Decision Making Under Uncertainty: Lecture 1—Introduction and Background

Lecture 1

Ryan Cory-Wright

Spring 2024

Outline of Lecture 1

- Module Organization
  - Administriva
  - Class Overview and Motivation
- Probability Bootcamp
  - Fundamentals of Probability
  - Modes of Convergence
  - Limit Theorems and Concentration Inequalities
- Optimization Bootcamp
  - What is Tractable?
  - Convex Conic Optimization
  - Integer Optimization

Module Organization

Administriva

Course Information

- Time: Mondays 10am-1pm (will have regular breaks during lectures)
    - If I forget to give us a break, please remind me!

Course Information

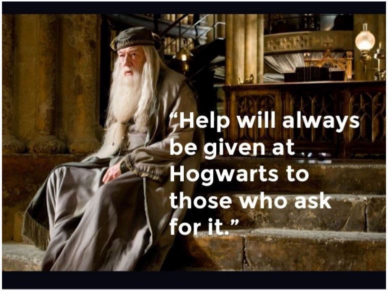
- Time: Mondays 10am-1pm (will have regular breaks during lectures)
  - If I forget to give us a break, please remind me!
- Place: ICBS LG19B

Course Information

- Time: Mondays 10am-1pm (will have regular breaks during lectures)
  - If I forget to give us a break, please remind me!
- Place: ICBS LG19B
- Module Leader: Ryan Cory-Wright, Business School Building, Room 393 (r.cory-wright@imperial.ac.uk, ryancorywright.github.io)

Course Information

- Time: Mondays 10am-1pm (will have regular breaks during lectures)
  - If I forget to give us a break, please remind me!
- Place: ICBS LG19B
- Module Leader: Ryan Cory-Wright, Business School Building, Room 393 (r.cory-wright@imperial.ac.uk, ryancorywright.github.io)
- Office hours: TBD (does 3-4 PM on Monday work for everyone?)



- Course materials: Distributed via Insendi.
- Suggested Prerequisites: Graduate-level courses in optimization and probability, or similar. Mathematical maturity.

Administration

Important dates:

- Homework 1: released week 2, due week 4 (10%)
- Homework 2: released week 4, due week 6 (10%)
- Homework 3: released week 6, due week 9 (10%)

Administration

Important dates:

- Homework 1: released week 2, due week 4 (10%)
- Homework 2: released week 4, due week 6 (10%)
- Homework 3: released week 6, due week 9 (10%)
- Critical paper review: paper you will be reviewing selected by week 2, to be presented in week $\in \{3, \ldots, 9\}$ (10%)

Administration

Important dates:

- Homework 1: released week 2, due week 4 (10%)
- Homework 2: released week 4, due week 6 (10%)
- Homework 3: released week 6, due week 9 (10%)
- Critical paper review: paper you will be reviewing selected by week 2, to be presented in week $\in \{3, \ldots, 9\}$ (10%)
- Quiz: 180 minutes in-class, week 7, on material from weeks 1–6 (30%)

Administration

Important dates:

- Homework 1: released week 2, due week 4 (10%)
- Homework 2: released week 4, due week 6 (10%)
- Homework 3: released week 6, due week 9 (10%)
- Critical paper review: paper you will be reviewing selected by week 2, to be presented in week $\in \{3, \ldots, 9\}$ (10%)
- Quiz: 180 minutes in-class, week 7, on material from weeks 1–6 (30%)
  - Suggestion: start HW3 after the quiz.
- Final project: short report due week 10, in-class presentation on project week 10 (30%)
  - Optional but highly encouraged: You should create a short project proposal outlining what you intend to do and hand it in week 6.
- Email policy: if you email me by the Friday before something is due, I'll aim to respond promptly for an assignment due on a Tuesday. However, no guarantees if you email later than that.

Project Ideas

- Incorporate decision-making under uncertainty into your research

Project Ideas

- Incorporate decision-making under uncertainty into your research
- Write a survey on a topic related to decision-making under uncertainty, and implement some methods related to this topic in a programming language of your choice

Project Ideas

- Incorporate decision-making under uncertainty into your research
- Write a survey on a topic related to decision-making under uncertainty, and implement some methods related to this topic in a programming language of your choice
- Explore a small idea related to decision-making under uncertainty

Project Ideas

- Incorporate decision-making under uncertainty into your research
- Write a survey on a topic related to decision-making under uncertainty, and implement some methods related to this topic in a programming language of your choice
- Explore a small idea related to decision-making under uncertainty
- Best-case scenario: When I was a Ph.D. student, some class projects turned into journal papers. E.g.,
    - Pareto Efficiency in Robust Optimization. D. Iancu and N. Trichakis. Management Science 60(1):130–147 (2014).
    - On polyhedral and second-order cone decompositions of semidefinite optimization problems. D. Bertsimas and R Cory-Wright. OR Letters 48(1):78–85 (2020).
    - Probabilistic guarantees in robust optimization. D. Bertsimas, D. Den Hertog, and J. Pauphilet. SIAM Journal on Optimization 31(4):2893–2920 (2021).

This Seems Quite Rigorous: Why Are we Working This Hard?

Fair question!

This Seems Quite Rigorous: Why Are we Working This Hard?

Fair question!

A goal of an MRes/PhD is to put you in a position where you could be competitive for an academic job when you graduate (or after you do a postdoc, depending on the field). This gets tougher each year, because competition from other institutions is getting fiercer.

This Seems Quite Rigorous: Why Are we Working This Hard?

Fair question!

A goal of an MRes/PhD is to put you in a position where you could be competitive for an academic job when you graduate (or after you do a postdoc, depending on the field). This gets tougher each year, because competition from other institutions is getting fiercer.

Realistically, to be competitive in the current (Operations) market, you need an accepted paper and a few papers with revisions, all in top journals (e.g. OR/MS), by the time you are in the final year of your PhD (feel free to ask questions about this in office hours). That means you need to be able to write papers that can get into top journals from early on in your degree.

This Seems Quite Rigorous: Why Are we Working This Hard?

Fair question!

A goal of an MRes/PhD is to put you in a position where you could be competitive for an academic job when you graduate (or after you do a postdoc, depending on the field). This gets tougher each year, because competition from other institutions is getting fiercer.

Realistically, to be competitive in the current (Operations) market, you need an accepted paper and a few papers with revisions, all in top journals (e.g. OR/MS), by the time you are in the final year of your PhD (feel free to ask questions about this in office hours). That means you need to be able to write papers that can get into top journals from early on in your degree.

To be in that position, you need to quickly pick up things that 15 years ago you might have learned across the first three years of an MRes/PhD. So, we will work hard this term to give you a good shot.

Grades and Philosophy Towards Amount of Content in Course

- Grades: They matter a lot at the undergraduate level, but I don't view them as important at the MRes/PhD level—you are here to learn how to do research, and you will be judged on how good your research is (I've *never* been asked for a transcript of my grad school grades, most faculty position applicants don't include their graduate level GPA). You should be here because you want to be here/learn because you want to learn.
- My philosophy in this class is to throw a lot of content at you, in hope some of it is useful. "Drinking from the firehose".
- We don't want to go so fast that you don't take anything in. So will periodically take temperature, adjust speed accordingly.
- Don't be alarmed if you feel that you are drowning at some points, grades will come out in the wash.

Who am I?

Bio:

- B.E (Hons) in Engineering Science, University of Auckland
- Ph.D. in Operations Research, MIT, advised by Dimitris Bertsimas
- Postdoctoral fellow, IBM Research (2022-23)
- Assistant Professor of Analytics and Operations at Imperial College Business School and Imperial-X since July 2023
- Hobbies: running (training for a marathon), cycling, skiing.

Who am I?

Bio:

- B.E (Hons) in Engineering Science, University of Auckland
- Ph.D. in Operations Research, MIT, advised by Dimitris Bertsimas
- Postdoctoral fellow, IBM Research (2022-23)
- Assistant Professor of Analytics and Operations at Imperial College Business School and Imperial-X since July 2023
- Hobbies: running (training for a marathon), cycling, skiing.

Research:

- Broadly interested in optimization (convex/mixed-integer/semidefinite/under uncertainty)
- And its applications in machine learning, statistics, renewable energy
- Recently involved in a collaboration with OCP (a large fertilizer manufacturer) to fully decarbonize their production system by investing $2 Bn USD in solar panels/batteries

Who am I?

Bio:

- B.E (Hons) in Engineering Science, University of Auckland
- Ph.D. in Operations Research, MIT, advised by Dimitris Bertsimas
- Postdoctoral fellow, IBM Research (2022-23)
- Assistant Professor of Analytics and Operations at Imperial College Business School and Imperial-X since July 2023
- Hobbies: running (training for a marathon), cycling, skiing.

Research:

- Broadly interested in optimization (convex/mixed-integer/semidefinite/under uncertainty)
- And its applications in machine learning, statistics, renewable energy
- Recently involved in a collaboration with OCP (a large fertilizer manufacturer) to fully decarbonize their production system by investing $2 Bn USD in solar panels/batteries
  - Using basically the techniques we learn in this class!

Who are you?

Good question! During next 10 mins, please write down on a piece of paper (or email me at r.cory-wright@imperial.ac.uk), the following:

Who are you?

Good question! During next 10 mins, please write down on a piece of paper (or email me at r.cory-wright@imperial.ac.uk), the following:

- Your name
- Your background in optimization and probability theory
- Why you are taking this class, and what you expect to get out of it
  - If you are auditing, whether you intend to complete assignments etc.
- Whether there is anything in the syllabus that you weren't expecting to learn, or anything that isn't in the syllabus that you were expecting to learn
- How many hours a week you are expecting to spend on each of: reading, homework, additional exercises, project
- Anything else you think I should know (e.g., "I'll be away in week 5 because I'll be at a conference").

Who are you?

Good question! During next 10 mins, please write down on a piece of paper (or email me at r.cory-wright@imperial.ac.uk), the following:

- Your name
- Your background in optimization and probability theory
- Why you are taking this class, and what you expect to get out of it
  - If you are auditing, whether you intend to complete assignments etc.
- Whether there is anything in the syllabus that you weren't expecting to learn, or anything that isn't in the syllabus that you were expecting to learn
- How many hours a week you are expecting to spend on each of: reading, homework, additional exercises, project
- Anything else you think I should know (e.g., "I'll be away in week 5 because I'll be at a conference").

Thanks! I'll aim to take feedback on board as I prep rest of module

Class Overview and Motivation

Why This Class?

- You have learned about optimization: a framework for, given a model of the world, making decisions that perform well for model

Why This Class?

- You have learned about optimization: a framework for, given a model of the world, making decisions that perform well for model
- But, this model of the world only exists *in our imagination*

Why This Class?

- You have learned about optimization: a framework for, given a model of the world, making decisions that perform well for model
- But, this model of the world only exists *in our imagination*
- In reality, we constructed this model using data, which may be uncertain. Why?

Why This Class?

- You have learned about optimization: a framework for, given a model of the world, making decisions that perform well for model
- But, this model of the world only exists *in our imagination*
- In reality, we constructed this model using data, which may be uncertain. Why?
    - Measurement error

Why This Class?

- You have learned about optimization: a framework for, given a model of the world, making decisions that perform well for model
- But, this model of the world only exists *in our imagination*
- In reality, we constructed this model using data, which may be uncertain. Why?
  - Measurement error
  - Implementation error

Why This Class?

- You have learned about optimization: a framework for, given a model of the world, making decisions that perform well for model
- But, this model of the world only exists *in our imagination*
- In reality, we constructed this model using data, which may be uncertain. Why?
  - Measurement error
  - Implementation error
  - Data might not have been observed yet

Why This Class?

- You have learned about optimization: a framework for, given a model of the world, making decisions that perform well for model
- But, this model of the world only exists *in our imagination*
- In reality, we constructed this model using data, which may be uncertain. Why?
    - Measurement error
    - Implementation error
    - Data might not have been observed yet
    - The future distribution may not look like the past

Why This Class?

- You have learned about optimization: a framework for, given a model of the world, making decisions that perform well for model
- But, this model of the world only exists *in our imagination*
- In reality, we constructed this model using data, which may be uncertain. Why?
  - Measurement error
  - Implementation error
  - Data might not have been observed yet
  - The future distribution may not look like the past
- If we want to guarantee that our optimization decisions perform well *in the real world*, we need to account for uncertainty in our models

Class Objectives and Structure

Objectives:

- Introduce you to three different modeling paradigms for decision-making under uncertainty

Class Objectives and Structure

Objectives:

- Introduce you to three different modeling paradigms for decision-making under uncertainty
- Provide background to explore the latest literature and apply it

Class Objectives and Structure

Objectives:

- Introduce you to three different modeling paradigms for decision-making under uncertainty
- Provide background to explore the latest literature and apply it
- (Help) prepare you to perform research in topics involving decision-making under uncertainty+ (Help) build background knowledge for performing research more broadly too

Class Objectives and Structure

Objectives:

- Introduce you to three different modeling paradigms for decision-making under uncertainty
- Provide background to explore the latest literature and apply it
- (Help) prepare you to perform research in topics involving decision-making under uncertainty+ (Help) build background knowledge for performing research more broadly too

Structure:

0. Background in Optimization and Probability (week 1)
1. Stochastic Optimization (weeks 2–4)
2. Robust Optimization (weeks 5–8)
3. Dynamic Optimization (weeks 9–10)

Paradigm 1: Stochastic Optimization

(Weeks 2–4)

- Model uncertainty by assuming uncertain parameters in optimization problem follow a joint probability distribution, which we know

Paradigm 1: Stochastic Optimization

(Weeks 2–4)

- Model uncertainty by assuming uncertain parameters in optimization problem follow a joint probability distribution, which we know
- Optimization over "random" parameters

Paradigm 1: Stochastic Optimization

(Weeks 2–4)

- Model uncertainty by assuming uncertain parameters in optimization problem follow a joint probability distribution, which we know
- Optimization over "random" parameters
- Typically (aim to) minimize expected cost with respect to a joint probability distribution
  - Given decision variables $\boldsymbol{x}$ in a known feasible region $\mathcal{X}$, uncertain parameters $\boldsymbol{\xi}$, and a known cost function $c(\boldsymbol{x}, \boldsymbol{\xi})$, solve

$$\min_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}[c(\boldsymbol{x}, \boldsymbol{\xi})]$$

Paradigm 1: Stochastic Optimization

(Weeks 2–4)

- Model uncertainty by assuming uncertain parameters in optimization problem follow a joint probability distribution, which we know
- Optimization over "random" parameters
- Typically (aim to) minimize expected cost with respect to a joint probability distribution
  - Given decision variables $\boldsymbol{x}$ in a known feasible region $\mathcal{X}$, uncertain parameters $\boldsymbol{\xi}$, and a known cost function $c(\boldsymbol{x}, \boldsymbol{\xi})$, solve

$$\min_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}[c(\boldsymbol{x}, \boldsymbol{\xi})]$$

- Appealing performance guarantees, but (a) might be hard to estimate joint probability distribution for $\boldsymbol{\xi}$, (b) probability theory can be intractable in high-dimensional settings (e.g., expectations hard to compute)

Paradigm 2: Robust Optimization

(Weeks 5–8)

- Model uncertainty by assuming nature selects uncertain parameters adversarially, but is bounded in her capacity to be adversarial

Paradigm 2: Robust Optimization

(Weeks 5–8)

- Model uncertainty by assuming nature selects uncertain parameters adversarially, but is bounded in her capacity to be adversarial
- Often yields a deterministic equivalent with a few more variables, using techniques from duality

Paradigm 2: Robust Optimization

(Weeks 5–8)

- Model uncertainty by assuming nature selects uncertain parameters adversarially, but is bounded in her capacity to be adversarial
- Often yields a deterministic equivalent with a few more variables, using techniques from duality
- More tractable than stochastic optimization, but also more conservative (why?)

Paradigm 2: Robust Optimization

(Weeks 5–8)

- Model uncertainty by assuming nature selects uncertain parameters adversarially, but is bounded in her capacity to be adversarial
- Often yields a deterministic equivalent with a few more variables, using techniques from duality
- More tractable than stochastic optimization, but also more conservative (why?)
- We will also look at distributionally robust optimization (DRO), which aims to combine performance guarantees of SO and tractability of RO. Need to understand SO and RO to understand DRO, so we look at SO and RO first

Paradigm 3: Dynamic Optimization

- Model uncertainty using a stagewise independence assumption to improve traceability

Paradigm 3: Dynamic Optimization

- Model uncertainty using a stagewise independence assumption to improve traceability
- Popular in some parts of Operations Research and Management Science, especially where things are predictably uncertain

Paradigm 3: Dynamic Optimization

- Model uncertainty using a stagewise independence assumption to improve traceability
- Popular in some parts of Operations Research and Management Science, especially where things are predictably uncertain
- Ali Aouad (LBS/MIT) is going to be running a full class on Dynamic Optimization this summer—we will briefly touch on it here, but I strongly encourage you to sign up for that class, especially if you would like to learn more about Dynamic Optimization

Aside: “Program” vs “Optimization Problem”

- Most classics texts refer to “Linear Programming”, “Stochastic Programming” etc. rather than “Linear Optimization Problem”

Aside: “Program” vs “Optimization Problem”

- Most classics texts refer to “Linear Programming”, “Stochastic Programming” etc. rather than “Linear Optimization Problem”
- “Program” originally meant “ordered list of events to take place/procedures to be followed/schedule”. Dantzig and others popularized term in the 50s, before computers were widely available.

Aside: "Program" vs "Optimization Problem"

- Most classics texts refer to "Linear Programming", "Stochastic Programming" etc. rather than "Linear Optimization Problem"
- "Program" originally meant "ordered list of events to take place/procedures to be followed/schedule". Dantzig and others popularized term in the 50s, before computers were widely available.
- Today, everyone and their dog who aren't research-active in Operations think "program" means "computer-stuff". So I and lots of others use "optimization problem" instead, and you should too!

Aside: “Program” vs “Optimization Problem”

- Most classics texts refer to “Linear Programming”, “Stochastic Programming” etc. rather than “Linear Optimization Problem”
- “Program” originally meant “ordered list of events to take place/procedures to be followed/schedule”. Dantzig and others popularized term in the 50s, before computers were widely available.
- Today, everyone and their dog who aren’t research-active in Operations think “program” means “computer-stuff”. So I and lots of others use “optimization problem” instead, and you should too!
- If you see a textbook or journal article that uses “program” rather than “optimization problem”, don’t worry, it means the same thing.

Common Threads

- Class requires knowledge of optimization and probability $\rightarrow$ rest of lecture reviews optimization and probability

Common Threads

- Class requires knowledge of optimization and probability $\rightarrow$ rest of lecture reviews optimization and probability
- Material in lecture isn't directly examinable, only to extent we use it in subsequent lectures
- Don't worry if you don't know all the material. You'll learn

Let's break for 5 minutes here.

Probability Bootcamp

Fundamentals of Probability

Don’t Panic

The notation/language in the next couple of slides might not be familiar.

> *The beginner should not be discouraged if he finds that he does not have the prerequisites for reading the prerequisites*
> *–Paul Halmos*

Don't Panic

The notation/language in the next couple of slides might not be familiar.

> *The beginner should not be discouraged if he finds that he does not have the prerequisites for reading the prerequisites*
> *–Paul Halmos*

- A class on probability is a prerequisite. But. . .

Don't Panic

The notation/language in the next couple of slides might not be familiar.

> *The beginner should not be discouraged if he finds that he does not have the prerequisites for reading the prerequisites*
> *–Paul Halmos*

- A class on probability is a prerequisite. But...
- If you need to, you can read up on this in chapter 1 of Probability and Random Processes by Grimmett and Stirzaker

Don't Panic

The notation/language in the next couple of slides might not be familiar.

> *The beginner should not be discouraged if he finds that he does not have the prerequisites for reading the prerequisites*
> *–Paul Halmos*

- A class on probability is a prerequisite. But...
- If you need to, you can read up on this in chapter 1 of Probability and Random Processes by Grimmett and Stirzaker
- We are about to go through the most useful conclusions of a first course on probability. So we will quote results, but not do any proofs

Probability Spaces: Gory Definitions

- Ω: a sample space of possible outcomes (e.g. from an experiment). Cardinality could be finite, countably infinite, or uncountably infinite

Probability Spaces: Gory Definitions

- Ω: a sample space of possible outcomes (e.g. from an experiment). Cardinality could be finite, countably infinite, or uncountably infinite
- $\mathcal{F}$ a sigma-algebra, or set of all subsets of Ω

Probability Spaces: Gory Definitions

- Ω: a sample space of possible outcomes (e.g. from an experiment). Cardinality could be finite, countably infinite, or uncountably infinite
- $\mathcal{F}$ a sigma-algebra, or set of all subsets of Ω
  - $\varnothing \in \mathcal{F}$
  - If $\boldsymbol{A} \in \mathcal{F}$ then $\mathcal{A}^c \in \mathcal{F}$
  - If $\boldsymbol{A}_i \in \mathcal{F}\ \forall \boldsymbol{i}$ then $\cup_i \boldsymbol{A}_i \in \mathcal{F}$

Probability Spaces: Gory Definitions

- Ω: a sample space of possible outcomes (e.g. from an experiment). Cardinality could be finite, countably infinite, or uncountably infinite
- $\mathcal{F}$ a sigma-algebra, or set of all subsets of Ω
    - $\varnothing \in \mathcal{F}$
    - If $\boldsymbol{A} \in \mathcal{F}$ then $\mathcal{A}^c \in \mathcal{F}$
    - If $\boldsymbol{A}_i \in \mathcal{F}\ \forall \boldsymbol{i}$ then $\cup_i \boldsymbol{A}_i \in \mathcal{F}$
- $\mathbb{P}$ a probability measure, which assigns a non-negative weight to each measurable subset $\boldsymbol{A} \in \mathcal{F}$ of Ω such that (Kolmogorov's Axioms)
    - $\mathbb{P}(\varnothing) = 0$
    - $\mathbb{P}(\Omega) = 1$
    - If $\boldsymbol{A}_i \in \mathcal{F}$ are disjoint then $\mathbb{P}(\cup_i \boldsymbol{A}_i) = \sum_i \mathbb{P}(\boldsymbol{A}_i)$

Probability Spaces: Gory Definitions

- Ω: a sample space of possible outcomes (e.g. from an experiment). Cardinality could be finite, countably infinite, or uncountably infinite
- $\mathcal{F}$ a sigma-algebra, or set of all subsets of Ω
  - $\varnothing \in \mathcal{F}$
  - If $\boldsymbol{A} \in \mathcal{F}$ then $\mathcal{A}^c \in \mathcal{F}$
  - If $\boldsymbol{A}_i \in \mathcal{F}\ \forall \boldsymbol{i}$ then $\cup_i \boldsymbol{A}_i \in \mathcal{F}$
- $\mathbb{P}$ a probability measure, which assigns a non-negative weight to each measurable subset $\boldsymbol{A} \in \mathcal{F}$ of Ω such that (Kolmogorov's Axioms)
  - $\mathbb{P}(\varnothing) = 0$
  - $\mathbb{P}(\Omega) = 1$
  - If $\boldsymbol{A}_i \in \mathcal{F}$ are disjoint then $\mathbb{P}(\cup_i \boldsymbol{A}_i) = \sum_i \mathbb{P}(\boldsymbol{A}_i)$

Probability Space

Let Ω be a sample space, $\mathcal{F}$ be a Sigma algebra, and $\mathbb{P}$ be a probability measure defined on $(\Omega, \mathcal{F})$. Then, we say that the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space

Probability Spaces: A Worked Example

Worked example of flipping a fair coin once: Outcomes are H and T

Probability Spaces: A Worked Example

Worked example of flipping a fair coin once: Outcomes are H and T

- $\Omega = \{H, T\}$

Probability Spaces: A Worked Example

Worked example of flipping a fair coin once: Outcomes are H and T

- $\Omega = \{H, T\}$
- $\mathcal{F} = \{\varnothing, \{H\}, \{T\}, \{H, T\}\}$

Probability Spaces: A Worked Example

Worked example of flipping a fair coin once: Outcomes are H and T

- $\Omega = \{H, T\}$
- $\mathcal{F} = \{\varnothing, \{H\}, \{T\}, \{H, T\}\}$
- $\mathbb{P} : \mathbb{P}(\varnothing) = 0, \mathbb{P}(H) = \mathbb{P}(T) = 1/2, \mathbb{P}(\{H, T\}) = 1$

Probability Spaces: A Worked Example

Worked example of flipping a fair coin once: Outcomes are H and T

- $\Omega = \{H, T\}$
- $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$
- $\mathbb{P} : \mathbb{P}(\emptyset) = 0, \mathbb{P}(H) = \mathbb{P}(T) = 1/2, \mathbb{P}(\{H, T\}) = 1$

What about if we flip a fair coin twice?

Why do we Need Sigma Algebras?

It's pretty intuitive that we need to define a sample space and a probability measure to talk about the possible outcomes of an experiment. But why do we need Sigma algebras?

Why do we Need Sigma Algebras?

It's pretty intuitive that we need to define a sample space and a probability measure to talk about the possible outcomes of an experiment. But why do we need Sigma algebras?

Answer: There are non-measurable subsets of $[0, 1]^n$, which are challenging to assign a probability to in a consistent way. So we screen them out by only assigning probabilities to measurable subsets. But this is mainly an issue when writing proofs, not in practice.

Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A random variable $\boldsymbol{X}$ is a real-valued function $\boldsymbol{X} : \Omega \to \mathbb{R}$ such that the set $\{\omega : \boldsymbol{X}(\omega) \leq c\}$ is $\mathcal{F}$-measurable for each $c \in \mathbb{R}$

Related Terminology

- We say that an event $\boldsymbol{A} \in \mathcal{F}$ occurs *almost surely* if it occurs with probability 1. Or equivalently, if the event does not occur with probability 0, i.e., $\mathbb{P}(\boldsymbol{A}^c) = 0$

Related Terminology

- We say that an event $\boldsymbol{A} \in \mathcal{F}$ occurs *almost surely* if it occurs with probability 1. Or equivalently, if the event does not occur with probability 0, i.e., $\mathbb{P}(\boldsymbol{A}^c) = 0$
  - But be careful with definitions! E.g., if we let $\boldsymbol{X}$ be a uniform random variable on $[0, 1]$ and $c \in [0, 1]$ be a constant then the event $\boldsymbol{X} \neq c$ almost surely occurs. But for any $\boldsymbol{X}$, we can pick some $d \in [0, 1]$ ex-post observing $\boldsymbol{X}$ such that $\boldsymbol{X} = d$

Modes of Convergence

Why do we Need to Know About Modes of Convergence?

- In stochastic optimization, we often want to talk about a situation where we have access to a finite number of data points, which we take to be i.i.d. observations from a stochastic process

Why do we Need to Know About Modes of Convergence?

- In stochastic optimization, we often want to talk about a situation where we have access to a finite number of data points, which we take to be i.i.d. observations from a stochastic process
- Accordingly, we would like to talk about how fast estimators converge towards the "true" stochastic process

Why do we Need to Know About Modes of Convergence?

- In stochastic optimization, we often want to talk about a situation where we have access to a finite number of data points, which we take to be i.i.d. observations from a stochastic process
- Accordingly, we would like to talk about how fast estimators converge towards the "true" stochastic process
- Modes of convergence provide us with a rigorous way of talking about the speed of convergence

Almost Sure Convergence

Almost Sure Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\{\boldsymbol{X}_i\}_{i \in \mathbb{N}}, \boldsymbol{X}$ be random variables. Suppose that $\boldsymbol{A} \in \mathcal{F}$ is a measurable set such that $\mathbb{P}(\boldsymbol{A}) = 1$ and for all $\omega \in \mathcal{A}$ we have

$$\boldsymbol{X}_i(\omega) \to \boldsymbol{X}(\omega) \text{ as } i \to \infty.$$

Then, we say that $\boldsymbol{X}_i \stackrel{a.s.}{\to} \boldsymbol{X}$.

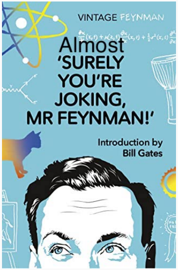
Almost Sure Convergence

Almost Sure Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}, \boldsymbol{X}$ be random variables. Suppose that $\boldsymbol{A} \in \mathcal{F}$ is a measurable set such that $\mathbb{P}(\boldsymbol{A}) = 1$ and for all $\omega \in \mathcal{A}$ we have

$$\boldsymbol{X}_i(\omega) \to \boldsymbol{X}(\omega) \text{ as } i \to \infty.$$

Then, we say that $\boldsymbol{X}_i \stackrel{a.s.}{\to} \boldsymbol{X}$.



Convergence in Probability

Convergence in Probability Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables. Suppose that for every $\epsilon > 0$ we have that

$$\lim_{i \to \infty} \mathbb{P}(|\boldsymbol{X}_i - \boldsymbol{X}| \geq \epsilon) = 0.$$

Then, we say that $\boldsymbol{X}_i \xrightarrow{p.} \boldsymbol{X}$.

Convergence in Probability

Convergence in Probability Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables. Suppose that for every $\epsilon > 0$ we have that

$$\lim_{i \to \infty} \mathbb{P}(|\boldsymbol{X}_i - \boldsymbol{X}| \geq \epsilon) = 0.$$

Then, we say that $\boldsymbol{X}_i \xrightarrow{p.} \boldsymbol{X}$.

Note: almost sure convergence implies convergence in probability, not necessarily other way around.

Convergence in Probability

Convergence in Probability Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables. Suppose that for every $\epsilon > 0$ we have that

$$\lim_{i \to \infty} \mathbb{P}(|\boldsymbol{X}_i - \boldsymbol{X}| \geq \epsilon) = 0.$$

Then, we say that $\boldsymbol{X}_i \overset{p.}{\to} \boldsymbol{X}$.

Note: almost sure convergence implies convergence in probability, not necessarily other way around. Can you think of a counterexample?

Convergence in Probability

Convergence in Probability Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables. Suppose that for every $\epsilon > 0$ we have that

$$\lim_{i \to \infty} \mathbb{P}(|\boldsymbol{X}_i - \boldsymbol{X}| \geq \epsilon) = 0.$$

Then, we say that $\boldsymbol{X}_i \overset{p.}{\to} \boldsymbol{X}$.

Note: almost sure convergence implies convergence in probability, not necessarily other way around. Can you think of a counterexample?

Let X_i be uniformly distributed on $[\frac{i}{2^k}, \frac{i+1}{2^k}]$ where k is such that $k \leq \log_2(n)$ and $2^k + i = n$. Then, $X_1 \sim \mathcal{U}[0,1]$, $X_2 \sim \mathcal{U}[0,1/2], X_3 \sim \mathcal{U}[1/2,1], X_4 \sim \mathcal{U}[0,1/4]$ etc.

Convergence in Probability

Convergence in Probability Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables. Suppose that for every $\epsilon > 0$ we have that

$$\lim_{i \to \infty} \mathbb{P}(|\boldsymbol{X}_i - \boldsymbol{X}| \geq \epsilon) = 0.$$

Then, we say that $\boldsymbol{X}_i \xrightarrow{p.} \boldsymbol{X}$.

Note: almost sure convergence implies convergence in probability, not necessarily other way around. Can you think of a counterexample?

Let X_i be uniformly distributed on $[\frac{i}{2^k}, \frac{i+1}{2^k}]$ where k is such that $k \leq \log_2(n)$ and $2^k + i = n$. Then, $X_1 \sim \mathcal{U}[0,1]$, $X_2 \sim \mathcal{U}[0,1/2], X_3 \sim \mathcal{U}[1/2,1], X_4 \sim \mathcal{U}[0,1/4]$ etc.

Thus, $\mathbb{P}(|X_i| > \epsilon) \to 0$, but $X_i(\omega) = 1$ infinitely often.

Convergence in Distribution

Convergence in Distribution Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables with CDFs F_i, F. Suppose that for every x where F_i is continuous we have that

$$\lim_{i \to \infty} F_i(x) = F(x)$$

Then, we say that $\boldsymbol{X}_i \xrightarrow{d} \boldsymbol{X}$.

Convergence in Distribution

Convergence in Distribution Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables with CDFs F_i, F. Suppose that for every x where F_i is continuous we have that

$$\lim_{i \to \infty} F_i(x) = F(x)$$

Then, we say that $\boldsymbol{X}_i \xrightarrow{d} \boldsymbol{X}$.

Note: Convergence in probability implies convergence in distribution, but not necessarily the other way around.

Convergence in Distribution

Convergence in Distribution Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables with CDFs F_i, F. Suppose that for every x where F_i is continuous we have that

$$\lim_{i \to \infty} F_i(x) = F(x)$$

Then, we say that $\boldsymbol{X}_i \xrightarrow{d} \boldsymbol{X}$.

Note: Convergence in probability implies convergence in distribution, but not necessarily the other way around. Can you think of a counterexample?

Convergence in Distribution

Convergence in Distribution Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables with CDFs F_i, F. Suppose that for every x where F_i is continuous we have that

$$\lim_{i \to \infty} F_i(x) = F(x)$$

Then, we say that $\boldsymbol{X}_i \xrightarrow{d} \boldsymbol{X}$.

Note: Convergence in probability implies convergence in distribution, but not necessarily the other way around. Can you think of a counterexample?

Let $X_1 \sim \mathcal{N}(0,1)$, $X_i = (-1)^i X_1$. Then, X_i's equal in distribution, but clearly do not converge in probability.

Limit Theorems and Concentration Inequalities

Laws of Large Numbers

Strong Law of Large Numbers

Let $\{\boldsymbol{X}_i\}_{i \in \mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[|\boldsymbol{X}_i|] < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mathbb{E}[\boldsymbol{X}_1] \tag{1}$$

Laws of Large Numbers

Strong Law of Large Numbers

Let $\{\boldsymbol{X}_i\}_{i \in \mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[|\boldsymbol{X}_i|] < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}[\boldsymbol{X}_1] \tag{1}$$

Weak Law of Large Numbers

Let $\{\boldsymbol{X}_i\}_{i \in \mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[|\boldsymbol{X}_i|] < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.} \mathbb{E}[\boldsymbol{X}_1] \tag{2}$$

Laws of Large Numbers

Strong Law of Large Numbers

Let $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[|\boldsymbol{X}_i|] < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mathbb{E}[\boldsymbol{X}_1] \tag{1}$$

Weak Law of Large Numbers

Let $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[|\boldsymbol{X}_i|] < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{p.}{\to} \mathbb{E}[\boldsymbol{X}_1] \tag{2}$$

There exist versions of both laws that hold under weaker assumptions than i.i.d.ness, e.g., pairwise independence

Central Limit Theorem

Central Limit Theorem

Let $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2. Then,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \tag{3}$$

What if we Have a Finite Amount of Data?

I hear you say "But in practice, we have access to a finite amount of training data, so we will never actually attain these limits!"

What if we Have a Finite Amount of Data?

I hear you say "But in practice, we have access to a finite amount of training data, so we will never actually attain these limits!"

Fair, SLLN/CLT provide good motivation for stochastic optimization, but we also need results that work with finite amounts of data

What if we Have a Finite Amount of Data?

I hear you say "But in practice, we have access to a finite amount of training data, so we will never actually attain these limits!"

Fair, SLLN/CLT provide good motivation for stochastic optimization, but we also need results that work with finite amounts of data

Berry-Esseen Theorem

Let $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[\boldsymbol{X}_i] = 0$, $\mathbb{E}[\boldsymbol{X}_i^2] = \sigma^2 < \infty$, $\mathbb{E}[|\boldsymbol{X}|^3] = \rho < \infty$. Then, define $Y_n := \frac{\sum_{i=1}^n X_i}{n}$ with CDF F_n. There exists some positive constant $C < 0.4748$ such that

$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^2\sqrt{n}}, \tag{4}$$

where Φ is the CDF of a standard normal

What if we Have a Finite Amount of Data?

I hear you say "But in practice, we have access to a finite amount of training data, so we will never actually attain these limits!"

Fair, SLLN/CLT provide good motivation for stochastic optimization, but we also need results that work with finite amounts of data

Berry-Esseen Theorem

Let $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[\boldsymbol{X}_i] = 0$, $\mathbb{E}[\boldsymbol{X}_i^2] = \sigma^2 < \infty$, $\mathbb{E}[|\boldsymbol{X}|^3] = \rho < \infty$. Then, define $Y_n := \frac{\sum_{i=1}^n X_i}{n}$ with CDF F_n. There exists some positive constant $C < 0.4748$ such that

$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^2\sqrt{n}}, \tag{4}$$

where Φ is the CDF of a standard normal

This result says that sample averages behave more and more like normal distributions as n gets larger

Markov's Inequality

For any non-negative random variable $\boldsymbol{X}$ and any $t \in \mathbb{R}_+$, we have

Markov's Inequality

$$\mathbb{P}(\boldsymbol{X} > t) \leq \min\left(1, \frac{\mathbb{E}[\boldsymbol{X}]}{t}\right) \quad (5)$$

Markov's Inequality

For any non-negative random variable $\boldsymbol{X}$ and any $t \in \mathbb{R}_+$, we have

Markov's Inequality

$$\mathbb{P}(\boldsymbol{X} > t) \leq \min\left(1, \frac{\mathbb{E}[\boldsymbol{X}]}{t}\right) \tag{5}$$

A weak but very general result about likelihood of "extreme" events

Markov's Inequality

For any non-negative random variable $\boldsymbol{X}$ and any $t \in \mathbb{R}_+$, we have

Markov's Inequality

$$\mathbb{P}(\boldsymbol{X} > t) \leq \min\left(1, \frac{\mathbb{E}[\boldsymbol{X}]}{t}\right) \tag{5}$$

A weak but very general result about likelihood of "extreme" events

How would we prove this?

Markov's Inequality

For any non-negative random variable $\boldsymbol{X}$ and any $t \in \mathbb{R}_+$, we have

Markov's Inequality

$$\mathbb{P}(\boldsymbol{X} > t) \leq \min\left(1, \frac{\mathbb{E}[\boldsymbol{X}]}{t}\right) \tag{5}$$

A weak but very general result about likelihood of "extreme" events

How would we prove this?

Let $\mathbb{E}[\boldsymbol{X}] = \mathbb{P}(\boldsymbol{X} > a)\mathbb{E}[\boldsymbol{X}|\boldsymbol{X} > a] + \mathbb{P}(\boldsymbol{X} \leq a)\mathbb{E}[\boldsymbol{X}|\boldsymbol{X} \leq a]$.

Therefore, $\mathbb{E}[\boldsymbol{X}] \geq \mathbb{P}(\boldsymbol{X} > a)\mathbb{E}[\boldsymbol{X}|\boldsymbol{X} > a] \geq a\mathbb{P}(\boldsymbol{X} > a)$

Chebyshev's Inequality

For any random variable $\boldsymbol{X}$ with finite variance σ^2 and expected value μ

Chebyshev's Inequality

$$\mathbb{P}(|\boldsymbol{X} - \mu| \geq t\sigma) \leq \min(1, \frac{1}{t^2}) \tag{6}$$

Chebyshev's Inequality

For any random variable $\boldsymbol{X}$ with finite variance σ^2 and expected value μ

Chebyshev's Inequality

$$\mathbb{P}(|\boldsymbol{X} - \mu| \geq t\sigma) \leq \min(1, \frac{1}{t^2}) \tag{6}$$

This is a slightly stronger result about the likelihood of extreme events

Hoeffding's Inequality

For a sequence of i.i.d. Bernoulli(p) random variables $\boldsymbol{X}_i$ we have

Hoeffding's Inequality

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \boldsymbol{X}_i - np\right| \geq nt\right) \leq 2\exp\left(\frac{-nt^2}{2}\right) \quad \forall t > 0 \tag{7}$$

Hoeffding's Inequality

For a sequence of i.i.d. Bernoulli(p) random variables $\boldsymbol{X}_i$ we have

Hoeffding's Inequality

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \boldsymbol{X}_i - np\right| \geq nt\right) \leq 2\exp\left(\frac{-nt^2}{2}\right) \quad \forall t > 0 \tag{7}$$

Conclusion: i.i.d.ness lets us concentrate uncertainty exponentially

McDiarmind's Inequality

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of n independent random variables $\mathcal{X}_i$ for each i, which almost surely have ranges $\mathcal{X}_i$. Let f satisfy the bounded differences condition

$$\sup_{\bar{x} \in \mathcal{X}_i} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \bar{x}, x_{i+1}, \ldots, x_n)| \leq c_i.$$

McDiarmind's Inequality

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of n independent random variables $\mathcal{X}_i$ for each i, which almost surely have ranges $\mathcal{X}_i$. Let f satisfy the bounded differences condition

$$\sup_{\bar{x} \in \mathcal{X}_i} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \bar{x}, x_{i+1}, \ldots, x_n)| \leq c_i.$$

Then, f satisfies the inequality:

McDiarmind's Inequality

$$\mathbb{P}\left(|f(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n) - \mathbb{E}[f\left(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n\right)]| \geq \epsilon\right) \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{8}$$

McDiarmind's Inequality

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of n independent random variables $\mathcal{X}_i$ for each i, which almost surely have ranges $\mathcal{X}_i$. Let f satisfy the bounded differences condition

$$\sup_{\bar{x} \in \mathcal{X}_i} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \bar{x}, x_{i+1}, \ldots, x_n)| \leq c_i.$$

Then, f satisfies the inequality:

McDiarmind's Inequality

$$\mathbb{P}\left(|f(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n) - \mathbb{E}[f\left(\boldsymbol{X}_1, \ldots \boldsymbol{X}_n\right)]| \geq \epsilon\right) \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{8}$$

Conclusion: functions of independent and bounded random variables concentrate exponentially

We just “covered” quite a lot of content!

Let’s break for 10 minutes here.

Optimization Bootcamp

Basic Terminology: What is an Optimization Problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) \tag{9}$$

$$\text{s.t.} \quad f_i(\boldsymbol{x}) \leq 0, \quad \forall i \in [m_1], \tag{10}$$

$$h_j(\boldsymbol{x}) = 0, \quad \forall j \in [m_2]. \tag{11}$$

Basic Terminology: What is an Optimization Problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad f(\boldsymbol{x}) \tag{9}$$

$$\text{s.t.} \quad f_i(\boldsymbol{x}) \leq 0, \quad \forall i \in [m_1], \tag{10}$$

$$h_j(\boldsymbol{x}) = 0, \quad \forall j \in [m_2]. \tag{11}$$

- Decision variables: $\boldsymbol{x} \in \mathbb{R}^n$ is the vector to be chosen
- Objective function: f is to be minimized
- Inequality constraints: f_i, equality constraints: h_j

Basic Terminology: What is an Optimization Problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) \tag{9}$$

$$\text{s.t.} \quad f_i(\boldsymbol{x}) \leq 0, \quad \forall i \in [m_1], \tag{10}$$

$$h_j(\boldsymbol{x}) = 0, \quad \forall j \in [m_2]. \tag{11}$$

- Decision variables: $\boldsymbol{x} \in \mathbb{R}^n$ is the vector to be chosen
- Objective function: f is to be minimized
- Inequality constraints: f_i, equality constraints: h_j

Variations: maximize objective, multiple objectives

How Expressive is Optimization?

We can phrase almost anything as an optimization problem.

E.g., Fermat's Last Theorem

$$\begin{aligned} \min_{x,y,z,n} \quad & (x^n + y^n - z^n)^2 \\ \text{s.t.} \quad & x, y, z \geq 1, n \geq 3, x, y, z, n \text{ Integer.} \end{aligned}$$

How Expressive is Optimization?

We can phrase almost anything as an optimization problem.

E.g., Fermat's Last Theorem

$$\begin{aligned} \min_{x,y,z,n} \quad & (x^n + y^n - z^n)^2 \\ \text{s.t.} \quad & x, y, z \geq 1, n \geq 3, x, y, z, n \text{ Integer.} \end{aligned}$$

A good question to ask is "what optimization problems can we solve?"

What is Tractable?

What Makes Optimization Tractable?

Attempt #1: linear optimization problems are tractable —we can solve them via the simplex method or IPMs

What Makes Optimization Tractable?

Attempt #1: linear optimization problems are tractable —we can solve them via the simplex method or IPMs

From "In Pursuit of the Traveling Salesman" by Bill Cook

News of the general linear-programming model, and the simplex algorithm for its solution, was delivered by George Dantzig in 1948 at a meeting held at the University of Wisconsin. The event was a defining moment for Dantzig, who has described often its proceedings.

What Makes Optimization Tractable?

Attempt #1: linear optimization problems are tractable —we can solve them via the simplex method or IPMs

From "In Pursuit of the Traveling Salesman" by Bill Cook

News of the general linear-programming model, and the simplex algorithm for its solution, was delivered by George Dantzig in 1948 at a meeting held at the University of Wisconsin. The event was a defining moment for Dantzig, who has described often its proceedings.

Like many good stories, repeated telling may have shifted a few details over the years, but all versions capture the spirit of a nervous rising star facing a large and distinguished group of mathematicians and economists. During the discussion following Dantzig's lecture, Harold Hotelling, great in both academic stature and physical size, rose from his seat, stated simply, "But we all know the world is nonlinear," and sat down. Dantzig was lost for a reply to such a sweeping criticism.

What Makes Optimization Tractable?

Attempt #1: linear optimization problems are tractable
—we can solve them via the simplex method or IPMs

From "In Pursuit of the Traveling Salesman" by Bill Cook

Suddenly another hand in the audience was raised. It was John von Neumann. "Mr. Chairman, Mr. Chairman," he said, "if the speaker does not mind, I would like to reply for him." Naturally I agreed. von Neumann said: "The speaker titled his talk 'linear programming' and carefully stated his axioms. If you have an application that satisfies the axioms, well use it. If it does not, then don't."

What Makes Optimization Tractable?

Attempt #1: linear optimization problems are tractable
—we can solve them via the simplex method or IPMs

From "In Pursuit of the Traveling Salesman" by Bill Cook

Suddenly another hand in the audience was raised. It was John von Neumann. "Mr. Chairman, Mr. Chairman," he said, "if the speaker does not mind, I would like to reply for him." Naturally I agreed. von Neumann said: "The speaker titled his talk 'linear programming' and carefully stated his axioms. If you have an application that satisfies the axioms, well use it. If it does not, then don't."

Conclusions:

- Dantzig and von Neumann are right: Linear is tractable

What Makes Optimization Tractable?

Attempt #1: linear optimization problems are tractable
—we can solve them via the simplex method or IPMs

From "In Pursuit of the Traveling Salesman" by Bill Cook

Suddenly another hand in the audience was raised. It was John von Neumann. "Mr. Chairman, Mr. Chairman," he said, "if the speaker does not mind, I would like to reply for him." Naturally I agreed. von Neumann said: "The speaker titled his talk 'linear programming' and carefully stated his axioms. If you have an application that satisfies the axioms, well use it. If it does not, then don't."

Conclusions:

- Dantzig and von Neumann are right: Linear is tractable
- Hotelling is right (Dantzig later admits as much): The world is nonlinear, and saying "linear is tractable" is not sufficient

What Makes Optimization Problems Tractable?

Attempt #2: Convex Optimization Problems Are Tractable

What Makes Optimization Problems Tractable?

Attempt #2: Convex Optimization Problems Are Tractable

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

What Makes Optimization Problems Tractable?

Attempt #2: Convex Optimization Problems Are Tractable

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

This quote is put up quite often at conferences ...

... usually, to justify using a heuristic on a non-convex problem

What Makes Optimization Problems Tractable?

Attempt #2: Convex Optimization Problems Are Tractable

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

This quote is put up quite often at conferences . . .

. . . usually, to justify using a heuristic on a non-convex problem

Let's take it to be true for a few slides

Convex Functions

Definition
A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if for each $x, y \in \mathbb{R}^n$ and every $\lambda \in [0, 1]$ we have that

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

If f is differentiable, an equivalent definition is:

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if for each $x, y \in \mathbb{R}^n$ we have that:

$$f(x) \geq f(y) + \nabla f(x)^\top (y - x).$$

If f is twice differentiable, an equivalent definition is:

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if its hessian, $\nabla^2 f(x)$ is positive semidefinite over the entire domain

Convex Functions

Definition
A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if for each $x, y \in \mathbb{R}^n$ and every $\lambda \in [0, 1]$ we have that

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y).$$

If f is differentiable, an equivalent definition is:

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if for each $x, y \in \mathbb{R}^n$ we have that:

$$f(x) \geq f(y) + \nabla f(x)^\top (y - x).$$

If f is twice differentiable, an equivalent definition is:

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if its hessian, $\nabla^2 f(x)$ is positive semidefinite over the entire domain

I like to think of this as "convex=holds water"

An Exercise in Convexity

Classify the following sets as convex or non-convex

- $\{\boldsymbol{x} : \boldsymbol{x} \in [0, 1]^n\}$

An Exercise in Convexity

Classify the following sets as convex or non-convex

- $\{\boldsymbol{x} : \boldsymbol{x} \in [0, 1]^n\}$
- $\{(x, \theta) \in \mathbb{R}^2 : \theta \geq x^3\}$
- Set of prime numbers

An Exercise in Convexity

Classify the following sets as convex or non-convex

- $\{\boldsymbol{x} : \boldsymbol{x} \in [0,1]^n\}$
- $\{(x,\theta) \in \mathbb{R}^2 : \theta \geq x^3\}$
- Set of prime numbers
- The dual cone $\mathcal{C}^* := \{\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{y} \leq 0 \ \forall \boldsymbol{x} \in \mathcal{C}\}$ for an arbitrary set $\mathcal{C}$.
- The polar set $\mathcal{C}^o := \{\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{y} \leq 1 \ \forall \boldsymbol{x} \in \mathcal{C}\}$ for an arbitrary set $\mathcal{C}$.
- Set of rank-one matrices $\{\boldsymbol{x}\boldsymbol{x}^\top : \boldsymbol{x} \in \mathbb{R}^n\}$

An Exercise in Convexity

Classify the following sets as convex or non-convex

- $\{\boldsymbol{x} : \boldsymbol{x} \in [0,1]^n\}$
- $\{(x,\theta) \in \mathbb{R}^2 : \theta \geq x^3\}$
- Set of prime numbers
- The dual cone $\mathcal{C}^* := \{\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{y} \leq 0 \ \forall \boldsymbol{x} \in \mathcal{C}\}$ for an arbitrary set $\mathcal{C}$.
- The polar set $\mathcal{C}^o := \{\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{y} \leq 1 \ \forall \boldsymbol{x} \in \mathcal{C}\}$ for an arbitrary set $\mathcal{C}$.
- Set of rank-one matrices $\{\boldsymbol{x}\boldsymbol{x}^\top : \boldsymbol{x} \in \mathbb{R}^n\}$

For more practice on this, Chapter 2 of Boyd and Vandenberghe (2004).

An Exercise in Convexity: Solutions

Classify the following sets as convex or non-convex

- $\{\boldsymbol{x} : \boldsymbol{x} \in [0,1]^n\}$: Convex
- $\{(x, \theta) \in \mathbb{R}^2 : \theta \geq x^3\}$
  - x^3 is quasiconvex but nonconvex, so this set is not convex.
- Set of prime numbers: Non-convex (3 and 5 are prime, 4 is not).
- The dual cone $\mathcal{C}^* := \{\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{y} \leq 0 \ \forall \boldsymbol{x} \in \mathcal{C}\}$ for an arbitrary set $\mathcal{C}$: convex (verify definition of convexity)
- The polar set $\mathcal{C}^o := \{\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{y} \leq 1 \ \forall \boldsymbol{x} \in \mathcal{C}\}$ for an arbitrary set $\mathcal{C}$: convex (verify definition of convexity)
- Set of rank-one matrices $\{\boldsymbol{x}\boldsymbol{x}^\top : \boldsymbol{x} \in \mathbb{R}^n\}$: Non-convex ($\boldsymbol{a}\boldsymbol{a}^\top$ and $\boldsymbol{b}\boldsymbol{b}^\top$ are rank one, but $\frac{1}{2}(\boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{b}\boldsymbol{b}^\top)$ may not be)

For more practice on this, Chapter 2 of Boyd and Vandenberghe (2004).

Convex Optimization: Why Do We Like Convex Functions?

$$\min_x f(x)$$
$$s.t. \quad g(x) \leq q$$

Convex optimization is relatively "easy" because of three key features:

Convex Optimization: Why Do We Like Convex Functions?

$$\min_{x} f(x)$$
$$s.t. \;\; g(x) \leq q$$

Convex optimization is relatively “easy” because of three key features:

- **Convex Feasible Set**: The feasible set $\{x \mid g(x) \leq q\}$ is a convex set, which has many good properties we like:

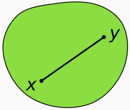
(a) A Convex Set **(b)** A Non-Convex Set

Convex Optimization: Why Do We Like Convex Functions?

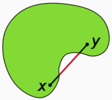
$$\min_x f(x)$$
$$s.t. \;\; g(x) \leq q$$

Convex optimization is relatively "easy" because of three key features:

- **Convex Feasible Set**: The feasible set $\{x \mid g(x) \leq q\}$ is a convex set, which has many good properties we like:



(a) A Convex Set



(b) A Non-Convex Set

- **Local-Global Correspondence**: A local optimum of $f(x)$ is guaranteed to be the global optimum (why?)

Convex Optimization: Why Do We Like Convex Functions?

$$\min_x f(x)$$
$$s.t. \;\; g(x) \leq q$$

Convex optimization is relatively "easy" because of three key features:

- **Convex Feasible Set**: The feasible set $\{x \mid g(x) \leq q\}$ is a convex set, which has many good properties we like:

(a) A Convex Set **(b)** A Non-Convex Set

- **Local-Global Correspondence**: A local optimum of $f(x)$ is guaranteed to be the global optimum (why?)
- **Strong Duality**: Convex optimization also satisfies strong duality (subject to a technical condition called Slater's condition)

Rockafeller, Revisited

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

What do we really think of this quote?

Rockafeller, Revisited

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

What do we really think of this quote?

- Any problem can be rewritten as a convex problem!

Rockafeller, Revisited

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

What do we really think of this quote?

- Any problem can be rewritten as a convex problem!
  - Rewrite in epigraph form
  - Replace feasible region with its convex hull

Rockafeller, Revisited

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

What do we really think of this quote?

- Any problem can be rewritten as a convex problem!
  - Rewrite in epigraph form
  - Replace feasible region with its convex hull
- Verifying the convexity of a function is NP-hard

Rockafeller, Revisited

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

What do we really think of this quote?

- Any problem can be rewritten as a convex problem!
  - Rewrite in epigraph form
  - Replace feasible region with its convex hull
- Verifying the convexity of a function is NP-hard
- Not all convex sets can be efficiently optimized over (copositive matrices, non-negative polynomials)

Rockafeller, Revisited

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

What do we really think of this quote?

- Any problem can be rewritten as a convex problem!
  - Rewrite in epigraph form
  - Replace feasible region with its convex hull
- Verifying the convexity of a function is NP-hard
- Not all convex sets can be efficiently optimized over (copositive matrices, non-negative polynomials)
- Some non-convex problems can be efficiently solved in practice (TSPs, computing the leading eigenvector of a matrix)

Rockafeller, Revisited

T. Rockafeller (1993)

"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

What do we really think of this quote?

- Any problem can be rewritten as a convex problem!
  - Rewrite in epigraph form
  - Replace feasible region with its convex hull
- Verifying the convexity of a function is NP-hard
- Not all convex sets can be efficiently optimized over (copositive matrices, non-negative polynomials)
- Some non-convex problems can be efficiently solved in practice (TSPs, computing the leading eigenvector of a matrix)
- Conclusion: Rockafeller is closer than attempt #1, but wrong

Attempt #3: What is Tractable?

This is a tricky question, especially because the "real" answer keeps changing as solvers and our algorithms improve. An attempt:

- A problem is theoretically tractable if it is solvable in polynomial time

Attempt #3: What is Tractable?

This is a tricky question, especially because the "real" answer keeps changing as solvers and our algorithms improve. An attempt:

- A problem is theoretically tractable if it is solvable in polynomial time
- A problem is practically tractable if it is solvable in a reasonable amount of time at instance sizes that we care about in practice

Attempt #3: What is Tractable?

This is a tricky question, especially because the "real" answer keeps changing as solvers and our algorithms improve. An attempt:

- A problem is theoretically tractable if it is solvable in polynomial time
- A problem is practically tractable if it is solvable in a reasonable amount of time at instance sizes that we care about in practice
- Generally speaking, polynomially solvable problems are tractable, integer problems are tractable, and polynomially solvable problems remain tractable if we introduce integer variables

Attempt #3: What is Tractable?

This is a tricky question, especially because the "real" answer keeps changing as solvers and our algorithms improve. An attempt:

- A problem is theoretically tractable if it is solvable in polynomial time
- A problem is practically tractable if it is solvable in a reasonable amount of time at instance sizes that we care about in practice
- Generally speaking, polynomially solvable problems are tractable, integer problems are tractable, and polynomially solvable problems remain tractable if we introduce integer variables
- NP-hard continuous problems are not (yet) practically tractable, but (in my opinion, others might disagree) will be in ten years time $\rightarrow$ Gurobi released a spatial branch-and-bound solver for them in 2019

Attempt #3: What is Tractable?

This is a tricky question, especially because the "real" answer keeps changing as solvers and our algorithms improve. An attempt:

- A problem is theoretically tractable if it is solvable in polynomial time
- A problem is practically tractable if it is solvable in a reasonable amount of time at instance sizes that we care about in practice
- Generally speaking, polynomially solvable problems are tractable, integer problems are tractable, and polynomially solvable problems remain tractable if we introduce integer variables
- NP-hard continuous problems are not (yet) practically tractable, but (in my opinion, others might disagree) will be in ten years time $\rightarrow$ Gurobi released a spatial branch-and-bound solver for them in 2019
- In the remaining part of this lecture, we'll go through classes of practically tractable problems that often show up in practice

Convex Conic Optimization

Linear Optimization

A generic linear optimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}, \boldsymbol{D}\boldsymbol{x} = \boldsymbol{d}.$$

Modeling power:

- Maximum of t linear functions: $t \ge c_i + \boldsymbol{d}_i^\top \boldsymbol{x} \ \forall i \in [t]$
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 \le t \iff \exists \boldsymbol{u} : -\boldsymbol{u} \le \boldsymbol{x} \le \boldsymbol{u}, \boldsymbol{e}^\top \boldsymbol{u} \le t$

Linear Optimization

A generic linear optimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{D}\boldsymbol{x} = \boldsymbol{d}.$$

Modeling power:

- Maximum of t linear functions: $t \geq c_i + \boldsymbol{d}_i^\top \boldsymbol{x} \ \forall i \in [t]$
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 \leq t \iff \exists \boldsymbol{u} : -\boldsymbol{u} \leq \boldsymbol{x} \leq \boldsymbol{u}, \boldsymbol{e}^\top \boldsymbol{u} \leq t$

Why is this useful?

- Can certify infeasibility of a linear system using Farkas's Lemma
- Can solve even massive LOs with modern solvers

Linear Optimization

A generic linear optimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{D}\boldsymbol{x} = \boldsymbol{d}.$$

Modeling power:

- Maximum of t linear functions: $t \geq c_i + \boldsymbol{d}_i^\top \boldsymbol{x} \ \forall i \in [t]$
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 \leq t \iff \exists \boldsymbol{u} : -\boldsymbol{u} \leq \boldsymbol{x} \leq \boldsymbol{u}, \boldsymbol{e}^\top \boldsymbol{u} \leq t$

Why is this useful?

- Can certify infeasibility of a linear system using Farkas's Lemma
- Can solve even massive LOs with modern solvers

How to solve?

- Mosek or Gurobi (simplex or interior point method)
- Exercise: What is the dual of this LO? (Do on board)

Conic Optimization

A generic conic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \in \mathcal{K}. \end{aligned}$$

Where $\mathcal{K}$ is a closed, convex pointed and solid cone

- Convex cone: $x, y \in \mathcal{K}$ implies $\lambda x + \mu y \in \mathcal{K}$ for all $\lambda, \mu \geq 0$
- Pointed: $\mathcal{K} \cap \{-\mathcal{K}\} = \{0\}$
- Solid: $\exists \boldsymbol{x} \in \mathcal{K}, \epsilon > 0 : \forall \boldsymbol{y}, \|\boldsymbol{x} - \boldsymbol{y}\| \leq \epsilon \implies \boldsymbol{y} \in \mathcal{K}$

Conic Optimization

A generic conic optimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \in \mathcal{K}.$$

Where $\mathcal{K}$ is a closed, convex pointed and solid cone

- Convex cone: $x, y \in \mathcal{K}$ implies $\lambda x + \mu y \in \mathcal{K}$ for all $\lambda, \mu \geq 0$
- Pointed: $\mathcal{K} \cap \{-\mathcal{K}\} = \{0\}$
- Solid: $\exists \boldsymbol{x} \in \mathcal{K}, \epsilon > 0 : \forall \boldsymbol{y}, \|\boldsymbol{x} - \boldsymbol{y}\| \leq \epsilon \implies \boldsymbol{y} \in \mathcal{K}$
- We usually want $\mathcal{K}$ to be an outer product of the non-negative orthant, second-order cone, semidefinite cone, exponential cone, and power cone, so that it can be solved using the Mosek solver

Conic Optimization

A generic conic optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \in \mathcal{K}.$$

Where $\mathcal{K}$ is a closed, convex pointed and solid cone

- Convex cone: $x, y \in \mathcal{K}$ implies $\lambda x + \mu y \in \mathcal{K}$ for all $\lambda, \mu \geq 0$
- Pointed: $\mathcal{K} \cap \{-\mathcal{K}\} = \{0\}$
- Solid: $\exists \boldsymbol{x} \in \mathcal{K}, \epsilon > 0 : \forall \boldsymbol{y}, \|\boldsymbol{x} - \boldsymbol{y}\| \leq \epsilon \implies \boldsymbol{y} \in \mathcal{K}$
- We usually want $\mathcal{K}$ to be an outer product of the non-negative orthant, second-order cone, semidefinite cone, exponential cone, and power cone, so that it can be solved using the Mosek solver
- Because some other convex cones are not tractable (copositive)

Second-Order Cone Optimization

A generic second-order cone problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{s.t.} \quad \|\boldsymbol{A}_i^\top \boldsymbol{x} + b_i\|_2 \leq c_i^\top \boldsymbol{x} + d_i, \ \forall i \in [m], \ \boldsymbol{D}\boldsymbol{x} = \boldsymbol{d}.$$

Second-Order Cone Optimization

A generic second-order cone problem:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \|\boldsymbol{A}_i^\top \boldsymbol{x} + b_i\|_2 \leq c_i^\top \boldsymbol{x} + d_i, \ \forall i \in [m], \ \boldsymbol{D}\boldsymbol{x} = \boldsymbol{d}. \end{aligned}$$

Modeling power:

- Linear inequalities
- Convex quadratics
- Portfolio risk and chance constraints

Second-Order Cone Optimization

A generic second-order cone problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{s.t.} \quad \|\boldsymbol{A}_i^\top \boldsymbol{x} + b_i\|_2 \leq c_i^\top \boldsymbol{x} + d_i, \ \forall i \in [m], \ \boldsymbol{D}\boldsymbol{x} = \boldsymbol{d}.$$

Modeling power:

- Linear inequalities
- Convex quadratics
- Portfolio risk and chance constraints

Why is this useful?

- Most general continuous problem we can solve to optimality at scale

Second-Order Cone Optimization

A generic second-order cone problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{s.t.} \quad \|\boldsymbol{A}_i^\top \boldsymbol{x} + b_i\|_2 \leq c_i^\top \boldsymbol{x} + d_i, \ \forall i \in [m], \ \boldsymbol{D}\boldsymbol{x} = \boldsymbol{d}.$$

Modeling power:

- Linear inequalities
- Convex quadratics
- Portfolio risk and chance constraints

Why is this useful?

- Most general continuous problem we can solve to optimality at scale

How to solve?

- Mosek or Gurobi (interior point method)

Modeling Power: Rotated Second-order Cone Constraints

A large class of problems can be cast as second-order cone problems since

$$\text{(a)}\quad x^2 \le yz, y, z \ge 0 \iff \left\| \begin{pmatrix} 2x \\ y - z \end{pmatrix} \right\|_2 \le y + z,$$

$$\text{(b)}\quad \boldsymbol{x}_i^\top \boldsymbol{P}_i \boldsymbol{x} + 2q_i^\top \boldsymbol{x} + r_i \le 0 \iff \left\| \boldsymbol{P}_i^{\frac{1}{2}} \boldsymbol{x} + \boldsymbol{P}_i^{\frac{-1}{2}} \boldsymbol{q}_i \right\|_2 \le \left(q_i^\top \boldsymbol{P}_i^{-1} q_i - r_i \right)^{\frac{1}{2}}$$

$$\text{(c)}\quad t \ge x^{\frac{3}{2}}, x \ge 0 \iff \exists s : 2st \ge x^2, \frac{1}{4}x \ge s^2$$

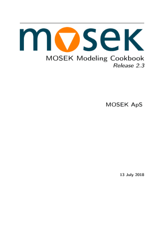
A large class of problems can be cast as second-order cone problems since

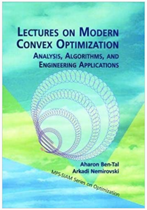
$$\text{(a)}\quad x^2 \le yz, y, z \ge 0 \iff \left\| \begin{pmatrix} 2x \\ y - z \end{pmatrix} \right\|_2 \le y + z,$$

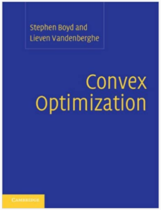
$$\text{(b)}\quad \boldsymbol{x}_i^\top \boldsymbol{P}_i \boldsymbol{x} + 2q_i^\top \boldsymbol{x} + r_i \le 0 \iff \left\| \boldsymbol{P}_i^{\frac{1}{2}} \boldsymbol{x} + \boldsymbol{P}_i^{\frac{-1}{2}} \boldsymbol{q}_i \right\|_2 \le \left(q_i^\top \boldsymbol{P}_i^{-1} q_i - r_i \right)^{\frac{1}{2}}$$

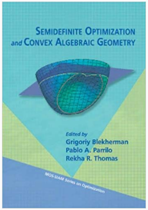
$$\text{(c)}\quad t \ge x^{\frac{3}{2}}, x \ge 0 \iff \exists s : 2st \ge x^2, \frac{1}{4}x \ge s^2$$

And many other problems! Good places to look are:









Semidefinite Cone Optimization

A generic semidefinite problem:

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathcal{S}_+^n} \quad & \langle \boldsymbol{C}, \boldsymbol{X} \rangle \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i, \ \forall i \in [m]. \end{aligned}$$

Semidefinite Cone Optimization

A generic semidefinite problem:

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathcal{S}_+^n} \quad & \langle \boldsymbol{C}, \boldsymbol{X} \rangle \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i, \ \forall i \in [m]. \end{aligned}$$

Modeling power:

- Linear matrix inequalities
- Eigenvalues and sums of eigenvalues

Semidefinite Cone Optimization

A generic semidefinite problem:

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathcal{S}^n_+} \quad & \langle \boldsymbol{C}, \boldsymbol{X} \rangle \\ \text{s.t.} \quad & \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i, \ \forall i \in [m]. \end{aligned}$$

Modeling power:

- Linear matrix inequalities
- Eigenvalues and sums of eigenvalues

Why is this useful?

- Most general problem can solve to optimality at moderate sizes
- Lots of non-convex problems admit quite tight semidefinite relaxations—useful for getting upper bounds

Semidefinite Cone Optimization

A generic semidefinite problem:

$$\min_{\boldsymbol{X} \in \mathcal{S}^n_+} \quad \langle \boldsymbol{C}, \boldsymbol{X} \rangle$$
$$\text{s.t.} \quad \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i, \ \forall i \in [m].$$

Modeling power:

- Linear matrix inequalities
- Eigenvalues and sums of eigenvalues

Why is this useful?

- Most general problem can solve to optimality at moderate sizes
- Lots of non-convex problems admit quite tight semidefinite relaxations—useful for getting upper bounds

How to solve?

- Mosek (interior point method).

Modeling Power: Semidefinite Constraints

A large class of problems can be cast as semidefinite problems since

(a) $\lambda_{\max}(\boldsymbol{X}) \leq t \iff \boldsymbol{X} \preceq t\mathbb{I}$

(b) $\displaystyle\sum_{i=1}^{k} \lambda_i(\boldsymbol{X}) \leq t \iff \exists \theta, \boldsymbol{U} : t \geq k\theta + \mathrm{tr}(\boldsymbol{U}), \theta\mathbb{I} + \boldsymbol{U} \preceq \boldsymbol{X}$

(c) $\|\boldsymbol{X}\|_* \leq t \iff \exists \boldsymbol{U}, \boldsymbol{V} : \begin{pmatrix} \boldsymbol{U} & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{V} \end{pmatrix} \succeq \mathbf{0}, 2t \geq \mathrm{tr}(\boldsymbol{U}) + \mathrm{tr}(\boldsymbol{V})$

Modeling Power: Semidefinite Constraints

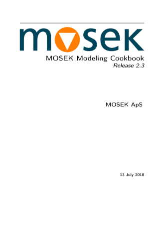
A large class of problems can be cast as semidefinite problems since

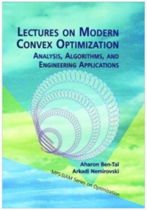
(a) $\lambda_{\max}(\boldsymbol{X}) \leq t \iff \boldsymbol{X} \preceq t\mathbb{I}$

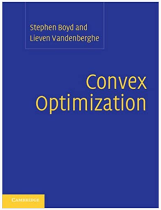
(b) $\sum_{i=1}^{k} \lambda_i(\boldsymbol{X}) \leq t \iff \exists \theta, \boldsymbol{U} : t \geq k\theta + \mathrm{tr}(\boldsymbol{U}), \theta\mathbb{I} + \boldsymbol{U} \preceq \boldsymbol{X}$

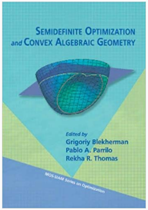
(c) $\|\boldsymbol{X}\|_* \leq t \iff \exists \boldsymbol{U}, \boldsymbol{V} : \begin{pmatrix} \boldsymbol{U} & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{V} \end{pmatrix} \succeq \mathbf{0}, 2t \geq \mathrm{tr}(\boldsymbol{U}) + \mathrm{tr}(\boldsymbol{V})$

And many other problems! Good places to look are:









Integer Optimization

Integer Optimization

Integer optimization generalizes linear optimization. For instance,

$$
\begin{aligned}
\min_{\boldsymbol{x} \in \mathbb{Z}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}.
\end{aligned}
$$

Integer Optimization

Integer optimization generalizes linear optimization. For instance,

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{Z}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

We can also have both continuous and discrete variables: mixed-integer optimization (MIO), mixed-integer conic optimization (MICO)

Integer Optimization

Integer optimization generalizes linear optimization. For instance,

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{Z}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

We can also have both continuous and discrete variables: mixed-integer optimization (MIO), mixed-integer conic optimization (MICO)

How to solve:

- Use Gurobi (branch-and-bound)
- Branch-and-cut with Gurobi and Mosek (if mixed-integer conic)
- Dantzig-Wolfe

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$.
How do we model the following?

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$. How do we model the following?

1. Exactly k of $x_1, x_2, \ldots x_n$ are equal to 1.

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$. How do we model the following?

1. Exactly k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n = k$

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$. How do we model the following?

1. Exactly k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n = k$
2. At most k of $x_1, x_2, \ldots x_n$ are equal to 1.

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$. How do we model the following?

1. Exactly k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n = k$
2. At most k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n \leq k$

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$. How do we model the following?

1. Exactly k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n = k$
2. At most k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n \leq k$
3. If $x_1 = 1$, then $y_1 = 1$.

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$. How do we model the following?

1. Exactly k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n = k$
2. At most k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n \leq k$
3. If $x_1 = 1$, then $y_1 = 1$.
   $x_1 \leq y_1$

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$. How do we model the following?

1. Exactly k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n = k$
2. At most k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n \leq k$
3. If $x_1 = 1$, then $y_1 = 1$.
   $x_1 \leq y_1$
4. If at least k of $x_1, x_2, \ldots x_n$ equals 1, then $y_1 = 1$.

Modeling: Logical Constraints

Assume $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ are binary decision variables $\{0, 1\}$. How do we model the following?

1. Exactly k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n = k$
2. At most k of $x_1, x_2, \ldots x_n$ are equal to 1.
   $x_1 + x_2 + x_3 + \ldots + x_n \leq k$
3. If $x_1 = 1$, then $y_1 = 1$.
   $x_1 \leq y_1$
4. If at least k of $x_1, x_2, \ldots x_n$ equals 1, then $y_1 = 1$.
   $x_1 + x_2 + x_3 + \ldots + x_n - (k - 1) \leq n * y_1$

Big-M Method for Modeling

Suppose we have a continuous variable a_i,

- How to model $\|\boldsymbol{a}\|_1 \leq 5$?

Big-M Method for Modeling

Suppose we have a continuous variable a_i,

- How to model $\|\boldsymbol{a}\|_1 \leq 5$?

$$\boldsymbol{e}^\top \boldsymbol{z} \leq 5,\ \boldsymbol{z} \geq \boldsymbol{a}, \boldsymbol{z} \geq -\boldsymbol{a}.$$

- How to model $\|\boldsymbol{a}\|_1 \geq 5$?

Big-M Method for Modeling

Suppose we have a continuous variable a_i,

- How to model $\|\boldsymbol{a}\|_1 \leq 5$?

$$\boldsymbol{e}^\top \boldsymbol{z} \leq 5,\ \boldsymbol{z} \geq \boldsymbol{a}, \boldsymbol{z} \geq -\boldsymbol{a}.$$

- How to model $\|\boldsymbol{a}\|_1 \geq 5$?

$$y_i \leq |a_i|,\ \sum_i y_i \geq 5.$$

Big-M Method for Modeling

Suppose we have a continuous variable a_i,

- How to model $\|\boldsymbol{a}\|_1 \leq 5$?

$$\boldsymbol{e}^\top \boldsymbol{z} \leq 5, \;\; \boldsymbol{z} \geq \boldsymbol{a}, \boldsymbol{z} \geq -\boldsymbol{a}.$$

- How to model $\|\boldsymbol{a}\|_1 \geq 5$?

$$y_i \leq |a_i|, \;\; \sum_i y_i \geq 5.$$

That is,

$$\sum_i y_i \geq 5, \;\; a_i \geq y_i - Mz_i, \;\; a_i \leq -y_i + M(1 - z_i), \;\; z_i \in \{0, 1\}.$$

We use M and z to model the "or" condition.

Big-M Method for Modeling

Suppose we have a continuous variable α_i such that if $z_i = 0$ then we must enforce $\alpha_i = 0$. What are two ways to model this?

Big-M Method for Modeling

Suppose we have a continuous variable α_i such that if $z_i = 0$ then we must enforce $\alpha_i = 0$. What are two ways to model this?

- We add the constraint:
  $-Mz_i \leq \alpha_i \leq Mz_i$, where M is a big number.

Big-M Method for Modeling

Suppose we have a continuous variable α_i such that if $z_i = 0$ then we must enforce $\alpha_i = 0$. What are two ways to model this?

- We add the constraint:
  $-Mz_i \leq \alpha_i \leq Mz_i$, where M is a big number.
  Question: how do you choose M practically?

Big-M Method for Modeling

Suppose we have a continuous variable α_i such that if $z_i = 0$ then we must enforce $\alpha_i = 0$. What are two ways to model this?

- We add the constraint:
  $-Mz_i \leq \alpha_i \leq Mz_i$, where M is a big number.
  Question: how do you choose M practically?
  We would like M to be as small as possible, provided it does not restrict the feasible region, i.e. consider $||\alpha||_\infty$.

Before You go... Readings

- Remind yourself of optimization, if it's not immediately familiar. Especially duality and convexity (chapters 2-5 in Boyd and Vandenberghe (2004))
- Remind yourself of probability theory, if it's unfamiliar. MIT OCW class 6.436J and the book by Grimmett and Stirzaker are good resources.
- Read this blog post by Ben Recht on different types of decision-making under uncertainty

> *...we turn to decision making where our current actions impact future decisions. These two weeks get a bit ridiculous because this topic could comprise a full graduate school curriculum. Should this lecture be about linear feedback systems,* **stochastic programming**, **robust optimization**, *model predictive control,* **dynamic programming**, *reinforcement learning, or combinatorial search? Uh, it sort of needs to be about all of them.*

– It seems this class is on a good track :-)

Let's wrap up here

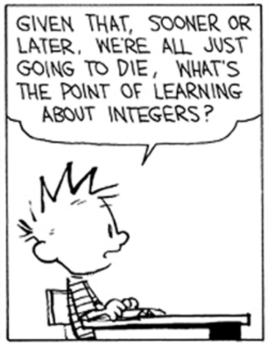


Figure 2: Calvin Explains Recent Advances in Integer Optimization

Thank you, and see you next week!