The Canadian Journal of Statistics
Vol. 00, No. 00, 2020, Pages 000–000
La revue canadienne de statistique

An efficient algorithm for Elastic I-optimal design of generalized linear models

Yiou LI[1] and Xinwei DENG[2*]

[1]*Department of Mathematical Sciences, DePaul University, Chicago, IL, U.S.A.*
[2]*Department of Statistics, Virginia Tech, Blacksburg, VA, U.S.A.*

Key words and phrases: Fast convergence, multiplicative algorithm, prediction-oriented, sequential algorithm.

MSC 2010: Primary 62K05; secondary 65K10, 62L99.

Abstract: The generalized linear models (GLMs) are widely used in statistical analysis and the related design issues are undoubtedly challenging. The state-of-the-art works mostly apply to design criteria on the estimates of regression coefficients. The prediction accuracy is usually critical in modern decision-making and artificial intelligence applications. It is of importance to study optimal designs from the prediction aspects for GLMs. In this work, we consider Elastic I-optimality as a prediction-oriented design criterion for GLMs, and develop an efficient algorithm for such EI-optimal designs. By investigating theoretical properties for the optimal weights of any set of design points and extending the general equivalence theorem to the EI-optimality for GLMs, the proposed efficient algorithm adequately combines the Fedorov–Wynn algorithm and the multiplicative algorithm. It achieves great computational efficiency with guaranteed convergence. Numerical examples are conducted to evaluate the feasibility and computational efficiency of the proposed algorithm. *The Canadian Journal of Statistics* 00: 000–000; 2020 © 2020 Statistical Society of Canada

Résumé: Largement utilisés pour l'analyse statistique, les modèles linéaires généralisés (GLM) présentent de grands défis pour la planification d'expériences. Les travaux de recherche de pointe en planification d'expériences portent habituellement sur des critères pour l'estimation des coefficients de régression. Pour les applications modernes de prise de décision et d'intelligence artificielle, la précision des prévisions est particulièrement critique, d'où l'importance d'étudier les plans d'expériences optimaux pour les GLM selon des critères de prévision. Les auteurs considèrent la I-optimalité élastique en tant que critère d'optimalité orienté sur les prévisions pour les GLM. Ils développent un algorithme efficace pour de tels plans EI-optimaux en combinant adéquatement l'algorithme de Fedorov-Wynn et l'algorithme multiplicatif. Cette harmonisation est rendue possible grâce à une investigation des propriétés théoriques des poids optimaux pour tout ensemble de points du plan d'expérience, et à une extension du théorème d'EI-optimalité pour les GLM. Les auteurs montrent que l'algorithme offre une grande efficacité computationnelle ainsi que des garanties de convergence. Ils présentent finalement des exemples numériques pour évaluer la faisabilité et l'efficacité de l'algorithme. *La revue canadienne de statistique* 00: 000–000; 2020 © 2020 Société statistique du Canada

1. INTRODUCTION

The generalized linear model (GLM) is a flexible generalization of linear models by relating the response to the predictors through a link function (Nelder & Wedderburn, 1972). The GLMs are

* *Author to whom correspondence may be addressed.*
E-mail: xdeng@vt.edu

widely used in many statistical analyses with different applications including business analytics, image analysis, bioinformatics and others. From a data collection perspective, it is of great importance to understand the optimal designs for the GLM, especially from a prediction perspective.

Suppose that an experiment has d explanatory variables $\boldsymbol{x} = [x_1, \dots, x_d]$, and let Ω_j be a measurable set of all possible levels of the jth explanatory variable. Common examples of Ω_j are $[-1, 1]$ and $\mathbb{R}$. The experimental region, Ω, is some measurable subset of $\Omega_1 \times \cdots \times \Omega_d$. In a GLM, the response variable $Y(\boldsymbol{x})$ is assumed to follow a particular distribution in the exponential family, such as the normal, binomial, Poisson and gamma distributions. A link function provides the relationship between the linear predictor, $\eta = \boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{x})$, and the mean of the response $Y(\boldsymbol{x})$,

$$\mu(\boldsymbol{x}) = \mathbb{E}[Y(\boldsymbol{x})] = h^{-1}\left(\boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{x})\right), \tag{1}$$

where $\boldsymbol{g} = [g_1, \dots, g_l]^T$ are the bases and $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_l]^T$ are the corresponding regression coefficients. Here $h : \mathbb{R} \to \mathbb{R}$ is the link function, and h^{-1} is the inverse function of h.

In this work, we consider an approximate design ξ as $\xi = \begin{Bmatrix} \boldsymbol{x}_1, & \dots, & \boldsymbol{x}_n \\ \lambda_1, & \dots, & \lambda_n \end{Bmatrix}$ with $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ if $i \neq j$. The $\lambda_i (\lambda_i \geq 0)$ represents the fraction of experiments that is to be carried out at design point $\boldsymbol{x}_i$ and $\sum_{i=1}^n \lambda_i = 1$. The Fisher information matrix of the GLM in Equation (1) is:

$$\mathsf{I}(\xi, \boldsymbol{\beta}) = \sum_{i=1}^{n} \lambda_i \boldsymbol{g}(\boldsymbol{x}_i) w(\boldsymbol{x}_i) \boldsymbol{g}^T(\boldsymbol{x}_i), \tag{2}$$

where $w(\boldsymbol{x}_i) = 1/\{\mathrm{var}(Y(\boldsymbol{x}_i))[h'(\mu(\boldsymbol{x}_i))]^2\}$. The notation $\mathsf{I}(\xi, \boldsymbol{\beta})$ emphasizes the dependence on the design ξ and regression coefficient $\boldsymbol{\beta}$.

For GLMs, the design issues tend to be much more challenging than those in linear models due to the complicated Fisher information matrix. The Fisher information matrix $\mathsf{I}(\xi, \boldsymbol{\beta})$ often depends on the regression coefficient $\boldsymbol{\beta}$ through the function $w(\boldsymbol{x}_i)$ in Equation (2). Since most optimal design criteria can be expressed as a function of the Fisher information matrix, a scientific understanding of locally optimal designs is often conducted under the assumption that some initial estimates of the model coefficients are available. Most works and algorithms on design of GLMs focus on the D-optimality or A-optimality for accurate estimation of regression coefficients. Atkinson & Woods (2015) provided a review of recent work on the designs of GLMs mainly based on D-optimality. Practically, design criteria related to model prediction accuracy are of large interest, especially in decision-making and artificial intelligence (e.g., Schein & Ungar, 2007; Settles, 2009; Bilgic, Mihalkova & Getoor, 2010). Under such a consideration, the I-optimality that aims at minimizing the average variance of prediction over the experimental region Ω (Atkinson, 2014) is often used in the literature. For the linear model, Haines (1987) proposed a simulated annealing algorithm to obtain exact I-optimal design. Meyer & Nachtsheim (1988, 1995) used simulated annealing and coordinate-exchange algorithms to construct exact I-optimal design. However, there are few works on I-optimality for GLMs in the literature.

In practical applications such as additive manufacturing (Sun et al., 2018), different regions of exploratory variables $\boldsymbol{x}$ present different features of interest. For example, one may be interested in the prediction of response over a subregion Ω_C of Ω instead of the whole experimental region Ω (Atkinson, 2014). For instance, the experimental region Ω could be the hypercube $[-1, 1]^d$. But it is likely that the responses corresponding to positive explanatory variables are of great importance, or even only those responses are of interest. Khuri et al. (2006) proposed the criterion that minimizes the mean-squared error of the prediction at a single explanatory variable value, where Ω_C contains a single point. Based on this motivation, we propose a so-called *Elastic I-optimality* that is more general and flexible than the classical I-optimality. The EI-optimality criterion aims at minimizing the integrated mean squared prediction error with respect to *a certain probability measure on the experimental region.* This criterion shares a similar spirit but is more general than the criterion in the work of Box & Draper (1963).

The contribution of this work is to study a general and flexible prediction-oriented optimality criterion, EI-optimality, to fill the gap in the theory of EI-optimal designs for GLMs, and to advance an efficient algorithm of constructing EI-optimal designs for GLMs. Specifically, we first establish the EI-optimality for GLMs and study theoretical properties of optimal fractions (i.e., weights) $\lambda_1, \ldots, \lambda_n$ given design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The resultant theoretical properties are not limited to EI-optimality, but also can be applied to other optimality criteria. Based on these theoretical investigations, we develop an efficient sequential algorithm for constructing EI-optimal designs for GLMs. The proposed algorithm sequentially adds points into design and optimizes their weights simultaneously using the multiplicative algorithm (Titterington, 1978). The convergence properties of the proposed algorithm are established. A key contribution of the proposed algorithm is to ensure the computational efficiency with sound theoretical convergence. Our theoretical results provide theoretical insights on the optimal weights, which is very crucial to achieve a fast convergence rate for the proposed algorithm. The advantages of the proposed algorithm are: (1) very easy to implement; (2) theoretically proven convergence; (3) computationally efficient; (4) suitable for other optimality criteria by simple modification.

The multiplicative algorithm, first proposed by Titterington (1978) and Silvey, Titterington & Torsney (1978), has been widely used in finding optimal designs of linear regression models. The main drawback of the multiplicative algorithm is that it requires a large candidate pool of points and updating the weights of all candidate points simultaneously can be computationally expensive. However, in our proposed sequential algorithm, guided by the theoretical results on optimal weights, we can apply the multiplicative algorithm to a much smaller set of points, which breaks the barrier of the original multiplicative algorithm and thus greatly improves the efficiency of the algorithm. The proposed algorithm is not only computationally efficient, but also very simple to implement. Furthermore, by employing the multiplicative algorithm, only nonnegative weights will be obtained and one does not need to deal with potential negative weights as in the algorithm proposed by Yang, Biedermann & Tang (2013). It is worth pointing out that the proposed algorithm can be easily extended to construct optimal design under other optimality criteria, like Φ_p-optimality.

The remainder of this work is organized as follows. The proposed prediction-oriented EI-optimality for GLMs is established in Section 2. Section 3 details the findings on computing optimal weights given any set of design points. The proposed sequential algorithm and its convergence properties are developed in Section 4. Numerical examples are conducted in Section 5 to evaluate the performance of the proposed algorithm. We conclude this work with some discussion in Section 6. All the proofs are reported in the Appendix.

2. THE EI-OPTIMALITY CRITERION FOR GLM

The GLM is a generalization of various statistical models, including linear, logistic and Poisson regression. For GLMs, making a prediction of response $Y(\boldsymbol{x})$ at given input $\boldsymbol{x}$ is always an important objective in many practical applications (e.g., Schein & Ungar, 2007; Settles, 2009; Bilgic, Mihalkova & Getoor, 2010). Thus, it is of great interest to adopt a prediction-oriented criterion for computing the optimal design. Following this idea, it is natural to consider the design based on mean response $\mu(\boldsymbol{x})$ that is square integrable with respect to some probability measure ν defined on $\mathbb{R}^d$. The associated probability distribution is given by $F_{\text{IMSE}}(\boldsymbol{x}) = \nu\left(\prod_{i=1}^{d}(-\infty, x_i]\right)$. Then, a general and flexible Elastic I-optimality criterion is defined as follows.

Definition 1. *The elastic integrated mean squared prediction error (EIMSE) is defined in terms of the difference between the true mean response, $\mu(\boldsymbol{x})$, and the fitted mean* *response $\hat{\mu}(\boldsymbol{x})$ as:*

$$EIMSE(\xi, \boldsymbol{\beta}, F_{IMSE}) = \mathbb{E}\left[\int_{\Omega} (\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}))^2 \, dF_{IMSE}(\boldsymbol{x})\right], \tag{3}$$

where F_{IMSE} is the probability distribution induced by probability measure ν.

The classical I-optimality for linear models is defined as the average variance of response over the experimental region Ω (Atkinson, 2014):

$$I(\xi) = \int_{\Omega} \mathrm{var}(Y|\boldsymbol{x}) d\boldsymbol{x} \Big/ \int_{\Omega} d\boldsymbol{x} = \mathbb{E}\left[\int_{\Omega} (\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}))^2 \, dF_{\mathrm{unif}}(\boldsymbol{x})\right],$$

where F_{unif} is the uniform distribution on Ω, $\hat{\mu}$ and μ are the fitted mean response and true mean response of the linear model, respectively. Obviously, the I-optimality $I(\xi)$ is a special case of EI-optimality in Equation (3) with ν chosen to be the uniform probability measure. In Atkinson, Donev & Tobias (2006) (Chapter 10.6), the authors briefly mentioned a similar criterion on average prediction variance involving a probability distribution, without theoretical investigation and efficient algorithms. Here, we formally propose the EI-optimality and discuss its advantages in detail. By introducing the probability measure ν, the EI-optimality is more flexible to assist practical applications in several scenarios: (1) the responses corresponding to $\boldsymbol{x}$ on a subregion Ω_C are of particular interest; (2) the responses corresponding to different values of $\boldsymbol{x}$ are of different importance; (3) the responses corresponding to a finite number of $\boldsymbol{x}$ values are of interest. It is worthwhile to point out that when ν is chosen to be the Dirac measure that puts a unit mass at $\boldsymbol{x}_0$ that maximizes $\mathbb{E}\left[(\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}))^2\right]$, i.e., $\mathbb{E}\left[(\hat{\mu}(\boldsymbol{x}_0) - \mu(\boldsymbol{x}_0))^2\right] = \sup_{\boldsymbol{x}\in\Omega} \mathbb{E}\left[(\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}))^2\right]$, the $\mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}})$ becomes the G-optimality (Atkinson, Donev & Tobias, 2006) that focuses on the maximum variance of the mean response.

Under the context of GLMs, the fitted mean response $\hat{\mu}(\boldsymbol{x}) = h^{-1}\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{g}(\boldsymbol{x})\right)$ can be expanded around the true mean response $\mu(\boldsymbol{x}) = h^{-1}\left(\boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{x})\right)$ using Taylor expansion, which is

$$\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}) = h^{-1}\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{g}(\boldsymbol{x})\right) - h^{-1}\left(\boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{x})\right) \approx \boldsymbol{c}(\boldsymbol{x})^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right),$$

with $\boldsymbol{c}(\boldsymbol{x}) = \left(\frac{\partial h^{-1}}{\partial \beta_1}(\boldsymbol{x}), \ldots, \frac{\partial h^{-1}}{\partial \beta_l}(\boldsymbol{x})\right)^T = \boldsymbol{g}(\boldsymbol{x})\left(\frac{dh^{-1}}{d\eta}\right)$. Here $\eta = \boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{x})$ is the linear predictor. Then, using the above first-order Taylor expansion, the elastic integrated mean squared error defined in Equation (3) could be approximated as,

$$\mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}}) = \mathbb{E}\left[\int_{\Omega} \boldsymbol{c}(\boldsymbol{x})^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \boldsymbol{c}(\boldsymbol{x}) dF_{\mathrm{IMSE}}\right].$$

In numerical analysis, the first-order Taylor expansion is often used to approximate the difference of a nonlinear function between adjacent points. Considering the design issue for GLMs, similar approaches were commonly used in the literature, such as the work in Schein & Ungar (2007) for logistic regression models.

Lemma 1. *For the GLMs in Equation (1), the Elastic IMSE $EI(\xi, \boldsymbol{\beta}, F_{IMSE})$ can be expressed as*

$$EI(\xi, \boldsymbol{\beta}, F_{IMSE}) = \mathrm{tr}\left(\mathsf{A}\mathsf{I}(\xi, \boldsymbol{\beta})^{-1}\right),$$

where the matrix $\mathsf{A} = \int_{\Omega} \boldsymbol{c}(\boldsymbol{x})\boldsymbol{c}(\boldsymbol{x})^T dF_{IMSE}(\boldsymbol{x}) = \int_{\Omega} \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x})\left[\frac{dh^{-1}}{d\eta}\right]^2 dF_{IMSE}(\boldsymbol{x})$ depends only on the regression coefficients, basis functions and the probability distribution F_{IMSE}, but not the *design. The Fisher information matrix* $\mathsf{I}(\xi, \boldsymbol{\beta})$*, defined in Equation (2), depends on regression coefficients, basis functions and design, but not on the probability distribution* F_{IMSE}.

Hereafter, we refer to the *EI-optimality* as the corresponding optimality criterion aiming at minimizing

$$\mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}}) = \mathrm{tr}\left(\mathsf{A}\mathsf{I}(\xi, \boldsymbol{\beta})^{-1}\right).$$

A design ξ^* is called an *EI-optimal design* if it minimizes $\mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}})$. In this work, we will focus on local EI-optimal designs given some initial estimate of regression coefficient $\boldsymbol{\beta}$. To simplify the notation, $\boldsymbol{\beta}$ will be omitted from the notation $\mathsf{I}(\xi, \boldsymbol{\beta})$ of the Fisher information matrix, and $\mathsf{I}(\xi)$ will be used instead.

3. ALGORITHMS FOR FINDING OPTIMAL WEIGHTS GIVEN DESIGN POINTS

In this section, we will investigate how to assign optimal weights to support points that minimize $\mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}})$ when the design points are given. One popular method in the literature (Yang, Biedermann & Tang, 2013) is the Newton–Raphson technique, which calculates the roots of the first-order partial derivatives of the criterion with respect to the weights. There are certain possible drawbacks of such a Newton–Raphson based method: first, it may result in weights outside $[0, 1]$ and thus further efforts are needed. Second, it requires the inversion of a Hessian matrix, which could be (numerically) singular. Moreover, as noted in Yang, Biedermann & Tang (2013), there is no guarantee of convergence. The problems of the Newton–Raphson method will be illustrated in the numerical examples in Section 5.

Specifically, we will derive a theorem on optimal weights given design points for Φ_p-optimality (Kiefer, 1974) in Section 3.1, and then will take the mathematical structure of EI-optimality as a slight variation of Φ_1-optimality. Guided by this theorem, we develop an efficient algorithm (Algorithm 1) to find the optimal weights given design points in Section 3.2. Note that Algorithm 1 can be used for both Φ_p-optimality and EI-optimality. Interestingly, this algorithm coincides with the well-known multiplicative algorithm, providing a good justification of applying multiplicative algorithm in our sequential algorithm in Section 4. There are several advantages of the multiplicative algorithm: simple to implement, guarantee of feasible weights, guarantee of convergence, and no Hessian matrix inversion.

3.1. Properties of Optimal Weights Given Design Points

Following the definition given in Kiefer (1985), the Φ_p-optimality is defined as

$$\Phi_p(\xi) = \left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-p}\right)\right]^{1/p}, \quad 0 < p < \infty,$$

with $\Phi_0(\xi)$ as D-optimality, $\Phi_\infty(\xi)$ as E-optimality and $\Phi_1(\xi)$ as A-optimality. Even more generally, one may be interested in several functions $\boldsymbol{f}(\boldsymbol{\beta}) = [f_1(\boldsymbol{\beta}), \ldots, f_q(\boldsymbol{\beta})]^T$ of regression coefficient $\boldsymbol{\beta}$. Then, the more general Φ_p-optimality is defined as (Kiefer, 1974)

$$\Phi_p(\xi) = \left(q^{-1}\mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}(\xi)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T\right]^p\right)^{1/p}, \quad 0 < p < \infty. \tag{4}$$

Given the fixed design points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, define $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_n]^T$ to be the weight vector with λ_i as the weight of the corresponding design point $\boldsymbol{x}_i$. We write the corresponding design as

$$\xi^\lambda = \left\{ \begin{array}{ccc} \boldsymbol{x}_1, & \ldots, & \boldsymbol{x}_n \\ \lambda_1, & \ldots, & \lambda_n \end{array} \right\}.$$

A superscript λ is added to emphasize that only the weight vector is changeable in the design under this situation. The optimal weight vector λ^* should be the one that minimizes $\Phi_p(\xi^\lambda)$ in Equation (4) with design points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ fixed.

Lemma 2. *The Φ_p is a convex function of weight vector λ, and*

$$\Phi_p(\xi^\lambda) = \left[q^{-1}\mathrm{tr}\left[\mathsf{I}^{-1}(\xi^\lambda)\mathsf{B}\right]^p\right]^{1/p},$$

where $\mathsf{B} = \left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)$ *is positive semidefinite with size $l \times l$ and rank $q \leq l$, with q as the length of $\boldsymbol{f}$.*

Consider another weight vector $\Delta\lambda = [\Delta\lambda_1, \dots, \Delta\lambda_n]^T$ with $0 \leq \Delta\lambda_i \leq 1, i = 1, \dots, n$ and $\sum_{i=1}^n \Delta\lambda_i = 1$. Then the convex combination $\tilde{\lambda} = (1-\alpha)\lambda + \alpha\Delta\lambda$ with $0 \leq \alpha \leq 1$ is also a feasible weight vector. Define the directional derivative of $\Phi_p(\xi^\lambda)$ in the direction of a new weight vector $\Delta\lambda = [\Delta\lambda_1, \dots, \Delta\lambda_n]$ as:

$$\psi(\Delta\lambda, \lambda) = \lim_{\alpha \to 0^+} \frac{\Phi_p(\xi^{\tilde{\lambda}}) - \Phi_p(\xi^\lambda)}{\alpha}.$$

Lemma 3. *The directional derivative of $\Phi_p(\xi^\lambda)$ in the direction of a new weight vector $\Delta\lambda = [\Delta\lambda_1, \dots, \Delta\lambda_n]^T$ can be calculated as,*

$$\psi(\Delta\lambda, \lambda) = \Phi_p(\xi^\lambda) - q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} \mathrm{tr}\left[\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^\lambda)^{-1}\mathsf{I}(\xi^{\Delta\lambda})\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right],$$

where $\mathsf{B} = \left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)$.

Remark 1. *Note that* $\mathsf{A} = \int_\Omega \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x})\left[\frac{dh^{-1}}{d\eta}\right]^2 dF_{IMSE}(\boldsymbol{x})$ *is an $l \times l$ positive semidefinite matrix. Although EI-optimality is not a member of Φ_p-optimality, since it has the same mathematical structure as $\Phi_1(\xi)$ with $\mathsf{B} = q\mathsf{A}$ defined in Lemma 1, the mathematical properties of Φ_p-optimality can be applied to EI-optimality.*

Corollary 1. *For EI-optimality, the directional derivative in the direction of a new weight vector $\Delta\lambda$ is*

$$\psi(\Delta\lambda, \lambda) = \mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{A}\right) - \sum_{i=1}^n \Delta\lambda_i w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^\lambda)^{-1}\mathsf{A}\mathsf{I}(\xi^\lambda)^{-1}\boldsymbol{g}(\boldsymbol{x}_i).$$

The following theorem provides a necessary and sufficient condition of optimal weight vector that minimizes $\Phi_p(\xi)$ when the design points are fixed.

Theorem 1. *Given a fixed set of design points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, the weight vector $\lambda^* = [\lambda_1^*, \dots, \lambda_n^*]^T$ minimizes $\Phi_p(\xi^\lambda)$ if and only if,*

$$\Phi_p(\xi^{\lambda^*}) = q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i),$$

for all design points $\boldsymbol{x}_i$ *with* $\lambda_i^* > 0$*; and*

$$\Phi_p(\xi^{\lambda^*}) \geq q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_j)\boldsymbol{g}(\boldsymbol{x}_j)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_j),$$

for all design points $\boldsymbol{x}_j$ *with* $\lambda_j^* = 0$.

Note that in Theorem 1, for an optimal weight vector, it is required to hold the equality for the nonzero weights. The following results provide a sufficient condition that a weight vector minimizes $\Phi_p(\xi^\lambda)$ and $\mathrm{EI}(\xi^\lambda, \boldsymbol{\beta}, F_{\mathrm{IMSE}})$, respectively.

Corollary 2. *(i) If a weight vector* $\boldsymbol{\lambda}^* = [\lambda_1^*, \ldots, \lambda_n^*]^T$ *satisfies*

$$\Phi_p(\xi^{\lambda^*}) = q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i),$$

for all design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$*, then* $\boldsymbol{\lambda}^*$ *minimizes* $\Phi_p(\xi^\lambda)$.

(ii) For EI-optimality, a sufficient condition that $\boldsymbol{\lambda}^*$ *minimizes* $EI(\xi^\lambda, \boldsymbol{\beta}, F_{IMSE})$ *is*

$$\mathrm{tr}(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{A}) = w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{A}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i), \quad for, \quad i = 1, \ldots, n.$$

The results in Theorem 1 and Corollary 2 provide a useful gateway to design an effective algorithm for finding the optimal weights given the design points.

3.2. Multiplicative Algorithm for Optimal Weights

As shown in Corollary 2, the solution of a system of equations

$$\Phi_p(\xi^{\lambda^*}) = q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i), \tag{5}$$

$i = 1, \ldots, n$, is a set of optimal weights that minimizes $\Phi_p(\xi^\lambda)$.

The weight of a design point $\boldsymbol{x}_i$ should be adjusted according to the values of the two sides of Equation (5). Note that for any weight vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]$, we have

$$\Phi_p(\xi^\lambda) = \sum_{i=1}^n \lambda_i q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^\lambda)^{-1}\boldsymbol{g}(\boldsymbol{x}_i).$$

Based on this observation, the weight λ_i of a design point $\boldsymbol{x}_i$ should be adjusted according to the value of $q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^\lambda)^{-1}\boldsymbol{g}(\boldsymbol{x}_i)$. Our strategy of finding optimal weights would be: if

$$\Phi_p(\xi^\lambda) < q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^\lambda)^{-1}\boldsymbol{g}(\boldsymbol{x}_i),$$

then the weight λ_i of $\boldsymbol{x}_i$ should be increased. On the other hand, if

$$\Phi_p(\xi^\lambda) > q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^\lambda)^{-1}\boldsymbol{g}(\boldsymbol{x}_i),$$

then the weight λ_i of $\boldsymbol{x}_i$ should be decreased. Thus, the ratio

$$\frac{q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right)^{p}\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)}{\Phi_p(\xi^{\lambda})}$$

would indicate a good adjustment for the current weight of a design point $\boldsymbol{x}_i$.

The details of the algorithm are described in Algorithm 1 as follows.

Algorithm 1

Assign a random weight vector $\lambda^0 = [\lambda_1^0, \ldots, \lambda_n^0]^T$, and $k = 0$.
while *change* $> 1e - 15$ and $k <$ *maxrun* **do**
  for $i = 1, \ldots, n$ **do**
    Update the weight of design point $\boldsymbol{x}_i$:

$$\begin{aligned}\lambda_i^{k+1} &= \lambda_i^k \frac{\left(\dfrac{q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p}\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)}{\Phi_p(\xi^{\lambda^k})}\right)^{\delta}}{\sum_{i=1}^n \lambda_i^k\left(\dfrac{q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p}\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)}{\Phi_p(\xi^{\lambda^k})}\right)^{\delta}} \\ &= \lambda_i^k \frac{\left[w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\right]^{\delta}}{\sum_{i=1}^n \lambda_i^k\left[w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\right]^{\delta}}, \end{aligned} \tag{6}$$

    change $= \max_{i=1,\cdots,n}(|\lambda_i^{k+1} - \lambda_i^k|)$.
    $k = k + 1$.
  end for
end while

We would like to remark that Algorithm 1 can be viewed as a well-known multiplicative algorithm proposed by Titterington (1978) and Silvey, Titterington & Torsney (1978). Originally, a heuristic explanation for the multiplicative algorithm is that $\lambda_i^{k+1} \propto \lambda_i^k \left(\left.\frac{\partial\Phi_p(\xi^{\lambda})}{\partial\lambda_i}\right|_{\lambda=\lambda^k}\right)^{\delta}$. Here, we obtain the same algorithm based on a sufficient condition of optimal weights given in Corollary 2. This may also explain why the multiplicative algorithm tends to converge slowly and result in many support points when there is a large candidate set. When the multiplicative algorithm is applied to a large candidate set, a very strong sufficient but not necessary condition is imposed on all the candidate points. This condition should be only imposed on the points with nonzero optimal weights. However, if the design points are appropriately chosen so that most of them have nonzero optimal weights, then the sufficient condition in Corollary 2 becomes almost a necessary condition for the optimal weights.

In Algorithm 1, there are two pre-specified parameters. One is a convergence parameter $\delta \in (0, 1)$ and the other one, *maxrun*, is the maximum number of iterations allowed. Following the suggestion by Fellman (1974) and Fiacco & Kortanek (1983), a convergence parameter of

$\delta = \frac{1}{2}$ is usually used for A-optimality. Since EI-optimality has similar mathematical structure to A-optimality, we choose $\delta = \frac{1}{2}$ in our algorithm. It is also observed that the computational performance of the proposed algorithm is quite robust regarding δ in *Example 2* of Section 5. We will show later in proposition 1 that Algorithm 1 converges to optimal weights monotonically. Thus, the algorithm is not very sensitive to the choice of maximum iterations allowed, *maxrun*. In all our numerical examples, we choose *maxrun* = 100, while the same parameter in Yang, Biedermann & Tang (2013) for Newton's method is 40. The following theorem provides a sufficient condition when the iterative formula in Algorithm 1 is feasible, and Corollary 3 shows that, for EI-optimality, one can ensure that every iteration in Algorithm 1 is always feasible by choosing appropriate basis functions.

Theorem 2. *Suppose* $\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B} \neq 0$ *in iteration k, then the iteration in Equation (6) is feasible, that is, the denominator*

$$\sum_{i=1}^{n} \lambda_i^k \left[w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T \mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B} \left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B} \right)^{p-1} \mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i) \right]^{\delta}$$

is positive.

For the EI-optimality, the iterative formula in Algorithm 1 becomes

$$\lambda_i^{k+1} = \lambda_i^k \frac{\left[w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T \mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{A}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i) \right]^{\delta}}{\sum_{i=1}^{n} \lambda_i^k \left[w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T \mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{A}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i) \right]^{\delta}},$$

with $\mathsf{A} = \int_\Omega \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x}) \left[\frac{dh^{-1}}{d\eta} \right]^2 dF_{\text{IMSE}}(\boldsymbol{x})$.

Corollary 3. *For the EI-optimality, with a nonsingular information matrix* $\mathsf{I}(\xi^{\lambda^k})$*, a positive definite* $\mathsf{A} = \int_\Omega \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x}) \left[\frac{dh^{-1}}{d\eta} \right]^2 dF_{IMSE}(\boldsymbol{x})$ *would ensure all the iterations in Equation (6) to be always feasible, that is, the denominator*

$$\sum_{i=1}^{n} \lambda_i^k \left[w(\boldsymbol{x}_i)\boldsymbol{g}^T(\boldsymbol{x}_i)\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{A}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i) \right]^{\delta}$$

is always positive.

For the Φ_p-optimality, the matrix $\mathsf{B} = \left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right)^T \left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right)$ is determined by $\boldsymbol{f}(\boldsymbol{\beta})$, the functions of $\boldsymbol{\beta}$ that are of interest, which are determined by the purpose of the experiment. But, for EI-optimality, the matrix $\mathsf{A} = \int_\Omega \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x}) \left[\frac{dh^{-1}}{d\eta} \right]^2 dF_{\text{IMSE}}(\boldsymbol{x})$ is always positive semi-definite and can be controlled by the experimenter before the optimal design is constructed. When a singular A is observed in the first step, one can choose the basis functions $g_1, \ldots, g_l$ carefully so that they are (nearly) orthogonal functions such that

$$\int_\Omega g_i(\boldsymbol{x})g_j(\boldsymbol{x}) \left[\frac{dh^{-1}}{d\eta} \right]^2 dF_{\text{IMSE}}(\boldsymbol{x}) \approx 0 \quad \text{for} \quad i \neq j.$$

By doing this, one can ensure the matrix A to be positive definite. In general, the Gram–Schmidt orthogonalization (Pursell & Trimble, 1991) can be used to construct uni/multi-variate orthogonal basis functions.

We can provide the convergence property of Algorithm 1, which generalizes Yu's work (Yu, 2010) to a broader class of optimality criteria, Φ_p-optimality.

Proposition 1 (Convergence of Algorithm 1). *Given the design points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ fixed, the weight vector $\boldsymbol{\lambda}^k$ obtained from Algorithm 1 monotonically converges to the optimal weight vector $\boldsymbol{\lambda}^*$ that minimizes Φ_p-optimality, as $k \to \infty$.*

4. PROPOSED SEQUENTIAL ALGORITHM OF CONSTRUCTING EI-OPTIMAL DESIGN

In this section, we will describe the proposed efficient algorithm for constructing EI-optimal design for GLMs. General equivalence theorems are one of the main theoretical tools to develop algorithms to construct optimal designs. Many Wynn–Federov type algorithms are developed based on general equivalence theorems (see Wynn, 1970, 1972; Whittle, 1973; Yang, Biedermann & Tang, 2013; Martín & Gutiérrez, 2015). Yang, Mandal & Majumdar (2016) developed an efficient algorithm to construct 2^k D-optimal factorial design with binary response based on a specialized version of general equivalence theorem on a pre-determined finite set of design points. We will first establish the general equivalence theorem for EI-optimality of GLMs in Section 4.1, which provides an intuitive way for choosing the support points to construct an EI-optimal design in a sequential fashion. Section 4.2 details the proposed algorithm and develops the convergence property of the proposed algorithm.

4.1. General Equivalence Theorem of EI-Optimality

As shown in Remark 1, the EI-optimality has a similar mathematical structure as the Φ_1-optimality, but the two criteria have different practical interpretations. The general equivalence theorem of Φ_p-optimality for GLMs has been established (Stufken & Yang, 2011), and it could be extended to EI-optimality for GLMs easily. We state the General Equivalence Theorem of EI-optimality for GLMs in the following Theorem 3, and the standard proof is omitted. The extended theoretical results facilitate the sequential choice of support points as that of the Fedorov–Wynn algorithm (Wynn, 1970; Fedorov, 1972).

Given two designs ξ and ξ', let the design $\tilde{\xi}$ be constructed as

$$\tilde{\xi} = (1-\alpha)\xi + \alpha\xi'.$$

Then, the derivative of $\mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}})$ in the direction of design ξ' is

$$\phi(\xi', \xi) = \lim_{\alpha \to 0^+} \frac{\mathrm{EI}(\tilde{\xi}, \boldsymbol{\beta}, F_{\mathrm{IMSE}}) - \mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}})}{\alpha}. \tag{7}$$

Lemma 4. *The $EI(\cdot, \boldsymbol{\beta}, F_{IMSE})$ is a convex function of the design ξ.*

With some algebra, we can obtain the directional derivative of $EI(\xi, \boldsymbol{\beta}, \Omega_c)$ in the direction of any design ξ' as,

$$\phi(\xi', \xi) = \lim_{\alpha \to 0^+} \frac{\mathrm{EI}(\tilde{\xi}, \boldsymbol{\beta}, F_{\mathrm{IMSE}}) - \mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}})}{\alpha} = \mathrm{tr}(\mathsf{I}(\xi)^{-1}\mathsf{A}) - \mathrm{tr}(\mathsf{I}(\xi')\mathsf{I}(\xi)^{-1}\mathsf{A}\mathsf{I}(\xi)^{-1}),$$

where $\mathsf{A} = \int_\Omega \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x}) \left[\frac{dh^{-1}}{d\eta}\right]^2 dF_{\text{IMSE}}(\boldsymbol{x})$. Moreover, we can also get the directional derivative of $\text{EI}(\xi, \boldsymbol{\beta}, F_{\text{IMSE}})$ in the direction of a single point $\boldsymbol{x}$ as

$$\phi(\boldsymbol{x}, \xi) = \text{tr}(\mathsf{I}(\xi)^{-1}\mathsf{A}) - w(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^T\mathsf{I}(\xi)^{-1}\mathsf{A}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x}).$$

Theorem 3 (General equivalence theorem). *The following three conditions of ξ^* are equivalent:*

1. *The design ξ^* minimizes*
$$EI(\xi, \boldsymbol{\beta}, \Omega_c) = \text{tr}(\mathsf{A}\mathsf{I}(\xi)^{-1}).$$

2. *The design ξ^* minimizes*
$$\sup_{\boldsymbol{x}\in\Omega} w(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x})\mathsf{I}(\xi)^{-1}\mathsf{A}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x}).$$

3. $w(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x})\mathsf{I}(\xi^*)^{-1}\mathsf{A}\mathsf{I}(\xi^*)^{-1}\boldsymbol{g}(\boldsymbol{x}) \le \text{tr}(\mathsf{I}(\xi^*)^{-1}\mathsf{A})$ *holds over the experimental region Ω, and the equality holds only at the points of support of the design ξ^*.*

According to Theorem 3, for an optimal design ξ^*, the directional derivative $\phi(\boldsymbol{x}, \xi^*)$ is nonnegative for any $\boldsymbol{x} \in \Omega$. It implies that for any nonoptimal design, there will be some directions in which the directional derivative $\phi(\boldsymbol{x}, \xi) < 0$. Given a current design ξ, to gain the maximal decrease in the EI-optimality criterion, we would choose a new support point $\boldsymbol{x}^*$ to be added into the design, if $\phi(\boldsymbol{x}^*, \xi) = \min_{\boldsymbol{x}\in\Omega} \phi(\boldsymbol{x}, \xi) < 0$. Then one can optimize the weights of all support points in the updated design, which is described in Algorithm 1 in Section 3. With this greedy search of the design points, we hope that most or all optimal weights in each iteration are nonzero, and that the multiplicative algorithm converges quickly. By iterating the selection of support point and the weight update of all support points, this two-step iterative procedure can be continued until $\min_{\boldsymbol{x}\in\Omega} \phi(\boldsymbol{x}, \xi) \ge 0$ for all $\boldsymbol{x} \in \Omega$, which means the updated design is EI-optimal. Such a sequential algorithm of constructing optimal designs, as described in Algorithm 2 in Section 4.2, follows similar spirits in the widely used Fedorov–Wynn algorithm (Wynn, 1970; Wynn, 1972), the multi-stage algorithm proposed by Yang, Biedermann & Tang (2013), and the combined algorithm proposed by Martín & Gutiérrez (2015). It is worth pointing out that when the region of explanatory variable $\boldsymbol{x}$ we are interested in, Ω_C is a subset of original experimental region Ω, the optimal design ξ^* is still defined and searched on the original experimental region Ω. As a result, the support points in the optimal design may locate outside Ω_C.

4.2. The Proposed Sequential Algorithm

As discussed in Section 3.2, the multiplicative algorithm tends to converge slowly under a large candidate set. The theoretical results in Theorem 3 provide insightful guidelines on sequential selection of design points. In combination with Algorithm 1 to find optimal weights for fixed design points in Section 3, we propose an efficient sequential algorithm to construct the EI-optimal design for GLMs. The details of the proposed sequential algorithm are summarized in Algorithm 2.

In practice, the stopping rule $\min_{\boldsymbol{x}\in\Omega} \phi(\boldsymbol{x}, \xi^r) \ge 0$ is impractical since it requires many iterations to make all the directional derivative values strictly positive (numerically it is unlikely to have exactly zero cases). To address this issue, we consider terminating the algorithm when the design efficiency is large enough, say close to 1.

Algorithm 2

1: Calculate matrix $\mathsf{A} = \int_\Omega \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x}) \left[\frac{dh^{-1}}{d\eta}\right]^2 dF_{\text{IMSE}}(\boldsymbol{x})$.
2: **if** cond($\mathsf{A} > 1e16$) **then**
3: Construct orthogonal basis $\boldsymbol{g}$ using Gram−Schmidt orthogonalization.
4: Calculate matrix A using the new orthogonal basis $\boldsymbol{g}$.
5: **end if**
6: Generate an N points candidate pool C using grid or Sobol sequence from experimental region Ω.
7: Choose an initial design points set $\mathcal{X}^0 = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{l+1}\}$ containing $l+1$ points.
8: Obtain optimal weights λ^0 of initial design points set $\mathcal{X}^0$ using Algorithm 1 and form the initial design $\xi^0 = \left\{\begin{matrix} \mathcal{X}^0 \\ \lambda^0 \end{matrix}\right\}$.
9: Calculate the lower bound of EI-efficiency of ξ^0,

$$\text{LEff}_{\text{EI}}(\xi^0|\xi^*) = \frac{\text{tr}(\mathsf{I}^{-1}(\xi^0)\mathsf{A})}{\max_{\boldsymbol{x}\in C} w(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^T\mathsf{I}(\xi^0)^{-1}\mathsf{A}\mathsf{I}(\xi^0)^{-1}\boldsymbol{g}(\boldsymbol{x})}.$$

10: Set $r = 1$.
11: **while** $\text{LEff}_{\text{EI}}(\xi^{r-1}|\xi^*) < reqeff$ and $r < maxiter$ **do**
12: Add the point $\boldsymbol{x}_r^* = \text{argmin}_{\boldsymbol{x}\in C}\phi(\boldsymbol{x}, \xi^{r-1})$ to the current design points set, i.e., $\mathcal{X}^r = \mathcal{X}^{r-1} \cup \boldsymbol{x}_r^*$, where $\phi(\boldsymbol{x}, \xi^{r-1})$ is the directional derivative expressed as

$$\phi(\boldsymbol{x}, \xi^{r-1}) = \text{tr}(\mathsf{I}(\xi^{r-1})^{-1}\mathsf{A}) - w(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^T\mathsf{I}(\xi^{r-1})^{-1}\mathsf{A}\mathsf{I}(\xi^{r-1})^{-1}\boldsymbol{g}(\boldsymbol{x}).$$

13: Obtain optimal weights λ^r of the current design points set $\mathcal{X}^r$ using Algorithm 1 and form the current design $\xi^r = \left\{\begin{matrix} \mathcal{X}^r \\ \lambda^r \end{matrix}\right\}$.
14: Calculate the lower bound of EI-efficiency of ξ^r,

$$\text{LEff}_{\text{EI}}(\xi^r|\xi^*) = \frac{\text{tr}(\mathsf{I}^{-1}(\xi^r)\mathsf{A})}{\max_{\boldsymbol{x}\in C} w(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^T\mathsf{I}(\xi^r)^{-1}\mathsf{A}\mathsf{I}(\xi^r)^{-1}\boldsymbol{g}(\boldsymbol{x})}.$$

15: $r = r + 1$.
16: **end while**

The efficiency of a design ξ relative to another design ξ' under a Φ_p-optimality is defined as (Pukelsheim, 2006)

$$\text{Eff}_{\Phi_p}(\xi|\xi') = \frac{\Phi_p(\xi')}{\Phi_p(\xi)}.$$

Asymptotically, the efficiency is essentially the ratio between the number of trials required to obtain the same amount of information measured by a specific Φ_p-optimality using design ξ' and ξ. The following proposition provides a lower bound $\text{LEff}_{\Phi_p}(\xi|\xi^*)$ of Φ_p-efficiency of a design ξ relative to the corresponding Φ_p-optimal design ξ^*, and this lower bound is used in the stopping rule in Algorithm 2.

Theorem 4. *With candidate pool C, the Φ_p-efficiency of any design ξ relative to Φ_p-optimal design ξ^*, $Eff_{\Phi_p}(\xi|\xi^*)$, is lower bounded by*

$$LEff_{\Phi_p}(\xi|\xi^*) = \frac{\Phi_p(\xi)}{\max_{\boldsymbol{x}\in C} q^{-1/p} \left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi)^{-1}\mathsf{B}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x}_i)}.$$

Specifically, for EI-optimality, the lower bound of a design ξ relative to the corresponding EI-optimal design ξ^* is

$$\mathrm{LEff}_{\mathrm{EI}}(\xi|\xi^*) = \frac{\mathrm{tr}(\mathsf{I}^{-1}(\xi)\mathsf{A})}{\max_{\boldsymbol{x}\in C} w(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^T\mathsf{I}(\xi)^{-1}\mathsf{A}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x})}.$$

Following the stopping criterion used in Harman, Filová & Richtárik (2019), when the lower bound $\mathrm{LEff}_{\mathrm{EI}}(\xi^r|\xi^*)$ of the efficiency $\mathrm{Eff}_{\mathrm{EI}}(\xi^r|\xi^*)$ reaches a user-specified threshold $reqeff$, there is no practical reason to continue the search. We choose $reqeff = 0.99$ in all the numerical examples. Another user-specified parameter is the maximum iterations allowed, *maxiter*. We choose $maxiter = 100$, and the proposed algorithm converges within 100 iterations in all the numerical examples conducted, while Newton's method fails to converge within 100 iterations for one case in *Example 2*. More details could be found in Section 5.

Note that the proposed sequential algorithm of constructing EI-optimal design for GLMs does not require the computation of a Hessian matrix inverse like Newton–Raphson methods do. Thus the algorithm can avoid the issue of singular Hessian matrices and can be more computationally efficient than the conventional methods. The sequential nature of the proposed algorithm also enables efficient search of optimal weights without updating weights for all candidate points. Martín & Gutiérrez (2015) considered a similar sequential algorithm that combines Whittle's method (Whittle, 1973) with one iteration of the multiplicative algorithm to update the weights after a new design point is added. As a result, the weights are not necessarily optimized in each iteration of the combined algorithm, and the proof of Theorem 5 cannot be applied to the combined algorithm. The convergence property of the combined algorithm for general Φ_p-optimality is not studied in Martín & Gutiérrez (2015).

Note that the sequential algorithm (Algorithm 2) can also be easily modified to achieve Φ_p-optimal designs, where the analytic formula of directional derivative $\phi(\boldsymbol{x}, \xi)$ for Φ_p-optimality has been studied before (Atkinson, Donev & Tobias, 2006; Stufken & Yang, 2011; Yang, Biedermann & Tang, 2013). Moreover, we also establish the convergence property of the proposed sequential algorithm as follows.

Theorem 5 (Convergence of Algorithm 2). *With a discrete design region C, the design constructed by Algorithm 2 converges to EI-optimal design ξ^* that minimizes $EI(\xi, \boldsymbol{\beta}, F_{IMSE})$, as $r \to \infty$, i.e.,*

$$\lim_{r\to\infty} EI(\xi^r, \boldsymbol{\beta}, F_{IMSE}) = EI(\xi^*, \boldsymbol{\beta}, F_{IMSE}).$$

When the experimental region Ω is continuous, a discretization would be needed to form the candidate pool C. We would like to point out that the choice of candidate pool C would affect the efficiency of Algorithm 2, and the computational time increases dramatically as the candidate pool gets large. We suggest using a grid as the candidate pool when the dimension of explanatory variables is low, and choosing the Sobol sequence (Sobol, 1967) as the candidate pool when the explanatory variable dimension is high. The Sobol sequence is a space-filling design that covers the experimental domain Ω well and is efficient when the dimension of the explanatory variable is high. To further improve the efficiency of the algorithm, a search strategy inspired by Yang

& Stufken (2012) could be employed. One can start with a more sparse Sobol sequence, and achieve the current best design using Algorithm 2. Then, we can further create denser and denser candidate pools in the neighbourhood of the support points in the current best design until there is no further improvement under the EI-optimality criterion.

5. NUMERICAL EXAMPLES

In this section, we will conduct numerical studies to evaluate the performance of the proposed sequential algorithm (Algorithm 2). As noted in Section 4.2, Algorithm 2 can be applied to construct Φ_p-optimal designs such as A-optimal design. To the best of our knowledge, there is no existing algorithm that can directly construct the EI-optimal design for GLMs. However, the Newton-type method (Yang, Biedermann & Tang, 2013) could be revised to construct EI-optimal designs. The proposed algorithm (Algorithm 2) will be compared with the Newton-type method in Yang, Biedermann & Tang (2013), which adopts Newton's method to update the weights of design points and is an efficient algorithm in the literature. The comparison will be conducted under different GLMs with various settings of variable dimensions. Both algorithms are implemented in MATLAB, and the code for Newton's method is converted from the SAS code kindly provided by the authors of Yang, Biedermann & Tang (2013). All codes were run on a MacBook Pro with a 2.4 GHz Intel Core i5 processor. The Newton-type method requires a well-conditioned Hessian matrix to update the weights, but the Hessian matrix could be numerically singular in a certain iteration which results in inaccurate weight updates. In this numerical example, the generalized inverse of Hessian matrix is used for the Newton-type method. The proposed algorithm always returns nonnegative weights, only eliminates the design points with almost zero weight, and does not require Hessian matrix inversion. A grid candidate pool of size $N = (s+1)^d$ with $s+1$ equally spaced points for each explanatory variable x_i, $i = 1, \dots, d$ is used in all numerical examples. Note that our Algorithm 2 does not require the candidate pool to be a grid. We would suggest using a space-filling design such as a Sobol sequence as the candidate pool for high-dimensional explanatory variable. The target efficiency lower bound $reqeff$ in the stopping criterion is chosen to be 0.99, and we consider two designs are both good enough if their efficiency lower bounds exceed 0.99. Thus, it is fair to compare the efficiency of the algorithms based on the computational time.

Example 1. The setting of this example follows the Example 2 in Yang, Biedermann & Tang (2013), which considers the linear model

$$
\begin{aligned}
&Y \sim \theta_1 + \theta_2 x_1 + \theta_3 x_1^2 + \theta_4 x_2 + \theta_5 x_1 x_2 + N(0, \sigma^2),\\
&\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5),
\end{aligned}
$$

where $\Omega = \{(2i/s - 1, j/s), i = 0, 1, \dots, s, j = 0, 1, \dots, s\}$, where s is the number of grid points in each variable and the total number of points in Ω is $N = (s+1)^2$. Note that since the experimental region is discrete, it is not necessary to discretize the experimental region. The proposed algorithm and the Newton-type method are compared under A- and EI-optimality. For EI-optimality, we choose $F_{\text{IMSE}} = F_{\text{unif}}$ on Ω.

Here we consider the same initial design, consisting of five randomly chosen points in Ω, for both Newton-type method and the proposed algorithm. For different values of s, Figure 1 reports the average computational time (in seconds) of the two algorithms and average efficiency $\text{Eff}(\xi^*_{\text{Proposed}} | \xi^*_{\text{Newton}})$ based on 10 randomly chosen initial designs, where ξ^*_{Proposed} and ξ^*_{Newton} are the optimal designs achieved using the proposed algorithm and Newton's method, respectively.

Figure 1 shows that the proposed algorithm is more efficient than the Newton-type method since it does not requires an additional remedy for negative weights, and the multiplicative algorithm is only performed on a small set of design points. It is apparent that the computational

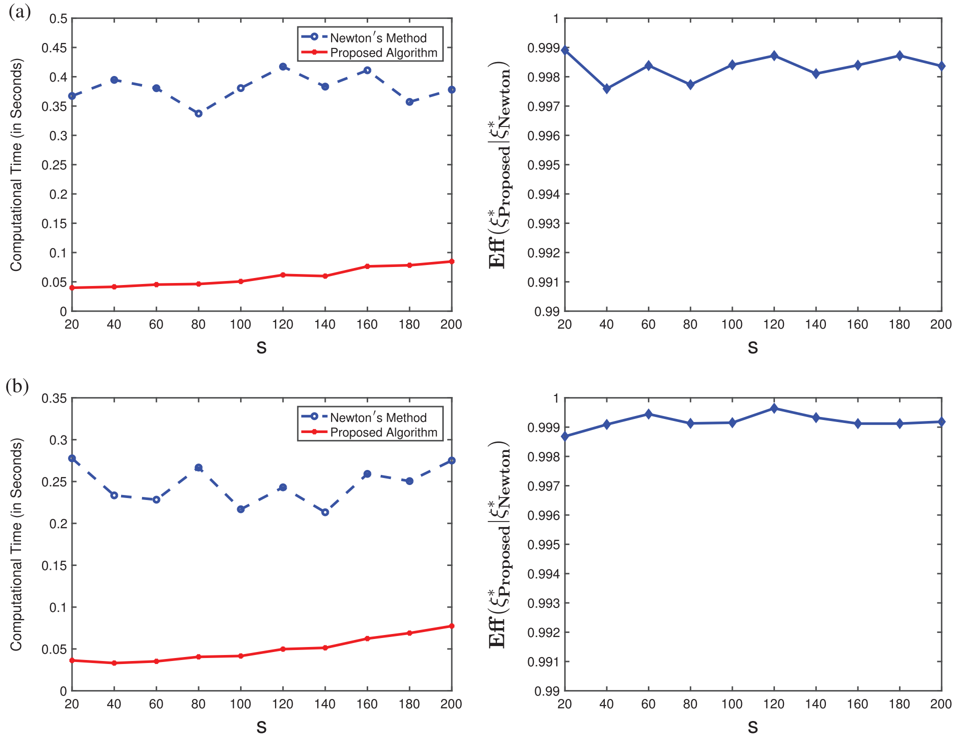

FIGURE 1: Computational time and relative efficiency $\text{Eff}(\xi^*_{\text{Proposed}}|\xi^*_{\text{Newton}})$ for A- and EI-optimal designs with linear regression model. (a) EI-optimality, candidate pool size $N = (s+1)^2$. (b) A-optimality, candidate pool size $N = (s+1)^2$.

time of Newton's method largely varies and significantly depends on the initial design points. In contrast, the proposed algorithm is very robust to initial design points and has relatively stable computational performance. Although the efficiencies of both designs exceed 0.99, the design achieved by Newton's method has a slightly smaller optimality criterion value, and this could be due to the quadratic convergence rate of Newton's method. As shown in Figure 1, the efficiency of the design achieved by the proposed algorithm relative to that achieved by Newton's method is always above 99.7%, and the proposed algorithm is about four times faster than Newton's method.

We further compare the above EI-optimal design with $F_{\text{IMSE}}(\boldsymbol{x}) = F_{\text{unif}}(\boldsymbol{x})$ to the EI-optimal design with a different $F_{\text{IMSE}}(\boldsymbol{x}) = F^{(1)}_{\text{arcsine}}(x_1)F^{(2)}_{\text{arcsine}}(x_2)$, where $F^{(1)}_{\text{arcsine}}(x_1)$ and $F^{(2)}_{\text{arcsine}}(x_2)$ are arcsine distributions on $[-1, 1]$ and $[0, 1]$, respectively. Different from the uniform distribution, an arcsine distribution on bounded support $[a, b]$ has probability density function $\rho(x) = \frac{1}{\pi\sqrt{(x-a)(x-b)}}$, $x \in [a, b]$, which puts more weight towards the interval ends. Figure 2 shows the support points of EI-optimal designs when F_{IMSE} is chosen to be the uniform distribution and the arcsine distribution. Under $F_{\text{IMSE}} = F_{\text{unif}}$ on Ω, the efficiency of EI-optimal design achieved using arcsine distribution relative to that achieved using uniform distribution is 0.9564. On the other hand, under $F_{\text{IMSE}} = F_{\text{arcsine}}$ on Ω, the efficiency of EI-optimal design achieved using uniform distribution relative to that achieved using arcsine distribution is 0.9595.

Example 2. This example considers the logistic regression model with a binary response variable. Assume that the domain of d-dimensional explanatory variable $\boldsymbol{x} = [x_1, \ldots, x_d]$ is standardized to be a unit hypercube $[-1, 1]^d$. With l basis functions $\mathbf{g}(\boldsymbol{x}) = [g_1(\boldsymbol{x}), \ldots, g_l(\boldsymbol{x})]^T$

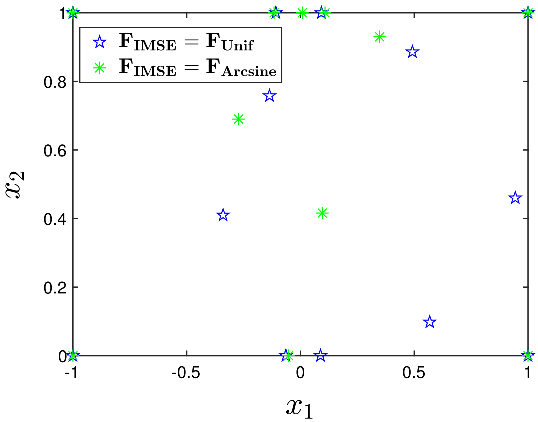

FIGURE 2: Support points of EI-optimal designs with $F_{\text{IMSE}} = F_{\text{unif}}$ and $F_{\text{IMSE}} = F_{\text{arcsine}}$.

and regression coefficients $\boldsymbol{\beta} = [\beta_1, \dots, \beta_l]^T$, the logistic regression model with binary response $Y \in \{0, 1\}$ is defined as:

$$Prob(Y = 1|\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{x})}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{x})}}.$$

Usually the basis functions $\boldsymbol{g}(\boldsymbol{x})$ are low-degree polynomials of explanatory variable $\boldsymbol{x}$, and in this example we consider linear predictors, that is, $\boldsymbol{g}(\boldsymbol{x}) = (1, \boldsymbol{x}^T)^T = (1, x_1, \dots, x_d)^T$. Given a design $\xi = \left\{ \begin{matrix} \boldsymbol{x}_1, & \dots, & \boldsymbol{x}_n \\ \lambda_1, & \dots, & \lambda_n \end{matrix} \right\}$, the EI-optimality criterion with some probability distribution F_{IMSE} is

$$\text{EI}(\xi, \boldsymbol{\beta}, F_{\text{IMSE}}) = \text{tr}\left(\mathsf{A}\mathsf{I}(\xi)^{-1}\right),$$

where $\mathsf{A} = \int_\Omega (1, \boldsymbol{x}^T)^T (1, \boldsymbol{x}^T) v(\boldsymbol{x}) dF_{\text{IMSE}}(\boldsymbol{x})$ with $v(\boldsymbol{x}) = e^{2(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}/(1 + e^{\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}})^4$, and $\mathsf{I}(\xi) = \sum_{i=1}^{n} \lambda_i w(\boldsymbol{x}_i)(1, \boldsymbol{x}_i^T)^T (1, \boldsymbol{x}_i^T)$ with $w(\boldsymbol{x}_i) = e^{\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i}/(1 + e^{\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i})^2$. In this example, we consider (case i) the classical I-optimality with F_{IMSE} being a uniform distribution on $\Omega_c = \Omega = [-1, 1]^d$, that is, $F'_{\text{IMSE}}(\boldsymbol{x}) = \frac{1}{2^d}, \boldsymbol{x} \in \Omega$ and $\mathsf{A} = \int_{[-1,1]^d} \frac{1}{2^d} (1, \boldsymbol{x}^T)^T (1, \boldsymbol{x}^T) v(\boldsymbol{x}) d\boldsymbol{x}$; and (case ii) F_{IMSE} being a uniform distribution on positive half hypercube $\Omega_c = [0, 1]^d \subset \Omega = [-1, 1]^d$, that is, $F'_{\text{IMSE}}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in \Omega_c, \\ 0, & \boldsymbol{x} \notin \Omega_c \end{cases}$ and $\mathsf{A} = \int_{[0,1]^d} (1, \boldsymbol{x}^T)^T (1, \boldsymbol{x}^T) v(\boldsymbol{x}) d\boldsymbol{x}$. To investigate the performance of the proposed algorithm when the explanatory variable dimension d gets large, we compute the EI-optimal design under the following scenarios:

(a) $d = 1$, $\boldsymbol{\beta} = [0.2, 1.6]^T$
(b) $d = 2$, $\boldsymbol{\beta} = [2, 1, -2.5]^T$.
(c) $d = 3$, $\boldsymbol{\beta} = [0.5, 1.6, -2.5, 2]^T$.

We explore three properties of the proposed algorithm: (1) efficiency compared to Newton's method; (2) choice of convergence rate δ in Algorithm 1 and (3) choice of candidate pool size.

• *Efficiency compared to Newton's method*

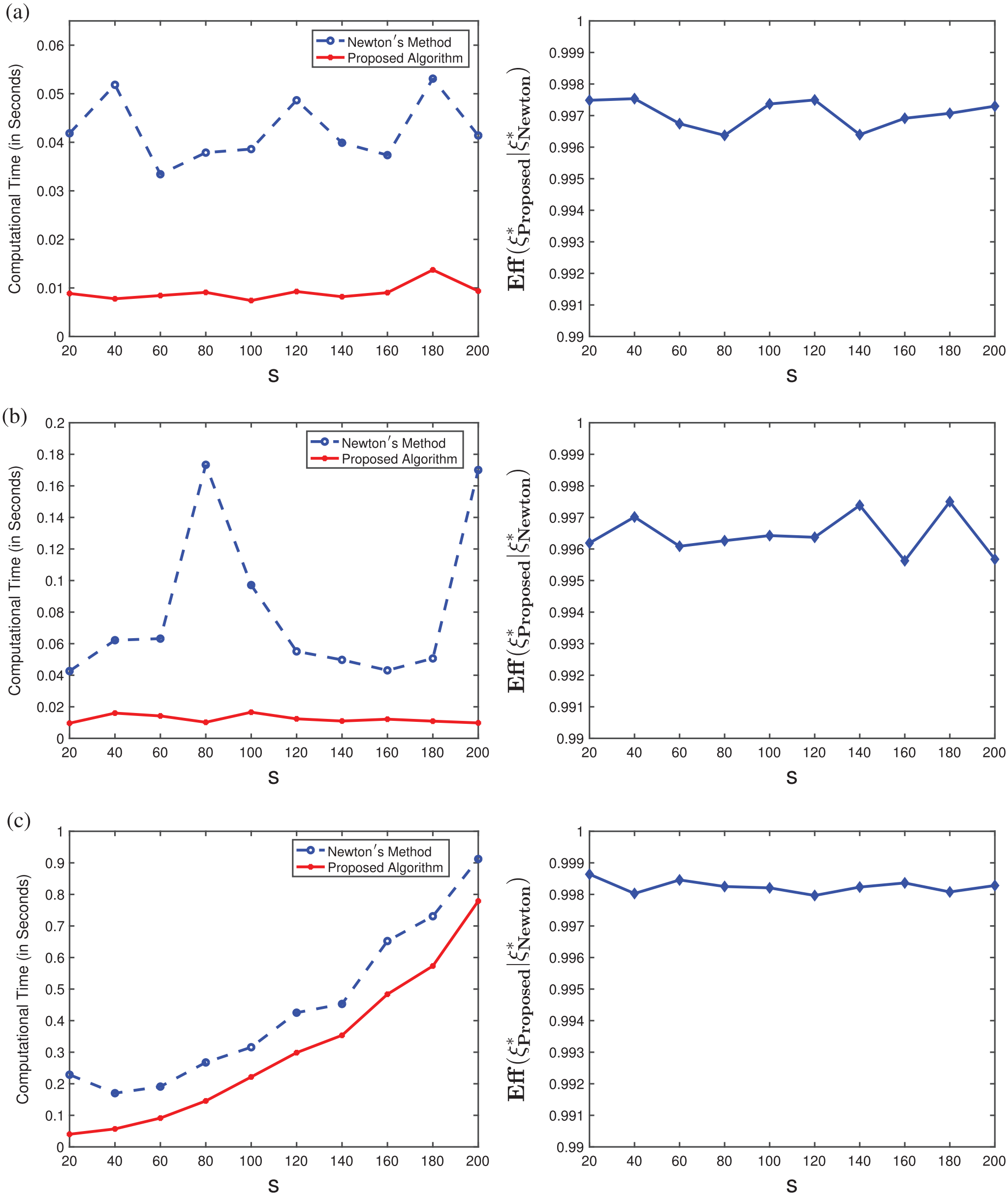

FIGURE 3: Computational time (in seconds) and relative efficiency Eff$(\xi^*_{\text{Proposed}}|\xi^*_{\text{Newton}})$ for EI-optimal designs with d-dimensional logistic regression model. (a) $d = 1$, case i, candidate pool size $N = s + 1$. (b) $d = 1$, case ii, candidate pool size $N = s + 1$. (c) $d = 2$, case i, candidate pool size $N = (s + 1)^2$.

The computational time comparison between the proposed algorithm and Newton's method is summarized in Figures 3 and 4. For $d = 3$ of case ii, there are situations in which Newton's method does not converge in 100 iterations, that is, the efficiency lower bound does not reach 0.99 in 100 iterations. In contrast, the proposed method is much more stable and meets the efficiency criterion in all scenarios. Thus, for $d = 3$, we only report the computational

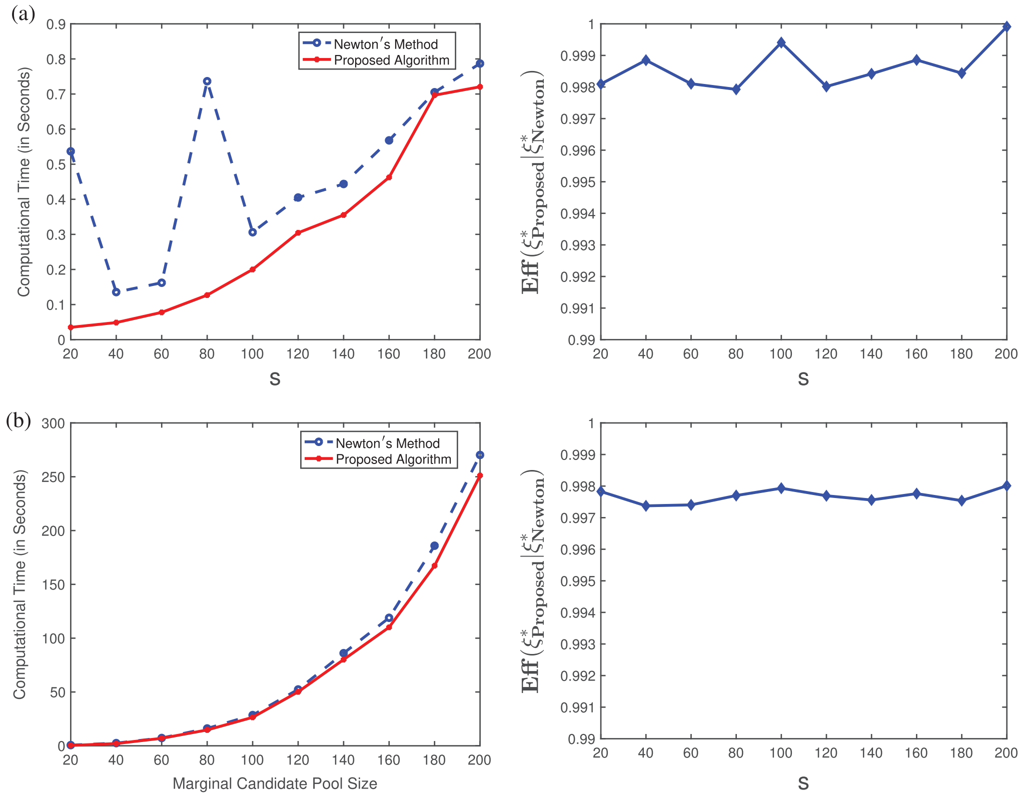

FIGURE 4: Computational time (in seconds) and relative efficiency Eff$(\xi^*_{\text{Proposed}}|\xi^*_{\text{Newton}})$ for EI-optimal designs with d-dimensional logistic regression model (continued). (a) $d = 2$, case ii, candidate pool size $N = (s + 1)^2$. (b) $d = 3$, case i, candidate pool size $N = (s + 1)^3$.

time comparison for case i where both algorithms converge within 100 iterations. Clearly, the proposed method is quite computationally efficient. When the candidate pool size gets larger, Figure 4b shows that the computational times of two algorithms become close with the proposed algorithm to be slightly faster. Note that both algorithms are Wynn–Fedorov type algorithms, and the computational time is dominated by evaluating the directional derivative over the large candidate pool. As the candidate pool gets large, the computational advantage of the proposed algorithm becomes less pronounced. However, it is worth remarking that even with similar computational times, the proposed algorithm still has other advantages such as guaranteed convergence and simple implementation.

- *Choice of convergence rate δ*
  As discussed in Section 3.2, the convergence parameter $\delta \in (0, 1)$ in Algorithm 1 is user defined and we choose $\delta = 0.5$ for EI-optimality. We further explore the computational time of the proposed algorithm with different choices of δ ranging from 0.05 to 0.95 in $d = 2$ in case i. The results are shown in Figure 5. It can be seen that the computational time of the proposed algorithm is quite robust to the choice of δ as long as δ is not too small.
- *Choice of candidate pool size*
  Since a discrete candidate pool C is required in Algorithm 2, discretization is needed when the experimental region is continuous. When using a grid as the candidate pool, its size can increase exponentially as the dimension d and the number of grid points in each dimension

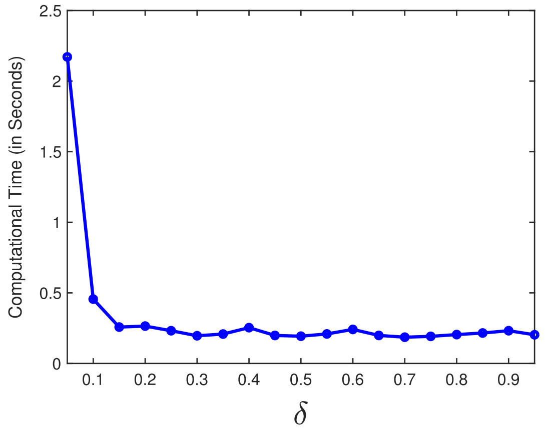

FIGURE 5: computational time (in seconds) of the proposed algorithm with different convergence parameter δ

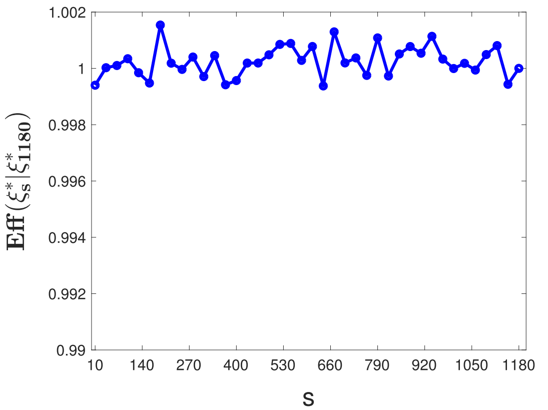

FIGURE 6: Relative efficiency of EI-optimal designs with different candidate pool sizes $N = (s + 1)^2$.

increases, and correspondingly the computational time of the proposed algorithm also increases dramatically. We explore how the size of candidate pool would influence the efficiency of the achieved design via $d = 2$ case i. The EI-optimal designs are constructed when the candidate pool size varies from $N = 10^2$ to $N = 1180^2$. The efficiencies of the optimal designs relative to the optimal design constructed using $N = 1180^2$ are reported in Figure 6. It is not surprising that the relative efficiency does not increase monotonically as the candidate pool size increases. One explanation is that the quality of the obtained design depends on whether the candidate pool contains the support points of the optimal design with a continuous experimental region, but not the size of the candidate pool. Although the candidate pool size ranges widely from $N = 10^2$ to $N = 1180^2$, the relative efficiencies are all very close to 1 (higher than 0.999), which indicates that a very fine grid candidate pool may not be necessary to construct an EI-optimal design. Based on this observation, we would suggest using a Sobol sequence as a set of candidate design points for our proposed method when the dimension of design region d is high.

Example 3. This example considers the Poisson regression models, which is a popular statistical tool to model count data in many applications (e.g., Gart, 1964; El-Shaarawi, Maul & Block, 1987). Like in *Example 2*, we assume that the domain of $d-$dimensional explanatory variable $\boldsymbol{x} = [x_1, \dots, x_d]$ is standardized to be a unit hypercube $\Omega = [-1, 1]^d$. With l basis functions $\boldsymbol{g}(\boldsymbol{x}) = [g_1(\boldsymbol{x}), \dots g_l(\boldsymbol{x})]^T$ and regression coefficients $\boldsymbol{\beta} = [\beta_1, \dots, \beta_l]^T$, the Poisson regression model with count response $Y \in \{0, 1, \dots, \}$ has the mean function:

$$\mu(\boldsymbol{x}) = \mathbb{E}[Y(\boldsymbol{x})] = e^{\boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{x})}.$$

Similar to *Example 2*, we consider linear predictors, i.e., $\boldsymbol{g}(\boldsymbol{x}) = (1, \boldsymbol{x}^T)^T = (1, x_1, \dots, x_d)^T$. Given a design $\xi = \begin{Bmatrix} \boldsymbol{x}_1, & \dots, & \boldsymbol{x}_n \\ \lambda_1, & \dots, & \lambda_n \end{Bmatrix}$, the classical I-optimality criterion with $F_{\text{IMSE}} = F_{\text{unif}}$ on $[-1, 1]^d$ is

$$\text{EI}(\xi, \boldsymbol{\beta}, F_{\text{unif}}) = \text{tr}\left(\mathsf{A}\mathsf{I}(\xi)^{-1}\right),$$

with $\mathsf{A} = \int \frac{1}{2^d}(1, \boldsymbol{x}^T)^T(1, \boldsymbol{x}^T)e^{2(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}d\boldsymbol{x}$ and $\mathsf{I}(\xi) = \sum_{i=1}^n \lambda_i e^{\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i}(1, \boldsymbol{x}_i^T)^T(1, \boldsymbol{x}_i^T)$. We consider the same scenarios as in *Example 2* for Poisson regression model and compute the corresponding classical I-optimal designs:

(a) $d = 1$, $\boldsymbol{\beta} = [0.2, 1.6]^T$
(b) $d = 2$, $\boldsymbol{\beta} = [2, 1, -2.5]^T$.
(c) $d = 3$, $\boldsymbol{\beta} = [0.5, 1.6, -2.5, 2]^T$.

Here we also use the grids to form the set of candidate design points. The computation comparison is shown in Figure 7. Similar to the results in *Example 2*, when the candidate pool size is moderate, the proposed algorithm outperforms the Newton's method by a large margin regarding the computational efficiency, and preserves a high design efficiency at the same time.

Example 4. In this example, we would like to provide some comparison between EI-optimal designs and other parameter estimation oriented D- and A-optimal designs through a real-world potato packing example in Woods et al. (2006). The experiment contains $d = 3$ quantitative variables—vitamin concentration in the prepackaging dip and levels of two gases in the packing atmosphere. The response is binary representing the presence or absence of liquid in the pack after 7 days. All explanatory variables are standardized, and the experimental region $\Omega = [-1, 1]^3$. We consider one of the candidate models used in the real study: a logit model with quadratic basis. The estimates of regression coefficients from the preliminary study are given in Table 1. $F_{\text{IMSE}} = F_{\text{unif}}$ on Ω is used to define EI-optimality.

Figure 8 shows the support points of D-, A- and EI-optimal Designs. The relative EI-efficiency of D- and A-optimal designs relative to the EI-optimal design are 80.64% and 85.44%, respectively. To some extent, this demonstrates the importance and necessity of accurate estimation of regression coefficients to make precise prediction, but D- or A-optimality may not be the most appropriate criterion to use when the prediction is of interest. The relative A- and D-efficiency of the EI-optimal design relative to the corresponding A- and D-optimal designs are 81.12% and 88.76%, respectively. Thus, it is important to choose the appropriate optimality criterion based on the purpose of the experiment.

6. DISCUSSION

In this work, we study a general and flexible prediction-oriented criterion EI-optimality for GLMs and advance an efficient sequential algorithm with sound convergence properties for

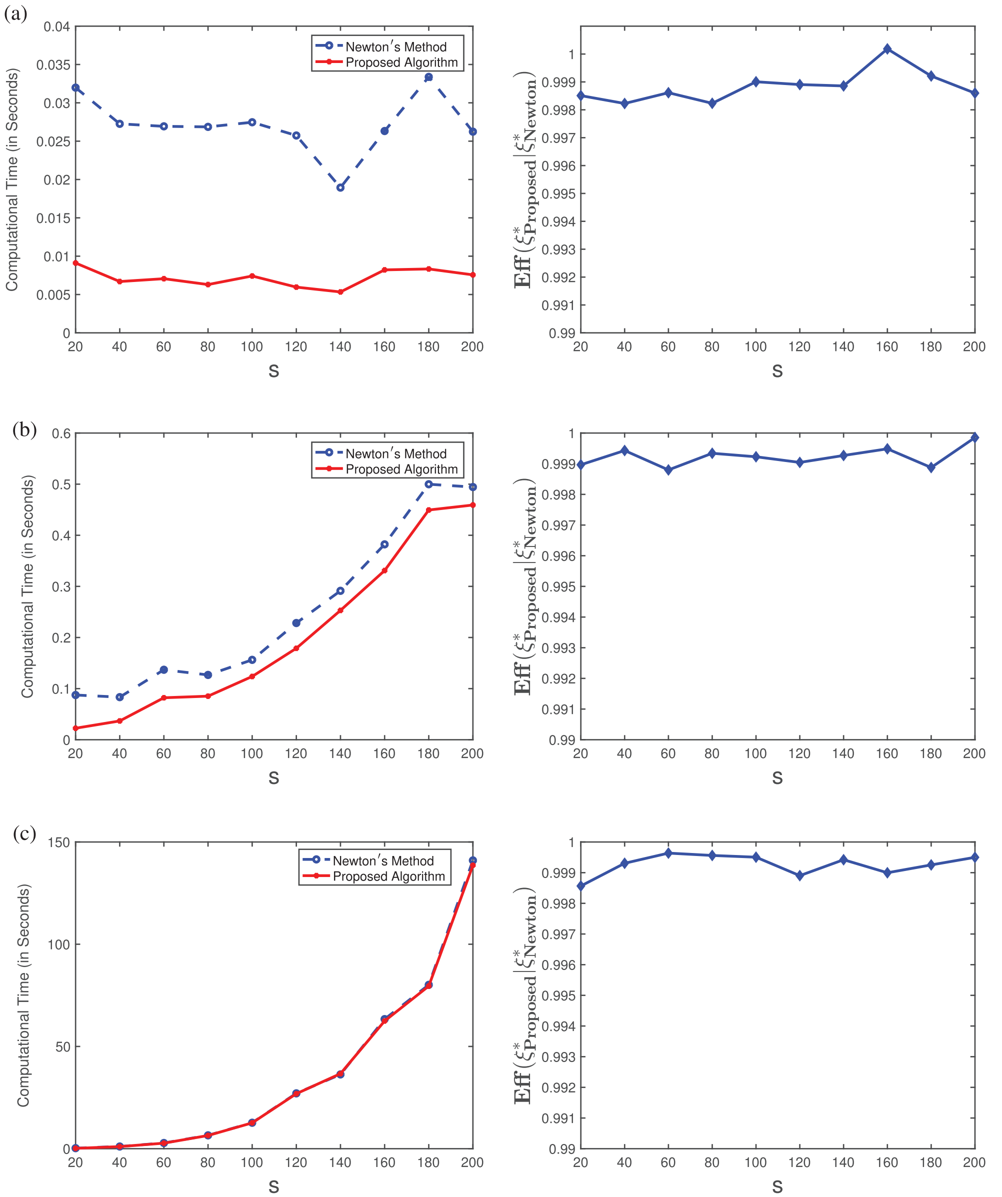

FIGURE 7: Computational time (in seconds) and relative efficiency $\text{Eff}(\xi^*_{\text{Proposed}}|\xi^*_{\text{Newton}})$ for EI-optimal designs with d-dimensional Poisson regression model. (a) $d = 1$, candidate pool size $N = s + 1$. (b) $d = 2$, candidate pool size $N = (s + 1)^2$. (c) $d = 3$, candidate pool size $N = (s + 1)^3$.

constructing EI-optimal designs. Through a deep investigation on the theoretical properties of the EI-optimality, we have obtained an insightful understanding of the proposed algorithm on how to sequentially choose the support point and update the weights of support points of the design. The computational advantages of the proposed algorithm over Newton's method are demonstrated through numerical examples with moderately sized candidate pools for various types of GLMs. Moreover, all the computations in the proposed algorithm are explicit and simple

TABLE 1: Logit model of potato packing example.

Term	Intercept	x_2	x_3	x_2x_3	x_1^2	x_2^2	x_3^2
Coefficient	−2.93	−0.52	−0.79	−0.66	0.94	0.79	1.82

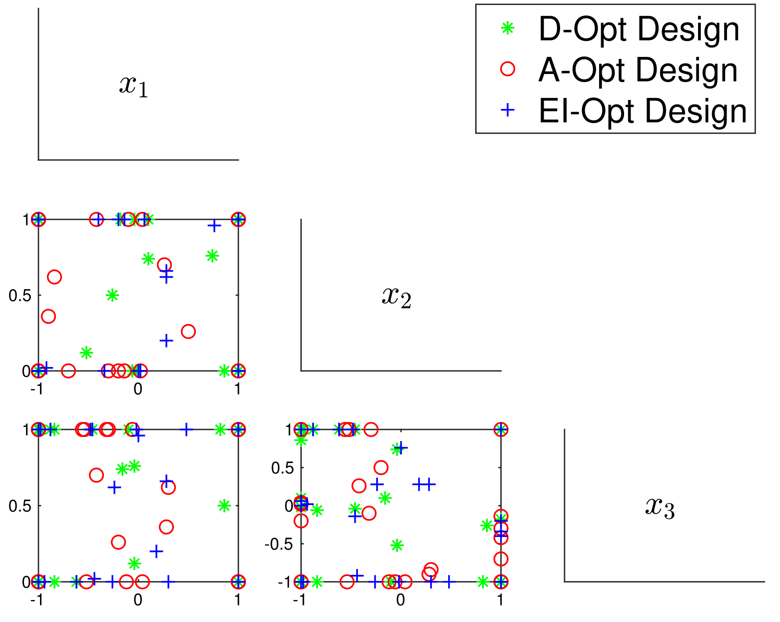

FIGURE 8: Support points of D-, A- and EI-optimal designs.

to implement. The proposed method exemplifies a good case on the integration of theory and computation to advance the development of new statistical methodology.

It is worth remarking that the proposed sequential algorithm (Algorithm 2) is not restricted to EI-optimality for constructing optimal designs. The proposed sequential algorithm can be extended to other optimality criteria when the directional derivative $\phi(\xi', \xi)$ of optimality criterion in Equation (7) exists. Although the convergence property of the proposed algorithm (Theorem 5) in Section 4.2 is stated in the context of EI-optimality, a proof on the convergence of the proposed algorithm for general Φ_p-optimality is provided in the Appendix. As the EI-optimality has the same mathematical structure as Φ_1-optimality, the convergence of the algorithm still holds. Note that this work focuses on the local EI-optimal designs for the GLMs, which depends on the given regression coefficients. The parameter dependence problem in the design of GLMs is an important yet challenging issue (Khuri et al., 2006; Woods et al., 2006). For instance, the EI-efficiency of the compromise design proposed by Woods et al. 2006 relative to the local EI-optimal design under three model specifications (see Woods et al.,, 2006) are 49%, 80% and 92% respectively. Bayesian or pseudo-Bayesian approaches are generally recognized as an appealing solution to address the parameter-dependence issue. For the proposed EI-optimality, we will further investigate the robust EI-optimal design for the GLMs. One possibility is to establish a tight upper bound for the integrated mean squared error to relax the dependency on the regression coefficients. Then we can modify the proposed algorithm in search of optimal designs based on the robust optimality criterion constructed according to the upper bound. Hickernell &

Liu (2002) developed some theoretical results in a similar direction for linear regression models under model uncertainty, and Li & Hickernell (2014) extended the results to linear regression models with gradient information.

We would like to remark that the proposed method may not have great computational efficiency when the candidate pool gets very large. One direction for tackling this issue is to better take advantage of the information on probability measure in the design region for constructing the design. For example, one can modify the selection of candidate pool based on the probability measure in the design region. Another interesting point is that the support points in the optimal design can be outside of the prediction region of interest, which could be a potential limitation of the proposed method.

For future research, it will be interesting to construct EI-optimal designs for the models with both quantitative and qualitative responses (Deng & Jin, 2015). Since the GLMs include both the linear regression for continuous response variables and the logistic regression for binary response, it would be interesting to study the EI-optimal designs under the consideration of jointly modelling both quantitative and qualitative responses. Finally, the proposed sequential algorithm is not restricted to find desirable designs for physical experiments. It can be applied towards finding space-filling designs in computer experiments (Deng, Hung & Lin, 2015).

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation and by College of Science and Health Faculty Summer Research Grant Program, DePaul University. The authors thank the anonymous referees for comments that led to improvements in the manuscript.

BIBLIOGRAPHY

Atkinson, A. C. (2014). Optimal design. *Wiley StatsRef: Statistics Reference Online*, 1–17.

Atkinson, A. C., Donev, A. N., & Tobias, R. D. (2006). *Optimum Experimental Designs, with SAS*. Oxford University Press.

Atkinson, A. C. & Woods, D. C. (2015). Designs for generalized linear models. *Handbook of Design and Analysis of Experiments*, Chapman & Hall/CRC Press, Boca Raton, 471–514.

Atwood, C. L. (1969). Optimal and efficient designs of experiments. *The Annals of Mathematical Statistics*, 1570–1602.

Bernstein, D. S. (2009). *Matrix Mathematics*, 2nd ed., Princeton University Press.

Bilgic, M., Mihalkova, L., & Getoor, L. (2010). Active learning for networked data. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 79–86.

Box, G. E. & Draper, N. R. (1963). The choice of a second order rotatable design. *Biometrika*, 335–352.

Deng, X., Hung, Y., & Lin, C. D. (2015). Design for computer experiments with qualitative and quantitative factors. *Statistica Sinica*, 1567–1581.

Deng, X. & Jin, R. (2015). QQ models: Joint modeling for quantitative and qualitative quality responses in manufacturing systems. *Technometrics*, 57, 320–331.

El-Shaarawi, A., Maul, A., & Block, J. -C. (1987). Application of Poisson regression to the analysis of bacteriological data. *Water Quality Research Journal*, 22, 298–307.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

Fedorov, V. V. & Hackl, P. (1997). *Model-Oriented Design of Experiments*, Vol. 125. Springer Science & Business Media.

Fellman, J. (1974). *On the Allocation of Linear Observations*. Societas Scientiarum Fennica.

Fiacco, A. V. & Kortanek, K. O. (Eds.) (1983). *A Moment Inequality and Monotonicity of an Algorithm*.

Gart, J. J. (1964). The analysis of Poisson regression with an application in virology. *Biometrika*, 51, 517–521.

Haines, L. M. (1987). The application of the annealing algorithm to the construction of exact optimal designs for linear-regression models. *Technometrics*, 29, 439–447.

Harman, R., Filová, L., & Richtárik, P. (2019). A randomized exchange algorithm for computing optimal approximate designs of experiments. *Journal of the American Statistical Association*, 1–30.

Hickernell, F. J. & Liu, M. (2002). Uniform designs limit aliasing. *Biometrika*, 89, 893–904.

Khuri, A. I., Mukherjee, B., Sinha, B. K., & Ghosh, M. (2006). Design issues for generalized linear models. *Technometrics*, 48, 520–529.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 849–879.

Kiefer, J. (1985). Optimal design: Variation in structure and performance under change of criterion. *Biometrika*, 62, 277–288.

Li, Y. & Hickernell, F. J. (2014). Design of experiments for linear regression models when gradient information is available. *Journal of Statistical Planning and Inference*, 144, 141–151.

Martín, R. M. & Gutiérrez, I. G. C. (2015). Combined algorithm to compute D-optimal designs. *Journal of Computational and Applied Mathematics*, 278, 248–257.

Meyer, R. K. & Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37, 60–69.

Meyer, R. K. & Nachtsheim, C. J. (1988). Simulated annealing in the construction of exact optimal design of experiments. *American Journal of Mathematical and Management Science*, 8, 329-359.

Nelder, J. & Wedderburn, R. (1972). Generalized Linear Models. *Journal of Royal Statistical Society –Series B*, 135, 370–384.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. SIAM.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. John Wiley & Sons, Inc.

Pursell, L. & Trimble, S. T. (1991). Gram–Schmidt orthogonalization by Gauss elimination. *The American Mathematical Monthly*, 98, 544–549.

Schein, A. I. & Ungar, L. H. (2007). Active learning for logistic regression: An evaluation. *Machine Learning*, 68, 235–265.

Settles, B. (2009). *Active learning literature survey*, Tech. rep., University of Wisconsin-Madison Department of Computer Sciences.

Silvey, S. D., Titterington, D. M., & Torsney, B. (1978). An algorithm for optimal designs on a finite design space. *Communications in Statistics –Theory and Methods*, 14, 1379–1389.

Sobol, I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7, 86–112.

Stufken, J. & Yang, M. (2011). Optimal designs for generalized linear models. *Design and Analysis of Experiments, Volume 3: Special Designs and Applications*, Vol. 810, 137.

Sun, H., Rao, P. K., Kong, Z. J., Deng, X., & Jin, R. (2018). Functional quantitative and qualitative models for quality modeling in a fused deposition modeling process. *IEEE Transactions on Automation Science and Engineering*, 15, 393–403.

Titterington, D. M. (1978). Estimation of correlation coefficients by ellipsoidal trimming. *Applied Statistics*, 27, 227–234.

Whittle, P. (1973). Some general points in the theory of optimal experimental design. *Journal of Royal Statistical Society –Series B*, 35, 123–130.

Woods, D. C., Lewis, S. M., Eccleston, J. A., & Russell, K. (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics*, 48, 284–292.

Wynn, H. (1970). The sequential generation of D-optimum experimental designs. *Annals of Mathematical Statistics, A*, 7, 1399–1412.

Wynn, H. P. (1972). Results in the theory and construction of D-optimum experimental designs. *Journal of Royal Statistical Society –Series B*, 34, 133–147.

Yang, J., Mandal, A., & Majumdar, D. (2016). Optimal designs for 2^k factorial experiments with binary response. *Statistica Sinica*, 385–411.

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.

Yang, M. & Stufken, J. (2012). Identifying locally optimal designs for nonlinear models: A simple extension with profound consequences. *The Annals of Statistics*, 40, 1665–1681.

Yu, Y. (2010). Monotonic convergence of a general algorithm for computing optimal designs. *The Annals of Statistics*, 38, 1593–1606.

APPENDIX

Proof of Lemma 1.

$$
\begin{aligned}
\mathrm{EI}(\xi, \boldsymbol{\beta}, F_{\mathrm{IMSE}}) &= \mathbb{E}\left[\int_\Omega \boldsymbol{c}(\boldsymbol{x})^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \boldsymbol{c}(\boldsymbol{x}) dF_{\mathrm{IMSE}}(\boldsymbol{x})\right] \\
&= \int_\Omega \boldsymbol{c}(\boldsymbol{x})^T \mathbb{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\right] \boldsymbol{c}(\boldsymbol{x}) dF_{\mathrm{IMSE}}(\boldsymbol{x}) \\
&\approx \int_\Omega \mathrm{tr}\left(\boldsymbol{c}(\boldsymbol{x})\boldsymbol{c}(\boldsymbol{x})^T \mathsf{I}(\xi, \boldsymbol{\beta})^{-1}\right) dF_{\mathrm{IMSE}}(\boldsymbol{x}) \\
&= \mathrm{tr}\left[\left(\int_\Omega \boldsymbol{c}(\boldsymbol{x})\boldsymbol{c}(\boldsymbol{x})^T dF_{\mathrm{IMSE}}(\boldsymbol{x})\right)\mathsf{I}(\xi, \boldsymbol{\beta})^{-1}\right] = \mathrm{tr}\left(\mathsf{A}\mathsf{I}(\xi, \boldsymbol{\beta})^{-1}\right).
\end{aligned}
$$

The approximation is provided by the fact that the estimated regression coefficient $\hat{\boldsymbol{\beta}}$ follows $N(\boldsymbol{\beta}, \mathsf{I}(\xi, \boldsymbol{\beta})^{-1})$ asymptotically. ∎

Proof of Lemma 2. For matrix $\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$ with size $q \times l$ and matrix $\mathsf{I}(\xi^\lambda)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T$ with size $l \times q$, and $q \leq l$, the eigenvalues of $\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}(\xi^\lambda)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T$ are the eigenvalues of $\mathsf{I}(\xi^\lambda)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$, with extra eigenvalues being 0 if there are any. Thus, the eigenvalues of $\left[\mathsf{I}(\xi^\lambda)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right]^p$ are the eigenvalues of $\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}(\xi^\lambda)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T\right]^p$ with extra eigenvalues being 0. Thus, we could rewrite

$$
\mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}(\xi^\lambda)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T\right]^p = \mathrm{tr}\left[\mathsf{I}(\xi^\lambda)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)\right]^p = \mathrm{tr}\left[\mathsf{I}^{-1}(\xi^\lambda)\mathsf{B}\right]^p, \tag{A1}
$$

where $\mathsf{B} = \left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)$ which is positive semidefinite with size $l \times l$ and rank $q \leq l$. Then, by Smith decomposition, there exists a nonsingular matrix S of size $l \times l$ such that

$$
\mathsf{B} = \mathsf{S}^T \begin{pmatrix} \boldsymbol{I}_q & \boldsymbol{0}_{q\times(l-q)} \\ \boldsymbol{0}_{(l-q)\times q} & \boldsymbol{0}_{(l-q)\times(l-q)} \end{pmatrix} \mathsf{S} = \mathsf{S}^T \begin{pmatrix} \boldsymbol{I}_q \\ \boldsymbol{0}_{(l-q)\times q} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_q & \boldsymbol{0}_{q\times(l-q)} \end{pmatrix} \mathsf{S},
$$

where $\boldsymbol{I}_q$ is the identity matrix of size $q \times q$. Thus, Equation (A1) could be written as

$$
\begin{aligned}
&\mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}(\xi^\lambda)^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T\right]^p \\
&= \mathrm{tr}\left[\mathsf{I}(\xi^\lambda)^{-1}\mathsf{S}^T \begin{pmatrix} \boldsymbol{I}_q \\ \boldsymbol{0}_{(l-q)\times q} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_q & \boldsymbol{0}_{q\times(l-q)} \end{pmatrix} \mathsf{S}\right]^p \\
&= \mathrm{tr}\left\{\left[\begin{pmatrix} \boldsymbol{I}_q & \boldsymbol{0}_{q\times(l-q)} \end{pmatrix}\left((\mathsf{S}^T)^{-1}\mathsf{I}(\xi^\lambda)\mathsf{S}^{-1}\right)^{-1}\begin{pmatrix} \boldsymbol{I}_q \\ \boldsymbol{0}_{(l-q)\times q} \end{pmatrix}\right]^{-1}\right\}^{-p}.
\end{aligned} \tag{A2}
$$

Consider two weights λ_1 and λ_2, and define $\lambda_3 = (1-a)\lambda_1 + a\lambda_2$. With $0 \le a \le 1$, λ_3 is still a feasible weight vector, and

$$(\mathsf{S}^T)^{-1}\mathsf{I}(\xi^{\lambda_3})\mathsf{S}^{-1} = (1-a)(\mathsf{S}^T)^{-1}\mathsf{I}(\xi^{\lambda_1})\mathsf{S}^{-1} + a(\mathsf{S}^T)^{-1}\mathsf{I}(\xi^{\lambda_2})\mathsf{S}^{-1}. \tag{A3}$$

By the theorem in section 3.13 of Pukelsheim (1993),

$$\left[\begin{pmatrix}\boldsymbol{I}_q & \mathbf{0}_{q\times(l-q)}\end{pmatrix}\left((\mathsf{S}^T)^{-1}\mathsf{I}(\xi^{\lambda})\mathsf{S}^{-1}\right)^{-1}\begin{pmatrix}\boldsymbol{I}_q & \mathbf{0}_{(l-q)\times q}\end{pmatrix}^T\right]^{-1} \tag{A4}$$

is matrix concave in $(\mathsf{S}^T)^{-1}\mathsf{I}(\xi^{\lambda})\mathsf{S}^{-1}$. Thus, by linearity of matrix $(\mathsf{S}^T)^{-1}\mathsf{I}(\xi^{\lambda})\mathsf{S}^{-1}$ in weight in Equation (A3), $\left[\begin{pmatrix}\boldsymbol{I}_q & \mathbf{0}_{q\times(l-q)}\end{pmatrix}\left((\mathsf{S}^T)^{-1}\mathsf{I}(\xi^{\lambda})\mathsf{S}^{-1}\right)^{-1}\begin{pmatrix}\boldsymbol{I}_q & \mathbf{0}_{(l-q)\times q}\end{pmatrix}^T\right]^{-1}$ is also concave in weight vector λ.

Since $\mathrm{tr}(\mathsf{C}^{-p})^{1/p}$ is nonincreasing and convex for any positive semidefinite matrix C (Fedorov & Hackl, 1997, see page 22), together with concavity of (A4), the composite function

$$\left[\mathrm{tr}\left\{\left[\begin{pmatrix}\boldsymbol{I}_q & \mathbf{0}_{q\times(l-q)}\end{pmatrix}\left((\mathsf{S}^T)^{-1}\mathsf{I}(\xi^{\lambda})\mathsf{S}^{-1}\right)^{-1}\begin{pmatrix}\boldsymbol{I}_q & \mathbf{0}_{(l-q)\times q}\end{pmatrix}^T\right]^{-1}\right\}^{-p}\right]^{1/p}$$

is a convex function of weight vector λ (Bernstein, 2009, see page 480). As a result, by Equation (A2), $\Phi_p(\xi^{\lambda})$ is convex in weight vector λ. ■

Proof of Lemma 3. Given $\tilde{\lambda} = (1-\alpha)\lambda + \alpha\Delta\lambda$, we have $\mathsf{I}(\xi^{\tilde{\lambda}}) = (1-\alpha)\mathsf{I}(\xi^{\lambda}) + \alpha\mathsf{I}(\xi^{\Delta\lambda})$. We still use Equation (A1) to rewrite $\Phi_p(\xi^{\lambda})$ as

$$\Phi_p(\xi^{\lambda}) = \left(q^{-1}\mathrm{tr}\left[\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right]^p\right)^{1/p},$$

where $\mathsf{B} = \left(\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T\left(\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)$.

For any positive semidefinite matrix C as a function of α, the derivative of its inverse C^{-1} can be calculated as $\frac{\partial \mathsf{C}^{-1}}{\partial\alpha} = -\mathsf{C}^{-1}\frac{\partial \mathsf{C}}{\partial\alpha}\mathsf{C}^{-1}$ (Bernstein, 2009). So, the derivative of $\mathsf{I}(\xi^{\tilde{\lambda}})^{-1}\mathsf{B}$ with respect to α can be expressed as,

$$\frac{\partial\left[\mathsf{I}(\xi^{\tilde{\lambda}})^{-1}\mathsf{B}\right]}{\partial\alpha} = \frac{\partial\mathsf{I}(\xi^{\tilde{\lambda}})^{-1})}{\partial\alpha}\mathsf{B} = -\mathsf{I}(\xi^{\tilde{\lambda}})^{-1}[\mathsf{I}(\xi^{\Delta\lambda}) - \mathsf{I}(\xi^{\lambda})]\mathsf{I}(\xi^{\tilde{\lambda}})^{-1}\mathsf{B}.$$

Then, the directional derivative of $\Phi_p(\xi^{\lambda})$ is

$$\begin{aligned}
\psi(\Delta\lambda, \lambda) &= \left.\frac{\partial\Phi_p(\xi^{\tilde{\lambda}})}{\partial\alpha}\right|_{\alpha=0} \\
&= q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^{\tilde{\lambda}})^{-1}\mathsf{B}\right)^p\right]^{1/p-1}\mathrm{tr}\left.\left[\left(\mathsf{I}(\xi^{\tilde{\lambda}})^{-1}\mathsf{B}\right)^{p-1}\left(-\mathsf{I}(\xi^{\tilde{\lambda}})^{-1}[\mathsf{I}(\xi^{\Delta\lambda}) - \mathsf{I}(\xi^{\lambda})]\mathsf{I}(\xi^{\tilde{\lambda}})^{-1}\mathsf{B}\right)\right]\right|_{\alpha=0} \\
&= q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right)^p\right]^{1/p-1}\mathrm{tr}\left[\left(\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right)^p - \left(\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{I}(\xi^{\Delta\lambda})\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right]
\end{aligned}$$

$$= \Phi_p(\xi^{\lambda}) - q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right)^p\right]^{1/p-1}\operatorname{tr}\left[\left(\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{I}(\xi^{\Delta\lambda})\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right].$$

■

Proof of Theorem 1. Since λ^* minimizes Φ_p, and Φ_p is a convex function of weight vector as proved in Lemma 2,

$$\begin{aligned}
&\psi(\Delta\lambda, \lambda^*)\\
&= \Phi_p(\xi^{\lambda^*}) - q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right)^p\right]^{1/p-1}\operatorname{tr}\left[\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{I}(\xi^{\Delta\lambda})\mathsf{I}(\xi^{\lambda})^{-1}\mathsf{B}\right]\\
&= \Phi_p(\xi^{\lambda^*})\\
&- \sum_{i}^{n}\Delta\lambda_i q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\\
&= \sum_{i=1}^{n}\Delta\lambda_i\left[\Phi_p(\xi^{\lambda^*})\right.\\
&\left.-q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\right] \geq 0,
\end{aligned}$$

for all feasible weight vectors $\Delta\lambda$.

Thus,

$$\Phi_p(\xi^{\lambda^*}) \geq q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i),$$

for $i = 1, \dots, n$.

Now we will show,

$$\Phi_p(\xi^{\lambda^*}) = q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i),$$

for all $\lambda_i^* > 0$.

Suppose there exists at least one $\boldsymbol{x}_j$ with $\lambda_j^* > 0$ such that

$$\Phi_p(\xi^{\lambda^*}) > q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_j)\boldsymbol{g}(\boldsymbol{x}_j)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_j).$$

Then, we have

$$\begin{aligned}
\Phi_p(\xi^{\lambda^*}) &= \sum_{i=1}^{n}\lambda_i^*\Phi_p(\xi^{\lambda^*})\\
&> \sum_{i=1}^{n}\lambda_i^* q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\\
&= q^{-1/p}\left[\operatorname{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1}\sum_{i=1}^{n}\lambda_i^* w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i)
\end{aligned}$$

$$= q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1}\mathrm{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p$$
$$= \Phi_p(\xi^{\lambda^*}),$$

which is a contradiction. So,

$$\Phi_p(\xi^{\lambda^*}) = q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^*})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^*})^{-1}\boldsymbol{g}(\boldsymbol{x}_i),$$

for design points $\boldsymbol{x}_i$ with $\lambda_i^* > 0$. ■

Proof of Corollary 1. This is a special case that $\mathsf{B} = q\mathsf{A} = q\int_\Omega \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x})\left[\frac{dh^{-1}}{d\eta}\right]^2 dF_{\mathrm{IMSE}}(\boldsymbol{x})$ and $p = 1$. Thus,

$$\begin{aligned}\psi(\Delta\lambda, \lambda) &= \mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{A}\right) - \mathrm{tr}\left[\mathsf{I}(\xi^\lambda)^{-1}\mathsf{I}(\xi^{\Delta\lambda})\mathsf{I}(\xi^\lambda)^{-1}\mathsf{A}\right]\\
&= \mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{A}\right) - \mathrm{tr}\left[\mathsf{I}(\xi^\lambda)^{-1}\left(\sum_i \Delta\lambda_i w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\right)\mathsf{I}(\xi^\lambda)^{-1}\mathsf{A}\right]\\
&= \mathrm{tr}\left(\mathsf{I}(\xi^\lambda)^{-1}\mathsf{A}\right) - \sum_{i=1}^n \Delta\lambda_i w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^\lambda)^{-1}\mathsf{A}\mathsf{I}(\xi^\lambda)^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\end{aligned}$$

■

Proof of Corollary 2. For (i), it directly follows the result in Theorem 1. For (ii), the proof follows exactly as the proof of Corollary 1. ■

Proof of Corollary 3. A positive definite A in EI-optimality would insure $\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B} \neq 0$ in Theorem 2 with $\mathsf{B} = q\mathsf{A}$ and $p = 1$. ■

Proof of Theorem 2. Since information matrix $\mathsf{I}(\xi^{\lambda^k})$ is positive definite, B is positive semidefinite, and $\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B} \neq 0$, we have

$$\begin{aligned}0 < \mathrm{tr}\left[\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right]^p &= \mathrm{tr}\left[\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left[\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right]^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{I}(\xi^{\lambda^k})\right]\\
&= \mathrm{tr}\left[\sum_{i=1}^n \lambda_i^k w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left[\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right]^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\right]\\
&= \sum_{i=1}^n \lambda_i^k \mathrm{tr}\left[w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left[\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right]^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\right]\\
&= \sum_{i=1}^n \lambda_i^k w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left[\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right]^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i).\end{aligned}$$

Thus, there exists some $\boldsymbol{x}_i$, such that

$$\lambda_i^k w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left[\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right]^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i) > 0,$$

and naturally, $\lambda_i^k \left(w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T \mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{A}\mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i) \right)^{\delta} > 0$, which leads to

$$\sum_{i=1}^{n} \lambda_i^k \left[w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T \mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p-1} \mathsf{I}(\xi^{\lambda^k})^{-1}\boldsymbol{g}(\boldsymbol{x}_i) \right]^{\delta} > 0.$$

■

Proof of Proposition 1. To prove Proposition 1, we first prove the following lemma.

Lemma 5. *Define* $\varphi(\mathsf{I}) = \phi_p(\xi) = \left(q^{-1}\mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \right]^p \right)^{1/p}$, $0 < p < \infty$ *as a function of Fisher information matrix* I*, then*

(a) $\varphi(\mathsf{I})$ *is a strictly convex function of* I.
(b) $\varphi(\mathsf{I})$ *is a decreasing function of* I.

Proof. The proof of (a) is very similar to that of Lemma 2, and is omitted here. For (b), for any $\mathsf{I}_1 \preceq \mathsf{I}_2$,

$$\mathsf{I}_1^{-1} \succeq \mathsf{I}_2^{-1} \Rightarrow \frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}_1^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \succeq \frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}_2^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T.$$

As $\mathrm{tr}(\mathsf{C}^r)$ with $r > 0$ is an increasing function of any positive definite matrix C (Bernstein, 2009),

$$\mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}_1^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \right]^p \geq \mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}_2^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \right]^p.$$

As a result,

$$\left(q^{-1}\mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}_1^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \right]^p \right)^{1/p} \geq \left(q^{-1}\mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}_2^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \right]^p \right)^{1/p},$$

i.e., $\varphi(\mathsf{I}_1) \geq \varphi(\mathsf{I}_2)$. ■

The proof of Proposition 1 is mainly based on the results in Theorems 1 and 2 in the paper by Yu (2010). Based on Lemma 5, it is known that

$$\varphi(\mathsf{I}) = \left(q^{-1}\mathrm{tr}\left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\mathsf{I}^{-1}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^T \right]^p \right)^{1/p}, \quad 0 < p < \infty$$

is a convex and decreasing function of I. Under Algorithm 1 with $0 < \delta < 1$, denote $\boldsymbol{\lambda}^k$ and $\boldsymbol{\lambda}^{k+1}$ to be the solutions on kth and $(k+1)$th iteration of the multiplicative algorithm, respectively. Since $\mathsf{I}(\xi^{\lambda^k})$, $\mathsf{I}(\xi^{\lambda^{k+1}})$, and B are positive definite, it is easy to see that

$$\mathsf{I}(\xi^{\lambda^k}) > 0, \quad \mathsf{I}(\xi^{\lambda^{k+1}}) > 0, \quad \text{and} \quad \mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1} \neq 0,$$

where $\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1}$ is a continuous function for I. Based on Theorem 1 in Yu (2010), we can have

$$\phi_p(\xi^{\lambda^{k+1}}) \le \phi_p(\xi^{\lambda^k}),$$

when $\lambda^{k+1} \neq \lambda^k$. Thus it shows the monotonicity of Algorithm 1.

Moreover, we also have that for $\varphi(\mathsf{I}) = \phi_p(\xi) = \left(q^{-1}\mathrm{tr}\left[\frac{\partial f(\beta)}{\partial \beta^T}\mathsf{I}^{-1}\left(\frac{\partial f(\beta)}{\partial \beta^T}\right)^T\right]^p\right)^{1/p}$,

$$w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}^T(\boldsymbol{x}_i)\frac{\partial \varphi(\mathsf{I})}{\partial \mathsf{I}}\bigg|_{\mathsf{I}=\mathsf{I}(\xi^{\lambda^k})} = w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}^T(\boldsymbol{x}_i)\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\left(\mathsf{I}(\xi^{\lambda^k})^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi^{\lambda^k})^{-1} \neq 0.$$

For the sequence $\mathsf{I}(\xi^{\lambda^k}), k = 1, 2, \ldots$ from Algorithm 1, its limit point is obviously nonsingular because of the positive definiteness. Combining the above statements with results in Lemma 5, all the required conditions in Theorem 2 of Yu (2010) are satisfied. Thus, using the results in Theorem 2 of Yu (2010), we have that all limit points of λ^k are global minimums of $\phi_p(\xi^\lambda)$, and the $\phi_p(\xi^\lambda)$ decreases monotonically to $\inf_\lambda \phi_p(\xi^\lambda)$ as $k \to \infty$.

Proof of Theorem 4. To prove Theorems 4 and 5, we first prove the following lemma.

Lemma 6. *For any design ξ and Φ_p-optimal design ξ^* that minimizes $\Phi_p(\xi)$, the following inequality holds:*

$$\min_{\boldsymbol{x}\in\Omega} \phi(\boldsymbol{x}, \xi) \le \phi(\xi^*, \xi) \le \Phi_p(\xi^*) - \Phi_p(\xi) \le 0,$$

where $\phi(\boldsymbol{x}, \xi)$ and $\phi(\xi^*, \xi)$ are the directional derivatives of $\Phi_p(\xi)$ in the direction of $\boldsymbol{x}$ and ξ^*, respectively.

Proof. The directional derivative of $\Phi_p(\xi)$ in the direction of the point $\boldsymbol{x}^* = \mathrm{argmin}_{\boldsymbol{x}\in\Omega}\phi(\boldsymbol{x}, \xi)$ is:

$$\begin{aligned}
&\min_{\boldsymbol{x}\in\Omega} \phi(\boldsymbol{x}, \xi) = \phi(\boldsymbol{x}^*, \xi)\\
&\qquad = \Phi_p(\xi) - q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}^*)\boldsymbol{g}(\boldsymbol{x}^*)^\top\mathsf{I}(\xi)^{-1}\mathsf{B}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x}^*)\\
&\qquad \le \Phi_p(\xi) - q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top\mathsf{I}(\xi)^{-1}\mathsf{B}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x})
\end{aligned} \tag{A5}$$

for any $\boldsymbol{x} \in \Omega$.

Denote the optimal design $\xi^* = \begin{Bmatrix} \boldsymbol{x}_1, & \ldots, & \boldsymbol{x}_n \\ \lambda_1^*, & \ldots, & \lambda_n^* \end{Bmatrix}$. With Inequality (A5), we have

$$\begin{aligned}
&\phi(\boldsymbol{x}^*, \xi)\\
&\le \sum_{i=1}^n \lambda_i^*\left(\Phi_p(\xi) - q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^\top\mathsf{I}(\xi)^{-1}\mathsf{B}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x}_i)\right)\\
&= \Phi_p(\xi) - \mathrm{tr}\left[\mathsf{I}(\xi^*)q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^p\right]^{1/p-1}\mathsf{I}(\xi)^{-1}\mathsf{B}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi)^{-1}\right]
\end{aligned}$$

$$= \phi(\xi^*, \xi). \tag{A6}$$

Furthermore, with the definition of directional derivative in the direction of the optimal design ξ^* and convexity of Φ_p, we have

$$\begin{aligned}
\phi(\xi^*, \xi) &= \lim_{\alpha \to 0} \frac{\Phi_p((1-\alpha)\xi + \alpha\xi^*) - \Phi_p(\xi)}{\alpha} \\
&\leq \lim_{\alpha \to 0} \frac{(1-\alpha)\Phi_p(\xi) + \alpha\Phi_p(\xi^*) - \Phi_p(\xi)}{\alpha} \\
&= \Phi_p(\xi^*) - \Phi_p(\xi)
\end{aligned} \tag{A7}$$

Combining inequality (A6) and (A7), we complete the proof that

$$\min_{\boldsymbol{x} \in \Omega} \phi(\boldsymbol{x}, \xi) \leq \phi(\xi^*, \xi) \leq \Phi_p(\xi^*) - \Phi_p(\xi) \leq 0.$$

■

In Equation (4), the reciprocal of Φ_p-optimality could be written as:

$$\frac{1}{\Phi_p(\xi)} = \left(q^{-1} \mathrm{tr} \left[\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \mathrm{I}(\xi)^{-1} \left(\frac{\partial \boldsymbol{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right)^T \right]^p \right)^{-1/p}, \quad 0 < p < \infty.$$

According to Pukelsheim (1993), Section 6.13, $\frac{1}{\Phi_p}$ is concave on the set of designs with positive definite information matrices. Using a similar approach to derive the lower bound of efficiency in (Atwood, 1969), with ξ and ξ^* fixed, define $t(\alpha) = \frac{1}{\Phi_p(\alpha\xi^* + (1-\alpha)\xi)}$. With simple algebra and directional derivative of Φ_p,

$$\begin{aligned}
\left.\frac{dt(\alpha)}{d\alpha}\right|_{\alpha=0} &= -\frac{1}{\Phi_p^2(\alpha\xi^* + (1-\alpha)\xi)} \left.\left(\frac{\partial \Phi_p(\alpha\xi^* + (1-\alpha)\xi)}{\partial \alpha} \right)\right|_{\alpha=0} \\
&= -\frac{1}{\Phi_p^2(\xi)} \phi(\xi^*, \xi),
\end{aligned}$$

where $\phi(\xi^*, \xi)$ is the directional derivative of Φ_p in the direction of ξ^* defined in Section 4.1.

According to Lemma 6 that $\phi(\boldsymbol{x}^*, \xi) \leq \phi(\xi^*, \xi) \leq \Phi_p(\xi^*) - \Phi_p(\xi) \leq 0$, we have

$$\begin{aligned}
&\frac{1}{\Phi_p(\xi^*)} - \frac{1}{\Phi_p(\xi)} = t(1) - t(0) \\
&\leq \left.\frac{dt(\alpha)}{d\alpha}\right|_{\alpha=\alpha^*}, \quad \text{where } \alpha^* \in (0,1) \text{ by the mean value theorem} \\
&\leq \left.\frac{dt(\alpha)}{d\alpha}\right|_{\alpha=0}, \quad \text{by concavity of } t \\
&= -\frac{1}{\Phi_p^2(\xi)} \phi(\xi^*, \xi)
\end{aligned}$$

$$\leq -\frac{1}{\Phi_p^2(\xi)}\phi(\boldsymbol{x}^*, \xi)$$

$$= \frac{\max_{\boldsymbol{x}\in C} q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi)^{-1}\mathsf{B}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x}_i) - \Phi_p(\xi)}{\Phi_p^2(\xi)}$$

That is, $\frac{\Phi_p(\xi)}{\Phi_p(\xi^*)} \leq \frac{\max_{\boldsymbol{x}\in C} q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi)^{-1}\mathsf{B}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x}_i)}{\Phi_p(\xi)}$, i.e.,

$$\mathrm{Eff}_{\Phi_p}(\xi|\xi^*) \geq \frac{\Phi_p(\xi)}{\max_{\boldsymbol{x}\in C} q^{-1/p}\left[\mathrm{tr}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^p\right]^{1/p-1} w(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)^T\mathsf{I}(\xi)^{-1}\mathsf{B}\left(\mathsf{I}(\xi)^{-1}\mathsf{B}\right)^{p-1}\mathsf{I}(\xi)^{-1}\boldsymbol{g}(\boldsymbol{x}_i)}.$$

Proof of Theorem 5. Here we prove the convergence that

$$\lim_{r\to\infty}\Phi_p(\xi^r) = \Phi_p(\xi^*)$$

of the proposed algorithm for a general Φ_p optimality, and that EI-optimality shares the same mathematical structure as Φ_1-optimality.

We will establish the argument by proof of contradiction with a similar proof as in the paper of Yang, Biedermann & Tang (2013). Assume that ξ^r does not converge to ξ^*, i.e.,

$$\lim_{r\to\infty}\Phi_p(\xi^r) - \Phi_p(\xi^*) > 0.$$

Since the support set of the rth iteration is a subset of the support set of the $(r+1)$th iteration, it is obvious that $\Phi_p(\xi^{r+1}) \leq \Phi_p(\xi^r)$ for all $r \geq 0$. Thus, there exists some $a > 0$, such that

$$\Phi_p(\xi^r) > \Phi_p(\xi^*) + a, \quad \text{for all } r.$$

According to Lemma 6, we can conclude $\phi(\boldsymbol{x}^*, \xi^r) \leq \phi(\xi^*, \xi^r) \leq \Phi_p(\xi^*) - \Phi_p(\xi^r) < -a$. It is quite obvious from the derivation of the directional derivative that $\Phi_p((1-\gamma)\xi + \gamma\xi')$ is infinitely differentiable with respect to $\gamma \in [0, 1]$. Thus, the second order derivative is bounded on γ, and denote the upper bound by U, with $U > 0$ as EI is a convex function.

Then, consider the design $\tilde{\xi}^{r+1} = (1-\gamma)\xi^r + \gamma\boldsymbol{x}_r^*$, with $\gamma \in [0, 1]$. Since the proposed algorithm achieves optimal weights in each iteration, we have,

$$\Phi_p(\xi^{r+1}) \leq \Phi_p(\tilde{\xi}^{r+1}).$$

Using the Taylor expansion of $\Phi_p(\tilde{\xi}^{r+1})$, we have

$$\begin{aligned}
\Phi_p(\xi^{r+1}) &\leq \Phi_p(\tilde{\xi}^{r+1}) \\
&= \Phi_p(\xi^r) + \gamma\phi(\boldsymbol{x}_r^*, \xi^r) + \frac{1}{2}\gamma^2 \left.\frac{\partial^2\Phi_p((1-\gamma)\xi^r + \gamma\boldsymbol{x}_r^*)}{\partial\gamma^2}\right|_{\gamma=\gamma'} \\
&< \Phi_p(\xi^r) - \gamma a + \frac{1}{2}\gamma^2 U,
\end{aligned}$$

where γ' is some value between 0 and 1. Consequently, we have

$$\Phi_p(\xi^{r+1}) - \Phi_p(\xi^r) < \frac{1}{2}U\left(\gamma - \frac{a}{U}\right)^2 - \frac{a^2}{2U}.$$

If $\frac{a}{U} \leq 1 \Leftrightarrow U \geq a$, choose $\gamma = \frac{a}{U}$, then we have

$$\Phi_p(\xi^{r+1}) - \Phi_p(\xi^r) < -\frac{a^2}{2U}.$$

If $\frac{a}{U} > 1 \Leftrightarrow U < a$, choose $\gamma = 1$, then we have

$$\Phi_p(\xi^{r+1}) - \Phi_p(\xi^r) < \frac{1}{2}U - a < 0.$$

Both situations will lead to $\lim_{r\to\infty} \Phi_p(\xi^r) = -\infty$, which contradicts with $\Phi_p(\xi^r) > 0$.

In summary, the assumption that ξ^r does not converge to ξ^* is not valid, and thus we prove that ξ^r converges to ξ^*. ■

Received 20 September 2019
Accepted 16 April 2020