Monotonic convergence of a general algorithm for computing optimal designs

Yaming Yu
Department of Statistics
University of California
Irvine, CA 92697, USA
yamingy@uci.edu

Abstract

Monotonic convergence is established for a general class of multiplicative algorithms introduced by Silvey et al. (1978) for computing optimal designs. A conjecture of Titterington (1978) is confirmed as a consequence. Optimal designs for logistic regression are used as an illustration.

Keywords: A-optimality; auxiliary variables; c-optimality; D-optimality; experimental design; generalized linear models; multiplicative algorithm.

1 A general class of algorithms

Optimal experimental design (approximate theory) is a well-developed area and we refer to Kiefer (1974), Silvey (1980), Pázman (1986), and Pukelsheim (1993) for general introduction and basic results. We consider computational aspects of optimal designs, focusing on a finite design space $\mathcal{X} = \{x_1, \ldots, x_n\}$. Suppose the probability density or mass function of the response is specified as $p(y|x, \theta)$, where $\theta = (\theta_1, \ldots, \theta_m)^\top$ is the parameter of interest. Let A_i denote the $m \times m$ expected Fisher information matrix from a unit assigned to x_i, with the (j, k) entry (the expectation is with respect to $p(y|x_i, \theta)$)

$$A_i(j, k) = E\left[\frac{\partial \log p(y|x_i, \theta)}{\partial \theta_j} \frac{\partial \log p(y|x_i, \theta)}{\partial \theta_k}\right].$$

The moment matrix, as a function of the design measure $w = (w_1, \ldots, w_n)$, is defined as

$$M(w) = \sum_{i=1}^{n} w_i A_i,$$

which is proportional to the Fisher information for θ when the number of units assigned to x_i is proportional to w_i. Here $w \in \bar{\Omega}$, and $\bar{\Omega}$ denotes the closure of $\Omega = \{w : \ w_i > 0, \ \sum_{i=1}^n w_i = 1\}$. Throughout we assume that A_i are well-defined and hence nonnegative definite. The set

$$\Omega_+ \equiv \{w \in \bar{\Omega} : \ M(w) > 0 \text{ (positive definite)}\}$$

is assumed nonempty. Our approach may conceivably extend to the case where $M(w)$ is allowed to be singular, by using generalized inverses, although we do not pursue this here.

Given an optimality criterion ϕ, defined on positive definite matrices, the goal is to maximize $\phi(M(w))$ with respect to $w \in \Omega_+$. Typical optimality criteria include

(i) the D-criterion $\phi_0(M) = \log\det(M)$,

(ii) the A-criterion $\phi_{-1}(M) = -tr(M^{-1})$,

(iii) more generally, the pth mean criterion $\phi_p(M) = -tr(M^p),\ p < 0$, and

(iv) the c-criterion $\phi_{-1,c}(M) = -c^\top M^{-1} c$, where c is a nonzero constant vector.

Often only a linear combination $K^\top\theta$, e.g., a subvector of θ, is of interest. The Fisher information for $K^\top\theta$ is naturally defined as $(K^\top M^{-1} K)^{-1}$, assuming invertibility (Pukelsheim, 1993). We may therefore consider the D- and A-criteria for $K^\top\theta$ defined respectively as

$$\begin{aligned} \phi_{0,K}(M) &= -\log\det(K^\top M^{-1} K); \\ \phi_{-1,K}(M) &= -tr(K^\top M^{-1} K). \end{aligned} \tag{1}$$

The c-criterion is a special case of $\phi_{-1,K}(M)$. Motivations for such optimality criteria are well-known. In a linear problem, the A-criterion seeks to minimize the sum of variances of the best linear unbiased estimators (BLUEs) for all coordinates of θ, while the c-criterion seeks to minimize the variance of the BLUE for $c^\top\theta$. Similar interpretations (with asymptotic arguments) apply to nonlinear problems.

In general $M(w)$ also depends on the unknown parameter θ, which complicates the definition of an optimality criterion. A simple solution is to maximize $\phi(M(w))$ with θ fixed at a prior guess θ^*; this leads to *local optimality* (Chernoff 1953). Local optimality may be criticized for ignoring uncertainty in θ. However, in a situation where real prior information is available, or where the dependence of M on θ is weak, it is nevertheless a viable approach, and has been adopted routinely (see, for example, Li and Majumdar 2008). Henceforth we assume a fixed θ^* and suppress the dependence of M on θ. Possible extensions are mentioned in Section 5.

Optimal designs do not usually come in closed form. As early as Wynn (1972), Fedorov (1972), Atwood (1973), and Wu and Wynn (1978), and as late as Torsney (2007), Harman and Pronzato (2007), and Dette et al. (2008), various procedures have been studied for numerical computation. We shall focus on the following multiplicative algorithm (Titterington 1976, 1978; Silvey et al. 1978), which is specified through a power parameter $\lambda \in (0, 1]$.

Algorithm I Set $\lambda \in (0, 1]$ and $w^{(0)} \in \Omega$. For $t = 0, 1, \ldots$, compute

$$w_i^{(t+1)} = w_i^{(t)} \frac{d_i^\lambda\left(w^{(t)}\right)}{\sum_{j=1}^n w_j^{(t)} d_j^\lambda\left(w^{(t)}\right)}, \quad i = 1, \ldots, n, \tag{2}$$

where

$$d_i(w) = tr(\phi'(M(w))A_i), \quad \phi'(M) \equiv \frac{\partial\phi(M)}{\partial M}.$$

Iterate until convergence.

For a heuristic explanation, observe that (2) is equivalent to

$$w_i^{(t+1)} \propto w_i^{(t)} \left(\left. \frac{\partial\phi(M(w))}{\partial w_i} \right|_{w=w^{(t)}} \right)^\lambda, \quad i = 1, \ldots, n. \tag{3}$$

The value of $\partial\phi(M(w))/\partial w_i$ indicates the amount of gain in information, as measured by ϕ, by a slight increase in w_i, the weight on the ith design point. So (3) can be seen as adjusting w so that relatively more weight is placed on design points whose increased weight may result in a larger gain in ϕ. If ϕ is increasing and concave, then a convenient convergence criterion, based on the general equivalence theorem (Kiefer and Wolfowitz, 1960; Whittle, 1973), is

$$\max_{1\leq i\leq n} d_i\left(w^{(t)}\right) \leq (1+\delta)\bar{d}\left(w^{(t)}\right), \tag{4}$$

where $\bar{d}(w) \equiv \sum_{i=1}^{n} w_i d_i(w)$ and δ is a small positive constant.

Algorithm I is remarkable in its generality. For example, little restriction is placed on the underlying model $p(y|x,\theta)$. Part of the reason, of course, is that we focus on Fisher information and local optimality, which essentially reduces the problem to a linear one.

There exists a large literature on Algorithm I and its relatives; see, for example, Titterington (1976, 1978), Silvey et al. (1978), Pázman (1986), Fellman (1989), Pukelsheim and Torsney (1991), Torsney and Mandal (2006), Harman and Pronzato (2007), Dette et al. (2008), and Torsney and Martín-Martín (2009). One feature that has attracted much attention is that Algorithm I appears to be monotonic, i.e., $\phi(M(w^{(t)}))$ increases in t, at least in some special cases. For example, when $\phi = \phi_0$ (for D-optimality) and $\lambda = 1$, Titterington (1976) and Pázman (1986) have shown monotonicity using clever probabilistic and analytic inequalities; see also Dette et al. (2008) and Harman and Trnovská (2009). Algorithm I is also known to be monotonic for $\phi = \phi_{-1,K}$ as in (1), assuming $\lambda = 1/2$ and A_i are rank-one (Fellman 1974; Torsney 1983). Monotonicity is important because convergence then holds under mild assumptions (see Section 4). Results in these special cases suggest a monotonic convergence theory for a broad class of ϕ, which is also supported by numerical evidence presented in some of the references above.

2 Main result

We aim to state general conditions on ϕ that ensure that Algorithm I converges monotonically. As a consequence certain known theoretical results are unified and generalized, and one particular conjecture (Titterington 1978) is confirmed. Define

$$\psi(M) \equiv -\phi(M^{-1}), \quad M > 0.$$

The functions ϕ and ψ are assumed to be differentiable on invertible matrices. Our conditions are conveniently stated in terms of ψ. As usual, for two symmetric matrices, $M_1 \leq (<) M_2$ means $M_2 - M_1$ is nonnegative (positive) definite.

- $\psi(M)$ is increasing:

$$0 < M_1 \leq M_2 \quad \Longrightarrow \quad \psi(M_1) \leq \psi(M_2), \tag{5}$$

  or, equivalently, $\psi'(M)$ is nonnegative definite for positive definite M.

- $\psi(M)$ is concave:

$$\alpha\psi(M_1) + (1-\alpha)\psi(M_2) \leq \psi(\alpha M_1 + (1-\alpha)M_2), \tag{6}$$

  for $\alpha \in [0,1]$, $M_1, M_2 > 0$. Equivalently,

$$\psi(M_2) \leq \psi(M_1) + tr(\psi'(M_1)(M_2 - M_1)), \quad M_1, M_2 > 0. \tag{7}$$

Condition (5) is usually satisfied by any reasonable information criterion (Pukelsheim 1993). Also note that, if (5) fails, then $\partial\phi(M(w))/\partial w_i$ on the right hand side of (3) is not even guaranteed to be nonnegative. The real restriction is the concavity condition (6). For example, (6) is not satisfied by $\psi_p(M) = -\phi_p(M^{-1})$ (the pth mean criterion) when $p < -1$. (It is usually assumed that $\phi(M)$, rather than $\psi(M)$, is concave.) Nevertheless, (6) *is* satisfied by a wide range of criteria, including the commonly used D-, A- or c-criteria (see Cases (i) and (ii) in the illustration of the main result below).

Our main result is as follows.

Theorem 1 (General monotonicity)**.** *Assume (5) and (6). Assume that in iteration (2), with* $0 < \lambda \leq 1$, *we have*

$$M(w^{(t)}) > 0, \quad \phi'(M(w^{(t)})) \neq 0, \quad \text{and} \quad M(w^{(t+1)}) > 0.$$

Then

$$\phi(M(w^{(t+1)})) \geq \phi(M(w^{(t)})).$$

In other words, under mild conditions which ensure that (2) is well-defined (specifically, the denominator in (2) is nonzero), (5) and (6) imply that (2) never decreases the criterion ϕ. Let us illustrate Theorem 1 with some examples. For simplicity, in (i)–(iv) we display formulae for $\lambda = 1$ only, although monotonicity holds for all $\lambda \in (0, 1]$.

(i) Take

$$\phi_p(M) = \begin{cases} \log\det M, & p = 0; \\ -tr(M^p), & p \in [-1, 0). \end{cases}$$

Then $\psi_p(M) \equiv -\phi_p(M^{-1})$ satisfies (5) and (6). By Theorem 1, Algorithm I is monotonic for $\phi = \phi_p,\ p \in [-1, 0]$. This generalizes the previously known cases $p = 0$ and $p = -1$ (with particular values of λ). The iteration (2) reads

$$w_i^{(t+1)} = w_i^{(t)} \frac{tr(M^{p-1}(w^{(t)})A_i)}{tr(M^p(w^{(t)}))}, \quad i = 1, \ldots, n.$$

(ii) More generally, given a full rank $m \times r$ matrix K $(r \leq m)$, consider

$$\psi_{p,K}(M^{-1}) \equiv -\phi_{p,K}(M) = \begin{cases} \log\det(K^\top M^{-1} K), & p = 0; \\ tr((K^\top M^{-1} K)^{-p}), & p \in [-1, 0). \end{cases}$$

Then $\psi_{p,K}(M)$ satisfies (5) and (6). By Theorem 1, Algorithm I is monotonic for $\phi = \phi_{p,K},\ p \in [-1, 0]$. The iteration (2) reads

$$w_i^{(t+1)} = w_i^{(t)} \left. \frac{tr(M^{-1}K(K^\top M^{-1}K)^{-p-1}K^\top M^{-1}A_i)}{tr\left((K^\top M^{-1}K)^{-p}\right)} \right|_{M=M(w^{(t)})}. \tag{8}$$

(iii) In particular, taking $r = 1,\ K = c$ (an $m \times 1$ vector) and $p = -1$ in Case (ii), we obtain that Algorithm I is monotonic for the c-criterion $\phi_{-1,c}$. The iteration (8) reduces to

$$w_i^{(t+1)} = w_i^{(t)} \frac{c^\top M^{-1}(w^{(t)})A_i M^{-1}(w^{(t)})c}{c^\top M^{-1}(w^{(t)})c}, \quad i = 1, \ldots, n.$$

As noted by a referee, with $p = -1$, the choice $\lambda = 1$ may lead to an oscillating behavior in the sense that $w^{(t)}$ alternates between two points at which $\phi_{-1,c}(M(w))$ takes the same value.

While this does not contradict Theorem 1, it suggests that other values of λ are more desirable for fast convergence. Following Fellman (1974) and Torsney (1983), a practical recommendation is $\lambda = 1/2$ in the $p = -1$ case.

(iv) Consider another example of Case (ii), with $p = 0$, $r = m - 1$ and $K = (0_r, I_r)^\top$. Henceforth 0_r denotes the $r \times 1$ vector of zeros, and I_r denotes the $r \times r$ identity matrix. Assume $A_i = x_i x_i^\top$, $x_i^\top = (1, z_i^\top)$ and z_i is $(m-1) \times 1$. This corresponds to a D-optimal design problem for $(\theta_2, \ldots, \theta_m)$ under the linear model

$$y|(x, \theta) \sim \mathrm{N}(x^\top \theta, \sigma^2), \quad x^\top = (1, z^\top),$$

where the parameter is $\theta = (\theta_1, \theta_2, \ldots, \theta_m)^\top$. That is, interest centers on all coefficients other than the intercept. Nevertheless, as far as the design measure w is concerned, the optimality criterion, $\phi_{0,K}(M)$, coincides with $\phi_0(M)$, i.e.,

$$-\log\det(K^\top M^{-1}(w) K) = \log\det M(w).$$

After some algebra, (8) reduces to

$$w_i^{(t+1)} = w_i^{(t)} \frac{(z_i - \bar{z})^\top M_c^{-1}(w^{(t)})(z_i - \bar{z})}{m-1}, \quad i = 1, \ldots, n, \tag{9}$$

where

$$\bar{z} = \sum_{i=1}^n w_i^{(t)} z_i; \quad M_c(w^{(t)}) = \sum_{i=1}^n w_i^{(t)} (z_i - \bar{z})(z_i - \bar{z})^\top.$$

Thus (9) satisfies $\det M(w^{(t+1)}) \geq \det M(w^{(t)})$.

Monotonicity of (9) has been conjectured since Titterington (1978), and considerable numerical evidence has accumulated over the years. Recently, extending the arguments of Pázman (1986), Dette et al. (2008) have obtained results which come very close to resolving Titterington's conjecture. Nevertheless, we have been unable to extend their arguments further. Instead we prove the general Theorem 1 using a different approach, and settle this conjecture as a consequence.

The proof of Theorem 1 is achieved by using a method of *auxiliary variables*. When a function $f(w)$ (e.g., $-\det M(w)$) to be minimized is complicated, we introduce a new variable Q and a function $g(w, Q)$ such that $\min_Q g(w, Q) = f(w)$ for all w, thus transforming the problem into minimizing $g(w, Q)$ over w and Q jointly. Then we may use an iterative conditional minimization strategy on $g(w, Q)$. This is inspired by the EM algorithm (Dempster et al. 1977; Meng and van Dyk 1997); in particular, see Csiszár and Tusnady's (1984) interpretation (see Yu (2008) for a related interpretation of the data augmentation algorithm).

In Section 3 we analyze Algorithm I using this strategy. Although attention is paid to the mathematics, our focus is on intuitively appealing interpretations, which may lead to further extensions of Algorithm I with the same desirable monotonicity properties. If the algorithm is monotonic, then convergence can be established under mild conditions (Section 4). Section 5 contains an illustration with optimal designs for a simple logistic regression model.

3 Explaining the monotonicity

A key observation is that the problem of maximizing $\phi(M(w))$, or, equivalently, minimizing $\psi(M^{-1}(w))$ can be formulated as a joint minimization over both the design and the estimator.

Specifically, let us compare the original Problem P1 with its companion P2. Throughout $A^{1/2}$ denotes the symmetric nonnegative definite (SNND) square root of an SNND matrix A.

Problem P1: Minimize $-\phi(M(w)) \equiv \psi((\sum_{i=1}^n w_i A_i)^{-1})$ over $w \in \Omega$.

Problem P2: Minimize

$$g(w, Q) \equiv \psi\left(Q\Delta_w Q^\top\right) \tag{10}$$

over $w \in \Omega$ and Q (an $m \times (mn)$ matrix), subject to $QG = I_m$, where

$$\Delta_w \equiv \mathrm{Diag}(w_1^{-1}, \ldots, w_n^{-1}) \otimes I_m; \quad G \equiv (A_1^{1/2}, \ldots, A_n^{1/2})^\top.$$

Though not immediately obvious, P1 and P2 are equivalent, and this may be explained in statistical terms as follows. In (10), $Q\Delta_w Q^\top$ is simply the variance matrix of a linear unbiased estimator, QY, of the $m \times 1$ parameter θ in the model

$$Y = G\theta + \epsilon, \quad \epsilon \sim \mathrm{N}(0,\, \Delta_w),$$

where Y is the $(mn) \times 1$ vector of observations. The constraint $QG = I_m$ ensures unbiasedness. (Note that G is full-rank since $M(w)$ is nonsingular by assumption.) Of course, the weighted least squares (WLS) estimator is the best linear unbiased estimator, having the smallest variance matrix (in the sense of positive definite ordering) and, by (5), the smallest ψ for that matrix. It follows that, for fixed w, $g(w, Q)$ is minimized by choosing QY as the WLS estimator:

$$g(w, \hat{Q}_{WLS}) = \inf_{QG=I_m} g(w, Q), \tag{11}$$

$$\hat{Q}_{WLS} = M^{-1}(w)\left(w_1 A_1^{1/2}, \ldots, w_n A_n^{1/2}\right). \tag{12}$$

However, from (10) and (12) we get

$$g(w, \hat{Q}_{WLS}) = \psi(M^{-1}(w)). \tag{13}$$

That is, P2 reduces to P1 upon minimizing over Q.

Since P2 is not immediately solvable, it is natural to consider the subproblems: (i) minimizing $g(w, Q)$ over Q for fixed w, and (ii) minimizing $g(w, Q)$ over w for fixed Q. Part (ii) is again formulated as a joint minimization problem. For a fixed $m \times (mn)$ matrix Q such that $QG = I_m$, let us consider Problems P3 and P4.

Problem P3: Minimize $g(w, Q)$ as in (10) over $w \in \Omega$.

Problem P4: Minimize the function

$$h(\Sigma, w, Q) = \psi(\Sigma) + tr\left(\psi'(\Sigma)\left(Q\Delta_w Q^\top - \Sigma\right)\right) \tag{14}$$

over $w \in \Omega$ and the $m \times m$ positive-definite matrix Σ.

The concavity assumption (7) implies that

$$h(\Sigma, w, Q) \geq \psi\left(Q\Delta_w Q^\top\right) \tag{15}$$

with equality when $\Sigma = Q\Delta_w Q^\top$, i.e., Problem P4 reduces to P3 upon minimizing over Σ.

Since P4 is not immediately solvable, it is natural to consider the subproblems: (i) minimizing $h(\Sigma, w, Q)$ over Σ for fixed w and Q, and (ii) minimizing $h(\Sigma, w, Q)$ over w for fixed Σ and Q. Part (ii), which amounts to minimizing

$$tr\left(\psi'(\Sigma) Q\Delta_w Q^\top\right) = tr\left(Q^\top \psi'(\Sigma) Q\Delta_w\right),$$

admits a closed-form solution: if we write $Q = (Q_1, \dots, Q_n)$ where each Q_i is $m \times m$, then w_i^2 should be proportional to $tr(Q_i^\top \psi'(\Sigma) Q_i)$. But algorithm I may not perform an exact minimization here; see (16).

Based on the above discussion, we can express Algorithm I as an iterative conditional minimization algorithm involving w, Q and Σ. At iteration t, define

$$\begin{aligned} Q^{(t)} &= (Q_1^{(t)}, \dots, Q_n^{(t)}); \\ Q_i^{(t)} &= w_i^{(t)} M^{-1}(w^{(t)}) A_i^{1/2}, \quad i = 1, \dots, n; \\ \Sigma^{(t)} &= Q^{(t)} \Delta_{w^{(t)}} Q^{(t)\top} = M^{-1}(w^{(t)}). \end{aligned}$$

Then we have

$$\begin{aligned} \psi(M^{-1}(w^{(t)})) &= g(w^{(t)}, Q^{(t)}) && \text{(by (13))} \\ &= h(\Sigma^{(t)}, w^{(t)}, Q^{(t)}) && \text{(by (14))} \\ &\geq h(\Sigma^{(t)}, w^{(t+1)}, Q^{(t)}) && \text{(see below)} \quad (16) \\ &\geq g(w^{(t+1)}, Q^{(t)}) && \text{(by (15), (10))} \quad (17) \\ &\geq \psi(M^{-1}(w^{(t+1)})) && \text{(by (11), (13)).} \quad (18) \end{aligned}$$

The choice of $w^{(t+1)}$ leads to (16) as follows. After simple algebra, the iteration (2) becomes

$$w_i^{(t+1)} = \frac{r_i^\lambda w_i^{1-2\lambda}}{\sum_{j=1}^n r_j^\lambda w_j^{1-2\lambda}}, \quad i = 1, \dots, n,$$

where

$$w_i \equiv w_i^{(t)}, \quad r_i \equiv tr\left(Q_i^{(t)\top} \psi'(\Sigma^{(t)}) Q_i^{(t)}\right).$$

Since $0 < \lambda \leq 1$, Jensen's inequality yields

$$\left(\sum_{i=1}^n \frac{r_i}{w_i}\right)^{1-\lambda} \geq \sum_{i=1}^n w_i \left(\frac{r_i}{w_i^2}\right)^{1-\lambda}; \quad \left(\sum_{i=1}^n \frac{r_i}{w_i}\right)^{\lambda} \geq \sum_{i=1}^n w_i \left(\frac{r_i}{w_i^2}\right)^{\lambda}.$$

That is,

$$\sum_{i=1}^n \frac{r_i}{w_i} \geq \left(\sum_{i=1}^n r_i^{1-\lambda} w_i^{2\lambda-1}\right)\left(\sum_{i=1}^n r_i^\lambda w_i^{1-2\lambda}\right).$$

Hence

$$\begin{aligned} tr\left(\psi'(\Sigma^{(t)}) Q^{(t)} \Delta_{w^{(t)}} Q^{(t)\top}\right) &= \sum_{i=1}^n \frac{r_i}{w_i^{(t)}} \\ &\geq \sum_{i=1}^n \frac{r_i}{w_i^{(t+1)}} = tr\left(\psi'(\Sigma^{(t)}) Q^{(t)} \Delta_{w^{(t+1)}} Q^{(t)\top}\right), \end{aligned}$$

which produces (16). Choosing $\lambda = 1/2$, i.e., $w_i^{(t+1)} \propto \sqrt{r_i}$, leads to exact minimization in (16); choosing $\lambda = 1$ yields equality in (16). But any choice of $w^{(t+1)}$ that decreases $h(\Sigma^{(t)}, w, Q^{(t)})$ at (16) would have resulted in the desired inequality

$$\psi(M^{-1}(w^{(t)})) \geq \psi(M^{-1}(w^{(t+1)})).$$

We may allow λ to change from iteration to iteration, and monotonicity still holds, as long as $\lambda \in (0, 1]$. See Silvey et al. (1978) and Fellman (1989) for investigations concerning the choice of λ. Also note that we assume $w_i^{(t)}, w_i^{(t+1)} > 0$ for all i. This is not essential, however, because (i) the possibility of $w_i^{(t)} = 0$ can be handled by restricting our analysis to all design points i such that $w_i^{(t)} > 0$, and (ii) the possibility of $w_i^{(t+1)} = 0$ can be handled by a standard limiting argument. Monotonicity holds as long as $M(w^{(t)})$ and $M(w^{(t+1)})$ are both positive definite, as noted in the statement of Theorem 1.

4 Global convergence

Monotonicity (Theorem 1) plays an important role in the following convergence theorem.

Theorem 2 (Global convergence). *Denote the mapping (2) by T.*

(a) Assume

$$\phi'(M(w)) \geq 0; \quad \phi'(M(w))A_i \neq 0, \quad w \in \Omega_+,\ i = 1, \ldots, n.$$

(b) Assume (2) is strictly monotonic, i.e.,

$$w \in \Omega_+,\ Tw \neq w \quad \Longrightarrow \quad \phi(M(Tw)) > \phi(M(w)). \tag{19}$$

(c) Assume ϕ is strictly concave and ϕ' is continuous on positive definite matrices.

(d) Assume that, if M (a positive definite matrix) tends to M^* such that $\phi(M)$ increases monotonically, then M^* is nonsingular.

Let $w^{(t)}$ be generated by (2) with $w_i^{(0)} > 0$ for all i. Then

(i) all limit points of $w^{(t)}$ are global maxima of $\phi(M(w))$ on Ω_+, and

(ii) as $t \to \infty$, $\phi(M(w^{(t)}))$ increases monotonically to $\sup_{w \in \Omega_+} \phi(M(w))$.

The proof of Theorem 2 is somewhat subtle. Standard arguments show that all limit points of $w^{(t)}$ are fixed points of the mapping T. This alone does not imply convergence to a global maximum, however, because there often exist sub-optimal fixed points on the boundary of Ω. (Global maxima occur routinely on the boundary also.) Our goal is therefore to rule out possible convergence to such sub-optimal points; details of the proof are presented in Yu (2009), an extended version of this paper. We shall comment on Conditions (a)–(d).

Condition (a) ensures that starting with $w^{(0)} \in \Omega_+$, all iterations are well-defined. Moreover, if $w_i^{(0)} > 0$ for all i, then $w_i^{(t)} > 0$ for all t and i. This highlights the basic idea that, in order to converge to a global maximum w^*, the starting value $w^{(0)}$ must assign positive weight to every support point of w^*. Such a requirement is not necessary for monotonicity. On the other hand, assigning weight to non-supporting points of w^* tends to slow the algorithm down. Hence methods that quickly eliminate non-optimal support points are valuable (Harman and Pronzato, 2007).

Condition (b) simply says that unless w is a fixed point, the mapping T should produce a better solution. Let us assume (5), (7) and Condition (a), so that Theorem 1 applies. Then, by checking the equality condition in (16), it is easy to see that Condition (b) is satisfied if $0 < \lambda < 1$. (The argument leading to (19) technically assumes that all coordinates of w are nonzero, but we can apply it to the appropriate subvector of w.) If $\lambda = 1$, then (16) reduces to an equality. However, by checking the equality conditions in (17) and (18), we can show that Condition (b) is satisfied if ψ is strictly increasing and strictly concave:

$$M_2 \geq M_1 > 0,\ M_1 \neq M_2 \implies \psi(M_1) < \psi(M_2); \tag{20}$$
$$M_1, M_2 > 0,\ M_1 \neq M_2 \implies \psi(M_2) < \psi(M_1) + tr(\psi'(M_1)(M_2 - M_1)). \tag{21}$$

Conditions (c) and (d) are technical requirements that concern ϕ alone. Condition (c) ensures uniqueness of the optimal moment matrix, which simplifies the analysis. Condition (d) ensures that positive definiteness of $M(w)$ is maintained in the limit. Conditions (c) and (d) are satisfied by $\phi = \phi_p$ with $p \leq 0$, for example.

Let us mention a typical example of Theorem 2.

Corollary 1. *Assume* $A_i \neq 0,\ w_i^{(0)} > 0,\ i = 1, \ldots, n,$ *and* $M(w^{(0)}) > 0$. *Then the conclusion of Theorem 2 holds for Algorithm I with* $\phi = \phi_0$.

Proof. Conditions (a), (c) and (d) are readily verified. Condition (b) is satisfied by (20) and (21). The claim follows from Theorem 2. □

When (20) or (21) fails, and $\lambda = 1$, it is often difficult to appeal to Theorem 2 because strict monotonicity (Condition (b)) may not hold. We illustrate this with an example where the monotonicity is not strict, and the algorithm does not converge (see also the remark in Case (iii) following Theorem 1). Consider iteration (9) ($\lambda = 1$) with $n = m = 2$ and design space $\mathcal{X} = \{x_i = (1, z_i)^\top,\ i = 1, 2\},\ z_1 = -z_2 = 1$. It is easy to show that, for any $w^{(t)} = (w_1, w_2) \in \Omega$, iteration (9) maps $w^{(t)}$ to $w^{(t+1)} = (w_2, w_1)$. Thus, unless $w_1 = w_2 = 1/2$ to begin with, the algorithm alternates between two distinct points. This appears to be a rare example, as (9) usually converges in practical situations.

5 Further remarks and illustrations

One can think of several reasons for the wide interest in Algorithm I and its relatives. Similar to the EM algorithm, Algorithm I is simple, easy to implement, and monotonically convergent for a large class of optimality criteria (although this was not proved in the present generality). Algorithm I is known to be slow sometimes. But it serves as a foundation upon which more effective variants can be built (see, e.g., Harman and Pronzato 2007, and Dette et al. 2008). While solving the conjectured monotonicity of (9) holds mathematical interest, our main contribution is a way of interpreting such algorithms as optimization on augmented spaces. This opens up new possibilities in constructing algorithms with the same desirable monotonic convergence properties.

As a numerical example, consider the logistic regression model

$$p(y|x, \theta) = \frac{\exp(yx^\top\theta)}{1 + \exp(x^\top\theta)}, \quad y = 0, 1.$$

The expected Fisher information for θ from a unit assigned to x_i is

$$A_i = x_i \frac{\exp(x_i^\top\theta)}{(1 + \exp(x_i^\top\theta))^2} x_i^\top.$$

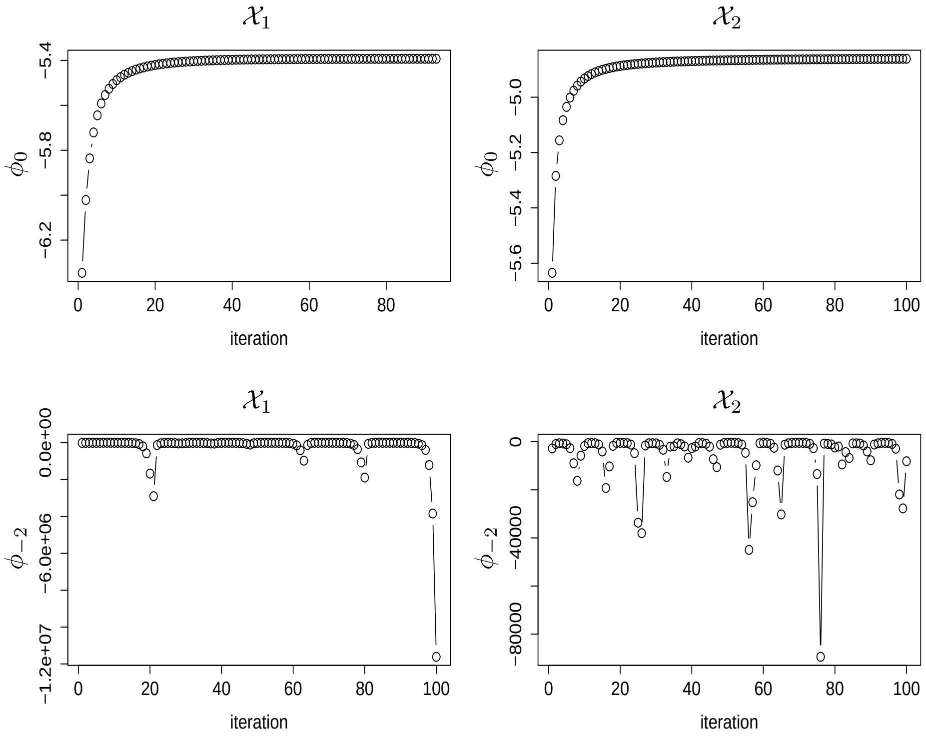

Figure 1: Values of $\phi_0 = \log\det M$ and $\phi_{-2} = -tr(M^{-2})$ for Algorithm I with design spaces $\mathcal{X}_1$ and $\mathcal{X}_2$.

We compute locally optimal designs with prior guess $\theta^* = (1,1)^\top$ $(m = 2)$, and design spaces

$$\mathcal{X}_1 = \{x_i = (1, i/20)^\top,\ i = 1, \dots, 20\};$$
$$\mathcal{X}_2 = \{x_i = (1, i/10)^\top,\ i = 1, \dots, 30\}.$$

The design criteria considered are ϕ_0 (for D-optimality) and ϕ_{-2}. We use Algorithm I with $\lambda = 1$, starting with equally weighted designs.

For ϕ_0, Corollary 1 guarantees monotonic convergence. This is illustrated by Figure 1, the first row, where $\phi_0 = \log\det M(w)$ is plotted against iteration t. Using the convergence criterion (4) with $\delta = 0.0001$, the number of iterations until convergence is 93 for $\mathcal{X}_1$ and 2121 for $\mathcal{X}_2$. The actual locally D-optimal designs are $w_1 = w_{20} = 0.5$ for $\mathcal{X}_1$ and $w_1 = w_{23} = 0.5$ for $\mathcal{X}_2$, as can be verified using the general equivalence theorem. This simple example serves to illustrate both the monotonicity of Algorithm I (when Theorem 1 applies) and its potential slow convergence.

For ϕ_{-2}, although Algorithm I can be implemented just as easily, Theorem 1 does not apply, because the concavity condition (7) no longer holds. Indeed, Algorithm I (with $\lambda = 1$) is not monotonic, as is evident from Figure 1, the second row, where $\phi_{-2} = -tr(M^{-2}(w))$ is plotted against iteration t. This shows the potential danger of using Algorithm I when monotonicity is not guaranteed.

Although Theorem 1 does not cover the ϕ_p criterion for $p < -1$, it is still possible that monotonicity holds for a smaller range of λ. Calculations in special cases lead to the conjecture

(Silvey et al. 1978) that Algorithm I is monotonic if $0 < \lambda \leq 1/(1-p)$. Theorem 1 provides further evidence for this conjecture, but new insights are needed to resolve it.

We have focused on local optimality. An alternative, *Bayesian optimality* (Chaloner and Larntz, 1989; Chaloner and Verdinelli, 1995), seeks to maximize the expected value of $\phi(M(\theta; w))$ over a prior distribution $\pi(\theta)$. The notation $M(\theta; w)$ emphasizes the dependence of the moment matrix on the parameter θ. It would be worthwhile to extend our strategy in Section 3 to Bayesian optimality, and we plan to report both theoretical and empirical evaluations of such extensions in future works.

Acknowledgement

The author would like to thank Don Rubin, Xiao-Li Meng, and David van Dyk for introducing him to the field of statistical computing. He is also grateful to Mike Titterington, Ben Torsney, and the referees for their valuable comments.

References

[1] C.L. Atwood, Sequences converging to D-optimal designs of experiments, Ann. Statist. 1 (1973) pp. 342–352.

[2] K. Chaloner and K. Larntz, Optimal Bayesian design applied to logistic regression experiments, J. Statist. Plann. Inference 21 (1989) pp. 191-208.

[3] K. Chaloner and I. Verdinelli, Bayesian experimental design: a review, Statist. Sci. 10 (1995) pp. 273-304.

[4] H. Chernoff, Locally optimal design for estimating parameters, Ann. Math. Statist. 24 (1953) pp. 586-602.

[5] I. Csiszár and G. Tusnady, Information geometry and alternating minimization procedures, Statistics & Decisions, Supplement Issue 1 (1984) pp. 205–237.

[6] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. Roy. Statist. Soc. B 39 (1977) pp. 1–38.

[7] H. Dette, A. Pepelyshev and A. Zhigljavsky, Improving updating rules in multiplicative algorithms for computing D-optimal designs, Computational Statistics & Data Analysis 53 (2008) pp. 312–320.

[8] V.V. Fedorov, Theory of Optimal Experiments (1972) Academic Press, New York.

[9] J. Fellman, On the allocation of linear observations (Thesis), Comment. Phys.-Math. 44 (1974) pp. 27–78.

[10] J. Fellman, An empirical study of a class of iterative searches for optimal designs, J. Statist. Planning Infer. 21 (1989) pp. 85–92.

[11] R. Harman and L. Pronzato, Improvements on removing nonoptimal support points in D-optimum design algorithms, Statist. Probab. Lett. 77 (2007) pp. 90-94.

[12] R. Harman and M. Trnovská, Approximate D-optimal designs of experiments on the convex hull of a finite set of information matrices, Mathematica Slovaca 59 (2009) pp. 693-704.

[13] J. Kiefer, General equivalence theory for optimum designs (approximate theory), Ann. Statist. 2 (1974) 849–879.

[14] J. Kiefer and J. Wolfowitz, The equivalence of two extremum problems, Canad. J. Math. 12 (1960) pp. 363–366.

[15] G. Li and D. Majumdar, D-optimal designs for logistic models with three and four parameters, Journal of Statistical Planning and Inference 138 (2008) pp. 1950–1959.

[16] X.-L. Meng and D. van Dyk, The EM algorithm – an old folk-song sung to a fast new tune (with discussion), J. Roy. Statist. Soc. B 59 (1997) pp. 511–567.

[17] A. Pázman, Foundations of Optimum Experimental Design, Reidel, Dordrecht (1986).

[18] F. Pukelsheim, Optimal Design of Experiments, John Wiley & Sons Inc, New York (1993).

[19] F. Pukelsheim and B. Torsney, Optimal weights for experimental designs on linearly independent support points, Ann. Statist. 19 (1991) pp. 1614-1625.

[20] S.D. Silvey, Optimal Design, Chapman & Hall, London (1980).

[21] S.D. Silvey, D.M. Titterington and B. Torsney, An algorithm for optimal designs on a finite design space, Commun. Stat. Theory Methods 14 (1978) pp. 1379-1389.

[22] D.M. Titterington, Algorithms for computing D-optimal design on finite design spaces. In Proc. of the 1976 Conf. on Information Science and Systems, John Hopkins University, 3 (1976) pp. 213-216.

[23] D.M. Titterington, Estimation of correlation coefficients by ellipsoidal trimming, Appl. Stat. 27 (1978) pp. 227-234.

[24] B. Torsney, A moment inequality and monotonicity of an algorithm. In: Kortanek, K.O. and Fiacco, A.V. (Eds.), Proceedings of the International Symposium on Semi-Infinite Programming and Appl., Lecture Notes in Economics and Mathematical Systems, 215. University of Texas at Austin (1983) pp. 249–260.

[25] B. Torsney, W-iterations and ripples therefrom, In: Pronzato, L., Zhigljavsky, A. (Eds.), Optimal Design and Related Areas in Optimization and Statistics, Springer-Verlag, N.Y (2007) pp. 1–12.

[26] B. Torsney and S. Mandal, Two classes of multiplicative algorithms for constructing optimizing distributions, Computational Statistics & Data Analysis 51 (2006) pp. 1591–1601.

[27] B. Torsney and R. Martín-Martín, Multiplicative algorithms for computing optimum designs, Journal of Statistical Planning and Inference 139 (2009) pp. 3947–3961.

[28] P. Whittle, Some general points in the theory of optimal experimental design. J. R. Statist. Soc. B 35 (1973) 123–130.

[29] C.F. Wu and H.P. Wynn, The convergence of general step-length algorithms for regular optimum design criteria, Annals of Statistics 6 (1978) 1273–1285.

[30] H.P. Wynn, Results in the theory and construction of D-optimum experimental designs, J. Roy. Statist. Soc. Ser. B 34 (1972) pp. 133-147.

[31] Y. Yu, A bit of information theory, and the data augmentation algorithm converges, IEEE Trans. Inform. Theory 54 (2008) pp. 5186–5188.

[32] Y. Yu, Monotonic convergence of a general algorithm for computing optimal designs (extended version), Technical Report (2009) arXiv:0905.2646