1 Outline

In this lecture, we study

- complexity of nonconvex optimization,
- sparse regression,
- low-rank matrix completion.

2 Introduction to Nonconvex Optimization

Figure 9.1 shows a nonconvex function with two variables. As we can see, the function has one

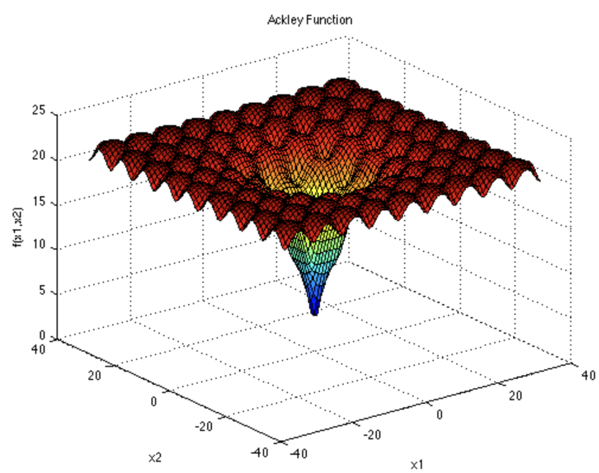

Figure 9.1: Ackley Function with Two Variables

global minimum but it has multiple local minima. Recall that for a convex function f, if $\nabla f(x) = 0$, then the corresponding x is a minimizer of the function f. Gradient-based methods for convex optimization basically seek for a solution x with $\nabla f(x) = 0$. However, for a nonconvex function f, a point x with $\nabla f(x) = 0$ does not guarantee global optimality as it can be a locally optimal solution.

In fact, finding the global minimum of a nonconvex function is a difficult task in general. To be more precise, the following theorem shows that there exists a smooth nonconvex function, to find the global minimum of which we need an exponential number of function evaluations.

Theorem 9.1. *For any $\beta > 0$, there exists a β-smooth function $f : [0,1]^d \to \mathbb{R}$ on $[0,1]^d$ such that any algorithm requires at least $(\beta/\epsilon)^{\Omega(d)}$ function queries to find an ϵ-optimal solution x with*

$$f(x) \leq \min_{x \in \mathbb{R}^d} f(x) + \epsilon.$$

Proof. We partition $[0,1]$ into k invervals of equal length, which gives rise to k^d boxes partitioning $[0,1]^d$. We can construct a function f such that a box contains a point x^* with $f(x^*) = -\epsilon$ but f has value 0 in the other boxes. This means that checking a box not containing x^* does not provide any information about the location of x^*. This means that to find the box containing x^*, we need at least $\Omega(k^d)$ function evalutions. To make function f smooth with parameter β, we can make f behaves like

$$f(x) \simeq f(x^*) + \frac{\beta}{2}\|x - x^*\|_2^2.$$

At the same time, to impose the condition that $f(x) = 0$ if x is not contained in the box with x^*, we can set $k = O(\sqrt{\beta/\epsilon})$. Therefore, we need at least $O\left((\sqrt{\beta/\epsilon})^d\right)$ function evaluations. $\square$

There exist several important applications of nonconvex optimization. In the remainder of this section, we provide an overview of some well-known nonconvex optimization problems.

2.1 Sparse Regression

Let us consider the following optimization problem.

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_0 \tag{9.1}$$

where

$$\|x\|_0 = |\{i \in [d] : \ x_i \neq 0\}|$$

counts the number of nonzero coordinates of x. Here, $\|x\|_0$ is called the ℓ_0-norm. Although the name of $\|x\|_0$ contains the term "norm", $\|x\|_0$ is a nonconvex function and thus it is not a norm. The optimization problem is referred to as **sparse regression** because the $\lambda\|x\|_0$ term encourages to use less variables of x.

Due to nonconvexity of $\|x\|_0$, it is often difficult to solve (9.1) efficiently. Motivated by this, we can approximate and replace the ℓ_0-norm by the ℓ_1-norm, which gives rise to LASSO:

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1. \tag{9.2}$$

We saw that the proximal gradient method solves (9.2) and it runs with

$$x_{t+1} = \mathrm{prox}_{\eta\lambda\|\cdot\|_1}(x_t - \eta\nabla f(x_t))$$

where

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 \quad \text{and} \quad \left(\mathrm{prox}_{\eta\lambda\|\cdot\|_1}(x)\right)_{i\in[d]} = \begin{cases} x - \eta\lambda, & \text{if } x \geq \eta\lambda, \\ 0, & \text{if } -\eta\lambda \leq x < \eta\lambda, \\ x + \eta\lambda, & \text{if } x < -\eta\lambda. \end{cases}$$

Recall that the prox operator $\mathrm{prox}_{\eta\lambda\|\cdot\|_1}(\cdot)$ is called the shrinkage operator or the **soft-thresholding** operator. In fact, we can apply proximal gradient descent to solve (9.1). Note that

$$\mathrm{prox}_{\eta\lambda\|\cdot\|_0}(x) = \operatorname*{argmin}_{u\in\mathbb{R}^d} \left\{\|u\|_0 + \frac{1}{2\eta\lambda}\|x - u\|_2^2\right\}.$$

By definition,

$$\left(\mathrm{prox}_{\eta\lambda\|\cdot\|_0}(x)\right)_{i\in[d]} = \begin{cases} x_i, & \text{if } x_i^2 \geq 2\eta\lambda, \\ 0, & \text{otherwise.} \end{cases}$$

Here, the operator $\text{prox}_{\eta\lambda\|\cdot\|_0}(\cdot)$ is referred to as the **hard-thresholding** operator. With the hard-thresholding operator, one may run proximal gradient descent with update rule

$$x_{t+1} = \text{prox}_{\eta\lambda\|\cdot\|_0}(x_t - \eta\nabla f(x_t)).$$

Unlike LASSO, however, proximal gradient descent applied to sparse regression does not necessarily converge to an optimal solution. This is again due to nonconvexity of the ℓ_0-norm.

Recent works [HM20, HMS22] consider mixed-integer programming formulations of (9.1) while earlier works including LASSO focused on convex approximations of (9.1).

2.2 Low-Rank Matrix Completion

Let us consider

$$\min_{X\in\mathbb{R}^{n\times p}} \quad \|D - X\|_F \quad \text{subject to} \quad \text{rank}(X) = k$$

where

- D is an $n \times p$ matrix,
- $\|A\|_F$ denotes the Frobenius norm, i.e., $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^p a_{ij}^2}$.

By the definition of the Frobenius norm, the problem is equivalent to

$$\min_{X\in\mathbb{R}^{n\times p}} \quad \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^p (d_{ij} - x_{ij})^2 \quad \text{subject to} \quad \text{rank}(X) = k.$$

A common application is movie recommendation where matrix D collects user preference data of n users rating p movies. In movie recommendation, the user rating matrix D is typically very sparse, as it is rare that an user rates all movies. Hence, the goal is to find a matrix X that completes the missing entries of D. The general hypothesis is that the true rating matrix $X \in \mathbb{R}^{n\times p}$ is generated by the product of an user-feature matrix $U \in \mathbb{R}^{n\times k}$ and a movie-feature matrix $V \in \mathbb{R}^{p\times k}$ over k features as follows.

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix} = \begin{bmatrix} \\ U \\ \\ \end{bmatrix} \begin{bmatrix} & V^\top & \end{bmatrix}.$$

Under such a hypothesis, matrix X has rank k.

As expected, the constraint $\text{rank}(X) = k$ defines a nonconvex set, so the matrix completion problem is nonconvex. A commonly used solution approach is based on relaxing the constraint $\text{rank}(X) = k$ by

$$\|X\|_* = \text{Trace}(\sqrt{X^\top X}) = \sum_{i=1}^{\min\{n,p\}} \sigma_i(X) \leq k$$

where $\|X\|_*$ denotes the **nuclear norm** of X and $\sigma_1(X), \ldots, \sigma_{\min\{n,p\}}(X)$ are the singular values of X.

2.3 Max-Cut

Given a graph $G = (V, E)$, the **max-cut** problem seeks to find a partition the vertex set V so that the number of edges crossing the partition is maximized. Here, a partition (V_1, V_2) of V consists of two sets V_1, V_2 satisfying $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \emptyset$, and the set of edges crossing the partition is basically $\{uv \in E : u \in V_1, v \in V_2\}$. For example, in Figure 9.2, there is a graph of 5 vertices partitioned into red and black vertices, and the edges highlighted are the ones crossing the partition.

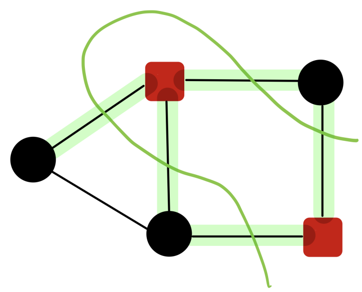

Figure 9.2: Edges crossing a partition

The problem can be formulated by the following (discrete) optimization problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{ij \in E} \frac{1 - x_i x_j}{2} \\ \text{subject to} \quad & x_i \in \{-1, 1\} \text{ for } i \in V. \end{aligned}$$

As long as $x_i \in \mathbb{R}$, $x_i \in \{-1, 1\}$ is equivalent to $x_i^2 = 1$. Hence, the formulation is equivalent to

$$\begin{aligned} \text{maximize} \quad & \sum_{ij \in E} \frac{1 - x_i x_j}{2} \\ \text{subject to} \quad & x_i^2 = 1 \text{ for } i \in V. \end{aligned}$$

Note that the contstraint $x_i \in \{-1, 1\}$ as well as $x_i^2 = 1$ are nonconvex constraints. The **vector relaxation** of the formulation is obtained by replacing x_i by vector $v_i \in \mathbb{R}^k$ as follows.

$$\begin{aligned} \text{maximize} \quad & \sum_{ij \in E} \frac{1 - v_i^\top v_j}{2} \\ \text{subject to} \quad & \|v_i\|_2 = 1 \text{ for } i \in V. \end{aligned}$$

Again, the constraint $\|v_i\|_2 = 1$ is still nonconvex.

Another relaxation is given as follows. Let $d = |V|$. Then we consider a $d \times d$ matrix X whose entry at ith row and jth column, X_{ij}, is $x_i x_j$. Then we have that $X = xx^\top$, which is the outer product of vector x and itself. In fact, X is of the form $X = xx^\top$ if and only if X is positive semidefinite and the rank of X is precisey 1. What this implies is that, the max-cut formulation can be rewritten as

$$\begin{aligned} \text{maximize} \quad & \sum_{ij \in E} \frac{1 - X_{ij}}{2} \\ \text{subject to} \quad & X_{ii} = 1 \text{ for } i \in V, \\ & X \succeq 0, \\ & \text{rank}(X) = 1. \end{aligned}$$

Here, the constrsint $\text{rank}(X) = 1$ is nonconvex. A common approach is to take out the nonconvex constraint and consider

$$\begin{aligned} \text{maximize} \quad & \sum_{ij \in E} \frac{1 - X_{ij}}{2} \\ \text{subject to} \quad & X_{ii} = 1 \text{ for } i \in V, \\ & X \succeq 0. \end{aligned}$$

This is often called the **semidefinite programming (SDP) relaxation** of max-cut. Here, the SDP relaxation is a convex optimization problem.

References

[HM20] Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Oper. Res.*, 68(5):1517–1537, sep 2020. 2.1

[HMS22] Hussein Hazimeh, Rahul Mazumder, and Ali Saab. Sparse regression at scale: branch-and-bound rooted in first-order optimization. *Mathematical Programming*, page 347–388, 2022. 2.1