Optimization

Optimization is the process of adjusting the parameters of a model to minimize the error of the model on the training data.

Minimizing a Function

- **Minimization** is the process of finding the input that results in the smallest value of the function;
- Given a function $f(x)$, the goal is to find the value of x that minimizes f.
- **Global minimum** is the smallest value of the function over its entire domain: for any $x \in \mathbb{R}^n$, $f(x^*) \leq f(x)$;
- **Local minimum** is the smallest value of the function over some small region of the domain: for any $||x - x^*|| \leq \epsilon$, $f(x^*) \leq f(x)$.

Convex Functions

- Function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if the line segment between any two points on the graph of the function lies on or above the graph:

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

- **Strictly convex** if the line segment lies strictly above the graph:

$$f(\lambda x + (1-\lambda)x') < \lambda f(x) + (1-\lambda) f(x')$$

- If f is **convex**, then any **local minimum** is also a **global minimum**;
- If f is **strictly convex**, then any **local minimum** is also the **unique global minimum**.

Gradients and Minimization

- A **gradient** represents the slope of the function in each dimension;
- The gradient points in the direction of the **greatest rate of increase of the function**;
- Given $f : \mathbb{R}^n \to \mathbb{R}$, the **gradient** of f is the vector of partial derivatives:

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \dots, \frac{\partial f(x)}{\partial x_n}\right]^T$$

- x^* is local minimizer $\implies \nabla f(x^*) = 0$.

Hessians and Convexity

- The **Hessian** is a matrix of second partial derivatives - the **gradient of the gradient**;
- The Hessian is a **measure of curvature** of the function;
- Given $f : \mathbb{R}^n \to \mathbb{R}$, the **Hessian** of f is the matrix of second partial derivatives:

$$H_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

- **Positive semidefinite** Hessian $\iff$ f is **convex**;
- **Positive definite** Hessian $\iff$ f is **strictly convex**.

More on Gradients

- Gradient of **quadratic form** $\nabla x^T A x = (A + A^T)x$;
- If A is **symmetric**, then $\nabla x^T A x = 2Ax$;
- Particular case: $f(x) = x^T x = ||x||^2$, then $\nabla f(x) = 2x$.
- If $f(x) = x^T b = b^T x$, then $\nabla f(x) = b$.
- If $g(x) = f(Ax), then \nabla g(x) = A^T \nabla f(Ax)$.
- If $g(x) = f(a \cdot x)$, then $\nabla g(x) = a \cdot \nabla f(a \cdot x)$.

Gradient Descent

Gradient descent is an iterative algorithm that starts with an initial guess x_0 and repeatedly moves in the direction of the negative gradient $\nabla f(x)$ until convergence.

$$x^{(t+1)} = x^{(t)} - \eta \nabla f(x^{(t)})$$

- η is the **step size** or **learning rate** - crucial for convergence and performance;
- **Stochastic gradient descent** is a variant of gradient descent that uses a **random sample** of the data at each iteration - a **mini-batch** - to estimate the gradient - it is **noisier**, but **faster**.