Machine learning and optimization in high dimensions

Motivation and introduction

Classical parametric statistics deal with a finite number of real parameters (d for dimension) estimated based on a sample, which size (denoted by n) tends to infinity. In this setting there exists sharp analyzes of asymptotic properties of common estimators. However they do not allow to take into account the following factors:

- Increasing number of descriptors: new technologies, digitization of personal data, sensor resolution.
- In many studies, the number of descriptors is bigger than the number of observations.
- In some applications, data is observed as a continuous flow rather than a fixed size sample.

These motivate to reconsider the classical setting taking into account the following ideas:

- If both d and n increase, powerful estimators are not necessarily the same as in the classical setting.
- Restrictions on the choice of the estimator: in this regime computation time becomes an issue.
- Constraints on n: distributed storage, flow (n is infinite), or simply n very big.

The goal of this course is to provide an overview of available approaches for these questions at the interplay between statistics and optimization. As a canonical example, we will present a complete description the LASSO estimator from statistical properties to numerical resolution for large scale problems.

The course requires comfortable knowledge of basic concepts in probability, statistics, linear algebra and analysis. Prior knowledge in convex analysis or optimization is not mandatory.

Content

Linear regression in high dimension. The linear model for regression will be used as a canonical example to illustrate most concepts and algorithms presented in the course.

- Introduction to sparsity, ℓ_0 penalization, concept of oracle bound.
- Sparse regression, algorithmic complexity (P versus NP), convexity and algorithmic consequences.
- Lasso estimator, first mention of oracle bounds and consistency results.
- Introduction to compressed sensing, RIP condition and decoding by linear programming.

Elements of convex analysis and optimization. Optimization is one of the crucial building block of data analysis and modeling applications for model parameter tuning based on a sample of observations.

- Conic programming hierarchy, Newton and interior point methods, motivation for simpler algorithms.
- Lagrangian duality and KKT conditions.
- Convex analysis, Fenchel conjugate, proximity operator, Legendre-Fenchel duality, optimality conditions.
- First order methods: subgradient descent, proximal splitting, complexity, acceleration.
- Application to oracle bounds and numerical resolution for the LASSO.

Stochastic approximation. Stochastic approximation algorithms are widespread to circumvent computational burden associated to large sample size n.

- Motivation for finite sums and population risk minimization.
- Convergence, convergence rate analysis, optimality notions for these algorithms, averaging.
- Relation and consequences for statistical estimation.

Random coordinate descent. Block coordinate approaches allow to work with large number of descriptors d. They also have an interesting duality interpretation with stochastic methods for finite sums.

- Presentation for the LASSO and convergence speed.
- Fenchel duality and primal dual interpretation, link with stochastic gradient. Detailed presentation for the support vector machine (SVM) for classification (SDCA method).

Dimension reduction and non convex optimization. Most optimization algorithms have natural extensions in nonconvex settings for which dimension reduction techniques provide a large number of applications.

- Matrix factorization, PCA, regularized PCA, dictionary learning.
- Dedicated optimization methods.