ESAIM: M2AN 47 (2013) 1657–1689
DOI: 10.1051/m2an/2013083

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

OPTIMAL UNCERTAINTY QUANTIFICATION FOR LEGACY DATA OBSERVATIONS OF LIPSCHITZ FUNCTIONS

T.J. Sullivan[1], M. McKerns[2], D. Meyer[4], F. Theil[4], H. Owhadi[5] and M. Ortiz[6]

Abstract. We consider the problem of providing optimal uncertainty quantification (UQ) – and hence rigorous certification – for partially-observed functions. We present a UQ framework within which the observations may be small or large in number, and need not carry information about the probability distribution of the system in operation. The UQ objectives are posed as optimization problems, the solutions of which are optimal bounds on the quantities of interest; we consider two typical settings, namely parameter sensitivities (McDiarmid diameters) and output deviation (or failure) probabilities. The solutions of these optimization problems depend non-trivially (even non-monotonically and discontinuously) upon the specified legacy data. Furthermore, the extreme values are often determined by only a few members of the data set; in our principal physically-motivated example, the bounds are determined by just 2 out of 32 data points, and the remainder carry no information and could be neglected without changing the final answer. We propose an analogue of the simplex algorithm from linear programming that uses these observations to offer efficient and rigorous UQ for high-dimensional systems with high-cardinality legacy data. These findings suggest natural methods for selecting optimal (maximally informative) next experiments.

Mathematics Subject Classification. 60E15, 62G99, 65C50, 90C26.

Received February 9, 2012.
Published online August 30, 2013.

Keywords and phrases. Uncertainty quantification, probability inequalities, non-convex optimization, Lipschitz functions, legacy data, point observations.

[1] Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK. Tim.Sullivan@warwick.ac.uk

[2] Center for Advanced Computing Research, California Institute of Technology, 1200 East California Boulevard, Mail Code 158-79, Pasadena, CA 91125, USA. mmckerns@caltech.edu

[3] Lehrstuhl für Numerische Mechanik, Technische Universität München, Boltzmannstrasse 15, 85747, Garching bei München, Germany. meyer@lnm.mw.tum.de

[4] Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK. f.theil@warwick.ac.uk

[5] Applied & Computational Mathematics and Control & Dynamical Systems, California Institute of Technology, Mail Code 9-94, 1200 East California Boulevard, Pasadena, CA 91125, USA. owhadi@caltech.edu

[6] Division of Engineering and Applied Science, California Institute of Technology, Mail Code 105-50, 1200 East California Boulevard, Pasadena, CA 91125, USA. ortiz@caltech.edu

Article published by EDP Sciences

© EDP Sciences, SMAI 2013

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

1. Introduction and outline

1.1. Introduction

In many settings – including the physical sciences, engineering, and finance – it is necessary to have a rigorous and also sharp/optimal quantitative understanding of the effects of uncertainties, which are often probabilistic in nature. Often, the available information about the system of interest comes in the form of *legacy data*, *i.e.* a data set that is provided "as is" and cannot be extended; the reasons for such restrictions may range from financial or practical difficulties to legal and ethical concerns. Uncertainty quantification (UQ) methods for addressing such problems must cope with this non-extensibility, the fact that the distribution of the legacy data may be unrelated to the probability distribution of the system in operation, and that the data set may be either very sparse or very large compared to the system's domain of operation. This paper approaches the UQ-with-legacy-data problem using the *Optimal UQ* framework proposed in [28], and thereby develops and illustrates that general framework in a specific setting.

In the Optimal UQ framework [28], UQ in the presence of both epistemic and aleatoric uncertainties [16,26,30] is posed as an optimization problem over all feasible scenarios that are consistent with the available information about the input uncertainties – those uncertainties may be infinite-dimensional in nature, and concern unknown or partially-known probability distributions and functions. In many cases, the corresponding infinite-dimensional optimization problem can be reduced to an equivalent finite-dimensional problem that allows for closed-form or numerical evaluation [28], Section 3.

Many UQ methods are not directly applicable if the available data are of legacy type. For example, in [17], it was proposed that rigorous certification of physical systems be performed using a concentration-of-measure inequality known as *McDiarmid's inequality* [18–20], also known as the *bounded differences inequality*. However, this method and its variants [1,13,36] require extensive data "on demand" in order to compute the McDiarmid diameter, which measures the system output variability and provides the concentration rate in McDiarmid's inequality. Section 3 of the present paper shows how the McDiarmid diameter of a Lipschitz function can be optimally bounded using legacy data observations of that function and (upper bounds on) the Lipschitz constants.

Relationships between the smoothness properties of a function f and bounds on deviation probabilities for f have been studied extensively. For Lipschitz functions, Talagrand's inequality [37] is a famous result in this area; a discussion of non-Lipschitz functions can be found in [40]. However, while such results do use the smoothness information, they do not use arbitrarily-located known values, *i.e.* point observations, of f. On the other hand, there are methods that use smoothness information and point observations to calculate the extreme values of f (notably, the algorithm of [12] does so without *a priori* knowledge of the Lipschitz constant), but these methods (a) direct further function evaluations, which are not permitted in the context of legacy data, and (b) do not appear to have been coupled to concentration-of-measure methods to produce probability-of-deviation inequalities. This last point is not surprising, since it is difficult to *prove* a general theorem that will make optimal or near-optimal use of data in advance of knowing those data.

Motivated by this, Section 4 shows how to *calculate* optimal upper bounds on the probability of deviations from the mean (or any linear function of the system's *a priori* unknown probability distribution) given the legacy data and (upper bounds on) the Lipschitz constants; this second approach forms part of a large and growing body of work concerning the calculation of optimal inequalities in probability theory – see *e.g.* [3,5,28] for some surveys and historical remarks on this topic. We find that the extremizers for our optimization problems tend to have a very simple, low-dimensional, singular structure. Furthermore, once this singular structure has been observed, even approximately, it can be exploited to greatly reduce the computational burden; see Remark 7.2 and Figure 11.

It is also shown that, in certain cases, additional information (in the form of new observations) may not propagate to the resulting bounds, or, dually, that the bounds may be determined by a relatively small "active" subset of a large data set. In Algorithm 5.5 we propose an analogue of the simplex algorithm in linear programming that uses these observations to offer efficient and rigorous UQ for high-dimensional systems with

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

high-cardinality legacy data. The motivating idea for this algorithm is to solve easier (less constrained) optimization problems when possible, and that the algorithm should terminate in a number of iterations of the same order as the number of relevant data points. In addition, in the case that the data set can be extended, the optimization formulation of the UQ objectives provides a natural notion of best next experiment (and hence maximally informative data set): it is the experiment that would induce the greatest change in the extreme value of the UQ optimization problem.

The methods and results of this paper are predicated upon having suitable information (or making assumptions) about the system of interest. As noted by Hoeffding [9], assumptions about the system of interest play a central and sensitive role in any statistical decision problem, even though the assumptions are often only approximations of reality. To illustrate the effect of information/assumptions, consider the following toy problem, which will be considered in further detail in Example 4.3 and treated numerically in subsection 7.2:

Example 1.1. Suppose that a measurable function $G\colon [0,1] \to \mathbb{R}$ is applied to a random variable X with unknown distribution on $[0,1]$, and the event $[G(X) \leq 0]$ is considered to constitute "failure". Given the values of G on some proper (usually finite) subset $\mathcal{O} \subsetneq [0,1]$, what is the optimal (*i.e.* least) upper bound $\widehat{P}$ on the failure probability $\mathbb{P}[G(X) \leq 0]$? (Note well that the points of $\mathcal{O}$ may be unrelated to the distribution of X, and so classical methods of statistical reasoning using the sample set $\{G(z) \mid z \in \mathcal{O}\}$ are inapplicable.)

With this information alone, the only rigorous upper bound that can be given is the trivial one: $\mathbb{P}[G(X) \leq 0] \leq \widehat{P} = 1$. Consider now the impact of two further pieces of information:

(I) G is Lipschitz continuous with Lipschitz constant 1, or *short*, *i.e.*

$$|G(x) - G(x')| \leq |x - x'| \text{ for all } x, x' \in [0,1],$$

and hence G is continuous on $[0,1]$, and by Rademacher's theorem is differentiable with $|G'(x)| \leq 1$ for Lebesgue-almost-every $x \in [0,1]$;

(II) some information about the distribution of X on $[0,1]$ or the distribution of $G(X)$ on $\mathbb{R}$, *e.g.* that $\mathbb{E}[G(X)] \geq m$ for some known m.

The first item of information does not generally provide any improvement on the trivial upper bound, since although it constrains the set of points $x \in [0,1]$ for which it is possible that $G(x) \leq 0$, it says nothing about the $\mathbb{P}$-measure of that set, unless it is found to be empty. However, taken together, $G|_{\mathcal{O}}$ and the two additional items of information *do* provide non-trivial bounds on $\mathbb{P}[G(X) \leq 0]$. Evaluating these bounds is an infinite-dimensional but well-posed optimization problem, which can be reduced to an equivalent finite-dimensional problem by the reduction theorems of [28]. Indeed, as will be shown later, if $\mathcal{O}$ consists of one point – *i.e.* we know one point $(z, G(z))$ that lies on the graph of G – then the least upper bound on $\mathbb{P}[G(X) \leq 0]$ can be given in closed form. This bound is given in (4.14), and surface and contour plots are given in Figure 1. Notably, the bound is both non-monotone and discontinuous with respect to the data point $(z, G(z))$.

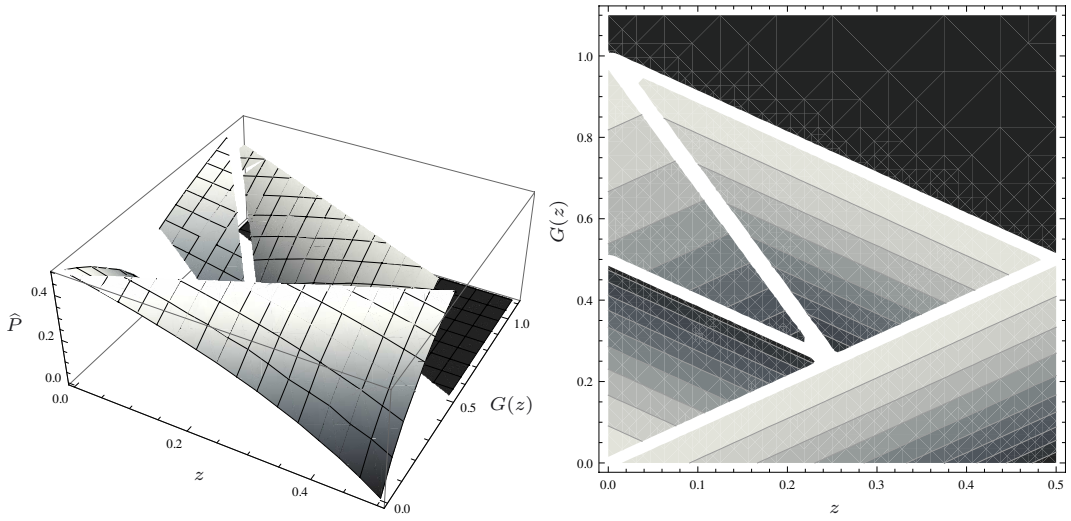
1.2. Outline

Section 2 establishes the notation and set-up of the problems of interest, and recalls a theorem of McShane [24] that will be useful later on.

Section 3 treats the determination of optimal upper bounds on McDiarmid diameters (*i.e.* L^∞ semi-norms on component-wise oscillations of a function of several independent inputs) using legacy data and Lipschitz constants. Such upper bounds can be used, together with McDiarmid's inequality and the mean performance of the system, to provide rigorous upper bounds on the system's probability of failure.

Section 4 treats the problem of directly and optimally bounding the probability of failure, *i.e.* finding the least upper bound that is consistent with the legacy data, the Lipschitz constants, and the specified mean performance. This problem is harder to solve than the problem of Section 3, but is still tractable, and has the advantage that it provides the optimal bound on the probability of failure given all the available information, whereas McDiarmid's inequality is non-optimal.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press



(a) Surface plot.

(b) Contour plot.

FIGURE 1. Plots of $\widehat{P}$, the least upper bound on $\mathbb{P}[G(X) \leq 0]$ given that $G\colon [0,1] \to \mathbb{R}$ has Lipschitz constant 1, mean $\frac{1}{2}$, and has $(z, G(z))$ on its graph, as a function of $(z, G(z)) \in [0, \frac{1}{2}] \times \mathbb{R}$. Note the discontinuity and non-monotonicity of $\widehat{P}$ as a function of $(z, G(z))$.

Section 5 discusses necessary and sufficient conditions for a data point to be relevant to the solution of the problems in Sections 3 and 4; put another way, this section concerns the identification of redundant information.

Section 6 contains some general remarks applicable to both Sections 3 and 4.

Section 7 gives the results of some example numerical implementations of the problems of Section 4. In this section, we see that many data points may be redundant in the sense of Section 5, and hence that optimal UQ for systems with large legacy data sets may be given by considering well-chosen small subsets of the larger data set.

Section 8 outlines some directions for generalization and future work.

2. Review and notation

2.1. Notation

Let $(\mathcal{X}_k, d_k)$ be a metric space for each $k \in \{1, \ldots, K\}$; prototypically, $\mathcal{X}_k = \mathbb{R}$ or $[a_k, b_k] \subseteq \mathbb{R}$ with the Euclidean distance $d_k(x, y) := |x - y|$. Let $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$. Let $G\colon \mathcal{X} \to \mathbb{R}$ be some function and suppose that, for each $k \in \{1, \ldots, K\}$, L_k is a global Lipschitz constant for G with respect to its kth argument: *i.e.*,

$$(x, x' \in \mathcal{X}, x^j = x'^j \text{ for } j \neq k) \implies |G(x) - G(x')| \leq L_k d_k(x^k, x'^k). \tag{2.1}$$

Define a quasi-metric $d_L\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$d_L(x, x') := \sum_{k=1}^{K} L_k d_k(x^k, x'^k). \tag{2.2}$$

If all L_k are strictly positive, then d_L is a metric. In the prototypical case, d_L is a rescaling of the ℓ^1 "Manhattan" metric on $\mathbb{R}^K$.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

Lemma 2.1. *A function $f\colon \mathcal{X} \to \mathbb{R}$ is Lipschitz with Lipschitz constant L_k in its k*th *argument if, and only if, it is short with respect to the metric d_L,* i.e.

$$|f(x) - f(x')| \leq d_L(x, x') \text{ for all } x, x' \in \mathcal{X}. \tag{2.3}$$

Proof. Suppose that f is short with respect to d_L, and let $k \in \{1, \ldots, K\}$. Let $x, x' \in \mathcal{X}$ differ only in their kth component. Then

$$|f(x) - f(x')| \leq \sum_{j=1}^{K} L_j d_j(x^j, x'^j) = L_k d_k(x^k, x'^k),$$

and so f is Lipschitz with Lipschitz constant L_k in its kth argument. Conversely, suppose that f is Lipschitz with Lipschitz constant L_k in its kth argument, and let $x, x' \in \mathcal{X}$. Then

$$\begin{aligned} |f(x) - f(x')| &\leq |f(x) - f(x'^1, x^2, \ldots, x^K)| \\ &\quad + |f(x'^1, x^2, \ldots, x^K) - f(x'^1, x'^2, x^3, \ldots, x^K)| \\ &\quad + \cdots + |f(x'^1, \ldots, x'^{K-1}, x^K) - f(x')| \\ &\leq L_1|x^1 - x'^1| + \cdots + L_K|x^K - x'^K| \\ &= d_L(x, x'), \end{aligned}$$

and so f is short with respect to d_L. □

$\mathcal{P}(\mathcal{X})$ denotes the set of all Borel probability measures on $\mathcal{X}$. The product of probability measures $\mu_k \in \mathcal{P}(\mathcal{X}_k)$ for $k \in \{1, \ldots, K\}$ will be denoted $\mu_1 \otimes \cdots \otimes \mu_K$ or $\bigotimes_{k=1}^K \mu_k$; the set of all such measures will be denoted $\bigotimes_{k=1}^K \mathcal{P}(\mathcal{X}_k)$. Recall that if $X = (X_1, \ldots, X_K)$ is an $\mathcal{X}$-valued random variable with law μ, then saying that the K components of X are independent is the same as saying that μ is a product measure $\bigotimes_{k=1}^K \mu_k$, where μ_k is the law of X_k on $\mathcal{X}_k$.

For $f\colon \mathcal{X} \to \mathbb{R}$, let $\mathcal{D}_k[f]$ be the *kth McDiarmid subdiameter* of f on $\mathcal{X}$:

$$\mathcal{D}_k[f] := \sup \left\{ |f(x) - f(x')| \,\middle|\, x, x' \in \mathcal{X}, x^j = x'^j \text{ for } j \neq k \right\}. \tag{2.4}$$

$\mathcal{D}_k[f]$ is a global sensitivity index that measures the sensitivity of f to changes in its kth argument. The *McDiarmid diameter* $\mathcal{D}[f]$ of f on $\mathcal{X}$ is defined by

$$\mathcal{D}[f] := \left(\sum_{k=1}^{K} \mathcal{D}_k[f]^2 \right)^{1/2}. \tag{2.5}$$

Each $\mathcal{D}_k[\cdot]$ (and, indeed, $\mathcal{D}[\cdot]$) is a semi-norm on the space of bounded real-valued functions on $\mathcal{X}$; any f that is constant in its kth argument has $\mathcal{D}_k[f] = 0$.

Clearly, if $f\colon \mathcal{X} \to \mathbb{R}$ is known to be d_L-short, then this information provides a (not necessarily sharp) upper bound on the McDiarmid diameter of f:

$$\mathcal{D}_k[f] \leq L_k \operatorname{diam}(\mathcal{X}_k, d_k) := L_k \sup_{x^k, x'^k \in \mathcal{X}_k} d_k(x^k, x'^k). \tag{2.6}$$

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

The McDiarmid diameter is useful because it places an upper bound on deviations of $f(X)$ from its mean value whenever X is an $\mathcal{X}$-valued random variable with independent components, as the following result shows:

Theorem 2.2 (McDiarmid's inequality [18–20])**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and, for $k \in \{1, \ldots, K\}$, let $X_k \colon \Omega \to \mathcal{X}_k$ be independent random variables. Suppose that $\mathbb{E}[|f(X)|]$ is finite. Then, for any $r > 0$,*

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq r] \leq \exp\left(-\frac{2r^2}{\mathcal{D}[f]^2}\right), \tag{2.7}$$

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \leq -r] \leq \exp\left(-\frac{2r^2}{\mathcal{D}[f]^2}\right). \tag{2.8}$$

The independence assumption in McDiarmid's inequality can be relaxed and replaced with some control on the martingale differences $\mathbb{E}[f(X)|\mathcal{F}_{i+1}] - \mathbb{E}[f(X)|\mathcal{F}_i]$ of $f(X)$ with respect to a suitable filtration $\mathcal{F}_\bullet$ of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Also, the mean and McDiarmid subdiameters can be used as inputs for sharper inequalities such as the *optimal McDiarmid inequality* of [28].

Given a measurable system of interest $G \colon \mathcal{X} \to \mathbb{R}$, let $\theta \in \mathbb{R}$ denote a possible value of G that is considered to be a *failure threshold*: the event $[G(X) \leq \theta]$ represents the failure of the system G, and the complementary event $[G(X) > \theta]$ represents the success of the system G. Under the assumption that the random inputs of G (*i.e.* the coordinate processes $X_1, \ldots, X_K$) are independent, McDiarmid's inequality implies that the probability of failure is bounded as follows:

$$\mathbb{P}[G(X) \leq \theta] \leq \exp\left(-\frac{2(\mathbb{E}[G(X)] - \theta)_+^2}{\mathcal{D}[G]^2}\right), \tag{2.9}$$

where, for $t \in \mathbb{R}$, $t_+ := \max\{0, t\}$. The inequality (2.9) can be rearranged in order to provide rigorous certification criteria for computational and physical systems of interest, subject to the determination of the mean system performance $\mathbb{E}[G(X)]$ and the McDiarmid diameter $\mathcal{D}[G]$; see *e.g.* [1,13,17]. Namely, if $p_* \in [0, 1]$ is the greatest probability of failure that can be accepted if the system is to be called safe, and it is known that $\mathbb{E}[G(X)] \geq m$ and $\mathcal{D}[G] \leq \widehat{D}$, then a sufficient condition for the safety of the system is the truth of the inequality

$$\frac{(m - \theta)_+}{\widehat{D}} \geq \sqrt{\log \sqrt{1/p_*}}. \tag{2.10}$$

2.2. UQ problem formulation

Suppose that the values of G are known only on some *observation set* $\mathcal{O} \subseteq \mathcal{X}$; that is, the restriction $G|_{\mathcal{O}}$ of G to $\mathcal{O}$ is known exactly. In applications, it is usually the case that $\mathcal{O}$ is a finite collection of points $\mathcal{O} = \{z_1, \ldots, z_N\} \subseteq \mathcal{X}$. Suppose also that constants $L_1, \ldots, L_K \geq 0$ are given such that G is known to be d_L-short. The main questions that this paper addresses are the following:

1. Section 3 shows how to use the observations $G|_{\mathcal{O}}$ and the Lipschitz constants $L = (L_1, \ldots, L_K)$ to provide an optimal (*i.e.* least) upper bound $\widehat{D}$ on the McDiarmid diameter $\mathcal{D}[G]$.
2. Section 4 shows how to use the data $G|_{\mathcal{O}}$, the constants L and the mean performance $\mathbb{E}[G(X)]$ to provide an optimal (*i.e.* least) upper bound $\widehat{P}$ on the probability of failure $\mathbb{P}[G(X) \leq \theta]$.
3. Section 5 considers the problem of determining which observations $z \in \mathcal{O}$ are relevant to the solutions of the problems in the previous two sections. Furthermore, one can consider the dual problem: if the data set $G|_{\mathcal{O}}$ could be extended, at what points of the input parameter space $\mathcal{X}$ should G be evaluated to gain maximally relevant information that will improve the bounds $\widehat{D}$ and $\widehat{P}$?

Remark 2.3 (Other UQ problems)**.** Although the exposition of this paper treats the *certification* problem of bounding $\mathbb{P}[G(X) \leq \theta]$, there are many other uncertainty quantification problems – *e.g.* verification, validation, and prediction [27] – to which this paper's methods are applicable. For example, G above may actually stand

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

for the difference between some physical system, H, and a model for that system, F; if the aim is to predict values of H using the simulation F with quantified error bounds, then this is tantamount to showing that $\mathbb{P}[\|H(X) - F(X)\| \geq \theta]$ is suitably small, where $\|\cdot\|$ is some "error norm" on (a subset of) parameter space $\mathcal{X}$. This certification-centric point of view is similar to that of [4], in which many reliability problems are placed in a unified framework, and that of [28].

In a different direction to the one taken in this paper, there are important questions of how to make optimal use of legacy data in the calibration and testing of models; for this problem, a particular difficulty is making best use of the data without *over-fitting* to the data [25]. The Bayesian perspective is a popular one in this area, and is receiving renewed attention in the context of Bayesian analysis for inverse problems on function spaces [35].

Remark 2.4 (Other regularity conditions)**.** In many practical applications, of course, the response function G is not known to be globally Lipschitz. In this paper attention is confined to the globally Lipschitz case as a representative example of a broad class of possible constraints. For example, it may be more appropriate to consider a Hölder-type constraint, which would correspond to an inequality of the form

$$|G(x) - G(x')| \leq d_L(x, x')^\alpha; \tag{2.11}$$

or a local Lipschitz constraint, which would correspond to an inequality of the form

$$|G(x) - G(x')| \leq \begin{cases} d_L(x, x'), & \text{if } d_L(x, x') < R, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.12}$$

The example of Subsection 7.3 will use just such a modified Lipschitz constraint, one suited to possibly discontinuous or multivalued functions. The minimum requirement on any proposed system of inequalities to constrain G is that the desired inequalities should hold whenever x and x' are elements of the observation set $\mathcal{O}$, and that the inequalities must constrain the values of G pointwise. So, for example, a constraint on the Sobolev $W^{k,p}(\mathbb{R}^d)$ norm of a function $G\colon \mathbb{R}^d \to \mathbb{R}$ would not be a suitable constraint if $kp < d$. It must be emphasized, though, that if no regularity assumptions are made, then no significant conclusions can be drawn from the data: regularity is essential if function values at finitely many isolated points are to be used to infer anything about function values elsewhere.

Remark 2.5 (Other types of observation)**.** In this paper the observations of G are pointwise evaluations of G at finitely many points of its domain. One could also consider more general observation operators, *e.g.* a continuous linear functional $\Lambda\colon W^{k,p}(\mathbb{R}^d) \to \mathbb{R}$, or a collection of such operators.

2.3. Extension of partially-defined functions

In what follows, in order to show that the upper bounds that are obtained are in fact the optimal upper bounds given the available information, it will be necessary to invoke the following extension theorem from metric space theory, which states that a real-valued Lipschitz function defined on any subset of a metric space can always be extended to the whole space without increasing the Lipschitz constant:

Theorem 2.6 (McShane's extension theorem [24])**.** *Let $(\mathcal{M}, \rho)$ be a metric space, let $E \subseteq \mathcal{M}$, and let $C \geq 0$. If $f\colon E \to \mathbb{R}$ satisfies*

$$|f(x) - f(x')| \leq C\rho(x, x') \text{ for all } x, x' \in E,$$

then there exists $\bar{f}\colon \mathcal{M} \to \mathbb{R}$ such that $\bar{f}|_E = f$ and

$$\left|\bar{f}(x) - \bar{f}(x')\right| \leq C\rho(x, x') \text{ for all } x, x' \in \mathcal{M}.$$

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

McShane's theorem also applies to the extension of Hölder continuous real-valued functions defined on a subset of a metric space; any continuous real-valued function with concave modulus of continuity can be extended to the whole space while preserving the modulus of continuity.

In the language of metric space theory, McShane's extension theorem says that the Euclidean line $(\mathbb{R}, |\cdot|)$ is an *injective metric space* [11]. The extension of *vector*-valued Lipschitz functions is a subtle topic: see *e.g.* the Kirszbraun–Valentine theorem [14, 39], which states that Lipschitz functions between Hilbert spaces can always be extended without increasing the Lipschitz constant, which is not generally true even for Lipschitz functions between finite-dimensional Banach spaces [8], p. 202. It is for this reason that this paper considers only scalar-valued performance measures G.

3. Optimal bounds on mcDiarmid diameters

For each $k \in \{1, \ldots, K\}$, an upper bound on the McDiarmid subdiameter $\mathcal{D}_k[G]$ can be obtained by an optimization problem. First observe that $\mathcal{D}_k[G]$ is the maximum value of the function

$$\operatorname{diff}_k G \colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_k \times \mathcal{X}_k \times \cdots \times \mathcal{X}_K \to \mathbb{R}$$

defined by

$$\begin{aligned}(\operatorname{diff}_k G)(x^1, \ldots, x^{k-1}, x^k, x'^k, x^{k+1}, \ldots, x^K) \\ := G(x^1, \ldots, x^{k-1}, x^k, x^{k+1}, \ldots, x^K) - G(x^1, \ldots, x^{k-1}, x'^k, x^{k+1}, \ldots, x^K).\end{aligned}$$

(Indeed, $\mathcal{D}_k[G]$ is also the negative of the minimum value of $\operatorname{diff}_k G$.) Therefore, an upper bound on $\mathcal{D}_k[G]$ consistent with the observations $G|_{\mathcal{O}}$ and the Lipschitz constant $L = (L_1, \ldots, L_K)$ is given by the solution of the following optimization problem in the $K + 3$ variables $x^1, \ldots, x^K, x'^k, y, y'$:

$$\left\{\begin{array}{ll} \text{maximize:} & |y - y'|; \\ \text{among:} & (x, y) \in \mathcal{X} \times \mathbb{R}, \\ & (x', y') \in \mathcal{X} \times \mathbb{R}; \\ \text{subject to:} & x^i = x'^i \text{ for all } i \in \{1, \ldots, K\} \setminus \{k\}: \\ & \quad |y - y'| \leq L_k d_k(x^k, x'^k); \\ & \text{for all } z \in \mathcal{O}: \\ & \quad |y - G(z)| \leq d_L(x, z), \\ & \quad |y' - G(z)| \leq d_L(x', z). \end{array}\right. \tag{3.1}$$

Note that (3.1) is not a linear programming problem: the feasible set for (x, y) and (x', y') is an intersection of double cones in $\mathcal{X} \times \mathbb{R}$, as illustrated in Figure 2. Note also that (3.1) is not a *cone program* in the sense of [7], Section 4.6.1; that term refers instead to the minimization of a linear objective function over a closed convex cone that contains no lines and has non-empty interior.

A point $(x, y) \in \mathcal{X} \times \mathbb{R}$ such that $|y - G(z)| \leq d_L(x, z)$ is said to be *feasible* with respect to the data point $(z, G(z))$; if this holds for all $z \in \mathcal{O}$, then (x, y) is said to be $G|_{\mathcal{O}}$*-feasible*.

Let $\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}}, d_L]$ (or simply $\widehat{D}_k$) denote the upper bound on $\mathcal{D}_k[G]$ that arises as the solution (extreme value) of the optimization problem (3.1). It is natural to ask whether or not $\widehat{D}_k$ is the *least* upper bound on $\mathcal{D}_k[G]$ given $G|_{\mathcal{O}}$ and L. In fact, this is the case, and the proof relies on McShane's extension theorem.

Theorem 3.1 (Optimality of $\widehat{D}_k$)**.** *Let*

$$\mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L) := \{g \colon \mathcal{X} \to \mathbb{R} \mid g \text{ is } d_L\text{-short and } g = G \text{ on } \mathcal{O}\} \tag{3.2}$$

denote the set of all functions on $\mathcal{X}$ that have Lipschitz constant L and interpolate the given values of G on $\mathcal{O}$. Then the maximum value $\widehat{D}_k$ of (3.1) is the optimal upper bound on $\mathcal{D}$ given $G|_{\mathcal{O}}$ and L in the sense that

$$\widehat{D}_k = \sup\{\mathcal{D}_k[g] \mid g \in \mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)\}. \tag{3.3}$$

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

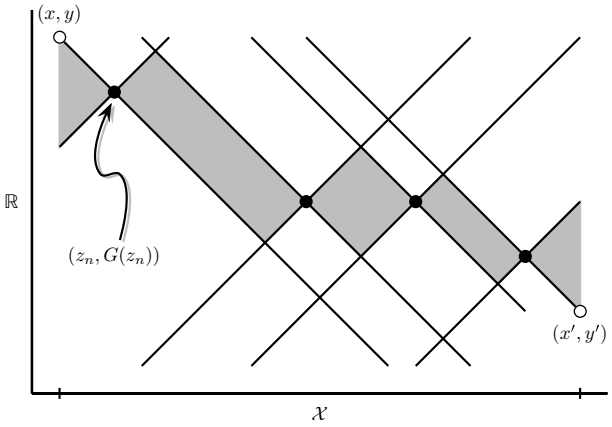


FIGURE 2. Shaded, the feasible set for (x, y) and (x', y') in $\mathcal{X} \times \mathbb{R}$ for the optimization problem (3.1) given the four observations represented by the four black dots. The white dots show the optimal values for (x, y) and (x', y'). Note that, as well as being constrained to lie in the shaded feasible set, (x, y) and (x', y') must also satisfy $|y - y'| \leq d_L(x, x') \equiv L_k d_k(x^k, x'^k)$.

Proof. Let S denote the supremum on the right-hand side of (3.3). Suppose that $\widehat{D}_k > S$. Then there exist points (x, y) and $(x', y') \in \mathcal{X} \times \mathbb{R}$ that satisfy the constraints in (3.1) and the inequality

$$S < |y - y'| \leq \widehat{D}_k.$$

Define $g\colon \mathcal{O} \cup \{x, x'\} \to \mathbb{R}$ by

$$\begin{aligned} g(z) &:= G(z) \text{ for each } z \in \mathcal{O}, \\ g(x) &:= y, \\ g(x') &:= y'. \end{aligned}$$

This g is d_L-short, and so McShane's extension theorem implies that g can be extended to some d_L-short $\bar{g}\colon \mathcal{X} \to \mathbb{R}$. Necessarily, $\bar{g}|_{\mathcal{O}} = g|_{\mathcal{O}} = G|_{\mathcal{O}}$. However, by construction, $\mathcal{D}_k[\bar{g}] \geq |y - y'| > S$, which contradicts the definition of S. Hence, by contradiction, $\widehat{D}_k \leq S$.

Now suppose that $\widehat{D}_k < S$. Then there exists some d_L-short $g\colon \mathcal{X} \to \mathbb{R}$ such that $g = G$ on $\mathcal{O}$ and $\mathcal{D}_k[g] > \widehat{D}_k$; hence, there exist points $x, x' \in \mathcal{X}$ that differ only in their kth component and such that

$$\widehat{D}_k < |g(x) - g(x')| \leq \mathcal{D}_k[g].$$

However, x and x' with $y := g(x)$ and $y' := g(x')$ satisfy the constraints in (3.1), and so $\widehat{D}_k \geq |g(x) - g(x')|$, which is a contradiction. Hence, $\widehat{D}_k \geq S$, which completes the proof. □

Remark 3.2. It is important to note that although $\widehat{D}_k$ is the optimal upper bound on $\mathcal{D}_k[G]$ given $G|_{\mathcal{O}}$ and L, and hence $\widehat{D} := \left(\widehat{D}_1^2 + \cdots + \widehat{D}_K^2\right)^{1/2} \geq \mathcal{D}[G]$, it is not generally true that $\widehat{D}$ is the optimal upper bound on $\mathcal{D}[G]$ given the same information ($G|_{\mathcal{O}}$ and L). The reason for this is that the (approximate) maximizers for, say, $\widehat{D}_1$ and $\widehat{D}_K$ may not be mutually consistent, *i.e.* d_L-short.

Note also that the upper bound

$$\mathbb{P}[G(X) \leq \theta] \leq \exp\left(-\frac{2(\mathbb{E}[G(X)] - \theta)_+^2}{\widehat{D}_1^2 + \cdots + \widehat{D}_K^2}\right)$$

is not the least upper bound on $\mathbb{P}[G(X) \leq \theta]$ given $\mathbb{E}[G(X)]$ and that $\mathcal{D}_k[G] \leq \widehat{D}_k$. The optimal such bound is given by the optimal McDiarmid inequality of [28], Section 4.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

3.1. Error bounds

In addition to the question of optimality, it is natural to ask whether or not solutions $\widehat{D}_k$ of (3.1) converge to the McDiarmid subdiameter $\mathcal{D}_k[G]$ as the number of observations increases to infinity. Unsurprisingly, the important quantity is not the number of observations, but rather the largest gap between them, as measured by the metric d_L. Define the *gap size* of the observation set $\mathcal{O}$ on $\mathcal{X}$ to be the (asymmetric) Hausdorff distance from $\mathcal{X}$ to $\mathcal{O}$ with respect to d_L, *i.e.*

$$\Gamma(\mathcal{X}, \mathcal{O}, d_L) := \sup_{x \in \mathcal{X}} d_L(x, \mathcal{O}) := \sup_{x \in \mathcal{X}} \inf_{z \in \mathcal{O}} d_L(x, z). \tag{3.4}$$

Theorem 3.3 (Error bound for $\widehat{D}_k$)**.** *For any $G \colon \mathcal{X} \to \mathbb{R}$ with finite McDiarmid subdiameter $\mathcal{D}_k[G]$ and any $\mathcal{O} \subseteq \mathcal{X}$,*

$$0 \leq \widehat{D}_k - \mathcal{D}_k[G] \leq 4\Gamma(\mathcal{X}, \mathcal{O}, d_L). \tag{3.5}$$

Proof. Theorem 3.1 shows that $\widehat{D}_k \geq \mathcal{D}_k[G]$, so it remains to show the effective "4Γ" part of the error estimate.

Let $\varepsilon > 0$ be arbitrary. Let (x, y) and $(x', y') \in \mathcal{X} \times \mathbb{R}$ satisfy the constraints in (3.1) and be ε-approximate maximizers for that problem, *i.e.*

$$\widehat{D}_k - \varepsilon \leq |y - y'| \leq \widehat{D}_k.$$

Then, even though the values $G(x)$ and $G(x')$ may be unknown since x and x' are not necessarily members of $\mathcal{O}$, the following estimate holds:

$$\begin{aligned} \widehat{D}_k &\leq |y - y'| + \varepsilon \\ &\leq |y - G(x)| + |G(x) - G(x')| + |G(x') - y'| + \varepsilon \\ &\leq |y - G(x)| + \mathcal{D}_k[G] + |G(x') - y'| + \varepsilon. \end{aligned}$$

By the definition of the gap size, there exists some $z \in \mathcal{O}$ such that $d_L(x, z) \leq \Gamma(\mathcal{X}, \mathcal{O}, d_L)$, and so

$$\begin{aligned} |y - G(x)| &\leq |y - G(z)| + |G(z) - G(x)| \\ &\leq d_L(x, z) + d_L(z, x) \\ &\leq 2\Gamma(\mathcal{X}, \mathcal{O}, d_L). \end{aligned}$$

Similarly, there exists $z' \in \mathcal{O}$ such that $d_L(x', z') \leq \Gamma(\mathcal{X}, \mathcal{O}, d_L)$, and so $|G(x') - y'| \leq 2g$. Therefore, $\widehat{D}_k \leq \varepsilon + 4\Gamma(\mathcal{X}, \mathcal{O}, d_L) + \mathcal{D}_k[G]$ and, since $\varepsilon > 0$ was arbitrary, the claim follows. □

3.2. Structure of the feasible set

The constrained optimization problem (3.1) entails exploration of the feasible set $\mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ of d_L-short extensions of the data $G|_{\mathcal{O}}$ to all of $\mathcal{X}$:

$$\mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L) := \{g \colon \mathcal{X} \to \mathbb{R} \mid g \text{ is } d_L\text{-short and } g = G \text{ on } \mathcal{O}\}.$$

Note that $\mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ not a linear space, but is a convex subset of the linear space of all real-valued functions on $\mathcal{X}$. Furthermore, by the Arzelà–Ascoli theorem, $\mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ is complete with respect to the uniform (supremum) norm, and is compact whenever $\mathcal{X}$ is compact. McShane's extension theorem (Thm. 2.6) is the assertion that, whenever $G|_{\mathcal{O}}$ has Lipschitz constant L on $\mathcal{O}$, $\mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ is non-empty. Theorem 3.1 states that the maximum value of $g \mapsto \mathcal{D}_k[g]$ over $g \in \mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ can be found by restricting attention to a finite-dimensional subset as described by the constraints in (3.1). Indeed, this search can be made even simpler than (3.1) suggests by considering structure of the problem for y and y' with fixed x and x'.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

For fixed $x \in \mathcal{X}$, define the least and greatest feasible values of $g(x)$ among $g \in \mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ by

$$Y^-(x, G|_{\mathcal{O}}, L) := \sup_{g \in \mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)} g(x) = \sup_{z \in \mathcal{O}} G(z) - d_L(x, z),$$
$$Y^+(x, G|_{\mathcal{O}}, L) := \inf_{g \in \mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)} g(x) = \inf_{z \in \mathcal{O}} G(z) + d_L(x, z),$$

Note that these quantities are easily calculated when $\mathcal{O}$ is a finite set. The pair $(y, y') \in \mathbb{R}^2$ is feasible (*i.e.* $y = g(x)$ and $y' = g(x')$ for some $x, x' \in \mathcal{X}$ that differ only in their kth component) if, and only if,

$$y \in \left[Y^-(x, G|_{\mathcal{O}}, L), Y^+(x, G|_{\mathcal{O}}, L)\right],$$
$$y' \in \left[Y^-(x', G|_{\mathcal{O}}, L), Y^+(x', G|_{\mathcal{O}}, L)\right], \text{ and}$$
$$|y - y'| \le d_L(x, x') = L_k d_k(x^k, x'^k).$$

So, for each (x, x'), the set of feasible (y, y') is a closed and convex polygon in $\mathbb{R}^2$. The maximum value of $|y - y'|$ over this polygon is $A(x, x')$, defined by

$$A(x, x') := \min \left\{ \begin{array}{c} d_L(x, x'), \\ Y^+(x, G|_{\mathcal{O}}, L) - Y^-(x', G|_{\mathcal{O}}, L), \\ Y^+(x', G|_{\mathcal{O}}, L) - Y^-(x, G|_{\mathcal{O}}, L) \end{array} \right\}.$$

The constrained optimization problem (3.1) is, therefore, equivalent to the following unconstrained (and, therefore, more easily solved) problem in $K + 1$ variables $x^1, \ldots, x^k, x'^k, \ldots x^K$:

$$\left\{ \begin{array}{ll} \text{maximize:} & A(x, x'); \\ \text{among:} & x \in \mathcal{X} \\ & x'^k \in \mathcal{X}_k \\ & x' := (x^1, \ldots, x^{k-1}, x'^k, x^{k+1}, \ldots, x^K). \end{array} \right. \tag{3.6}$$

3.3. Examples

As a simple example that can be solved explicitly, consider an affine function $G\colon \mathcal{X} := [0, 1]^K \to \mathbb{R}$:

$$G(x) = a_0 + \sum_{k=1}^{K} a_k x^k \tag{3.7}$$

for some constants $a_0, a_1, \ldots, a_K \in \mathbb{R}$. Suppose that the observation set $\mathcal{O}$ consists of a $N_1 \times \cdots \times N_K$ rectangular grid of equally-spaced points of $[0, 1]^K$, with observations at the corners of the cube. Given $L_k \ge |a_k|$, the gap size for this observation set is

$$\Gamma(\mathcal{X}, \mathcal{O}, d_L) = \sum_{k=1}^{K} \frac{L_k}{2(N_k - 1)}. \tag{3.8}$$

The exact McDiarmid subdiameters of G satisfy $\mathcal{D}_k[G] = |a_k|$. On the other hand, $\widehat{D}_k$, the least upper bound on $\mathcal{D}_k[G]$ given the observations $G|_{\mathcal{O}}$ and the Lipschitz constants $L_1, \ldots, L_K$ but *not* the information that G is affine,[7] is given by

$$\widehat{D}_k = |a_k| + \sum_{i=1}^{K} \frac{L_i - |a_i|}{N_i - 1}. \tag{3.9}$$

In this case, the error $\widehat{D}_k - \mathcal{D}_k[G]$ is approximately half the upper bound given by Theorem 3.3 if $L_k \gg |a_k|$, and vanishes if $L_k = |a_k|$.

[7] If G is known to be affine and its values are given at $K+1$ points in general position in $[0, 1]^K$, then G is determined everywhere.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

4. Optimal bounds on probabilities

In this section, in the spirit of [5, 28], the emphasis is on providing optimal bounds on the probability of failure $\mathbb{P}[G(X) \leq \theta]$ rather than bounds on the McDiarmid diameter $\mathcal{D}[G]$. Theorem 3.1 shows that the optimization problem (3.1) determines the optimal upper bound on each McDiarmid subdiameter $\mathcal{D}_k[G]$, and hence – given that $\mathbb{E}[G(X)] \geq m$ and *via* McDiarmid's inequality (2.9) – an upper bound on the probability of failure $\mathbb{P}[G(X) \leq \theta]$. However, this bound is not necessarily the sharpest one given the available information, namely that G is d_L-short, its inputs are independent, and that $G|_{\mathcal{O}}$ and $\mathbb{E}[G(X)]$ are as given. The optimal upper bound on the probability of failure given this information is denoted by $\widehat{P}[\mathcal{X}, G|_{\mathcal{O}}, L, m]$ (or simply $\widehat{P}$) and is given by

$$\widehat{P} := \sup_{(g,\mu) \in \mathcal{A}} \mu[g \leq \theta], \tag{4.1}$$

where

$$\mathcal{A} := \left\{ (g, \mu) \left| \begin{array}{c} g \colon \mathcal{X} \to \mathbb{R} \text{ is } d_L\text{-short}, \\ \mu = \mu_1 \otimes \cdots \otimes \mu_K \in \bigotimes_{k=1}^{K} \mathcal{P}(\mathcal{X}_k), \\ g = G \text{ on } \mathcal{O}, \text{ and } \mathbb{E}_\mu[g] \geq m \end{array} \right. \right\}, \tag{4.2}$$

i.e.

$$\mathcal{A} := \left\{ (g, \mu) \left| \begin{array}{l} \mu = \mu_1 \otimes \cdots \otimes \mu_K \in \bigotimes_{k=1}^{K} \mathcal{P}(\mathcal{X}_k), \\ g \in \mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L), \text{ and } \mathbb{E}_\mu[g] \geq m \end{array} \right. \right\}.$$

This infinite-dimensional optimization problem over coupled $g \in \mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ and $\mu \in \mathcal{P}(\mathcal{X})$ is more numerically tractable that it may seem. The next subsection shows that, for each g, the extreme values can be found by searching only among measures μ that have a particularly simple structure; furthermore, this simple structure simplifies the search over g as well.

Remark 4.1. Note that while the examples below have only two constraints on the measure μ, namely the product structure and that $\mathbb{E}_\mu[g] \geq m$, *any* combination of information on independence, non-independence, correlations and generalized moments can be used in the same way. For further discussion, see the general theory expounded in [28] and the remarks in Subsection 8.1.

4.1. Finite-dimensional reduction theorem

Given two points $x_0, x_1 \in \mathcal{X}$, let $\mathcal{C}(x_0, x_1)$ denote the discrete cube in $\mathcal{X}$ that has x_0 and x_1 as its "opposite corners":

$$\mathcal{C}(x_0, x_1) := \left\{ x_\varepsilon := \left(x^k_{\varepsilon_k} \right)_{k=1}^{K} \in \mathcal{X} \,\middle|\, \varepsilon \in \{0,1\}^K \right\}. \tag{4.3}$$

The elements of $\mathcal{C}(x_0, x_1)$ are indexed by the elements of the Hamming cube $\{0,1\}^K$: for $\varepsilon \in \{0,1\}^K$, $x_\varepsilon \in \mathcal{C}(x_0, x_1)$ is the point whose kth component is the same as the k^{th} component of x_0 if $\varepsilon_k = 0$, and the same as the kth component of x_1 if $\varepsilon_k = 1$.

Recall that a topological space $\mathcal{Z}$ is said to be a *Radon space* if it is separable and every Borel probability measure on $\mathcal{Z}$ is inner regular [31, 41]; that is, $\mathcal{Z}$ is a Radon space if it has a countable dense subset and, for every $\mu \in \mathcal{P}(\mathcal{Z})$ and every Borel-measurable set $B \subseteq \mathcal{Z}$,

$$\mu(B) = \sup\{\mu(K) \mid K \subseteq B \text{ and } K \text{ is compact}\}. \tag{4.4}$$

In particular, any continuous Hausdorff image of a separable and completely metrizable space (a *Suslin space*) is a Radon space. Compact subsets of Euclidean space $\mathbb{R}^n$ are Radon spaces, whereas a simple example of a non-inner-regular probability measure (and hence a non-Radon space) is $[0,1]$ with the topology of convergence from the right (see [33], Example 51) and uniform (Lebegsue) measure.

Under the mild technical assumption that each $(\mathcal{X}_k, d_k)$ is a Radon space, the reduction theorems of [28] imply that, for each d_L-short $g \colon \mathcal{X} \to \mathbb{R}$, the extreme value in (4.1) is obtained among product probability

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

measures μ such that each marginal distribution μ_k has support on at most two points of $\mathcal{X}_k$ – *i.e.* μ_k is a convex combination of at most two Dirac measures (point masses). That is, it is sufficient to search over probability measures of the form

$$\mu = \bigotimes_{k=1}^{K} \mu_k = \bigotimes_{k=1}^{K} \left(p_k \delta_{x_0^k} + (1 - p_k) \delta_{x_1^k} \right) \tag{4.5}$$

that are supported on $\mathcal{C}(x_0, x_1)$ for some $x_0, x_1 \in \mathcal{X}$; x_0, x_1 and p are parameters with respect to which we must optimize.

It is a simple matter of combinatorics to convert the product representation (4.5) into the sum representation

$$\mu = \sum_{\varepsilon \in \{0,1\}^K} \left(\prod_{k=1}^{K} (p_k)^{1-\varepsilon_k} (1 - p_k)^{\varepsilon_k} \right) \delta_{x_\varepsilon} \tag{4.6}$$

using the indexing scheme (4.3). If μ is any such measure and r is any real-valued measurable function defined on any superset of $\mathcal{C}(x_0, x_1)$, then $\mathbb{E}_\mu[r]$ exists and depends only upon the points x_ε, the values $y_\varepsilon := g(x_\varepsilon)$ and the weights p_k. The sum representation (4.6) makes the calculation of $\mathbb{E}_\mu[r]$ very easy:

$$\mathbb{E}_\mu[r] = \sum_{\varepsilon \in \{0,1\}^K} \left(\prod_{k=1}^{K} (p_k)^{1-\varepsilon_k} (1 - p_k)^{\varepsilon_k} \right) r(x_\varepsilon). \tag{4.7}$$

In particular, given $g \colon \mathcal{X} \to \mathbb{R}$, the mean and probability of failure for g are easily calculated using (4.7) with $r = g$ and $r = \mathbf{1}[g \leq \theta]$ respectively.

As the following theorem shows, a search over the finite-dimensional collection of feasible x_0, x_1, $\{y_\varepsilon \mid \varepsilon \in \{0,1\}^K\}$ and $p \in [0,1]^K$ has the same extreme values as the infinite-dimensional problem (4.1)–(4.2), where "feasible" means being d_L-short, extending $G|_{\mathcal{O}}$, and having the right mean value:

Theorem 4.2 (Optimality/finite-dimensional reduction)**.** *Suppose that $(\mathcal{X}_k, d_k)$ is a Radon space for each $k \in \{1, \ldots, K\}$. Let $\mathcal{A}$ be given by (4.2) and let*

$$\mathcal{A}_\Delta := \left\{ (g, \mu) \;\middle|\; \begin{array}{c} \text{for some } x_0, x_1 \in \mathcal{X}, \\ g \colon \mathcal{C}(x_0, x_1) \cup \mathcal{O} \to \mathbb{R} \text{ is } d_L\text{-short}, \\ \mu = \bigotimes_{k=1}^{K} \mu_k \in \mathcal{P}(\mathcal{C}(x_0, x_1)) \cap \bigotimes_{k=1}^{K} \mathcal{P}(\mathcal{X}_k), \\ g = G \text{ on } \mathcal{O}, \text{ and } \mathbb{E}_\mu[g] \geq m \end{array} \right\}. \tag{4.8}$$

Then

$$\dim(\mathcal{A}_\Delta) = 2 \sum_{k=1}^{K} \dim(\mathcal{X}_k) + 2^K + K, \tag{4.9}$$

$$\sup_{(g,\mu) \in \mathcal{A}} \mu[g \leq \theta] = \sup_{(g,\mu) \in \mathcal{A}_\Delta} \mu[g \leq \theta], \tag{4.10}$$

$$\inf_{(g,\mu) \in \mathcal{A}} \mu[g \leq \theta] = \inf_{(g,\mu) \in \mathcal{A}_\Delta} \mu[g \leq \theta]. \tag{4.11}$$

Proof. Assertion (4.9) follows from the fact that an element of $\mathcal{A}_\Delta$ is determined by a choice of $x_0 \in \mathcal{X}$, $x_1 \in \mathcal{X}$, $p \in [0,1]^K$ as in (4.5) or (4.6), and a choice of $g(x)$ for each of the 2^K points of $\mathcal{C}(x_0, x_1)$.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

To prove (4.10), let $S := \sup_{(g,\mu)\in\mathcal{A}} \mu[g \leq \theta]$. Then

$$\begin{aligned} S &= \sup \left\{ \mu[g \leq \theta] \left| \begin{array}{c} \text{for some } x_0, x_1 \in \mathcal{X}, \\ g\colon \mathcal{X} \to \mathbb{R} \text{ is } d_L\text{-short}, \\ \mu = \bigotimes_{k=1}^K \mu_k \in \mathcal{P}(\mathcal{C}(x_0,x_1)) \cap \bigotimes_{k=1}^K \mathcal{P}(\mathcal{X}_k), \\ g = G \text{ on } \mathcal{O}, \text{ and } \mathbb{E}_\mu[g] \geq m \end{array} \right. \right\} \\ &\leq \sup \left\{ \mu[g \leq \theta] \left| \begin{array}{c} \text{for some } x_0, x_1 \in \mathcal{X}, \\ g\colon \mathcal{C}(x_0,x_1) \cup \mathcal{O} \to \mathbb{R} \text{ is } d_L\text{-short}, \\ \mu = \bigotimes_{k=1}^K \mu_k \in \mathcal{P}(\mathcal{C}(x_0,x_1)) \cap \bigotimes_{k=1}^K \mathcal{P}(\mathcal{X}_k), \\ g = G \text{ on } \mathcal{O} \text{ and } \mathbb{E}_\mu[g] \geq m \end{array} \right. \right\} \\ &= \sup_{(g,\mu)\in\mathcal{A}_\Delta} \mu[g \leq \theta]. \end{aligned}$$

The first equality follows from the reduction theorem [28], Theorem 3.1 and Corollary 3.4 and the inequality follows from the fact that only the values of g on the discrete cube $\mathcal{C}(x_0, x_1)$ are germane to the probability of failure and the mean constraint; the final equality holds true by definition of the right-hand side.

To see that this inequality must, in fact, be an equality, suppose for a contradiction that $S < \sup_{(g,\mu)\in\mathcal{A}_\Delta} \mu[g \leq \theta]$. Then there exist some $x_0, x_1 \in \mathcal{X}$, $p \in [0,1]^K$ and a d_L-short $g\colon \mathcal{C}(x_0,x_1) \cup \mathcal{O} \to \mathbb{R}$ such that $g = G$ on $\mathcal{O}$, $\mathbb{E}_\mu[g] \geq m$ and $\mu[g \leq \theta] > S$. By McShane's extension theorem, there exists an extension of g to a d_L-short function $\bar{g}\colon \mathcal{X} \to \mathbb{R}$; necessarily, this extension has $\bar{g} = G$ on $\mathcal{O}$, $\mathbb{E}_\mu[\bar{g}] = \mathbb{E}_\mu[g] \geq m$ and $\mu[\bar{g} \leq \theta] = \mu[g \leq \theta] > S$, *i.e.* $(\mu, \bar{g}) \in \mathcal{A}$. Hence, $S < \mu[\bar{g} \leq \theta] \leq S$, which is a contradiction.

This establishes (4.10); the proof of (4.11) is similar, and is omitted. □

Theorem 4.2 shows that the infinite-dimensional optimization problem (4.1) is equivalent to (*i.e.* has the same extreme value as) the following finite-dimensional optimization problem, where now y_ε is written in place of $g(x_\varepsilon)$:

$$\left\{ \begin{array}{ll} \text{maximize:} & \displaystyle\sum_{\varepsilon\in\{0,1\}^K} \left(\prod_{k=1}^K (p_k)^{1-\varepsilon_k} (1-p_k)^{\varepsilon_k} \right) \mathbf{1}[y_\varepsilon \leq \theta]; \\ \text{among:} & x_0, x_1 \in \mathcal{X}, \\ & y\colon \{0,1\}^K \to \mathbb{R}, \\ & p \in [0,1]^K; \\ \text{subject to:} & \text{for all } \varepsilon, \varepsilon' \in \{0,1\}^K, \varepsilon \neq \varepsilon'\text{:} \\ & \quad |y_\varepsilon - y_{\varepsilon'}| \leq d_L(x_\varepsilon, x_{\varepsilon'}); \\ & \text{for all } \varepsilon \in \{0,1\}^K, z \in \mathcal{O}\text{:} \\ & \quad |y_\varepsilon - G(z)| \leq d_L(x_\varepsilon, z); \\ & \displaystyle\sum_{\varepsilon\in\{0,1\}^K} \left(\prod_{k=1}^K (p_k)^{1-\varepsilon_k} (1-p_k)^{\varepsilon_k} \right) y_\varepsilon \geq m. \end{array} \right. \tag{4.12}$$

See Figure 3 for a schematic illustration of the problem (4.12). The problem (4.12) has high dimension: assuming that $\dim(\mathcal{X}_k) = 1$ for each k, (4.12) is a problem in $3K + 2^K$ unknowns with $2^{K-1}(2^K - 1) + |\mathcal{O}|2^K + 1$ distinct constraints. However, as will be seen in Section 5, many of these constraints are redundant or non-binding. Furthermore, we have numerical evidence that in some cases not all of the 2^K support points of the measure μ need to be considered: see the remarks in Section 7 about "dimensional collapse" and Figure 10.

4.2. Error bounds

As with the McDiarmid diameters, it is natural to ask how much of an over-estimate $\widehat{P}$ is of the true probability of failure $\mathbb{P}[G(X) \leq \theta]$. Such an error estimate for the maximization problem (4.12) is naturally

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

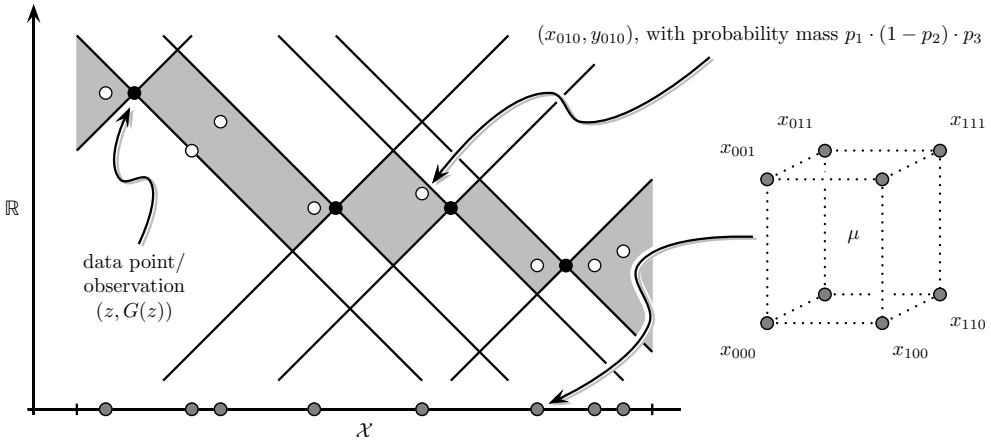


FIGURE 3. A schematic illustration of the variables in the optimization problem (4.12). The black dots show the fixed locations of the legacy observations $G|_{\mathcal{O}}$. The grey dots show the movable locations of the 2^K support points x_ε, $\varepsilon \in \{0,1\}^K$, of the discrete product measure μ on $\mathcal{X}$. The white dots show some feasible values $(x_\varepsilon, y_\varepsilon)$. The marginal distribution μ_k on $\mathcal{X}_k$ assigns mass p_k to x_0^k and mass $1-p_k$ to x_1^k; the mass of x_ε is determined by (4.6).

provided by solving the corresponding *minimization* problem. That is, the double inequality

$$\inf_{(g,\mu)\in\mathcal{A}} \mu[g \le \theta] \le \mathbb{P}[G(X) \le \theta] \le \sup_{(g,\mu)\in\mathcal{A}} \mu[g \le \theta]$$

is *ipso facto* the sharpest such inequality on the probability of failure given the available information encoded in $\mathcal{A}$ (*i.e.* $G|_{\mathcal{O}}$, L and $\mathbb{E}[G(X)] \ge m$). It is not possible, on the basis of this information, to rule out the possibility that

$$\inf_{(g,\mu)\in\mathcal{A}} \mu[g \le \theta] = \mathbb{P}[G(X) \le \theta].$$

Hence, the upper bound on $\widehat{P} - -\mathbb{P}[G(X) \le \theta]$ is simply

$$\widehat{P} - \mathbb{P}[G(X) \le \theta] \le \sup_{(g,\mu)\in\mathcal{A}} \mu[g \le \theta] - \inf_{(g,\mu)\in\mathcal{A}} \mu[g \le \theta], \tag{4.13}$$

and this inequality is sharp, given the information encoded in $\mathcal{A}$.

4.3. Prototypical example

The next example, Example 4.3, in which (4.12) is solved explicitly for one observation of a function on the unit interval, illustrates two very important points: the least upper bound on the probability of failure, $\widehat{P}[\mathcal{X}, G|_{\mathcal{O}}, L, m]$, can depend discontinuously and non-monotonically on the observed data $G|_{\mathcal{O}}$. It may be useful to first observe that

$$\begin{aligned}
&\sup\left\{ \mu((-\infty,0]) \,\middle|\, \begin{array}{c} \mu \in \mathcal{P}(\mathbb{R}),\ \mathbb{E}_{Y\sim\mu}[Y] \ge m, \\ \mu \text{ supported on an interval of length} \le R \end{array} \right\} \\
&\quad = \sup\left\{ \mu((-\infty,0]) \,\middle|\, \begin{array}{c} \mu = p\delta_{y_0} + (1-p)\delta_{y_1} \in \mathcal{P}(\mathbb{R}), \\ y_0, y_1 \in \mathbb{R},\ p \in [0,1], \\ py_0 + (1-p)y_1 \ge m, \\ |y_0 - y_1| \le R \end{array} \right\} \\
&\quad = \left(1 - \frac{m_+}{R}\right)_+,
\end{aligned}$$

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

and that the maximizer satisfies $y_0 = 0$, $y_1 = R$. The heuristic to bear in mind is that the event $[y_0 = 0]$ can be assigned high probability if the value y_1 can be chosen to be sufficiently greater than the prescribed mean m.

Example 4.3. Suppose that a function $G\colon [0,1] \to \mathbb{R}$ with Lipschitz constant $L > 0$ is observed at a single point, *i.e.* $\mathcal{O} = \{z\}$ for some $z \in [0,1]$. By symmetry, it is enough to consider the case that $z \in [0, \frac{1}{2}]$; for simplicity, suppose that $G(z) > 0$; also, it is no loss of generality to set the failure threshold to be $\theta := 0$.

Suppose it is known that $\mathbb{E}[G(X)] \geq m \in \mathbb{R}$; necessarily, it must hold that $|G(z) - m| \leq L|1 - z|$, otherwise the data and the mean and Lipschitz constraints are mutually contradictory. The least upper bound $\widehat{P}$ on $\mathbb{P}[G(X) \leq 0]$ given the observation $(z, G(z))$, that $\mathbb{E}[G(X)] \geq m$, and the Lipschitz constant L, is given in five cases:

$$\widehat{P} = \begin{cases} \left(1 - \frac{m_+}{L - (Lz - G(z))}\right)_+, & \text{if } G(z) \leq Lz, \\ \left(1 - \frac{m_+}{L - (Lz + G(z))}\right)_+, & \text{if } Lz < G(z) \leq L|\frac{1}{2} - z|, \\ \left(1 - \frac{2m_+}{L + (G(z) - Lz)}\right)_+, & \text{if } L|\frac{1}{2} - z| < G(z) \leq L|1 - 3z|, \\ \left(1 - \frac{m_+}{Lz + G(z)}\right)_+, & \text{if } G(z) > L\max\{z, 1 - 3z\}, \\ 0, & \text{if } G(z) > L|1 - z|. \end{cases} \tag{4.14}$$

The five cases are shown in Figure 4; surface and contour plots of $\widehat{P}$ as a function of the observed data $(z, G(z))$ were given in the introduction in Figure 1. Note well that $\widehat{P}$ is neither continuous nor monotone with respect to $(z, G(z))$: the boundaries among the five cases define "critical lines" in data space, across which there are stark changes in the conclusions that may be inferred from the observed data. Note also that the maximizers for (4.12) may be non-unique: *e.g.* in Figure 4a, which corresponds to the first case in (4.14), the maximum is attained by any (x_0, y_0), (x_1, y_1) and p satisfying

$$\begin{aligned} x_0 &\in [0, z - G(z)/L], & y_0 &= 0, \\ x_1 &= 1, & y_1 &= L - Lz + G(z), \\ p &= \left(1 - \tfrac{m_+}{|y_1 - y_0|}\right)_+. \end{aligned}$$

There is a similar lack of uniqueness in Figure 4(d). On the other hand, the maximizers in Figures 4(b) and (c) are unique.

Note that, for any single observation $(z, G(z))$, the least upper bound on the McDiarmid diameter, $\widehat{D}[G]$, is simply L, and that the bound (4.14) is in each case an improvement on both McDiarmid's inequality

$$\mathbb{P}[G(X) \leq 0] \leq \exp\left(-2m_+^2 \Big/ \widehat{D}[G]^2\right) = \exp\left(-2m_+^2 / L^2\right)$$

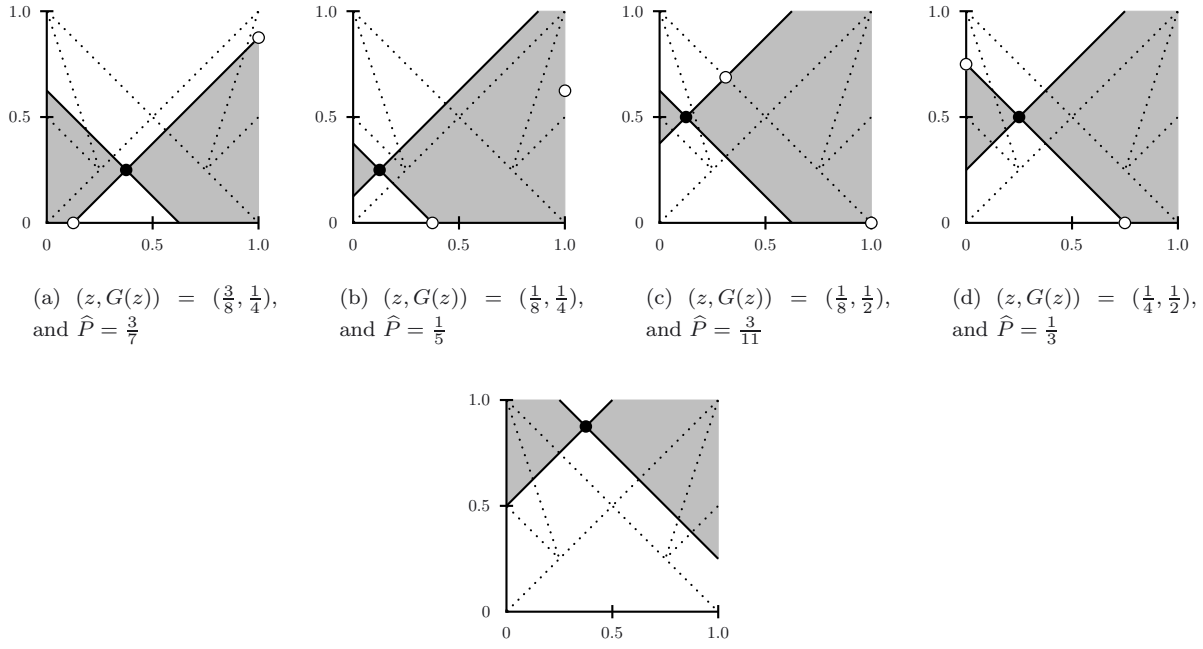
and on the $K = 1$ optimal McDiarmid inequality [28], Section 4

$$\mathbb{P}[G(X) \leq 0] \leq \left(1 - \frac{m_+}{\widehat{D}[G]}\right)_+ = \left(1 - \frac{m_+}{L}\right)_+.$$

5. Redundant and non-binding observations

In many applications, the aim is not to understand the behaviour of G on the whole of the input parameter space $\mathcal{X}$, but only on some subset $V \subseteq \mathcal{X}$, or on the elements of a partition $\mathcal{X} = \biguplus_{j=1}^{J} V_j$ of $\mathcal{X}$ [36]. The observation set $\mathcal{O}$ may lie entirely within V, or only partially lie within V, or lie entirely outside V. Heuristically, it seems reasonable that the points of $\mathcal{O}$ that are "nearest" to V should be the most important ones, but it is not immediately obvious what "nearest" means.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press



(a) $(z, G(z)) = (\frac{3}{8}, \frac{1}{4})$, and $\widehat{P} = \frac{3}{7}$

(b) $(z, G(z)) = (\frac{1}{8}, \frac{1}{4})$, and $\widehat{P} = \frac{1}{5}$

(c) $(z, G(z)) = (\frac{1}{8}, \frac{1}{2})$, and $\widehat{P} = \frac{3}{11}$

(d) $(z, G(z)) = (\frac{1}{4}, \frac{1}{2})$, and $\widehat{P} = \frac{1}{3}$

(e) $(z, G(z)) = (\frac{3}{8}, \frac{7}{8})$, and $\widehat{P} = 0$

FIGURE 4. Illustration of the maximizers in Example 4.3 with $L = 1$. The dotted lines show the boundaries of the various cases in $(z, G(z))$ data space. The black dot shows the data point, and the white dots the positions of (x_0, y_0) and (x_1, y_1) that maximize the probability of failure; in each case, $\widehat{P} = \left(1 - \frac{m_+}{|y_1 - y_0|}\right)_+$. Note that failure is impossible in case (e).

However, the formulation of the UQ objectives as optimization problems provides a natural notion of information content. Instead of calculating, for example, information-theoretic entropies, we simply make use of notions of relevancy that are natural to the optimization-theoretic context: the relevant data points are the precisely the ones that correspond to non-trivial constraints, or rather, determine the extreme value of the optimization problem.

Even if the aim is to understand the behaviour of G on all of $\mathcal{X}$ rather than a subset $V \subseteq \mathcal{X}$, the problems (3.1) and (4.1)–(4.2) are highly constrained, and their solution is much simplified by elimination of redundant constraints/observations. To that end, this section discusses two notions of redundancy/relevancy for data points and other constraints [10]:

- *redundant* constraints do not change the feasible set in the problems (3.1) and (4.1)–(4.2);
- *non-binding* constraints may change the feasible set in the problems (3.1) and (4.1)–(4.2), and may even change the extremizer, but do not change the extreme value.

Clearly, every redundant constraint is non-binding, but not *vice versa*. With this point of view, the problem of finding "nearest data points" becomes one of finding minimal data sets $\mathcal{O}$ that are *redundancy-free*.

5.1. Redundant Lipschitz constraints

In problem (4.12), many of the 2^{2K} Lipschitz constraints of the form

$$|y_\varepsilon - y_{\varepsilon'}| \leq d_L(x_\varepsilon, x_{\varepsilon'}) \tag{5.1}$$

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

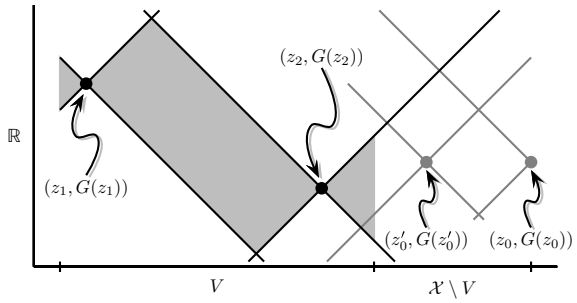


FIGURE 5. The observation at $z_0 \in \mathcal{X} \setminus V$ is redundant on V with respect to $\mathcal{O} := \{z_1, z_2\}$, since its feasible cone contains the set of all $G|_{\mathcal{O}}$-feasible points in $V \times \mathbb{R}$. Contrarily, the observation at z_0' is relevant on V with respect to $\mathcal{O}$.

are redundant constraints. First, (5.1) is obviously satisfied when $\varepsilon = \varepsilon'$, so there are at most $2^{2K} - 2^K = 2^K(2^K - 1)$ non-redundant constraints of the form (5.1). Secondly, (5.1) is symmetric under interchange of ε and ε', and so there are at most $2^K(2^K - 1)/2 = 2^{K-1}(2^K - 1)$ non-redundant constraints of the form (5.1); it suffices to endow $\{0,1\}^K$ with some total order $\preceq$ (*e.g.* lexicographic ordering) and only verify (5.1) for $\varepsilon \prec \varepsilon'$.

A third source of redundancy is neatly encapsulated in Lemma 2.1: in order to verify that (5.1) holds for all $\varepsilon, \varepsilon' \in \{0,1\}^K$ (*i.e.* to show that $g|_{\mathcal{C}(x_0,x_1)}$ is d_L-short, where $y_\varepsilon = g(x_\varepsilon)$), it is necessary and sufficient to check that (5.1) holds when ε and ε' differ in precisely one entry. Geometrically, this corresponds to checking (5.1) not between arbitrary vertices of the cube $\mathcal{C}(x_0, x_1)$ but only along edges joining adjacent vertices. There are $K2^K$ such edges, and so symmetry considerations yield the following result:

Theorem 5.1 (Relevant Lipschitz constraints)**.** *A constraint of the form* (5.1) *in problem* (4.12) *is relevant only if $\varepsilon \prec \varepsilon'$ and $\varepsilon_k \neq \varepsilon_k'$ for precisely one $k \in \{1, \ldots, K\}$; otherwise, it is redundant. Hence, there are at most $K2^{K-1}$ non-redundant constraints of the form* (5.1)*.*

5.2. Redundant data points

Given $V \subseteq \mathcal{X}$ and $\mathcal{O} \subseteq \mathcal{X}$ such that $G|_{\mathcal{O}}$ is known, an observation $(z_0, G(z_0)) \in \mathcal{X} \times \mathbb{R}$ is said to be *redundant on V with respect to $\mathcal{O}$* if, for all $(x, y) \in V \times \mathbb{R}$,

$$\left.\begin{array}{c}\text{for all } z \in \mathcal{O}, \\ |y - G(z)| \leq d_L(x, z)\end{array}\right\} \implies |y - G(z_0)| \leq d_L(x, z_0), \tag{5.2}$$

and say that it is *relevant* otherwise. That is, a redundant observation is one for which the induced constraint in (3.1) (or (4.1)–(4.2) or (4.12)) is automatically satisfied whenever the constraints induced by $\mathcal{O}$ are satisfied; put another way, the set of $G|_{\mathcal{O}}$-feasible points in $V \times \mathbb{R}$ is contained in the cone of $G|_{\{z_0\}}$-feasible points in $V \times \mathbb{R}$. See Figure 5 for an illustration.

Proposition 5.2 shows that every (non-isolated) data point $z \in \mathcal{O} \cap V$ is relevant; only data points $z \in \mathcal{O} \setminus V$ may be redundant. Furthermore, Theorem 5.3 shows that every point $z \in \mathcal{O} \setminus V$ that is sufficiently far away from V is redundant.

Proposition 5.2 (Relevant data points)**.** *Let $V \subseteq \mathcal{X}$, $\mathcal{O} \subseteq \mathcal{X}$, and $G|_{\mathcal{O}}$ be given, and suppose that d_L is a metric. If $z_0 \in \mathcal{O} \cap V$ and z_0 is an isolated point of $\mathcal{O}$, then z_0 is relevant on V with respect to $\mathcal{O} \setminus \{z_0\}$.*

Proof. Let z' be the closest point of $\mathcal{O} \setminus \{z_0\}$ to z_0 (if there is more than one such point, then choose any such point). Then any value

$$y \in [G(z') - d_L(z_0, z'), G(z') + d_L(z_0, z')]$$

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

is feasible with respect to $\mathcal{O} \setminus \{z_0\}$. Since z_0 is an isolated point of $\mathcal{O}$ and d_L is a metric, this interval has non-zero length. However, the such y that is feasible with respect to $\mathcal{O}$ is $G(z_0)$. Hence, z_0 supplies a non-trivial constraint and is relevant on V with respect to $\mathcal{O} \setminus \{z_0\}$. (Note that if V is, say, a subset of $\mathbb{R}^K$ with non-empty interior, then this argument can be applied on a neighbourhood of z_0, thereby demonstrating relevancy of z_0 to a non-trivial set.) □

The next result gives a sufficient condition for observations $z_0 \in \mathcal{O} \setminus V$ to be redundant. Say that $y \in \mathcal{X}$ is *between* $x \in \mathcal{X}$ *and* $z \in \mathcal{X}$ if

$$d_L(x, z) = d_L(x, y) + d_L(y, z), \tag{5.3}$$

and that y is *between* $V \subseteq \mathcal{X}$ *and* $W \subseteq \mathcal{X}$ if (5.3) holds for every $x \in V$ and $z \in W$. Note well that in the prototypical case that d_L is the ℓ^1 Manhattan metric on $\mathbb{R}^K$, the set of points between x and z is not the Euclidean line segment joining them, but the closed convex hull $\overline{\text{co}}(\mathcal{C}(x, z))$, *i.e.* the compact cuboid with faces perpendicular to the coordinate axes and x and z as its opposite corners.

Theorem 5.3 (Redundant data points)**.** *Let $V \subseteq \mathcal{X}$, $\mathcal{O} \subseteq \mathcal{X}$, L and $G|_{\mathcal{O}}$ be given. Fix $z_0 \in \mathcal{O} \setminus V$. Suppose that $p \in \mathcal{X}$ is between V and z_0, and that there exist $z', z'' \in \mathcal{O} \cap V$ satisfying*

$$G(z') + d_L(z', p) \leq G(z_0) + d_L(z_0, p), \tag{5.4}$$
$$G(z'') - d_L(z'', p) \geq G(z_0) - d_L(z_0, p). \tag{5.5}$$

Then z_0 is redundant on V with respect to $\mathcal{O} \cap V$.

Proof. Let $(x, y) \in V \times \mathbb{R}$ be a feasible point with respect to $G|_{\mathcal{O} \cap V}$, *i.e.*

$$|y - G(z)| \leq d_L(x, z) \text{ for each } z \in \mathcal{O} \cap V,$$

and suppose for a contradiction that $|y - G(z_0)| > d_L(x, z_0) > 0$. If $y > G(z_0)$, then the assumption *ad absurdum* implies that $y > G(z_0) + d_L(x, z_0)$. Hence,

$$\begin{aligned}
|y - G(z')| &\geq y - G(z') \\
&> G(z_0) + d_L(x, z_0) - G(z') \\
&\geq d_L(z', p) - d_L(z_0, p) + d_L(x, z_0) && \text{by (5.4)} \\
&= d_L(z', p) + d_L(x, p) && \text{since } p \text{ is between } V \text{ and } z_0 \\
&\geq d_L(x, z') && \text{by the triangle inequality,}
\end{aligned}$$

which contradicts the feasibility of (x, y) with respect to $G|_{\mathcal{O} \cap V}$. Similarly, if $y < G(z_0)$, then (5.5) implies that

$$|y - G(z'')| > d_L(x, z''),$$

which is again a contradiction. This completes the proof. □

If the closure $\overline{V}$ of V is a compact rectangular box $\prod_{k=1}^{K} [\alpha^k, \beta^k] \subseteq \mathbb{R}^K$, then, for each $z_0 \in \mathcal{O} \setminus V$, there is a natural choice for the point p with respect to which conditions (5.4) and (5.5) can be checked: the unique point $P_{z_0, V} \in \overline{V}$ that is closest to z_0, where

$$P_{x,V}^k := \begin{cases} \alpha^k, & \text{if } x^k < \alpha^k, \\ x^k, & \text{if } \alpha^k \leq x^k \leq \beta^k, \\ \beta^k, & \text{if } x^k > \beta^k. \end{cases} \tag{5.6}$$

It is easy to see that $P_{z_0, V}$ is between V and z_0. This choice of p validates the heuristic that observations far away from V ought to be redundant, since (5.4) and (5.5) are certain to hold when V is bounded and $d_L(z_0, V)$ is large enough.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

5.3. Non-binding data points

A more interesting notion of the information content of the data points $(z, G(z))$ is not relevancy but *bindingness*. Whereas redundancy concerns the set of feasible points for an optimization problem, a *non-binding* constraint (or data point) is one that perhaps changes the feasible set but does not change the extreme value of the problem.

Given $\mathcal{O} \subseteq \mathcal{X}$ such that $G|_{\mathcal{O}}$ is known, an observation $(z_0, G(z_0)) \in \mathcal{X} \times \mathbb{R}$ is said to be

- *non-binding for $\widehat{D}_k$ with respect to $\mathcal{O}$* if

$$\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}\cup\{z_0\}}, L] = \widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}}, d_L];$$

- *non-binding for $\widehat{P}$ with respect to $\mathcal{O}$* if

$$\widehat{P}[\mathcal{X}, G|_{\mathcal{O}\cup\{z_0\}}, L, m] = \widehat{P}[\mathcal{X}, G|_{\mathcal{O}}, L, m].$$

Otherwise, an observation is said to be *binding*. Note well that the inclusion of a binding observation *strictly* changes the extreme value of the optimization problems, not just the set of extremizers.

Clearly, if including an observation at z_0 does not change the feasible set for, say, the $\widehat{D}_k$ problem (3.1), then including it does not change the extreme value of (3.1); that is, every redundant data point is non-binding, and every binding data point is relevant. The converse implications, however, are false: in general, there are data points that do change the feasible set for the optimization problems for $\widehat{D}_k$ and $\widehat{P}$, but do not change the extreme values. See Figure 6 for some illustrations based upon the earlier Example 4.3. See also Figure 7, which illustrates the set of all second data points $(z_2, G(z_2)) \in [0,1] \times \mathbb{R}$ that are redundant with respect to the first data point from Example 4.3.

A sufficient (but not necessary) condition for the extreme value of an optimization problem to be unchanged upon the introduction of a new constraint is that the extremizer of the original problem is feasible with respect to the new constraint. This, a sufficient condition for a data point to be non-binding is provided by the following result:

Proposition 5.4 (Non-binding data points)**.** *Let $\mathcal{O} \subseteq \mathcal{X}$, $z_0 \in \mathcal{X}$, L and $G|_{\mathcal{O}\cup\{z_0\}}$ be given.*

1. *Let $(\bar{x}, \bar{y}, \bar{x}', \bar{y}')$ be a maximizer for (3.1) with observations $\mathcal{O}$. If*

$$|\bar{y} - G(z_0)| \leq d_L(\bar{x}, z_0) \text{ and } |\bar{y}' - G(z_0)| \leq d_L(\bar{x}', z_0), \tag{5.7}$$

   then z_0 is non-binding and $\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}\cup\{z_0\}}, L] = \widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}}, d_L]$.
2. *Let $(\bar{x}_0, \bar{x}_1, \bar{y}, \bar{p})$ be a maximizer for (4.12) with observations $\mathcal{O}$. If*

$$|\bar{y}_\varepsilon - G(z_0)| \leq d_L(\bar{x}_\varepsilon, z_0) \text{ for all } \varepsilon \in \{0,1\}^K, \tag{5.8}$$

   then z_0 is non-binding and $\widehat{P}[\mathcal{X}, G|_{\mathcal{O}\cup\{z_0\}}, L, m] = \widehat{P}[\mathcal{X}, G|_{\mathcal{O}}, L, m]$.

Proof. Since $\mathcal{O} \subseteq \mathcal{O}\cup\{z_0\}$, every (x, y, x', y') that is feasible for (3.1) with observations $\mathcal{O}\cup\{z_0\}$ is also feasible for (3.1) with observations $\mathcal{O}$. Hence

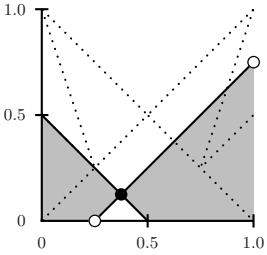
$$\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}}, d_L] \geq \widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}\cup\{z_0\}}, L].$$

Now let $(\bar{x}, \bar{y}, \bar{x}', \bar{y}')$ be a maximizer for (3.1) with observations $\mathcal{O}$ and suppose that (5.7) holds; then $(\bar{x}, \bar{y}, \bar{x}', \bar{y}')$ satisfies the criteria to be a feasible point for (3.1) with observations $\mathcal{O}\cup\{z_0\}$, and has the same objective function value $|\bar{y} - \bar{y}'|$. Hence,

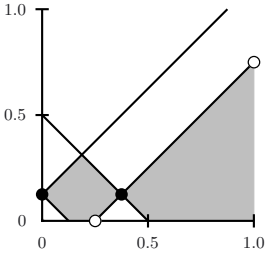
$$\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}}, d_L] \leq \widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}\cup\{z_0\}}, L],$$

and the claim for $\widehat{D}_k$ follows. The proof of the claim for $\widehat{P}$ is analogous. □

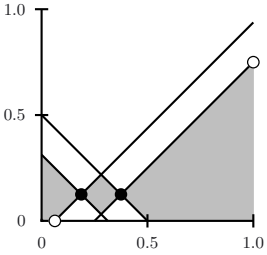
https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press



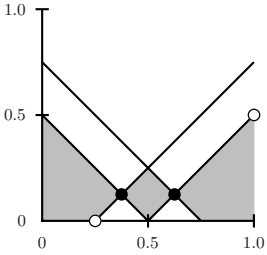
(a) (Non-unique) maximizer for the probability of failure with one data point at $(\frac{3}{8}, \frac{1}{8})$.



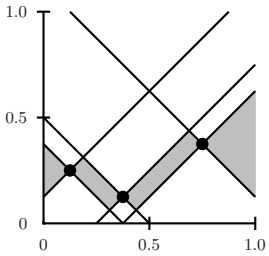
(b) A non-binding new data point; the maximizer does not change. *Cf.* Figure 7(a).



(c) A non-binding new data point; the maximizer changes but the maximum value does not.



(d) A binding new data point: the maximizer and maximum value both change.



(e) Two binding new data points that together render failure impossible.

FIGURE 6. Additional binding and non-binding data points for the one-dimensional Example 4.3. As before, black dots show data points and white dots the locations of maximizing (x_0, y_0) and (x_1, y_1), with $\widehat{P} = \left(1 - \frac{m_+}{|y_1 - y_0|}\right)_+$.

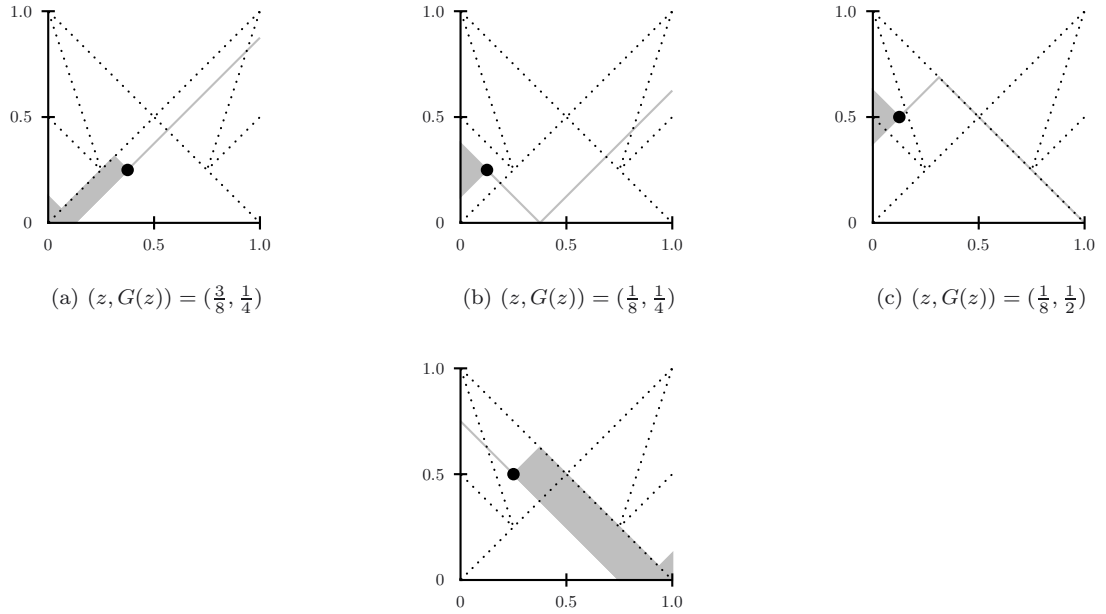
Note well that the converse of Proposition 5.4 is false in general: the introduction of a new data point may render some of the previous (non-unique) maximizers infeasible but still fail to change the maximum value of the problem.

Nevertheless, the simple algebraic conditions of Proposition 5.4 suggest a practical method for calculating $\widehat{D}_k$ or $\widehat{P}$ if the data set $\mathcal{O} = \{z_1, \ldots, z_N\}$ is a large finite set that is believed to contain many redundant points. The idea is to introduce the data points one at a time and only solve (3.1) (for $\widehat{D}_k$) or (4.12) (for $\widehat{P}$) when strictly necessary. In the following algorithm, $\mathcal{O}_i \subseteq \mathcal{O}$ will denote the data points (constraints) that are enforced at iteration i, while $\widetilde{\mathcal{O}}_i \subseteq \mathcal{O}$ will denote those that are potentially binding and will be checked for feasibility at iteration i. Note well that, in general, $\mathcal{O}_i \cup \widetilde{\mathcal{O}}_i \subsetneq \mathcal{O}$.

Algorithm 5.5. Initialize with $\mathcal{O}_0 = \varnothing$ and $\widetilde{\mathcal{O}}_0 = \mathcal{O}$. Then, for $i = 1, 2, \ldots,$

(1) For each $z \in \widetilde{\mathcal{O}}_{i-1}$, calculate $\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}_{i-1} \cup \{z\}}, L]$.
(2) Let $\mathcal{M} \subseteq \widetilde{\mathcal{O}}_{i-1}$ be the set of maximizers of $z \mapsto \widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}_{i-1} \cup \{z\}}, L]$ among $z \in \widetilde{\mathcal{O}}_{i-1}$.
(3) Set $\mathcal{O}_i := \mathcal{O}_{i-1} \cup \mathcal{M}$ and calculate $\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}_i}, L]$.
(4) Let $\widetilde{\mathcal{O}}_i$ consist of those $z \in \mathcal{O} \setminus \mathcal{O}_i$ such that the extremizer for $\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}_i}, L]$ is infeasible with respect to $(z, G(z))$ (*i.e.* fails (5.7)), and hence is possibly binding.
(5) Terminate if $\widetilde{\mathcal{O}}_i = \varnothing$.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press



(a) $(z, G(z)) = (\frac{3}{8}, \frac{1}{4})$ (b) $(z, G(z)) = (\frac{1}{8}, \frac{1}{4})$ (c) $(z, G(z)) = (\frac{1}{8}, \frac{1}{2})$

(d) $(z, G(z)) = (\frac{1}{4}, \frac{1}{2})$

FIGURE 7. In grey, those locations for the second data point in the one-dimensional Example 4.3 that are non-binding with respect to the first point (the black dot); *cf.* (a)–(d) of Figure 4.

The algorithm for $\widehat{P}$ is analogous, with (5.8) in place of (5.7).

In the numerical examples that have been considered so far, it has been observed that relatively few elements of $\mathcal{O}$ determine $\widehat{D}_k$ or $\widehat{P}$, even though, in principle, every element of $\mathcal{O}$ could supply a binding constraint. This situation is somewhat analogous to the simplex algorithm in linear programming: in the theoretical worst case, the simplex method can take exponential time [15], but it "usually" requires polynomial time in practice. We will reserve detailed numerical analysis of this algorithm for a future work.

6. FURTHER REMARKS

6.1. Feasible lipschitz constants

Given $\mathcal{O} \subseteq \mathcal{X}$ and the associated observations $G|_{\mathcal{O}}$, let $\mathrm{Lip}(G|_{\mathcal{O}})$ denote the set of Lipschitz constants for G that are consistent with the observations $G|_{\mathcal{O}}$, *i.e.*

$$\mathrm{Lip}(G|_{\mathcal{O}}) := \left\{ L \in \mathbb{R}^K \,\middle|\, \begin{array}{c} \text{for all } z, z' \in \mathcal{O}, \\ |G(z) - G(z')| \leq d_L(z, z') \end{array} \right\}. \tag{6.1}$$

It is easy to check that, for any given $\mathcal{O} \subseteq \mathcal{X}$ and $G|_{\mathcal{O}}$, $\mathrm{Lip}(G|_{\mathcal{O}})$ is a convex subset of $\mathbb{R}^K$. This remains the case if additional inequality constraints on the L_k are supplied: *e.g.* if it is required that $\ell_k^- \leq L_k \leq \ell_k^+$, then

$$\mathrm{Lip}'(G|_{\mathcal{O}}) := \left\{ L \in \mathrm{Lip}(G|_{\mathcal{O}}) \,\middle|\, \ell_k^- \leq L_k \leq \ell_k^+ \text{ for each } k \in \{1, \ldots, K\} \right\}$$

is a convex set.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

It is not immediately clear what one should regard as the "smallest" element of $\mathrm{Lip}(G|_{\mathcal{O}})$. However, recall that Theorem 3.3 shows that the gap size Γ of the data set with respect to d_L controls the error $\widehat{D}_k - \mathcal{D}_k[G]$:

$$0 \leq \widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}}, d_L] - \mathcal{D}_k[G] \leq 4\Gamma(\mathcal{X}, \mathcal{O}, d_L).$$

Therefore, it makes sense to search among the feasible Lipschitz constants $L \in \mathrm{Lip}(G|_{\mathcal{O}})$ for one L^* that minimizes the gap size. Unfortunately, this is not a *convex minimization problem* in the sense of [7], Section 4.2, since $\Gamma(\mathcal{X}, \mathcal{O}, d_L)$ is not a convex function of L: for each $x \in \mathcal{X}$, $d_L(x, \mathcal{O})$ is a concave function of L, and a supremum of a family of concave functions can be badly behaved. $\widehat{D}_k[\mathcal{X}, G|_{\mathcal{O}}, d_{L^*}]$ is then the upper bound on $\mathcal{D}_k[G]$ that has the tightest error estimate that can be justified by the data $G|_{\mathcal{O}}$ alone; of course, further data might invalidate this scenario.

6.2. Sensitivity and robustness analysis

In some applications, there may be doubt about the correct values for the Lipschitz constants $L_1, \ldots, L_K$. Such doubt necessarily propagates to doubt about the validity of the bounds $\widehat{D}_k$ and $\widehat{P}$: however, it does not do so in an entirely uncontrolled fashion. It is possible to perform a (local or global) sensitivity/robustness analysis of $\widehat{D}_k$ and $\widehat{P}$ with respect to $L_1, \ldots, L_K$ and thereby determine which Lipschitz constants strongly control the values of $\widehat{D}_k$ and $\widehat{P}$; the key Lipschitz constants can be identified for further, more detailed, research; the less important ones can be (relatively) safely accepted as they stand.

Notably, as in the optimal concentration-of-measure inequalities of McDiarmid and Hoeffding type [28], Section 4 some L_k may turn out to have *zero* influence on $\widehat{D}_k$ and $\widehat{P}$. Indeed, by rescaling arguments, it is easy to see that just as $\widehat{D}_k$ and $\widehat{P}$ may be discontinuous as functions of the observed data $G|_{\mathcal{O}}$ (as in Example 4.3), $\widehat{D}_k$ and $\widehat{P}$ may be discontinuous as functions of L.

7. Numerical examples

This section covers the numerical calculation of $\widehat{P}$ in two example cases. The first case (Sect. 7.2) is a validation exercise, in which the closed-form results of Example 4.3 are replicated numerically. The second case (Sect. 7.3) is a more involved calculation, in which the response function is a function of three variables and the data set comes from an archive of impact engineering experiments.

7.1. Overview of the numerical method

A description of the OUQ algorithm, as implemented in the *mystic* framework [21], can be found in [22,23]. In those earlier implementations of OUQ, it was the case that the response function was known/modelled exactly, and so it was only necessary to numerically represent the unknown probability measure μ. To implement the "Legacy OUQ" method of this paper, it was necessary to extend the existing OUQ algorithm in the following ways:

- *Mystic*'s `product_measure` class, which provides a numerical representation of a probability measure μ of the form (4.5)/(4.6), was extended to associate to each of the support points of a product measure μ a scalar value, thereby providing a numerical representation of a pair $(g, \mu) \in \mathcal{A}_\Delta$ as in (4.8). Such an object will be referred to as a `scenario` and denoted `X`; typically, `X` is stored in the "compressed" form of (x_0, x_1, p, y) as used in (4.12) and elsewhere, but is sometimes converted into other representations.
- A `dataset` class, which numerically represents the observed data $G|_{\mathcal{O}}$ and the cone structure that comes from the Lipschitz constants, was added. As alluded to in the previous bullet point, a `scenario` object `X` can be regarded as a `dataset` object by "forgetting" the probabilistic structure and remembering only the points in input parameter space and their associated output values. Below, the legacy data set $G|_{\mathcal{O}}$ will be denoted `data`.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

- Methods were added to both of these classes to allow for efficient calculation of d_L distances (and hence whether or not a given `scenario` object `X` is d_L-short with respect to itself and `data`) and integrals with respect to μ as in (4.7).

The overall structure of the optimization calculations is that of an outer and an inner optimization loop. The outer loop generates the next population of candidate `scenario` objects `X` to which the objective function `F` (the probability-of-failure functional) will be applied. The inner loop applies the constraints (bounds, mean, and shortness) to those generated candidates `X` so that `F` is only ever evaluated on `scenario` objects `X'=C(X)` that satisfy the constraints imposed by `C`.

The outer optimization loop, as described in [22], is used with the "expanded solver interface" described in [23]. A differential evolution solver [29, 34] was used with termination condition `ChangeOverGenerations`, population size `npop` $= 32$, `ngen` $= 100$, and `tol` $= 10^{-6}$; that is, the calculations used populations of 32 candidates and terminated when the best objective function value had shown no improvement greater than 10^{-6} for 100 consecutive iterations of the outer loop. The objective function value `F(X)`, when `X` represents (g, μ), is the probability of failure for g under μ as defined in (4.7) with $r(x) := \mathbf{1}[g(x) \leq \theta]$. The optimizer generates values for the weights and positions of the measure points in each coordinate direction. For Legacy OUQ, the optimizer must also generate scalar values $y = g(x)$ for each point x in the support of the product measure μ.

In *mystic*, constraints are solved explicitly through algebraic or numerical means. A *constraints solver* `C` is built to impose the set of constraints on the candidate `scenario` generated by the outer loop optimizer at each iteration. Constraints solvers are functions that map any (not necessary feasible) `scenario` object `X` to a `scenario` object `X'=C(X)` that satisfies all of the required constraints. Thus, only valid solutions to the constraints equations are seen by the objective function `F`. Effectively, the value of the objective function value evaluated by the outer loop optimizer at each step is `F(C(X))`. In contrast, standard optimizers use penalty functions `P` (and often dynamic multipliers `k`) so that the objective function `F` as evaluated by the optimizer is in fact `F(X)+k*P(X)`; this approach corrupts the structure of the problem by severing an explicit connection to the constraints.

The constraints function used in the Legacy OUQ algorithm first builds the `scenario` object `X` from the optimizer-generated inputs to the objective function. A first constraints solver `C'` is then applied: this ensures that the weights of each of the underlying discrete measures sum to `1.0`. A second constraints solver `C''` is then applied, which imposes the mean constraint $\mathbb{E}_\mu[g] \geq m$; this is done through *mystic*'s `impose_mean` function, which, in our example, shifts the coordinates of `X` so that `X` has the desired mean. At this point, the candidate `scenario` objects `X` generated by the optimizer have passed through the constraints solvers `C'` and `C''`, and only provide the objective function `F` with valid solutions `X'=C*(X)=C''(C'(X))` of the given bounds and mean constraints. If the resulting candidate `scenario` object `X'` is not d_L-short with respect to itself and to the legacy data `data`, *i.e.* the inequality

$$|g(x) - g(x')| \leq d_L(x, x')$$

fails for some x in the support of μ and some x' either in the support of μ or in $\mathcal{O}$, then *mystic*'s `set_feasible` function is used in a third constraints solver `C` to impose the desired shortness on the `scenario` object `X'`. Unlike for `C'` and `C''`, the constraints in `C` can not be imposed algebraically. Instead, the application of `C` is an inner optimization loop.

The details of how *mystic* checks for shortness and how feasibility is imposed on a `scenario` object are worth a little further discussion.

The check for shortness of a scenario `X` with respect to the legacy data `data` is done by first converting `X` into a `dataset` object with the `load` method, and then applying the `is_short` function, which calculates the a 2-dimensional array `dist` with elements $|y - y'| - d_L(x, x')$ for each combination of x, x' from the two collections of support points (here, the legacy data set `data` and the scenario `X` regarded as a data set). The result is a matrix corresponding to the distances required for shortness, where all distances less than a given tolerance `short_tol` are treated as acceptably close to zero; if all entries of the matrix `dist` are at most `short_tol`, then,

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

modulo that tolerance, `X` is d_L-short with respect to `data`; otherwise, the positivity of the matrix `dist` provides a numerical measure of the failure of shortness. Shortness of `X` with respect to itself is calculated similarly.

Shortness is imposed through an inner optimization loop that solves for a candidate `scenario` object `X'` for which `dist<=short_tol`. Similarly to the outer optimization loop, this inner optimization loop uses a differential evolution solver – however, the termination condition used in the inner loop is `VTR` [21], and solver parameters are set to `npop` $= 40$ and `tol` $= 10^{-9}$. The constraints solver `C*` described above is reused by the inner optimization loop to ensure that the constraints on the weights and mean are also respected by `C`. For shortness, the objective function for the inner loop is the sum over all elements of the matrix `max(0.0, dist-short_tol)`. When the inner loop terminates, a candidate `scenario` object `X'=C(X)` is produced that satisfies all constraints imposed by the solver `C` (and thus also `C*`).

The solution produced by the outer optimization loop is a `scenario` object `C(X)` that both satisfies all of the above constraints and maximizes the probability of failure `F(C(X))`.

7.2. One data oint in one dimension

As a first exercise in applying the protocol, we numerically replicate the exact values for $\widehat{P}$ in Example 4.3. Numerical convergence plots are given in Figure 8. In this subsection and the next, $\widehat{P}_n$ denotes the optimizer's best approximation to $\widehat{P}$ after n outer loop iterations.

It may be useful to note that the dimensionality of the problem can be slightly reduced, and more accurate results obtained more quickly, if instead of searching over

$$(x_0, x_1, y_0, y_1, p) \in [0,1]^2 \times \mathbb{R}^2 \times [0,1],$$

one instead forces (x_1, y_1) to be a failure, and therefore searches over

$$(x_0, x_1, y_0, y_1, p) \in [0,1]^2 \times \mathbb{R} \times \{0\} \times [0,1].$$

The same value for $\widehat{P}$ is attained using either approach; if $y = 0$ is not a feasible value for any $x \in [0,1]$, then the optimizer detects this fact and reports that the feasible set is empty, from which we infer that the maximum probability of failure is zero.

7.3. Three-dimensional example

This subsection reports the results of implementing the above method for obtaining optimal bounds on probabilities using a data set generated by physical experiments. These experiments were performed at the California Institute of Technology's Small Particle Hypervelocity Impact Range (SPHIR) facility. A brief description of the experimental setup is given in the next two paragraphs; the essential mathematical point is that Table 1 forms the legacy data for a function G of three real-valued inputs with smoothness given by (7.1).

In these experiments, a solid steel ball of diameter 0.07 inches is fired at an aluminium plate of thickness h. The projectile impacts the plate at an angle α away from the plate normal (referred to as the *obliquity* of the impact), and at a speed v. This impact event may result in the plate being perforated[8] by the projectile.

The impact event is very complicated, with many physical processes happening at very high rates; the experimental diagnostics and the numerical modelling of the entire event are beyond the scope of this paper, and further details can be found in [1, 13]. To simplify matters and focus on the relevant mathematics, this example selects a single, scalar, "post mortem" quantity of interest: after the impact event, the cross-sectional area $G(h, \alpha, v)$ (in mm^2) of the perforation in the plate is measured using an optical scanner and recorded, with the obvious convention that failure to perforate means that $G(h, \alpha, v) = 0$.

[8] To be precise, *perforation* (also known as *complete penetration*) means that the impact event has caused a hole in the plate that passes fully from one side of the plate to the other; a topologist would say that the plate has changed topology from genus 0 to genus ≥ 1. The opposite situation, in which the plate is merely "dented" by the projectile, is referred to as a *penetration* or *partial penetration*. The shorthand terms "a complete" and "a partial" are in common use.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

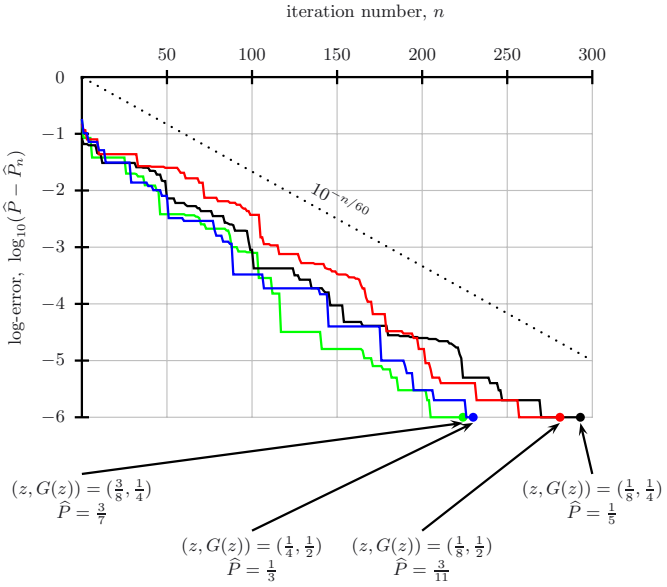


FIGURE 8. Log-linear plot illustrating typical numerical convergence of the approximate maxima $\widehat{P}_n$ as a function of the number n of outer loop iterations in the numerical implementation of Example 4.3. Note the approximate convergence rate of $|\widehat{P}_n - \widehat{P}| \approx 10^{-(1+n/60)}$. After the last iteration shown in each plot, $|\widehat{P}_n - \widehat{P}| \leq 10^{-6}$, *i.e.* the two are equal up to the convergence tolerance.

The results of a series of such impact tests are given in Table 1, which forms the legacy data set $G|_{\mathcal{O}}$ for this example. The protocol described above is now applied over the parameter space

$$(h, \alpha, v) \in \mathcal{X} := [0.062, 0.125]\,\text{in} \times [0, 30]\,\text{deg} \times [2300, 3200]\,\text{m} \cdot \text{s}^{-1}.$$

The data are, in fact, multi-valued (two distinct perforation areas were observed for the same input triplet (h, α, v)). Therefore, the response function is not Lipschitz continuous, and so we apply a natural generalization of the above protocol using the following "Lipschitz with tolerance" constraint:

$$|G(h, \alpha, v) - G(h', \alpha', v')| \leq d_L((h, \alpha, v), (h', \alpha', v')) + T, \tag{7.1}$$

where

$$L := (L_h, L_\alpha, L_v), \qquad T := 1.0\,\text{mm}^2,$$

$$L_h := 175.0\,\text{mm}^2/\text{in}, \qquad L_\alpha := 0.075\,\text{mm}^2/\text{deg}, \qquad L_v := 0.1\,\text{mm}^2/(\text{m} \cdot \text{s}^{-1}).$$

Condition (7.1) is satisfied by the observed data in Table 1, and we assume that it remains valid for the system in operation. We also assume that the system in operation will be exposed to random (h, α, v) taking values in $\mathcal{X}$, with independent components, and such that $\mathbb{E}[G(h, \alpha, v)] \geq 11.0\,\text{mm}^2$.

In this example, the "failure" event is that the perforation area $G(h, \alpha, v)$ falls below some threshold area θ. Figure 9 shows the computed least upper bound on $\mathbb{P}[G(h, \alpha, v) \leq \theta]$ for $\theta \in \{0, 1, \ldots, 12\}\,\text{mm}^2$. As expected, the least upper bound on $\mathbb{P}[G(h, \alpha, v) \leq \theta]$ is indeed 1 when $\theta \geq m$ and decreases as $m - \theta$ increases.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

TABLE 1. Hypervelocity impact legacy data. Note that this data set corresponds to a multi-valued function: see shots A62 and A77.

ID	plate thickness h/in	impact obliquity α/deg	impact speed v/m · s^{-1}	perforation area $G(h,\alpha,v)$/mm^2
A48	0.062	0.0	2288.0	7.73
A49	0.125	30.0	2840.0	13.38
A50	0.125	0.0	2556.0	11.83
A51	0.062	30.0	2329.0	6.31
A52	0.062	0.0	2363.0	7.78
A53	0.125	0.0	2326.0	9.26
A54	0.125	30.0	3235.0	15.98
A55	0.062	0.0	2686.0	9.86
A56	0.062	30.0	2728.0	11.35
A57	0.062	30.0	2627.0	12.09
A58	0.125	30.0	2531.0	11.24
A60	0.125	0.0	2363.0	9.93
A61	0.062	0.0	2707.0	9.96
A62	0.062	30.0	2756.0	11.07
A63	0.062	0.0	2614.0	9.02
A64	0.125	0.0	2439.0	10.52
A65	0.062	0.0	2485.0	8.56
A66	0.125	0.0	2607.0	12.46
A67	0.125	30.0	3036.0	15.36
A68	0.125	30.0	2325.0	8.15
A69	0.062	30.0	2702.0	10.81
A70	0.062	30.0	2473.0	9.52
A71	0.121	30.0	2520.0	9.47
A72	0.121	0.0	2439.0	10.19
A73	0.121	30.0	2366.0	9.42
A74	0.121	30.0	2402.0	8.68
A75	0.062	30.0	2413.0	9.19
A77	0.062	30.0	2756.0	11.32
A78	0.121	30.0	2432.0	10.00
A79	0.062	30.0	2393.0	9.29
A80	0.121	30.0	2479.0	9.53
A81	0.060	30.0	2356.0	8.27

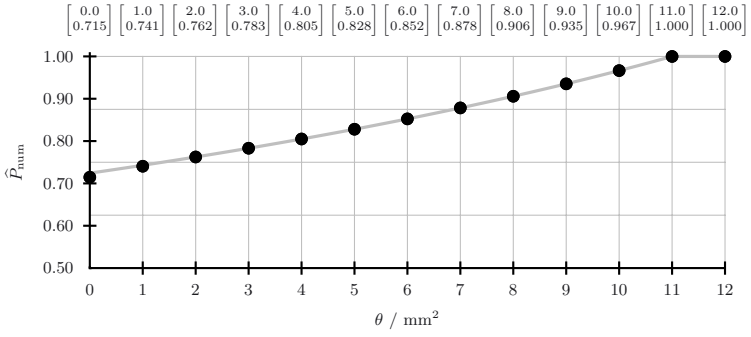


FIGURE 9. Numerical results for the least upper bound on $\mathbb{P}[G(h,\alpha,v) \le \theta]$ for various θ. Note the close agreement with the Markov bound (7.2) (grey line): for $\theta \ge 2.0\,\text{mm}^2$, the difference is less than the change-over-generations criterion of 10^{-6}. Convergence plots are given in Figure 11.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

Remark 7.1 (Markov bound and non-binding data)**.** One interesting feature of Figure 9 is that the numerical results demonstrate very close agreement with the Markov bound

$$\mathbb{P}[G(h,\alpha,v) \leq \theta] \leq \frac{M-m}{M-\theta}, \tag{7.2}$$

where

$$M := \sup_{(h,\alpha,v)\in\mathcal{X}} \inf_{z\in\mathcal{O}} \big(G(z) + d_L(z,(h,\alpha,v)) + T\big) \approx 39.895\,\mathrm{mm}^2 \tag{7.3}$$

with maximizer at

$$(h_M,\alpha_M,v_M) \approx (0.062\,\mathrm{in}, 0.0\,\mathrm{deg}, 3138.6\,\mathrm{m}\cdot\mathrm{s}^{-1}) \tag{7.4}$$

is the largest perforation area that can be realised anywhere in $\mathcal{X}$ subject to the data and the Lipschitz constraints. (We note in passing that efficient algorithms for finding extrema of Lipschitz functions are an area of independent interest: see *e.g.* [12].) Indeed, for $\theta \geq 2.0\,\mathrm{mm}^2$, the difference between the computed $\widehat{P}$ and Markov's bound is dominated by the numerical convergence criterion (less than $\texttt{tol} = 10^{-6}$ change over $\texttt{ngen} = 10^2$ consecutive generations).

This observation shows that most of the data set (*i.e.* those data points that do not determine M) consists of non-binding data points; indeed, only the constraints corresponding to data points A54 and A67 in Table 1 hold as equalities at $((h_M,\alpha_M,v_M),M)$. Put another way, the other 30 data points carry no information about $\widehat{P}$, and could have been ignored. Also, this finding suggests that the best next experiment to reduce the gap between $\widehat{P}$ and $\mathbb{P}[G(h,\alpha,v) \leq \theta]$ would be to determine $G(h_M,\alpha_M,v_M)$, since if it is discovered that in fact $G(h_M,\alpha_M,v_M) \ll M$, then $\widehat{P}$ will decrease considerably.

However, for $\theta \leq 1.0\,\mathrm{mm}^2$, a significant difference (10^{-2} or greater) is observed between the computed $\widehat{P}$ and Markov's bound; this order-10^{-2} difference was confirmed using runs with an extended convergence criterion (less than $\texttt{tol} = 10^{-6}$ change over $\texttt{ngen} = 10^3$ consecutive generations). This suggests that data points other than A54 and A67 supply relevant data in these cases, and that it is no longer feasible to have all the μ-probability mass located at $((h_M,\alpha_M,v_M),M)$ and $((h',\alpha',v'),\theta)$.

It is worth noting, though, that working with only the two relevant data points did not result in a statistically significant shortening of the algorithmic run-time. Instead, significant – even dramatic – reductions in computational cost resulted from reducing the dimension of the optimization problem rather than its constraints, as discussed in the next remark. This is not unexpected: problem (4.12) is a problem in $\sim 2^K$ unknowns with $\sim K2^K + |\mathcal{O}|2^K$ distinct constraints, so it is unsurprising that K has a much greater effect on computational cost than $|\mathcal{O}|$.

Remark 7.2 (Dimensional collapse)**.** An interesting empirical observation about the solutions of the optimization problem is that, during the course of the calculation, the approximate maximizers appear to undergo a kind of "dimensional collapse", as illustrated in Figure 10. That is, the extremizing measure μ does not have support on the 8 distinct points of a non-degenerate discrete cube $\mathcal{C}(x_0,x_1)$; instead, the support of the measure collapses to just one point in the h and α marginals. This indicates that the uncertainty in the impact velocity v is the dominant uncertainty in this problem.

Furthermore, once this "dimensional collapse" phenomenon has been observed, even approximately, it is natural to try the calculation of $\widehat{P}$ using $1\times1\times2$ product measures instead of $2\times2\times2$ product measures; this approach always produces valid lower bounds on $\widehat{P}$ and, as Figure 11 shows, can greatly reduce the computational burden. In this way, lower bounds on the solution of a large OUQ problem can be found relatively quickly by considering lower-dimensional sub-problems.

The automated implementation of this heuristic for general OUQ problems with $n_1 \times \cdots \times n_K$ product measures, in which dimensional collapse events are diagnosed "on the fly" during an optimization and then enforced as additional simplifying constraints, and the resulting improvements to computational efficiency, will be the topic of a future paper.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

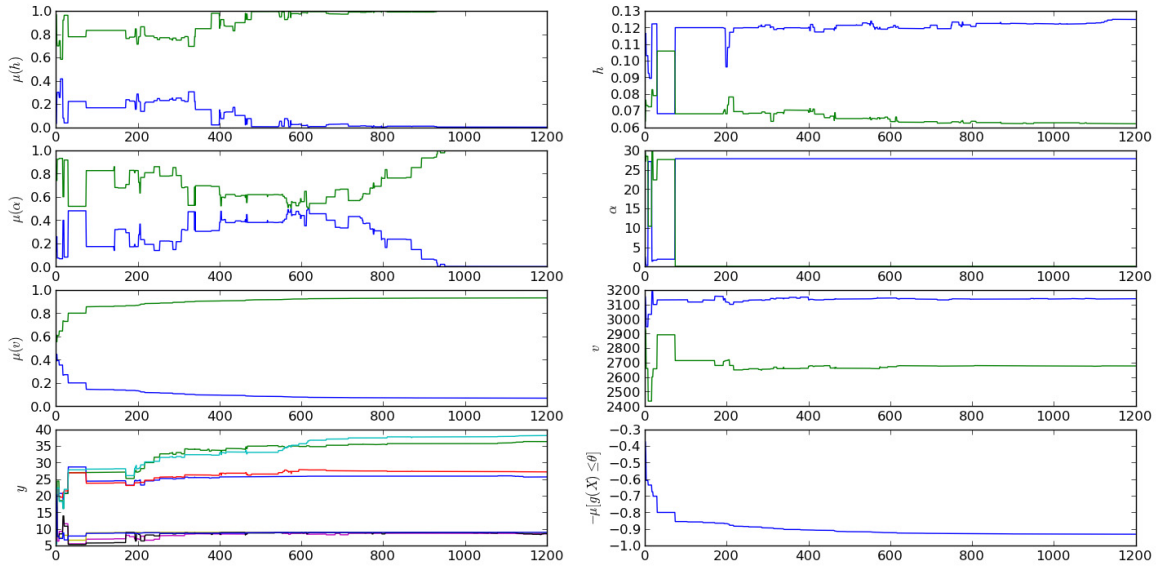


FIGURE 10. Illustration of the dimensional collapse phenomenon for the approximate maximizers for $\theta = 9.0\,\mathrm{mm}^2$ in Figure 9. The first three rows show the μ-probability (left column) and position (right column) of the h, α and v coordinates of the support of μ. The bottom-left figure shows the y-values, and the bottom-right the negative of $\mu[g(X) \leq \theta]$, *i.e.* $-\widehat{P}_n$. In the later iterations, μ is effectively a $1 \times 1 \times 2$, not a $2 \times 2 \times 2$, product measure.

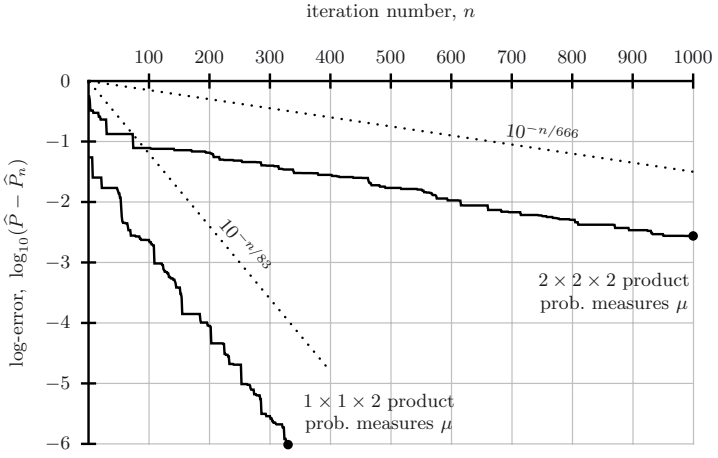


FIGURE 11. Log-linear plot illustrating typical numerical convergence for the approximate maximum $\widehat{P}_n$ for $\theta = 9.0\,\mathrm{mm}^2$ in Figure 9 at full $2 \times 2 \times 2$ dimensionality and reduced $1 \times 1 \times 2$ dimensionality. Note the improvement to the convergence rate obtained by operating at reduced dimensionality.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

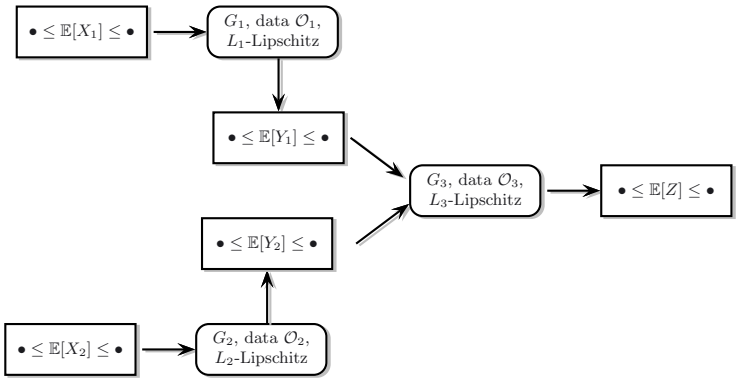


FIGURE 12. For $i \in \{1, 2\}$, bounds on the expected value of X_i can be propagated through a system G_i that is known on $\mathcal{O}_i$ and has Lipschitz constant L_i to yield optimal bounds on the expectation of some output quantity Y_i. The bounds on Y_1 and Y_2 can then be propagated through a third system G_3, and so on.

8. GENERALIZATIONS

8.1. Additional statistical information

The approach of Section 4 is open to a great deal of generalization, much more so than that of Section 3. In principle, *any* information about G and $\mathbb{P}$ can be used to define a set of admissible scenarios $\mathcal{A}$ for the optimization problem (4.1)–(4.2); also, the objective function can be more general than the probability of failure. Let $r \colon \mathcal{X} \to \mathbb{R}$ be measurable. As shown in [28], if $\mathcal{A}$ is described by independence constraints and inequalities of the form

$$\begin{aligned} \mathbb{E}_{\mu}\left[\varphi_i'\right] &\leq 0, & &\text{for } i \in \{1, \ldots, n'\}, \\ \mathbb{E}_{\mu_k}\left[\varphi_i^{(k)}\right] &\leq 0, & &\text{for } k \in \{1, \ldots, K\},\ i \in \{1, \ldots, n_k\}, \end{aligned}$$

for given measurable functions $\varphi_i \colon \mathcal{X} \to \mathbb{R}$ and $\varphi_i^{(k)} \colon \mathcal{X}_k \to \mathbb{R}$, then, to extremize $\mathbb{E}_{\mu}[r]$ over $\mu \in \mathcal{A}$, it is sufficient to search over measures $\mu = \bigotimes_{k=1}^{K} \mu_k \in \mathcal{A}$ with μ_k supported on at most $n' + n_k + 1$ points of $\mathcal{X}_k$; this paper made use only of the case $r = \mathbf{1}[f \leq \theta]$, $n' = 1$, $\varphi_1' = m - f$, $n_k \equiv 0$. In particular, independence assumptions can be relaxed, and information about the moments and correlations of the input random variables X_k can be included in the definition of $\mathcal{A}$. If such information is used, then a reduced upper bound on the probability of failure is obtained, but at the cost of solving a higher-dimensional optimization problem.

Since, in general, the same methods can be used to provide optimal bounds on $\mathbb{E}_{\mu}[r]$ for any quantity of interest r, the methods of this paper can be used to optimally propagate uncertainties through a hierarchy (directed acyclic graph) of partially-observed input-output relationships, as in [38]. See Figure 12 for a schematic illustration.

8.2. Measurement uncertainty

Bounded measurement uncertainty can also be incorporated in the inequality constraints. More precisely, suppose that an error of up to $\pm\delta$ is associated to the observed value $G(z)$, and an error of up to δ' with respect to the metric d_L is associated to the corresponding input parameter value z. Then the observed datum is not $(z, G(z))$ but rather some $\left(\widetilde{z}, \widetilde{G}(\widetilde{z})\right) \in \mathcal{X} \times \mathbb{R}$ such that

$$d_L(z, \widetilde{z}) \leq \delta' \text{ and } \left|G(z) - \widetilde{G}(\widetilde{z})\right| \leq \delta.$$

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

In this situation, the Lipschitz constraints of the form

$$|y - G(z)| \leq d_L(x, z) \tag{8.1}$$

generalize to

$$\big|y - \gamma\big| \leq d_L(x, \zeta), \tag{8.2}$$

where $(\zeta, \gamma) \in \mathcal{X} \times \mathbb{R}$ is a new optimization variable that plays the rôle of the imperfectly-observed input-output pair $(z, G(z))$, and, therefore, is constrained to satisfy

$$d_L(\zeta, \widetilde{z}) \leq \delta' \text{ and } \left|\gamma - \widetilde{G}(\widetilde{z})\right| \leq \delta. \tag{8.3}$$

Note that, geometrically, (8.2)–(8.3) corresponds to a pointed double cone with a movable vertex that must remain close to $\big(\widetilde{z}, \widetilde{G}(\widetilde{z})\big)$, whereas (7.1) corresponds to a fixed and blunt double cone. Note that, as in the simple situation of Example 4.3, the bounds $\widehat{D}_k$ and $\widehat{P}$ may be discontinuous as functions of δ and δ'.

If specific statistical information is available about the measurement uncertainty (*e.g.* Gaussian scatter), then confidence intervals can be used in the above procedure. The resulting bounds on $\mathbb{P}[G(X) \leq \theta]$ will be probabilistic in nature, and will become looser as the required level of confidence increases.

8.3. Model-based certification

In many applications, although the real response function $G\colon \mathcal{X} \to \mathbb{R}$ cannot be easily exercised, there may be a *model* $F\colon \mathcal{X} \to \mathbb{R}$ for G that can be used instead. Quantitative relationships between G and F can be used to define sets of admissible scenarios as before. For example, suppose that it is known that

$$\|G - F\|_\infty := \sup_{x \in \mathcal{X}} |G(x) - F(x)| \leq C_V, \tag{8.4}$$

where $C_V \geq 0$ is some constant resulting from an exercise in *model validation*. Then, compared with the admissible set $\mathcal{A}$ of (4.2), the corresponding set $\mathcal{A}_F$ that uses also the model F and the information (8.4) is

$$\mathcal{A}_F := \left\{ (g, \mu) \,\middle|\, \begin{array}{c} g\colon \mathcal{X} \to \mathbb{R} \text{ is } d_L\text{-short}, \\ \mu = \mu_1 \otimes \cdots \otimes \mu_K \in \bigotimes_{k=1}^K \mathcal{P}(\mathcal{X}_k), \\ \|g - F\|_\infty \leq C_V,\ g = G \text{ on } \mathcal{O}, \text{ and } \mathbb{E}_\mu[g] \geq m \end{array} \right\} \subseteq \mathcal{A}.$$

Hence, in the $\mathcal{A}_F$-analogue of the reduced problem (4.12), the model F and (8.4) induce additional constraints of the form

$$|y_\varepsilon - F(x_\varepsilon)| \leq C_V \text{ for each } \varepsilon \in \{0, 1\}^K.$$

As remarked above, $\widehat{D}_k$ and $\widehat{P}$ may be discontinuous as functions of C_V.

Other quantitative measures of model validity can be used in similar ways. Without going into detail, we note that the uniform norm in (8.4) is too strong for many applications, particularly those in which F or G may have discontinuities: in such cases, $\|F - G\|_\infty$ being small requires that F and G have *approximately* the same discontinuities in $\mathbb{R}$ at *exactly* the same locations in $\mathcal{X}$, which is a very strong requirement. Therefore, metrics that allow "wiggle room" in both $\mathcal{X}$ and $\mathbb{R}$, *e.g.* the various Skorohod metrics [6,32], are expected to be of use in this area. For example, it may be reasonable to assume that the distance between the graphs of F and G as subsets of $\mathcal{X} \times \mathbb{R}$ is small enough that, for some $C'_V \geq 0$,

$$\sup_{x \in \mathcal{X}} \inf_{x' \in \mathcal{X}} \max\{d_L(x, x'), |G(x) - F(x')|\} \leq C'_V; \tag{8.5}$$

i.e. every point on the graph of G lies within distance C'_V of some point on the graph of F. (Note well that the roles of F and G in (8.5) are not symmetric.) In this case, the corresponding constraint satisfied by any feasible $(x_\varepsilon, y_\varepsilon) \in \mathcal{X} \times \mathbb{R}$ is that

$$\inf_{\substack{x' \in \mathcal{X} \\ d_L(x_\varepsilon, x') \leq C'_V}} |y_\varepsilon - F(x')| \leq C'_V.$$

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

8.4. Set-valued lipschitz functions

In many applications (*e.g.* inverse problems, which are often ill-posed), the system of interest cannot be accurately represented as a single-valued function $G\colon \mathcal{X} \to \mathbb{R}$. For example, the system outcome may depend on so-called *unknown unknowns*, which can be neither controlled nor even observed, but have the effect that $G(x)$ is not a uniquely determined real number for each fixed $x \in \mathcal{X}$. One resolution to this problem is to treat G as a partially-observed *set-valued* function $G\colon \mathcal{X} \rightsquigarrow \mathbb{R}$, *i.e.* an operation that assigns to each $x \in \mathcal{X}$ a (possibly empty) subset of $\mathbb{R}$. There is a notion of Lipschitz continuity for set-valued functions [2]: for metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, a set-valued function $G\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ is said to be a *set-valued Lipschitz function* with Lipschitz constant $L \geq 0$ if, for all $x, x' \in \mathcal{X}$,

$$G(x) \subseteq \left\{ y \in \mathbb{R} \,\middle|\, \operatorname{dist}(y, G(x')) := \inf_{y' \in G(x')} d_{\mathcal{Y}}(y, y') \leq L d_{\mathcal{X}}(x, x') \right\}, \tag{8.6}$$

that is, $G(x)$ is a subset of the uniform $Ld_{\mathcal{X}}(x, x')$-neighbourhood of $G(x')$; or, equivalently, the Hausdorff distance between the sets $G(x)$ and $G(x')$ is at most $Ld_{\mathcal{X}}(x, x')$.

It would be an interesting and natural extension of the present work to consider set-valued response functions. Indeed, the set of single-valued Lipschitz extensions $\mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ as defined in (3.2) defines a set-valued function $\widetilde{G}\colon \mathcal{X} \rightsquigarrow \mathbb{R}$ by

$$\widetilde{G}(x) := \{ g(x) \mid x \in \mathcal{X}, g \in \mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L) \}.$$

$\widetilde{G}$ is a set-valued Lipschitz function, with Lipschitz constant 1 with respect to the metric d_L, and $\mathcal{E}(\mathcal{X}, G|_{\mathcal{O}}, d_L)$ is the collection of Lipschitz selections [2], Section 9.4.3 of $\widetilde{G}$. In this paper, since G is assumed to be single-valued, the sets $\widetilde{G}(x)$ are all convex; in the general situation, this need not be the case.

Acknowledgements. Portions of this work were supported by the US Department of Energy NNSA under award DE-FC52-08NA28613 through the California Institute of Technology's ASC/PSAAP Center for the Predictive Modeling and Simulation of High Energy Density Dynamic Response of Materials. We thank the California Institute of Technology PSAAP Center's Experimental Science Group – in particular, M. Adams, J.M. Mihaly and A. Rosakis – for the data set in Table 1. Finally, we thank three anonymous referees for their helpful comments.

References

[1] M. Adams, A. Lashgari, B. Li, M. McKerns, J.M. Mihaly, M. Ortiz, H. Owhadi, A.J. Rosakis, M. Stalzer T.J. Sullivan, Rigorous model-based uncertainty quantification with application to terminal ballistics. Part II: Systems with uncontrollable inputs and large scatter. *J. Mech. Phys. Solids* **60** (2011) 1002–1019.

[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Modern Birkhäuser Classics, Birkhäuser Boston Inc., Boston, MA (2009), Reprint of the 1990 edition [MR1048347].

[3] I. Babuška, F. Nobile and R. Tempone, Reliability of computational science. *Numer. Methods Partial Differ. Eq.* **23** (2007) 753–784.

[4] R. E. Barlow and F. Proschan, Mathematical Theory of Reliability, in vol. 17 of *Classics in Applied Mathematics*. Society Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1996). With contributions by L. C. Hunter, Reprint of the 1965 original [MR 0195566].

[5] D. Bertsimas and I. Popescu, Optimal inequalities in probability theory: a convex optimization approach. *SIAM J. Optim.* **15** (2005) 780–804.

[6] P. Billingsley, Convergence of Probability Measures, 2nd edn., *Wiley Series in Probability and Statistics: Probability and Statistics*. John Wiley and Sons Inc., New York (1999). http://dx.doi.org/10.1002/9780470316962. MR 1700749 (2000e:60008)

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Cambridge (2004).

[8] H. Federer, *Geometric Measure Theory*, Die Grundlehren der Mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York (1969).

[9] W. Hoeffding, The role of assumptions in statistical decisions. *Proc. of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, 1954–1955 (Berkeley and Los Angeles). University of California Press (1956) 105–114.

[10] A. Holder, *Mathematical Programming Glossary*, INFORMS Computing Society, http://glossary.computing.society.informs.org (2006). Originally authored by H. J. Greenberg, 1999–2006.

[11] J.R. Isbell, Six theorems about injective metric spaces, *Comment. Math. Helv.* **39** (1964), 65–76.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press

[12] D.R. Jones, C.D. Perttunen and B.E. Stuckman, Lipschitzian optimization without the Lipschitz constant. *J. Optim. Theory Appl.* **79** (1993) 157–181.

[13] A.A. Kidane, A. Lashgari, B. Li, M. McKerns, M. Ortiz, H. Owhadi, G. Ravichandran, M. Stalzer and T.J. Sullivan, Rigorous model-based uncertainty quantification with application to terminal ballistics. Part I: Systems with controllable inputs and small scatter. *J. Mech. Phys. Solids* **60** (2011) 983–1001.

[14] M.D. Kirszbraun, Über die zusammenziehende und Lipschitzsche Transformationen. *Fund. Math.* **22** (1934) 77–108.

[15] V. Klee and G.J. Minty, How good is the simplex algorithm?, Inequalities, III, in *Proc. Third Sympos.* (Univ. California, Los Angeles, Calif., 1969; dedicated to the memory of Theodore S. Motzkin). Academic Press, New York (1972) 159–175.

[16] P. Limbourg, Multi-objective optimization of problems with epistemic uncertainty, Evolutionary Multi-Criterion Optimization, in *Lect. Notes Comput. Sci.,* of vol. 3410, edited by C.A. Coello Coello, A. Hernández Aguirre and E. Zitzler. Springer Berlin/Heidelberg (2005) 413–427.

[17] L.J. Lucas, H. Owhadi and M. Ortiz, Rigorous verification, validation, uncertainty quantification and certification through concentration-of-measure inequalities. *Comput. Methods Appl. Mech. Engrg.* **197** (2008) 51–52, 4591–4609.

[18] C. McDiarmid, On the method of bounded differences, Surveys in combinatorics, *London Math. Soc.* in vol. 141 of *Lecture Note Ser.* Cambridge Univ. Press, Cambridge (1989) 148–188.

[19] C. McDiarmid, Centering sequences with bounded differences, *Combin. Probab. Comput.* **6** (1997) 79–86,

[20] C. McDiarmid, Concentration, Probabilistic Methods for Algorithmic Discrete Mathematics. In vol. 16 of *Algorithms Combin.* Springer, Berlin (1998) 195–248.

[21] M. McKerns, P. Hung and M. Aivazis, *Mystic: A simple model-independent inversion framework* (2009).

[22] M. McKerns, H. Owhadi, C. Scovel, T.J. Sullivan and M. Ortiz, *The optimal uncertainty algorithm in the mystic framework*, Caltech CACR Technical Report, August 2010, available at http://arxiv.org/pdf/1202.1055v1.

[23] M.M. McKerns, L. Strand, T.J. Sullivan, A. Fang and M.A.G. Aivazis, Building a framework for predictive science. *Proc. of the 10th Python in Science Conference* (SciPy 2011), edited by S. van der Walt and J. Millman (2011) 67–78. Available at http://jarrodmillman.com/scipy2011/pdfs/mckerns.pdf.

[24] E.J. McShane, Extension of range of functions. *Bull. Amer. Math. Soc.* **40** (1934) 837–842.

[25] R. Morrison, C. Bryant, G. Terejanu, K. Miki and S. Prudhomme, Optimal data split methodology for model validation, *Proc. of World Congress on Engrg and Comput. Sci.* (2011) vol. II, 1038–1043.

[26] W.L. Oberkampf, J.C. Helton, C.A. Joslyn, S.F. Wojtkiewicz and S. Ferson, Challenge problems: Uncertainty in system response given uncertain parameters. *Reliab. Eng. Sys. Safety* **85** (2004) 11–19.

[27] W.L. Oberkampf, T.G. Trucano and C. Hirsch, Verification, validation and predictive capability in computational engineering and physics. *Appl. Mech. Rev.* **57** (2004) 345–384.

[28] H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns and M. Ortiz, Optimal Uncertainty Quantification. *SIAM Rev.* To appear.

[29] K.V. Price, R.M. Storn and J.A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization, *Natural Comput. Ser.* Springer-Verlag, Berlin (2005).

[30] C.J. Roy and W.L. Oberkampf, *A complete framework for verification, validation and uncertainty quantification in scientific computing*, 48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition (2010).

[31] L. Schwartz, *Radon Measures on Arbitrary Topological Spaces and Cylindrical Measures*, Published for the Tata Institute of Fundamental Research, Bombay by Oxford University Press, London (1973). Tata Institute of Fundamental Research Studies in Mathematics, No. 6.

[32] A.V. Skorohod, Limit theorems for stochastic processes, *Teor. Veroyatnost. i Primenen.* (*Theor. Probab. Appl.*) **1** (1956), 289–319.

[33] L.A. Steen and J.A. Seebach, Jr., *Counterexamples in Topology*, 2nd edn. Springer-Verlag, New York (1978).

[34] R. Storn and K. Price, Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **11** (1997) 341–359.

[35] A.M. Stuart, Inverse problems: a Bayesian perspective. *Acta Numer.* **19** (2010) 451–559.

[36] T. J. Sullivan, U. Topcu, M. McKerns and H. Owhadi, Uncertainty quantification via codimension-one partitioning. *Int. J. Numer. Meth. Engng.* **85** (2011) 1499–1521.

[37] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces. Inst. Hautes Études Sci. Publ. Math.* (1995) 73–205.

[38] U. Topcu, L. J. Lucas, H. Owhadi and M. Ortiz, Rigorous uncertainty quantification without integral testing. *Reliab. Eng. Sys. Safety* **96** (2011) 1085–1091.

[39] F.A. Valentine, A Lipschitz condition preserving extension for a vector function. *Amer. J. Math.* **67** (1945) 83–93.

[40] V.H. Vu, Concentration of non-Lipschitz functions and applications, *Random Structures Algorithms* **20** (2002) 262–316.

[41] M.L. Wage, The product of Radon spaces, *Uspekhi Mat. Nauk* **35** (1980) 151–153, International Topology Conference (Moscow State Univ., Moscow, 1979), Translated from the English by A.V. Arhangel′skiĭ.

https://doi.org/10.1051/m2an/2013083 Published online by Cambridge University Press