Chapter 0:
Notation & Basics

In this chapter we introduce some basic concepts and notation necessary for Machine Learning.

0.1 Notation

- Vectors $\mathbf{v} \in \mathbb{R}^d$ are denoted by bold letters whereas scalars $s \in \mathbb{R}$ are denoted by normal letters.

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_d \end{pmatrix}$$

- Denote matrices $A \in \mathbb{R}^{m \times n}$ by normal letters.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & & \ddots & \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{pmatrix}$$

- We define $\mathbf{0}$ (resp. $\mathbf{1}$) as the vector full of zeros (resp. full of ones) with appropriate dimension to the context.

$$\mathbf{0} := \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{1} := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

- If $\mathbf{v} \in \mathbb{R}^d$, then we define $\mathbf{v}^T \in \mathbb{R}^{(1 \times d)}$ as its transpose

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_d \end{pmatrix}, \mathbf{v}^T := \begin{pmatrix} v_1, \dots, v_d \end{pmatrix}$$

  and if $A \in \mathbb{R}^{m \times n}$, then we define $A^T \in \mathbb{R}^{n \times m}$ as its transpose.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & & \ddots & \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{pmatrix}, A^T := \begin{pmatrix} a_{1,1} & a_{2,1} & \dots & a_{m,1} \\ a_{1,2} & a_{2,2} & \dots & a_{m,2} \\ \vdots & & \ddots & \\ a_{1,n} & a_{2,n} & \dots & a_{m,n} \end{pmatrix}$$

- The scalar product of two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ is defined by

$$\langle \mathbf{v}, \mathbf{w} \rangle := \mathbf{v}^\top \mathbf{w} = \sum_{i=1}^{d} v_i w_i.$$

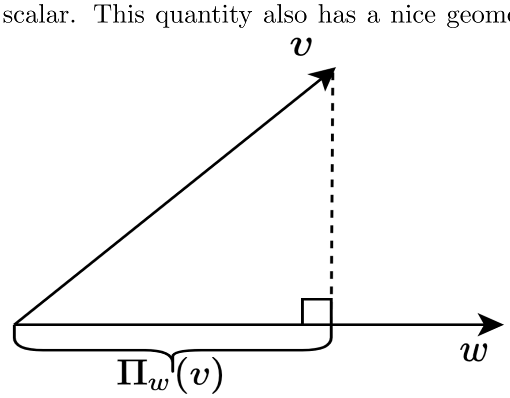
- The norm of a vector $\mathbf{v} \in \mathbb{R}^d$ is denoted by

$$\|\mathbf{v}\| := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{\mathbf{v}^T \mathbf{v}} = \sqrt{\sum_{i=1}^{d} v_i^2}.$$

- Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, then denote $v \leq w :\iff \forall i = 1 \ldots d : v_i \leq w_i$.
- For a set S we denote the cardinality of S by $|S|$.

0.2 Scalar Projection

Definition: Scalar Projection
Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ with $\mathbf{w} \neq \mathbf{0}$. The scalar projection of $\mathbf{v}$ onto $\mathbf{w}$ is defined as $\Pi_w(v) := \dfrac{\mathbf{w}^T \mathbf{v}}{\|\mathbf{w}\|}$.We observe that the scalar projection is not a vector, but a scalar. This quantity also has a nice geometric interpretation as seen below[1].



0.3 Hyperplanes

Definition: Affine Linear Function
An (affine-)linear function is a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^d (\mathbf{w} \neq \mathbf{0})$ and $b \in \mathbb{R}$.

Example:

$$\begin{aligned} f : \mathbb{R}^2 &\to \mathbb{R} \\ (x_1, x_2) &\mapsto 3x_1 + 4x_2 - 7 \end{aligned}$$

is affine linear as it is of the form $f(x) = \begin{pmatrix} 3 \\ 4 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 7$

[1] If you are looking for a proof `https://www.youtube.com/watch?v=LyGKycYT2v0`

Definition: Hyperplane
A hyperplane is a subset $H \subset \mathbb{R}^d$ *defined as* $H := \left\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 0\right\}$, *where* f *is affine linear.*

Let H be a hyperplane defined by the affine-linear function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$.

Prop 0.1: *The vector* $\mathbf{w}$ *is orthogonal to* H, *meaning that:* $\forall \mathbf{x}_1, \mathbf{x}_2 \in H$ *it holds* $\mathbf{w}^\top (\mathbf{x}_1 - \mathbf{x}_2) = 0$.

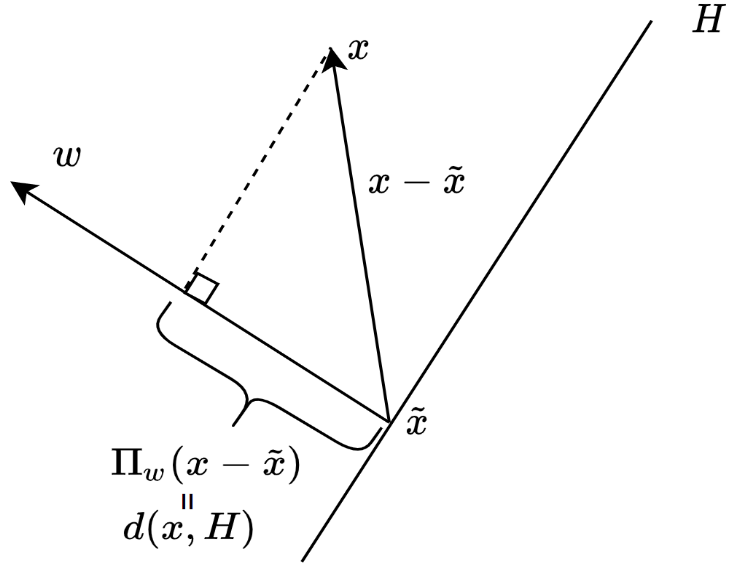
Proof :
Let $\mathbf{x}_1, \mathbf{x}_2 \in H = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^\top \mathbf{x} + b = 0\right\}$.
Then $\mathbf{w}^T \mathbf{x}_1 + b = 0$ and $\mathbf{w}^T \mathbf{x}_2 + b = 0$.
Thus $\mathbf{w}^T \mathbf{x}_1 + b = \mathbf{w}^T \mathbf{x}_2 + b$, or equivalently $\mathbf{w}^T (\mathbf{x}_1 - \mathbf{x}_2) = 0$. $\square$

Prop 0.2: *The signed distance of a point* $\mathbf{x}$ *to* H *is given by*
$d(\mathbf{x}, H) \overset{def.}{:=} \operatorname{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right) \min_{\tilde{x} \in H} \|x - \tilde{x}\| \overset{!}{=} \dfrac{1}{\|\mathbf{w}\|} \left(\mathbf{w}^\top \mathbf{x} + b\right)$

Proof :
Let $\tilde{\mathbf{x}}$ be an arbitrary element of H. By our previous proof, we know that $\mathbf{w}$ is orthogonal to our hyperplane. First we notice from the picture below that

$$\forall \tilde{\mathbf{x}} \in H : \Pi_{\mathbf{w}}(\mathbf{x} - \tilde{\mathbf{x}}) = \operatorname{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right) \min_{\tilde{\mathbf{x}} \in H} \|\mathbf{x} - \tilde{\mathbf{x}}\|$$

where $\operatorname{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right)$ shows on which side of the hyperplane $\mathbf{x}$ lies.



So $d(\mathbf{x}, H) = \Pi_w(\mathbf{x} - \tilde{\mathbf{x}}) \overset{\text{def.}}{=} \dfrac{\mathbf{w}^T(\mathbf{x} - \tilde{\mathbf{x}})}{\|\mathbf{w}\|} = \dfrac{\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\tilde{\mathbf{x}}}{\|\mathbf{w}\|} \overset{(\star)}{=} \dfrac{\mathbf{w}^\top\mathbf{x} + b}{\|\mathbf{w}\|}.$

Where in $(\star)$ we use: $\tilde{\mathbf{x}} \in H \Rightarrow \mathbf{w}^T\tilde{\mathbf{x}} + b = 0 \Longleftrightarrow -\mathbf{w}^T\tilde{\mathbf{x}} = +b$. $\square$

0.4 Eigenvalues

Definition: Eigenvalue

Let $A \in \mathbb{R}^{d\times d}$. $\lambda \in \mathbb{R}$ is called an eigenvalue of A if there is a vector $\mathbf{x} \in \mathbb{R}^d\backslash\{0\}$ such that $A\mathbf{x} = \lambda\mathbf{x}$. In that case $\mathbf{x}$ is an eigenvector corresponding to the eigenvalue λ. The set

$$\operatorname{Eig}(A, \lambda) := \left\{\mathbf{x} \in \mathbb{R}^d : A\mathbf{x} = \lambda\mathbf{x}\right\}$$

is called the eigenspace corresponding to the eigenvalue λ.

Intuitively an eigenvector preserves its direction under the linear transformation A but not necessarily its magnitude.

Example: Eigenvalue Calculation

$$B = \begin{pmatrix} -6 & 3 \\ 4 & 5 \end{pmatrix}$$

Let us try to find the eigenvalues of the matrix B. Let I denote the Identity matrix.
For $\lambda \in \mathbb{R}$ and $A \in \mathbb{R}^{d\times d}$ it holds:

$$\begin{aligned}
&\lambda\,\text{Eigenvalue of } A \Leftrightarrow \exists\,\mathbf{x} \neq 0 \in \mathbb{R}^d\,\text{with}: \; A\mathbf{x} = \lambda\mathbf{x} \\
&\Leftrightarrow \exists\,\mathbf{x} \neq 0 \in \mathbb{R}^d\,\text{with}: \; A\mathbf{x} = \lambda I\mathbf{x} \\
&\Leftrightarrow \exists\,\mathbf{x} \neq 0 \in \mathbb{R}^d\,\text{with}: \; \lambda I\mathbf{x} - A\mathbf{x} = 0 \\
&\Leftrightarrow \dim\operatorname{Ker}(\lambda I - A) > 0 \\
&\Leftrightarrow \dim\operatorname{Im}(\lambda I - A) < d \\
&\Leftrightarrow \lambda I - A\,\text{not invertible} \\
&\Leftrightarrow \det(\lambda I - A) = 0
\end{aligned}$$

Let's use this insight to calculate the eigenvalues of B.

$$\det\begin{pmatrix} \lambda + 6 & 3 \\ 4 & \lambda - 5 \end{pmatrix} = 0 \iff (\lambda{+}6)(\lambda{-}5){-}12 = 0 \iff \lambda^2{+}\lambda{-}42 = 0 \iff (\lambda{+}7)(\lambda{-}6) = 0$$

Apparently the eigenvalues of A are -7 and 6. We can also find the corresponding eigenvectors by resubstituting these values back into $A\mathbf{x} = \lambda\mathbf{x}$.

0.5 Positive definite matrices

Definition: Positive definite matrix

Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix ($A^T = A$). We call
A positive definite $:\Longleftrightarrow \forall\mathbf{x} \neq 0 \in \mathbb{R}^d: \; \mathbf{x}^T A\mathbf{x} > 0.$
A positive semi definite $:\Longleftrightarrow \forall\mathbf{x} \neq 0 \in \mathbb{R}^d: \; \mathbf{x}^T A\mathbf{x} \geq 0.$

0.6 Gradient

Definition: Gradient
Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. We define the gradient by:

$$\nabla f = \operatorname{grad} f =: \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)^T$$

We observe that ∇f is again a function. Each component of the gradient tells us how fast our function is changing in each direction.
To see how fast the change is at a point p at direction v, we would multiply $\nabla f(p) \cdot \mathbf{v}$.
Observe that this scalar product is maximized if $\mathbf{v}$ is parallel to $\nabla f(p)$ which shows that ∇f shows in the direction of the steepest ascent.

0.7 Hessian Matrix

Definition: Hessian Matrix Let $f : \mathbb{R}^d \to \mathbb{R}$. If all second partial derivatives of f exist and are continuous over the domain of the function, then the Hessian matrix H is defined by:

$$\mathrm{H}_f(x) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{i,j=1,\ldots,d} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix}$$

Example: Hessian Matrix
Let $f : \mathbb{R}^2 \to \mathbb{R}, f(x, y) = x^3 + y^3 - 3xy$.
Let us try to calculate the Hessian Matrix. First we need to find the partial derivatives.
We have :

$$\frac{\partial f}{\partial x}(x, y) = 3x^2 - 3y$$
$$\frac{\partial f}{\partial y}(x, y) = 3y^2 - 3x$$

We know that

$$\frac{\partial^2 f}{\partial x \partial x} = \frac{\partial^2 f}{\partial x \partial x} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = 6x$$
$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = -3$$
$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = -3$$
$$\frac{\partial^2 f}{\partial y \partial y} = \frac{\partial^2 f}{\partial y \partial y} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = 6y$$

$$\text{So } H = \begin{pmatrix} 6x & -3 \\ -3 & 6y \end{pmatrix}.$$

Properties of Hessian Matrix

- The Hessian Matrix of a convex function is positive semi definite.
- If the Hessian is positive-definite at $\mathbf{x}$, then f attains an isolated local minimum at $\mathbf{x}$.
- If the Hessian is negative-definite at x, then f attains an isolated local maximum at $\mathbf{x}$.