Optimization Methods (CS1.404)
Spring 2024

Naresh Manwani

Machine Learning Lab, IIIT-H

February 5th, 2024

Descent Direction Methods

- We consider the unconstrained minimization problem as follows:

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

  where we assume that f is continuously differentiable over $\mathbb{R}^n$.
- In many cases, it might be very difficult to solve the equation $\nabla f(\mathbf{x}) = \mathbf{0}$ to find the stationary points.
- Even if it is possible to find the solutions of $\nabla f(\mathbf{x}) = \mathbf{0}$, if there are infinitely many solutions, finding the one corresponding to a local minima might be as difficult problem as original optimization problem.
- Due to these reasons, instead of finding the stationary points analytically, we consider adopting an iterative algorithm to find them.
- Iterative algorithms to find the stationary points are of the following form:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k, \quad k = 0, 1, 2, \ldots,$$

  where $\mathbf{d}_k$ is the so-called direction t_k is the stepsize.

Descent Direction

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function over $\mathbb{R}^n$. A vector $\mathbf{d} \in \mathbb{R}^n$ $(\mathbf{d} \neq \mathbf{0})$ is said a **descent direction** of f at $\mathbf{x}$ if the directional derivative of f at $\mathbf{x}$ along the direction $\mathbf{d}$ is negative, i.e.,

$$\nabla f(\mathbf{x})^T \mathbf{d} < 0$$

Remark: Taking small enough steps along descent directions lead to a decrease of the function f.

Descent Property of Descent Directions

Lemma

Let f be a continuously differentiable function over an open set S of $\mathbb{R}^n$ and let $\mathbf{x} \in S$. Suppose that $\mathbf{d}$ is a descent direction of f at $\mathbf{x}$. Then there exist $\epsilon > 0$ such that

$$f(\mathbf{x} + \alpha \mathbf{d}) < f(\mathbf{x})$$

for any $\alpha \in (0, \epsilon]$.

Schematic Descent Directions Method

Schematic Descent Directions Method

- **Initialization:** Pick $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrarily
- **General Step:** For any $k = 0, 1, 2, \ldots$, set
  1. Pick a descent direction $\mathbf{d}_k$.
  2. Find a step size t_k satisfying $f(\mathbf{x}_k + t_k\mathbf{d}_k) < f(\mathbf{x}_k)$.
  3. Set $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k$.
  4. STOP if the stopping condition is satisfied and Output $\mathbf{x}_{k+1}$. Else go to Step (1).

Challanges:

1. How to choose the initial point $\mathbf{x}_0$?
2. How to choose the descent direction $\mathbf{d}_k$?
3. How to choose the stepsize t_k?
4. What should be the stopping condition?
5. Does the algorithm converge? If yes, then how fast does it converge? Does the convergence depend on $\mathbf{x}_0$?

Stopping Condition

1. Stopping condition for a minimization problem is $\nabla f(\mathbf{x}_k) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}_k)$ is positive semi-definite.
2. A practical stopping condition is $\|\nabla f(\mathbf{x}_k)\| \leq \epsilon$.
3. Other stopping conditions

$$\|\nabla f(\mathbf{x}_k)\| < \epsilon(1 + |f(\mathbf{x}_k)|)$$
$$\frac{f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})}{|f(\mathbf{x}_k)|} \leq \epsilon$$

Finding Step Size t_k

- Step size t_k is chosen in such a way that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$.
- The method of finding step size is called line search, since it a minimization of one dimensional function $g(t) = f(\mathbf{x}_k + t\mathbf{d}_k)$.
- Four popular choices for step size selection are as follows:
  - **Constant Step size:** $t_k = \eta,\ \forall k$. It is very simple approach, but it is unclear how to choose η. A large value of η might cause the algorithm to be nondecreasing and small η can cause very slow convergence.
  - **Diminishing Step Size:** $\alpha_k \to 0$, $\sum_{k=1}^{\infty} \alpha_k = \infty$. For example, $\alpha_k = \frac{1}{k}$.
    - Descent not guaranteed at each step; only later when becomes small.
    - $\sum_{k=1}^{\infty} \alpha_k = \infty$ imposed to guarantee progress does not become too slow.
    - Good theoretical guarantees, but unless the right sequence is chosen, can also be a slow method.

Finding Step Size t_k

- **Exact Line Search:** Here, t_k is the minimizer of f along the ray $\mathbf{x}_k + t\mathbf{d}_k$.

$$t_k = \arg\min_{t \geq 0} \; f(\mathbf{x}_k + t\mathbf{d}_k)$$

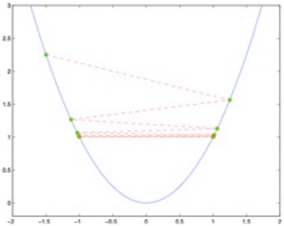
  It is an attractive approach, but it is not always possible to find the exact minimizer of $g(t) = f(\mathbf{x}_k + t\mathbf{d}_k)$.

- **Inexact Line Search:** This method iteratively finds t_k which minimizes f along the ray $\mathbf{x}_k + t\mathbf{d}_k$. It finds good enough step size which ensures sufficient decrease.

Example 1: How line search methods fails !

Large Step Sizes

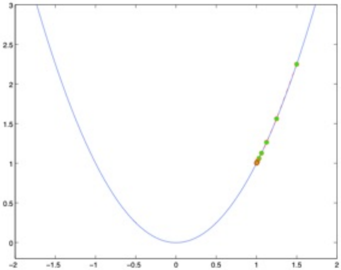
- The objective function $f(x) = x^2$. Global minimizer is $x^* = 0$ and optimal value of $f(x^*) = 0$.
- Iterates $x_{k+1} = x_k + \alpha_k d_k$ generated by the descent directions $d_k = (-1)^k$ and steps $\alpha_k = 2 + 3/2^k$ from $x_0 = 2$.
- $\{x\} = \{2, -3/2, 5/4, -9/8, \ldots\}$. As $k \to \infty$, x_k will oscillate between $+1$ and -1. Thus, the sequence x_k, $k = 1, 2, 3, \ldots$ does not converge.
- $\{f\} = \{4, 9/4, 25/16, 81/64, \ldots\}$. Thus, function value decreases in each iteration. As $k \to \infty$, $f(x_k)$ will remain close to 1.
- **Key reason is small decrease in function values relative to the step length.**



Example 2: How line search methods fail !

Small Step Sizes

- The objective function $f(x) = x^2$. Global minimizer is $x^* = 0$ and optimal value of $f(x^*) = 0$.
- Iterates $x_{k+1} = x_k + \alpha_k d_k$ generated by the descent directions $d_k = -1$, $\forall k$ and steps $\alpha_k = 1/2^k$ from $x_0 = 2$.
- $\{x\} = \{2, 3/2, 5/4, 9/8, \ldots\}$. As $k \to \infty$, x_k will converge to $+1$. But, $\lim_{k\to\infty} x_k \neq x^*$.
- $\{f\} = \{4, 9/4, 25/16, 81/64, \ldots\}$. Thus, function value decreases in each iteration. As $k \to \infty$, $f(x_k)$ will remain close to 1.
- **Key reason is step sizes are too small compared to the initial rate of decrease of f.**



Sufficient Decrease Condition

Lemma

Let f be a continuously differentiable function over $\mathbb{R}^n$ and let $\mathbf{x} \in \mathbb{R}^n$. Suppose that $\mathbf{d} \in \mathbb{R}^n$ ($\mathbf{d} \neq \mathbf{0}$) is a descent direction of $\mathbf{d}$ at $\mathbf{x}$ and let $\alpha \in (0, 1)$. Then there exist $\epsilon > 0$ such that the inequality

$$f(\mathbf{x}) - f(\mathbf{x} + t\mathbf{d}) \geq -\alpha t \nabla f(\mathbf{x})^T \mathbf{d}$$

holds for all $t \in [0, \epsilon]$.

Armijo Line Search Method

- Armijo inexact line search condition stipulates that α_k should first of all give sufficient decrease in the objective function f , as measured by the following inequality:

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \leq f(\mathbf{x}_k) + c_1 \alpha_k \nabla f(\mathbf{x}_k)^T \mathbf{d}_k$$

  for some constant $c_1 \in (0, 1)$.
- Thus, the reduction in f should be proportional to both the step length α_k and the directional derivative $\nabla f(\mathbf{x}_k)\mathbf{d}_k$.