Gradient Estimation Schemes for Noisy Functions

Brekelmans, R.C.M.; Driessen, L.; Hamers, H.J.M.; den Hertog, D.

Publication date:
2003

CentER

Discussion Paper

No. 2003–12

GRADIENT ESTIMATION SCHEMES FOR NOISY FUNCTIONS

By Ruud Brekelmans, Lonneke Driessen, Herbert Hamers and Dick den Hertog

February 2003

ISSN 0924-7815

Tilburg University

Gradient estimation schemes for noisy functions

Ruud Brekelmans[1], Lonneke Driessen[2], Herbert Hamers[3], Dick den Hertog[3,4]

[1] Tilburg University, CentER Applied Research

[2] Centre for Quantitative Methods BV, Eindhoven

[3] Tilburg University, Department of Econometrics and OR, P.O. Box 90153,

5000 LE Tilburg, The Netherlands, e-mail: D.denHertog@uvt.nl

[4] Corresponding author.

February 7th, 2003

Abstract

In this paper we analyze different schemes for obtaining gradient estimates when the underlying function is noisy. Good gradient estimation is e.g. important for nonlinear programming solvers. As an error criterion we take the norm of the difference between the real and estimated gradients. This error can be split up into a deterministic and a stochastic error. For three finite difference schemes and two Design of Experiments (DoE) schemes we analyze both the deterministic and the stochastic errors. We also derive optimal step sizes for each scheme, such that the total error is minimized. Some of the schemes have the nice property that this step size also minimizes the variance of the error. Based on these results we show that to obtain good gradient estimates for noisy functions it is worthwhile to use DoE schemes. We recommend to implement such schemes in NLP solvers.

Key words: Design of Experiments, finite differences, gradient estimation, noisy functions

1 Introduction

We are interested in a function $f : \mathbb{R}^n \to \mathbb{R}$ and more specifically its gradient $\nabla f(x)$. The function f is not explicitly known and we cannot observe it exactly. All observations are the result of function evaluations, which are subject to certain perturbation errors.

Hence, for a fixed $x \in \mathbb{R}^n$ we observe an approximation

$$g(x) = f(x) + \varepsilon(x). \tag{1}$$

The error term $\varepsilon(x)$ represents a random component. We assume that the error terms in (1) are i.i.d. random errors with $E[\varepsilon(x)] = 0$ and $V[\varepsilon(x)] = \sigma^2$, hence the error terms do not depend on x. Note that g can also be a computer simulation model. Even deterministic simulation models are often noisy due to all kind of numerical errors.

In this paper we analyze both finite difference schemes and Design of Experiments (DoE) schemes for obtaining gradient estimations. In all these schemes the gradient is estimated by observing the function value in several points in the neighborhood of x, using finite step sizes h. We compare the resulting errors made in the gradient estimations due to both the presence of noise and the deterministic approximation error ('lack of fit'). It will appear that DoE schemes are worthwhile alternatives for finite difference schemes in the case of noisy functions. Moreover, we will derive efficient step sizes for the different schemes, such that the total error (sum of deterministic and stochastic error) is minimized. We will compare these step sizes to those which minimize the variance of the total error.

Gradients play an important role in all kind of optimization techniques. In most non-linear programming (NLP) codes, first-order or even second-order derivatives are used. Sometimes these derivatives can be calculated symbolically: in recent years automatic differentiation has been developed; see e.g. [7] and [3]. Although this is becoming more and more popular, there are still many optimization techniques in which finite differencing is used to approximate the derivatives. In almost every NLP code such finite difference schemes are implemented.

Finite difference schemes have also been applied to problems with stochastic functions. Kiefer and Wolfowitz [8] were the first to describe the so-called stochastic (quasi) gradients; see also [2]. Methods based on stochastic quasi gradients are still subject of much research; for an overview see [6]. So, although finite difference schemes originate from obtaining gradient estimations for deterministic functions, they are also applied to stochastic functions.

Also in the field of Design of Experiments (DoE), schemes are available for obtaining gradient estimations. Some popular schemes are full or fractional factorial schemes, including Plackett-Burman schemes. Contrary to finite differencing, these schemes take

noise into account. The schemes are such that, for example, the variance of the estimators is as small as possible. However, most DoE schemes assume a special form of the underlying model, e.g. polynomial, and lack of fit is usually not taken into account.

In [4] and [5] also lack of fit is taken into account besides the noise. In those papers it is analyzed what happens when the postulated linear (resp. quadratic) model is misspecified, due to the true model structure being of second (resp. third) order. In these two papers new DoE schemes are derived by minimizing the integrated mean squared error for either the predictor or the gradient. However, we think that such estimations are less valuable for optimization purposes since the integrated mean squared error is not a good measure for the gradient in one point. Moreover, the underlying assumption in those papers is still that the real model is quadratic (in [4]) or third order (in [5]) which is not necessarily true.

The remainder of this paper is organized as follows. In Section 2 we analyze three finite difference schemes for obtaining gradient estimations. In Section 3 we do the same for two DoE schemes. In Section 4 we compare the errors of all the five schemes. We end with some conclusions in Section 5.

2 Gradient estimation using finite differencing

2.1 Forward finite differencing

One classical approach to estimate the gradient of f is to apply forward finite differencing (FFD) to the approximating function g, defined in (1). In this scheme, an estimator of the partial derivative, $\frac{\partial f(x)}{\partial x_i}$ $(i = 1, \ldots, n)$, is obtained by

$$\hat{\beta}_i^{FFD}(h) = \frac{g(x + he_i) - g(x)}{h}, \qquad h > 0, \tag{2}$$

where h is the step size and e_i is the i-th unit vector. Using (1) and Taylor's formula, we can rewrite the estimator as

$$\begin{aligned} \hat{\beta}_i^{FFD} &= \frac{f(x + he_i) - f(x) + \varepsilon(x + he_i) - \varepsilon(x)}{h} && (3) \\ &= \frac{\partial f(x)}{\partial x_i} + \frac{1}{2} h e_i^T \nabla^2 f(x + \zeta h e_i) e_i + \frac{\varepsilon(x + he_i) - \varepsilon(x)}{h}, && (4) \end{aligned}$$

in which $0 \leq \zeta \leq 1$. We are now interested in how good this estimator is. Note that

$$E[\hat{\beta}_i^{FFD}] = \frac{\partial f(x)}{\partial x_i} + O(h) \tag{5}$$

$$\text{VAR}[\hat{\beta}_i^{FFD}] = \frac{2\sigma^2}{h^2}. \tag{6}$$

The estimators $\hat{\beta}_i^{FFD}$ and $\hat{\beta}_j^{FFD}$ are correlated, because both depend on the random error $\varepsilon(x)$:

$$\begin{aligned}
\text{Cov}[\hat{\beta}_i^{FFD}, \hat{\beta}_j^{FFD}] &= E\left((\hat{\beta}_i^{FFD} - E[\hat{\beta}_i^{FFD}])(\hat{\beta}_j^{FFD} - E[\hat{\beta}_j^{FFD}])\right) \\
&= \frac{1}{h^2} E\left((\varepsilon(x + he_i) - \varepsilon(x))(\varepsilon(x + he_j) - \varepsilon(x))\right) \\
&= \frac{1}{h^2} E\left(\varepsilon(x)^2\right) \\
&= \frac{\sigma^2}{h^2}, \qquad i \neq j.
\end{aligned}$$

However, we are not only interested in the errors of the individual derivatives, but more in the error made in the resulting estimated gradient. A logical measure for the quality of our gradient estimation is the mean squared error:

$$E\left(\left\|\hat{\beta}^{FFD} - \nabla f(x)\right\|^2\right).$$

Not only the expectation is important, but also the variance

$$VAR\left(\left\|\hat{\beta}^{FFD} - \nabla f(x)\right\|^2\right),$$

since high variance means that we run the risk that the error in a real situation is much higher (or lower) than expected. Suppose for example that two simulation schemes have the same expected mean squared error, then we prefer the scheme with the lowest variance. The variance can also be used in determining the optimal step size h, as we will see in Section 4. By defining the deterministic error

$$\text{error}_d^{FFD} = \begin{pmatrix} \frac{f(x+he_1)-f(x)}{h} \\ \vdots \\ \frac{f(x+he_n)-f(x)}{h} \end{pmatrix} - \nabla f(x)$$

and the stochastic error

$$\text{error}_s^{FFD} = \begin{pmatrix} \frac{\varepsilon(x+he_1)-\varepsilon(x)}{h} \\ \vdots \\ \frac{\varepsilon(x+he_n)-\varepsilon(x)}{h} \end{pmatrix}$$

we get

$$E\left(\left\|\hat{\beta}^{FFD} - \nabla f(x)\right\|^2\right) = \left\|\text{error}_d^{FFD}\right\|^2 + E\left(\left\|\text{error}_s^{FFD}\right\|^2\right).$$

From (3) we easily derive that

$$\left\|\text{error}_d^{FFD}\right\|^2 \leq \frac{1}{4}nh^2D_2^2,$$

in which D_2 is the maximal second order derivative of $f(x)$. Let us now analyze the stochastic error. The first part of the following theorem is well-known in the literature; see ([10]).

Theorem 1 *For FFD we have*

$$\begin{aligned} E\left(\left\|error_s^{FFD}\right\|^2\right) &= \frac{2n\sigma^2}{h^2} \\ VAR\left(\left\|error_s^{FFD}\right\|^2\right) &= \frac{n}{h^4}\left[n(M_4-\sigma^4)+M_4+3\sigma^4\right] \\ VAR\left(\left\|\hat{\beta}^{FFD}-\nabla f(x)\right\|^2\right) &\leq VAR\left(\left\|error_s^{FFD}\right\|^2\right)+2n\sigma^2D_2^2 \end{aligned}$$

in which M_4 is the fourth moment of $\varepsilon(x)$ in (1), i.e. $M_4 = E(\varepsilon(x)^4)$.

Proof. By defining $\varepsilon_i = \varepsilon(x+he_i)$, $i = 1, ..., n$, and $\varepsilon_0 = \varepsilon(x)$, we have for forward finite differencing

$$\begin{aligned} E(\left\|\text{error}_s^{FFD}\right\|^2) &= \frac{1}{h^2}E\left(\sum_i(\varepsilon_i-\varepsilon_0)^2\right) = \frac{1}{h^2}E\left(\sum_i(\varepsilon_i^2+\varepsilon_0^2-2\varepsilon_i\varepsilon_0)\right) \quad (7) \\ &= \frac{2n\sigma^2}{h^2}, \quad (8) \end{aligned}$$

which proves the first part of the theorem. Considering the second part, we have

$$VAR(\left\|\text{error}_s^{FFD}\right\|^2) = E(\left\|\text{error}_s^{FFD}\right\|^4) - E^2(\left\|\text{error}_s^{FFD}\right\|^2). \quad (9)$$

Let us now concentrate on the first term of the right-hand side of (9):

$$\begin{aligned} E(\left\|\text{error}_s^{FFD}\right\|^4) &= \frac{1}{h^4}E\left[\sum_i(\varepsilon_i^2+\varepsilon_0^2-2\varepsilon_i\varepsilon_0)\sum_j(\varepsilon_j^2+\varepsilon_0^2-2\varepsilon_j\varepsilon_0)\right] \\ &= \frac{1}{h^4}[E\left(\sum\varepsilon_i^2\sum\varepsilon_j^2\right)+E\left(\sum\varepsilon_i^2\sum\varepsilon_0^2\right)-2E\left(\sum\varepsilon_i^2\sum\varepsilon_j\varepsilon_0\right) \\ &\quad +E\left(\sum\varepsilon_0^2\sum\varepsilon_j^2\right)+E\left(\sum\varepsilon_0^2\sum\varepsilon_0^2\right)-2E\left(\sum\varepsilon_0^2\sum\varepsilon_j\varepsilon_0\right) \\ &\quad -2E\left(\sum\varepsilon_i\varepsilon_0\sum\varepsilon_j^2\right)-2E\left(\sum\varepsilon_i\varepsilon_0\sum\varepsilon_0^2\right)+4E\left(\sum\varepsilon_i\varepsilon_0\sum\varepsilon_j\varepsilon_0\right)] \\ &= \frac{1}{h^4}[(nM_4+n(n-1)\sigma^4)+n^2\sigma^4+0+n^2\sigma^4+n^2M_4+0+0+0+4n\sigma^4] \\ &= \frac{1}{h^4}[n^2(M_4+3\sigma^4)+n(M_4+3\sigma^4)] \end{aligned}$$

Substituting this result and the square of (8) into (9), we have the second part of the theorem. To prove the third part, first observe that

$$
\begin{aligned}
VAR\left(\left\|\hat{\beta}^{FFD}-\nabla f(x)\right\|^2\right) &= VAR(\left\|\text{error}_d^{FFD}+\text{error}_s^{FFD}\right\|^2) \\
&= E(\left\|\text{error}_d^{FFD}+\text{error}_s^{FFD}\right\|^4)-E^2(\left\|\text{error}_d^{FFD}+\text{error}_s^{FFD}\right\|^2) \\
&= \left\|\text{error}_d^{FFD}\right\|^4+E(\left\|\text{error}_s^{FFD}\right\|^4)+2\left\|\text{error}_d^{FFD}\right\|^2 E(\left\|\text{error}_s^{FFD}\right\|^2 \\
&\quad +4E((\text{error}_d^{FFD})^T\text{error}_s^{FFD})^2 \\
&\quad +4E\left(\left\|\text{error}_s^{FFD}\right\|^2(\text{error}_d^{FFD})^T\text{error}_s^{FFD}\right) \\
&\quad +4\left\|\text{error}_d^{FFD}\right\|^2 E(\left(\text{error}_d^{FFD}\right)^T\text{error}_s^{FFD}) \\
&\quad -\left\|\text{error}_d^{FFD}\right\|^4-2\left\|\text{error}_d^{FFD}\right\|^2 E(\left\|\text{error}_s^{FFD}\right\|^2) \\
&\quad -E^2(\left\|\text{error}_s^{FFD}\right\|^2) \\
&= VAR(\left\|\text{error}_s^{FFD}\right\|^2)+4E((\text{error}_d^{FFD})^T\text{error}_s^{FFD})^2+ \\
&\quad 4E\left(\left\|\text{error}_s^{FFD}\right\|^2(\text{error}_d^{FFD})^T\text{error}_s^{FFD}\right) \\
&= VAR(\left\|\text{error}_s^{FFD}\right\|^2)+4\sum(\text{error}_d^{FFD})_i^2E(\text{error}_s^{FFD})_i^2+ \\
&\quad 4\sum(\text{error}_d^{FFD})_iE(\text{error}_s^{FFD})_i^3. \qquad (10)
\end{aligned}
$$

Further note that

$$\sum(\text{error}_d^{FFD})_i^2E(\text{error}_s^{FFD})_i^2\leq n\left(\frac{1}{4}h^2D_2^2\right)\left(\frac{2\sigma^2}{h^2}\right)=\frac{1}{2}n\sigma^2D_2^2 \qquad (11)$$

and

$$\sum(\text{error}_d^{FFD})_iE(\text{error}_s^{FFD})_i^3=0, \qquad (12)$$

since

$$E(\text{error}_s^{FFD})_i^3=\frac{1}{h^3}E(\varepsilon_i-\varepsilon_0)^3=\frac{1}{h^2}E\left(\sum_i(\varepsilon_i^3-3\varepsilon_i^2\varepsilon_0+3\varepsilon_i\varepsilon_0^2-\varepsilon_0^3)\right)=0. \qquad (13)$$

Finally, substituting (11) and (12) into (10) results into the third part of the theorem ■

2.2 Central finite differencing

A variant of the forward finite differencing (FFD) is the central finite differencing (CFD) approach. In this scheme, an estimation of the partial derivative, $\frac{\partial f(x)}{\partial x_i}$ $(i=1,\ldots,n)$, is obtained by

$$\hat{\beta}_i^{CFD}(h)=\frac{g(x+he_i)-g(x-he_i)}{2h}, \qquad h>0, \qquad (14)$$

where h is the step size and e_i is the i-th unit vector. Using (1) we can rewrite the estimate as

$$\begin{aligned}
\hat{\beta}_i^{CFD} &= \frac{f(x+he_i) - f(x-he_i) + \varepsilon(x+he_i) - \varepsilon(x-he_i)}{2h} && (15)\\
&= \frac{\partial f(x)}{\partial x_i} + \frac{h^2}{12}\nabla^3 f(x+\zeta_1 he_i)[e_i,e_i,e_i] + \frac{h^2}{12}\nabla^3 f(x+\zeta_2 he_i)[e_i,e_i,e_i] && (16)\\
&\quad + \frac{\varepsilon(x+he_i) - \varepsilon(x-he_i)}{2h}, && (17)
\end{aligned}$$

where the last equality follows from Taylor's formula

$$f(x+he_i) = f(x) + h\frac{\partial f(x)}{\partial x_i} + \frac{1}{2}h^2 e_i^T \nabla^2 f(x) e_i + \frac{h^3}{6}\nabla^3 f(x+\zeta_1 he_i)[e_i,e_i,e_i]$$

in which $0 \leq \zeta_1 \leq 1$, and

$$f(x-he_i) = f(x) - h\frac{\partial f(x)}{\partial x_i} + \frac{1}{2}h^2 e_i^T \nabla^2 f(x) e_i - \frac{h^3}{6}\nabla^3 f(x+\zeta_2 he_i)[e_i,e_i,e_i]$$

in which $0 \leq \zeta_2 \leq 1$. Let us first analyze the individual derivatives:

$$E[\hat{\beta}_i^{CFD}] = \frac{\partial f(x)}{\partial x_i} + O(h^2) \tag{18}$$

and

$$\mathrm{VAR}[\hat{\beta}_i^{CFD}] = \frac{\sigma^2}{2h^2}. \tag{19}$$

Contrary to the FFD estimations, the estimations $\hat{\beta}_i^{CFD}$ and $\hat{\beta}_j^{CFD}$ are not correlated:

$$\begin{aligned}
\mathrm{Cov}[\hat{\beta}_i^{CFD}, \hat{\beta}_j^{CFD}] &= E\left((\hat{\beta}_i^{CFD} - E[\hat{\beta}_i^{CFD}])(\hat{\beta}_j^{CFD} - E[\hat{\beta}_j^{CFD}])\right)\\
&= \frac{1}{h^2}E[(\varepsilon(x+he_i) - \varepsilon(x-he_i))(\varepsilon(x+he_j) - \varepsilon(x-he_j))]\\
&= 0, \qquad i \neq j.
\end{aligned}$$

We now analyze the mean squared error criterion

$$E\left(\left\|\hat{\beta}^{CFD} - \nabla f(x)\right\|^2\right),$$

and its variance

$$VAR\left(\left\|\hat{\beta}^{CFD} - \nabla f(x)\right\|^2\right).$$

By defining

$$\text{error}_d^{CFD} = \begin{pmatrix} \frac{f(x+he_1)-f(x-he_1)}{2h} \\ \vdots \\ \frac{f(x+he_n)-f(x-he_n)}{2h} \end{pmatrix} - \nabla f(x)$$

and

$$\text{error}_s^{CFD} = \begin{pmatrix} \frac{\varepsilon(x+he_1)-\varepsilon(x-he_1)}{2h} \\ \vdots \\ \frac{\varepsilon(x+he_n)-\varepsilon(x-he_n)}{2h} \end{pmatrix}$$

we get

$$E\left(\left\|\hat{\beta}^{CFD} - \nabla f(x)\right\|^2\right) = \left\|\text{error}_d^{CFD}\right\|^2 + E\left(\left\|\text{error}_s^{CFD}\right\|^2\right)$$

From (15) it is easy to verify that

$$\left\|\text{error}_d^{CFD}\right\|^2 \leq \frac{1}{36} n h^4 D_3^2,$$

in which D_3 is the maximal third order derivative of $f(x)$. Let us now analyze the stochastic error. The first part of the following theorem is well-known in the literature; see ([10]).

Theorem 2 *For CFD we have:*

$$\begin{aligned} E\left(\left\|error_s^{CFD}\right\|^2\right) &= \frac{n\sigma^2}{2h^2} \\ VAR\left(\left\|error_s^{CFD}\right\|^2\right) &= \frac{n}{8h^4}\left[M_4 + \sigma^4\right] \\ VAR\left(\left\|\hat{\beta}^{CFD} - \nabla f(x)\right\|^2\right) &\leq VAR(\left\|error_s^{CFD}\right\|^2) + \frac{1}{18} n h^2 \sigma^2 D_3^2. \end{aligned}$$

Proof. By defining $\varepsilon_i = \varepsilon(x + he_i)$, $\varepsilon_{-i} = \varepsilon(x - he_i)$, $i = 1, ..., n$, and $\varepsilon_0 = \varepsilon(x)$ we have for CFD

$$\begin{aligned} E(\left\|\text{error}_s^{CFD}\right\|^2) &= \frac{1}{4h^2} E\left(\sum_i (\varepsilon_i - \varepsilon_{-i})^2\right) = \frac{1}{4h^2} E\left(\sum_i (\varepsilon_i^2 + \varepsilon_{-i}^2 - 2\varepsilon_i \varepsilon_{-i})\right) \quad (20) \\ &= \frac{n\sigma^2}{2h^2}, \quad (21) \end{aligned}$$

which proves the first part of the theorem. For the variance we have:

$$VAR(\left\|\text{error}_s^{CFD}\right\|^2) = E(\left\|\text{error}_s^{CFD}\right\|^4) - E^2(\left\|\text{error}_s^{CFD}\right\|^2) \quad (22)$$

Let us now concentrate on the first term of the right-hand side in (22):

$$\begin{aligned}
E(\left\|\text{error}_s^{CFD}\right\|^4) &= \frac{1}{16h^4} E \sum_i (\varepsilon_i^2 + \varepsilon_{-i}^2 - 2\varepsilon_i\varepsilon_{-i}) \sum_j (\varepsilon_j^2 + \varepsilon_{-j}^2 - 2\varepsilon_j\varepsilon_{-j}) \\
&= \frac{1}{16h^4}[E\left(\sum \varepsilon_i^2 \sum \varepsilon_j^2\right) + E\left(\sum \varepsilon_i^2 \sum \varepsilon_{-j}^2\right) - 2E\left(\sum \varepsilon_i^2 \sum \varepsilon_j\varepsilon_{-j}\right) \\
&\quad + E\left(\sum \varepsilon_{-i}^2 \sum \varepsilon_j^2\right) + E\left(\sum \varepsilon_{-i}^2 \sum \varepsilon_{-j}^2\right) - 2E\left(\sum \varepsilon_{-i}^2 \sum \varepsilon_j\varepsilon_{-j}\right) \\
&\quad - 2E\left(\sum \varepsilon_i\varepsilon_{-i} \sum \varepsilon_j^2\right) - 2E\left(\sum \varepsilon_i\varepsilon_{-i} \sum \varepsilon_{-j}^2\right) + 4E\left(\sum \varepsilon_i\varepsilon_{-i} \sum \varepsilon_j\varepsilon_{-j}\right)] \\
&= \frac{1}{16h^4}[(nM_4 + n(n-1)\sigma^4) + n^2\sigma^4 + 0 + n^2\sigma^4 + (nM_4 + n(n-1)\sigma^4) \\
&\quad + 0 + 0 + 0 + 4n\sigma^4] \\
&= \frac{1}{8h^4}\left(nM_4 + n(2n+1)\sigma^4\right)
\end{aligned}$$

Substitution of this result and the square of (21) into formula (22) proves the second part of the theorem. The last part of the theorem follows similar as in the proof of the last part of the previous theorem:

$$\begin{aligned}
VAR\left(\left\|\hat{\beta}^{CFD} - \nabla f(x)\right\|^2\right) &= VAR(\left\|\text{error}_s^{CFD}\right\|^2) + 4\sum (\text{error}_d^{CFD})_i^2 E(\text{error}_s^{CFD})_i^2 + \\
&\quad 4\sum (\text{error}_d^{CFD})_i E(\text{error}_s^{CFD})_i^3 \\
&\leq VAR(\left\|\text{error}_s^{CFD}\right\|^2) + 4n(\frac{1}{36}h^4 D_3^2)(\frac{\sigma^2}{2h^2}) + 0 \\
&= VAR(\left\|\text{error}_s^{CFD}\right\|^2) + \frac{1}{18}nh^2\sigma^2 D_3^2.
\end{aligned}$$

This concludes the proof. ■

The result of this theorem can be simply checked for a special case. Suppose that all $\varepsilon(x)$ are standard normal distributed. Then by normalizing the stochastic error through the variance (see (19)), we know that

$$\frac{2h^2}{\sigma^2}\left\|\text{error}_s^{CFD}\right\|^2 \tag{23}$$

is the sum of n squared stochastic normally distributed variables, since $\left(\text{error}_s^{CFD}\right)_i = \varepsilon_i - \varepsilon_{-i}$ is normally distributed. Hence (23) is $\chi^2(n)$ distributed, with expectation n and variance $2n$. So, we get

$$E(\left\|\text{error}_s^{CFD}\right\|^2) = n\frac{\sigma^2}{2h^2},$$

which is exactly the result of the theorem. Furthermore,

$$\text{VAR}(\left\|\text{error}_s^{CFD}\right\|^2) = 2n\frac{\sigma^4}{4h^4} = \frac{n\sigma^4}{2h^4},$$

which also agrees with the result of the theorem, since for a normal distribution we have $M_4 = 3\sigma^4$.

2.3 Replicated central finite differencing

To decrease the stochastic error one can repeat central finite differencing K times. We call this replicated central finite differencing (RCFD). Of course the deterministic error will not change by doing replications. The next theorem shows the expectation and variance of the resulting stochastic error.

Theorem 3 *For RCFD we have:*

$$\begin{aligned} E\left(\left\|error_s^{RCFD}\right\|^2\right) &= \frac{n\sigma^2}{2h^2K} \\ VAR\left(\left\|error_s^{RCFD}\right\|^2\right) &= \frac{n}{8h^4K^3}\left[M_4+(4K-3)\sigma^4\right] \\ VAR\left(\left\|\hat{\beta}^{RCFD}-\nabla f(x)\right\|^2\right) &\leq VAR(\left\|error_s^{RCFD}\right\|^2)+\frac{1}{18K}nh^2\sigma^2D_3^2. \end{aligned}$$

Proof. By defining $\varepsilon_{ik}=\varepsilon_k(x+he_i)$, $\varepsilon_{-i,k}=\varepsilon_k(x-he_i)$, $i=1,...,n,\ \ k=1,...,K$ and $\varepsilon_{0k}=\varepsilon_k(x)$, where k denotes the k-th replicate, we have for RCFD

$$E(\left\|\text{error}_s^{RCFD}\right\|^2) = \frac{1}{4h^2K^2}E\left(\sum_i\left(\sum_k(\varepsilon_{ik}-\varepsilon_{-i,k})\right)^2\right) \tag{24}$$

$$= \frac{1}{4h^2K^2}E\left(\sum_{i,k,l}(\varepsilon_{ik}\varepsilon_{il}-\varepsilon_{-i,k}\varepsilon_{i,l}+\varepsilon_{-i,k}\varepsilon_{-i,l}-\varepsilon_{i,k}\varepsilon_{-i,l})\right) \tag{25}$$

$$= \frac{n\sigma^2}{2h^2K}, \tag{26}$$

which proves the first part of the theorem. For the variance we have:

$$\text{VAR}(\left\|\text{error}_s^{RCFD}\right\|^2) = E(\left\|\text{error}_s^{RCFD}\right\|^4) - E^2(\left\|\text{error}_s^{RCFD}\right\|^2) \tag{27}$$

Let us now concentrate on the first term of the right-hand side in (27):

$$\begin{aligned}
E(\|\text{error}_s^{RCFD}\|^4) &= \frac{1}{16h^4K^4}E\left(\sum_{i,k,l}(\varepsilon_{ik}\varepsilon_{il}-\varepsilon_{-i,k}\varepsilon_{i,l}+\varepsilon_{-i,k}\varepsilon_{-i,l}-\varepsilon_{i,k}\varepsilon_{-i,l})\right)^2 \\
&= \frac{1}{16h^4K^4}[E\left(\sum_{i,k,l}\varepsilon_{ik}\varepsilon_{il}\right)^2+E\left(\sum_{i,k,l}\varepsilon_{ik}\varepsilon_{il}\sum_{i,k,l}\varepsilon_{-i,k}\varepsilon_{-i,l}\right) \\
&\quad +E\left(\sum_{i,k,l}\varepsilon_{-i,k}\varepsilon_{i,l}\right)^2+E\left(\sum_{i,k,l}\varepsilon_{-i,k}\varepsilon_{i,l}\sum_{i,k,l}\varepsilon_{i,k}\varepsilon_{-i,l}\right) \\
&\quad +E\left(\sum_{i,k,l}\varepsilon_{-i,k}\varepsilon_{-i,l}\sum_{i,k,l}\varepsilon_{ik}\varepsilon_{il}\right)+E\left(\sum_{i,k,l}\varepsilon_{-i,k}\varepsilon_{-i,l}\right)^2 \\
&\quad +E\left(\sum_{i,k,l}\varepsilon_{i,k}\varepsilon_{-i,l}\sum_{i,k,l}\varepsilon_{-i,k}\varepsilon_{i,l}\right)+E\left(\sum_{i,k,l}\varepsilon_{i,k}\varepsilon_{-i,l}\right)^2] \\
&= \frac{1}{16h^4K^4}[[KnM_4+K^2n(n-1)\sigma^4+3nK(K-1)\sigma^4]+[K^2n^2\sigma^4] \\
&\quad +[K^2n\sigma^4]+[K^2n\sigma^4] \\
&\quad +[K^2n^2\sigma^4]+[KnM_4+K^2n(n-1)\sigma^4+3nK(K-1)\sigma^4] \\
&\quad +[K^2n\sigma^4]+[K^2n\sigma^4]] \\
&= \frac{1}{16h^4K^4}[2KnM_4+(2K^2n(n-1)+6nK(K-1)+2K^2n^2+4K^2n)\sigma^4]
\end{aligned}$$

Substitution of this result and the square of formula (26) into formula (27) proves the second part of the theorem. Finally, the third part can be derived almost identical as in the proof of the previous theorem. ■

3 Gradient estimation using DoE

3.1 Plackett-Burman

We now analyze Design of Experiments (DOE) schemes for estimating the gradient. Let us start with the Plackett-Burman scheme. Suppose that we have a set of vectors $d_k \in \mathbb{R}^n$ $(k = 1, \ldots, N)$ with $\|d_k\| = 1$ and that we observe $g(x + hd_k)$ for fixed $x \in \mathbb{R}^n$ and $h > 0$. Define the matrix

$$X := \begin{pmatrix} 1 & hd_1^T \\ \vdots & \vdots \\ 1 & hd_N^T \end{pmatrix}. \tag{28}$$

Now suppose that N, with $n+1 \leq N \leq n+4$, is a multiple of four. Then the Plackett-Burman scheme can be written as

$$H = \begin{pmatrix} 1 & p_1^T \\ \vdots & \vdots \\ 1 & p_N^T \end{pmatrix},$$

where $p_k \in \{-1, 1\}^n$. This so-called Hadamard matrix has the property $H^T H = NI$, where I is the identity matrix. For more information, see [1] or [8] and for an example see the Appendix.

Now let the vectors d_k in (28) be defined by

$$d_k = \frac{p_k}{\sqrt{n}}, \qquad k = 1, \ldots, N.$$

It then follows that

$$X^T X = \operatorname{diag}\left(N,\ \tfrac{h^2 N}{n}, \ldots, \tfrac{h^2 N}{n}\right).$$

The vector containing the function value of f at x and the gradient can be estimated by the OLS estimator

$$\begin{aligned} \begin{pmatrix} \hat{\beta}_0^{PB} \\ \hat{\beta}^{PB} \end{pmatrix} &= (X^T X)^{-1} X^T \begin{pmatrix} g(x + hd_1) \\ \vdots \\ g(x + hd_N) \end{pmatrix} \\ &= (X^T X)^{-1} X^T \begin{pmatrix} f(x + hd_1) \\ \vdots \\ f(x + hd_N) \end{pmatrix} + (X^T X)^{-1} X^T \begin{pmatrix} \varepsilon(x + hd_1) \\ \vdots \\ \varepsilon(x + hd_N) \end{pmatrix}. \end{aligned}$$

First note that

$$\begin{aligned} &E[\hat{\beta}_0^{PB}] = f(x) + O(h^2), \qquad V[\hat{\beta}_0^{PB}] = \tfrac{1}{n+1}\sigma^2 \\ &E[\hat{\beta}_i^{PB}] = \tfrac{\partial f(x)}{\partial x_i} + O(\sqrt{n}h), \quad V[\hat{\beta}_i^{PB}] = \tfrac{n\sigma^2}{(n+1)h^2}, \qquad i = 1, \ldots, n. \end{aligned} \tag{29}$$

Furthermore, since the columns of X are orthogonal, we have

$$\operatorname{Cov}[\hat{\beta}_i^{PB}, \hat{\beta}_j^{PB}] = 0, \qquad i \neq j.$$

Now defining D as the X matrix excluding the first column, and

$$\begin{aligned} \text{error}_d^{PB} &= \frac{n}{h^2 N} D^T \begin{pmatrix} f(x+hd_1) \\ \vdots \\ f(x+hd_N) \end{pmatrix} - \nabla f(x) \\ \text{error}_s^{PB} &= \frac{n}{h^2 N} D^T \begin{pmatrix} \varepsilon(x+hd_1) \\ \vdots \\ \varepsilon(x+hd_N) \end{pmatrix} \end{aligned}$$

we have

$$E\left(\left\|\hat{\beta}^{PB} - \nabla f(x)\right\|^2\right) = \left\|\text{error}_d^{PB}\right\|^2 + E\left(\left\|\text{error}_s^{PB}\right\|^2\right).$$

Let us now concentrate on the deterministice error. Using Taylor's formula

$$f(x+hd_k) = f(x) + hd_k^T \nabla f(x) + \frac{h^2}{2} d_k^T \nabla^2 f(x+\zeta h d_k) d_k.$$

in which $0 \leq \zeta \leq 1$, it is easy to derive that

$$\left\|\text{error}_d^{PB}\right\|^2 \leq \frac{n^2 h^2 D_2^2}{4}$$

in which D_2 is an overall upper bound for the second order derivative. Concerning the expectation and the variance of the stochastic error we have the following theorem.

Theorem 4 *For Plackett-Burman designs we have:*

$$\begin{aligned} E\left(\left\|error_s^{PB}\right\|^2\right) &= \frac{n^2\sigma^2}{Nh^2} \\ VAR\left(\left\|error_s^{PB}\right\|^2\right) &= \frac{n^4}{N^3 h^4}\left(M_4 + (\frac{2N}{n} - 3)\sigma^4\right) \\ VAR\left(\left\|\hat{\beta}^{PB} - \nabla f(x)\right\|^2\right) &\leq VAR(\left\|error_s^{PB}\right\|^2) + \frac{n^3\sigma^2 D_2^2}{N} \end{aligned}$$

Proof. For the Plackett-Burman schemes we have:

$$\text{error}_s^{PB} = \frac{n}{h^2 N} D^T \nu = \frac{\sqrt{n}}{hN} P^T \nu$$

in which P is the H matrix excluding the first column, and

$$\nu = \begin{pmatrix} \varepsilon(x+hd_1) \\ \vdots \\ \varepsilon(x+hd_N)) \end{pmatrix}.$$

We can now derive the following:

$$\begin{aligned} E(\left\|\text{error}_s^{PB}\right\|^2) &= \frac{n}{h^2N^2}E(\left\|P^T\nu\right\|^2) \quad (30)\\ &= \frac{n}{h^2N^2}nN\sigma^2 \\ &= \frac{n^2\sigma^2}{Nh^2} \quad (31)\end{aligned}$$

which proves the first part of the theorem. For the variance we have:

$$\text{VAR}(\left\|\text{error}_s^{PB}\right\|^2) = E(\left\|\text{error}_s^{PB}\right\|^4) - E^2(\left\|\text{error}_s^{PB}\right\|^2). \quad (32)$$

Let us now concentrate on the first term of the right-hand side of (32):

$$\begin{aligned} E(\left\|\text{error}_s^{PB}\right\|^4) &= \frac{n^2}{h^4N^4}E\left(\sum_j\left(\sum_i(P_{ij}\varepsilon_i)\sum_k(P_{kj}\varepsilon_k)\right)\right)^2 \\ &= \frac{n^2}{h^4N^4}E\left(\sum_{i,j,k}P_{ij}P_{kj}\varepsilon_i\varepsilon_k\right)^2 \\ &= \frac{n^2}{h^4N^4}E\left(\sum_{i,j,k,r,s,t}P_{ij}P_{kj}P_{sr}P_{tr}\varepsilon_i\varepsilon_k\varepsilon_s\varepsilon_t\right) \\ &= \frac{n^2}{h^4N^4}[2E\left(\sum_{i,j,k\neq i,r}P_{ij}P_{kj}P_{ir}P_{kr}\varepsilon_i^2\varepsilon_k^2\right) \\ &\quad +E\left(\sum_{i,j,s\neq i,r}P_{ij}P_{ij}P_{sr}P_{sr}\varepsilon_i^2\varepsilon_s^2\right) \\ &\quad +E\left(\sum_{i,j,r}P_{ij}P_{ij}P_{ir}P_{ir}\varepsilon_i^4\right)], \end{aligned}$$

where the last equality holds since the expectations of the terms $\varepsilon_i\varepsilon_k\varepsilon_s\varepsilon_t$, $\varepsilon_i^3\varepsilon_k$ and $\varepsilon_i^2\varepsilon_k\varepsilon_s$ are zero. We now concentrate on the three terms in the last equality. For the first term we have

$$\begin{aligned} E\left(\sum_{i,j,k\neq i,r}P_{ij}P_{kj}P_{ir}P_{kr}\varepsilon_i^2\varepsilon_k^2\right) &= \left(\sum_{i,j,r\neq j}P_{ij}P_{ir}\sum_{k\neq i,}P_{kj}P_{kr}\right)\sigma^4 + \left(\sum_{i,j}P_{ij}P_{ij}\sum_{k\neq i,}P_{kj}P_{kj}\right)\sigma^4 \\ &= \sum_{i,j,r\neq j}P_{ij}P_{ir}(-P_{ij}P_{ir}) + n(N-1)\sigma^4 \\ &= -n(n-1)N + n(N-1)\sigma^4 \\ &= nN(N-n)\sigma^4. \end{aligned}$$

Moreover, for the second term it holds

$$E\left(\sum_{i,j,s\neq i,r}P_{ij}P_{ij}P_{sr}P_{sr}\varepsilon_i^2\varepsilon_s^2\right) = n^2N(N-1)\sigma^4.$$

For the third term we have:

$$E\left(\sum_{i,j,r} P_{ij}P_{ij}P_{ir}P_{ir}\varepsilon_i^4\right) = n^2 N M_4.$$

Substituting these results and the square of (31) into (32), we have proved the second part of the theorem. The third part of the theorem follows similar as in the proof of the last part of Theorem 1:

$$\begin{aligned} VAR\left(\left\|\hat{\beta}^{PB} - \nabla f(x)\right\|^2\right) &= VAR(\left\|\text{error}_s^{BP}\right\|^2) + 4\sum(\text{error}_d^{PB})_i^2 E(\text{error}_s^{PB})_i^2 + \\ &\quad 4\sum(\text{error}_d^{PB})_i E(\text{error}_s^{PB})_i^3 \\ &\leq VAR(\left\|\text{error}_s^{PB}\right\|^2) + 4n(\frac{1}{4}nh^2D_2^2)(\frac{n\sigma^2}{Nh^2}) + 0 \\ &= VAR(\left\|\text{error}_s^{PB}\right\|^2) + \frac{n^3\sigma^2D_2^2}{N}. \end{aligned}$$

This concludes the proof. ■

3.2 Factorial designs

Factoral designs are based on the same principle as Plackett-Burman scheme, but now $N = 2^n$ for full factorial designs and $N = 2^{n-p}$, $p \leq n$, for fractional factorial designs; for more information see [1] or [8], and for an example see the Appendix.

For the deterministic error we can derive a better bound than for Plackett-Burman schemes. Again we have

$$\text{error}_d^{FD} = \frac{n}{h^2N}D^T\begin{pmatrix} f(x+hd_1) \\ \vdots \\ f(x+hd_N) \end{pmatrix} - \nabla f(x).$$

Now using Taylor's formula

$$f(x+hd_k) = f(x) + hd_k^T\nabla f(x) + \frac{h^2}{2}d_k^T\nabla^2 f(x)d_k + \frac{h^3}{6}\nabla^3 f(x+\zeta hd_k)[d_k,d_k,d_k], \quad (33)$$

in which $0 \leq \zeta \leq 1$, and using the fact that in factorial designs for each vector d_k there exists exactly one other vector d_j in the factorial design scheme such that $d_k = -d_j$, we obtain by adding these two vectors:

$$|f(x+hd_k) - f(x+hd_j)| \leq 2hd_k^T\nabla f(x) + \frac{h^3}{3}D_3$$

in which D_3 is an overall upper bound for the third order derivative. Combining all $N/2$ pairs of vectors we get

$$\begin{aligned} \left\|error_d^{FD}\right\|^2 &= \left\| \frac{n}{h^2 N} D^T \begin{pmatrix} f(x+hd_1) \\ \vdots \\ f(x+hd_N) \end{pmatrix} - \nabla f(x) \right\|^2 \\ &\leq \frac{n^2 h^4 D_3^2}{36}. \end{aligned}$$

Concerning the stochastig error we can derive the following results.

Theorem 5 *For factorial designs we have:*

$$\begin{aligned} E\left(\left\|error_s^{FD}\right\|^2\right) &= \frac{n^2\sigma^2}{Nh^2} \\ VAR\left(\left\|error_s^{FD}\right\|^2\right) &= \frac{n^4}{N^3h^4}\left(M_4 + (\frac{2N}{n} - 3)\sigma^4\right) \\ VAR\left(\left\|\hat{\beta}^{FD} - \nabla f(x)\right\|^2\right) &\leq VAR(\left\|error_s^{FD}\right\|^2) + \frac{1}{9N} n^3 h^2 \sigma^2 D_3^2. \end{aligned}$$

Proof. Concerning the first and second part we can derive the same results as for Plackett-Burman designs in the same way. We therefore omit the proof of these parts. The third part of the theorem follows similar as in the proof of the last part of Theorem 1:

$$\begin{aligned} VAR\left(\left\|\hat{\beta}^{FD} - \nabla f(x)\right\|^2\right) &= VAR(\left\|\text{error}_s^{FD}\right\|^2) + 4\sum(\text{error}_d^{FD})_i^2 E(\text{error}_s^{FD})_i^2 + \\ &\quad 4\sum(\text{error}_d^{FD})_i E(\text{error}_s^{FD})_i^3. \\ &\leq VAR(\left\|\text{error}_s^{FD}\right\|^2) + 4n(\frac{1}{36} n h^4 D_3^2)(\frac{n\sigma^2}{Nh^2}) + 0. \\ &= VAR(\left\|\text{error}_s^{FD}\right\|^2) + \frac{1}{9N} n^3 h^2 \sigma^2 D_3^2. \end{aligned}$$

■

4 Comparison of the five schemes

In the previous sections we have derived both the deterministic and the stochastic estimation errors for several schemes; see Table 1. The deterministic errors are increasing in the step size h, while the stochastic errors are decreasing in h. The expressions for the total error are convex functions in h. It is straightforward to calculate the optimal step

sizes for each scheme such that the total error is minimized. The results are mentioned in the last column of Table 1.

Of course, usually we do not know the values for σ, D_2 and D_3. However, for a practical problem we might estimate these values by sampling. Moreover, these optimal step sizes give some indication; e.g., the step sizes are increasing in σ and decreasing in N, D_2, and D_3, which agrees with our intuition.

	$\#eval$	$\|error_d\|^2$	$E(\|error_s\|^2)$	$opt.\ h_e$
Forward FD	$n{+}1$	$\frac{1}{4}nh^2D_2^2$	$\frac{2n\sigma^2}{h^2}$	$\sqrt[4]{8\frac{\sigma^2}{D_2^2}}$
Central FD	$2n$	$\frac{1}{36}nh^4D_3^2$	$\frac{n\sigma^2}{2h^2}$	$\sqrt[6]{9\frac{\sigma^2}{D_3^2}}$
Replicated CFD	$2nK$	$\frac{1}{36}nh^4D_3^2$	$\frac{n\sigma^2}{2h^2K}$	$\sqrt[6]{9\frac{\sigma^2}{KD_3^2}}$
Plackett-Burman	$n{+}1{\leq}N{\leq}n{+}4$	$\frac{1}{4}n^2h^2D_2^2$	$\frac{n^2\sigma^2}{Nh^2}$	$\sqrt[4]{4\frac{\sigma^2}{ND_2^2}}$
Factorial	$N{=}2^{n-p}$	$\frac{1}{36}n^2h^4D_3^2$	$\frac{n^2\sigma^2}{Nh^2}$	$\sqrt[6]{18\frac{\sigma^2}{ND_3^2}}$

Table 1: Overview of the number of evaluations and the errors for both finite difference and DoE schemes, and the optimal step sizes such that the total error is minimized.

	$VAR(\|error_s\|^2)$	$VAR(\|error_d{+}error_s\|^2)$	$opt.\ h_v$
Forward FD	$\frac{n}{h^4}[n(M_4{-}\sigma^4){+}M_4{+}3\sigma^4]$	$\frac{n}{h^4}[n(M_4{-}\sigma^4){+}M_4{+}3\sigma^4]{+}2n\sigma^2D_2^2$	—
Central FD	$\frac{n}{8h^4}(M_4{+}\sigma^4)$	$\frac{n}{8h^4}(M_4{+}\sigma^4){+}\frac{1}{18}nh^2\sigma^2D_3^2$	$\sqrt[6]{\frac{9(M_4+\sigma^4)}{2\sigma^2D_3^2}}$
Replicated CFD	$\frac{n}{8h^4K^3}[M_4{+}(4K{-}3)\sigma^4]$	$\frac{n}{8h^4K^3}[M_4{+}(4K{-}3)\sigma^4]{+}\frac{1}{18K}nh^2\sigma^2D_3^2$	$\sqrt[6]{\frac{9(M_4+(4K-3)\sigma^4)}{2K^2\sigma^2D_3^2}}$
Plackett-Burman	$\frac{n^4}{N^3h^4}\left(M_4{+}(\frac{2N}{n}{-}3)\sigma^4\right)$	$\frac{n^4}{N^3h^4}\left(M_4{+}(\frac{2N}{n}{-}3)\sigma^4\right){+}\frac{n^3\sigma^2D_2^2}{N}$	–
Factorial	$\frac{n^4}{N^3h^4}\left(M_4{+}(\frac{2N}{n}{-}3)\sigma^4\right)$	$\frac{n^4}{N^3h^4}\left(M_4{+}(\frac{2N}{n}{-}3)\sigma^4\right){+}\frac{1}{9N}n^3h^2\sigma^2D_3^2$	$\sqrt[6]{\frac{18n(M_4+(\frac{2N}{n}-3)\sigma^4)}{N^2\sigma^2D_3^2}}$

Table 2: Overview of the variances of the error vectors for both finite difference and DoE schemes and the optimal step sizes to minimize the variance.

From the literature we know that CFD gives a much lower deterministic error than FFD. Concerning the stochastic error we see from the table that the CFD scheme is four times better than FFD. However, the number of evaluations is two times more. To save evaluations, we can use a Plackett-Burman design: its number of evaluations is similar to the FFD scheme, but the stochastic error is two times lower; the deterministic error, however, is n times higher. Full or fractional factorial designs have a much

lower deterministic error than Plackett-Burman schemes. The stochastic error is similar, but since the number of evaluations is higher than for a Plackett-Burman scheme the stochastic error can be made much lower by increasing N. However this results in more evaluations. Observe also that the deterministic errors for Plackett-Burman and factorial schemes are independent of the number of evaluations, N. For the factorial schemes this also means that we can decrease the stochastic error by increasing N, without affecting the deterministic error. Concerning the variances of the stochastic errors it appears that CFD, Plackett-Burman and factorial schemes are much better than FFD. When comparing RCFD and factorial schemes it appears that the results are similar, since for a good comparison we have to take $N = 2nK$. Note, however, that in the case of numerical noise, e.g. in many deterministic simulation, RCFD is not applicable, since replicates will lead to the same outcomes. For such cases factorial schemes are useful.

In Table 2 we have listed the variance of the stochastic errors and the total errors. Note that in the calculations for the optimal step sizes h_e in Table 1 the variances of the errors are not taken into account. One can also determine a different step size by e.g. minimizing the expected error plus a certain number times the standard deviation. It can easily be verified that this will increase the optimal step sizes h. In the last column of Table 2 we have calculated the optimal step size such that the total variance is minimized. This calculation is not possible for FFD and Plackett-Burman since those variances are decreasing functions in h. The optimal stepsizes h_v for the other schemes resemble the corresponding h_e. Suppose for example that all $\varepsilon(x)$ are standard normal distributed, then it can easily be verified that $h_v = \sqrt[6]{2}h_e \approx 1.1h_e$, since then $M_4 = 3\sigma^4$. This means that the step size h_e which minimizes the total error equals approximately the step size which minimizes the upper bound for the variance of the error. This property is an advantage of the schemes CFD, RCFD and FD above CFD and PB.

In this paper we focus on the estimation of gradients. However, note that CFD, Plackett-Burman, and factorial schemes also deliver better estimations for the function value. These better estimations can also be valuable for NLP solvers.

Concerning the amount of work needed to calculate the gradient estimation, we emphasize that the estimations based on the DoE schemes need nN additions/subtractions and n multiplications, while FFD and CFD need n additions/subtractions and n multiplications and RCFD needs nK additions/subtractions and n multiplications. So, the extra amount of work needed in DOE schemes is limited

5 Conclusions

In the previous sections we have discussed several methods for estimating the gradient of a function that is subject to i.i.d. random errors. The error that we make when estimating the gradient can be split into two parts: a deterministic error and a stochastic error. The deterministic error arises because we do not observe the function exactly at x, but in the neighborhood of x using finite step sizes h. The stochastic error arises because of the noise. We have derived upper bounds for both the deterministic and stochastic errors. Based on these upper bounds we have discussed the advantages and disadvantages of three finite difference schemes and two DoE schemes.

The conclusion is that when the underlying function is indeed noisy the (fractional or full) factorial DoE schemes are useful to reduce the stochastic error. Such schemes do not vary the variables one at a time, but vary all variables simultaneously. The errors for factorial schemes are exactly the same as for replicated central finite differences, but in case of numerical noise we can use factorial schemes while replicates are meaningless. Plackett-Burman schemes are useful when the evaluations are expensive. The stochastic errors of these schemes are two times lower than FFD, but the deterministic error is higher. Moreover, our error analysis indicates how to choose the step size h. It also shows that for CFD, RCFD and FD-schemes the step sizes which minimizes the total error, also minimizes the variance of the error. The DoE schemes can be easily included in the NLP solvers to estimate gradients.

Acknowledgement. We would like to thank our colleague Jack Kleijnen for his useful remarks on an earlier version of this paper, and Gul Gurkan for providing us with relevant literature.

References

[1] Box, G.E.P., W.G. Hunter, and J.S. Hunter (1987), Statistics for Experimenters, Wiley, New York.

[2] Blum, J.R. (1954), Multidimensional Stochastic Approximation Methods, Annals of Mathematical Statistics 25, 737-744.

[3] Dixon, L.C.W. (1994), On Automatic Differentiation and Continuous Optimization, NATO Advanced Study Institutes Series 434, 501-512.

[4] Donohue, J.M., E.C. Houck and R.H. Myers (1993), Simulation Designs for Controlling Second-Order Bias in First-Order Response Surfaces, Operations Research 41, 880-902.

[5] Donohue, J.M., E.C. Houck and R.H. Myers (1995), Simulation Designs for the Estimation of Quadratic Response Surface Gradients in the Presence of Model Misspecification, Management Science 41 (2), 244-262.

[6] Ermoliev, Y. (1980) Stochastic Quasigradient Methods, in: Y. Ermoliev and R.J-B. Wets, eds., Numerical Techniques for Stochastic Optimization, Springer Verlag, Chapter 6.

[7] Griewank, A. (1989) On Automatic Differentiation, in: M. Iri and K. Tanabe, eds., Mathematical Programming, KTK Scientific Publishers, Tokyo, 83-107.

[8] Kiefer, J. and J. Wolfowitz (1952), Stochastic Estimation of a Regression Function, Annals of Mathematical Statistics 23, 462-466.

[9] Montgomery, D.C. (1984), Design and Analysis of Experiments, 2nd edition, Wiley, New York.

[10] Zazanis, M.A. and R. Suri (1993) Convergence Rates of Finite-Difference Sensitivity Estimates for Stochastic Systems, Operations research 41 (4), 694-703.

Appendix: DoE schemes

In Table 3, four evaluation schemes are given for $n = 4$. Note that for Plackett-Burman we have $N = 8$, which means that 8 evaluations are needed. In this case the number of evalations for Plackett-Burman is the same as for CFD; in general, however, the number of evaluations needed by CFD is more. Moreover, it is easy to verify that the orthogonality property holds for this specific full factorial and Plackett-Burman scheme. In fact, Plackett-Burman schemes were developed to reduce the number of evaluations, but such that the orthogonality property still holds. There is no need for tabulating the DoE schemes, since there is a simple procedure for generating such schemes.

	FFD				CFD				Plackett-Burman				Full factorial			
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
1	1	0	0	0	1	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
2	0	1	0	0	0	1	0	0	-1	1	-1	1	-1	-1	-1	1
3	0	0	1	0	0	0	1	0	-1	-1	1	1	-1	-1	1	-1
4	0	0	0	1	0	0	0	1	-1	1	1	-1	-1	-1	1	1
5					-1	0	0	0	-1	-1	-1	-1	-1	1	-1	-1
6					0	-1	0	0	-1	1	-1	1	-1	1	-1	1
7					0	0	-1	0	-1	-1	1	1	-1	1	1	-1
8					0	0	0	-1	-1	1	1	-1	-1	1	1	1
9													1	-1	-1	-1
10													1	-1	-1	1
11													1	-1	1	-1
12													1	-1	1	1
13													1	1	-1	-1
14													1	1	-1	1
15													1	1	1	-1
16													1	1	1	1

Table 3: Evaluation schemes for $n = 4$ factors.