1 Outline

In this lecture, we study

- Newton's method for equality constrained minimization
- Barrier method.

2 Newton's method for equality constrained minimization

Let us consider the following convex optimization problem with equality constraints.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b. \end{aligned} \tag{24.1}$$

Here, $Ax = b$ consists of affine constraints, and the objective function f is convex and twice continuously differentiable. Recall that for the unconstrained setting, Newton's method proceeds with the update rule

$$x_{t+1} \in \operatorname*{argmin}_x \left\{ f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2}(x - x_t)^\top \nabla^2 f(x_t)(x - x_t) \right\}$$

from which we deduce

$$x_{t+1} = x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t).$$

Here, the descent direction $d = -\nabla^2 f(x_t)^{-1} \nabla f(x_t)$ can be directly computed by

$$d \in \operatorname*{argmin}_x \left\{ f(x_t) + \nabla f(x_t)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_t) d \right\}$$

because $x_{t+1} = x_t + d$. Based on this, we may extend Newton's method to the equality constrained problem. Basically, the direction d for the update rule can be computed as an optimal solution to the following optimization problem

$$\begin{aligned} \text{minimize} \quad & f(x_t) + \nabla f(x_t)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_t) d \\ \text{subject to} \quad & A(x_t + d) = b. \end{aligned} \tag{24.2}$$

Here, if this optimization problem has a solution, then $x_t + d$ is indeed a feasible solution to (24.1). In fact, we can characterize such a direction d by the KKT conditions. Note that the associated Lagrangian is given by

$$L(d, \mu) = f(x_t) + \nabla f(x_t)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_t) d + \mu^\top (A(x_t + d) - b).$$

Then, since f is convex and the constraints are all affine, it follows from the KKT conditions that d is an optimal solution to (24.2) if and only if there exists μ such that

$$\begin{aligned} \nabla f(x_t) + \nabla^2 f(x_t) d + A^\top \mu &= 0, \\ A(x_t + d) &= b. \end{aligned}$$

Subject to $Ax_t = b$, this can be expressed as the following matrix system.

$$\begin{bmatrix} \nabla^2 f(x_t) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \mu \end{bmatrix} = \begin{bmatrix} -\nabla f(x_t) \\ 0 \end{bmatrix}.$$

Here, the matrix

$$\begin{bmatrix} \nabla^2 f(x_t) & A^\top \\ A & 0 \end{bmatrix}$$

is referred to as the KKT matrix.

3 Barrier method

In this section we consider the following constrained convex minimization problem.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0, \quad i = 1, \dots, m, \\ & Ax = b. \end{aligned} \tag{24.3}$$

Comparing this setting and (24.1), we have additional inequality constraints $g_i(x) \le 0$ for $i \in [m]$. Suppose that (24.3) satisfies Slater's condition. As an example of (24.3), we consider linear programs of the form

$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{subject to} \quad & p_i^\top x \le q_i, \quad i = 1, \dots, m, \\ & Ax = b. \end{aligned} \tag{24.4}$$

In the last section, we dealt with the equality constrained setting, motivated by which we consider the following equivalent setting of (24.3).

$$\begin{aligned} \text{minimize} \quad & f(x) + \sum_{i=1}^m I_{\mathbb{R}_-}(g_i(x)) \\ \text{subject to} \quad & Ax = b \end{aligned} \tag{24.5}$$

where $\mathbb{R}_- = \{x \in \mathbb{R} : \ x \le 0\}$ and $I_{\mathbb{R}_-}$ is the associated indicator function. Here, the indicator function $I_{\mathbb{R}_-}$ is non-smooth. One way of dealing with this is to approximate the indicator function, for which we consider so-called *barrier functions*. There are two common examples for barrier functions as follows.

$$\begin{aligned} \text{log-barrier :} \quad & \psi(x) = -\sum_{i=1}^m \log(-g_i(x)), \\ \text{inverse :} \quad & \psi(x) = -\sum_{i=1}^m \frac{1}{g_i(x)}. \end{aligned}$$

The important property of barrier function $\psi(x)$ is that as $g_i(x)$ approaches 0, $\psi(x)$ gets arbitrarily large and goes to $+\infty$. Note that both functions are convex if $g_1, \dots, g_m$ are convex. In this section, we focus on the log-barrier function. For the linear program given by (24.4), the corresponding log-barrier function is given by

$$\psi(x) = -\sum_{i=1}^m \log(q_i - p_i^\top x).$$

Before we discuss some specific properties of the log-barrier function, we explain the general outline of the barrier method and related concepts. The basic idea is to consider

$$\begin{aligned} \text{minimize} \quad & f(x) + \frac{1}{t}\psi(x) \\ \text{subject to} \quad & Ax = b \end{aligned} \tag{24.6}$$

where ψ is the barrier function and $t > 0$ is a parameter that we increase over time.

3.1 Central path

Suppose for now that (24.6) has a unique optimal solution. Note that (24.6) is equivalent to

$$\begin{aligned} \text{minimize} \quad & tf(x) + \psi(x) \\ \text{subject to} \quad & Ax = b \end{aligned} \tag{24.7}$$

In fact, the uniqueness can be guaranteed for many of the important applications as the negative log function $-\log x$ is strictly convex. For example, linear programs and quadratic programs. Let

$$x^\star(t) = \operatorname*{argmin}_x \{tf(x) + \psi(x) : \; Ax = b\}.$$

Here, the set consists of the optimal solutions for varying values of t

$$\{x^\star(t) : \; t > 0\}$$

is referred to as the *central path*. Note that each point $x^\star(t)$ is a feasible solution to (24.3), and therefore, the central path is fully contained in the feasible region of the original optimization problem (24.3). Figure 24.1[1] shows the central path for a linear program, Here, the dotted contours

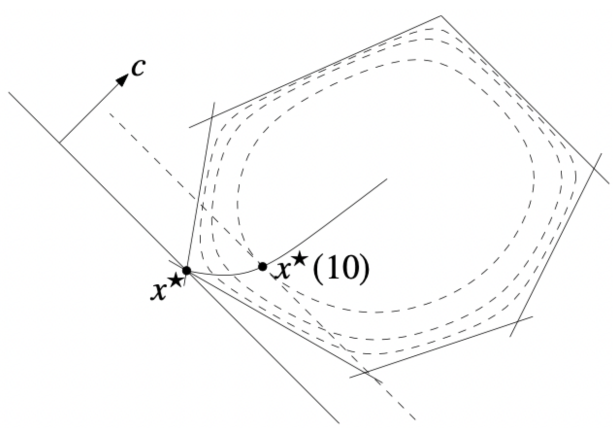

Figure 24.1: Central path for a linear program

correspond to the log-barrier function. Interestingly, the hyperplane $c^\top x = c^\top x^\star(t)$ containing $x^\star(t)$ with direction c is tangent to the contour containing $x^\star(t)$. This can be seen from characterizing the central path with the KKT conditions.

[1]The figure is taken from the lecture slides of Stanford University's EE364a: Convex Optimization by Boyd and Vandenberghe.

Note that the gradient of the log-barrier function is given by

$$\nabla\psi(x) = -\sum_{i=1}^{m} \frac{1}{g_i(x)} \nabla g_i(x).$$

As the Lagrangian of (24.7) is given by

$$L(x, \mu) = tf(x) + \psi(x) + \mu^\top (Ax - b),$$

the KKT conditions state that $x^\star(t)$ is optimal to (24.7) if and only if there exists $\mu^\star$ such that

$$\begin{aligned} t\nabla f(x^\star(t)) - \sum_{i=1}^{m} \frac{1}{g_i(x^\star(t))} \nabla g_i(x^\star(t)) + A^\top \mu^\star = 0, \\ g_i(x^\star(t)) < 0, \quad i = 1, \ldots, m, \\ Ax^\star(t) = b. \end{aligned}$$

For a linear program with an equality constraint, i.e. $A = 0$ and $b = 0$, the characterization of $x^\star(t)$ states that

$$t \cdot c = -\nabla\psi(x^\star(t)) = \sum_{i=1}^{m} \frac{1}{p_i^\top x - q_i} p_i.$$

Note that the direction of the tangent hyperplane at $x^\star(t)$ is given by $\nabla\psi(x^\star(t))$ and it is a scaling of the objective direction c.

3.2 Duality gap

By definition, $x^\star(t)$ is feasible to (24.3) by definition. We may construct a feasible dual solution associated with $x^\star(t)$. Let $\lambda_i^\star(t)$ and $\mu^\star(t)$ be defined as

$$\lambda_i^\star(t) = -\frac{1}{t \cdot g_i(x^\star(t))}, \quad i = 1, \ldots, m, \qquad \mu^\star(t) = \frac{\mu^\star}{t}.$$

By definition, it follows that

$$\begin{aligned} \nabla f(x^\star(t)) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^\star(t)) + A^\top \mu^\star(t) = 0, \\ \lambda_i^\star(t) > 0, \quad i = 1, \ldots, m. \end{aligned}$$

This implies that

$$\begin{aligned} L(x^\star(t), \lambda^\star(t), \mu^\star(t)) &= f(x^\star(t)) + \sum_{i=1}^{m} \lambda_i^\star(t) g_i(x^\star(t)) + \mu^\star(t)^\top (Ax^\star(t) - b) \\ &= \min_x \left\{ f(x) + \sum_{i=1}^{m} \lambda_i^\star(t) g_i(x) + \mu^\star(t)^\top (Ax - b) \right\} \\ &= q(\lambda^\star(t), \mu^\star(t)) \end{aligned}$$

where $L(x, \lambda, \mu)$ is the Lagrangian function for (24.3). Furthermore,

$$f(x^\star(t)) - q(\lambda^\star(t), \mu^\star(t)) = -\sum_{i=1}^{m} \lambda_i^\star(t) g_i(x^\star(t)) - \mu^\star(t)^\top (Ax^\star(t) - b) = \frac{m}{t}.$$

Since the Lagrangian dual function $q(\lambda, \mu)$ provides a lower bound on the optimal value of (24.3), it follows that

$$f(x^\star(t)) - \min\{f(x): \ g_i(x) \le 0, \ i = 1, \dots, m, \ Ax = b\} \le \frac{m}{t}.$$

This suggests an algorithm for solving (24.3).

3.3 Implementing the barrier method

Suppose that the desired accuracy for solving (24.3) is ϵ. In other words, we want to find a feasible solution x such that

$$f(x) - \min\{f(x): \ g_i(x) \le 0, \ i = 1, \dots, m, \ Ax = b\} \le \epsilon.$$

In this case, we may choose $t = m/\epsilon$ and obtain $x^\star(m/\epsilon)$ by applying the barrier method. However, when ϵ is tiny, solving (24.7) with huge $t = m/\epsilon$ can be numerically unstable. Hence, in practice, we incrementally increase the value of t instead of setting it to a large value upfront. Here is the general template.

1. Initialize $t^0 > 0$ and $\alpha > 1$.
2. Obtain $x^0 = x^\star(t^0)$.
3. For $k = 1, 2, 3, \dots$, repeat the following.
   - Set $t^k = \alpha t^{k-1}$.
   - Apply Newton's method initialized at x^{k-1} to obtain $x^k = x^\star(t^k)$.
   - Break if $m/t^k \le \epsilon$.

We may easily deduce the convergence analysis of the barrier method. Suppose that k is the smallest number such that $m/t^k \le \epsilon$. This means that

$$\frac{m}{\alpha^{k-1}t^0} \ge \epsilon,$$

which in turn implies that

$$k \le 1 + \frac{1}{\log \alpha} \log \frac{m}{t^0 \epsilon} = O\left(\log \frac{m}{\epsilon}\right).$$

3.4 Perturbed KKT conditions

Recall that $\lambda_i^\star(t)$ and $\mu^\star(t)$ defined as

$$\lambda_i^\star(t) = -\frac{1}{t \cdot g_i(x^\star(t))}, \quad i = 1, \dots, m, \qquad \mu^\star(t) = \frac{\mu^\star}{t}$$

together with $x^\star(t)$ satisfy $\nabla f(x^\star(t)) + \sum_{i=1}^m \lambda_i \nabla g_i(x^\star(t)) + A^\top \mu = 0$. By definition, $(x, \lambda, \mu) = (x^\star(t), \lambda^\star(t), \mu^\star(t))$ satisfies

$$\begin{aligned} \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla g_i(x) + A^\top \mu = 0, \\ \lambda_i g_i(x) = -\frac{1}{t}, \quad i = 1, \dots, m, \\ g_i(x) \le 0, \quad i = 1, \dots, m \\ Ax = b, \\ \lambda_i \ge 0, \quad i = 1, \dots, m. \end{aligned} \tag{24.8}$$

Here, the only difference between this system and the KKT conditions is the condition $\lambda_i g_i(x) = -1/t$ for $i \in [m]$. In fact, as $t \to +\infty$, the condition gets close to the complementary slackness condition $\lambda_i g_i(x) = 0$ for $i \in [m]$. For this reason, the conditions (24.8) are referred to as the *perturbed KKT conditions*.