MA 598 - Topics on optimization algorithms, Spring 2023

- Instructor: Xiangxiong Zhang
- Webpage:

  https://www.math.purdue.edu/∼zhan1966/teaching/598/598_2023S.html
- Selected topics from the following reference books:
  - Beck, Introduction to Nonlinear Optimization
  - Nocedal and Wright, Numerical Optimization
  - Nesterov, Introductory Lectures on Convex Optimization A Basic Course
  - Ryu and Yin, Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators

Plan for this semester

There are many different types of optimization problems, but we mainly focus on **the convergence** of algorithms minimizing a convex function $f(x)$ with a large scale:

- Part I: some classical algorithms for minimizing a smooth function $f(x)$ such as gradient descent, accelerated gradient descent, Newton's method, quasi Newton methods, etc.
- Part II: algorithms for composite optimization of minimizing $f(x) + g(x)$ where $f(x)$ and $g(x)$ are both convex, but at least one of them is not differentiable, e.g.,

  $$\min \|x\|_1 + \|Ax - b\|_2^2$$

  where $\|x\|_1 = \sum_i |x_i|$.
- Part III: stochastic type algorithms, such as stochastic gradient descent.

Examples

- Part I: for $\min_x f(x)$, the gradient descent method is

$$x_{k+1} = x_k - \eta_k \nabla f(x_k)$$

  When and why does gradient descent converge? How fast does it converge?

  Prerequisites for Part I:

  - Calculus: gradient, Hessian, Taylor Theorem...
  - Linear algebra: eigenvalues, singular values and etc.

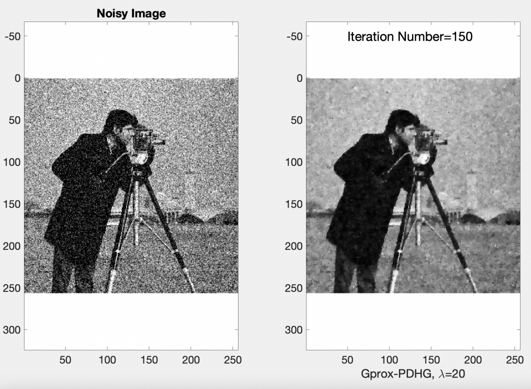
- Part II: we will introduce subderivatives, proximal operator, and algorithms using the subderivatives. We will use monotonicity of operators to prove convergence.

$$\min \|x\|_1 + \|Ax - b\|_2^2$$

- Part II: here is another example of nonsmooth convex optimization for denoising a given noisy image A via TV (total variation) norm minimization

$$\min_x \|x\|_{TV} + \lambda \|x - A\|_2^2,$$

where $\|x\|_{TV} = \sum_{i,j} \sqrt{|x_{i,j} - x_{i+1,j}|^2 + |x_{i,j} - x_{i,j+1}|^2}$



The algorithm PDHG will be covered in part II, and the paper on Gprox-PDHG (knowledge on Hilbert spaces are needed) would be a good choice for the final presentation.
Large scale means: if the dimension of x is n then only $\mathcal{O}(n)$ storage is acceptable. What is n^2?

Examples

- Part III: for minimizing $f(x) := \sum_{i=1}^{N} f_i(x)$, the full gradient is $\nabla f(x) = \sum_{i=1}^{N} \nabla f_i(x)$, we can use the stochastic gradient like

$$\nabla_S f(x) := \sum_{i \in S} \nabla f_i(x)$$

  where S is a random small subset of $\{1, 2, \cdots, N\}$. The stochastic gradient descent can be defined as:

$$x_{k+1} = x_k - \eta_k \nabla_{S_k} f(x_k).$$

  In order to analyze the convergence, we need some probability knowledge, which will be introduced.

An example where N is too large: recommendation systems for customers rating products (movies, merchandise, etc).

Focuses and learning outcomes of this course

- We focus on analysis of classical algorithms, i.e., why and how fast they converge. Applications will be barely mentioned, though questions about applications are always welcome.
- A final presentation/report (depending on our schedule) is required by reading a paper and/or implementing some classical/novel algorithms. Examples of possible choices of papers:
  - Convergence of nonlinear conjugate gradient method.
  - Convergence analysis of Adam.
  - Stochastic gradient Langevin dynamics.
- Learning outcome: by the end of the semester, I expect you to ??