Gradient Method

SIE 449/549: Optimization for Machine Learning

Afrooz Jalilzadeh

Department of Systems and Industrial Engineering
University of Arizona

Least Square

- We are given a linear system of the form

$$Ax = b$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$

- If $m = n$ and A has a full column rank, then $x =$

- Assume the system is **overdetermined** ($m > n$), A has a full column rank (rank(A)=n)
  $\implies$ system is usually *inconsistent* (has no solution)
  $\implies$ find an approximate solution

Least Square

- The squared objective function is given by

$$f(x) = x^T A^T Ax - 2b^T Ax + \|b\|^2$$

- A is full column rank $\implies$ for any $x \in \mathbb{R}^n$, $\nabla^2 f(x) =$
- Hence, the unique stationary point

is the optimal solution of (LS). x_{LS} is the *least square solution* or the *least square estimate* of the system $Ax = b$.

Example

Example 1

Consider the inconsistent system

$$\begin{aligned} x_1 + 2x_2 &= 0 \\ 2x_1 + x_2 &= 1 \\ 3x_1 + 2x_2 &= 1 \end{aligned}$$

Example

MATLAB

```
A=[1,2;2,1;3,2];
b=[0;1;1];
A\b
```

ans =
0.5769
-0.3077

Python

```
import numpy as np
A = np.array([[1,2], [2,1], [3,2]])
b = np.array([0,1,1])
xls = np.linalg.lstsq(A,b)
print(xls[0])
```

[0.5769 -0.3077]

Gradient Method

- To solve $\min \|Ax - b\|^2$, we need to compute $(A^T A)^{-1}$
- Computing the inverse of a matrix might be computationally expensive
- Alternative Approach: Gradient Method
- Objective: Find an optimal solution of the problem

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

  where f is continuously differentiable over $\mathbb{R}^n$
- The iterative algorithms that we will consider are of the form

$$x_{k+1} = x_k + \alpha_k d_k$$

  d_k is the *direction* and α_k is the *stepsize*

Descent Direction

Definition 1

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

$$f'(x; d) = \nabla f(x)^T d < 0.$$

Lemma 2

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

$$f(x + \alpha d) < f(x)$$

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

Schematic Descent Direction Method

Algorithm 1 Schematic Descent Direction Method

Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily
for $k = 0, 1, 2, \dots$ **do**
  pick a descent direction d_k
  find a stepsize α_k satisfying $f(x_k + \alpha_k d_k) < f(x_k)$
  set $x_{k+1} = x_k + \alpha_k d_k$
  if a stopping criteria is satisfied, then STOP and x_{k+1} is the output
end for

Of course, many details are missing in the above schematic algorithm:

1. What is the starting point?
2. What is the stopping criteria?
3. How to choose the descent direction?
4. What stepsize should be taken?

Descent Direction Method

1. What is the starting point?
2. What is the stopping criteria?
3. How to choose the descent direction? Is $d_k = -\nabla f(x)$ a descent direction?

- $d_k = -\nabla f(x)$ is also the steepest direction.

Gradient Method

Lemma 3

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be a continuously differentiable function, and let* $x \in \mathbb{R}^n$ *be a non-stationary point* ($\nabla f(x) \neq 0$)*. Then* $d = -\dfrac{\nabla f(x)}{\|\nabla f(x)\|}$ *is the optimal solution of*

$$\min_{d \in \mathbb{R}^n} \{\nabla f(x)^T d : \|d\| = 1\}$$

Gradient Method

Algorithm 2 Gradient Method

Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily
for $k = 0, 1, 2, \ldots$ **do**
  find a stepsize α_k satisfying $f(x_k + \alpha_k d_k) < f(x_k)$
  set $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$
  if $\|\nabla f(x_{k+1})\| \leq \epsilon$ then STOP and x_{k+1} is the output
end for

4. What stepsize should be taken? How about constant stepsize $\alpha_k = \alpha$?

- A large constant might cause the algorithm to be nondecreasing
- A small constant can cause slow convergence of the method

Gradient Method

Example 2

Solve $\min_x f(x) = x^2$ *using the gradient method with initial point* $x_0 = 4$ *for 20 iterations with step sizes* $\alpha = 0.1$ *and* $\alpha = 10$*. What do you observe?*

Lipschitz Continuity

Definition 4

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. f is **Lipschitz** continuous if there exists $L > 0$ such that

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

- f has a **Lipschitz continuous gradient** if there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

- $f(x) = a^T x$

- $f(x) = x^T Ax$, where A is symmetric

Lipschitz Continuity

Lemma 5 (Descent Lemma)

Let function f be continuously differentiable whose gradient is Lipschitz continuous with constant L. Then, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|^2$$

Corollary 6

Let function f be continuously differentiable whose gradient is Lipschitz with constant L. Then, for $y = x - \alpha \nabla f(x)$ we have

$$f(y) \leq f(x) + \left(\frac{\alpha^2 L}{2} - \alpha \right) \|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}^n$$

Gradient Method

- In gradient method, we have $x_{k+1} = x_k - \alpha \nabla f(x_k)$, using Corollary 6, we have that:

Algorithm 3 Gradient Method

Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily
for $k = 0, 1, 2, \ldots$ **do**
  select $\alpha < 2/L$
  set $x_{k+1} = x_k - \alpha \nabla f(x_k)$
  if $\|\nabla f(x_{k+1})\| \leq \epsilon$ then STOP and x_{k+1} is the output
end for

Convergence of Gradient Method

- Let's run gradient method for T iterations

Algorithm 4 Gradient Method

Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily
for $k = 1, \dots, T$ **do**
  select $\alpha < 2/L$
  set $x_{k+1} = x_k - \alpha \nabla f(x_k)$
end for

- We show that $\|\nabla f(x_k)\| \to 0$ as $k \to \infty$
- When f is convex and $\alpha = 1/L$, gradient method has convergence rate of $\mathcal{O}(1/T)$

Lemma 7

For any $x, y, z \in \mathbb{R}^n$*:* $\langle x - y, y - z \rangle = \frac{1}{2}\left(\|x - z\|^2 - \|y - z\|^2 - \|x - y\|^2\right)$.

Convergence of Gradient Method

Theorem 8

Let function f be continuously differentiable whose gradient is Lipschitz with constant L. Also, let $\{x_k\}_{k\geq 0}$ be the sequence generated by Gradient method with step-size $\alpha < 2/L$ and initial point $x_0 \in \mathbb{R}^n$, then

$$\|\nabla f(x_k)\| \to 0, \quad \text{as } k \to \infty$$

Convergence of Gradient Method

Theorem 9

Let function f be convex and continuously differentiable whose gradient is Lipschitz with constant L. Also, let $\{x_k\}_{k\geq 0}$ be the sequence generated by Gradient method with step-size $\alpha = 1/L$, then $f(x_T) - f(x^*) \leq \frac{L}{2T}\|x_0 - x^*\|^2$, for all $T \geq 1$.