Convex and Distributed Optimization

Franck Iutzeler

Jérôme Malick

Thomas Ropars

from **LJK**, the applied maths and computer science lab

and **LIG**, the Grenoble Informatics Laboratory

To reach us: `firstname.name@univ-grenoble-alpes.fr`

To find this presentation: `http://ljk.imag.fr/membres/Jerome.Malick/CDO.pdf`

>>> Context: Big Data

- We have entered the **Big Data** area...
- Huge amounts of data are collected, routinely and continuously
  - Consumer and people data (phone calls and text, social media, email, surveillance cameras, web activity...)
  - Scientific data (biological, genomic, astronomical,...)
- Challenges in the whole chain of data processing
  from data collection to computation, analysis, interpretation

Illustration: images reconstruction in radio-astronomy

example maybe usual for you

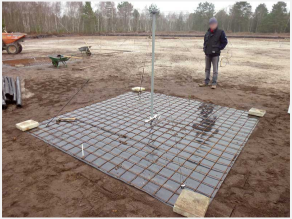
technology of the past

>>> Example of big data in science

Illustration: images reconstruction in radio-astronomy

example maybe usual for you

technology of the past



technology of the **future** !!

software-telescope

- large, flexible, and cheap networks
- huge data flow, huge numerical treatment
- with in particular: large-scale optimization problems

>>> Data Analysis

Goals of data analysis

- Extract meaning from data: understand statistical properties, learn important features and fundamental structures in the data.
- Use this knowledge to make decisions or predictions about other data.

Highly multidisciplinary area
with foundations in statistics and computer science (artificial intelligence, machine learning, databases, parallel systems...)

and **Optimization** here ?

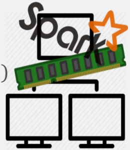
$$\begin{cases} \text{minimize} \quad f(x) & \text{(objective function)} \\ \quad x \in X \subset \mathbb{R}^n & \text{(constraints)} \end{cases}$$

- Optimization provides a toolkit of modeling and algorithmic techniques
- This branch of applied maths is being revolutionized by its interactions with data analysis (computational statistics and machine learning)
- Ongoing challenges because of increasing **scale and complexity** of data analysis applications

>>> Large-scale optimization for data analysis

Need for **scalable optimization algorithms**...

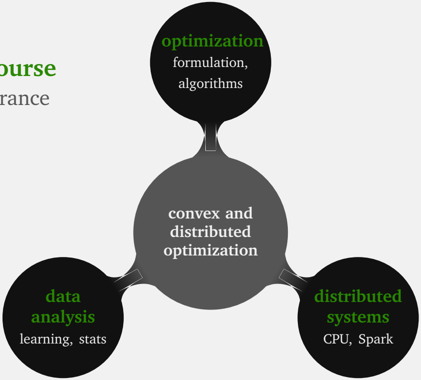
Leveraging on the **new distributed systems**



- Hardware improvements
  - Explosion of available computing resources (data centers, cloud)
  - Improvement of multicore infrastructure and networks
- Software improvements
  - developed by a wide scientific community and powerful industrial partners (Google, Facebook, Twitter)

Positioning of this course

original and unique in France



optimization

formulation, algorithms

convex and distributed optimization

data analysis

learning, stats

distributed systems

CPU, Spark

>>> This course is not...

topic positioning

- not a course on distributed systems
  but we manipulate the hottest technologies of this domain
- not a standard optimization course
  rather a data analysis-related optimization course
- not a course on stats or machine learning
  but we discuss standard learning problems

contents

- not a maths course – but requires some maths agility
- not an algorithmic course – but requires some programming skills

prerequisites

- basic programming skills in Python (check-out online tutorials if necessary)
- basic knowledge in matrix calculus (matrix operations, norms) and differential calculus (definition and manipulation of gradients...)
- basic ideas on optimization (e.g. definition of convex functions, convex sets...)
  check-out the Refresher course on matrix analysis and optimization

>>> This course is...

Extended subtitle could be:

algorithmic aspects of optimization
for data analysis applications

Three objectives of the course:

- present optimization algorithms that scale up to high dimensions:
  stochastic, incremental, coordinate, random, and distributed algorithms
- implement them efficiently on data problems
  with high-level tools currently used in big data companies
- provide a complementary viewpoint on data analysis from an optimization perspective

>>> Tutorials on standard machine learning problems

Core of this course:

3 tutorials on machines:

- Tutorial 1: parsing and manipulating data
- Tutorial 2: sparse logistic regression
- Tutorial 3: matrix factorization for recommender systems

++ increasing programming difficulty and mathematical technicality...

Objectives of the tutorials:

- understand the basics of optimization algorithms in large-scale settings
- review learning applications and interpret numerical results
- programming: play with the hottest big data technologies

>>> Programming environnement

we work on Jupyter notebooks with Python,
Spark for computation, and Docker for installation

Spark (v2.0.1, october 2016)

- open-source distributed computing framework
- high-level paradigm (higher than MPI, OpenMP...) that automatically adapts to underlying hardware infrastructure
- is becoming the main big data technology (with thousand of developers)
- adopted by Twitter, Facebook, Google, Amazon...

Docker

- open-source project that automates
  the deployment of applications inside software "containers"
- container ≃ small virtual machines = provides an environment with a full OS and all softwares and libraries needed
- our docker contains a linux system + python, pyspark, jupyter...
- nothing else to install and everyone has the same soft environment

>>> Schedule

	Monday (3h)	Tuesday (3h)
Week 1	**today:** Presentation of the course Quick recalls on Optimization	
Week 2	Course on optimization 1 Incremental algorithms	Tutorial stochastic gradient
Week 3	Overview of distributed computing Introduction to Spark	Tutorial data preprocessing
Week 4	Tutorial application to classification	Course on optimization 2 Distributed algorithms
Week 5	Tutorial recommendation systems	Final Tutorial computation on cluster

Note: course Amphi D tutorial : E301 +E202

Note: sessions in January about article study

>>> Evaluation (master students only !)

Report on tutorials (by group of < 3)

- report on the accomplished work on tutorials **2 & 3**
  (with tables, plots, comments... but no code !)
- with highlights on chosen aspects

  Examples: learning (interpretation of results, other models...), maths (proof of related results, theoretical analysis of special cases,...) or numerical extensions
- in a very open format – before christmas break

Presentation of a research article (by group of ≤ 3)

- list of various articles
  (theoretical, algorithmical, computations, or applications-oriented)
- oral presentation of ~ 8 mins
- again in a very open format – beginning of Januray

Find our own way to valorize your work !

Final note is a convex combination : 2/3 report + 1/3 article

>>> Get ready for the tutorials

Before the first tutorial: get ready !!

We recommend you to work on your own machine (...)

but it is at your own risk... we can still provide little support for linux users...

- install Docker CE (Community Edition)
  for Ubuntu:
  `https://docs.docker.com/engine/installation/linux/docker-ce/ubuntu/`
- run the command `docker run hello-world` to check your install
- In a second time: take the docker image that contains all necessary material at the following link   to be given

Other useful Links:

- Python/Numpy's documentation
  `http://docs.scipy.org/doc/numpy-1.11.0/reference/`
- Spark documentation
  `http://spark.apache.org/docs/latest/`