Chapter 10
Sensitivity Analysis and Model Reduction

> Le doute n'est pas un état bien agréable, mais l'assurance est un état ridicule.
>
> Voltaire

The topic of this chapter is *sensitivity analysis*, which may be broadly understood as understanding how $f(x_1, \ldots, x_n)$ depends upon variations not only in the x_i individually, but also combined or correlated effects among the x_i. There are two broad classes of sensitivity analyses: *local* sensitivity analyses study the sensitivity of f to variations in its inputs at or near a particular base point, as exemplified by the calculation of derivatives; *global* sensitivity analyses study the 'average' sensitivity of f to variations of its inputs across the domain of definition of f, as exemplified by the McDiarmid diameters and Sobol′ indices introduced in Sections 10.3 and 10.4 respectively.

A closely related topic is that of *model order reduction*, in which it is desired to find a new function $\tilde{f}$, a function of many fewer inputs than f, that can serve as a good approximation to f. Practical problems from engineering and the sciences can easily have models with millions or billions of inputs (degrees of freedom). Thorough exploration of such high-dimensional spaces, e.g. for the purposes of parameter optimization or a Bayesian inversion, is all but impossible; in such situations, it is essential to be able to resort to some kind of proxy $\tilde{f}$ for f in order to obtain results of any kind, even though their accuracy will be controlled by the accuracy of the approximation $\tilde{f} \approx f$.

T.J. Sullivan, *Introduction to Uncertainty Quantification*, Texts in Applied Mathematics 63, DOI 10.1007/978-3-319-23395-6_10

10.1 Model Reduction for Linear Models

Suppose that the model mapping inputs $x \in \mathbb{C}^n$ to outputs $y = f(x) \in \mathbb{C}^m$ is actually a linear map, and so can be represented by a matrix $A \in \mathbb{C}^{m\times n}$. There is essentially only one method for the dimensional reduction of such linear models, the *singular value decomposition* (SVD).

Theorem 10.1 (Singular value decomposition)**.** *Every matrix $A \in \mathbb{C}^{m\times n}$ can be factorized as $A = U\Sigma V^*$, where $U \in \mathbb{C}^{m\times m}$ is unitary (i.e. $U^*U = UU^* = I$), $V \in \mathbb{C}^{n\times n}$ is unitary, and $\Sigma \in \mathbb{R}^{m\times n}_{\geq 0}$ is diagonal. Furthermore, if A is real, then U and V are also real.*

Remark 10.2. The existence of an SVD-like decomposition for an operator A between Hilbert spaces is essentially the definition of A being a compact operator (cf. Definition 2.48).

The columns of U are called the *left singular vectors* of A; the columns of V are called the *right singular vectors* of A; and the diagonal entries of Σ are called the *singular values* of A. While the singular values are unique, the singular vectors may fail to be. By convention, the singular values and corresponding singular vectors are ordered so that the singular values form a decreasing sequence

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0.$$

Thus, the SVD is a decomposition of A into a sum of rank-1 operators:

$$A = U\Sigma V^* = \sum_{j=1}^{\min\{m,n\}} \sigma_j u_j \otimes v_j = \sum_{j=1}^{\min\{m,n\}} \sigma_j u_j \langle v_j, \cdot \rangle.$$

The singular values and singular vectors are closely related to the eigenpairs of self-adjoint and positive semi-definite matrices A^*A:

(a) If $m < n$, then the eigenvalues of A^*A are $\sigma_1^2, \ldots, \sigma_m^2$ and $n - m$ zeros, and the eigenvalues of AA^* are $\sigma_1^2, \ldots, \sigma_m^2$.
(b) If $m = n$, then the eigenvalues of A^*A and of AA^* are $\sigma_1^2, \ldots, \sigma_n^2$.
(c) If $m > n$, then the eigenvalues of A^*A are $\sigma_1^2, \ldots, \sigma_n^2$ and the eigenvalues of AA^* are $\sigma_1^2, \ldots, \sigma_n^2$ and $m - n$ zeros.

In all cases, the eigenvectors of A^*A are the columns of V, i.e. the right singular vectors of A, and the eigenvectors of AA^* are the columns of U, i.e. the left singular vectors of A.

The appeal of the SVD is that it can be calculated in a numerically stable fashion (e.g. by bidiagonalization via Householder reflections, followed by a

variant of the QR algorithm for eigenvalues), and that it provides optimal low-rank approximation of linear operators in a sense made precise by the next two results:

Theorem 10.3 (Courant–Fischer minimax theorem)**.** *For $A \in \mathbb{C}^{m\times n}$ and a subspace $E \subseteq \mathbb{C}^n$, let*

$$\left\|A|_E\right\|_2 := \sup_{x\in E\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2} \equiv \sup_{x\in E\setminus\{0\}} \frac{\langle x, A^*Ax\rangle^{1/2}}{\|x\|_2}$$

be the operator 2-norm of A restricted to E. Then the singular values of A satisfy, for $k = 1, \ldots, \min\{m, n\}$,

$$\sigma_k = \inf_{\substack{\text{subspaces } E \text{ s.t.} \\ \operatorname{codim} E = k-1}} \left\|A|_E\right\|_2 = \inf_{\substack{\text{subspaces } E \text{ s.t.} \\ \operatorname{codim} E \le k-1}} \left\|A|_E\right\|_2.$$

Proof. Let A have SVD $A = U\Sigma V^*$, and let $v_1, \ldots, v_n$ be the columns of V, i.e. the eigenvectors of A^*A. Then, for any $x \in \mathbb{C}^n$,

$$x = \sum_{j=1}^n \langle x, v_j\rangle v_j, \qquad \|x\|^2 = \sum_{j=1}^n |\langle x, v_j\rangle|^2,$$
$$A^*Ax = \sum_{j=1}^n \sigma_j^2 \langle x, v_j\rangle v_j, \qquad \langle x, A^*Ax\rangle = \sum_{j=1}^n \sigma_j^2 |\langle x, v_j\rangle|^2.$$

Let $E \subseteq \mathbb{C}^n$ have $\operatorname{codim} E \le k-1$. Then the k-dimensional subspace spanned by $v_1, \ldots, v_k$ has some $x \neq 0$ in common with E, and so

$$\langle x, A^*Ax\rangle = \sum_{j=1}^k \sigma_j^2 |\langle x, v_j\rangle|^2 \ge \sigma_k^2 \sum_{j=1}^k |\langle x, v_j\rangle|^2 = \sigma_k^2 \|x\|^2.$$

Hence, $\sigma_k \le \left\|A|_E\right\|$ for any E with $\operatorname{codim} E \le k-1$.

It remains only to find some E with $\operatorname{codim} E = k-1$ for which $\sigma_k \ge \left\|A|_E\right\|$. Take $E := \operatorname{span}\{v_k, \ldots, v_n\}$. Then, for any $x \in E$,

$$\langle x, A^*Ax\rangle = \sum_{j=k}^n \sigma_j^2 |\langle x, v_j\rangle|^2 \le \sigma_k^2 \sum_{j=k}^n |\langle x, v_j\rangle|^2 = \sigma_k^2 \|x\|^2,$$

which completes the proof. □

Theorem 10.4 (Eckart–Young low-rank approximation theorem)**.** *Given $A \in \mathbb{C}^{m\times n}$, let $A_k \in \mathbb{C}^{m\times n}$ be the matrix formed from the first k singular vectors and singular values of A, i.e.*

$$A_k := \sum_{j=1}^k \sigma_j u_j \otimes v_j. \tag{10.1}$$

Then

$$\sigma_{k+1} = \|A - A_k\|_2 = \inf_{\substack{X \in \mathbb{C}^{m\times n} \\ \operatorname{rank} X \le k}} \|A - X\|_2.$$

Hence, as measured by the operator 2-norm,
(a) A_k is the best approximation to A of rank at most k; and
(b) if $A \in \mathbb{C}^{n\times n}$, then A is invertible if and only if $\sigma_n > 0$, and σ_n is the distance of A from the set of singular matrices.

Proof. Let $\mathcal{M}_k$ denote the set of matrices in $\mathbb{C}^{m\times n}$ with rank $\le k$, and let $X \in \mathcal{M}_k$. Since $\operatorname{rank} X + \dim \ker X = n$, it follows that $\operatorname{codim} \ker X \le k$. By Theorem 10.3,

$$\sigma_{k+1} \le \sup_{\substack{x\in E \\ \operatorname{codim} E \le k}} \frac{\|Ax\|_2}{\|x\|_2}.$$

Hence,

$$\sigma_{k+1} \le \sup_{x\in \ker X} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x\in \ker X} \frac{\|(A-X)x\|_2}{\|x\|_2} \le \|A - X\|_2.$$

Hence $\sigma_{k+1} \le \inf_{X\in\mathcal{M}_k} \|A - X\|_2$.

Now consider A_k as given by (10.1), which certainly has $\operatorname{rank} A_k \le k$. Now,

$$A - A_k = \sum_{j=k+1}^{r} \sigma_j u_j \otimes v_j,$$

where $r := \operatorname{rank} A$. Write $x \in \mathbb{C}^n$ as $x = \sum_{j=1}^{n} \langle x, v_j\rangle v_j$. Then

$$(A - A_k)x = \sum_{j=k+1}^{r} \sigma_j u_j \langle v_j, x\rangle,$$

and so

$$\begin{aligned} \|(A - A_k)x\|_2^2 &= \sum_{j=k+1}^{r} \sigma_j^2 |\langle v_j, x\rangle|^2 \\ &\le \sigma_{k+1}^2 \sum_{j=k+1}^{r} |\langle v_j, x\rangle|^2 \\ &\le \sigma_{k+1}^2 \|x\|_2^2 \end{aligned}$$

Hence, $\|A - A_k\|_2 \le \sigma_{k+1}$. □

See Chapter 11 for an application of the SVD to the analysis of sample data from random variables, a discrete variant of the Karhunen–Loève

expansion, known as *principal component analysis* (PCA). Simply put, when A is a matrix whose columns are independent samples from some stochastic process (random vector), the SVD of A is the ideal way to fit a linear structure to those data points. One may consider nonlinear fitting and dimensionality reduction methods in the same way, and this is known as *manifold learning*. There are many nonlinear generalizations of the SVD/PCA: see the bibliography for some references.

10.2 Derivatives

One way to understand the dependence of $f(x_1, \ldots, x_n)$ upon $x_1, \ldots, x_n$ near some nominal point $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ is to estimate the partial derivatives of f at $\bar{x}$, i.e. to approximate

$$\frac{\partial f}{\partial x_i}(\bar{x}) := \lim_{h \to 0} \frac{f(\bar{x}_1, \ldots, \bar{x}_i + h, \ldots, \bar{x}_n) - f(\bar{x})}{h}.$$

For example, for a function f of a single real variable x, and with a fixed step size $h > 0$, the derivative of f at $\bar{x}$ may be approximated using the *forward difference*

$$\frac{\mathrm{d}f}{\mathrm{d}x}(\bar{x}) \approx \frac{f(\bar{x} + h) - f(\bar{x})}{h}$$

or the *backward difference*

$$\frac{\mathrm{d}f}{\mathrm{d}x}(\bar{x}) \approx \frac{f(\bar{x}) - f(\bar{x} - h)}{h}.$$

Similarly, the second derivative of f might be approximated using the *second order central difference*

$$\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(\bar{x}) \approx \frac{f(\bar{x} + h) - 2f(\bar{x}) + f(\bar{x} - h)}{h^2}.$$

Ultimately, approximating the derivatives of f in this way is implicitly a polynomial approximation: polynomials coincide with their Taylor expansions, their derivatives can be computed exactly, and we make the approximation that $f \approx p \implies f' \approx p'$. Alternatively, we can construct a randomized estimate of the derivative of f at $\bar{x}$ by random sampling of x near $\bar{x}$ (i.e. x not necessarily of the form $x = \bar{x} + he_i$), as in the *simultaneous perturbation stochastic approximation* (SPSA) method of Spall (1992).

An alternative paradigm for differentiation is based on the observation that many numerical operations on a computer are in fact polynomial operations, so they can be differentiated accurately using the *algebraic* properties of differential calculus, rather than the *analytical* definitions of those objects.

A simple algebraic structure that encodes first derivatives is the concept of dual numbers, the abstract algebraic definition of which is as follows:

Definition 10.5. The *dual numbers* $\mathbb{R}_\epsilon$ are defined to be the quotient of the polynomial ring $\mathbb{R}[x]$ by the ideal generated by the monomial x^2.

In plain terms, $\mathbb{R}_\epsilon = \{x_0 + x_1\epsilon \mid x_0, x_1 \in \mathbb{R}\}$, where $\epsilon \neq 0$ has the property that $\epsilon^2 = 0$ (ϵ is said to be *nilpotent*). Addition and subtraction of dual numbers is handled componentwise; multiplication of dual numbers is handled similarly to multiplication of complex numbers, except that the relation $\epsilon^2 = 0$ is used in place of the relation $i^2 = -1$; however, there are some additional subtleties in division, which is only well defined when the real part of the denominator is non-zero, and is otherwise multivalued or even undefined. In summary:

$$\begin{aligned}(x_0 + x_1\epsilon) + (y_0 + y_1\epsilon) &= (x_0 + y_0) + (x_1 + y_1)\epsilon,\\ (x_0 + x_1\epsilon) - (y_0 + y_1\epsilon) &= (x_0 - y_0) + (x_1 - y_1)\epsilon,\\ (x_0 + x_1\epsilon)(y_0 + y_1\epsilon) &= x_0y_0 + (x_0y_1 + x_1y_0)\epsilon\end{aligned}$$

$$\frac{x_0 + x_1\epsilon}{y_0 + y_1\epsilon} = \begin{cases} \dfrac{x_0}{y_0} + \dfrac{y_0x_1 - x_0y_1}{y_0^2}\epsilon, & \text{if } y_0 \neq 0,\\ \dfrac{x_1}{y_1} + z\epsilon, & \text{for any } z \in \mathbb{R} \text{ if } x_0 = y_0 = 0,\\ \text{undefined}, & \text{if } y_0 = 0 \text{ and } x_0 \neq 0.\end{cases}$$

A helpful representation of $\mathbb{R}_\epsilon$ in terms of 2×2 real matrices is given by

$$x_0 + x_1\epsilon \longleftrightarrow \begin{bmatrix} x_0 & x_1 \\ 0 & x_0 \end{bmatrix} \quad \text{so that} \quad \epsilon \longleftrightarrow \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

One can easily check that the algebraic rules for addition, multiplication, etc. in $\mathbb{R}_\epsilon$ correspond exactly to the usual rules for addition, multiplication, etc. of 2×2 matrices.

Automatic Differentiation. A useful application of dual numbers is *automatic differentiation*, which is a form of exact differentiation that arises as a side-effect of the algebraic properties of the nilpotent element ϵ, which behaves rather like an infinitesimal in non-standard analysis. Given the algebraic properties of the dual numbers, any polynomial $p(x) := p_0 + p_1x + \cdots + p_nx^n \in \mathbb{R}[x]_{\leq n}$, thought of as a function $p\colon \mathbb{R} \to \mathbb{R}$, can be extended to a function $p\colon \mathbb{R}_\epsilon \to \mathbb{R}_\epsilon$. Then, for any $x_0 + x_1\epsilon \in \mathbb{R}_\epsilon$,

$$\begin{aligned} p(x_0 + x_1\epsilon) &= \sum_{k=0}^{n} p_k (x_0 + x_1\epsilon)^k \\ &= \left(\sum_{k=0}^{n} p_k x_0^k \right) + \left(p_1 x_1 \epsilon + 2p_2 x_0 x_1 \epsilon + \cdots + n p_n x_0^{n-1} x_1 \epsilon \right) \\ &= p(x_0) + p'(x_0) x_1 \epsilon. \end{aligned}$$

Thus the derivative of a real polynomial at x is exactly the coefficient of ϵ in its dual-number extension $p(x+\epsilon)$. Indeed, by considering Taylor series, it follows that the same result holds true for any analytic function (see Exercise 10.1). Since many numerical functions on a computer are evaluations of polynomials or power series, the use of dual numbers offers accurate symbolic differentiation of such functions, once those functions have been extended to accept dual number arguments and return dual number values. Implementation of dual number arithmetic is relatively straightforward for many common programming languages such as C/C++, Python, and so on; however, technical problems can arise when interfacing with legacy codes that cannot be modified to operate with dual numbers.

Remark 10.6. (a) An attractive feature of automatic differentiation is that complicated compositions of functions can be differentiated exactly using the chain rule

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

and automatic differentiation of the functions being composed.

(b) For higher-order derivatives, instead of working in a number system for which $\epsilon^2 = 0$, one works in a system in which ϵ^3 or some other higher power of ϵ is zero. For example, to obtain automatic second derivatives, consider

$$\mathbb{R}_{\epsilon,\epsilon^2} = \{x_0 + x_1\epsilon + x_2\epsilon^2 \mid x_0, x_1, x_2 \in \mathbb{R}\}$$

with $\epsilon^3 = 0$. The derivative at x of a polynomial p is again the coefficient of ϵ in $p(x + \epsilon)$, and the second derivative is twice (i.e. 2! times) the coefficient of ϵ^2 in $p(x + \epsilon)$.

(c) Analogous dual systems can be constructed for any commutative ring R, by defining the dual ring to be the quotient ring $R[x]/(x^2)$ — a good example being the ring of square matrices over some field. The image of x under the quotient map then has square equal to zero and plays the role of ϵ in the above discussion.

(d) Automatic differentiation of vector-valued functions of vector arguments can be accomplished using a nilpotent vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ with the property that $\epsilon_i \epsilon_j = 0$ for all $i, j \in \{1, \ldots, n\}$; see Exercise 10.3.

The Adjoint Method. A common technique for understanding the impact of uncertain or otherwise variable parameters on a system is the so-called

adjoint method, which is in fact a cunning application of the implicit function theorem (IFT) from multivariate calculus:

Theorem 10.7 (Implicit function theorem)**.** *Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be Banach spaces, let $W \subseteq \mathcal{X} \times \mathcal{Y}$ be open, and let $f \in \mathcal{C}^k(W; \mathcal{Z})$ for some $k \geq 1$. Suppose that, at $(\bar{x}, \bar{y}) \in W$, the partial Fréchet derivative $\frac{\partial f}{\partial y}(\bar{x}, \bar{y}) \colon \mathcal{Y} \to \mathcal{Z}$ is an invertible bounded linear map. Then there exist open sets $U \subseteq \mathcal{X}$ about $\bar{x}$, $V \subseteq \mathcal{Y}$ about $\bar{y}$, with $U \times V \subseteq W$, and a unique $\varphi \in \mathcal{C}^k(U; V)$ such that*

$$\{(x, y) \in U \times V \,|\, f(x, y) = f(\bar{x}, \bar{y})\} = \{(x, y) \in U \times V \,|\, y = \varphi(x)\},$$

i.e. the contour of f through $(\bar{x}, \bar{y})$ is locally the graph of φ. Furthermore, U can be chosen so that $\frac{\partial f}{\partial y}(x, \varphi(x))$ is boundedly invertible for all $x \in U$, and the Fréchet derivative $\frac{\mathrm{d}\varphi}{\mathrm{d}x}(x) \colon \mathcal{X} \to \mathcal{Y}$ of φ at any $x \in U$ is the composition

$$\frac{\mathrm{d}\varphi}{\mathrm{d}x}(x) = -\left(\frac{\partial f}{\partial y}(x, \varphi(x))\right)^{-1} \left(\frac{\partial f}{\partial x}(x, \varphi(x))\right). \tag{10.2}$$

We now apply the IFT to derive the adjoint method for sensitivity analysis. Let $\mathcal{U}$ and Θ be (open subsets of) Banach spaces. Suppose that uncertain parameters $\theta \in \Theta$ and a derived quantity $u \in \mathcal{U}$ are related by an implicit function of the form $F(u, \theta) = 0$; to take a very simple example, suppose that $u \colon [-1, 1] \to \mathbb{R}$ solves the boundary value problem

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{\theta}\frac{\mathrm{d}}{\mathrm{d}x}u(x)\right) &= (x-1)(x+1), & -1 < x < 1, \\ u(x) &= 0, & x \in \{\pm 1\}. \end{aligned}$$

Suppose also that we are interested in understanding the effect of changing θ upon the value of a quantity of interest $q \colon \mathcal{U} \times \Theta \to \mathbb{R}$. To be more precise, the aim is to understand the derivative of $q(u, \theta)$ with respect to θ, with u depending on θ via $F(u, \theta) = 0$, at some nominal point $(\bar{u}, \bar{\theta})$.

Observe that, by the chain rule,

$$\frac{\mathrm{d}q}{\mathrm{d}\theta}(\bar{u}, \bar{\theta}) = \frac{\partial q}{\partial u}(\bar{u}, \bar{\theta})\frac{\partial u}{\partial \theta}(\bar{u}, \bar{\theta}) + \frac{\partial q}{\partial \theta}(\bar{u}, \bar{\theta}). \tag{10.3}$$

Note that (10.3) only makes sense if u can be locally expressed as a differentiable function of θ near $(\bar{u}, \bar{\theta})$: by the IFT, a sufficient condition for this is that F is continuously Fréchet differentiable near $(\bar{u}, \bar{\theta})$ with $\frac{\partial F}{\partial u}(\bar{u}, \bar{\theta})$ invertible. Using this insight, the partial derivative of the solution u with respect to the parameters θ can be eliminated from (10.3) to yield an expression that uses only the partial derivatives of the explicit functions F and q.

To perform this elimination, observe that the total derivative of F vanishes everywhere on the set of $(u, \theta) \in \mathcal{U} \times \Theta$ such that $F(u, \theta) = 0$ (or, indeed, on any level set of F), and so the chain rule gives

$$\frac{\mathrm{d}F}{\mathrm{d}\theta} = \frac{\partial F}{\partial u}\frac{\partial u}{\partial \theta} + \frac{\partial F}{\partial \theta} \equiv 0.$$

Therefore, since $\frac{\partial F}{\partial u}(\bar{u}, \bar{\theta})$ is invertible,

$$\frac{\partial u}{\partial \theta}(\bar{u}, \bar{\theta}) = -\left(\frac{\partial F}{\partial u}(\bar{u}, \bar{\theta})\right)^{-1} \frac{\partial F}{\partial \theta}(\bar{u}, \bar{\theta}), \tag{10.4}$$

as in (10.2) in the conclusion of the IFT. Thus, (10.3) becomes

$$\frac{\mathrm{d}q}{\mathrm{d}\theta}(\bar{u}, \bar{\theta}) = -\frac{\partial q}{\partial u}(\bar{u}, \bar{\theta})\left(\frac{\partial F}{\partial u}(\bar{u}, \bar{\theta})\right)^{-1} \frac{\partial F}{\partial \theta}(\bar{u}, \bar{\theta}) + \frac{\partial q}{\partial \theta}(\bar{u}, \bar{\theta}), \tag{10.5}$$

which, as desired, avoids explicit reference to $\frac{\partial u}{\partial \theta}$.

Equation (10.4) can be re-written as

$$\frac{\partial q}{\partial \theta}(\bar{u}, \bar{\theta}) = \lambda \frac{\partial F}{\partial \theta}(\bar{u}, \bar{\theta})$$

where the linear functional $\lambda \in \mathcal{U}'$ is the solution to

$$\lambda \frac{\partial F}{\partial u}(\bar{u}, \bar{\theta}) = -\frac{\partial q}{\partial u}(\bar{u}, \bar{\theta}), \tag{10.6}$$

or, equivalently, taking the adjoint (conjugate transpose) of (10.6),

$$\left(\frac{\partial F}{\partial u}(\bar{u}, \bar{\theta})\right)^{*} \lambda^{*} = -\left(\frac{\partial q}{\partial u}(\bar{u}, \bar{\theta})\right)^{*}, \tag{10.7}$$

which is known as the *adjoint equation*. This is a powerful tool for investigating the dependence of q upon θ, because we can now compute $\frac{\mathrm{d}q}{\mathrm{d}\theta}$ without ever having to work out the relationship between θ and u or its derivative $\frac{\partial u}{\partial \theta}$ explicitly — we only need partial derivatives of F and q with respect to θ and u, which are usually much easier to calculate. We then need only solve (10.6)/(10.7) for λ, and then substitute that result into (10.5).

Naturally, the system (10.6)/(10.7) is almost never solved by explicitly computing the inverse matrix; instead, the usual direct (e.g. Gaussian elimination with partial pivoting, the QR method) or iterative methods (e.g. the Jacobi or Gauss–Seidel iterations) are used. See Exercise 10.4 for an example of the adjoint method for an ODE.

Remark 10.8. Besides their local nature, the use of *partial* derivatives as sensitivity indices suffers from another problem well known to students of multivariate differential calculus: a function can have well-defined partial derivatives that all vanish, yet not be continuous, let alone locally constant. The standard example of such a function is $f\colon \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) := \begin{cases} \dfrac{xy}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

This function f is discontinuous at $(0,0)$, since approaching $(0,0)$ along the line $x = 0$ gives

$$\lim_{\substack{x=0 \\ y\to 0}} f(x,y) = \lim_{y\to 0} f(0,y) = \lim_{y\to 0} 0 = 0$$

but approaching $(0,0)$ along the line $x = y$ gives

$$\lim_{y=x\to 0} f(x,y) = \lim_{x\to 0} \frac{x^2}{2x^2} = \frac{1}{2} \neq 0.$$

However, f has well-defined partial derivatives with respect to x and y at every point in $\mathbb{R}^2$, and in particular at the origin:

$$\frac{\partial f}{\partial x}(x,y) = \begin{cases} \dfrac{y^3 - x^2 y}{(x^2+y^2)^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0), \end{cases}$$

$$\frac{\partial f}{\partial y}(x,y) = \begin{cases} \dfrac{x^3 - x y^2}{(x^2+y^2)^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Such pathologies do not arise if the partial derivatives are themselves continuous functions. Therefore, before placing much trust in the partial derivatives of f as local sensitivity indices, one should check that f is $\mathcal{C}^1$.

10.3 McDiarmid Diameters

Unlike the partial derivatives of the previous section, which are local measures of parameter sensitivity, this section considers global 'L^∞-type' sensitivity indices that measure the sensitivity of a function of n variables or parameters to variations in those variables/parameters individually.

Definition 10.9. The i^{th} *McDiarmid subdiameter* of $f\colon \mathcal{X} := \prod_{i=1}^n \mathcal{X}_i \to \mathbb{K}$ is defined by

$$\mathcal{D}_i[f] := \sup\big\{ |f(x) - f(y)| \big| x, y \in \mathcal{X} \text{ such that } x_j = y_j \text{ for } j \neq i \big\};$$

equivalently, $\mathcal{D}_i[f]$ is

$$\sup \left\{ |f(x) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \,\middle|\, \begin{array}{c} x = (x_1, \ldots, x_n) \in \mathcal{X} \\ \text{and } x_i' \in \mathcal{X}_i \end{array} \right\}.$$

The *McDiarmid diameter* of f is

$$\mathcal{D}[f] := \sqrt{\sum_{i=1}^{n} \mathcal{D}_i[f]^2}.$$

Remark 10.10. Note that although the two definitions of $\mathcal{D}_i[f]$ given above are obviously mathematically equivalent, they are very different from a computational point of view: the first formulation is 'obviously' a constrained optimization problem in $2n$ variables with $n-1$ constraints (i.e. 'difficult'), whereas the second formulation is 'obviously' an unconstrained optimization problem in $n+1$ variables (i.e. 'easy').

Lemma 10.11. *For each $j = 1, \ldots, n$, $\mathcal{D}_j[\cdot]$ is a seminorm on the space of bounded functions $f\colon \mathcal{X} \to \mathbb{K}$, as is $\mathcal{D}[\cdot]$.*

Proof. Exercise 10.5. □

The McDiarmid subdiameters and diameter are useful not only as sensitivity indices, but also for providing a rigorous upper bound on deviations of a function of independent random variables from its mean value:

Theorem 10.12 (McDiarmid's bounded differences inequality)**.** *Let $X = (X_1, \ldots, X_n)$ be any random variable with independent components taking values in $\mathcal{X} = \prod_{i=1}^{n} \mathcal{X}_i$, and let $f\colon \mathcal{X} \to \mathbb{R}$ be absolutely integrable with respect to the law of X and have finite McDiarmid diameter $\mathcal{D}[f]$. Then, for any $t \geq 0$,*

$$\mathbb{P}\big[f(X) \geq \mathbb{E}[f(X)] + t\big] \leq \exp\left(-\frac{2t^2}{\mathcal{D}[f]^2}\right), \tag{10.8}$$

$$\mathbb{P}\big[f(X) \leq \mathbb{E}[f(X)] - t\big] \leq \exp\left(-\frac{2t^2}{\mathcal{D}[f]^2}\right), \tag{10.9}$$

$$\mathbb{P}\big[|f(X) - \mathbb{E}[f(X)]| \geq t\big] \leq 2\exp\left(-\frac{2t^2}{\mathcal{D}[f]^2}\right). \tag{10.10}$$

Corollary 10.13 (Hoeffding's inequality)**.** *Let $X = (X_1, \ldots, X_n)$ be a random variable with independent components, taking values in the cuboid $\prod_{i=1}^{n}[a_i, b_i]$. Let $S_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for any $t \geq 0$,*

$$\mathbb{P}\big[S_n - \mathbb{E}[S_n] \geq t\big] \leq \exp\left(-\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

and similarly for deviations below, and either side, of the mean.

McDiarmid's and Hoeffding's inequalities are just two examples of a broad family of inequalities known as *concentration of measure* inequalities. Roughly put, the concentration of measure phenomenon, which was first noticed by Lévy (1951), is the fact that a function of a high-dimensional

random variable with many independent (or weakly correlated) components has its values overwhelmingly concentrated about the mean (or median). An inequality such as McDiarmid's provides a rigorous certification criterion: to be sure that $f(X)$ will deviate above its mean by more than t with probability no greater than $\varepsilon \in [0,1]$, it suffices to show that

$$\exp\left(-\frac{2t^2}{\mathcal{D}[f]^2}\right) \leq \varepsilon$$

i.e.

$$\mathcal{D}[f] \leq t\sqrt{\frac{2}{\log \varepsilon^{-1}}}.$$

Experimental effort then revolves around determining $\mathbb{E}[f(X)]$ and $\mathcal{D}[f]$; given those ingredients, the certification criterion is mathematically rigorous. That said, it is unlikely to be the *optimal* rigorous certification criterion, because McDiarmid's inequality is not guaranteed to be sharp. The calculation of optimal probability inequalities is considered in Chapter 14.

To prove McDiarmid's inequality first requires a lemma bounding the moment-generating function of a random variable:

Lemma 10.14 (Hoeffding's lemma)**.** *Let X be a random variable with mean zero taking values in $[a,b]$. Then, for $t \geq 0$,*

$$\mathbb{E}\left[e^{tX}\right] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

Proof. By the convexity of the exponential function, for each $x \in [a,b]$,

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}.$$

Therefore, applying the expectation operator,

$$\mathbb{E}\left[e^{tX}\right] \leq \frac{b}{b-a}e^{ta} + \frac{a}{b-a}e^{tb} = e^{\phi(t)}.$$

Observe that $\phi(0) = 0$, $\phi'(0) = 0$, and $\phi''(t) \leq \frac{1}{4}(b-a)^2$. Hence, since exp is an increasing and convex function,

$$\mathbb{E}\left[e^{tX}\right] \leq \exp\left(0 + 0t + \frac{(b-a)^2}{4}\frac{t^2}{2}\right) = \exp\left(\frac{t^2(b-a)^2}{8}\right). \qquad \square$$

We can now give the proof of McDiarmid's inequality, which uses Hoeffding's lemma and the properties of conditional expectation outlined in Example 3.22.

Proof of McDiarmid's inequality (Theorem 10.12). Let $\mathscr{F}_i$ be the σ-algebra generated by $X_1, \ldots, X_i$, and define random variables $Z_0, \ldots, Z_n$ by $Z_i := \mathbb{E}[f(X)|\mathscr{F}_i]$. Note that $Z_0 = \mathbb{E}[f(X)]$ and $Z_n = f(X)$. Now consider the conditional increment $(Z_i - Z_{i-1})|\mathscr{F}_{i-1}$. First observe that

$$\mathbb{E}[Z_i - Z_{i-1}|\mathscr{F}_{i-1}] = 0,$$

so that the sequence $(Z_i)_{i\geq 0}$ is a *martingale*. Secondly, observe that

$$L_i \leq \left(Z_i - Z_{i-1}\middle|\mathscr{F}_{i-1}\right) \leq U_i,$$

where

$$\begin{aligned} L_i &:= \inf_{\ell} \mathbb{E}[f(X)|\mathscr{F}_{i-1}, X_i = \ell] - \mathbb{E}[f(X)|\mathscr{F}_{i-1}], \\ U_i &:= \sup_{u} \mathbb{E}[f(X)|\mathscr{F}_{i-1}, X_i = u] - \mathbb{E}[f(X)|\mathscr{F}_{i-1}]. \end{aligned}$$

Since $U_i - L_i \leq \mathcal{D}_i[f]$, Hoeffding's lemma implies that

$$\mathbb{E}\left[e^{s(Z_i - Z_{i-1})}\,\middle|\,\mathscr{F}_{i-1}\right] \leq e^{s^2\mathcal{D}_i[f]^2/8}. \tag{10.11}$$

Hence, for any $s \geq 0$,

$$\begin{aligned} &\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq t] \\ &\quad = \mathbb{P}\left[e^{s(f(X)-\mathbb{E}[f(X)])} \geq e^{st}\right] \\ &\quad \leq e^{-st}\mathbb{E}\left[e^{s(f(X)-\mathbb{E}[f(X)])}\right] && \text{by Markov's ineq.} \\ &\quad = e^{-st}\mathbb{E}\left[e^{s\sum_{i=1}^n Z_i - Z_{i-1}}\right] && \text{as a telescoping sum} \\ &\quad = e^{-st}\mathbb{E}\left[\mathbb{E}\left[e^{s\sum_{i=1}^n Z_i - Z_{i-1}}\middle|\mathscr{F}_{n-1}\right]\right] && \text{by the tower rule} \\ &\quad = e^{-st}\mathbb{E}\left[e^{s\sum_{i=1}^{n-1} Z_i - Z_{i-1}}\mathbb{E}\left[e^{s(Z_n - Z_{n-1})}\middle|\mathscr{F}_{n-1}\right]\right] \end{aligned}$$

since $Z_0, \ldots, Z_{n-1}$ are $\mathscr{F}_{n-1}$-measurable, and

$$\leq e^{-st}e^{s^2\mathcal{D}_n[f]^2/8}\mathbb{E}\left[e^{s\sum_{i=1}^{n-1} Z_i - Z_{i-1}}\right]$$

by (10.11). Repeating this argument a further $n-1$ times shows that

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq t] \leq \exp\left(-st + \frac{s^2}{8}\mathcal{D}[f]^2\right). \tag{10.12}$$

The right-hand side of (10.12) is minimized by $s = 4t/\mathcal{D}[f]^2$, which yields McDiarmid's inequality (10.8). The inequalities (10.9) and (10.10) follow easily from (10.8). □

10.4 ANOVA/HDMR Decompositions

The topic of this section is a variance-based decomposition of a function of n variables that goes by various names such as the *analysis of variance* (ANOVA), the *functional ANOVA*, the *high-dimensional model representation* (HDMR), or the *integral representation*. As before, let $(\mathcal{X}_i, \mathscr{F}_i, \mu_i)$ be a probability space for $i = 1, \dots, n$, and let $(\mathcal{X}, \mathscr{F}, \mu)$ be the product space. Write $\mathcal{N} := \{1, \dots, n\}$, and consider a ($\mathscr{F}$-measurable) function of interest $f\colon \mathcal{X} \to \mathbb{R}$. Bearing in mind that in practical applications n may be large (10^3 or more), it is of interest to efficiently identify

- which of the x_i contribute in the most dominant ways to the variations in $f(x_1, \dots, x_n)$,
- how the effects of multiple x_i are cooperative or competitive with one another,
- and hence construct a surrogate model for f that uses a lower-dimensional set of input variables, by using only those that give rise to dominant effects.

The idea is to write $f(x_1, \dots, x_n)$ as a sum of the form

$$\begin{aligned} f(x_1, \dots, x_n) &= f_\varnothing + \sum_{i=1}^{n} f_{\{i\}}(x_i) + \sum_{1 \le i < j \le n} f_{\{i,j\}}(x_i, x_j) + \dots \\ &= \sum_{I \subseteq \mathcal{N}} f_I(x_I). \end{aligned} \tag{10.13}$$

Experience suggests that 'typical real-world systems' f exhibit only low-order cooperativity in the effects of the input variables $x_1, \dots, x_n$. That is, the terms f_I with $|I| \gg 1$ are typically small, and a good approximation of f is given by, say, a second-order expansion,

$$f(x_1, \dots, x_n) \approx f_\varnothing + \sum_{i=1}^{n} f_{\{i\}}(x_i) + \sum_{1 \le i < j \le n} f_{\{i,j\}}(x_i, x_j).$$

Note, however, that low-order cooperativity does not necessarily imply that there is a small set of significant variables (it is possible that $f_{\{i\}}$ is large for most $i \in \{1, \dots, n\}$), nor does it say anything about the linearity or non-linearity of the input-output relationship. Furthermore, there are many HDMR-type expansions of the form given above; orthogonality criteria can be used to select a particular HDMR representation.

Recall that, for $I \subseteq \mathcal{N}$, the conditional expectation operator

$$f \mapsto \mathbb{E}_\mu[f(x_1, \dots, x_n) | x_i, i \in I] = \int_{\prod_{i \in I} \mathcal{X}_i} f(x_1, \dots, x_n) \, \mathrm{d} \bigotimes_{i \in I} \mu_i(x_i)$$

is an orthogonal projection operator from $L^2(\mathcal{X}, \mu; \mathbb{R})$ to the set of square-integrable measurable functions that are independent of x_i for $i \in I$, i.e. that depend only on x_i for $i \in \mathcal{N} \setminus I$. Let

$$P_\varnothing f := \mathbb{E}_\mu[f]$$

and, for non-empty $I \subseteq \mathcal{N}$,

$$P_I f := \mathbb{E}_\mu[f(x_1, \ldots, x_n)|x_i, i \notin I] - \sum_{J \subsetneq I} P_J f.$$

The functions $f_I := P_I f$ provide a decomposition of f of the desired form (10.13). By construction, we have the following:

Theorem 10.15 (ANOVA)**.** *For each $i \subseteq \mathcal{N}$, the linear operator P_I is an orthogonal projection of $L^2(\mathcal{X}, \mu; \mathbb{R})$ onto*

$$F_I := \left\{ f \;\middle|\; \begin{array}{c} f \text{ is independent of } x_j \text{ for } j \notin I \\ \text{and, for } i \in I,\ \int_0^1 f(x)\,\mathrm{d}\mu_i(x_i) = 0 \end{array} \right\} \subseteq L^2(\mathcal{X}, \mu; \mathbb{R}).$$

Furthermore, the linear operators P_I are idempotent, commutative and mutually orthogonal, i.e.

$$P_I P_J f = P_J P_I f = \begin{cases} P_I f, & \text{if } I = J, \\ 0, & \text{if } I \neq J, \end{cases}$$

and form a resolution of the identity:

$$\sum_{I \subseteq \mathcal{N}} P_I f = f.$$

Thus, $L^2(\mathcal{X}, \mu; \mathbb{R}) = \bigoplus_{I \subseteq \mathcal{N}} F_I$ is an orthogonal decomposition of $L^2(\mathcal{X}, \mu; \mathbb{R})$, so Parseval's formula implies the following decomposition of the variance $\sigma^2 := \|f - P_\varnothing f\|^2_{L^2(\mu)}$ of f:

$$\sigma^2 = \sum_{I \subseteq \mathcal{D}} \sigma_I^2, \tag{10.14}$$

where

$$\sigma_\varnothing^2 := 0,$$
$$\sigma_I^2 := \int_{\mathcal{X}} (P_I f)(x)^2\,\mathrm{d}\mu(x).$$

Two commonly used ANOVA/HDMR decompositions are *random sampling HDMR*, in which μ_i is uniform measure on $[0, 1]$, and Cut-HDMR, in which an expansion is performed with respect to a reference point $\bar{x} \in \mathcal{X}$, i.e. μ is the unit Dirac measure $\delta_{\bar{x}}$:

$$
\begin{aligned}
f_{\varnothing}(x) &= f(\bar{x}),\\
f_{\{i\}}(x) &= f(\bar{x}_1,\ldots,\bar{x}_{i-1},x_i,\bar{x}_{i+1},\ldots,\bar{x}_n) - f_{\varnothing}(x)\\
f_{\{i,j\}}(x) &= f(\bar{x}_1,\ldots,\bar{x}_{i-1},x_i,\bar{x}_{i+1},\ldots,\bar{x}_{j-1},x_j,\bar{x}_{j+1},\ldots,\bar{x}_n)\\
&\quad - f_{\{i\}}(x) - f_{\{j\}}(x) - f_{\varnothing}(x)\\
&\vdots
\end{aligned}
$$

Note that a component function f_I of a Cut-HDMR expansion vanishes at any $x \in \mathcal{X}$ that has a component in common with $\bar{x}$, i.e.

$$f_I(x) = 0 \quad \text{whenever } x_i = \bar{x}_i \text{ for some } i \in I.$$

Hence,

$$f_I(x) f_J(x) = 0 \quad \text{whenever } x_k = \bar{x}_k \text{ for some } k \in I \cup J.$$

Indeed, this orthogonality relation defines the Cut-HDMR expansion.

Sobol′ Sensitivity Indices. The decomposition of the variance (10.14) given by an HDMR/ANOVA decomposition naturally gives rise to a set of sensitivity indices for ranking the most important input variables and their cooperative effects. An obvious (and naïve) assessment of the relative importance of the variables x_I is the variance component σ_I^2, or the normalized contribution σ_I^2/σ^2. However, this measure neglects the contributions of those x_J with $J \subseteq I$, or those x_J such that J has some indices in common with I. With this in mind, Sobol′ (1990) defined sensitivity indices as follows:

Definition 10.16. Given an HDMR decomposition of a function f of n variables, the *lower and upper Sobol′ sensitivity indices* of $I \subseteq \mathcal{N}$ are, respectively,

$$\underline{\tau}_I^2 := \sum_{J \subseteq I} \sigma_J^2, \text{ and } \overline{\tau}_I^2 := \sum_{J \cap I \neq \varnothing} \sigma_J^2.$$

The *normalized lower and upper Sobol′ sensitivity indices* of $I \subseteq \mathcal{N}$ are, respectively,

$$\underline{s}_I^2 := \underline{\tau}_I^2/\sigma^2, \text{ and } \overline{s}_I^2 := \overline{\tau}_I^2/\sigma^2.$$

Since $\sum_{I \subseteq \mathcal{N}} \sigma_I^2 = \sigma^2 = \|f - f_{\varnothing}\|_{L^2}^2$, it follows immediately that, for each $I \subseteq \mathcal{N}$,

$$0 \leq \underline{s}_I^2 \leq \overline{s}_I^2 \leq 1.$$

Note, however, that while Theorem 10.15 guarantees that $\sigma^2 = \sum_{I \subseteq \mathcal{N}} \sigma_I^2$, in general Sobol′ indices satisfy no such additivity relation:

$$1 \neq \sum_{I \subseteq \mathcal{N}} \underline{s}_I^2 < \sum_{I \subseteq \mathcal{N}} \overline{s}_I^2 \neq 1.$$

The decomposition of variance (10.14), and sensitivity indices such as the Sobol′ indices, can also be used to form approximations to f with lower-dimensional input domain: see Exercise 10.8.

10.5 Active Subspaces

The global sensitivity measures discussed above, such as Sobol′ indices and McDiarmid diameters, can be used to identify a collection of important input *parameters* for a given response function. By way of contrast, the active subspace method seeks to identify a collection of important *directions* that are not necessarily aligned with the coordinate axes.

In this case, we take as the model input space $\mathcal{X} = [-1, 1]^n \subseteq \mathbb{R}^n$, and $f \colon \mathcal{X} \to \mathbb{R}$ is a function of interest. Suppose that, for each $x \in \mathcal{X}$, both $f(x) \in \mathbb{R}$ and $\nabla f(x) \in \mathbb{R}^n$ can be easily evaluated — note that evaluation of $\nabla f(x)$ might be accomplished by many means, e.g. finite differences, automatic differentiation, or use of the adjoint method. Also, let $\mathcal{X}$ be equipped with a probability measure μ. Informally, an active subspace for f will be a linear subspace of $\mathbb{R}^n$ for which f varies a lot more on average (with respect to μ) along directions in the active subspace than along those in the complementary inactive subspace.

Suppose that all pairwise products of the partial derivatives of f are integrable with respect to μ. Define $C = C(\nabla f, \mu) \in \mathbb{R}^{n \times n}$ by

$$C := \mathbb{E}_{X \sim \mu}\big[(\nabla f(X))(\nabla f(X))^{\mathsf{T}}\big]. \tag{10.15}$$

Note that C is symmetric and positive semi-definite, so it diagonalizes as

$$C = W \Lambda W^{\mathsf{T}},$$

where $W \in \mathbb{R}^{n \times n}$ is an orthogonal matrix whose columns $w_1, \ldots, w_n$ are the eigenvectors of C, and $\Lambda \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal entries $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, which are the corresponding eigenvalues of C. A quick calculation reveals that the eigenvalue λ_i is nothing other than the mean-squared value of the directional derivative in the direction w_i:

$$\lambda_i = w_i^{\mathsf{T}} C w_i = w_i^{\mathsf{T}} \mathbb{E}_{\mu}\big[(\nabla f)(\nabla f)^{\mathsf{T}}\big] w_i = \mathbb{E}_{\mu}\big[(\nabla f \cdot w_i)^2\big]. \tag{10.16}$$

In general, the eigenvalues of C may be any non-negative reals. If, however, some are clearly 'large' and some are 'small', then this partitioning of the eigenvalues and observation (10.16) can be used to define a new coordinate system on $\mathbb{R}^n$ such that in some directions f values 'a lot' and on others it varies 'only a little'. More precisely, write Λ and W in block form as

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}, \quad \text{and} \quad W = \begin{bmatrix} W_1 & W_2 \end{bmatrix}, \tag{10.17}$$

where $\Lambda_1 \in \mathbb{R}^{k\times k}$ and $W_1 \in \mathbb{R}^{n\times k}$ with $k \leq n$; of course, the idea is that $k \ll n$, and that $\lambda_k \gg \lambda_{k+1}$. This partitioning of the eigenvalues and eigenvectors of C defines new variables $y \in \mathbb{R}^k$ and $z \in \mathbb{R}^{n-k}$ by

$$y := W_1^\mathsf{T} x, \quad \text{and} \quad z := W_2^\mathsf{T} x. \tag{10.18}$$

so that $x = W_1 y + W_2 z$. Note that the (y, z) coordinate system is simply a rotation of the original x coordinate system. The k-dimensional subspace spanned by $w_1, \ldots, w_k$ is called the *active subspace* for f over $\mathcal{X}$ with respect to μ. The heuristic requirement that f should vary mostly in the directions of the active subspace is quantified by the eigenvalues of C:

Proposition 10.17. *The mean-squared gradients of f with respect to the active coordinates $y \in \mathbb{R}^k$ and inactive coordinates $z \in \mathbb{R}^{n-k}$ satisfy*

$$\begin{aligned} \mathbb{E}_\mu\big[(\nabla_y f)^\mathsf{T}(\nabla_y f)\big] &= \lambda_1 + \cdots + \lambda_k, \\ \mathbb{E}_\mu\big[(\nabla_z f)^\mathsf{T}(\nabla_z f)\big] &= \lambda_{k+1} + \cdots + \lambda_n. \end{aligned}$$

Proof. By the chain rule, the gradient of $f(x) = f(W_1 y + W_2 z)$ with respect to y is given by

$$\begin{aligned} \nabla_y f(x) &= \nabla_y f(W_1 y + W_2 z) \\ &= W_1^\mathsf{T} \nabla_x f(W_1 y + W_2 z) \\ &= W_1^\mathsf{T} \nabla_x f(x). \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}_\mu\big[(\nabla_y f)^\mathsf{T}(\nabla_y f)\big] &= \mathbb{E}_\mu\big[\operatorname{tr}\big((\nabla_y f)(\nabla_y f)^\mathsf{T}\big)\big] \\ &= \operatorname{tr}\mathbb{E}_\mu\big[(\nabla_y f)(\nabla_y f)^\mathsf{T}\big] \\ &= \operatorname{tr}\big(W_1^\mathsf{T}\mathbb{E}_\mu\big[(\nabla_x f)(\nabla_x f)^\mathsf{T}\big]W_1\big) \\ &= \operatorname{tr}\big(W_1^\mathsf{T} C W_1\big) \\ &= \operatorname{tr}\Lambda_1 \\ &= \lambda_1 + \cdots + \lambda_k. \end{aligned}$$

This proves the claim for the active coordinates $y \in \mathbb{R}^k$; the proof for the inactive coordinates $z \in \mathbb{R}^{n-k}$ is similar. □

Proposition 10.17 implies that a function for which $\lambda_{k+1} = \cdots = \lambda_n = 0$ has $\nabla_z f = 0$ μ-almost everywhere in $\mathcal{X}$. Unsurprisingly, for such functions, the value of f depends only on the active variable y and not upon the inactive variable z:

Proposition 10.18. *Suppose that μ is absolutely continuous with respect to Lebesgue measure on $\mathcal{X}$, and suppose that $f\colon \mathcal{X} \to \mathbb{R}$ is such that $\lambda_{k+1} = \cdots = \lambda_n = 0$. Then, whenever $x_1, x_2 \in \mathcal{X}$ have equal active component, i.e. $W_1^\mathsf{T} x_1 = W_1^\mathsf{T} x_2$, it follows that $f(x_1) = f(x_2)$ and $\nabla_x f(x_1) = \nabla_x f(x_2)$.*

Proof. The gradient $\nabla_z f$ being zero everywhere in $\mathcal{X}$ implies that $f(x_1) = f(x_2)$. To show that the gradients are equal, assume that x_1 and x_2 lie in the interior of $\mathcal{X}$. Then for any $v \in \mathbb{R}^n$, let

$$x_1' = x_1 + hv, \quad \text{and} \quad x_2' = x_2 + hv,$$

where $h \in \mathbb{R}$ is small enough that x_1' and x_2' lie in the interior of $\mathcal{X}$. Note that $W_1^\mathsf{T} x_1' = W_1^\mathsf{T} x_2'$, and so $f(x_1') = f(x_2')$. Then

$$\begin{aligned} c &= v \cdot (\nabla_x f(x_1) - \nabla_x f(x_2)) \\ &= \lim_{h \to 0} \frac{(f(x_1') - f(x_1)) - (f(x_2') - f(x_2))}{h} \\ &= 0. \end{aligned}$$

Simple limiting arguments can be used to extend this result to x_1 or $x_2 \in \partial\mathcal{X}$. Since $v \in \mathbb{R}^n$ was arbitrary, it follows that $\nabla_x f(x_1) = \nabla_x f(x_2)$. □

Example 10.19. In some cases, the active subspace can be identified exactly from the form of the function f:

(a) Suppose that f is a *ridge function*, i.e. a function of the form $f(x) := h(a \cdot x)$, where $h\colon \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R}^n$. In this case, C has rank one, and the eigenvector defining the active subspace is $w_1 = a/\|a\|$, which can be discovered by a single evaluation of the gradient anywhere in $\mathcal{X}$.

(b) Consider $f(x) := h(x \cdot Ax)$, where $h\colon \mathbb{R} \to \mathbb{R}$ and $A \in \mathbb{R}^{n \times n}$ is symmetric. In this case,

$$C = 4A\mathbb{E}[(h')^2 x x^\mathsf{T}]A^\mathsf{T},$$

where $h' = h'(x \cdot Ax)$ is the derivative of h. Provided h' is non-degenerate, $\ker C = \ker A$.

Numerical Approximation of Active Subspaces. When the expected value used to define the matrix C and hence the active subspace decomposition is approximated using Monte Carlo sampling, the active subspace method has a nice connection to the singular value decomposition (SVD). That is, suppose that $x^{(1)}, \ldots, x^{(M)}$ are M independent draws from the probability measure μ. The corresponding Monte Carlo approximation to C is

$$C \approx \widehat{C} := \frac{1}{M} \sum_{m=1}^{M} \nabla f(x^{(m)}) \nabla f(x^{(m)})^\mathsf{T}.$$

The eigendecomposition of $\widehat{C}$ as $\widehat{C} = \widehat{W}\widehat{\Lambda}\widehat{W}^{\mathsf{T}}$ can be computed as before. However, if

$$G := \frac{1}{\sqrt{M}}\left[\nabla f(x^{(1)}) \quad \cdots \quad \nabla f(x^{(M)})\right] \in \mathbb{R}^{n\times M},$$

then $\widehat{C} = GG^{\mathsf{T}}$, and an SVD of G is given by $G = \widehat{W}\widehat{\Lambda}^{1/2}V^{\mathsf{T}}$ for some orthogonal matrix V. In practice, the eigenpairs $\widehat{W}$ and $\widehat{\Lambda}$ from the finite-sample approximation $\widehat{C}$ are used as approximations of the true eigenpairs W and Λ of C.

The SVD approach is more numerically stable than an eigendecomposition, and is also used in the technique of *principal component analysis* (PCA). However, PCA applies the SVD to the rectangular matrix whose columns are samples of a vector-valued response function, and posits a linear model for the data; the active subspace method applies the SVD to the rectangular matrix whose columns are the gradient vectors of a scalar-valued response function, and makes no linearity assumption about the model.

Example 10.20. Consider the Van der Pol oscillator

$$\ddot{u}(t) - \mu(1 - u(t)^2)\dot{u}(t) + \omega^2 u(t) = 0,$$

with the initial conditions $u(0) = 1$, $\dot{u}(0) = 0$. Suppose that we are interested in the state of the oscillator at time $T := 2\pi$; if $\omega = 1$ and $\mu = 0$, then $u(T) = u(0) = 1$. Now suppose that $\omega \sim \text{Unif}([0.8, 1.2])$ and $\mu \sim \text{Unif}([0, 5])$; a contour plot of $u(T)$ as a function of ω and μ is shown in Figure 10.1(a).

Sampling the gradient of $u(T)$ with respect to the normalized coordinates

$$x_1 := 2\frac{\omega - 0.8}{1.2 - 0.8} - 1 \in [-1, 1]$$
$$x_2 := 2\frac{\mu}{5} - 1 \in [-1, 1]$$

gives an approximate covariance matrix

$$\mathbb{E}\left[\nabla_x u(T)(\nabla_x u(T))^{\mathsf{T}}\right] \approx \widehat{C} = \begin{bmatrix} 1.776 & -1.389 \\ -1.389 & 1.672 \end{bmatrix},$$

which has the eigendecomposition $\widehat{C} = \widehat{W}\widehat{\Lambda}\widehat{W}^{\mathsf{T}}$ with

$$\widehat{\Lambda} = \begin{bmatrix} 3.115 & 0 \\ 0 & 0.3339 \end{bmatrix} \quad \text{and} \quad \widehat{W} = \begin{bmatrix} 0.7202 & 0.6938 \\ -0.6938 & 0.7202 \end{bmatrix}.$$

Thus — at least over this range of the ω and μ parameters — this system has an active subspace in the direction $w_1 = (0.7202, -0.6938)$ in the normalized

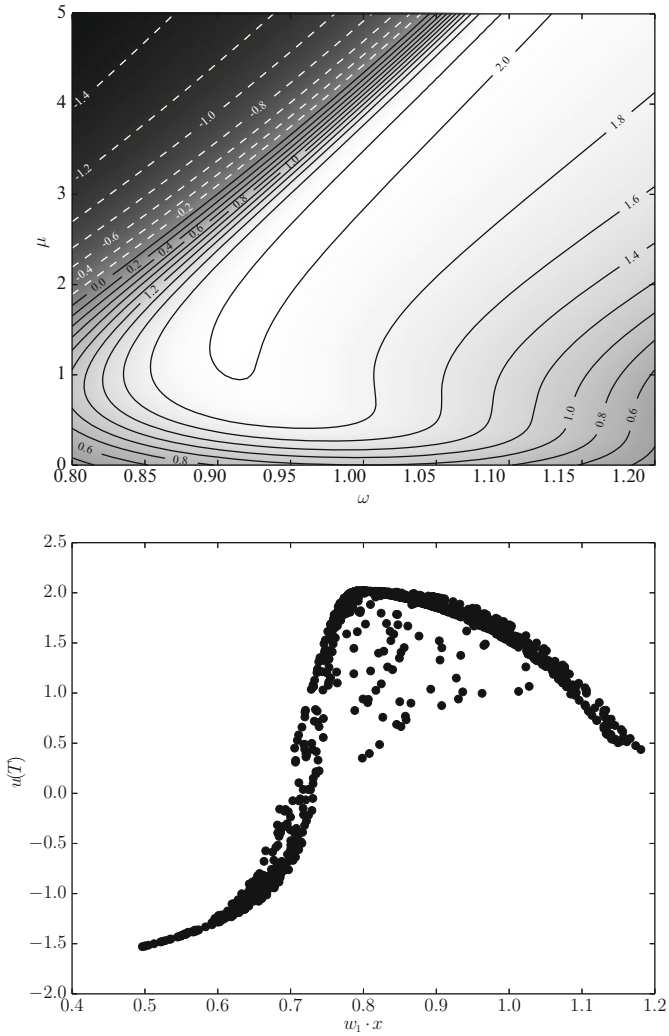

Fig. 10.1: Illustration of Example 10.20. Subfigure (a) shows contours of the state at time $T = 2\pi$ of a Van der Pol oscillator with initial state 1.0 and velocity 0.0, as a function of natural frequency ω and damping μ. This system has an active subspace in the $(0.144, -1.735)$ direction; roughly speaking, 'most' of the contours are perpendicular to this direction. Subfigure (b) shows a projection onto this directions of 1000 samples of $u(T)$, with uniformly distributed ω and μ, in the style of Exercise 10.9; this further illustrates the almost one-dimensional nature of the system response.

x-coordinate system. In the original (ω, μ)-coordinate system, this active subspace lies in the $(0.144, -1.735)$ direction.

Applications of Active Subspaces. The main motivation for determining an active subspace for $f\colon \mathcal{X} \to \mathbb{R}$ is to then approximate f by a function F of the active variables alone, i.e.

$$f(x) = f(W_1 y + W_2 z) \approx F(W_1 y).$$

Given such an approximation, $F \circ W_1$ can be used as a proxy for f for the purposes of optimization, optimal control, forward and inverse uncertainty propagation, and so forth.

10.6 Bibliography

General references for sensitivity analysis, in theory and in application, include the book of Saltelli et al. (2008) and the two-volume monograph of Cacuci et al. (2003, 2005). Smith (2014) discusses local and global sensitivity analysis in Chapters 14 and 15 respectively.

Detailed treatment of the singular value decomposition, including proof of Theorem 10.1, can be found in many texts on numerical linear algebra, such as those of Golub and Van Loan (2013, Section 2.4) and Trefethen and Bau (1997, Lectures 4, 5, and 31). See also Stewart (1993) for some historical discussion of the SVD. Jolliffe (2002) gives a general introduction to principal component analysis, and de Leeuw (2013) gives a survey of the history of PCA and its nonlinear generalizations.

The book of Griewank and Walther (2008) and the article of Neidinger (2010) are good references for the theory, implementation, and applications of automatic differentiation. See Krantz and Parks (2013, Section 3.4) for the proof and discussion of the implicit function theorem in Banach spaces.

McDiarmid's inequality appears in McDiarmid (1989), although the underlying martingale results go back to Hoeffding (1963) and Azuma (1967); see Lucas et al. (2008) for some UQ studies using McDiarmid diameters. General presentations of the concentration-of-measure phenomenon, including geometrical considerations such as isoperimetric inequalities, can be found in Ledoux (2001) and Ledoux and Talagrand (2011).

In the statistical literature, the analysis of variance (ANOVA) method originates with Fisher and Mackenzie (1923). The ANOVA decomposition was generalized by Hoeffding (1948) to functions in $L^2([0,1]^d)$ for $d \in \mathbb{N}$; for $d = \infty$, see Owen (1998). That generalization can easily be applied to L^2 functions on any product domain, and leads to the functional ANOVA of Sobol′ (1993) and Stone (1994). In the mathematical chemistry literature, the HDMR was popularized by Rabitz and Alış (1999). The treatment of ANOVA/HDMR in

this chapter also draws upon the presentations of Beccacece and Borgonovo (2011), Holtz (2011), and Hooker (2007).

The name "active subspace method" appears to have been coined by Russi (2010). Further discussion of active subspace methods, including numerical implementation issues, can be found in Constantine (2015).

10.7 Exercises

Exercise 10.1. Consider a power series $f(x) := \sum_{n \in \mathbb{N}_0} a_n x^n$, thought of as a function $f \colon \mathbb{R} \to \mathbb{R}$, with radius of convergence R. Show that the extension $f \colon \mathbb{R}_\epsilon \to \mathbb{R}_\epsilon$ of f to the dual numbers satisfies

$$f(x + \epsilon) = f(x) + f'(x)\epsilon$$

whenever $|x| < R$. Hence show that, if $g \colon \mathbb{R} \to \mathbb{R}$ is an analytic function, then $g'(x)$ is the coefficient of ϵ in $g(x + \epsilon)$.

Exercise 10.2. An example partial implementation of dual numbers in Python is as follows:

```
class DualNumber(object):
  def __init__(self, r, e):
    # Initialization of real and infinitesimal parts.
    self.r = r
    self.e = e
  def __repr__(self):
    # How to print dual numbers
    return str(self.r) + " + " + str(self.e) + " * e"
  def __add__(self, other):
    # Overload the addition operator to allow addition of
    # dual numbers.
    if not isinstance(other, DualNumber):
      new_other = DualNumber(other, 0)
    else:
      new_other = other
    r_part = self.r + new_other.r
    e_part = self.e + new_other.e
    return DualNumber(r_part, e_part)
```

Following the template of the overloaded addition operator, write analogous methods `def __sub__(self, other)`, `def __mul__(self, other)`, and `def __div__(self, other)` for this `DualNumber` class to overload the subtraction, multiplication and division operators. The result will be that any numerical function you have written using the standard arithmetic operations

`+`, `-`, `*`, and `/` will now accept `DualNumber` arguments and return `DualNumber` values in accordance with the rules of dual number arithmetic.

Once you have done this, the following function will accept a function `f` as its argument and return a new function `f_prime` that is the derivative of `f`, calculated using automatic differentiation:

```
def AutomaticDerivative(f):
  # Accepts a function f as an argument and returns a new
  # function that is the derivative of f, calculated using
  # automatic differentiation.
  def f_prime(x):
    f_x_plus_eps = f(DualNumber(x, 1))
    deriv = f_x_plus_eps.e
    return deriv
  return f_prime
```

Test this function using several functions of your choice, and verify that it correctly calculates the derivative of a product (the Leibniz rule), a quotient and a composition (the chain rule).

Exercise 10.3. Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a polynomial or convergent power series

$$f(x) = \sum_{\alpha} c_\alpha x^\alpha$$

in $x = (x_1, \ldots, x_n)$, where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ are multi-indices, $c_\alpha \in \mathbb{R}^m$, and $x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. Consider the dual vectors over $\mathbb{R}^n$ obtained by adjoining a vector element $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ such that $\epsilon_i \epsilon_j = 0$ for all $i, j \in \{1, \ldots, n\}$. Show that

$$f(x + \epsilon) = \sum_{\alpha} c_\alpha \sum_{i=1}^{n} \alpha_i x^{\alpha - e_i} \epsilon_i$$

and hence that $\frac{\partial f}{\partial x_i}(x)$ is the coefficient of ϵ_i in $f(x + \epsilon)$.

Exercise 10.4. Consider an ODE of the form $\dot{u}(t) = f(u(t); \theta)$ for an unknown $u(t) \in \mathbb{R}$, where $\theta \in \mathbb{R}$ is a vector of parameters, and $f\colon \mathbb{R}^2 \to \mathbb{R}$ is a smooth vector field. Define the local sensitivity of the solution u about a nominal parameter value $\theta^* \in \mathbb{R}$ to be the partial derivative $s := \frac{\partial u}{\partial \theta}(\theta^*)$. Show that this sensitivity index s evolves according to the adjoint equation

$$\dot{s}(t) = \frac{\partial f}{\partial u}\big(u(t; \theta^*); \theta^*\big) s(t) + \frac{\partial f}{\partial \theta}\big(u(t; \theta^*); \theta^*\big).$$

Extend this result to a vector-valued unknown $u(t)$, and vector of parameters $\theta = (\theta_1, \ldots, \theta_n)$.

Exercise 10.5. Show that, for each $j = 1, \ldots, n$, the McDiarmid subdiameter $\mathcal{D}_j[\,\cdot\,]$ is a seminorm on the space of bounded functions $f\colon \mathcal{X} \to \mathbb{K}$, as is the McDiarmid diameter $\mathcal{D}[\,\cdot\,]$. What are the null-spaces of these seminorms?

Exercise 10.6. Define, for constants $a, b, c, d \in \mathbb{R}$, $f\colon [0,1]^2 \to \mathbb{R}$ by

$$f(x_1, x_2) := a + bx_1 + cx_2 + dx_1x_2.$$

Show that the ANOVA decomposition of f (with respect to uniform measure on the square) is

$$\begin{aligned} f_{\varnothing} &= a + \tfrac{b}{2} + \tfrac{c}{2} + \tfrac{d}{2}, \\ f_{\{1\}}(x_1) &= \big(b + \tfrac{d}{2}\big)\big(x_1 - \tfrac{1}{2}\big), \\ f_{\{2\}}(x_2) &= \big(c + \tfrac{d}{2}\big)\big(x_2 - \tfrac{1}{2}\big), \\ f_{\{1,2\}}(x_1, x_2) &= d\big(x_1 - \tfrac{1}{2}\big)\big(x_2 - \tfrac{1}{2}\big). \end{aligned}$$

Exercise 10.7. Let $f\colon [-1,1]^2 \to \mathbb{R}$ be a function of two variables. Sketch the vanishing sets of the component functions of f in a Cut-HDMR expansion through $\bar{x} = (0,0)$. Do the same exercise for $f\colon [-1,1]^3 \to \mathbb{R}$ and $\bar{x} = (0,0,0)$, taking particular care with second-order terms like $f_{\{1,2\}}$.

Exercise 10.8. For a function $f\colon [0,1]^n \to \mathbb{R}$ with variance σ^2, suppose that the input variables of f have been ordered according to their importance in the sense that $\sigma^2_{\{1\}} \geq \sigma^2_{\{2\}} \geq \cdots \geq \sigma^2_{\{n\}} \geq 0$. The *truncation dimension* of f with proportion $\alpha \in [0,1]$ is defined to be the least $d_{\mathrm{t}} = d_{\mathrm{t}}(\alpha) \in \{1, \ldots, n\}$ such that

$$\sum_{\varnothing \neq I \subseteq \{1,\ldots,d_{\mathrm{t}}\}} \sigma_I^2 \geq \alpha\sigma^2,$$

i.e. the first d_{t} inputs explain a proportion α of the variance of f. Show that

$$f_{d_{\mathrm{t}}}(x) := \sum_{I \subseteq \{1,\ldots,d_{\mathrm{t}}\}} f_I(x_I)$$

is an approximation to f with error $\big\|f - f_{d_{\mathrm{t}}}\big\|^2_{L^2} \leq (1-\alpha)\sigma^2$. Formulate and prove a similar result for the *superposition dimension* d_{s}, the least $d_{\mathrm{s}} = d_{\mathrm{s}}(\alpha) \in \{1, \ldots, n\}$ such that

$$\sum_{\substack{\varnothing \neq I \subseteq \{1,\ldots,n\} \\ \#I \leq d_{\mathrm{s}}}} \sigma_I^2 \geq \alpha\sigma^2,$$

Exercise 10.9. Building upon the notion of a *sufficient summary plot* developed by Cook (1998), Constantine (2015, Section 1.3) offers the following "quick and dirty" check for a one-dimensional active subspace for $f\colon [-1,1]^n \to \mathbb{R}$ that can be evaluated a limited number — say, M — times with the available resources:

(a) Draw M samples $x_1, \ldots, x_M \in [-1,1]^n$ according to some probability distribution on the cube, e.g. uniform measure.
(b) Evaluate $f(x_m)$ for $m = 1, \ldots, M$.

(c) Find $(a_0, a_1, \ldots, a_n) \in \mathbb{R}^{1+n}$ to minimize

$$J(a) := \frac{1}{2} \left\| \begin{bmatrix} 1 & x_1^\mathsf{T} \\ \vdots & \vdots \\ 1 & x_M^\mathsf{T} \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} - \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \right\|_2^2 .$$

is minimal. Note that this step can be interpreted as forming a linear statistical regression model.

(d) Let $a' := (a_1, \ldots, a_n)$, and define a unit vector $w \in \mathbb{R}^n$ by $w := a'/\|a'\|_2$.

(e) Produce a scatter plot of the points $(w \cdot x_m, f(x_m))$ for $m = 1, \ldots, M$. If this scatter plot looks like the graph of a single-valued function, then this is a good indication that f has a one-dimensional active subspace in the w direction.

One interpretation of this procedure is that it looks for a rotation of the domain $[-1, 1]^n$ such that, in this rotated frame of reference, the graph of f looks 'almost' like a curve — though it is not necessary that f be a *linear* function of $w \cdot x$. Examine your favourite model f for a one-dimensional active subspace in this way.