CentER

Discussion Paper

No. 2003–101

GRADIENT ESTIMATION USING LAGRANGE INTERPOLATION POLYNOMIALS

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

October 2003

ISSN 0924-7815

TILBURG UNIVERSITY

Gradient estimation using Lagrange interpolation polynomials

Ruud Brekelmans
Tilburg University, Center Applied Research, Tilburg,

Lonneke Driessen
Centre for Quantitative Methods BV, Eindhoven,

Herbert Hamers [1]
Tilburg University, Department of Econometrics and OR, Tilburg,

Dick den Hertog
Tilburg University, Department of Econometrics and OR, Tilburg.

October 28, 2003

Abstract:

In this paper we use Lagrange interpolation polynomials to obtain good gradient estimations. This is e.g. important for nonlinear programming solvers. As an error criterion we take the mean squared error. This error can be split up into a deterministic and a stochastic error. We analyze these errors using (N times replicated) Lagrange interpolation polynomials. We show that the mean squared error is of order $N^{-1+\frac{1}{2d}}$ if we replicate the Lagrange estimation procedure N times and use $2d$ evaluations in each replicate. As a result the order of the mean squared error converges to N^{-1} if the number of evaluation points increases to infinity. Moreover, we show that our approach is also useful for deterministic functions in which numerical errors are involved. Finally, we consider the case of a fixed budget of evaluations. For this situation we provide an optimal division between the number of replicates and the number of evaluations in a replicate.

Keywords: Gradient estimation, noisy function, Lagrange interpolation

[1] Corresponding author: Herbert Hamers, Tilburg University, Department of Econometrics and OR, P.O.Box 90153, 5000 LE Tilburg, The Netherlands, e-mail: H.J.M.Hamers@uvt.nl

1 Introduction

In this paper we estimate the gradient $\bigtriangledown f(x)$ of a function $f : I\!R^n \rightarrow I\!R$. The function f is not explicitly known and we cannot observe it exactly. All observations are the result of an evaluation of the function, which is subject to certain perturbations. These perturbations can be of stochastic nature (e.g. in discrete-event simulation) or numerical nature (e.g. deterministic simulation models are often noisy due to numerical errors).

Obviously, gradients play an important role in all kind of optimisation techniques. In most non-linear programming (NLP) codes first-order and even second-order derivatives are used. Sometimes these derivatives can be calculated symbolically: in recent years automatic differentiation has been developed, see e.g. Griewank (1989). Although this is becoming more and more popular, there are still many optimisation solvers which use e.g. finite differencing to obtain a good approximation of the gradient. See e.g. Gill et al. (1981) or Dennis and Schnabel (1989).

Finite differences schemes have also been applied and analysed for problems with stochastic functions. Kiefer and Wolfowitz (1952) were the first to describe the so-called stochastic (quasi)gradients; see also Blum (1954). Methods based on stochastic quasi gradients are still subject of much research; for an overview see Ermoliev (1980). It was shown that the estimation error by using optimal stepsizes is $O(N^{-\frac{1}{2}})$ for forward finite differencing and $O(N^{-\frac{2}{3}})$ for central finite differencing, in which N is the number of replicates; see Glynn (1989), Zazanis and Suri (1988), L'Ecuyer and Perron (1990) and L'Ecuyer (1991).

In this paper we will improve these convergence rates by extending the finite difference method. Instead of using two evaluations for each dimension, we use $2d$ evaluations. We use Langrange interpolation polynomials to obtain a good point estimate of the gradient of a function $f : I\!R^n \rightarrow I\!R$. More precisely, each partial derivative is estimated using an interpolating function $h(x) = a_0 + a_1 x + a_2 x^2 + ... + a_{2d-1} x^{2d-1}$ that equals f in $2d$ evaluated points in one coordinate direction of f, with d a positive integer. Then $h'(0) = a_1$ is an estimate for this partial derivative. We consider the errors in the gradient estimation both due the deterministic approximation error ('lack of fit') and the presence of noise. We provide bounds for both the deterministic and the stochastic error. We show that the convergence rate is $N^{-1+\frac{1}{2d}}$, where N is the number of replicates of the Lagrange interpolation. This improves the above mentioned convergence rates for finite differencing when $d \geq 2$. Note that $d = 1$, resulting into a linear Lagrange interpolation function, corresponds to the central finite difference method. Moreover, we provide some results in case we have a deterministic function in which numerical errors are involved. Finally, given a fixed budget of evaluations, we provide an optimal division between the number of replicates (N) and the number of evaluations in such a replicate ($2d$).

This paper is organized as follows. Section 2 discusses the estimate of the gradient using Lagrange polynomials. The replicated Lagrange polynomials and the behavior of the mean squared error are considered in Section 3. In Section 4 we consider the error of the gradient estimation if the function is deterministic. The optimal division between the number of replicates and the number of evaluations in such a replicate, if there is a fixed budget of evaluations, is discussed in Section 5. An illustrative example is provided in Section 6.

2 Gradient estimation of stochastic noisy functions using Lagrange polynomials

In this section we estimate the gradient of a $2d$ times continuously differentiable function $f : I\!R^n \rightarrow I\!R$ that is subject to stochastic noise using Lagrange interpolation polynomials. We provide an upper bound for the mean squared error.

Let $f : I\!R^n \rightarrow I\!R$ be a function subjected to stochastic noise. Hence, for a fixed $y \in I\!R^n$ we observe

$$g(y) = f(y) + \epsilon(y). \tag{1}$$

The error term $\epsilon(y)$ represents a random component. In this paper we assume that the error terms in (1) are i.i.d. random errors with $E[\epsilon(y)] = 0$ and $V[\epsilon(y)] = \sigma^2$. This assumption implies that the error terms do not depend on y. Note that g can also be a computer simulation model.

We will approximate $\frac{\partial f(y)}{\partial y_i}$,$(i = 1, ..., n)$ in a point $y \in I\!R^n$ using the approximation function g, defined in (1). Without loss of generality we take $y = (0, ..., 0)^T$. For convenience, let $I = \{-d, ..., -1, 1, ..., d\}$. Next, the function g is evaluated in the grid points $y_v^i = vhe_i$ for all $v \in I$, where $h > 0$ and e_i is the i-th unit vector of dimension n. Observe that the grid points are equidistant on each side of zero and that this distance is given by h (see Figure 1.1).

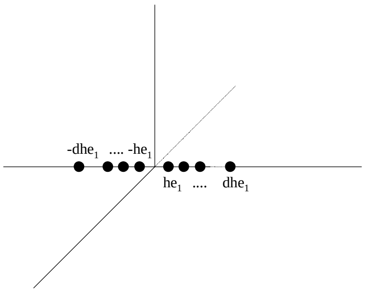

Figure 1.1: The $2d$ grid points for some h.

Now, take the interpolating polynomial $h_i : I\!R \rightarrow I\!R$ defined as

$$h_i(x) = a_0 + a_1 x + a_2 x^2 + ... + a_{2d-1} x^{2d-1} \tag{2}$$

that is exact in the evaluated points, i.e., according to (1) it holds that

$$h_i(x_v^i) = g(y_v^i) \;\; \text{for all } v \in I, \tag{3}$$

where $x_v^i = e_i^T y_v^i$. Obviously, $h_i'(0) = a_1$ is an estimate of $\frac{\partial f(0)}{\partial y_i}$.

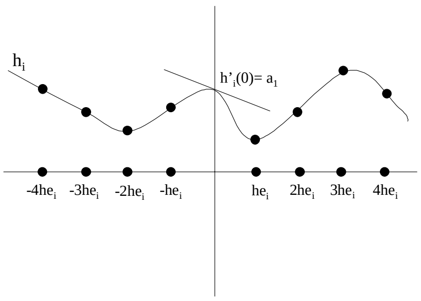

Figure 1.2: Estimate of gradient using interpolating polynomial.

Using the Lagrange functions $l_{v,i} : I\!R \rightarrow I\!R$ defined as

$$l_{v,i}(x) = \Pi_{u \in I \setminus \{v\}} \frac{x - x^i_u}{x^i_v - x^i_u},$$

for any $v \in I$, (2) can be rewritten into

$$h_i(x) = \sum_{v \in I} l_{v,i}(x) g(y^i_v). \tag{4}$$

Hence, the derivative of $h_i(x)$ equals

$$h'_i(x) = \sum_{v \in I} \left[l_{v,i}(x) g(y^i_v) \sum_{u \in I \setminus \{v\}} \frac{1}{x - x^i_u} \right]. \tag{5}$$

From (5) it follows that the estimate of the partial derivative is a linear combination of the evaluations. Observe that the corresponding coefficients only depend on the $2d$ evaluation points. Table 1.1 provides the coefficients for $2d = 2, 4, 6, 8, 10$, respectively. The example in Section 6 will illustrate the use of the coefficients in Table 1.1.

2d=2		2d=4		2d=6		2d=8		2d=10	
v = ih	coeff g(y^i_v)	v = ih	coeff g(y^i_v)	v = ih	coeff g(y^i_v)	v = ih	coeff g(y^i_v)	v = ih	coeff g(y^i_v)
-1	-0.5	-2	0.0833	-3	-0.0167	-4	0.0036	-5	-0.0008
1	0.5	-1	-0.6667	-2	0.1500	-3	-0.0381	-4	0.0099
		1	0.6667	-1	-0.7500	-2	0.2000	-3	-0.0595
		2	-0.0833	1	0.7500	-1	-0.8000	-2	0.2381
				2	-0.1500	1	0.8000	-1	-0.8333
				3	0.0167	2	-0.2000	1	0.8333
						3	0.0381	2	-0.2381
						4	-0.0036	3	0.0595
								4	-0.0099
								5	0.0008

Table 1.1: Coefficients to generate estimate partial derivative.

Obviously, we are interested in the quality of $h'_i(0)$ as estimate of the partial derivative $\frac{\partial f(0)}{\partial y_i}$. Therefore we define

$$h'_{i,1}(x) = \sum_{v \in I} \left[l_{v,i}(x) f(y^i_v) \sum_{u \in I \setminus \{v\}} \frac{1}{x - x^i_u} \right] \tag{6}$$

and

$$h'_{i,2}(x) = \sum_{v \in I} \left[l_{v,i}(x) \epsilon(y^i_v) \sum_{u \in I \setminus \{v\}} \frac{1}{x - x^i_u} \right]. \tag{7}$$

It follows that

$$h'_i(x) = h'_{i,1}(x) + h'_{i,2}(x). \tag{8}$$

A well-known measure for the quality of the estimate of the partial derivative $\frac{\partial f(0)}{\partial y_i}$ by $h'_i(0)$ is the mean squared error:

$$E \left(h'_i(0) - \frac{\partial f(0)}{\partial y_i} \right)^2 .$$

By defining the deterministic error

$$\left(\text{error}^{h'_i}_d \right)^2 = \left(h'_{i,1}(0) - \frac{\partial f(0)}{\partial y_i} \right)^2$$

and the stochastic error

$$\left(\text{error}^{h'_i}_s \right)^2 = E(\ h'_{i,2}(0))^2$$

we get, because $E[\epsilon(x)] = 0$, that

$$E \left(\ h'_i(0) - \frac{\partial f(0)}{\partial y_i}\ \right)^2 = \left(\text{error}^{h'_i}_d \right)^2 + \left(\text{error}^{h'_i}_s \right)^2 . \tag{9}$$

From (9) we learn that the mean squared error is the sum of the deterministic and the stochastic error. The following Lemma provides an upper bound for the deterministic error.

Lemma 2.1 For the Lagrange estimate we have

$$\left(error^{h'_i}_d \right)^2 \leq M^2_{2d} C^2_1(d) h^{4d-2},$$

where $C_1(d) = \frac{2}{(2d)!} \sum_{q=1}^{d} \left[q^{2d-1} \Pi_{r \in I \setminus \{q\}}\ \frac{|r|}{|r-q|} \right]$ and M_{2d} is an upper bound for the $2d$ order derivative of f.

PROOF:

For an upper bound of the deterministic error we use the Kowalewski's exact remainder for polynomial interpolation (cf. Davis (1975), pp. 72):

$$f_i(x) - h_{i,1}(x) = \frac{1}{(2d-1)!} \sum_{v \in I} l_{v,i}(x) \int_{x^i_v}^{x} (x^i_v - t)^{2d-1} f^{2d}(t) dt, \tag{10}$$

where f_i is the slice function of f taking the i^{th} component as variable. Taking the derivative to x on both sides of (10), substituting $x = 0$ and using $\mid f^{2d}(y) \mid \leq M_{2d}$ we obtain

$$error_d^{h'_i} \leq \frac{M_{2d}}{(2d)!} \sum_{v \in I} \left[\mid l'_{v,i}(0) \mid (x_v^i - 0)^{2d} \right],$$

where $l'_{v,i}(0) = \Pi_{u \in I \setminus \{v\}} \left[\frac{0 - x_u^i}{x_v^i - x_u^i} \right] \cdot \left[\sum_{u \in I \setminus \{v\}} \frac{1}{0 - x_u^i} \right]$.
Because

$$\mid l'_{v,i}(0) \mid = \left| \Pi_{u \in I \setminus \{v\}} \frac{0 - x_u^i}{x_v^i - x_u^i} \right| \frac{1}{\mid x_v^i \mid}.$$

and $x_u^i = hu$ for all $u \in I$, we have

$$\begin{aligned} error_d^{h'_i} &\leq \frac{M_{2d}}{(2d)!} 2 \sum_{q=1}^{d} \left[(qh)^{2d} \Pi_{r \in I \setminus \{q\}} \frac{\mid r \mid h}{\mid r - q \mid h} \cdot \frac{1}{qh} \right] \\ &= M_{2d} C_1(d) h^{2d-1}, \end{aligned}$$

which completes the proof. □

The next Lemma shows, as illustrated in Figure 1.3, that $C_1(d)$ converges to zero. Hence, $error_d^{h'_i}$ will also converge to zero, if M_{2d} is bounded.

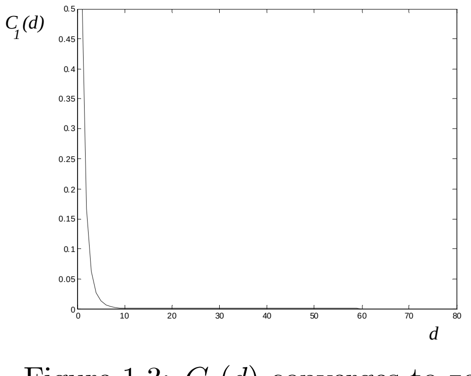

Figure 1.3: $C_1(d)$ converges to zero.

Lemma 2.2 Let $C_1(d) = \frac{2}{(2d)!} \sum_{q=1}^{d} \left[q^{2d-1} \Pi_{r \in I \setminus \{q\}} \frac{|r|}{|r-q|} \right]$. Then the following two statements hold:

$$(i) \;\; C_1(d) \leq 2d \left(\frac{3}{4 - \epsilon} \right)^d,$$

with $\epsilon > 0$ small,

$$(ii) \quad C_1(d) \to 0 \quad \text{if} \quad d \to \infty.$$

Proof: It is sufficient to prove (i). First observe that $C_1(d)$ can be rewritten into

$$C_1(d) = \frac{2(d!)^2}{(2d)!} \sum_{q=1}^{2d-1} \frac{q^{2d-1}}{(d+q)!(d-q)!}.$$

Let $a_d = \frac{(2d)!}{2(d!)^2}$. Then $a_{d+1} = \frac{(d+1)^2}{(2d+2)(2d+1)} a_d$. Hence, there exists a small $\epsilon > 0$ such that $a_d \geq (4-\epsilon) a_{d+1}$ for large d. This implies that there is a constant c such that for large d we have

$$a_d \geq c(4-\epsilon)^d. \tag{11}$$

Let $b_d = \sum_{q=1}^{d} \frac{q^{2d-1}}{(d+q)!(d-q)!}$. Then for each $q = 1, ..., d$ we have

$$\begin{aligned}
\frac{q^{2d-1}}{(d+q)!(d-q)!} &\leq q^{-1} \frac{q^{2d-1}}{\left(\frac{d+q}{3}\right)^{d+q} \left(\frac{d-q}{3}\right)^{d-q}} \\
&= 3^{2d} \left(\frac{d+q}{q}\right)^{-d-q} \left(\frac{d-q}{q}\right)^{-d+q} \\
&= 3^{2d} \left[\left(\frac{1+x}{x}\right)^{-1-x} \left(\frac{1-x}{x}\right)^{-1+x} \right]^d \\
&\leq 3^{2d} \cdot \left(\frac{1}{3}\right)^d = 3^d
\end{aligned}$$

where the first inequality follows from Stirlings formula and that $q^{-1} \leq 1$. In the second inequality we use that the continuous and concave function $z : (0,1] \to I\!R$ defined by $z(x) = \left(\frac{1+x}{x}\right)^{-1-x} \left(\frac{1-x}{x}\right)^{-1+x}$ is upper bounded by $\frac{1}{3}$. Hence, we can conclude that

$$b_d \leq d3^d. \tag{12}$$

From (11) and (12) it follows that for large d we have

$$C_1(d) \leq 2d\left(\frac{3}{4-\epsilon}\right)^d,$$

which completes the proof. □

The following Lemma provides an expression for the stochastic error.

Lemma 2.3 For the Lagrange estimate we have

$$\left(error_s^{h_i'}\right)^2 = C_2(d)\frac{\sigma^2}{h^2},$$

with $C_2(d) = 4\sum_{q=1}^{d}\left(\Pi_{r\in I\setminus\{q\}}\ \frac{|r|}{|r-q|}\frac{1}{q}\right)^2$.

Proof: We obtain

$$\begin{aligned}
\left(error_s^{h_i'}\right)^2 &= E(h_{i,2}'(0))^2 \\
&= E\left(\sum_{v\in I}\Pi_{u\in I\setminus\{v\}}\frac{0-x_u^i}{x_v^i-x_u^i}\epsilon(x_v^i)\sum_{u\in I\setminus\{v\}}\frac{1}{0-x_u^i}\right)^2 \\
&= \sigma^2\sum_{v\in I}\left(\Pi_{u\in I\setminus\{v\}}\frac{0-x_u^i}{x_v^i-x_u^i}\sum_{u\in I\setminus\{v\}}\frac{1}{0-x_u^i}\right)^2 \\
&= 4\frac{\sigma^2}{h^2}\sum_{q=1}^{d}\left(\Pi_{r\in I\setminus\{q\}}\ \frac{\mid r\mid}{\mid r-q\mid}\frac{1}{q}\right)^2
\end{aligned}$$

which completes the proof. □

The next lemma shows, as Figure 1.4 suggests, that $C_2(d)$ is upper bounded.

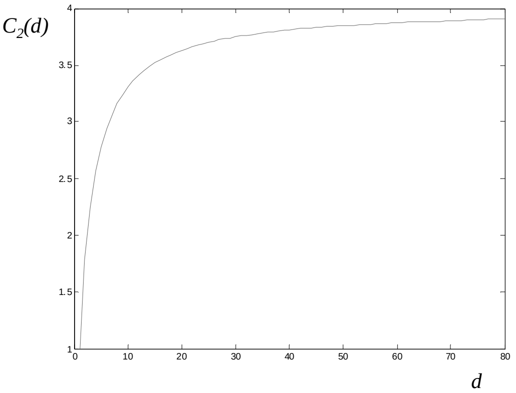

Figure 1.4: The behavior of $C_2(d)$ if the number of evaluation points increases.

Lemma 2.4 Let $C_2(d) = 4\sum_{q=1}^{d}\left(\Pi_{r\in I\setminus\{q\}}\ \frac{|r|}{|r-q|}\frac{1}{q}\right)^2$. Then $C_2(d) \le \frac{2}{3}\pi^2$ for all d.

Proof: Observe that $C_2(d) = 4\sum_{q=1}^{2d-1}\left(\frac{(d!)^2}{(d+q)!(d-q)!)}\frac{1}{q}\right)^2$. Because $\frac{(d!)^2}{(d+q)!(d-q)!)} \le 1$ for all q we have that $C_2(d) \le 4\sum_{q=1}^{2d-1}\frac{1}{q^2} \le 4\cdot\frac{1}{6}\pi^2 = \frac{2}{3}\pi^2$, which completes the proof. □

3 Derivative estimation of stochastic noisy functions using replicates

In this section we estimate the gradient of a $2d$ continuous differentiable function $f : I\!R^n \rightarrow I\!R$ that is subject to stochastic noise by replicating the Lagrange estimation of the previous sections. We investigate the mean squared error.

The following lemmata with respect to the deterministic and stochastic error follow straightforward from Lemma 2.1 and Lemma 2.3, respectively. Obviously, the upper bound for the deterministic error will not change in case of replicates.

Lemma 3.1 For the Lagrange estimation with N replicates we have

$$\left(error_d^{h_i'}\right)^2 \leq M_{2d}^2 C_1^2(d) h^{4d-2}.$$

Evidently, the stochastic error in case of replicates is decreased by a factor N, the number of replicates.

Lemma 3.2 For the Lagrange estimation with N replicates we have

$$\left(error_s^{h_i'}\right)^2 = C_2(d) \frac{\sigma^2}{Nh^2}.$$

In the final part of this section we determine the step size h that minimizes the mean squared error. From Lemma 3.1 and 3.2 it follows that the mean squared error, as a function of h, is upper bounded by

$$UMSE(h) = M_{2d}^2 C_1^2(d) h^{4d-2} + C_2(d) \frac{\sigma^2}{Nh^2}. \tag{13}$$

The following Theorem states the optimal step size and shows that the minimum mean squared error converges to N^{-1} if d goes to infinity.

Theorem 3.3 Let $UMSE(h)$ be defined as in (13). Then :

(i) The optimal step size h^* is $h^* = (PN)^{\frac{-1}{4d}}$ with $P = \left(\frac{C_2(d)\sigma^2}{M_{2d}^2 C_1^2(d)(2d-1)} \right)^{-1}$,

(ii) The minimum of $UMSE$ is $UMSE(h^*)^{\frac{-1}{4d}} = M_{2d}^{\frac{1}{d}} \sigma^{2-\frac{1}{d}} C_3(d) N^{-1+\frac{1}{2d}}$

with $C_3(d) = (C_1(d))^{\frac{1}{2d}} (C_2(d))^{1-\frac{1}{2d}} (2d-1)^{\frac{1}{2d}} \left(\frac{2d}{2d-1} \right)$,

$(iii)\ \ C_3(d) \leq 0.9\pi^2 \quad \text{for large } d,$

$(iv)\ \ UMSE(h^*) \rightarrow \mathcal{O}(N^{-1}) \quad \text{if} \quad d \rightarrow \infty.$

Proof:

The proof of (i) and (ii) is straightforward and (iv) results from (ii) and (iii). We will prove (iii). From Lemma 2.2 (i) it follows that $C_1(d)^{\frac{1}{2d}} = (2d)^{\frac{1}{2d}} \left(\frac{3}{4-\epsilon}\right)^{\frac{1}{2}}$. Because $(2d)^{\frac{1}{2d}}$ converges to 1, we have that $(2d)^{\frac{1}{2d}} \leq 1.1$ for large d and $\left(\frac{3}{4-\epsilon}\right)^{\frac{1}{2}} \leq 1$. Hence,

$$C_1(d)^{\frac{1}{2d}} \leq 1.1 \ \text{ if d is large.} \tag{14}$$

Obviously, it holds that

$$C_2(d)^{1-\frac{1}{2d}} \leq \frac{2}{3}\pi^2. \tag{15}$$

Because both $(2d-1)^{\frac{1}{2d}}$ and $\frac{2d}{2d-1}$ converge to 1 we have that both terms are upper bounded by 1.1 if d is large. Combining this last observation with (14) and (15) we obtain

$$C_3(d) \leq 1.1 \cdot \frac{2}{3}\pi^2 \cdot 1.1 \cdot 1.1 < 0.9\pi^2.$$

□

In Figure 3.1 the behavior of $C_3(d)$ is illustrated.

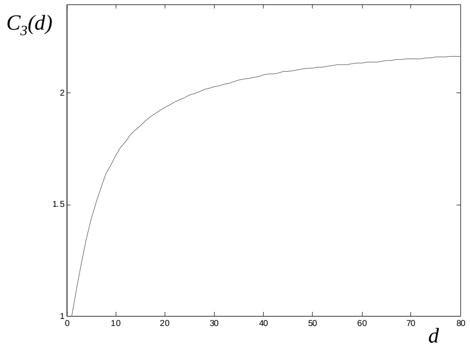

Figure 3.1: The behavior of $C_3(d)$.

Table 3.1 provides the UMSE for some specific values of d. Observe that already for small

d the best results in forward finite differencing ($\mathcal{O}(N^{-\frac{1}{2}})$) and central finite differencing ($\mathcal{O}(N^{-\frac{2}{3}})$) are improved. In fact, for $d = 1$ our result is identical to forward finite differencing.

d	$UMSE$
1	$1 \cdot M_2 \sigma N^{-\frac{1}{2}}$
2	$2.10(M_4)^{\frac{1}{2}} \sigma^{\frac{3}{2}} N^{-\frac{3}{4}}$
10	$1.68(M_{20})^{\frac{1}{10}} \sigma^{\frac{19}{10}} N^{-\frac{19}{20}}$
20	$1.94(M_{40})^{\frac{1}{20}} \sigma^{\frac{39}{20}} N^{-\frac{39}{40}}$
50	$2.12(M_{100})^{\frac{1}{50}} \sigma^{\frac{99}{50}} N^{-\frac{99}{100}}$

Table 3.1: The $UMSE$ for some values of d.

4 Gradient estimation of numerically noisy functions using Lagrange polynomials

In this section we estimate the gradient of a $2d$ times continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ that is subjected to numerical noise using Lagrange polynomials.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is subjected to numerical noise. Hence, for a fixed $y \in \mathbb{R}^n$ we observe

$$g(y) = f(y) + \epsilon(y),$$

where $\epsilon(y)$ is the fixed, unknown numerical error. To estimate the gradient of f we take the same approach as in section 1.2. Let the function h, $h'_{i,1}$ and $h'_{i,2}$ be defined as in (4), (6)and (7), respectively.

Then the total error of the estimate of the partial derivative is equal to

$$\left| \frac{\partial f(0)}{\partial y_i} - h'_i(0) \right| . \tag{16}$$

We define the deterministic model error by

$$\left| \frac{\partial f(0)}{\partial y_i} - h'_{i,1}(0) \right|$$

and the numerical error by

$$\left|h'_{i,2}(0)\right|.$$

We get, by using (8), the following upper bound for the total error

$$\left|\frac{\partial f(0)}{\partial y_i} - h'_i(0)\right| \leq \left|\frac{\partial f(0)}{\partial y_i} - h'_{i,1}(0)\right| + \left|h'_{i,2}(0)\right|. \tag{17}$$

Similarly to section 2.1 we can provide upper bounds for the deterministic model and the numerical error. The proofs of the following two Lemmata are omitted because they are almost identical to the proofs of Lemma 2.1 and 2.3, respectively.

Lemma 4.1 For the Lagrange estimate we have

$$\left|\frac{\partial f(0)}{\partial y_i} - h'_{i,1}(0)\right| \leq M_{2d}C_1(d)h^{2d-1}.$$

Lemma 4.2 For the Lagrange estimate we have

$$\left|h'_{i,2}(0)\right| \leq C_2(d)^{\frac{1}{2}}\frac{K}{h},$$

where K is an upper bound of ϵ.

In the final part of this section we determine the step size h that minimizes the total error. From 4.1 and 4.2 it follows that the total error TE, as a function of h, is upper bounded by

$$UTE(h) = M_{2d}C_1(d)h^{2d-1} + C_2(d)^{\frac{1}{2}}\frac{K}{h}. \tag{18}$$

The next Theorem provides the step size that minimizes the total error.

Theorem 4.3

(i) The optimal step size h^* is $h^* = \left(\frac{C_2(d)^{\frac{1}{2}}K}{(2d-1)^{2d-1}M_{2d}C_1(d)}\right)^{-\frac{1}{2d}}$,

(ii) The minimum of UTE is

$$UTE(h^*) = M_{2d}^{1-\frac{1}{2d}}C_1(d)^{\frac{1}{2d}}C_2(d)^{\frac{1}{2}-\frac{1}{4d}}K^{1-\frac{1}{2d}}(2d-1)^{\frac{1}{2d}}\left(\frac{2d}{2d-1}\right).$$

The proof is straightforward and is therefore omitted.

Observe that for the special case $d = 1$ that the result in Theorem 4.3 is similar to the result obtained in Gill et al. (1981), pp.340, for the forward finite-difference approximation.

5 Grid points versus replicates

In this section we provide an optimal division between the number of grid points and replicates in case the number of evaluations is fixed.

Let B be the total number of evaluations available, N the number of replicates and $2d$ the number of evaluations per replicate. The problem to solve is the following:

$$\begin{aligned} \text{minimize} \quad & UMSE(KN^{\frac{-1}{4d}}) = C_3(d)\left(\frac{M_{2d}}{\sigma}\right)^{\frac{1}{d}}\sigma^2 N^{-1+\frac{1}{2d}} \\ \text{subject to} \quad & B = 2dN, \\ & d, N \text{ positive integers.} \end{aligned} \tag{19}$$

In Table 5.1 we provide the optimal division between d and N for some values of B and a specific ratio of $\frac{M_{2d}}{\sigma}$.

σ = 1, M_{2d} = 1			
B	d	error	N
24	2	0.2879	6
804	3	0.0207	134
21984	4	0.0013	2748
386720	5	0.0001	38672
5461476	6	9.85E-06	455123

σ = 0.1, M_{2d} = 1			
B	d	error	N
4	2	0.0349	1
12	3	0.0148	2
240	4	0.0012	30
3880	5	0.0001	388
54660	6	9.85E-0.6	4555

σ = 0.01, M_{2d} = 1			
B	d	error	N
4	2	0.0011	1
12	3	0.0003	2
24	4	0.0001	3
40	5	0.0001	4
600	6	9.03E-06	50

σ = 10, M_{2d} = 1			
B	d	error	N
2376	2	0.2901	594
79932	3	0.0208	13322

Table 5.1: The optimal division between d and N at a fixed number of evaluations B

In the upper left cell of Table 4.1 we have chosen $\sigma = 1$ and $M_{2d} = 1$. This cell illustrates that for a fixed budget $B = 24$ till $B = 803$ it is optimal to evaluate 4,($i.e., d = 2$), points in each replicate. Obviously, in this case the number of replicates is determined by the quotient of the budget and 4. From $B = 804$ till $B = 21984$ it turns out that it is optimal to evaluate 6 points in each replicate. For example, if $B = 6000$ then we take $d = 3$, which equals 6 evaluations, and 1000 replicates. The other three cells of Table 4.1 present the results for different ratios of σ and M_{2d}. Observe that the error decreases if σ decreases. Moreover, the turning points to increase the number of grid points are also decreased if σ is decreased. For example, if $\sigma = 1$, then we turn to 6 grid points if $B = 804$, whereas if $\sigma = 0.01$ we already increase to 6 grid points if $B = 12$.

6 An illustrative example

The function under consideration is $f(y) = -1 + e^y$. We observe the function $g(y) = f(y) + \epsilon(y)$, where $\epsilon(y)$ is normal distributed with expectation $\mu = 0$ and standard deviation $\sigma = 0.01$. For the true derivative of the function f we have $f'(0) = 1$.

In Table 6.1 we compare the performance of CFD and our method, denoted by L. More precisely, we compare the average absolute error for CFD ($\mid e_{CFD} \mid$) and our method ($\mid e_L \mid$) for several simulation budgets. These averages are based on 1000 replications. The optimal values for d are derived from Table 4.1. The number of replications for our method, N_L, then equals $\frac{B}{2d}$, while for CFD we have $N_{CFD} = \frac{B}{2}$. The optimal step size h_L is calculated with formula $h_L = (PN)^{\frac{-1}{4d}}$ with $P = \left(\frac{C_2(d)\sigma^2}{M_{2d}^2 C_1^2(d)(2d-1)}\right)^{-1}$. For CFD we used the formula $h_{CFD} = \sqrt[6]{\frac{9\sigma^2}{M_3^2}}$ (cf. Brekelmans et al. (2003)) to determine the optimal step size. In both methods we used $M_{2d} = 1$ and $M_3 = 1$, respectively.

B	d	N_L	N_{CFD}	h_L	h_{CFD}	$\lvert e_L \rvert$	$\lvert e_{CFD} \rvert$	$\lvert e_{CFD} \rvert / \lvert e_L \rvert$
32	4	4	16	7.62E-01	1.96E-01	2.14E-04	6.40E-03	30
100	5	10	50	8.25E-01	1.62E-01	7.10E-05	4.38E-03	62
500	5	50	250	7.61E-01	1.24E-01	3.10E-05	2.56E-03	83
1200	6	100	600	8.18E-01	1.07E-01	1.45E-05	1.91E-03	132

Table 6.1: Comparison of the error and optimal step sizes for different simulation budgets between our method and CFD.

The last column of Table 6.1 shows that the absolute difference between the estimated derivative and the real derivative is smaller for our method than for CFD, and the larger the budget, the bigger the gap between the two methods. Hence, in a stochastic setting our method reduces the average absolute error.

Now let us look at the deterministic situation. CFD needs only two function evaluations, and using replications is useless as evaluating the same point more than once results in exactly the same function value each time. Table 6.2 shows the added value of our method when the evaluation budget is not limited to two evaluations only. We carried out the calculations for different values of h, namely $h = 0.01, h = 0.05$, and $h = 0.1$. The table shows that the error reduces significantly as can be expected from Lemma 2.1. The machine accuracy yields that for $h = 0.01$ we can evaluate at most 6 point, whereas in the cases $h = 0.05$ and $h = 0.1$ we can evaluate at most 10 points.

d	h =0.01	h=0.05	h=0.1
1 (=CFD)	1.67E-05	4.17E-04	1.67E-03
2	3.33E-10	2.08E-07	3.34E-06
3	4.89E-15	1.12E-10	7.16E-09
4		6.14E-14	1.59E-11
5		4.44E-16	3.69E-14

Table 6.2: Comparison of the error in deterministic setting between our method and CFD.

Acknowledgement: The authors thank Henk Norde for the proof of Lemma 2.2.

References

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

Brekelmans R., Driessen L., Hamers H., Den Hertog D., (2003) Gradient estimation schemes for noisy functions, CentER Discussion Paper 2003-12, Tilburg University, The Netherlands.

Davis P.J. (1975) Interpolation and Approximation, Dover Publications, New York.

Dennis J.E. and Schnabel R.B. (1989) A view of unconstrained optimization, in Handbook of Operations Research and Management Science, Volume 1, Optimization, G.L. Nemhauser et al. (eds), Elsevier Science Publishers B.V., North-Holland, Amsterdam.

L'Ecuyer P. (1991), An overview of derivative estimation, in: B.L. Nelson et al. (eds.), Proceedings of the 1991 Winter Simulation Conference, 207-217.

L'Ecuyer P. and Perron G.(1990), On the convergence rates of IPA and FDC derivative estimators for finite-horizon stochastic systems. Manuscript.

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

Gill P.E., Murray W. and Wright (1981), Practical Optimization, Academic Press, London.

Glynn P.W. (1989), Optimization of stochastic systems via simulation, in: E.A. MacNair et al. (eds.), Proceedings of the 1989 Winter Simulation Conference, 90-105.

Griewank A. (1989), On automatic differentiation, in: M.Iri and K Tanabe, eds., Mathematical Programming, KTK Scientific Publishers, Tokyo, 83-107.

Kiefer J. and Wolfowitz J. (1952), Stochastic estimation of a regression function, Annals of Mathematical Statistics, 23, 462-466.

Zazanis M.A. and Suri R. (1988), Comparison of perturbation analysis with conventional sensitivity estimates for stochastic systems. Manuscript.