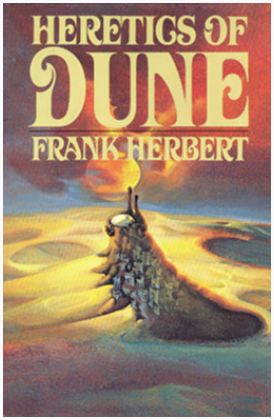
Optimal Distributionally Robust Uncertainty Quantification

Tim Sullivan[1,2]

SFB 1294 Colloquium, University of Potsdam, DE

8 December 2017

[1]Free University of Berlin, DE
[2]Zuse Institute Berlin, DE



"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."

Overview

Overview

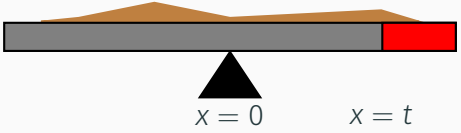
- In uncertainty quantification, one is usually faced with the challenge of quantifying the impact of some uncertainty or random variability (often modelled as a probability distribution μ) on a particular system of interest (often modelled as a response function g).
- This talk is an introduction to uncertainty quantification under a particularly severe form of uncertainty: 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.

- The framework these lectures describe is very general, and in particular the measure μ might be interpreted in a Bayesian or frequentist fashion.
- In the robust Bayesian analysis paradigm (Berger, 1994; Owhadi et al., 2015a,b), varying μ corresponds to changing one's prior or likelihood model.
- With such examples 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 are grounded in optimization theory, and have a particularly strong analogy to finite-dimensional linear programming, even though $\mathcal{A}$ will typically be infinite-dimensional.

Example (Balancing a Seesaw)

You are given 1kg of sand to arrange however you wish on a seesaw (= the real line). Your challenge is to make the region $x \geq t$, $t \geq 0$, as heavy as possible subject to two constraints:

- the centre of mass of the sand (and seesaw) must be at $x = 0$; and
- all the sand must be contained in a region of length $\leq L$ (with $L \geq t$).

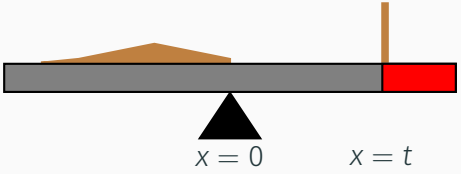


- Optimal distributionally robust UQ can be seen as the extension of the same basic idea to complicated settings: no hope of a pen-and-paper solution, but can compute a numerical solution.

Example (Balancing a Seesaw)

You are given 1kg of sand to arrange however you wish on a seesaw (= the real line). Your challenge is to make the region $x \geq t$, $t \geq 0$, as heavy as possible subject to two constraints:

- the centre of mass of the sand (and seesaw) must be at $x = 0$; and
- all the sand must be contained in a region of length $\leq L$ (with $L \geq t$).

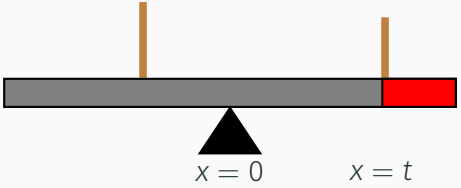


- Optimal distributionally robust UQ can be seen as the extension of the same basic idea to complicated settings: no hope of a pen-and-paper solution, but can compute a numerical solution.

Example (Balancing a Seesaw)

You are given 1kg of sand to arrange however you wish on a seesaw (= the real line). Your challenge is to make the region $x \geq t$, $t \geq 0$, as heavy as possible subject to two constraints:

- the centre of mass of the sand (and seesaw) must be at $x = 0$; and
- all the sand must be contained in a region of length $\leq L$ (with $L \geq t$).

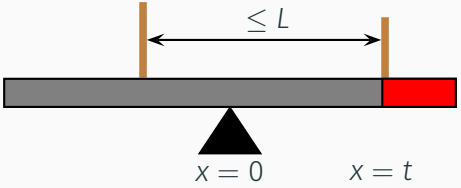


- Optimal distributionally robust UQ can be seen as the extension of the same basic idea to complicated settings: no hope of a pen-and-paper solution, but can compute a numerical solution.

Example (Balancing a Seesaw)

You are given 1kg of sand to arrange however you wish on a seesaw (= the real line). Your challenge is to make the region $x \geq t$, $t \geq 0$, as heavy as possible subject to two constraints:

- the centre of mass of the sand (and seesaw) must be at $x = 0$; and
- all the sand must be contained in a region of length $\leq L$ (with $L \geq t$).

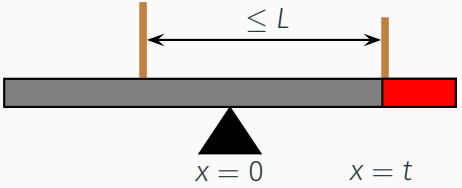


- Optimal distributionally robust UQ can be seen as the extension of the same basic idea to complicated settings: no hope of a pen-and-paper solution, but can compute a numerical solution.

Example (Balancing a Seesaw)

You are given 1kg of sand to arrange however you wish on a seesaw (= the real line). Your challenge is to make the region $x \geq t$, $t \geq 0$, as heavy as possible subject to two constraints:

- the centre of mass of the sand (and seesaw) must be at $x = 0$; and
- all the sand must be contained in a region of length $\leq L$ (with $L \geq t$).



$$\max = 1 - \frac{t}{L}$$

- Optimal distributionally robust UQ can be seen as the extension of the same basic idea to complicated settings: no hope of a pen-and-paper solution, but can compute a numerical solution.

Overview

- 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?
- The answer 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

Suppose that it is desired to give bounds on the CDF 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 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 g, what are optimal bounds on the cumulative distribution function of Y, i.e., for $t \in \mathbb{R}$, what are

$$\inf_{\mu\in\mathcal{A}} \mathbb{P}_{X\sim\mu}[g(X) \leq t] \text{ and } \sup_{\mu\in\mathcal{A}} \mathbb{P}_{X\sim\mu}[g(X) \leq t]?. \tag{1}$$

Example

- We will show that these extremal values can be found by solving an optimisation problem involving at most eight optimisation 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.
- In general, this problem is a non-convex global optimisation 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.
- Problem (1) reduces to a nonlinear family of linear programs, parametrised by $\{x_i\}_{i=0}^3$.

Example

$$\begin{aligned}
\text{extremise: } & \sum_{i=0}^{3} w_i \mathbb{1}[g(x_i) \leq t]\,; \\
\text{w.r.t.: } & x_0, \ldots, x_3 \in \mathbb{R} \text{ and } w_0, \ldots, w_3 \geq 0\,; \\
\text{subject to: } & \sum_{i=0}^{3} w_i = 1\,, \\
& 0 \leq \sum_{i=0}^{3} w_i \mathbb{1}[x_i \leq a] \leq 0.1\,, \\
& 0.8 \leq \sum_{i=0}^{3} w_i \mathbb{1}[a < x_i \leq b] \leq 1.0\,, \\
& 0 \leq \sum_{i=0}^{3} w_i \mathbb{1}[x_i > b] \leq 0.1\,.
\end{aligned}$$

Overview

- By way of contrast, the *Maximum Entropy* approach seeks to approximate $Q(\mu^\dagger)$, knowing only that $\mu^\dagger \in \mathcal{A}$, by selecting a particular 'generic' representative of the class $\mathcal{A}$.

Definition

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 (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

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

The only constraints on p are the natural ones that $p_i \geq 0$ and that $\sum_{i=1}^m p_i = 1$. Using the method of Lagrange multipliers, the unique extremiser of $\mathrm{H}(p)$ among $\{p \in \mathbb{R}^m \mid p_i \geq 0, \sum_{i=1}^m p_i = 1\}$ is the uniform distribution $p_1 = \cdots = p_m = \frac{1}{m}$.

Example (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 $\mu = \rho\,\mathrm{d}x$ to maximise

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

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 $\int_{\mathbb{R}} (x-m)^2\rho(x)\,\mathrm{d}x = s^2$. The method of Lagrange multipliers yields that the maximum entropy distribution on $\mathbb{R}$ of with mean m and variance s^2 is $\mathcal{N}(m, s^2)$, with entropy

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

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

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

  for given convex functions $\varphi_1, \dots, \varphi_n \colon \mathbb{R}^m \to \mathbb{R}$.
- An explicit formula for the maximum entropy distribution is rarely available, but in this situation we can quickly and reliably compute it.

- Entropy is really *relative* entropy (Kullback–Leibler divergence) with respect to uniform measure. Why privilege this measure? What about settings that don't admit a uniform reference measure?
- Not all classes of probability measures contain maximum entropy distributions:
  - 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.
  - 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 etropy bounded above but there is no distribution which attains the maximal entropy.
- The MaxEnt Principle is an application-blind selection mechanism. It asserts that the correct course of action 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.
- MaxEnt distributions are atypically smooth and light-tailed, as the next exercise illustrates, whereas many important applications involve distributions that have heavy tails.

Overview

- 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 particularly 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

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

- The challenge is computing $\underline{Q}(\mathcal{A})$ and $\overline{Q}(\mathcal{A})$!

- Suppose that the sample space $\mathcal{X} = \{1, \ldots, K\}$ is a finite set equipped with the discrete topology.
- 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:

- 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.
- The extreme values of this linear programming problem are found in the extremal set $\text{ext}(\mathcal{A})$.

- This simple insight, that expectation is linear in the probability measure, and so extreme values over a class of measures should be found at extreme points, can be exploited to great effect in the study of distributional robustness problems for general sample spaces $\mathcal{X}$.
- This insight generalises to much more general sample spaces, so long as the feasible set $\mathcal{A}$ of probability measures is 'sufficiently like a polytope'.
- What would appear to be an intractable optimisation problem over an infinite-dimensional set of measures is in fact equivalent to a tractable finite-dimensional problem.
- Thus, the aim of this section (and in the next two sections) is to find a finite-dimensional subset $\mathcal{A}_\Delta \subseteq \mathcal{A}$ such that

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

- The first step 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 test the truth of finitely many inequalities, such moment classes are appealing 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

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

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

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:

Definition

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

$$\mu(B) = \int_{\mathrm{ext}(\mathcal{A})} \nu(B)\, \mathrm{d}p(\nu) \quad \text{for all measurable } B \subseteq X. \tag{2}$$

The measure p is said to represent the barycentre μ.

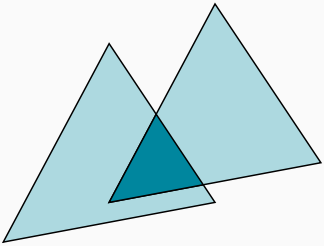
- Recall that a d-dimensional simplex is the closed convex hull of $d+1$ points $p_0, \ldots, d_d$ such that $p_1 - p_0, \ldots, p_d - p_0$ are linearly independent.
- The right infinite-dimensional analogue for distributional robustness is a Choquet simplex in $\mathcal{M}_{\pm}(X)$.
- There is a cumbersome definition using orderings and cones on vector spaces, but there is a convenient alternative definition that is much more amenable to visual intuition, and more easily checked in practice:

Definition

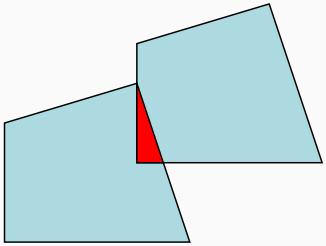
A homothety of a real topological vector space $\mathcal{V}$ is the composition of a positive dilation with a translation, i.e. a function $f\colon \mathcal{V} \to \mathcal{V}$ of the form $f(x) = \alpha x + v$, for fixed $\alpha > 0$ and $v \in \mathcal{V}$.

Theorem (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.



A triangle is indeed a simplex.



Not a simplex, by Choquet–Kendall.

Theorem (Winkler, 1988)

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and let $S \subseteq \mathcal{M}_1(\mathcal{F})$ be a Choquet simplex such that $\operatorname{ext}(S)$ *consists of Dirac measures. Fix measurable functions $\varphi_1, \dots, \varphi_N \colon \mathcal{X} \to \mathbb{R}$ and $c_1, \dots, c_N \in \mathbb{R}$ and let*

$$\mathcal{A} \coloneqq \left\{ \mu \in S \;\middle|\; \begin{array}{c} \text{for } n = 1, \dots, N, \\ \varphi_n \in L^1(\mathcal{X}, \mu) \text{ and } \mathbb{E}_\mu[\varphi_n] \leq c_n \end{array} \right\}.$$

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

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

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

The important point for us is that, when $\mathcal{X}$ is pseudo-Radon, Winkler's theorem applies to $S = \mathcal{M}_1(\mathcal{X})$, so $\operatorname{ext}(\mathcal{A}) \subseteq \mathcal{A} \cap \Delta_N(\mathcal{X})$, where

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

denotes convex combinations of $\leq N + 1$ unit Dirac measures on $\mathcal{X}$.

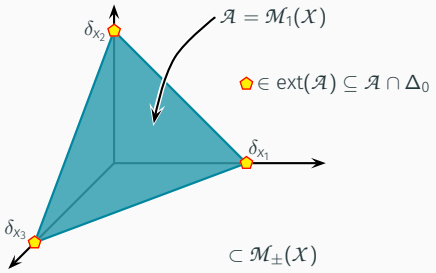


Figure 1: Heuristic justification of Winkler's theorem. Observe that the extreme points of the moment class $\mathcal{A}$ consist of convex combinations of at most $1+$ the number of constraints defining $\mathcal{A}$ point masses.

The important point for us is that, when $\mathcal{X}$ is pseudo-Radon, Winkler's theorem applies to $S = \mathcal{M}_1(\mathcal{X})$, so $\operatorname{ext}(\mathcal{A}) \subseteq \mathcal{A} \cap \Delta_N(\mathcal{X})$, where

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

denotes convex combinations of $\leq N + 1$ unit Dirac measures on $\mathcal{X}$.

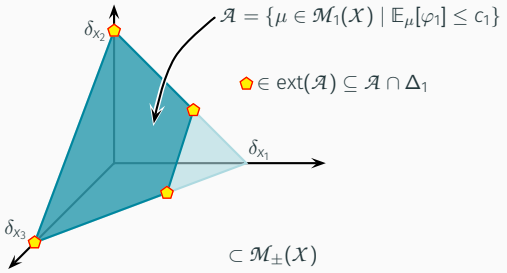


Figure 1: Heuristic justification of Winkler's theorem. Observe that the extreme points of the moment class $\mathcal{A}$ consist of convex combinations of at most 1+ the number of constraints defining $\mathcal{A}$ point masses.

The important point for us is that, when $\mathcal{X}$ is pseudo-Radon, Winkler's theorem applies to $S = \mathcal{M}_1(\mathcal{X})$, so $\operatorname{ext}(\mathcal{A}) \subseteq \mathcal{A} \cap \Delta_N(\mathcal{X})$, where

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

denotes convex combinations of $\leq N+1$ unit Dirac measures on $\mathcal{X}$.

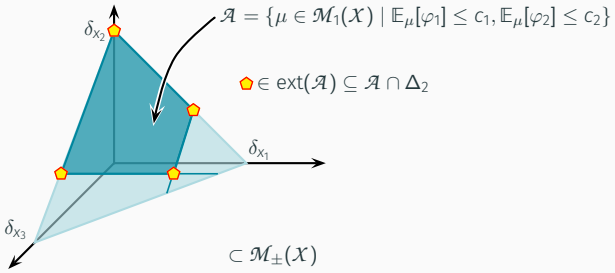


Figure 1: Heuristic justification of Winkler's theorem. Observe that the extreme points of the moment class $\mathcal{A}$ consist of convex combinations of at most 1+ the number of constraints defining $\mathcal{A}$ point masses.

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

Definition

For $\mathcal{A} \subseteq \mathcal{M}_1(X)$, a function $F\colon \mathcal{A} \to \mathbb{R} \cup \{\pm\infty\}$ is said to be measure affine if, for all barycentres $\mu \in \mathcal{A}$ represented by $p \in \mathcal{M}_1(\mathrm{ext}(A))$, F is p-integrable with

$$F(\mu) = \int_{\mathrm{ext}(\mathcal{A})} F(\nu)\, \mathrm{d}p(\nu). \tag{3}$$

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

Lemma

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

Theorem

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

In summary, we now have the following:

Theorem (Reduction for measures)

Let $\mathcal{X}$ be a pseudo-Radon space and let $\mathcal{A} \subseteq \mathcal{M}_1(\mathcal{X})$ be a moment class:

$$\mathcal{A} := \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid \mathbb{E}_\mu[\varphi_n] \leq 0 \text{ for } n = 1, \dots, N\}$$

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

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

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

- The Reduction Theorem is good news from a computational standpoint for two reasons:
  - 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 optimisation techniques that can be implemented on a computer.
  - 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.
- 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 optimisation problems over $\mathcal{A}_\Delta$ in the Reduction Theorem can only be solved approximately, using the tools of numerical global optimisation. However, some of the classical inequalities of basic probability theory can be obtained in closed form by this approach.

Example (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

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

This $\mathcal{A}$ is a moment class with a single moment constraint. By the Reduction Theorem, 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 $0 \leq x_0 \leq x_1 < \infty$, with masses w_0, w_1 respectively.

Example (Markov's inequality)

- In order to satisfy the mean constraint that $\mathbb{E}[X] \leq m$, we must have $x_0 \leq m$.
- If $x_1 > t$ and the mean constraint is satisfied, then moving the mass w_1 at x_1 to $x_1' := t$ 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 $x_1 = t$.
- By similar reasoning, it is sufficient to consider two-point distributions with $x_0 = 0$.
- 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 .....

Example (Markov's inequality)

- If $\mathbb{E}_{X\sim\mu}[X] < m$, change masses to

$$w_1' := m/t > w_1,$$
$$w_0' := 1 - m/t < w_0,$$

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.

Overview

- The kinds of constraints on measures (or, if you prefer, random variables) that can be considered in the Reduction Theorem 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.

Exercise

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 is typically not convex.

- Fortunately, a cunning application of Fubini's theorem resolves this non-convexity difficulty.
- Fubini's theorem says that integration (expectation) against a product measure can be performed as an iterated integral:

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

  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 the Reduction Theorem to cope with independence constraints coupled with moment constraints on the marginal and joint distributions.
- N.B.: unlike the previous Reduction Theorem, we do *not* now say that $\mathcal{A}_\Delta = \mathrm{ext}(\mathcal{A})$, only that the optimisation problem has the same extreme values over $\mathcal{A}_\Delta$ and $\mathcal{A}$.

Theorem (Reduction for product measures)

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

$$\mathcal{A} := \left\{ \mu = \bigotimes_{k=1}^{K} \mu_k \in \bigotimes_{k=1}^{K} \mathcal{M}_1(X_k) \;\middle|\; \begin{array}{c} \mathbb{E}_\mu[\varphi_n] \le 0 \text{ for } n = 1, \dots, N, \\ \mathbb{E}_{\mu_1}[\varphi_{1,n}] \le 0 \text{ for } n = 1, \dots, N_1, \\ \vdots \\ \mathbb{E}_{\mu_K}[\varphi_{K,n}] \le 0 \text{ for } n = 1, \dots, N_K \end{array} \right\}$$

for prescribed measurable functions $\varphi_n \colon X \to \mathbb{R}$ and $\varphi_{k,n} \colon X_k \to \mathbb{R}$. Let

$$\mathcal{A}_\Delta := \left\{ \mu \in \mathcal{A} \,\middle|\, \mu_k \in \Delta_{N+N_k}(X_k) \right\}.$$

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

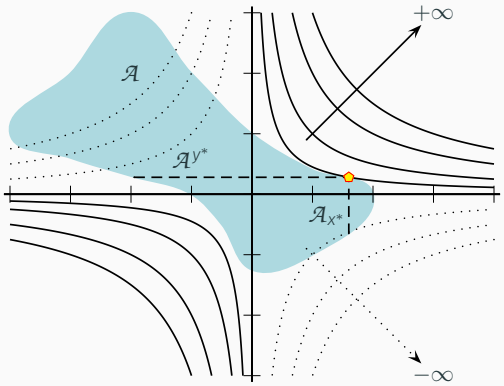


Figure 2: Optimisation of a bilinear form B over a non-convex set $\mathcal{A} \subseteq \mathbb{R}^2$ that has convex cross-sections. The black curves show contours of $B(x, y) = xy$. Note that the maximum value of B over $\mathcal{A}$ is found at a point (x^*, y^*) such that x^* and y^* are both extreme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

Example

The essential features of the Reduction Theorem for Independence are captured by optimising bilinear form on $\mathbb{R}^2$ over a set $\mathcal{A} \subseteq \mathbb{R}^2$ with convex cross-sections, i.e. such that the sections

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

are convex sets for each $x, y \in \mathbb{R}$. 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, 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 Reduction Theorem for Independence, the optimal point can be found by either maximising $\max_{x \in \mathcal{A}^y} B(x, y)$ with respect to y, or maximising $\max_{y \in \mathcal{A}_x} B(x, y)$ with respect to x.

- In the Reduction Theorem for Independence, a measure $\mu \in \mathcal{A}_\Delta$ is of the form

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

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

$$w_i \coloneqq w_{1,i_1} w_{2,i_2} \dots w_{K,i_K} \geq 0,$$
$$x_i \coloneqq \left(x_{1,i_1}, \dots x_{K,i_K}\right) \in \mathcal{X}.$$

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

- As noted earlier, 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

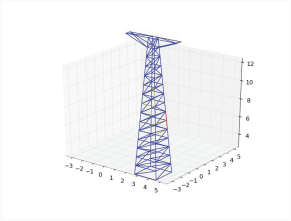
$$\#\operatorname{supp}(\mu) = \prod_{k=1}^{K}(1 + N + N_k).$$

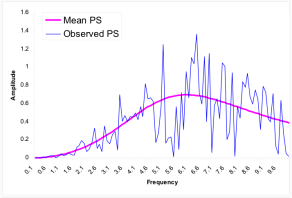
  In such cases, it is usually necessary to sacrifice exact integration against μ for the sake of computational cost and resort to Monte Carlo averages against μ.
- However, it is often found in practice that the $\mu^* \in \mathcal{A}_\Delta$ that extremises $Q(\mu^*)$ does not have support on as many distinct points of $\mathcal{X}$ as the Reduction Theorem for Independence permits as an upper bound, and that not all of the constraints determining $\mathcal{A}$ hold as equalities.
- 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.

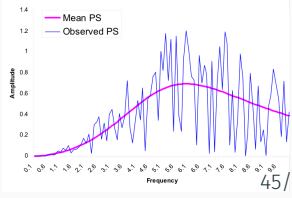
- Consider the survivability of a truss structure under an random earthquake of known intensity drawn from an incompletely specified probability distribution.
- Extending the commonly-used shape function technique, consider a random ground motion u, with the constraint that the mean power spectrum is the Matsuda–Asano shape function s_{MA}:

$$\mathbb{E}_{u\sim\mu}\left[|\hat{u}(\omega)|^2\right] = s_{\mathrm{MA}}(\omega) \propto \frac{\omega_{\mathrm{g}}^2\omega^2 e^{M_{\mathrm{L}}}}{(\omega_{\mathrm{g}}^2-\omega^2)^2 + 4\xi_{\mathrm{g}}^2\omega_{\mathrm{g}}^2\omega^2}.$$

- We used 200 3d Fourier modes, leading to a 1200-dimensional OUQ problem.







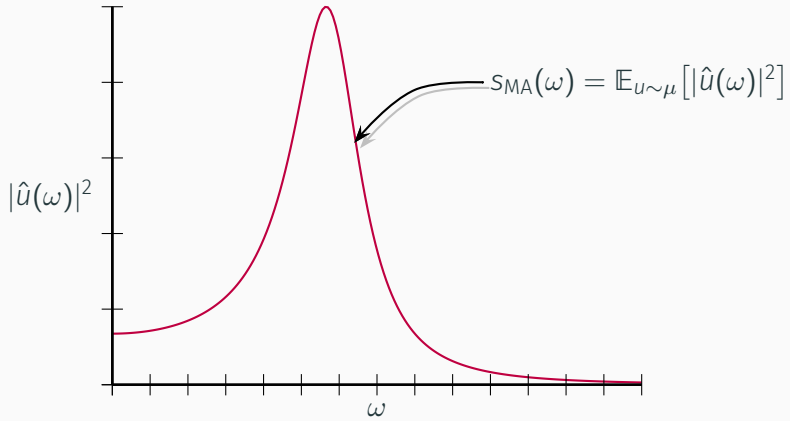


Figure 3: One mean constraint on each independent random Fourier mode $\hat{u}(\omega)$ (i.e. that $\mathbb{E}_{u\sim\mu}\left[|\hat{u}(\omega)|^2\right] = s_{\mathrm{MA}}(\omega)$) $\implies$ we get to pretend that $u(\omega)$ takes at most two distinct values which together satisfy this mean constraint.

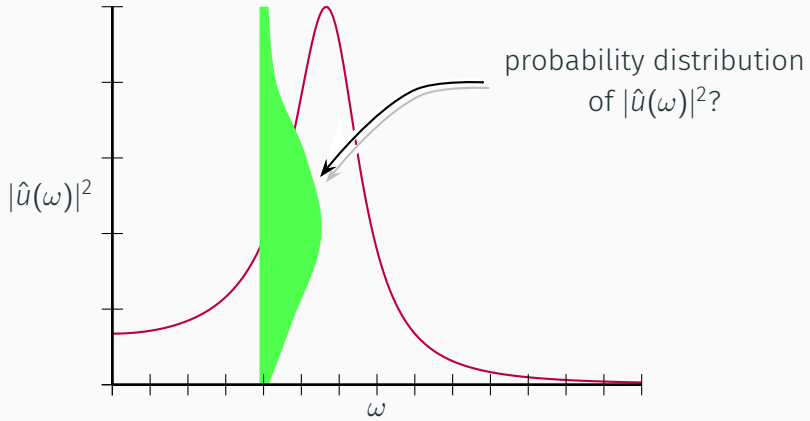


Figure 3: One mean constraint on each independent random Fourier mode $\hat{u}(\omega)$ (i.e. that $\mathbb{E}_{u\sim\mu}\left[|\hat{u}(\omega)|^2\right] = s_{\mathrm{MA}}(\omega)$) $\implies$ we get to pretend that $u(\omega)$ takes at most two distinct values which together satisfy this mean constraint.

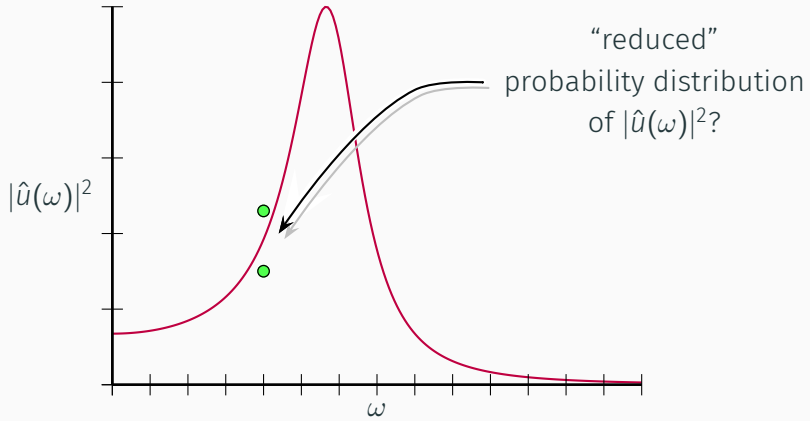


Figure 3: One mean constraint on each independent random Fourier mode $\hat{u}(\omega)$ (i.e. that $\mathbb{E}_{u\sim\mu}\left[|\hat{u}(\omega)|^2\right] = s_{\mathrm{MA}}(\omega)$) $\implies$ we get to pretend that $u(\omega)$ takes at most two distinct values which together satisfy this mean constraint.

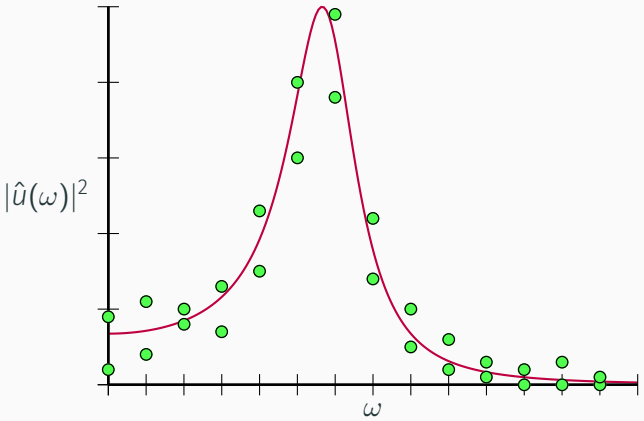


Figure 3: One mean constraint on each independent random Fourier mode $\hat{u}(\omega)$ (i.e. that $\mathbb{E}_{u\sim\mu}\left[|\hat{u}(\omega)|^2\right] = s_{\mathrm{MA}}(\omega)$) $\implies$ we get to pretend that $u(\omega)$ takes at most two distinct values which together satisfy this mean constraint.

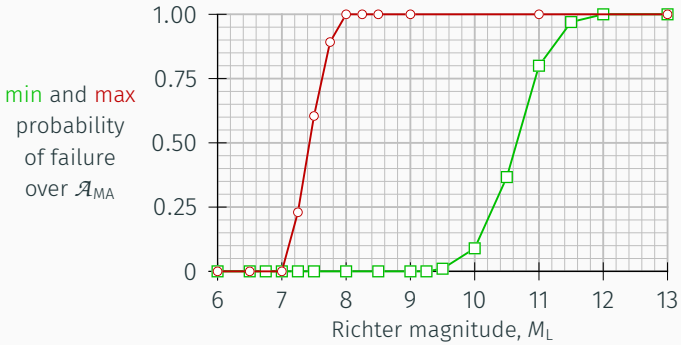


Figure 4: The minimum and maximum probability of failure as a function of Richter magnitude, M_{L}, where the ground motion u is constrained to have $\mathbb{E}_{\mu}[|\hat{u}|^2] =$ the Matsuda–Asano shape function s_{MA} with natural frequency ω_{g} and natural damping ξ_{g} taken from the 24 Jan. 1980 Livermore earthquake. Each data point required $\approx$ 1 day on 44+44 AMD Opterons (*shc* and *foxtrot* at Caltech). The forward model used 200 Fourier modes for a 3-dimensional ground motion u.

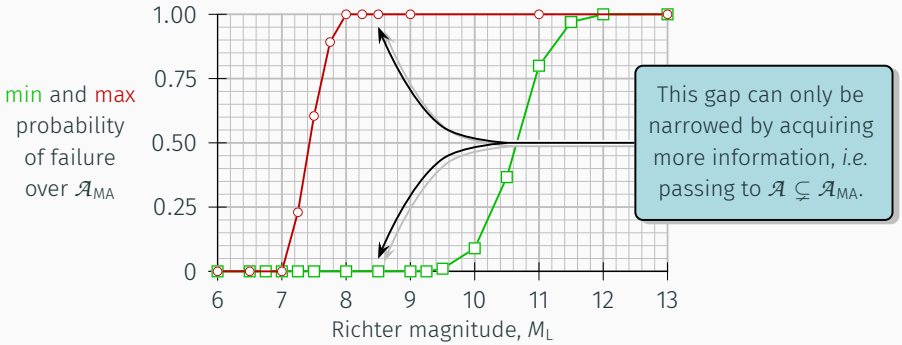


Figure 4: The minimum and maximum probability of failure as a function of Richter magnitude, M_{L}, where the ground motion u is constrained to have $\mathbb{E}_{\mu}[|\hat{u}|^2] =$ the Matsuda–Asano shape function s_{MA} with natural frequency ω_{g} and natural damping ξ_{g} taken from the 24 Jan. 1980 Livermore earthquake. Each data point required ≈ 1 day on 44+44 AMD Opterons (*shc* and *foxtrot* at Caltech). The forward model used 200 Fourier modes for a 3-dimensional ground motion u.

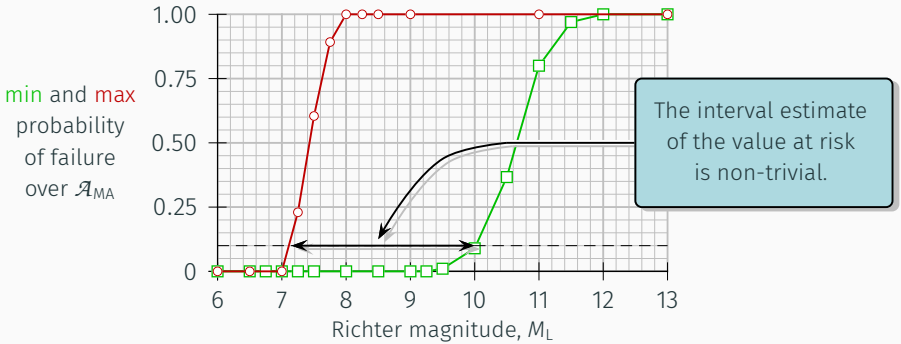


Figure 4: The minimum and maximum probability of failure as a function of Richter magnitude, M_{L}, where the ground motion u is constrained to have $\mathbb{E}_{\mu}[|\hat{u}|^2] =$ the Matsuda–Asano shape function s_{MA} with natural frequency ω_{g} and natural damping ξ_{g} taken from the 24 Jan. 1980 Livermore earthquake. Each data point required $\approx$ 1 day on 44+44 AMD Opterons (*shc* and *foxtrot* at Caltech). The forward model used 200 Fourier modes for a 3-dimensional ground motion u.

Overview

- 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 $Q(g,\mu) = \mathbb{E}_{X\sim\mu}[q(X, g(X))]$, and seek the extreme values

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

- 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 the Reduction Theorem for Independence, then

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

- The passage to discrete measures μ often enables us to finite-dimensionalise the search over g, because only the values of g on the finite set $\operatorname{supp}(\mu)$ 'matter' in computing $\mathbb{E}_{X\sim\mu}[q(X, g(X))]$.
- Instead of optimising 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:
  - 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 Lebesgue-almost everywhere. It makes no sense to restrict such 'functions' to a finite set $\operatorname{supp}(\mu)$. We have to insist that $\mathcal{G}$ is a space of functions with pointwise-defined values.
  - Conversely, it is sometimes difficult to verify whether a vector y is $g|_{\operatorname{supp}(\mu)}$ for some $g \in \mathcal{G}$; we need functions that can be extended from $\operatorname{supp}(\mu)$ to all of $\mathcal{X}$.

Theorem (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 $f\colon E \to \mathcal{H}$ is α-Hölder on E for some $0 < \alpha \leq 1$:

$$\|f(x) - f(y)\|_{\mathcal{H}} \leq d(x,y)^{\alpha} \quad \text{for all } x, y \in E. \tag{4}$$

Then there exists $F\colon \mathcal{X} \to \mathcal{H}$ such that $F|_E = f$ and (4) 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 without increasing the Hölder norm

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

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!

Example

Let $E \coloneqq \{(1,-1),(-1,1),(1,1)\} \subseteq \mathbb{R}^2$ 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 of E, $F((0,0))$ must lie within $\|\cdot\|_2$-distance 1 of all three points of $f(E)$. .....

Example

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 $\mathrm{Lip}(F) \geq \frac{2}{\sqrt{3}}$.

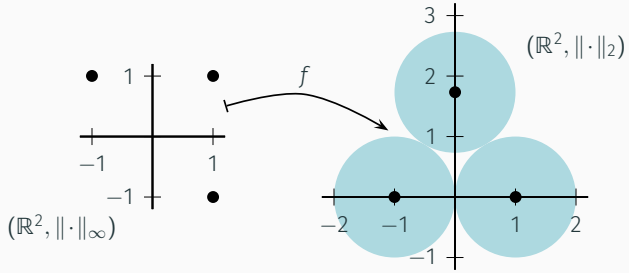


Figure 5: The function f that takes the three points on the left (equipped with $\|\cdot\|_\infty$) to the three points on the right (equipped with $\|\cdot\|_2$) has $\mathrm{Lip}(f) = 1$, but has no 1-Lipschitz extension F to $(0, 0)$, let alone the whole plane $\mathbb{R}^2$, since $F((0, 0))$ would have to lie in the (empty) intersection of the three discs.

Theorem (Reduction for measures and functions)

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 $\mathcal{A}_g = \{\mu \in \mathcal{M}_1(\mathcal{X}) \mid (g, \mu) \in \mathcal{A}\}$ is a moment class of the form considered in the Reduction Theorem for Independence. Let

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

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

Example

Suppose that $g^\dagger : [0,1] \to \mathbb{R}$ is known to have Lipschitz constant $\mathrm{Lip}(g^\dagger) \leq L$. Suppose also that the inputs of $g^\dagger$ are distributed according to $\mu^\dagger \in \mathcal{M}_1([0,1])$, and it is known that

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

and $g^\dagger$ is known on $\mathcal{O} \subset [0,1]$. Hence, the corresponding feasible set is

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

Suppose that our quantity of interest is the probability of output values below 0, i.e. $q(x,y) = \mathbb{1}[y \leq 0]$.

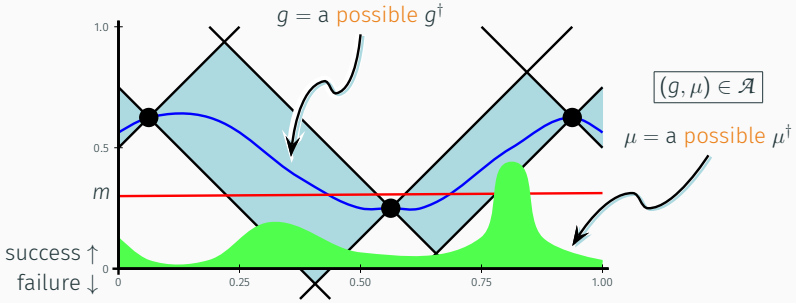
Example

By the Reduction Theorem for Independence, extremes of $Q(g,\mu) \coloneqq \mathbb{P}_{X\sim\mu}[g(X)\leq 0]$ are solutions of

$$
\begin{aligned}
\text{extremise: } & \sum_{i=0}^{1} w_i \mathbb{1}[y_i \leq 0]; \\
\text{w.r.t.: } & w_0, w_1 \geq 0, \\
& x_0, x_1 \in [0,1], \\
& y_0, y_1 \in \mathbb{R}; \\
\text{subject to: } & \sum_{i=0}^{1} w_i = 1, && |y_i - y_j| \leq L|x_i - x_j| \text{ for } i,j \in \{0,1\}, \\
& \sum_{i=0}^{1} w_i y_i \geq m, && |y_i - g^\dagger(z)| \leq L|x_i - z| \text{ for } i \in \{0,1\}, z \in \mathcal{O}.
\end{aligned}
$$

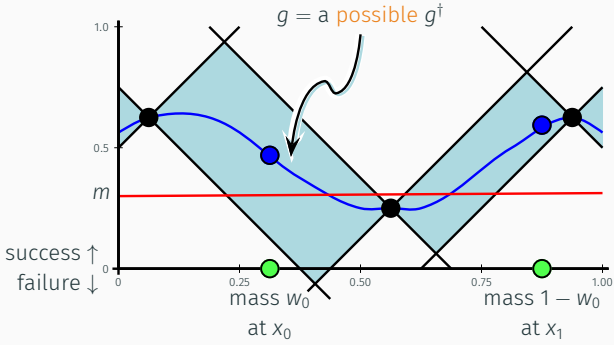
The original problem entails optimizing over an infinite-dimensional collection of (g, μ) that could be $(g^\dagger, \mu^\dagger)$. In the reduced problem, we only have to move around and re-weight two Dirac measures (point masses) and the values of g over those two points.

infinite-dimensional problem $\rightsquigarrow$ equivalent 5-dimensional problem!



The original problem entails optimizing over an infinite-dimensional collection of (g, μ) that could be $(g^\dagger, \mu^\dagger)$. In the reduced problem, we only have to move around and re-weight two Dirac measures (point masses) and the values of g over those two points.

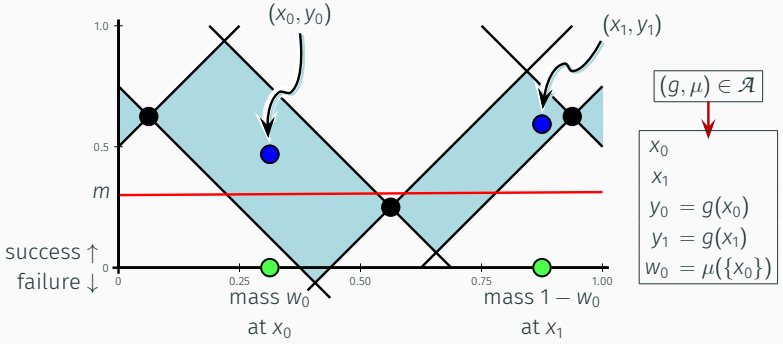
infinite-dimensional problem $\rightsquigarrow$ equivalent 5-dimensional problem!



$\boxed{(g, \mu) \in \mathcal{A}}$

The original problem entails optimizing over an infinite-dimensional collection of (g, μ) that could be $(g^\dagger, \mu^\dagger)$. In the reduced problem, we only have to move around and re-weight two Dirac measures (point masses) and the values of g over those two points.

infinite-dimensional problem $\rightsquigarrow$ equivalent 5-dimensional problem!



- The case of a single observation in 1d can be solved explicitly.
- Suppose that you have one observation $(z, g^\dagger(z)) \in [0, \frac{1}{2}] \times \mathbb{R}$ of a function $g^\dagger \colon [0,1] \to \mathbb{R}$ with Lipschitz constant $L \geq 0$.
- Explicit piecewise and discontinuous least upper bound on $\mathbb{P}_{X \sim \mu^\dagger}[g^\dagger(X) \leq 0]$ given L, $(z, g^\dagger(z))$, and that $\mathbb{E}_{X \sim \mu^\dagger}[g^\dagger(X)] \geq m$:

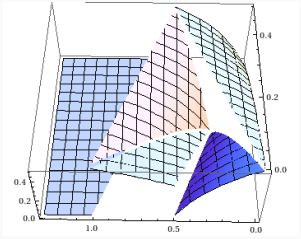


Figure 6: Surface plot of the least upper bound $\overline{P}$ on $\mathbb{P}_{\mu^\dagger}[g^\dagger \leq 0]$, as a function of the observed data point $(z, g^\dagger(z))$.

The previous example generalises to product measures on inputs, and generic legacy evaluations of $g^\dagger$ on some observation set $O \subset \mathcal{X}$. What does this feasible set 'look like', e.g. in the $2 \times 2 \times \ldots$ case?

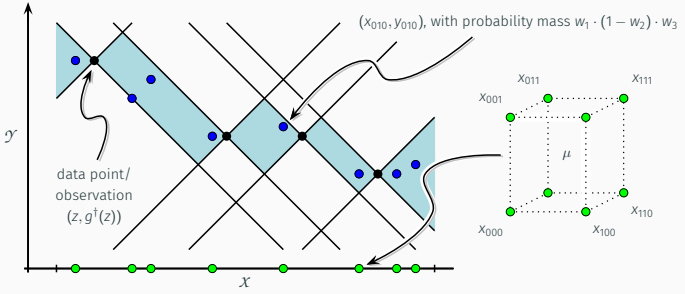


Figure 7: The black dots are the fixed locations of the legacy observations $g^\dagger|_O$. With one mean-value constraint, the grey dots are the movable locations of the 2^K support points x_i, $i \in \{0,1\}^K$, of the discrete product measure $\mu = \mu_1 \otimes \cdots \otimes \mu_K$ on $\mathcal{X}$. The white dots show some feasible values (x_i, y_i). The marginal distribution μ_k on $\mathcal{X}_k$ assigns mass w_k to x_0^k and mass $1 - w_k$ to x_1^k.

- Legacy data = 32 data points (steel-on-aluminium shots A48–A81, less two mis-fires) from summer 2010 at Caltech's SPHIR facility:

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

  Output $g^\dagger(h, \alpha, v) =$ the induced perforation area in mm^2; the data set contains results between $6.31\,\text{mm}^2$ and $15.36\,\text{mm}^2$.
- Failure event is $[g^\dagger(h, \alpha, v) \leq \theta]$, for various values of θ.
- Constrain the mean perf. area: $\mathbb{E}_{\mu^\dagger}[g^\dagger(h, \alpha, v)] \geq m := 11.0\,\text{mm}^2$.
- Modified Lipschitz constraint (multi-valued data):

$$L = \left(\frac{175.0}{\text{in}}, \frac{0.075}{\text{deg}}, \frac{0.1}{\text{m/s}}\right) \text{mm}^2$$

$$|y - y'| \leq \sum_{k=1}^{3} L_k |x_k - x'_k| + 1.0\,\text{mm}^2.$$

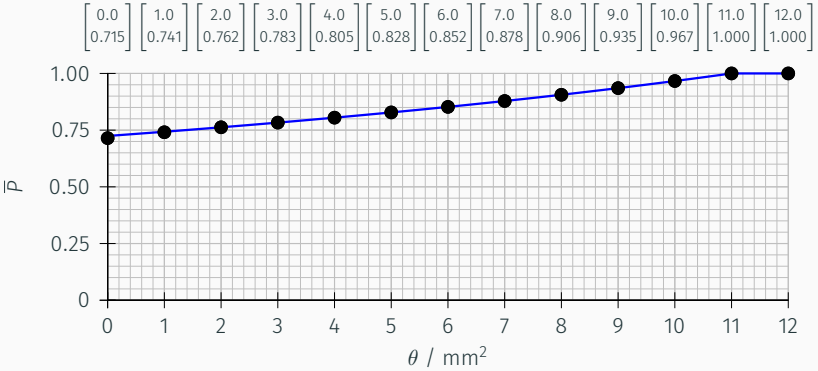


Figure 8: Maximum probability that perforation area is $\leq \theta$, for various θ, with the data and assumptions of the previous slide, including mean perforation area $\mathbb{E}[g^\dagger(h, \alpha, v)] \geq 11.0\,\text{mm}^2$. For $\theta \geq 2\,\text{mm}^2$, the results are within 10^{-6} of Markov's bound, which indicates that 2 binding data points are those that constrain the maximum of the response function; the other 30 are non-binding.

- Consider a bounded function $g\colon \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K \to \mathbb{R}$.
- The k^{th} McDiarmid subdiameter of g (a seminorm):

$$\begin{aligned}\mathcal{D}_k[g] &:= \sup\left\{ |g(x) - g(x')| \;\middle|\; \begin{array}{c} x, x' \in \mathcal{X} \\ x_i = x'_i \text{ for } i \neq k \end{array} \right\} \\ &= \sup\left\{ \begin{array}{l} |g(\ldots, x_{k-1}, x_k, x_k, \ldots) \\ \quad -g(\ldots, x_{k-1}, x'_k, x_k, \ldots)| \end{array} \;\middle|\; \begin{array}{c} x_i \in \mathcal{X}_i \text{ for } i = 1, \ldots, K, \\ x'_k \in \mathcal{X}_k \end{array} \right\}.\end{aligned}$$

- McDiarmid's concentration-of-measure inequality (McDiarmid, 1989) says that if $X = (X_k)_{k=1}^K$ has independent components X_k taking values in $\mathcal{X}_k$, then

$$\mathbb{P}[g(X) \leq 0] \leq \exp\left(-\frac{2 \max\{0, \mathbb{E}[g(X)]\}^2}{\sum_{k=1}^K \mathcal{D}_k[g]^2} \right).$$

Example (McDiarmid)

Consider the following admissible set of response functions and product measures on their inputs

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

Let $m_+ \coloneqq \max\{0, m\}$. This $\mathcal{A}_{\mathrm{McD}}$ is the admissible set corresponding to the assumptions of McDiarmid's inequality, which is the upper bound

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

Example (McDiarmid)

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.

- For $K = 1$,

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

Example (McDiarmid)

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.

- For $K = 2$,

$$\overline{Q}(\mathcal{A}_{\text{McD}}) = \begin{cases} 0, & \text{if } D_1 + D_2 \leq m_+, \\ \frac{(D_1+D_2-m_+)^2}{4D_1D_2}, & \text{if } |D_1 - D_2| \leq m_+ \leq D_1 + D_2, \\ 1 - \frac{m_+}{\max\{D_1,D_2\}}, & \text{if } 0 \leq m_+ \leq |D_1 - D_2|. \end{cases}$$

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 inequality $\mathcal{D}_2[g] \leq D_2$ is non-binding information and does not affect the global uncertainty.

Example (McDiarmid)

- 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 optimisation problem for $\overline{Q}(\mathcal{A}_{\text{McD}})$.
- It is also possible to show that if we additionally specify that f is linear (Hoeffding's inequality), then this is non-binding information in the case $K = 2$ but is binding information in the case $K \geq 3$.

- The phenomenon observed in the legacy data hypervelocity impact example, and in the $K = 2$ solution of the optimal McDiarmid inequality, occurs in many contexts: not all of the constraints that specify $\mathcal{A}$ necessarily hold as binding or active constraints at the extremizing solution $(g^*, \mu^*) \in \mathcal{A}$.
- 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!
- 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 $\mathcal{A}'$.

Overview

- 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.
- The approach of UQ advocated here falls under the umbrella of imprecise probability, the origins of which date back to very early works on probability like those of Boole (1854) and Keynes (1921). More recent foundations and expositions for imprecise probability have been put forward by Walley (1991), Kuznetsov (1991), Weichselberger (2000), and by Dempster (1967) and Shafer (1976).
- 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.

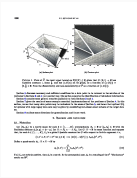
- Convex optimisation approaches to distributional robustness and optimal probability inequalities are also considered by Bertsimas and Popescu (2005). There is also an extensive literature on the related topic of majorization, for which see the book of Marshall et al. (2011).
- A standard short reference on Choquet theory is the book of Phelps (2001). The Choquet–Kendall theorem 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 optimisation of measure affine functionals, are due to von Weizsäcker and Winkler (1979/80, 1980) and Winkler (1988). Karr (1983) proved similar results under additional topological and continuity assumptions.

- Theorem 17 and the Lipschitz version of Theorem 21 can be found in Owhadi et al. (2013) and Sullivan et al. (2013) respectively. Theorem 19 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:
  - applications to hypervelocity impact: Owhadi et al. (2013), Sullivan et al. (2013), and Kamga et al. (2014);
  - applications to seismic safety certification: Owhadi et al. (2013);
  - application to power grid optimisation: Han et al. (2015);
  - applications to the robustness of Bayesian inference: Owhadi et al. (2015a,b).
- Corresponding software can be found in the examples section of the *mystic* optimisation framework at

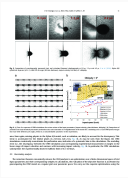
`http://github.com/uqfoundation/mystic`

Acknowledgements

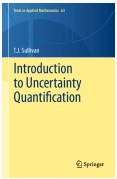
Owhadi et al. (2013)



Sullivan et al. (2013)



Kamga et al. (2014)



Sullivan (2015)

These lectures are an abridged selection of material from Sullivan (2015), and draw upon joint work with Paul-Hervé Kamga, Lan Nguyen, Mike McKerns, Dominik Meyer, Michael Ortiz, Houman Owhadi, Clint Scovel, and Florian Theil in 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.

Bibliography I

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. Theory and applications, A Wiley-Interscience Publication.

J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994. doi: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. doi:10.1137/S1052623401399903.

G. Boole. *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberley, London, 1854.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

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

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

A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38:325–339, 1967.

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

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

Bibliography II

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. doi: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.

A. F. Karr. Extreme points of certain sets of probability measures, with applications. *Math. Oper. Res.*, 8(1):74–85, 1983. doi:10.1287/moor.8.1.74.

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

J. M. Keynes. *A Treatise on Probability*. Macmillan and Co., London, 1921.

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

V. P. Kuznetsov. *Intervalnye statisticheskie modeli*. "Radio i Svyaz'", Moscow, 1991.

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. doi: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. doi:10.1090/S0002-9904-1934-05978-0.

Bibliography III

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

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.

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. doi: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. doi:10.1137/130938633.

R. R. Phelps. *Lectures on Choquet's Theorem*, volume 1757 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, second edition, 2001. doi:10.1007/b76887.

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

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.

J. E. Smith. Generalized Chebychev inequalities: theory and applications in decision analysis. *Oper. Res.*, 43(5):807–825, 1995. doi:10.1287/opre.43.5.807.

T. J. Sullivan. *Introduction to Uncertainty Quantification*, volume 63 of *Texts in Applied Mathematics*. Springer, 2015. doi:10.1007/978-3-319-23395-6.

Bibliography iv

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. doi:10.1051/m2an/2013083.

F. A. Valentine. A Lipschitz condition preserving extension for a vector function. *Amer. J. Math.*, 67(1):83–93, 1945. doi: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. doi: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.

P. Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1991.

K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *Internat. J. Approx. Reason.*, 24 (2-3):149–170, 2000. doi:10.1016/S0888-613X(00)00032-3.

G. Winkler. Extreme points of moment sets. *Math. Oper. Res.*, 13(4):581–587, 1988. doi:10.1287/moor.13.4.581.