Optimization Methods	Date: *02-March-2018*
Instructor: *Sujit Prakash Gujar*	Scribes: Binu Jasim T Chaitanya Patel

Lecture 15: Nonlinear Programming - Unconstrained Optimization

Recap. *So far in the first section of this course we covered linear programming and integer linear programming and various techniques to solve them. To briefly summarize, we have learned that;*

- Linear Programming problems can be solved using the Simplex Method. It may take exponential time in the worst case.
- The Ellipsoid method can solve Linear Programs theoretically in polynomial time. But for practical purposes, Simplex method is more efficient than the Ellipsoid method.
- Solving Integer Programs is a hard problem because the solution lies at the integer points near the surface of the feasible region and most of the volume of feasible region is concentrated at the surface.
- For a specific kind of Integer program, where constraints are totally unimodular, we can find the optimal solution in polynomial time.
- For some other Integer programs, we can make use of LP relaxation. Using its solution, we may find the solution or approximate the solution to the original Integer program.
  - Bipartite Matching: We can get optimal integer solution from the solution of LP relaxation in polynomial time.
  - Konig's Theorem: Bipartite Matching and Minimum Vertex Cover Problem are dual of each other.
  - Traveling Salesman Problem
- Game Theory: Using LP duality, we proved that Nash equilibrium exists for two player zero sum games and can be computed in polynomial time by solving a linear program. For general games, the proof of existence of Nash equilibrium is non-trivial and computationally a hard (PPAD-complete) problem.

1 Introduction to Nonlinear Programming

So far in this course we have seen linear programming, where both the objective function and the constraints are linear. Now we'll study a different kind of optimization problems where the objective function or constraints can be nonlinear.

First we'll study unconstrained nonlinear programming. i.e. the optimization of nonlinear objective functions without any constraints. We will use the following problem as an example of an unconstrained nonlinear program.

1.1 Fermat-Weber Problem

Assume that an organization has m facilities to be served. We have to decide a supplying point such that sum of the distances from the supplying point to served centers is minimized.

Formally, for a given set of m facilities $c_1, c_2, \cdots, c_m$, with each $c_i \in \mathbb{R}^n$, the supplying point is computed as $\hat{x} = \min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \|x - c_i\|_2$. If we have different importance to different facilities (that is some facilities need to be served faster than the other), then the problem can be written as deciding the center which minimizes the *weighted* summation of the distances of all the given points from the center.

$$\hat{x} = \min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} w_i \|x - c_i\|_2$$

This is an example of a nonlinear unconstrained optimization problem as the objective is nonlinear and it has no constraints.

Sometimes we don't care about the total distance but we want to serve all the centers in a reasonable amount of time i.e. we want to minimize the maximum time taken to serve any center. This optimization problem can be written as follows:

$$\hat{x} = \min_{x} \ \max_{i} \|x - c_i\|_2$$

we can reformulate the above unconstrained optimization problem as a constrained nonlinear optimization problem as shown below;

$$\begin{aligned} &\min \delta \\ &s.t. \ \ \|x - c_i\|_2 \leq \delta \ \ \forall i \end{aligned}$$

1.2 More Examples

Financial portfolio optimization is an example of non linear optimization. One problem in financial portfolio optimization is to select the best portfolio out of the set of all portfolios being considered, according to some objective. The objective typically maximizes factors such as expected return, and minimizes costs like financial risk. Deciding to invest in k different stocks such that the sum of variances of utilities is minimized is one typical application.

Other examples of nonlinear optimization include support vector machines and the popular deep learning methods.

1.3 Unconstrained Optimization Problem

Standard unconstrained nonlinear optimization problem can be written as follows:

$$\min_{x \in \mathbb{R}^n} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a nonlinear function (which need not be a convex function in general).

2 Concept of Local Minimum and Global Minimum

For a function $f : \mathbb{R}^n \to \mathbb{R}$, we formally define its global minimum and local minimum as follows:

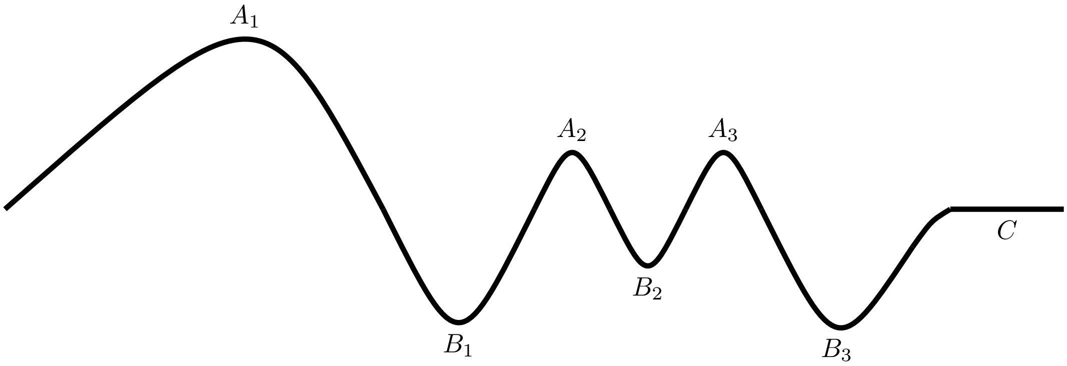

Figure 1: The points A_1, A_2 and A_3 are local maxima and the points B_1, B_2 and B_3 are local minima. A_1 is the global maximum, while both B_1 and B_3 are global minima. The point C (and points around it in the horizontal flat line) is both a local minimum and a local maximum.

Definition 1 (Global Minimum)**.** *A point x^* is called a global minimum of f if*

$$f(x^*) \;\leq\; f(x) \;\; \forall x \in \mathbb{R}^n$$

Definition 2 (Strict Global Minimum)**.** *A point x^* is called a strict global minimum of f if*

$$f(x^*) \;<\; f(x) \;\; \forall x \in \mathbb{R}^n,\; x \neq x^*$$

Definition 3 (Local Minimum)**.** *A point x^* is called a local minimum of f if $\;\exists\, \epsilon > 0$ such that*

$$f(x^*) \;\leq\; f(x) \quad \forall x \in \mathbb{R}^n,\; \|x - x^*\| < \epsilon$$

Definition 4 (Strict Local Minimum)**.** *A point x^* is called a strict local minimum of f if $\;\exists\, \epsilon > 0$*

$$f(x^*) \;<\; f(x) \quad \forall x \in \mathbb{R}^n,\; \|x - x^*\| < \epsilon,\; x \neq x^*$$

Note that a point which is both a local minimum and a local maximum can't be a strict minimum or a strict maximum.

3 First Order Necessary Condition

3.1 Intuition of First Order Condition

Consider a function f which is continuous and differentiable,

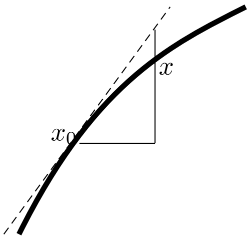

Figure 2: Linear approximation of a curve in small neighborhood is a straight line

A function is approximately linear in a very small neighborhood of any point lying on that function. The slope of that line is approximated as the slope of the tangent at that point.

$$f(x) \approx f(x_0) + (x - x_0)\, f'(x_0)$$

For multidimensional functions $f : \mathbb{R}^n \to \mathbb{R}$, the first order approximation is given as follows;

$$f(x) \approx f(x_0) + \delta^T\, \nabla f(x_0)$$
$$\delta = x - x_0 \in \mathbb{R}^n$$
$$\nabla f(x_0) = \begin{bmatrix} \dfrac{\partial f(x_0)}{\partial x_1} \\ \dfrac{\partial f(x_0)}{\partial x_2} \\ \vdots \\ \dfrac{\partial f(x_0)}{\partial x_n} \end{bmatrix}$$

($\nabla f(x_0)$ is called the *gradient* of $f(x)$ at x_0).

If x^* is a local minimum, then $f(x^*)$ should be the least value in a small neighborhood around x^*

$$\therefore f(x) - f(x^*) \approx \nabla f(x^*)^T \delta \geq 0$$

If $\delta = (\epsilon, 0, 0, 0, \cdots)^T$ then

$$\epsilon\, \frac{\partial f(x^*)}{\partial x_1} \geq 0$$

We can do this for all dimensions. Thus for $1 \leq i \leq n$

$$\epsilon\, \frac{\partial f(x^*)}{\partial x_i} \geq 0 \tag{1}$$

But since x^* is a local minimum of f, above derivation is true for any arbitrary $\epsilon \in \mathbb{R}$. Thus it should be true for $-\epsilon$ also.

$$-\epsilon\, \frac{\partial f(x^*)}{\partial x_i} \geq 0 \tag{2}$$

From Eq. 1 and 2,

$$\frac{\partial f(x^*)}{\partial x_i} = 0$$
$$\Rightarrow \nabla f(x^*) = 0$$

(Note that this is not a formal proof as we used the the first order approximation to arrive at this).

$\nabla f(x^*) = 0$ is not a sufficient condition to be a local minimum. For example x^* could be a local maximum also.

3.2 Proof of First Order Necessary Condition

Let $d \in \mathbb{R}^n$ be a non zero vector and α be a positive real value, then define $g : \mathbb{R} \to \mathbb{R}$, as a function of α as

$$g(\alpha) = f(x^* + \alpha d) \tag{3}$$

Its right derivative is given by

$$g^{'}(\alpha) = \lim_{\alpha \to 0^+} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha}$$

Because x^* is a local minimum, $f(x^*)$ should be the least value in a small neighborhood around x^*. So for sufficiently small (positive) α, $f(x^* + \alpha d) \geq f(x^*)$.

$$\therefore g^{'}(\alpha) \geq 0$$

Differentiating Eq. 3 w.r.t. α using chain rule and using the above result, we get,

$$\begin{aligned} g^{'}(\alpha) &= d^T \nabla f(x^* + \alpha d) \geq 0 \\ \therefore g^{'}(0) &= d^T \nabla f(x^*) \geq 0 \end{aligned} \tag{4}$$

But since $f(x^*)$ is a local minimum, $f(x^*) \leq f(x^* + \alpha d)$ holds true for all points around x^* at a sufficiently small radius. Hence it should be true for $-d$ as well (α can be made arbitrarily small for large d). Hence

$$\begin{aligned} g^{'}(\alpha) &= \lim_{\alpha \to 0} \frac{f(x^* - \alpha d) - f(x^*)}{\alpha} \geq 0 \\ \therefore g^{'}(\alpha) &= -d^T \nabla f(x^* - \alpha d) \geq 0 \\ \therefore g^{'}(0) &= -d^T \nabla f(x^*) \geq 0 \end{aligned} \tag{5}$$

From 4 and 5 we get $\nabla f(x^*) = 0$ (since d is not zero). Hence the proof.

4 Second Order Necessary Condition

A function $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable if $\frac{d}{dx}(\frac{df}{dx})$ exists.

For a twice differentiable function, $f : \mathbb{R}^n \to \mathbb{R}$, its Hessian matrix $\nabla^2 f(x)$ is defined as

$$\nabla^2 f(x) = \begin{bmatrix} \dfrac{\partial^2 f(x)}{\partial x_1^2} & \dfrac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 f(x)}{\partial x_n \partial x_1} & \dfrac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

Using Taylor series expansion of a function at the local minimum point x^*, we define a new function $g : \mathbb{R} \to \mathbb{R}$ as

$$g(\alpha) = f(x^* + \alpha d) = f(x^*) + \alpha\ d^T \nabla f(x^*) + \frac{\alpha^2}{2} d^T\ \nabla^2 f(x^*)\ d + o(\alpha^2) \tag{6}$$

where $o(\alpha^2)$ sums up all the higher order terms which are negligible for small α. i.e.

$$\lim_{\alpha \to 0} \frac{o(\alpha^2)}{\alpha^2} = 0$$

Notice that in Eq. 6, the first order term $d^T \nabla f(x^*) = 0$ by the first order necessary condition. Also since x^* is a local minimum $f(x^* + \alpha d) - f(x^*) \geq 0$ for sufficiently small α.

So from Eq. 6 we get

$$\frac{\alpha^2}{2} d^T \ \nabla^2 f(x^*) \ d + o(\alpha^2) \geq 0$$

dividing by α^2 and taking the limit $\alpha \to 0$, we obtain the second order necessary condition,

$$d^T \ \nabla^2 f(x^*) \ d \geq 0 \tag{7}$$

A matrix A is called positive semi definite (psd) if $x^T A x \geq 0 \ \forall x \in \mathbb{R}^n$

we can see that the Hessian ($\nabla^2 f(x^*)$) is positive semi definite at x^*. This is the second order necessary condition for local minima.
Note 1: The Identity matrix is a positive semi definite matrix because $x^T I x = x^T x = ||x||^2 \geq 0$.

Note 2: The function $f(x) = |x|$ has a local minimum at $x^* = 0$, but $\nabla f(x^*) \neq 0$. This doesn't violate our first order necessary condition because $f(x) = |x|$ is not differentiable at $x = 0$.

In the next class we will see some sufficient conditions for local minima/maxima.

References

[1] Dimitri P. Bertsekas *Nonlinear Programming*, Athena Scientific, ISBN: 978-1-886529-05-2