\eta}_i - \eta_i)}}{\prod_{i=1}^{d} \int_{-\infty}^{\infty} e^{-2r_i \cosh(u_i)} du_i}.
\]

Finally, the observed Fisher information can be written as

\[
I(D_n) = -\frac{\partial^2 \log f(\hat{\eta}|r)}{\partial \theta^2} \bigg|_{\theta = \hat{\theta}} = 2 \sum_{i=1}^{d} r_i g(x_i) g(x_i)^T
\]

since \(\sinh(0) = 0\) and \(\cosh(0) = 1\).
2 Simulation study for the MLE

This section reproduces the results for the simulation from the main text for the MLE of $\beta$, denoted $\hat{\beta}$, in place of the WLSE. For this series of simulations, the MLE similar to the WLSE. In nearly every case considered the OAD was better with respect to inference, conditional MSE, and Power.

Figure 1: Simulation results, from 10,000 iterations, for $c^T \text{MSE}[\beta]c$, $c^T J_{\lambda}^{-1}c$ and power (left to right) are presented for both a STR, with $v = 1$, and a GHR, with $v = 1/4$. In columns 1 and 2 the $c^T M_{c}^{-1}c$ is plotted as the dotted line. The first row is a STR with $p = 2$; the second row is a GHR $p = 2$, the third row is a STR $p = 4$ and the fourth row is a GHR $p = 4$. Sample sizes ranged from $n = 20$ to 120 and $\beta = 1_5$. 

2.5
Figure 2: Simulation results, from 10,000 iterations, for $|J_{A}|^{-1/p}$, $|MSE[\beta]|^{1/p}$ and power (left to right) are presented for a STR with $v = 1$ and a GHR with $v = 1/4$ models. In columns 1 and 2 the $|M_{D}^{j}|^{-1/p}$ is plotted as a dotted line. The first row is a STR with $r = 1$; second row is a GHR $r = 1$, third row is a STR $r = 2$ and the fourth row is a GHR $r = 2$. Sample sizes ranged from $n = 30$ to 120 and $\beta = 1_{p}$.
Figure 3: Simulation results, from 10,000 iterations, for $\text{Tr}\{J_A^{-1}\}$, $\text{Tr}\{MSE[\beta]^{-1}\}$ and power (left to right) are presented for a STR with $v = 1$ and a GHR with $v = 1/4$ models. In columns 1 and 2 the $\text{Tr}[M_A^{-1}]$ is plotted as a dotted line. The first row is a STR with $r = 1$; second row is a GHR $r = 1$, third row is a STR $r = 2$ and the fourth row is a GHR $r = 2$. Sample sizes ranged from $n = 30$ to 120 and $\beta = 1_\rho$. 