ECE 490: Introduction to Optimization **Fall 2018**

Lecture 18

Optimization with Equality Constraints

Lecturer: Bin Hu, Date:11/1/2018

In this lecture, we will talk about optimization problems with equality constraints:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h_j(x) = 0,\ j = 1, \ldots, l \end{aligned} \tag{18.1}$$

We will talk about how to address linear equality constraints $Ax - b = 0$ and explain the main difficulties for handling general h_j.

18.1 Optimality Conditions and Lagrange Multipliers

For unconstrained optimization, we know that any local min x^* of a differentiable f has to satisfy $\nabla f(x^*) = 0$. Then the optimization problem can be viewed as an equation solving task. For constrained optimization, we also have optimality conditions. We assume that h_j is differentiable for all j. A point x is said to be a regular point for (18.1) if $\nabla h_1(x)$, $\nabla h_2(x)$, ..., $\nabla h_{l-1}(x)$, and $\nabla h_l(x)$ are linearly independent. Now we are ready to state the main optimality condition for (18.1).

Theorem 18.1. *Suppose x^* is a local min and a regular point for* (18.1)*. Then there exist unique scalars λ_1^*, λ_2^*, ..., λ_l^* such that*

$$\nabla f(x^*) + \sum_{j=1}^{l} \lambda_j^* \nabla h_j(x^*) = 0. \tag{18.2}$$

Here the scalars λ_1^*, λ_2^*, ..., λ_l^* are called Lagrange multipliers. The above condition is part of the Lagrange multiplier theorem. The regularity condition on x^* is important. Otherwise there may not exist any λ_1^*, λ_2^*, ..., λ_l^* such that (18.2) is satisfied. Here is an example.

Example: Consider the following minimization problem

$$\begin{aligned} \text{minimize} \quad & x_1 + x_2 \\ \text{subject to} \quad & (x_1 - 1)^2 + x_2^2 = 1 \\ & (x_1 - 2)^2 + x_2^2 = 4 \end{aligned}$$

This problem has only one feasible point $(x_1, x_2) = (0, 0)$. Therefore, the objective function is minimized at this feasible point, and we have $(x_1^*, x_2^*) = (0, 0)$. We have

$$\nabla f(x^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ \nabla h_1(x^*) = \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \ \nabla h_2(x^*) = \begin{bmatrix} -4 \\ 0 \end{bmatrix}$$

Clearly there does not exist λ_1^* and λ_2^* such that $\nabla f(x^*) + \lambda_1^* \nabla h_1(x^*) + \lambda_2^* \nabla h_2(x^*) = 0$. The issue here is that $\nabla h_1(x^*)$ and $\nabla h_2(x^*)$ are linearly dependent in this case.

So (18.2) only provides a necessary condition for regular local minimum points. Finding a local min that is not regular is a difficult task since we may not have a well-defined optimality condition for these points in the first place.

Exceptions: Suppose we only have a linear equality constraint $Ax - b = 0$. If A is full rank, then any feasible point is actually regular. However, an amazing fact is that the regularity assumption on x^* can be dropped and we can still find Lagrangian multipliers when A is not full rank. In this case, we do not have uniqueness of the Lagrangian multipliers. There may exist many choices of λ_1^*, λ_2^*, ..., λ_{l-1}^*, and λ_l^* such that (18.2) holds if A is not full rank. Just think that you can discard the redundant equality constraints and assign the associated Lagrangian multipliers to be 0.

18.2 Lagrangian Function

Now we define the Lagrangian function as $L(x, \lambda) = f(x) + \lambda^\mathsf{T} h(x)$ where λ and h are given as

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_l \end{bmatrix}, \ h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_l(x) \end{bmatrix}$$

Then any regular local min x^* and associated unique Lagrange multiplier λ^* will form a stationary point for the Lagrangian function $L(x, \lambda)$. Specifically, we have

$$\nabla_x L(x^*, \lambda^*) = 0 \tag{18.3}$$
$$\nabla_\lambda L(x^*, \lambda^*) = 0 \tag{18.4}$$

Notice (18.3) is equivalent to (18.2), and (18.4) just restates the fact $h(x^*) = 0$.

Therefore, we can find all the regular local minimum points of (18.1) if we know all the stationary points of $L(x, \lambda)$.

18.3 Duality

The duality theory is built upon the following inequality:

$$\max_{\lambda\in\mathbb{R}^l} D(\lambda) = \max_{\lambda\in\mathbb{R}^l}\min_{x\in\mathbb{R}^p} L(x,\lambda) \le \min_{x\in\mathbb{R}^p}\max_{\lambda\in\mathbb{R}^l} L(x,\lambda) = \min_{x:h(x)=0} f(x) \tag{18.5}$$

where $D(\lambda) := \min_{x\in\mathbb{R}^p} L(x,\lambda)$ is the so-called dual function. Now we explain the above statement:

1. $\max_{\lambda\in\mathbb{R}^l} D(\lambda) = \max_{\lambda\in\mathbb{R}^l}\min_{x\in\mathbb{R}^p} L(x,\lambda)$: This follows from the definition of the dual function.

2. $\max_{\lambda\in\mathbb{R}^l}\min_{x\in\mathbb{R}^p} L(x,\lambda) \le \min_{x\in\mathbb{R}^p}\max_{\lambda\in\mathbb{R}^l} L(x,\lambda)$: This follows from the fact that we have $L(x,\lambda) \le \max_{\lambda\in\mathbb{R}^l} L(x,\lambda)$ for any x and λ. Consequently, we have $\min_{x\in\mathbb{R}^p} L(x,\lambda) \le \min_{x\in\mathbb{R}^p}\max_{\lambda\in\mathbb{R}^l} L(x,\lambda)$ which directly leads to the desired inequality.

3. $\min_{x\in\mathbb{R}^p}\max_{\lambda\in\mathbb{R}^l} L(x,\lambda) = \min_{x:h(x)=0} f(x)$: This is a direct consequence of the following relation:

$$\max_{\lambda\in\mathbb{R}^l} L(x,\lambda) = \begin{cases} f(x) & \text{if } h(x)=0 \\ +\infty & \text{Otherwise} \end{cases} \tag{18.6}$$

   Notice if we do not have $h(x)=0$, then we can always choose some arbitrarily large λ to make $f(x)+\lambda^{\mathsf{T}}h(x)$ go to infinity.

Concavity of dual function. A remarkable property of the dual function is that it is always concave no matter what f we have. Please verify this fact by yourself. The only inequality you need to prove the concavity of D is $\min_{x\in\mathbb{R}^p}\{a(x)+b(x)\} \ge \min_{x\in\mathbb{R}^p} a(x) + \min_{x\in\mathbb{R}^p} b(x)$. Since the dual function is concave, we can always apply the gradient ascent method to maximize the dual function when it is differentiable. Based on (18.5), we can obtain a lower bound for (18.1) by maximizing the dual function. Sometimes the lower bound obtained in this way may be too conservative and is not that useful.

Strong duality. If the inequality in (18.5) holds as an equality, i.e. $\max_{\lambda\in\mathbb{R}^l}\min_{x\in\mathbb{R}^p} L(x,\lambda) = \min_{x\in\mathbb{R}^p}\max_{\lambda\in\mathbb{R}^l} L(x,\lambda)$, then we have the so-called strong duality. Strong duality means that the global max of the dual function is equal to the global min of the primal problem (18.1). Strong duality is firmly related to the saddle point of the Lagrangian function $L(x,\lambda)$.

Saddle point of Lagrangian We call (x^*,λ^*) a saddle point of $L(x,\lambda)$ if the following condition holds

$$L(x^*,\lambda) \le L(x^*,\lambda^*) \le L(x,\lambda^*). \tag{18.7}$$

When we have strong duality, we know (x^*,λ^*) is a saddle point of $L(x,\lambda)$ where x^* is a global min of the primal problem (18.1) and λ^* is a global max of the dual function $D(\lambda)$. Consequently, a large family of saddle point algorithms can be used to solve (18.1) when we have strong duality.

When do we have strong duality? In general, the proof of strong duality is difficult and case-dependent. When we only have linear equality constraints for (18.1) and the objective function f is convex, the strong duality is automatically guaranteed. This means that we can apply various saddle point methods to solve (18.1) with convex f and affine h.

18.4 Algorithms for Linear Equality Constraints

When f is convex and the only constraint is $Ax = b$, we have strong duality and all we need to do is to find the saddle point of $L(x, \lambda)$. When f is strongly-convex and A is full row rank, one can further prove that the saddle point of L exists and is unique. Now we can apply an iterative algorithm that performs gradient descent on x and gradient ascent on λ after initializing from some point. This leads to the following algorithm:

$$x_{k+1} = x_k - \alpha \nabla_x L(x_k, \lambda_k)$$
$$\lambda_{k+1} = \lambda_k + \eta(Ax_k - b)$$

where $x_k \in \mathbb{R}^p$ and $\lambda_k \in \mathbb{R}^l$ are vectors at step k. This is called the discrete-time primal dual gradient algorithm. When f is smooth and strongly convex, the method converges linearly.

Instead of performing gradient descent on x, we can also minimize the Lagrangian function over x. This leads to the so-called dual ascent algorithm:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^p} L(x, \lambda_k)$$
$$\lambda_{k+1} = \lambda_k + \eta(Ax_{k+1} - b)$$

Sometimes, there are other structures in the problem and more efficient algorithms are available. In next lecture, we will talk about the augmented Lagrangian method and the alternating direction of multiplier method (ADMM).