# Implicit Bias and Generalisation on Linear Models


1041243

A dissertation submitted for the degree of

*MSc in Mathematical Sciences*

Trinity 2020

# Abstract

Surprisingly, many modern overparametrised machine learning methods are able to generalise well on unseen data, even if trained to a perfect fit of noisy training data. Attention to this phenomenon has been inspired by neural networks which have recently shown stunning success and gained immense popularity. A theory towards an explanation is an implicit bias or implicit regularisation hidden somewhere in the learning process. This motivates our work, which studies implicit bias on linear models in an overparametrised setting and its implications to generalisation. We examine implicit bias of projected mirror descent and extend with some new results to more constraint sets. We also investigate generalisation of the minimum Euclidean norm solution in linear regression and compare it to a new mirror descent interpolator.

# Contents

# Chapter 1

# Introduction

## 1.1 Motivation and Literature Review

State-of-the-art deep neural networks are often overparametrised - they have more parameters than the size of the training dataset. In image recognition, the best performing networks have hundreds of millions of parameters (for example Xie et al. [58]). It is common to train them to perfectly fit training data (interpolate), nonetheless, they have excellent performance on unseen data (generalisation). In an influential paper, Zhang et al. [59] showed that, strikingly, neural networks can interpolate even on randomly corrupted data and still exhibit admirable generalisation. This clashes with common statistical wisdom which postulates a trade-off between the fit to the training data and generalisation (Hastie et al. [22, p. 221]), especially for complex prediction rules. The qualitative explanation is that, when interpolating, they "learn" the underlying noise in the data and hence generalise badly, which is called overfitting (Mohri et al. [35, p. 8]). Put concisely, neural networks seem to be sometimes immune to overfitting. This is surprising.

Instead, a double descent behaviour has been observed in neural networks (Belkin et al. [8], [9]). Here, the aforementioned classical trade-off is extended by a monotonically decreasing curve.

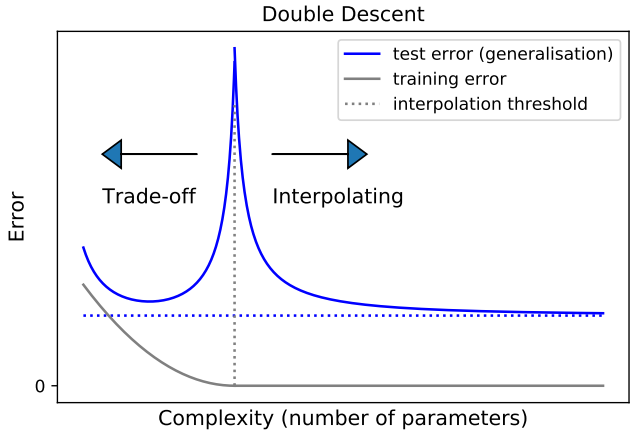


Figure 1.1

Through experiments, Zhang et al. [59] showed that standard explicit regularisation (e.g. weight decay, dropout) is not sufficient to explain this mystifying phenomenon. Instead, a candidate is the implicit bias of the optimisation algorithm - usually stochastic gradient descent. Given a set of minimisers of an optimisation problem, which minimisers does the algorithm "pick" or is biased towards? Converging towards minimisers with specific properties is called the implicit bias of an algorithm. Can implicit bias provide some sort of (implicit) regularisation, which helps finding an explanation? This has been observed in simpler settings. Liang and Rakhlin [32, Figure 1, p. 2] empirically present that, in kernel ridge regression, the best generalisation is sometimes achieved by the minimum norm interpolating solution - i.e. when the explicit ridge regularisation is turned off. Notably, the minimum norm interpolator is the implicit bias of (stochastic) gradient descent. They refer to this as implicit regularisation. Belkin et al. [10, Section 3] also observe strong performance of interpolators in kernel methods. Several lines of research have investigated implicit bias and regularisation; in linear models with unique root losses (Gunasekar et al. [19]), linear models with logistic-like losses on separable and nonseparable data (Soudry et al. [49], Nacson et al. [37] and Ji and Telgarsky [26], respectively), in matrix factorisation (Gunasekar et al. [20], Arora et al. [2]) and in neural networks (Neyshabur et al. [41],[39],[40], Gunasekar et al. [18]).

However, can implicit bias explain the mentioned phenomena of neural networks? The way to tackle this now becomes clearer. First, we investigate what the implicit bias is and then examine how it influences out-of-sample error (generalisation). In order to fully understand this phenomenon, one needs to first understand it in the simplest setting. Therefore, this work studies only linear models with unique root losses (including linear regression).

The analysis of Liang and Rakhlin [32, pp. 5–6] displays that good generalisation of the kernel minimum norm interpolator happens in a high-dimensional setting under suitable properties of the kernel. Bartlett et al. [6, Section 3.1] isolate a setting of "benign overfitting" in which good generalisation is possible even in linear regression. Similarly, their setting is highly overparametrised. Noticeably, large overparametrisation seems to appear simultaneously with this effect. Indeed, Muthukumar et al. [36, Theorem 1] elegantly identify a fundamental limit on how well any interpolator in linear regression can generalise, which can be small only for high overparametrisation [36, Corollary 1]. However, the situation is not so simple, Hastie et al. [21] suggest that when overparametrisation is too large then the minimum norm solution, the interpolator that we have access to through gradient descent, does not always generalise well - see [21, Figure 2, p. 10].

The following central questions provide goals towards which this project is directed.

**Question 1.** Assuming an algorithm converges, can we characterise its implicit bias?

We tackle this for linear models, a unique root loss and projected mirror descent (Nemirovski and Yudin [38]) - a generalisation of gradient descent.

**Question 2.** Assuming we can characterise the limit of an algorithm, what can we infer about generalisation?

We study this in linear regression and the minimum Euclidean norm interpolator, which is the implicit bias of gradient descent.

Importantly, our motivation is not to argue that interpolating is necessarily better than explicit regularisation, certainly not in linear regression (see [21, Section 6]). The objective is that understanding interpolation and implicit bias in a simpler setting gives insights to the hypothesis that the aforementioned phenomena of neural networks can be explained through implicit regularisation.

## 1.2 A Note on Organisation and Originality

The work could be split into two subparts - implicit bias and generalisation. In both, we first understand, reproduce and review results from relevant literature and then generalise or extend through own ideas.

A significant part of this work is, to our best knowledge, original (20/60 pages). Specifically, we could identify three contributions. Implicit bias results of projected mirror descent with inequality constraint sets (Propositions 2.13, 2.15 and Appendix B), convergence of gradient descent on linear models with an increasing nonlinearity (Lemma 2.17 and Proposition 2.18) and a generalisation analysis of a particular mirror descent interpolator (Section 3.3). All present ideas are heavily inspired by the read literature and, importantly, by Professor Rebeschini, the supervisor of this project. The author would like to express sincere thanks for his help, patience and guidance.

Because of the word limit, we had to carefully decide what to include in the final work. We decided to leave out some proofs (but include the intuition and statements) to not deprive of a thorough discussion of other important topics.

## 1.3 Problem Setup

We denote deterministic quantities by lower-case letters, random variables by upper-case or Greek letters, matrices with bold font and sets by capital italic letters, as follows. The following is inspired by Rebeschini [46] and Gunasekar et al. [19].

We define a *dataset* to be a set of independent and identically distributed (i.i.d.) random variables $\{Z_1, \dots, Z_n\} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ with *features* $X_i \in \mathcal{X}$ and *labels* $Y_i \in \mathcal{Y}$, for some sets $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. We "store" the features in a random matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with rows $X_i$, and the labels in a random vector $Y \in \mathbb{R}^n$ with entries $Y_i \in \mathbb{R}$. The deterministic covariance matrix of the features is denoted by $\mathbf{c} = \mathbb{E}(X_i X_i^T) \in \mathbb{R}^{d \times d}$. In machine learning, one cares about when there is a

relationship between $X_i$ and $Y_i$. Namely that there exists $f : \mathcal{X} \to \mathcal{Y}$ such that

$$Y_i = f(X_i) + \xi_i,$$

where, we assume, $\{\xi_1, \ldots, \xi_n\}$ are i.i.d. Gaussian random variables with mean 0 and variance $\sigma^2 \in \mathbb{R}_{>0}$, denoted by $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, which are independent of $\{X_1, \ldots, X_n\}$. We call $\xi = (\xi_1, \ldots, \xi_n)$ a *noise vector*. The objective is to "learn" the function $f$. For this, one introduces a set $\mathcal{H}$, called the *hypothesis class* and a *loss function* $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and seeks to find

$$f^* \in \arg\min_{h \in \mathcal{H}} r(h) := \arg\min_{h \in \mathcal{H}} \mathbb{E}\, \ell(h, \tilde{Z}), \tag{1.1}$$

where $\tilde{Z}$ is a random variable which follows the same distribution as $Z_i = (X_i, Y_i)$, is independent from $\{Z_1, \ldots, Z_n\}$ and

$$r(h) := \mathbb{E}\, \ell(h, \tilde{Z})$$

is called the *population risk* of $h$. $r(h) - r(f^*)$ is called the *excess risk* of $h$. Often in practice, the distribution of the data and noise is unknown and hence (1.1) is impossible to solve directly. An alternative is to search for

$$F^* \in \arg\min_{h \in \mathcal{H}} R(h) := \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i), \tag{1.2}$$

(Vapnik, [55, Section 1.7]) where

$$R(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$$

is called the *empirical risk* of $h$. Then, one tries to bound

$$r(F^*) - r(f^*). \tag{1.3}$$

in expectation or probability, which is sometimes possible by notions of complexity of $\mathcal{H}$ (Bartlett and Mendelson [5], Bartlett et al. [4]). Importantly, $\mathcal{H}$ itself is an input of the problem because $f \in \mathcal{H}$ is apriori unknown.

What we introduced so far is general and also covers neural networks. This work studies a simpler setting, where we assume

$$\mathcal{H} = \{\mathbb{R}^d \ni x \mapsto \langle w, x \rangle \ : \ w \in \mathbb{R}^d\}$$

is the set of *linear models*. We denote $w \in \mathbb{R}^d$ to represent the unique function of $\mathcal{H}$ that it defines. We also assume that $\ell : \mathcal{H} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a loss function with a unique root.

**Definition 1.1.** $\ell : \mathcal{H} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a loss function *with a unique root* if there exists a continuously differentiable $\phi : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ such that the following are satisfied.

- $\ell(w,(x,y)) = \phi(\langle w,x\rangle, y)$ for all $w,x \in \mathbb{R}^d$ and $y \in \mathbb{R}$.
- $\phi(\hat{y},y) = 0$ if and only if $\hat{y} = y$.
- $\lim_{t\to\infty} \phi(\hat{y}_t, y) = 0$ if and only if $\lim_{t\to\infty} \hat{y}_t = y$ for any sequence $(\hat{y}_t)_{t=1}^{\infty}$.

We abuse notation to write $\ell = \phi$ and $\ell(w,(x,y)) = \ell(\langle w,x\rangle, y)$. The most important example is the squared error loss $\ell(w,(x,y)) = (y - \langle w,x\rangle)^2$, but also the Huber loss (Huber [24], Hastie et al. [22, p. 349]), Log-cosh loss (Zhang et al. [60, (12)]) or Tangent loss (Masnadi-Shirazi et al. [34, (25)]). The minimisation problem (1.2) reduces to

$$\arg\min_{w\in\mathbb{R}^d} R(w) = \arg\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(w^T X_i, Y_i). \tag{1.4}$$

As $\ell$ has a unique root, $R(w) = 0$ if and only if

$$w \in \mathcal{G} := \{w \in \mathbb{R}^d : \mathbf{X}w = Y\}.$$

$\mathcal{G}$ is the *set of solutions*, *interpolators* or *minimisers* and we always assume $\mathcal{G} \neq \emptyset$ for all realisations of the data. We also introduce

$$\mathcal{M} := \mathrm{Span}\{X_i : i \in \{1,\dots,n\}\} = \mathrm{Im}(\mathbf{X}^T).$$

A common algorithm of choice for solving problem (1.4) is gradient descent, defined as follows [46, Lecture 9, p. 3].

**Algorithm 1:** Gradient Descent

**Input:** initialisation $W_1 = w_1$, stopping time $T$ and stepsize $(\eta_t)_{t\in\{1,\dots,T\}}$;
**for** $t = 1,\dots,T$ **do**
  $W_{t+1} = W_t - \eta_t \nabla R(W_t)$
**end**

# Chapter 2

# Implicit Bias

## 2.1 Projected Mirror Descent

The results of this chapter hold exactly the same if one treats all present quantities deterministically, as in the implicit bias literature. We had to make a choice, regarding notation, between consistency of the document and consistency with the literature. We chose the former and we use the notation as introduced.

The following exposition is adapted from Cesa-Bianchi and Lugosi [14, Chapter 11], Bubeck [12, Chapter 4], [13, Chapter 5] and Rebeschini [46, Lecture 10]. In the rest of the work, let $\mathcal{D} \subseteq \mathbb{R}^d$ be an open convex set endowed with a norm $\|\cdot\|_{\mathcal{D}}$, let $\mathcal{C} \subseteq \mathbb{R}^d$ be closed and convex, and let $\mathcal{C} \cap \mathcal{D}$ have nonempty interior. See Appendix A for a review of convexity.

**Definition 2.1.** $\Phi : \mathcal{D} \to \mathbb{R}$ is a *mirror map* if the following hold.

1. $\Phi$ is strictly convex and continuously differentiable.
2. $\nabla\Phi : \mathcal{D} \to \mathbb{R}^d$ is surjective.
3. If $w_\infty \in \partial\mathcal{D}$ and
$$\mathcal{D} \ni w_t \underset{t\to\infty}{\longrightarrow} w_\infty$$
then
$$\lim_{t\to\infty} \|\nabla\Phi(w_t)\|_{\mathcal{D}} = \infty.$$
4. $\Phi$ is continuously extendable to $\overline{\mathcal{D}}$.

Without surjectivity of $\nabla\Phi$, $\Phi$ is called a *Legendre* function (Rockafellar [47, p. 258], [14, p. 294]). The last condition assures $\Phi$ is a *closed convex function* on $\overline{\mathcal{D}}$ [11, Appendix A.3.3, pp. 639-640], which will be needed later.

**Lemma 2.2.** If $\Phi : \mathcal{D} \to \mathbb{R}$ is a mirror map then $\nabla\Phi : \mathcal{D} \to \mathbb{R}^d$ is injective.

*Proof.* By a standard property of strict convexity (Proposition A.3),
$$\Phi(y) > \Phi(x) + \nabla\Phi(x)^T(y - x)$$
$$\Phi(x) > \Phi(y) + \nabla\Phi(y)^T(x - y),$$

for all $x \neq y \in \mathcal{D}$. If we assume $\nabla\Phi(x) = \nabla\Phi(y)$, then adding the equations gives a contradiction. □

**Definition 2.3.** The *Bregman divergence* of a mirror map $\Phi : \mathcal{D} \to \mathbb{R}$ is a function $D_\Phi : \mathcal{D}^2 \to \mathbb{R}$ with

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \nabla\Phi(x)^T(y - x).$$

The *(Bregman) projection* of $x \in \mathcal{D}$ onto $\mathcal{C}$ is defined by

$$\prod_{\mathcal{C}}^{\Phi}(x) = \operatorname*{arg\,min}_{y \in \mathcal{C} \cap \mathcal{D}} D_\Phi(y, x).$$

The projection is well-defined and unique by Bauschke and Borwein [7, Theorem 3.12 (iii)]. For this, we needed $\Phi$ to be closed convex on $\overline{\mathcal{D}}$ .

**Lemma 2.4.** The function $f : \mathcal{D} \ni y \mapsto D_\Phi(y, x)$ is continuously differentiable and stricly convex for all $x \in \mathcal{D}$.

*Proof.* $f$ is continuously differentiable, because $\Phi$ is. By repeatedly using Proposition A.3 we have

$$\begin{aligned} \nabla D_\Phi(z, x)^T(y - z) &= \nabla\Phi(z)^T(y - z) - \nabla\Phi(x)^T(y - z) \\ &< \Phi(y) - \Phi(z) - \nabla\Phi(x)^T(y - z) \\ &= D_\Phi(y, x) - D_\Phi(z, x), \end{aligned}$$

for all $x, y \neq z \in \mathcal{D}$. □

Given a mirror map $\Phi$, we define *projected mirror descent* on $\mathcal{D}$ with constraint set $\mathcal{C}$, initialisation $w_1$ and stepsize $(\eta_t)_{t\in\mathbb{N}}$ as follows [38], [46, Lecture 10].

**Algorithm 2:** Projected Mirror Descent

**Input:** $W_1 = w_1$, $(\eta_t)_{t\in\mathbb{N}}$;
**for** $t \in \mathbb{N}$ **do**
  $\nabla\Phi(\widetilde{W}_{t+1}) = \nabla\Phi(W_t) - \eta_t \nabla R(W_t)$ ;
  $W_{t+1} = \prod_{\mathcal{C}}^{\Phi}(\widetilde{W}_{t+1})$
**end**

There always exists a unique $\widetilde{W}_{t+1} \in \mathcal{D}$ such that

$$\nabla\Phi(\widetilde{W}_{t+1}) = \nabla\Phi(W_t) - \eta_t \nabla R(W_t),$$

because $\nabla\Phi : \mathcal{D} \to \mathbb{R}^d$ is a bijection (by Definition 2.1 and Lemma 2.2). In practice, one also specifies a stopping time $T$, but here we study the limit as $t \to \infty$. We also consider unconstrained mirror descent, which is the above algorithm without the projection step.

**Example 2.5.** The simplest example of a mirror map is $\Phi : \mathbb{R}^d \to \mathbb{R}$ with

$$\Phi(w) = \frac{1}{2}\|w\|_2^2.$$

Here, $\nabla\Phi(w) = w$ and

$$\begin{aligned} D_\Phi(y, x) &= \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|x\|_2^2 - x^T(y - x) \\ &= \frac{1}{2}\|y - x\|_2^2. \end{aligned}$$

Algorithm 2 reduces to projected gradient descent, where the projection is

$$\prod_{\mathcal{C}}^{\Phi}(x) = \arg\min_{y \in \mathcal{C}} \|y - x\|_2,$$

which is the standard projection in $\mathbb{R}^d$ (see Theorem C.11).

**Example 2.6.** Another important example is

$$\mathcal{D} = \mathbb{R}^d_{>0} := \{w \in \mathbb{R}^d : w_i > 0 \text{ for all } i \in \{1, \dots, d\}\}$$

with the negative entropy mirror map

$$\Phi(w) = \sum_{i=1}^{d} w_i \log(w_i).$$

which is continuously extendable to $\mathbb{R}^d_{\geq 0}$ by 0. Moreover,

$$\nabla\Phi(w)_i = 1 + \log(w_i)$$

for all $i \in \{1, \dots, d\}$ and

$$\begin{aligned} D_\Phi(y, x) &= \sum_{i=1}^{d} y_i \log(y_i) - \sum_{i=1}^{d} x_i \log(x_i) - \sum_{i=1}^{d} (1 + \log(x_i))(y_i - x_i) \\ &= \sum_{i=1}^{d} y_i \log(\frac{y_i}{x_i}) + \sum_{i=1}^{d} x_i - y_i. \end{aligned}$$

The gradient update

$$\nabla\Phi(\widetilde{W}_{t+1}) = \nabla\Phi(W_t) - \eta_t \nabla R(W_t)$$

is equivalent to

$$\begin{aligned} \log(\widetilde{W}_{t+1})_i &= \log(W_t)_i - \eta_t \nabla R(W_t)_i \\ (\widetilde{W}_{t+1})_i &= (W_t)_i e^{-\eta_t \nabla R(W_t)_i} \end{aligned} \tag{2.1}$$

for all $i \in \{1, \dots, d\}$. This is related to exponentiated gradient descent (Kivinen and Warmuth [28]), which will be illustrated later.

**Example 2.7.** If $\mathbf{c} \in \mathbb{R}^{d\times d}$ is a positive definite symmetric matrix then $\Phi : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\Phi(w) = \frac{1}{2} w^T \mathbf{c} w$$

is a mirror map with

$$\nabla\Phi(w) = \mathbf{c}w$$

and

$$\begin{aligned} D_\Phi(y,x) &= \Phi(y) - \Phi(x) - \nabla\Phi(x)^T(y-x) \\ &= \frac{1}{2}(y^T\mathbf{c}y - x^T\mathbf{c}x - 2x^T\mathbf{c}(y-x)) \\ &= \frac{1}{2}(y-x)^T\mathbf{c}(y-x). \end{aligned}$$

The gradient update is

$$\widetilde{W}_{t+1} = W_t - \eta_t \mathbf{c}^{-1}\nabla R(W_t),$$

which is an example of natural gradient descent (Amari [1]).

Examples of Legendre functions ($\nabla\Phi$ not necessarily surjective) include:

- $\frac{1}{2}\|w\|_p^2 = \frac{1}{2}(\sum_{i=1}^d |w_i|^p)^{\frac{2}{p}}$ for $p \geq 2$,
- $\sum_{i=1}^d e^{w_i}$,
- $\sum_{i=1}^d \cosh(w_i)$.

**Definition 2.8.** Given an algorithm with iterates $W_t \in \mathbb{R}^d$ to minimise (1.4), we say that the *algorithm converges* if $\lim_{t\to\infty} R(W_t) = 0$.

All convergence in this chapter is pointwise everywhere on the underlying probability space (if treating $W_t$ as random, because it depends on $\mathbf{X}, Y$). Because $\ell$ has a unique root (Definition 1.1), $R(W_t) \to 0$ is equivalent to

$$\begin{aligned} \lim_{t\to\infty} \frac{1}{n}\sum_{i=1}^n \ell(\langle W_t, X_i\rangle, Y_i) = 0 &\iff \lim_{t\to\infty}\langle W_t, X_i\rangle = Y_i \qquad \forall i \in \{1,\dots,n\} \\ &\iff \langle W_\infty, X_i\rangle = Y_i \qquad \forall i \in \{1,\dots,n\} \\ &\iff W_\infty \in \mathcal{G}, \end{aligned} \tag{2.2}$$

where $W_\infty := \lim_{t\to\infty} W_t$. Sometimes we say the *algorithm converges to* $W_\infty$. For projected mirror descent on $\mathcal{D}$, we assume

$$W_\infty \in \mathcal{G} \cap \mathcal{D}, \tag{2.3}$$

because $W_\infty \in \partial\mathcal{D}$ is a pathological case. Indeed, if $W_t \to W_\infty \in \partial\mathcal{D}$, then $\|\nabla\Phi(W_t)\|_{\mathcal{D}} \to \infty$ by Definition 2.1. This would not happen in practice, because we are controlling $\|\nabla\Phi(W_t)\|_{\mathcal{D}}$ by the stepsize $\eta_s$ via

$$\nabla\Phi(\widetilde{W}_{t+1}) = \nabla\Phi(W_t) - \eta_s\nabla R(W_t)$$

and we want it to converge. If $\mathcal{C}$ is bounded and $\mathcal{C} \cap \mathcal{D}$ is closed in $\mathbb{R}^d$ then $\|\nabla\Phi(W_t)\|_{\mathcal{D}} \to \infty$ is, in fact, impossible because $\nabla\Phi$ is continuous and hence $\sup_{w\in\mathcal{C}\cap\mathcal{D}} \|\nabla\Phi(w)\|_{\mathcal{D}}$ is finite.

In what follows, we characterise the implicit bias of projected mirror descent for a unique root loss. First, we warm up by known results of Gunasekar et al. [19] in the unconstrained case and for $\mathcal{C} = \{w \in \mathbb{R}^d : \mathbf{g}w = h\}$ with some fixed $\mathbf{g} \in \mathbb{R}^{m\times d}$, $h \in \mathbb{R}^m$. Then, we extend to some new inequality constraint sets $\mathcal{C}$ and, finally, we touch on quite general

$$\mathcal{C} = \{w \in \mathbb{R}^d : c_j(w) = 0, \widetilde{c}_i(w) \leq 0 \text{ for } j \in J, i \in I\}.$$

**Proposition 2.9.** [19, Theorem 1] If unconstrained mirror descent with initialisation $w_1$ and stepsize $(\eta_t)_{t\in\mathbb{N}}$ converges to $W_\infty$, then

$$W_\infty = \operatorname*{arg\,min}_{w\in\mathcal{G}} D_\Phi(w, w_1). \tag{2.4}$$

We use the notation

$$\operatorname*{arg\,min}_{w\in\mathcal{G}} D_\Phi(w, w_1)$$

when we really mean

$$\operatorname*{arg\,min}_{w\in\mathcal{G}\cap\mathcal{D}} D_\Phi(w, w_1).$$

This is justified because $w \mapsto D_\Phi(w, w_1)$ is only defined on $\mathcal{D}$.

*Proof.* (2.4) is a convex minimisation problem (see Definition A.5), because

$$f : w \mapsto D_\Phi(w, w_1) = \Phi(w) - \Phi(w_1) - \nabla\Phi(w_1)^T(w - w_1)$$

is continously differentiable and convex by Lemma 2.4, and

$$\mathcal{G} = \{w \in \mathbb{R}^d : \mathbf{X}w = Y\} = \{w \in \mathbb{R}^d : c_i(w) = 0 \text{ for all } i \in \{1, \dots, n\}\},$$

where $c_i(w) = \langle w, X_i\rangle - Y_i$, is convex. Moreover,

$$\begin{aligned}\nabla f(w) &= \nabla\Phi(w) - \nabla\Phi(w_1),\\ \nabla c_i(w) &= X_i.\end{aligned}$$

Therefore, $W_\infty$ is the unique solution of (2.4) if and only if the KKT conditions of (2.4) are satisfied (see Theorem A.8), which is equivalent to existence of $\mu \in \mathbb{R}^n$ such that

$$\nabla\Phi(W_\infty) - \nabla\Phi(w_1) = \sum_{i=1}^n \mu_i X_i = \mathbf{X}^T\mu \qquad \text{and} \qquad W_\infty \in \mathcal{G} \cap \mathcal{D}. \tag{2.5}$$

We are left to prove that $W_\infty$ satisfies (2.5). Indeed, by definition of unconstrained mirror descent (Algorithm 2),

$$\begin{aligned}\nabla\Phi(W_{t+1}) - \nabla\Phi(W_t) &= -\eta_t \nabla R(W_t)\\ &= -\eta_t \nabla\big(\frac{1}{n}\sum_{i=1}^n \ell(\langle W_t, X_i\rangle, Y_i)\big)\\ &= -\eta_t \frac{1}{n}\sum_{i=1}^n \partial_1\ell(\langle W_t, X_i\rangle, Y_i)X_i,\end{aligned} \tag{2.6}$$

for all $t \in \mathbb{N}$. Consider the finite dimensional vector space

$$\mathcal{M} = \text{Span}\{X_i : i \in \{1, \dots, n\}\} = \text{Im}(\mathbf{X}^T).$$

Equation (2.6) shows $\nabla\Phi(W_{t+1}) - \nabla\Phi(W_t) \in \mathcal{M}$ and by linearity of $\mathcal{M}$, hence

$$\nabla\Phi(W_{t+1}) - \nabla\Phi(w_1) = \sum_{s=1}^{t} \nabla\Phi(W_{s+1}) - \nabla\Phi(W_s) \in \mathcal{M}.$$

Now, $W_{t+1}$ converges to $W_\infty$ by assumption, $\Phi$ is continuously differentiable and $\mathcal{M}$ is closed in $\mathbb{R}^d$ (because $\mathcal{M}$ is finite dimensional, by Kreyszig [30, Theorem 2.4-3]). Thus, we have

$$\nabla\Phi(W_\infty) - \nabla\Phi(w_1) = \lim_{t\to\infty} \nabla\Phi(W_{t+1}) - \nabla\Phi(w_1) \in \mathcal{M}.$$

This is exactly the first condition of (2.5). As the algorithm converges and by assumption (2.3) we have $W_\infty \in \mathcal{G} \cap \mathcal{D}$. This finishes the proof that

$$W_\infty = \arg\min_{w\in\mathcal{G}} D_\Phi(w, w_1).$$

□

In particular, if $w_1 = \arg\min_{\mathcal{D}} \Phi(w)$ then $\nabla\Phi(w_1) = 0$ and

$$W_\infty = \arg\min_{w\in\mathcal{G}} \Phi(w). \tag{2.7}$$

This makes precise the claim about implicit bias of gradient descent from the introduction. Indeed, in the setting of Example 2.5, (2.7) tells us that if unconstrained gradient descent initialised at 0 converges to $W_\infty$, then

$$W_\infty = \arg\min_{w\in\mathcal{G}} \frac{1}{2}\|w\|_2^2.$$

In other words, gradient descent initialised at 0 is implicitly biased towards the interpolator with the smallest $\|\cdot\|_2$ norm.

Strikingly, Proposition 2.9 is completely independent of the loss function, given that it has a unique root. It depends only on the data and the algorithm.

Further, we consider a constraint set $\mathcal{C} = \{w \in \mathbb{R}^d : \mathbf{g}w = h\}$ for some fixed $\mathbf{g} \in \mathbb{R}^{m\times d}$ and $h \in \mathbb{R}^m$ with $\mathcal{C} \cap \mathcal{D} \neq \emptyset$ (note $\mathcal{C}$ is convex and closed). We refer to this as *affine constraints*.

**Proposition 2.10.** [19, Theorem 1a)] If mirror descent with constraint set $\mathcal{C} = \{w \in \mathbb{R}^d : \mathbf{g}w = h\}$, initialisation $w_1$ and stepsizes $(\eta_t)_{t\in\mathbb{N}}$ converges to $W_\infty$, then

$$W_\infty = \arg\min_{w\in\mathcal{G}\cap\mathcal{C}} D_\Phi(w, w_1). \tag{2.8}$$

*Proof.* We refer to some parts of the previous proof, to reduce repetition. Similarly, (2.8) is a convex minimisation problem, so by KKT conditions (Theorem A.8), $W_\infty$ is its unique solution if and only if there exist $\delta \in \mathbb{R}^m$, $\mu \in \mathbb{R}^n$ such that

$$\nabla\Phi(W_\infty) - \nabla\Phi(w_1) = \sum_{i=1}^{n} \mu_i X_i + \sum_{i=1}^{m} \delta_i \mathbf{g}_i \qquad \text{and} \qquad W_\infty \in \mathcal{G} \cap \mathcal{C} \cap \mathcal{D} \tag{2.9}$$

The difference here is that we have projections. By definition of Algorithm 2,

$$\begin{aligned} W_{t+1} = \prod_{\mathcal{C}}^{\Phi}(\widetilde{W}_{t+1}) &= \operatorname*{arg\,min}_{y \in \mathcal{C}} D_\Phi(y, \widetilde{W}_{t+1}) \\ &= \operatorname*{arg\,min}_{\{w : \mathbf{g}w = h\}} \Phi(y) - \Phi(\widetilde{W}_{t+1}) - \nabla\Phi(\widetilde{W}_{t+1})^T(y - \widetilde{W}_{t+1}), \end{aligned} \tag{2.10}$$

for all $t \in \mathbb{N}$. Therefore, by applying the KKT conditions to this minimisation subproblem (2.10), there exists $\mu_t \in \mathbb{R}^m$ such that

$$\begin{aligned} \nabla D_\Phi(W_{t+1}, \widetilde{W}_{t+1}) &= \mathbf{g}^T \mu_t \\ \nabla\Phi(W_{t+1}) - \nabla\Phi(\widetilde{W}_{t+1}) &= \mathbf{g}^T \mu_t, \end{aligned}$$

where by definition,

$$\nabla\Phi(\widetilde{W}_{t+1}) = \nabla\Phi(W_t) - \eta_t \nabla R(W_t).$$

Combining this gives

$$\nabla\Phi(W_{t+1}) - \nabla\Phi(W_t) = \mathbf{g}^T \mu_t - \eta_t \nabla R(W_t).$$

Now, we only repeat steps analogous to the previous proof. As before, $\nabla R(W_t) \in \mathcal{M}$ (equation (2.6)) so that

$$\nabla\Phi(W_{t+1}) - \nabla\Phi(W_t) \in \operatorname{Span}\{X_i, \mathbf{g}_i : i \in \{1, \dots, n\}\}$$

and by using a telescoping sum

$$\nabla\Phi(W_{t+1}) - \nabla\Phi(w_1) \in \operatorname{Span}\{X_i, \mathbf{g}_i : i \in \{1, \dots, n\}\}.$$

As $\operatorname{Span}\{X_i, \mathbf{g}_i : i \in \{1, \dots, n\}\}$ is closed in $\mathbb{R}^d$ (finite dimensional) and by continuity of $\nabla\Phi$ we have

$$\nabla\Phi(W_\infty) - \nabla\Phi(w_1) \in \operatorname{Span}\{X_i, \mathbf{g}_i : n \in \{1, \dots, n\}\}.$$

Moreover, $W_\infty \in \mathcal{C}$ because $W_t \in \mathcal{C}$ for all $t \in \mathbb{N}$ (as $W_t$ is defined by the projection onto $\mathcal{C}$) and as $w \mapsto \mathbf{g}w$ is continuous. $W_\infty \in \mathcal{G} \cap \mathcal{D}$ follows by assumption. This proves that $W_\infty$ satisfies (2.9) and therefore is the unique solution to (2.8). □

An application of interest is exponentiated gradient descent [28], whose update step is

$$W_t \mapsto W_{t+1},$$

with

$$(W_{t+1})_i = \frac{(W_t)_i e^{-\eta_t \nabla R(W_t)_i}}{\sum_{j=1}^d (W_t)_j e^{-\eta_t \nabla R(W_t)_j}} \tag{2.11}$$

for all $i \in \{1, \dots, d\}$. In Example 2.6 we had $\mathcal{D} = \mathbb{R}^d_{>0}$ and $\Phi(w) = \sum_{i=1}^d w_i \log(w_i)$, where the unconstrained gradient update (equation (2.1)) was

$$(\widetilde{W}_{t+1})_i = (W_t)_i e^{-\eta_t \nabla R(W_t)_i}. \tag{2.12}$$

Now, if we choose $\mathcal{C} = \{w \in \mathbb{R}^d : \mathbf{1}^T w = 1\}$, where $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^d$, then $\mathcal{C} \cap \mathcal{D}$ is the probability simplex

$$\mathcal{C} \cap \mathcal{D} = \{w \in (0,1]^d : \sum_{i=1}^d w_i = 1\}$$

and one can show that, in this setting, the projection step of mirror descent is

$$\mathcal{D} \ni w \mapsto \frac{w}{\|w\|_1} = \frac{w}{\sum_{i=1}^d w_i} \in \mathcal{C} \cap \mathcal{D}. \tag{2.13}$$

Combining (2.13) and (2.12) shows that projected mirror descent with this mirror map and constraint set is precisely exponentiated gradient descent (2.11). Hence, by choosing $\mathbf{g} = \mathbf{1} \in \mathbb{R}^d$ and $h = 1 \in \mathbb{R}$, Proposition 2.10 characterises implicit bias of exponentiated gradient descent on $\mathbb{R}^d_{>0}$.

**Remark 2.11.** In practice, computing $\nabla R(W_t)$ can be expensive for large $n$, which is why stochastic versions of algorithms are popular. We discuss the stochastic version where we randomly sample a single data point at every iteration. However, the following would similarly work for other versions (e.g. sampling a mini-batch). Recall

$$R(w) = \frac{1}{n} \sum_{i=1}^n \ell(\langle w, X_i \rangle, Y_i)$$

and denote

$$\nabla R(w)^i = \partial_1 \ell(\langle w, X_i \rangle, Y_i) X_i, \tag{2.14}$$

for any $i \in \{1, \dots, n\}$, so that

$$\nabla R(w) = \frac{1}{n} \sum_{i=1}^n \nabla R(w)^i.$$

Let $S = \{X_1, Y_1, \dots, X_n, Y_n\}$, $\sigma(S)$ be the $\sigma$-algebra generated by $S$ and $(I_k)_{k \in \mathbb{N}}$ be a sequence of i.i.d. random variables which are uniformly distributed on $\{1, \dots, n\}$. Moreover, let $I_k$ be independent of $\sigma(S)$ for all $k \in \mathbb{N}$. Then stochastic (projected) mirror descent is defined as follows [46, Lecture 11].

**Algorithm 3:** Stochastic Projected Mirror Descent

**Input:** $W_1 = w_1$ and $(\eta_t)_{t \in \mathbb{N}}$;
**for** $t \in \mathbb{N}$ **do**
  $\nabla \Phi(\widetilde{W}_{t+1}) = \nabla \Phi(W_t) - \eta_t \nabla R^{I_t}(W_t)$ ;
  $W_{t+1} = \prod_{\mathcal{C}}^{\Phi}(\widetilde{W}_{t+1})$
**end**

The results of this section, only use the property of $\nabla R(W_t)$ that

$$\nabla R(W_t) \in \text{Span}\{X_i : i \in \{1, \dots, n\}\}.$$

However, (2.14) implies

$$\nabla R^{I_t}(W_t) \in \text{Span}\{X_i : i \in \{1, \dots, n\}\}$$

also. Hence, the results hold exactly the same in stochastic projected mirror descent.

We have characterised implicit bias of unconstrained mirror descent and with affine constraints. This naturally raises the following ultimate question. Given any convex, closed constraint set $\mathcal{C}$, can we characterise the implicit bias (if it exists) of projected mirror descent? Indeed, this problem is suggested as further research and an open problem by Gunasekar et al. [19].

Our goal is to provide an idea when this is possible for convex sets of the form

$$\mathcal{C} = \{w \in \mathbb{R}^d : c_j(w) = 0, \widetilde{c}_i(w) \leq 0 \text{ for } j \in J, i \in I\}.$$

In what follows, we tackle this for implicit bias of the type

$$W_\infty = \operatorname*{arg\,min}_{w \in \mathcal{G} \cap \mathcal{C}} D_\Phi(w, w_1). \tag{2.15}$$

The rest of this section is, as far as we know, original work.

We start by searching for constraint sets where (2.15) holds. A natural candidate is

$$\mathcal{C} = \{w \in \mathcal{D} : \Phi(w) \leq r\},$$

which is closed and convex because $\Phi$ is convex and continuously extendable to $\overline{\mathcal{D}}$. When $\Phi(w) = \|w\|_2^2/2$, this corresponds to gradient descent projected onto an Euclidean ball.

**Example 2.12.** Let $\mathcal{C} = \overline{B}_1(0) = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$ and $\Phi(w) = \|w\|_2^2/2$. Then there exists an initialisation $w_1$ and stepsize $(\eta_t)_{t \in \mathbb{N}}$ with projected gradient descent converging to $W_\infty \in \mathcal{G} \cap \mathcal{C}$, but

$$W_\infty \neq W^\dagger := \operatorname*{arg\,min}_{\mathcal{G}, \|w\|_2 \leq 1} \|w - w_1\|_2^2.$$

A simple example is a one element dataset $X = (1, 1) \in \mathbb{R}^d, Y = 1 \in \mathbb{R}$, initialisation $w_1 = (3, 1)$ and stepsize $\eta_t = p^t$ where $p \in [1/2, 1)$, see Figure 2.1. Moreover, $W_\infty$ can be empirically seen to depend on the stepsize (on $p \in [1/2, 1)$).

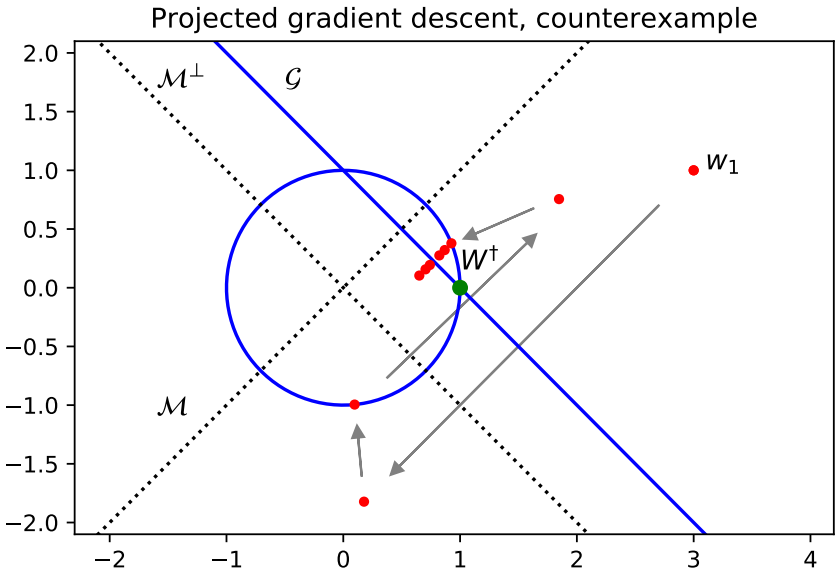


Figure 2.1

However, if $w_1 = (3,3)$ or $w_1 = (0,0)$, the expected result holds, which raises the following question. Are there initialisations for which the implicit bias result holds? Is it possible to characterise them? The key property in Propositions 2.9 and 2.10 was that the gradient update $\nabla\Phi(W_t) \mapsto \nabla\Phi(W_{t+1})$ was always in the same subspace. Here, the projection step makes this property break. In gradient descent projected to $\overline{B}_r(0)$ the projection step is

$$W_{t+1} = \begin{cases} \widetilde{W}_{t+1} & \text{if } \|\widetilde{W}_{t+1}\|_2 \leq r \\ \frac{\widetilde{W}_{t+1}}{\|\widetilde{W}_{t+1}\|_2} r & \text{otherwise,} \end{cases}$$

where

$$\widetilde{W}_{t+1} = W_t - \eta_t \nabla R(W_t).$$

We want $\frac{\widetilde{W}_{t+1}}{\|\widetilde{W}_{t+1}\|_2} r - W_t$ to be in the same subspace for all $t \in \mathbb{N}$. This provides intuition for the following result, which generalizes this idea to mirror descent.

**Proposition 2.13.** If projected mirror descent with constraint set $\mathcal{C} = \{w \in \mathcal{D} : \Phi(w) \leq r\}$ such that $\mathcal{G} \cap \{w \in \mathcal{D} : \Phi(w) < r\} \neq \emptyset$, initialisation $w_1$ and stepsizes $(\eta_t)_{t\in\mathbb{N}}$ converges to $W_\infty$ and

$$\nabla\Phi(w_1) \in \operatorname{Span}\{X_i : i \in \{1,\dots,n\}\}$$

then

$$W_\infty = \operatorname*{arg\,min}_{w \in \mathcal{G}\cap\mathcal{C}} D_\Phi(w, w_1).$$

Conversely, if

$$\nabla\Phi(w_1) \notin \operatorname{Span}\{X_i : i \in \{1,\dots,n\}\}$$

and $\exists t \in \mathbb{N}$ such that $W_t \neq \widetilde{W}_t$ then

$$W_\infty \neq \operatorname*{arg\,min}_{w \in \mathcal{G} \cap \mathcal{C}} D_\Phi(w, w_1) \text{ or } \Phi(W_\infty) = r.$$

This proposition provides an almost if and only if condition on the initialisation of the algorithm under which the implicit bias result holds. The condition $\exists t \in \mathbb{N}$ such that $W_t \neq \widetilde{W}_t$ only assures that the whole algorithm is not identical to unconstrained mirror descent, which is already covered by Proposition 2.9.

*Proof.* By definition,

$$\begin{aligned} W_{t+1} &= \operatorname*{arg\,min}_{w \in \mathcal{C}} D_\Phi(w, \widetilde{W}_{t+1}) \\ &= \operatorname*{arg\,min}_{\Phi(w) \leq r} D_\Phi(w, \widetilde{W}_{t+1}), \end{aligned} \tag{2.16}$$

for $t \in \mathbb{N}$. We apply KKT conditions (Theorem A.8) to the convex minimisation problem (2.16). Slater's condition (Definition A.7), holds as $\mathcal{C} \cap \mathcal{D}$ has nonempty interior by assumption. Hence, there exists $\lambda_t \geq 0$ such that

$$\begin{aligned} \nabla D_\Phi(W_{t+1}, \widetilde{W}_{t+1}) &= -\lambda_t \nabla \Phi(W_{t+1}) \\ \nabla \Phi(W_{t+1}) - \nabla \Phi(\widetilde{W}_{t+1}) &= -\lambda_t \nabla \Phi(W_{t+1}) \\ \nabla \Phi(W_{t+1}) &= \frac{1}{1+\lambda_t} \nabla \Phi(\widetilde{W}_{t+1}) \\ &= \frac{1}{1+\lambda_t} \big( \nabla \Phi(W_t) - \eta_t \nabla R(W_t) \big). \end{aligned} \tag{2.17}$$

Therefore, inductively we have

$$\nabla \Phi(W_{t+1}) = \prod_{s=1}^{t} \frac{1}{1+\lambda_s} \nabla \Phi(w_1) - \xi_t, \tag{2.18}$$

where $\xi_t \in \mathcal{M}$ because $\nabla R(W_s) \in \mathcal{M}$ for all $s \in \{1, \ldots, t\}$. Moreover,

$$\prod_{s=1}^{t} \frac{1}{1+\lambda_s}$$

is nonincreasing and bounded in $(0, 1]$ because $\lambda_s \geq 0$ for all $s \in \mathbb{N}$. Therefore, there exists $\lambda_\infty \geq 0$ such that

$$\lim_{t \to \infty} \prod_{s=1}^{t} \frac{1}{1+\lambda_s} = \frac{1}{1+\lambda_\infty}. \tag{2.19}$$

As $\mathcal{M} = \operatorname{Span}\{X_i : i \in \{1, \ldots, n\}\}$ is closed, $\nabla \Phi$ is continuous and $W_t \to W_\infty \in \mathcal{G} \cap \mathcal{D}$, there exists $\xi \in \mathcal{M}$ such that

$$\nabla \Phi(W_\infty) = \lim_{t \to \infty} \nabla \Phi(W_{t+1}) = \frac{1}{1+\lambda_\infty} \nabla \Phi(w_1) - \xi. \tag{2.20}$$

Moreover, $W_\infty \in \mathcal{C}$ because $W_t \in \mathcal{C}$ for all $t \in \mathbb{N}$ and $\Phi$ is continuous. The KKT conditions for the convex minimisation problem

$$\operatorname*{arg\,min}_{w \in \mathcal{G} \cap \mathcal{C}} D_\Phi(w, w_1) = \operatorname*{arg\,min}_{\mathbf{X}w = Y, \Phi(w) \leq r} D_\Phi(w, w_1)$$

give that $W^\dagger = \arg\min_{w \in \mathcal{G} \cap \mathcal{C}} D_\Phi(w, w_1)$ if and only if $W^\dagger \in \mathcal{G} \cap \mathcal{C} \cap \mathcal{D}$ and there exist $\lambda \geq 0$, $\mu \in \mathbb{R}^n$ such that

$$\nabla\Phi(W^\dagger) - \nabla\Phi(w_1) = \sum_{i=1}^n \mu_i X_i - \lambda \nabla\Phi(W^\dagger) \qquad \text{and} \qquad \lambda(\Phi(W^\dagger) - r) = 0$$

$$\nabla\Phi(W^\dagger) = \frac{1}{1+\lambda}\big(\nabla\Phi(w_1) + \sum_{i=1}^n \mu_i X_i\big) \qquad \text{and} \qquad \lambda(\Phi(W^\dagger) - r) = 0.$$

Slater's condition holds because $\mathcal{G} \cap \{w \in \mathcal{D} : \Phi(w) < r\} \neq \emptyset$. We already know that $W_\infty \in \mathcal{G} \cap \mathcal{C} \cap \mathcal{D}$, hence by equation (2.20), $W_\infty = \arg\min_{w \in \mathcal{G} \cap \mathcal{C}} D_\Phi(w, w_1)$ if and only if there exist $\lambda \geq 0$, $\mu \in \mathbb{R}^n$ such that

$$\nabla\Phi(w_1)\big(\frac{1}{1+\lambda_\infty} - \frac{1}{1+\lambda}\big) = \frac{1}{1+\lambda}\sum_{i=1}^n \mu_i X_i + \xi \tag{2.21}$$

and

$$\lambda(\Phi(W_\infty) - r) = 0. \tag{2.22}$$

Now, if $\nabla\Phi(w_1) \in \mathcal{M}$, then after choosing $\lambda = 0$ there exists $\mu \in \mathbb{R}^n$ such that (2.21),(2.22) are satisfied because

$$1/(1+\lambda_\infty)\nabla\Phi(w_1) - \xi \in \mathcal{M},$$

which proves the first implication. Conversely, assume $\nabla\Phi(w_1) \notin \mathcal{M}$. Then (2.21) cannot be satisfied unless

$$\frac{1}{1+\lambda_\infty} - \frac{1}{1+\lambda} = 0 \tag{2.23}$$

because

$$\frac{1}{1+\lambda}\sum_{i=1}^n \mu_i X_i - \xi \in \mathcal{M}.$$

(2.23) is true only if $\lambda = \lambda_\infty$, as $\lambda, \lambda_\infty \in [0, \infty)$. Hence, for (2.22) we need $\lambda = \lambda_\infty = 0$ or $\Phi(W_\infty) - r = 0$.

If $\lambda = \lambda_\infty = 0$, then (2.19) implies $\lambda_t = 0$ for all $t \in \mathbb{N}$ which is equivalent to

$$W_{t+1} = \widetilde{W}_{t+1} \qquad \forall t \in \mathbb{N}$$

by equation (2.17) and injectivity of $\nabla\Phi$ (Lemma 2.2). This proves the second implication. □

The second implication says that, provided $\Phi(W_\infty) \neq r$, if the algorithm does a single projection we immediately know that it will not converge to the minimum Bregman divergence classifier with respect to $w_1$.

**Remark 2.14.** The proof also implies that if $\Phi(W_\infty) = r$, then $W_\infty = W^\dagger$ (by choosing $\lambda = \lambda_\infty$). However, we do not apriori know if $\Phi(W_\infty) = r$, so this is not very helpful if we want to determine the limit from the initial conditions.

Is the condition $\Phi(W_\infty) \neq r$ actually needed in Proposition 2.13? It is indeed not immediate to find a nontrivial example where $\Phi(W_\infty) = r$, $\nabla\Phi(w_1) \notin \text{Span}\{X_i : i \in \{1, \dots, n\}\}$ and

$$W_\infty = \underset{w \in \mathcal{G} \cap \mathcal{C}}{\arg\min}\, D_\Phi(w, w_1).$$

It is also intriguing to ask whether Proposition 2.13 is "continuous" with respect to $\nabla\Phi(w_1)$. We tackle these questions in Appendix B for projected gradient descent, along with one more implicit bias result .

Can we generalise further? We would like to reason about convex sets of the form

$$\mathcal{C} = \{w \in \mathbb{R}^d : c_j(w) = 0, \widetilde{c}_i(w) \leq 0 \text{ for } j \in J, i \in I\}.$$

Proposition 2.13 brings us closer to understanding

$$\mathcal{C} = \{w \in \mathbb{R}^d : c(w) \leq 0\},$$

for convex $c$. To generalise, it is often beneficial to understand particular examples. Therefore, we examine Example 2.6 and the $\|\cdot\|_1$ ball,

$$\mathcal{C} = \{w \in \mathbb{R}^d : \|w\|_1 \leq 1\} \cap \mathbb{R}^d_{>0} = \{w \in \mathbb{R}^d_{>0} : \mathbf{1}^T w \leq 1\}.$$

Here, an analogous result holds.

**Proposition 2.15.** If projected mirror descent on $\mathcal{D} = \mathbb{R}^d_{>0}$ with the negative entropy mirror map, constraint set $\mathcal{C} = \{w \in \mathbb{R}^d : \|w\|_1 \leq r\}$ such that $\mathcal{G} \cap \{w \in \mathbb{R}^d_{>0} : \mathbf{1}^T w < r\} \neq \emptyset$, initialisation $w_1$ and stepsizes $(\eta_t)_{t \in \mathbb{N}}$ converges to $W_\infty$ and

$$\mathbf{1} \in \text{Span}\{X_i : i \in \{1, \dots, n\}\}$$

then

$$W_\infty = \underset{w \in \mathcal{G} \cap \mathcal{C}}{\arg\min}\, D_\Phi(w, w_1).$$

Conversely, if

$$\mathbf{1} \notin \text{Span}\{X_i : i \in \{1, \dots, n\}\}$$

and $\exists t \in \mathbb{N}$ such that $W_t \neq \widetilde{W_t}$ then

$$W_\infty \neq \underset{w \in \mathcal{G} \cap \mathcal{C}}{\arg\min}\, D_\Phi(w, w_1) \text{ or } \|W_\infty\|_1 = r.$$

The negative entropy mirror map is of interest, but actually, the mirror map can be arbitrary here (as long as $\mathcal{D} = \mathbb{R}^d_{>0}$). Importantly, this is not a corollary of Proposition 2.13 because $\mathcal{C}$ is not defined in terms of $\Phi$. However, similarities

between Proposition 2.13 and Proposition 2.15 are evident. In both, $\nabla\Phi(w_1)$ and $\mathbf{1}$ seem to be crucial, while also corresponding to $\nabla c(w_1)$ in

$$\mathcal{C} = \{w \in \mathcal{D} : c(w) \leq 0\},$$

where $c \in \{\Phi, w \mapsto \mathbf{1}^T w\}$. Does a generalised result holds with $\nabla\Phi(w_1)$ and $\mathbf{1}$ replaced by $\nabla c(w_1)$?

Unfortunately, the answer is no. If $c : \mathbb{R}^d \to \mathbb{R}$ is a convex continuously differentiable function and $\mathcal{C} = \{w \in \mathbb{R}^d : c(w) \leq r\}$, then the analogue to

$$\begin{aligned} \nabla\phi(w_1) &\in \text{Span}\{X_i : i \in \{1, \dots, n\}\} \\ \mathbf{1} &\in \text{Span}\{X_i : i \in \{1, \dots, n\}\} \end{aligned}$$

which we had in Propositions 2.13 and 2.15 is

$$\nabla c(W_t) \in \text{Span}\{X_i : i \in \{1, \dots, n\}\} \text{ for all } t \in \mathbb{N}, \tag{2.24}$$

which then similarly implies

$$W_\infty = \underset{w \in \mathcal{G} \cap \mathcal{C}}{\arg\min}\, D_\Phi(w, w_1).$$

This is not very interesting, because (2.24) depends on the behaviour of the iterates and hence is not a reasonable condition to satisfy apriori. This reduced to something reasonable for $\{w \in \mathcal{D} : \Phi(w) \leq r\}$ because there, as a special case,

$$\nabla\Phi(W_t) \in \text{Span}(\nabla\Phi(w_1)) + \text{Span}\{X_i : i \in \{1, \dots, n\}\} \tag{2.25}$$

for all $t \in \mathbb{N}$, by equation (2.18). Hence (2.24) was equivalent to

$$\nabla\Phi(w_1) \in \text{Span}\{X_i : i \in \{1, \dots, n\}\}.$$

In Proposition 2.15 we had

$$\nabla c(W_t) = r\mathbf{1},$$

for all $t \in \mathbb{N}$. In summary, for

$$\mathcal{C} = \{w \in \mathbb{R}^d : c(w) \leq 0\},$$

implicit bias of the form

$$W_\infty = \underset{w \in \mathcal{G} \cap \mathcal{C}}{\arg\min}\, D_\Phi(w, w_1)$$

does not generally hold just based on properties of the initialisation. For special functions $c : \mathbb{R}^d \to \mathbb{R}$, when (2.25) is simplified it can hold. For example when $c$ is affine (Proposition 2.15) or defined in terms of $\Phi$ (Prosposition 2.13).

We do not provide precise proofs for Proposition 2.15 and (2.25) because of repetition of technique and the word limit.

What can we say about

$$\mathcal{C} = \{w \in \mathbb{R}^d : c_j(w) = 0, \widetilde{c}_i(w) \leq 0 \text{ for } j \in J, i \in I\}?$$

Here, $I, J$ are index sets and for $\mathcal{C}$ to be convex, as standard in convex optimisation (Boyd and Vandenberghe [11, Chapter 1]), we assume $c_j$ to be affine and $\widetilde{c}_i$ convex (this is not strictly necessary, see quasiconvexity in [11, Chapter 3]). Affine $c_j$ are covered by Proposition 2.10, so this essentially reduces to

$$C = \{w \in \mathbb{R}^d : \widetilde{c}_i(w) \leq 0 \text{ for } i \in I\}. \tag{2.26}$$

An analysis of (2.26) is only a generalisation of

$$\mathcal{C} = \{w \in \mathbb{R}^d : c(w) \leq 0\}.$$

Similarly, one can show

$$W_\infty = \arg\min_{w \in \mathcal{G} \cap \mathcal{C}} D_\Phi(w, w_1)$$

does not hold under reasonable conditions.

Importantly, however, this answers the question for implicit bias of the form

$$W_\infty = \arg\min_{w \in \mathcal{G} \cap \mathcal{C}} D_\Phi(w, w_1),$$

but does not rule out existence of a different implicit bias for specific examples of

$$\mathcal{C} = \{w \in \mathbb{R}^d : c(w) \leq 0\}.$$

## 2.2 Convergence of Gradient Descent

So far, we have studied implicit bias assuming that an algorithm converges. But does it actually converge? This is a fundamental question. We connect it with implicit bias and prove convergence of unconstrained gradient descent to the minimum norm solution. We also extend with an original result, to our best knowledge, by adding an increasing differentiable nonlinearity.

In this section, let the hypothesis class be

$$\mathcal{H} = \{\mathbb{R}^d \ni x \mapsto \psi(\langle w, x \rangle) \, : \, w \in \mathbb{R}^d\},$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is an increasing differentiable function, referred to as nonlinearity. We still assume

$$\mathcal{G} = \{w \in \mathbb{R}^d : \psi(\mathbf{X}w) = Y\} \neq \emptyset.$$

For now, we make the restrictive assumption that there exist $0 < \gamma \leq \Gamma$ with

$$\gamma \leq \psi'(x) \leq \Gamma \tag{2.27}$$

for all $x \in \mathbb{R}$. For example, the Leaky ReLU nonlinearity,

$$\psi(x) = \begin{cases} x & x \geq 0 \\ 0.01x & x < 0, \end{cases}$$

satisfies this. However, for many common nonlinearities,

$$\begin{aligned} \text{Sigmoid}(x) &= \frac{1}{1+e^{-x}} \\ \text{ReLU}(x) &= \max(0, x) \\ \text{Softplus}(x) &= \log(1+e^{x}) \\ \text{Tanh}(x) &= \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} m \end{aligned}$$

(Teh [53, p. 51]) it is false because

$$\lim_{x \to -\infty} \psi'(x) = 0.$$

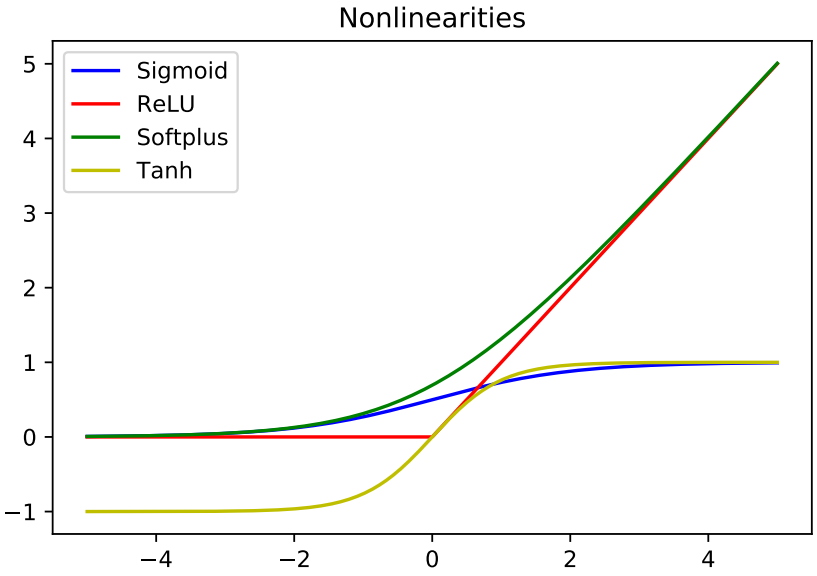


Figure 2.2

Our contribution is in removing this restriction, later on. The following is due to Oymak and Soltanolkotabi [43] and inspired by Rebeschini [46, Lecture 14].

**Proposition 2.16.** [43, Theorem 4.1] If there exist $0 < \gamma \leq \Gamma$ such that $\gamma \leq \psi'(x) \leq \Gamma$ for all $x \in \mathbb{R}$ then unconstrained gradient descent initialized at $w_1$ with constant stepsize $\eta \leq \frac{n}{2\Gamma^2 \|\mathbf{X}\|_2^2}$ converges to

$$W_\infty = \arg\min_{w \in \mathcal{G}} \|w_1 - w\|_2.$$

*Proof.* $\psi$ is increasing and therefore invertible. Hence

$$\begin{aligned}\mathcal{G} &= \{w \in \mathbb{R}^d : \psi(\mathbf{X}w) = Y\} \\ &= \{w \in \mathbb{R}^d : \mathbf{X}w = \psi^{-1}(Y)\}\end{aligned}$$

where $\psi$ and $\psi^{-1}$ act on vectors elementwise. Denote

$$W^\dagger = \arg\min_{w \in \mathcal{G}} \|w_1 - w\|_2.$$

$\mathcal{G}$ is nonempty (by assumption), closed and convex. Hence

$$W^\dagger = P_{\mathcal{G}}(w_1),$$

where $P_{\mathcal{G}} : \mathbb{R}^d \to \mathcal{G}$ is the projection map defined by the Projection Theorem (Theorem C.11). Moreover, by Proposition C.14,

$$W^\dagger := \mathbf{X}^\dagger \psi^{-1}(Y) + w_1 - \mathbf{X}^\dagger \mathbf{X} w_1, \tag{2.28}$$

where $\mathbf{X}^\dagger \in \mathbb{R}^{d \times n}$ is the Moore-Penrose pseudoinverse [44] (see Definition C.6 and an explicit form in Proposition C.7). The update of gradient descent reads

$$\begin{aligned}W_{t+1} &= W_t - \eta_t \nabla R(W_t) \\ &= W_t - \eta_t \frac{2}{n} \mathbf{X}^T \mathrm{diag}(\psi'(\mathbf{X}W_t))(\psi(\mathbf{X}W_t) - Y),\end{aligned} \tag{2.29}$$

where we used the chain rule applied to $R(w) = \frac{1}{n}\|\psi(\mathbf{X}w) - Y\|_2^2$ and that the derivative of the map $\mathbb{R}^n \ni z \mapsto \psi(z) = (\psi(z_1), \ldots, \psi(z_n)) \in \mathbb{R}^n$ is

$$\begin{pmatrix} \psi'(z_1) & 0 & 0 & \ldots \\ 0 & \psi'(z_2) & 0 & \ldots \\ & & \ddots & \\ 0 & 0 & \ldots & \psi'(z_n) \end{pmatrix} =: \mathrm{diag}(\psi'(z)).$$

As $W^\dagger \in \mathcal{G}$, we have

$$\psi(\mathbf{X}W_t) - Y = \psi(\mathbf{X}W_t) - \psi(\mathbf{X}W^\dagger)$$

and by the Mean Value Theorem applied to $\psi$ componentwise, there exists $\xi \in \mathbb{R}^n$ such that

$$\psi(\mathbf{X}W_t) - \psi(\mathbf{X}W^\dagger) = \begin{pmatrix} \psi(\mathbf{X}W_t)_1 - \psi(\mathbf{X}W^\dagger)_1 \\ \psi(\mathbf{X}W_t)_2 - \psi(\mathbf{X}W^\dagger)_2 \\ \vdots \\ \psi(\mathbf{X}W_t)_n - \psi(\mathbf{X}W^\dagger)_n \end{pmatrix} = \begin{pmatrix} \psi'(\xi_1)(\mathbf{X}(W_t - W^\dagger))_1 \\ \psi'(\xi_2)(\mathbf{X}(W_t - W^\dagger))_2 \\ \vdots \\ \psi'(\xi_n)(\mathbf{X}(W_t - W^\dagger))_n \end{pmatrix},$$

that is

$$\psi(\mathbf{X}W_t) - \psi(\mathbf{X}W^\dagger) = \mathrm{diag}(\psi'(\xi))\mathbf{X}(W_t - W^\dagger).$$

Moreover,

$$\xi \in \mathcal{S}^t := \prod_{i=1}^{n} \{\mu_i (\mathbf{X} W_t)_i + (1-\mu_i)(\mathbf{X} W^\dagger)_i : \mu_i \in [0,1]\}. \tag{2.30}$$

Therefore equation (2.29) reduces to

$$\begin{aligned} W_{t+1} &= W_t - \eta_t \mathbf{M}_t (W_t - W^\dagger) \\ W_{t+1} - W^\dagger &= (I_d - \eta_t \mathbf{M}_t)(W_t - W^\dagger), \end{aligned} \tag{2.31}$$

where

$$\mathbf{M}_t = \frac{2}{n} \mathbf{X}^T \mathrm{diag}(\psi'(\mathbf{X} W_t)) \mathrm{diag}(\psi'(\xi_t)) \mathbf{X} \in \mathbb{R}^{d \times d} \tag{2.32}$$

is symmetric positive semi-definite (because $\psi'(x) > 0$). To prove convergence we show that (2.31) is a contraction. For this, we need $\|I_d - \eta_t \mathbf{M}_t\|_2 \in (0,1)$ (see Appendix C). However, $I_d - \eta_t \mathbf{M}_t$ is not necessarily positive semi-definite and its largest eigenvalue is 1 (because $d > n$ and $\emptyset \neq \mathrm{Ker}(\mathbf{X}) \subseteq \mathrm{Ker}(\mathbf{M}_t)$). To make $I_d - \eta_t \mathbf{M}_t$ positive definite we just make $\eta_t$ small enough, but getting the eigenvalues of $I_d - \eta_t \mathbf{M}_t$ to be less than 1 is not immediate. For this we transform equation (2.31) into

$$\widehat{W}_{t+1} - \widehat{W}^\dagger = (I_r - \eta_t \widehat{\mathbf{M}}_t)(\widehat{W}_t - \widehat{W}^\dagger), \tag{2.33}$$

where $\|\widehat{W}_t - \widehat{W}^\dagger\|_2 = \|W_t - W^\dagger\|_2$ and $\widehat{\mathbf{M}}_t$ has positive eigenvalues uniformly lower bounded by $\delta \in (0,1)$. After this, we are done because $\|I_r - \eta_t \widehat{\mathbf{M}}_t\|_2 \in (0, 1 - \eta_t \delta)$ and hence

$$\|W_{t+1} - W^\dagger\|_2 = \|\widehat{W}_{t+1} - \widehat{W}^\dagger\|_2 \leq \|I_r - \eta_t \widehat{\mathbf{M}}_t\|_2 \|\widehat{W}_t - \widehat{W}^\dagger\|_2 \leq \tag{2.34}$$

$$\ldots \leq \prod_{s=1}^{t} (1 - \eta_s \delta) \|\widehat{w}_1 - \widehat{W}^\dagger\|_2. \tag{2.35}$$

The right hand side converges to zero if we do not decrease $\eta_t$ too fast.

Now, we derive equation (2.33). If $\mathbf{X} = \mathbf{U}_{1:r} \mathbf{\Sigma}_{1:r} \mathbf{V}_{1:r}^T$ is the compact SVD of $\mathbf{X}$ (see Theorem C.5), define $\widehat{\mathbf{X}} = \mathbf{X} \mathbf{V}_{1:r} \in \mathbb{R}^{n \times r}$, $\widehat{\mathbf{M}}_t = \mathbf{V}_{1:r}^T \mathbf{M}_t \mathbf{V}_{1:r} \in \mathbb{R}^{r \times r}$ and $\widehat{w} = \mathbf{V}_{1:r}^T w \in \mathbb{R}^r$, for any $w \in \mathbb{R}^d$.

**Claim 1.** For all $t \in \mathbb{N}$, $W_t - W^\dagger \in \mathrm{Im}(\mathbf{X}^T)$.

By definition of gradient descent,

$$W_t = w_1 - \xi_t,$$

where $\xi_t = \sum_{s=1}^{t-1} \eta_s \nabla R(W_s) \in \mathrm{Im}(\mathbf{X}^T)$. Hence, by (2.28),

$$\begin{aligned} W_t - W^\dagger &= w_1 - \xi_t - (\mathbf{X}^\dagger \psi^{-1}(Y) + w_1 - \mathbf{X}^\dagger \mathbf{X} w_1) \\ &= \mathbf{X}^\dagger \mathbf{X} w_1 - \mathbf{X}^\dagger \psi^{-1}(Y) - \xi_t. \end{aligned}$$

As $\mathbf{X}^\dagger = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^\dagger$ (see Proposition C.7), the last equation shows $W_t - W^\dagger \in \mathrm{Im}(\mathbf{X}^T)$.

Moreover, as $w \mapsto \mathbf{V}_{1:r}\mathbf{V}_{1:r}^T w = \mathbf{X}^\dagger \mathbf{X} w$ is the projection onto $\mathrm{Im}(\mathbf{X}^T)$ (by Propositions C.12 and C.13), the claim implies

$$\mathbf{V}_{1:r}\mathbf{V}_{1:r}^T(W_t - W^\dagger) = W_t - W^\dagger.$$

Therefore, multiplying equation (2.31) by $\mathbf{V}_{1:r}^T$ on the left gives

$$\begin{aligned}
\widehat{W}_{t+1} - \widehat{W}^\dagger &= (\widehat{W}_t - \widehat{W}^\dagger) - \eta_t \mathbf{V}_{1:r}^T \mathbf{M}_t (W_t - W^\dagger) \\
&= (\widehat{W}_t - \widehat{W}^\dagger) - \eta_t \mathbf{V}_{1:r}^T \mathbf{M}_t \mathbf{V}_{1:r}\mathbf{V}_{1:r}^T (W_t - W^\dagger) \\
&= (I_r - \eta_t \widehat{\mathbf{M}}_t)(\widehat{W}_t - \widehat{W}^\dagger),
\end{aligned}$$

which is exactly equation (2.33). Moreover,

$$\begin{aligned}
\|\widehat{W}_t - \widehat{W}^\dagger\|_2^2 &= (\widehat{W}_t - \widehat{W}^\dagger)^T(\widehat{W}_t - \widehat{W}^\dagger) \\
&= (W_t - W^\dagger)^T \mathbf{V}_{1:r}\mathbf{V}_{1:r}^T (W_t - W^\dagger) \\
&= (W_t - W^\dagger)^T (W_t - W^\dagger) \\
&= \|W_t - W^\dagger\|_2^2.
\end{aligned}$$

Now, we prove $\|I_d - \eta_t \widehat{\mathbf{M}}_t\|_2 \in (0, 1 - \eta_t \delta)$. For this we need the following claim.

**Claim 2.** The eigenvalues of $\widehat{\mathbf{X}}^T\widehat{\mathbf{X}} \in \mathbb{R}^{r\times r}$ are the $r$ positive eigenvalues of $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{d\times d}$.

As $\mathbf{X} = \mathbf{U}_{1:r}\mathbf{\Sigma}_{1:r}\mathbf{V}_{1:r}^T$, $\widehat{\mathbf{X}} = \mathbf{X}\mathbf{V}_{1:r}$ and $\mathbf{U}_{1:r}^T\mathbf{U}_{1:r} = \mathbf{V}_{1:r}^T\mathbf{V}_{1:r} = I_r$ (as $\mathbf{U}, \mathbf{V}$ are orthogonal),

$$\begin{aligned}
\mathbf{X}^T\mathbf{X} &= \mathbf{V}_{1:r}\mathbf{\Sigma}_{1:r}^2\mathbf{V}_{1:r} \\
\widehat{\mathbf{X}}^T\widehat{\mathbf{X}} &= \mathbf{\Sigma}_{1:r}^2,
\end{aligned}$$

which proves the claim.

Now, using the bound $0 < \gamma \le \psi'(x) \le \Gamma$ and Proposition C.1, we have for any $v \in \mathbb{R}^r$

$$\begin{aligned}
v^T\widehat{\mathbf{M}}_t v &= \frac{2}{n} v^T \widehat{\mathbf{X}}^T \mathrm{diag}(\psi'(\mathbf{X}W_t))\mathrm{diag}(\psi'(\xi_t))\widehat{\mathbf{X}} v \\
&\le \sum_i^n \frac{2\Gamma^2}{n}(\widehat{\mathbf{X}}v)_i^2 = \frac{2\Gamma^2}{n} v^T\widehat{\mathbf{X}}^T\widehat{\mathbf{X}} v \le \frac{2\Gamma^2}{n}\lambda_{\max}(\widehat{\mathbf{X}}^T\widehat{\mathbf{X}}) v^T I_r v \qquad (2.36)
\end{aligned}$$

and similarly

$$v^T\widehat{\mathbf{M}}_t v \ge \frac{2\gamma^2}{n} v^T\widehat{\mathbf{X}}^T\widehat{\mathbf{X}} v \ge \frac{2\gamma^2}{n}\lambda_{\min}(\widehat{\mathbf{X}}^T\widehat{\mathbf{X}}) v^T I_n v. \qquad (2.37)$$

By Claim 2,

$$\lambda_{\max}(\widehat{\mathbf{X}}^T\widehat{\mathbf{X}}) = \lambda_{\max}(\mathbf{X}^T\mathbf{X}) = \|\mathbf{X}\|_2^2$$
$$\lambda_{\min}(\widehat{\mathbf{X}}^T\widehat{\mathbf{X}}) = \lambda_r(\mathbf{X}^T\mathbf{X}) =: \lambda_{\min} > 0.$$

Putting this together, we have

$$(1 - \eta_t \frac{2\Gamma^2}{n}\|\mathbf{X}\|_2^2) \leq \frac{v^T(I_r - \eta_t\widehat{\mathbf{M}}_t)v}{\|v\|_2^2} \leq (1 - \eta_t\frac{2\gamma^2}{n}\lambda_{\min}) < 1.$$

Finally, choosing stepsize $\eta_t < \frac{n}{2\Gamma^2\|\mathbf{X}\|_2^2}$, the lower bound implies $I_n - \eta_t\widehat{\mathbf{M}}_t$ is positive definite and the upper bound (with Propositions (C.1) and (C.3)) implies

$$\|I_r - \eta_t\widehat{\mathbf{M}}_t\|_2 < 1 - \eta_t\delta,$$

where $\delta = \frac{2\gamma^2}{n}\lambda_{\min} > 0$. Therefore, (2.35) finishes the proof if we assume that

$$\lim_{t\to\infty}\prod_{s=1}^{t}(1 - \eta_s\frac{2\gamma^2}{n}\lambda_{\min}) = 0. \tag{2.38}$$

If $\eta_t < \frac{n}{2\Gamma^2\|\mathbf{X}\|_2^2}$ is constant, this is true. □

As mentioned,

$$0 < \gamma \leq \psi'(x) \leq \Gamma \tag{2.39}$$

for all $x \in \mathbb{R}$ is not satisfied by standard nonlinearities. However, for strictly increasing nonlinearities convergence still holds, although (2.39) is not satisfied. We prove this in Proposition 2.18. The rest of this section is, as far as we know, original.

In the following, we only assume $\psi' > 0$. Let $(W_t)_{t\in\mathbb{N}}$ be iterates of gradient descent initialised at $w_1$.

**Lemma 2.17.** There exist $0 < \gamma \leq \Gamma$ such that if $\eta_t \leq \frac{n}{2\Gamma^2\|\mathbf{X}\|_2^2}$ for all $t \in \mathbb{N}$, then

$$\|W_t - W^\dagger\|_2 \leq \|w_1 - W^\dagger\|_2$$

and

$$\gamma \leq \psi'(\xi_i) \leq \Gamma$$

for all $\xi \in \mathcal{S}^t$, $t \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$.

*Proof.* As before, $W^\dagger = P_{\mathcal{G}}(w_1)$. Define

$$r := \|\mathbf{X}\|_2(\|W^\dagger\|_2 + \|W^\dagger - w_1\|_2)$$

and

$$\gamma := \inf\{\psi'(y) : |y| \leq r, y \in \mathbb{R}\}$$
$$\Gamma := \sup\{\psi'(y) : |y| \leq r, y \in \mathbb{R}\}$$

$\{y \in \mathbb{R} : |y| \leq r\}$ is a compact set so that $\Gamma \geq \gamma > 0$ because $\psi'$ is continuous. We proceed proving the lemma by induction. Recall the definition

$$\mathcal{S}^t := \prod_{i=1}^{n}\{\mu_i(\mathbf{X}W_t)_i + (1 - \mu_i)(\mathbf{X}W^\dagger)_i : \mu_i \in [0, 1]\}.$$

- Assume t=1.

For any $\xi \in \mathcal{S}^1$ and $i \in \{1, \dots, n\}$ we have

$$\begin{aligned}
|\xi_i| &\leq |(\mathbf{X}w_1)_i| + |(\mathbf{X}W^\dagger)_i| \\
&\leq \|\mathbf{X}w_1\|_2 + \|\mathbf{X}W^\dagger\|_2 \\
&\leq \|\mathbf{X}\|_2 \max(\|w_1\|_2, \|W^\dagger\|_2) \\
&= \|\mathbf{X}\|_2 \max(\|W^\dagger + (w_1 - W^\dagger)\|_2, \|W^\dagger\|_2) \\
&\leq \|\mathbf{X}\|_2 \max(\|W^\dagger\|_2 + \|w_1 - W^\dagger\|_2, \|W^\dagger\|_2) \\
&= r.
\end{aligned}$$

Therefore, by definition of $\Gamma \geq \gamma > 0$,

$$\gamma \leq \psi'(\xi_i) \leq \Gamma.$$

- Assume that the statement holds for some $t \in \mathbb{N}$. We prove it for $t+1$.

By equation (2.34),

$$\|W_{t+1} - W^\dagger\|_2 = \|\widehat{W}_{t+1} - \widehat{W}^\dagger\|_2 \leq \|I_r - \eta_t \widehat{\mathbf{M}}_t\|_2 \|\widehat{W}_t - \widehat{W}^\dagger\|_2 \tag{2.40}$$

$$= \|I_r - \eta_t \widehat{\mathbf{M}}_t\|_2 \|W_t - W^\dagger\|_2. \tag{2.41}$$

We show that $\|I_r - \eta_t \widehat{\mathbf{M}}_t\|_2 \leq 1$. As $\widehat{\mathbf{M}}_t$ is positive semi-definite ($\psi' > 0$ and (2.32)), it is enough to show $\eta_t \widehat{\mathbf{M}}_t$ has eigenvalues bounded by 1. Indeed, by definition of $\widehat{\mathbf{M}}_t$,

$$v^T \widehat{\mathbf{M}}_t v = \frac{2}{n} v^T \widehat{\mathbf{X}}^T \mathrm{diag}(\psi'(\mathbf{X}W_t)) \mathrm{diag}(\psi'(\xi_t)) \widehat{\mathbf{X}} v.$$

for $v \in \mathbb{R}^r$. By construction in (2.30), $\xi_t \in \mathcal{S}^t$ and also $\mathbf{X}W_t \in \mathcal{S}^t$, hence by inductive assumption

$$\psi'(\mathbf{X}W_t)_i \psi'(\xi_t)_i \leq \Gamma^2$$

for all $i \in \{1, \dots, n\}$. Therefore, if $\eta_t \leq \frac{n}{2\Gamma^2 \|\mathbf{X}\|_2^2}$, then $\eta_t \widehat{\mathbf{M}}_t$ has eigenvalues in $[0, 1]$. Hence

$$\|I_r - \eta_t \widehat{\mathbf{M}}_t\|_2 \leq 1.$$

Plugging this into (2.41) and using the inductive assumption again gives

$$\begin{aligned}
\|W_{t+1} - W^\dagger\|_2 &\leq \|I_r - \eta_t \widehat{\mathbf{M}}_t\|_2 \|W_t - W^\dagger\|_2 \\
&\leq \|w_1 - W^\dagger\|_2.
\end{aligned} \tag{2.42}$$

Moreover, as before, for any $\xi_{t+1} \in \mathcal{S}^{t+1}$ and $i \in \{1, \dots, n\}$,

$$\begin{aligned}
|(\xi_{t+1})_i| &\leq |(\mathbf{X}w_{t+1})_i| + |(\mathbf{X}W^\dagger)_i| \\
&\leq \|\mathbf{X}W_{t+1}\|_2 + \|\mathbf{X}W^\dagger\|_2 \\
&\leq \|\mathbf{X}\|_2 \max(\|W_{t+1}\|_2, \|W^\dagger\|_2) \\
&\leq \|\mathbf{X}\|_2 (\|W^\dagger\|_2 + \|W_{t+1} - W^\dagger\|_2) \\
&\leq r,
\end{aligned}$$

where the last inequality follows from (2.42). Therefore, by definition of $\Gamma \geq \gamma > 0$,

$$\gamma \leq \psi'(\xi_{t+1})_i \leq \Gamma.$$

This finishes the proof by induction. □

This lemma allows us to remove the restriction from Proposition 2.16 and give us a strictly stronger result.

**Proposition 2.18.** For unconstrained gradient descent on linear models with an increasing, continuously differentiable nonlinearity $\psi$, initialized at $w_1$, there exists $\Gamma > 0$ such that if $\eta \leq \frac{n}{2\Gamma^2\|\mathbf{X}\|_2^2}$ is constant then the algorithm converges to

$$W_\infty = \arg\min_{w \in \mathcal{G}} \|w_1 - w\|_2.$$

*Proof.* By Lemma 2.17, there exist $0 < \gamma \leq \Gamma$ such that if $\eta_t \leq \frac{n}{2\Gamma^2\|\mathbf{X}\|_2^2}$ for all $t \in \mathbb{N}$, then

$$\gamma \leq \psi'(\xi_i) \leq \Gamma$$

for all $\xi \in \mathcal{S}^t$, $t \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$. The proof of Propoosition 2.16 only uses this, not necessarily

$$\gamma \leq \psi'(x) \leq \Gamma$$

for all $x \in \mathbb{R}$. Therefore, the proof follows identically to proof of Propoosition 2.16. □

Here, constant stepsize $\eta \leq \frac{n}{2\Gamma^2\|\mathbf{X}\|_2^2}$ can be replaced by (2.38).

# Chapter 3

# Generalisation

## 3.1 The Fundamental Price of Interpolation

Why do we care about implicit bias? Apart from intrinsic interest, it also provides an attempt to explain the mystifying phenomenon that neural networks have very good performance even when they fit training data perfectly. In this section, we try to understand performance (generalisation) in relation to fitting training data perfectly, in linear regression.

Fitting the data perfectly, or interpolating, means that the output of the algorithm, $\hat{W}$, satisfies $\hat{W} \in \mathcal{G}$. The first question that we tackle, is whether good generalisation is even possible for $\hat{W} \in \mathcal{G}$. Our finding is that a neccessary condition is high overparametrisation. The following exposition is adapted from Muthukumar et al. [36].

In this chapter, we use the squared error loss

$$\ell : \mathbb{R}^2 \ni (x, y) \mapsto (x - y)^2,$$

we assume $\text{rank}(\mathbf{X}) = n < d$, (hence also $\mathcal{G} = \{w \in \mathbb{R}^d : \mathbf{X}w = Y\} \neq \emptyset$), $\mathbb{E}(X_i X_i^T) = \mathbf{c} \in \mathbb{R}^{d \times d}$ is positive definite and that there exists $w^* \in \mathbb{R}^d$ (unknown) such that

$$Y_i = w^{*T} X_i + \xi_i \tag{3.1}$$

for all $i \in \{1, \ldots, n\}$, where $\xi = (\xi_1, \ldots, \xi_n) \sim \mathcal{N}(0, \sigma^2 I_n)$. Our goal (recall Chapter 1) is to minimise the population risk,

$$r(w) = \mathbb{E}(w^T \widetilde{X} - \widetilde{Y})^2,$$

or the excess risk

$$r(w) - r(w^*).$$

**Definition 3.1.** We define the *best interpolator*, given our data to be

$$\begin{aligned} W_b &= \arg\min_{w \in \mathcal{G}} r(w) \\ &= \arg\min_{w \in \mathcal{G}} \mathbb{E}(w^T \widetilde{X} - \widetilde{Y})^2 \end{aligned}$$

It is important to conceptually distinguish between $W_b$ and $w^*$. While $w^*$ is the true parameter, $W_b$ is the best interpolator given the information that we possess (the data). Also, $w^* \notin \mathcal{G}$ with probability 1, because of the noise $\xi$. The population risk $r(w)$ satisfies

$$\begin{aligned} r(w) = \mathbb{E}(w^T \widetilde{X} - \widetilde{Y})^2 &= \mathbb{E}(w^T \widetilde{X} - (w^{*T} \widetilde{X} + \widetilde{\xi}))^2 \\ &= \mathbb{E}((w - w^*)^T \widetilde{X} + \widetilde{\xi})^2 \\ &= (w - w^*)^T \mathbf{c} (w - w^*) + \sigma^2, \end{aligned} \tag{3.2}$$

becuase $\widetilde{\xi}$ is centered and independent of $\widetilde{X}$. As $\mathbf{c} \in \mathbb{R}^{d \times d}$ is positive definite (hence, diagonalisable with positive eigenvalues), there exists invertible, symmetric $\mathbf{c}^{\frac{1}{2}}$ with

$$\mathbf{c}^{\frac{1}{2}} \mathbf{c}^{\frac{1}{2}} = \mathbf{c}.$$

Therefore,

$$r(w) = \|\mathbf{c}^{\frac{1}{2}}(w - w^*)\|_2^2 + \sigma^2. \tag{3.3}$$

**Remark 3.2.** Here, an interesting observation arises. Let

$$\Phi(w) = \frac{1}{2} w^T \mathbf{c} w,$$

which is a mirror map on $\mathbb{R}^d$ (Example 2.7). Moreover,

$$\begin{aligned} D_\Phi(w, w^*) &= \frac{1}{2}(w - w^*)^T \mathbf{c} (w - w^*) \\ &= \frac{1}{2}(r(w) - \sigma^2) \\ &= \frac{1}{2}(r(w) - r(w^*)), \end{aligned}$$

which is half of the excess risk. Therefore, an alternative definition of $W_b$ is the Bregman projection of $w^*$ onto $\mathcal{G}$,

$$W_b = \prod_{\mathcal{C}}^{\Phi}(w^*) = \operatorname*{arg\,min}_{w \in \mathcal{G}} D_\Phi(w, w^*), \tag{3.4}$$

This natural connection is intriguing with respect to the implicit bias of mirror descent in Proposition 2.9.

The following proposition characterises the excess risk of $W_b$ independently of $w^*$.

**Proposition 3.3.** [36, Theorem 1] The excess risk of $W_b$ is

$$r(W_b) - r(w^*) = \xi^T (\mathbf{X} \mathbf{c}^{-1} \mathbf{X}^T)^{-1} \xi, \tag{3.5}$$

*Proof.* By definition,

$$\begin{aligned} W_b &= \operatorname*{arg\,min}_{w \in \mathcal{G}} r(w) \\ &= \operatorname*{arg\,min}_{\{w \in \mathbb{R}^d : \mathbf{X} w = Y\}} \|\mathbf{c}^{\frac{1}{2}}(w - w^*)\|_2^2. \end{aligned}$$

Let $z = \mathbf{c}^{\frac{1}{2}}(w - w^*)$, then

$$\begin{aligned} \mathbf{X}w = Y \qquad &\iff \qquad \mathbf{X}(\mathbf{c}^{-\frac{1}{2}}z + w^*) = Y \\ &\iff \qquad \mathbf{X}\mathbf{c}^{-\frac{1}{2}}z = \xi \\ &\iff \qquad \mathbf{B}z = \xi, \end{aligned}$$

where $\mathbf{B} = \mathbf{X}\mathbf{c}^{-\frac{1}{2}}$. Hence, if

$$Z_b := \underset{\{z \in \mathbb{R}^d : \mathbf{B}z = \xi\}}{\arg\min} \|z\|_2^2, \tag{3.6}$$

then

$$W_b = \mathbf{c}^{-\frac{1}{2}} Z_b + w^*.$$

Moreover, (3.6) implies that $Z_b$ is the best approximate solution to the equation $\mathbf{B}z = \boldsymbol{\xi}$ by Penrose [45, Definition], which satisfies

$$\begin{aligned} Z_b &= \mathbf{B}^\dagger \xi \\ &= (\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger \xi, \end{aligned}$$

by Theorem C.8. Hence

$$W_b = \mathbf{c}^{-\frac{1}{2}}(\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger \xi + w^*. \tag{3.7}$$

Now, because $\text{rank}(\mathbf{X}) = n$, $\mathbf{X}\mathbf{c}^{-\frac{1}{2}} \in \mathbb{R}^{n \times d}$ has independent rows so

$$\begin{aligned} (\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger &= (\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^T (\mathbf{X}\mathbf{c}^{-\frac{1}{2}}(\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^T)^{-1} \\ &= \mathbf{c}^{-\frac{1}{2}}\mathbf{X}^T(\mathbf{X}\mathbf{c}^{-1}\mathbf{X}^T)^{-1}. \end{aligned} \tag{3.8}$$

by Proposition C.7. Plugging this into (3.3) gives

$$r(W_b) - r(w^*) = \xi^T(\mathbf{X}\mathbf{c}^{-1}\mathbf{X}^T)^{-1}\xi.$$

□

Importantly, this proposition shows that

$$\xi^T(\mathbf{X}\mathbf{c}^{-1}\mathbf{X}^T)^{-1}\xi \tag{3.9}$$

provides a fundamental limit to how well an interpolator can perform, because $W_b$ is the interpolator with the smallest excess risk.

We make the following important observation, which will be of use later. If we define $\mathbf{Z} = \mathbf{X}\mathbf{c}^{-\frac{1}{2}} \in \mathbb{R}^{n \times d}$ then each row $Z_i$ of $\mathbf{Z}$ has identity covariance matrix, because

$$\begin{aligned} \mathbb{E}(Z_i Z_i^T) &= \mathbb{E}(\mathbf{c}^{-\frac{1}{2}} X_i (\mathbf{c}^{-\frac{1}{2}} X_i)^T) \\ &= \mathbf{c}^{-\frac{1}{2}} \mathbb{E}(X_i X_i^T) \mathbf{c}^{-\frac{1}{2}} \\ &= \mathbf{c}^{-\frac{1}{2}} \mathbf{c} \mathbf{c}^{-\frac{1}{2}} \\ &= I_d. \end{aligned} \tag{3.10}$$

Moreover,

$$\begin{aligned} r(W_b) - r(w^*) &= \xi^T (\mathbf{X}\mathbf{c}^{-1}\mathbf{X}^T)^{-1}\xi \\ &= \xi^T (\mathbf{Z}\mathbf{Z}^T)^{-1}\xi \qquad (3.11) \\ &\geq \frac{\|\xi\|_2^2}{\lambda_{\max}(\mathbf{Z}\mathbf{Z}^T)}, \qquad (3.12) \end{aligned}$$

by Proposition C.1 and as $\mathbf{Z}\mathbf{Z}^T$ is positive definite.

When is nearly ideal generalisation possible for an interpolator? As (3.9) is a fundamental lower bound, this can happen only if $\xi^T(\mathbf{X}\mathbf{c}^{-1}\mathbf{X}^T)^{-1}\xi$ is small. Hence, now we analyse $\xi^T(\mathbf{X}\mathbf{c}^{-1}\mathbf{X}^T)^{-1}\xi$. We do this for a class of random variables called sub-Gaussian. We show that good generalisation is possible only if $d >> n$, for which we will employ concentration inequalities and random matrix theory. The following results are adapted from Vershynin [56], Wainwright [57, Chapter 2], Jin et al. [27] and Rebeschini [46].

**Definition 3.4.** A random variable $X \in \mathbb{R}^d$ is *sub-Gaussian* with variance proxy $\sigma^2 > 0$ if for all $v \in \mathbb{R}^d$,

$$\mathbb{E}\, e^{\langle v, X - \mathbb{E}(X)\rangle} \leq e^{\|v\|_2^2 \sigma^2/2}.$$

A number of equivalent definitions are in [56, Lemma 5.5, Definition 5.22]. If $X_i$ are Gaussian with covariance matrix $\mathbf{c}$ then they are sub-Gaussian with variance proxy $\lambda_{\max}(\mathbf{c})$, because

$$\mathbb{E}\, e^{\langle v, X - \mathbb{E}\, X\rangle} = e^{v^T \mathbf{c} v/2} \leq e^{\|v\|_2^2 \lambda_{\max}(\mathbf{c})/2}.$$

**Definition 3.5.** A random variable $X \in \mathbb{R}$ is *sub-exponential* with parameters $(\sigma^2, c) \in \mathbb{R}_{>0}$ if for all $t \in (-1/c, 1/c)$,

$$\mathbb{E}\, e^{t(X - \mathbb{E}\, X)} \leq e^{t^2\sigma^2/2}.$$

This is also generalisable to higher dimensions [56, Remark 5.2.3].

**Lemma 3.6.** If a random variable $X \in \mathbb{R}$ is centered Gaussian with variance $\sigma^2$ then $X^2$ is sub-exponential with parameters $(4\sigma^4, 4\sigma^2)$.

*Proof.* Let $t \in \mathbb{R}$. Note that $Y = \frac{X^2}{\sigma^2}$ follows the Gamma$(\frac{1}{2}, \frac{1}{2})$ distribution. Hence,

$$\begin{aligned} \mathbb{E}\, e^{t(X^2 - \mathbb{E}(X^2))} &= \mathbb{E}\, e^{t\sigma^2(Y - \mathbb{E}(Y))} \\ &= \mathbb{E}\, e^{t\sigma^2(Y-1)} \\ &= \sqrt{\frac{1}{2\Gamma(1/2)}} e^{-t\sigma^2} \int_0^\infty y^{-1/2} e^{y(t\sigma^2 - 1/2)} dy \\ &= \sqrt{\frac{1}{1 - 2t\sigma^2}} e^{-t\sigma^2}, \end{aligned}$$

which follows by a gamma integral computation. Then, to show that,

$$\mathbb{E}\, e^{t(X^2 - \mathbb{E}(X^2))} \leq e^{4t^2\sigma^4/2},$$

one equivalently shows

$$1 \leq e^{4t^2\sigma^4+2t\sigma^2}(1-2t\sigma^2),$$

which is easy to check to hold for $t \in (-\frac{1}{4\sigma^2}, \frac{1}{4\sigma^2})$ by ordinary calculus. □

**Lemma 3.7.** For any $\delta \in (0,1)$,

$$\|\xi\|_2^2 \geq n\sigma^2(1-\delta),$$

with probability at least $1 - e^{-n\sigma^2/8}$.

*Proof.* The entries of $\xi$ are independent centered Gaussian variables with variance $\sigma^2$. Therefore, by Lemma 3.6,

$$\begin{aligned}
\mathbb{E}\, e^{t(\|\xi\|_2^2 - \mathbb{E}\,\|\xi\|_2^2)} &= \mathbb{E}\, e^{t(\sum_{i=1}^n \xi_i^2 - n\sigma^2)} \\
&= \prod_{i=1}^n \mathbb{E}\, e^{t(\xi_i^2-\sigma^2)} \\
&\leq \prod_{i=1}^n e^{t^2 4\sigma^4/2} \\
&= e^{4nt^2\sigma^4/2} \qquad (3.13)
\end{aligned}$$

for $t \in (-\frac{1}{4\sigma^2}, \frac{1}{4\sigma^2})$, so $\|\xi\|_2^2$ is sub-exponential with parameters $(4n\sigma^4, 4\sigma^2)$. Now, we employ a standard technique to bound the tail probability of $\|\xi\|_2^2$. If $\delta \in (0,1)$, then

$$\begin{aligned}
\mathbb{P}(\|\xi\|_2^2 - \mathbb{E}\|\xi\|_2^2 \leq -n\sigma^2\delta) &= \mathbb{P}(\|\xi\|_2^2 - n\sigma^2 \leq -n\sigma^2\delta) \\
&= \mathbb{P}(e^{t(\|\xi\|_2^2 - n\sigma^2)} \geq e^{-tn\sigma^2\delta})
\end{aligned}$$

for $t \in (-\frac{1}{4\sigma^2}, 0)$. By Markov's inequality and (3.13),

$$\begin{aligned}
\mathbb{P}(e^{t(\|\xi\|_2^2 - n\sigma^2)} \geq e^{-tn\sigma^2\delta}) &\leq e^{tn\sigma^2\delta}\mathbb{E}\, e^{t(\|\xi\|_2^2 - n\sigma^2)} \\
&\leq e^{tn\sigma^2\delta + 4nt^2\sigma^4/2}.
\end{aligned}$$

Minimising the right side over $t \in (-\frac{1}{4\sigma^2}, 0)$ one finds a minimum at

$$t^* = -\frac{\delta}{4\sigma^2},$$

which gives

$$\begin{aligned}
\mathbb{P}(\|\xi\|_2^2 - n\sigma^2 \leq -n\sigma^2\delta) &\leq e^{-n\delta^2/4 + n\delta^2/8} \\
&= e^{-n\delta^2/8}.
\end{aligned}$$

Hence

$$\mathbb{P}(\|\xi\|_2^2 \geq n\sigma^2(1-\delta)) \geq 1 - e^{-n\delta^2/8}.$$

□

To further lower bound

$$r(W_b) - r(w^*) \geq \frac{\|\xi\|_2^2}{\lambda_{\max}(\mathbf{Z}\mathbf{Z}^T)}, \tag{3.14}$$

we need to upper bound $\lambda_{\max}(\mathbf{Z}\mathbf{Z}^T)$. For this we use random matrix theory. The following is [56, Theorem 5.39] adapted using [56, Lemma 5.5].

**Theorem 3.8.** If $\mathbf{Z} \in \mathbb{R}^{n\times d}$ has independent sub-Gaussian rows with identity covariance matrix, variance proxy $\sigma^2$ and mean $\mu$, then for every $t \geq 0$ we have

$$\sqrt{n} - C\sqrt{d} - t \leq \sqrt{\lambda_{\min}(\mathbf{Z}\mathbf{Z}^T)} \leq \sqrt{\lambda_{\max}(\mathbf{Z}\mathbf{Z}^T)} \leq \sqrt{n} + C\sqrt{d} + t \tag{3.15}$$

with probability at least $1 - 2e^{-ct^2}$. Here, $C = C_{\sigma^2,\mu}, c = c_{\sigma^2,\mu} > 0$ depend only on $\sigma^2$ and $\mu$.

Putting together (3.14), observation (3.10), Lemma 3.7 and Theorem 3.8 proves the following result.

**Proposition 3.9.** [36, Corrolary 1.2] If $X_i$ are sub-Gaussian with variance proxy $\sigma^2$, $\mathbb{E}(X_i) = \mu$ and $\mathbb{E}(X_i X_i^T) = \mathbf{c}$ then for every $\delta \in (0,1)$,

$$r(W_b) - r(w^*) \geq \frac{n\sigma^2(1-\delta)}{(C\sqrt{d} + 2\sqrt{n})^2}$$

with probability at least $1 - e^{-cn} - e^{-n\delta^2/8}$. Here, $C = C_{\sigma^2,\mu,\mathbf{c}}, c = c_{\sigma^2,\mu,\mathbf{c}} > 0$ depend only on $\sigma^2$, $\mu$ and $\mathbf{c}$.

*Proof.* Assume $X_i$ are sub-Gaussian. By (3.12),

$$r(W_b) - r(w^*) \geq \frac{\|\xi\|_2^2}{\lambda_{\max}(\mathbf{Z}\mathbf{Z}^T)}.$$

By Lemma 3.7,

$$\|\xi\|_2^2 \geq n\sigma^2(1-\delta)$$

with probability at least $1 - e^{-n\delta^2/8}$. Let $Z_i$ be rows of $\mathbf{Z}$. Then

$$\begin{aligned}\mathbb{E}\, e^{\langle v, Z_i - \mathbb{E}\, Z_i\rangle} = e^{\langle \mathbf{c}^{-\frac{1}{2}} v, X_i - \mathbb{E}\, X_i\rangle} &\leq e^{v^T \mathbf{c}^{-1} v \sigma^2/2} \\ &\leq e^{\|v\|_2^2 \sigma^2 / 2\lambda_{\min}(\mathbf{c})}\end{aligned}$$

for all $v \in \mathbb{R}^d$, so $Z_i$ are sub-Gaussian with variance proxy $\sigma^2/\lambda_{\min}(\mathbf{c})$. $\mathbb{E}(Z_i Z_i^T) = I_d$ by observation (3.10), so $\mathbf{Z}$ satisfies assumptions of Theorem 3.8. Choosing $t = \sqrt{n}$ gives

$$\lambda_{\max}(\mathbf{Z}\mathbf{Z}^T) \leq (2\sqrt{n} + C\sqrt{d})^2$$

with probability at least $1 - 2e^{-cn}$. Therefore,

$$r(W_b) - r(w^*) \geq \frac{n\sigma^2(1-\delta)}{(C\sqrt{d} + 2\sqrt{n})^2}$$

with probability at least

$$(1 - e^{-n\delta^2/8})(1 - 2e^{-cn}) \geq 1 - e^{-n\delta^2/8} - e^{-cn},$$

by independence of $\xi$ and $\mathbf{Z}$. □

What does Proposition 3.9 tell us about the relationship between good generalisation and overparametrisation? As $d > n$, there exists $\gamma > 1$ with $d = \gamma n$. Choose $\delta = \frac{1}{2}$ and assume the dataset is large, say $n \geq \max(100/c, 3200)$. Then

$$1 - e^{-cn} - e^{-n\delta^2/8} \geq 1 - 2e^{-100} > 1 - 10^{-43}.$$

Therefore, with very high probability, for all interpolators $\hat{W} \in \mathcal{G}$,

$$\begin{aligned} r(\hat{W}) - r(w^*) &\geq r(W_b) - r(w^*) \\ &\geq \frac{n\sigma^2(1-\delta)}{(C\sqrt{d} + 2\sqrt{n})^2} \\ &= \frac{\sigma^2}{2(C\sqrt{\gamma} + 2)^2}. \end{aligned}$$

Hence, the only way $r(W_b) - r(w^*)$ can be very close to 0 is if $\gamma$ is large. In other words, for near-ideal generalisation of an interpolator to be possible, large overparametrisation ($d >> n$) is necessary, if the features are sub-Gaussian.

But is it sufficient? The lower bound of Theorem 3.8,

$$\sqrt{n} - C\sqrt{d} - t \leq \sqrt{\lambda_{\min}(\mathbf{Z}\mathbf{Z}^T)}, \tag{3.16}$$

is vacuous for $d >> n$. Indeed, lower bounding the least singular value is a research topic on its own (Tao [52]). However, if $X_i$ are Gaussian, an elegant trick can be used.

If $X_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathbf{c})$, then $Z_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_d)$ by observation (3.10). An uncorrelated Gaussian vector has independent entries (Hogg et al. [23, p. 159]). Therefore, also rows of $\mathbf{Z}^T \in \mathbb{R}^{d \times n}$ are independent, so we can apply Theorem 3.8 to $\mathbf{Z}^T \in \mathbb{R}^{d \times n}$ and hence switch $d$ and $n$ in (3.16). Moreover, one can show $C = c = 1$ [56, Corrolary 5.35]. Therefore, the same argument proves the following.

**Proposition 3.10.** [36, Corrolary 2.2, Corrolary 1.2] If $X_i$ are centered Gaussian then for all $\delta \in (0, 1)$,

$$\frac{n\sigma^2(1+\delta)}{(\sqrt{d} - 2\sqrt{n})^2} \geq r(W_b) - r(w^*) \geq \frac{n\sigma^2(1-\delta)}{(\sqrt{d} + 2\sqrt{n})^2},$$

with probability at least $1 - e^{-n} - e^{-n\delta^2/8}$.

Setting $d = \gamma n$ with $\gamma > 1$ gives

$$\frac{\sigma^2(1+\delta)}{(\sqrt{\gamma} - 2)^2} \geq r(W_b) - r(w^*) \geq \frac{\sigma^2(1-\delta)}{(\sqrt{\gamma} + 2)^2},$$

with probability at least $1 - e^{-n} - e^{-n\delta^2/8}$. Hence, this is a powerful statement about linear regression with Gaussian features. In words: Near-ideal generalisation of an interpolator is possible if and only if high overparametrisation is present.

## 3.2 The Minimum Norm Interpolator

However, the problem is that we cannot access $W_b$. Therefore, we still do not know whether good generalisation of an interpolator is actually achievable. This is what we focus on now.

We do have access to $W_{\ell_2}$ - the minimiser with smallest $\|\cdot\|_2$ norm, because of the implicit bias of gradient descent (Proposition 2.16). Several excellent papers have recently discussed the generalisation properties of $W_{\ell_2}$. Bartlett et al. [6, Theorem 1] prove sharp upper and lower bounds on $r(W_{\ell_2}) - r(w^*)$ in high probability and provide exact regimes of the covariance matrix under which this quantity is small [6, Theorem 6]. Importantly, their setting is non-asymptotic, i.e. holds for finite $n$ and $d$. Hastie et al. [21] analyse $\mathbb{E}_\xi(r(W_{\ell_2}) - r(w^*))$ and characterise its limit as a function of $\gamma$, where $p/n \to \gamma$ as $n \to \infty, p \to \infty$. We present the latter approach.

**Lemma 3.11.** [21, Lemma 1] $\mathcal{R}(W_{\ell_2}) := \mathbb{E}_\xi(r(w) - r(w^*))$ satisfies the bias variance decomposition

$$\begin{aligned}\mathcal{R}(W_{\ell_2}) &= \mathcal{R}(\mathbf{X}^\dagger \mathbf{X} w^*) + \mathbb{E}_\xi(\xi^T \mathbf{X}^{\dagger T} \mathbf{c} \mathbf{X}^\dagger \xi) \\ &= B_{\ell_2} + V_{\ell_2},\end{aligned} \tag{3.17}$$

where

$$\begin{aligned}B_{\ell_2} &= P_{\mathrm{Ker}(\mathbf{X})}(w^*)^T \mathbf{c} P_{\mathrm{Ker}(\mathbf{X})}(w^*) \\ V_{\ell_2} &= \frac{\sigma^2}{n} \mathrm{Tr}(\hat{\mathbf{C}}^\dagger \mathbf{c})\end{aligned} \tag{3.18}$$

and $Tr(\cdot)$ is the trace, $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ and $P_{\mathrm{Ker}(\mathbf{X})}(w^*) = (I - \mathbf{X}^\dagger \mathbf{X})(w^*)$ is the projection onto $\mathrm{Ker}(\mathbf{X})$ (see Propositions C.12, C.14).

*Proof.* By (3.2),

$$\mathcal{R}(W_{\ell_2}) = \mathbb{E}_\xi (w^* - W_{\ell_2})^T \mathbf{c} (w^* - W_{\ell_2}).$$

As $W_{\ell_2} = P_{\mathcal{G}}(0) = \mathbf{X}^\dagger Y = \mathbf{X}^\dagger(\mathbf{X}w^* + \xi)$ (Proposition C.14),

$$\begin{aligned}\mathcal{R}(W_{\ell_2}) = {} & w^{*T}(I - \mathbf{X}^\dagger \mathbf{X})^T \mathbf{c} (I - \mathbf{X}^\dagger \mathbf{X}) w^* + \\ & \mathbb{E}_\xi(2 w^{*T} (I - \mathbf{X}^\dagger \mathbf{X})^T \mathbf{c} \mathbf{X}^\dagger \xi) + \mathbb{E}_\xi((\mathbf{X}^\dagger \xi)^T \mathbf{c} \mathbf{X}^\dagger \xi).\end{aligned}$$

The first term is $B_{\ell_2}$. The second term is 0 because $\mathbb{E}(\xi) = 0$. The last term is $V_{\ell_2}$ and to simplify it we use the trace trick

$$v^T \mathbf{c} v = \mathrm{Tr}(v^T \mathbf{c} v) = \mathrm{Tr}(\mathbf{c} v v^T) = \mathrm{Tr}(v v^T \mathbf{c})$$

with $v = \mathbf{X}^\dagger \xi$, linearity of expectation, $\mathbb{E}(\xi \xi^T) = \sigma^2 I_d$ and Proposition C.7,

$$\begin{aligned}\mathbb{E}_\xi((\mathbf{X}^\dagger \xi)^T \mathbf{c} \mathbf{X}^\dagger \xi) &= \mathbb{E}_\xi \mathrm{Tr}(\mathbf{X}^\dagger \xi \xi^T \mathbf{X}^{\dagger T} \mathbf{c}) \\ &= \mathrm{Tr}(\mathbf{X}^\dagger \mathbb{E}(\xi \xi^T) \mathbf{X}^{\dagger T} \mathbf{c}) \\ &= \sigma^2 \mathrm{Tr}(\mathbf{X}^\dagger \mathbf{X}^{\dagger T} \mathbf{c}) \\ &= \frac{\sigma^2}{n} \mathrm{Tr}((\frac{1}{n}\mathbf{X}^T \mathbf{X})^\dagger \mathbf{c}) \\ &= V_{\ell_2}.\end{aligned}$$

□

Now, we use random matrix theory to analyse $B_{\ell_2}$ and $V_{\ell_2}$ as $\frac{d}{n} \to \gamma \in (1, \infty)$ with $n \to \infty, d \to \infty$. First, we provide some intuition for the reader. For a symmetric matrix $\mathbf{m} \in \mathbb{R}^{d \times d}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ we define its *spectral distribution* by

$$F_{\mathbf{m}}(x) = \frac{1}{d} \sum_{i=1}^{d} \mathbb{1}_{[\lambda_i, \infty)}(x), \tag{3.19}$$

for $x \in \mathbb{R}$. The associated Lebesgue-Stieltjes measure (see Kolmogorov [29, Chapter 10]),

$$\mu_{\mathbf{m}} = \frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_i}, \tag{3.20}$$

is defined by

$$\mu_{\mathbf{m}}(\mathcal{S}) = \frac{1}{d} |\{i \in \{1, \ldots, d\} : \lambda_i \in \mathcal{S}\}|,$$

for any Borel measurable $\mathcal{S} \subseteq \mathbb{R}$. We assume that $X_i$ have i.i.d. entries with 0 mean and variance 1. The discussion in the i.i.d. setting then provides the outline for the general case. Hence $\mathbf{c} = I_d$ and

$$\begin{aligned} V_{\ell_2} &= \frac{\sigma^2}{n} \mathrm{Tr}(\hat{\mathbf{C}}^{\dagger}) \\ &= \sigma^2 \sum_{i=1}^{n} \frac{1}{\lambda_i}, \end{aligned}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ are the non-zero eigenvalues of $\mathbf{X}^T\mathbf{X}$ (and $\mathbf{X}\mathbf{X}^T$). The important connection with (3.19) is that, if $\mathbf{K} = \frac{1}{d}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}$, then

$$\int_{\lambda_{i+1}/d}^{\lambda_i/d} f(x) \mu_{\mathbf{K}}(dx) = \frac{1}{n} f(\frac{\lambda_i}{d}),$$

by definition (3.20). Thus,

$$\begin{aligned} V_{\ell_2} = \sigma^2 \sum_{i=1}^{n} \frac{1}{\lambda_i} &= \sum_{i=1}^{n-1} \frac{\sigma^2 n}{d} \int_{\lambda_{i+1}/d}^{\lambda_i/d} \frac{1}{x} \mu_{\mathbf{K}}(dx) + \frac{\sigma^2 n}{d} \int_{0}^{\lambda_n/d} \frac{1}{x} \mu_{\mathbf{K}}(dx) \\ &= \frac{\sigma^2 n}{d} \int_{\mathbb{R}} \frac{1}{x} \mu_{\mathbf{K}}(dx). \end{aligned} \tag{3.21}$$

Then we are nearly done, because standard random matrix theory asserts that $F_{\mathbf{K}}$ weakly converges and we can compute the limiting integral. Now, we formalise these arguments.

**Theorem 3.12.** [3, Theorem 2] If $\mathbf{X} \in \mathbb{R}^{n \times d}$ has centered i.i.d. entries with variance 1 and

$$\frac{n}{d} \longrightarrow y \in (0, 1)$$

as $n \to \infty, d \to \infty$, then with probability 1,

$$\lambda_{\min}(\frac{1}{d}\mathbf{X}\mathbf{X}^T) \longrightarrow (1-\sqrt{y})^2$$
$$\lambda_{\max}(\frac{1}{d}\mathbf{X}\mathbf{X}^T) \longrightarrow (1+\sqrt{y})^2.$$

Applying this theorem with $y = \frac{1}{\gamma} \in (0,1)$ gives that, for sufficiently large $n, d$,

$$\lambda_{\min}(\frac{1}{d}\mathbf{X}\mathbf{X}^T) > \frac{1}{2}(1-\sqrt{1/\gamma})^2 > 0.$$

Now, we use a theorem of Marčenko and Pastur [33] (see Götze and Tikhomirov [17] and Bai and Yin [3] for more accessible references) to reason about the limit of (3.21).

**Theorem 3.13.** If $\mathbf{X} \in \mathbb{R}^{n\times d}$ has centered i.i.d. entries with variance 1, $\mathbb{E}(X_{i,j}^4) \in \mathbb{R}$ and

$$\frac{d}{n} \longrightarrow \gamma \in (0,\infty)$$

as $n \to \infty, d \to \infty$, then there exists a distribution $F_\gamma$ such that

$$F_{\mathbf{K}} = F_{\frac{\mathbf{X}\mathbf{X}^T}{d}} \longrightarrow F_\gamma,$$

weakly, with probability 1. That is,

$$\int_{\mathbb{R}} f(x)\mu_{\mathbf{K}}(dx) \longrightarrow \int_{\mathbb{R}} f(x)\mu_\gamma(dx)$$

as $n \to \infty, d \to \infty$, for all bounded continuous $f : \mathbb{R} \to \mathbb{R}$, with probability 1. Here, $\mu_\gamma$ is the Lebesgue-Stieltjes measure of $F_\gamma$.

In both theorems, the almost-sure limit is with respect to the probability space of $X_1, \ldots, X_n$. Assuming $\mathbb{E}(X_{i,j}^4) \in \mathbb{R}$, equation (3.21) implies

$$V_{\ell_2} = \sigma^2 \frac{n}{d} \int_{\mathbb{R}} \frac{1}{x}\mu_{\mathbf{K}}(dx) \longrightarrow \frac{\sigma^2}{\gamma} \int_{\mathbb{R}} \frac{1}{x}\mu_\gamma(dx) \tag{3.22}$$

with probability 1. This is because, by Theorem 3.12 for all $n, d$ large enough,

$$\mu_{\mathbf{K}}(\mathcal{S}) := \mu_{\mathbf{K}}\Big[\frac{(1-\sqrt{1/\gamma})^2}{2}, \infty\Big) = 1$$

and hence also

$$\mu_\gamma(\mathcal{S}) := \mu_\gamma\Big[\frac{(1-\sqrt{1/\gamma})^2}{2}, \infty\Big) = 1.$$

The function $\psi : \mathcal{S} \ni x \mapsto \frac{1}{x}$ is continuous and bounded and hence can be extended onto $\mathbb{R} \setminus \mathcal{S}$ (by a constant) to a continuous and bounded function $\mathbb{R} \to \mathbb{R}$ and we can apply Theorem 3.13 to get (3.22).

The last step is computing the integral in (3.22). This is a technicality involving the Stieltjes transform of $\mu_\gamma$, for which we refer to the proof of [21, Theorem 1], which shows

$$\int_{\mathbb{R}} \frac{1}{x} \mu_\gamma(dx) = \frac{\gamma}{\gamma - 1}.$$

Therefore, we have proved

$$V_{\ell_2} \longrightarrow \frac{\sigma^2}{\gamma - 1} \tag{3.23}$$

with probability 1. [21, Lemma 2] similarly finds the limit of $B_{\ell_2}$ (assuming $\mathbb{E}|X_{i,j}|^{8+\eta} \in \mathbb{R}$, for some $\eta > 0$). Putting this together, we obtain the following theorem.

**Theorem 3.14.** [21, Theorem 2] If $\mathbf{X} \in \mathbb{R}^{n \times d}$ has centered i.i.d. entries with variance 1, $\mathbb{E}|X_{i,j}|^{8+\eta} \in \mathbb{R}$, for some $\eta > 0$, and $\frac{d}{n} \to \gamma \in (1, \infty)$ as $n \to \infty, d \to \infty$ then

$$\mathcal{R}(W_{\ell_2}) \longrightarrow \mathcal{R}_\gamma(W_{\ell_2}) := \|w^*\|_2^2 (1 - \frac{1}{\gamma}) + \frac{\sigma^2}{\gamma - 1} \tag{3.24}$$

with probability 1.

If $\mathbf{c} \neq I_d$, $k \leq \lambda_{\min}(\mathbf{c}) \leq \lambda_{\max}(\mathbf{c}) \leq K$ for some $K \geq k > 0$, uniformly in $n, d \in \mathbb{N}$ and $F_{\mathbf{c}}$ converges weakly, then a generalised result is [21, Theorem 3], where

$$\mathcal{R}(W_{\ell_2}) \longrightarrow \lim_{z \to 0^+} \frac{\mathbb{E}_P \|w^*\|_2^2}{\gamma v(z)} + \sigma^2 \big( \frac{v'(z)}{v(z)^2} - 1 \big), \tag{3.25}$$

where $P$ is a prior on $w^*$, $v$ is the companion Stieltjes transform (for a definition see Dobriban and Wager [15, Section 1.4]) of the limit of $F_{\hat{\mathbf{C}}}$ given by a (generalised) Marčenko-Pastur theorem (Ledoit and Peche [31, Theorem 1.1]). $\mathbb{E}(Z_{i,j})^{12} \in \mathbb{R}$ is also required.

Now, we illustrate these results. Define the signal-to-noise ratio, SNR $:= \|w^*\|_2^2 / \sigma^2$. Then, $\mathcal{R}_\gamma(W_{\ell_2})$ depends on SNR as follows.

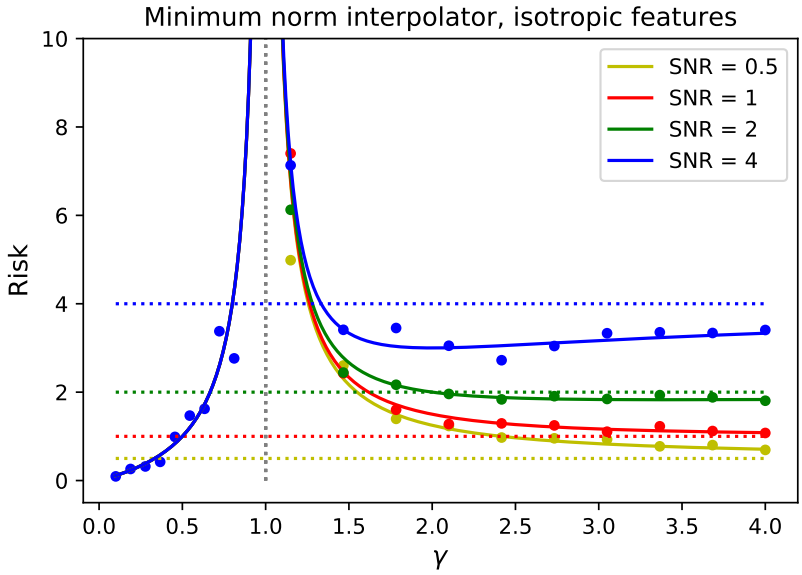


Figure 3.1: Plot of $\mathcal{R}_\gamma(W_{\ell_2})$ (lines), $r(W_{\ell_2}) - r(w^*)$ (points) and $\mathcal{R}(0) = \|w^*\|_2^2$ (dotted lines) with $\sigma^2 = 1$, $\|w^*\|_2^2 \in \{0.5, 1, 2, 4\}$, $n = 200$, $d = \lfloor \gamma n \rfloor$, $X_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

We also included $\gamma \in (0, 1)$, where one can show $\mathcal{R}_\gamma(W_{\ell_2}) = \sigma^2 \frac{\gamma}{1-\gamma}$ [21, Theorem 1] under the assumptions of Theorem 3.14. Figure 3.1 shows a version of the double descent phenomenon (Figure 1.1), which has also been observed in neural networks ([8, Appendix C]). Curiously,

$$\mathcal{R}_\gamma(W_{\ell_2}) \longrightarrow \|w^*\|_2^2$$

as $\gamma \to \infty$, where $\|w^*\|_2^2 = \mathcal{R}(0)$ is the *null risk*. Therefore, for very large overparametrisation, $W_{\ell_2}$ performs similarly to $w = 0$, i.e. as if we did no learning at all.

When SNR $\leq 1$, we have gradually better performance as $\gamma \to \infty$, which is analogous to neural networks, where scaling up often leads to better performance (Tan and Le [51]). However, $\mathcal{R}_\gamma(W_{\ell_2}) > \mathcal{R}(0)$, i.e. $W_{\ell_2}$ is asymptotically always worse than $w = 0$.

When SNR $> 1$, $\mathcal{R}_\gamma(W_{\ell_2}) < \mathcal{R}(0)$ for $\gamma > \frac{\text{SNR}}{\text{SNR}-1}$ and we see a bias-variance trade-off in the interpolating regime;

- For $1 < \gamma$ close to 1, $B_{\ell_2} \approx 0$ and $\mathcal{R}_\gamma(W_{\ell_2}) \approx V_{\ell_2}$.
- For $1 << \gamma$, $V_{\ell_2} \approx 0$ and $\mathcal{R}_\gamma(W_{\ell_2}) \approx B_{\ell_2}$.

A sweet spot of minimal $\mathcal{R}_\gamma(W_{\ell_2})$ is at $\gamma = \frac{\sqrt{\text{SNR}}}{\sqrt{\text{SNR}}-1}$.

In summary, we observe some phenomena of neural networks (double descent, decreasing risk with overparametrisation) also in linear regression for the implicit bias interpolator of (stochastic) gradient descent. However, caution in the analogy must be taken. For example, [15] and [21, Section 6] show that the asymptotic risk of ridge regression is better for all $\gamma > 0$, in this setting. In contrast, interpolation for

neural networks is often considered best practice [8, p. 2].

We showed how the implicit bias of gradient descent on linear models affects generalisation. This provides grounding for a study in more complicated settings. Whether implicit bias can explain generalisation of neural networks is unknown. However, this is a step towards an answer.

## 3.3 A Mirror Descent Interpolator

Is $W_{\ell_2}$ the best interpolator that we have access to? Can we do better if we have some additional knowledge about the problem? In this section, we reconnect ideas of the whole work to discuss a new mirror descent interpolator. To our best knowledge, this is original work. Assume that we know the covariance matrix $\mathbf{c} = \mathbb{E}(X_i X_i^T)$, which is a nontrivial assumption. Recall Remark 3.2, where the Bregman divergence of $\Phi(w) = w^T \mathbf{c} w$ (assuming $\lambda_{\min}(\mathbf{c}) > 0$) arises naturally as

$$D_\Phi(w, w^*) = \frac{1}{2}(w - w^*)^T \mathbf{c}(w - w^*) = \frac{1}{2}(r(w) - r(w^*)). \tag{3.26}$$

Hence, instead of

$$W_{\ell_2} = \arg\min_{w \in \mathcal{G}} \|w\|^2,$$

we examine

$$\begin{aligned} W_{\mathbf{c}} &= \arg\min_{w \in \mathcal{G}} D_\Phi(w, 0) \\ &= \arg\min_{w \in \mathcal{G}} w^T \mathbf{c} w, \end{aligned}$$

which is the limit of mirror descent initialised at 0 (Proposition 2.9). Because of (3.26), we expect $W_{\mathbf{c}}$ to have interesting generalisation properties. Similarly to Proposition 3.3,

$$\begin{aligned} W_{\mathbf{c}} &= \arg\min_{\mathbf{X}w = Y} \|\mathbf{c}^{\frac{1}{2}} w\|_2^2 \\ &= \arg\min_{\mathbf{X}\mathbf{c}^{-\frac{1}{2}}(\mathbf{c}^{\frac{1}{2}} w) = Y} \|\mathbf{c}^{\frac{1}{2}} w\|_2^2 \\ &= \mathbf{c}^{-\frac{1}{2}} (\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger Y. \end{aligned}$$

Comparing it to $W_{\ell_2} = \mathbf{X}^\dagger Y$ and the best interpolator $W_b$ (Definition 3.1, equation (3.7)), we have

$$\begin{aligned} W_{\ell_2} &= \mathbf{X}^\dagger \xi + \mathbf{X}^\dagger \mathbf{X} w^*, \\ W_{\mathbf{c}} &= \mathbf{c}^{-\frac{1}{2}} (\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger \xi + \mathbf{c}^{-\frac{1}{2}} (\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger \mathbf{X} w^* \\ W_b &= \mathbf{c}^{-\frac{1}{2}} (\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger \xi + w^* \end{aligned}$$

Noticeably, $W_b$ and $W_{\mathbf{c}}$ have the same noise term and hence $\mathcal{R}(W_{\mathbf{c}})$ and $\mathcal{R}(W_b)$ have the same variance term in the bias-variance decompositions

$$\begin{aligned}\mathcal{R}(W_{\ell_2}) &= B_{\ell_2} + V_{\ell_2}\\ \mathcal{R}(W_{\mathbf{c}}) &= B_{\mathbf{c}} + V_{\mathbf{c}}\\ \mathcal{R}(W_b) &= B_b + V_b,\end{aligned}$$

analogous to Lemma 3.11. Here,

$$\begin{aligned}B_b &= \mathcal{R}(w^*) = 0\\ B_{\mathbf{c}} &= \mathcal{R}(\mathbf{c}^{-\frac{1}{2}}(\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger \mathbf{X}w^*)\end{aligned}$$

and

$$\begin{aligned}\mathcal{R}(W_b) = V_b = V_{\mathbf{c}} &= \mathbb{E}_\xi(\xi^T(\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^{\dagger^T}\mathbf{c}^{-\frac{1}{2}}\mathbf{c}\mathbf{c}^{-\frac{1}{2}}(\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger\xi)\\ &= \mathbb{E}_\xi(\xi^T(\mathbf{Z}\mathbf{Z}^T)^\dagger\xi),\end{aligned}$$

where $\mathbf{Z} = \mathbf{X}\mathbf{c}^{-\frac{1}{2}}$. That is, $W_{\mathbf{c}}$ enjoys having the variance term of the best interpolator $W_b$ and $V_{\mathbf{c}}$ is completely independent of $\mathbf{c}$. Hence, we expect $W_{\mathbf{c}}$ to perform better than $W_{\ell_2}$ in noisy settings, e.g. when $\sigma^2 > \|w^*\|_2^2$.

**Proposition 3.15.** If $X_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathbf{c})$ then

$$\mathcal{R}(W_b) = V_{\mathbf{c}} \longrightarrow \frac{\sigma^2}{\gamma - 1}$$

and if there exists $K \geq 0$ such that $\mathcal{R}(0) = w^{*T}\mathbf{c}w^* \leq K$ for all $n, d \in \mathbb{N}$, then

$$\frac{B_{\mathbf{c}}}{\mathcal{R}(0)} \longrightarrow 1 - \frac{1}{\gamma}$$

with probability 1, as $\frac{d}{n} \to \infty$ with $n \to \infty, d \to \infty$.

Notice the simplicity of this result compared to its analogue for $W_{\ell_2}$ in (3.25).

*Proof.* Firstly,

$$\begin{aligned}V_{\mathbf{c}} = \mathbb{E}_\xi(\xi^T(\mathbf{Z}\mathbf{Z}^T)^\dagger\xi) &= \operatorname{Tr}\mathbb{E}_\xi(\mathbf{Z}^\dagger\xi\xi^T\mathbf{Z}^{T\dagger})\\ &= \sigma^2\operatorname{Tr}(\mathbf{Z}^\dagger\mathbf{Z}^{\dagger^T})\\ &= \sigma^2\operatorname{Tr}(\mathbf{Z}^T\mathbf{Z})^\dagger\end{aligned}$$

and we already proved (equation (3.23)) that

$$\sigma^2\operatorname{Tr}(\mathbf{Z}^T\mathbf{Z})^\dagger \longrightarrow \frac{\sigma^2}{\gamma - 1},$$

as $\frac{d}{n} \to \infty$ with probability 1, because $Z_{i,j} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$. Secondly,

$$\begin{aligned}B_{\mathbf{c}} &= w^{*T}(I - \mathbf{c}^{-\frac{1}{2}}(\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger\mathbf{X})^T\mathbf{c}(I - \mathbf{c}^{-\frac{1}{2}}(\mathbf{X}\mathbf{c}^{-\frac{1}{2}})^\dagger\mathbf{X})w^*\\ &= \mathbf{c}^{\frac{1}{2}}w^{*^T}(I - \mathbf{Z}^\dagger\mathbf{Z})^T(I - \mathbf{Z}^\dagger\mathbf{Z})\mathbf{c}^{\frac{1}{2}}w^*\\ &= P_{\operatorname{Ker}(\mathbf{Z})}(\mathbf{c}^{\frac{1}{2}}w^*)^T P_{\operatorname{Ker}(\mathbf{Z})}(\mathbf{c}^{\frac{1}{2}}w^*).\end{aligned}$$

Noting the connection to (3.18) and as $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$,

$$\frac{B_{\mathbf{c}}}{\mathcal{R}(0)} = \frac{B_{\mathbf{c}}}{w^{*T}\mathbf{c}w^*} \longrightarrow 1 - \frac{1}{\gamma} \tag{3.27}$$

holds with probability 1. We exclude the proof. It follows analogous arguments to [21, Appendix A.1]. □

Now, we compare the performance of $W_{\mathbf{c}}, W_{\ell_2}$ and $W_b$ for an *autoregressive regime* of $\mathbf{c}$. Here, $\mathbf{c} = \mathbf{c}^\rho \in \mathbb{R}^{d\times d}$, with

$$\mathbf{c}^\rho_{i,j} := \rho^{|i-j|}, \tag{3.28}$$

where $\rho \in [0,1)$ and $i,j \in \{1,\dots,d\}$. [21, Appendix 6] numerically compute $\mathcal{R}_\gamma(W_{\ell_2})$ for $\mathbf{c} = \mathbf{c}^\rho$ (through (3.25)) and illustrate that $\mathcal{R}_\gamma(W_{\ell_2})$ behaves similarly to Figure 3.1 and

$$\mathcal{R}_\gamma(W_{\ell_2}) \to \|w^*\|_2^2 \tag{3.29}$$

as $\gamma \to \infty$, if $\|w^*\|_2^2$ is kept fixed. Let

$$\mathcal{R}_\gamma(W_{\mathbf{c}}) := \mathcal{R}(0)(1 - \frac{1}{\gamma}) + \frac{\sigma^2}{\gamma - 1} \tag{3.30}$$

and

$$\mathcal{R}_\gamma(W_b) := \frac{\sigma^2}{\gamma - 1}.$$

First, we empirically convince the reader about (3.27).

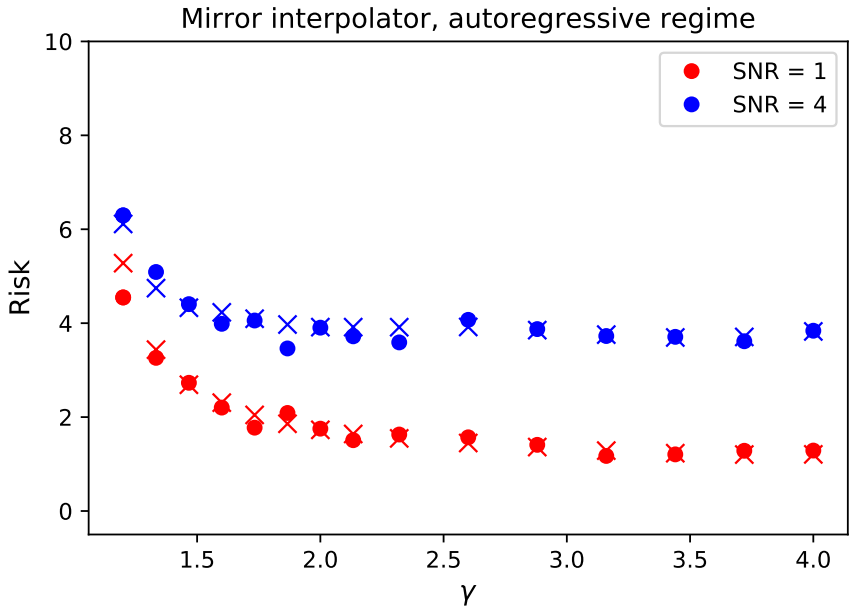


Figure 3.2: Plot of $r(W_{\mathbf{c}}) - r(w^*)$ (points) and $\mathcal{R}_\gamma(W_{\mathbf{c}})$ (crosses) for $\mathbf{c} = \mathbf{c}^\rho$ (3.28), $\rho = 0.9$, $n = 200, d = \lfloor \gamma n \rfloor$, $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{c})$, $\sigma^2 = 1, \|w^*\|_2^2 \in \{1, 4\}$.

Now, we compare $\mathcal{R}_\gamma(W_{\ell_2})$, $\mathcal{R}_\gamma(W_{\mathbf{c}})$ and $\mathcal{R}_\gamma(W_b)$.

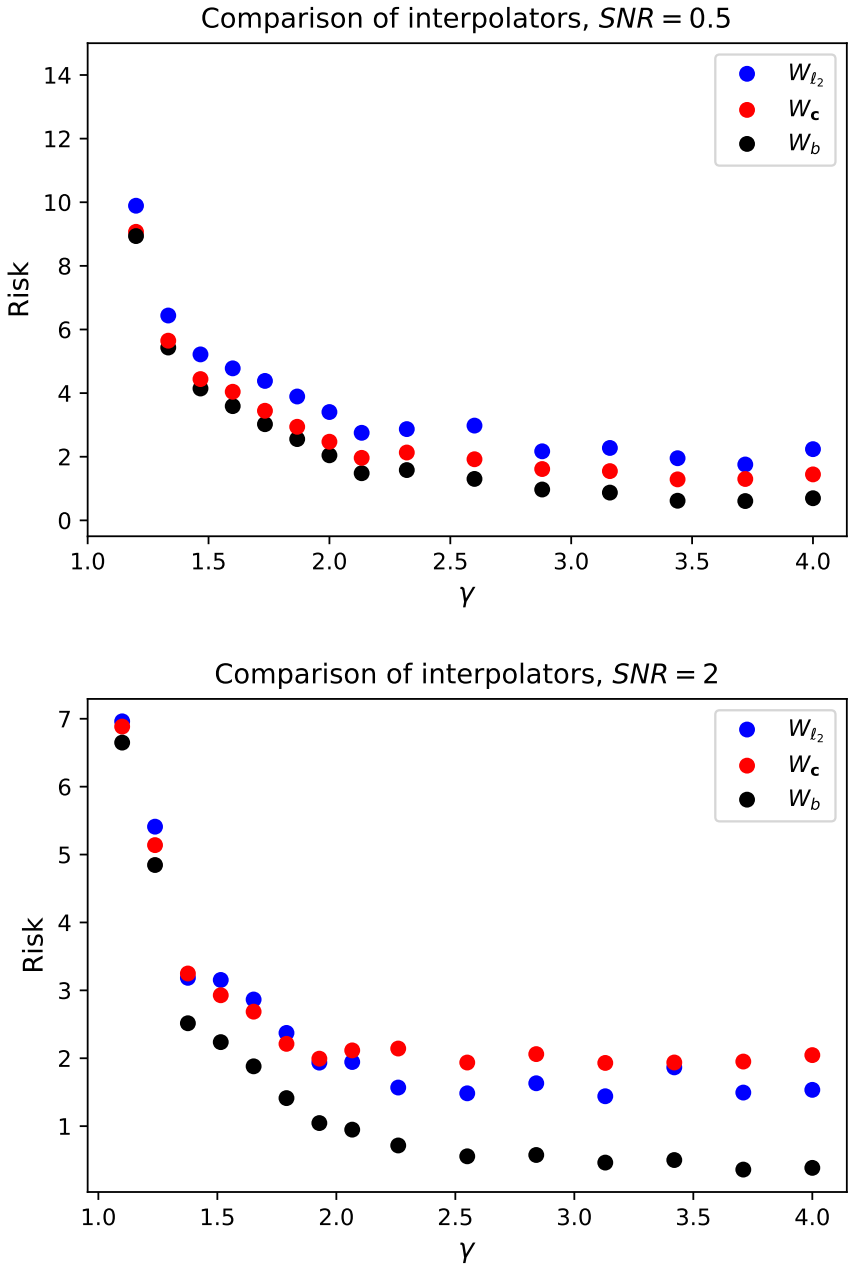


Figure 3.3: Plot of $r(W) - r(w^*)$ for $W \in \{W_{\ell_2}, W_{\mathbf{c}}, W_b\}$, $\mathbf{c} = \mathbf{c}^\rho$ (3.28), $\rho = 0.9$, $n = 200, d = \lfloor \gamma n \rfloor$, $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{c})$, $(\|w^*\|_2^2, \sigma^2) \in \{(1,2),(2,1)\}$.

When SNR $= 0.5$, the noise has a larger effect and $W_{\mathbf{c}}$ is better than $W_{\ell_2}$, as expected, because $W_{\mathbf{c}}$ has the variance term of the best possible interpolator $W_b$.

When SNR $= 2$, $W_{\mathbf{c}}$ does not perform better. For $1 \leq \gamma$ close to 1, the variance term is dominant which is noticeable in the plot. Moreover, we know

$$\mathcal{R}_\gamma(W_{\ell_2}) \longrightarrow \|w^*\|_2^2$$
$$\mathcal{R}_\gamma(W_{\mathbf{c}}) \longrightarrow \mathcal{R}(0) = w^* \mathbf{c} w^*$$

as $\gamma \to \infty$, by (3.29) and (3.30), so we can deduce $w^* \mathbf{c} w^* > \|w^*\|_2^2$ from the plot.

This suggests that performance of $W_{\mathbf{c}}$ can depend on the alignment of $w^*$ (here, initialised randomly) with eigenvectors of $\mathbf{c}$.

Here, only $\|w^*\|_2^2$ is fixed, not $\mathcal{R}(0) = w^*\mathbf{c}w^*$. In the first plot, however, $\mathcal{R}(0) \approx \|w^*\|_2^2 < \sigma^2$ and hence

$$\mathcal{R}_\gamma(W_{\ell_2}) \geq \mathcal{R}_\gamma(W_{\mathbf{c}}) := \mathcal{R}(0)(1 - \frac{1}{\gamma}) + \frac{\sigma^2}{\gamma - 1} \geq \mathcal{R}(0),$$

for all $\gamma > 1$, so both $W_{\ell_2}$ and $W_{\mathbf{c}}$ performed worse than $w = 0$. Noticeably, $W_{\mathbf{c}}$ did better than $W_{\ell_2}$ when SNR $< 1$ because of its variance term, but simultaneously, did no learning, because SNR $< 1$ implied $\mathcal{R}(0) < \sigma^2$. Can $W_{\mathbf{c}}$ generalise better than $W_{\ell_2}$ even when SNR $> 1$ and $\mathcal{R}(W_{\mathbf{c}}) < \mathcal{R}(0)$ (i.e. it learns)?

Yes. $W_{\mathbf{c}}$ can also do well when $\mathbf{c}$ is ill-behaved because

$$V_{\mathbf{c}} = \mathbb{E}_\xi(\xi^T(\mathbf{Z}\mathbf{Z}^T)^\dagger\xi)$$

is independent of $\mathbf{c}$, while

$$V_{\ell_2} = \mathbb{E}_\xi(\xi^T(\mathbf{X}^\dagger)^T\mathbf{c}\mathbf{X}^\dagger\xi)$$

is not. Consider

$$c_{i,j} = \begin{cases} \rho^{|i-j|}, & \text{if } i \neq j \\ \log(i + j + e) & \text{if } i = j \end{cases} \tag{3.31}$$

and let $w^*$ have only first 10 entries nonzero (so that $\mathcal{R}(0)$ is bounded as $d \to \infty$). In the following example, SNR $> 1$, $\mathcal{R}(W_{\mathbf{c}}) < \mathcal{R}(0)$ and $W_{\mathbf{c}}$ performs significantly better than $W_{\ell_2}$.

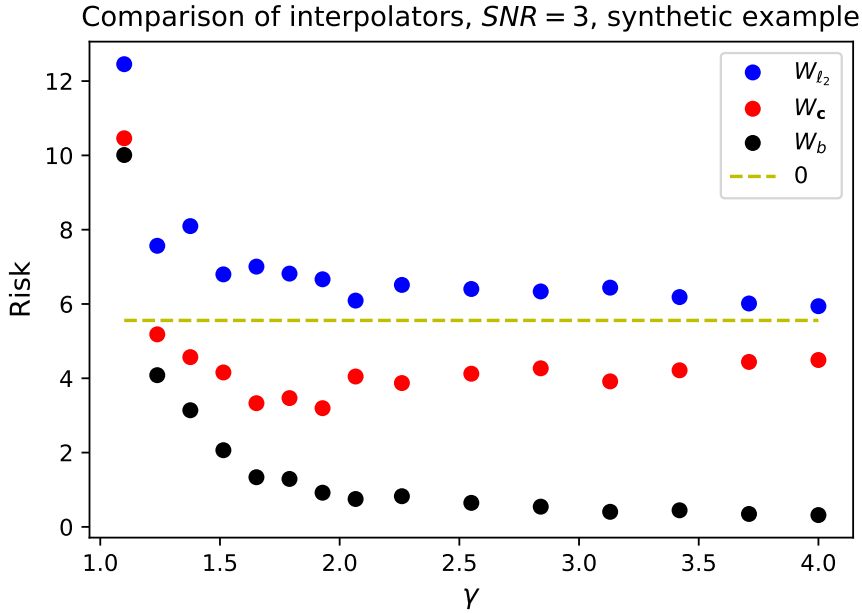


Figure 3.4: Plot of $r(W) - r(w^*)$ for $W \in \{W_{\ell_2}, W_{\mathbf{c}}, W_b, 0\}$ in regime (3.31) for $\rho = 0.9$, $n = 200, d = \lfloor \gamma n \rfloor$, $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{c})$, $\|w^*\|_2^2 = 3$, $\sigma^2 = 1$. Only $w_1, \ldots, w_{10}$ are nonzero. Hence $\mathcal{R}(0)$ is fixed.

See Appendix D for more experiments with some irregular regimes of $\mathbf{c}$ and an example where $W_{\ell_2}$ generalises nearly perfectly.

In summary, by knowledge of implicit bias of mirror descent, and by exploiting properties of a particular learning problem (knowing $\mathbf{c}$) we constructed an estimator, $W_{\mathbf{c}}$, to suit it. A final verdict is not clear, but $W_{\mathbf{c}}$ seems to perform better than $W_{\ell_2}$ in noisy settings. This not only shows how knowledge of implicit bias can explain generalisation, but also assist in enhancing it.

# Chapter 4

# Conclusion

The first goal of this project was to characterise the implicit bias of projected mirror descent for a unique root loss on linear models. We have analysed known results and extended with results of our own, namely, to inequality constraint sets. We also provided an extension of a convergence result for gradient descent.

The second goal was to discuss implications of implicit bias to generalisation in linear regression. First, we answered when is good generalisation of an interpolator even possible and second, whether it is actually achievable. We thoroughly examined the minimum norm interpolator - the implicit bias of (stochastic) gradient descent - and computed its asymptotic risk. Finally, we connected implicit bias and generalisation to construct a new interpolator - the implicit bias of a particular mirror descent - and evaluated its performance.

Our motivating example, neural networks, cannot be yet explained through this study. However, understanding linear regression is necessary first. It is an intriguing question whether implicit bias of mirror descent can be exploited to construct good interpolators for neural networks, just as we did here, in linear regression.

# Appendices

# Appendix A

# The Karush-Kuhn-Tucker (KKT) Conditions

The following is adapted from Nocedal and Wright [42, Chapter 12], Boyd and Vandenberghe [11, Sections 4.2, 5.5] and Humpherys and Jarvis [25].

**Definition A.1.** $\mathcal{D} \subseteq \mathbb{R}^d$ is *convex* if for all $x, y \in \mathcal{D}$ and $t \in [0,1]$,

$$tx + (1-t)y \in \mathcal{D}.$$

Let $\mathcal{C}, \mathcal{D} \subseteq \mathbb{R}^d$, $f : \mathcal{D} \to \mathbb{R}$ with $\mathcal{D}$ open and convex.

**Definition A.2.** $f$ is *convex* if for all $x, y \in \mathcal{D}$ and all $t \in [0,1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

f is strictly convex if for all $x \neq y \in \mathcal{D}$ and $t \in [0,1]$,

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y).$$

**Proposition A.3.** If $f$ is differentiable, then it is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x),$$

for all $x, y \in \mathcal{C}$. For strict convexity, the inequality is strict.

**Definition A.4.** We say that finding $w \in \mathcal{D}$ such that for all $z \in \mathcal{D}$,

$$f(w) \leq f(z),$$

is a *minimisation (or optimisation) problem.* We denote it

$$\arg\min_{w \in \mathcal{D}} f(w).$$

If there exist $c_j, \widetilde{c}_i : \mathcal{S} \to \mathbb{R}$ with $j \in J$, $i \in I$, where $I$, $J$ are finite index sets and

$$\mathcal{C} = \{w \in \mathcal{S} : c_j(w) = 0, \widetilde{c}_i(w) \leq 0 \text{ for all } j \in J, i \in I\},$$

where $\mathcal{C} \cap \mathcal{D} \neq \emptyset$. Then

$$\arg\min_{w \in \mathcal{C}} f(w) := \arg\min_{w \in \mathcal{C} \cap \mathcal{D}} f(w) \tag{A.1}$$

is a minimisation problem with constraint set $\mathcal{C}$. In this work, we always have $\mathcal{S} \in \{\mathbb{R}^d, \mathcal{D}\}$, $f, c_j, \widetilde{c}_i$ are continuously differentiable and $\mathcal{C}$ is closed.

**Definition A.5.** If $\mathcal{S}$ is convex, $f, \widetilde{c}_i$ are convex and $c_j$ are affine, i.e. there exist $a_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}$ such that

$$c_j(w) = a_j^T w - b_j,$$

for all $i \in I$, $j \in J$, then

$$\mathcal{C} = \{w \in \mathcal{S} : c_j(w) = 0, \widetilde{c}_i(w) \leq 0 \text{ for all } j \in J, i \in I\}$$

is a convex set and (A.1) is a *convex minimisation problem.*

**Definition A.6.** For a constrained minimisation problem (A.1) we define the *Lagrangian funtion*

$$\mathcal{L} : \mathcal{D} \times \mathbb{R}^{|J|} \times \mathbb{R}^{|I|}_{\geq 0} \longrightarrow \mathbb{R}^d,$$

by

$$\mathcal{L}(w, \mu, \lambda) = \nabla f(x) + \sum_{j \in J} \nabla c_j(w)\mu_j + \sum_{i \in I} \nabla \widetilde{c}_i(w)\lambda_i,$$

where $\mathbb{R}^{|I|}_{\geq 0} = \{\lambda \in \mathbb{R}^{|I|} : \lambda_i \geq 0 \text{ for all } i \in I\}$

**Definition A.7.** The constrained minimisation problem (A.1) satisfies *Slater's condition* if there exists $w \in \mathcal{C} \cap \mathcal{D}$ such that $\widetilde{c}_i(w) < 0$ for all $i \in I$.

**Theorem A.8.** If the constrained minimisation problem (A.1) is convex and satisfies Slater's condition, then $w^\dagger \in \mathcal{D}$ is its unique solution if and only if there exist $\mu \in \mathbb{R}^{|J|}$, $\lambda \in \mathbb{R}^{|I|}_{\geq 0}$ such that

$$\nabla \mathcal{L}(w^\dagger, \mu, \lambda) = 0, \tag{A.2}$$

$$\lambda_i \widetilde{c}_i(w^\dagger) = 0 \text{ for all } i \in I, \tag{A.3}$$

$$w^\dagger \in \mathcal{C} \cap \mathcal{D}. \tag{A.4}$$

These are the *KKT conditions.* (A.3) is called *complementary slackness.* If $I = \emptyset$, this holds without Slater's condition.

This is proved in [25, Chapter 15].

# Appendix B

# More Implicit Bias

In relation to Proposition 2.13, we provide an example, along with intuition, where $\Phi(W_\infty) = r$, $\nabla\Phi(w_1) \notin \text{Span}\{X_i : i \in \{1, \dots, n\}\} = \mathcal{M}$ and

$$W_\infty = \underset{w \in \mathcal{G} \cap \mathcal{C}}{\arg\min} D_\Phi(w, w_1).$$

It illustraters that $\Phi(W_\infty) \neq r$ in the second implication of Proposition 2.13 is required, and inspires more results about projected gradient descent.

**Example B.1.** Consider projected gradient descent. By the Projection Theorem (C.11) for any $w \in \mathbb{R}^d$,

$$w = P_{\mathcal{M}}(w) + P_{\mathcal{M}^\perp}(w),$$

where and $P_{\mathcal{M}}, P_{\mathcal{M}^\perp}$ are projection maps. Note

$$\widetilde{W}_{t+1} = W_t - \eta_t \nabla R(W_t)$$

does not change the projection onto $\mathcal{M}^\perp$, because $\nabla R(W_t) \in \mathcal{M}$. Therefore, the projection step

$$W_{t+1} = \begin{cases} \widetilde{W}_{t+1} & \text{if } \|\widetilde{W}_{t+1}\|_2 \leq r \\ \frac{\widetilde{W}_{t+1}}{\|\widetilde{W}_{t+1}\|_2} r & \text{if } \|\widetilde{W}_{t+1}\|_2 > r \end{cases}$$

induces

$$P_{\mathcal{M}^\perp}(W_{t+1}) = \begin{cases} P_{\mathcal{M}^\perp}(W_t) & \text{if } \|\widetilde{W}_{t+1}\|_2 \leq r \\ P_{\mathcal{M}^\perp}(W_t) \frac{r}{\|\widetilde{W}_{t+1}\|_2} & \text{if } \|\widetilde{W}_{t+1}\|_2 > r. \end{cases}$$

Hence

$$\|P_{\mathcal{M}^\perp}(W_{t+1})\|_2 \leq \|P_{\mathcal{M}^\perp}(W_t)\|_2 \tag{B.1}$$

for all $t \in \mathbb{N}$ and so, to get $W_\infty = W^\dagger := \arg\min_{w \in \mathcal{G} \cap \mathcal{C}} \|w - w_1\|_2$, we need a learning rate so that $\|P_{\mathcal{M}^\perp}(W_t)\|_2$ decreases but never surpasses $\|P_{\mathcal{M}^\perp}(W^\dagger)\|_2$. If $\mathcal{C} = \overline{B}_1(0), X = (1, 1) \in \mathbb{R}^2, Y = 1, w_1 = (3, 1)$ and $(\eta_t)_{t \geq 2}$ which always halves (or quarters) the distance between $W_t$ and $\mathcal{G}$ then it works (see Figure B.1).

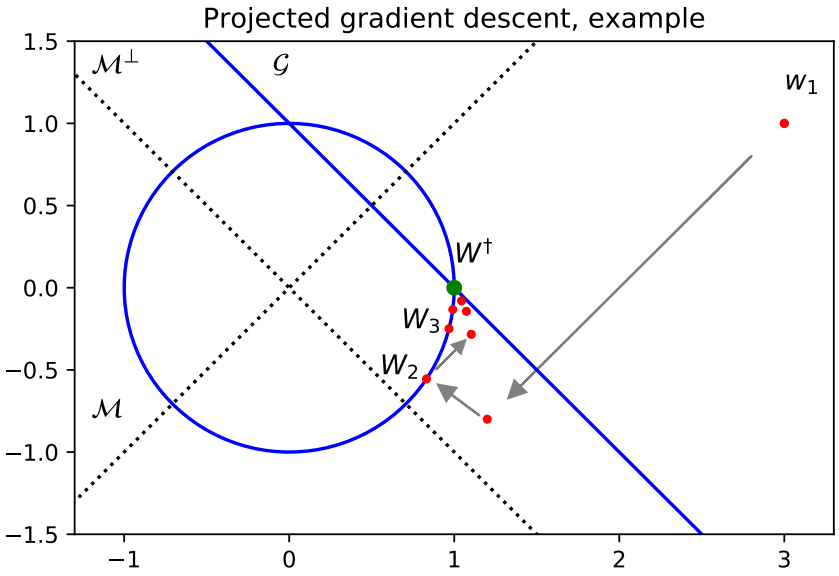


Figure B.1

Notice that

$$w_1 \notin \text{Span}\{(1,1)\} = \mathcal{M}$$

and

$$\lim_{t\to\infty} W_t = (1,0) = W^\dagger = \arg\min_{w\in\mathcal{G}\cap\mathcal{C}} \|w - w_1\|_2.$$

The specific learning rate was

$$\eta_1 = 0.3$$
$$\eta_t = \frac{3}{4}\left(\frac{1 - (W_t)_0 - (W_t)_1}{\|\nabla R(W_t)\|_2}\right)$$

which is $\frac{3}{4}$ of the distance of $W_t$ from $\mathcal{G}$ for points on $\{w \in \mathbb{R}^2 : \|w\|_2 = 1\}$.

A natural question is whether Proposition 2.13 is "continuous" with respect to the initialisation. What if $\nabla\Phi(w_1) \notin \text{Span}\{X_i : i \in \{1,\dots,n\}\}$ but is "close"? Is $W_\infty$ then also close to $\arg\min_{w\in\mathcal{G}\cap\mathcal{C}} D_\Phi(w, w_1)$? Equation (B.1) inspires an answer, for projected gradient descent.

**Proposition B.2.** If projected gradient descent with $\mathcal{C} = \overline{B}_r(0)$, initialisation $w_1$ and stepsize $(\eta_t)_{t\in\mathbb{N}}$ converges to $W_\infty$ and

$$W^\dagger := \arg\min_{w\in\mathcal{G}\cap\mathcal{C}} \|w - w_1\|_2$$

then

$$\|W_\infty - W^\dagger\|_2 \leq 2\|P_{\mathcal{M}^\perp} w_1\|_2$$

This is a stronger result than the first implication of Proposition 2.13, in projected gradient descent, because it quantifies how far we are from the implicit bias based on the initialisation. If $w_1 \in \mathcal{M}$ then it reduces to Proposition 2.13.

*Proof.* For any $a, b \in \mathcal{G}$,

$$\mathbf{X}a = Y = \mathbf{X}b.$$

Hence $a - b \in \mathrm{Ker}(\mathbf{X}) = \mathrm{Im}(\mathbf{X}^T)^\perp = \mathcal{M}^\perp$ (Proposition C.10) which implies

$$P_{\mathcal{M}}(a - b) = P_{\mathcal{M}}(a) - P_{\mathcal{M}}(b) = 0.$$

As $W_\infty, W^\dagger \in \mathcal{G}$, hence

$$\begin{aligned} \|W_\infty - W^\dagger\|_2 &= \|P_{\mathcal{M}^\perp}(W_\infty) - P_{\mathcal{M}^\perp}(W^\dagger)\|_2 \\ &\leq \|P_{\mathcal{M}^\perp}(W_\infty)\|_2 + \|P_{\mathcal{M}^\perp}(W^\dagger)\|_2 \end{aligned}$$

Equation (B.1) implies

$$\|P_{\mathcal{M}^\perp}(W_\infty)\|_2 < \|P_{\mathcal{M}^\perp}(w_1)\|_2,$$

so we only need to prove

$$\|P_{\mathcal{M}^\perp}(W^\dagger)\|_2 \leq \|P_{\mathcal{M}^\perp}(w_1)\|_2.$$

Indeed,

$$\begin{aligned} \|W^\dagger\|_2 &= \| \operatorname*{arg\,min}_{w \in \mathcal{G}, \|w\|_2 \leq r} \|w - w_1\|_2 \|_2 \\ &\leq \| \operatorname*{arg\,min}_{w \in \mathcal{G}} \|w - w_1\|_2 \|_2 \\ &= \|P_{\mathcal{G}}(w_1)\|_2. \end{aligned} \tag{B.2}$$

$W^\dagger, P_{\mathcal{G}}(w_1) \in \mathcal{G}$, hence as before

$$P_{\mathcal{M}}(W^\dagger) = P_{\mathcal{M}}(P_{\mathcal{G}}(w_1))$$

and

$$\begin{aligned} \|W^\dagger\|_2^2 &= \|P_{\mathcal{M}}(W^\dagger)\|_2^2 + \|P_{\mathcal{M}^\perp}(W^\dagger)\|_2^2 \\ &= \|P_{\mathcal{M}}(P_{\mathcal{G}}(w_1))\|_2^2 + \|P_{\mathcal{M}^\perp}(W^\dagger)\|_2^2 \end{aligned}$$

Hence (B.2) implies

$$\|P_{\mathcal{M}^\perp}(W^\dagger)\|_2 \leq \|P_{\mathcal{M}^\perp}(P_{\mathcal{G}}(w_1))\|_2.$$

Finally, it can be checked that

$$P_{\mathcal{M}^\perp}(P_{\mathcal{G}}(w_1)) = P_{\mathcal{M}^\perp}(w_1)$$

because $w_1 - P_{\mathcal{G}}(w_1) \in \mathcal{M}$ (see Propositiona C.14 and C.7, part 5). □

Recall Remark 2.14. If

$$\Phi(W_\infty) = r \tag{B.3}$$

then in Proposition 2.13, we automatically have

$$W_\infty = \underset{w \in \overline{B}_r(0) \cap \mathcal{G}}{\arg\min}\ D_\Phi(w, w_1).$$

However, we do not apriori know (B.3). So, what if we forced it? What if we project to

$$\{w \in \mathbb{R}^d : \Phi(w) = r\}$$

instead of

$$\{w \in \mathbb{R}^d : \Phi(w) \leq r\}?$$

For general mirror descent, $\mathcal{C} = \{w \in \mathbb{R}^d : \Phi(w) = r\}$ is not convex and projections

$$\prod_{\mathcal{C}}^{\Phi}(x) = \underset{y \in \mathcal{C} \cap \mathcal{D}}{\arg\min}\, D_\Phi(y, x),$$

are problematic. However, in projected gradient descent this is well-defined, if $x \neq 0$, by

$$x \longmapsto \frac{x}{\|x\|_2}.$$

**Proposition B.3.** For projected gradient descent with constraint set $\mathcal{C} = \partial\overline{B}_r(0) = \{w \in \mathbb{R}^d : \|w\|_2 = r\}$, initialised at $w_1 \notin \mathcal{G}$, there exists $K > 0$ dependent on $\nabla R$, $w_1$, $r$ such that if $\eta_t < K$ for all $t \in \mathbb{N}$ and the algorithm converges to $W_\infty$ then

$$W_\infty = \underset{w \in \mathcal{G} \cap \mathcal{C}}{\arg\min} \|w - w_1\|_2. \tag{B.4}$$

The extra restrictions are only to assure that the algorithm never reaches 0, where the projection is ill-defined.

*Proof.* As $\{w_1\} \cup \partial\overline{B}_r(0)$ is compact and $\nabla R$ is continuous, there exists $k > 0$ such that

$$\|\nabla R(w)\|_2 \leq k$$

for all $w \in \{w_1\} \cup \partial\overline{B}_r(0)$. The iterates, by definition, satisfy

$$W_t \in \{w_1\} \cup \partial\overline{B}_r(0)$$

for all $t \in \mathbb{N}$. Therefore, if

$$\eta_t < K := \frac{\min(r, \|w_1\|_2)}{k},$$

then $\widetilde{W_t} = W_t - \eta_t \nabla R(W_t) \neq 0$ for all $t \geq 2$ and the projection is always well-defined.

The rest follows by KKT conditions (Theorem A.8). Indeed, as

$$W_{t+1} = \underset{\|w\|_2^2 = r^2}{\arg\min} \|w - \widetilde{W}_{t+1}\|_2^2,$$

there exists $\delta_t \in \mathbb{R}$ with

$$\begin{aligned} W_{t+1}(1-\delta_t) &= \widetilde{W}_{t+1} \\ &= W_t - \eta_t \nabla R(W_t), \end{aligned}$$

where $\delta_t \neq 1$ as $\widetilde{W}_{t+1} \neq 0$. Then inductively,

$$W_{t+1} = \prod_{s=1}^{t} \frac{1}{1-\delta_s} w_1 - \xi_t \in \mathrm{Span}(w_1) + \mathcal{M}$$

Similarly, the KKT conditions for (B.4) imply that $W^\dagger = \arg\min_{w \in \mathcal{G} \cap \mathcal{C}} \|w - w_1\|_2$ if and only if $W^\dagger \in \mathcal{G} \cap \mathcal{C}$ and there exist $\delta \in \mathbb{R}$, $\mu \in \mathbb{R}^n$ with

$$W^\dagger(1-\delta) = w_1 + \mathbf{X}^T \mu. \tag{B.5}$$

Because $\mathrm{Span}(w_1) + \mathcal{M}$ is closed, using inductive arguments gives

$$W_\infty \in (\mathrm{Span}(w_1) + \mathcal{M}) \cap \mathcal{C} \cap \mathcal{G}.$$

Existence of $\delta \in \mathbb{R}$, $\mu \in \mathbb{R}^n$ required by (B.5) now follows. □

# Appendix C

# Linear Algebra

The following is adapted from Nocedal and Wright [42, Appendix A], Trefethen [54, Chapter 1], and Penrose [44], [45].

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and let $\mathbf{m} \in \mathbb{R}^{k \times k}$ be any symmetric matrix with eigenvalues $\lambda_{\max}(\mathbf{m}) = \lambda_1 \geq \cdots \geq \lambda_k = \lambda_{\min}(\mathbf{m})$.

**Proposition C.1.** For all $v \in \mathbb{R}^n \setminus \{0\}$,

$$\lambda_{\min}(\mathbf{m}) \leq \frac{v^T \mathbf{m} v}{\|v\|_2^2} \leq \lambda_{\max}(\mathbf{m}). \tag{C.1}$$

**Definition C.2.** $\mathbf{m}$ is *positive semi-definite* if

$$\lambda_{\min}(\mathbf{m}) \geq 0$$

and *positive definite* if

$$\lambda_{\min}(\mathbf{m}) > 0.$$

**Proposition C.3.**

$$\|\mathbf{X}\|_2 := \sup\{\frac{\|\mathbf{X}v\|_2}{\|v\|_2} : v \in \mathbb{R}^m \setminus \{0\}\} = \sqrt{\lambda_{\max}(\mathbf{X}^T\mathbf{X})}$$

If $d = n$ and $\mathbf{X}$ is symmetric positive semi-definite then

$$\|\mathbf{X}\|_2 = \lambda_{\max}(\mathbf{X}).$$

**Definition C.4.** $\mathbf{X}$ is *orthogonal* if $n = d$ and

$$\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{X}^T = I.$$

**Theorem C.5.** There exist $\mathbf{\Sigma} \in \mathbb{R}^{n \times d}$ and orthogonal $\mathbf{U} \in \mathbb{R}^{n \times n}$, $\mathbf{V} \in \mathbb{R}^{d \times d}$ such that

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{C.2}$$

Moreover, $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Sigma}$ can be chosen so that

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$$
$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T,$$

where

$$\mathbf{\Sigma}_{ij} = \begin{cases} \sqrt{\lambda_i(\mathbf{X}^T\mathbf{X})} & \text{if } i = j \leq r \\ 0 & \text{otherwise} \end{cases}$$

and $r \leq \min(n, d)$ is the rank of $\mathbf{X}$. Alternatively,

$$\mathbf{X} = \mathbf{U}_{1:r}\mathbf{\Sigma}_{1:r}\mathbf{V}_{1:r}^T, \tag{C.3}$$

where $\mathbf{g}_{1:r}$ denotes only including the first $r$ columns of $\mathbf{g} \in \{\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}\}$. (C.2) is the *singular value decomposition* (SVD) and (C.3) is the *compact SVD* of $\mathbf{X}$.

**Definition C.6.** [44, Theorem 1] There exists a unique matrix $\mathbf{X}^\dagger \in \mathbb{R}^{d\times n}$ which satisfies

1. $\mathbf{X}\mathbf{X}^\dagger\mathbf{X} = \mathbf{X}$,
2. $\mathbf{X}^\dagger\mathbf{X}\mathbf{X}^\dagger = \mathbf{X}^\dagger$,
3. $(\mathbf{X}\mathbf{X}^\dagger)^T = \mathbf{X}\mathbf{X}^\dagger$,
4. $(\mathbf{X}^\dagger\mathbf{X})^T = \mathbf{X}^\dagger\mathbf{X}$,

called the *Moore-Penrose pseudoinverse*.

**Proposition C.7.** The Moore-Penrose pseudoinverse satisfies the following.

1. If $\mathbf{\Sigma} = \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$ is diagonal then $\mathbf{\Sigma}^\dagger = \text{diag}(1/\lambda_1, \ldots, 1/\lambda_r, 0, \ldots, 0)$.
2. If $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD of $\mathbf{X}$, then $\mathbf{X}^\dagger = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T$.
3. If $\mathbf{X} = \mathbf{U}_{1:r}\mathbf{\Sigma}_{1:r}\mathbf{V}_{1:r}^T$ is the compact SVD of $\mathbf{X}$, then $\mathbf{X}^\dagger = \mathbf{V}_{1:r}\mathbf{\Sigma}_{1:r}^{-1}\mathbf{U}_{1:r}^T$.
4. $\mathbf{X}^\dagger = (\mathbf{X}^T\mathbf{X})^\dagger\mathbf{X}^T$.
5. $\mathbf{X}^\dagger = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^\dagger$.

**Theorem C.8.** [45, Theorem] If $Y \in \mathbb{R}^n$, then $W_{\ell_2} = \mathbf{X}^\dagger Y$ is the *best approximate solution* to the equation $\mathbf{X}w = Y$. That is, for all $w \in \mathbb{R}^d$, either

$$\|\mathbf{X}w - Y\|_2 > \|\mathbf{X}W_{\ell_2} - Y\|_2$$

or

$$\|\mathbf{X}w - Y\|_2 = \|\mathbf{X}W_{\ell_2} - Y\|_2 \text{ and } \|w\|_2 \geq \|W_{\ell_2}\|_2$$

**Definition C.9.** If $\mathcal{H}$ is a Hilbert space and $\mathcal{A} \subseteq \mathcal{H}$,

$$\mathcal{A}^\perp := \{y \in \mathcal{H} : \langle \eta, y\rangle_{\mathcal{H}} = 0 \text{ for all } \eta \in \mathcal{A}\}.$$

is the *orthogonal complement* of $\mathcal{A}$.

**Proposition C.10.** The following hold.

1. $\mathrm{Ker}(\mathbf{X})^{\perp} = \mathrm{Im}(\mathbf{X}^T)$
2. $\mathrm{Im}(\mathbf{X})^{\perp} = \mathrm{Ker}(\mathbf{X}^T)$
3. $\mathrm{Ker}(\mathbf{X}^T\mathbf{X}) = \mathrm{Ker}(\mathbf{X})$
4. $\mathrm{Im}(\mathbf{X}^T\mathbf{X}) = \mathrm{Im}(\mathbf{X}^T)$

See Strang [50].

**Theorem C.11.** If $\mathcal{H}$ is a Hilbert space and $\mathcal{A} \subseteq \mathcal{H}$ is nonempty, convex and closed then there exists $P_{\mathcal{A}} : \mathcal{H} \to \mathcal{A}$, called the *projection* onto $\mathcal{A}$, such that for all $x \in \mathcal{H}$

$$P_{\mathcal{A}}(x) = \arg\min_{y \in \mathcal{A}} \|x - y\|_{\mathcal{H}}.$$

Moreover, $\zeta = P_{\mathcal{A}}(x)$ if and only if

$$\langle x - \zeta, \eta - \zeta \rangle_{\mathcal{H}} \leq 0$$

for all $\eta \in \mathcal{A}$. If $\mathcal{A}$ is a vector subspace of $\mathcal{H}$ then $\zeta = P_{\mathcal{A}}(x)$ if and only if

$$\zeta \in \mathcal{A} \text{ and } x - \zeta \in \mathcal{A}^{\perp}.$$

Moreover, for all $x \in \mathcal{H}$,

$$x = P_{\mathcal{A}}(x) + P_{\mathcal{A}^{\perp}}(x)$$

and

$$\|x\|_{\mathcal{H}}^2 = \|P_{\mathcal{A}}(x)\|_{\mathcal{H}}^2 + \|P_{\mathcal{A}^{\perp}}(x)\|_{\mathcal{H}}^2.$$

See Rudin [48, Theorem 4.11] and Domokos et al. [16, Theorem 1.1].

**Proposition C.12.** The maps $\mathbb{R}^d \ni w \mapsto \mathbf{X}\mathbf{X}^{\dagger}w$ and $\mathbb{R}^d \ni w \mapsto \mathbf{X}^{\dagger}\mathbf{X}w$ are the projections onto $\mathrm{Im}(\mathbf{X})$ and $\mathrm{Im}(\mathbf{X}^T)$, respectively.

*Proof.* Clearly, $\mathbf{X}\mathbf{X}^{\dagger}w \in \mathrm{Im}(\mathbf{X})$ and for any $z \in \mathbb{R}^d$,

$$\langle w - \mathbf{X}\mathbf{X}^{\dagger}w, \mathbf{X}z \rangle = w^T\mathbf{X}z - w^T(\mathbf{X}\mathbf{X}^{\dagger})^T\mathbf{X}w = 0,$$

by properties 4 and 1 of Definition C.6. Hence $w - \mathbf{X}\mathbf{X}^{\dagger}w \in \mathrm{Im}(\mathbf{X})^{\perp}$.

Similarly, by Definition C.6, $\mathbf{X}^{\dagger}\mathbf{X}w = (\mathbf{X}^{\dagger}\mathbf{X})^T w = \mathbf{X}^T{\mathbf{X}^{\dagger}}^T w \in \mathrm{Im}(\mathbf{X}^T)$ and

$$\begin{aligned}\langle w - \mathbf{X}^{\dagger}\mathbf{X}w, \mathbf{X}^T z \rangle &= w^T\mathbf{X}^T z - w^T\mathbf{X}^T{\mathbf{X}^{\dagger}}^T\mathbf{X}^T z \\ &= w^T\mathbf{X}^T z - w^T(\mathbf{X}\mathbf{X}^{\dagger}\mathbf{X})^T z \\ &= 0.\end{aligned}$$

By Theorem C.11, we are done. □

**Proposition C.13.** If $\mathbf{X} = \mathbf{U}_{1:r}\mathbf{\Sigma}_{1:r}\mathbf{V}_{1:r}^T$ is the compact SVD of $\mathbf{X}$ then

$$\mathbf{X}^\dagger\mathbf{X} = \mathbf{V}_{1:r}\mathbf{V}_{1:r}^T$$

and

$$\mathbf{X}\mathbf{X}^\dagger = \mathbf{U}_{1:r}\mathbf{U}_{1:r}^T.$$

*Proof.* This follows from orthogonality of $\mathbf{U}, \mathbf{V}$ and Proposition C.7. □

**Proposition C.14.** Let $\mathcal{G} = \{w \in \mathbb{R}^d : \mathbf{X}w = Y\}$ be nonempty. Then

$$P_{\mathcal{G}} : \mathbb{R}^d \ni w \longmapsto \mathbf{X}^\dagger Y + w - \mathbf{X}^\dagger\mathbf{X}w$$

is the projection $P_{\mathcal{G}} : \mathbb{R}^d \to \mathcal{G}$.

*Proof.* Let $\zeta = \mathbf{X}^\dagger Y + w - \mathbf{X}^\dagger\mathbf{X}w$. Clearly, $\mathcal{G}$ is convex and closed. As $\mathcal{G}$ is nonempty, Theorem C.8 implies

$$\mathbf{X}\mathbf{X}^\dagger Y = Y.$$

As $\mathrm{Ker}(\mathbf{X})^\perp = \mathrm{Im}(\mathbf{X}^T)$ (Proposition C.10), $w \mapsto w - \mathbf{X}^\dagger\mathbf{X}w$ is the projection onto $\mathrm{Ker}(\mathbf{X})$ by Proposition C.12 and Theorem C.11. Therefore,

$$\mathbf{X}\zeta = Y$$

and $\zeta \in \mathcal{G}$. Now, we show

$$\langle w - \zeta, z - \zeta\rangle = 0$$

for all $z \in \mathcal{G}$. As $z - \mathbf{X}^\dagger Y \in \mathrm{Ker}(\mathbf{X})$ and $w - \mathbf{X}^\dagger\mathbf{X}w \in \mathrm{Ker}(\mathbf{X})$,

$$z - \zeta \in \mathrm{Ker}(\mathbf{X}).$$

Finally, $w - \zeta \in \mathrm{Im}(\mathbf{X}^T)$, because $\mathbf{X}^\dagger\mathbf{X}w = P_{\mathrm{Im}(\mathbf{X}^T)}(w)$ and $\mathbf{X}^\dagger Y = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^\dagger Y \in \mathrm{Im}(\mathbf{X}^T)$. Thus

$$\langle w - \zeta, z - \zeta\rangle = 0$$

because $w - \zeta \in \mathrm{Im}(\mathbf{X}^T) = \mathrm{Ker}(\mathbf{X})^\perp$ and $z - \zeta \in \mathrm{Ker}(\mathbf{X})$. □

# Appendix D

# More Experiments

We present more comparisons of $W_{\ell_2}, W_{\mathbf{c}}$ and $W_b$ for random matrices generated as follows.

```
import numpy as np

def make_spd_matrix(d,eigenval = '0.5',sdfactor = 3.5, equicor = 'false', ro = 0.9):
#Generates a random dxd symmetric positive definite matrix c, whose every
#eigenvalue has mean = eigenval. Meaning of sdfactor is clear from the context.
#If equicorrelated = true, then it generates an 'equicorrelated' matrix with
#parameter rho. See Below.

    if eigenval == '0.5':
        A = np.random.rand(d,d) #dxd matrix with entries i.i.d. from Uniform(0,1)

        U,s,V = np.linalg.svd(A.transpose().dot(A)) #diagonal form of A^TA

        #Replaces each eigenvalue by a Uniform(0,1) random variable
        c = np.dot(U,np.diag(np.random.rand(d))).dot(U.transpose())

        #The matrices produced here typically have off-diagonal entries which
        #are between 1/10 - 1/1000 of the diagonal entries for n=200,
        #d = int(gamma*d). This range is smaller for smaller n.

    if eigenval == '1':
        A = np.random.rand(d,d) #dxd matrix with entries i.i.d. from Uniform(0,1)

        U,s,V = np.linalg.svd(A.transpose().dot(A)) #diagonal form of A^TA

        #Replaces each eigenvalue by a Uniform(0.5,1.5) random variable
        c = np.dot(U,np.diag(np.random.rand(d)+0.5)).dot(U.transpose())

        #The matrices produced here typically have off-diagonal entries which
        #are between 1/10 - 1/1000 of the diagonal entries for n=200,
        #d = int(gamma*d). This range is smaller for smaller n.

    if eigenval == 'log':
        A = np.random.rand(d,d) #dxd matrix with entries i.i.d. from Uniform(0,1)

        U,s,V = np.linalg.svd(A.transpose().dot(A)) #diagonal form of A^TA

        #Takes the diagonal form of A^TA and replaces each eigenvalue by a normal
        #random variable with mean log(d) and standard deviation np.log(d)/sdfactor.
        #Note there is a small probability that some eigenvalue will be negative,
        #and an error is produced later in generating data.

        mean_m = np.log(d)*np.ones(d)
        cov_m = np.log(d)/sdfactor*np.eye(d)
        s_new = np.random.multivariate_normal(mean_m,cov_m) #new eigenvalues

        c = np.dot(U,np.diag(s_new)).dot(U.transpose())

        #If sdfactor = 3.5 the off-diagonal entries are around 1/10-1/100 of
        #the diagonal entries for n=200, d = int(gamma*d).
        #This range is smaller for smaller n.

        #If sdfactor = 6 the off-diagonal entries are around 1/100 - 1/1000 of
        #the diagonal entries for n=200, d = int(gamma*d).
        #This range is smaller for smaller n.

        #Here, the produced matrices are not far from identity*log(d) matrices.

    if equicor == 'true':
        #matrix with diagonal entries 1 and off diagonal entries rho in [0,1)
        c = (1-ro)*np.eye(d) + ro*np.ones((d,d))

    return c
```

If $\lambda_i(\mathbf{c}) \overset{\text{i.i.d}}{\sim} \text{Uniform}(0+e, 1+e)$ for $e \in \{0.5, 1.5\}$, we observe the regularizing effect of $V_{\mathbf{c}}$. Whether $B_{\mathbf{c}}$ or $B_{\ell_2}$ is better is unclear. We observe little learning ($\mathcal{R}(W) \geq \mathcal{R}(0)$) possibly because small $\lambda_i(\mathbf{c})$ imply small $\mathcal{R}(0)$.

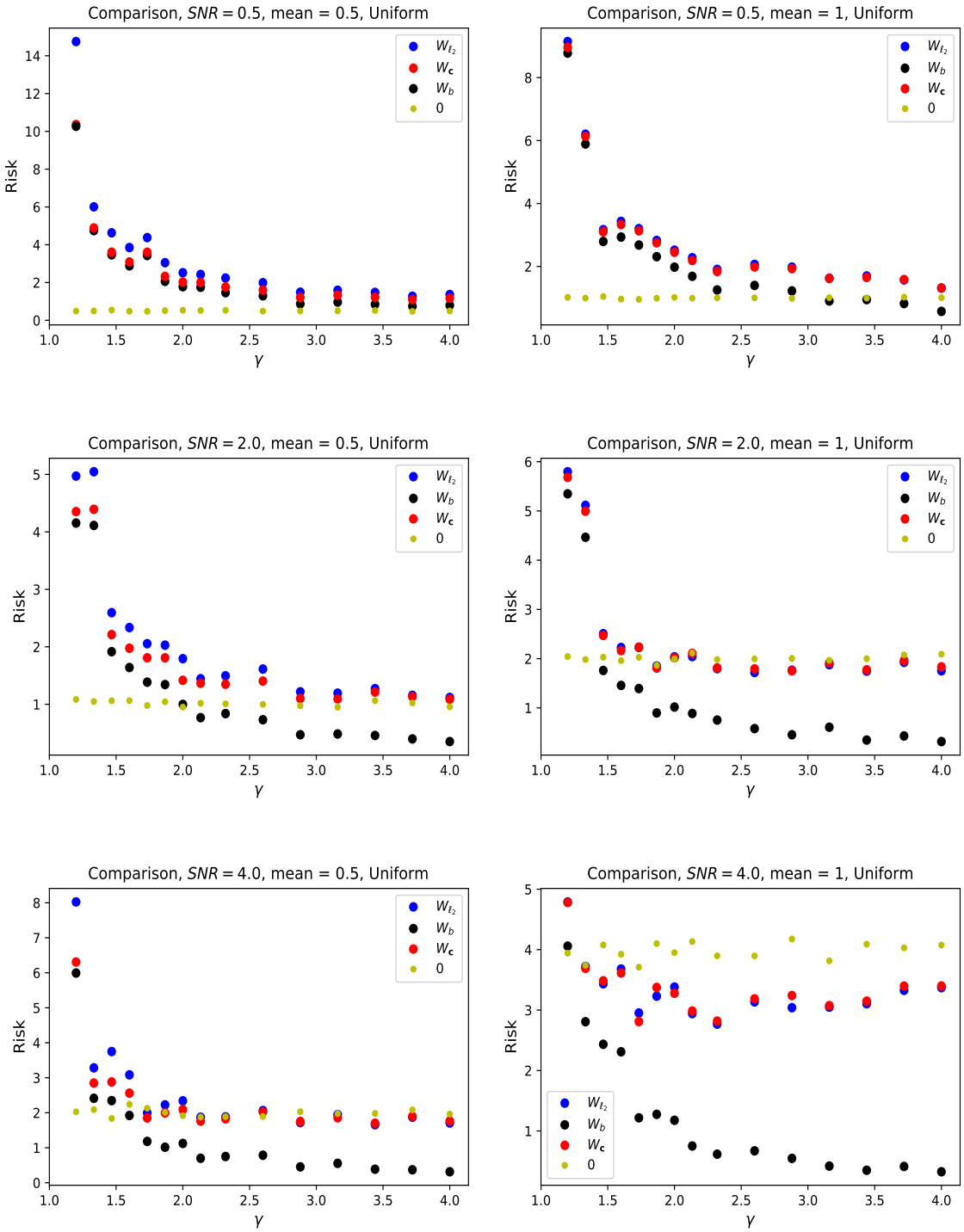


Figure D.3: Plot of $r(W) - r(w^*)$ for $W \in \{W_{\ell_2}, W_{\mathbf{c}}, W_b, 0\}$, SNR $\in \{0.5, 2, 4\}$ in the "eigenval $\in \{0.5, 1\}$" regimes, i.e. $\lambda_i(\mathbf{c}) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0 + \text{eigenval}, 1 + \text{eigenval})$.

If $\lambda_i(\mathbf{c}) \overset{\text{i.i.d}}{\sim} \mathcal{N}(\log(d), (\log(d)/\text{sd})^2)$ for $\text{sd} \in \{3.5, 6\}$, we observe learning ($\mathcal{R}(W) < \mathcal{R}(0)$) and $W_{\ell_2}$, $W_{\mathbf{c}}$ seem to perform similarly. Possibly, because $\mathbf{c} \approx \log(d) I_d$.

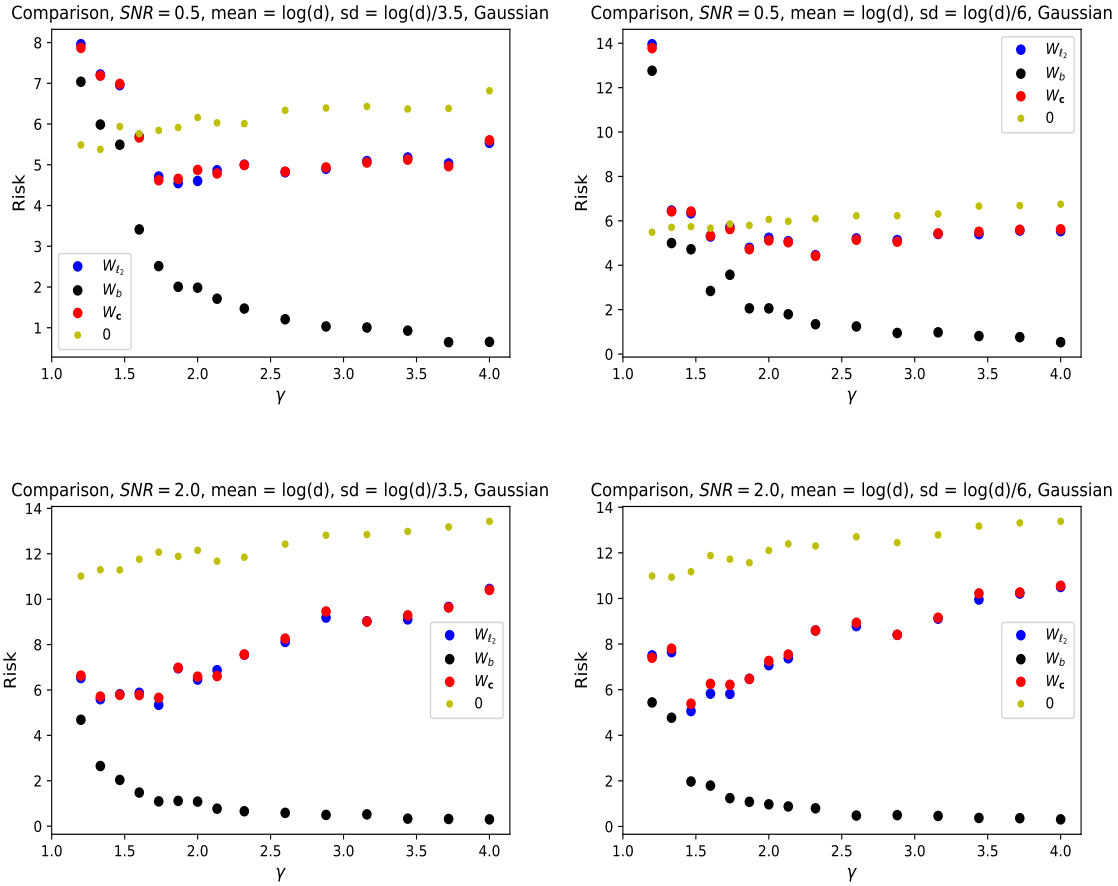


Figure D.5: Plot of $r(W) - r(w^*)$ for $W \in \{W_{\ell_2}, W_{\mathbf{c}}, W_b, 0\}$, SNR $\in \{0.5, 2\}$ in the "eigenval = log" regime, i.e. $\lambda_i(\mathbf{c}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\log(d), (\log(d)/\text{sdfactor})^2)$, sdfactor $\in \{3.5, 6\}$.

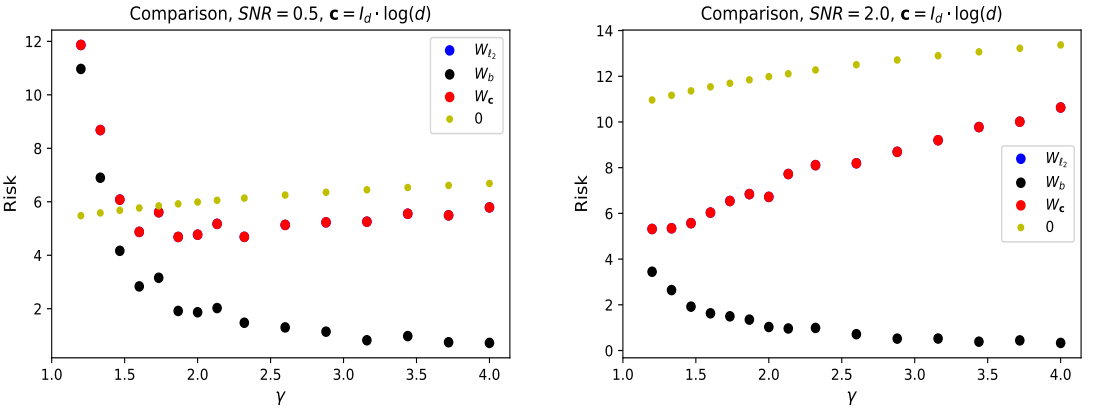


Figure D.6: Plot of $r(W) - r(w^*)$ for $W \in \{W_{\ell_2}, W_{\mathbf{c}}, W_b, 0\}$, SNR $\in \{0.5, 2\}$, $\mathbf{c} = \log(\text{d}) I_d$.

If $\mathbf{c} = (1-\rho)I_d + \rho\mathbf{1}\mathbf{1}^T$, $\rho \in [0,1)$, then [21, Corrolary 2] shows

$$\mathcal{R}_\gamma(W_{\ell_2}) = \|w^*\|_2^2(1-\rho)(1-\frac{1}{\gamma}) + \frac{\sigma^2}{\gamma-1}.$$

Hence, $W_{\ell_2}$ has asymptotically the variance term of $W_b, W_{\mathbf{c}}$. Here, $W_{\mathbf{c}}$ loses its advantage and $W_{\ell_2}$ does remarkably similarly to $W_b$ (nearly perfectly). Moreover, $\mathbf{c}$ has an eigenvector $\mathbf{1}$ with eigenvalue $\rho d + (1-\rho)$ and $d-1$ eigenvalues $1-\rho$. Hence $\mathcal{R}_\gamma(W_{\mathbf{c}}), \mathcal{R}(0)$ are very sensitive to $w^*$ being "close" to $\mathbf{1}$.

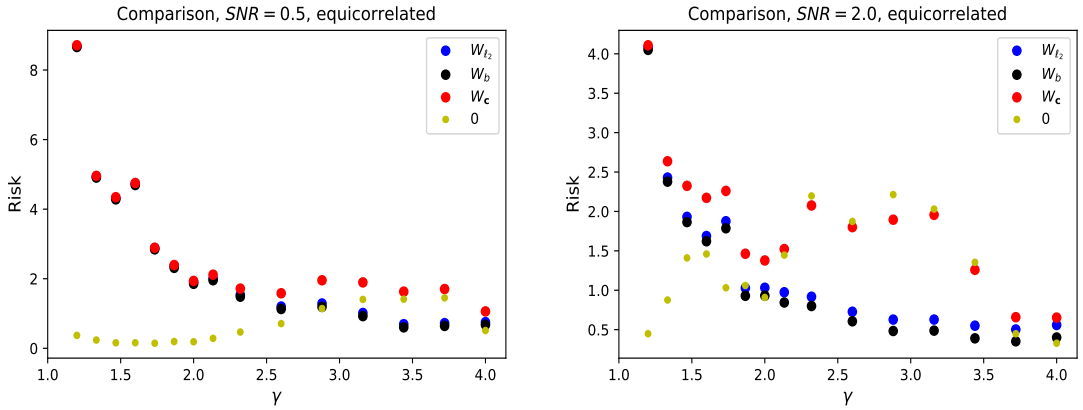


Figure D.7: Plot of $r(W) - r(w^*)$ for $W \in \{W_{\ell_2}, W_{\mathbf{c}}, W_b, 0\}$, SNR $\in \{0.5, 2\}$, $\mathbf{c} = (1-\rho)I_d + \rho\mathbf{1}\mathbf{1}^T$, $\rho = 0.9$.

# Bibliography


[1] S. Amari, "Natural Gradient Works Efficiently in Learning", *Neural Computation* (2) 10 (1998) pp. 251–276.

[2] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit Regularization in Deep Matrix Factorization", *Advances in Neural Information Processing Systems 32* (Curran Associates, Vancouver, 2019) 7413-7424.

[3] Z. D. Bai and Y. Q. Yin, "Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix", *The Annals of Probability* (3) 21 (1993) pp. 1275–1294.

[4] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities", *Annals of Statistics* (4) 33 (2005) pp. 1497–1537.

[5] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results", *Journal of Machine Learning Research* 3 (2003) pp. 463–482.

[6] P. Bartlett, P. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression", *Proceedings of the National Academy of Sciences* (2020).

[7] H. Bauschke and J. Borwein, "Legendre Functions and the Method of Random Bregman Projections", *Journal of Convex Analysis* (1) 4 (1997) pp. 27–67.

[8] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off", *Proceedings of the National Academy of Sciences* (32) 116 (2019) pp. 15849–15854.

[9] M. Belkin, D. Hsu, and J. Xu, *Two models of double descent for weak features*, Preprint, 2019, arXiv: `1903.07571`.

[10] M. Belkin, S. Ma, and S. Mandal, "To Understand Deep Learning We Need to Understand Kernel Learning", *Proceedings of the 35th International Conference on Machine Learning* (PMLR, Stockholm, 2018) 541-549.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, New York, 2004).

[12] S. Bubeck, "Convex Optimization: Algorithms and Complexity", *Foundations and Trends in Machine Learning* (3–4) 8 (2015) pp. 231–357.

[13] S. Bubeck, *Introduction to Online Optimization*, Lecture Notes, Princetion University, 2011, Last Visited: 10.04.2020, `http://sbubeck.com/BubeckLectureNotes.pdf`, 2011.

[14] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games* (Cambridge University Press, New York, 2006).

[15] E. Dobriban and S. Wager, "High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification", *The Annals of Statistics* (1) 46 (2015) pp. 247–279.

[16] A. Domokos, J. Ingram, and M. Marsh, "Projections onto closed convex sets in Hilbert spaces", *Acta Mathematica Hungarica* (1) 152 (2017) pp. 114–129.

[17] F. Götze and A. Tikhomirov, "Rate of convergence in probability to the Marchenko-Pastur law", *Bernoulli* (3) 10 (2004) pp. 503–548.

[18] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit Bias of Gradient Descent on Linear Convolutional Networks", *Advances in Neural Information Processing Systems 31* (Curran Associates, Montreal, 2018) 9461-9471.

[19] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing Implicit Bias in Terms of Optimization Geometry", *Proceedings of the 35th International Conference on Machine Learning* (PMLR, Stockholm, 2018) 1832-1841.

[20] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit Regularization in Matrix Factorization", *Advances in Neural Information Processing Systems 30* (Curran Associates, Long Beach, 2017) 6151-6159.

[21] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*, Preprint, 2019, arXiv: 1903.08560.

[22] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction,* 2nd edn (Springer, New York, 2009).

[23] R. V. Hogg, D. L. Zimmerman, and E. A. Tanis, *Probability and Statistical Inference,* 9th edn (Pearson, New Jersey, 2015).

[24] P. J. Huber, "Robust Estimation of a Location Parameter", *Annals of Mathematical Statistics* (1) 35 (1964) pp. 73–101.

[25] J. Humpherys and T. J. Jarvis, *Foundations of Applied Mathematics, Volume 2: Algorithms, Approximation, Optimization,* (SIAM, Philadelphia, 2020).

[26] Z. Ji and M. Telgarsky, "The implicit bias of gradient descent on nonseparable data", *Proceedings of the 32nd Conference on Learning Theory* (PMLR, Pheonix, 2019) 1772-1798.

[27] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, *A Short Note on Concentration Inequalities for Random Vectors with SubGaussian Norm*, Preprint, 2019, arXiv: 1902.03736.

[28] J. Kivinen and M. K. Warmuth, "Exponentiated Gradient versus Gradient Descent for Linear Predictors", *Information and Computation* (1) 132 (1997) pp. 1–63.

[29] A. N. Kolmogorov and S. V. Fomin, *Introductory real analysis* (Prentice-Hall, New Jersey, 1975).

[30] E. Kreyszig, *Introductory functional analysis with applications* (Wiley, New York, 1989).

[31] O. Ledoit and S. Peche, "Eigenvectors of some large sample covariance matrix ensembles", *Probability Theory and Related Fields* (1-2) 151 (2009) pp. 233–264.

[32] T. Liang and A. Rakhlin, "Just Interpolate: Kernel "Ridgeless" Regression Can Generalize", *The Annals of Statistics*, to appear.

[33] V. Marčenko and L. Pastur, "Distribution of eigenvalues for some sets of random matrices", *Mathematics of the USSR-Sbornik* 1 (1967) pp. 457–483.

[34] H. Masnadi-Shirazi, V. Mahadevan, and N. Vasconcelos, "On the design of robust classifiers for computer vision", *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, San Francisco, 2010) 779-786.

[35] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning,* 2nd edn (The MIT Press, Cambridge, MA, 2012).

[36] V. Muthukumar, K. Vodrahalli, and A. Sahai, "Harmless interpolation of noisy data in regression", *2019 IEEE International Symposium on Information Theory* (2019) pp. 2299–2303.

[37] M. Nacson, J. Lee, S. Gunasekar, N. Srebro, and D. Soudry, *Convergence of Gradient Descent on Separable Data*, Preprint, 2018, arXiv: 1803.01905.

[38] A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization* (Wiley Interscience, New York, 1983).

[39] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring Generalization in Deep Learning", *Advances in Neural Information Processing Systems 30* (Curran Associates, Long Beach, 2017) 5947-5956.

[40] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro, *Geometry of Optimization and Implicit Regularization in Deep Learning*, Preprint, 2017, arXiv: 1705.03071.

[41] B. Neyshabur, R. Tomioka, and N. Srebro, "In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning", *3rd International Conference on Learning Representations* (ICLR, San Diego, 2015).

[42] J. Nocedal and S. J. Wright, *Numerical Optimization*, second, New York, USA: Springer, 2006.

[43] S. Oymak and M. Soltanolkotabi, "Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?", *Proceedings of the 36th International Conference on Machine Learning* (PMLR, Long Beach, 2019) 4951-4960.

[44] R. Penrose, "A generalized inverse for matrices", *Mathematical Proceedings of the Cambridge Philosophical Society* (3) 51 (1955) pp. 406–413.

[45] R. Penrose, "On best approximate solutions of linear matrix equations", *Mathematical Proceedings of the Cambridge Philosophical Society* (1) 52 (1956) pp. 17–19.

[46] P. Rebeschini, *Algorithmic Foundations of Learning*, Lecture Notes, University of Oxford, 2018, Last Visited: 14.04.2020, `http://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/19/`.

[47] R. T. Rockafellar, *Convex Analysis* (Princetion University Press, Princeton, 1970).

[48] W. Rudin, *Real and Complex Analysis,* 3rd edn (McGraw-Hill, New York, 1987).

[49] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The Implicit Bias of Gradient Descent on Separable Data", *Journal of Machine Learning Research* (1) 19 (2018) pp. 2822–2878.

[50] G. Strang, "The Fundamental Theorem of Linear Algebra", *The American Mathematical Monthly* (9) 100 (1993) pp. 848–855.

[51] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", *Proceedings of the 36th International Conference on Machine Learning* (PMLR, Long Beach, 2019) 6105-6114.

[52] T. Tao and V. Vu, "Random Matrices: the Distribution of the Smallest Singular Values", *Geometric and Functional Analysis* (1) 20 (2010) pp. 260–297.

[53] Y. W. Teh, *Advanced Topics in Statistical Machine Learning*, Lecture Notes, University of Oxford, 2020, Last Visited: 28.04.2020, `https://github.com/ywteh/advml2020/blob/master/notes.pdf`.

[54] L. N. Trefethen and D. Bau, *Numerical Linear Algebra* (SIAM, Philadelphia, 1997).

[55] V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer-Verlag, Berlin, 1995).

[56] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices", *Compressed Sensing: Theory and Applications* (Cambridge University Press, Cambridge, MA, 2012).

[57] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge University Press, Cambridge, UK, 2019).

[58] Q. Xie, E. H. Hovy, M. Luong, and Q. V. Le, *Self-training with Noisy Student improves ImageNet classification*, Preprint, 2019, arXiv: `1911.04252`.

[59] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, Preprint, 2016, arXiv: `1611.03530`.

[60] N. Zhang, S.-L. Shen, A. Zhou, and Y.-S. Xu, "Investigation on performance of neural networks using quadratic relative error cost function", *IEEE Access* 7 (2019) pp. 106642–106652.