0.1 Gradient descent

- *Algorithm:* Gradient Descent(algo. 1)
- *Input:* A vector
- *Complexity:* None.
- *Data structure compatibility:* None
- *Common applications:* Artificial intelligence

Problem. Gradient descent

The purpose of optimization is to minimize our target function, for example,

$$\min f(x) \tag{0.1.1}$$

However, for some large scale function, analytical solution $x^\star = (A^\top A)^{-1} A^T Y$ is unsolvable and time consuming ($\mathcal{O}(n^3)$). So the descent method is raised to solve optimization.

Description

Descent Method

To solve equation 0.1.1, a sequence of points $x^{(k)}$ is produced to approach the optimum.

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \quad \text{with} \quad f\left(x^{(k+1)}\right) < f\left(x^{(k)}\right) \tag{0.1.2}$$

The general descent method is shown in Algorithm 1

Algorithm 1: General descent method

Input : a starting point $x \in$ **dom** f
Output: optimal x to minimize $f(x)$

while *Stopping criterion is not satisfied* **do**
  Determine a descent direction Δx.
  Line search. Choose a step size $t > 0$.
  Update. $x = x + t\Delta x$
end while
return x

Gradient Descent Algorithm

In this section, we will decide the direction Δx in Algorithm1. From convexity

$$f\left(x^{(k+1)}\right) \le f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right) \Delta x^{(k)} \tag{0.1.3}$$

So,

$$f\left(x^{(k+1)}\right) < f\left(x^{(k)}\right) \Rightarrow \nabla f\left(x^{(k)}\right) \Delta x^{(k)} < 0 \tag{0.1.4}$$

A natural choice is gradient: $\Delta x^{(k)} = \nabla f\left(x^{(k)}\right)$.

Line Search: Backtracking

This section will find the step size t. t can be described by

$$t = \text{argmin}_t f(x + t\Delta x) \tag{0.1.5}$$

We use two parameters $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$. Starting from $t = 1$, repeat $t = \beta t$ until

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^{\mathrm{T}} \Delta x \tag{0.1.6}$$

See Figure 1.

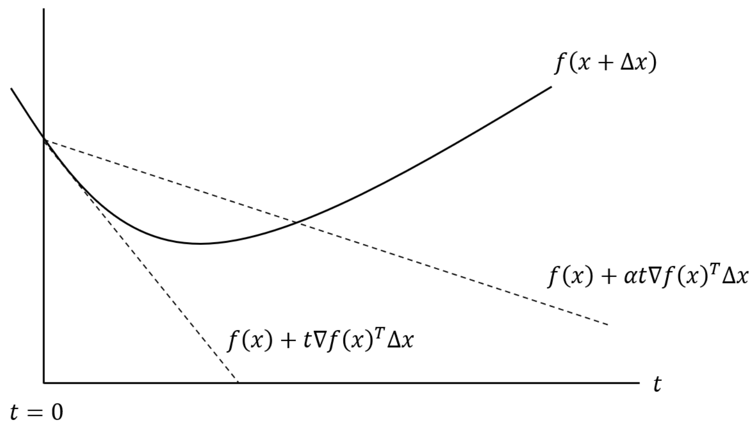

Figure 1: Backtracking

Constrained Optimization

Consider the following optimization problems that include inequality constraints,

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \cdots, m \\ & Ax = b \end{array} \tag{0.1.7}$$

We use central path. The equivalent problem of problem 0.1.7 is

$$\begin{array}{ll} \text{minimize} & tf_0(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array} \tag{0.1.8}$$

which has the same minimizers. We assume problem 0.1.8 can be solved by GD and has the unique solution for any $t > 0$. We use $x^\star(t)$ to denote the solution to problem 0.1.8, and we call it central point. We define the set of $x^\star(t)$ the central path. $x^\star(t)$ should satisfy,

$$Ax^\star(t) = b. \quad f_i(x^\star(t)) < 0, \quad i = 1, 2, \cdots, mao \tag{0.1.9}$$

There exits $\hat{\nu} \in R^p$

$$0 = t\nabla f_0(x^\star(t)) + \nabla \phi(x^\star(t)) + A^T \hat{\nu} \tag{0.1.10}$$

$$= t\nabla f_0(x^\star(t)) + \sum_{i=1}^{m} \frac{1}{-f_i(x^\star(t))} + A^T \hat{\nu} \tag{0.1.11}$$

A geometric explanation is shown in Figure 2.

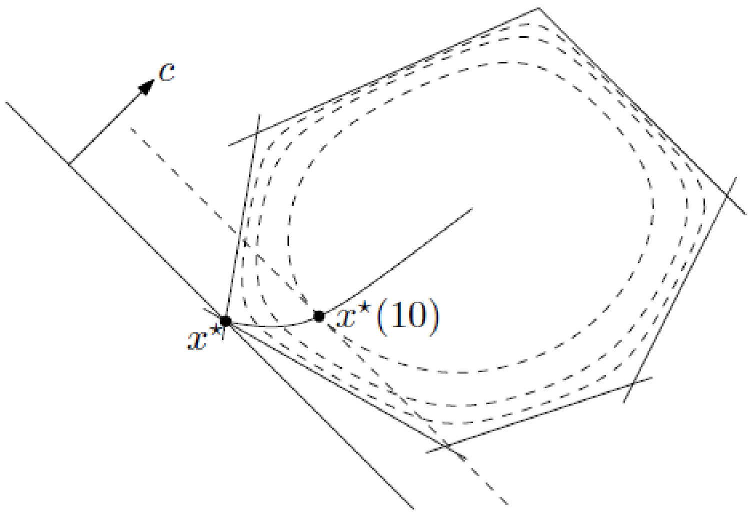

Figure 2: Central path for an LP with n = 2 and m = 6