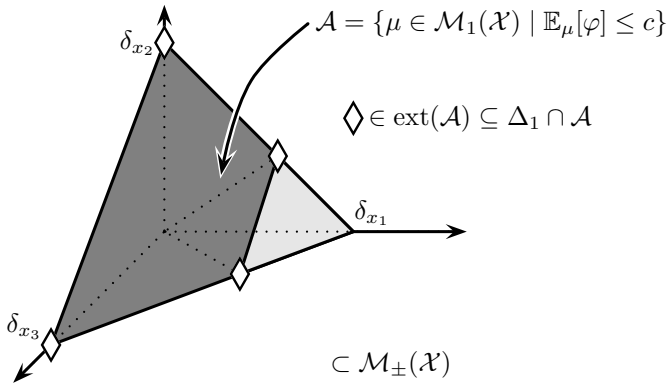
Optimal Distributionally Robust Uncertainty Quantification

T. J. Sullivan
Free University of Berlin and Zuse Institute Berlin
Takustrasse 7, D-14195 Berlin-Dahlem, Germany
sullivan@zib.de



Lecture notes for the École Thématique sur les Incertitudes en Calcul Scientifique / Research School on Uncertainty in Scientific Computing held at the Centre de séminaire Séolane, Barcelonette, France, 6–10 June 2016.

This page intentionally left almost blank.

Contents

1 Introduction

> Technology, in common with many other activities, tends toward avoidance of risks by investors. Uncertainty is ruled out if possible. [P]eople generally prefer the predictable. Few recognize how destructive this can be, how it imposes severe limits on variability and thus makes whole populations fatally vulnerable to the shocking ways our universe can throw the dice.
>
> *Heretics of Dune*
> Frank Herbert

In uncertainty quantification, one is usually faced with the challenge of quantifying the impact of some uncertainty or random variability (often modelled as a probability distribution μ) on a particular system of interest (often modelled as a response function g). These lecture notes are an introduction to uncertainty quantification under a particularly severe form of uncertainty, namely uncertainty about μ and g themselves. This kind of uncertainty can arise very easily: we may be conducting simulations using computational or numerical versions of μ and g that differ in some way from their 'real' counterparts, or there may be non-negligible uncertainty about what the 'real' μ and g actually are. Nevertheless, the challenge is to provide rigorous and useful information about the system.

Example 1.1. Suppose that a company manufactures a particular part P for use in an aeroplane, and we are interested in how some property X of the part P affects some performance metric, or *quantity of interest*, $q(X)$. For simplicity, suppose that both X and $q(X)$ are real-valued. Consider the following hierarchy of complexity.

(a) At the simplest level, every instance of the part P is completely identical and in accordance with the design specification $x_0 \in \mathbb{R}$. In this case, X is a constant, $X = x_0$. We just need to evaluate the function q on this constant x_0, and then we are done.

(b) At the next level, we accept the due to manufacturing imperfections, sometimes $X \neq x_0$. Nevertheless, quality control procedures are such that, in some sense, $|X - x_0| \leq \delta$. Note that, in this case, we profess no knowledge about which values of X are more or less likely than others, except that those further than δ from x_0 are ruled out as impossible. To determine the corresponding worst- and best-case performance of the aeroplane, we need to minimise and maximise $q(x)$ over all x with $|x - x_0| \leq \delta$.

(c) At the next level, perhaps after extensive statistical studies of the manufacturing process, we even have a model for X as a random variable: X is distributed according to some probability distribution μ. Then, the challenge is to determine quantities like the mean performance, i.e. the expected value $\mathbb{E}_{X\sim\mu}[q(X)]$ and perhaps also measures of variability such as the standard deviation. Perhaps some outcome, such as $q(X) \leq t$, is considered to be 'failure', in which case we might care about the probability of failure, $\mathbb{P}_{X\sim\mu}[q(X) \leq t]$.

(d) However, if we are honest, the most realistic situation is that we are somewhere between the last two cases: we *partially* understand the probability distribution μ based on a *limited* amount of sample information. For example, we may know that $|X - x_0| \leq \delta/2$ with probability at least 99% under μ. This does not completely characterise μ, since it tells us nothing about how the probability mass of μ is distributed within the interval $[x_0 - \delta/2, x_0 + \delta/2]$, but it does significantly constrain what μ can do compared to case (b) above: μ can no longer put all its probability mass at $x_0 - \delta$, for example. The art is to use this information to give *bounds* on e.g. the probability of failure $\mathbb{P}_{X\sim\mu}[q(X) \leq t]$ that are rigorously true and also useful: it is no use to give the true but trivial bounds $0 \leq \mathbb{P}_{X\sim\mu}[q(X) \leq t] \leq 1$ unless there really are admissible μ that realise these bounds. Thus, we are arrive at the need to be able to *optimise* over measures μ.

Exercise 1.2. Adapt this discussion to your own field of interest. Perhaps X no longer describes some property of a manufactured part for an aeroplane, but the operating circumstances (e.g. traffic and meteorological and geological stresses) of a road bridge or road tunnel.

With examples like this in mind, it makes sense to develop mathematical theory and computational tools to allow us to explore admissible sets (or 'feasible sets') $\mathcal{A}$ for what μ and g could be. The tools that we use will be grounded in optimization theory, and will have a particularly strong connection to the now-classical theory of finite-dimensional linear programming, even though $\mathcal{A}$ will typically be infinite-dimensional.

Acknowledgements. These notes are an abridged selection of material from Sullivan (2015), and draw upon joint work with Owhadi et al. (2013), Sullivan et al. (2013), and Kamga et al. (2014). Those collaborations are gratefully acknowledged, as is the support of the Free University of Berlin within the Excellence Initiative of the German Research Foundation.

2 General Notation and Terminology

- $\mathbb{N} \coloneqq \{1, 2, 3, \dots\}$ denotes the natural numbers starting at 1.
- $\mathbb{R}$ denotes the real number system, with its usual arithmetic operations, absolute value $|\cdot|$, etc. For $m \in \mathbb{N}$, $\mathbb{R}^m$ denotes the vector space of m-tuples of real numbers, with its usual operations of vector addition and scalar multiplication.
- Calligraphic letters $\mathcal{X}$ and $\mathcal{Y}$ denote space for 'inputs' and 'outputs' of a response function $g\colon \mathcal{X} \to \mathcal{Y}$. For example, they may be finite sets, $\mathbb{R}$, $\mathbb{R}^m$, or a 'nice' subset of those. More precisely, $\mathcal{X}$ is a complete and separable metric space, equipped with its Borel σ-algebra $\mathscr{B}(\mathcal{X})$ (generated by the open sets), and the same applies for $\mathcal{Y}$. Sets $E \in \mathscr{B}(\mathcal{X})$ are called Borel-measurable, or simply measurable — such sets are the only ones for which it is permitted to consider notions of 'length', 'area', 'measure', or 'probability'.

- $\mathcal{M}_+(\mathcal{X})$ denotes the set of all non-negative σ-additive measures on $\mathcal{X}$, i.e. countably additive set functions $\mu\colon \mathscr{B}(\mathcal{X}) \to [0, +\infty]$ with

$$\mu(\varnothing) = 0 \quad \text{and} \quad \mu\left(\bigcup_{n\in\mathbb{N}} E_n\right) = \sum_{n\in\mathbb{N}} \mu(E_n)$$

whenever the $E_n \in \mathscr{B}(\mathcal{X})$ are pairwise disjoint. $\mathcal{M}_1(\mathcal{X}) \subset \mathcal{M}_+(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$, i.e. those μ for which $\mu(\mathcal{X}) = 1$. $\mathcal{M}_\pm(\mathcal{X}) \supset \mathcal{M}_+(\mathcal{X})$ denotes the vector space of signed measures on $\mathcal{X}$. $\mathcal{M}_\pm(\mathcal{X})$ can be given the total variation norm

$$\|\mu\|_{\mathrm{TV}} \coloneqq \sup\left\{ \sum_{n=1}^{N} |\mu(E_n)| \,\middle|\, E_1, \ldots, E_N \subseteq \mathcal{X} \text{ are pairwise disjoint} \right\}.$$

In particular, the total variation distance between two probability measures μ and ν (i.e. the total variation norm of their difference) is twice the greatest absolute difference in the two probability values that μ and ν assign to any measurable event E:

$$d_{\mathrm{TV}}(\mu, \nu) \equiv \|\mu - \nu\|_{\mathrm{TV}} = 2 \sup\bigl\{ |\mu(E) - \nu(E)| \,\big|\, E \in \mathscr{F} \bigr\}.$$

- When X is a random variable taking values in some space $\mathcal{X}$ and $\mu \in \mathcal{M}_1(\mathcal{X})$, the notation $X \sim \mu$ is read 'X is distributed according to μ' and means that the probability that X takes a value in some measurable set $E \subseteq \mathcal{X}$ is exactly the μ-probability mass $\mu(E) \in [0, 1]$.
- $q\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a function, which is considered to be a quantity of interest. In particular, we are interested in the values of $q(X, f(X))$ when X is distributed according to some known or partially known probability measure μ on $\mathcal{X}$.
- When $\mu \in \mathcal{M}_1(\mathcal{X})$ and $f\colon \mathcal{X} \to \mathbb{R}$, we write $\mathbb{E}_{X\sim\mu}[f(X)]$ or simply $\mathbb{E}_\mu[f]$ for the expected value (Lebesgue integral) of f against μ:

$$\mathbb{E}_{X\sim\mu}[f(X)] \equiv \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x).$$

- For $a \in \mathcal{X}$, $\delta_x \in \mathcal{M}_+(\mathcal{X})$ denotes the Dirac measure or unit point mass centred at a. This is the probability measure

$$\delta_a(E) \coloneqq \begin{cases} 1, & \text{if } a \in E, \\ 0, & \text{if } a \notin E. \end{cases}$$

Integration against δ_a is a simple matter of point evaluation:

$$\mathbb{E}_{X\sim\delta_a}[f(X)] \equiv \int_{\mathcal{X}} f(x)\,\mathrm{d}\delta_a(x) \equiv f(a).$$

- $\mathbb{I}[P]$ is shorthand for the indicator function of a logical statement P, the function that takes the value 1 if P is true and 0 if P is false. The similar notation $\mathbb{I}_E$ denotes the indicator function of a set E: $\mathbb{I}_E(x) = \mathbb{I}[x \in E]$ evaluates to 1 if x is in E and to 0 otherwise.
- The support of a probability measure μ on $\mathcal{X}$ is denoted $\mathrm{supp}(\mu)$ and is defined to be the smallest closed set C so that $\mu(C) = 1$, or equivalently $\mu(\mathcal{X} \setminus C) = 0$. Thus, for example, $\mathrm{supp}(\delta_a) = \{a\}$, a Gaussian measure $\mathcal{N}(m, s^2)$ is supported on all of $\mathbb{R}$, and the lognormal distribution is supported on the half-line $[0, \infty)$.

3 Convex and Linear Optimization Theory

The topic of this section is *convex optimization*. As will be seen, convexity is a powerful property that makes optimization problems tractable to a much greater extent than any amount of smoothness (which still permits local minima) or low-dimensionality can do.

The general form of a constrained optimization problem is

$$\begin{aligned} \text{extremise: } & f(x) \\ \text{with respect to: } & x \in \mathcal{X} \\ \text{subject to: } & g_i(x) \in E_i \quad \text{for } i = 1, 2, \dots, \end{aligned}$$

where $\mathcal{X}$ is some set; $f \colon \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ is a function called the *objective function*; and, for each i, $g_i \colon \mathcal{X} \to \mathcal{Y}_i$ is a function and $E_i \subseteq \mathcal{Y}_i$ some subset. By 'extremise' we mean either to *minimise* (make as small as possible) or to *maximise* (make are large as possible). The conditions $\{g_i(x) \in E_i \mid i = 1, 2, \dots\}$ are called *constraints*, and a point $x \in \mathcal{X}$ for which all the constraints are satisfied is called *feasible*; the set of feasible points,

$$\{x \in \mathcal{X} \mid g_i(x) \in E_i \text{ for } i = 1, 2, \dots\},$$

is called the *feasible set*. If there are no constraints, so that the problem is a search over all of $\mathcal{X}$, then the problem is said to be *unconstrained*. In the case of a minimization problem, the objective function f is also called the *cost function* or *energy*; for maximization problems, the objective function is also called the *utility function*.

In this section, $\mathcal{X}$ will be a real normed vector space, such as $\mathbb{R}^m$ with its usual Euclidean norm; later on, when considering distributional robustness problems, it will be the space of signed measures, equipped with the total variation norm. Given two points x_0 and x_1 of $\mathcal{X}$ and $t \in [0, 1]$, x_t will denote the *convex combination*

$$x_t \coloneqq (1 - t)x_0 + tx_1.$$

More generally, given points $x_0, \dots, x_n$ of a vector space, a sum of the form

$$\alpha_0 x_0 + \cdots + \alpha_n x_n$$

is called a *linear combination* if the $\alpha_i \in \mathbb{R}$ are any scalars, an *affine combination* if their sum is 1, and a *convex combination* if they are non-negative and sum to 1.

Definition 3.1. (a) A subset $K \subseteq \mathcal{X}$ is a *convex set* if, for all $x_0, x_1 \in K$ and $t \in [0, 1]$, $x_t \in K$; it is said to be *strictly convex* if $x_t \in \mathring{K}$ whenever x_0 and x_1 are distinct points of $\bar{K}$ and $t \in (0, 1)$.

(b) An *extreme point* of a convex set K is a point of K that cannot be written as a non-trivial convex combination of distinct elements of K; the set of all extreme points of K is denoted $\operatorname{ext}(K)$.

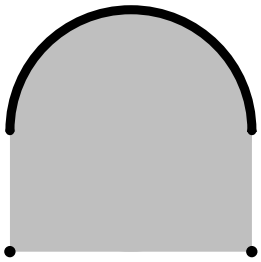
(c) The *convex hull* $\operatorname{co}(S)$ (resp. *closed convex hull* $\overline{\operatorname{co}}(S)$) of $S \subseteq \mathcal{X}$ is defined to be the intersection of all convex (resp. closed and convex) subsets of $\mathcal{X}$ that contain S.

Example 3.2. (a) The square $[-1, 1]^2$ is a convex subset of $\mathbb{R}^2$, but is not strictly convex, and its extreme points are the four vertices $(\pm 1, \pm 1)$.

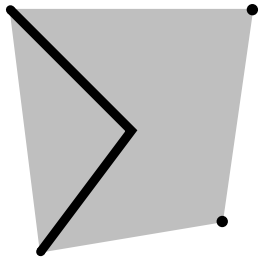
(b) The closed unit disc $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$ is a strictly convex subset of $\mathbb{R}^2$, and its extreme points are the points of the unit circle $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$.

(c) If $p_0, \dots, p_d \in \mathcal{X}$ are distinct points such that $p_1 - p_0, \dots, p_d - p_0$ are linearly independent, then their (closed) convex hull is called a *d-dimensional simplex*. The points $p_0, \dots, p_d$ are the extreme points of the simplex.

(d) See Figure 3.1 for further examples.



(a) A convex set (grey) and its set of extreme points (black).



(b) A non-convex set (black) and its convex hull (grey).

Figure 3.1: Convex sets, extreme points, and convex hulls of some subsets of the plane $\mathbb{R}^2$.

Example 3.3. $\mathcal{M}_1(\mathcal{X})$ is a convex subset of the space of all (signed) Borel measures on $\mathcal{X}$. The extremal probability measures are the *zero-one measures*, i.e. those for which, for every measurable set $E \subseteq \mathcal{X}$, $\mu(E) \in \{0, 1\}$. Furthermore, as will be discussed later on, if $\mathcal{X}$ is, say, a Polish space, then the zero-one measures (and hence the extremal probability measures) on $\mathcal{X}$ are the Dirac point masses. Indeed, in this situation,

$$\mathcal{M}_1(\mathcal{X}) = \overline{\text{co}}\big(\{\delta_x \mid x \in \mathcal{X}\}\big) \subseteq \mathcal{M}_{\pm}(\mathcal{X}).$$

The principal reason to confine attention to normed spaces[3.1] $\mathcal{X}$ is that it is highly inconvenient to have to work with spaces for which the following 'common sense' results do not hold:

Theorem 3.4 (Kreĭn–Milman)**.** *Let $K \subseteq \mathcal{X}$ be compact and convex. Then K is the closed convex hull of its extreme points.*

Theorem 3.5 (Choquet–Bishop–de Leeuw)**.** *Let $K \subseteq \mathcal{X}$ be compact and convex, and let $c \in K$. Then there exists a probability measure p supported on $\text{ext}(K)$ such that, for all affine functions f on K,*

$$f(c) = \int_{\text{ext}(K)} f(e)\,\mathrm{d}p(e).$$

The point c in Theorem 3.5 is called a *barycentre* of the set K, and the probability measure p is said to *represent* the point c. Informally speaking, the Kreĭn–Milman and Choquet–Bishop–de Leeuw theorems together ensure that a compact, convex subset K of a topologically respectable space is entirely characterised by its set of extreme points in the following sense: every point of K can be obtained as an average of extremal points of K, and, indeed, the value of any affine function at any point of K can be obtained as an average of its values at the extremal points in the same way.

Definition 3.6. Let $K \subseteq \mathcal{X}$ be convex. A function $f\colon K \to \mathbb{R} \cup \{\pm\infty\}$ is a *convex function* if, for all $x_0, x_1 \in K$ and $t \in [0, 1]$,

$$f(x_t) \leq (1 - t)f(x_0) + tf(x_1), \tag{3.1}$$

and is called a *strictly convex function* if, for all distinct $x_0, x_1 \in K$ and $t \in (0, 1)$,

$$f(x_t) < (1 - t)f(x_0) + tf(x_1).$$

[3.1] Or, more generally, Hausdorff, locally convex, topological vector spaces.

The inequality (3.1) defining convexity can be seen as a special case — with $X \sim \mu$ supported on two points x_0 and x_1 — of the following result:

Theorem 3.7 (Jensen). *Let $(\Theta, \mathscr{F}, \mu)$ be a probability space, let $K \subseteq \mathcal{X}$ and $f\colon K \to \mathbb{R} \cup \{\pm\infty\}$ be convex, and let $X \in L^1(\Theta, \mu; \mathcal{X})$ take values in K. Then*

$$f\big(\mathbb{E}_\mu[X]\big) \leq \mathbb{E}_\mu\big[f(X)\big], \tag{3.2}$$

where $\mathbb{E}_\mu[X] \in \mathcal{X}$ is defined by the relation $\langle \ell \,|\, \mathbb{E}_\mu[X] \rangle = \mathbb{E}_\mu[\langle \ell \,|\, X \rangle]$ for every $\ell \in \mathcal{X}'$. Furthermore, if f is strictly convex, then equality holds in (3.2) if and only if X is μ-almost surely constant.

It is straightforward to see that $f\colon K \to \mathbb{R} \cup \{\pm\infty\}$ is convex (resp. strictly convex) if and only if its *epigraph*

$$\operatorname{epi}(f) \coloneqq \{(x, v) \in K \times \mathbb{R} \mid v \geq f(x)\}$$

is a convex (resp. strictly convex) subset of $K \times \mathbb{R}$. Furthermore, twice-differentiable convex functions are easily characterised in terms of their second derivative (Hessian):

Theorem 3.8. *Let $f\colon K \to \mathbb{R}$ be twice continuously differentiable on an open, convex set K. Then f is convex if and only if $\mathrm{D}^2 f(x)$ is positive semi-definite for all $x \in K$. If $\mathrm{D}^2 f(x)$ is positive definite for all $x \in K$, then f is strictly convex, though the converse is false.*

Convex functions have many convenient properties with respect to minimization and maximization:

Theorem 3.9. *Let $f\colon K \to \mathbb{R}$ be a convex function on a convex set $K \subseteq \mathcal{X}$. Then*

(a) any local minimiser of f in K is also a global minimiser;
(b) the set $\arg\min_K f$ of global minimisers of f in K is convex;
(c) if f is strictly convex, then it has at most one global minimiser in K;
(d) if K is also compact, then f has the same maximum values on K and $\operatorname{ext}(K)$.

Proof. (a) Suppose that x_0 is a local minimiser of f in K that is not a global minimiser: that is, suppose that x_0 is a minimiser of f in some open neighbourhood N of x_0, and also that there exists $x_1 \in K \setminus N$ such that $f(x_1) < f(x_0)$. Then, for sufficiently small $t > 0$, $x_t \in N$, but convexity implies that

$$f(x_t) \leq (1-t)f(x_0) + tf(x_1) < (1-t)f(x_0) + tf(x_0) = f(x_0),$$

which contradicts the assumption that x_0 is a minimiser of f in N.

(b) Suppose that $x_0, x_1 \in K$ are global minimisers of f. Then, for all $t \in [0,1]$, $x_t \in K$ and

$$f(x_0) \leq f(x_t) \leq (1-t)f(x_0) + tf(x_1) = f(x_0).$$

Hence, $x_t \in \arg\min_K f$, and so $\arg\min_K f$ is convex.

(c) Suppose that $x_0, x_1 \in K$ are distinct global minimisers of f, and let $t \in (0,1)$. Then $x_t \in K$ and

$$f(x_0) \leq f(x_t) < (1-t)f(x_0) + tf(x_1) = f(x_0),$$

which is a contradiction. Hence, f has at most one minimiser in K.

(d) Suppose that $c \in K \setminus \operatorname{ext}(K)$ has $f(c) > \sup_{\operatorname{ext}(K)} f$. By Theorem 3.5, there exists a probability measure p on $\operatorname{ext}(K)$ such that, for all affine functions ℓ on K,

$$\ell(c) = \int_{\operatorname{ext}(K)} \ell(x) \,\mathrm{d}p(x).$$

i.e. $c = \mathbb{E}_{X\sim p}[X]$. Then Jensen's inequality implies that

$$\mathbb{E}_{X\sim p}\big[f(X)\big] \geq f(c) > \sup_{\mathrm{ext}(K)} f,$$

which is a contradiction. Hence, since $\sup_K f \geq \sup_{\mathrm{ext}(K)} f$, f must have the same maximum value on $\mathrm{ext}(K)$ as it does on K. ■

Remark 3.10. Note well that Theorem 3.9 does not assert the existence of minimisers, which requires non-emptiness and compactness of K, and lower semicontinuity of f. For example:

- the exponential function on $\mathbb{R}$ is strictly convex, continuous and bounded below by 0 yet has no minimiser;
- the interval $[-1, 1]$ is compact, and the function $f\colon [-1,1] \to \mathbb{R} \cup \{\pm\infty\}$ defined by

$$f(x) \coloneqq \begin{cases} x, & \text{if } |x| < \frac{1}{2}, \\ +\infty, & \text{if } |x| \geq \frac{1}{2}, \end{cases}$$

is convex, yet f has no minimiser — although $\inf_{x\in[-1,1]} f(x) = -\frac{1}{2}$, there is no x for which $f(x)$ attains this infimal value.

Definition 3.11. A *convex optimization problem* (or *convex program*) is a minimization problem in which the objective function and all constraints are equalities or inequalities with respect to convex functions.

Remark 3.12. (a) Beware of the common pitfall of saying that a convex program is simply the minimization of a convex function over a convex set. Of course, by Theorem 3.9, such minimization problems are nicer than general minimization problems, but bona fide convex programs are an even nicer special case.

(b) In practice, many problems are not obviously convex programs, but can be transformed into convex programs by e.g. a cunning change of variables. Being able to spot the right equivalent problem is a major part of the art of optimization.

It is difficult to overstate the importance of convexity in making optimization problems tractable. Indeed, it has been remarked that lack of convexity is a much greater obstacle to tractability than high dimension. There are many powerful methods for the solution of convex programs, with corresponding standard software libraries such as `cvxopt`. For example, *interior point methods* explore the interior of the feasible set in search of the solution to the convex program, while being kept away from the boundary of the feasible set by a *barrier function*. The discussion that follows is only intended as an outline; for details, see Boyd and Vandenberghe (2004, Chapter 11).

Consider the convex program

$$\begin{aligned} \text{minimise: } & f(x) \\ \text{with respect to: } & x \in \mathbb{R}^n \\ \text{subject to: } & c_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m, \end{aligned}$$

where the functions $f, c_1, \ldots, c_m\colon \mathbb{R}^n \to \mathbb{R}$ are all convex and differentiable. Let F denote the feasible set for this program. Let $0 < \mu \ll 1$ be a small scalar, called the *barrier parameter*, and define the *barrier function* associated to the program by

$$B(x;\mu) \coloneqq f(x) - \mu \sum_{i=1}^{m} \log c_i(x).$$

Note that $B(\,\cdot\,;\mu)$ is strictly convex for $\mu > 0$, that $B(x;\mu) \to +\infty$ as $x \to \partial F$, and that $B(\,\cdot\,;0) = f$; therefore, the unique minimiser x^*_μ of $B(\,\cdot\,;\mu)$ lies in $\mathring{F}$ and (hopefully) converges

to the minimiser of the original problem as $\mu \to 0$. Indeed, using arguments based on convex duality, one can show that

$$f(x_\mu^*) - \inf_{x \in F} f(x) \leq m\mu.$$

The strictly convex problem of minimizing $B(\cdot\,; \mu)$ can be solved approximately using Newton's method. In fact, however, one settles for a partial minimization of $B(\cdot\,; \mu)$ using only one or two steps of Newton's method, then decreases μ to μ', performs another partial minimization of $B(\cdot\,; \mu')$ using Newton's method, and so on in this alternating fashion.

Theorem 3.9 has the following immediate corollary for the minimization and maximization of affine functions on convex sets:

Corollary 3.13. *Let $\ell \colon K \to \mathbb{R}$ be a continuous affine function on a non-empty, compact, convex set $K \subseteq \mathcal{X}$. Then*

$$\operatorname{ext}\{\ell(x) \mid x \in K\} = \operatorname{ext}\{\ell(x) \mid x \in \operatorname{ext}(K)\}.$$

That is, ℓ has the same minimum and maximum values over both K and the set of extreme points of K.

Definition 3.14. A *linear program* is an optimization problem of the form

$$\begin{aligned} \text{extremise: } & f(x) \\ \text{with respect to: } & x \in \mathbb{R}^p \\ \text{subject to: } & g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, q, \end{aligned}$$

where the functions $f, g_1, \ldots, g_q \colon \mathbb{R}^p \to \mathbb{R}$ are all affine functions. Linear programs are often written in the *canonical form*

$$\begin{aligned} \text{maximise: } & c \cdot x \\ \text{with respect to: } & x \in \mathbb{R}^n \\ \text{subject to: } & Ax \leq b \\ & x \geq 0, \end{aligned}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are given, and the two inequalities are interpreted componentwise. (Conversion to canonical form, and in particular the introduction of the non-negativity constraint $x \geq 0$, is accomplished by augmenting the original $x \in \mathbb{R}^p$ with additional variables called *slack variables* to form the extended variable $x \in \mathbb{R}^n$.)

Note that the feasible set for a linear program is an intersection of finitely many half-spaces of $\mathbb{R}^n$, i.e. a *polytope*. This polytope may be empty, in which case the constraints are mutually contradictory and the program is said to be *infeasible*. Also, the polytope may be unbounded in the direction of c, in which case the extreme value of the problem is infinite.

We finish this section with some terminology concerning constraints:

Definition 3.15. For a given constrained optimization problem, a constraint is said to be

(a) *redundant* if it does not change the feasible set, and *non-redundant* or *relevant* otherwise;
(b) *non-binding* if it does not change the extreme value, and *binding* otherwise;
(c) *active* if it is an inequality constraint that holds as an equality at the extremiser, and *inactive* otherwise.

Example 3.16. Consider $f \colon \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) \coloneqq y$. Suppose that we wish to minimize f over the unbounded w-shaped region

$$W \coloneqq \{(x, y) \in \mathbb{R}^2 \mid y \geq (x^2 - 1)^2\}.$$

Over W, f takes the minimum value 0 at $(x, y) = (\pm 1, 0)$. Note that the inequality constraint $y \geq (x^2 - 1)^2$ is an active constraint. The additional constraint $y \geq 0$ would be redundant with respect to this feasible set W, and hence also non-binding. The additional constraint $x > 0$ would be non-redundant, but also non-binding, since it excludes the previous minimiser at $(x, y) = (-1, 0)$ but not the one at $(x, y) = (1, 0)$. Similarly, the additional equality constraint $y = (x^2 - 1)^2$ would be non-redundant and non-binding.

The importance of these concepts for UQ lies in the fact that many UQ problems are, in part or in whole, optimization problems: a good example is the calibration of parameters in a model in order to best explain some observed data. Each piece of information about the problem (e.g. a hypothesis about the form of the model, such as a physical law) can be seen as a constraint on that optimization problem. It is easy to imagine that each additional constraint may introduce additional difficulties in computing the parameters of best fit. Therefore, it is natural to want to exclude from consideration those constraints (pieces of information) that are merely complicating the solution process, and not actually determining the optimal parameters, and to have some terminology for describing the various ways in which this can occur.

4 Motivation and Notation for Distributional Robustness

To begin with, we will suppress all reference to uncertain response functions and focus only on uncertain probability measures. The reasons for doing so will become clearer later, but in essence handling the measures first will enable huge reductions in the complexity of the response function problem.

Suppose that we are interested in the value $Q(\mu^\dagger)$ of some *quantity of interest* that is a functional of a partially known probability measure $\mu^\dagger$ on a space $\mathcal{X}$. (Here we use the common notation of having daggers — † — denote the 'truth'.) Very often, $Q(\mu^\dagger)$ arises as the expected value with respect to $\mu^\dagger$ of some function $q \colon \mathcal{X} \to \mathbb{R}$, so the objective is to determine

$$Q(\mu^\dagger) \equiv \mathbb{E}_{X \sim \mu^\dagger}[q(X)].$$

Now suppose that $\mu^\dagger$ is known only to lie in some subset $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$. How should we try to understand or approximate $Q(\mu^\dagger)$?

In the absence of any further information about which $\mu \in \mathcal{A}$ are more or less likely to be $\mu^\dagger$, and particular if the consequences of planning based on an inaccurate estimate of $Q(\mu^\dagger)$ are very high, it makes sense to adopt a posture of 'healthy conservatism' and compute bounds on $Q(\mu^\dagger)$ that are as tight as justified by the information that $\mu^\dagger \in \mathcal{A}$, but no tighter, i.e. to find

$$\underline{Q}(\mathcal{A}) \coloneqq \inf_{\mu \in \mathcal{A}} Q(\mu) \text{ and } \overline{Q}(\mathcal{A}) \coloneqq \sup_{\mu \in \mathcal{A}} Q(\mu).$$

When $Q(\mu)$ is the expected value with respect to μ of some function $q \colon \mathcal{X} \to \mathbb{R}$, the objective is to determine

$$\underline{Q}(\mathcal{A}) \coloneqq \inf_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q] \text{ and } \overline{Q}(\mathcal{A}) \coloneqq \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q].$$

The inequality

$$\underline{Q}(\mathcal{A}) \leq Q(\mu^\dagger) \leq \overline{Q}(\mathcal{A})$$

is, by construction, the sharpest possible bound on $Q(\mu^\dagger)$ given only information that $\mu^\dagger \in \mathcal{A}$: any wider inequality would be unnecessarily pessimistic, with one of its bounds not attained; any narrower inequality would ignore some feasible scenario $\mu \in \mathcal{A}$ that could be $\mu^\dagger$. The obvious question is, can $\underline{Q}(\mathcal{A})$ and $\overline{Q}(\mathcal{A})$ be computed?

Naturally, the answer to this question depends upon the form of the admissible set $\mathcal{A}$. These notes focus upon admissible sets $\mathcal{A}$ of a particular but very accessible type, those

specified by equality or inequality constraints on expected values of test functions, otherwise known as *generalised moment classes*.

Example 4.1. As an example of this paradigm, suppose that it is desired to give bounds on the quality of some output $Y = g(X)$ of a manufacturing process in which the probability distribution of the inputs X is partially known. For example, quality control procedures may prescribe upper and lower bounds on the cumulative distribution function of X, but not the exact CDF of X, e.g.

$$\begin{aligned} 0 \leq \mathbb{P}_{X\sim\mu^\dagger}[-\infty < X \leq a] &\leq 0.1 \\ 0.8 \leq \mathbb{P}_{X\sim\mu^\dagger}[a < X \leq b] &\leq 1.0 \\ 0 \leq \mathbb{P}_{X\sim\mu^\dagger}[b < X \leq \infty] &\leq 0.1. \end{aligned}$$

Let $\mathcal{A}$ denote the (infinite-dimensional) set of all probability measures μ on $\mathbb{R}$ that are consistent with these three inequality constraints. Given the input-to-output map f, what are optimal bounds on the cumulative distribution function of Y, i.e., for $t \in \mathbb{R}$, what are

$$\inf_{\mu\in\mathcal{A}} \mathbb{P}_{X\sim\mu}[f(X) \leq t] \text{ and } \sup_{\mu\in\mathcal{A}} \mathbb{P}_{X\sim\mu}[f(X) \leq t]?. \tag{4.1}$$

The results of this section will show that these extremal values can be found by solving an optimization problem involving at most eight optimization variables, namely four possible values $x_0, \ldots, x_3 \in \mathbb{R}$ for X, and the four corresponding probability masses $w_0, \ldots, w_3 \geq 0$ that sum to unity. More precisely, we minimise or maximise

$$\sum_{i=0}^{3} w_i \mathbb{I}[f(x_i) \leq t]$$

subject to the the constraints

$$\begin{aligned} 0 \leq \sum_{i=0}^{3} w_i \mathbb{I}[x_i \leq a] &\leq 0.1 \\ 0.8 \leq \sum_{i=0}^{3} w_i \mathbb{I}[a < x_i \leq b] &\leq 1.0 \\ 0 \leq \sum_{i=0}^{3} w_i \mathbb{I}[x_i > b] &\leq 0.1. \end{aligned}$$

In general, this problem is a non-convex global optimization problem that can only be solved approximately. However, for fixed positions $\{x_i\}_{i=0}^3$, the optimal weights $\{w_i\}_{i=0}^3$ can be determined quickly and accurately using the tools of linear programming. Thus, the problem (4.1) reduces to a nonlinear family of linear programs, parametrised by $\{x_i\}_{i=0}^3$.

5 Maximum Entropy Distributions

Suppose that we are interested in the value $Q(\mu^\dagger)$ of some *quantity of interest* that is a functional of a partially known probability measure $\mu^\dagger$ on a space $\mathcal{X}$. Very often, $Q(\mu^\dagger)$ arises as the expected value with respect to $\mu^\dagger$ of some function $q\colon \mathcal{X} \to \mathbb{R}$, so the objective is to determine

$$Q(\mu^\dagger) \equiv \mathbb{E}_{X\sim\mu^\dagger}[q(X)].$$

Now suppose that $\mu^\dagger$ is known only to lie in some subset $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$. How should we try to understand or approximate $Q(\mu^\dagger)$? One approach is the following *MaxEnt Principle*:

Definition 5.1. The *Principle of Maximum Entropy* states that if all one knows about a probability measure μ is that it lies in some set $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$, then one should take μ to be the element $\mu^{\mathrm{ME}} \in \mathcal{A}$ of maximum entropy.

There are many heuristics underlying the MaxEnt Principle, including appeals to equilibrium thermodynamics and attractive derivations due to Wallis and Jaynes (2003). If entropy is understood as being a measure of uninformativeness, then the MaxEnt Principle can be seen as an attempt to avoid bias by selecting the 'least biased' or 'most uninformative' distribution.

Example 5.2 (Unconstrained maximum entropy distributions)**.** If $\mathcal{X} = \{1, \dots, m\}$ and $p \in \mathbb{R}^m_{>0}$ is a probability measure on $\mathcal{X}$, then the entropy of p is

$$H(p) \coloneqq -\sum_{i=1}^{m} p_i \log p_i. \tag{5.1}$$

The only constraints on p are the natural ones that $p_i \geq 0$ and that $S(p) \coloneqq \sum_{i=1}^m p_i = 1$. Temporarily neglect the inequality constraints and use the method of Lagrange multipliers to find the extrema of $H(p)$ among all $p \in \mathbb{R}^m$ with $S(p) = 1$; such p must satisfy, for some $\lambda \in \mathbb{R}$,

$$0 = \nabla H(p) - \lambda \nabla S(p) = -\begin{bmatrix} 1 + \log p_1 + \lambda \\ \vdots \\ 1 + \log p_m + \lambda \end{bmatrix}.$$

It is clear that any solution to this equation must have $p_1 = \cdots = p_m$, for if p_i and p_j differ, then at most one of $1 + \log p_i + \lambda$ and $1 + \log p_j + \lambda$ can equal 0 for the same value of λ. Therefore, since $S(p) = 1$, it follows that the unique extremiser of $H(p)$ among $\{p \in \mathbb{R}^m \mid S(p) = 1\}$ is $p_1 = \cdots = p_m = \frac{1}{m}$. The inequality constraints that were neglected initially are satisfied, and are not active constraints, so it follows that the uniform probability measure on $\mathcal{X}$ is the unique maximum entropy distribution on $\mathcal{X}$.

A similar argument using the calculus of variations shows that the unique maximum entropy probability distribution on an interval $[a, b] \subsetneq \mathbb{R}$ is the uniform distribution $\frac{1}{|b-a|}\,\mathrm{d}x$.

Example 5.3 (Constrained maximum entropy distributions)**.** Consider the set of all probability measures μ on $\mathbb{R}$ that have mean m and variance s^2; what is the maximum entropy distribution in this set? Consider probability measures μ that are absolutely continuous with respect to Lebesgue measure, having density ρ. Then the aim is to find μ to maximise

$$H(\rho) = -\int_{\mathbb{R}} \rho(x) \log \rho(x)\,\mathrm{d}x,$$

subject to the constraints that $\rho \geq 0$, $\int_{\mathbb{R}} \rho(x)\,\mathrm{d}x = 1$, $\int_{\mathbb{R}} x\rho(x)\,\mathrm{d}x = m$ and $\int_{\mathbb{R}}(x - m)^2\rho(x)\,\mathrm{d}x = s^2$. Introduce Lagrange multipliers $c = (c_0, c_1, c_2)$ and the Lagrangian

$$F_c(\rho) \coloneqq H(\rho) + c_0 \int_{\mathbb{R}} \rho(x)\,\mathrm{d}x + c_1 \int_{\mathbb{R}} x\rho(x)\,\mathrm{d}x + c_2 \int_{\mathbb{R}} (x - m)^2 \rho(x)\,\mathrm{d}x.$$

Consider a perturbation $\rho + t\sigma$; if ρ is indeed a critical point of F_c, then, regardless of σ, it must be true that

$$\frac{\mathrm{d}}{\mathrm{d}t} F_c(\rho + t\sigma)\bigg|_{t=0} = 0.$$

This derivative is given by

$$\frac{\mathrm{d}}{\mathrm{d}t} F_c(\rho + t\sigma)\bigg|_{t=0} = \int_{\mathbb{R}} \sigma(x)\big[-\log \rho(x) - 1 + c_0 + c_1 x + c_2 (x - m)^2\big]\,\mathrm{d}x.$$

Since it is required that $\frac{\mathrm{d}}{\mathrm{d}t} F_c(\rho + t\sigma)\big|_{t=0} = 0$ for every σ, the expression in the brackets must vanish, i.e.

$$\rho(x) = \exp(-c_0 + 1 - c_1 x - c_2 (x - m)^2).$$

Since $\rho(x)$ is the exponential of a quadratic form in x, μ must be a Gaussian of some mean and variance, which, by hypothesis, are m and s^2 respectively, i.e.

$$\begin{aligned} c_0 &= 1 - \log\bigl(1/\sqrt{2\pi s^2}\bigr), \\ c_1 &= 0, \\ c_2 &= \tfrac{1}{2s^2}. \end{aligned}$$

Thus, the maximum entropy distribution on $\mathbb{R}$ of with mean m and variance s^2 is $\mathcal{N}(m, s^2)$, with entropy

$$H(\mathcal{N}(m, s^2)) = \frac{1}{2} \log(2\pi e s^2).$$

Discrete Entropy and Convex Programming. In discrete settings, the entropy of a probability measure $p \in \mathcal{M}_1(\{1, \ldots, m\})$ with respect to the uniform measure as defined in (5.1) is a strictly convex function of $p \in \mathbb{R}^m_{>0}$. Therefore, when p is constrained by a family of convex constraints, finding the maximum entropy distribution is a convex program:

$$\begin{aligned} \text{minimise: } & \sum_{i=1}^{m} p_i \log p_i \\ \text{with respect to: } & p \in \mathbb{R}^m \\ \text{subject to: } & p \geq 0 \\ & p \cdot 1 = 1 \\ & \varphi_i(p) \leq 0 \quad \text{for } i = 1, \ldots, n, \end{aligned}$$

for given convex functions $\varphi_1, \ldots, \varphi_n \colon \mathbb{R}^m \to \mathbb{R}$. This is useful because an explicit formula for the maximum entropy distribution, such as in Example 5.3, is rarely available. Therefore, the possibility of efficiently computing the maximum entropy distribution, as in this convex programming situation, is very attractive.

Exercise 5.4. Suppose that a six-sided die (with the six sides bearing 1 to 6 spots) has been tossed $N \gg 1$ times and that the sample average number of spots is 4.5, rather than 3.5 as one would usually expect. Assume that this sample average is, in fact, the true average.

(a) What, according to the Principle of Maximum Entropy, is the correct probability distribution on the six sides of the die given this information?
(b) What are the optimal lower and upper probabilities of each of the 6 sides of the die given this information?

Remark 5.5. Note well that not all classes of probability measures contain maximum entropy distributions:

(a) The class of all absolutely continuous $\mu \in \mathcal{M}_1(\mathbb{R})$ with mean 0 but arbitrary variance contains distributions of arbitrarily large entropy.
(b) The class of all absolutely continuous $\mu \in \mathcal{M}_1(\mathbb{R})$ with mean 0 and second and third moments equal to 1 has all entropies bounded above but there is no distribution which attains the maximal entropy.

Remark 5.6. There are some philosophical, mathematical, and practical objections to the use of the Principle of Maximum Entropy:

(a) The MaxEnt Principle is an application-blind selection mechanism. It asserts that the correct course of action when faced with a collection $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ and an unknown $\mu^\dagger \in \mathcal{A}$ is to select a *single* representative $\mu^{\mathrm{ME}} \in \mathcal{A}$ and to make the approximation

$Q(\mu^\dagger) \approx Q(\mu^{\mathrm{ME}})$ regardless of what Q is. This is in contrast to hierarchical and optimization-based methods later in this chapter. Furthermore, MaxEnt distributions are typically 'nice' (exponentially small tails etc.), whereas many practical problems with high consequences involve heavy-tailed distributions.

(b) Recalling that in fact all entropies are *relative* entropies (Kullback–Leibler divergences), the result of applying the MaxEnt Principle is dependent upon the reference measure chosen, and many complex systems do not admit a uniform measure for use as a reference measure. Thus, the MaxEnt Principle would appear to depend upon an ad hoc choice of reference measure.

(c) MaxEnt distributions are almost atypically smooth and light-tailed, as the next exercise illustrates, whereas many important applications involve distributions that have heavy tails.

Exercise 5.7. Suppose $\mathcal{X} \subseteq \mathbb{R}$ is closed, and we seek a maximum entropy distribution subject to N constraints of the form $\mathbb{E}_{X\sim\mu}[\varphi_n(X)] = c_n$, for $n = 1, \ldots, N$, where the φ_n are known measurable functions and the c_n are known real constants. Using the Lagrange multiplier theorem, show that, if such a MaxEnt distribution exists and has positive probability density function ρ in $\mathcal{X}$, the ρ is given by

$$\rho(x) = \frac{1}{Z} \exp\left(\sum_{n=1}^{N} \lambda_n \varphi_n(x)\right),$$

where $Z > 0$ and the $\lambda_n \in \mathbb{R}$ are constants to be determined. Thus — if the φ_n are, say, smooth and bounded — MaxEnt distributions are smooth with exponentially small tails at infinity.

Exercise 5.8. Let $\mathcal{P}^k$ denote the set of probability measures μ on $\mathbb{R}$ with finite moments up to order $k \geq 0$, i.e.

$$\mathcal{P}^k := \left\{ \mu \in \mathcal{M}_1(\mathbb{R}) \,\middle|\, \int_{\mathbb{R}} x^k \,\mathrm{d}\mu(x) < \infty \right\}.$$

Show that $\mathcal{P}^k$ is a 'small' subset of $\mathcal{P}^\ell$ whenever $k > \ell$ in the sense that, for every $\mu \in \mathcal{P}^k$ and every $\varepsilon > 0$, there exists $\nu \in \mathcal{P}^\ell \setminus \mathcal{P}^k$ with $d_{\mathrm{TV}}(\mu, \nu) < \varepsilon$. Hint: follow the example of the Cauchy distribution

$$\rho_{\mathrm{Cauchy}}(x) = \frac{1}{\pi} \frac{1}{1 + x^2},$$

which only has finite moments of order strictly less than 1, to construct a 'standard' probability measure with polynomial moments of order ℓ and no higher, and consider convex combinations of this 'standard' measure with μ.

6 Distributional Robustness

As before, suppose that we are interested in the value $Q(\mu^\dagger)$ of some *quantity of interest* that is a functional of a partially-known probability measure $\mu^\dagger$ on a space $\mathcal{X}$, and that $\mu^\dagger$ is known only to lie in some subset $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$. In the absence of any further information about which $\mu \in \mathcal{A}$ are more or less likely to be $\mu^\dagger$, and particular if the consequences of planning based on an inaccurate estimate of $Q(\mu^\dagger)$ are very high, it makes sense to adopt a posture of 'healthy conservatism' and compute bounds on $Q(\mu^\dagger)$ that are as tight as justified by the information that $\mu^\dagger \in \mathcal{A}$, but no tighter, i.e. to find

$$\underline{Q}(\mathcal{A}) := \inf_{\mu\in\mathcal{A}} Q(\mu) \text{ and } \overline{Q}(\mathcal{A}) := \sup_{\mu\in\mathcal{A}} Q(\mu).$$

As discussed earlier, the inequality

$$\underline{Q}(\mathcal{A}) \leq Q(\mu^\dagger) \leq \overline{Q}(\mathcal{A})$$

is the sharpest possible bound on $Q(\mu^\dagger)$ given only information that $\mu^\dagger \in \mathcal{A}$. However, can $\underline{Q}(\mathcal{A})$ and $\overline{Q}(\mathcal{A})$ be computed?

Finite Sample Spaces. Suppose that the sample space $\mathcal{X} = \{1, \ldots, K\}$ is a finite set equipped with the discrete topology. Then the space of measurable functions $f\colon \mathcal{X} \to \mathbb{R}$ is isomorphic to $\mathbb{R}^K$ and the space of probability measures μ on $\mathcal{X}$ is isomorphic to the unit simplex in $\mathbb{R}^K$; integrating f against μ is simply taking the Euclidean dot product of the two K-vector representations. If the available information on $\mu^\dagger$ is that it lies in the set

$$\mathcal{A} \coloneqq \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_n] \leq c_n \text{ for } n = 1, \ldots, N\}$$

for known measurable functions $\varphi_1, \ldots, \varphi_N\colon \mathcal{X} \to \mathbb{R}$ and values $c_1, \ldots, c_N \in \mathbb{R}$, then the problem of finding the extreme values of $\mathbb{E}_\mu[q]$ among $\mu \in \mathcal{A}$ reduces to linear programming:

$$\begin{aligned} \text{extremise: } & p \cdot q \\ \text{with respect to: } & p \in \mathbb{R}^K \\ \text{subject to: } & p \geq 0 \\ & p \cdot 1 = 1 \\ & p \cdot \varphi_n \leq c_n \text{ for } n = 1, \ldots, N. \end{aligned}$$

Note that the feasible set $\mathcal{A}$ for this problem is a convex subset of $\mathbb{R}^K$; indeed, $\mathcal{A}$ is a *polytope*, i.e. the intersection of finitely many closed half-spaces of $\mathbb{R}^K$. Furthermore, as a closed subset of the probability simplex in $\mathbb{R}^K$, $\mathcal{A}$ is compact. Therefore, by Corollary 3.13, the extreme values of this problem are certain to be found in the extremal set $\operatorname{ext}(\mathcal{A})$. This insight can be exploited to great effect in the study of distributional robustness problems for general sample spaces $\mathcal{X}$.

Remarkably, when the feasible set $\mathcal{A}$ of probability measures is sufficiently like a polytope, it is not necessary to consider finite sample spaces. What would appear to be an intractable optimization problem over an infinite-dimensional set of measures is in fact equivalent to a tractable finite-dimensional problem. Thus, the aim of this section is to find a finite-dimensional subset $\mathcal{A}_\Delta$ of $\mathcal{A}$ with the property that

$$\operatorname*{ext}_{\mu \in \mathcal{A}} Q(\mu) = \operatorname*{ext}_{\mu \in \mathcal{A}_\Delta} Q(\mu).$$

To perform this reduction, it is necessary to restrict attention to probability measures, topological spaces, and functionals that are sufficiently well-behaved.

Extreme Points of Moment Classes. The first step in this reduction is to classify the extremal measures in sets of probability measures that are prescribed by inequality or equality constraints on the expected value of finitely many arbitrary measurable test functions, so-called *moment classes*. Since, in finite time, we can only verify — even approximately, numerically — the truth of finitely many inequalities, such moment classes are appealing feasible sets from an epistemological point of view because they conform to the dictum of Karl Popper (1963) that "Our knowledge can be only finite, while our ignorance must necessarily be infinite."

Definition 6.1. A Borel measure μ on a topological space $\mathcal{X}$ is called *inner regular* if, for every Borel-measurable set $E \subseteq \mathcal{X}$,

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

A *pseudo-Radon space* is a topological space on which every Borel probability measure is inner regular. A *Radon space* is a separable, metrisable, pseudo-Radon space.

Example 6.2. (a) Lebesgue measure (n-dimensional volume) on Euclidean space $\mathbb{R}^n$ (restricted to the Borel σ-algebra $\mathscr{B}(\mathbb{R}^n)$, if pedantry is the order of the day) is an inner regular measure. Similarly, Gaussian measure is an inner regular probability measure on $\mathbb{R}^n$.

(b) Indeed, every Polish space (i.e. every separable and completely metrisable topological space) is a pseudo-Radon space. Thus, almost all of the spaces that one meets in 'practical' discussions — compact rectangular boxes in $\mathbb{R}^n$, the whole of $\mathbb{R}^n$, separable Banach and Hilbert spaces of functions — are suitable for the UQ theory that we are building here.

(c) However, there are some special cases where the inner regularity assumptions fail. For example, Lebesgue/Gaussian measures on $\mathbb{R}$ equipped with the topology of one-sided convergence are not inner regular measures: see Exercise 6.3 below if you are interested in the details.

Exercise 6.3. Consider the topology $\mathcal{T}$ on $\mathbb{R}$ generated by the basis of open sets $[a, b)$, where $a, b \in \mathbb{R}$.

1. Show that this topology generates the same σ-algebra on $\mathbb{R}$ as the usual Euclidean topology does. Hence, show that Gaussian measure is a well-defined probability measure on the Borel σ-algebra of $(\mathbb{R}, \mathcal{T})$.
2. Show that every compact subset of $(\mathbb{R}, \mathcal{T})$ is a countable set.
3. Conclude that Gaussian measure on $(\mathbb{R}, \mathcal{T})$ is not inner regular and that $(\mathbb{R}, \mathcal{T})$ is not a pseudo-Radon space.

Compare the following definition of a barycentre (a centre of mass) for a set of probability measures with the conclusion of the Choquet–Bishop–de Leeuw theorem (Theorem 3.5):

Definition 6.4. A *barycentre* for a set $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ is a probability measure $\mu \in \mathcal{M}_1(\mathcal{X})$ such that there exists $p \in \mathcal{M}_1(\mathrm{ext}(\mathcal{A}))$ such that

$$\mu(B) = \int_{\mathrm{ext}(\mathcal{A})} \nu(B) \, \mathrm{d}p(\nu) \quad \text{for all measurable } B \subseteq \mathcal{X}. \tag{6.1}$$

The measure p is said to *represent* the barycentre μ.

Recall that a d-dimensional simplex is the closed convex hull of $d + 1$ points $p_0, \ldots, d_d$ such that $p_1 - p_0, \ldots, p_d - p_0$ are linearly independent. The next ingredient in the analysis of distributional robustness is an appropriate infinite-dimensional generalization of the notion of a simplex — a *Choquet simplex* — as a subset of the vector space of signed measures on a given measurable space. One way to define Choquet simplices is through orderings and cones on vector spaces, but this definition can be somewhat cumbersome. Instead, the following geometrical description of Choquet simplices, illustrated in Figure 6.1, is much more amenable to visual intuition, and more easily checked in practice:

Definition 6.5. A *homothety* of a real topological vector space $\mathcal{V}$ is the composition of a positive dilation with a translation, i.e. a function $f \colon \mathcal{V} \to \mathcal{V}$ of the form $f(x) = \alpha x + v$, for fixed $\alpha > 0$ and $v \in \mathcal{V}$.

Theorem 6.6 (Choquet–Kendall). *A convex subset S of a topological vector space $\mathcal{V}$ is a Choquet simplex if and only if the intersection of any two homothetic images of S is empty, a single point, or another homothetic image of S.*

With these definitions, the extreme points of moment sets of probability measures can be described by the following theorem:

Theorem 6.7 (Winkler, 1988). *Let $(\mathcal{X}, \mathscr{F})$ be a measurable space and let $S \subseteq \mathcal{M}_1(\mathscr{F})$ be a Choquet simplex such that* $\mathrm{ext}(S)$ *consists of Dirac measures. Fix measurable functions $\varphi_1, \ldots, \varphi_n \colon \mathcal{X} \to \mathbb{R}$ and $c_1, \ldots, c_n \in \mathbb{R}$ and let*

$$\mathcal{A} := \left\{ \mu \in S \;\middle|\; \begin{array}{c} \text{for } i = 1, \ldots, n, \\ \varphi_i \in L^1(\mathcal{X}, \mu) \text{ and } \mathbb{E}_\mu[\varphi_i] \leq c_i \end{array} \right\}.$$

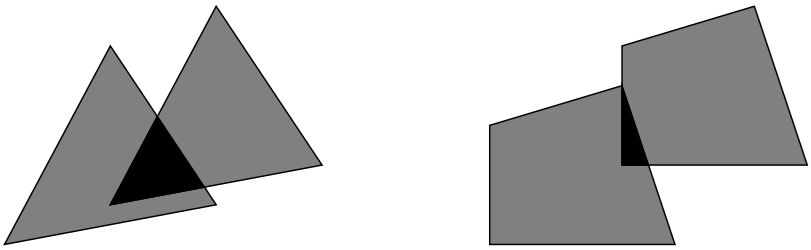


Figure 6.1: By the Choquet–Kendall theorem (Theorem 6.6), like finite-dimensional simplices, Choquet simplices S in a vector space $\mathcal{V}$ are characterised by the property that the intersection of any two homothetic images of S, $(\alpha_1 S + v_1) \cap (\alpha_1 S + v_2)$, with $\alpha_1, \alpha_2 > 0$ and $v_1, v_2 \in \mathcal{V}$, is either empty, a single point, or another homothetic image of S. This property holds for the simplex (a), but not for the non-simplicial convex set (b).

Then $\mathcal{A}$ is convex and its extremal set satisfies

$$\operatorname{ext}(\mathcal{A}) \subseteq \mathcal{A}_\Delta \coloneqq \left\{ \mu \in \mathcal{A} \;\middle|\; \begin{array}{c} \mu = \sum_{i=1}^m w_i \delta_{x_i}, \\ 1 \leq m \leq n+1, \text{ and} \\ \text{the vectors } (\varphi_1(x_i), \ldots, \varphi_n(x_i), 1)_{i=1}^m \\ \text{are linearly independent} \end{array} \right\};$$

Furthermore, if all the moment conditions defining $\mathcal{A}$ are equalities $\mathbb{E}_\mu[\varphi_i] = c_i$ instead of inequalities $\mathbb{E}_\mu[\varphi_i] \leq c_i$, then $\operatorname{ext}(\mathcal{A}) = \mathcal{A}_\Delta$.

The proof of Winkler's theorem is rather technical, and is omitted. The important point for our purposes is that, when $\mathcal{X}$ is a pseudo-Radon space, Winkler's theorem applies with $S = \mathcal{M}_1(\mathcal{X})$, so $\operatorname{ext}(\mathcal{A}) \subseteq \mathcal{A} \cap \Delta_n(\mathcal{X})$, where

$$\Delta_N(\mathcal{X}) \coloneqq \left\{ \mu = \sum_{i=0}^N w_i \delta_{x_i} \in \mathcal{M}_1(\mathcal{X}) \;\middle|\; \begin{array}{c} w_0, \ldots, w_N \geq 0, \\ w_0 + \cdots + w_N = 1, \\ x_0, \ldots, x_N \in \mathcal{X} \end{array} \right\}$$

denotes the set of all convex combinations of at most $N+1$ unit Dirac measures on the space $\mathcal{X}$. Pictures like Figure 6.2 should make this an intuitively plausible claim, at least in the case that $\mathcal{X}$ is a finite set.

Optimization of Measure Affine Functionals. Having understood the extreme points of moment classes, the next step is to show that the optimization of suitably nice functionals on such classes can be exactly reduced to optimization over the extremal measures in the class.

Definition 6.8. For $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$, a function $F \colon \mathcal{A} \to \mathbb{R} \cup \{\pm\infty\}$ is said to be *measure affine* if, for all $\mu \in \mathcal{A}$ and $p \in \mathcal{M}_1(\operatorname{ext}(A))$ for which (6.1) holds, F is p-integrable with

$$F(\mu) = \int_{\operatorname{ext}(\mathcal{A})} F(\nu) \, \mathrm{d}p(\nu). \tag{6.2}$$

As always, the reader should check that the terminology 'measure affine' is a sensible choice by verifying that when $\mathcal{X} = \{1, \ldots, K\}$ is a finite sample space, the restriction of any affine function $F \colon \mathbb{R}^K \cong \mathcal{M}_\pm(\mathcal{X}) \to \mathbb{R}$ to a subset $A \subseteq \mathcal{M}_1(\mathcal{X})$ is a measure affine function in the sense of Definition 6.8.

An important and simple example of a measure affine functional is an evaluation functional, i.e. the integration of a fixed measurable function q:

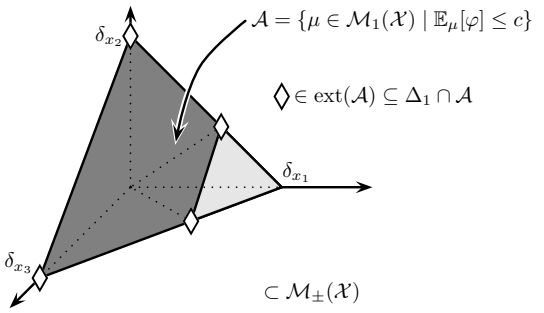


Figure 6.2: Heuristic justification of Winkler's classification of extreme points of moment sets (Theorem 6.7). Observe that the extreme points of the dark grey set $\mathcal{A}$ consist of convex combinations of at most 2 point masses, and $2 = 1 +$ the number of constraints defining $\mathcal{A}$.

Proposition 6.9. *If q is bounded either below or above, then $\mu \mapsto \mathbb{E}_\mu[q]$ is a measure affine map.*

Proof. First consider the case that $q = \mathbb{I}_E$ is the indicator function of a measurable set $E \subseteq \mathcal{X}$. Suppose that μ is a barycentre for $\mathcal{A}$ and that $p \in \mathcal{M}_1(\mathrm{ext}(\mathcal{A}))$ represents μ, i.e.

$$\mu(B) = \int_{\mathrm{ext}(\mathcal{A})} \nu(B) \, \mathrm{d}p(\nu) \quad \text{for all measurable } B \subseteq \mathcal{X}.$$

For $B = E$, this is the statement that

$$\mathbb{E}_\mu[\mathbb{I}_E] = \int_{\mathrm{ext}(\mathcal{A})} \mathbb{E}_\nu[\mathbb{I}_E] \, \mathrm{d}p(\nu),$$

which is (6.2). To complete the proof, verify the claim for q a linear combination of indicator functions, then for a sequence of such functions increasing to a function that is bounded above (resp. decreasing to a function that is bounded below), and apply the monotone class theorem — see Exercise 6.10. ■

Exercise 6.10. Complete the proof of Proposition 6.9: verify the claim for q a linear combination of indicator functions, then for a sequence of such functions increasing to a function that is bounded above (resp. decreasing to a function that is bounded below), and finish by applying the monotone class theorem.

Proposition 6.11. *Let $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ be convex and let F be a measure affine function on $\mathcal{A}$. Then F has the same extreme values on $\mathcal{A}$ and* $\mathrm{ext}(\mathcal{A})$.

Proof. Without loss of generality, consider the maximization problem; the proof for minimization is similar. Let $\mu \in \mathcal{A}$ be arbitrary and choose a probability measure $p \in \mathcal{M}_1(\mathrm{ext}(\mathcal{A}))$ with barycentre μ. Then, it follows from the barycentric formula (6.2) that

$$F(\mu) \leq \sup_{\nu \in \mathrm{supp}(p)} F(\nu) \leq \sup_{\nu \in \mathrm{ext}(\mathcal{A})} F(\nu). \tag{6.3}$$

First suppose that $\sup_{\mu \in \mathcal{A}} F(\mu)$ is finite. Necessarily, $\sup_{\nu \in \mathrm{ext}(\mathcal{A})} F(\nu)$ is also finite, but it remains to show that the two suprema are equal. Let $\varepsilon > 0$ be arbitrary. Let μ^* be $\frac{\varepsilon}{2}$-suboptimal for the problem of maximizing F over $\mathcal{A}$, i.e. $F(\mu^*) \geq \sup_{\mu \in \mathcal{A}} F(\mu) - \frac{\varepsilon}{2}$, and

let ν^* be $\frac{\varepsilon}{2}$-suboptimal for the problem of maximizing F over $\operatorname{ext}(\mathcal{A})$. Then

$$\begin{aligned} F(\nu^*) &\geq \sup_{\nu \in \operatorname{ext}(\mathcal{A})} F(\nu) - \frac{\varepsilon}{2} \\ &\geq F(\mu^*) - \frac{\varepsilon}{2} && \text{by (6.3) with } \mu = \mu^* \\ &\geq \sup_{\mu \in \mathcal{A}} F(\mu) - \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, $\sup_{\mu \in \mathcal{A}} F(\mu) = \sup_{\nu \in \operatorname{ext}(\mathcal{A})} F(\nu)$, and this proves the claim in this case.

In the case that $\sup_{\mu \in \mathcal{A}} F(\mu) = +\infty$, let $C, \varepsilon > 0$. Then there exists some $\mu^* \in \mathcal{A}$ such that $F(\mu^*) \geq C + \varepsilon$. Then, regardless of whether or not $\sup_{\nu \in \operatorname{ext}(\mathcal{A})} F(\nu)$ is finite, (6.3) with $\mu = \mu^*$ implies that there is some $\nu^* \in \operatorname{ext}(\mathcal{A})$ such that

$$F(\nu^*) \geq F(\mu^*) - \varepsilon \geq C + \varepsilon - \varepsilon = C.$$

However, since $C > 0$ was arbitrary, it follows that in fact $\sup_{\nu \in \operatorname{ext}(\mathcal{A})} F(\nu) = +\infty$, and this completes the proof. ■

In summary, we now have the following:

Theorem 6.12. *Let $\mathcal{X}$ be a pseudo-Radon space and let $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ be a moment class of the form*

$$\mathcal{A} \coloneqq \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_j] \leq 0 \text{ for } j = 1, \ldots, N\}$$

for prescribed measurable functions $\varphi_j \colon \mathcal{X} \to \mathbb{R}$. Then the extreme points of $\mathcal{A}$ are given by

$$\begin{aligned} \operatorname{ext}(\mathcal{A}) \subseteq \mathcal{A}_\Delta &\coloneqq \mathcal{A} \cap \Delta_N(\mathcal{X}) \\ &= \left\{ \mu \in \mathcal{M}_1(\mathcal{A}) \;\middle|\; \begin{array}{c} \text{for some } w_0, \ldots, w_N \in [0,1],\ x_0, \ldots, x_N \in \mathcal{X}, \\ \mu = \sum_{i=0}^N w_i \delta_{x_i} \\ \sum_{i=0}^N w_i = 1, \\ \text{and } \sum_{i=0}^N w_i \varphi_j(x_i) \leq 0 \text{ for } j = 1, \ldots, N \end{array} \right\}. \end{aligned}$$

Hence, if q is bounded either below or above, then $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$ and $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$.

Proof. Winkler's theorem (Theorem 6.7) implies that $\operatorname{ext}(\mathcal{A}) \subseteq \mathcal{A}_\Delta$. Since q is bounded on at least one side, Proposition 6.9 implies that $\mu \mapsto F(\mu) \coloneqq \mathbb{E}_\mu[q]$ is measure affine. The claim then follows from Proposition 6.11. ■

Remark 6.13. (a) Theorem 6.12 is good news from a computational standpoint for two reasons:

(i) Since any feasible measure in $\mathcal{A}_\Delta$ is completely described by $N+1$ scalars and $N+1$ points of $\mathcal{X}$, the reduced set of feasible measures is a finite-dimensional object — or, at least, it is as finite-dimensional as the space $\mathcal{X}$ is — and so it can in principle be explored using the finite-dimensional numerical optimization techniques that can be implemented on a computer.

(ii) Furthermore, since the probability measures in $\mathcal{A}_\Delta$ are finite sums of Dirac measures, expectations against such measures can be performed exactly using finite sums — there is no quadrature error.

(b) That said, when $\mu \in \mathcal{A}_\Delta$ has $\#\operatorname{supp}(\mu) \gg 1$, as may be the case with problems exhibiting independence structure like those considered below, it may be cheaper to integrate against a discrete measure $\mu = \sum_{i=0}^N \alpha_i \delta_{x_i} \in \mathcal{A}_\Delta$ in a Monte Carlo fashion, by drawing some number $1 \ll M \ll \#\operatorname{supp}(\mu)$ of independent samples from μ (i.e. x_i with probability α_i).

In general, the optimization problems over $\mathcal{A}_\Delta$ in Theorem 6.12 can only be solved approximately, using the tools of numerical global optimization. However, some of the classical inequalities of basic probability theory can be obtained in closed form by this approach.

Example 6.14 (Markov's inequality)**.** Suppose that X is a non-negative real-valued random variable with mean $\mathbb{E}[X] \leq m > 0$. Given $t \geq m$, what is the least upper bound on $\mathbb{P}[X \geq t]$?

To answer this question, observe that the given information says that the distribution $\mu^\dagger$ of X is some (and could be any!) element of $\mathcal{A}$, where

$$\mathcal{A} \coloneqq \left\{ \mu \in \mathcal{M}_1([0,\infty)) \,\middle|\, \mathbb{E}_{X\sim\mu}[X] \leq m \right\}.$$

This $\mathcal{A}$ is a moment class with a single moment constraint. By Theorem 6.12, the least upper bound on $\mathbb{P}_{X\sim\mu}[X \geq t]$ among $\mu \in \mathcal{A}$ can be found by restricting attention to the set $\mathcal{A}_\Delta$ of probability measures with support on at most two points $x_0, x_1 \in [0,\infty)$, with masses w_0, w_1 respectively.

Assume without loss of generality that the two point masses are located at x_0 and x_1 with $0 \leq x_0 \leq x_1 < \infty$. Now make a few observations:

(a) In order to satisfy the mean constraint that $\mathbb{E}[X] \leq m$, we must have $x_0 \leq m$.

(b) If $x_1 > t$ and the mean constraint is satisfied, then moving the mass w_1 at x_1 to $x_1' \coloneqq t$ does not decrease the objective function value $\mathbb{P}_{X\sim\mu}[X \geq t]$ and the mean constraint is still satisfied. Therefore, it is sufficient to consider two-point distributions with $x_1 = t$.

(c) By similar reasoning, it is sufficient to consider two-point distributions with $x_0 = 0$.

(d) Finally, suppose that $x_0 = 0$, $x_1 = t$, but that

$$\mathbb{E}_{X\sim\mu}[X] = w_0 x_0 + w_1 x_1 = w_1 t < m.$$

Then we may change the masses to

$$\begin{aligned} w_1' &\coloneqq m/t > w_1, \\ w_0' &\coloneqq 1 - m/t < w_0, \end{aligned}$$

keeping the positions fixed, thereby increasing the objective function value $\mathbb{P}_{X\sim\mu}[X \geq t]$ while still satisfying the mean constraint.

Putting together the above observations yields that

$$\sup_{\mu\in\mathcal{A}} \mathbb{P}_{X\sim\mu}[X \geq t] = \frac{m}{t},$$

with the maximum being attained by the two-point distribution

$$\left(1 - \frac{m}{t}\right)\delta_0 + \frac{m}{t}\delta_t.$$

This result is exactly Markov's inequality from basic probability theory.

Exercise 6.15 (Bounded random variables)**.** Calculate by hand, as a function of $t \in \mathbb{R}$, $D \geq 0$ and $m \in \mathbb{R}$,

$$\sup_{\mu\in\mathcal{A}} \mathbb{P}_{X\sim\mu}[X \leq t],$$

where

$$\mathcal{A} \coloneqq \left\{ \mu \in \mathcal{M}_1(\mathbb{R}) \,\middle|\, \begin{array}{l} \mathbb{E}_{X\sim\mu}[X] \geq m, \text{ and} \\ \operatorname{diam}(\operatorname{supp}(\mu)) \leq D \end{array} \right\}.$$

Physically, this exercise corresponds to the following challenge: you have 1 kg of idealised infinitely divisible sand, which you can arrange on a horizontal beam ($\mathbb{R}$) however you like, and must place as much sand as possible in the region $x \geq t$ while ensuring that the beam balances about a point $\geq m$ and all sand is contained in a region at most D in length. Hint: the answer is the same for

$$\mathcal{A}_= \coloneqq \left\{ \mu \in \mathcal{M}_1(\mathbb{R}) \,\middle|\, \begin{array}{l} \mathbb{E}_{X\sim\mu}[X] = m, \text{ and} \\ \operatorname{diam}(\operatorname{supp}(\mu)) \leq D \end{array} \right\}.$$

Exercise 6.16 (Chebyshev's inequality)**.** Calculate by hand, as a function of $t \in \mathbb{R}$, $s \geq 0$ and $m \in \mathbb{R}$,

$$\sup_{\mu \in \mathcal{A}} \mathbb{P}_{X \sim \mu}[X - m \geq st],$$

and

$$\sup_{\mu \in \mathcal{A}} \mathbb{P}_{X \sim \mu}[|X - m| \geq st],$$

where

$$\mathcal{A} \coloneqq \left\{ \mu \in \mathcal{M}_1(\mathbb{R}) \,\middle|\, \begin{array}{c} \mathbb{E}_{X \sim \mu}[X] \leq m, \text{ and} \\ \mathbb{E}_{X \sim \mu}[|X - m|^2] \leq s^2 \end{array} \right\}.$$

Hint: reduce the search space to a search over probability measures supported on at most three points in $\mathbb{R}$, and then model your reasoning on Example 6.14.

7 Independence

The kinds of constraints on measures (or, if you prefer, random variables) that can be considered in Theorem 6.12 include values for, or bounds on, functions of one or more of those random variables: e.g. the mean of X_1, the variance of X_2, the covariance of X_3 and X_4, and so on. However, one commonly encountered piece of information that is not of this type is that X_5 and X_6 are independent random variables, i.e. that their joint distibution is a product measure. The problem here is that sets of product measures can fail to be convex, so the reduction to extreme points cannot be applied directly.

For measures μ_1 on $\mathcal{X}_1$ and μ_2 on $\mathcal{X}_2$, $\mu_1 \otimes \mu_2$ denotes their product, which is the measure on $\mathcal{X}_1 \times \mathcal{X}_2$ defined by

$$(\mu_1 \otimes \mu_2)(E_1 \times E_2) \coloneqq \mu_1(E_1)\mu_2(E_2)$$

i.e the measure of a 'rectangle' is the product of the measures of its 'sides'. This formula is then extended to non-rectangular subsets of $\mathcal{X}_1 \times \mathcal{X}_2$ by σ-additivity. In this sense, 'area measure' is the product of 'length measure' with itself. As remarked in the previous paragraph, random variables X_1 and X_2 with marginal distributions μ_1 and μ_2 are independent exactly when the joint distribution of (X_1, X_2) is the product measure $\mu_1 \otimes \mu_2$.

Exercise 7.1. Let λ denote uniform measure on the unit interval $[0, 1] \subsetneq \mathbb{R}$. Show that the line segment in $\mathcal{M}_1([0, 1]^2)$ joining the measures $\lambda \otimes \delta_0$ and $\delta_0 \otimes \lambda$ contains measures that are not product measures. Hence show that a set $\mathcal{A}$ of product probability measures like that considered in Theorem 7.2 is typically not convex.

Fortunately, a cunning application of Fubini's theorem resolves this difficulty. Fubini's theorem is the result that ensures that integration (expectation) against a product measure can be performed as an iterated integral:

$$\begin{aligned} \mathbb{E}_{(X_1, X_2) \sim \mu_1 \otimes \mu_2}[f(X_1, X_2)] &= \mathbb{E}_{X_1 \sim \mu_1}\big[\mathbb{E}_{X_2 \sim \mu_2}[f(X_1, X_2)]\big] \\ &= \mathbb{E}_{X_2 \sim \mu_2}\big[\mathbb{E}_{X_1 \sim \mu_1}[f(X_1, X_2)]\big], \end{aligned}$$

at least for integrands $f \colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ that are measurable and bounded either below or above. Using Fubini's theorem, we can extend Theorem 6.12 to cope with independence constrains coupled with moment constraints on the marginal and joint distributions. Note well, though, that unlike Theorem 6.12, Theorem 7.2 does *not* say that $\mathcal{A}_\Delta = \operatorname{ext}(\mathcal{A})$; it only says that the optimization problem has the same extreme values over $\mathcal{A}_\Delta$ and $\mathcal{A}$.

Theorem 7.2. *Let $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ be a moment class of the form*

$$\mathcal{A} \coloneqq \left\{ \mu = \bigotimes_{k=1}^{K} \mu_k \in \bigotimes_{k=1}^{K} \mathcal{M}_1(\mathcal{X}_k) \,\middle|\, \begin{array}{c} \mathbb{E}_{\mu}[\varphi_j] \leq 0 \text{ for } j = 1, \ldots, N, \\ \mathbb{E}_{\mu_1}[\varphi_{1j}] \leq 0 \text{ for } j = 1, \ldots, N_1, \\ \vdots \\ \mathbb{E}_{\mu_K}[\varphi_{Kj}] \leq 0 \text{ for } j = 1, \ldots, N_K \end{array} \right\}$$

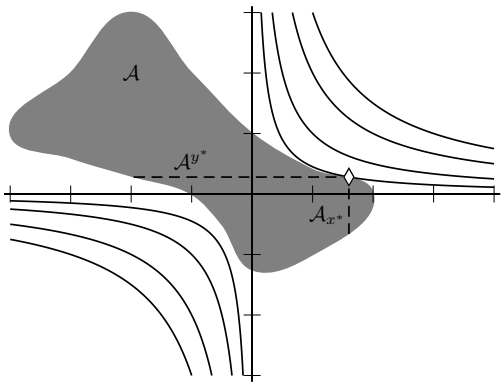


Figure 7.1: Optimization of a bilinear form B over a non-convex set $\mathcal{A} \subseteq \mathbb{R}^2$ that has convex cross-sections. The black curves show level sets of $B(x, y) = xy$. Note that the maximum value of B over $\mathcal{A}$ is found at a point (x^*, y^*) (marked with a diamond) such that x^* and y^* are both extreme points of the corresponding sections $\mathcal{A}^{y^*}$ and $\mathcal{A}_{x^*}$ respectively.

---

for prescribed measurable functions $\varphi_j \colon \mathcal{X} \to \mathbb{R}$ and $\varphi_{kj} \colon \mathcal{X} \to \mathbb{R}$. Let

$$\mathcal{A}_\Delta \coloneqq \{\mu \in \mathcal{A} \,|\, \mu_k \in \Delta_{N+N_k}(\mathcal{X}_k)\}\,.$$

Then, if q is bounded either above or below, $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$ and $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$.

Proof. Let $\varepsilon > 0$ and let $\mu^* \in \mathcal{A}$ be $\frac{\varepsilon}{K+1}$-suboptimal for the maximization of $\mu \mapsto \mathbb{E}_\mu[q]$ over $\mu \in \mathcal{A}$, i.e.

$$\mathbb{E}_{\mu^*}[q] \geq \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q] - \frac{\varepsilon}{K+1}.$$

By Fubini's theorem,

$$\mathbb{E}_{\mu_1^* \otimes \cdots \otimes \mu_K^*}[q] = \mathbb{E}_{\mu_1^*}\left[\mathbb{E}_{\mu_2^* \otimes \cdots \otimes \mu_K^*}[q]\right]$$

By the same arguments used in the proof of Theorem 6.12, μ_1^* can be replaced by some probability measure $\nu_1 \in \mathcal{M}_1(\mathcal{X}_1)$ with support on at most $N + N_1$ points, such that $\nu_1 \otimes \mu_2^* \otimes \cdots \otimes \mu_K^* \in \mathcal{A}$, and

$$\mathbb{E}_{\nu_1}\left[\mathbb{E}_{\mu_2^* \otimes \cdots \otimes \mu_K^*}[q]\right] \geq \mathbb{E}_{\mu_1^*}\left[\mathbb{E}_{\mu_2^* \otimes \cdots \otimes \mu_K^*}[q]\right] - \frac{\varepsilon}{K+1} \geq \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q] - \frac{2\varepsilon}{K+1}.$$

Repeating this argument a further $K - 1$ times yields $\nu = \bigotimes_{k=1}^K \nu_k \in \mathcal{A}_\Delta$ such that

$$\mathbb{E}_\nu[q] \geq \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q] - \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, it follows that

$$\sup_{\mu \in \mathcal{A}_\Delta} \mathbb{E}_\mu[q] = \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[q].$$

The proof for the infimum is similar. ■

Example 7.3. A simple two-dimensional optimization problem that illustrates the essential features of Theorem 7.2 is that of optimizing a bilinear form on $\mathbb{R}^2$ over a non-convex set with convex cross-sections. Suppose that $\mathcal{A} \subseteq \mathbb{R}^2$ is such that, for each $x, y \in \mathbb{R}$, the sections

$$\begin{aligned} \mathcal{A}_x &= \{y \in \mathbb{R} \mid (x, y) \in \mathcal{A}\}, \quad \text{and} \\ \mathcal{A}^y &= \{x \in \mathbb{R} \mid (x, y) \in \mathcal{A}\} \end{aligned}$$

are convex sets. Note that this does not imply that $\mathcal{A}$ itself is convex, as illustrated in Figure 7.1. Let $B\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be a bilinear functional: for definiteness, consider $B(x,y) = xy$. Since $\mathcal{A}$ is not convex, its extremal set is undefined, so it does not even make sense to claim that B has the same extreme values on $\mathcal{A}$ and $\operatorname{ext}(\mathcal{A})$. However, as can be seen in Figure 7.1, the extreme values of B over $\mathcal{A}$ *are* found at points (x^*, y^*) for which $x^* \in \operatorname{ext}(\mathcal{A}^{y^*})$ and $y^* \in \operatorname{ext}(\mathcal{A}_{x^*})$. Just as in the Fubini argument in the proof of Theorem 7.2, the optimal point can be found by either maximizing $\max_{x\in\mathcal{A}^y} B(x,y)$ with respect to y, or maximizing $\max_{y\in\mathcal{A}_x} B(x,y)$ with respect to x.

Remark 7.4. (a) In the context of Theorem 7.2, a measure $\mu \in \mathcal{A}_\Delta$ is of the form

$$\mu = \bigotimes_{k=1}^{K} \sum_{i_k=0}^{N+N_k} w_{ki_k}\delta_{x_{ki_k}} = \sum_{\boldsymbol{i}=(0,\dots,0)}^{(N+N_1,\dots,N+N_K)} w_{\boldsymbol{i}}\delta_{x_{\boldsymbol{i}}}$$

where, for a multi-index $\boldsymbol{i} \in \{0,\dots,N+N_1\}\times\cdots\times\{0,\dots,N+N_K\}$,

$$w_{\boldsymbol{i}} \coloneqq w_{1i_1}w_{2i_2}\dots w_{Ki_K} \geq 0,$$
$$x_{\boldsymbol{i}} \coloneqq \left(x_{1i_1},\dots x_{Ki_K}\right) \in \mathcal{X}.$$

Note that this means that the support of μ is a rectangular grid in $\mathcal{X}$.

(b) As noted in Remark 6.13(b), the support of a discrete measure $\mu \in \mathcal{A}_\Delta$, while finite, can be very large when K is large: the upper bound is

$$\#\operatorname{supp}(\mu) = \prod_{k=1}^{K}(1+N+N_k).$$

In such cases, it is usually necessary to sacrifice exact integration against μ for the sake of computational cost and resort to Monte Carlo averages against μ.

(c) However, it is often found in practice that the $\mu^* \in \mathcal{A}_\Delta$ that extremises $Q(\mu^*)$ does not have support on as many distinct points of $\mathcal{X}$ as Theorem 7.2 permits as an upper bound, and that not all of the constraints determining $\mathcal{A}$ hold as equalities. That is, there are often many inactive and non-binding constraints, and only those that are active and binding truly carry information about the extreme values of Q.

(d) Finally, note that this approach to UQ is non-intrusive in the sense that if we have a deterministic solver for $g\colon \mathcal{X}\to\mathcal{Y}$ and are interested in $\mathbb{E}_{X\sim\mu^\dagger}[q(g(X))]$ for some quantity of interest $q\colon \mathcal{Y}\to\mathbb{R}$, then the deterministic solver can be used 'as is' at each support point x of $\mu \in \mathcal{A}_\Delta$ in the optimization with respect to μ over $\mathcal{A}$.

8 Functional and Distributional Robustness

In addition to epistemic uncertainty about probability measures, applications often feature epistemic uncertainty about the functions involved. For example, if the system of interest is in reality some function $g^\dagger$ from a space $\mathcal{X}$ of inputs to another space $\mathcal{Y}$ of outputs, it may only be known that $g^\dagger$ lies in some subset $\mathcal{G}$ of the set of all (measurable) functions from $\mathcal{X}$ to $\mathcal{Y}$; furthermore, our information about $g^\dagger$ and our information about $\mu^\dagger$ may be coupled in some way, e.g. by knowledge of $\mathbb{E}_{X\sim\mu^\dagger}[g^\dagger(X)]$. Therefore, we now consider admissible sets of the form

$$\mathcal{A} \subseteq \left\{ (g,\mu) \;\middle|\; \begin{array}{c} g\colon \mathcal{X}\to\mathcal{Y} \text{ is measurable} \\ \text{and } \mu \in \mathcal{M}_1(\mathcal{X}) \end{array} \right\},$$

quantities of interest of the form $Q(g,\mu) = \mathbb{E}_{X\sim\mu}[q(X,g(X))]$ for some measurable function $q\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$, and seek the extreme values

$$\underline{Q}(\mathcal{A}) \coloneqq \inf_{(g,\mu)\in\mathcal{A}} \mathbb{E}_{X\sim\mu}[q(X,g(X))] \text{ and } \overline{Q}(\mathcal{A}) \coloneqq \sup_{(g,\mu)\in\mathcal{A}} \mathbb{E}_{X\sim\mu}[q(X,g(X))].$$

Obviously, if for each $g\colon \mathcal{X} \to \mathcal{Y}$ the set of $\mu \in \mathcal{M}_1(\mathcal{X})$ such that $(g, \mu) \in \mathcal{A}$ is a moment class of the form considered in Theorem 7.2, then

$$\operatorname*{ext}_{(g,\mu)\in\mathcal{A}} \mathbb{E}_{X\sim\mu}[q(X, g(X))] = \operatorname*{ext}_{\substack{(g,\mu)\in\mathcal{A} \\ \mu\in\bigotimes_{k=1}^{K}\Delta_{N+N_k}(\mathcal{X}_k)}} \mathbb{E}_{X\sim\mu}[q(X, g(X))].$$

In principle, though, although the search over μ is finite-dimensional for each g, the search over g is still infinite-dimensional. However, the passage to discrete measures often enables us to finite-dimensionalise the search over g, since, in some sense, only the values of g on the finite set $\operatorname{supp}(\mu)$ 'matter' in computing $\mathbb{E}_{X\sim\mu}[q(X, g(X))]$.

The idea is quite simple: instead of optimizing with respect to $g \in \mathcal{G}$, we optimise with respect to the finite-dimensional vector $y = g|_{\operatorname{supp}(\mu)}$. However, this reduction step requires some care:

(a) Some 'functions' do not have their values defined pointwise, e.g. 'functions' in Lebesgue and Sobolev spaces, which are actually equivalence classes of functions modulo equality almost everywhere. If isolated points have measure zero, then it makes no sense to restrict such 'functions' to a finite set $\operatorname{supp}(\mu)$. These difficulties are circumvented by insisting that $\mathcal{G}$ be a space of functions with pointwise-defined values.

(b) In the other direction, it is sometimes difficult to verify whether a vector y indeed arises as the restriction to $\operatorname{supp}(\mu)$ of some $g \in \mathcal{G}$; we need functions that can be extended from $\operatorname{supp}(\mu)$ to all of $\mathcal{X}$. Suitable extension properties are ensured if we restrict attention to smooth enough functions between the right kinds of spaces.

Theorem 8.1 (Minty, 1970). *Let $(\mathcal{X}, d)$ be a metric space, let $\mathcal{H}$ be a Hilbert space, let $E \subseteq \mathcal{X}$, and suppose that $f\colon E \to \mathcal{H}$ satisfies*

$$\|f(x) - f(y)\|_{\mathcal{H}} \leq d(x, y)^{\alpha} \quad \text{for all } x, y \in E \tag{8.1}$$

with Hölder constant $0 < \alpha \leq 1$. Then there exists $F\colon \mathcal{X} \to \mathcal{H}$ such that $F|_E = f$ and (8.1) holds for all $x, y \in \mathcal{X}$ if either $\alpha \leq \frac{1}{2}$ or if $\mathcal{X}$ is an inner product space with metric given by $d(x, y) = k^{1/\alpha}\|x - y\|$ for some $k > 0$. Furthermore, the extension can be performed so that $F(\mathcal{X}) \subseteq \overline{\operatorname{co}}(f(E))$, and hence without increasing the Hölder norm

$$\|f\|_{\mathcal{C}^{0,\alpha}} \coloneqq \sup_x \|f(x)\|_{\mathcal{H}} + \sup_{x\neq y} \frac{\|f(x) - f(y)\|_{\mathcal{H}}}{d(x, y)^{\alpha}},$$

where the suprema are taken over E or $\mathcal{X}$ as appropriate.

Minty's extension theorem includes as special cases the Kirszbraun–Valentine theorem (which assures that Lipschitz functions between Hilbert spaces can be extended without increasing the Lipschitz constant) and McShane's theorem (which assures that scalar-valued continuous functions on metric spaces can be extended without increasing a prescribed convex modulus of continuity). However, the extensibility property fails for Lipschitz functions between Banach spaces, even finite-dimensional ones, as shown by the following example of Federer (1969, p. 202):

Example 8.2. Let $E \subseteq \mathbb{R}^2$ be given by $E \coloneqq \{(1, -1), (-1, 1), (1, 1)\}$ and define $f\colon E \to \mathbb{R}^2$ by

$$f((1, -1)) \coloneqq (1, 0), \quad f((-1, 1)) \coloneqq (-1, 0), \quad \text{and } f((1, 1)) \coloneqq (0, \sqrt{3}).$$

Suppose that we wish to extend this f to $F\colon \mathbb{R}^2 \to \mathbb{R}^2$, where E and the domain copy of $\mathbb{R}^2$ are given the metric arising from the maximum norm $\|\cdot\|_{\infty}$ and the range copy of $\mathbb{R}^2$ is given the metric arising from the Euclidean norm $\|\cdot\|_2$. Then, for all distinct $x, y \in E$,

$$\|x - y\|_{\infty} = 2 = \|f(x) - f(y)\|_2,$$

so f has Lipschitz constant 1 on E. What value should F take at the origin, $(0, 0)$, if it is to have Lipschitz constant at most 1? Since $(0, 0)$ lies at $\|\cdot\|_{\infty}$-distance 1 from all three points

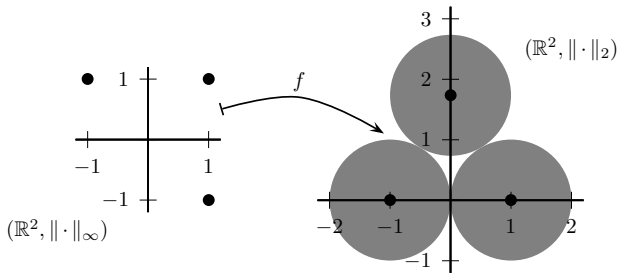


Figure 8.1: Illustration of Example 8.2. The function f that takes the three points on the left (equipped with $\|\cdot\|_\infty$) to the three points on the right (equipped with $\|\cdot\|_2$) has Lipschitz constant 1, but has no 1-Lipschitz extension F to $(0,0)$, let alone the whole plane $\mathbb{R}^2$, since $F((0,0))$ would have to lie in the (empty) intersection of the three grey discs.

of E, $F((0,0))$ must lie within $\|\cdot\|_2$-distance 1 of all three points of $f(E)$. However, there is no such point of $\mathbb{R}^2$ within distance 1 of all three points of $f(E)$, and hence any extension of f to $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ must have $\mathrm{Lip}(F) > 1$; indeed, any such F must have $\mathrm{Lip}(F) \geq \frac{2}{\sqrt{3}}$. See Figure 8.1.

Theorem 8.3. *Let $\mathcal{G}$ be a collection of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ such that, for every finite subset $E \subseteq \mathcal{X}$ and $g\colon E \to \mathcal{Y}$, it is possible to determine whether or not g can be extended to an element of $\mathcal{G}$. Let $\mathcal{A} \subseteq \mathcal{G} \times \mathcal{M}_1(\mathcal{X})$ be such that, for each $g \in \mathcal{G}$, the set of $\mu \in \mathcal{M}_1(\mathcal{X})$ such that $(g,\mu) \in \mathcal{A}$ is a moment class of the form considered in Theorem 7.2. Let*

$$\mathcal{A}_\Delta \coloneqq \left\{ (y,\mu) \;\middle|\; \begin{array}{c} \mu \in \bigotimes_{k=1}^K \Delta_{N+N_k}(\mathcal{X}_k), \\ y \text{ is the restriction to } \mathrm{supp}(\mu) \text{ of some } g \in \mathcal{G}, \\ \text{and } (g,\mu) \in \mathcal{A} \end{array} \right\}.$$

Then, if q is bounded either above or below, $\underline{Q}(\mathcal{A}) = \underline{Q}(\mathcal{A}_\Delta)$ and $\overline{Q}(\mathcal{A}) = \overline{Q}(\mathcal{A}_\Delta)$.

Exercise 8.4. Prove Theorem 8.3.

Example 8.5. Suppose that $g^\dagger\colon [-1,1] \to \mathbb{R}$ is known to have Lipschitz constant $\mathrm{Lip}(g^\dagger) \leq L$. Suppose also that the inputs of $g^\dagger$ are distributed according to $\mu^\dagger \in \mathcal{M}_1([-1,1])$, and it is known that

$$\mathbb{E}_{X\sim\mu^\dagger}[X] = 0 \quad \text{and} \quad \mathbb{E}_{X\sim\mu^\dagger}[g^\dagger(X)] \geq m > 0.$$

Hence, the corresponding feasible set is

$$\mathcal{A} = \left\{ (g,\mu) \;\middle|\; \begin{array}{c} g\colon [-1,1] \to \mathbb{R} \text{ has Lipschitz constant} \leq L, \\ \mu \in \mathcal{M}_1([-1,1]), \mathbb{E}_{X\sim\mu}[X] = 0, \text{ and } \mathbb{E}_{X\sim\mu}[g(X)] \geq m \end{array} \right\}.$$

Suppose that our quantity of interest is the probability of output values below 0, i.e. $q(x,y) = \mathbb{I}[y \leq 0]$. Then Theorem 8.3 ensures that the extreme values of

$$Q(g,\mu) = \mathbb{E}_{X\sim\mu}[\mathbb{I}[g(X) \leq 0]] = \mathbb{P}_{X\sim\mu}[g(X) \leq 0]$$

are the solutions of

$$\begin{aligned}
\text{extremise:} \quad & \sum_{i=0}^{2} w_i \mathbb{I}[y_i \leq 0] \\
\text{with respect to:} \quad & w_0, w_1, w_2 \geq 0 \\
& x_0, x_1, x_2 \in [-1, 1] \\
& y_0, y_1, y_2 \in \mathbb{R} \\
\text{subject to:} \quad & \sum_{i=0}^{2} w_i = 1 \\
& \sum_{i=0}^{2} w_i x_i = 0 \\
& \sum_{i=0}^{2} w_i y_i \geq m \\
& |y_i - y_j| \leq L|x_i - x_j| \text{ for } i, j \in \{0, 1, 2\}.
\end{aligned}$$

Example 8.6 (McDiarmid)**.** Consider the following admissible set of response functions and product measures on their inputs

$$\mathcal{A}_{\text{McD}} = \left\{ (g, \mu) \;\middle|\; \begin{array}{c} g \colon \mathcal{X} \to \mathbb{R} \text{ has } \mathcal{D}_k[g] \leq D_k, \\ \mu = \bigotimes_{k=1}^{K} \mu_k \in \mathcal{M}_1(\mathcal{X}), \\ \text{and } \mathbb{E}_{X \sim \mu}[g(X)] = m \end{array} \right\}.$$

Let $m_+ \coloneqq \max\{0, m\}$. This $\mathcal{A}_{\text{McD}}$ is the admissible set corresponding to the assumptions of a concentration-of-measure inequality known as McDiarmid's inequality (McDiarmid, 1989), which is the upper bound

$$\overline{Q}(\mathcal{A}_{\text{McD}}) \coloneqq \sup_{(g,\mu) \in \mathcal{A}_{\text{McD}}} \mathbb{P}_\mu[g(X) \leq 0] \leq \exp\left(-\frac{2m_+^2}{\sum_{k=1}^{K} D_k^2} \right).$$

Perhaps not surprisingly given its general form, McDiarmid's inequality is not the *least* upper bound on $\mathbb{P}_\mu[g(X) \leq 0]$; the actual least upper bound can be calculated using the reduction theorems. The proofs are lengthy, and the results are dependent upon K (Owhadi et al., 2013).

(a) For $K = 1$,

$$\overline{Q}(\mathcal{A}_{\text{McD}}) = \begin{cases} 0, & \text{if } D_1 \leq m_+, \\ 1 - \frac{m_+}{D_1}, & \text{if } 0 \leq m_+ \leq D_1. \end{cases} \tag{8.2}$$

(b) For $K = 2$,

$$\overline{Q}(\mathcal{A}_{\text{McD}}) = \begin{cases} 0, & \text{if } D_1 + D_2 \leq m_+, \\ \frac{(D_1 + D_2 - m_+)^2}{4 D_1 D_2}, & \text{if } |D_1 - D_2| \leq m_+ \leq D_1 + D_2, \\ 1 - \frac{m_+}{\max\{D_1, D_2\}}, & \text{if } 0 \leq m_+ \leq |D_1 - D_2|. \end{cases} \tag{8.3}$$

Note that in the third case, $\min\{D_1, D_2\}$ does not contribute to the least upper bound on $\mathbb{P}_\mu[g(X) \leq 0]$. In other words, if most of the uncertainty is contained in the first variable (i.e. $m_+ + D_2 \leq D_1$), then the uncertainty associated with the second variable does not affect the global uncertainty; the inequality $\mathcal{D}_2[g] \leq D_2$ is non-binding information, and a reduction of the global uncertainty requires a reduction in D_1.

(c) Similar, but more complicated, results are possible for $K \geq 3$, and there are similar 'screening effects' in which only a few of the diameter constraints supply binding information to the optimization problem for $\overline{Q}(\mathcal{A}_{\text{McD}})$.

Dominant Uncertainties and Screening Effects. The phenomenon observed in the $K = 2$ solution of the optimal McDiarmid inequality (8.3) occurs in many contexts: not all of the constraints that specify $\mathcal{A}$ necessarily hold as binding or active constraints at the extremizing solution $(g^*, \mu^*) \in \mathcal{A}$. That is, the best- and worst-case predictions for the quantity of interest $Q(g^\dagger, \mu^\dagger)$ are controlled by only a few pieces of input information, and the others have not just little impact, but none at all! Far from being undesirable, this phenomenon is actually very useful, since it can be used to direct future information-gathering activities, such as expensive experimental campaigns, by attempting to acquire information (and hence pass to a smaller feasible set $\mathcal{A}' \subsetneq \mathcal{A}$) that will modify the binding/active constraints for the previous problem, i.e. invalidate the previous extremiser in $\mathcal{A}$ and lead to a new extremiser in $\mathcal{A}'$. In this way, we hence pass from the optimal bounds given the information in $\mathcal{A}$

$$\underline{Q}(\mathcal{A}) \leq Q(g^\dagger, \mu^\dagger) \leq \overline{Q}(\mathcal{A})$$

to improved optimal bounds given the information in $\mathcal{A}'$

$$\underline{Q}(\mathcal{A}) < \underline{Q}(\mathcal{A}') \leq Q(g^\dagger, \mu^\dagger) \leq \overline{Q}(\mathcal{A}') < \overline{Q}(\mathcal{A}).$$

Exercise 8.7. Calculate by hand, as a function of $t \in \mathbb{R}$, $m \in \mathbb{R}$, $z \in [0, 1]$ and $v \in \mathbb{R}$,

$$\sup_{(g,\mu) \in \mathcal{A}} \mathbb{P}_{X \sim \mu}[g(X) \leq t],$$

where

$$\mathcal{A} := \left\{ (g, \mu) \;\middle|\; \begin{array}{c} g \colon [0,1] \to \mathbb{R} \text{ has Lipschitz constant 1,} \\ \mu \in \mathcal{M}_1([0,1]), \, \mathbb{E}_{X \sim \mu}[g(X)] \geq m, \\ \text{and } g(z) = v \end{array} \right\}.$$

Numerically verify your calculations.

9 Numerical Implementation

The inequalities of Markov and Chebyshev are elementary deviation inequalities for random variables under very simple assumptions, or rather with very simple information. The approach to uncertainty quantification discussed in these notes, namely optimizing over families of probability measures and functions, can be seen as the calculation of situation-specific probabilistic inequalities. Even after applying the reduction theorems (Theorems 6.12, 7.2, 8.3), we will usually have no hope of expressing the solutions in closed form and we must resort to the tools of numerical optimization.

As remarked earlier, if the input space $\mathcal{X}$ is a finite set and the constraints are all moment constraints, then the optimization problem is already a finite-dimensional linear programming problem to determine the worst- or best-case weights w_i. Such problems can be solved quickly, accurately, and reliably using many off-the-shelf software packages for linear or convex optimization.

In general, though, the optimization problems are nonconvex and highly constrained global optimization problems — a numerical nightmare! This is not to say that the situation is hopeless, only that it must usually be solved in a more careful and time-consuming 'offline' mode, whereas linear and convex programming is amenable to 'real time' solutions.

Without going into details, it is worth discussing some the characteristics of the optimization problems to be solved, and hence the requirements of any putative software implementation. For purely illustrative purposes, pseudo-code will be presented in a Python-like syntax.

The central data objects are representatives of the discrete function-measure pairs $(g, \mu) \in \mathcal{A}_\Delta$. The basic object is a measure μ supported at a single point $x \in \mathcal{X}$, with weight $w \in \mathbb{R}$, and a corresponding value y representing $g(x)$. This kind of object could be implemented as a class, or even a simple dictionary:

```
p0 = {"position": x, "weight": w, "value": y}
```

A float or double-precision float may be a suitable data type for `w`; array data types should be used for `x` and `y`. After treating single-point measures, one needs to be able to consider measures supported on finitely many points, which could again be implemented as classes, or just as lists:

```
mu = [p0, p1, ..., pn]
```

Assuming that this `mu` is indeed a probability measure (the weights are all non-negative and sum to unity), the expected value of a function `f` on $\mathcal{X}$ is simply

```
def expected_value(f, mu):
  return sum(f(p["position"]) * p["weight"] for p in mu)
```

and the expected value of a quantity of interest, $\mathbb{E}_{X\sim\mu}[q(X, g(X))]$, as discussed earlier, would be

```
def expected_value_of_qoi(q, mu):
  return sum(q(p["position"], p["value"]) * p["weight"] for p in mu)
```

The question of what to do if `mu` has not been normalised is actually a gateway to the most important aspect of optimal UQ problems: the treatment of constraints. For example, to ensure that `mu` remains normalised as a probability measure throughout the computation, even when the optimiser subjects its components to e.g. some random walk in parameter space, it is very useful to have a function that *imposes* the constraint that all the point masses are non-negative and sum to unity. Here is one possibility, which reflects any negative masses in the origin (truncation is another popular choice) and then renormalises:

```
def make_probability_measure(mu):
  for p in mu:
    p["weight"] = abs(p["weight"])
  total_weight = sum(p["weight"] for p in mu)
  if not total_weight == 0.0:
     for p in mu:
       p["weight"] = p["weight"] / total_weight
     return mu
  else:
     return <some error message>
```

Similar functions should be written for the imposition of target means, variances, etc. When it is not possible to write functions that will perform such transformations exactly, there is a choice: either

- add *penalty terms* to the objective function, so that failure to satisfy the constraints is heavily penalised; or
- continue to insist that trial points `mu` satisfy the constraints, but impose the constraints implicitly rather than explicitly, by operating an *inner optimisation loop* that minimises non-satisfaction of the constraints down to zero.

The penalty function approach is often quick to implement and computationally cheap to run for simple constraints, but becomes unwieldy when many constraints are involved; it also has the disadvantage of corrupting the problem structure and allowing the objective function to be evaluated off the feasible set. The inner optimisation approach has complementary advantages and disadvantages: the problem structure is respected, but potentially at a high computational cost. Which approach is 'better' is highly problem-dependent.

One final remark about the imposition of constraints is that implementations should remain *flexible* so that new constraints can be added and old ones removed, in order to explore the effect of new items of information or other hypothetical scenarios upon the

optimisation problem. This means, as a design choice, that the optimisation strategy and the representation of constraints should be kept separate unless there is a very good reason to couple them and use an optimiser that is only good for particular kinds of constraints.

Exercise 9.1. Write a program to solve Example 6.14 and Exercises 6.15 and 6.16 numerically using the number of support points given by the reduction theorem (Theorem 6.12). Once your program is working well, allow it to run on a collection of support point *more numerous* than Theorem 6.12 says is sufficient — what do you observe?

A final remark about implementation is that — depending upon the optimisation strategy — one often observes that the point masses `p` comprising `mu` degenerate in the course of the calculation, either by weights degenerating to a numerical `0.0` or positions 'colliding'. This degeneracy is actually a good thing: the optimiser is discovering that not all the available degrees of freedom are needed to solve the problem — as in the last part of the previous exercise. Therefore, efficient implementations will monitor the optimiser for this kind of degeneracy event, and then restart the calculation in a lower-dimensional search space, e.g. with one less point mass, corresponding to removing the point mass with weight `0.0`.

10 Background and Literature

The principle of maximum entropy was proposed by Jaynes (1957a,b), appealing to a correspondence between statistical mechanics and information theory. On the basis of this principle and Cox's theorem (Cox, 1946, 1961), Jaynes (2003) developed a comprehensive viewpoint on probability theory, viewing it as the natural extension of Aristotelian logic.

Berger (1994) makes the case for distributional robustness, with respect to priors and likelihoods, in Bayesian inference. Smith (1995) provides theory and several practical examples for generalised Chebyshev inequalities in decision analysis. Boyd and Vandenberghe (2004, Section 7.2) cover some aspects of distributional robustness under the heading of nonparametric distribution estimation, in the case in which it is a convex problem. Convex optimization approaches to distributional robustness and optimal probability inequalities are also considered by Bertsimas and Popescu (2005). There is also an extensive literature on the related topic of majorization, for which see the book of Marshall et al. (2011).

A standard short reference on Choquet theory is the book of Phelps (2001). Theorem 6.6 was proved first by Choquet under the additional assumption that the simplex is compact; the assumption was later dropped by Kendall (1962). For linear programming in infinite-dimensional spaces, with careful attention to what parts of the analysis are purely algebraic and what parts require topology / order theory, see Anderson and Nash (1987).

The classification of the extreme points of moment sets, and the consequences for the optimization of measure affine functionals, are due to von Weizsäcker and Winkler (1979/80, 1980) and Winkler (1988). Theorem 7.2 and the Lipschitz version of Theorem 8.3 can be found in Owhadi et al. (2013) and Sullivan et al. (2013) respectively. Theorem 8.1 is due to Minty (1970), and generalises earlier results by McShane (1934), Kirszbraun (1934), and Valentine (1945). The optimal version of McDiarmid's inequality is given by Owhadi et al. (2013, Section 5.1.1).

Applications of the methodology discussed in these notes can be found in various papers:

1. applications to hypervelocity impact Owhadi et al. (2013), Sullivan et al. (2013), and Kamga et al. (2014);
2. applications to seismic safety certification: Owhadi et al. (2013);
3. application to power grid optimisation: Han et al. (2015);
4. applications to the robustness of Bayesian inference: Owhadi et al. (2015a,b).

Corresponding software can be found in the examples section of the *mystic* optimization framework at

`http://github.com/uqfoundation/mystic`

References

E. J. Anderson and P. Nash. *Linear Programming in Infinite-Dimensional Spaces*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Ltd., Chichester, 1987. ISBN 0-471-91250-6. Theory and applications, A Wiley-Interscience Publication.

J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994. ISSN 1133-0686. doi: 10.1007/BF02562676. URL http://dx.doi.org/10.1007/BF02562676. With comments and a rejoinder by the author.

D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: a convex optimization approach. *SIAM J. Optim.*, 15(3):780–804 (electronic), 2005. ISSN 1052-6234. doi: 10.1137/S1052623401399903. URL http://dx.doi.org/10.1137/S1052623401399903.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. ISBN 0-521-83378-7.

R. T. Cox. Probability, frequency and reasonable expectation. *Amer. J. Phys.*, 14:1–13, 1946. ISSN 0002-9505.

R. T. Cox. *The Algebra of Probable Inference*. The Johns Hopkins Press, Baltimore, Md, 1961.

H. Federer. *Geometric Measure Theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

S. Han, M. Tao, U. Topcu, H. Owhadi, and R. M. Murray. Convex optimal uncertainty quantification. *SIAM J. Optim.*, 25(3):1368–1387, 2015. ISSN 1052-6234. doi: 10.1137/13094712X. URL http://dx.doi.org/10.1137/13094712X.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev. (2)*, 106:620–630, 1957a.

E. T. Jaynes. Information theory and statistical mechanics. II. *Phys. Rev. (2)*, 108:171–190, 1957b.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003. ISBN 0-521-59271-2. doi: 10.1017/CBO9780511790423. URL http://dx.doi.org/10.1017/CBO9780511790423. Edited and with a foreword by G. Larry Bretthorst.

P.-H. T. Kamga, B. Li, M. McKerns, L. H. Nguyen, M. Ortiz, H. Owhadi, and T. J. Sullivan. Optimal uncertainty quantification with model uncertainty and legacy data. *J. Mech. Phys. Solids*, 72:1–19, 2014. doi: 10.1016/j.jmps.2014.07.007. URL http://dx.doi.org/10.1016/j.jmps.2014.07.007.

D. G. Kendall. Simplexes and vector lattices. *J. London Math. Soc.*, 37:365–371, 1962. ISSN 0024-6107.

M. D. Kirszbraun. Über die zusammenziehende und Lipschitzsche Transformationen. *Fund. Math.*, 22:77–108, 1934.

A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and its Applications*. Springer Series in Statistics. Springer, New York, second edition, 2011. ISBN 978-0-387-40087-7. doi: 10.1007/978-0-387-68276-1. URL http://dx.doi.org/10.1007/978-0-387-68276-1.

C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

E. J. McShane. Extension of range of functions. *Bull. Amer. Math. Soc.*, 40(12):837–842, 1934. ISSN 0002-9904. doi: 10.1090/S0002-9904-1934-05978-0. URL http://dx.doi.org/10.1090/S0002-9904-1934-05978-0.

G. J. Minty. On the extension of Lipschitz, Lipschitz–Hölder continuous, and monotone functions. *Bull. Amer. Math. Soc.*, 76:334–339, 1970. ISSN 0002-9904.

H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz. Optimal Uncertainty Quantification. *SIAM Rev.*, 55(2):271–345, 2013. doi: 10.1137/10080782X. URL http://dx.doi.org/10.1137/10080782X.

H. Owhadi, C. Scovel, and T. J. Sullivan. Brittleness of Bayesian inference under finite information in a continuous world. *Electron. J. Stat.*, 9:1–79, 2015a. ISSN 1935-7524. doi: 10.1214/15-EJS989. URL http://dx.doi.org/10.1214/15-EJS989.

H. Owhadi, C. Scovel, and T. J. Sullivan. On the brittleness of Bayesian inference. *SIAM Rev.*, 57(4):566–582, 2015b. ISSN 0036-1445. doi: 10.1137/130938633. URL http://dx.doi.org/10.1137/130938633.

R. R. Phelps. *Lectures on Choquet's Theorem*, volume 1757 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, second edition, 2001. ISBN 3-540-41834-2. doi: 10.1007/b76887. URL http://dx.doi.org/10.1007/b76887.

K. R. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, 1963.

J. E. Smith. Generalized Chebychev inequalities: theory and applications in decision analysis. *Oper. Res.*, 43(5):807–825, 1995. ISSN 0030-364X. doi: 10.1287/opre.43.5.807. URL http://dx.doi.org/10.1287/opre.43.5.807.

T. J. Sullivan. *Introduction to Uncertainty Quantification*, volume 63 of *Texts in Applied Mathematics*. Springer, 2015. ISBN 978-3-319-23394-9; 978-3-319-23395-6. doi: 10.1007/978-3-319-23395-6. URL http://dx.doi.org/10.1007/978-3-319-23395-6.

T. J. Sullivan, M. McKerns, D. Meyer, F. Theil, H. Owhadi, and M. Ortiz. Optimal uncertainty quantification for legacy data observations of Lipschitz functions. *ESAIM Math. Model. Numer. Anal.*, 47(6):1657–1689, 2013. ISSN 0764-583X. doi: 10.1051/m2an/2013083. URL http://dx.doi.org/10.1051/m2an/2013083.

F. A. Valentine. A Lipschitz condition preserving extension for a vector function. *Amer. J. Math.*, 67(1):83–93, 1945. ISSN 0002-9327. doi: 10.2307/2371917. URL http://dx.doi.org/10.2307/2371917.

H. von Weizsäcker and G. Winkler. Integral representation in the set of solutions of a generalized moment problem. *Math. Ann.*, 246(1):23–32, 1979/80. ISSN 0025-5831. doi: 10.1007/BF01352023. URL http://dx.doi.org/10.1007/BF01352023.

H. von Weizsäcker and G. Winkler. Noncompact extremal integral representations: some probabilistic aspects. In *Functional Analysis: Surveys and Recent Results, II (Proc. Second Conf. Functional Anal., Univ. Paderborn, Paderborn, 1979)*, volume 68 of *Notas Mat.*, pages 115–148. North-Holland, Amsterdam, 1980.

G. Winkler. Extreme points of moment sets. *Math. Oper. Res.*, 13(4): 581–587, 1988. ISSN 0364-765X. doi: 10.1287/moor.13.4.581. URL http://dx.doi.org/10.1287/moor.13.4.581.