Deep Learning - Foundations and Concepts

Chapter 7. Gradient Descent

nonlineark@github

February 23, 2025

Outline

Gradient and stationary points

Theorem

Let the function $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at $a \in \mathbb{R}^n$:

1. Near a the function f increases fastest in the direction of $\nabla f(a) \in \mathbb{R}^n$.
2. The rate of increase in f is measured by the length of $\nabla f(a)$.
3. If f has a local extremum at a then $\nabla f(a) = 0$.

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at $a \in \mathbb{R}^n$. Then a is said to be a stationary point for f if $Df(a) = 0$, or, equivalently, $\nabla f(a) = 0$.

Gradient and stationary points

It's easy to see the correctness of the theorem. Since the rate of increase of f at the point a in an arbitray direction $v \in \mathbb{R}^n$ is given by the directional derivative at v, we have:

$$|Df(a)v| = |v^T \nabla f(a)| \leq ||\nabla f(a)||||v||$$

where we have used the Cauchy-Schwarz inequality. The rate of increase is maximal if v is a positive scalar multiple of $\nabla f(a)$. For the third claim, let's define g_j as:

$$g_j(t) = f(a_1, \ldots, a_{j-1}, t, a_{j+1}, \ldots, a_n)$$

then $g_j'(a_j) = D_j f(a)$. Since f has a local extremum at a, g_j also has a local extremum at a_j. Thus $g_j'(a_j) = 0$ for $1 \leq j \leq n$, and we have $\nabla f(a) = 0$.

Gradient and stationary points

During training, we want to optimize the weights and biases $w \in \mathbb{R}^W$ by using a chosen error function $E(w)$. From the previous theorem we see that, its smallest value will occur at a point in weight space such that:

$$\nabla E(w) = 0$$

But:

- Global minimum vs. local minimum.
- For any point w that is a local minimum, there will generally be other points in weight space that are equivalent minima (weight-space symmetries).

Local quadratic approximation

Threorem

Let U be a convex open subset of $\mathbb{R}^n$ and let $a \in U$ be a stationary point for $f \in C^2(U)$. Then we have the following assertions:

1. If $Hf(a)$ is positive definite, then f has a local strict minimum at a.
2. If $Hf(a)$ is negative definite, then f has a local strict maximum at a.

where $Hf(a)$ is the Hessian of f at a.

Local quadratic approximation

We only prove the first claim. Using Taylor expansion, we see that:

$$\begin{aligned} f(a+h) &= f(a) + h^T \nabla f(a) + \frac{1}{2} h^T H f(a) h + R_2(a,h) \\ &= f(a) + \frac{1}{2} h^T H f(a) h + R_2(a,h) \end{aligned}$$

where $\lim_{h \to 0} \frac{R_2(a,h)}{||h||^2} = 0$. Since $f \in C^2(U)$, $Hf(a)$ is a self-adjoint operator. Let λ be its smallest eigenvalue. Because $Hf(a)$ is positive definite, $\lambda > 0$. Notice that:

- $h^T H f(a) h \geq \lambda ||h||^2$.
- There is $\delta > 0$, such that $\frac{|R_2(a,h)|}{||h||^2} < \frac{\lambda}{4}$ for $||h|| < \delta$.

For $||h|| < \delta$, we have:

$$f(a+h) - f(a) = \frac{1}{2} h^T H f(a) h + R_2(a,h) > \frac{\lambda}{2} ||h||^2 - \frac{\lambda}{4} ||h||^2 = \frac{\lambda}{4} ||h||^2$$

From the previous theorem, we see that: A necessary and sufficient condition for w^* to be a local minimum of the error function $E(w)$ is that the gradient of $E(w)$ should vanish at w^* and the Hessian matrix evaluated at w^* should be positive definite.

Figure: Geometry of the error surface in the neighbourhood of a minimum w^*