C20 Robust Optimization

Kostas Margellos

University of Oxford

Logistics

- Who: Kostas Margellos, Control Group, IEB 50.16
  contact: `kostas.margellos@eng.ox.ac.uk`
- When: 4 lectures,
  weeks 7 & 8 – Wed & Fri @12pm
- Where: Thom Building, LR1 on Wed – LR2 on Fri
- Other info:
  - There is 1 example class: week 1 TT – Wed 28/4 @2-4pm (LR6)
  - Lecture slides available on WebLearn and on my webpage
  - Teaching style: Mix of slides and whiteboard
- Suggested textbook: "Introduction to the Scenario Approach", by Marco Campi and Simone Garatti, SIAM 2019

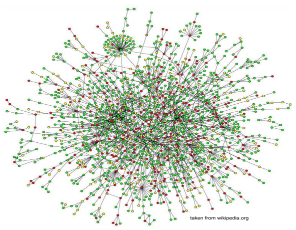
Motivation

- Social networks

- Power networks

- Robotic networks

- Biological networks



I believe we do not know anything for certain, but everything probably.

– Christiaan Huygens, 1629 – 1695

Objectives of the second part of this class

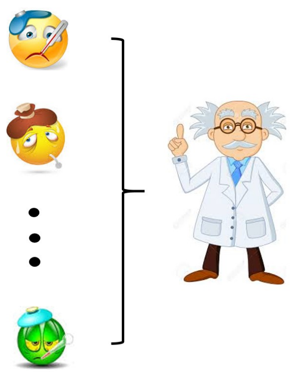
- Big picture
  - Decision making in the presence of uncertainty
  - Related to: Randomized/stochastic and robust optimization
  - Convex optimization ... and a bit of Statistical Learning Theory
- What it is actually about
  1. Introduce data based optimization
  2. Make decisions under uncertainty and accompany them with performance certificates
  3. New toolkit: easy implementation – difficulty comes in the math

How to deal with uncertainty?

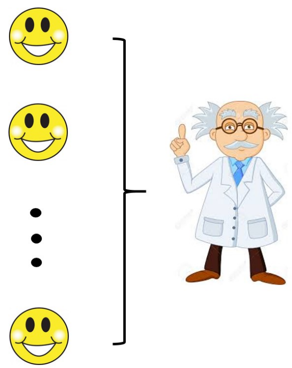
- There are many ways
  - Deterministic: Just stick with the forecasts
    Simple but agnostic!
  - Robust: Consider the worst-case
    Offers immunization but conservative!
- Let the DATA speak

'After careful consideration of all 437 charts, graphs, and metrics, I've decided to throw up my hands, hit the liquor store, and get snockered. Who's with me?!"

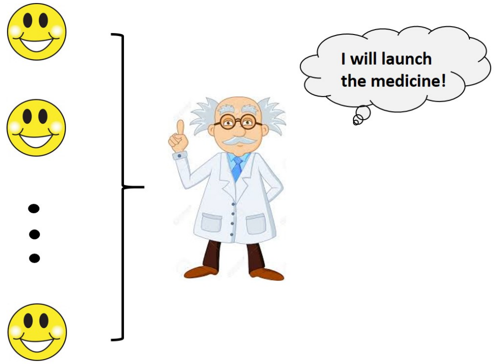
Motivation - The doctor's problem



Motivation - The doctor's problem

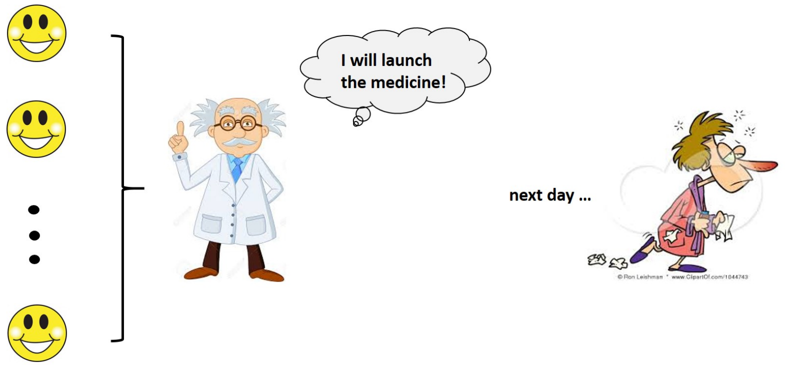


Motivation - The doctor's problem



Motivation - The doctor's problem

Motivation - The doctor's problem



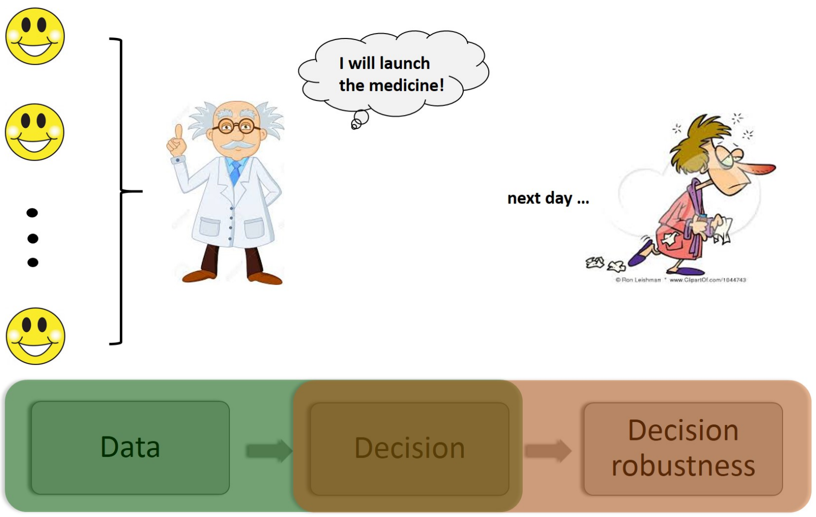
Data → Decision → Decision robustness

Motivation - The doctor's problem

Data → Decision → Decision robustness

Learning

Motivation - The doctor's problem



Learning **Generalization**

Probably Approximately Correct Learning

- Introduction to a particular notion of "learnability"
- Quantification of the notion of "generalization"
- Strong links with statistical learning theory

Terminology by means of an example

1. Consider the most popular random experiment: **coin tossing**
   - Random variable $\delta \in \{\text{Head, Tail}\}$
   - Toss a fair coin 100 times, multi-sample: $\delta_1, \ldots, \delta_{100}$ multi-extraction, instances of our random variable
   - Calculate the frequency of getting a head (empirical head probability)

$$\widehat{\mathbb{P}}_{(\delta_1,\ldots,\delta_{100})} = \frac{\#\ \text{Heads}}{\#\ \text{coin tosses}}$$

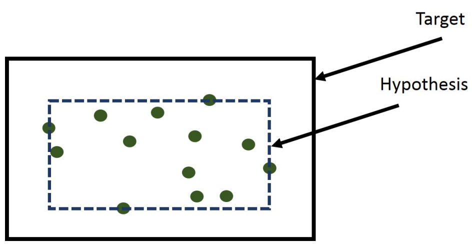
2. Repeat it the experiment 50 times
   - You will get 50 different $\widehat{\mathbb{P}}_{(\delta_1,\ldots,\delta_{100})}$: $0.55, 0.47, 0.53, \ldots$
   - $\widehat{\mathbb{P}}_{(\delta_1,\ldots,\delta_{100})}$ is itself random!
   - How likely it is that $|\widehat{\mathbb{P}}_{(\delta_1,\ldots,\delta_{100})} - 0.5|$ is very small?

Learning & Generalization question

How **many times** shall you toss the coin initially so that the empirical head probability is very close to 0.5 for most of the 50 trials?

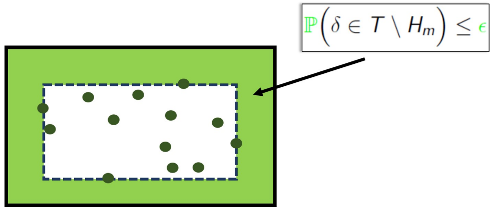
Learning

- Target set T
  - T is not known, but we are given samples $\delta_1, \ldots, \delta_m$ contained in T
  - *Example:* Consider T to be an axis-aligned rectangle
- Hypothesis H_m (also a set)
  - Depends on multi-sample $\delta_1, \ldots, \delta_m$
  - Provides an approximation of T
  - *Example:* Smallest axis-aligned rectangle that contains the samples



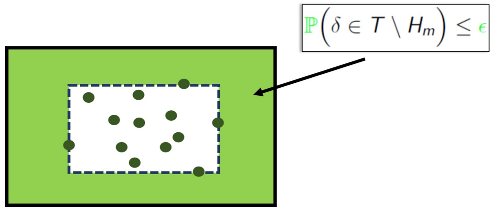
Generalization – Probably Approximately Correct Learning

- **Approximately:** T and H_m very close
  - How likely is it that H_m does not contain another sample δ (extracted according to $\mathbb{P}$)?
  - Depends on the "distance" $\mathbb{P}\Big(\delta \in T \setminus H_m\Big)$
  - ☺ if $\mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon$ (shaded region)
- **Probably:** T and H_m very close for most of the multi-samples
  - H_m is itself random as it depends on the samples
  - What is the probability that $\mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon$?
  - In other words, for "how many" of the multi-samples is this the case?



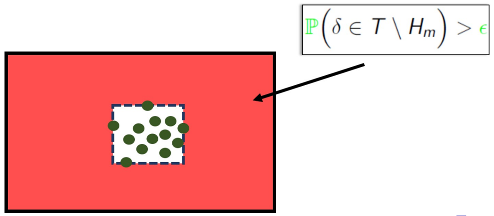
Generalization – Probably Approximately Correct Learning

- **Approximately:** T and H_m very close
  - How likely is it that H_m does not contain another sample δ (extracted according to $\mathbb{P}$)?
  - Depends on the "distance" $\mathbb{P}\Big(\delta \in T \setminus H_m\Big)$
  - ☺ if $\mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon$ (shaded region)
- **Probably:** T and H_m very close for most of the multi-samples
  - H_m is itself random as it depends on the samples
  - What is the probability that $\mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon$?
  - In other words, for "how many" of the multi-samples is this the case?



Generalization – Probably Approximately Correct Learning

- **Approximately:** T and H_m very close
  - How likely is it that H_m does not contain another sample δ (extracted according to $\mathbb{P}$)?
  - Depends on the "distance" $\mathbb{P}\Big(\delta \in T \setminus H_m\Big)$
  - ☺ if $\mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon$ (shaded region)
- **Probably:** T and H_m very close for most of the multi-samples
  - H_m is itself random as it depends on the samples
  - What is the probability that $\mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon$?
  - In other words, for "how many" of the multi-samples is this the case?



Generalization

- In the doctor's problem: Doctor would be satisfied if ...
  - Medicine cures patients with probability at least $1 - \epsilon$
    ... or, probability that a new patient δ is not cured, is at most ϵ
  - If this holds with probability at least $1 - q(m, \epsilon)$ with respect to the $\delta_1, \dots, \delta_m$ trial patients

Problem

Find conditions for the existence of some $q(m, \epsilon)$ such that

$$\mathbb{P}^m\Big\{\delta_1, \dots, \delta_m : \ \mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon\Big\} \ge 1 - q(m, \epsilon)$$

and $\lim_{m \to \infty} q(m, \epsilon) = 0$.

- Probability T and H_m being different *at most* ϵ, occurs with confidence *at least* 1-$q(m, \epsilon)$
- We have implicitly assumed that $T \supseteq H_m$; this is for simplicity, otherwise we should use $\mathbb{P}\Big(\delta \in (T \setminus H_m) \cup (H_m \setminus T)\Big)$

Generalization

Problem

Find conditions for the existence of some $q(m,\epsilon)$ such that

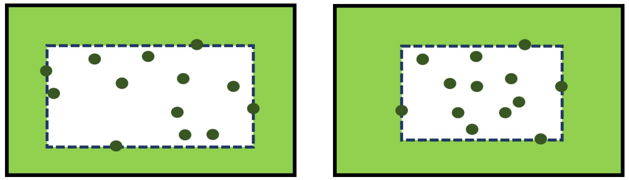
$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m:\ \mathbb{P}\Big(\delta\in T\smallsetminus H_m\Big)\le\epsilon\Big\}\ge 1-q(m,\epsilon)$$

and $\lim_{m\to\infty} q(m,\epsilon)=0$.

- Probability of a "new" δ: $\mathbb{P}$
- Probability of an m multisample $\delta_1,\dots,\delta_m$: $\mathbb{P}\times\dots\times\mathbb{P}=\mathbb{P}^m$ product probability as all samples are independent from each other
- **Confidence** $1-q(m,\epsilon)$. It depends on the number of samples m and the **violation level** ϵ, The more samples we are provided, the closer it is to 1, i.e. $\lim_{m\to\infty} q(m,\epsilon)=0$

Generalization - sufficient condition

- Observation
  - For any m multi-sample often *only* a subset of them matters



- Axis-aligned rectangle example
  - The hypothesis H_m is determined only by the samples on the facets
  - Different multi-samples, but always 4 are needed to determine the hypothesis (but for degenerate cases)!

Generalization - sufficient condition

- Fix $d<m$
- Denote by $C_d\subset\{\delta_1,\dots,\delta_m\}$ a subset of the multi-sample with cardinality d, i.e. $|C_d|=d$
- Let H_d bet the hypothesis constructed using **only** the samples in C_d

Compression set

C_d with $|C_d|=d<m$ is called a compression set if

$$\mathbb{1}_{H_d}(\delta_i)=\mathbb{1}_T(\delta_i),\quad \text{for all } i=1,\dots,m$$

- Hypothesis H_d agrees with the target T on all samples, i.e. existence of a compression set $\Leftrightarrow$ **Empirical generalization**
- Indicator function

$$\mathbb{1}_T(\delta)=\begin{cases}1 & \text{if } \delta\in T\\ 0 & \text{otherwise}\end{cases}$$

Generalization - sufficient condition

Compression set

Assume that for any m multi-sample there exists C_d with $|C_d|=d<m$ such that

$$\mathbb{1}_{H_d}(\delta_i)=\mathbb{1}_T(\delta_i),\quad \text{for all } i=1,\dots,m$$

C_d is then called a compression set.

- Existence of a compression set $\Leftrightarrow$ **Empirical generalization**
  - We approximate T with H_d using only d samples
  - This hypothesis agrees with T on all other samples as well, i.e. approximation error on the samples is zero
  - We do *not* need to know C_d; we only care that such a set exists

Recall our problem ...

Problem

Find conditions for the existence of some $q(m,\epsilon)$ such that

$$\mathbb{P}^m\Big\{\delta_1,\ldots,\delta_m:\ \mathbb{P}\Big(\delta\in T\setminus H_m\Big)\le\epsilon\Big\}\ge 1-q(m,\epsilon)$$

and $\lim_{m\to\infty} q(m,\epsilon)=0$.

Generalization

Theorem

If a **compression set** C_d **with cardinality** d exists, then

$$\mathbb{P}^m\Big\{\delta_1,\ldots,\delta_m:\ \mathbb{P}\Big(\delta\in T\setminus H_m\Big)\le\epsilon\Big\}\ge 1-q(m,\epsilon)$$

with $q(m,\epsilon)=\binom{m}{d}(1-\epsilon)^{m-d}$.

- Hypothesis probably approximately correct (PAC) learns target
- We do not care about C_d but only about d
- It holds $\lim_{m\to\infty} q(m,\epsilon)=0$

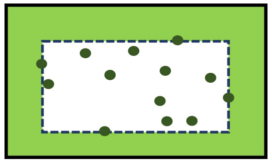
$$\lim_{m\to\infty} q(m,\epsilon)=\binom{m}{d}(1-\epsilon)^{m-d}$$

$$\le\lim_{m\to\infty}\Big(\frac{me}{d}\Big)^d(1-\epsilon)^{m-d}=0$$

First term increases polynomially; second term tends to zero exponentially fast (dominant)

Intermediate summary

- **Learning:** Approximate target T with hypothesis H_m
- **Generalization:** Find confidence $1-q(m,\epsilon)$ such that hypothesis is an ϵ-good approximation of the target, i.e. $\mathbb{P}\Big(\delta\in T\setminus H_m\Big)\le\epsilon$



- **Compression:** Only the important samples (the $d=4$ boundary ones in the rectangle example)
- Produces the same hypothesis with the one that would be obtained if all samples were used, i.e. $H_d=H_m$
- Target T and hypothesis H_d agree on all samples, i.e. approximation error on the samples is zero

Intermediate summary

Theorem

If a **compression set** C_d **with cardinality** d exists, then

$$\mathbb{P}^m\Big\{\delta_1,\ldots,\delta_m:\ \mathbb{P}\Big(\delta\in T\setminus H_m\Big)\le\epsilon\Big\}\ge 1-q(m,\epsilon)$$

with $q(m,\epsilon)=\binom{m}{d}(1-\epsilon)^{m-d}$, where $\lim_{m\to\infty} q(m,\epsilon)=0$.

- Hypothesis probably approximately correct (PAC) learns target
- We do not care about C_d but only about d
- It is a distribution-free result; holds true for any underlying (possibly unknown) distribution, as long as data are independently extracted
- **If a compression set exists:**
  H_m and T fully agree on the samples $\Rightarrow$ ϵ-agree for another δ.
  Empirical generalization $\Rightarrow$ Probabilistic generalization

On the cardinality of the compression set

Theorem

If a **compression set** C_d **with cardinality** d exists, then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon\Big\} \ge 1 - q(m,\epsilon)$$

with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$.

- Does the cardinality d of the compression set matter?

$$\lim_{d\to m} 1 - q(m,\epsilon) = 1 - \lim_{d\to m}\binom{m}{d}(1-\epsilon)^{m-d} = 0$$

- As the compression "increases" the confidence $1 - q(m,\epsilon)$ tends to 1 $\Rightarrow$ result trivial (not useful) as we claim that H_m is an ϵ-good approximation of T with positive probability!
- The smaller the compression the more useful the result!

Generalization – Stronger statement

Theorem

If there exists a **unique compression set** C_d **with cardinality** d, then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon\Big\} \ge 1 - q(m,\epsilon)$$

with $q(m,\epsilon) = \sum_{k=0}^{d-1}\binom{m}{k}\epsilon^k(1-\epsilon)^{m-k}$.

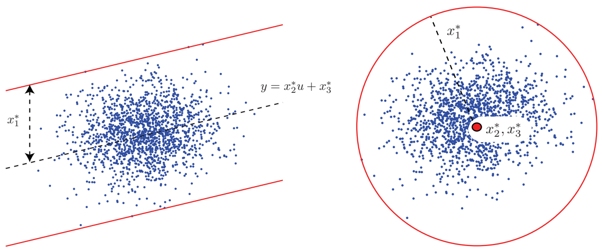
- Stronger assumption $\Longrightarrow$ stronger statement
- For the same m and $\epsilon \in (0,1)$,

$$\sum_{k=0}^{d-1}\binom{m}{k}\epsilon^k(1-\epsilon)^{m-k} < \binom{m}{d}(1-\epsilon)^{m-d},$$

i.e. we can claim the probabilistic result with higher confidence $1 - q(m,\epsilon)$

Generalization – Stronger statement

- Minimum width strip vs. minimum radius disk (assume continuous distribution) – figures taken from [Campi & Garatti, 2008]



- Does there exist a **unique** compression set with cardinality $d = 3$?
- For both problems a compression set with 3 exists, i.e. $\Rightarrow d = 3$
- For the disk problem, this not unique (hence the result not tight):
  - For almost all multi-samples, only 2 matter – the most isolated ones
  - Take the 2 most isolated samples and pick 1 from all inbetween samples $\Rightarrow$ many compression sets with cardinality 3

Generalization – Complementary statements

- Probability T and H_m being different higher than ϵ, occurs with confidence *at most* $q(m,\epsilon)$

Theorem

If a **compression set** C_d **with cardinality** d exists, then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \mathbb{P}\Big(\delta \in T \setminus H_m\Big) > \epsilon\Big\} \le q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}.$$

Theorem

If there exists a **unique compression set** C_d **with cardinality** d, then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \mathbb{P}\Big(\delta \in T \setminus H_m\Big) > \epsilon\Big\} \le q(m,\epsilon) = \sum_{k=0}^{d-1}\binom{m}{k}\epsilon^k(1-\epsilon)^{m-k}.$$

From learning to optimization under uncertainty

- Uncertain scenario programs
- Probabilistic guarantees on constraint satisfaction
- The convex case (a compression set exists)

Optimization under uncertainty

- Uncertain program

$$\begin{aligned} \min_{x \in \mathbb{R}^{n_x}} \quad & c^\top x \\ \text{subject to:} \quad & \\ & g(x,\delta) \le 0, \text{ for all } \delta \in \Delta \end{aligned}$$

- Description of the uncertainty
  - Uncertain vector $\delta \in \mathbb{R}^{n_\delta}$, distributed according to $\mathbb{P}$
  - Δ denotes the set of values δ can take with non-zero probability
- Finite number of decision variables $x \in \mathbb{R}^{n_x}$ but infinite constraints (one per element of Δ, and Δ might be a continuous set)
- Either Δ is unknown, or infinite constraints $\Longrightarrow$ In general not solvable!

Data based optimization

- Uncertain scenario program

$$\begin{aligned} \min_{x \in \mathbb{R}^{n_x}} \quad & c^\top x \\ \text{subject to:} \quad & \\ & g(x,\delta_i) \le 0, \text{ for all } i = 1,\dots,m \end{aligned}$$

- Description of the uncertainty
  - Represent uncertainty $\delta \in \mathbb{R}^{n_\delta}$, by an m multi-sample $(\delta_1,\dots,\delta_m)$
  - All samples are independent from each other from the same distribution
- Finite number of decision variables $x \in \mathbb{R}^{n_x}$ and finite number of constraints (one per sample δ_i)
- Solvable! Denote by x_m^* its minimizer

Data based optimization as a learning problem

- Uncertain program

$$\begin{aligned} \min_{x \in \mathbb{R}^{n_x}} \quad & c^\top x \\ \text{subject to:} \quad & \\ & g(x,\delta_i) \le 0, \text{ for all } i = 1,\dots,m \end{aligned}$$

- Connections with learning – Learn the uncertainty space Δ

Target set	$T = \Delta$, (i.e. $\mathbb{1}_T(\delta) = 1, \ \forall \delta \in \Delta$)
Decision	Minimizer $\Rightarrow x_m^*$
Hypothesis	$H_m = \left(\delta \in \Delta : \ g(x_m^*,\delta) \le 0\right)$

- Hypothesis: The set of δ's for which x_m^* remains feasible
- In other words, the subset of the uncertainty space for which constraint satisfaction is ensured for x_m^*

Data based optimization as a learning problem

- Uncertain program

$$\min_{x\in\mathbb{R}^{n_x}} \; c^\top x$$
$$\text{subject to:} \quad g(x,\delta_i) \le 0, \text{ for all } i = 1,\dots,m$$

- Connections with learning – Learn the uncertainty space Δ

Target set	$T = \Delta$, (i.e. $\mathbb{1}_T(\delta) = 1, \ \forall \delta \in \Delta$)
Decision	Minimizer $\Rightarrow x_m^*$
Hypothesis	$H_m = \Big(\delta \in \Delta : \ g(x_m^*, \delta) \le 0\Big)$

- Approximation error = Probability of constraint violation for x_m^*

$$\mathbb{P}\Big(\delta \in T \setminus H_m\Big) = \mathbb{P}\Big(\delta \in \Delta : \ g(x_m^*, \delta) > 0\Big)$$

Data based optimization – Generalization

Theorem (the abstract version)

If a **compression set** C_d **with cardinality** d exists, then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \ \mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon\Big\} \ge 1 - q(m,\epsilon)$$

with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$.

Theorem (the optimization version)

If a **compression set** C_d **with cardinality** d exists, then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \ \mathbb{P}\Big(\delta \in \Delta : \ g(x_m^*, \delta) > 0\Big) \le \epsilon\Big\} \ge 1 - q(m,\epsilon)$$

with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$.

Scenario vs. Uncertain programs

Probabilistic feasibility

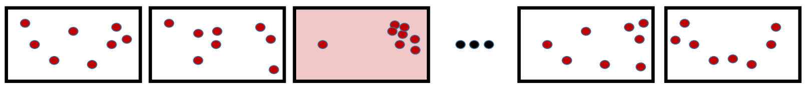
Data based program

$$\min_{x\in\mathbb{R}^{n_x}} \; c^\top x$$
subject to $\quad \rightarrow x_m^*$
$$g(x,\delta_i) \le 0, \ \forall i = 1,\dots,m$$

Robust program

$$\min_{x\in\mathbb{R}^{n_x}} \; c^\top x$$
subject to
$$g(x,\delta) \le 0, \ \forall \delta \in \Delta$$

- Is x_m^* feasible for the uncertain program? No!
- Is this true for any m multi-sample? Yes, with confidence $1 - q(m,\epsilon)$



Scenario vs. Uncertain programs

Probabilistic feasibility

Data based program

$$\min_{x\in\mathbb{R}^{n_x}} \; c^\top x$$
subject to $\quad \rightarrow x_m^*$
$$g(x,\delta_i) \le 0, \ \forall i = 1,\dots,m$$

Robust program

$$\min_{x\in\mathbb{R}^{n_x}} \; c^\top x$$
subject to
$$g(x,\delta) \le 0, \ \forall \delta \in \Delta$$

- The link is our theorem: Probabilistic robustness
  With certain confidence, the probability that a new δ appears and x_m^* (generated based on $\delta_1,\dots,\delta_m$) violates the corresponding constraint, i.e. $g(x_m^*,\delta) > 0$, is at most ϵ

If a **compression set** C_d **with cardinality** d exists, then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \ \mathbb{P}\Big(\delta \in \Delta : \ g(x_m^*, \delta) > 0\Big) \le \epsilon\Big\} \ge 1 - \binom{m}{d}(1-\epsilon)^{m-d}$$

Convex uncertain programs

$$\min_{x \in \mathbb{R}^{n_x}} \quad c^\top x$$
subject to:
$$g(x, \delta_i) \le 0, \text{ for all } i = 1, \ldots, m$$

- For any $\delta \in \Delta$, $g(x, \delta)$ is convex in x
- Existence of a compression set: Minimizer with d samples coincides with minimizer with m samples, i.e. $x_d^* = x_m^*$ so that $H_d = H_m$

For convex programs a compression set always exists:

- $d \le \#$ decision variables n_x
- If $d = n_x$ then result is "tight" (i.e. non-conservative)
- This bound is based on the notion of support constraints (very close to the active constraints)

Probabilistic feasibility for convex scenario programs

Theorem – Convex scenario programs

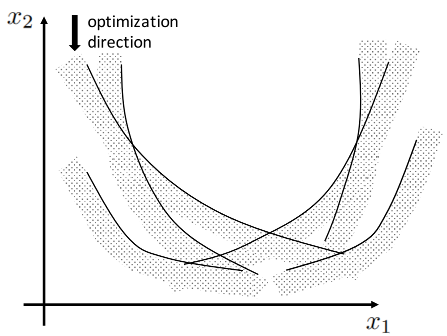
Let d be the # of decision variables in a convex scenario program. Then

$$\mathbb{P}^m \Big\{ \delta_1, \ldots, \delta_m : \ \mathbb{P}\Big(\delta \in \Delta : \ g(x_m^*, \delta) > 0 \Big) \le \epsilon \Big\} \ge 1 - q(m, \epsilon)$$

with $q(m, \epsilon) = \binom{m}{d} (1 - \epsilon)^{m-d}$.

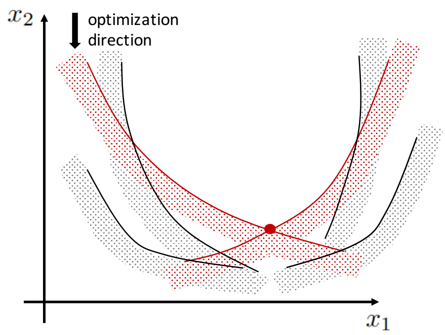
- Cardinality of the compression set d is equal to the # of decision variables in a convex scenario program
- Convex scenario programs with different objective and constraint function could share the same feasibility guarantees if they have the same number of decision variables
  $\Rightarrow$ only for some of them the confidence bound would be tight!

Compression set: 2D example



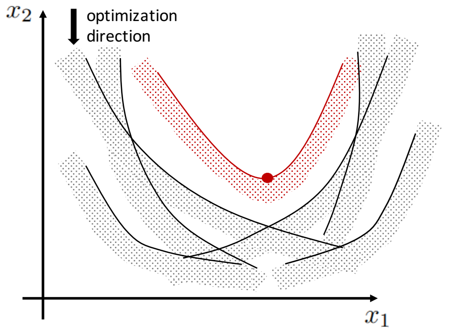
- Example with two decision variables x_1, x_2
- Objective: minimize x_2 (see optimization direction)
- Feasibility region *outside* the shaded part

Compression set: 2D example



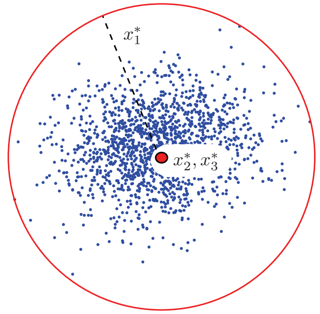
- Compression set cardinality $d = n_x$
- Compression set = Two active constraints
  $\Rightarrow$ If any of the two red constraints is removed the solution drifts to a lower value (intersection of the remaining red with a lower constraint)
- Compression set coincides with "red" constraints $\Longrightarrow x_{\text{red}}^* = x_m^*$

Compression set: 2D example



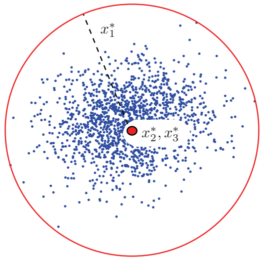
- Compression set cardinality $d \leq n_x$ (always)
- Compression set = One active constraint
  $\Rightarrow$ If any of the other constraints are removed the solution remains unaltered; only the red constraint is needed
- We again have that $x^*_{\text{red}} = x^*_m$

Example



- $m = 1650$ points (u_i, y_i) are given – the underlying distribution is unknown
- Consider the disk with the smallest radius that contains all of them
- **What guarantees can you offer that this disk contains 99% of all possible points extracted from the same distribution (other than the data points)?**

Example (cont'd)

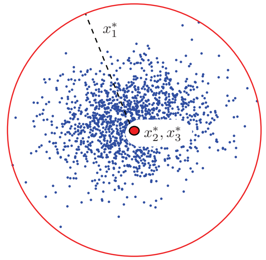


- Construct the minimum radius disk program (d=3 decision variables)

$$\min_{x_1,x_2,x_3} \; x_1$$
$$\text{subject to: } \sqrt{(y_i - x_3)^2 + (u_i - x_2)^2} \leq x_1, \text{ for all } i = 1, \ldots, 1650$$

- All samples should be within the x_1 radius disk; (x_2, x_3) parametrize its center
- Decision variables: x_1, x_2, x_3; Samples: $\delta_i = (u_i, y_i)$, $i = 1, \ldots, 1650$

Example (cont'd)



- Construct the minimum radius disk program (d=3 decision variables)

$$\min_{x_1,x_2,x_3} \; x_1$$
$$\text{subject to: } \sqrt{(y_i - x_3)^2 + (u_i - x_2)^2} \leq x_1, \text{ for all } i = 1, \ldots, 1650$$

- Disk should contain 99% of new points $\delta = (u, y) \Rightarrow \epsilon = 0.01$
- Hence the "guarantee" is the confidence
  $1 - q(1650, 0.01) = 1 - \binom{1650}{3}(1 - 0.01)^{1650-3}$

Convex scenario programs

- Relationship between compression set and support constraints
- Bound on the cardinality of the compression set (Helly's Theorem)
- Distribution of the probability of constraint violation

Convex scenario programs

$$\min_{x \in \mathbb{R}^{n_x}} \quad c^\top x$$
subject to:
$$g(x, \delta_i) \le 0, \text{ for all } i = 1, \dots, m$$

- For any $\delta \in \Delta$, $g(x, \delta)$ is convex in x

Definition: Compression set

A set $C_d \subset \{\delta_1, \dots, \delta_m\}$ with $|C_d| = d < m$ is a compression set if

$$x_d^* = x_m^*,$$

i.e. the minimizer with d samples is the same with the minimizer with all samples.

Definition: Support constraints

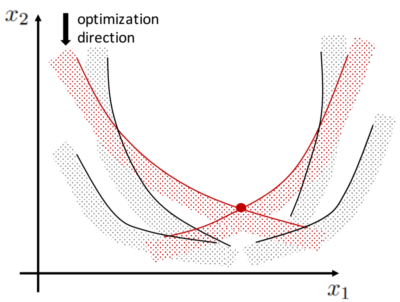
A constraint $k \in \{1, \dots, m\}$ is of support if

$$x^*_{\{\delta_1, \dots, \delta_m\} \setminus \delta_k} \neq x_m^*,$$

i.e. if we remove the k-th constraint, the solution with the remaining ones changes.

Compression set vs. Support constraints

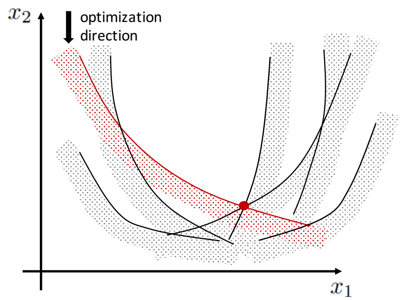
Non-degenerate problems: support constraints = compression set



- If any of the "red" constraints is removed, then the solution changes
  $\Rightarrow$ "red" constraints are support constraints
- Solving the problem **only** with the "red" constraints is the same with the solution if all constraints are taken into account

Compression set vs. Support constraints

Degenerate problems (constraints accumulate at single points):
support constraints $\subset$ compression set



- **Only if** the "red" constraints is removed, then the solution changes
  $\Rightarrow$ only "red" constraint is support constraint
- Solving the problem **only** with the "red" constraints is **not** the same with the solution if all constraints are taken into account
  $\Rightarrow$ Need to include one of the other active ones in the compression set

Compression set vs. Support constraints

Facts: Compression set for convex scenario programs

1. It always exists and has cardinality is $d \leq n_x$,
   i.e. at most equal to the # of decision variables
2. For non-degenerate problems: support constraints = compression set
3. For degenerate problems: support constraints $\subset$ compression set
4. For any convex problem: support constraints $\subseteq$ active constraints

- We will assume that any given scenario program is non-degenerate
  Compression set = Support constraints
- In case of a degenerate problem we could slightly perturb the constraints (constraint "heating")
- For continuous probability distributions (in fact distributions that admit density) convex degenerate problems occur with probability zero

Compression set for non-degenerate convex problems

Theorem: Bound on compression set cardinality

For non-degenerate convex scenario programs, for a compression set C_d it holds

1. $|C_d| = d \leq n_x$ (# of decision variables)
2. ... or equivalently, since compression set = support constraints
   # support constraints $\leq n_x$

We will make use of the following theorem

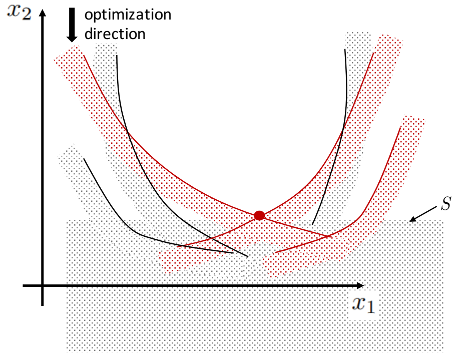
Helly's theorem (fundamental result in convex analysis)

Consider any finite number of convex sets in $\mathbb{R}^{n_x}$. If every collection of $n_x + 1$ sets has a non-empty intersection, then all of them have a non-empty intersection.

How is this relevant?

Proof

- We will apply Helly's theorem with $n_x = 2$ (similarly for higher n_x)
- Consider the family of sets including
  - m sets: each set is the feasibility region for each constraint (non-shaded part of each parabola)
  - set S: shaded region *not* including x_m^*, i.e. all points that have a lower value than x_m^* (i.e. $c^\top x < c^\top x_m^*$)



Proof (cont'd)

1. For the sake of contradiction assume that a third support constraint exists (e.g. lower red one in the figure)
2. To apply Helly's theorem take any $n_x + 1 = 3$ sets from our collection and show that they have a non-empty intersection

Case A: Take any $n_x + 1 = 3$ sets the parabolic ones.
As the overall problem is feasible, by construction their intersection is non-empty

Case B: Take now 2 of the parabolic sets and S.

- As we have assumed 3 support constraints, one of them will be missing from the intersection
- As a support constraint is missing, then the solution changes from x_m^*, hence it will be in S (it includes points such that $c^\top x < c^\top x_m^*$)
- Therefore, any such collection will also have non-empty intersection

Proof (cont'd)

3. For any case, any collection of $n_x + 1 = 3$ sets has non-empty intersection
4. By Helly's theorem, any group of 3 sets has a non-empty intersection $\Longrightarrow$ all of them should have a non-empty intersection
5. However, by construction S has empty intersection with the feasibility region (non-shaded epigraph), as it includes all points with strictly lower cost (infeasible solutions) $\Longrightarrow$ contradiction

Only $d \le n_x = 2$ support constraints may exist!

Stronger version for convex scenario programs

For convex scenario programs we can always have a stronger version!

Let $d = n_x$, i.e. the # of decision variables in a convex scenario program. Then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m:\ \mathbb{P}\Big(\delta\in\Delta:\ g(x_m^*,\delta)>0\Big)\le\epsilon\Big\}\ge 1-q(m,\epsilon)$$

with $q(m,\epsilon)=\sum_{k=0}^{d-1}\binom{m}{k}\epsilon^k(1-\epsilon)^{m-k}$.

- Existence of a *unique* compression set is a sufficient condition for the stronger generalization result (see Lecture 2)
- For non-degenerate convex problems a unique compression set can always be constructed (possibly upon some lexicographic order to select among multiple ones)
- It can be shown that stronger bound holds even for degenerate convex scenario programs (via a constraint "heating and cooling" procedure)

Stronger version – Different interpretation

For convex scenario programs we can always have a stronger version

Let $d = n_x$, i.e. the # of decision variables in a convex scenario program.

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m:\ \mathbb{P}\Big(\delta\in\Delta:\ g(x_m^*,\delta)\le 0\Big)>1-\epsilon\Big\}\ge 1-q(m,\epsilon)$$

with $q(m,\epsilon)=\sum_{k=0}^{d-1}\binom{m}{k}\epsilon^k(1-\epsilon)^{m-k}$.

- **Different interpretation:** Fix confidence $\beta\in(0,1)$ and violation level $\epsilon\in(0,1)$. Determine the number of samples needed to guarantee that, with confidence at least $1-\beta$, the probability of constraint satisfaction for x_m^* is at least $1-\epsilon$.
- Set $\beta\ge q(m,\epsilon)$, and find an m that satisfies

$$\sum_{k=0}^{d-1}\binom{m}{k}\epsilon^k(1-\epsilon)^{m-k}\le\beta$$

Stronger version – Different interpretation

For convex scenario programs we can always have a stronger version

Let $d = n_x$, i.e. the # of decision variables in a convex scenario program. Then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m:\ \mathbb{P}\Big(\delta\in\Delta:\ g(x_m^*,\delta)\le 0\Big)>1-\epsilon\Big\}\ge 1-q(m,\epsilon)$$

with $q(m,\epsilon)=\sum_{k=0}^{d-1}\binom{m}{k}\epsilon^k(1-\epsilon)^{m-k}$.

- **Different interpretation:** Fix confidence $\beta\in(0,1)$ and violation level $\epsilon\in(0,1)$. Determine the number of samples needed to guarantee that, with confidence at least $1-\beta$, the probability of constraint satisfaction for x_m^* is at least $1-\epsilon$.
- A sufficient condition for m is given by

$$m\ge\frac{2}{\epsilon}\Big(d-1+\ln\frac{1}{\beta}\Big)$$

Proof of explicit bound for number of samples m

1. By the Chernoff bound we can bound the "binomial tail" by

$$q(m,\epsilon) \le e^{-\frac{(m\epsilon-d+1)^2}{2m\epsilon}}, \text{ for any } m\epsilon > d$$

2. We determine a sequence of sufficient conditions for $q(m,\epsilon) \le \beta$:

$$e^{-\frac{(m\epsilon-d+1)^2}{2m\epsilon}} \le \beta \Leftarrow \frac{(m\epsilon-d+1)^2}{2m\epsilon} \ge \ln\frac{1}{\beta} \quad \text{[taking logarithm]}$$

$$\Leftarrow \frac{1}{2}m\epsilon + \frac{(d-1)^2}{2m\epsilon} + 1 - d \ge \ln\frac{1}{\beta} \quad \text{[expanding the square]}$$

$$\Leftarrow \frac{1}{2}m\epsilon + 1 - d \ge \ln\frac{1}{\beta} \quad \text{[dropping the red term since} \ge 0]$$

3. Solving with respect to m

$$m \ge \frac{2}{\epsilon}\left(d - 1 + \ln\frac{1}{\beta}\right)$$

Distribution of the probability of constraint violation

- Strong confidence bound $\Leftrightarrow$ the distribution is bounded by a binomial
- How tight is the strong confidence bound?
- Bound on the expected value of the probability of violation
- Robust control synthesis by means of an example

Distribution of the probability of constraint violation

- For a random variable X, its distribution is characterized by $\mathrm{Prob}\{X \le x\}$, where x is the valuation of the random variable
- For our probabilistic feasibility result
  - Random variable: Probability of constraint violation

$$X = \mathbb{P}\Big(\delta \in \Delta : g(x_m^*, \delta) > 0\Big), \text{ and value: } x = \epsilon$$

  - Probability distribution of $X \le x$, i.e. "probability of the probability"

$$\mathbb{P}\Big(\delta \in \Delta : g(x_m^*, \delta) > 0\Big) \le \epsilon$$

- Can we characterize the probability distribution of the probability of constraint violation? This is our generalization theorem!

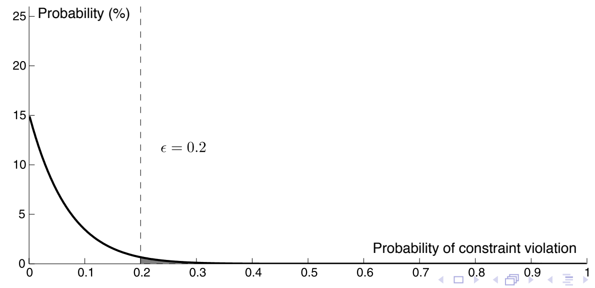
Distribution of the probability of constraint violation

The distribution of $\mathbb{P}\Big(\delta \in \Delta : g(x_m^*, \delta) > 0\Big)$ is bounded by a binomial!

- By our generalization statement, it is bounded by

$$1 - \sum_{k=0}^{d-1} \binom{m}{k} \epsilon^k (1-\epsilon)^{m-k}, \quad \text{[non-shaded area in figure below]}$$

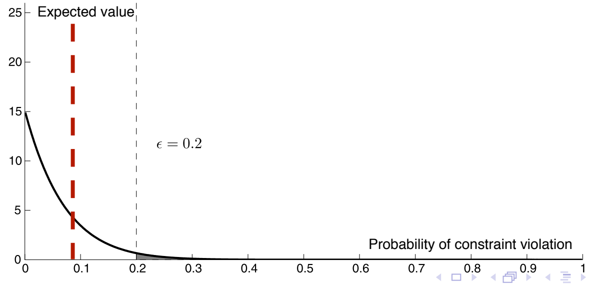
the (tail of the) cumulative distribution of a binomial random variable
- Figure below shows the binomial distribution for $d = 1$ and $m = 15$



Distribution of the probability of constraint violation

The distribution of $\mathbb{P}\Big(\delta \in \Delta: \; g(x_m^*, \delta) > 0\Big)$ is bounded by a binomial!

1. When is it **equal to** the tail of the cumulative distribution of a binomial random variable?
$$1 - \sum_{k=0}^{d-1} \binom{m}{k} \epsilon^k (1-\epsilon)^{m-k}, \quad \text{[non-shaded area in figure below]}$$
2. What can we say about its expected value?



Distribution of the probability of constraint violation

- We will show that our strong theorem can hold with equality, i.e. the confidence $1 - \sum_{k=0}^{d-1} \binom{m}{k} \epsilon^k (1-\epsilon)^{m-k}$ is tight
- We will do so by means of an example

Example with tight confidence bound

Assume that samples are extracted from a uniform distribution in $[0,1]$, and consider the scenario program
$$\min_{x \in \mathbb{R}} \; x$$
$$\text{subject to } \; \delta_i \le x, \; \text{for all } i = 1, \dots, m$$

- Convex scenario program with $n_x = 1$
- Objective function: $c^\top x = x$
- Constraint function: $g(x, \delta) = \delta - x$

Distribution of the probability of constraint violation

1. Denote by x_m^* its minimizer, and notice that this is equal to the maximum sample, i.e.
$$x_m^* = \max_{i=1,\dots,m} \delta_i$$
2. What is the probability of constraint violation?
$$\mathbb{P}\Big(\delta \in \Delta: \; g(x_m^*, \delta) > 0\Big) = \mathbb{P}\Big(\delta \in \Delta: \; \delta > x_m^*\Big)$$
$$= 1 - x_m^* \quad \text{[since } \mathbb{P} \text{ uniform in } [0,1]]$$
3. We will show that (our complementary generalization statement)
$$\mathbb{P}^m\Big\{\delta_1, \dots, \delta_m: \; \mathbb{P}\Big(\delta \in \Delta: \; \delta > x_m^*\Big) > \epsilon\Big\} = (1-\epsilon)^m,$$
i.e. the the strong bound for $d = n_x$.
Note that this holds with equality, hence it is tight! Problems where the strong bound holds with equality are called fully-supported

Distribution of the probability of constraint violation

- To see this, notice that
$$\mathbb{P}^m\Big\{\delta_1, \dots, \delta_m: \; \mathbb{P}\Big(\delta \in \Delta: \; \delta > x_m^*\Big) > \epsilon\Big\}$$
$$= \mathbb{P}^m\Big\{\delta_1, \dots, \delta_m: \; 1 - \max_i \delta_i > \epsilon\Big\}$$
$$= \mathbb{P}^m\Big\{\delta_1, \dots, \delta_m: \; \max_i \delta_i < 1 - \epsilon\Big\}$$
$$= \mathbb{P}^m\Big\{\delta_1, \dots, \delta_m: \; \delta_i < 1 - \epsilon, \; \text{for all } i = 1, \dots, m\Big\}$$
- Second step: we used the fact that $\mathbb{P}\Big(\delta \in \Delta: \; \delta > x_m^*\Big) = 1 - x_m^*$
- Third step: if the maximum is below $1 - \epsilon$, then each sample is as well

Distribution of the probability of constraint violation

- Samples are independent, so probability of "intersection" is the product of individual probabilities

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m:\ \mathbb{P}\Big(\delta\in\Delta:\ \delta>x_m^*\Big)>\epsilon\Big\}$$
$$=\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m:\ \delta_i<1-\epsilon,\ \text{for all } i=1,\dots,m\Big\}$$
$$=\Pi_{i=1}^m\ \mathbb{P}\Big\{\delta_i<1-\epsilon\Big\}$$

- Since the probability is uniform, each individual probability is given by

$$\mathbb{P}\Big\{\delta_i<1-\epsilon\Big\}=1-\epsilon$$

- Putting everything together

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m:\ \mathbb{P}\Big(\delta\in\Delta:\ \delta>x_m^*\Big)>\epsilon\Big\}=(1-\epsilon)^m$$

Expected probability of constraint violation

Expected probability of constraint violation – Convex scenario programs

Let $d=n_x$, i.e. the # of decision variables in a convex scenario program. Then

$$\mathbb{E}_{\sim\mathbb{P}^m}\Big[\mathbb{P}\Big(\delta\in\Delta:\ g(x_m^*,\delta)>0\Big)\Big]\le\frac{d}{m+1}$$

- $\mathbb{E}_{\sim\mathbb{P}^m}$ denotes the expected value operator associated with the probability $\mathbb{P}^m$ of extracting $(\delta_1,\dots,\delta_m)$
- We no longer have two layers of probability, but rather a bound on the expectation $\mathbb{E}_{\sim\mathbb{P}^m}$
- From the "probability of the probability" to "expectation of the probability"

Expected probability of constraint violation

Expected probability of constraint violation – Convex scenario programs

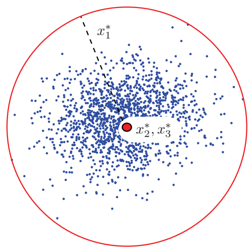
Let $d=n_x$, i.e. the # of decision variables in a convex scenario program. Then

$$\mathbb{E}_{\sim\mathbb{P}^m}\Big[\mathbb{P}\Big(\delta\in\Delta:\ g(x_m^*,\delta)>0\Big)\Big]\le\frac{d}{m+1}$$

- **Explicit bound on the number of samples:** Fix a violation level $\rho\in(0,1)$. Determine the number of samples needed to guarantee that the expected value of the probability of constraint violation for x_m^* is at most ρ.
- A sufficient condition for $\mathbb{E}_{\sim\mathbb{P}^m}\Big[\mathbb{P}\Big(\delta\in\Delta:\ g(x_m^*,\delta)>0\Big)\Big]\le\rho$

$$\frac{d}{m+1}\le\rho\ \Leftrightarrow\ m\ge\frac{d}{\rho}-1$$

Example: Minimum radius disk problem revisited



- Construct the minimum radius disk program (d=3 decision variables)

$$\min_{x_1,x_2,x_3}\ x_1$$
$$\text{subject to: } \sqrt{(y_i-x_3)^2+(u_i-x_2)^2}\le x_1,\ \text{for all } i=1,\dots,1650$$

- How high is the expected value of the probability that the minimum radius disk will **not** contain a new point $\delta=(u,y)$?

$$\mathbb{E}_{\sim\mathbb{P}^m}\Big[\mathbb{P}\Big(\delta=(u,y):\ \sqrt{(y-x_3^*)^2+(u-x_2^*)^2}>x_1^*\Big)\Big]\le\frac{d}{m+1}=\frac{3}{1651}$$

Robust state feedback control design

Problem specifications

Consider the family of systems

$$\dot{x} = A(\delta_i)x + B(\delta_i)u, \ \ i = 1, \ldots, m,$$

where δ_i's are independent samples extracted from $\mathbb{P}$.

1. Design a gain matrix K such that $u = Kx$ renders the closed loop system asymptotically stable.
2. Provide guarantees that the constructed K will stabilize a new system $\dot{x} = A(\delta)x + B(\delta)u$ (for some new δ).

- Uncertainty enters the problem data, i.e. the elements of A and B depend on δ_i
- We need that the same K stabilizes all systems, *not* a different feedback matrix per system

Robust state feedback control design (cont'd)

- Consider the closed loop system, once $u = Kx$ has been applied
- We have a family of closed loop systems:

$$\dot{x} = \big(A(\delta_i) + B(\delta_i)K\big)x, \ \ \text{for all } i = 1, \ldots, m$$

- Restatement of the problem:
  Find K such that $A(\delta_i) + B(\delta_i)K$ is Hurwitz for all $i = 1, \ldots, m$.

Recall Lyapunov's stability condition

A matrix A is Hurwitz **if and only if** there exists $P = P^\top > 0$ such that

$$PA^\top + AP < 0 \qquad \text{[Linear Matrix Inequality (LMI)]}$$

Note that this is a equivalent to the more standard $A^\top P + PA < 0$

$\Longrightarrow$ Apply Lyapunov's LMI to the family of closed-loop systems

Robust state feedback control design (cont'd)

Three step procedure:

1. Lyapunov's stability LMI for the closed loop family of systems, i.e. with $A(\delta_i) + B(\delta_i)K$ in place of A

$$P\big(A(\delta_i) + B(\delta_i)K\big)^\top + \big(A(\delta_i) + B(\delta_i)K\big)P < 0, \ \ \forall i = 1, \ldots, m$$

   which leads to

$$PA(\delta_i)^\top + (PK^\top)B(\delta_i)^\top + A(\delta_i)P + B(\delta_i)(KP) < 0, \ \ \forall i = 1, \ldots, m$$

2. Set $Z = KP$ (recall that P is symmetric) and find P and Z such that

$$PA(\delta_i)^\top + Z^\top B(\delta_i)^\top + A(\delta_i)P + B(\delta_i)Z < 0, \ \ \forall i = 1, \ldots, m$$

3. Compute the gain matrix by $K = ZP^{-1}$, for all $i = 1, \ldots, m$

Robust state feedback control design (cont'd)

- How to find P and Z such that

$$PA(\delta_i)^\top + Z^\top B(\delta_i)^\top + A(\delta_i)P + B(\delta_i)Z < 0, \ \ \forall i = 1, \ldots, m$$

- By means of an optimization (in fact feasibility problem)

$$\begin{aligned} \min_{P,Z} \ \ & 0 \qquad \text{[any constant would work]} \\ \text{subject to} \ \ & PA(\delta_i)^\top + Z^\top B(\delta_i)^\top + A(\delta_i)P + B(\delta_i)Z < 0, \\ & \qquad \text{for all } i = 1, \ldots, m \end{aligned}$$

- Convex scenario program as LMIs are convex constraints!
  Let P^* and Z^* denote its minimizers, and construct $K^* = Z^*(P^*)^{-1}$

Robust state feedback control design (cont'd)

- Consider a new δ that gives rise to the system

$$\dot{x} = A(\delta)x + B(\delta)u$$

Determine the confidence with which the probability that K^* renders the new system unstable is at most equal to a given level ϵ

Probabilistic guarantees

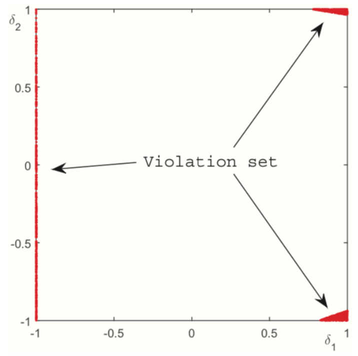
1. Consider a given number of samples m and a violation level $\epsilon \in (0,1)$.
2. Count the number of decision variables in $P \in \mathbb{R}^{n_x \times n_x}$ and $Z \in \mathbb{R}^{n_x \times n_x}$, i.e. $d = 2n_x^2$
3. With confidence at least $1 - \sum_{k=0}^{d-1} \binom{m}{k} \epsilon^k (1-\epsilon)^{m-k}$,

$$\mathbb{P}\Big(\delta : \; P^* A(\delta)^\top + (Z^*)^\top B(\delta)^\top + A(\delta)P^* + B(\delta)Z^* > 0\Big) \leq \epsilon$$

or equivalently, the probability that $K^* = Z^*(P^*)^{-1}$ renders a new system/plant (induced by the new sample δ) unstable is at most ϵ.

Robust state feedback control design (cont'd)

- Red regions illustrate the set of new δ's for which x_m^* violates the constraints
- Example[1] refers to a 2-dimensional uncertainty vector δ



[1] Figure taken from "Introduction to the scenario approach", by M. Campi & S. Garatti, SIAM 2018

Robust state feedback control design (cont'd)

Guarantees on the expected probability of constraint violation

Let $n_x = 2$. Determine the number of samples m such that the expected value of the probability that $K^* = Z^*(P^*)^{-1}$ renders a new system/plant unstable is at most 0.05.

- We want

$$\mathbb{E}_{\sim\mathbb{P}^m}\Big[\mathbb{P}\Big(\delta : \; P^* A(\delta)^\top + (Z^*)^\top B(\delta)^\top + A(\delta)P^* + B(\delta)Z^* > 0\Big)\Big] \leq 0.05$$

- Set $\rho = 0.05$. A sufficient condition for this to hold is given by

$$m \geq \frac{d}{\rho} - 1,$$

where $d = 2n_x^2$ denotes the number of decision variables in $P \in \mathbb{R}^{n_x \times n_x}$ and $Z \in \mathbb{R}^{n_x \times n_x}$

- We thus have that $m \geq \frac{8}{0.05} - 1 = 159$ samples need to be extracted

Appendix

Appendix: Proof of the main PAC learning theorem

Theorem

If a **compression set** C_d **with cardinality** d exists, then

$$\mathbb{P}^m\Big\{\delta_1,\ldots,\delta_m : \mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon\Big\} \ge 1 - q(m,\epsilon)$$

with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$.

Proof

- We assume existence of C_d for any m multi-sample; it will also exist with confidence $1 - q(m,\epsilon)$, i.e.

Fix $\epsilon \in (0,1)$. We will equivalently show that

$$\mathbb{P}^m\Big\{\delta_1,\ldots,\delta_m : \exists\ C_d \text{ such that } \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \text{ for all } i = 1,\ldots,m \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon \Big\} \le q(m,\epsilon)$$

where $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$.

- "Yellow" events: empirical generalization and probabilistic generalization, respectively
- First event: Zero disagreement between H_d and T on the samples; Second event: ϵ disagreement in probability

Proof (cont'd)

Equivalently, we have that

$$\mathbb{P}^m\Big\{\bigcup_{C_d}\Big\{\delta_1,\ldots,\delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i),\ \forall i \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon \Big\}\Big\}$$

$$\le \sum_{C_d}\mathbb{P}^m\Big\{\delta_1,\ldots,\delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i),\ \forall i \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon \Big\}$$

- Existence of a compression set C_d is equivalent to taking the "union"
- Union is taken with respect to all potential compression sets C_d sets, each one containing d samples
- Subadditivity property: Probability of the "union" of events smaller than or equal to the "sum" of the individual probability of each event

Proof (cont'd)

- Without loss of generality let $C_d = \{\delta_1,\ldots,\delta_m\}$ and

$$\bar{\Delta} = \Big\{\delta_1,\ldots,\delta_d : \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon\Big\}$$
$$= \Big\{\delta_1,\ldots,\delta_d : \mathbb{P}\Big(\delta : \mathbb{1}_{H_d}(\delta) \neq \mathbb{1}_T(\delta)\Big) > \epsilon\Big\}$$

- Since H_d is constructed based on $\delta_1,\ldots,\delta_d$, notice that

$$\mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \quad \text{for all } i = 1,\ldots,d$$

Pick a "new" δ

$$\mathbb{P}\Big\{\delta : \mathbb{1}_{H_d}(\delta) = \mathbb{1}_T(\delta) \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon \ \Big|\ \delta_1,\ldots,\delta_d \in \bar{\Delta}\Big\}$$
$$= \mathbb{P}\Big\{\delta : \mathbb{1}_{H_d}(\delta) = \mathbb{1}_T(\delta) \ \Big|\ \delta_1,\ldots,\delta_d \in \bar{\Delta}\Big\} \le 1 - \epsilon$$

- The equality follows from the fact that second "yellow" event is independent of δ; the inequality follows from the definition of $\bar{\Delta}$

Proof (cont'd)

- Pick a "new" δ

$$\mathbb{P}\Big\{\delta : \mathbb{1}_{H_d}(\delta) = \mathbb{1}_T(\delta) \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon \;\Big|\; \delta_1, \dots, \delta_d \in \bar{\Delta}\Big\} \le 1 - \epsilon$$

Bernoulli trials: $m - d$ independent extractions $\delta_{d+1}, \dots, \delta_m$; condition on $\delta_1, \dots, \delta_d \in \bar{\Delta}$

$$\mathbb{P}^{m-d}\Big\{\delta_{d+1}, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i) \text{ for all } i = d+1, \dots, m \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon \;\Big|\; \delta_1, \dots, \delta_d \in \bar{\Delta}\Big\}$$

$$= \prod_{i=d+1}^{m} \mathbb{P}\Big\{\delta_i : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i) \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon \;\Big|\; \delta_1, \dots, \delta_d \in \bar{\Delta}\Big\} \le (1-\epsilon)^{m-d}$$

Proof (cont'd)

Deconditioning ...

$$\mathbb{P}^m\Big\{\delta_1, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \ \forall i \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon\Big\}$$

$$= \int_{\bar{\Delta}} \mathbb{P}^{m-d}\Big\{\delta_{d+1}, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i) \text{ for all } i = d+1, \dots, m \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon \;\Big|\; \delta_1, \dots, \delta_d \in \bar{\Delta}\Big\} \, d\mathbb{P}(d\delta_1, \dots, d\delta_d)$$

$$\le (1-\epsilon)^{m-d}$$

- The equality is due to the definition of the conditional probability (for continuous distributions)
- The inequality follows from the obtained Bernoulli trials bound

Proof (cont'd)

Deconditioning ...

$$\mathbb{P}^m\Big\{\delta_1, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \ \forall i \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon\Big\}$$

$$\le (1-\epsilon)^{m-d}$$

Desired statement was shown to be upper-bounded by

$$\sum_{C_d} \mathbb{P}^m\Big\{\delta_1, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \ \forall i \text{ and } \mathbb{P}\Big(\delta \in T \setminus H_d\Big) > \epsilon\Big\}$$

$$\le \sum_{C_d} (1-\epsilon)^{m-d} \qquad \left[\binom{m}{d} \text{ terms in the summation}\right]$$

$$= \binom{m}{d}(1-\epsilon)^{m-d}$$

Summary

Generalization theorem for abstract problems

If a **compression set C_d with cardinality d** exists, then

$$\mathbb{P}^m\Big\{\delta_1, \dots, \delta_m : \mathbb{P}\Big(\delta \in T \setminus H_m\Big) \le \epsilon\Big\} \ge 1 - q(m, \epsilon)$$

with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$, where $\lim_{m \to \infty} q(m, \epsilon) = 0$.

- Hypothesis probably approximately correct (PAC) learns target
- We do not care about C_d but only about d
- It is a distribution-free result; holds true for any underlying (possibly unknown) distribution, as long as data are independently extracted
- **Stronger version:** If the compression set is unique, then $q(m,\epsilon) = \sum_{k=0}^{d-1} \binom{m}{k} \epsilon^k (1-\epsilon)^{m-k}$

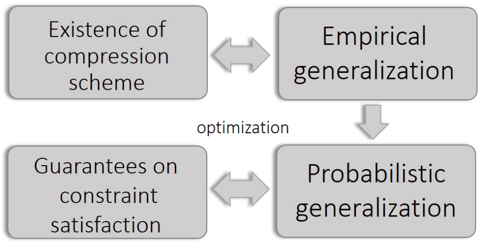
Summary

Probabilistic feasibility – Convex scenario programs

Let $d = n_x$, i.e. the # of decision variables in a convex scenario program. Then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \mathbb{P}\Big(\delta \in \Delta : \ g(x_m^*, \delta) > 0\Big) \le \epsilon\Big\} \ge 1 - q(m,\epsilon)$$

with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$.



Summary

Probabilistic feasibility – Convex scenario programs (stronger version)

Let $d = n_x$, i.e. the # of decision variables in a convex scenario program. Then

$$\mathbb{P}^m\Big\{\delta_1,\dots,\delta_m : \mathbb{P}\Big(\delta \in \Delta : \ g(x_m^*, \delta) > 0\Big) \le \epsilon\Big\} \ge 1 - q(m,\epsilon)$$

with $q(m,\epsilon) = \sum_{k=0}^{d-1}\binom{m}{k}\epsilon^k(1-\epsilon)^{m-k}$.

- **Explicit bound on the number of samples:** Fix confidence $\beta \in (0,1)$ and violation level $\epsilon \in (0,1)$. Determine the number of samples needed to guarantee that, with confidence at least $1-\beta$, the probability of constraint satisfaction for x_m^* is at least $1-\epsilon$.

$$m \ge \frac{2}{\epsilon}\Big(d - 1 + \ln\frac{1}{\beta}\Big)$$

Summary

Expected probability of constraint violation – Convex scenario programs

Let $d = n_x$, i.e. the # of decision variables in a convex scenario program. Then

$$\mathbb{E}_{\sim\mathbb{P}^m}\Big[\mathbb{P}\Big(\delta \in \Delta : \ g(x_m^*, \delta) > 0\Big)\Big] \le \frac{d}{m+1}$$

- **Explicit bound on the number of samples:** Fix a violation level $\rho \in (0,1)$. Determine the number of samples needed to guarantee that the expected value of the probability of constraint violation for x_m^* is at most ρ.

$$m \ge \frac{d}{\rho} - 1$$

Thank you for your attention!
Questions?

Contact at:
kostas.margellos@eng.ox.ac.uk

Slides can be found on WebLearn and here:
https://sites.google.com/site/margellosk/home/teaching