Gradient Descent: How and Why

March 8, 2025

Outline

Gradient Descent: Definition

First, the definition:

Topic: Gradient Descent

For function $f(\cdot)$, $x_c \in \text{Range}(f)$, and some learning rate $r > 0$, in practice, we can find a point x_{new}

$$x_{new} = x_c - r\nabla f(x_c),$$

such that $f(x_{new}) < f(x_c)$, where $\nabla f(x_c)$ is the gradient of $f(\cdot)$ at x_c.

We are now presenting a gentle introduction to the Gradient Descent Method in the following slides.

Preliminary: Descent direction

1. Definition: If $f(x_c + \alpha d) < f(x_c)$, $\forall 0 < \alpha < \bar{\alpha}$ given $\bar{\alpha} > 0$, then d is the descent direction.
2. Derived lemma: For gradient descent under the descent direction d, $\langle d, \nabla f(x_c) \rangle < 0$ and learning rate r, we have the equation

$$x_{new} = x_c + rd.$$

Here we present a heuristic *proof*:
Intuitively, from first order Taylor, for $0 < r < \bar{\alpha}$,

$$f(x_c + rd) \leq f(x_c) + \nabla f(x_c)^T (rd),$$

if $\langle d, \nabla f(x_c) \rangle < 0$, we immediately have

$$f(x_c + \alpha d) \leq f(x_c) + (< 0) < f(x_c).$$

Why Gradient Descent?

3. How do we find the best descent direction? By Lagrange Multiplier (KKT, Karush–Kuhn–Tucker)
We optimize the following program to find the optimal (normalized) descent direction that maximally decreases the value of the function for $x_{new} = x_c + rd$.
First, given the First-order Taylor, we have

$$\phi(r) = f(x_c + rd) = f(x_c) + r\nabla f(x_c)^T d + o(|r|),$$

the problem is hence reduced to finding the d such that $\nabla f(x_c)^T d$ is minimal.
Or formally,

$$\begin{aligned} \min \quad & \nabla f(x_c)^T d \\ \text{s.t.} \quad & ||d|| = 1 \end{aligned}$$

Why Gradient Descent?

Equivalently, the program from the last slide can be written as

$$\begin{aligned} \min \quad & \nabla f(x_c)^T d \\ \text{s.t.} \quad & d^T d = 1 \end{aligned}$$

According to KKT, we can write its dual as

$$\min \max L(d, \lambda) = \nabla f(x_c)^T d + \lambda(d^T d - 1), s.t. \lambda \geq 0$$

and the gradient of the Lagrangian

$$\nabla_d L(d, \lambda) = \nabla f(x_c) + 2\lambda d = 0,$$

$$\nabla_\lambda L(d, \theta) = d^T d - 1 = 0.$$

If the program has an optimal solution, from ∇_d, we get $d = -\frac{1}{2\lambda}\nabla f(x_c)$, and $d^T d = 1$ by ∇_λ. Also,

$$d = -\frac{\nabla f(x_c)}{||\nabla f(x_c)||}, \quad \lambda = \frac{||\nabla f(x_c)||}{2}.$$

Hence the negative gradient is the optimal descent direction.

Further Discussions

- Gauss-Newton Method
  As we discussed before, the Gradient Descent only takes care of the first-order term in Taylor expansion, if we add the second term into the Taylor expansion, we now have

  $$\phi(r) = f(x_c + rd) = f(x_c) + r\nabla f(x_c)^T d + \frac{1}{2} r^2 d^T Hessian(f(x_c))d + o(r^2).$$

  Gauss-Newton deals with this case, in which we ignore the first-order term.
- Learning Rate Selection and Update
  For simple functions $f(\cdot)$, we can plug $x_c + rd$ into the function $f(\cdot)$ and calculate the optimal r given the known vector $d = \nabla f(x_c)$ and fixed x_c; in practice, we can also find a suitable root by linear searching the minimum $x_c + rd$ given a vector for selected discretized values $r \in (0, 1)$. To validify such selection, we can use Wolfe Condition.