Sequential Experimental Designs for Generalized Linear Models

Technical Report RP-SOR-0607

Hovav A. Dror* and David M. Steinberg**

Department of Statistics and Operations Research

Raymond and Beverly Sackler Faculty of Exact Sciences

Tel Aviv University

Ramat Aviv 69978

Israel

Email: *hovavdror@gmail.com **dms@post.tau.ac.il

July 2006

Abstract: We consider the problem of experimental design when the response is modeled by a generalized linear model (GLM) and the experimental plan can be determined sequentially. Most prior research on this problem has been limited to the case of one-factor, binary response experiments, which are encountered in dose-response studies and sensitivity testing. We suggest a new procedure for the sequential choice of observations that improves on existing methods in four important ways: (1) it can be applied to multi-factor experiments and is not limited to the one-factor setting; (2) it can be used with any GLM, not just binary responses; (3) both fully sequential and group sequential settings are treated; and (4) the experimenter is not constrained to specify a single model and can use the prior to reflect uncertainty as to the link function and the form of the linear predictor. Our procedure is based on a D-optimality criterion, and on a Bayesian analysis that exploits a discretization of the parameter space to efficiently represent the posterior distribution. In the one-factor setting, a simulation study shows that our method is superior in efficiency to commonly used procedures, such as the"Bruceton" test (Dixon and Mood, 1948), the Langlie (1965) test or Neyer's (1994) procedure. We also present a comparison of results obtained with the new algorithm versus the "Bruceton" method on an actual sensitivity test conducted recently at an industrial plant.

KEY WORDS: Dose-response; Sensitivity tests; Binary Response; Poisson; D-optimal; Bayesian

1. INTRODUCTION

Efficient experimental designs for generalized linear models (GLM's) depend on the unknown coefficients, so two experiments having the same model but different coefficient values will typically require different designs. Recently, Dror and Steinberg (in press) suggested a method to construct robust D-optimal experimental designs for generalized linear models, which is based on clustering local D-optimal designs.

This paper extends our ideas to sequential designs. These can be fully sequential plans, in which the experimental plan is revised after each observation; or, they can be group sequential in which the experimental plan is revised after each batch of k observations. As in Dror and Steinberg (in press) we consider models with multivariate explanatory variables and allow the prior distribution to describe uncertainty over possible coefficient values, and also ambiguity of the proper linear predictor - enabling the design to assist in determining the necessity of certain interactions, or between higher and lower order models.

We limit our discussion to parametric models and to the estimation of a set of coefficients through a D-optimality criterion. Note that for multivariate models common optimality criteria that are limited to the estimation of a single percentile (see, for example Wu, 1985) are of lesser relevance, as usually there does not exist a single point, and sometimes not even a continuous surface, where the probability of response is fixed.

Source code for the algorithms and examples throughout this paper is available at http://www.math.tau.ac.il/~dms/GLM_Design.

2. PRIOR WORK AND ITS LIMITATIONS

Most work on sequential design for generalized linear models has focused on the rather simple case of fully sequential designs for a one-factor experiment with a binary response, also known as "sensitivity tests" or "dose-response" studies. Typical applications are experiments aimed at learning about the sensitivity of a new explosive, as a function of the strength of a shock

(possibly tested through dropping explosives from different heights); or, the toxicity of a drug administered at different doses.

A common example for a univariate model with a binary response is a logistic function, such that the expected response given a dose level x is $p(x;\mu,\theta) = 1/[1+\exp\{-\theta(x-\mu)\}]$, with (μ,θ) unknowns. Chaloner and Larntz (1989) used an example of this nature where the uncertainty over the coefficient values was described by a uniform distribution. We use a small variation of their example where $\mu \sim U[-1,1]$ and $\theta \sim U[6,18]$. Figure 1 presents the expected response $p(x;\mu,\theta)$ as a function of x for a sample of possible (μ,θ) values from the specified prior.

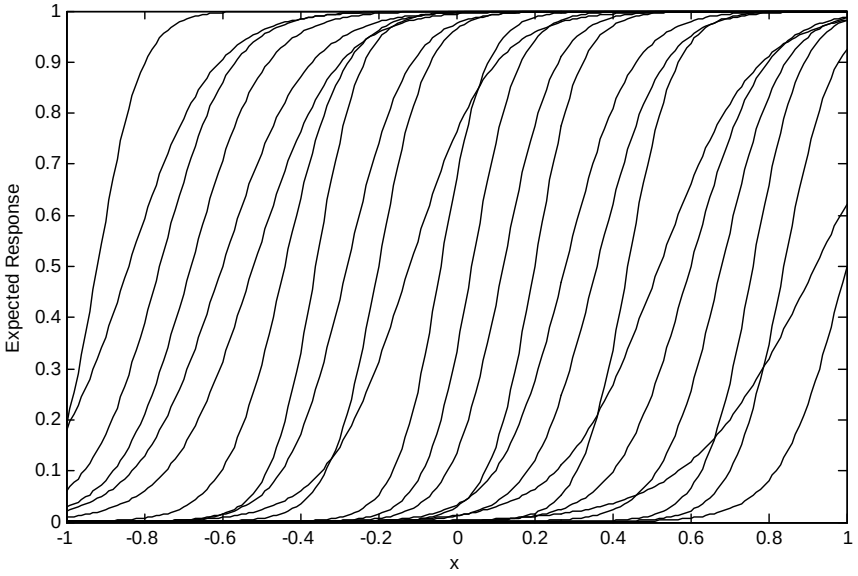

Figure 1: A sample of possible logistic probability curves
in an example based on Chaloner and Larntz (1989).

Most existing design algorithms for GLM's are based on a guess of the parameter values. In the given example the centroid of the parameter space, $\mu = 0,\ \theta = 12$, is a good starting point. The most commonly used procedure for sensitivity tests is the "Bruceton" up and down method (Dixon and Mood, 1948). This algorithm uses a guess of the mean μ, and a step size based on a guess for the dispersion $\sigma = 1/\theta$ (to be exact, the Bruceton method is defined for a probit model, and the step size should be chosen accordingly). A successor to the Bruceton method was suggested by Langlie (1965) and is based on estimates of lower and upper limits for μ. These procedures may give reasonable results when the estimates are good, but in the

common case where the prior estimates suffer from non-negligible uncertainty, they waste a lot of experimental effort. Furthermore, assuming we would like to estimate both parameters, it seems inviting to base the design on the D-optimality criterion, which minimizes the confidence ellipsoid for the unknowns.

Abdelbasit and Plackett (1983) discussed sequential designs for one-factor models with a binary response and suggested beginning the sequential design by observing a full local D-optimal design for the centroid (in the example above the local D-optimal design for the centroid places equal weight on two points: $x = \pm 0.13$), so that both unknown parameters can be estimated from the data; after completing two observations they advised repeating the process with the new, data-based, estimates.

The method of Abdelbasit and Plackett (1983) has several disadvantages. Assume, for example, that we have performed the first observation at $x = 0.13$, and that we observed a "no response". Although it is true that one observation is not sufficient to compute maximum likelihood estimates (MLE) from the data alone (one observation and two parameters), the experimenter should be aware that even a single observation may contribute much information, which can be usefully reflected by an immediate change in the experimental plan. A large portion of the parameter space can be "ruled out" by this single observation, as many of the prior models have an expected response that is very close to 1 at $x = 0.13$. The dashed lines in Figure 2 represent such models, from the sample of models presented in Figure 1. Clearly, using the second half of the local D-optimal design, $x = -0.13$, as suggested by Abdelbasit and Plackett (1983), cannot be a very efficient choice.

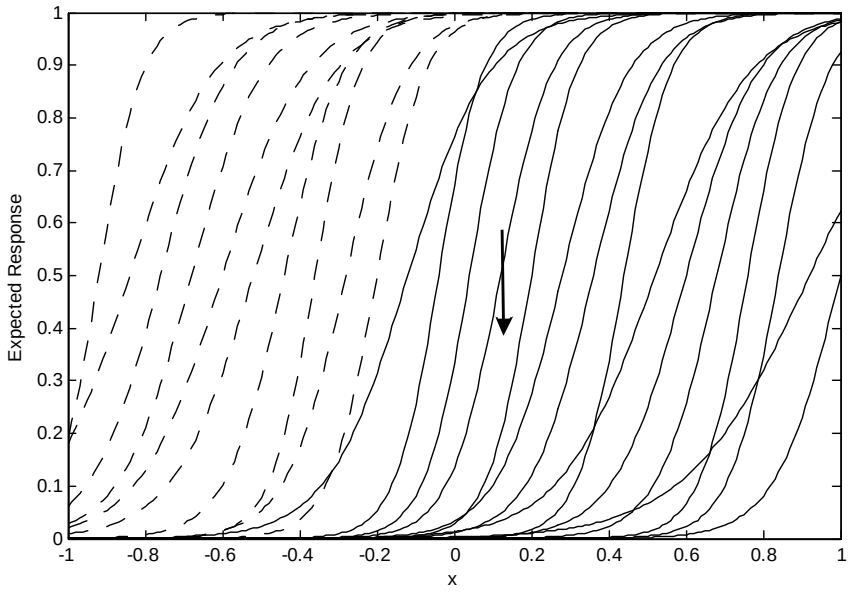

Figure 2: After a single observation of "no response" at $x = 0.13$ we may already have substantial information reducing the parameter space. Dashed lines represent models from the sample presented in Figure 1 that are now unlikely to be the true model.

In multivariate designs there are more parameters to be estimated, and therefore a larger delay from the beginning of data accumulation to achieving data-based maximum likelihood estimates and putting the information obtained into use. If, for example, we would have a first-order model with 3 explanatory variables and their interactions, we would need at least 8 observations before we start using the information gathered to improve the sequential design.

Furthermore, even after completing a set of observations that is equal in size to the number of unknown parameters, we might still not be able to compute a maximum likelihood estimate. This would be the case, for example, in the single factor experiment, if we got a non-response at $x = -0.13$, too. Note that given a non-response at $x = 0.13$ and our prior distribution, a non-response at $x = -0.13$ is very likely. Therefore, the delay between obtaining relevant information and using it to improve the design may be even longer than described before.

Even when we can compute maximum likelihood estimates, they may suffer from severe bias due to the small sample. This is especially true when applying regression techniques to generalized linear models with small samples - the estimates are often highly inaccurate. The result will be the usage of an inefficient experimental plan following a local D-optimal

experimental design for the biased parameter vector.

Neyer (1994) addressed these issues, suggesting a D-optimality based sensitivity test. He suggested a three part procedure, in which the first part is designed to "close in" on the region of interest, the second is devoted to determining unique estimates of the parameters, and only then, in its third and final part, the method uses a local D-optimal design based on the maximum likelihood estimate for the parameters. Handling cases where the MLE's are "wild" estimates is done by restricting the values of the parameters at each step. The scheme of Neyer's (1994) procedure, then, is to minimize the time until we can get good estimates from the data, and then use a local D-optimal design.

It seems there is potential to better utilize the information and increase the efficiency by using a D-optimality criterion beginning from the first observation. More important, a method is needed that does not have the restrictions of sensitivity tests and that extends the functionality of the procedure from univariate cases to the treatment of multiple predictors, from the fully sequential design to any partition of the experiment to batches of size k observations, and from a binary response to any generalized linear model.

3. METHODOLOGY

We separate our method into two parts. The first concerns utilizing the experimental information gathered, for which Bayesian tools are applied. The second concerns the process of choosing the next observation given the current knowledge.

3.1 Parameter Space Discretization

Using a posterior distribution would enable a researcher to update the plan after each observation, and avoid highly biased estimates at early stages. However, updating a prior on the basis of a smaller number of observations than the number of estimated parameters is a difficult task, and the resulting posterior will typically have a complex form. To overcome this difficulty we suggest a Bayesian approximation that exploits a discretization of the parameter space. Figure 3 helps motivate the idea, using the example discussed earlier. The left part of the figure is a discrete candidate set that represents the parameter space. For each vector from this set, (μ_i, θ_i), we can calculate the likelihood given a non-response observation at $x = 0.13$. The

right part of the figure illustrates our posterior knowledge, by presenting only the vectors with likelihood value above a predetermined threshold (in the specific case, vectors with likelihood of no response at $x = 0.13$ greater than 0.05).

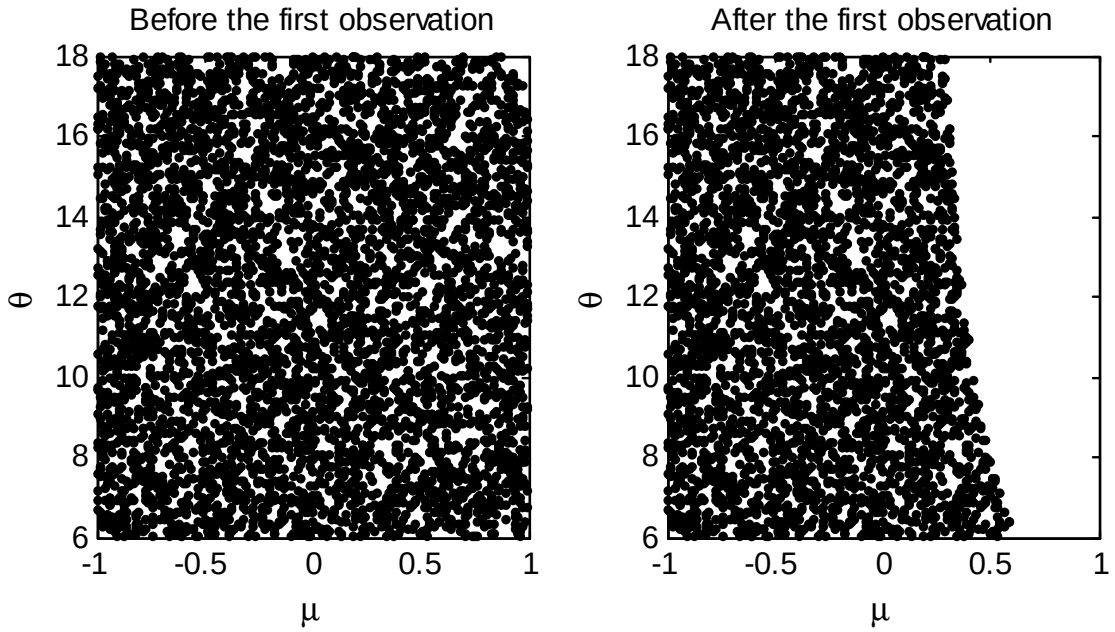

Figure 3: Discrete approximation to the posterior information, for the one-factor logistic model, following a "no response" at $x = 0.13$.

A useful representation of the posterior is easily obtained by methods similar to those used for the production of Figure 3. We first represent the prior distribution by a large sample of discrete vectors (say $N = 10,000$) based on a random or quasi-random sample (such as Niederreiter, 1988) from the same distribution, and compute for each of these vectors the likelihood of the results recorded: for a parameter vector (μ_i, θ_i) and observations $y_1, \ldots, y_k$ taken at $x_1, \ldots, x_k$ the likelihood is $L(\mu_i, \theta_i, y_1(x_1), \ldots, y_k(x_k)) = \prod_{j=1}^{k} \frac{\exp(\theta_i(x_j-\mu_i)y_j)}{1+\exp(\theta_i(x_j-\mu_i))}$ for a logit link. Although we use the logit link throughout this example, the method is suitable for any other link function. For the common case of the probit link, the likelihood is $L(\mu_i, \theta_i, y_1(x_1), \ldots, y_k(x_k)) = \prod_{j=1}^{k} \Phi(\theta_i(x_j - \mu_i))^{y_j} (1 - \Phi(\theta_i(x_j - \mu_i)))^{1-y_j}$. Furthermore, we would like our method to be suitable for multivariate cases, and indeed the discrete parametrization is not limited to a specific number of coefficients or explanatory variables; in the general case where $\boldsymbol{\beta}_i,\ i = 1, \ldots, N$ are vectors of any dimension representing a prior distribution for a coefficient vector $\boldsymbol{\beta}$, and assuming observations $y_1, \ldots, y_k$, taken at $\mathbf{x}_1, \ldots, \mathbf{x}_k$ (with $\mathbf{x}_j$ of any dimension) we can calculate

the likelihood $L(\boldsymbol{\beta}_i, y_1(\mathbf{x}_1), \ldots, y_k(\mathbf{x}_k))$. Now, denoting the prior distribution by $\pi(\boldsymbol{\beta})$, the posterior distribution is given by $f(\boldsymbol{\beta}|y_1(\mathbf{x}_1), \ldots, y_k(\mathbf{x}_k)) \propto L(\boldsymbol{\beta}, y_1(\mathbf{x}_1), \ldots, y_k(\mathbf{x}_k))\pi(\boldsymbol{\beta})$. Rather than computing the posterior, we characterize it in terms of weighted summaries of the prior, using the initial sample to represent the prior and the normalized likelihood as a weight function,
$w_i = L(\boldsymbol{\beta}_i, y_1(\mathbf{x}_1), \ldots, y_k(\mathbf{x}_k)) / \sum_{i=1}^{N} L(\boldsymbol{\beta}_i, y_1(\mathbf{x}_1), \ldots, y_k(\mathbf{x}_k))$. Our representation is reminiscent of importance sampling, in that we simulate the discrete points from the prior, and then weight by the ratio of the posterior to the prior, that is, by the likelihood.

3.2 Choosing a Location for the Next Observation

Ideally, we might consider a weighted D-optimality criterion for choosing each successive point, $\max E\{\log|M(\boldsymbol{\beta})|\}$ where $M(\boldsymbol{\beta})$ is the information matrix for $\boldsymbol{\beta}$, the expectation is taken with respect to the current posterior distribution of $\boldsymbol{\beta}$, and the maximum is over possible locations for the next design point. The criterion is the same one used for non-sequential designs by, for example, Chaloner and Larntz (1989), Woods, Lewis, Eccleston and Russell (2006) and Dror in Steinberg (in press). We could proceed by computing an approximate version of the expectation using the prior and the weights, $E\{\log|M(\boldsymbol{\beta})|\} \approx \sum_{i=1}^{N} w_i \log|M(\boldsymbol{\beta}_i)|$. For instance, consider the sensitivity test example discussed earlier, with a design region $x \in [-1, 1]$ and k observations, $y_1(\mathbf{x}_1), \ldots, y_k(\mathbf{x}_k)$. The global optimization problem for choosing the next point is then

$$x_{k+1} = \underset{z \in [-1,1]}{\arg\max} \sum_{i=1}^{N} w_i \log|M(\mu_i, \theta_i, x_1, \ldots, x_k, z)|. \tag{1}$$

where $M(\mu_i, \theta_i, x_1, \ldots, x_k, z) = F^T W F$, $F^T = \begin{bmatrix} 1 & \ldots & 1 & 1 \\ x_1 & \ldots & x_k & z \end{bmatrix}$, W is a diagonal matrix with $W_{m,m} = \frac{\exp(\theta_i(x_m - \mu_i))}{(1+\exp(\theta_i(x_m-\mu_i)))^2}$ for the m-th diagonal element, and w_i are the weights calculated as described before.

The global optimization for finding x_{k+1} is theoretically attractive but difficult to implement. Solving the global optimization problem for any specific coefficient vector $\boldsymbol{\beta}_i$ is a non-trivial problem, and global optimization over the weighted sum of a large set of such determinants is highly complex. Simulated annealing or other heuristics can be used to search for a promising

x_{k+1} candidate. But, it is common practice to be satisfied with optimization of the information matrix attributed to one promising coefficient vector. In fact, all the methods discussed before, including those of Abdelbasit and Plackett (1983) and Neyer (1994), utilize one guess or estimate of the parameters at each step, which simplifies the optimization above into:

$$x_{k+1} = \underset{z\in[-1,1]}{\arg\max} \log \left| M\left(\hat{\mu}, \hat{\theta}, x_1, \ldots, x_k, z\right)\right|. \tag{2}$$

Using our discrete representation of the posterior distribution, a natural choice for a single parameter vector is the weighted median taken for each of the parameters (that is the value for each parameter, say $\hat{\mu}$ for μ, where half the weight belongs to values smaller than or equal to $\hat{\mu}$ and half to larger values; if $\mu_i,\ i = 1,\ldots,N$ are the N discrete possible μ values in our sample, and if w_i is the weight attributed to μ_i then the weighted median for μ is found by sorting (μ_i, w_i) in ascending μ_i order and choosing $\hat{\mu} = \mu_{(k)}$ such that $\sum_{i=1}^{k} w_{(i)} \geq \frac{1}{2}$ and $\sum_{i=k}^{N} w_{(i)} \geq \frac{1}{2}$).

Returning to the multivariate case, a solution to this simplified optimization problem,

$$\mathbf{x}_{k+1} = \underset{\mathbf{z}\in[-1,1]^p}{\arg\max} \log \left| M\left(\hat{\boldsymbol{\beta}}, \mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{z}\right)\right|. \tag{3}$$

is given in Dror and Steinberg (in press) who describe an algorithm for the construction of local D-optimal designs. The algorithm is based on performing an exchange algorithm on a transformed regression matrix, $\tilde{F} = FW^{1/2}$. This algorithm is suitable to construct and augment local D-optimal designs for models of high dimension with any GLM response, and is not limited to the univariate case or to a binary response. Source code for implementation of the algorithm is available at http://www.math.tau.ac.il/~dms/GLM_Design.

The idea of using a single "best estimate" parameter vector is reasonable since, asymptotically, as the best guess becomes precise, the global search and the local D-optimal design for the best guess become one. This method is simple and quite efficient, yet a few refinements can make it even more efficient with an emphasis on the initial experimental phase, when the parameter estimates may still be poor. Often, the solution to the local D-optimal augmentation (3) is not unique; there may be a few location alternatives for the next observation that are equivalent in terms of their increment to the determinant of the information matrix for the given parameter estimates. It can than be useful to revert to the full weighted sum criterion to break the tie.

That is, rank all points $\mathbf{z}$ that maximize (3), in terms of $\sum_{i=1}^{N} w_i \log |M(\boldsymbol{\beta}_i, \mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{z})|$, so that if $\mathbf{z}_1, \mathbf{z}_2$ are two candidates points which are both local D-optimal augmentations for $\hat{\boldsymbol{\beta}}$,

$$\mathbf{x}_{k+1} = \arg\max_{\mathbf{z} \in \mathbf{z}_1, \mathbf{z}_2} \sum_{i=1}^{N} w_i \log |M(\boldsymbol{\beta}_i, \mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{z})| . \tag{4}$$

This idea may be used even if the one-point D-optimal augmentation of the design is unique. In a sensitivity test, consider a two point augmentation for $\left(\hat{\mu}, \hat{\theta}\right)$, symbolized as (x_{L1}, x_{L2}). Even if using x_{L1} is superior to x_{L2} for the local parameter vector $\left(\hat{\mu}, \hat{\theta}\right)$, the difference in their donation may not be large, and x_{L2} may be a good candidate that will be discovered the preferred choice after evaluating it in relation to the overall posterior distribution.

This raises the question how far forward we should augment to create different candidates for the final comparison. In sensitivity experiments, a local D-optimal design consists of 2 support points (see, for example, Abdelbasit and Plackett, 1983); a three point augmentation usually places the third point as a repetition of one of the first two points, or very close to it. It therefore seems sufficient to use a 2-point augmentation. For the multivariate case one cycle length of a local D-optimal design is advised, as discussed later.

In the initial stages of the experiment the prior may be widely spread and we often find that neither of the local D-optimal points for $\hat{\boldsymbol{\beta}}$ are very good at optimizing (1). In those cases the median of $\{x_{L1}, x_{L2}\}$ is more appropriate, assisting in "slicing" the parameter space in a manner similar to a binary search. This way, in the example discussed in the previous section, the first observation will be placed correctly in the center of the design region, $x = 0$, as should be foreseen given the ambiguity about the parameters, and not at one of the local D-optimal points, $x = \pm 0.13$.

Summing up the ideas, we begin by finding a one cycle length local D-optimal augmentation for $\hat{\boldsymbol{\beta}}$ denoted as $\{\mathbf{x}_{L1}, \ldots, \mathbf{x}_{Lm}\}$ using the algorithm suggested in Dror and Steinberg (in press), and then choose $\mathbf{x}_{k+1}$ through a comparison of the candidate points $\{\mathbf{x}_{L1}, \ldots, \mathbf{x}_{Lm}, \text{median}(\mathbf{x}_{L1}, \ldots, \mathbf{x}_{Lm})\}$, evaluated over the full posterior represented by the weighted prior sample:

$$\mathbf{x}_{k+1} = \arg\max_{\mathbf{z} \in \{\mathbf{x}_{L1}, \ldots, \mathbf{x}_{Lm}, \text{median}(\mathbf{x}_{L1}, \ldots, \mathbf{x}_{Lm})\}} \sum_{i=1}^{N} w_i \log |M(\boldsymbol{\beta}_i, \mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{z})| . \tag{5}$$

As noted earlier, the parameter vector $\hat{\boldsymbol{\beta}}$ is the weighted median over a sample from the

prior distribution of the parameters.

3.3 Multivariate Models

Multivariate models are of great importance as often there is more than one factor affecting the expected response. In explosives sensitivity tests, for example, it could be that the height of drop, the ambient temperature and the angle of impact all affect the detonation probability. In pharmacological experiments, one might want to test a mix of doses from more than one substance or drug. Unlike the Bruceton (Dixon and Mood, 1948), Langlie (1965) or Neyer (1994) procedures, which treat only sequential designs for models with one explanatory variable, our suggested method can be applied "as is" to multivariate models.

One point worth expanding on is the choice of horizon the experimenter should augment forward to create candidate points for comparison. It was noted before that we advise a horizon that is one cycle length of a local D-optimal design; but, a good size for a (pre-planned) local D-optimal design is not trivial (see Dror and Steinberg, in press). To choose a cycle length we suggest comparing the relative efficiency (normalized to the number of points) of local D-optimal designs with different size, all created for the centroid of the prior distribution using the algorithm described in Dror and Steinberg (in press). Observing the different efficiencies can give good directions of a suitable choice.

As an example we use a crystallography experiment, based on Woods *et al.* (2006), in which four explanatory variables (rate of agitation during mixing, volume of composition, temperature and evaporation rate) affect the probability that a new product is formed. The factors have been coded so that the design space is $[-1, 1]^4$. As in Dror and Steinberg (in press) we focus on the prior distribution in Table 1, with all coefficients mutually independent and uniformly distributed (parameter space $\mathcal{B}_3$ in Table 1 of the original paper Woods *et al.,* 2006).

Table 1: Coefficient ranges from Woods *et al.* (2006) crystallography experiment.

Parameter	Range
β_0	$[-3, 3]$
β_1	$[4, 10]$
β_2	$[5, 11]$
β_3	$[-6, 0]$
β_4	$[-2.5, 3.5]$

We choose one candidate parameter vector, say $\boldsymbol{\beta} = (0, 7, 8, -3, 0.5)'$, construct local D-optimal designs for experiments of length $n = 5, ..., 24$, normalize their determinant according to the experimental size, and plot their relative D-efficiency. Figure 4 was created this way. It is clear from the figure that an augmentation of 8 points forward is a good choice, since a local D-optimal design with 8 points is as efficient as larger designs (in terms of information contributed per experimental point).

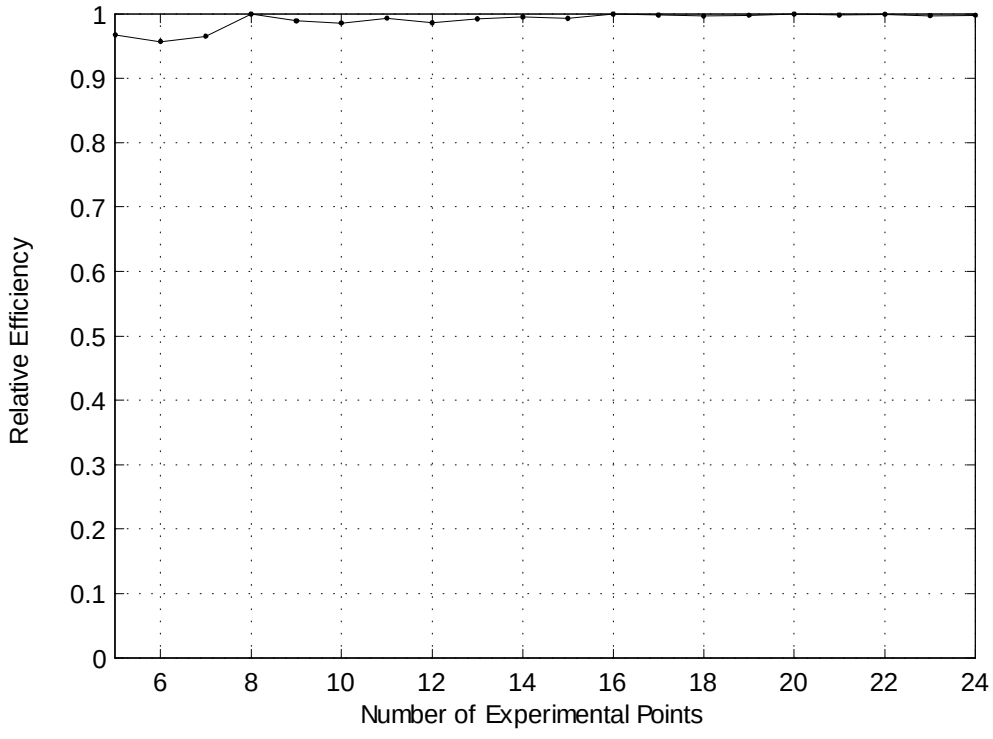

Figure 4: Choosing a cycle-length for the crystallography experiment.

3.4 Group Sequential Designs

Between the two extremes of fixing the entire experiment in advance, regardless of the observations accumulated, and the fully sequential approach in which we update the plan after each observation, there is the possibility of a group sequential design. The term group sequential refers to cases where we obtain data in groups of $k > 1$ observations (with $k = 1$ being the fully sequential case). An example is a situation where we can update the experimental plan from day to day, but must plan the k observations for each day in advance. Another example is when we put a batch of experimental units together in an oven, so the specifications for each member of the batch must be pre-determined.

In Dror and Steinberg (in press), we suggested using K-means clustering in order to create a robust one-stage (predetermined) design for generalized linear models. We now extend the ideas of that article in a way that makes efficient group sequential designs possible. Previously, for fully sequential designs, we chose one representative parameter vector from the prior distribution, using a weighted median, with weights reflecting the results thus far. A simple extension is to choose k vectors using a weighted K-means procedure. That is, we perform a K-means procedure on the discrete sample representing the prior distribution, using a weighted distance measure when calculating the centroids, with weights that are determined by the likelihoods. Similarly, we can choose a sub-sample from the discrete sample, with points selected using probabilities proportional to their weights, and then use a regular K-means procedure on this sample.

After performing clustering we have k parameter vectors. Following the ideas of the former sections, we find one cycle length D-optimal augmentation for each of them, and so get a candidate set of locations from all these designs and their medians. If each local D-optimal augmentation consists of m points, then we have $(m+1)\,k$ candidate points for the choice of the next k observations. To choose the best k locations out of the $(m+1)\,k$ candidates we can mimic an exchange algorithm procedure: choose k candidates at random, and try to improve the weighted sum of the log determinants of the information matrices by exchanging the initial points with alternatives from the larger $(m+1)\,k$ set. The exchange algorithm for a single information matrix has the benefit of a simple updating formula. No such formula is available for our criterion, based on the weighted sum of log determinants. However, given our relatively

small set of candidate design points, direct computation can be used to assess exchanges. The computational burden can be reduced further by a simplification - limiting the candidate set for each row only to the $(m+1)$ candidates produced by the same local D-optimal design (for one of the K-means clustering centroids).

3.5 Non-Binary Responses

We find it important to emphasize that our proposed method can be applied "as is" to models with any response that fits a generalized linear model, and is not limited to a binary response. This includes, of course, the common case of Poisson count models.

3.6 Robustness

Often prior to the experiment the researcher faces uncertainty not only over the possible coefficient values for a specific parametric model, but also over the model itself. There could be different alternative for the link function. There might be several possible linear predictors, expressing indecision as to whether the true model is of first or second order, or whether it should include certain interactions. Woods *et al.* (2006) and Dror and Steinberg (in press) confronted these issues for one-stage designs. Yet, for sequential designs current procedures require the experimenter to choose one model in advance.

Our proposed method can take into account different possibilities with only minor modifications. First, create a discrete representation of a prior for each of the possible models. The numbers of points in each prior should be in proportion to the a-priori weight given to the assumption that this is the right model. Next, calculate likelihoods for all the parameter vectors in all the models, and choose the next observation (or group of observations), as described previously, for each model separately. The result is a group of candidates, and one can once again use an exchange-algorithm mimic to choose the next observation (or group of k observations) so as to maximize the weighted sum of log determinants for the information matrices - over the full prior (including the different models).

4. EXAMPLES

We demonstrate the advantages of the new algorithm through a series of examples. First, we demonstrate better efficiency achieved in sensitivity tests, with comparative results from an actual sensitivity test conducted recently at an industrial plant, followed by a comprehensive comparison of techniques via a Monte-Carlo simulation. Then, for capabilities which extend beyond the limits of the currently available algorithms, we compare results between one-stage algorithms, as presented in Dror and Steinberg (in press) and the new method. Examples include the crystallography experiment introduced earlier, for both fully sequential and group sequential applications, and a Poisson count model with robustness to the choice of different linear predictors.

4.1 Industrial Plant Experimentation

We depict a sensitivity experiment that took place in June 2006, at a military industrial plant. It was performed twice: first following a standard format in use at the plant, which is based on the "Bruceton" method of Dixon and Mood (1948); then, using our new algorithm. The experiment's objectives were to estimate the sensitivity curve in general, and in particular to verify a manufacturer's statement that the explosives will not detonate at 12V (being a safe voltage), and will detonate ("all fire") at 25V. Quantitatively, the requirement was to show, using probit regression, that the probability of detonation at 12V is under 5% and the probability of detonation at 25V is above 95% (that is - the 95% probit confidence interval for the expected response should be below 5% at 12V and over 95% at 25V).

The experimenters began with limited prior knowledge. They could not say what voltage would provoke a response from half the observations (the mean of the response curve), or even if this value is within the specified range of 12-25V; not much was known about the dispersion as well, with the possibility of a very slow increase from the "no-fire" to "all-fire" zone, to a very steep curve.

Together with the plant engineers, we formulated the following prior distribution, which reflects this (lack of) information. The prior takes $(x-\mu)/\sigma$ as the parametrization and a log-normal distribution for both the mean and dispersion, with $\mu \sim$lognormal$(\log(17), 0.5^2)$ and $\sigma \sim$lognormal$(\log(0.7), 1^2)$. Figure 5 presents a sample of 150 curves given possible values

for (μ, σ) sampled from the prior distribution.

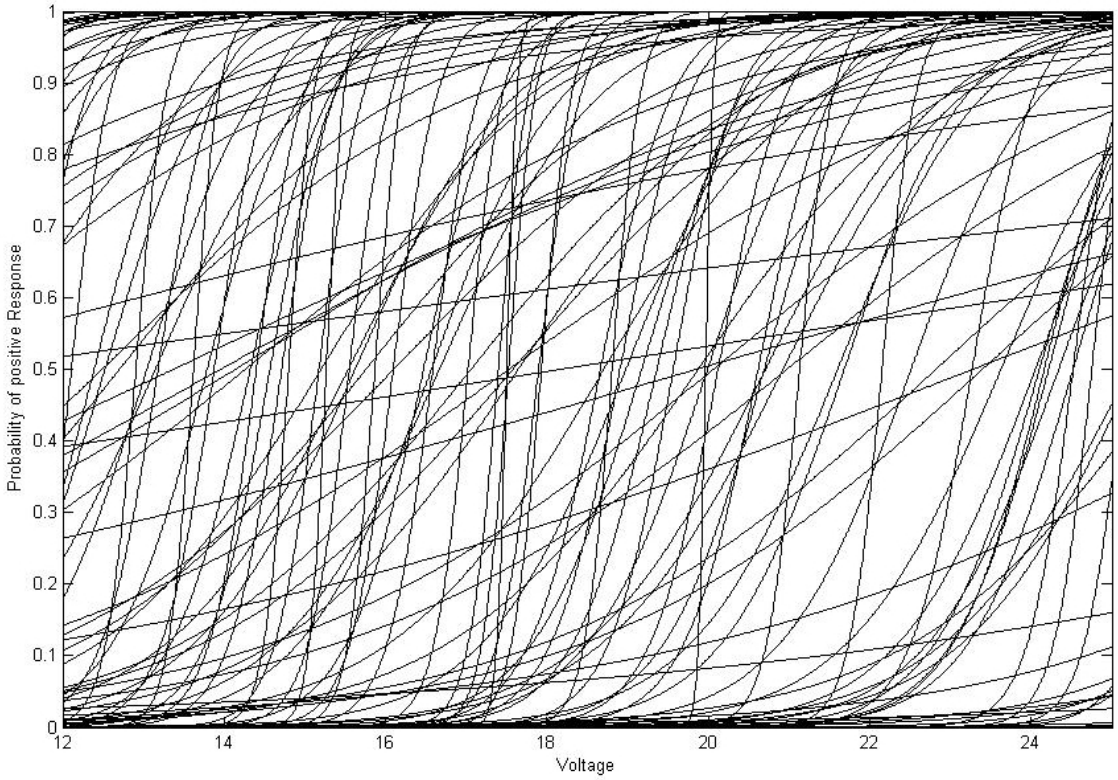

Figure 5: Sensitivity experiment - a sample of possible curves from the prior distribution.

It is seen that there are both very steep curves (with a small fraction of a volt separating the "no-fire" to "all-fire" zone) and very flat curves offering the possibility that there is a non-negligible chance of detonation at 12V AND of no-detonation at 25V. The industrial plant's procedure began with a "screening" phase following the "Bruceton" procedure with a rather large step size of 1V. After 9 observations, when the experimenters felt they knew the right region, they began a new "Bruceton" procedure with a smaller step size of 0.25V; this second "Bruceton" continued for 31 trials, so a total of 40 runs were performed. For the new algorithm 20 trials were allocated. The locations at which the observations were taken by each method, and their outcomes are presented in Figure 6.

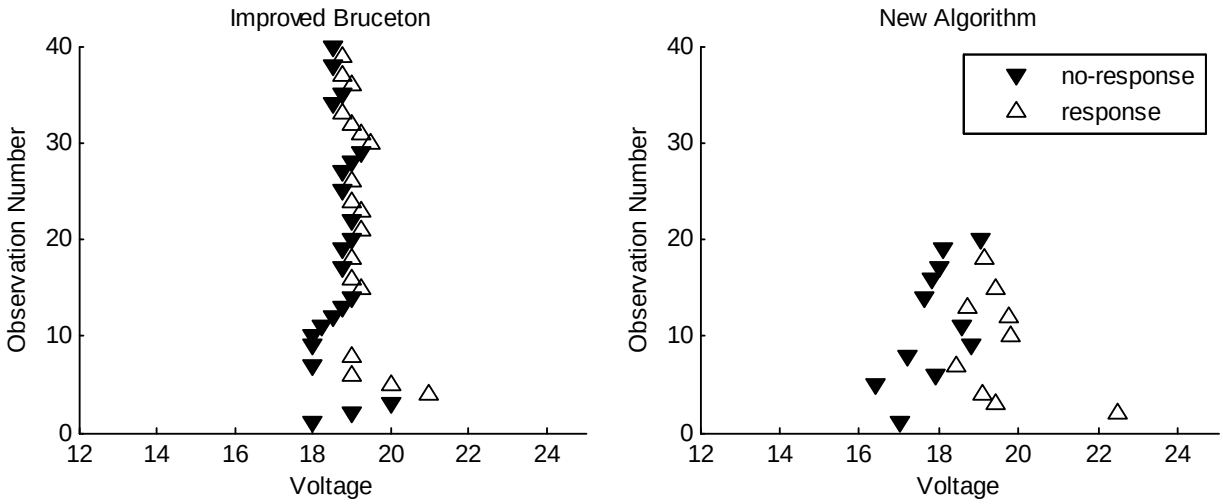

Figure 6: Sensitivity experiment - record of observations' location and outcome.

Figure 7 compares the analysis of outcomes of both methods after 20 trials. The curves in the figure represent the 95% confidence intervals for the expected response as a function of voltage value, as produced by a "probit" regression.

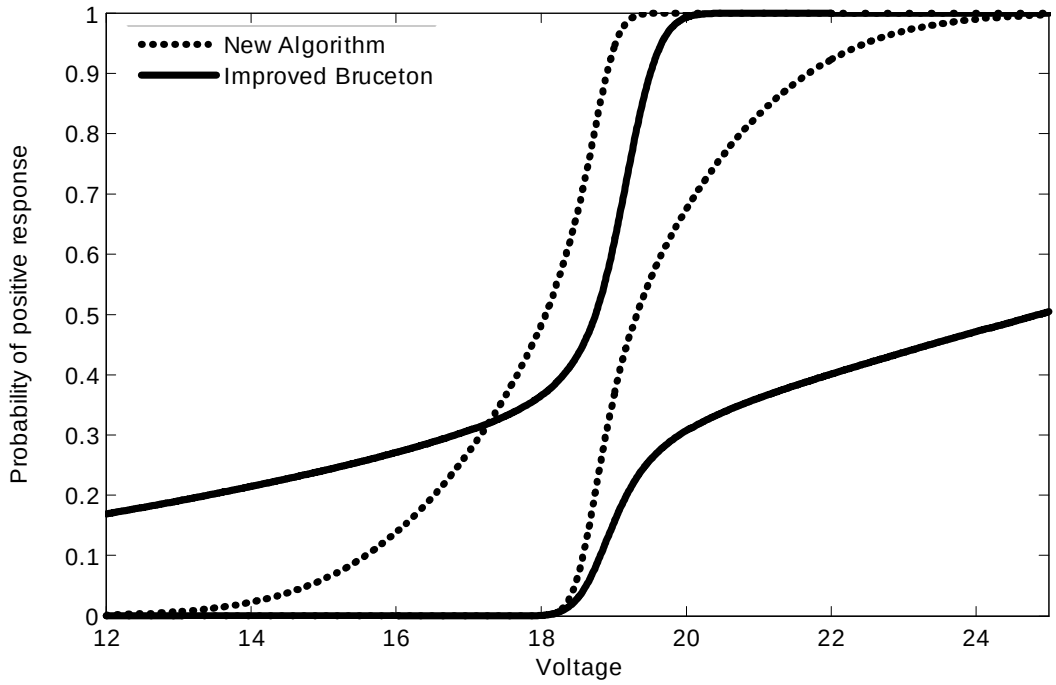

Figure 7: Comparison of methods after 20 observations.

It is seen that the confidence intervals are much smaller for the new method. In fact, the

new method has already reached the objective after 20 observations, while the "Bruceton" plan is far from it (not reaching the objective at either 12V or 25V). In total there were 40 observations using the "Bruceton" method. Figure 8 is similar to Figure 7, after completing all 40 observations (versus the outcome of the 20 observation for the new algorithm).

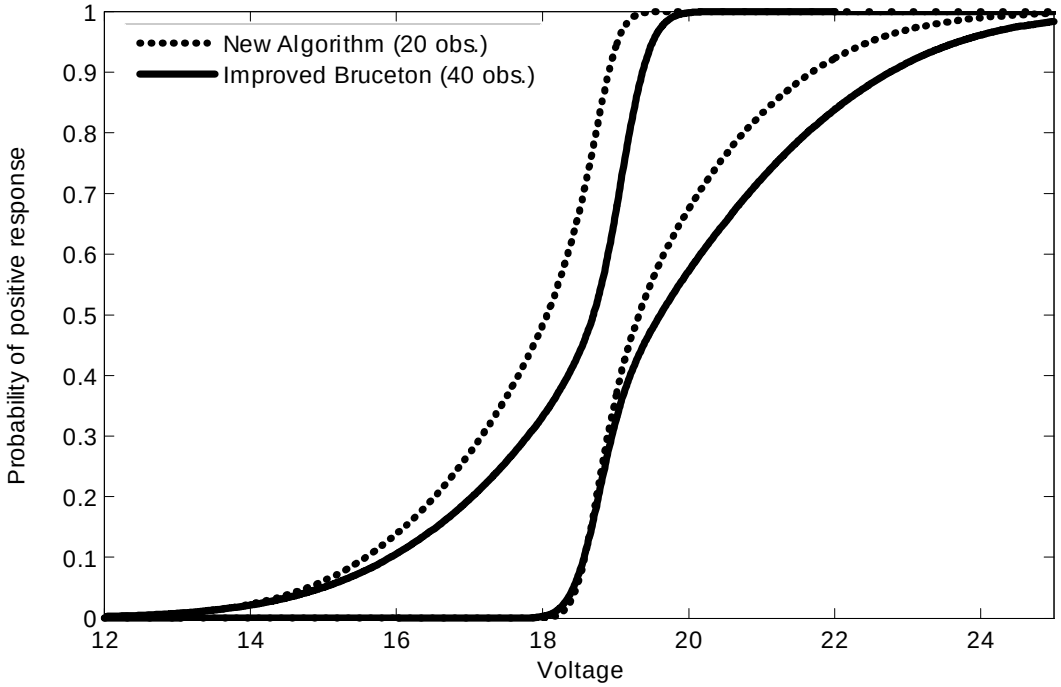

Figure 8: Comparison of methods.

Bruceton after 40 observations vs. the new method with 20 observations.

Even in the extreme comparison where the Bruceton method benefits from twice as many observations (40) as the new method (20), the latter succeeds to provide better results. Observe the strip where the probit 95% confidence interval is above 0.05 and below 0.95 for the probability of detonation; the Bruceton method yields a strip of 15V-23.5V, compared to the narrower strip yielded by the new algorithm: 15V-22.5V.

4.2 Monte-Carlo based comparison

A more comprehensive comparison between the two methods is available through Monte-Carlo experimentation. We compare the Bruceton method (Dixon and Mood, 1948), Neyer's (1994)

procedure and our proposed algorithm.

We conducted the comparison for various possible "true" models for a sensitivity test similar to the one performed in practice. For each "true model" we carried out 100 repetitions of an experiment with each method. In each repetition the location for each observation was determined according to the algorithm's rules and the outcome of the observation was generated at random according to the "true" model. Figure 9 presents the different "true" models that were used for the comparison. The figure shows that we consider cases with both true mean smaller than 12V or higher than 25V, very steep curves and very flat curves. The bold curve emphasizes the case where the center of the prior distribution discussed above (or the "best guess" for the parameter values) is in fact the true model.

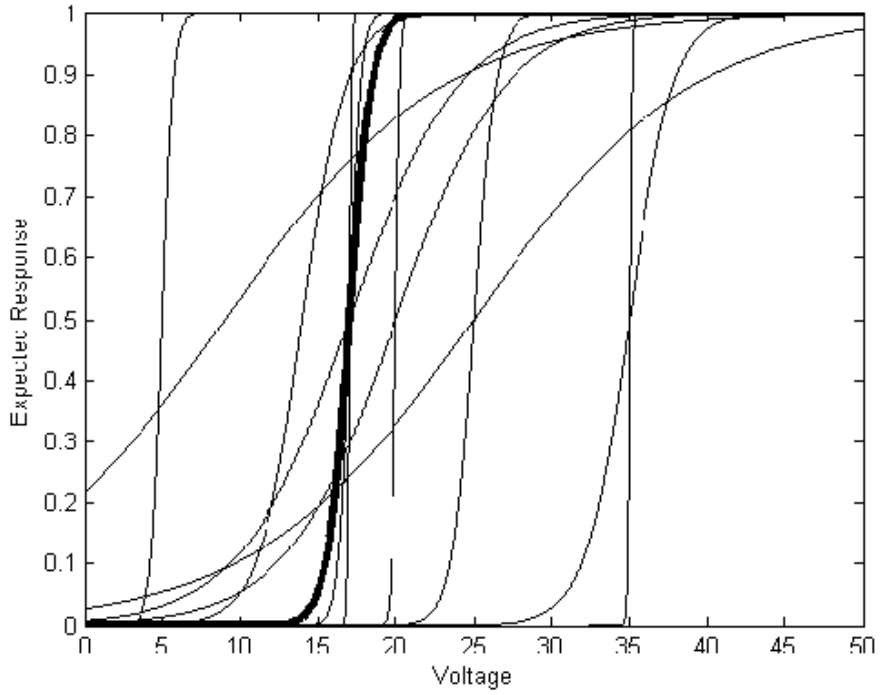

Figure 9: "True" models for the Monte-Carlo comparison of techniques
(bold curve represents the case where the best guess is in fact the true model).

Each of the methods requires a slightly different input from the user. For our method a prior distribution is needed. This allows the experimenter to balance between cases where he feels he has a good guess of the parameters prior to the experiment, and therefore would like

to have high efficiency when his guess will be reinforced by the data, versus cases in which his best guess is of low fidelity so robustness to true parameter values that are far from the initial estimate is required. We will assume the latter, and use the same prior as before: $\mu \sim$lognormal$(\log(17), 0.5^2)$ and $\sigma \sim$lognormal$(\log(0.7), 1^2)$.

The Bruceton method (Dixon and Mood, 1948) requires an initial guess for μ and a step size, based on a probit model guess for σ. We will use the center of our prior distribution as the guess, so $\hat{\mu} = 17$, and the chosen step size is 1.28V, after adjusting the difference of σ for the logit and probit models. Neyer's (1994) algorithm requires guesses for a lower and upper bound for the mean, which we will assume $\mu_{\min} = 12,\ \mu_{\max} = 25$, and a guess of the standard deviation, for which we will use the center of the prior distribution $\hat{\sigma} = 0.7$.

We have conducted experiments of length 16 and 48 observations. Table 2 presents the median D-efficiency (out of 100 repetitions for each true parameter vector and each method), and Table 3 presents the 5% quantile of the D-efficiencies for the same cases.

Table 2: Median D-Efficiency for different techniques.

			16 observations			48 observations		
Case #	True μ	True σ	Bruceton	Neyer	New	Bruceton	Neyer	New
1	17	0.07	0.002	0.40	0.31	0.002	0.74	0.63
2	35	0.07	0.0005	0.11	0.22	0.002	0.32	0.45
3	20	0.14	0.18	0.59	0.48	0.19	0.78	0.76
4	5	0.35	0.37	0.58	0.59	0.70	0.82	0.77
5	17	0.35	0.67	0.77	0.69	0.67	0.86	0.84
6	17	0.7	0.85	0.80	0.77	0.86	0.88	0.88
7	25	0.7	0.64	0.69	0.74	0.79	0.84	0.85
8	14	1.4	0.77	0.74	0.83	0.79	0.86	0.89
9	35	1.4	0.25	0.61	0.67	0.64	0.83	0.85
10	17	3.5	0.41	0.61	0.72	0.50	0.86	0.88
11	20	3.5	0.43	0.49	0.68	0.54	0.86	0.86
12	9	7	0.31	0.51	0.59	0.42	0.86	0.85
13	25	7	0.33	0.44	0.57	0.43	0.85	0.83

Table 3: 5% quantile of D-Efficiencies for different techniques.

			16 observations			48 observations		
Case #	True μ	True σ	Bruceton	Neyer	New	Bruceton	Neyer	New
1	17	0.07	0.002	0.37	0.25	0.002	0.65	0.50
2	35	0.07	0.0001	0.11	0.18	0.002	0.28	0.41
3	20	0.14	0.15	0.47	0.36	0.18	0.63	0.56
4	5	0.35	0.28	0.45	0.45	0.64	0.71	0.64
5	17	0.35	0.63	0.61	0.48	0.65	0.69	0.66
6	17	0.7	0.76	0.67	0.61	0.83	0.76	0.76
7	25	0.7	0.51	0.55	0.57	0.74	0.67	0.72
8	14	1.4	0.57	0.62	0.60	0.69	0.64	0.73
9	35	1.4	0.14	0.51	0.51	0.57	0.62	0.72
10	17	3.5	0.30	0.27	0.36	0.40	0.54	0.63
11	20	3.5	0.27	0.25	0.41	0.40	0.49	0.71
12	9	7	0.15	0.16	0.23	0.30	0.51	0.61
13	25	7	0.16	0.15	0.25	0.30	0.39	0.60

Figure 10 offers a graphical representation of the results in Table 2, comparing the median D-efficiencies for the three techniques in experiments with 16 observations. The lower part of Figure 10 shows the values of the true parameters, in the same order as in Table 2; case number 1, for example, is the parameter vector in the first row: $(\mu, \sigma) = (17, 0.07)$ and case number 6 is the best guess, or center of the prior distribution, with $(\mu, \sigma) = (17, 0.7)$. The upper part is the median D-efficiencies. For example, it is seen that in case number 10, $(\mu, \sigma) = (17, 3.5)$, the median D-efficiencies are 0.41, 0.61,0.72 for the Bruceton, Neyer and the new algorithm, respectively.

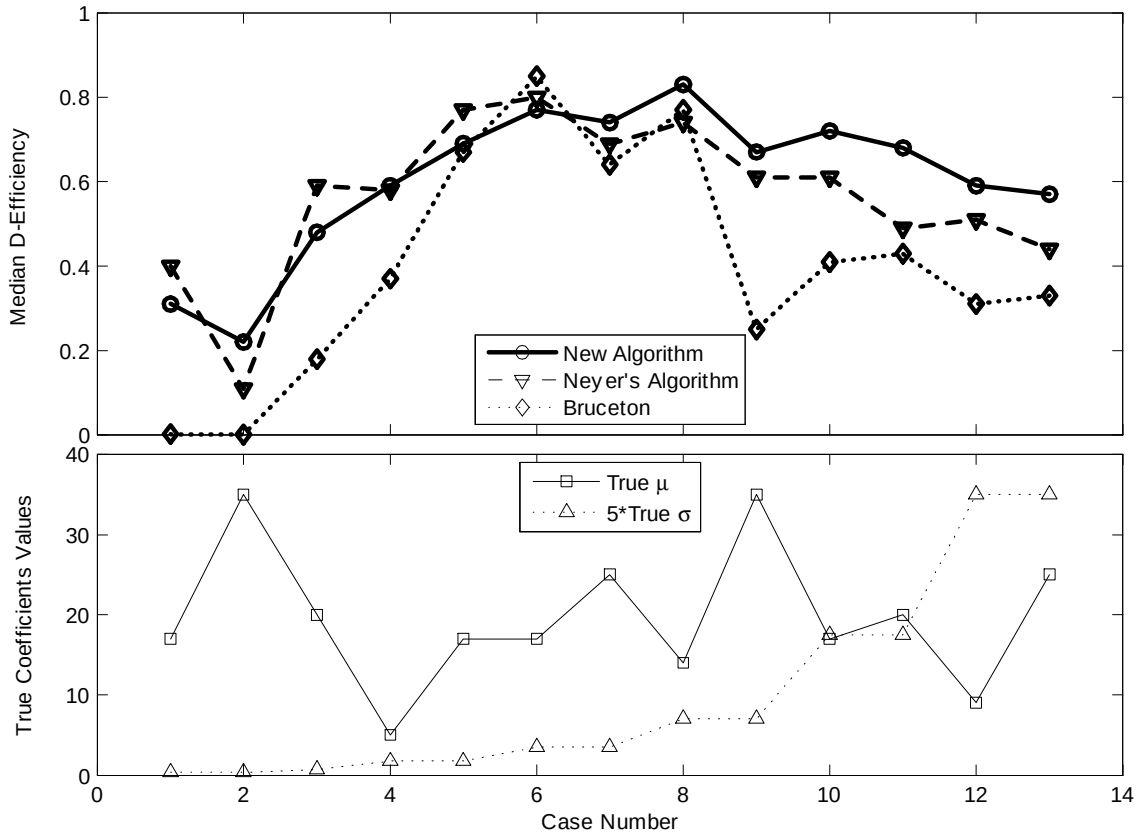

Figure 10: Comparison of median D-efficiencies for 16-run sensitivity experiments.

It is clear from Tables 2,3 and Figure 10 that the Bruceton method (Dixon and Mood, 1948) is inferior in its efficiency to the other techniques. When compared to Neyer's (1994) procedure, our proposed algorithm handles better the uncertainty as described by the prior distribution: D-efficiencies for Neyer's (1994) algorithm are roughly equal to or slightly better than the new algorithm for cases where the true σ is smaller than its guess (or at the center of the prior distribution), but the proposed algorithm is significantly superior for the cases where the true σ is larger than its best guess. Choosing a different prior with a larger emphasis on smaller values of σ (for example by deliberately choosing a biased prior distribution for σ such as $\sigma \sim$lognormal$\left(\log\left(\frac{0.7}{2}\right), 1^2\right)$) allows the proposed algorithm to be generally superior to Neyer's (1994) algorithm for both large and small true σ values, but is less desirable as it does not exploit the efficiency for large true σ values to its maximum potential.

The comparison so far assumed vague knowledge of the parameters as expressed by the prior

distribution. If we assume that higher fidelity prior information is available, say reducing the variance value in the prior so that $\mu \sim$lognormal$(\log(17), 0.2^2)$ and $\sigma \sim$lognormal$(\log(0.7), 0.75^2)$, then the efficiency of the design is increased. Note that this prior still has non-negligible mass for μ below 12 and above 25 and also allows various σ values. As can be seen in Table 4, for the typical case when it is possible to supply reasonable estimates for the parameters, so the prior is not completely vague, the proposed algorithm is even more efficient when compared to Neyer's (1994) algorithm, and all the more so if compared to older test methods such as Bruceton (Dixon and Mood 1948).

Table 4: Median D-Efficiency for different techniques - adding a narrow prior.

					New Algorithm	
observations	True μ	True σ	Bruceton	Neyer	Wide Prior	Narrow Prior
16	18	0.37	0.72	0.76	0.77	**0.83**
	17	0.7	0.85	0.80	0.75	**0.83**
	3	1.75	0.71	0.71	0.76	**0.71**
48	18	0.37	0.74	0.86	0.87	**0.89**
	17	0.7	0.86	0.89	0.88	**0.91**
	3	1.75	0.74	0.86	0.89	**0.85**

4.3 Multivariate models

Better efficiency is only one of our algorithm's advantages, perhaps not even the key benefit. The following examples demonstrate how our algorithm extends functionality to multivariate models, group sequential designs and to any generalized linear model.

Lacking alternative sequential algorithms for multivariate models we cannot provide a comparison of performance the way we did with univariate sensitivity tests. Dror and Steinberg (in press) suggested a robust one-stage experiment for the crystallography experiment described earlier. They evaluated the design versus a sample of 10,000 vectors from the prior, calculating the local D-efficiency of the design versus each of them, reporting the median and minimum local D-efficiency. We applied the sequential algorithm on the same prior, using a Monte-Carlo evaluation as before, with a small modification; each of the 10,000 iterations was performed

with a different true coefficient value, sampled randomly from the prior in Table 1. This allows a legitimate comparison of the median local D-efficiency achieved through the robust one-stage design and the proposed sequential design. Table 5 displays the comparison

Table 5: D-Efficiency comparison of robust one-stage design versus the proposed multivariate sequential procedure.

	16 observations		48 observations	
	One stage	Sequential	One stage	Sequential
Median D-Efficiency	0.42	0.64	0.42	0.83
5% quantile D-Efficiency	0.26*	0.46	0.31*	0.73
Minimum D-Efficiency	0.10	0.24	0.18	0.56

(*) 5% quantile for this case was estimated by a single robust design, which has above average median and minimum efficiency

As expected, performing the experiment sequentially considerably improves the efficiency. Note that the 5% quantile D-Efficiency for the sequential design is better than the median D-Efficiency for robust one-stage designs.

4.4 Group Sequential Designs

We continue with the crystallography example, and present in Table 6 the addition of a group sequential design, with each group including 16 observations.

Table 6: D-Efficiencies for the Group Sequential procedure versus one stage and fully sequential.

	16 observations				48 observations
	One stage	Group Sequential	Sequential	One Stage	Group Sequential
Median D-Efficiency	0.42	**0.42**	0.64	0.42	**0.73**
5% quantile D-Efficiency	0.26*	**0.26**	0.46	0.31*	**0.62**
Minimum D-Efficiency	0.10		0.24	0.18	

We see that even for a large group of 16 observations, so that the 48 observations are conducted in only 3 stages, the group sequential procedure's D-Efficiency is closer to the fully sequential results than to the one-stage robust design's efficiency.

For the 16-observation experiment, the group sequential design is in fact a one-stage design. It is desired that the proposed sequential procedure would provide an efficiency as close as possible to a one-stage design if all the experiment is grouped together into a single batch. Indeed, Table 6 shows that our group sequential algorithm succeeded to reproduce the one-stage results obtained in Dror and Steinberg (in press).

4.5 Non-Binary Responses and Robustness

As a concluding example we will use a model with a Poisson count as response, 5 explanatory variables, and two competing linear predictors - one with interactions and the other a standard first-order model. This could represent, for example, a clinical trial where the response is the bacterial count a specified time after treatment, and the explanatory variables are doses from 5 types of drugs. We will use as a baseline an equivalent problem concerned with wave soldering defects, as discussed and analyzed, for a one-stage design, in Dror and Steinberg (in press). The prior, which includes two possible models (with and without interactions) is displayed in Table 7, with the assumption of independent normal distributions for the parameters.

Table 7: Prior coefficient estimates for two models for the wave soldering example.

	First-order		With Interactions	
Term	Estimate	S.E.	Estimate	S.E.
Intercept	-1.52	0.21	-2.35	0.69
x_1	-4.30	0.20	-5.53	0.94
x_2	-1.79	0.16	-2.99	0.82
x_3	-3.39	0.24	-3.95	0.59
x_4	-0.28	0.32	-0.86	0.54
x_5	0.23	0.30	0.41	0.36
x_1x_2			-2.07	1.32
x_1x_3			-1.13	0.98

Dror and Steinberg (in press) displayed a histogram of local D-Efficiencies for the robust design versus 20,000 parameter vectors taken from this prior (half for a first-order model, and half for the model with interactions). Figure 11 compares this histogram to a histogram of local D-efficiencies achieved by applying the sequential procedure described above to the same problem.

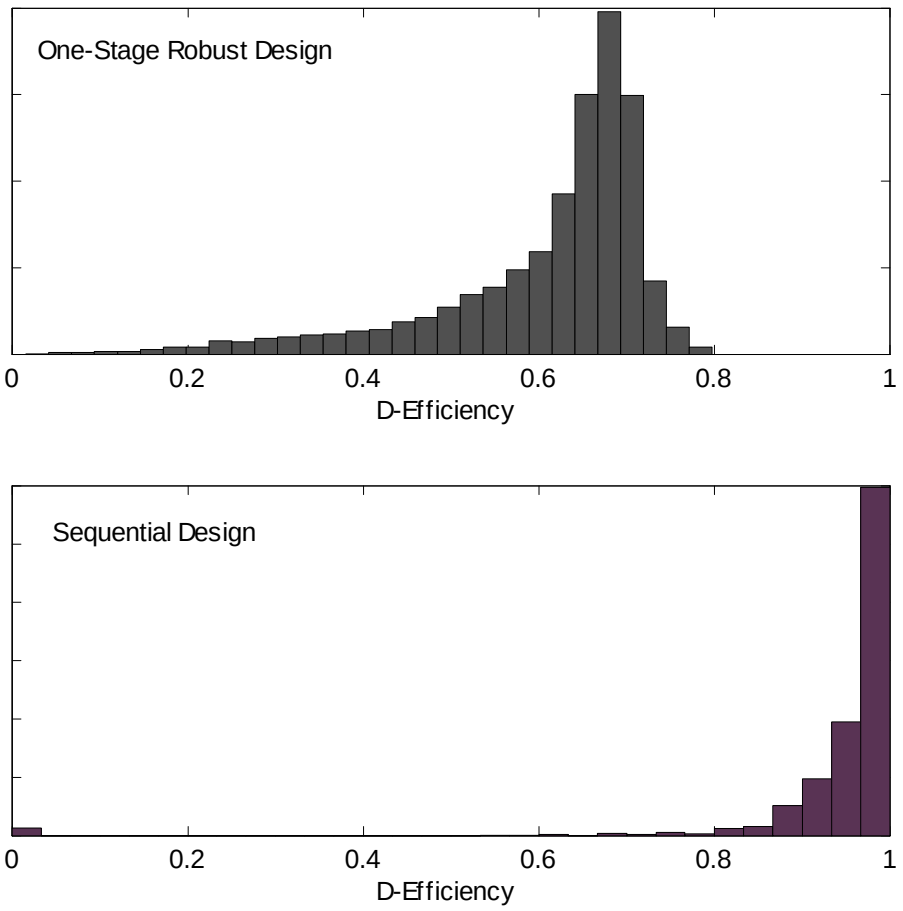

Figure 11: Sequential versus One-Stage robust design.

Remark 1 *A small portion of the runs ended with a zero local D-efficiency. This may happen when the true model contains interactions, but their effects are non-significant. In such a setting the algorithm may choose a design that cannot estimate the interactions, and therefore - although being quite efficient for estimating the expected response - have zero local D-efficiency for the true model.*

It is clear that the sequential design is superior to the one-stage design, with a median

D-Efficiency value of 0.98 versus 0.67 for the one-stage design, and 5% quantile of 0.85 for the sequential design versus 0.30 for the one-stage design.

5. CONCLUSIONS

An algorithm has been suggested that outperforms available procedures for sensitivity tests, and enables simple extension to the treatment of more complex models. These include multi-factor experiments, group sequential designs, responses for any generalized linear model, and the possibility to consider several competing alternatives for the true model.

The disadvantages of the "one variable at a time" attitude to the design of experiments are well known, both for its implied result of inflation in required sample size and for its ignorance of interaction effects. The proposed procedure enables proper sequential designs for multivariate models, avoiding this common misstep.

REFERENCES

1. Abdelbasit, K. M., and Plackett, R. L. (1983), "Experimental Designs for Binary Data," *Journal of the American Statistical Association*, 78, 90-98.

2. Chaloner, K. and Larntz, K. (1989), "Optimal Bayesian Design Applied to Logistic Regression Experiments," *Journal of Statistical Planning and Inference*, 21, 191-208.

3. Dixon, J. W., and Mood, A. M. (1948), "A Method for Obtaining and Analyzing Sensitivity Data," *Journal of the American Statistical Association*, 43, 109-126.

4. Dror, H. A. and Steinberg D. M. (in press), "Robust Experimental Design for Multivariate Generalized Linear Models," *Technometrics.*

5. Langlie, H. J. (1965), "A Reliability Test Method For 'One-Shot' Items," Technical Report U-1792, Aeronutronic Division of Ford Motor Company, Newport Beach, California.

6. Neyer B. (1994), "A D-optimality-Based Sensitivity Test," *Technometrics*, 36, 61-70.

7. Niederreiter, H. (1988), "Low-Discrepancy and Low-Dispersion Sequences," *Journal of Number Theory*, 30, 51-70.

8. Woods, D.C., Lewis, S.M., Eccleston, J.A. and Russell, K.G. (2006), "Designs for Generalized Linear Models with Several Variables and Model Uncertainty," *Technometrics*, 48, 284-292.

9. Wu, C.F.J. (1985), "Efficient Sequential Designs with Binary Data," *Journal of the American Statistical Association*, 80, 974-984.