# Gradients

To understand loss minimization problem (and later to implement the gradient descent algorithm) we will often need to compute gradients of functions with **multiple** inputs and **single** outputs. Specifically, given a function $f : \mathbb{R}^d \to \mathbb{R}$ , the gradient $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is **a function** defined:

$$\nabla f(\vec{x}) = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \\ \vdots \\ \partial f / \partial x_d \end{bmatrix} .$$

So, the gradient takes in a vector $\vec{x}$ and returns a column vector of all partial derivatives of $f$ at $\vec{x}$.

When $f$ is differentiable, we must have that $\nabla f(\vec{x}) = \vec{0}$ whenever $\vec{x}$ is an extreme point (e.g. minimizer or maximizer) of $f$.

## Some Properties of Gradients

When calculating gradients for different loss functions, here are some basic properties to keep in mind:

- **Linearity**:
    - If $h(\vec{x}) = f(\vec{x}) + g(\vec{x})$, then $\nabla h(\vec{x}) = \nabla f(\vec{x}) + \nabla g(\vec{x})$.
    - If $h(\vec{x}) = f(c\vec{x})$ for some scalar $c$, then $\nabla h(\vec{x}) = c\nabla f(\vec{x})$.
- **Multi-dimensional chain rule**:
- Suppose $h : \mathbb{R}^d \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$, and $g : \mathbb{R}^d \to \mathbb{R}^n$.
- Now suppose $h(\vec{x}) = f(g(\vec{x}))$ .
- Let $g_1(\vec{x}), \ldots, g_n(\vec{x})$ denote each component of the function $g(\vec{x})$. So each $g_i(\vec{x})$ is a function from $\mathbb{R}^d \to \mathbb{R}$ and $g(\vec{x}) = [g_1(\vec{x}); \ldots; g_n(\vec{x})]$.
- Let $\partial f / \partial [g(\vec{x})]_j$ denote the $j^{\text{th}}$ partial derivative of $f$, evaluated at $g(\vec{x})$.
- The chain rule tells us that $\frac{\partial h}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial f}{\partial [g(\vec{x})]_j} \cdot \frac{\partial g_j}{\partial x_i}$

The multidimensional chain rule can seem a bit complicated when you first use it, but it's really just a generalization of what you already know from single variable calculus. See this article from Khan Academy for a more in depth review.

Roughly, the chain rule just tells us that, if a function $h$ depends on inputs $z_1, \ldots, z_n$ and each $z_i$ depends on other inputs $x_1, \ldots, x_d$, then $\frac{\partial h}{\partial x_i} = \sum \frac{\partial h}{\partial z_j} \cdot \frac{\partial z_j}{\partial x_i}$.

# Gradient Practice

Here are some examples of functions and their gradients:

- **Function**: $f(\vec{x}) = \vec{a}^T \vec{x} = \langle \vec{a}, \vec{x} \rangle$ for some fixed vector $\vec{a}$.

  **Gradient**: $\nabla f(\vec{x}) = \vec{a}$.

  - Proof: write $\vec{a}^T \vec{x} = \sum_{i=1}^{d} a_i x_i$, from which it's clear that $\frac{\partial}{\partial x_i}(\vec{a}^T \vec{x}) = a_i$.
- **Function**: $f(\vec{x}) = \|\vec{x}\|_2^2$.

  **Gradient**: $\nabla f(\vec{x}) = 2\vec{x}$.

  - Proof: write $\|\vec{x}\|_2^2 = \sum_{i=1}^{d} x_i^2$, from which it's clear that $\frac{\partial}{\partial x_i}(\|\vec{x}\|_2^2) = 2x_i$.
- **Function**: $f(\vec{x}) = g(A\vec{x})$ where $A$ is a $n \times d$ matrix and $g$ is some function from $\mathbb{R}^n \to \mathbb{R}$.

  **Gradient**: $\nabla f(\vec{x}) = A^T \nabla g(A\vec{x})$.

  - Proof: Let $k(\vec{x}) = Ax$. For $j = 1, \ldots, n$ the $j^{\text{th}}$ entry of $k(\vec{x})$ is $k_j(\vec{x}) = \langle A_j, x \rangle$, where $A_j$ is the $j^{\text{th}}$ row of $A$. From chain rule we have that $\frac{\partial f}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial g}{\partial [k(\vec{x})]_j} \cdot \frac{\partial k_j}{\partial x_i}$

  - $\frac{\partial k_j}{\partial x_i} = A_{j,i}$ where $A_{j,i}$ is the entry in $A$'s $j^{\text{th}}$ row and $i^{\text{th}}$ column.

  - Substituting we have:

  - $\frac{\partial f}{\partial x_i} = \sum_{j=1}^{n} A_{j,i} \frac{\partial g}{\partial [k(\vec{x})]_j}$ which we can obeserve is equal to:
    $\frac{\partial f}{\partial x_i} = \langle A_{:,i}, \nabla g(k(\vec{x})) \rangle = \langle A_{:,i}, \nabla g(A\vec{x}) \rangle$

    where $A_{:,i}$ denotes the $i^{\text{th}}$ column of $A$.

  - So if we stack $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d}$ into a column vector to for $\nabla f(\vec{x})$ we get $\nabla f(\vec{x}) = A^T \nabla g(A\vec{x})$
    .

**This last one is a good one to just memorize! It will come up again and again!**

# Application to Multiple Linear Regression Sqaured Loss

Now that we have some basic identities, let's try to compute the gradient of the following function from $\mathbb{R}^d \to \mathbb{R}$:

$L(\vec{\beta}) = \|\vec{y} - X\vec{\beta}\|_2^2$.

Here $\vec{y}$ is a length $n$ column vector, $X$ is our $n \times d$ data matrix, $\beta$ is a column vector of $d$ parameters and $L$ is the squared loss.

**Question:** What the gradient $\nabla L(\vec{\beta})$?

**Solution:**

First note that

$$L(\vec{\beta}) = \|\vec{y} - X\vec{\beta}\|_2^2 = \langle \vec{y} - X\vec{\beta}, \vec{y} - X\vec{\beta} \rangle = \langle \vec{y}, \vec{y} \rangle + \langle X\vec{\beta}, X\vec{\beta} \rangle - 2\langle \vec{y}, X\vec{\beta} \rangle.$$

So, by **linearity**,

$$\nabla L(\vec{\beta}) = \nabla \langle \vec{y}, \vec{y} \rangle + \nabla \langle X\vec{\beta}, X\vec{\beta} \rangle - 2\nabla \langle \vec{y}, X\vec{\beta} \rangle.$$

Let's figure out each term seperately:

- $\nabla \langle \vec{y}, \vec{y} \rangle = \vec{0}$ because $\langle \vec{y}, \vec{y} \rangle$ does not depend oon $\beta$ at all (which is what we're computing partial derivatives with respect to).

- $\nabla \langle X\vec{\beta}, X\vec{\beta} \rangle = \nabla \|X\vec{\beta}\|_2^2$. We can evaluate this gradient using the first and last example in our gradient practice section: it's equal to $\|X\vec{\beta}\|_2^2 = X^T \nabla \|\vec{z}\|_2^2$ where $\vec{z} = X\vec{\beta}$.

  So we have $\|X\vec{\beta}\|_2^2 = X^T(2\vec{z}) = 2X^T X\vec{\beta}$.

- Finally, we note that $\langle \vec{y}, X\vec{\beta} \rangle = \vec{y}^T X\beta = \langle X^T \vec{y}, \beta \rangle$ (here I'm using that $(\vec{y}^T X)^T = X^T \vec{y}$).

  So $\nabla \langle \vec{y}, X\vec{\beta} \rangle = \nabla \langle X^T \vec{y}, \beta \rangle = X^T \vec{y}$ using example 1 from the previous section.

Putting it all together, we get that

$$\nabla L(\vec{\beta}) = 0 + 2X^T X\vec{\beta} - 2X^T \vec{y}$$

$$\nabla L(\vec{\beta}) = 2X^T(X\vec{\beta} - \vec{y})$$