

Logistics

C20 Robust Optimization

Lecture 1

Kostas Margellos

University of Oxford



Michaelmas Term 2024

C20 Robust Optimization

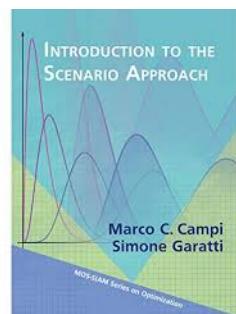
November 9, 2024

1 / 27

References

- [Campi & Garatti \(2019\)](#)
Introduction to the Scenario Approach
SIAM (*some figures are taken from that book*).

- [Margellos, Prandini & Lygeros \(2015\)](#)
On the Connection Between Compression Learning and Scenario Based Single-Stage and Cascading Optimization Problems,
IEEE Transactions on Automatic Control, 60(10), 2716-2721.



- **Who:** Kostas Margellos, Control Group, IEB 50.16
contact: kostas.margellos@eng.ox.ac.uk
- **When:** 4 lectures,
weeks 7 & 8 – Thu, Fri @4pm
- **Where:** LR2
- **Other info:**
 - ▶ 2 example classes: early Trinity Term; date to be announced
 - ▶ Lecture slides available on Canvas
 - ▶ Teaching style: Mix of slides and whiteboard writing

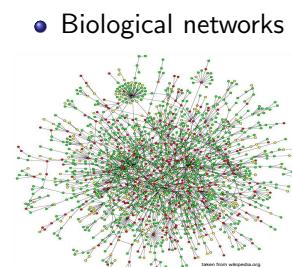
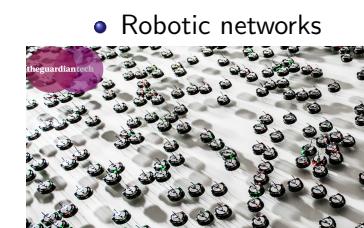
Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

2 / 27

Motivation



Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

3 / 27

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

4 / 27

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

– Christiaan Huygens, 1629 – 1695



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



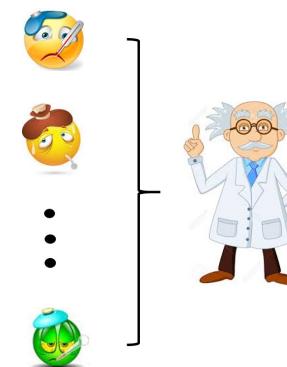
'After careful consideration of all 437 charts, graphs, and metrics,
I've decided to throw up my hands, hit the liquor store,
and get snookered. Who's with me?!"

Objectives of the second part of this class

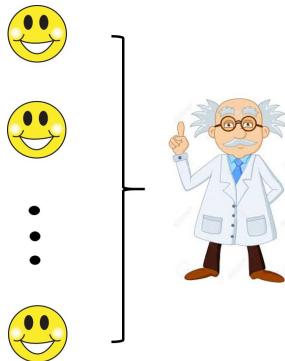
- Big picture
 - Decision making in the presence of uncertainty
 - Related to: Randomized/stochastic and robust optimization
 - Convex optimization ... and a bit of Statistical Learning Theory
- What it is actually about
 - ① Introduce data based optimization
 - ② Make decisions under uncertainty and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



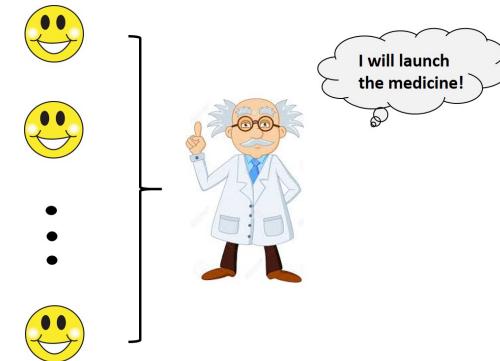
Motivation - The doctor's problem



Motivation - The doctor's problem



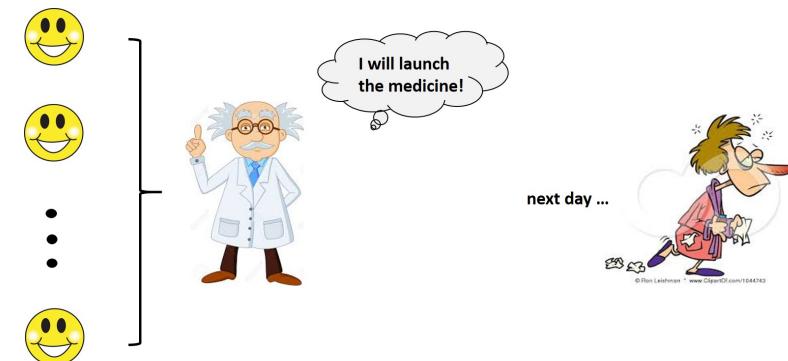
Motivation - The doctor's problem



Motivation - The doctor's problem



Motivation - The doctor's problem

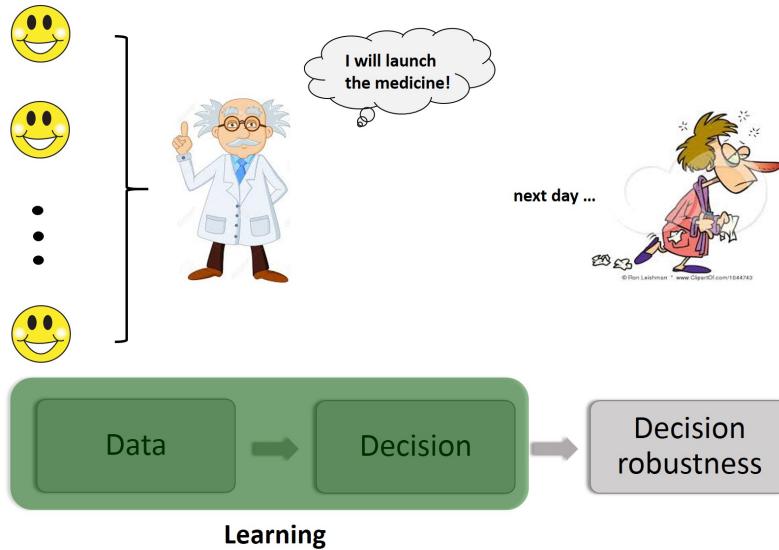


Data

Decision

Decision
robustness

Motivation - The doctor's problem



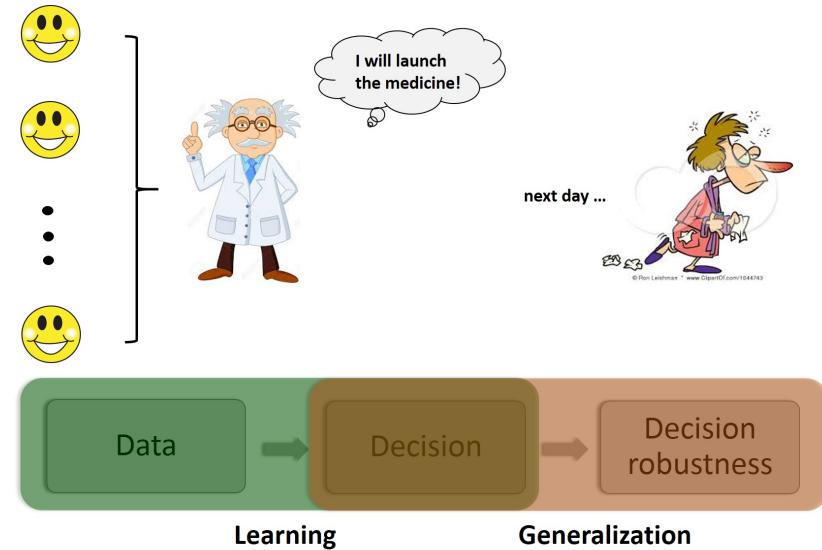
Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

13 / 27

Motivation - The doctor's problem



Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

14 / 27

Probably Approximately Correct Learning

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

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

9 / 27

Terminology by means of an example

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

$$\widehat{\mathbb{P}}_{(\delta_1, \dots, \delta_{100})} = \frac{\# \text{ Heads}}{\# \text{ coin tosses}}$$
- ➋ Repeat it the experiment 50 times
 - You will get 50 different $\widehat{\mathbb{P}}_{(\delta_1, \dots, \delta_{100})}$: 0.55, 0.47, 0.53, ...
 - $\widehat{\mathbb{P}}_{(\delta_1, \dots, \delta_{100})}$ is itself random!
 - How likely it is that $|\widehat{\mathbb{P}}_{(\delta_1, \dots, \delta_{100})} - 0.5|$ 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?

Michaelmas Term 2024

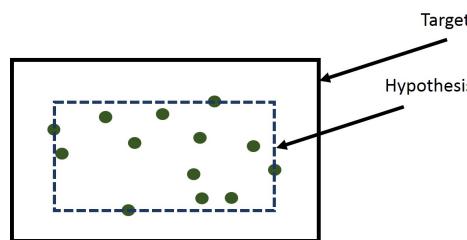
C20 Robust Optimization

November 9, 2024

10 / 27

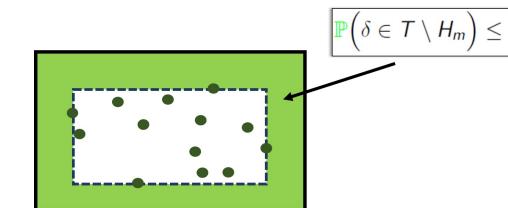
Learning

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



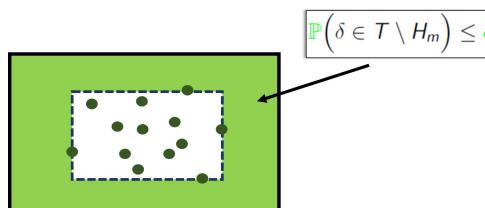
Generalization – Probably Approximately Correct Learning

- **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” $\mathbb{P}(\delta \in T \setminus H_m)$
 - ☺ if $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$ (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 $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$?
 - In other words, for “how many” of the multi-samples is this the case?



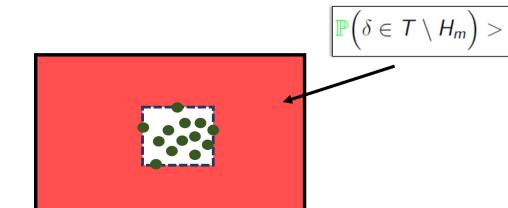
Generalization – Probably Approximately Correct Learning

- **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” $\mathbb{P}(\delta \in T \setminus H_m)$
 - ☺ if $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$ (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 $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$?
 - 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” $\mathbb{P}(\delta \in T \setminus H_m)$
 - ☺ if $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$ (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 $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$?
 - 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

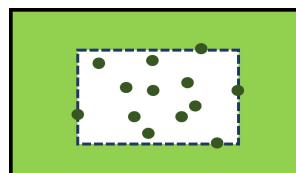
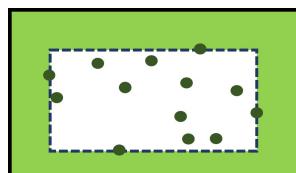
Problem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

- 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 $\mathbb{P}(\delta \in (T \setminus H_m) \cup (H_m \setminus T))$



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

Problem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

- 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. $\lim_{m \rightarrow \infty} q(m, \epsilon) = 0$

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

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

$$H_d = H_m$$

C_d is then called a compression set.

- Hypothesis H_d based on samples in C_d is the same with the hypothesis H_m , that would have been obtained with all samples

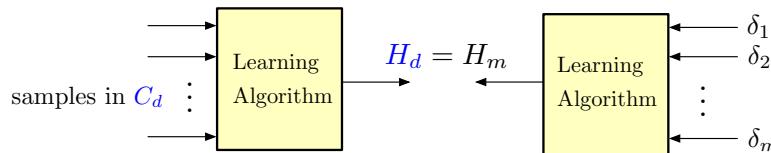
Generalization - sufficient condition

Compression set

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

$$H_d = H_m$$

C_d is then called a compression set.



Generalization - sufficient condition

Compression set (more general definition)

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

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

C_d is then called a compression set.

- 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 (more general definition)

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

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

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

$$\mathbb{P}^m\left\{\delta_1, \dots, \delta_m : \mathbb{P}\left(\delta \in T \setminus H_m\right) \leq \epsilon\right\} \geq 1 - q(m, \epsilon)$$

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

Generalization

Theorem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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 $\lim_{m \rightarrow \infty} q(m, \epsilon) = 0$

$$\begin{aligned} \lim_{m \rightarrow \infty} q(m, \epsilon) &= \lim_{m \rightarrow \infty} \binom{m}{d} (1 - \epsilon)^{m-d} \\ &\leq \lim_{m \rightarrow \infty} \left(\frac{me}{d} \right)^d (1 - \epsilon)^{m-d} = 0 \end{aligned}$$

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

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

21 / 27

Generalization

Theorem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

- Does the cardinality d of the compression set matter?
Assume we were allowing the trivial case $d = m$

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

- As the compression “increases” the confidence $1 - q(m, \epsilon)$ tends to 0
⇒ result trivial (not useful) as we claim that H_m is an ϵ -good approximation of T with non-negative probability!
- The smaller the compression the more useful the result!

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

22 / 27

Generalization – Stronger statement

Theorem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

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

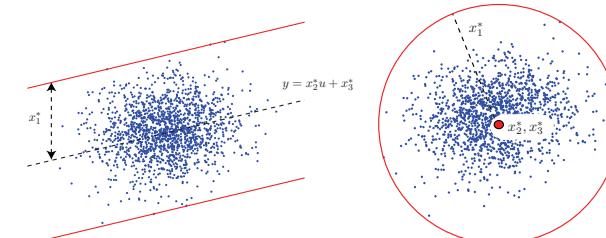
$$\sum_{k=0}^{d-1} \binom{m}{k} \epsilon^k (1 - \epsilon)^{m-k} < \binom{m}{d} (1 - \epsilon)^{m-d},$$

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

Michaelmas Term 2024 C20 Robust Optimization November 9, 2024 23 / 27

Generalization – Stronger statement

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



- In both problems 3 samples are sufficient
⇒ compression set with $d = 3$ exists
- For the disk problem, for almost all multi-samples we only need 2 samples: Take the two most isolated samples as the disk's diameter
⇒ only 2 matter, the third could be arbitrary (it falls inside the disk)
- In the disk problem, multiple compression sets of cardinality $d = 3$!

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

24 / 27

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

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

Theorem

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

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

Summary

Theorem

If a compression set C_d with cardinality d exists, then

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

Existence of compression scheme

Empirical generalization

Probabilistic generalization

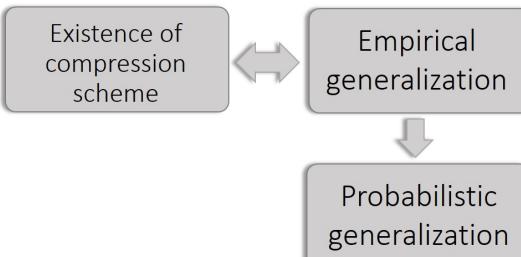
Summary

Theorem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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



Thank you for your attention!
Questions?

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

C20 Robust Optimization

Lecture 2

Kostas Margellos

University of Oxford



Michaelmas Term 2024

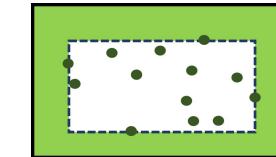
C20 Robust Optimization

November 9, 2024

1 / 23

Recap – Learning & Generalization

- **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. $\mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon$



- **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

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

2 / 23

Recap – Generalization

Theorem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

- 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

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

3 / 23

Generalization

Theorem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

- Does the cardinality d of the compression set matter?
Assume we were allowing the trivial case $d = m$
- $\lim_{d \rightarrow m} 1 - q(m, \epsilon) = 1 - \lim_{d \rightarrow m} \binom{m}{d} (1 - \epsilon)^{m-d} = 0$
- As the compression “increases” the confidence $1 - q(m, \epsilon)$ tends to 0
 \Rightarrow result trivial (not useful) as we claim that H_m is an ϵ -good approximation of T with non-negative probability!
- The smaller the compression the more useful the result!

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

4 / 23

Generalization – Stronger statement

Theorem

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

- Stronger assumption \implies 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

$$1 - q(m, \epsilon)$$

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

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

- 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
 \implies In general not solvable!

Data based optimization

- Uncertain scenario program

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

- 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

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

- 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	$H_m = \{\delta \in \Delta : g(x_m^*, \delta) \leq 0\}$

- 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

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

- 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	$H_m = \{\delta \in \Delta : g(x_m^*, \delta) \leq 0\}$

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

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

Data based optimization – Generalization

Theorem (the abstract version)

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

Scenario vs. Uncertain programs

Probabilistic feasibility

Data based program

$$\begin{aligned} \min_{x \in \mathbb{R}^{n_x}} \quad & c^T x \\ \text{subject to} \quad & \rightarrow x_m^* \\ & g(x, \delta_i) \leq 0, \forall i = 1, \dots, m \end{aligned}$$

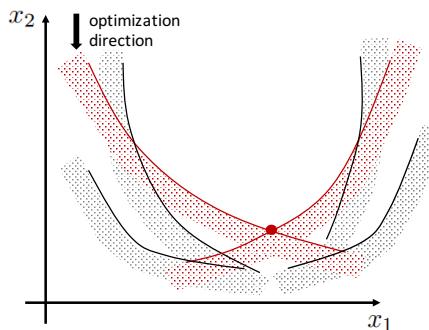
Robust program

$$\begin{aligned} \min_{x \in \mathbb{R}^{n_x}} \quad & c^T x \\ \text{subject to} \quad & g(x, \delta) \leq 0, \forall \delta \in \Delta \end{aligned}$$

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

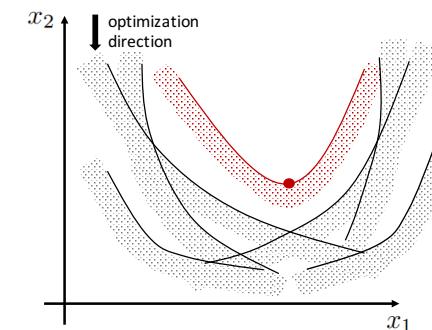


Compression set: 2D example



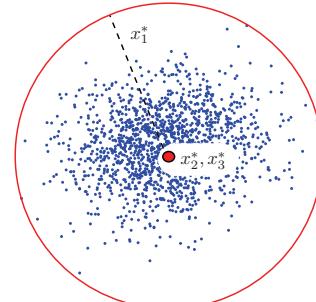
- Compression set cardinality $d = n_x$
 - Compression set = Two active constraints
⇒ If any of the two red constraints is removed the solution drifts to a lower value (intersection of the remaining red with a lower constraint)
 - Compression set coincides with “red” constraints $\implies x_{\text{red}}^*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

Compression set: 2D example



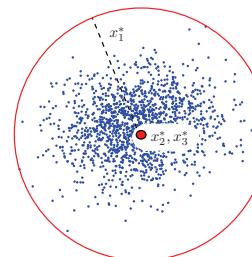
- Compression set cardinality $d \leq n_x$ (always)
 - Compression set = One active constraint
⇒ If any of the other constraints are removed the solution remains unaltered; only the red constraint is needed
 - We again have that $x_{red}^* = x_m^*$

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)



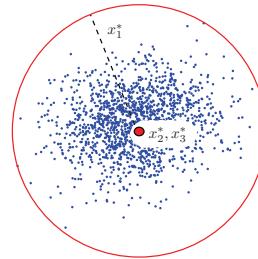
- Construct the minimum radius disk program ($d=3$ decision variables)

$$\min_{x_1, x_2, x_3} x_1$$

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

- 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: $\delta_i = (u_i, y_i), i = 1, \dots, 1650$

Example (cont'd)



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

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

- Disk should contain 99% of new points $\delta = (u, y) \Rightarrow \epsilon = 0.01$
 - Hence the “guarantee” is the confidence

$$1 - q(1650, 0.01) = 1 - \left(\frac{1650}{3}\right)(1 - 0.01)^{1650-3}$$

Summary

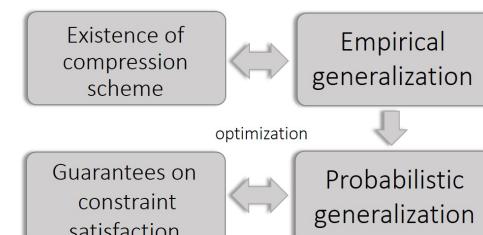
Theorem – Convex scenario programs

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P} \left(\delta \in \Delta : g(x_m^*, \delta) > 0 \right) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

Could we also have a stronger version? See Lecture 3



Thank you for your attention!
Questions?

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

C20 Robust Optimization

Lecture 3

Kostas Margellos



Recap: Probabilistic feasibility

Theorem – Convex scenario programs

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P} \left(\delta \in \Delta : g(x_m^*, \delta) > 0 \right) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

- Existence of a compression set \Leftrightarrow Empirical generalization

Subset of the samples that leads to $x_d^* = x_m^*$

- Empirical generalization \Rightarrow Probabilistic generalization

\Leftrightarrow Feasibility guarantees

i.e. ϵ -probability of constraint violation

- For convex scenario programs: $d \leq \#$ of decision variables

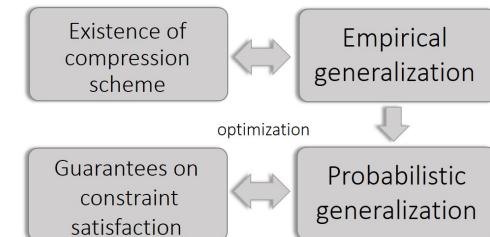
Recap: Probabilistic feasibility

Theorem – Convex scenario programs

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P} \left(\delta \in \Delta : g(x_m^*, \delta) > 0 \right) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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



Convex scenario programs

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

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

- 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

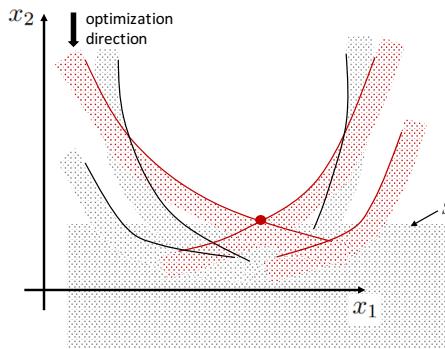
A constraint $k \in \{1, \dots, m\}$ is of support if

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

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

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
 \implies 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)
 \implies contradiction

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



Michaelmas Term 202

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.

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P} \left(\delta \in \Delta : g(x_m^*, \delta) \leq 0 \right) > 1 - \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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 $\beta \geq q(m, \epsilon)$, and find an m that satisfies

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

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P} \left(\delta \in \Delta : g(x_m^*, \delta) \leq 0 \right) > 1 - \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

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

Proof of explicit bound for number of samples m

- ① By the Chernoff bound we can bound the “binomial tail” by

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

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

$$\begin{aligned} e^{-\frac{(m\epsilon - d+1)^2}{2m\epsilon}} \leq \beta &\iff \frac{(m\epsilon - d+1)^2}{2m\epsilon} \geq \ln \frac{1}{\beta} \quad [\text{taking logarithm}] \\ &\iff \frac{1}{2}m\epsilon + \frac{(d-1)^2}{2m\epsilon} + 1 - d \geq \ln \frac{1}{\beta} \quad [\text{expanding the square}] \\ &\iff \frac{1}{2}m\epsilon + 1 - d \geq \ln \frac{1}{\beta} \quad [\text{dropping the red term since } \geq 0] \end{aligned}$$

- ③ Solving with respect to m

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

Distribution of the probability of constraint violation

- For a random variable X , its distribution is characterized by $\text{Prob}\{X \leq x\}$, where x is the valuation of the random variable

- For our probabilistic feasibility result

- Random variable: Probability of constraint violation

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

- Probability distribution of $X \leq x$, i.e. “probability of the probability”

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

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

Distribution of the probability of constraint violation

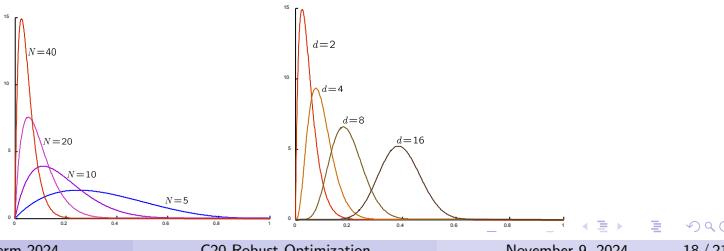
The distribution of $\mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0)$ is bounded by a binomial!

- By our generalization statement, it is bounded by

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

the tail (complementary cumulative) distribution of a binomial random variable

- Density examples for different m, d (with thanks to S. Garatti)



Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

18 / 21

Distribution of the probability of constraint violation

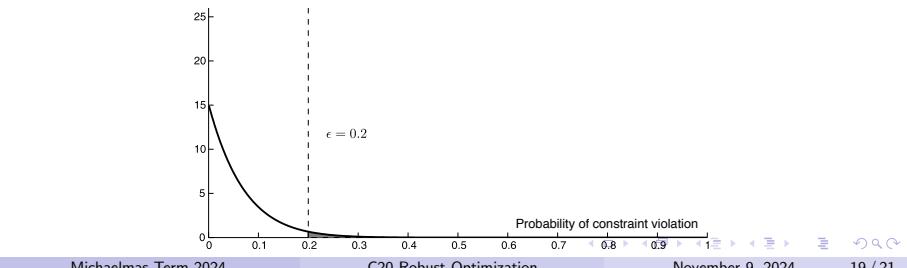
The distribution of $\mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0)$ is bounded by a binomial!

- By our generalization statement, it is bounded by

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

the tail (complementary cumulative) distribution of a binomial random variable

- Density for $d = 1$ and $m = 15$



Summary

Main result for convex scenario programs

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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 $1 - \epsilon$.

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

Thank you for your attention!
Questions?

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

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

20 / 21

Michaelmas Term 2024

C20 Robust Optimization

November 9, 2024

21 / 21

C20 Robust Optimization

Lecture 4

Kostas Margellos

University of Oxford



Tightness and expected probability of constraint violation

- 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

Recap

Stronger generalization statement for convex scenario programs

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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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 $1 - \epsilon$.

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

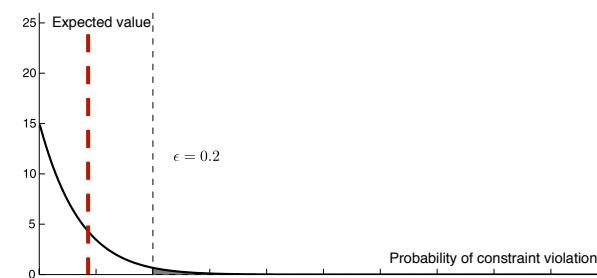
Distribution of the probability of constraint violation

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

- When is it **equal to** the tail distribution of a binomial random variable?

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

- What can we say about its expected value?



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

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

- $\mathbb{E}_{\sim \mathbb{P}^m}$ 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

Sketch of proof (cont'd)

- The expected value of $\mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0)$ will thus be upper-bounded by the expected value associated with the binomial tail

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

- This tail coincides with the cumulative distribution of the so called beta distribution with d and $m-d+1$ degrees of freedom
- Its density function (by differentiating it with respect to ϵ) is given by

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

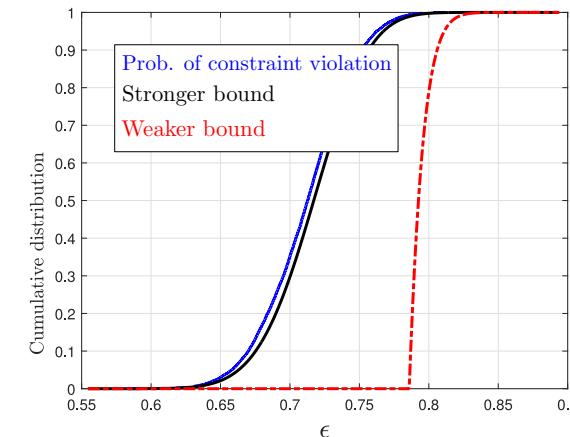
- The expected value associated with the binomial tail can be shown to be (by repeated integration by parts) $\frac{d}{m+1}$, i.e.,

$$\int_0^1 \epsilon f(\epsilon) d\epsilon = \frac{d}{m+1}$$

Expected probability of constraint violation

Sketch of proof

- The cumulative distribution of the probability of constraint violation is lower-bounded by the binomial bounds (say the strong one)



Expected probability of constraint violation

Expected probability of constraint violation – Convex scenario programs

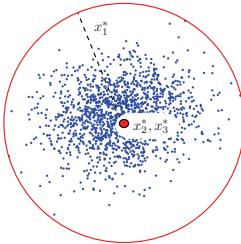
Let $d = n_x$, i.e. the # of decision variables in a convex scenario program. Then

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

- **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 $\mathbb{E}_{\sim \mathbb{P}^m} [\mathbb{P}(\delta \in \Delta : g(x_m^*, \delta) > 0)] \leq \rho$

$$\frac{d}{m+1} \leq \rho \Leftrightarrow m \geq \frac{d}{\rho} - 1$$

Example: Minimum radius disk problem revisited



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

$$\min_{x_1, x_2, x_3} x_1$$

subject to: $\sqrt{(y_i - x_3)^2 + (u_i - x_2)^2} \leq x_1$, 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 $\delta = (u, y)$?

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

Robust state feedback control design

Problem specifications

Consider the family of systems (each with n_x states and n_u inputs)

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

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

- Design a gain matrix K such that $u = Kx$ renders the closed loop system asymptotically stable.
- 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 ;
- 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:
- $$\dot{x} = (A(\delta_i) + B(\delta_i)K)x, \text{ for all } i = 1, \dots, m$$
- Restatement of the problem:
Find K such that $A(\delta_i) + B(\delta_i)K$ is Hurwitz for all $i = 1, \dots, m$.

Recall Lyapunov's stability condition

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

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

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

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

Robust state feedback control design (cont'd)

Three step procedure:

- 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

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

which leads to

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

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

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

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

Robust state feedback control design (cont'd)

- How to find P and Z such that

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

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

$$\min_{P,Z} 0 \quad [\text{any constant would work}]$$

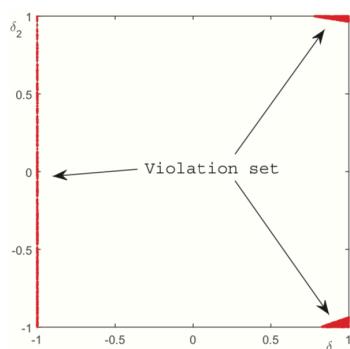
subject to $PA(\delta_i)^\top + Z^\top B(\delta_i)^\top + A(\delta_i)P + B(\delta_i)Z < 0$,
for all $i = 1, \dots, m$

- 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)

- Red regions illustrate the set of new δ 's for which x_m^* violates the constraints
- Example¹ refers to a 2-dimensional uncertainty vector δ



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

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 stable is at least $1 - \epsilon$

Probabilistic guarantees

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

$$\mathbb{P}\left(\delta : P^* A(\delta)^\top + (Z^*)^\top B(\delta)^\top + A(\delta)P^* + B(\delta)Z^* < 0\right) > 1 - \epsilon$$

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

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

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

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

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

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

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

Thank you for your attention!
Questions?

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

C20 Robust Optimization Appendix

Kostas Margellos
University of Oxford



Appendix: Proof of the main PAC learning theorem

Theorem

If a **compression set** \mathcal{C}_d with cardinality d exists, then

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{P}(\delta \in T \setminus H_m) \leq \epsilon \right\} \geq 1 - q(m, \epsilon)$$

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

Proof

- We assume existence of \mathcal{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

$$\mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \begin{array}{l} \exists \mathcal{C}_d \text{ such that } \mathbb{1}_{\mathcal{H}_d}(\delta_i) = \mathbb{1}_T(\delta_i), \text{ for all } i = 1, \dots, m \\ \text{and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \end{array} \right\} \leq q(m, \epsilon)$$

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

$$\begin{aligned} & \mathbb{P}^m \left\{ \bigcup_{\substack{\delta_1, \dots, \delta_m \\ \mathcal{C}_d}} : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \forall i \text{ and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \right\} \\ & \leq \sum_{\mathcal{C}_d} \mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \forall i \text{ and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \right\} \end{aligned}$$

- Existence of a compression set \mathcal{C}_d is equivalent to taking the “union”
- Union is taken with respect to all potential compression sets \mathcal{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)

- Pick a “new” δ

$$\mathbb{P} \left\{ \delta : \mathbb{1}_{H_d}(\delta) = \mathbb{1}_T(\delta) \text{ and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \right\} \leq 1 - \epsilon$$

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

$$\begin{aligned} & \mathbb{P}^{m-d} \left\{ \delta_{d+1}, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i) \text{ for all } i = d+1, \dots, m \right. \\ & \quad \left. \text{and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \right\} \\ &= \prod_{i=d+1}^m \mathbb{P} \left\{ \delta_i : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i) \text{ and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \right\} \\ &\leq (1 - \epsilon)^{m-d} \end{aligned}$$

Proof (cont'd)

- Without loss of generality let $\mathcal{C}_d = \{\delta_1, \dots, \delta_m\}$ and

$$\begin{aligned} \bar{\Delta} &= \left\{ \delta_1, \dots, \delta_d : \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \right\} \\ &= \left\{ \delta_1, \dots, \delta_d : \mathbb{P}(\delta : \mathbb{1}_{H_d}(\delta) \neq \mathbb{1}_T(\delta)) > \epsilon \right\} \end{aligned}$$

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

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

Pick a “new” δ

$$\begin{aligned} & \mathbb{P} \left\{ \delta : \mathbb{1}_{H_d}(\delta) = \mathbb{1}_T(\delta) \text{ and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \right\} \\ &= \mathbb{P} \left\{ \delta : \mathbb{1}_{H_d}(\delta) = \mathbb{1}_T(\delta) \right\} \leq 1 - \epsilon \end{aligned}$$

- 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)

Deconditioning ...

$$\begin{aligned} & \mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \forall i \text{ and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \right\} \\ &= \int_{\bar{\Delta}} \mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i) \text{ for all } i = 1, \dots, m \right. \\ & \quad \left. \text{and } \mathbb{P}(\delta \in T \setminus H_d) > \epsilon \mid \delta_1, \dots, \delta_d \in \bar{\Delta} \right\} d\mathbb{P}(d\delta_1, \dots, d\delta_d) \\ &\leq (1 - \epsilon)^{m-d} \end{aligned}$$

- The equality is due to the definition of the conditional probability
- The inequality follows from the obtained Bernoulli trials bound, since the conditional probability is equal to the derived expression for \mathbb{P}^{m-d}

Proof (cont'd)

Deconditioning ...

$$\begin{aligned} \mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \forall i \text{ and } \mathbb{P} \left(\delta \in T \setminus H_d \right) > \epsilon \right\} \\ \leq (1 - \epsilon)^{m-d} \end{aligned}$$

Desired statement was shown to be upper-bounded by

$$\begin{aligned} \sum_{C_d} \mathbb{P}^m \left\{ \delta_1, \dots, \delta_m : \mathbb{1}_{H_d}(\delta_i) = \mathbb{1}_T(\delta_i), \forall i \text{ and } \mathbb{P} \left(\delta \in T \setminus H_d \right) > \epsilon \right\} \\ \leq \sum_{C_d} (1 - \epsilon)^{m-d} \quad \left[\binom{m}{d} \text{ terms in the summation} \right] \\ = \binom{m}{d} (1 - \epsilon)^{m-d} \end{aligned}$$

Thank you for your attention!
Questions?

Contact at:

kostas.margellos@eng.ox.ac.uk