COMM 616: Modern Optimization with Applications in ML and OR 2024-25 Fall

Lecture 1: Overview

Instructor: Jiajin Li September 3rd, 2024

1 Basic Notions in Optimization

In this class, we will consider a class of mathematical optimization problems that express the form:

$$\inf_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}). \tag{P}$$

- Objective function $F:\mathbb{R}^d\to\overline{\mathbb{R}}$: Extended value functions. We can always reformulate the constrained problem as an unconstrained one by using an extended value function as the objective. That is,

$$\inf_{\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^d} f(\boldsymbol{x}) \quad\Leftrightarrow\quad \inf_{\boldsymbol{x}\in\mathbb{R}^d} \underbrace{f(\boldsymbol{x})+\mathbb{1}_{\mathcal{X}}(\boldsymbol{x})}_{F(\boldsymbol{x})},$$

  where the indicator function $\mathbb{1}_{\mathcal{X}}$ of a set $\mathcal{X}\subseteq\mathbb{R}^d$ is defined through $\mathbb{1}_{\mathcal{X}}(\boldsymbol{x})=1$ if $\boldsymbol{x}\in\mathcal{X}$; $=0$ otherwise.

- Feasible set: $\left\{\boldsymbol{x}\in\mathbb{R}^d : F(\boldsymbol{x})<+\infty\right\}$
- Decision variable: $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)$

As the above formulation suggests, we are interested in a **global minimizer** of (P) and the **optimal value** of (P).

Definition 1 (Global Minimizer)**.** *A point $\boldsymbol{x}^\star\in\mathcal{X}$ such that $f(\boldsymbol{x}^\star)\le f(\boldsymbol{x})$ for all $\boldsymbol{x}\in\mathbb{R}^d$.*

Fact 2. *Such an $\boldsymbol{x}^\star$ may not exist, see Example 3.*

Example 3.

$$\inf_{x\in\mathbb{R}} F(x) := \frac{1}{x}+\mathbb{1}_{\mathbb{R}_{++}}(x).$$

Definition 4 (Optimal Value)**.** *The optimal value of (P) is defined to be the greatest lower bound or infimum of the set $\{F(\boldsymbol{x}):\boldsymbol{x}\in\mathbb{R}^d\}$ and is denoted by*

$$v^\star=\inf\{F(\boldsymbol{x}):\boldsymbol{x}\in\mathbb{R}^d\}.$$

If $\boldsymbol{x}^\star$ is a global minimizer of (P), then $v^\star=F(\boldsymbol{x}^\star)$.

Let us revisit Example (3). We have $v^\star=0$, but no corresponding $\boldsymbol{x}^\star$ exists.
Most machine learning tasks (e.g., training deep neural networks) make it impossible to find a global minimizer in finite time. Instead, we introduce a relaxed notion: the *local minimizer*.

Definition 5 (Local Minimizer)**.** *A point $\boldsymbol{x}'\in\mathbb{R}^d$ such that for some $\epsilon>0$, we have $F(\boldsymbol{x}')\le F(\boldsymbol{x})$ for all $\boldsymbol{x}\in\mathcal{B}(\boldsymbol{x}',\epsilon)$. Here,*

$$\mathcal{B}(\boldsymbol{x}',\epsilon)=\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}-\boldsymbol{x}'\|_2\le\epsilon\}$$

is the Euclidean ball of radius $\epsilon>0$ centered at x'.

Remark 6. *Later in this class, we will see that it is necessary to further relax the concept of a local minimizer to a "critical point" or "stationary point".*

2 Big Picture of this Course

(i) Characterize the optimal solution or critical point of (P).

(ii) Study how the **structures** of F affect the our ability to solve (P).

(a) How exploit the structures to design **iterative methods** to address (P)?

i. Pick up an initial point $\boldsymbol{x}_0 \in \mathbb{R}^d$.

ii. Design the iterative scheme as

$$\boldsymbol{x}_{k+1} = \mathcal{H}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k), \quad \forall k \in \mathbb{N}.$$

E.g., the simplest method is the Gradient Descent (GD):

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla F(\boldsymbol{x}_k),$$

where $\eta > 0$ is the step size. The key structure that allows us to apply GD is the differentiability of F.

(b) How these structures impact the **convergence analysis** of the proposed algorithm?

Question: How many iterations we have to run to find an ϵ-global minimizer / ϵ-ciritical point of (P)?

Let us revisit Gradient Descent (GD) to examine how the structural properties impact the convergence analysis. To obtain quantitative results, it is clear that the differentiability of F is not sufficient to answer the **Question**. Then, we will discuss two cases:

i. To find an ϵ-global minimizer, we require both the convexity of F and the Lipschitz continuity of its gradient to achieve an iteration complexity of $\mathcal{O}(\epsilon^{-1})$ [2, Chapter 2].

ii. If we remove the convexity assumption of F, we can only achieve an iteration complexity of $\mathcal{O}(\epsilon^{-2})$ to obtain an ϵ-critical point, where the gradient norm of F is less than ϵ [1, Chapter 10].

2.1 Topics Covered in This Course

For more details, please refer to the **Information Sheet**.

Part I Optimization Theory — Introduce fundamental optimization tools and concepts to characterize the optimal solution and critical points of (P), as well as the structure of (P).

- Convex analysis and duality theory
- Strong convexity; Error bound condition; Kurdyka-Lojasiewicz (KL) property; Weak convexity
- Smoothness
- Nonsmoothness and Subdifferential (generalized gradient)

Remark 7. *In general, duality theory can be used either to characterize the optimal solution of* (P) *or to reformulate the problem in different ways. Additionally, convexity and smoothness conditions are key factors guiding algorithm design and convergence analysis. Specifically, convexity conditions characterize the lower bound of F, while smoothness conditions characterize the upper bound of F.*

Part II First-order Methods for Structured Optimization Problems — Discuss various algorithms under different structures and how these structures impact the convergence analysis. More importantly, can we establish a unified approach to proving their results?

	Smooth	Structured Nonsmooth	Nonsmooth
strongly convex	GD	PGD	Subgradient, Proximal Point Algorithms (PPA)
convex	GD	PGD	Subgradient, PPA
error bound conditions	GD	PGD	Subgradient, PPA
KŁ conditions	GD	PGD	Subgradient, PPA
weakly convex	—	—	(Stochastic) subgradient, PPA

Table 1: Summary of different function classes: Structured Smooth: Minimizing a smooth function combined with a convex nonsmooth function; PGD: Proximal (Projected) Gradient Descent; PPA: Proximal Point Algorithms. Additionally, note that a smooth function is inherently a weakly convex function. It is important to recognize that these two types of conditions are not entirely orthogonal.

We will also consider two other types of structured nonsmooth problems in this course: (i) polyhedral constrained minimization problems and (ii) minimax optimization problems. Both can be regarded as structured nonsmooth problems with specific structures that we can further exploit.

Part III: Applications in Machine Learning and Operation Research — We will utilize the optimization techniques from Parts I and II to address real-world problems. Some tentative topics we plan to cover in this class include:

(i) Optimal transport and Gromov-Wasserstein problems in machine learning and data science

(ii) Tractable reformulations and algorithm design for distributionally robust optimization

(iii) High-dimensional statistics problems, such as LASSO, ...

(iv) ...

References

[1] Amir Beck. *First-Order Methods in Optimization*. SIAM, 2017. 2

[2] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013. 2