# MATHvsMACHINE: THE BOOK

LOUIS DE THANHOFFER DE VOLCSEY

**Abstract**

In this book, we propose a new way of studying of machine learning, by providing a new -mathematically rigourous- definition to the foundational concepts of machine learning and showing how the known techniques and results fit inside this framework.

---

## Contents



---

## 0 Overview

To begin, we'll discuss our definition of supervised learners and describe how linear regression satisfies our definition

## 1 Supervised Learners

We begin our study of machine learning by suggesting a definition for supervised learners.
Recall that supervised learning (roughly) corresponds to the following paradigm: one is given *data* which consists of *features* and their corresponding *labels*. A supervised learner now assigns to any new feature a new label in a way that respects the given data as well as possible.

Let's look at this idea in a little more detail:
To begin, we denote the sets of features and labels by $\mathfrak{X}$ and $\mathfrak{y}$ respectively. We'll also define the

*dataspace* $\mathfrak{D}$, which consists of all possible dataset, each of which is a finite subset of $\mathfrak{X} \times \mathfrak{y}$ (the given features with their assigned labels). Third, we introduce a *hypothesis space* $\mathfrak{H} \subset \mathfrak{y}^{\mathfrak{X}}$ which contains all the possible ways of assigning a label to a new feature.
A supervised learner now assigns a choice a hypothesis given a dataset in an *optimal* way. In other words, we wish to construct an assignment from the dataspace to the hypothesis space

$$h : \mathfrak{D} \longrightarrow \mathfrak{H} : \Delta \mapsto h_\Delta$$

such that $h_\Delta$ *fits* the data $\Delta$ optimally.
To formalize this optimality condition, we introduce a cost function which assigns a real number given any choice of data and hypothesis:

$$c : \mathfrak{D} \times \mathfrak{H} \longrightarrow \mathbb{R} : (\Delta, f) \mapsto c(\Delta, f)$$

The idea that the choice $h_\Delta$ of hypothesis fits the data best is now translated into $h_\Delta$ minimizing the cost:

$$c(\Delta, h_\Delta) = \min_{f \in \mathfrak{H}} c(\Delta, f) \tag{1}$$

Leading us to the following definition:

**Definition 1.0.1.** A supervised learner $\mathfrak{L}$ (or simply learner) is a tuple $(\mathfrak{X}, \mathfrak{y}, \mathfrak{D}, \mathfrak{H}, c, h)$ where $\mathfrak{D}$ consists of finite subsets of $\mathfrak{X} \times \mathfrak{y}$, $\mathfrak{H} \subset \mathfrak{y}^{\mathfrak{X}}$ and $h : \mathfrak{H} \longrightarrow \mathfrak{D}$, $c : \mathfrak{D} \times \mathfrak{H} \longrightarrow \mathbb{R}$ are functions satisfying the learning condition

$$c(\Delta, h_\Delta) = \min_{f \in \mathfrak{H}} c(\Delta, f)$$

We say that $\mathfrak{L}$ has *learned* the hypothesis $h_\Delta$ from the data $\Delta$.
The functions $c_\Delta \overset{\text{def}}{=} c(\Delta, -) : \mathfrak{H} \longrightarrow \mathbb{R}$ are the *cost functions* of $\mathfrak{L}$

To make our exposition a little more transparent we'll introduce a bit of notation:

**Notation 1.** For any dataset $\Delta \subset \mathfrak{X} \times \mathfrak{y}$, we definte $\Delta_{\mathfrak{X}} \overset{\text{def}}{=} \left\{(x)_{(x,y)\in\Delta}\right\}$ and define $\Delta_{\mathfrak{y}}$ similarly.
For a hypothesis $f \in \mathfrak{H}$, we let $f(\Delta_{\mathfrak{X}}) = \left\{f(x)_{(x,y)\in\Delta}\right\}$

In the rest of our discussion, we will encounter a few interesting properties that a learner can possess:
For one, it will be convenient to study learners where the cost function really only depends on the values of the hypothesis on the features of the dataset as follows:

**Definition 1.0.2.** A learner $\mathfrak{L}$ is *regular* if for any hypotheses $f, g \in \mathfrak{H}$ and any dataset $\Delta \in \mathfrak{D}$

$$g(\Delta_{\mathfrak{X}}) = f(\Delta_{\mathfrak{X}}) \implies c(\Delta, f) = c(\Delta, g)$$

Additionally, it will be convenient to give a specific name to learners where the function $h_\Delta$ is unique:

**Definition 1.0.3.** We say that a learner is *sharp* if $h_\Delta = \arg\min_{f \in \mathfrak{H}} c(\Delta, f)$ for any dataset $\Delta \in \mathfrak{D}$.
In particular $h_\Delta$ is fully determined by the cost function $c$

The first few chapters in this book are dedicated to proving how some of the main learning algorithms that are being used today can indeed be interpreted in the context of Definition 1.0.1. Along the way we shall give a clean interpretation of some of these algorithms and build on the theory of learners just described...

## 2 Regression: Linear Learners

We begin our study of learners with one of the most ubiquitous learning algorithms: *linear regression*.
in this context, we endow the label space $\mathfrak{y}$ with the structure of a finite-dimensional inner product space. $\mathfrak{y}$ is thus equipped with a norm in particular. Before we continue our study of linear regression we recall the following standard consructions of inner product spaces:

**Lemma 2.0.1.** *Let $V$ and $W$ be inner product spaces with orthonormal bases $\big(v_1, \dots v_m\big) \in V$ and $\big(w_1, \dots, w_n\big) \subset W$. Then*

1. *$V^* \overset{\text{def}}{=} \text{Hom}_{\mathbb{R}}(V, \mathbb{R})$ is an inner product space with $\langle f\,, g\,\rangle \overset{\text{def}}{=} \langle v\,, v'\,\rangle$ if $\langle v\,, -\,\rangle = f$ and $\langle -\,, v'\,\rangle = g$ and orthonormal basis given by the maps $\Big(\langle v_1\,, -\,\rangle, \dots, \langle v_m\,, -\,\rangle\Big)$*
2. *$V \times W$ is an inner product space with $\langle (v, w)\,, (v', w')\,\rangle \overset{\text{def}}{=} \langle v\,, w\,\rangle + \langle v'\,, w'\,\rangle$ and orthonormal basis $\Big((v_i, w_j)\Big)_{i,j}$*
3. *$V \otimes W$ is an inner product space with $\langle v \otimes w\,, v' \otimes w\,\rangle \overset{\text{def}}{=} \langle v\,, w\,\rangle \cdot \langle v'\,, w'\,\rangle$ and orthonormal basis $\Big((v_i \otimes w_j)\Big)_{i,j}$*
4. *$\text{Hom}(V, W)$ is an inner product space with an orthonormal basis given by maps $e_{i,j}$ defined as*

$$e_{i,j}(v_k) = \begin{cases} 0, & \text{if } k \neq i \\ w_j, & \text{otherwise} \end{cases} \tag{2}$$

*Proof.* Items (1) through (3) consist of routine calculations. To show item (4), recall that the function

$$\iota : V^* \otimes W \longrightarrow \text{Hom}_{\mathbb{R}}(V, W)$$

which assigns to $f \otimes w$ the linear map $\iota_{f \otimes w}(v) \overset{\text{def}}{=} f(v) \cdot w$ is an isomorphism. Now, items (1) and (3) together imply that $V^* \otimes W$ is indeed an inner product space with orthonormal basis $\big\{\langle v_i\,, -\,\rangle \otimes w_j\big\}_{i,j}$. Now simply note that the map associated to $\langle v_i\,, -\,\rangle \otimes w_j$ under the function $\iota$ is exactly $e_{i,j}$ ∎

Returning to our discussion above, we consider a finite-dimensional inner product space $\mathfrak{y}$ of labels and $\mathfrak{X}$ any set of features. Now, for any finite dataset $\Delta \subset \mathfrak{X} \times \mathfrak{y}$, the space $\mathfrak{y}^{\Delta}$ in turn carries an inner product by Lemma 2.0.1, so that for any map $f \in \mathfrak{y}^{\mathfrak{X}}$, we can define a cost as follows:

$$c(\Delta, f) = ||y - f(x)||_{\mathfrak{y}^{\Delta}} \overset{\text{def}}{=} \sqrt{\sum_{\Delta} ||y - f(x)||^2}$$

The main result of this chapter will be to show that this cost indeed describes a learner under the right conditions. To this end, we will make the following definition:

**Definition 2.0.2.** Let $\mathfrak{H} \subset \mathfrak{y}^{\mathfrak{X}}$ and $\Delta \subset \mathfrak{X} \times \mathfrak{y}$. Then we say that $\Delta$ separates $\mathfrak{H}$ if for any $f, g \in \mathfrak{H}$:

$$f|_{\Delta_{\mathfrak{X}}} = g|_{\Delta_{\mathfrak{X}}} \implies f = g$$

The main result of this chapter is:

**Theorem 2.0.3.** *Let $\mathfrak{y}$ be a finite-dimensional inner product space, $\mathfrak{X}$ any set and let $\mathfrak{H} \subset \mathfrak{y}^{\mathfrak{X}}$ be a finite-dimensional subspace of $\mathfrak{y}^{\mathfrak{X}}$. Assume that any $\Delta \in \mathfrak{D}$ separates $\mathfrak{H}$ and let $c(\Delta, f) = ||y - f(x)||_{\mathfrak{y}^{\Delta}}$. Then $(\mathfrak{X}, \mathfrak{y}, \mathfrak{D}, \mathfrak{H}, c)$ defines a sharp learner (as in Definition 1.0.3)*

**Definition 2.0.4.** We say that a learner is linear if it is of the above form

One particular type of linear learner will be of particular interest, as it allows us to be a little more explicit with certain constructions: if we assume that $\mathfrak{X}$ itself carries the structure of a finite-dimensional inner product space, and put $\mathfrak{H} \stackrel{\text{def}}{=} \text{Hom}_{\mathbb{R}}(\mathfrak{X}, \mathfrak{y})$ as well as consider finite datasets $\Delta \subset \mathfrak{X} \times \mathfrak{y}$ such that $\Delta_{\mathfrak{X}}$ spans the whole of $\mathfrak{X}$ (where we recall our use of the notation 1), then it's easy to see that the conditions of Theorem 2.0.3 are satisfied so that we indeed obtain an example of a linear learner:

**Definition 2.0.5.** A *Euclidean learner* is a sharp (linear) learner where $\mathfrak{X}, \mathfrak{y}$ are finite-dimensional inner product spaces,

$$\mathfrak{D} = \left\{ \Delta \subset \mathfrak{X} \times \mathfrak{y} \;\middle|\; |\Delta| \neq \infty, \text{ and } \text{span}(\Delta_{\mathfrak{X}}) = \mathfrak{X} \right\}, \; \mathfrak{H} = \text{Hom}_{\mathbb{R}}(\mathfrak{X}, \mathfrak{y})$$

and

$$c : \mathfrak{D} \times \mathfrak{H} \longrightarrow \mathbb{R} : (\Delta, f) \mapsto ||\big(y - f(x)\big)_{(x,y)\in\Delta}||_{\mathfrak{y}^{\Delta}}$$

## 2.1. Projections and the Normal equation

We will prove Theorem 2.0.3 by showing that the learned hypothesis $h_{\Delta}$ can be constructed as the projection onto the image of a certain linear map $\text{ev}_{\Delta} : \mathfrak{H} \longrightarrow \mathfrak{y}^{\Delta}$. It makes sense to refresh the reader on projections onto images of linear maps: The starting point of this construction begin the following lemma:

**Lemma 2.1.1.** *Let $W \subset V$ be a subspace of the finite-dimensional inner product space $V$. Let $v \in V$ and $w \in W$ Then the following are equivalent:*

1. $(v - w) \perp W$
2. $\min_{u \in W} ||v - u|| = ||v - w||$

*Proof.* Assume that $w$ satisfies the first condition and let $u \in W$. Then $v - w \perp w - u$, and by the Pythagorean theorem, we have:

$$||v - u||^2 = ||u - w||^2 + ||w - u||^2 \geq ||v - w||^2$$

Proving the second condition.
The converse can be proved through a fun trick: consider the function:

$$\phi : \mathbb{R} \longrightarrow \mathbb{R} : t \mapsto ||v - w + tu||^2$$

Since $||v-w||^2$ is minimal, $\phi$ has a minimum at $t = 0$. Moreover, $\phi$ is differentiable so that $\phi'(0) = 0$. It's also easy to see that

$$\phi(t) = ||v - w||^2 + 2 \cdot t \langle v - w \,, u \rangle + t^2 ||u||^2$$

Hence $0 = \phi'(0) = 2 \langle v - w \,, u \rangle$, and $v - w \perp u$ as required ∎

**Lemma 2.1.2.** *Let $W \subset V$ be a subspace of a finite-dimnsional inner product space. Then there exists a unique linear map $\pi : V \longrightarrow W$ where $\pi(v)$ is the unique vector in $W$ satisfying*

$$\min_{w \in W} ||v - w|| = ||v - \pi(v)||$$

*Proof.* By the above lemma we need to show that for any $v \in V$, there exists a unique $\pi(v) \in W$ such that $v - \pi(v) \perp W$. To this end, we look at the following map $f \in W^*$

$$f : W \longrightarrow \mathbb{R} : w \mapsto \langle v \,, w \rangle$$

Since $W^*$ is in turn an inner product space, $f$ can be written in the form $\langle \pi(v) \,, - \rangle$ for some unique $\pi(v) \in W$. The result now follows ∎

The lemma above motivates the following definition:

**Definition 2.1.3.** Let $W \subset V$ be a subspace of a finite dimensional inner product space. Then the unique map $\pi : V \longrightarrow W$ defined by

$$||v - \pi(v)|| = \min_{w \in W} ||v - w||$$

is *the projection of $V$ onto $W$*

as mentioned, In the context of linear regression, we will be interested in projections onto subspaces that are can be described as the image of a linear maps. In this setting, one can give a more explicit description of the projection $\pi$:

**Lemma 2.1.4.** *Let $f \in \mathrm{Hom}(V, W)$ and and $w \in W$ Then the following are equivalent:*

1. *$f(v)$ is the projection of $w$ onto the subspace $\mathrm{im}(f) \subset W$*
2. *The vector $v \in V$ satisfies $(f^* \circ f)(v) = f^* w$*

*Proof.* $f(v)$ is the projection of $w$ onto $\mathrm{im}(f)$ if and only if $\langle w - f(v), f(v') \rangle$ for any $v' \in V$ by Lemma 2.1.1. Now,

$$\langle w - f(v), f(v') \rangle = \langle f^*(w) - f^* f(v), v' \rangle$$

This last expression is 0 if and only if $f^*(w) - f^* f(v) = 0$ since the inner product is nondegenerate ∎

The above lemma justifies the following definition:

**Definition 2.1.5.** Let $f \in \mathrm{Hom}(V, W)$ and $w \in W$.
We say that $v \in V$ satisfies the normal equation if and only if

$$(f^* \circ f)(v) = f^* w$$

In this terminology, we can restate lemma 2.1.4 as follows:

**Lemma 2.1.6.** *Let $f \in \mathrm{Hom}(V, W)$ and $w \in W$ Then the following are equivalent:*

1. *$||w - f(v)|| = \min_{u \in V} ||w - f(u)||$*
2. *$v$ is a solution to the normal equation $(f^* \circ f)(v) = f^* w$*

It is finding the solutions finding the solutions to this equation that we are interested in. It turns out that one can give an explicit description of them using the so-called *Moore-Penrose pseudo-inverse*. Since this construction seems to be a little less well covered in standard linear algebra literature, we'll discuss in detail below:

## 2.2. the (Moore-Penrose) Pseudo-inverse

For the rest of this section, we will let $V, W$ be finite-dimensional vector spaces and $f \in \mathrm{Hom}_{\mathbb{R}}(V, W)$.
It is well-known that $f$ does not have an inverse in general. There is however a natural generalization of the notion of inverse which can be defined for *any* map: a *pseudo-inverse*. More precisely, if $f$ either has a nonzero kernel or if the image of $f$ is not the whole of $W$, then the inverse of $f$ will not exist. One natural way to remediate this issue is to consider complements for both subspaces and write

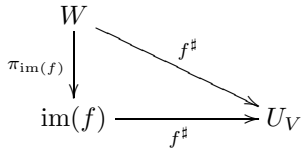
$$V \overset{\text{def}}{=} \ker(f) \oplus U_V \text{ and } W \overset{\text{def}}{=} \mathrm{im}(f) \oplus U_W$$

It's easy to see that restricting $f$ to appropriate subspaces now does produce an invertible map as follows:

**Lemma 2.2.1.** *the map $f : U_V \longrightarrow \mathrm{im}(f)$ is an isomorphism.*

We'll denote the inverse of $f$ on $U_V$ by $f^\sharp : \mathrm{im}(f) \longrightarrow U_V$. A pseudo-inverse is now the natural lift of $f^\sharp$ to the whole of $W$:

**Lemma 2.2.2.** *There exists a unique map $f^\sharp : W \longrightarrow U_V$ making the following diagram commute:*




*Proof.* The commutativity of the diagram means that for $u \in U_V$, we have

$$f^\sharp(w) \stackrel{\mathrm{def}}{=} u \iff f^\sharp(\pi_{\mathrm{im}(f)}(w)) = u \iff \pi_{\mathrm{im}(f)}(w) = f(u)$$

Where the second equivalence follows from the fact that $f^\sharp$ is the inver of $f$ on $U_V$.
The claim will thus follow if we show that the above assignment is indeed a well-defined linear map. To this end assume that $u, u' \in U_V$ satisfy $f(u') = \pi_{\mathrm{im}(f)}(w) = f(u)$.
Then $u - u' \in \ker(f)$, hence $u - u' \in \ker(f) \cap U_V$ in particular. Now since $\ker(f) \oplus U_V = V$, we have $u - u' = 0$, so that $u = u'$, showing the well-definedness.
We leave the linearity to the reader. ∎

It will be helpful to note that the map $f^\sharp \in \mathrm{Hom}(W, V)$ can also be characterized by $\mathrm{im}(f^\sharp) \subset U_V$ and $f \circ f^\sharp = \pi_{\mathrm{im}(f)}$.
To give the map $f^\sharp$ a name, we first let $\Lambda(f)$ denote the set

$$\Lambda(f) \stackrel{\mathrm{def}}{=} \{(U_V, U_W) |\ \ker(f) \oplus U_V = V \text{ and } \mathrm{im}(f) \oplus U_W = W\}$$

and conclude from Lemma 2.2.2 that there is a assignment:

$$\Phi : \Lambda(f) \longrightarrow \mathrm{Hom}_{\mathbb{R}}(W, V) : (U_V, U_W) \mapsto f^\sharp$$

where $f^\sharp \in \mathrm{Hom}_{\mathbb{R}}(W, V)$ is the unique map satisfying

$$f \circ f^\sharp = \pi_{\mathrm{im}(f)} \text{ and } \mathrm{im}(f^\sharp) \subset U_V$$

Let's denote the image of $\Phi$ by $\Pi(f)$. Summarizing the discussion, we make the following:

**Definition 2.2.3.** Let $(U_V, U_W) \in \Lambda(f)$. Then the pseudo-inverse of $(U_V, U_W, f)$ is the map $\Phi(f)$.
We say that $g \in \mathrm{Hom}_{\mathbb{R}}(W, V)$ is a pseudo-inverse to $f$ if $g \in \Pi(f)$

We can give a slightly different description of pseudo-inverses by describing them on the 2 components in the decomposition $\mathrm{im}(f) \oplus U_W = W$:

**Lemma 2.2.4.** *Let $(U_V, U_W)$ in $\Lambda(f)$. Then the following are equivalent:*

1. *$f^\sharp$ is the pseudo-inverse to $(U_V, U_W, f)$*
2. *$f^\sharp|_{\mathrm{im}(f)}$ is the inverse to $f : U_V \longrightarrow \mathrm{im}(f)$ and $f^\sharp|_{U_W} = 0$*

*Proof.* Since the pseudo=inverse to $(U_V, U_W, f)$ is unique, it suffices to show that the pseudo-inverse indeed satisfies the conditions of (2). The fact that $f^\sharp|_{\mathrm{im}(f)}$ is the inverse of $f|_{U_V}$ follows from

$$(f \circ f^\sharp)|_{\mathrm{im}(f)} = \big(\pi_{\mathrm{im}(f)}\big)|_{\mathrm{im}(f)} = \mathrm{Id}\,|_{\mathrm{im}(f)}$$

Moreover, if $w \in U_W$, then $\pi_{\mathrm{im}(f)}(w) = 0$ since $\mathrm{im}(f) \oplus U_W$. Hence $f^\sharp(w) = f^\sharp(\pi_{\mathrm{im}(f)}(w)) = 0$ by Lemma 2.2.2 ∎

Our next order of business is to give an explicit description of the set $\Pi(f)$ of pseudo-inverses to $f$. We begin by showing that we can describe the complements $U_V$ and $U_W$ solely by using the maps $f$ and $f^\sharp$:

**Lemma 2.2.5.** *Let $f^\sharp$ be the pseudo-inverse to $(U_V, U_W, f)$. Then $U_V = \mathrm{im}(f^\sharp)$ and $U_W = \ker(f^\sharp)$*

*Proof.* We have $\mathrm{im}(f^\sharp) \subset U_V$ by Definition 2.2.3 . Moreover, $f^\sharp$ is a composition of surjections and hence itself surjective, proving the first claim.
To prove the second claim, note that the second condition of Lemma 2.2.4 immediately implies that $U_W \subset \ker(f^\sharp)$. We can also show the other inclusion by assuming that $w \in W$ satisfies $f^\sharp(w) = 0$, in which case $\pi_{\mathrm{im}(f)}(w) = f(f^\sharp(w)) = f(0) = 0$, implying that $w$ lies in the component $U_W$ of the decomposition $\mathrm{im}(f) \oplus U_W = W$ as required ∎

Taking the above lemma one step further allows us to describe the set $\Pi(f)$ of pseudo-inerses as promised:

**Lemma 2.2.6.** *Let $f \in \mathrm{Hom}(V, W)$. Then the following are equivalent:*

1. $g \in \Pi(f)$
2. $(f \circ g)|_{\mathrm{im}(f)} = \mathrm{Id}$ *and* $(g \circ f)|_{\mathrm{im}(g)} = \mathrm{Id}$

*Proof.* Let $g$ be a pseudo-inverse to $f$ and define $U_V \overset{\text{def}}{=} \mathrm{im}(g)$ and $U_W \overset{\text{def}}{=} \ker(f)$. Then Lemma 2.2.5 shows that $g$ is in fact the pseudo-inverse to the triple $\big(U_V, U_W, f\big)$. Now, since $g|_{\mathrm{im}(\,f)}$ is the inverse to $f|_{U_V}$ by Lemma 2.2.4, we have $(f \circ g)|_{\mathrm{im}(f)} = \mathrm{Id}$ and $(g \circ f)|_{\mathrm{im}(g)} = (g \circ f)|_{U_V} = \mathrm{Id}$.
Conversely, assume that $g$ satisfies the conditions in (2).
We begin by showing that $\big(\mathrm{im}(g), \ker(g)\big) \in \Lambda(f)$. Let's show that $\mathrm{im}(f) \oplus \ker(g) = W$ by way of example. Indeed, first note that $\mathrm{im}(f) \cap \ker(g) = 0$, as any $w$ in this intersection must satisfy $w = (f \circ g)(w) = f(0) = 0$. Moreover, if we write $w = \big(w - f(g(w))\big) + f(g(w))\big)$, we see that trivially $f(g(w)) \in \mathrm{im}(f)$ and

$$g(w - f(g(w))) = g(w) - (g(f(g(w)) = g(w) - g(w) = 0$$

so that $\big(w - f(g(w))\big) \in \ker(g)$. This indeed shows that $\mathrm{im}(f) \oplus \ker(g) = W$. The proof of $\mathrm{im}(g) \oplus \ker(f) = V$ is completely analogous, allowing us to conclude that $(\mathrm{im}(g), \ker(g)) \in \Lambda(f)$.
It now remains to show that $g$ is indeed a pseudo-inverse to the triple $(\mathrm{im}(g), \ker(f), f)$. By Lemma 2.2.4, it suffices to show that $g|_{\mathrm{im}(f)}$ is the inverse to $f|_{\mathrm{im}(g)}$ and that $g|_{\ker(g)} = 0$. The first claim follows immediately from the fact that $g$ is a left inverse to $f : \mathrm{im}(g) \longrightarrow W$ and the second claim is trivial. ∎

In order to summarize the previous 2 lemmas, we introduce the following assignment, which is well-defined by Lemma 2.2.5

$$\Psi : \Pi(f) \longrightarrow \Lambda(f) : g \mapsto \big(\mathrm{im}(g), \ker(g)\big)$$

We now have:

**Lemma 2.2.7.** *Let $f \in \mathrm{Hom}(V, W)$. Then:*

- $\Pi(f) = \big\{ g \in \mathrm{Hom}(W, V) \;\big|\; (f \circ g)|_{\mathrm{im}(f)} = \mathrm{Id} \;\; \textit{and} \;\; (g \circ f)|_{\mathrm{im}(g)} = \mathrm{Id} \big\}$
- *The assignments $\Phi$ and $\Psi$ define 1:1 correspondences between $\Lambda(f)$ and $\Pi(f)$*

*Proof.* The first claim simply restates Lemma 2.2.6. To prove the second, we note that $\Psi \circ \Phi = \mathrm{Id}$ by Lemma 2.2.5. Moreover, $\Phi$ is surjective by definition, implying that $\Phi \circ \Psi = \mathrm{Id}$ as well ∎

We finish our discussion of pseudo-inverses by discussing a special choice of pseudo-inverse in $\Pi(f)$ that one can make if the vector spaces $V$ and $W$ are equipped with inner products. Indeed, recall the following standard result:

**Lemma 2.2.8.** *Let $U \subset V$ be a subspace of a finite dimensional inner product space. Then $U \oplus U^{\perp} = V$*

This leads us to the following Definition:

**Definition 2.2.9.** Let $V, W$ be finite-dimensional inner product spaces and let $f \in \mathrm{Hom}_{\mathbb{R}}(V, W)$. Then the *Moore-Penrose pseudo-inverse* is the pseudo-inverse to the triple $(\ker(f)^{\perp}, \mathrm{im}(f)^{\perp}, f)$.
We will denote it by $f^{+}$

It turns out that we can give a very satisfying description of Moore-Penrose pseudo-inverses:

**Lemma 2.2.10.** *Let $V, W$ be finite-dimensional inner product spaces and $f \in \mathrm{Hom}(V, W)$. Then the following are equivalent:*

1. *$g$ is the Moore-Penrose pseudo-inverse $f^{+}$ to $f$*
2. *$g$ is a pseudo-inverse to $f$ and $g \circ f$ and $f \circ g$ are self-adjoint linear maps*
3. *$f$ and $g$ satisfy $f \circ g \circ f = f$, $g \circ f \circ g = g$, $(g \circ f)^{*} = g \circ f$ and $(f \circ g)^{*} = f \circ g$*

*Proof.* The equivalence (2) $\iff$ (3) is simply a restatement of Lemma 2.2.7.
We now prove (2) $\implies$ (1):
Assume that $g$ is a pseudo-inverse to $f$ and that $g \circ f$ and $f \circ g$ are both self-adjoint. then Lemma 2.2.5 implies that $g$ is the pseudo-inverse to the triple $\big(\mathrm{im}(g), \ker(f), f\big)$. The claim will thus follow if we show that $\mathrm{im}(g) = \ker(f)^{\perp}$ and $\ker(g) = \mathrm{im}(f)^{\perp}$. By way of example, we will prove the former equality: First note that since $\mathrm{im}(g) \oplus \ker(f) = V$, it suffices to show that $\mathrm{im}(g) \perp \ker(f)$. Indeed, for $w \in W$ and $v \in \ker(f)$, we have:

$$\langle v\,, g(w)\,\rangle = \langle v\,, (g \circ f)(g(w))\,\rangle = \langle (g \circ f)^{*}(v)\,, g(w)\,\rangle = \langle (g \circ f)(v)\,, g(w)\,\rangle = \langle g(0)\,, g(w)\,\rangle = 0$$

The proof of $\ker(g) = \mathrm{im}(f)^{\perp}$ is analogous.
Finally, we show (1) $\implies$ (2):
Assume that $g$ is the Moore Penrose pseudo-inverse to $f$. Ie $g$ is the pseudo-inverse to the triple $(\ker(f)^{\perp}, \mathrm{im}(f)^{\perp}, f)$. We will show that $(f \circ g)$ is self-adjoint and leave the other claim to the reader. To this end, let $v, v' \in V$. Then

$$\begin{aligned}\langle v\,, g(f(v'))\,\rangle &= \Big\langle v - g(f(v)) + g(f(v)), g(f(v')) - v' + v' \Big\rangle \\ &= \Big\langle v - g(f(v)), g(f(v')) \Big\rangle + \Big\langle g(f(v)), g(f(v')) - v' \Big\rangle + \Big\langle g(f(v)), v' \Big\rangle\end{aligned}$$

Now, since $f \circ g \circ f = f$, we conclude that $v - g(f(v))$ and $g(f(v')) - v'$ lie in $\ker(f)$. Moreover, since $\ker(f) = \mathrm{im}(g)^{\perp}$, we conclude that

$$\langle v - g(f(v))\,, g(f(v'))\,\rangle = \langle g(f(v))\,, g(f(v')) - v'\,\rangle = 0$$

So that

$$\langle v\,, g(f(v'))\,\rangle = \langle g(f(v))\,, v'\,\rangle$$

implying that $f \circ g = (f \circ g)^{*}$. The equality $g \circ f = (g \circ f)^{*}$ is completely analogous. ∎

As mentioned in the introduction of this section, our main motivation for studying the Moore-Penrose pseudo-inverse, is to provide a description of the projection of a vector onto the image of a linear map. We begin with the following preparatory lemma:

**Lemma 2.2.11.** *Let $W = \mathrm{im}(f) \oplus \mathrm{im}(f)^\perp$ and $\pi_{\mathrm{im}(f)} : W \longrightarrow \mathrm{im}(f)$ be the canonical map. Then $\pi_{\mathrm{im}(f)}$ coincides with the projection onto the subspace $\mathrm{im}(f) \subset W$ in the sense of Definition 2.1.3*

*Proof.* Let $w \in W$. Then the decomposition $\mathrm{im}(f) \oplus \mathrm{im}(f)^\perp = W$ is given explicitely by $w = \big(w - \pi_{\mathrm{im}(f)}(w)\big) + \pi_{\mathrm{im}(f)}(w)$, implying that $w - \pi_{\mathrm{im}(f)(w)} \in \mathrm{im}(f)^\perp$. The result now follows from Lemma 2.1.1 ∎

**Lemma 2.2.12.** *Let $V, W$ be finite-dimensional inner product spaces and $f \in \mathrm{Hom}(V, W)$. Let $v \in V$ and $w \in W$. Finally denote the Moore-Penrose inverse of $f$ by $f^+$. Then the following are equivalent:*

1. *$f(v)$ is the projection of $w$ onto the subspace $\mathrm{im}(f)$*
2. *$v$ satisfies the normal equation $(f^* \circ f)(v) = f^*(w)$*
3. *$v$ lies in the affine subspace $f^+(w) + \ker(f)$*

*Proof.* The equivalence of (1) $\iff$ (2) is simply a restatement of Lemma 2.1.4.
To show the equivalence of (1) $\iff$ (3), we first note that $f(f^+w)) = \pi_{\mathrm{im}(f)}$, where $\pi_{im(f)}$ is the projection onto the subspace $\mathrm{im}(f) \subset W$ by Lemma 2.2.11. This shows that the vector $f^+(w) \in V$ indeed satisfies the condition (1). Next, assume (1), so that $v \in V$ satisfies $f(v) = \pi_{\mathrm{im}(f)}(v)$ and write $v = f^+(w) + v'$. Then

$$f(v) = \pi_{\mathrm{im}(f)(w)} \iff f(f^+(w) + v') = \pi_{\mathrm{im}(f)}(w) \iff \pi_{\mathrm{im}(f)}(w) + f(v') = \pi_{\mathrm{im}(f)}(w) \iff v' \in \ker(f)$$

This proves the claim ∎

This lemma has an interesting corollary which allows us to write the Moore-Penrose even more explicitly which will play an important role later on:

**Corollary 2.2.13.** *Let $V$ be a finite dimensional vector space and $W$ a finite dimensional inner product space. Let $f \in \mathrm{Hom}(V, W)$ be injective and choose* any *inner product on $V$. Then*

$$f^+ = (f^* \circ f)^{-1} \circ f^*$$

*Proof.* Since $f$ is injective (so that $\ker(f) = 0$), $f^+$ is the pseudo-inverse to the triple $(V, \mathrm{im}(f)^\perp, f)$ by Definition 2.2.9. It follows immediately that this condition is independent of the inner product on $V$. To prove the formula, simply note that $f^* \circ f$ is invertible if $f$ is injective and apply the second criterium of Lemma 2.2.12 ∎

We finish this section by giving a more explicit description of this map after introducing coordinates:
To this end, let $\{v_1, \ldots, v_n\}$ be an orthonormal basis for $V$ and $\{w_1, \ldots, w_m\}$ be an orthonormal basis for $W$. We denote by $M : \mathrm{Hom}(V, W) \longrightarrow \mathrm{Mat}_{n \times m}(\mathbb{R})$ the isomorphism that assigns to any $f \in \mathrm{Hom}_{\mathbb{R}}(V, W)$ its associated matrix $M_f$.

**Lemma 2.2.14.** *For any $f$, we have $M_{f^+} = \big(M_f\big)^+$ where $\big(M_f\big)^+$ is the unique matrix $M$ satisfying*

- *$M \cdot M_f \cdot M = M$ and $M_f \cdot M \cdot M_f = M_f$*
- *$M \cdot M_f$ and $M_f \cdot M$ are symmetric*

*Proof.* The first claim follows from the compatibility between composition of maps and multiplication of matrices. The second follows from the fact that the inner product is the standard one since the bases are orthonormal ∎

## 2.3. Proving the Main Theorem

After having reviewed the necessary linear algebraic, we now have the ingredients to prove the promised main theorem of logistic regression, which we recall below for the reader's convenience:

**Theorem 2.3.1** (linear regression). *Let $\mathfrak{y}$ be a finite dimensional inner product space, $\mathfrak{X}$ any set. Let $\mathfrak{H} \subset \mathfrak{y}^{\mathfrak{X}}$ be a finite dimensional subspace and*

$$\mathfrak{D} = \left\{ \Delta \subset \mathfrak{X} \times \mathfrak{y} \;\middle|\; |\Delta| \neq \infty \text{ and } \Delta \text{ separates } \mathfrak{H} \right\}.$$

*and*

$$c : \mathfrak{D} \times \mathfrak{H} \longrightarrow \mathfrak{R} : (\Delta, f) \mapsto ||(y - f(x))_{(x,y)\in\Delta}||_{\mathfrak{y}^\Delta}$$

*Then $\left(\mathfrak{X}, \mathfrak{y}, \mathfrak{D}, \mathfrak{H}, c\right)$ is a sharp learner.*

*Proof.* Let $\Delta \in \mathfrak{D}$.
Since $\mathfrak{H}$ is finite dimensional, we can choose an inner product on it. Moreover, the space $\mathfrak{y}^\Delta$ is canonically a finite dimensional inner product space by Lemma 2.0.1
We now consider the map:

$$\mathrm{ev}_\Delta : \mathfrak{H} \longrightarrow \mathfrak{y}^\Delta : f \mapsto f(\Delta_{\mathfrak{X}})$$

(where we used notation 1). It is easy to see that this map is linear. Moreover, the fact that $\Delta$ separates $\mathfrak{H}$ is equivalent to $ev_\Delta$ being injective
WE note now that $\mathrm{ev}_\Delta$ allows us to rewrite the cost function $c(\Delta, f)$ as follows:

$$c(\Delta, f) = ||\big(y - f(x)\big)_{(x,y)\in\Delta}||_{\mathfrak{y}^\Delta} = ||\Delta_{\mathfrak{y}} - \mathrm{ev}_\Delta(f)||_{\mathfrak{y}^\Delta}$$

It follows that $f$ that minimizes the cost $c(\Delta, f)$ iff $f$ minimizes the distance between $\Delta_{\mathfrak{y}}$ and $\mathrm{ev}_\Delta(f)$. By Lemma 2.1.2, this is in turn equivalent to requiring that $\mathrm{ev}_\Delta(f)$ is the projection of the vector $\Delta_{\mathfrak{y}}$ onto the image of the map $\mathrm{ev}_\Delta : \mathfrak{H} \longrightarrow \mathfrak{y}^\Delta$. We can now describe this image using Moore-Penrose inverses: indeed, let

$$h_\Delta \stackrel{\text{def}}{=} \mathrm{ev}_\Delta^+(\Delta_{\mathfrak{y}})$$

Then Lemma 2.2.12 implies that $\mathrm{ev}_\Delta(f)$ is the projection of $\Delta_{\mathfrak{y}}$ onto the image of $\mathrm{ev}_\Delta^+$ if and only if $f$ lies in the affine subspace $h_\Delta + \ker(\mathrm{ev}_\Delta) \subset \mathfrak{y}^\Delta$. We now recall that $\mathrm{ev}_\Delta$ is in fact injective so that finally $f = h_\Delta$ ∎

As a corollary, we can show that the Euclidean learner introduced in Definition 2.0.5 indeed is a learner provide an explicit formula for the learned hypothesis $h_\Delta$

**Corollary 2.3.2.** *Let $\mathfrak{X}$ and $\mathfrak{y}$ be finite dimensional inner product spaces,*

$$\mathfrak{D} = \left\{ \Delta \subset \mathfrak{X} \times \mathfrak{y} \;\middle|\; |\Delta| \neq \infty, \text{ and } span(\Delta_{\mathfrak{X}}) = \mathfrak{X} \right\},\ \mathfrak{H} = \mathrm{Hom}_{\mathbb{R}}(\mathfrak{X}, \mathfrak{y})$$

*and*

$$c : \mathfrak{D} \times \mathfrak{H} \longrightarrow \mathbb{R} : (\Delta, f) \mapsto ||\big(y - f(x)\big)_{(x,y)\in\Delta}||_{\mathfrak{y}^\Delta}$$

*Then $(\mathfrak{X}, \mathfrak{y}, \mathfrak{D}, \mathfrak{H}, c)$ is a sharp linear learner. Moreover,*

$$h_\Delta = \mathrm{ev}_\Delta^+\left(\Delta_{\mathfrak{y}}\right) = \left( \left( \mathrm{ev}_\Delta^* \circ \mathrm{ev}_\Delta \right)^{-1} \circ \mathrm{ev}_\Delta^* \right)(\Delta_{\mathfrak{y}})$$

*where $\mathrm{ev}_\Delta : \mathfrak{H} \longrightarrow \mathfrak{y}^\Delta : f \mapsto f(\Delta_{\mathfrak{y}})$*

*Proof.* By Theorem 2.0.3, we simply need to note that $\mathfrak{H} = \mathrm{Hom}_{\mathbb{R}}(\mathfrak{X}, \mathfrak{y}$ is finite dimensional and that each dataset $\Delta$ indeed separates $\mathfrak{H}$ since $\Delta_{\mathfrak{X}}$ spans the whole of $\mathfrak{X}$, so that $f$ is fully characterized on $\Delta_{\mathfrak{X}}$. The formula for $h_\Delta$ follows from the proof of Theorem 2.0.3, where we showed that $h_\Delta = ev^+(\Delta_y)$ and corollary 2.2.13 since $\mathrm{ev}_\Delta$ is injective ∎

## 2.4. Coordinates for Euclidean Learners

In the last section of this chapter, we will once and for all fix a Euclidean learner $\mathfrak{L}$ (as defined in 2.0.5) and introduce coordinates on the feature space $\mathfrak{X}$ and label space $\mathfrak{y}$. This will allow us to represent the learned hypothesis $h_\Delta$ by a matrix. We will show that this matrix indeed coincides with what is classically referred to as the *regression matrix*.

Let us pick orthonormal bases $\mathcal{E} \stackrel{\text{def}}{=} (v_1, \dots, v_m)$ for $\mathfrak{X}$ and $\mathcal{F} \stackrel{\text{def}}{=} (w_1, \dots w_n)$ for $\mathfrak{y}$. We will denote the associated coordinate maps by $\text{co}_\mathcal{E}$ and $\text{co}_\mathcal{F}$ respectively.

Now to any map $f \in \text{Hom}_\mathbb{R}(\mathfrak{X}, \mathfrak{y})$, we'll associate the matrix $M_f \in \text{Mat}_{n\times m}(\mathbb{R})$ whose $i$-th column is given by $\text{co}_\mathcal{F}(v_i)$. It is well known that this matrix satisfies

$$\text{co}_\mathcal{F}\big(f(v)\big) = M_f \cdot \text{co}_\mathcal{E}(v)$$

Our goal in this section is to compute the matrix $M_{h_\Delta}$ associated to the learned hypothesis $h_\Delta$ of the dataset $\Delta$. More precisely, we will prove the following theorem:

**Theorem 2.4.1.** *Let $\Delta = \big((x_1, y_1), \dots, (x_d, y_d)\big) \in \mathfrak{D}$ be a dataset.*
*Let $X = [\text{co}_\mathcal{E}(x_i)_i] \in \text{Mat}_{m\times d}(\mathbb{R})$ and $Y = [\text{co}_\mathcal{F}(y_j)_j] \in \text{Mat}_{d\times n}(\mathbb{R})$.*
*Then the matrix corresponding to the learned hypothesis $h_\Delta \in \text{Hom}_\mathbb{R}(\mathfrak{X}, \mathfrak{y})$ is given by*

$$M_{h_\Delta} = Y \cdot X^+,$$

*where $X^+$ is the Moore-Penrose pseudo-inverse of $X$ (see 2.2.14).*
*In other words, for any feature $v \in \mathfrak{X}$, we have*

$$\text{co}_\mathcal{F}\big(h_\Delta(v)\big) = \big(Y \cdot X^+\big) \cdot \text{co}_\mathcal{E}(v)$$

We will prove this using a series of lemma's.

Since $h_\Delta = \text{ev}_\Delta(\Delta_\mathfrak{y})$ by corollary 2.3.2, it's worth reinterpreting the map $\text{ev}_\Delta : \text{Hom}(\mathfrak{X}, \mathfrak{y}) \longrightarrow \mathfrak{y}^d$ as a map between matrix spaces through the use of the orthonormal bases $\mathcal{E}$ and $\mathcal{F}$. To this end, we consider the isomorphism $M : \text{Hom}(\mathfrak{X}, \mathfrak{y}) \longrightarrow \text{Mat}_{n\times m}(\mathbb{R})$which assigns to $f$ its matrix representation $M_f$, after the choice of bases $\mathcal{E} \subset \mathfrak{X}$ and $\mathcal{F} \subset \mathfrak{y}$. We will also consider the map $\alpha_P$ the multiplication by a matrix $P$ on the right

**Lemma 2.4.2.** *Let $X$ denote the matrix $\big[\text{co}_\mathcal{E}(x_i)_i\big]$. Then the diagram:*

$$\begin{array}{ccc} \text{Hom}(\mathfrak{X}, \mathfrak{y}) & \xrightarrow{\text{ev}_\Delta} & \mathfrak{y}^d \\ \Big\downarrow{\scriptstyle M_{(-)}} & & \Big\downarrow{\scriptstyle (\text{co}_\mathcal{F})^d} \\ \text{Mat}_{n\times m}(\mathbb{R}) & \xrightarrow[\alpha_X]{} & \text{Mat}_{d\times n}(\mathbb{R}) \end{array}$$

*is commutative*

*Proof.* This is really just a restatement of the various definitions:
Let $f \in \text{Hom}(\mathfrak{X}, \mathfrak{y})$. Then $\big(ev_\Delta \circ \text{co}_\mathcal{F}^d\big)(f)$ is the matrix $\big[(\text{co}_\mathcal{F}(f(x_i))_i\big]$, whose $i$-th column is the vector $\text{co}_\mathcal{F}\big(f(x_i)\big) \in \mathbb{R}^n$.
Going around the other way, we obtain the matrix $\alpha_X \circ M_f = M_f \cdot X$, whose $i$-th column is $M_F \cdot \big[\text{co}_\mathcal{E}(x_i)\big]$. Now, by the definition of $M_f$, we have $M_f \cdot [\text{co}_\mathcal{E}(x_i)] = \text{co}_\mathcal{F}\big(f(x_i)\big)$, proving the claim

∎

Next, we will use the above lemma to describe the Moore-Penrose pseudo-inverse of $\text{ev}_\Delta$ in terms of the maps $M$, $\alpha_X$ and $\big(\text{co}_\mathcal{F}\big)^d$. We'll need a little result from Euclidean geometry to do this.. recall that a map $f$ on an inner product space $V$ is *orthogonal* if it preserves the inner product:

$$\langle u\,, v\,\rangle = \langle f(u)\,, f(v)\,\rangle$$

This is equivalent to the adjoint coinciding with the inverse: $f^* = f^{-1}$. We now have the following linear algebraic lemma:

**Lemma 2.4.3.** *Let $V$ and $W$ be finite dimensional inner product spaces.*
*Let $f \in \mathrm{Hom}(V,W)$ and let $\phi \in \mathrm{Hom}_{\mathbb{R}}(V,V)$ and $\psi \in \mathrm{Hom}_{\mathbb{R}}(W,W)$ be orthogonal maps. Then the Moore-Penrose inverse of $\psi \circ f \circ \phi \in \mathrm{Hom}(V,W)$ is given by*

$$(\psi \circ f \circ \phi)^+ = \phi^{-1} \circ f^+ \circ \psi^{-1}$$

*Proof.* We need to check the 4 conditions of criterium (3) in Lemma 2.2.10. First

$$(\phi^{-1} \circ f^+ \circ \psi^{-1}) \circ (\psi \circ f \circ \phi) \circ (\phi^{-1} \circ f^+ \circ \psi^{-1}) = (\phi^{-1} \circ f^+ \circ f \circ f^+ \circ \psi^{-1}) = \phi^{-1} \circ f^+ \circ \psi^{-1}$$

The second condition is analogous.
To prove the third condition, we invoke that $\phi$ and $\psi$ are orthogonal, so that $\phi^{-1^*} = \phi$ and $\psi^{-1^*} = \psi$. We now compute:

$$\Big((\phi^{-1} \circ f^+ \circ \psi^{-1}) \circ (\psi \circ f \circ \phi)\Big)^* = \Big(\phi^{-1} \circ f^+ \circ f \circ \phi\Big)^* = \phi^* \circ (f^+ \circ f)^* \circ \phi^{-1^*} = \phi \circ (f^+ \circ f) \circ \phi$$

Where the last line follows from the orthogonality of $\phi$ and the fact that $f^+ \circ f$ is self-adjoint since $f^+$ is the Moore-Penrose pseudo-inverse to $f$.
The fourth condition is proven analogously.

∎

To use the above lemma we'll need to explain how $\mathrm{co}_{\mathcal{F}}$ and $M$ are indeed orthogonal maps. For the benefit of the reader we first recap the inner products involved in the commutative diagram of Lemma 2.4.2:

- The bilinear form on $\mathrm{Hom}(\mathfrak{X}, \mathfrak{y})$ does not have a general description, however Lemma 2.0.1 implies that the maps
$$e_{i,j}(v_k) = \begin{cases} 0, & \text{if } k \neq i \\ w_j, & \text{otherwise} \end{cases} \tag{3}$$
form an orthonormal basis

- The space $\mathfrak{y}^d$ has the inner product defined by $\langle (y_1, \ldots, y_n), (y'_1, \ldots y'_n) \rangle = \sum_i \langle y_i, y'_i \rangle$ and orthonormal basis $\big\{(w_{i_1}, \ldots w_{i_m})\big\}_{i_1, \ldots i_m}$ by Lemma 2.0.1

- The matrix spaces are endowed with the *Frobenius inner product*: for matrices $A$ and $B$, it's defined as $\langle A, B \rangle = \sum_{i,j} A_{i,j} \cdot B_{i,j}$. It follows immediately that the elementary matrices $\big\{E_{i,j}\big\}_{i,j}$ form an orthonormal basis for these spaces.
Another description of this inner product will be useful later on. Indeed, we have:
$$\mathrm{Tr}(A \cdot B^t) = \sum_k (A \cdot B^t)_{k,k} = \sum_k (\sum_j A_{k,j} B^t_{j,k}) = \sum_k (\sum_j A_{k,j} B_{k,j}) = \langle A, B \rangle$$
This allows us to prove certain properties. For example, for any matrix $P$, we have
$$\langle A, B \cdot P \rangle = \mathrm{Tr}\left(A \cdot (B \cdot P)^t\right) = \mathrm{Tr}(A \cdot P^t \cdot B^t) = \left\langle (A \cdot P^t), B \right\rangle$$

We now have:

**Lemma 2.4.4.** *The maps $M$ and $(\mathrm{co}_{\mathcal{F}})^d$ are orthogonal*

*Proof.* Indeed, it suffices to show that both maps $M$ and $\left(\mathrm{co}_{\mathcal{F}}\right)^d$ send an orthonormal basis to an orthonormal basis. Now, the map $M$ sends the orthonormal basis $\left\{e_{i,j}\right\}_{i,j}$ to the orthonormal basis of elementary matrices $\left\{E_{i,j}\right\}_{i,j}$.
The map $(\mathrm{co}_{\mathcal{F}})^d$ in turn sends the orthonormal basis $\left\{(w_{i_1},\dots w_{i_m})\right\}_{i_1,\dots i_m}$ to the elementary matrices $\left\{E_{i,j}\right\}_{i,j}$ as well. ∎

Combining both lemma's hence lets us write the Moore-Penrose inverse of $\mathrm{ev}_{\Delta}^{+}$ as

$$\mathrm{ev}_{\Delta}^{+} = \left(M^{-1} \circ \alpha_X \circ \mathrm{co}_{\mathcal{F}}^d\right)^{+} = M^{-1} \circ \left(\alpha_X\right)^{+} \circ \mathrm{co}_{\mathcal{F}}^d$$

We now wish to simplify the map $\left(\alpha_X\right)^{+}$. To this end we note that for any matrix $P \in \mathrm{Mat}_{d\times m}(\mathbb{R})$, we can associate its Moore-Penrose inverse as the inverse following example 2.2.14

**Lemma 2.4.5.** *For any $P \in \mathrm{Mat}_{d\times m}(\mathbb{R})$, let $\alpha_P$ denote the right multplication by $P$*

$$\alpha_P : \mathrm{Mat}_{n\times m}(\mathbb{R}) \longrightarrow \mathrm{Mat}_{d\times n}(\mathbb{R}) : A \mapsto A \cdot P$$

*Then we have $\left(\alpha_P\right)^{+} = \alpha_{P^+}$*

*Proof.* We simply need to show that $\alpha_{P^+}$ satisfies the conditions of Moore-Penrose pseudo-inverse by checking (3) of Lemma 2.2.10.
The first two conditions are absolutely trivial.
To prove the third condition, we need to show that the map $\alpha_P \circ \alpha_{P^+}$ (which coincides with $\alpha_{P^+\cdot P}$) is self-adjoint. To this end, we let $A, B \in \mathrm{Mat}_{n\times m}(\mathbb{R})$. The discussion before Lemma 2.4.4 yields:

$$\langle A\,, \alpha_{P^+\cdot P}(B)\,\rangle = \mathrm{Tr}\left(A \cdot \left(B \cdot P^+ \cdot P\right)^t\right) = \mathrm{Tr}\left(A \cdot (P^+ \cdot P)^t \cdot B\right) = \mathrm{Tr}\left(A \cdot (P^+ \cdot P) \cdot B^t\right) = \langle \alpha_{P^+\cdot P}(A)\,, B\,\rangle$$

Where the 3rd equality follows from the fact that the matrix $P^+ \cdot P$ is symmetric by Lemma 2.2.14.
The proof of condition (4) is completely analogous. ∎

The above lemma allows us to rewrite the Moore-Penrose pseudo-inverse one step further

$$\mathrm{ev}_{\Delta}^{+} = M^{-1} \circ \alpha_{X^+} \circ \left(\mathrm{co}_{\mathcal{F}}\right)^d$$

We can now prove the main theorem of this section:

*proof of Theorem 2.4.1.* We want to show that $M_{h_\Delta} = Y \cdot X^+$. Now, we know from Corollary 2.3.2 that $h_\Delta = \mathrm{ev}_{\Delta}^{+}\left(\Delta_{\mathfrak{y}}\right) = \mathrm{ev}_{\Delta}^{+}\left(y_1,\dots y_d\right)$. Moreover, the above discussion shows that

$$\mathrm{ev}_{\Delta}^{+} = M^{-1} \circ \alpha_{X^+} \circ \left(\mathrm{co}_{\mathcal{F}}\right)^d$$

So that applying the set of features $(y_1,\dots y_d)$ to the left hand side followed by the map $M$ yields

$$M_{h_\Delta} = \left(\alpha_{X^+} \circ \mathrm{co}_{\mathcal{F}}^d\right)(y_1,\dots,y_d) = \alpha_{X^+}\left(\left[\mathrm{co}_{\mathcal{F}}(y_1)\dots \mathrm{co}_{\mathcal{F}}(y_d)\right]\right) = \alpha_{X^+}(Y) = Y \cdot X^+$$

as required ∎

# 3 Gradient Descent: Convex Learners

# 4 Topology: Closed Learners

In this section, we discuss a natural way to extend the hypothesis space of a learner through the use of topology.
More precisely, we will endow the feature space $\mathfrak{X}$ and $\mathfrak{y}$ with a topology and assume that in our setting, the hypothesis space $\mathfrak{H}$ is a subset of $\mathcal{C}(\mathfrak{X},\mathfrak{y})$, the space of continuous function (endowed with the usual compact-open topology).
In this section, we intend to naturally construct a learner whose hypothesis space is the closure of $\mathfrak{H}$ in $\mathcal{C}(\mathfrak{X},\mathfrak{y})$. It is rather clear how one would go about this: we'll first extend the cost functions $c_\Delta : \mathfrak{H} \longrightarrow \mathbb{R}$ to $\overline{\mathfrak{H}}$. We next argue why this cost function still attains a unique minimum for a hypothesis $h_\Delta \in \overline{\mathfrak{H}}$, which in turn allows us to construct the required hypothesis function $\mathfrak{D} \longrightarrow \overline{\mathfrak{H}}$.

## 4.1. Some Necessary Facts on Topology

#### 4.1.1. Topologies on function spaces

**4.1.2. Extending continuous functions** As alluded to in the introduction to this section, we will show how to extend the hypothesis space. Since for any hypothesis we'll need to define a cost function, it will be crucial to understand exactly when continuous functions have unique extensions.

## 4.2. Topological Learners

As mentioned in the introduction, our goal in this section is to endow feature- and label space with topologies in such a way that the associated cost functions become continuous. To this end we make the following definition

**Definition 4.2.1.** A topological learner is a learner where $\mathfrak{X}$ and $\mathfrak{y}$ are topological spaces such that $\mathfrak{H} \subset \mathcal{C}(\mathfrak{X},\mathfrak{y})$ and $\mathfrak{y}$ satisfies the following admissibility condition:

For any dataset $\Delta$ and hypothesis $f \in \mathfrak{H}$, there exists an open set $U \subset \mathfrak{y}$ such that

$$f = \underset{g(\Delta_{\mathfrak{X}})\subset U}{\arg\max} \left\{ |c(\Delta, g)| \right\}$$

By way of example we show how any regular learner can trivially be viewed as a topological learner:

**Lemma 4.2.2.** *Let $\mathfrak{L}$ be any regular learner (as in Definition 1.0.2). Then $\mathfrak{L}$ is topological when $\mathfrak{X}$ and $\mathfrak{y}$ are endowed with the discrete topologies*

*Proof.* First, it is clear that $\mathfrak{H} \subset \mathcal{C}(\mathfrak{X},\mathfrak{y})$ in this case, since $\mathfrak{X}$ is discrete. Next, we need to show that the topology on $\mathfrak{y}$ is indeed admissible. Now, for any $f \in \mathfrak{H}$, let $U = f(\Delta_{\mathfrak{X}})$. Then $g(\Delta) \subset U \iff f(\Delta) = g(\Delta)$,and so $c(\Delta, f) = c(\Delta, g)$ implying that $c(\Delta, -)$ is in fact constant. ✗

**Lemma 4.2.3.** *Let $\mathcal{L}$ be a topological learner. Then the cost functions*

$$c_\Delta : \mathfrak{H} \longrightarrow \mathbb{R}$$

*are continuous*

*Proof.* Since the centered intervals $]-\epsilon, \epsilon[$ form a subbase, it suffices to show that their inverse images are continuous, i.e. the set $O = \left\{ f \in \mathfrak{H} \mid |c(\Delta, f)| < \epsilon \right\}$ is open in $\mathfrak{H}$. We will first show that this set is open in $\mathcal{C}(\mathfrak{X},\mathfrak{y})$.
Indeed, since the sets $\Gamma(K,U)$ for $K \subset \mathfrak{X}$ compact and $U \subset \mathfrak{y}$ open form a subbase, we'll need to show that for any $f \in O$ there exists such a $K$ and $U$ such that $f \in \Gamma(K,U) \subset O$. To this end, we let

$K \stackrel{\text{def}}{=} \Delta_{\mathfrak{X}}$ (which is finite and compact in particular). Then the previous claim now translates into the following: if $\epsilon$ and $f$ satisfy $c(\Delta, f) < \epsilon$, then there exists an open set $U$ such that $f(\Delta_{\mathfrak{X}}) \subset U$ and for any $g$ such that $g(\Delta_{\mathfrak{X}}) \subset U$, we necessarily have $c(\Delta, g), \epsilon$.
Now, let $U$ be the open set guaranteed by the admissibility condition from Definition 4.2.1. Then $f(\Delta_{\mathfrak{X}}) \subset U$, trivially and any other $g$, such that $g(\Delta_{\mathfrak{X}}) \subset U$, necessarily satisfies

$$|c(\Delta, g)| \leq |c(\Delta, f)| < \epsilon$$

Hence the claim ∎

**Lemma 4.2.4.** *Let $\mathfrak{L}$ be a linear learner. Then $\mathfrak{L}$ is topological when $\mathfrak{X}, fy$ are considered with their natural topologies*

*Proof.* Recall that in the context of linear learners, $\mathfrak{X}$ and $\mathfrak{y}$ are finite-dimensional inner product spaces and the hypothesis space is $\mathrm{Hom}_{\mathbb{R}}(\mathfrak{X}, \mathfrak{y})$. We now let $\Delta$ be a dataset and $f \in \mathfrak{H}$ and wish to show that there exists an open set in $\mathfrak{y}$ such that

$$f = \underset{g(\Delta_{\mathfrak{X}}) \subset U}{\arg\max} \left\{ |c(\Delta, g)| \right\}$$

∎

**Lemma 4.2.5.** *Let $\mathfrak{L}$ be a topological learner, and let $\overline{\mathfrak{H}} \subset \mathcal{C}(\mathfrak{X}, \mathfrak{y})$ denote the closure of $\mathfrak{H}$ in the compact-open topology. Then each cost function $c_\Delta$ has unique continuous extension*

$$c_\Delta : \overline{\mathfrak{H}} \longrightarrow \mathbb{R}$$

*Proof.* ∎

# 5 Application: Neural Learners