1 Overview

In this course we will cover convex and combinatorial optimization and emphasize the deep connections between these two areas. Convex optimization is a central tool for solving large-scale problems which in recent years has had a profound impact on statistical machine learning, data analysis, mathematical finance, signal processing, control, approximation algorithms, as well as many other areas. In a single sentence, the premise of this course can be summarized as follows:

Convex optimization is an elegant mathematical theory which provides fundamental tools for reasoning about and solving problems across a broad range of areas in the data sciences.

The first part of the course will be dedicated the theory of convex optimization and its direct applications. The second part will focus on advanced techniques in combinatorial optimization using machinery developed in the first part. Let’s start.

2 An Example from Statistical Machine Learning

Let’s assume that we want to predict a person’s shoe size from data on people’s height and shoe sizes. We’ll collect a set of points $(\alpha_1, \beta_1), \ldots, (\alpha_N, \beta_N)$, and build a statistical model to make a prediction. One popular model for prediction is linear regression: we assume that the shoe size is simply a linear function of the height with additive coefficient being zero, $\beta = \alpha x$ and fit the parameter $x$ which minimizes the residual sum of squares. Our objective is then to find $x^* = \text{argmin} \sum_{i=1}^N (\beta_i - \alpha_i x)^2$. In this case, to find $x$ we can simply take the derivates of $(\beta_i - \alpha_i x)^2$:

$$\frac{\partial (\beta_i - x\alpha_i)^2}{\partial x} = 2(\beta_i - x\alpha_i)(-\alpha_i) = 2x\alpha_i^2 - 2\beta_i\alpha_i$$

The second derivative here is $2\alpha_i^2$ which is positive, thus the critical point is a local minimum, and therefore $x^*$ is the solution $x$ for which: $\sum_{i=1}^N (2x\alpha_i^2 - 2\beta_i\alpha_i) = 0$. Rearranging, we get that $x^* = \frac{\sum_{i=1}^N \alpha_i\beta_i}{\sum_{i=1}^N \alpha_i^2}$.

The above example is a very simple case of linear regression which is one of the most well-studied models for prediction and classification. More generally, given a $d$-dimensional point of data $\alpha \in \mathbb{R}^d$, the model assumes that the output $\beta \in \mathbb{R}$ is a linear function of $\alpha$:

$$\beta = x_0 + \sum_{i=1}^d \alpha_j x_j$$
For convenience, it is common to include $x_0$ in the vector of coefficients (by considering $\alpha$ as a $d+1$ dimensional point with the first entry set to 1), and write the linear model as the inner product: $\beta = \alpha^T x$. Given a set of observations $(\alpha_1, \beta_1), \ldots, (\alpha_N, \beta_N)$ the goal is fit a parameter $x$ which minimizes the residual sum of squares:

$$RSS(x) = \sum_{i=1}^{N} (\beta_i - \alpha^T x)^2$$  \hspace{1cm} (3)

In a similar manner to the way we found the optimal solution to the example above, the solution has a closed form solution here too. In practice, more sophisticated methods are used such as the LASSO where we seek a solution under constraints.

$$\min \sum_{i=1}^{N} (\beta_i - \alpha^T x)^2$$  \hspace{1cm} (4)

$$\text{s.t. } \sum_{j=1}^{d} |x_j| \leq t$$  \hspace{1cm} (5)

For the LASSO method the optimal solution can no longer be expressed in closed form. In order to apply such techniques we must develop algorithms that fit the parameters appropriately. More importantly, if we want to design methods such as LASSO, we must understand what kind of statistical models one can fit parameters for.

### 2.1 Optimization Problems

The example of fitting parameters is a special case of an optimization problem: minimizing (or maximizing) a function $f$ under some constraints. We usually write optimization problems in the following format:

$$\min f(x)$$  \hspace{1cm} (6)

$$\text{s.t. } g_i(x) \leq b_i \quad \forall i \in \{1, \ldots, m\}$$  \hspace{1cm} (7)

Where $g_i(x)$ encode the constraints. In the LASSO case, there was only one constraint $g(x) = \sum_{j=1}^{d} x_j \leq t$. The set of points which respect the constraints is the feasible set of the optimization problem.

### 3 Convex Optimization

If we take a look at the RSS function we aim to minimize in regression models, we will see it follows a particular structure. For convenience, let’s plot such a function with a single data point, and for simplicity set $\beta = 1/2, \alpha = 1$. The plot will look something like to figure below.

As we can see, the RSS function has a convex shape – any two points on the curve sit below the line connecting them. Appropriately, functions that have such structure are called convex functions.
Definition 1 (Convex set). A set $S$ is called a **convex set** if any two points in $S$ contain their line, i.e. for any $x, y \in S$ we have that $\lambda x + (1 - \lambda)y \in S$ for any $\lambda \in [0, 1]$.

Definition 2 (Convex function). For a convex set $S \subseteq \mathbb{R}^n$, we say that a function $f : S \rightarrow \mathbb{R}$ is **convex on** $S$ if for any two points $x, y \in S$ and any $\lambda \in [0, 1]$ we have that:

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

3.1 Why Convex Optimization?

Linear regression is just one example of the applications of convex functions. Throughout the course we will see how convex functions provide reasonable models for many real-world objectives and have numerous applications in data science (note that linear functions are a special case of convex functions). Importantly, as we will learn in this course, convex functions can be minimized under various constraints in a computationally efficient manner. Why? For convex functions, local optima are also global optima, and this property is extremely helpful for finding optimal solutions. Let’s formalize this argument.

Definition 3 (Global optimum). Let $S$ be a set of points, $f : S \rightarrow \mathbb{R}$. For a minimization problem, a point $x^* \in S$ is called **(global) optimum** if for any point $x \in S$ we have that $f(x^*) \leq f(x)$.

Definition 4 (Local optimum). Let $S$ be a set of points, $f : S \rightarrow \mathbb{R}$. For a minimization problem, a point $x^* \in S$ is called **local optimum** if $\exists \epsilon$ s.t. for any point $x \in S$ for which $\|x^* - x\|_2 \leq \epsilon$ we have that $f(x^*) \leq f(x)$.

Theorem 5. Let $S$ be a convex set and $f : S \rightarrow \mathbb{R}$ a convex function. If $x^*$ be a local optimum of $\min\{f(x) : x \in S\}$, then $x^*$ is a global optimum.

Proof. Let $x^*$ be a local optimum and assume for purpose of contradiction that there exists another point $y \in S$ s.t. $f(y) \leq f(x^*)$. Consider any strict convex combination $z = \lambda x^* + (1 - \lambda)y$, $\lambda \in (0, 1)$. From convexity: $f(z) \leq \lambda f(x^*) + (1 - \lambda)f(y) < \lambda f(x^*) + (1 - \lambda)f(x^*) = f(x^*)$. Since $z$ can be any convex combination, this contradicts $x^*$ being a local optimum. \qed
3.2 Concave Optimization

Note that if we take the mirror image of the plot above, we have a concave function: any two points on the curve lie above the line. That is, \( f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) \). So, \( f : S \to \mathbb{R} \) is convex function if and only if then \(-f(\cdot)\) is a concave function. Thus, if one knows how to minimize a convex function, one can maximize a concave function.

4 Combinatorial Optimization

Consider the following problem. We are given sets \( T_1, \ldots, T_n \) of elements and are asked to find a family of \( k \) sets whose union has the largest cardinality. That is:

\[
\begin{align*}
\max & \ |\cup_{i \in S} T_i| \\
\text{s.t.} & \ |S| \leq k
\end{align*}
\]

If the sets are all disjoint the problem is easy: pick the \( k \) sets with the highest cardinality. When the sets overlap, solving the problem exactly is well known to be NP-hard, which for the time being we'll interpret as computationally intractable. Interestingly, although we do not know how to find the optimal solution efficiently, we know how to develop algorithms that give us solutions that are provably “close to optimal”.

4.1 Submodular functions

As we will discuss in the last part of this course, the above problem has a special combinatorial structure. Notice that the objective of maximizing coverage has a diminishing returns property: the more sets we include into our solution, the marginal contribution of adding a new set decreases. This property is called submodularity and it is a key property which we will exploit. Although submodular functions were first studied in the 70s, due to new developments, there has been an exploding interest in algorithms, mechanism design, machine learning, data mining, and many other application domains.

As we will learn in this course, there is a deep connection between convex and submodular optimization. At this point, we will leave the connection implicit by pointing to the following intuitive observation: note that concave functions exhibit the diminishing returns property which defines submodular functions. In some sense, one can view concave functions as the continuous analogue of submodular functions.

5 Roadmap

The agenda for the course is the following. We will begin with linear optimization (which is a special case of convex optimization), and cover fundamental concepts like duality, and describe algorithms for solving linear optimization problems (roughly 3 to 4 weeks). We will then generalize
the concepts and develop algorithms for convex optimization problems (roughly 4 weeks). In the last part of the course, we will transition into combinatorial optimization, and develop a theory for combinatorial optimization which is based on the machinery we developed for convex optimization.