Optim Lett (2015) 9:845–865
DOI 10.1007/s11590-014-0815-x

ORIGINAL PAPER

The calculus of simplex gradients

Rommel G. Regis

Received: 19 January 2014 / Accepted: 6 October 2014 / Published online: 17 October 2014

Abstract Simplex gradients are widely used in derivative-free optimization. This article clarifies some of the properties of simplex gradients and presents calculus rules similar to that of an ordinary gradient. For example, the simplex gradient does not depend on the order of sample points in the underdetermined and determined cases but this property is not true in the overdetermined case. Moreover, although the simplex gradient is the gradient of the corresponding linear model in the determined case, this is not necessarily true in the underdetermined and overdetermined cases. However, the simplex gradient is the gradient of an alternative linear model that is required to interpolate the reference data point. Also, the negative of the simplex gradient is a descent direction for any interpolating linear function in the determined and underdetermined cases but this is again not necessarily true for the linear regression model in the overdetermined case. In addition, this article reviews a previously established error bound for simplex gradients. Finally, this article treats the simplex gradient as a linear operator and provides formulas for the simplex gradients of products and quotients of two multivariable functions and a power rule for simplex gradients.

Keywords Simplex gradient · Derivative-free optimization · Black-box optimization · Linear interpolation and regression · Minimum norm least squares solution · Moore–Penrose pseudoinverse

1 Introduction

In the fully determined case, a simplex gradient of a function is the gradient of a linear model that interpolates data points on the surface of the function that corresponds to a

R. G. Regis (✉)
Department of Mathematics, Saint Joseph's University, Philadelphia,
PA 19131, USA
e-mail: rregis@sju.edu

maximal set of affinely independent points (i.e., a set of points in a simplex). The notion of a simplex gradient is widely used in derivative-free optimization. For example, it is used in the analysis of optimization methods for noisy problems that utilize function values on a sequence of simplices such as Nelder–Mead and implicit filtering (Bortz and Kelley [2], Kelley [8], Conn et al. [5]). Moreover, Custódio et al. [6] analyzed sequences of simplex gradients computed for nonsmooth functions in the context of direct search methods of the directional type such as Generalized Pattern Search (GPS) (Torczon [12]) and Mesh Adaptive Direct Search (MADS) (Audet and Dennis [1]). Simplex gradients have also been used to enhance the performance of pattern search by using them to reorder the objective function evaluations associated with the various poll directions (Custódio and Vicente [7]). In addition, in the benchmarking of derivative-free optimization algorithms, the data profile of a solver is defined as the percentage of problems solved for a given number of simplex gradient estimates (Moré and Wild [9]). More recently, Regis [10] used underdetermined simplex gradients to develop an initialization strategy for surrogate-based, high-dimensional expensive black-box optimization.

The purpose of this article is to clarify some of the properties of the simplex gradient and present calculus rules similar to that of an ordinary gradient. In particular, the simplex gradient does not depend on the order of sample points in the underdetermined and determined cases but this property does not hold in the overdetermined case. Moreover, as expected, the simplex gradient is the gradient of the corresponding linear interpolation model in the determined case. However, it is *not* necessarily the gradient of the linear model corresponding to the minimum norm least squares solution of the associated linear system in both the underdetermined and overdetermined cases. It turns out, though, that the simplex gradient is the gradient of an alternative linear model that is required to interpolate the reference data point. Also, in the underdetermined and determined cases, the negative of the simplex gradient with respect to a set of data points is shown to be a descent direction for any linear model that interpolates these data points. However, in the overdetermined case, the negative of the simplex gradient is *not* necessarily a descent direction for the corresponding linear regression model. Next, a previously established error bound for simplex gradients is reviewed. Furthermore, a convenient notation for the simplex gradient is introduced that treats it as a linear operator and some calculus rules such as product and quotient rules are proved. Finally, although the simplex gradient does not seem to satisfy a general chain rule, a power rule for simplex gradients is also proved.

2 Preliminaries

2.1 Definition of a simplex gradient

Throughout this article, f and g are functions from $\mathbb{R}^d$ to $\mathbb{R}$. Let $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ be an ordered set of $k+1$ points in $\mathbb{R}^d$, where $k \geq 1$. Define

$$S(\mathcal{X}) := [x_1 - x_0 \quad \ldots \quad x_k - x_0] \in \mathbb{R}^{d \times k} \quad \text{and} \quad \delta_f(\mathcal{X}) := \begin{bmatrix} f(x_1) - f(x_0) \\ \vdots \\ f(x_k) - f(x_0) \end{bmatrix} \in \mathbb{R}^k.$$

First, consider the case where $\mathcal{X}$ consists of $k+1$ affinely independent points (and so, $k \leq d$). When $k = d$ (the determined case), $S(\mathcal{X})$ is invertible and the *simplex gradient of f with respect to $\mathcal{X}$*, denoted by $\nabla_s f(\mathcal{X})$, is given by

$$\nabla_s f(\mathcal{X}) = S(\mathcal{X})^{-T} \delta_f(\mathcal{X}).$$

When $k < d$ (the underdetermined case), the *simplex gradient of f with respect to $\mathcal{X}$* is the minimum 2-norm solution of the linear system

$$S(\mathcal{X})^T \nabla_s f(\mathcal{X}) = \delta_f(\mathcal{X}),$$

which is given by $\nabla_s f(\mathcal{X}) = S(\mathcal{X})(S(\mathcal{X})^T S(\mathcal{X}))^{-1}\delta_f(\mathcal{X})$. In this case, note that $S(\mathcal{X})^T S(\mathcal{X})$ is symmetric and positive definite (and hence nonsingular) since $S(\mathcal{X})$ has full (column) rank. Moreover, $\nabla_s f(\mathcal{X})$ is a linear combination of $x_1 - x_0, \ldots, x_k - x_0$ since $\nabla_s f(\mathcal{X}) = S(\mathcal{X})v$, where $v = (S(\mathcal{X})^T S(\mathcal{X}))^{-1}\delta_f(\mathcal{X}) \in \mathbb{R}^k$.

Next, let $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ be an ordered set of $k+1$ distinct points in $\mathbb{R}^d$ that contains a *proper* subset of $d+1$ affinely independent points, and so, $k > d$ (the overdetermined case). In this case, $\{[1, x_0^T], [1, x_1^T], \ldots, [1, x_k^T]\}$ contains a subset of $d+1$ linearly independent points and it can be assumed that this subset contains $[1, x_0^T]$. Hence, $\mathcal{X}$ is guaranteed to have a proper subset of $d+1$ affinely independent points that includes x_0. This implies that $S(\mathcal{X})$ has full (row) rank, and so, $S(\mathcal{X})S(\mathcal{X})^T$ is symmetric and positive definite. Now the *simplex gradient of f with respect to $\mathcal{X}$* is the least squares solution of the linear system

$$S(\mathcal{X})^T \nabla_s f(\mathcal{X}) = \delta_f(\mathcal{X}),$$

which is given by $\nabla_s f(\mathcal{X}) = (S(\mathcal{X})S(\mathcal{X})^T)^{-1}S(\mathcal{X})\delta_f(\mathcal{X})$.

Custódio et al. [6] notes that the definitions of the simplex gradient for the three cases above can be combined into a single definition by considering a reduced SVD of $S(\mathcal{X})^T$ as shown next.

Definition 1 Let $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ be an ordered set of $k+1$ points in $\mathbb{R}^d$ where $k \geq 1$. Suppose $S(\mathcal{X})$ has full rank so that $\text{rank}(S(\mathcal{X})) = \min\{d, k\}$. Then the *simplex gradient of f with respect to $\mathcal{X}$* is given by

$$\nabla_s f(\mathcal{X}) = V(\mathcal{X})\Sigma(\mathcal{X})^{-1}U(\mathcal{X})^T \delta_f(\mathcal{X}),$$

where $U(\mathcal{X})\Sigma(\mathcal{X})V(\mathcal{X})^T$ is a reduced SVD of $S(\mathcal{X})^T$.

Note that $V(\mathcal{X})\Sigma(\mathcal{X})^{-1}U(\mathcal{X})^T$ is a reduced SVD of the Moore-Penrose pseudoinverse of $S(\mathcal{X})^T$, and so, the simplex gradient of f with respect to $\mathcal{X}$ can also be expressed as:

$$\nabla_s f(\mathcal{X}) = (S(\mathcal{X})^T)^\dagger \delta_f(\mathcal{X}) = (S(\mathcal{X})^\dagger)^T \delta_f(\mathcal{X}),$$

where $A^\dagger$ denotes the Moore–Penrose pseudoinverse of the matrix A.

2.2 Linear interpolation and regression

Consider a set $\mathcal{X} = \{x_0, x_1, \ldots, x_k\}$ of $k+1$ points in $\mathbb{R}^d$. If the points in $\mathcal{X}$ are affinely independent (which implies $k \leq d$), then there exists a linear function (an infinite number if $k < d$) that interpolates the data points $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_k, f(x_k))\}$. More precisely, if $m(x) = c_0 + c^T x$, where $c = [c_1, \ldots, c_d]^T$, is a linear polynomial in d variables that interpolates these data points, then

$$\begin{bmatrix} 1 & x_0^T \\ 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_k^T \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_d \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_k) \end{bmatrix}. \tag{1}$$

For convenience, define the $(k+1) \times (d+1)$ interpolation matrix $L(\mathcal{X})$ and the column vector $F(\mathcal{X})$ as follows:

$$L(\mathcal{X}) := \begin{bmatrix} 1 & x_0^T \\ 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_k^T \end{bmatrix} \quad \text{and} \quad F(\mathcal{X}) := \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_k) \end{bmatrix}$$

Equation (1) then becomes

$$L(\mathcal{X}) \begin{bmatrix} c_0 \\ c \end{bmatrix} = F(\mathcal{X}). \tag{2}$$

When $k > d$ and $L(\mathcal{X})$ has full (column) rank, Eq. (2) can be solved in a least squares sense. For convenience, the following definition from Conn, Scheinberg and Vicente [5] is used below.

Definition 2 Let $\mathcal{X} = \{x_0, x_1, \ldots, x_k\}$ be a set of $k+1$ points in $\mathbb{R}^d$. When $k = d$ (determined case), the set $\mathcal{X}$ is said to be *poised for linear interpolation in* $\mathbb{R}^d$ if $L(\mathcal{X})$ is nonsingular. When $k > d$ (overdetermined case), the set $\mathcal{X}$ is said to be *poised for linear regression in* $\mathbb{R}^d$ if $L(\mathcal{X})$ has full (column) rank.

If $\mathcal{X}$ consists of exactly $d+1$ affinely independent points (determined case), then $L(\mathcal{X})$ is nonsingular (and so $\mathcal{X}$ is poised for linear interpolation in $\mathbb{R}^d$) and the coefficients of the linear model are given by $\begin{bmatrix} c_0 \\ c \end{bmatrix} = L(\mathcal{X})^{-1} F(\mathcal{X})$. If $\mathcal{X}$ consists of $k+1$ affinely independent points, where $k < d$ (underdetermined case), then $L(\mathcal{X})$ has full row rank and the minimum 2-norm solution to Eq. (1) is given by $\begin{bmatrix} c_0 \\ c \end{bmatrix} = L(\mathcal{X})^T (L(\mathcal{X}) L(\mathcal{X})^T)^{-1} F(\mathcal{X})$. If $k > d$ and $\mathcal{X}$ contains $d+1$ affinely independent points (overdetermined case), then $L(\mathcal{X})$ has full column rank (and so $\mathcal{X}$ is poised for linear regression in $\mathbb{R}^d$) and the least squares solution to Eq. (1) is given

by $\begin{bmatrix} c_0 \\ c \end{bmatrix} = (L(\mathcal{X})^T L(\mathcal{X}))^{-1} L(\mathcal{X})^T F(\mathcal{X})$. As before, $\begin{bmatrix} c_0 \\ c \end{bmatrix} = L(\mathcal{X})^\dagger F(\mathcal{X})$ in all these cases. For a comprehensive treatment of the geometry of sample sets of points for interpolation (determined and underdetermined cases) and regression (overdetermined case) in the context of derivative-free optimization, the reader is referred to the papers by Conn, Scheinberg and Vicente ([3,4]) and Scheinberg and Toint [11].

The first proposition below relates poisedness for linear interpolation or regression with the existence of the simplex gradient in the determined and overdetermined cases.

Proposition 1 *Let $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ be an ordered set of $k+1$ points in $\mathbb{R}^d$ with $k \geq d$. Then $L(\mathcal{X})$ has full (column) rank if and only if $S(\mathcal{X})$ has full (row) rank.*

Proof Note that $\text{rank}(L(\mathcal{X})) = d+1$ if and only if $L(\mathcal{X})$ has $d+1$ linearly independent rows. Since the first nonzero row of any matrix can be extended to a maximal set of linearly independent rows of that matrix, it follows that $L(\mathcal{X})$ has full (column) rank if and only if $L(\mathcal{X})$ has $d+1$ linearly independent rows that include $[1, x_0^T]$, and this is true if and only if $\mathcal{X}$ has a subset of $d+1$ affinely independent points that include x_0. This implies that $L(\mathcal{X})$ has full (column) rank if and only if $S(\mathcal{X})$ has d linearly independent columns, or equivalently, $\text{rank}(S(\mathcal{X})) = d$. □

The above proposition says that $\mathcal{X}$ is poised for linear interpolation or linear regression (depending on whether $k = d$ or $k > d$) if and only if the simplex gradient $\nabla_s f(\mathcal{X})$ is defined.

3 Basic properties of simplex gradients

When $k \leq d$ (the determined and underdetermined cases), the following proposition, which was proved in Regis [10], shows that the simplex gradient $\nabla_s f(\mathcal{X})$ does not depend on the order of the points in $\mathcal{X}$. A slightly different proof from the one in Regis [10] is included below to make this article self-contained.

Proposition 2 *Suppose $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ is an ordered set of $k+1$ affinely independent points in $\mathbb{R}^d$, where $1 \leq k \leq d$. Let α be a permutation of the indices $\{0, 1, \ldots, k\}$ and let $\mathcal{X}_\alpha = \langle x_{\alpha(0)}, x_{\alpha(1)}, \ldots, x_{\alpha(k)} \rangle$. Then $\nabla_s f(\mathcal{X}_\alpha) = \nabla_s f(\mathcal{X})$.*

Proof First, consider the case where $\alpha(0) = 0$. Then

$$S(\mathcal{X}_\alpha) = [x_{\alpha(1)} - x_0 \quad \ldots \quad x_{\alpha(k)} - x_0] = S(\mathcal{X})P \qquad \text{and} \qquad \delta_f(\mathcal{X}_\alpha)^T = \delta_f(\mathcal{X})^T P,$$

for some permutation matrix P. Now

$$\begin{aligned}
\nabla_s f(\mathcal{X}_\alpha) &= S(\mathcal{X}_\alpha)(S(\mathcal{X}_\alpha)^T S(\mathcal{X}_\alpha))^{-1} \delta_f(\mathcal{X}_\alpha) \\
&= S(\mathcal{X})P((S(\mathcal{X})P)^T S(\mathcal{X})P)^{-1}(P^T \delta_f(\mathcal{X})) \\
&= S(\mathcal{X})P(P^T S(\mathcal{X})^T S(\mathcal{X})P)^{-1} P^T \delta_f(\mathcal{X})) \\
&= S(\mathcal{X})PP^{-1}(S(\mathcal{X})^T S(\mathcal{X}))^{-1}(P^T)^{-1} P^T \delta_f(\mathcal{X})) \\
&= S(\mathcal{X})(S(\mathcal{X})^T S(\mathcal{X}))^{-1} \delta_f(\mathcal{X}) = \nabla_s f(\mathcal{X}).
\end{aligned}$$

The previous argument shows that if the points in $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ are permuted except for the reference point x_0, then the simplex gradient remains the same. Next, observe that any permutation of $\{0, 1, \ldots, k\}$ that maps 0 to another element can be obtained from a permutation that fixes 0 by means of a single transposition. Hence, it only remains to show that the simplex gradient is preserved by a permutation of $\{0, 1, \ldots, k\}$ that switches 0 and another element while holding the other elements fixed.

Let α be a permutation of $\{0, 1, \ldots, k\}$ such that $\alpha(0) = j \neq 0, \alpha(j) = 0$, and that fixes all other elements. Note that $S(\mathcal{X}_\alpha) = [x_{\alpha(1)} - x_{\alpha(0)} \quad \ldots \quad x_{\alpha(k)} - x_{\alpha(0)}]$ can be transformed to $S(\mathcal{X}) = [x_1 - x_0 \quad \ldots \quad x_k - x_0]$ by applying a series of elementary column operations to $S(\mathcal{X}_\alpha)$. To see this, begin by multiplying the jth column of $S(\mathcal{X}_\alpha)$ by -1. The result is also given by $S(\mathcal{X}_\alpha)M$, where M is the elementary matrix obtained by replacing the jth diagonal entry of I_k by -1. Next, for each $i = 1, \ldots, k, \ i \neq j$, perform an elementary column operation that consist of adding the jth column of $S(\mathcal{X}_\alpha)M$ to the ith column and storing the result in the latter column. The result is some permutation of the columns of $S(\mathcal{X})$, and so,

$$S(\mathcal{X}_\alpha)ME_1E_2 \ldots E_{k-1}P = S(\mathcal{X}),$$

where P is a permutation matrix and $E_1, E_2, \ldots, E_{k-1}$ are the elementary matrices obtained by adding the jth column of I_k to the other columns and storing the results in those columns.

Let $F = ME_1E_2 \ldots E_{k-1}P$. Then $S(\mathcal{X}_\alpha)F = S(\mathcal{X})$ and F is nonsingular because it is the product of nonsingular matrices. Observe that

$$F^T\delta_f(\mathcal{X}_\alpha) = (ME_1E_2 \ldots E_{k-1}P)^T\delta_f(\mathcal{X}_\alpha) = P^T E_{k-1}^T \ldots E_2^T E_1^T M^T \delta_f(\mathcal{X}_\alpha) = \delta_f(\mathcal{X}).$$

Hence,

$$\begin{aligned}
\nabla_s f(\mathcal{X}) &= S(\mathcal{X})(S(\mathcal{X})^T S(\mathcal{X}))^{-1}\delta_f(\mathcal{X}) = S(\mathcal{X}_\alpha)F\ (\ (S(\mathcal{X}_\alpha)F)^T(S(\mathcal{X}_\alpha)F)\ )^{-1}\delta_f(\mathcal{X}) \\
&= S(\mathcal{X}_\alpha)F\ (\ F^T S(\mathcal{X}_\alpha)^T S(\mathcal{X}_\alpha)F\ )^{-1}\delta_f(\mathcal{X}) \\
&= S(\mathcal{X}_\alpha)FF^{-1}\ (\ S(\mathcal{X}_\alpha)^T S(\mathcal{X}_\alpha)\ )^{-1}(F^T)^{-1}\delta_f(\mathcal{X}) \\
&= S(\mathcal{X}_\alpha)(S(\mathcal{X}_\alpha)^T S(\mathcal{X}_\alpha))^{-1}\delta_f(\mathcal{X}_\alpha) = \nabla_s f(\mathcal{X}_\alpha).
\end{aligned}$$

□

When $k > d$ (the overdetermined case), $\nabla_s f(\mathcal{X})$ depends on the order of the points in $\mathcal{X}$ as can be seen from the following examples in $\mathbb{R}$ and $\mathbb{R}^2$.

Example 1 Consider $x_0 = -1, x_1 = 0, x_2 = 1$ and suppose $f(x_0) = 2$, $f(x_1) = 1$, $f(x_2) = 3$. Let $\mathcal{X} = \langle x_0, x_1, x_2 \rangle$ and consider the permutation of $\{0, 1, 2\}$ given by $\alpha = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \end{pmatrix}$. Then $\mathcal{X}_\alpha = \langle x_1, x_2, x_0 \rangle$. Note that $S(\mathcal{X})$ and $S(\mathcal{X}_\alpha)$ have full rank and

$$\begin{aligned}
\nabla_s f(\mathcal{X}) &= (S(\mathcal{X})S(\mathcal{X})^T)^{-1}S(\mathcal{X})\delta_f(\mathcal{X}) \\
&= \left(\begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 2 \end{bmatrix}^T\right)^{-1}\begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/5 \end{bmatrix}
\end{aligned}$$

and

$$\nabla_s f(\mathcal{X}_\alpha) = \left(\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix}^T \right)^{-1} \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/2 \end{bmatrix},$$

and so, $\nabla_s f(\mathcal{X}) \neq \nabla_s f(\mathcal{X}_\alpha)$.

Example 2 Consider $x_0 = [0, 0]^T, x_1 = [1, 0]^T, x_2 = [0, 1]^T, x_3 = [1, 2]^T$ and suppose $f(x_0) = 1, f(x_1) = 2, f(x_2) = 0, f(x_3) = 1$. Let $\mathcal{X} = \langle x_0, x_1, x_2, x_3 \rangle$ and consider the permutation of $\{0, 1, 2, 3\}$ given by $\alpha = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 0 & 1 \end{pmatrix}$. Then $\mathcal{X}_\alpha = \langle x_3, x_2, x_0, x_1 \rangle$. Note that $S(\mathcal{X})$ and $S(\mathcal{X}_\alpha)$ have full rank and

$$\begin{aligned} \nabla_s f(\mathcal{X}) &= (S(\mathcal{X})S(\mathcal{X})^T)^{-1} S(\mathcal{X}) \delta_f(\mathcal{X}) \\ &= \left(\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix}^T \right)^{-1} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 7/6 \\ -2/3 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \nabla_s f(\mathcal{X}_\alpha) &= \left(\begin{bmatrix} -1 & -1 & 0 \\ -1 & -2 & -2 \end{bmatrix} \begin{bmatrix} -1 & -1 & 0 \\ -1 & -2 & -2 \end{bmatrix}^T \right)^{-1} \begin{bmatrix} -1 & -1 & 0 \\ -1 & -2 & -2 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 4/3 \\ -5/9 \end{bmatrix}, \end{aligned}$$

and so, $\nabla_s f(\mathcal{X}) \neq \nabla_s f(\mathcal{X}_\alpha)$.

For convenience, call the first point in the ordered set $\mathcal{X}$ the *reference point*. The next proposition shows that, when $k > d$ (the overdetermined case), the simplex gradient is not affected by changing the order of the sample points that are not the reference point. That is, $\nabla_s f(\mathcal{X})$ depends only on which point in $\mathcal{X}$ is used as the reference point and not on the order of the other sample points. Note that although the DFO book by Conn et al. [5] and other papers (e.g., Custódio and Vicente [7]) did not explicitly mention the dependence of the simplex gradient on the reference point, the notation $\nabla_s f(x_0)$ in the book and in these other papers suggests that the authors were aware of this dependence.

Proposition 3 *Suppose $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ is an ordered set of $k + 1$ points in $\mathbb{R}^d$, with $k > d$, that contains a proper subset of $d + 1$ affinely independent points. Let α be a permutation of the indices $\{0, 1, \ldots, k\}$ such that $\alpha(0) = 0$ and let $\mathcal{X}_\alpha = \langle x_{\alpha(0)}, x_{\alpha(1)}, \ldots, x_{\alpha(k)} \rangle$. Then $\nabla_s f(\mathcal{X}_\alpha) = \nabla_s f(\mathcal{X})$.*

Proof Since $\alpha(0) = 0$, it follows that

$$S(\mathcal{X}_\alpha) = [x_{\alpha(1)} - x_0 \quad \ldots \quad x_{\alpha(k)} - x_0] = S(\mathcal{X})P \qquad \text{and} \qquad \delta_f(\mathcal{X}_\alpha)^T = \delta_f(\mathcal{X})^T P,$$

for some permutation matrix P. Now

$$\begin{aligned}\nabla_s f(\mathcal{X}_\alpha) &= (S(\mathcal{X}_\alpha)S(\mathcal{X}_\alpha)^T)^{-1}S(\mathcal{X}_\alpha)\delta_f(\mathcal{X}_\alpha)\\ &= (S(\mathcal{X})P(S(\mathcal{X})P)^T)^{-1}S(\mathcal{X})P(P^T\delta_f(\mathcal{X}))\\ &= (S(\mathcal{X})(PP^T)S(\mathcal{X})^T)^{-1}S(\mathcal{X})(PP^T)\delta_f(\mathcal{X}))\\ &= (S(\mathcal{X})S(\mathcal{X})^T)^{-1}S(\mathcal{X})\delta_f(\mathcal{X}) = \nabla_s f(\mathcal{X}).\end{aligned}$$

□

4 Simplex gradients and linear interpolation and regression models

Next, this section analyzes the relationship between the simplex gradient and the gradient of the corresponding linear model. When $\mathcal{X}$ consists of exactly $d+1$ affinely independent points, the following result mentioned in Conn et al. [5] shows that the simplex gradient $\nabla_s f(\mathcal{X})$ is the gradient of the unique linear function that interpolates the points in $\mathcal{X}$ and their function values.

Proposition 4 *Let $\mathcal{X} = \{x_0, x_1, \ldots, x_d\}$ be a set of $d+1$ affinely independent points in $\mathbb{R}^d$ (and so $\mathcal{X}$ is poised for linear interpolation in $\mathbb{R}^d$). Then $\nabla_s f(\mathcal{X})$ is the gradient of the unique linear function that interpolates the data points $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_d, f(x_d))\}$.*

Proof Let $m(x) = c_0 + c^T x$, where $c_0 \in \mathbb{R}$ and $c \in \mathbb{R}^d$, be the unique linear function that interpolates the data points $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_d, f(x_d))\}$. Then

$$c_0 + c^T x_i = f(x_i) \quad \text{for} \quad i = 0, 1, \ldots, d. \tag{3}$$

Subtracting $c_0 + c^T x_0 = f(x_0)$ from each of the equations in (3) for $i = 1, \ldots, d$ gives

$$c^T(x_i - x_0) = f(x_i) - f(x_0) \quad \text{for} \quad i = 1, \ldots, d.$$

Hence, c satisfies

$$c^T S(\mathcal{X}) = \delta_f(\mathcal{X})^T, \quad \text{or equivalently,} \quad S(\mathcal{X})^T c = \delta_f(\mathcal{X}).$$

Since $\mathcal{X}$ consists of $d+1$ affinely independent points, $S(\mathcal{X})^T$ is nonsingular and

$$\nabla m(x) = c = S(\mathcal{X})^{-T}\delta_f(\mathcal{X}) = \nabla_s f(\mathcal{X}).$$

□

Next, consider the overdetermined case. Let $\mathcal{X} = \langle x_0, x_1, \ldots, x_k\rangle$ be an ordered set of $k+1$ distinct points in $\mathbb{R}^d$ that contains a proper subset of $d+1$ affinely independent points (and so $k > d$) and let $m(x) = c_0 + c^T x$ be the linear regression model for

the data points $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_k, f(x_k))\}$. Recall that the gradient $\nabla m(x) = c$ is obtained from the least squares solution of

$$\begin{bmatrix} 1 & x_0^T \\ 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_k^T \end{bmatrix} \begin{bmatrix} c_0 \\ c \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_k) \end{bmatrix} \tag{4}$$

while the simplex gradient $\nabla_s f(\mathcal{X})$ is the least squares solution of

$$\begin{bmatrix} (x_1 - x_0)^T \\ (x_2 - x_0)^T \\ \vdots \\ (x_k - x_0)^T \end{bmatrix} \nabla_s f(\mathcal{X}) = \begin{bmatrix} f(x_1) - f(x_0) \\ f(x_2) - f(x_0) \\ \vdots \\ f(x_k) - f(x_0) \end{bmatrix}. \tag{5}$$

The examples below show that $\nabla_s f(\mathcal{X})$ is not necessarily equal to $\nabla m(x) = c$ in the overdetermined case. In fact, Example 4 below shows that none of the simplex gradients $\nabla_s f(\mathcal{X})$ using all possible reference points have to equal $\nabla m(x) = c$.

Example 3 Consider $\mathcal{X}$ and $F(\mathcal{X})$ from Example 1: $x_0 = -1, x_1 = 0, x_2 = 1$ and $f(x_0) = 2, f(x_1) = 1, f(x_2) = 3$. Then $\nabla_s f(\mathcal{X}) = [1/5]$. Now the coefficients of the linear regression model $m(x) = c_0 + c^T x$ are given by

$$\begin{aligned} \begin{bmatrix} c_0 \\ c \end{bmatrix} &= (L(\mathcal{X})^T L(\mathcal{X}))^{-1} L(\mathcal{X})^T F(\mathcal{X}) \\ &= \left(\begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}^T \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}^T \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix} \end{aligned}$$

Hence, $\nabla m(x) = c = [1/2] \neq \nabla_s f(\mathcal{X})$.

Example 4 Consider $\mathcal{X}$ and $F(\mathcal{X})$ from Example 2: $x_0 = [0, 0]^T, x_1 = [1, 0]^T, x_2 = [0, 1]^T, x_3 = [1, 2]^T$ and $f(x_0) = 1, f(x_1) = 2, f(x_2) = 0, f(x_3) = 1$. Then $\nabla_s f(\mathcal{X}) = \begin{bmatrix} 7/6 \\ -2/3 \end{bmatrix}$. Now the coefficients of the linear regression model $m(x) = c_0 + c^T x$ are given by

$$\begin{aligned} \begin{bmatrix} c_0 \\ c \end{bmatrix} &= (L(\mathcal{X})^T L(\mathcal{X}))^{-1} L(\mathcal{X})^T F(\mathcal{X}) \\ &= \left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}^T \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 4/5 \\ 13/10 \\ -3/5 \end{bmatrix} \end{aligned}$$

Hence, $\nabla m(x) = c = [13/10, -3/5]^T \neq \nabla_s f(\mathcal{X})$.

By similar calculations, the simplex gradients obtained by using x_1, x_2 and x_3 as reference points are $[11/9, -5/9]^T$, $[3/2, -2/3]^T$ and $[4/3, -5/9]^T$, respectively. Note that none of these simplex gradients are equal to $\nabla m(x) = c = [13/10, -3/5]^T$.

The fact that Proposition 4 does not hold for the overdetermined case is not really surprising considering that Examples 1 and 2 and Proposition 3 showed that $\nabla_s f(\mathcal{X})$ depends on which point in $\mathcal{X}$ is chosen as the reference point whereas the linear regression model and its gradient are fixed for a given $\mathcal{X}$ containing a subset of $d+1$ affinely independent points.

Finally, in the underdetermined case ($k < d$), $\nabla_s f(\mathcal{X})$ is also *not* the gradient of the linear model whose coefficients are the minimum 2-norm solution to Eq. (1) as can be seen from the following counterexample.

Example 5 Consider $x_0 = [1, 0, 0]^T$, $x_1 = [0, 1, 0]^T$, $x_2 = [0, 0, 1]^T$ and suppose $f(x_0) = 2$, $f(x_1) = 0$, $f(x_2) = 1$. Let $\mathcal{X} = \langle x_0, x_1, x_2 \rangle$. Then

$$\begin{aligned}\nabla_s f(\mathcal{X}) &= S(\mathcal{X})(S(\mathcal{X})^T S(\mathcal{X}))^{-1}\delta_f(\mathcal{X}) \\ &= \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \left(\begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}^T \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} -2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.\end{aligned}$$

On the other hand, the minimum 2-norm solution to Eq. (1) is given by

$$\begin{aligned}\begin{bmatrix} c_0 \\ c \end{bmatrix} &= L(\mathcal{X})^T (L(\mathcal{X})L(\mathcal{X})^T)^{-1} F(\mathcal{X}) \\ &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}^T \left(\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}^T \right)^{-1} \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3/4 \\ 5/4 \\ -3/4 \\ 1/4 \end{bmatrix}.\end{aligned}$$

Again, $c = [5/4, -3/4, 1/4]^T \neq \nabla_s f(\mathcal{X})$.

The next proposition shows that, in the underdetermined and determined cases, $-\nabla_s f(\mathcal{X})$ is a descent direction for any linear function that interpolates the points in $\mathcal{X}$ and their function values. In particular, although $\nabla_s f(\mathcal{X})$ is not necessarily equal to the gradient of the linear model corresponding to the minimum 2-norm solution to Eq. (1) in the underdetermined case, this proposition says that $-\nabla_s f(\mathcal{X})$ is always a descent direction for this linear model.

Proposition 5 *Suppose $\mathcal{X} = \{x_0, x_1, \ldots, x_k\}$ ($k \leq d$) is a set of $k+1$ affinely independent points in $\mathbb{R}^d$. If $f(x_0), f(x_1), \ldots, f(x_k)$ are not all equal, then $-\nabla_s f(\mathcal{X})$ is a descent direction for any linear function that interpolates the data points $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_k, f(x_k))$.*

Proof Let $g(x) = c_0 + c^T x$ be any linear function that interpolates the data points $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_k, f(x_k))$, where $c_0 \in \mathbb{R}$ and $c \in \mathbb{R}^d$. Then $c_0 +$

$c^T x_i = f(x_i)$ for $i = 0, 1, \ldots, k$, and so, $c^T(x_i - x_0) = f(x_i) - f(x_0)$ for $i = 1, \ldots, k$. Hence, $c^T S(\mathcal{X}) = \delta_f(\mathcal{X})^T$. Now for any $x \in \mathbb{R}^d$,

$$\begin{aligned}\nabla g(x)^T(-\nabla_s f(\mathcal{X})) &= -c^T S(\mathcal{X})(S(\mathcal{X})^T S(\mathcal{X}))^{-1}\delta_f(\mathcal{X}) \\ &= -\delta_f(\mathcal{X})^T(S(\mathcal{X})^T S(\mathcal{X}))^{-1}\delta_f(\mathcal{X}).\end{aligned}$$

Since $S(\mathcal{X})$ has full column rank, it follows that $S(\mathcal{X})^T S(\mathcal{X})$ and its inverse are both symmetric and positive definite. Moreover, since not all the $f(x_i)$'s are equal, it follows that $\delta_f(\mathcal{X})$ is not the zero vector. Hence, $\nabla g(x)^T(-\nabla_s f(\mathcal{X})) < 0$ for any $x \in \mathbb{R}^d$, and so, $-\nabla_s f(\mathcal{X})$ is a descent direction for $g(x) = c_0 + c^T x$ from any point $x \in \mathbb{R}^d$. □

On the other hand, in the overdetermined case, the following example shows that $-\nabla_s f(\mathcal{X})$ is not always a descent direction for the corresponding linear regression model.

Example 6 Consider the ordered set of sample points and their function values: $\mathcal{X} = \langle x_0, x_1, x_2\rangle = \langle 0, 1, 2\rangle$ and $f(x_0) = 2$, $f(x_1) = 1$, $f(x_2) = 9/4$. Then the coefficients of the linear regression model $m(x) = c_0 + c^T x$ are given by

$$\begin{aligned}\begin{bmatrix} c_0 \\ c \end{bmatrix} &= (L(\mathcal{X})^T L(\mathcal{X}))^{-1} L(\mathcal{X})^T F(\mathcal{X}) \\ &= \left(\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}^T \begin{bmatrix} 2 \\ 1 \\ 9/4 \end{bmatrix} = \begin{bmatrix} 13/8 \\ 1/8 \end{bmatrix}\end{aligned}$$

Hence, $\nabla m(x) = c = [1/8]$. The simplex gradient is given by

$$\begin{aligned}\nabla_s f(\mathcal{X}) &= (S(\mathcal{X})S(\mathcal{X})^T)^{-1} S(\mathcal{X})\delta_f(\mathcal{X}) \\ &= \left(\begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 2 \end{bmatrix}^T\right)^{-1} \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 1/4 \end{bmatrix} = \begin{bmatrix} -1/10 \end{bmatrix}\end{aligned}$$

Note that $\nabla m(x)^T(-\nabla_s f(\mathcal{X})) = [1/8]^T[1/10] = 1/80 > 0$ for any $x \in \mathbb{R}$, and so, $-\nabla_s f(\mathcal{X})$ is *not* a descent direction for $m(x)$ from any $x \in \mathbb{R}$.

Examples 3, 4 and 5 above showed that the simplex gradient $\nabla_s f(\mathcal{X})$ is not necessarily the gradient of the corresponding linear model $m(x) = c_0 + c^T x$ for the data points $\mathcal{X} = \langle x_0, x_1, \ldots, x_k\rangle$ in the underdetermined and overdetermined cases. In particular, Examples 3 and 4 correct the statement on page 33 of the DFO book by Conn et al. [5] where it is stated that, in the overdetermined case, $\nabla_s f(\mathcal{X})$ is also the gradient of the linear regression model for the data points $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_k, f(x_k))\}$. That statement in the DFO book [5] has also been rephrased in the errata for Theorem 2.13 of the book where it is stated that the simplex gradient is the gradient of the following alternative linear model:

$$m(x) = f(x_0) + c^T(x - x_0) = f(x_0) + (x - x_0)^T c, \tag{6}$$

where $c = [c_1, \ldots, c_d]^T$ are the coefficients to be determined. One main difference between the original linear model $m(x) = c_0 + c^T x$ and the above alternative linear model is that the original model has $d + 1$ coefficients to be determined (c_0 and the components of c) while the alternative model has only d coefficients. Moreover, the alternative linear model is required to interpolate the reference data point $(x_0, f(x_0))$. The coefficients of the linear model (6) are obtained by finding a minimum norm least squares solution to the following linear system:

$$\begin{cases} f(x_0) + (x_1 - x_0)^T c = f(x_1) \\ \qquad\qquad \vdots \qquad\qquad\quad \vdots \\ f(x_0) + (x_k - x_0)^T c = f(x_k) \end{cases}$$

This linear system is equivalent to:

$$\begin{bmatrix} (x_1 - x_0)^T \\ (x_2 - x_0)^T \\ \vdots \\ (x_k - x_0)^T \end{bmatrix} c = \begin{bmatrix} f(x_1) - f(x_0) \\ f(x_2) - f(x_0) \\ \vdots \\ f(x_k) - f(x_0) \end{bmatrix},$$

which is precisely the linear system that yields the simplex gradient $\nabla_s f(\mathcal{X})$. Thus, the gradient of the alternative linear model $m(x) = f(x_0) + c^T (x - x_0)$ is indeed the simplex gradient $\nabla_s f(\mathcal{X})$.

5 Error bounds for linear interpolation and regression models

Throughout this section, let $\mathcal{X} = \{x_0, x_1, \ldots, x_k\}$ be a set of sample points in $\mathbb{R}^d$ and assume that f is continuously differentiable in an open domain Ω containing the closed ball $B(x_0, \Delta) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq \Delta\}$, where $\Delta = \max_{1 \leq i \leq k} \|x_i - x_0\|$. Further, assume that ∇f is Lipschitz continuous in Ω with constant $\nu > 0$.

The following result from Conn et al. [5] (Theorem 2.11 in [5]) provides an error bound for the gradient of the linear interpolation model $m(x) = c_0 + c^T x$ in the determined case. Since $\nabla m(x)$ is identical to the simplex gradient $\nabla_s f(\mathcal{X})$ in the determined case, Proposition 6 also provides an error bound for the simplex gradient.

Proposition 6 *Assume that the set $\mathcal{X} = \langle x_0, x_1, \ldots, x_d \rangle \subset \mathbb{R}^d$ is poised for linear interpolation in $\mathbb{R}^d$ and suppose that the conditions mentioned above hold. Then the gradient of the linear interpolation model satisfies, for all points $x \in B(x_0, \Delta)$, an error bound of the form*

$$\|\nabla f(x) - \nabla m(x)\| \leq \kappa_{eg} \Delta,$$

where $\kappa_{eg} = \nu(1 + d^{1/2} \|\widehat{S}(\mathcal{X})^{-T}\|/2)$ and $\widehat{S}(\mathcal{X}) = S(\mathcal{X})/\Delta$.

The following proposition extends Theorems 2.11 and 2.13 in Conn et al. [5] (with the correction from the errata for the book). This result was essentially established in

[3], [4], and [8] but the statement of the proposition was derived from Custódio and Vicente [7] and Custódio et al. [6]. This proposition provides an error bound for the gradient of the alternative linear model $m(x) = f(x_0) + c^T(x - x_0)$, which is also the simplex gradient $\nabla_s f(\mathcal{X})$. While Theorems 2.11 and 2.13 in [5] cover the determined and overdetermined cases, respectively, this proposition covers all three cases. The proof uses the same arguments as in the proofs of Theorems 2.11 and 2.13 in [5] and is included for completeness.

Proposition 7 *Let $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ be an ordered set of $k+1$ points in $\mathbb{R}^d$ such that $S(\mathcal{X})$ has full rank and assume that the conditions mentioned above hold. Consider the alternative minimum norm least squares linear model $m(x) = f(x_0) + c^T(x - x_0)$ for the data points $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_k, f(x_k))\}$ so that $\nabla m(x) = c = \nabla_s f(\mathcal{X})$. Then*

$$\|\widehat{S}(\mathcal{X})^T(\nabla f(x_0) - \nabla m(x))\| \leq k^{1/2}\frac{\nu}{2}\Delta,$$

where $\widehat{S}(\mathcal{X}) = S(\mathcal{X})/\Delta$. Moreover, when $k \geq d$ (determined and overdetermined cases),

$$\|\nabla f(x_0) - \nabla m(x)\| \leq k^{1/2}\frac{\nu}{2}\|(\widehat{S}(\mathcal{X})^T)^\dagger\|\Delta,$$

and the gradient of this alternative linear model satisfies, for all points $x \in B(x_0, \Delta)$, an error bound of the form

$$\|\nabla f(x) - \nabla m(x)\| \leq \kappa_{eg}\Delta,$$

where $\kappa_{eg} = \nu(1 + k^{1/2}\|(\widehat{S}(\mathcal{X})^T)^\dagger\|/2)$.

Proof Define the vector $r(\mathcal{X}) = [r_1(\mathcal{X}), \ldots, r_k(\mathcal{X})]^T \in \mathbb{R}^k$ by

$$\begin{aligned} r(\mathcal{X}) :&= S(\mathcal{X})^T(\nabla f(x_0) - c) = S(\mathcal{X})^T\nabla f(x_0) - \delta_f(\mathcal{X}) \\ &= \begin{bmatrix} (x_1 - x_0)^T\nabla f(x_0) - (f(x_1) - f(x_0)) \\ \vdots \\ (x_k - x_0)^T\nabla f(x_0) - (f(x_k) - f(x_0)) \end{bmatrix}. \end{aligned}$$

By the integral form of the mean value theorem,

$$f(x_i) - f(x_0) = \int_0^1 (x_i - x_0)^T\nabla f(x_0 + t(x_i - x_0))\, dt, \qquad i = 1, \ldots, k.$$

From this equation, it follows that for $i = 1, \ldots, k$,

$$\begin{aligned}
|r_i(\mathcal{X})| &= |(x_i - x_0)^T \nabla f(x_0) - (f(x_i) - f(x_0))| \\
&= \left| \int_0^1 (x_i - x_0)^T \nabla f(x_0) dt - \int_0^1 (x_i - x_0)^T \nabla f(x_0 + t(x_i - x_0))\, dt \right| \\
&= \left| \int_0^1 (x_i - x_0)^T \left[\nabla f(x_0) - \nabla f(x_0 + t(x_i - x_0)) \right] dt \right| \\
&\leq \int_0^1 \left| (x_i - x_0)^T \left[\nabla f(x_0) - \nabla f(x_0 + t(x_i - x_0)) \right] \right| dt \\
&\leq \int_0^1 \|x_i - x_0\| \|\nabla f(x_0) - \nabla f(x_0 + t(x_i - x_0))\|\, dt \\
&\leq \int_0^1 \|x_i - x_0\| \nu \|t(x_i - x_0)\|\, dt = \nu \|x_i - x_0\|^2 \int_0^1 t\, dt \\
&= \frac{\nu}{2} \|x_i - x_0\|^2 \leq \frac{\nu}{2} \Delta^2.
\end{aligned}$$

Hence,

$$\|r(\mathcal{X})\| = \left(\sum_{i=1}^k r_i(\mathcal{X})^2 \right)^{1/2} \leq \left(\sum_{i=1}^k \left(\frac{\nu}{2} \Delta^2 \right)^2 \right)^{1/2} = \left(k \left(\frac{\nu}{2} \Delta^2 \right)^2 \right)^{1/2} = k^{1/2} \frac{\nu}{2} \Delta^2.$$

Now

$$\|S(\mathcal{X})^T (\nabla f(x_0) - c)\| = \|r(\mathcal{X})\| \leq k^{1/2} \frac{\nu}{2} \Delta^2,$$

and so,

$$\|\widehat{S}(\mathcal{X})^T (\nabla f(x_0) - c)\| \leq k^{1/2} \frac{\nu}{2} \Delta.$$

When $k \geq d$, $\widehat{S}(\mathcal{X})^T$ has full column rank, and so, $(\widehat{S}(\mathcal{X})^T)^\dagger$ is a left inverse of $\widehat{S}(\mathcal{X})^T$. In this case,

$$\begin{aligned}
\|\nabla f(x_0) - c\| &= \|(\widehat{S}(\mathcal{X})^T)^\dagger \widehat{S}(\mathcal{X})^T (\nabla f(x_0) - c)\| \\
&\leq \|(\widehat{S}(\mathcal{X})^T)^\dagger\| \|\widehat{S}(\mathcal{X})^T (\nabla f(x_0) - c)\| \leq k^{1/2} \frac{\nu}{2} \|(\widehat{S}(\mathcal{X})^T)^\dagger\| \Delta.
\end{aligned}$$

Finally, when $k \geq d$, note that for all $x \in B(x_0, \Delta)$,

$$\begin{aligned}
\|\nabla f(x) - c\| &\leq \|\nabla f(x) - \nabla f(x_0)\| + \|\nabla f(x_0) - c\| \\
&\leq \nu \|x - x_0\| + k^{1/2} \frac{\nu}{2} \|(\widehat{S}(\mathcal{X})^T)^\dagger\| \Delta \leq \left(\nu + k^{1/2} \|(\widehat{S}(\mathcal{X})^T)^\dagger\| \frac{\nu}{2} \right) \Delta.
\end{aligned}$$

□

The corollary below provides the error bound for the simplex gradient as stated in Custódio and Vicente [7] and Custódio et al. [6]. Note that this is essentially the first part of Proposition 7 stated in terms of the reduced SVD of $\widehat{S}(\mathcal{X})^T = S(\mathcal{X})^T/\Delta$. Moreover, as in the first part of Proposition 7, the gradient ∇f is evaluated at x_0 and not just at any $x \in B(x_0, \Delta)$. An error bound involving $\nabla f(x)$ for all $x \in B(x_0, \Delta)$ that is similar to the one in Proposition 7 can be obtained when $k \geq d$.

Corollary 1 *Let $\mathcal{X} = \langle x_0, x_1, \ldots, x_k \rangle$ be an ordered set of $k+1$ points in $\mathbb{R}^d$ such that $S(\mathcal{X})$ has full rank. Moreover, assume that the conditions mentioned above hold. Further, let $\widehat{U}(\mathcal{X})\widehat{\Sigma}(\mathcal{X})\widehat{V}(\mathcal{X})^T$ be a reduced SVD of $\widehat{S}(\mathcal{X})^T = S(\mathcal{X})^T/\Delta$. Then*

$$\|\widetilde{V}(\mathcal{X})^T[\nabla f(x_0) - \nabla_s f(\mathcal{X})]\| \leq \left(k^{1/2}\frac{\nu}{2}\|\widehat{\Sigma}(\mathcal{X})^{-1}\|\right)\Delta,$$

where $\widetilde{V}(\mathcal{X}) = I_d$ if $k \geq d$ and $\widetilde{V}(\mathcal{X}) = \widehat{V}(\mathcal{X})$ if $k < d$.

Proof From the previous proposition,

$$\|\widehat{U}(\mathcal{X})\widehat{\Sigma}(\mathcal{X})\widehat{V}(\mathcal{X})^T(\nabla f(x_0) - \nabla_s f(\mathcal{X}))\| \leq k^{1/2}\frac{\nu}{2}\Delta.$$

Since the columns of $\widehat{U}(\mathcal{X})$ are orthonormal, it follows that

$$\|\widehat{\Sigma}(\mathcal{X})\widehat{V}(\mathcal{X})^T(\nabla f(x_0) - \nabla_s f(\mathcal{X}))\| \leq k^{1/2}\frac{\nu}{2}\Delta.$$

Moreover, since $S(\mathcal{X})$ has full rank, it follows that $\widehat{\Sigma}(\mathcal{X})$ is nonsingular, and so,

$$\begin{aligned}\|\widehat{V}(\mathcal{X})^T(\nabla f(x_0) - \nabla_s f(\mathcal{X}))\| &= \|\widehat{\Sigma}(\mathcal{X})^{-1}\widehat{\Sigma}(\mathcal{X})\widehat{V}(\mathcal{X})^T(\nabla f(x_0) - \nabla_s f(\mathcal{X}))\| \\ &\leq \|\widehat{\Sigma}(\mathcal{X})^{-1}\|\|\widehat{\Sigma}(\mathcal{X})\widehat{V}(\mathcal{X})^T(\nabla f(x_0) - \nabla_s f(\mathcal{X}))\| \\ &\leq k^{1/2}\frac{\nu}{2}\|\widehat{\Sigma}(\mathcal{X})^{-1}\|\Delta.\end{aligned}$$

When $k \geq d$, $\widehat{V}(\mathcal{X})^T$ is an orthogonal matrix, and so,

$$\|\nabla f(x_0) - \nabla_s f(\mathcal{X})\| \leq k^{1/2}\frac{\nu}{2}\|\widehat{\Sigma}(\mathcal{X})^{-1}\|\Delta.$$

□

6 The simplex gradient as a linear operator and calculus rules

The purpose of this section is to explore what calculus rules for ordinary gradients also hold for simplex gradients. It begins with a convenient way of defining the simplex gradients of a function. As before, f and g are functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Definition 3 Let $x_0 \in \mathbb{R}^d$ and let $Y \in \mathbb{R}^{d \times k}$ be a matrix with full rank, i.e., $\text{rank}(Y) = \min\{d, k\}$. The *simplex gradient of f at x_0 with respect to the matrix Y*, denoted by $\nabla_Y f(x_0)$, is the minimum 2-norm least-squares solution to the system:

$$Y^T \nabla_Y f(x_0) = \delta_{f,Y}(x_0) := \begin{bmatrix} f(x_0 + y_1) - f(x_0) \\ \vdots \\ f(x_0 + y_k) - f(x_0) \end{bmatrix},$$

where $y_1, \ldots, y_k$ are the columns of Y, which can also be expressed as:

$$\nabla_Y f(x_0) = (Y^\dagger)^T \delta_{f,Y}(x_0),$$

where $Y^\dagger$ is the Moore–Penrose pseudoinverse of Y.

In the previous definition, $\nabla_Y f(x_0) = Y^{-T} \delta_{f,Y}(x_0)$ when $k = d$, $\nabla_Y f(x_0) = Y(Y^T Y)^{-1} \delta_{f,Y}(x_0)$ when $k < d$, and $\nabla_Y f(x_0) = (YY^T)^{-1} Y \delta_{f,Y}(x_0)$ when $k > d$. Note that $\nabla_Y f(x_0) = \nabla_s f(\langle x_0, x_0 + y_1, \ldots, x_0 + y_k \rangle)$, where $y_1, \ldots, y_k$ are the columns of Y. Moreover, if $Y = hI_d$, where h is a positive constant, then $\nabla_Y f(x_0) = \frac{1}{h} \delta_{f,Y}(x_0)$ is simply the finite-difference gradient of f at x_0 with fixed step size h.

Example 7 Suppose $f(x)$ is a linear function, say $f(x) = c_0 + c^T x$, where $c_0 \in \mathbb{R}$ and $c \in \mathbb{R}^d$, and $Y \in \mathbb{R}^{d \times k}$ has full rank. Then, for any $x \in \mathbb{R}^d$, $\nabla_Y f(x) = (Y^\dagger)^T \delta_{f,Y}(x) = (Y^\dagger)^T Y^T c = (YY^\dagger)^T c$. Furthermore, if Y is nonsingular, then $\nabla_Y f(x) = c$ for any $x \in \mathbb{R}^d$.

The following proposition is an immediate consequence of Propositions 2 and 3.

Proposition 8 *Let $Y \in \mathbb{R}^{d \times k}$ be a matrix with full rank. Then for any permutation matrix $P \in \mathbb{R}^{k \times k}$,*

$$\nabla_Y f(x) = \nabla_{(YP)} f(x).$$

Next, let $y_0 = 0$ and let $y_1, \ldots, y_k$ be the columns of Y. Furthermore, let α be a permutation of the indices $\{0, 1, \ldots, k\}$ and define $Y_\alpha \in \mathbb{R}^{d \times k}$ to be the matrix whose columns are $y_{\alpha(1)} - y_{\alpha(0)}, \ldots, y_{\alpha(k)} - y_{\alpha(0)}$. When $k \leq d$ (underdetermined and determined cases only), the simplex gradient has the more general property that for any $x \in \mathbb{R}^d$,

$$\nabla_Y f(x) = \nabla_{Y_\alpha} f(x + y_{\alpha(0)}).$$

The next proposition shows that the simplex gradient is a linear operator on the space of functions from $\mathbb{R}^d$ to $\mathbb{R}$.

Proposition 9 *Let $Y \in \mathbb{R}^{d \times k}$ be a matrix with full rank. Then for any $x \in \mathbb{R}^d$ and for any constant c,*

(a) $\nabla_Y (f + g)(x) = \nabla_Y f(x) + \nabla_Y g(x)$, *and*

(b) $\nabla_Y(cf)(x) = c\nabla_Y f(x)$.

Proof For any $x \in \mathbb{R}^d$,

$$\begin{aligned}\nabla_Y(f+g)(x) &= (Y^\dagger)^T\delta_{f+g,Y}(x) = (Y^\dagger)^T\begin{bmatrix}(f+g)(x+y_1)-(f+g)(x)\\ \vdots\\ (f+g)(x+y_k)-(f+g)(x)\end{bmatrix}\\ &= (Y^\dagger)^T\left(\begin{bmatrix}f(x+y_1)-f(x)\\ \vdots\\ f(x+y_k)-f(x)\end{bmatrix}+\begin{bmatrix}g(x+y_1)-g(x)\\ \vdots\\ g(x+y_k)-g(x)\end{bmatrix}\right)\\ &= (Y^\dagger)^T\delta_{f,Y}(x)+(Y^\dagger)^T\delta_{g,Y}(x) = \nabla_Y f(x)+\nabla_Y g(x)\end{aligned}$$

Moreover, for any $x \in \mathbb{R}^d$ and any constant c,

$$\begin{aligned}\nabla_Y(cf)(x) &= (Y^\dagger)^T\delta_{cf,Y}(x) = (Y^\dagger)^T\begin{bmatrix}(cf)(x+y_1)-(cf)(x)\\ \vdots\\ (cf)(x+y_k)-(cf)(x)\end{bmatrix}\\ &= c(Y^\dagger)^T\begin{bmatrix}f(x+y_1)-f(x)\\ \vdots\\ f(x+y_k)-f(x)\end{bmatrix} = c\nabla_Y f(x).\end{aligned}$$

□

The next proposition provides a product rule for simplex gradients. For convenience, $\mathrm{diag}(a_1,\ldots,a_k)$ denotes a diagonal matrix whose diagonal entries are $a_1,\ldots,a_k$.

Proposition 10 *Let $Y \in \mathbb{R}^{d\times k}$ be a matrix with full rank. Then for any $x \in \mathbb{R}^d$,*

$$\nabla_Y(fg)(x) = f(x)\nabla_Y g(x) + diag(g(x+y_1),\ldots,g(x+y_k))\nabla_Y f(x).$$

Proof For any $x \in \mathbb{R}^d$,

$$\begin{aligned}\nabla_Y(fg)(x) &= (Y^\dagger)^T\delta_{fg,Y}(x) = (Y^\dagger)^T\begin{bmatrix}f(x+y_1)g(x+y_1)-f(x)g(x)\\ \vdots\\ f(x+y_k)g(x+y_k)-f(x)g(x)\end{bmatrix}\\ &= (Y^\dagger)^T\left(\mathrm{diag}(g(x+y_1),\ldots,g(x+y_k))\begin{bmatrix}f(x+y_1)-f(x)\\ \vdots\\ f(x+y_k)-f(x)\end{bmatrix}\right.\\ &\quad\left.+f(x)\begin{bmatrix}g(x+y_1)-g(x)\\ \vdots\\ g(x+y_k)-g(x)\end{bmatrix}\right)\end{aligned}$$

Since a diagonal matrix commutes with any other matrix (assuming the products are defined), it follows that

$$\begin{aligned}\nabla_Y(fg)(x) &= \text{diag}(g(x+y_1),\ldots,g(x+y_k))(Y^\dagger)^T \begin{bmatrix} f(x+y_1)-f(x) \\ \vdots \\ f(x+y_k)-f(x) \end{bmatrix} \\ &\quad + f(x)(Y^\dagger)^T \begin{bmatrix} g(x+y_1)-g(x) \\ \vdots \\ g(x+y_k)-g(x) \end{bmatrix} \\ &= \text{diag}(g(x+y_1),\ldots,g(x+y_k))\nabla_Y f(x) + f(x)\nabla_Y g(x).\end{aligned}$$

□

The next proposition provides a quotient rule for simplex gradients.

Proposition 11 *Let $Y \in \mathbb{R}^{d\times k}$ be a matrix with full rank. Then for any $x \in \mathbb{R}^d$ for which $g(x), g(x+y_1),\ldots,g(x+y_k)$ are all nonzero,*

$$\nabla_Y\left(\frac{f}{g}\right)(x) = \text{diag}\left(\frac{1}{g(x+y_1)},\ldots,\frac{1}{g(x+y_k)}\right)\left[\frac{g(x)\nabla_Y f(x) - f(x)\nabla_Y g(x)}{g(x)}\right].$$

Proof By the previous proposition,

$$\begin{aligned}\nabla_Y f(x) &= \nabla_Y\left(\frac{f}{g}\cdot g\right) \\ &= \text{diag}(g(x+y_1),\ldots,g(x+y_k))\nabla_Y\left(\frac{f}{g}\right)(x) + \left(\frac{f}{g}\right)(x)\nabla_Y g(x).\end{aligned}$$

Solving for $\nabla_Y\left(\frac{f}{g}\right)(x)$ gives

$$\begin{aligned}\nabla_Y\left(\frac{f}{g}\right)(x) &= \text{diag}(g(x+y_1),\ldots,g(x+y_k))^{-1}\left[\nabla_Y f(x) - \left(\frac{f(x)}{g(x)}\right)\nabla_Y g(x)\right] \\ &= \text{diag}\left(\frac{1}{g(x+y_1)},\ldots,\frac{1}{g(x+y_k)}\right)\left[\frac{g(x)\nabla_Y f(x) - f(x)\nabla_Y g(x)}{g(x)}\right].\end{aligned}$$

□

Corollary 2 *Let $Y \in \mathbb{R}^{d\times k}$ be a matrix with full rank. Then for any $x \in \mathbb{R}^d$ for which $f(x), f(x+y_1),\ldots,f(x+y_k)$ are all nonzero,*

$$\nabla_Y\left(\frac{1}{f}\right)(x) = \frac{-1}{f(x)}\text{diag}\left(\frac{1}{f(x+y_1)},\ldots,\frac{1}{f(x+y_k)}\right)\nabla_Y f(x).$$

There does not seem to be a general chain rule for simplex gradients. However, it is possible to derive a version of the power rule for simplex gradients as shown next.

Proposition 12 *Let $Y \in \mathbb{R}^{d\times k}$ be a matrix with full rank. Then for any $x \in \mathbb{R}^d$ and for any positive integer n,*

$$\nabla_Y [f(x)]^n = \left[\sum_{i=1}^{n} [f(x)]^{n-i} diag(f(x+y_1), \ldots, f(x+y_k))^{i-1}\right] \nabla_Y f(x).$$

Proof Proceed by induction on n. The equation is obviously true for $n = 1$. Next, assume that the equation is true for $n = \ell$ for some integer $\ell \geq 1$, i.e.,

$$\nabla_Y [f(x)]^\ell = \left[\sum_{i=1}^{\ell} [f(x)]^{\ell-i} \mathrm{diag}(f(x+y_1), \ldots, f(x+y_k))^{i-1}\right] \nabla_Y f(x).$$

Now

$$\begin{aligned}
\nabla_Y [f(x)]^{\ell+1} &= \nabla_Y ([f(x)]^\ell f(x)) \\
&= [f(x)]^\ell \nabla_Y f(x) + \mathrm{diag}(f(x+y_1), \ldots, f(x+y_k)) \nabla_Y [f(x)]^\ell \\
&= [f(x)]^\ell \nabla_Y f(x) + \mathrm{diag}(f(x+y_1), \ldots, f(x+y_k)) \\
&\qquad \left[\sum_{i=1}^{\ell} [f(x)]^{\ell-i} \mathrm{diag}(f(x+y_1), \ldots, f(x+y_k))^{i-1}\right] \nabla_Y f(x) \\
&= \left[[f(x)]^\ell I_k + \sum_{i=1}^{\ell} [f(x)]^{\ell-i} \mathrm{diag}(f(x+y_1), \ldots, f(x+y_k))^{i}\right] \nabla_Y f(x)
\end{aligned}$$

Replacing i by $i-1$ in the previous sum gives

$$\begin{aligned}
\nabla_Y [f(x)]^{\ell+1} &= \Bigg[[f(x)]^\ell I_k + \sum_{i=2}^{\ell+1} [f(x)]^{\ell+1-i} \mathrm{diag}(f(x+y_1), \ldots, f(x \\
&\qquad +y_k))^{i-1}\Bigg] \nabla_Y f(x) \\
&= \left[\sum_{i=1}^{\ell+1} [f(x)]^{\ell+1-i} \mathrm{diag}(f(x+y_1), \ldots, f(x+y_k))^{i-1}\right] \nabla_Y f(x)
\end{aligned}$$

Hence, the equation is also true for $n = \ell + 1$ and the induction is complete. □

Corollary 3 *Let $Y \in \mathbb{R}^{d\times k}$ be a matrix with full rank. Then for any $x \in \mathbb{R}^d$ and any positive integer n,*

$$\nabla_Y [f(x)]^{-n} = -\left[\sum_{i=1}^{n} [f(x)]^{-i} diag(f(x+y_1), \ldots, f(x+y_k))^{i-n-1}\right] \nabla_Y f(x).$$

Proof By Corollary 2 and Proposition 12,

$$\begin{aligned}\nabla_Y\left(\frac{1}{[f(x)]^n}\right) &= \frac{-1}{[f(x)]^n}\text{diag}\left(\frac{1}{f(x+y_1)^n},\ldots,\frac{1}{f(x+y_k)^n}\right)\nabla_Y[f(x)]^n\\ &= -[f(x)]^{-n}\text{diag}(f(x+y_1),\ldots,f(x+y_k))^{-n}\\ &\quad\left[\sum_{i=1}^{n}[f(x)]^{n-i}\text{diag}(f(x+y_1),\ldots,f(x+y_k))^{i-1}\right]\nabla_Y f(x)\\ &= -\left[\sum_{i=1}^{n}[f(x)]^{-i}\text{diag}(f(x+y_1),\ldots,f(x+y_k))^{i-n-1}\right]\nabla_Y f(x)\end{aligned}$$

□

Also, there does not seem to be a chain rule for function composition involving exponential functions. However, the following proposition gives a rule for the simplex gradient of an exponential function.

Proposition 13 *Let $Y \in \mathbb{R}^{d\times k}$ be a matrix with full rank. Then for any $x \in \mathbb{R}^d$ and for any positive integer n,*

$$\nabla_Y e^{f(x)} = e^{f(x)}(Y^\dagger)^T(e^{\delta_{f,Y}(x)} - 1_{k\times 1}),$$

where $1_{k\times 1}$ is a vector of all 1's and the exponentiation is taken componentwise.

Proof

$$\begin{aligned}\nabla_Y e^{f(x)} &= (Y^\dagger)^T\begin{bmatrix}e^{f(x+y_1)}-e^{f(x)}\\ \vdots\\ e^{f(x+y_k)}-e^{f(x)}\end{bmatrix} = e^{f(x)}(Y^\dagger)^T\begin{bmatrix}e^{f(x+y_1)-f(x)}-1\\ \vdots\\ e^{f(x+y_k)-f(x)}-1\end{bmatrix}\\ &= e^{f(x)}(Y^\dagger)^T(e^{\delta_{f,Y}(x)} - 1_{k\times 1}).\end{aligned}$$

□

7 Summary and conclusions

This article clarified some of the properties of simplex gradients that were previously not explicitly mentioned in the literature. In particular, the simplex gradient was shown to be independent of the order of the points in the underdetermined and determined cases but it depends only on which point is used as the reference point in the overdetermined case. Moreover, although the simplex gradient and the gradient of the corresponding linear model are equal in the determined case, this property is not true for the underdetermined and overdetermined cases. However, the simplex gradient turns out to be the gradient of an alternative linear model that has one less coefficient than the original model and that requires interpolation at the reference point. The negative of the simplex gradient was also shown to be a descent direction for any interpolating

linear function in the determined and underdetermined cases but this property does not hold for the overdetermined case. In addition, a previously established error bound for simplex gradients was reviewed. Finally, calculus rules for simplex gradients (similar to those for ordinary gradients) were also proved.

Acknowledgments The author would like to thank the three anonymous referees. Their comments and suggestions greatly improved this article. In particular, the alternative linear model in Sect. 4 was suggested by one of the referees.

References

1. Audet, C., Dennis Jr, J.E.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**(2), 188–217 (2006)
2. Bortz, D.M., Kelley, C.T.: The simplex gradient and noisy optimization problems. In: Borggaard, J., et al. (eds.) Computational Methods for Optimal Design and Control, Progress in Systems and Control Theory, vol. 24, pp. 77–90. Springer, New York (1998)
3. Conn, A.R., Scheinberg, K., Vicente, L.N.: Geometry of interpolation sets in derivative free optimization. Math. Progr. **111**(1–2), 141–172 (2008a)
4. Conn, A.R., Scheinberg, K., Vicente, L.N.: Geometry of sample sets in derivative-free optimization: polynomial regression and underdetermined interpolation. IMA J. Numer. Anal. **28**(4), 721–748 (2008b)
5. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. MOS/SIAM book series on optimization. SIAM, Philadelphia (2009)
6. Custódio, A.L., Dennis Jr, J.E., Vicente, L.N.: Using simplex gradients of nonsmooth functions in direct search methods. IMA J. Numer. Anal. **28**(4), 770–784 (2008)
7. Custódio, A.L., Vicente, L.N.: Using sampling and simplex derivatives in pattern search methods. SIAM J. Optim. **18**(2), 537–555 (2007)
8. Kelley, C.T.: Iterative Methods for Optimization. SIAM, Philadelphia (1999)
9. Moré, J.J., Wild, S.M.: Benchmarking derivative-free optimization algorithms. SIAM J. Optim. **20**(1), 172–191 (2009)
10. Regis, R.G.: An initialization strategy for high-dimensional surrogate-based expensive black-box optimization. In: Zuluaga, L.F., Terlaky, T. (eds.) Selected Contributions from the MOPTA 2012 Conference Series. Springer Proceedings in Mathematics and Statistics, vol 62, pp. 51–85 (2013)
11. Scheinberg, K., Toint, P.L.: Self-correcting geometry in model-based algorithms for derivative-free unconstrained optimization. SIAM J. Optim. **20**(6), 3512–3532 (2010)
12. Torczon, V.: On the convergence of pattern search algorithms. SIAM J. Optim. **7**(1), 1–25 (1997)