EPFL

Thèse n° 10 589

Deep Learning Theory Through the Lens of Diagonal Linear Networks

Présentée le 7 juin 2024

Faculté informatique et communications
Laboratoire de théorie en apprentissage automatique
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Scott William PESME

Acceptée sur proposition du jury

Prof. M. Jaggi, président du jury
Prof. N. H. B. Flammarion, directeur de thèse
Prof. N. Srebro, rapporteur
Prof. F. Bach, rapporteur
Prof. L. Chizat, rapporteur

École
polytechnique
fédérale
de Lausanne

2024

"*What a useful thing a pocket-map is!*" I remarked.

"*That's another thing we've learned from your Nation,*" said Mein Herr, "*map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?*"

"*About six inches to the mile.*"

"*Only six inches!*" exclaimed Mein Herr. "*We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!*"

"*Have you used it much?*" I enquired.

"*It has never been spread out, yet,*" said Mein Herr: "*the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well.*"

from Lewis Carroll, *Sylvie and Bruno Concluded, Chapter XI, London, 1895*

Acknowledgements

I was initially skeptical about including personal acknowledgments in a document that would be available online and could be read by strangers, especially since I could directly address them to the people involved. I changed my mind for two reasons. Firstly, this exercise forced me to perform the rather difficult exercise of articulating my heartfelt gratitude. Secondly, these acknowledgments offer a small glimpse into my personal life, reminding the reader that a PhD is not just about equations and dry math, but also about relationships and friendship.

Similar to the content of this manuscript on learning trajectories, I have had the privilege of encountering fantastic people who have guided and helped me complete a PhD which I am proud of. One could say that I have been *implicitly biased* towards where I am today: I have unintentionally encountered people who have positively affected my life trajectory.

I am immensely grateful to have had Nicolas as my PhD advisor. Through our ups and downs, we managed to build a strong relationship and I sincerely believe that I was extremely lucky to be under your supervision. I doubt I could have matched as well with another PhD supervisor. I am very happy with the path you set me on and very proud of the work we did together. Thank you for the trust you gave me along the way and for the genuine kindness you showed during the difficult times.

Loucas joined the lab as a post-doc shortly after the start of my PhD and quickly became my 'academic mentor' whom I looked up to and could always count on. He helped me navigate through the very noisy research landscape. Thank you for always taking the time to help me with my projects and, more importantly, for guiding me through my professional and personal doubts. I hope to one day transmit as much guidance to someone as you have done for me.

I am sincerely thankful to have journeyed through the PhD path alongside numerous colleagues from all around the world, many of whom are now good friends. While I cannot list all the individuals who have been part of the TMLbiO team (I would be afraid of forgetting someone!), I trust that you will recognise yourselves in these lines. Besides the breathtaking view of the lake, the INJ third floor was always filled with open-minded and kind people, with whom I loved discussing with. A special thanks to JB and Aditya, whose friendship I cherish and with whom I had many mind-opening discussions. I express my gratitude (and apologies) to my co-authors Mathieu, Hristo, and Radu, who had to put up with my bizarre taste of doing things last minute when writing papers. Finally, meeting Pierre after our lab moved into the INR building was undoubtedly the highlight of the last months of my PhD. Your open-mindedness, kindness, and intelligence inspire me every day, and I can't wait for our future projects.

During my PhD, I had the fantastic opportunity to do a five-month exchange at RIKEN AIP in Tokyo, and I would like to express my gratitude to Professor Taiji Suzuki for kindly hosting me in his lab. I also thank Sophie, Sho, Pierre-Louis, Kishan, and Wei, whom I had the pleasure of meeting at Tokyo University and with whom I shared extraordinary experiences (in every sense of the word!). And of course, a huge thanks to my close and dear friend Michael, who became my day and night companion for three months straight, as we explored the wonders of Japan together and redid the world. No doubt we will have new exciting adventures in the future.

I am very grateful to have made many friends through the various activities Lausanne has to offer: slacklining, running, climbing, ski touring, mountaineering, acroyoga, chilling by the lake... With a special mention to the Frouzes! Spending time with all of you guys helped me keep things in perspective and stay grounded to the earth.

Thank you to the (numerous!) roommates which I've had over the years. Coming back home and sitting down for a spontaneous meal or discussion was always a real pleasure.

To my family, and especially to my sisters and brothers, thank you for sticking together during the rough times.

To Alice, for your joy, attentiveness, and contagious smile.

And to my parents, for their unconditional love.

Lausanne, May 23, 2024

Abstract

In this PhD manuscript, we explore optimisation phenomena which occur in complex neural networks through the lens of 2-layer diagonal linear networks. This rudimentary architecture, which consists of a two layer feedforward linear network with a diagonal inner weight matrix, has the advantage of revealing interesting training characteristics while keeping the theoretical analysis clean and insightful.

The manuscript is composed of four parts. The first serves as a general introduction to the depicted architecture, it provides results on the optimisation trajectory of gradient flow, upon which the rest of the manuscript is built. The second part focuses on saddle-to-saddle dynamics. Taking the initialisation scale of the gradient flow to zero, we prove and describe the existence of an asymptotic learning trajectory where coordinates are learnt incrementally. In the third part we focus on the effect of various hyperparameters (namely the batch-size, the stepsize and the momentum parameter) on the solution which is recovered by the corresponding gradient method. The fourth and last part takes a slightly different point of view. An underlying mirror-descent structure emerges when analysing gradient descent on diagonal linear networks and slightly more complex architectures. This consequently encourages a deeper understanding of mirror-descent trajectories. In this context, we prove the convergence of the mirror flow in the linear classification setting towards a maximum margin separating hyperplane.

Keywords: theory of deep learning, diagonal linear networks, implicit regularisation, non-convex optimisation, mirror descent

Résumé

Dans ce manuscrit de thèse, nous explorons des phénomènes d'optimisation qui se produisent dans des réseaux de neurones complexes à travers le prisme des réseaux linéaires diagonaux à deux couches. Cette architecture rudimentaire, qui consiste en un réseau linéaire à deux couches avec une matrice de poids interne diagonale, présente l'avantage de révéler des caractéristiques d'entraînement intéressantes tout en conservant une analyse théorique claire et instructive.

Le manuscrit est composé de quatre parties. La première sert d'introduction générale à l'architecture présentée, fournissant des résultats sur la trajectoire d'optimisation du flot de gradient, sur laquelle repose le reste du manuscrit. La deuxième partie se concentre sur les dynamiques selle à selle. En prenant l'échelle d'initialisation du flot de gradient à zéro, nous prouvons et décrivons l'existence d'une trajectoire d'apprentissage asymptotique où les coordonnées sont apprises de manière incrémentale. Dans la troisième partie, nous nous concentrons sur l'effet de différents hyper-paramètres (notamment le 'batch-size', le pas de gradient et le paramètre de momentum) sur la solution retrouvée par la méthode de gradient correspondante. La quatrième et dernière partie adopte un point de vue légèrement différent. Une structure sous-jacente de descente de miroir émerge lors de l'analyse de la descente de gradient sur les réseaux linéaires diagonaux ainsi que sur des architectures légèrement plus complexes. Cela encourage par conséquent une compréhension plus approfondie des trajectoires de descente de miroir. Dans ce contexte, nous prouvons la convergence du flot de miroir vers un hyperplan séparateur de marge maximale dans le cadre de la classification linéaire.

Mots-clés : théorie de l'apprentissage profond, réseaux linéaires diagonaux, régularisation implicite, optimisation non convexe, descente de miroir

Contents

Chapter 1

Introduction

1.1 The deep-learning success story

The breakthroughs in deep learning have begun to reshape various aspects of our societies. In healthcare, these advancements involve training models to enhance clinical diagnostics, aid in drug discovery, as well as provide personalised treatment plans [Esteva et al., 2019, Jumper et al., 2021]. In the transportation sector, deep learning has facilitated the development of autonomous vehicles, improved traffic management systems, and popularised ride-sharing services. In the sector of content creation, generative models now produce texts, music, images, and videos which are sufficiently realistic to pose challenges to traditional industries [Brown et al., 2020, Ramesh et al., 2021, OpenAI, 2024]. These powerful models are not only employed daily by businesses and governments but have also become accessible to everyone through web interfaces, such as conversational agents like ChatGPT, as well as image and video creation platforms like DALL·E and Sora.

The performance of these models challenges our human-centered understanding of truth, intelligence, reasoning, and creativity. In text and image creation, these models blur the distinction between human and machine-generated content, making it increasingly difficult to differentiate between reality and the realm of virtuality [Hsu, 2024]. The computer program AlphaGo, developed to play the game of Go, was described as 'creative' by experts after defeating professional player Lee Sedol [Metz, 2016] and language models can now tackle complex reasoning spanning mathematics, puzzle-solving, vision, and psychology [Bubeck et al., 2023].

These achievements can largely be attributed to three main factors. Firstly, the availability of tremendous amounts of data. To provide a sense of scale, current language models are trained using all publicly available text on the Internet, equating to roughly a trillion 'words', equivalent to a million copies of *War and Peace* by Leo Tolstoy. Secondly, there has been a significant increase in computational power, thanks to the continuous miniaturisation of transistors and the development of powerful and energy-efficient processors, along with the emergence of GPUs and TPUs ideal for training deep neural networks. These advancements have allowed for the training of increasingly larger models. Lastly, the development of novel architectures, alongside the accumulation of intricate expertise in the training of these models, has significantly enhanced their performance and accessibility. The spread of open-source culture and the establishment of competitive benchmarks like ImageNet [Deng et al., 2009] have also played an important role. Also note that the complexity of the engineering processes involved in model training, such as data preprocessing, parallel computing, infrastrucutre mangagement, hyperparameter tuning and model monitoring, underscores the requirement of specialised skills to train large models. This fact is reflected in the generous salaries associated with such skills [Eckert, 2023].

1.2 A brief history of deep learning

Understanding the role of theory in this success story is a natural question. Interestingly, it appears as a somewhat controversial topic as pioneers in deep learning, like Yann LeCun, have critiqued what they call a *'mathematical hypnosis'* [Lecun, 2019] and LeCun writes that *'blind trust in theoretical results that turned out to be irrelevant is a major reason why neural nets were dismissed between 1995 and 2010'* [Lecun, 2023].

In the 1990s, convolutional neural networks were successfully trained with the recent use of backpropagation along with stochastic gradient descent, outperforming other techniques in tasks such as handwritten character recognition [LeCun et al., 1998]. These promising empirical results, however, lacked solid theoretical foundations and interpretability. Worse, they seemed to contradict prevailing mathematical beliefs: classical generalisation bounds predicted catastrophic overfitting, and deep networks were seen as excessively complex, given the proven approximation universality of single hidden layer networks [Hornik et al., 1989]. Additionally, their training relied on many heuristics, and in Vapnik's own words: *'the designers of neural networks compensate the mathematical shortcomings with the high art of engineering'* [Vapnik, 1999, p.171].

This perspective led Vladimir Vapnik, a pioneer in statistical learning theory, to take the following bet in 1995: *'by 2005, no one in his right mind will use neural nets that are essentially like those used in 1995'* [Lecun, 2019]. In contrast to the advancement of neural networks, Vapnik and colleagues pursued a 'bottom-up approach',[1]constructing the Support Vector Machines (SVM), which are *designed* to provide guaranteed generalisation performances [Boser et al., 1992, Cortes and Vapnik, 1995]. While theoretically appealing, SVMs struggled to scale efficiently with the increasing volume of available data.

The 'SVM vs Deep Neural Networks' debate was (at least temporaly) settled in 2012 with the appartition of the AlexNet architecture [Krizhevsky et al., 2012], which made a significant breakthrough by largely outperforming other models in the ImageNet Large Scale Visual Recognition Challenge. Nevertheless, the outstanding results achieved by deep neural networks still largely remain mysterious and Vapnik's original theoretical skepticism remains pertinent.

1.3 The necessity of a ('good') theory.

Given the achievements of deep learning, largely independent of theoretical foundations, one may rightfully question the necessity or desirability of a 'theory of deep learning'. However, considering the title of this PhD, it seems imperative to advocate for the importance of such a theory. To begin, we must try to establish what constitutes a 'good' theory.

'Nothing is more practical than a good theory.' [2]

Adopting this saying, a good theory must therefore be practical (similar to a pocket-map!) and provide a framework and tools which facilitates the understanding and conceptualisation of existing empirical observations. Why is this essential in the context of deep learning? We propose three main reasons:

[1]following Vapnik's philosophy, he chose the second part of Hegel's formula *'Whatever is real is rational, and whatever is rational is real.'* (see Chapter 9.6 in Vapnik [1999] for his enlightening discussion on the two different point of views a theoretician can adopt when facing a natural science phenomenon)

[2]commonly attributed to the social psychologist Kurt Lewin. Note that this saying is at the heart of Vapnik's book on statistical learning theory [Vapnik, 1999].

- **Sustainability:** A better understanding of models should lead to an optimisation of their structure, of their training and of the use of data. Consequently and hopefully leading to:[3] *(i)* reduce model sizes, *(ii)* minimise training data and storage requirements, *(iii)* decrease the number of required training iterations. These benefits, in turn, translate into lower energy consumption for training, data storage, and hardware manufacturing as well as reduced material extraction including rare metals, used in hardware production.

- **Interpretability:** Understanding the criteria behind a model's predictions and ensuring robustness guarantees is crucial for ethical, fairness, and security considerations. This is particularly critical in domains like healthcare, finance, education, and robotics, where trust and accountability are required.

- **A better understanding of intelligence:** Language and image generation models challenge our conceptions of intelligence. Machines now navigate territories once reserved to conscious beings. A better understanding of these seemingly intelligent machines will undoubtedly shed light on various aspects of our own.

1.4 What are the deep learning mysteries?

To understand why the success of deep learning is so intriguing, we must first introduce its fundamental principles. Deep learning, in essence, follows a straightforward approach.

Deep learning approach:

Input

(Big) training set $(x_1, y_1), \dots, (x_n, y_n)$

(Large and deep) architecture, initialised at $f_{w_0} : \mathbb{R}^d \to \mathcal{Y}$

Training procedure

Minimise the empirical loss $\sum_i \ell(y_i, f_w(x_i))$ using a gradient method and heuristics

Output

Trained weights $w^\star$ and prediction function $x \mapsto f_{w^\star}(x)$

We expand on various aspects of the approach.

A diversity of architectures. There are various types of architectures, each corresponding to different parameterisations f_w. Examples include multilayer perceptrons (MLP), convolutional neural networks (CNN), and residual neural networks (ResNets). Despite their differences, they all involve stacking multiple layers, hence the term 'deep'. The final transformation involves a series of matrix operations and nonlinear transformations. For instance, the AlexNet architecture [Krizhevsky et al., 2012] is composed of a total of 8 trainable layers: 5 convolutional layers and 3 fully connected layers, with a total of 60 million parameters. The latest major architectural development is the Transformer architecture [Vaswani et al., 2017] which has had a profound impact on various learning tasks since 2017.

[3] this comment is only applicable if there is no rebound effect and these benefits are not immediately exploited to develop more powerful models.

Training procedure. Successfully training the model is crucial. At its core, this involves using a gradient method to minimise the training loss, on top of which a wide range of techniques has been developed through trial and error, forming a toolbox of common practices which have significantly enhanced the test performances. These include batch normalisation, dropout, adaptive learning rates, the use of momentum, data cleaning, data augmentation, weight decay, early stopping etc. Years of trial and error have resulted in an accumulation of these engineering tricks which have led to the current state-of-the-art results.

Deep learning philosophy. There is a notable inclination towards the development of increasingly larger models, coupled with the use of much bigger datasets and the allocation of substantial computational resources for model training. This inclination results from the observation that over the years, this philosophy has been consistently demonstrated to yield significant advancements in performance. For instance, contemporary models like GPT have approximately 10^{11} trainable parameters, whereas comparatively 'older' models like ResNet18 have around 10^6 parameters.

1.4.1 What could possibly go wrong?

In short, many things! In fact, we've become accustomed with their success to the point that it is tempting to forget all the things which *a priori* could have gone wrong. As written in Vapnik [1999] (Chapter 5.11): *'From the formal point of view one cannot guarantee that neural networks generalize well, since according to theory, in order to control generalisation ability one should control two factors: the value of the empirical risk and the value of the confidence interval. Neural networks, however, cannot control either two.'* We expand on the two main concerns highlighted by Vapnik, which still largely hold today:

- **Optimisation:** The objective function which is minimised is non-convex and (potentially) has many local minima which have high training loss values in which the gradient method could get trapped. Furthermore, the training procedure could (potentially) exhibit significant instabilities such as exploding or vanishing gradients which would make the search for appropriate hyperparameters prohibitively complicated.

- **Generalisation:** Even if the training process does converge to a global minimum[4], there is no inherent assurance that the resulting solution will generalise well due to the absence of theoretical results ensuring this. The traditional learning theory framework, originally designed to offer such guarantees, fails to furnish meaningful bounds in the context of deep neural networks.

These two concerns are fully justified. And yet, deep learning *works*! To understand why, we must go beyond classical non-convex and generalisation results:

- **Loss landscape analysis:** Relying solely on general (and hence worst-case) results concerning non-convex functions cannot be sufficient. Instead, we must take a deeper look at the specific loss landscape: are there many local minima, and if so, why doesn't the gradient method converge to them? How does this depend on the hyperparameters?

- **Implicit regularisation:** Regarding the generalisation concern, the concept of *algorithmic implicit regularisation* has emerged: if overfitting is benign, it must be because the

[4]note that some of the comments may not apply in the context of Large Language Models (LLM). The training sets have become so large that overfitting them is impossible. In this thesis, we focus on the 'pre'-LLM paradigm, where overfitting is common.

optimisation process is implicitly biased towards solutions which have good generalisation properties for the considered real-world prediction tasks.

We expand the discussion by examining several influential empirical studies which demonstrate the necessity of rethinking traditional approaches.

1.4.2 Theoretical insights from empirical observations

Several empirical studies, which conduct *controlled experiments*, offer precious insights into the types of theoretical guarantees we can or cannot expect. Such works have significantly influenced recent theoretical research.

Neural networks can memorise random pixels. In their seminal work, Zhang et al. [2017] show that a convolutional neural network trained on images consisting entirely of random pixels can perfectly fit the data with zero training error. The same phenomenon occurs with real world images which have random labels. *This indicates that the structure of the inputs and outputs is not the key factor explaining successful optimisation.* Furthermore, this implies that 'easy optimisation' and 'good generalisation' are two separate phenomena.

Larger models lead to better test performance. In another seminal work, Neyshabur et al. [2014] observe that for a simple single hidden layer network, increasing the size of the hidden layer consistently reduces the test loss, even after reaching the width at which training interpolation begins. However, traditional measures of model complexity, such as the VC-dimension or the Rademacher complexity, *increase* with the width. *This suggests that these complexity measures are not adequate for explaining generalisation performance.*

Data dependent generalisation bounds. Zhang et al. [2017] observe that a convolutional neural network can overfit an image dataset where the labels are fully random, resulting in a training loss of 0 and poor test accuracy (by construction due to the random labeling). *This indicates that any uniform bound explaining good generalisation properties cannot hold in the case of random labels.* Such bounds must therefore take into account the specific characteristics of the dataset.

The training algorithm must be taken into account. Liu et al. [2020a] show that there exist bad global minima of the training loss, i.e. there exists neural networks that perfectly fit the training set but exhibit poor generalisation performance. Therefore, when good generalisation occurs, it must be because the standard training algorithm has led us towards a solution that enjoys favorable generalisation properties for the task at hand.

Training hyperparameters influence test performance. While it's commonly expected that training hyperparameters affect the convergence speed of the training loss, recent observations suggest they also impact the generalisation performance of the trained model. Notably, the *initialisation scale* of the weights plays an important role in generalisation [Woodworth et al., 2020b], with large initialisations leading to what is known as a 'lazy regime', characterised by decreased test performance [Chizat et al., 2019, Fort et al., 2020]. Empirical findings by Masters and Luschi [2018] reveal that increasing the *step size* improves test performance for mini-batch stochastic gradient descent (SGD), while the opposite trend occurs for large-batch SGD. Moreover, Geiping et al. [2022] note a performance gap between GD and SGD in favor of SGD, and similar trends are corroborated in the study by Keskar et al. [2017]. Additionally, the use of

momentum in deep learning training is a common practice known to enhance generalisation performance [Sutskever et al., 2013, Leclerc and Madry, 2020].

Explicit regularisation alone does not fully account for generalisation performance. While ℓ_2-regularisation, often referred to as weight decay, is a common technique used to enhance network training and testing performance, it does not appear as fundamental as non-regularised networks still perform very well as observed in Zhang et al. [2017].

We can now finally discuss the approach we employ to try and clarify some of these aspects.

1.5 Goal, outline and contibutions of the thesis

The objective of this thesis is to try to elucidate some of the empirical findings mentioned above: particularly regarding the global convergence of gradient methods and their implicit regularisation.

To do so, we will extensively utilise the simplest neural network model available: the 2-layer diagonal linear network (DLN). Despite its simplicity, this model surprisingly exhibits training characteristics mentioned previously, including global convergence and non-trivial implicit regularisation dependent on factors such as initialisation, step size, batch size, and the presence of momentum. Consequently, this model serves as an ideal proxy for gaining a deeper understanding of complex phenomenons.

The thesis is outlined as follows.

Part I: In Chapter 2, we give a gentle introduction to the concept of implicit regularisation by illustrating its manifestation even for linear parametrisations. Following this, in Chapter 3, we introduce the 2-layer diagonal linear network architecture and present key results upon which the rest of the thesis is built. Lastly, in Chapter 4, we present general convergence results on mirror flows, which are extensively leveraged when dealing with diagonal linear networks. It's important to note that the results presented in these chapters are primarily synthesised from existing works *and not novel contributions*: we acknowledge the works which establish these results in Chapter 5.

Part II: The aim of this section is to extend beyond the asymptotic characterisation of the solution recovered by gradient flow (GF) over the 2-layer diagonal linear network in the regression setting. We provide a full description of the trajectory in the limit of vanishing initialisation. We show that the limiting flow successively jumps from a saddle of the training loss to another until reaching the minimum ℓ_1-norm solution. Starting from the zero vector, coordinates are successively activated until the minimum ℓ_1-norm solution is recovered, revealing an incremental learning.

Part III: The objective of this section is to explore the influence of hyperparameters on the solution recovered by gradient methods over the 2-layer diagonal linear network. Chapter 7 investigates the impact of stochastic noise using a continuous-time model of stochastic gradient descent (SGD). We demonstrate that this model allows us to show the beneficial effects of stochastic sampling noise for sparse recovery. Chapter 8 extends this investigation by directly analysing the discrete recursion, without resorting to a continuous-time model. We prove that while large step sizes enhance sparse recovery for SGD, employing large step sizes in the gradient descent recursion is detrimental. Finally, Chapter 9, we leverage a continuous-time approach to

analyse momentum gradient descent. This approach identifies an intrinsic quantity $\lambda = \frac{\gamma}{(1-\beta)^2}$ that uniquely defines the optimisation path and offers a simple acceleration rule applicable beyond the diagonal network architecture. We then explain that mild values of λ facilitate the recovery of sparse solutions when training diagonal linear networks in the regression setting.

Part IV: Finally in the last part, we examine the continuous-time counterpart of mirror descent, namely mirror flow, on classification problems which are linearly separable. Such problems are minimised 'at infinity' and have many possible solutions; we study which solution is preferred by the algorithm depending on the mirror potential. For exponential tailed losses and under mild assumptions on the potential, we show that the iterates converge in direction towards a ϕ_∞-maximum margin classifier. The function ϕ_∞ is the *horizon function* of the mirror potential and characterises its shape 'at infinity'. When the potential is separable, a simple formula allows to compute this function. We analyse several examples of potentials and provide numerical experiments highlighting our results.

Contributions beyond this thesis. The chapters in this thesis are a selection of works which are related to diagonal linear networks or to implicit regularisation. The author also contributed to two other projects on other topics.

In Pesme and Flammarion [2020], we consider a regression setting with gaussian inputs and where the outputs have potentially been corrupted by an oblivious adversary. In this setting, we show that performing SGD on the ℓ_1 loss converges to the true parameter at a fast and adaptive rate.

In Pesme et al. [2020], we consider the underparametrised least-squares setting and explore two convergence-diagnostic methods to automatically decrease the stepsize when saturation of the loss is reached. We show that relying on the inner product between consecutive stochastic gradients cannot lead to an adequate convergence diagnostic. We then propose another simple strategy which is based on the distance travelled by the iterates.

Part I

Implicit regularisation, diagonal linear networks and mirror flows

Chapter 2

Warm-up with linear parametrisations

In order to set the stage and illustrate how algorithmic implicit regularisation can be exhibited, we begin by focusins on the simplest parametrisation: considering a weight vector $w \in \mathbb{R}^d$, we define the function $f_w(x) = \langle w, x \rangle$. To maintain consistency in notation throughout subsequent sections, we use the symbol 'β' instead of 'w' to parametrise this class of functions. Our 'neural network' architecture is thus:

$$\{f_\beta : x \mapsto \langle \beta, x \rangle, \beta \in \mathbb{R}^d\}.$$

Clearly this class of function is limited in its ability to learn or extract complex features from the data, as it can only perform linear combination of the input coordinates. However, it serves as a useful framework for exploring questions related to implicit regularisation.

In the following, we differentiate between the regression and classification settings, as these settings are relatively different.

2.1 Linear regression

We set ourselves in the *overparametrised* setting in which the number of trainable parameters (here equal to the dimension d) is larger than the number of samples. This setting is also sometimes referred to as the *under-determined* setting. In this case, there are multiple ways of interpolating our data as $y_i = \langle \beta^\star, x_i \rangle$. In order to rigourously ensure that this is indeed the case, we put the following non-restrictive assumption on the inputs.

Assumption 1 (Overparametrised setting)**.** *d is larger than n and the features $x_1, \dots, x_n \in \mathbb{R}^d$ are linearly independent.*

Provided that $d \geq n$, this assumption holds almost surely if the samples are drawn from a distribution which is absolutely continuous, i.e. which doesn't assign any probability to sets that have no size according to the Lebesgue measure. Importantly, Assumption 1 ensures that the span of the columns of the feature matrix[1] $X \in \mathbb{R}^{n \times d}$ is equal to $\mathbb{R}^n$. Consequently, letting $\mathcal{S}$ denote the set of *interpolators* which perfectly fit the training data:

$$\mathcal{S} \coloneqq \{\beta^\star \in \mathbb{R}^d, \langle \beta^\star, x_i \rangle = y_i \ \forall i \in [n]\},$$

we are guaranteed that $\mathcal{S}$ is not empty. In fact, $\mathcal{S}$ is an affine space of dimension $(d-n)$ equal to $\beta^\star + \text{span}(x_1, \dots, x_n)^\perp$, where $\beta^\star$ is any arbitrary element of $\mathcal{S}$. We will also use the terminology of *interpolator* to denote an element of $\mathcal{S}$, we also say of such a vector that it *fits* the dataset.

Consequences for learning. In this setting, it does not make sense to speak of **the** empirical risk minimiser since any element of $\mathcal{S}$ minimises it. One might then wonder if it makes sense to find a generalisation bound which holds *for any* interpolator in $\mathcal{S}$. We argue that it does not.

[1] the feature matrix $X \in \mathbb{R}^{n \times d}$ denotes the matrix whose i^{th} row is input x_i.

In fact, we expect 'most'[2] elements of $\mathcal{S}$ to perform extremely bad! Indeed, intuitively consider the set $\{\beta^\star \in \mathcal{S}, \|\beta^\star\|_2 \geq \text{'large constant'}\}$: we expect the infinitely many elements of this set to generalise catastrophically. This is formalised in the following proposition.

Proposition 1 (No free lunch)**.** *Assume that the samples $(x_i, y_i)_{1\leq i\leq n}$ are sampled from an underlying 'true' distribution which is such that the true expectations $\mathbb{E}[\|x\|_2^2]$ and $\mathbb{E}[y\|x\|_2]$ are finite and that $\mathbb{E}[xx^\top]$ is positive definite. Then, if $n < d$ it holds that:*

$$\sup_{\beta^\star\in\mathcal{S}} \mathbb{E}[(y - \langle\beta^\star, x\rangle)^2] = +\infty.$$

Proof. The proof follows from the fact that the (well-defined) true risk $\mathbb{E}[(y-\langle\beta, x\rangle)^2]$ is coercive and that $\mathcal{S}$ is unbounded. □

Remark. *If the data samples almost surely lie in a low dimensional subspace of dimension d_{eff}, then the previous proposition still holds but under the condition $n < d_{\text{eff}}$.*

In words, it is hopeless to expect all elements of $\mathcal{S}$ to generalise well. However, we can still hope that some of them will! In fact, the solutions recovered by the common training procedures usually perform pretty well. We thus seek to understand what solutions are recovered by specific training algorithms.

Taking the training algorithm into account. The method at the core of deep learning training is the gradient descent (GD) algorithm and all other training methods can be seen as more sophisticated variations of this method: stochastic gradient descent (SGD), the use of momentum, adaptive stepsizes etc. In the linear regression setting with a square residual penalty[3], the empirical train loss which we seek to minimise is the well-known quadratic loss:

$$L(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle\beta, x_i\rangle)^2. \qquad \text{(Quadratic loss)}$$

Importantly, notice that $L(\beta^\star) = 0$ if and only if $y_i = \langle\beta^\star, x_i\rangle$ for all $i \in [n]$. Since our discussion readily holds for mini-batch SGD with any batch-size (thus interpolating between SGD and GD). We directly state the mini-batch SGD recursion (which we abreviate as (S)GD):

$$\beta_{k+1} = \beta_k - \gamma_k \nabla L_{B_k}(\beta_k) \quad \text{where} \quad L_{B_k}(\beta) \coloneqq \frac{1}{2b}\sum_{i\in B_k}(y_i - \langle\beta, x_i\rangle)^2, \qquad \text{((S)GD)}$$

where γ_k corresponds to the stepsize at iteration k, $B_k \subset [n]$ the mini-batch of $b \in [n]$ distinct samples sampled uniformly and independently at each iteration. Classical SGD and full-batch GD are special cases with $b = 1$ and $b = n$, respectively. At first sight, we could expect the iterates β_k to explore the whole $\mathbb{R}^d$ space.[4] However, notice that for any sampled batch $B_k \subset [n]$, we have that $\nabla L_{B_k}(\beta) \in \text{span}(x_1, \dots, x_n)$. Summing the (S)GD we then immediately get the following key observation:

$$\beta_k \in \beta_0 + \text{span}(x_1, \dots, x_n). \qquad \text{(Key observation)}$$

This observation is crucial and it means that the iterates β_k remain in the n-dimensional affine space $\beta_0 + \text{span}(x_1, \dots, x_n)$ which is considerably smaller than $\mathbb{R}^d$ when $d \gg n$. This simple observation naturally leads to the implicit regularisation problem associated to (S)GD.

[2] the definition of 'most' is ill-defined as $\mathcal{S}$ is infinite.
[3] note that the implicit regularisation results still hold for a very large variety of other penalty losses $\ell(y, \hat{y})$
[4] this would be the case if we added random isotropic noise at each iteration

Proposition 2 (Implicit reguralisation of (S)GD in the regression setting)**.** *For any initialisation $\beta_0 \in \mathbb{R}^d$ and batch size $b \in [n]$ and for any stepsize sequence $(\gamma_k)_k$.*

- *if the loss converges towards 0, then the iterates converge towards the following interpolator:*

$$\beta^{(\mathrm{S})\mathrm{GD}} = \underset{\beta^\star \in \mathcal{S}}{\arg\min} \ \|\beta^\star - \beta_0\|_2^2.$$

- *if a strictly positive and constant $\gamma_k = \gamma \leq \sup_{B \subset [n]} \frac{1}{b} \sum_{i \in B} x_i x_i^\top$ is used, then $L(\beta_k) \overset{a.s.}{\to} 0$.*

Proof. Assume that the loss converges towards 0, then the convergence of the iterates follows from the fact that the restriction of L to $\beta_0 + \mathrm{span}(x_1, \dots, x_n)$ is strongly convex. Let $\beta^{(\mathrm{S})\mathrm{GD}}$ denote this interpolator, and notice that it must also belong to $\beta_0 + \mathrm{span}(x_1, \dots, x_n)$. For any other interpolator $\beta^\star \in \mathcal{S}$, the Pythagorean theorem concludes the proof:

$$\begin{aligned} \|\beta^\star - \beta_0\|^2 &= \|\beta^\star - \beta^{(\mathrm{S})\mathrm{GD}}\|_2^2 + \|\beta^{(\mathrm{S})\mathrm{GD}} - \beta_0\|_2^2 \\ &> \|\beta^{(\mathrm{S})\mathrm{GD}} - \beta_0\|_2^2. \end{aligned}$$

The proof of the almost sure convergence of the loss towards 0 under the stepize constraint can be found in Even [2024] (Proposition 1.1.4). □

In words, the recovered solution corresponds to the ℓ_2-projection of the initialisation on the set of interpolators. If the initialisation is chosen such that $\beta_0 = 0$, then the recovered solution corresponds to the minimum ℓ_2-norm interpolator, which is known to have favorable generalisation properties in various settings [Bartlett et al., 2020].

Remark. *Importantly, note that the recovered solution does not depends on the stepsize, nor on the batch size. GD and SGD, when they converge, always converge towards the same solution! It can easily be shown that this is also the case for momentum (stochastic) gradient descent when initialised such that $\beta_1 = \beta_0$.*

Remark. *Note that the hessian of the training loss at any interpolator $\beta^\star$ is the same. Hoping for generalisation bounds involving any type of flatness definition is therefore hopeless here.*

2.2 Linear classification

We now switch our focus towards linear classification. In this setting the picture is surprisingly quite different, and many of the methods which we used in the previous section cannot easily be transferred. Similarly to the regression setting, we are interested in the case where there exists different ways of classifying our dataset, we therefore put the following assumption our dataset.

Assumption 2 (Separable data)**.** *The dataset is linearly separable: there exists $\beta^\star \in \mathbb{R}^d$ such that $y_i \langle \beta^\star, x_i \rangle > 0$ for all $i \in [n]$.*

Similar to Assumption 1, the previous assumption holds a.s. as soon as $d > n$ if the samples are sampled from a continuous probability distribution. However note that Assumption 2 can hold even when $d \gg n$. Contrary to the regression setting, our prediction function is not $x \mapsto \langle \beta, x \rangle$ which takes its values in $\mathbb{R}$, but $x \mapsto \mathrm{sgn}(\langle \beta, x \rangle)$ which takes its values in $\{-1, 1\}$. Since this function is invariant up to positive rescalings of β. It is natural to consider the set following set of '*separating directions*':

$$\mathcal{S} = \{\bar{\beta}^\star \in \mathbb{R}^d, \|\bar{\beta}^\star\| = 1 \text{ and } y_i \langle \bar{\beta}^\star, x_i \rangle > 0, \forall i \in [n]\}.$$

The norm which appears in the definition of $\mathcal{S}$ is arbitrary and this set is non-empty due to Assumption 2, in fact it has infinitely many elements since $\bar{\beta} \mapsto \min_i y_i \langle \bar{\beta}, x_i \rangle$ is continuous on the $\|\cdot\|$ unit sphere. As in the regression setting, we cannot expect all of them to generalise well. Instead, we focus on the solutions recovered by gradient methods to investigate whether they enjoy a particular structure.

Taking the training algorithm into account. As in the regression setting, we consider mini-batch stochastic gradient descent Equation ((S)GD) but with a logistic loss[5] which is well suited to classification tasks:

$$L(\beta) = \sum_{i=1}^{n} \ln\left(1 + \exp(-y_i \langle \beta, x_i \rangle)\right). \qquad \text{(Logistic loss)}$$

Observe that due to Assumption 2, $\arg\min L$ is an empty set even though $\min L = 0$. This is due to the fact that L is minised 'at infinity'. We therefore expect de (S)GD iterates β_k to diverge to infinity, we can nonetheless consider the normalised iterates $\bar{\beta}_k := \frac{\beta_k}{\|\beta_k\|}$, where the norm is arbitrary, as done in the following proposition.

Proposition 3 (Implicit regularisation of (S)GD in the linear classification setting)**.** *For any initialisation β_0 and batch size $b \in [b]$ and sufficiently small constant step-size γ. The normalised iterates $\bar{\beta}_k := \frac{\beta_k}{\|\beta_k\|}$ converge towards a vector proportional to the ℓ_2-max margin solution:*

$$\bar{\beta}^{(\mathrm{S})\mathrm{GD}} \propto \underset{\min_i y_i \langle \bar{\beta}, x_i \rangle \geq 1}{\arg\min} \|\bar{\beta}\|_2,$$

where $\propto$ denotes positive proportionality.

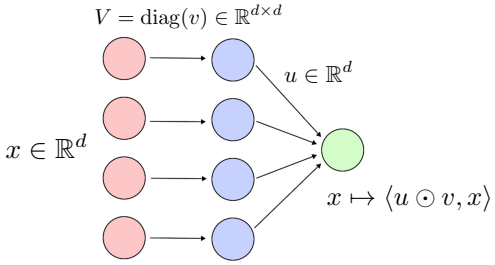
The proof of the previous proposition is not as intuitive as in the regression setting, but still *heavily relies on the fact that the updates remain in the span of the data.* We refer to Soudry et al. [2018], Nacson et al. [2019a] for the proofs for GD ans SGD respectively and to Part IV for a more compact proof in the case of gradient flow.

While we can already illustrate the concept of implicit regularisation through the linear parametrisation by showing that standard gradient descent methods converge towards minimal ℓ_2-norm solutions, it's important to note that the hyperparameters *do not* impact the recovered solution, *unlike what is observed in more complex architectures.* To theoretically explore this phenomenon further, we move on to a slightly more complex architecture.

[5] note that the results do not change for a large variety of other penalty losses which have exponential tails.

Chapter 3

Diagonal linear networks



Armed with our insights from the linear parametrisation and in an effort to understand the training dynamics of neural networks, we consider a 2-layer diagonal linear network (DLN) which corresponds to reparametrising the vector β as:

$$\beta_w = u \odot v \quad \text{where} \quad w = (u, v) \in \mathbb{R}^{2d}\,. \tag{DLN}$$

This parametrisation can be interpreted as a simple neural network $x \mapsto \langle u, \sigma(\mathrm{diag}(v)x)\rangle$ where u are the output weights, the diagonal matrix $\mathrm{diag}(v)$ represents the inner weights, and the activation σ is the identity function.[1]. We refer to $w = (u, v) \in \mathbb{R}^{2d}$ as the *weights* and to $\beta \coloneqq u \odot v \in \mathbb{R}^d$ as the *prediction parameter*.

This parametrisation may initially appear disappointing since the prediction function *remains linear in the input* x. However this simple multiplicative parametrisation leads to training behaviours which closely resemble those observe in more complex networks.

With the parametrisation (DLN), the loss function F over the weights $w = (u, v) \in \mathbb{R}^{2d}$ is defined as

$$F(w) \coloneqq L(u \odot v), \tag{3.1}$$

where L can either correspond to a regression or classification loss. In this chapter, we focus on the continuous-time *model* of gradient descent, namely gradient flow (GF):

$$\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t. \tag{GF}$$

Remark. *Continuous-time models enable to use calculus tools like differentiation and integration, making the computations much simpler. However, due to the (worst-case) catastrophic discretisation bounds, continuous time results cannot a priori be directly transferred to their discrete time counter-parts. However: (i) they turn out to be surprisingly good models in the sense that the countinuous and discrete trajectories often stay close, even for non-convex losses, and (ii) the developed mathematical tools and insights can often be adapted and transferred to the discrete world.*

[1] note that this parametrisation is strictly equivalent to another which also appears in the literature and which writes β as $(w_+^2 - w_-^2)/2$. This is because (w_+, w_-) can be seen as a 45° rotation of (u, v) (see Proposition 33)

3.1 Regression setting

In this section, we consider the regression setting where the data satisfies Assumption 1. Considering the sum of squared residuals, the training loss writes:

$$F(w) \coloneqq \frac{1}{2n}\sum_{i=1}^{n}(\langle u \odot v, x_i\rangle - y_i)^2. \tag{3.2}$$

Loss landscape. F is a non-convex function and this can be problematic from an optimisation perspective as we could end stuck at a local-minima which has a high training loss. However the following proposition shows the non-convexity of F is rather benign.

Proposition 4 (Benign non-convexity)**.** *The loss function F defined in eq. (3.2) is such that:*

- *It does not have any local maxima and its local minima must be global.*
- *If $w_\circlearrowleft = (u_\circlearrowleft, v_\circlearrowleft)$ is a saddle point of F, then $\beta_\circlearrowleft \coloneqq u_\circlearrowleft \odot v_\circlearrowleft$ satisfies:*

$$\beta_\circlearrowleft = \operatorname*{arg\,min}_{\beta_i=0 \text{ for } i\notin \mathrm{supp}(\beta_\circlearrowleft)} L(\beta),$$

  where $\mathrm{supp}(\beta_\circlearrowleft) = \{i \in [d], \beta_\circlearrowleft[i] \neq 0\}$ corresponds to the support of $\beta_\circlearrowleft$.

- *Conversely, for a subset $I_\circlearrowleft \subset [d]$, let $\beta_\circlearrowleft \in \operatorname*{arg\,min}_{\beta_i=0 \text{ for } i\notin I_\circlearrowleft} L(\beta)$. Then provided that $\beta_\circlearrowleft \notin \mathcal{S}$ and for $\lambda \in \mathbb{R}^d_{\neq 0}$, the following vector is a saddle point of F:*

$$w_\circlearrowleft \coloneqq (\mathrm{sgn}(\beta_\circlearrowleft)\sqrt{|\beta_\circlearrowleft|}/\lambda, \odot\sqrt{|\beta_\circlearrowleft|} \odot \lambda).$$

- *The global solutions of F corresponds to the set:*

$$\Big\{\big(\mathrm{sign}(\beta^\star)\sqrt{|\beta^\star|} \odot \lambda, \sqrt{|\beta^\star|}/\lambda\big) \text{ for } \beta^\star \in \mathcal{S} \text{ and } \lambda \in \mathbb{R}^d_{\neq 0}\Big\}.$$

Proof. First point. Notice that the image of the mapping $(u, v) \mapsto u \odot v$ of any neighbourhood of a point $(u_0, v_0) \in R^{2d}$ is a neighbourhood of $u_0 \odot v_0$ in $\mathbb{R}^d$. Consequently, if $(u_\circlearrowleft, v_\circlearrowleft)$ were a local extrema of F, then $u_\circlearrowleft \odot v_\circlearrowleft$ would be a local extrema of the convex loss L, which is absurd. The proof of the last point is straightforward. We refer to Appendix A.2 for the proof of points 2 and 3. □

It is very unlikely for any gradient method to get *exactly* stuck at a saddle point, however they can significantly slow down the optimisation due to the vanishing gradient.[2] Also note the very particular structure of the 'saddle' predictors $\beta_\circlearrowleft$: they correspond to *sparse* vectors that minimise the (convex) loss L over its non-zero coordinates. Also not that due to Assumption 1, for all subset $I \subset [d]$ *of size larger or equal than n*, there must exist a vector $\beta^\star \in \mathcal{S}$ such that $\mathrm{supp}(\beta^\star) \subset I$. Consequently, the saddle points $w_\circlearrowleft$ of F map to at most $\sum_{k=0}^{n-1}\binom{d}{k}$ different vectors $\beta_\circlearrowleft$.

[2] this also holds for the stochastic gradients because $\nabla F_i(\beta_\circlearrowleft)|_{\mathrm{supp}(\beta_\circlearrowleft)^\complement} = 0$ for any partial loss F_i, $i \in [n]$.

Gradient flow dynamics. We now turn to the gradient flow dynamics and seek to: *(i)* prove that the iterates indeed converge towards an interpolator, *(ii)* characterise the recovered solution. The chain rule applied to the training loss immediately gives that the gradient flow writes:

$$\dot{u}_t = -\nabla L(\beta_t) \odot v_t \quad \text{and} \quad \dot{v}_t = -\nabla L(\beta_t) \odot u_t.$$

Since our prediction function $x \mapsto \langle u_t \odot v_t, x \rangle$ at time t only depends on $\beta_t := u_t \odot v_t$, we are interested in characterising towards which vector do the predictors β_t converge to. An intermediate natural question to is then:

What type of flow do the predictors $(\beta_t)_{t \geq 0}$ follow?

An underlying mirror flow. Taking the derivative of β_t we immediately get that:

$$\mathrm{d}\beta_t = -(u_t^2 + v_t^2) \odot \nabla L(\beta_t)\mathrm{d}t. \tag{3.3}$$

It may seem that we are stuck as it is impossible to write $u^2 + v^2$ as only depending on $u \odot v$. However, the invariances of the mapping $(u, v) \mapsto u \odot v$ provides us a quantity which is conserved along the flow [Marcotte et al., 2024], namely the (absolute) balancedness $\Delta_t := \frac{1}{2}|u_t^2 - v_t^2|$. Indeed simply notice that $\dot{\Delta}_t = 0$ and therefore Δ_t is conserved and equal to $\Delta := |u_0^2 - v_0^2|/2$ which we assume to have non-zero coordinates.[3] Equation 3.3 then rewrites:

$$\mathrm{d}\beta_t = -2\sqrt{\beta_t^2 + \Delta^2} \odot \nabla L(\beta_t)\mathrm{d}t. \tag{3.4}$$

A sharp eye then recognises a mirror flow. Indeed, let $\phi_\Delta : \mathbb{R}^d \to \mathbb{R}$ denote the following Δ-hypentropy function:

$$\phi_\Delta(\beta) = \frac{1}{2}\sum_{i=1}^{d} \left(\beta_i \operatorname{arcsinh}\left(\frac{\beta_i}{\Delta_i}\right) - \sqrt{\beta_i^2 + \Delta_i^2} + \Delta_i \right). \tag{Δ-Hypentropy}$$

This strictly convex function is such that $\nabla^2 \phi_\Delta(\beta)$ is a diagonal matrix which has the vector $\frac{1}{2\sqrt{\beta^2+\Delta^2}}$ as its diagonal and equation 3.4 then naturally rewrites:

$$\mathrm{d}\nabla\phi_\Delta(\beta_t) = -\nabla L(\beta_t)\mathrm{d}t. \tag{MF}$$

This corresponds to the continuous version of mirror descent, namely *mirror flow*, and where the potential is the function ϕ_Δ.

What have we gained? It might seem that we have complicated the situation due to the appearance of the potential. *Que nenni*: we have moved from a *non-convex* gradient flow to a *convex* mirror flow and we can now leverage classical convex optimisation tools from the literature to prove the convergence of the iterates. Moreover, we now identify the new key observation which highlights which quantity remains in the span of the data. Indeed, integrating Equation (MF) we get:

$$\nabla\phi_\Delta(\beta_t) \in \nabla\phi_\Delta(\beta_0) + \operatorname{span}(x_1, \ldots, x_n). \tag{Key observation}$$

This geometrical observation leads to the following implicit regularisation problem.

[3] If initially $u_{i,0} = \pm v_{i,0}$ for some coordinate $i \in [d]$, then $u_{i,t} = \pm v_{i,t}$, $\forall t \geq 0$. Hence, imposing $u_0^2 - v_0^2 \neq 0$ corresponds to working with $2d$ distinct weights.

Proposition 5 (Implicit regularisation of GF over DLNs, regression setting)). *The iterates β_t converge towards an interpolator $\beta^{\mathrm{GF}} \in \mathcal{S}$ and:*

$$\beta^{\mathrm{GF}} = \operatorname*{arg\,min}_{\beta^\star \in \mathcal{S}} D_{\phi_\Delta}(\beta^\star, \beta_0),$$

where D_{ϕ_Δ} denotes the Bregman divergence.

Proof. The convergence of the iterates towards an interpolator β^{GF} can readily be adapted from Bauschke et al. [2017] and we refer to Chapter 4. For the implicit regularisation, we use the key observation along with the fact that for any different interpolator $\beta^\star \in \mathcal{S}$, we have that $\beta^\star \in \beta^{\mathrm{GF}} + \operatorname{span}(x_1, \dots, x_n)^\perp$. The (Bregman) Pythagorean theorem concludes the proof:

$$\begin{aligned} D_{\phi_\Delta}(\beta^\star, \beta_0) &= D_{\phi_\Delta}(\beta^\star, \beta^{\mathrm{GF}}) + D_{\phi_\Delta}(\beta^{\mathrm{GF}}, \beta_0) + \langle \nabla\phi_\Delta(\beta^{\mathrm{GF}}) - \nabla\phi_\Delta(\beta_0), \beta^\star - \beta^{\mathrm{GF}} \rangle \\ &= D_{\phi_\Delta}(\beta^\star, \beta^{\mathrm{GF}}) + D_{\phi_\Delta}(\beta^{\mathrm{GF}}, \beta_0) \\ &> D_{\phi_\Delta}(\beta^{\mathrm{GF}}, \beta_0), \end{aligned}$$

where the strict inequality is due to the strict convexity of the potential. □

Note the resemblance with Proposition 2: the recovered solution is the Bregman projection of the initialisation β_0 onto the set of interpolators. However, note that the projection depends on the potential ϕ_Δ which itself depends on the initial balancedness Δ!

Role of the initialisation. Note that β^{GF} depends on the weight initialisation (u_0, v_0) through the initial predictor $\beta_0 = u_0 \odot v_0$ and initial balancedness $\Delta = |u_0^2 - v_0^2|/2$. There are therefore $\mathbb{R}^{2d}$ distinct parameters which impact the recovered solution. In order to simplify the discussion and highlight the key aspects, we reduce the dependency of β^{GF} to a single parameter $\alpha > 0$ by considering the following initialisation scheme:

$$u_0 = \sqrt{2}\alpha\mathbf{1} \quad \text{and} \quad v_0 = 0. \qquad \text{(Simplified initialisation)}$$

In which case $\beta_0 = 0$, $\Delta = \alpha^2\mathbf{1}$ and Proposition 5 simply becomes:

$$\beta_\alpha^{\mathrm{GF}} = \operatorname*{arg\,min}_{\beta^\star \in \mathcal{S}} \phi_\alpha(\beta^\star),$$

where we overload the notation $\phi_\alpha \coloneqq \phi_{\alpha^2\mathbf{1}}$ from the definition of the Δ-Hypentropy and index the recovered solution $\beta_\alpha^{\mathrm{GF}}$ with α to highlight the dependency. A simple asymptotic expansion of $\phi_\alpha(\beta)$ shows that for fixed β:

$$\phi_\alpha(\beta) \underset{\alpha\to 0}{\sim} \ln(1/\alpha) \cdot \|\beta\|_1 \quad \text{and} \quad \phi_\alpha(\beta) \underset{\alpha\to\infty}{\sim} \frac{1}{4\alpha^2}\|\beta\|_2^2. \tag{3.5}$$

In words, the level lines of ϕ_α resemble to those of the ℓ_1-norm for small initialisatons and to those of the ℓ_2-norm for large initialisations (see Figure 3.1). This observation can be made rigourous and leads to the following proposition.

Proposition 6. *The recovered solutions $(\beta_\alpha^{\mathrm{GF}})_{\alpha>0}$ are bounded and:*

$$\beta_\alpha^{\mathrm{GF}} \underset{\alpha\to\infty}{\longrightarrow} \operatorname*{arg\,min}_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_2 \qquad \textit{and} \qquad \beta_\alpha^{\mathrm{GF}} \underset{\alpha\to 0}{\longrightarrow} \operatorname*{arg\,min}_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_1.$$

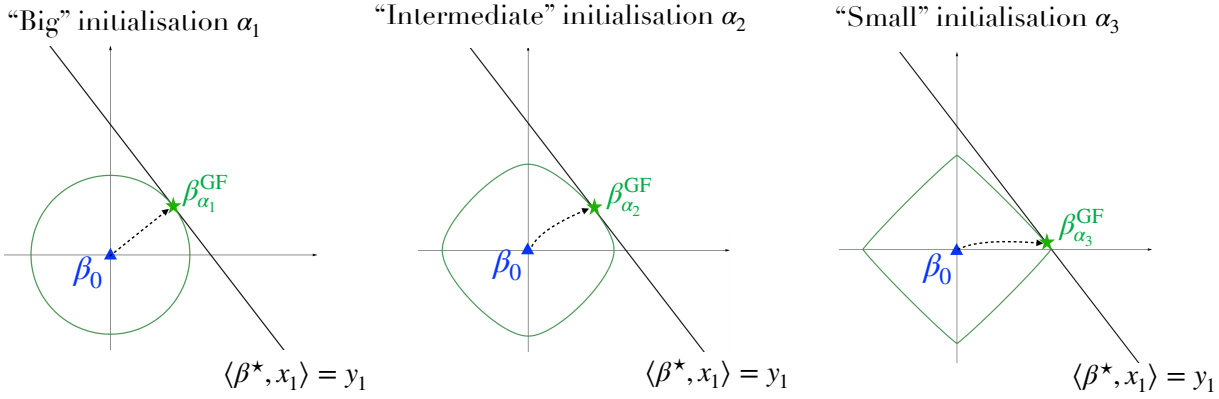


Figure 3.1: Simple illustration of Proposition 6 in 2d with a single sample x_1.

Note that the previous proposition is slightly sloppy on the notations and we clarify them: from the strict convexity of the ℓ_2 unit ball, $\arg\min_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_2$ is only composed of a single element and the first limit is well defined. However $\arg\min_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_1$ may not be a set composed of a unique element: in this case the limit in Proposition 6 must be understood in the sense that all limit points of $(\beta_\alpha^{\mathrm{GF}})_\alpha$ for $\alpha \to 0$ belong to the set. Also note that there there exists non-restrictive assumptions on the samples which guarantee the uniqueness of the $\arg\min$. For instance, the general position assumption [Dossal, 2012] holds a.s. for data sampled from a absolutely continuous distribution and ensures the uniqueness of the minimum ℓ_1-norm interpolator.

Proof of Proposition 6. **Boundedness.** Let $\beta_1^\star \in \mathcal{S}$ be an interpolator. We will show that $S_\leq^\alpha := \{\beta^\star \in S, \phi_\alpha(\beta^\star) \leq \phi_\alpha(\beta_1^\star)\}$ is bounded independently of $\alpha > 0$. The proof is then done since $\beta_\infty^\alpha = \arg\min_{\beta^\star \in \mathcal{S}_\leq^\alpha} \phi_\alpha(\beta^\star)$. For $z \in \mathbb{R}$, let $\varphi_\alpha(z) := \frac{1}{2}(z\operatorname{arcsinh}\left(\frac{z}{\alpha}\right) - \sqrt{z^2+\alpha^2} + \alpha)$ be the real function such that $\phi_\alpha(\beta) = \sum_i \varphi_\alpha(\beta_i)$. Let $\beta^\star \in S_\leq^\alpha$ and observe that

$$\varphi_\alpha(\|\beta^\star\|_\infty) = \max_i \varphi_\alpha(\beta_i^\star) \leq \phi_\alpha(\beta^\star) \leq \phi_\alpha(\beta_1^\star).$$

The first equality is because φ_α is an even and positive function. Now noticing that φ_α is a bijection over $\mathbb{R}_{\geq 0}$, we denote by $\varphi_\alpha^{-1} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ its inverse over $\mathbb{R}_{\geq 0}$, such that $\|\beta^\star\|_\infty \leq \varphi_\alpha^{-1}(\phi_\alpha(\beta_1^\star))$. It remains to show that $\varphi_\alpha^{-1}(\phi_\alpha(\beta_1^\star))$ is bounded independently of α. To show this, consider $h_\alpha := \varphi_\alpha^{-1} \circ \phi_\alpha$ and notice that

$$\nabla h_\alpha(\beta) = \left(\frac{\varphi_\alpha'(\beta_i)}{\varphi_\alpha'(\varphi_\alpha^{-1}(\phi_\alpha(\beta)))}\right)_{1 \leq i \leq d} \in [0,1]^d$$

Therefore h_α is Lipschitz with a Lipschitz constant independent of α. Consequently one can bound $h_\alpha(\beta_1^\star)$ independently of α, which concludes the proof.

Asymptotic convergence: $\alpha \to \infty$. For a fixed β we have that $\phi_\alpha(\beta) \underset{\alpha \to \infty}{\sim} \frac{1}{4\alpha^4}\|\beta\|_2^2$ and a simple function analysis shows that $4\alpha^2\phi_\alpha$ uniformly converges towards $\|\cdot\|_2^2$ on all compact subsets of $\mathbb{R}^d$. Since $(\beta_\alpha^{\mathrm{GF}})_{\alpha>0}$ is bounded, we can extract a converging subsequence: $\beta_{\alpha_k}^{\mathrm{GF}} \to \beta_\infty^{\mathrm{GF}}$ as $\alpha_k \to \infty$. Now let $\beta^\star \in \mathcal{S}$, we have that:

$$\left(4\alpha_k^2\phi_{\alpha_k}(\beta_{\alpha_k}^{\mathrm{GF}}) - \|\beta_{\alpha_k}^{\mathrm{GF}}\|_2^2\right) + \|\beta_{\alpha_k}^{\mathrm{GF}}\|_2^2 = 4\alpha_k^2\phi_{\alpha_k}(\beta_{\alpha_k}^{\mathrm{GF}}) \leq 4\alpha_k^2\phi_{\alpha_k}(\beta^\star)$$

The difference on the left hand side goes to zero by uniform convergence and we get that $\|\beta_\infty^{\mathrm{GF}}\|_2 \leq \|\beta^\star\|_2$. Therefore $\beta_\infty^{\mathrm{GF}} = \arg\min_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_2$, since this holds for any extraction, we can conclude the proof. The case $\alpha \to 0$ follows the exact same reasoning. □

3.2 Classification setting

We switch again to the classification setting and keep the discussion succinct as many of results and proof techniques resemble those of the regression setting. We consider a linearly separable dataset (Assumption 2) and we let F correspond to the 2-layer diagonal linear network classification training loss:

$$F(w) = L(u \odot v) = \sum_{i=1}^{n} \ln\big(1 + \exp(-y_i \langle u \odot v, x_i \rangle)\big), \tag{3.6}$$

where L here corresponds to the Logistic loss.

The following proposition is the equivalent of Proposition 7 and its proof follows the exact same lines as in the regression setting.

Proposition 7 (Benign non-convexity)**.** *The loss function F defined in eq. (3.6) is such that:*

- *It does not have any local maxima nor minima.*
- *If $w_\circlearrowleft = (u_\circlearrowleft, v_\circlearrowleft)$ is a saddle point of F, then $\beta_\circlearrowleft := u_\circlearrowleft \odot v_\circlearrowleft$ satisfies:*

$$\beta_\circlearrowleft \in \underset{\beta_i = 0 \text{ for } i \notin \mathrm{supp}(\beta_\circlearrowleft)}{\arg\min} L(\beta),$$

  where $\mathrm{supp}(\beta_\circlearrowleft) = \{i \in [d], \beta_\circlearrowleft[i] \neq 0\}$ corresponds to the support of $\beta_\circlearrowleft$.
- *Conversely, for a subset $I_\circlearrowleft \subset [d]$, let $\beta_\circlearrowleft$ be an element of $\underset{\beta_i = 0 \text{ for } i \notin I_\circlearrowleft}{\arg\min} L(\beta)$ provided this set is non-empty. Then for $\lambda \in \mathbb{R}^d_{\neq 0}$ the following vector is a saddle point of F:*

$$w_\circlearrowleft := (\mathrm{sgn}(\beta_\circlearrowleft)\sqrt{|\beta_\circlearrowleft|} \odot \lambda, \odot\sqrt{|\beta_\circlearrowleft|}/\lambda).$$

In words, due to the non-existence of local minima, we expect gradient methods to exhibit global convergence.

Gradient flow dynamics. From an initialisation (u_0, v_0), we consider the gradient flow over the loss F from eq. (3.6):

$$\dot{u}_t = -\nabla L(\beta_t) \odot v_t \quad \text{and} \quad \dot{v}_t = -\nabla L(\beta_t) \odot u_t.$$

Notice that from here, the computations are exactly the same as in the regression setting as they do not depend L. We therefore get that the iterates β_t follow a mirror flow on the Δ-Hypentropy potential ϕ_Δ:

$$\mathrm{d}\nabla\phi_\Delta(\beta_t) = -\nabla L(\beta_t)\mathrm{d}t.$$

However, since L is minised 'at infinity', we expect the mirror flow to diverge to infinity and we consider the normalised iterates $\bar{\beta}_t := \frac{\beta_t}{\|\beta_t\|}$.

Proposition 8 (Implicit Regularisation of GF over DLNs (Classification Setting))**.** *For any initialisation (u_0, v_0), the loss converges towards 0, the iterates β_t diverge to infinity and the normalised iterates $\bar{\beta}_t := \frac{\beta_t}{\|\beta_t\|}$ converge towards a vector proportional to the ℓ_1-max margin:*

$$\bar{\beta}^{\mathrm{GF}} \propto \underset{\min_i y_i \langle \bar{\beta}, x_i \rangle \geq 1}{\arg\min} \|\bar{\beta}\|_1,$$

provided that the minimisation problem has a unique minimiser and where $\propto$ denotes positive proportionality.

The convergence of the loss towards 0 is a consequence of classical mirror descent techniques and we refer to Chapter 4. The implicit regularisation can be seen as a consequence of Theorem 4.4 from Lyu and Li [2020]. However their result covers a much more general framework and has nothing to do with mirror flows. The object of Part IV is precisely to leverage the mirror flow structure to prove this result. Similar to the regression setting, $\arg\min_{\min_i y_i\langle\bar{\beta},x_i\rangle\geq 1} \|\bar{\beta}\|_1$ is not necessarily be unique, however there exists non-restrictive assumptions on the dataset which ensure its uniqueness.

Remark. *Notice that contrary to the regression setting, the implicit regularisation in the classification setting does not depend on the initialisation of the weights. This is due to the fact that the iterates diverge and that the initialisation is asymptotically 'forgotten'. We refer to the work of Moroshko et al. [2020] for insights into the impact of the initialisation in finite-time considerations.*

Chapter 4

Standalone mirror flow results

The aim of this chapter is to collect and adapt results related to mirror-descent and apply them to gradient flow, with the intention of offering a straightforward picture of the existing 'mirror flow toolbox'.

Starting from $\beta_0 \in \mathbb{R}^d$, we consider the following differential equation:

$$\mathrm{d}\nabla\phi(\beta_t) = -\nabla L(\beta_t)\mathrm{d}t \tag{MF}$$

Note that the existence of a global solution of (MF) over $\mathbb{R}_{\geq 0}$ is *a priori* not obvious. To ensure this, we must put several assumptions over the loss L as well as the potential ϕ.

Assumption 3. *The loss L is twice continuously differentiable, convex and bounded below.*

We additionally require the following assumptions on the potential.

Assumption 4. *The potential $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfies:*

1. *ϕ is twice continuously differentiable, strictly convex and coercive.*
2. *$\nabla\phi$ is coercive: $\lim_{\|\beta\|\to\infty} \|\nabla\phi(\beta)\| = +\infty$.*
3. *$\nabla^2\phi(\beta)$ is positive-definite for all $\beta \in \mathbb{R}^d$.*
4. *for every $c \in \mathbb{R}_{\geq 0}$ and $\beta_2 \in \mathbb{R}^d$, the sub-level set $\{\beta_1 \in \mathbb{R}^d, D_\phi(\beta_2, \beta_1) \leq c\}$ is bounded.*

These assumptions are common when considering mirror descent [Bauschke et al., 2017], except for the third point which is only required in the continuous time framework. Crucially, these assumptions ensure the existence and uniqueness of a solution to (MF).

Lemma 1. *Under Assumption 3 and 4, for any initialisation $\beta_0 \in \mathbb{R}^d$, there exists a unique global solution over $\mathbb{R}_{\geq 0}$ to (MF) such that $\beta_{t=0} = \beta_0$.*

Proof. From Assumption 4, we have that ϕ is differentiable, strictly convex and its gradient is coercive. Consequently, $\nabla\phi$ is bijective over $\mathbb{R}^d$ (see Rockafellar [1970], Theorem 26.6). Furthermore, the Fenchel conjugate ϕ^* is differentiable over $\mathbb{R}^d$ and $(\nabla\phi)^{-1} = \nabla\phi^*$.

To prove the existence and uniqueness of a global solution of (MF), we first consider the following differential equation:

$$\mathrm{d}u_t = -\nabla L(\nabla\phi^*(u_t))\mathrm{d}t, \tag{4.1}$$

with initial condition $u_{t=0} = \nabla\phi^*(\beta_0)$.

Since L is $\mathcal{C}^2$, ∇L is Lipschitz on all compact sets. Furthermore, since ϕ is $\mathcal{C}^2$ and $\nabla^2\phi$ is positive definite, $\nabla\phi^* = (\nabla\phi)^{-1}$ is $\mathcal{C}^1$ and therefore Lipschitz on all compact sets. Hence $\nabla L \circ \nabla\phi^*$ is Lipschitz on all compact sets and from the Picard-Lindelöf theorem, there exists a unique maximal (*i.e.* which cannot be extended) solution (u_t) satisfying eq. (E.1) and such

that $u_{t=0} = \nabla\phi^*(\beta_0)$. We denote $[0, T_{\max})$ the intersection of this maximal interval of definition (which must be open) and $\mathbb{R}_{\geq 0}$. Our goal is now to prove that $T_{\max} = +\infty$. To do so, we assume that $T_{\max}$ is finite and we will show that this leads to a contradiction due to the fact that the iterates β_t cannot diverge in finite time. Let $\beta_t \coloneqq \nabla\phi^*(u_t)$ and notice that β_t is therefore the unique solution satisfying (MF) over $[0, T_{\max})$ with $\beta_{t=0} = \beta_0$.

Bounding the trajectory of β_t over $[0, T_{\max})$. Pick any $\beta \in \mathbb{R}^d$ and notice that by convexity of L:

$$\frac{\mathrm{d}}{\mathrm{d}t} D_\phi(\beta, \beta_t) = -\langle \nabla L(\beta_t), \beta_t - \beta \rangle \leq -(L(\beta_t) - L(\beta)) \leq L(\beta) - L_{\min}.$$

Where $L_{\min}$ is a lower bound on the loss. Integrating from 0 to $t < T_{\max}$ we get:

$$\begin{aligned} D_\phi(\beta, \beta_t) &\leq t \cdot (L(\beta) - L_{\min}) + D_\phi(\beta, \beta_0) \\ &\leq T_{\max} \cdot (L(\beta) - L_{\min}) + D_\phi(\beta, \beta_0) \end{aligned}$$

Therefore, due to Assumption 4, the iterates β_t are bounded over $[0, T_{\max})$. The proof from here is standard (see e.g. Attouch et al. [2000], Theorem 3.1): from eq. (E.1) we get that $\dot{u}_t$ is bounded over $[0, T_{\max})$ and $\sup_{t\in[0,T_{\max})} \|\dot{u}_t\| =: C < +\infty$ which means that $\|u_t - u_{t'}\| \leq C|t - t'|$. Hence $\lim_{t\to T_{\max}} u_t =: u_\infty$ must exist. Applying the Picard-Lindelöf again at time $T_{\max}$ with initial condition u_∞ violates the initial maximal interval assumption. Therefore $T_{\max} = +\infty$ which concludes the proof. □

Now that we have ensured the existence and uniqueness of a solution, we can prove the convergence / divergence of the iterates and of the loss. To do so, we require additional assumptions on the potential ϕ, as in Bauschke et al. [2017]:[1]

Assumption 5. *The function $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfies the following assumptions:*

(i) if $(\beta_t)_{t\geq 0}$ converges to some β_∞, then $D_\phi(\beta_\infty, \beta_t) \to 0$.

(ii) Reciprocally, if $\beta_\infty \in \mathbb{R}^d$ and if $(\beta_t)_{t\geq 0}$ is such that $D_\phi(\beta_\infty, \beta_t) \to 0$, then $\beta_t \to \beta_\infty$.

We now state convergence results.

Proposition 9. *The losses converges to zero:*

$$L(\beta_t) \to \min_{\beta\in\mathbb{R}^d} L(\beta).$$

Concerning the behaviour of the iterates, we distinguish two scenarios:

- ***Losses with an attained minimum.*** *If $\arg\min_{\beta\in\mathbb{R}^d} L(\beta)$ is not empty, then the iterates β_t must converge towards an element of this set.*
- ***Losses with a minimum attained at infinity.*** *However if $\arg\min_{\beta\in\mathbb{R}^d} L(\beta)$ is empty, them the iterates β_t must diverge: $\|\beta_t\| \to \infty$ as $t \to \infty$.*

Proof. For the proof, we assume without loss over generality that $\min_\beta L = 0$.

[1] these additional assumptions are only required for losses which have an attained minimum

The loss decreases. Note that:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}L(\beta_t) &= -\langle \nabla L(\beta_t), \dot{\beta}_t\rangle \\ &= -\langle [\nabla^2\phi(\beta_t)]^{-1}\nabla L(\beta_t), \nabla L(\beta_t)\rangle \\ &\leq -\lambda_{\max}(\nabla^2\phi(\beta_s))^{-1}\|\nabla L(\beta_t)\|_2^2 \\ &\leq 0,\end{aligned}$$

where the inequality is by convexity of the potential ϕ. From here we distinguish the two scenarios.

1. Case where $\arg\min L$ is non-empty. Consider the Bregman divergence between an arbitrary minimiser $\beta^\star \in \arg\min L$ and β_t and notice that it is decreasing:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}D_\phi(\beta^\star, \beta_t) &= \langle \frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(\beta_t), \beta_t - \beta^\star\rangle \\ &= -\langle \nabla L(\beta_t), \beta_t - \beta^\star\rangle \\ &\leq -L(\beta_t) \qquad (4.2) \\ &\leq 0\end{aligned}$$

where the first inequality is by the convexity of the loss. Therefore, integrating inequality 4.2 and using that the loss is decreasing:

$$L(\beta_t) \leq \frac{1}{t}\int_0^t L(\beta_s)\,\mathrm{d}s \leq \frac{D_\phi(\beta^\star, \beta_0) - D_\phi(\beta^\star, \beta_t)}{t} \leq \frac{D_\phi(\beta^\star, \beta_0)}{t} \underset{t\to+\infty}{\longrightarrow} 0. \qquad (4.3)$$

The iterates converge towards an interpolator β_∞. Let $\beta^\star \in \arg\min L$. Since $\frac{\mathrm{d}}{\mathrm{d}t}D_\phi(\beta^\star, \beta_t) \leq 0$, we have that $D_\phi(\beta^\star, \beta_t)$ is decreasing and upper-bounded by $D_\phi(\beta^\star, \beta_0)$. Since it is a positive quantity, it must converge. Moreover, from Assumption 4, we have that $\beta \mapsto D_\phi(\beta^\star, \beta)$ has bounded sub-levels sets. Therefore, the iterates are bounded and we can extract a convergent subsequence: let β_∞ be such that $\beta_{t_k} \longrightarrow \beta_\infty$ as $k \to \infty$. Since $L(\beta_t) \to 0$ as $t \to \infty$, $L(\beta_{t_k})$ also converges to 0 as $k \to \infty$. By continuity of L, $L(\beta^\star) = 0$ and $\beta_\infty \in \arg\min L$. This means that *(a)* $D_\phi(\beta_\infty, \beta_t)$ converges and *(b)* it converges towards the same limit as $D_\phi(\beta_\infty, \beta_{t_k})$ which is 0 by Assumption 5. Finally, again from Assumption 5, β_t converges towards the solution β_∞ which concludes the proof.

2. Case where $\arg\min L$ is empty. Consider the Bregman divergence between an arbitrary point β and β_t and notice that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}D_\phi(\beta, \beta_t) &= \langle \frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(\beta_t), \beta_t - \beta\rangle \\ &= -\langle \nabla L(\beta_t), \beta_t - \beta\rangle \\ &\leq -(L(\beta_t) - L(\beta)) \qquad (4.4)\end{aligned}$$

where the inequality is by convexity of the loss. Integrating and due to the decrease of the loss, we get that:

$$L(\beta_t) \leq \frac{1}{t}\int_0^t L(\beta_s)\,\mathrm{d}s \leq L(\beta) + \frac{D_\phi(\beta, \beta_0) - D_\phi(\beta, \beta_t)}{t} \leq L(\beta) + \frac{D_\phi(\beta, \beta_0)}{t}$$

Since this is true for all point β, we get that

$$L(\beta_t) \leq \inf_{\beta \in \mathbb{R}^d} L(\beta) + \frac{D_\phi(\beta, \beta_0)}{t}. \tag{4.5}$$

It remains to show that the right hand term goes to 0 as t goes to infinity. To show this, for a sequence $0 < \varepsilon_k \to 0$, consider a sequence $\beta^{(k)}$ such that $L(\beta^{(k)}) \leq \varepsilon_k$ (it must exist since $\min L = 0$) and let $t_k \coloneqq \max(D_\phi(\beta^{(k)}, \beta_0)/\varepsilon_k, k)$. Notice that $t_k \to \infty$ and plugging $\beta^{(k)}$ in the r.h.s of eq. (4.5) we get that and $L(\beta_{t_k}) \leq 2\varepsilon_k \to 0$ as $k \to \infty$. Since the loss is decreasing we get that $L(\beta_t) \to 0$ as $t \to \infty$.

It remains to show that the iterates must diverge. Assume they did not: $\|\beta_t\| \not\to \infty$, then we could extract a convergent subsequence: $\beta_{t_k} \to \beta_\infty$ as $t_k \to \infty$. This then leads to $L(\beta_\infty) = 0$, which is absurd since $\arg\min L$ is empty.

□

To conclude the chapter, we provide several convergence rates of the loss. Their proofs are transparent from the proof of Proposition 9.

Proposition 10.

- ***Case where*** $\arg\min L$ ***is non-empty.*** *It holds that*

$$L(\beta_t) - L(\beta^\star) \leq \frac{D_\phi(\beta^\star, \beta_0)}{t},$$

where $\beta^\star$ any is element of $\arg\min L$. Furthermore, if the loss L satisfies a μ-PL inequality: $\frac{1}{2}\|\nabla L(\beta)\|_2^2 \geq \mu(L(\beta) - L(\beta^\star))$. Then we get the following expontential convergence:

$$\begin{aligned} L(\beta_t) - L(\beta^\star) &\leq L(\beta_0) \cdot \exp\Big(-2\mu \int_0^t \lambda_{\max}(\nabla^2 \phi(\beta_s))^{-1} \mathrm{d}s \Big) \\ &\leq L(\beta_0) \cdot \exp\Big(-\frac{2\mu}{\lambda} t \Big), \end{aligned}$$

where $\lambda \coloneqq \sup_{\beta \in \mathcal{B}} \lambda_{\max}(\nabla^2 \phi(\beta))$ with $\mathcal{B} \coloneqq \{\beta \in \mathbb{R}^d, D_\phi(\beta^\star, \beta) \leq D_\phi(\beta^\star, \beta_0)\}$.

- ***Case where*** $\arg\min L$ ***is empty.*** *In this case, it holds that:*

$$L(\beta_t) \leq \inf_{\beta \in \mathbb{R}^d} L(\beta) + \frac{D_\phi(\beta, \beta_0)}{t}.$$

Remark. *In the case where $\arg\min L$ is empty, the quantity $\inf_{\beta \in \mathbb{R}^d} L(\beta) + \frac{D_\phi(\beta, \beta_0)}{t}$ can be optimised for a given couple (L, ϕ). The obtained rate depends on how fast L decreases and phi grows in the directions of minimisation.*

Chapter 5

Related works

We conclude the introduction by giving an overview of the numerous related works that prove, underpin or are associated with the results presented in Part I.

Implicit regularisation. When the set of solutions is an affine space and the loss is convex, results describing the recovered solution can already be found in Lemaire [1996] (Corollary 2.2) for gradient flow and in Alvarez [2000] (Proposition 2.5) for the (continuous-time) heavy-ball method on quadratics. In the context of deep learning, the necessity of understanding algorithmic implicit regularisation was already put forward in Neyshabur et al. [2014]. Neyshabur et al. [2015] recalls that the (stochastic) gradient descent is inherently linked to the ℓ_2 geometry. In the case of matrix sensing with the $UU^\top$ multiplicative parametrisation, Gunasekar et al. [2017] proves that gradient flow with a vanishing initialisation recovers the minimum trace norm solution when the input matrices commute (which is equivalent to looking at the $\beta = u^2$ parametrisation in a regression setting). In the linear classification setting, the seminal paper Soudry et al. [2018] shows that gradient descent selects the ℓ_2-max-margin classifier amongst all classifiers. Gunasekar et al. [2018a] provides the implicit regularisation of mirror descent in the overparametrised linear regression setting. We refer to Vardi [2023] for a comprehensive overview of implicit regularisation results in deep learning.

Diagonal Linear Networks. *(Sparse recovery litterature)* Hoff [2017] is one of the first works which explicitly considers the Hadamard product parametrisation $\beta = u \odot v$ in order to promote sparsity. Results concerning the non-existence of local extrema can already be found there. Still in the context of promoting sparsity, Zhao et al. [2019] considers the same Hadamard parametrisation and shows that SGD converges towards the minimum ℓ_1-norm solution for vanishing initialisation and sufficiently small stepsizes. Results concerning the saddle points of the loss are also given in this same paper. Similar results were concurrently obtained by Vaskevicius et al. [2019]. In the same spirit of achieving top performance sparse recovery, Poon and Peyré [2021] leverages the same parametrisation for a wide range of sparse recovery tasks. *(Machine learning literature)* In parallel, the machine learning community started to leverage the same parametrisation as a toy model to gain insight into neural network optimisation. Gunasekar et al. [2017] considers the $\beta = u^2$ parametrisation as a simple case of matrix multiplication when the input matrices are co-diagonalisable. Woodworth et al. [2020b] coins the term of diagonal linear network for the parametrisation $w_+^2 - w_-^2$ which is equivalent to the $u \odot v$ parametrisation, and shows how the initialisation scale is the key parameter impacting the recovered solution for gradient flow. However, the underlying mirror flow structure is not *explicitly* exhibited. Slightly after, Gunasekar et al. [2021] makes this explicit link between gradient flow on diagonal linear networks and mirror flow with the hypentropy potential over the quadratic loss. In parallel, Amid and Warmuth [2020b] shows the equivalence of gradient flow over the u^2 parametrisation and mirror flow with the entropy potential. Still in parallel, Vaskevicius et al. [2020b] observed that gradient descent over the $u \odot v$ parametrisation closely resembles to the EG$\pm$ algorithm

(with no normalisation) [Kivinen and Warmuth, 1997], which was known to match mirror descent with the hypentropy potential since Ghai et al. [2020].

Mirror Descent. Mirror descent, also known as Bregman gradient descent or Bregman iteration was originally proposed by Nemirovski [1979], Nemirovski and Yudin [1983] for minimising convex functions. This method recently regained in popularity when the *'relative smoothness'* framework was brought to light: the idea is to choose a potential which is adapted to the curvature of the loss in order to get improved convergence rates [Bauschke et al., 2017]. All the convergence results on mirror flow we provide Chapter 4 are adapted from this last paper. We refer to Dragomir [2021] (Chapter 1.5) for a nice history of Bregman methods.

Part II

Full trajectory characterisation

Chapter 6

Saddle-to-saddle dynamics

6.1 Preface

This chapter follows Pesme and Flammarion [2023].

Summary We fully describe the trajectory of gradient flow over 2-layer diagonal linear networks for the regression setting in the limit of vanishing initialisation. We show that the limiting flow successively jumps from a saddle of the training loss to another until reaching the minimum ℓ_1-norm solution. We explicitly characterise the visited saddles as well as the jump times through a recursive algorithm reminiscent of the LARS algorithm used for computing the Lasso path. Starting from the zero vector, coordinates are successively activated until the minimum ℓ_1-norm solution is recovered, revealing an incremental learning. Our proof leverages a convenient arc-length time-reparametrisation which enables to keep track of the transitions between the jumps. Our analysis requires negligible assumptions on the data, applies to both under and overparametrised settings and covers complex cases where there is no monotonicity of the number of active coordinates. We provide numerical experiments to support our findings.

Co-author Nicolas Flammarion.

6.2 Introduction

Strikingly simple algorithms such as gradient descent are driving forces for deep learning and have led to remarkable empirical results. Nonetheless, understanding the performances of such methods remains a challenging and exciting mystery: (i) their global convergence on highly non-convex losses is far from being trivial and (ii) the fact that they lead to solutions which generalise well [Zhang et al., 2017] is still not fully understood.

To explain this second point, a major line of work has focused on the concept of implicit regularisation: amongst the infinite space of zero-loss solutions, the optimisation process must be implicitly biased towards solutions which have good generalisation properties for the considered real-world prediction tasks. Many papers have therefore shown that gradient methods have the fortunate property of asymptotically leading to solutions which have a well-behaving structure [Neyshabur, 2017, Gunasekar et al., 2017, Chizat and Bach, 2020].

Aside from these results which mostly focus on characterising the asymptotic solution, a slightly different point of view has been to try to describe the full trajectory. Indeed it has been experimentally observed that gradient methods with small initialisations have the property of learning models of increasing complexity across the training of neural networks [Kalimeris et al., 2019]. This behaviour is usually referred to as *incremental learning* or as a *saddle-to-saddle process* and describes learning curves which are piecewise constant: the training process makes very little progress for some time, followed by a sharp transition where a new "feature" is

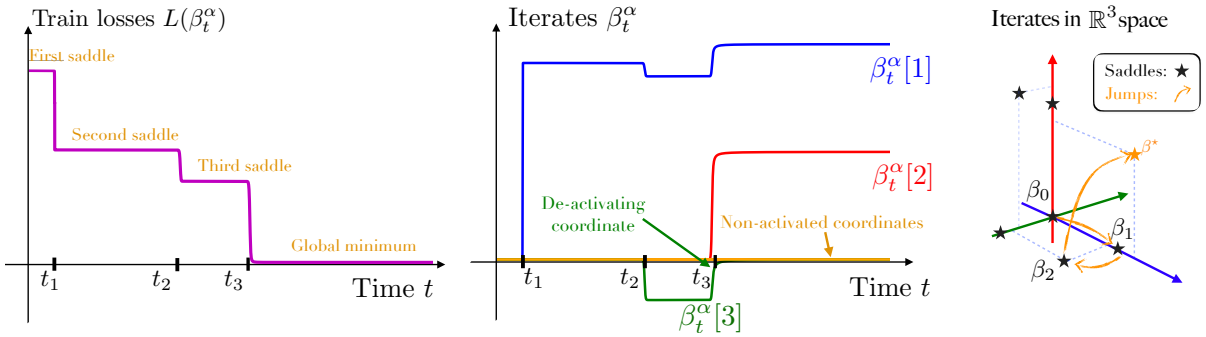


Figure 6.1: Gradient flow $(\beta_t^\alpha)_t$ with small initialisation scale α over a 2-layer diagonal linear network (for the precise experimental setting, see Appendix A.1). *Left:* Training loss across time, the learning is piecewise constant. *Middle:* The magnitudes of the coordinates are plotted across time: the process is piecewise constant. *Right:* In the $\mathbb{R}^3$ space in which the iterates evolve (the remaining coordinates stay at 0), the iterates jump from a saddle of the training loss to another. The jumping times t_i as well as the visited saddles β_i are entirely predicted by our theory.

suddenly learned. In terms of optimisation trajectory, this corresponds to the iterates "jumping" from a saddle of the training loss to another.

Several settings exhibiting such dynamics for small initialisation have been considered: matrix and tensor factorisation [Razin et al., 2021, Jiang et al., 2022], simplified versions of diagonal linear networks [Gissin et al., 2020, Berthier, 2022], linear networks [Gidel et al., 2019, Saxe et al., 2019, Jacot et al., 2021], 2-layer neural networks with orthogonal inputs [Boursier et al., 2022], learning leap functions with 2-layer neural networks [Abbe et al., 2023] and matrix sensing [Arora et al., 2019, Li et al., 2021, Jin et al., 2023]. However, all these results require restrictive assumptions on the data or only characterise the first jump. Obtaining a complete picture of the saddle-to-saddle process by describing all the visited saddles and jump times is mathematically challenging and still missing. We intend to fill this gap by considering diagonal linear networks which are simplified neural networks that have received significant attention lately [Woodworth et al., 2020b, Vaskevicius et al., 2019, HaoChen et al., 2021, Pesme et al., 2021, Even et al., 2023] as they are ideal proxy models for gaining a deeper understanding of complex phenomenons such as saddle-to-saddle dynamics.

6.2.1 Informal statement of the main result

In this chapter, we provide a full description of the trajectory of gradient flow over 2-layer diagonal linear networks in the limit of vanishing initialisation. The main result is informally presented here.

Theorem 1 (Main result, informal)**.** *In the regression setting and in the limit of vanishing initialisation, the trajectory of gradient flow over a 2-layer diagonal linear network converges towards a limiting process which is piecewise constant: the iterates successively jump from a saddle of the training loss to another, each visited saddle and jump time can recursively be computed through an algorithm (Algorithm 1) reminiscent of the LARS algorithm for the Lasso.*

The incremental learning stems from the particular structure of the saddles as they correspond to minimisers of the training loss with a constraint on the set of non-zero coordinates. The saddles therefore correspond to sparse vectors which partially fit the dataset. For simple datasets, a consequence of our main result is that **the limiting trajectory successively**

starts from the zero vector and successively learns the support of the sparse ground truth vector until reaching it. However, we make minimal assumptions on the data and our analysis also holds for complex datasets. In that case, the successive active sets are not necessarily increasing in size and coordinates can deactivate as well as activate until reaching the minimum ℓ_1-norm solution (see Figure 6.1 (middle) for an example of a deactivating coordinate). The regression setting and the diagonal network architecture are introduced in Section 8.3. Section 6.4 provides an intuitive construction of the limiting saddle-to-saddle dynamics and presents the algorithm that characterises it. Our main result regarding the convergence of the iterates towards this process is presented in Section 6.5 and further discussion is provided in Section 6.6.

6.3 Problem setup and leveraging the mirror structure

6.3.1 Setup

Linear regression. We study a linear regression problem with inputs $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ and outputs $(y_1, \dots, y_n) \in \mathbb{R}^n$. We consider the typical quadratic loss:

$$L(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (\langle \beta, x_i \rangle - y_i)^2 \,. \tag{6.1}$$

We make no assumption on the number of samples n nor the dimension d. The only assumption we make on the data throughout the chapter is that the inputs $(x_1, \dots, x_n)$ are in *general position*. In order to state this assumption, let $X \in \mathbb{R}^{n \times d}$ be the feature matrix whose i^{th} row is x_i and let $\tilde{x}_j \in \mathbb{R}^n$ be its j^{th} column for $j \in [d]$.

Assumption 6 (General position)**.** *For any $k \leq \min(n, d)$ and arbitrary signs $\sigma_1, \dots, \sigma_k \in \{-1, 1\}$, the affine span of any k points $\sigma_1 \tilde{x}_{j_1}, \dots, \sigma_k \tilde{x}_{j_k}$ does not contain any element of the set $\{\pm \tilde{x}_j, j \neq j_1, \dots, j_k\}$.*

This assumption is slightly technical but is standard in the Lasso literature [Tibshirani, 2013]. Note that it is not restrictive as it is almost surely satisfied when the data is drawn from a continuous probability distribution [Tibshirani, 2013, Lemma 4]. Letting $\mathcal{S} = \arg\min_\beta L(\beta)$ denote the affine space of solutions, Assumption 6 ensures that the minimisation problem $\min_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_1$ has a unique minimiser which we denote $\beta^\star_{\ell_1}$ and which corresponds to the minimum ℓ_1-norm solution.

2-layer diagonal linear network. In an effort to understand the training dynamics of neural networks, we consider a 2-layer diagonal linear network which corresponds to writing the regression vector β as

$$\beta_w = u \odot v \quad \text{where} \quad w = (u, v) \in \mathbb{R}^{2d} \,. \tag{6.2}$$

This parametrisation can be interpreted as a simple neural network $x \mapsto \langle u, \sigma(\mathrm{diag}(v)x) \rangle$ where u are the output weights, the diagonal matrix $\mathrm{diag}(v)$ represents the inner weights, and the activation σ is the identity function. We refer to $w = (u, v) \in \mathbb{R}^{2d}$ as the *weights* and to $\beta := u \odot v \in \mathbb{R}^d$ as the *prediction parameter*. With the parametrisation (8.1), the loss function F over the parameters $w = (u, v) \in \mathbb{R}^{2d}$ is defined as:

$$F(w) := L(u \odot v) = \frac{1}{2n} \sum_{i=1}^{n} (\langle u \odot v, x_i \rangle - y_i)^2 \,. \tag{6.3}$$

Though this reparametrisation is simple, the associated optimisation problem is non-convex and highly non-trivial training dynamics already occur. The critical points of the function F exhibit a very particular structure, as highlighted in the following proposition proven in Appendix A.2.

Proposition 11. *All the critical points w_c of F which are not global minima, i.e., $\nabla F(w_c) = \mathbf{0}$ and $F(w_c) > \min_w F(w)$, are necessarily saddle points (*i.e., *not local extrema). They map to parameters $\beta_c = u_c \odot v_c$ which satisfy $|\beta_c| \odot \nabla L(\beta_c) = \mathbf{0}$ and:*

$$\beta_c \in \underset{\beta[i]=0 \text{ for } i \notin \mathrm{supp}(\beta_c)}{\arg\min} L(\beta) \tag{6.4}$$

where $\mathrm{supp}(\beta_c) = \{i \in [d], \beta_c[i] \neq 0\}$ corresponds to the support of β_c.

The optimisation problem in eq. (6.4) states that the saddle points of the train loss F correspond to **sparse vectors that minimise the loss function L over its non-zero coordinates**. This property already shows that the saddle points possess interesting properties from a learning perspective. In the following we loosely use the term of 'saddle' to refer to points $\beta_c \in \mathbb{R}^d$ solution of eq. (6.4) **that are not saddles of the convex loss function L.** We adopt this terminology because they correspond to points $w_c \in \mathbb{R}^{2d}$ that are indeed saddles of the non-convex loss F.

Gradient Flow and necessity of "accelerating" time. We minimise the loss F using gradient flow:

$$\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t\,, \tag{6.5}$$

initialised at $u_0 = \sqrt{2}\alpha\mathbf{1} \in \mathbb{R}^d_{>0}$ with $\alpha > 0$, and $v_0 = \mathbf{0} \in \mathbb{R}^d$. This initialisation results in $\beta_0 = \mathbf{0} \in \mathbb{R}^d$ independently of the chosen weight initialisation scale α. We denote $\beta^\alpha_t := u^\alpha_t \odot v^\alpha_t$ the prediction iterates generated from the gradient flow to highlight its dependency on the initialisation scale α[1]. The origin $\mathbf{0} \in \mathbb{R}^{2d}$ is a critical point of the function F and taking the initialisation $\alpha \to 0$ therefore arbitrarily slows down the dynamics. In fact, it can be easily shown for any fixed time t, that $(u^\alpha_t, v^\alpha_t) \to \mathbf{0}$ as $\alpha \to 0$, indicating that the iterates are stuck at the origin. Therefore if we restrict ourselves to a finite time analysis, there is no hope of exhibiting the observed saddle-to-saddle behaviour. To do so, we must find an appropriate bijection $\tilde{t}_\alpha$ in $\mathbb{R}_{\geq 0}$ which "accelerates" time (*i.e.* $\tilde{t}_\alpha(t) \underset{\alpha\to 0}{\longrightarrow} +\infty$ for all t) and consider the accelerated iterates $\beta^\alpha_{\tilde{t}_\alpha(t)}$ which can escape the saddles. Finding this bijection becomes very natural once the mirror structure is unveiled.

6.3.2 Leveraging the mirror flow structure

While the iterates $(w^\alpha_t)_t$ follow a gradient flow on the non-convex loss F, it is shown in Azulay et al. [2021] that the iterates β^α_t follow a mirror flow on the convex loss L with potential ϕ_α and initialisation $\beta^\alpha_{t=0} = \mathbf{0}$:

$$\mathrm{d}\nabla\phi_\alpha(\beta^\alpha_t) = -\nabla L(\beta^\alpha_t)\mathrm{d}t, \tag{6.6}$$

where ϕ_α is the hyperbolic entropy function [Ghai et al., 2020] defined as:

$$\phi_\alpha(\beta) = \frac{1}{2}\sum_{i=1}^{d}\left(\beta_i \mathrm{arcsinh}(\frac{\beta_i}{\alpha^2}) - \sqrt{\beta_i^2 + \alpha^4} + \alpha^2\right). \tag{6.7}$$

[1] We point out that the trajectory of β^α_t exactly matches that of another common parametrisation $\beta_w := \frac{1}{2}(w_+^2 - w_-^2)$, with initialisation $w_{+,0} = w_{-,0} = \alpha\mathbf{1}$.

Unveiling the mirror flow structure enables to leverage convex optimisation tools to prove convergence of the iterates to a global minimiser $\beta^\star_\alpha$ as well as a simple proof of the implicit regularisation problem it solves. As shown by Woodworth et al. [2020b], in the overparametrised setting where $d > n$ and where there exists an infinite number of global minima, the limit $\beta^\star_\alpha$ is the solution of the problem:

$$\beta^\star_\alpha = \operatorname*{arg\,min}_{y_i = \langle x_i, \beta\rangle, \forall i} \phi_\alpha(\beta). \tag{6.8}$$

Furthermore, a simple function analysis shows that ϕ_α behaves as a rescaled ℓ_1-norm as α goes to 0, meaning that the recovered solution $\beta^\star_\alpha$ converges to the minimum ℓ_1-norm solution $\beta^\star_{\ell_1} = \arg\min_{y_i=\langle x_i,\beta\rangle} \|\beta\|_1$ as α goes to 0 (see Wind et al. [2023] for a precise rate). To bring to light the saddle-to-saddle dynamics which occurs as we take the initialisation to 0, we make substantial use of the nice mirror structure from eq. (6.6).

Appropriate time rescaling. To understand the limiting dynamics of β^α_t, it is natural to consider the limit $\alpha \to 0$ in eq. (6.6). However, the potential ϕ_α is such that $\phi_\alpha(\beta) \sim \ln(1/\alpha)\|\beta\|_1$ for small α and therefore degenerates as $\alpha \to 0$. Similarly, for $\beta \neq \mathbf{0}$, $\|\nabla\phi_\alpha(\beta)\| \to \infty$ as $\alpha \to 0$. The formulation from eq. (6.6) is thus not appropriate to take the limit $\alpha \to 0$. We can nonetheless obtain a meaningful limit by considering the opportune time acceleration $\tilde{t}_\alpha(t) = \ln(1/\alpha)\cdot t$ and looking at the accelerated iterates

$$\tilde{\beta}^\alpha_t \coloneqq \beta^\alpha_{\tilde{t}_\alpha(t)} = \beta^\alpha_{\ln(1/\alpha)t}. \tag{6.9}$$

Indeed, a simple chain rule leads to the "accelerated mirror flow": $\mathrm{d}\nabla\phi_\alpha(\tilde{\beta}^\alpha_t) = -\ln\left(\frac{1}{\alpha}\right)\nabla L(\tilde{\beta}^\alpha_t)\mathrm{d}t$. The accelerated iterates $(\tilde{\beta}^\alpha_t)_t$ follow a mirror descent with a rescaled potential:

$$\mathrm{d}\nabla\tilde{\phi}_\alpha(\tilde{\beta}^\alpha_t) = -\nabla L(\tilde{\beta}^\alpha_t)\mathrm{d}t, \qquad \text{where} \qquad \tilde{\phi}_\alpha \coloneqq \frac{1}{\ln(1/\alpha)}\cdot\phi_\alpha, \tag{6.10}$$

with $\tilde{\beta}_{t=0} = \mathbf{0}$ and where ϕ_α is defined eq. (6.7). Our choice of time acceleration ensures that the rescaled potential $\tilde{\phi}_\alpha$ is non-degenerate as the initialisation goes to 0 since $\tilde{\phi}_\alpha(\beta) \underset{\alpha\to 0}{\sim} \|\beta\|_1$.

6.4 Intuitive construction of the limiting flow and saddle-to-saddle algorithm

In this section, we aim to give a comprehensible construction of the limiting flow. We therefore choose to provide intuition over pure rigor, and defer the full and rigorous proof to the Appendix A.5. The technical crux of our analysis is to demonstrate the existence of a piecewise constant limiting process towards which the iterates $\tilde{\beta}^\alpha$ converge to. The convergence result is deferred to the following Section 6.5. **In this section we assume this convergence and refer to this piecewise constant limiting process as** $(\tilde{\beta}^\circ_t)_t$. Our goal is then to determine the jump times $(t_1, \dots, t_p)$ as well as the saddles $(\beta_0, \dots, \beta_p)$ which fully define this process.

To do so, it is natural to examine the limiting equation obtained when taking the limit $\alpha \to 0$ in eq. (6.10). We first turn to its integral form which writes:

$$-\int_0^t \nabla L(\tilde{\beta}^\alpha_s)\mathrm{d}s = \nabla\tilde{\phi}_\alpha(\tilde{\beta}^\alpha_t). \tag{6.11}$$

Provided the convergence of the flow $\tilde{\beta}^\alpha$ towards $\tilde{\beta}^\circ$, the left hand side of the previous equation converges to $-\int_0^t \nabla L(\tilde{\beta}^\circ_s)\mathrm{d}s$. For the right hand side, recall that $\tilde{\phi}_\alpha(\beta) \overset{\alpha\to 0}{\sim} \|\beta\|_1$, it is

therefore natural to expect the right hand side of eq. (6.11) to converge towards an element of $\partial\|\tilde{\beta}^\circ_t\|_1$, where we recall the definition of the subderivative of the ℓ_1-norm as:

$$\partial\|\tilde{\beta}\|_1 = \{1\} \text{ if } \tilde{\beta} > 0, \quad \{-1\} \text{ if } \tilde{\beta} < 0, \quad [-1,1] \text{ if } \tilde{\beta} = 0.$$

The arising key equation which must satisfy the limiting process $\tilde{\beta}^\circ$ is then, for all $t \geq 0$:

$$-\int_0^t \nabla L(\tilde{\beta}^\circ_s)\mathrm{d}s \in \partial\|\tilde{\beta}^\circ_t\|_1. \tag{6.12}$$

We show that **this equation uniquely determines the piecewise constant process** $\tilde{\beta}^\circ$ by imposing the number of jumps p, the jump times as well as the saddles which are visited between the jumps. Indeed the relation described in eq. (6.12) provides 4 restrictive properties that enable to construct $\tilde{\beta}^\circ$. To state them, let $s_t = -\int_0^t \nabla L(\tilde{\beta}^\circ_s)\mathrm{d}s$ and notice that it is continuous and piecewise linear since $\tilde{\beta}^\circ$ is piecewise constant. For each coordinate $i \in [d]$, it holds that:

(K1) $s_t[i] \in [-1,1]$ **(K2)** $s_t[i] = 1 \Rightarrow \tilde{\beta}^\circ_t[i] \geq 0$ and $s_t[i] = -1 \Rightarrow \tilde{\beta}^\circ_t[i] \leq 0$
(K3) $s_t[i] \in (-1,1) \Rightarrow \tilde{\beta}^\circ_t[i] = 0$ **(K4)** $\tilde{\beta}^\circ_t[i] > 0 \Rightarrow s_t[i] = 1$ and $\tilde{\beta}^\circ_t[i] < 0 \Rightarrow s_t[i] = -1$

To understand how these conditions lead to the algorithm which determines the jump times and the visited saddles, we present a 2-dimensional example for which we can walk through each step. The general case then naturally follows from this simple example.

6.4.1 Construction of the saddle-to-saddle algorithm with an illustrative $2d$ example.

Let us consider $n = d = 2$ and data matrix $X \in \mathbb{R}^{2\times 2}$ such that $X^\top X = ((1, 0.2), (0.2, -0.2))$. We consider $\beta^\star = (-0.2, 2) \in \mathbb{R}^2$ and outputs $y = X\beta^\star$. This setting is such that the loss L has $\beta^\star$ as its unique minimum and $L(\beta^*) = 0$. Furthermore the non-convex loss F has 3 saddles which map to: $\beta_{c,0} := (0,0) = \arg\min_{\beta_i=0,\forall i} L(\beta)$, $\beta_{c,1} := (0.2, 0) = \arg\min_{\beta[2]=0} L(\beta)$ and $\beta_{c,2} := (0, 1.6) = \arg\min_{\beta[1]=0} L(\beta)$. The loss function L is sketched in Figure 6.2 (*Left*). Notice that by the definition of $\beta_{c,1}$ and $\beta_{c,2}$, the gradients of the loss at these points are orthogonal to the axis they belong to. When running gradient flow with a small initialisation over our diagonal linear network, we obtain the plots illustrated Figure 6.2 (*Middle and Right*). We observe three jumps: the iterates jump from the saddle at the origin to $\beta_{c,1}$ at time t_1, then to $\beta_{c,2}$ at time t_2 and finally to the global minimum $\beta^\star$at time t_3.

Let us show how eq. (6.12) enables us to theoretically recover this trajectory. A simple observation which we will use several times below is that for any $t' > t$ such that $\tilde{\beta}^\circ$ is constant equal to β over the time interval (t, t'), the definition of s enables to write that $s_{t'} = s_t - (t' - t) \cdot \nabla L(\beta)$.

Zeroth saddle: The iterates are at the saddle at the origin: $\tilde{\beta}^\circ_t = \beta_0 := \beta_{c,0}$ and therefore $s_t = -t \cdot \nabla L(\beta_0)$. Our key equation eq. (6.12) is verified since $s_t = -t \cdot \nabla L(\beta_0) \in \partial\|\beta_0\|_1 = [-1,1]^d$. However the iterates cannot stay at the origin after time $t_1 := 1/\|\nabla L(\beta_0)\|_\infty$ which corresponds to the time at which the first coordinate of s_t hits $+1$: $s_{t_1}[1] = 1$. If the iterates stayed at the origin after t_1, (K1) for $i = 1$ would be violated. The iterates must hence jump.

First saddle: The iterates can only jump to a point different from the origin which maintains eq. (6.12) valid. We denote this point as β_1. Notice that:

- $s_{t_1}[2] = -t_1 \cdot \nabla L(\beta_0)[2] \in (-1, 1)$ and since s_t is continuous, we must have $\beta_1[2] = 0$ (K3).

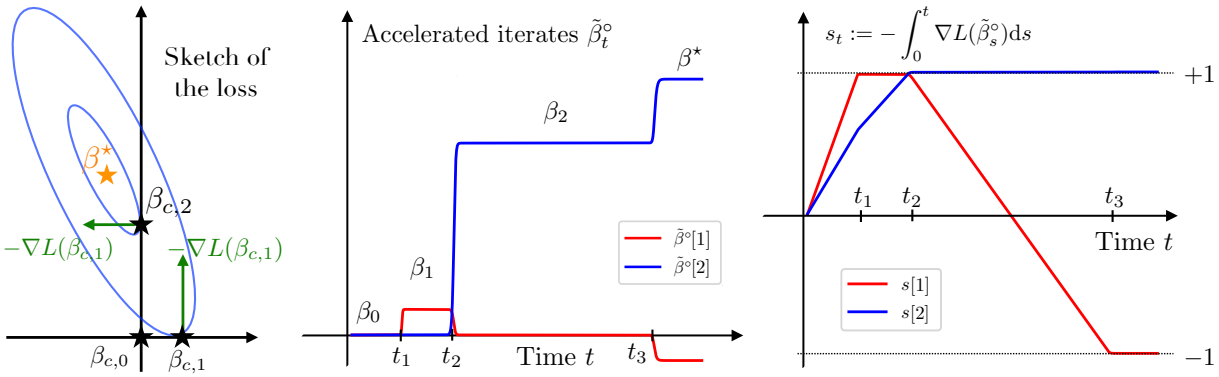


Figure 6.2: *Left*: Sketch of the $2d$ loss. *Middle and right*: Outputs of gradient flow with small initialisation scale: the iterates are piecewise constant and s_t is piecewise linear across time. We refer to the main text for further details.

- $s_{t_1}[1] = 1$ and hence for $t \geq t_1$, $s_t[1] = 1 - (t - t_1)\nabla L(\beta_1)[1]$. We cannot have $\nabla L(\beta_1)[1] < 0$ (K1), and neither $\nabla L(\beta_1)[1] > 0$ since otherwise $s_t[1] \in (-1, 1)$ and $\beta_1 = \mathbf{0}$ (K3).

The two conditions $\beta_1[2] = 0$ and $\nabla L(\beta_1)[1] = 0$ **uniquely defines** β_1 **as equal to** $\beta_{c,1}$. We now want to know if and when the iterates jump again. We saw that $s_t[1]$ remains at the value $+1$. However since β_1 is not a global minimum, $\nabla L(\beta_1)[2] \neq 0$ and $s_t[2]$ hits $+1$ at time t_2 defined such that $-(t_1 \nabla L(\beta_0) + (t_2 - t_1)\nabla L(\beta_1))[2] = 1$. The iterates must jump otherwise (K1) would break.

The iterates cannot jump to $\beta^\star$ yet! As the second coordinate of the iterates can activate, one could expect the iterates to be able to jump to the global minimum. However note that s_t is a continuous function and that s_{t_2} is equal to the vector $(1, 1)$. If the iterates jumped to the global minimum, then the first coordinate of the iterates would change sign from $+0.2$ to -0.2. Due to (K4) this would lead s_t jumping from $+1$ to -1, violating its continuity.

Second saddle: We denote as β_2 the point to which the iterates jump. s_{t_2} is now equal to the vector $(1, 1)$ and therefore *(i)* $\beta_2 \geq 0$ (coordinate-wise) from (K2 and K3) and the continuity of s. Since $s_t = s_{t_2} - (t - t_2)\nabla L(\beta_2)$, we must also have: *(ii)* $\nabla L(\beta_2) \geq 0$ from (K1) *(iii)* for $i \in \{1, 2\}$, if $\beta_2[i] \neq 0$ then $\nabla L(\beta_2)[i] = 0$ from (K4). The three conditions *(i)*, *(ii)* and *(iii)* precisely correspond to the optimality conditions of the following problem:

$$\underset{\beta[1]\geq 0, \beta[2]\geq 0}{\arg\min} \quad L(\beta).$$

The unique minimiser of this problem is $\beta_{c,2}$, hence $\beta_2 = \beta_{c,2}$, which means that the first coordinate deactivates. Similar to before, (K1) is valid until the time t_3 at which the first coordinate of $s_t = s_{t_2} - (t - t_2)\nabla L(\beta_2)$ reaches -1 due to the fact that $\nabla L(\beta_2)[1] > 0$.

Global minimum: We follow the exact same reasoning as for the second saddle. We now have s_{t_3} equal to the vector $(-1, 1)$ and the iterates must jump to a point β_3 such that *(i)* $\beta_3[1] \leq 0$, $\beta_3[2] \geq 0$ (K2 and K3), *(ii)* $\nabla L(\beta_3)[1] \leq 0$, $\nabla L(\beta_3)[2] \geq 0$ (K1), *(iii)* for $i \in \{1, 2\}$, if $\beta_3[i] \neq 0$ then $\nabla L(\beta_3)[i] = 0$ (K4). Again, these are the optimality conditions of the following problem:

$$\underset{\beta[1]\leq 0, \beta[2]\geq 0}{\arg\min} \quad L(\beta).$$

$\beta^\star$ is the unique minimiser of this problem and $\beta_3 = \beta^\star$. For $t \geq t_3$ we have $s_t = s_{t_3}$ and eq. (6.12) is satisfied for all following times: the iterates do not have to move anymore.

6.4.2 Presentation of the full saddle-to-saddle algorithm

We can now provide the full algorithm (Algorithm 1) which computes the jump times $(t_1, \dots, t_p)$ and saddles $(\beta_0 = \mathbf{0}, \beta_1, \dots, \beta_p)$ as the values and vectors such that the associated piecewise constant process satisfies eq. (6.12) for all t. This algorithm therefore defines our limiting process $\tilde{\beta}^\circ$.

Algorithm 1: Successive saddles and jump times of $\lim_{\alpha\to 0} \tilde{\beta}^\alpha$

Initialise: $(t, \beta, s) \leftarrow (0, \mathbf{0}, \mathbf{0})$;
while $\nabla L(\beta) \neq \mathbf{0}$ **do**

$\mathcal{A} \leftarrow \{j \in [d], \nabla L(\beta)(j) \neq 0\}$

$\Delta \leftarrow \inf\left\{\delta > 0 \text{ s.t. } \exists i \in \mathcal{A},\ s(i) - \delta \nabla L(\beta)(i) = \pm 1\right\}$

$(t,\ s) \leftarrow (t + \Delta,\ s - \Delta \cdot \nabla L(\beta))$

$\beta \leftarrow \arg\min\ L(\beta)$ where $\beta \in \left\{\beta \in \mathbb{R}^d \text{ s.t. } \begin{array}{l} \beta_i \geq 0 \text{ if } s(i) = +1 \\ \beta_i \leq 0 \text{ if } s(i) = -1 \\ \beta_i = 0 \text{ if } s(i) \in (-1, 1) \end{array}\right\}$

Output: Successive values of β and t

Algorithm 1 in words. The algorithm is a concise representation of the steps we followed in the previous section to construct $\tilde{\beta}^\circ$. We explain each step in words below. Starting from $k = 0$, assume we enter the loop number k at the saddle β_k computed in the previous loop:

- The set $\mathcal{A}_k$ contains the set of coordinates "which are unstable": by having a non-zero derivative, the loss could be decreased by moving along each one of these coordinates and one of these coordinates will have to activate.
- The time gap Δ_k corresponds to the time spent at the saddle β_k. It is computed as being the elapsed time just before (K1) breaks if the coordinates do not jump.
- We update $t_{k+1} = t_k + \Delta_k$ and $s_{k+1} = s_k - \Delta_k \nabla L(\beta_k)$: t_{k+1} corresponds to the time at which the iterates leave the saddle β_k and s_{k+1} constrains the signs of the next saddle β_{k+1}
- The solution β_{k+1} of the constrained minimisation problem is the saddle to which the flow jumps to at time t_{k+1}. The optimality conditions of this problem are such that eq. (6.12) is maintained for $t \geq t_{k+1}$.

Various comments on Algorithm 1. First we point out that any solution β_c of the constrained minimisation problem which appears in Algorithm 1 also satisfies

$$\beta_c = \underset{\beta[i]=0 \text{ for } i \notin \mathrm{supp}(\beta_c)}{\arg\min} L(\beta),$$

as in eq. (6.4): the algorithm hence indeed outputs saddles as expected. Up until now we have never checked whether the algorithm's constrained minimisation problem has a unique minimum. This is crucial otherwise the assignment step would be ill-defined. Showing the uniqueness is non-trivial and is guaranteed thanks to the general position Assumption 6 on the data (see Proposition 26 in Appendix A.4.1). In this same proposition, we also show that the algorithm terminates in at most $\min\left(2^d, \sum_{k=0}^n \binom{d}{k}\right)$ steps, that the loss strictly decreases at each step and that the final output β_p is the minimum ℓ_1-norm solution. These last two properties are expected

given the fact that the algorithm arises as being the limit process of $\tilde{\beta}^\alpha$ which follows the mirror flow eq. (6.10).

Links with the LARS algorithm for the Lasso. Recall that the Lasso problem [Tibshirani, 1996, Chen et al., 2001] is formulated as:

$$\beta^\star_\lambda = \arg\min_{\beta\in\mathbb{R}^d} \; L(\beta) + \lambda\|\beta\|_1. \tag{6.13}$$

The optimality condition of eq. (6.13) writes $-\nabla L(\beta^\star_\lambda) \in \lambda\partial\|\beta^\star_\lambda\|_1$. Now notice the similarity with eq. (6.12): the two would be equivalent with $\lambda = 1/t$ if the integration on the left hand side of eq. (6.12) did not average over the whole trajectory but only on the final iterate, in which case $-\int_0^t \nabla L(\tilde{\beta}^\circ_t)\mathrm{d}s = -t\cdot\nabla L(\tilde{\beta}^\circ_t)$. Though the difference is small, the trajectories of our limiting trajectory $\tilde{\beta}^\circ$ and the lasso path $(\beta^\star_\lambda)_\lambda$ are quite different: one has jumps, whereas the other is continuous. Nonetheless, the construction of Algorithm 1 shares many similarities with that of the Least Angle Regression (LARS) algorithm [Efron et al., 2004] (originally named the Homotopy algorithm [Osborne et al., 2000]) which is used to compute the Lasso path. A notable difference however is the fact that each step of our algorithm depends on the whole trajectory through the vector s, whereas the LARS algorithm can be started from any point on the path.

6.4.3 Outputs of the algorithm under a RIP and gap assumption on the data.

Unlike previous results on incremental learning, complex behaviours can occur when the feature matrix is ill designed: several coordinates can activate and deactivate at the same time (see Appendix A.1 for various cases). However, if the feature matrix satisfies the $2r$-restricted isometry property (RIP) [Candès et al., 2006] and there exists an r-sparse solution $\beta^\star$, the visited saddles can be easily approximated using Algorithm 1. We provide the precise characterisation below.

Sparse regression with RIP and gap assumption. ***(RIP)*** Assume that there exists an r-sparse vector $\beta^\star$ such that $y_i = \langle x_i, \beta^\star\rangle$. Furthermore we assume that the feature matrix $X \in \mathbb{R}^{n,d}$ satisfies the $2r$-restricted isometry property with constant $\tilde{\varepsilon} < \sqrt{2}-1 < 1/2$: i.e. for all submatrix X_s where we extract any $s \leq 2r$ columns of X, the matrix $X_s^\top X_s/n$ of size $s\times s$ has all its eigenvalues in the interval $[1-\tilde{\varepsilon}, 1+\tilde{\varepsilon}]$. ***(Gap assumption)*** Furthermore we assume that the r-sparse vector $\beta^\star$ has coordinates which have a "sufficient gap'. W.l.o.g we write $\beta^\star = (\beta^\star_1, \dots, \beta^\star_r, 0, \dots, 0)$ with $|\beta^\star_1| \geq \cdots \geq |\beta^\star_r| > 0$ and we define $\lambda := \min_{i\in[r]}(|\beta^\star_i| - |\beta^\star_{i+1}|) \geq 0$ which corresponds to the smallest gap between the entries of $|\beta^\star|$. We assume that $5\tilde{\varepsilon}\|\beta^\star\|_2 < \lambda/2$ and we let $\varepsilon := 5\tilde{\varepsilon}$.

A classic result from compressed sensing (see Candes [2008, Theorem 1.2]) is that the $2r$-restricted isometry property with constant $\sqrt{2}-1$ ensures that the minimum ℓ_0-minimisation problem has a unique r-sparse solution which is $\beta^\star$. This means that Algorithm 1 will have $\beta^\star$ as final output and the following proposition shows that we can precisely characterise each of its outputs when the data satisfies the previous assumptions.

Proposition 12. *Under the restricted isometry property and the gap assumption stated right above, Algorithm 1 terminates in r-loops and outputs:*

$$\begin{aligned}
&\beta_1 = (\beta_1[1], 0, \dots, 0) && \text{with} \quad \beta_1[1] \in \big[\beta^\star_1 - \varepsilon\|\beta^\star\|, \beta^\star_2 + \varepsilon\|\beta^\star\|\big]\\
&\beta_2 = (\beta_2[1], \beta_2[2], 0, \dots, 0) && \text{with} \quad \begin{cases} \beta_2[1] \in \big[\beta^\star_1 - \varepsilon\|\beta^\star\|, \beta^\star_1 + \varepsilon\|\beta^\star\|\big] \\ \beta_2[2] \in [\beta^\star_2 - \varepsilon\|\beta^\star\|, \beta^\star_2 + \varepsilon\|\beta^\star\|] \end{cases}\\
&\vdots\\
&\beta_{r-1} = (\beta_{r-1}[1], \dots, \beta_{r-1}[r-1], 0, \dots, 0) && \text{with} \quad \beta_{r-1}[i] \in \big[\beta^\star_i - \varepsilon\|\beta^\star\|, \beta^\star_i + \varepsilon\|\beta^\star\|\big]\\
&\beta_r = \beta^\star = (\beta^\star_1, \dots, \beta^\star_r, 0, \dots, 0),
\end{aligned}$$

at times $t_1, \dots, t_r$ *such that* $t_i \in \left[\frac{1}{|\beta_i^\star| + \varepsilon\|\beta^\star\|}, \frac{1}{|\beta_i^\star| - \varepsilon\|\beta^\star\|}\right]$ *and where* $\|\cdot\|$ *denotes the* ℓ_2 *norm.*

Informally, this means that the algorithm terminates in exactly r loops and outputs jump times and saddles roughly equal to $t_i = 1/|\beta_i^\star|$ and $\beta_i = (\beta_1^\star, \cdots, \beta_i^\star, 0, \dots, 0)$. Therefore, in simple settings, the support of the sparse vector is learnt a coordinate at a time, without any deactivations. We refer to Appendix A.4.2 for the proof.

6.5 Convergence of the iterates towards the process defined by Algorithm 1

We are now fully equipped to state our main result which formalises the convergence of the accelerated iterates towards the limiting process $\tilde{\beta}^\circ$ which we built in the previous section.

Theorem 2. *Let the saddles* $(\beta_0 = \mathbf{0}, \beta_1, \dots, \beta_{p-1}, \beta_p = \beta_{\ell_1}^\star)$ *and jump times* $(t_0 = 0, t_1, \dots, t_p)$ *be the outputs of Algorithm 1 and let* $(\tilde{\beta}_t^\circ)_t$ *be the piecewise constant process defined as follows:*

$$\textit{(Saddles)} \qquad \tilde{\beta}_t^\circ = \beta_k \qquad \text{for } t \in (t_k, t_{k+1}) \text{ and } 0 \le k \le p, \quad t_{p+1} = +\infty.$$

The accelerated flow $(\tilde{\beta}_t^\alpha)_t$ *defined in eq. (6.9) uniformly converges towards the limiting process* $(\tilde{\beta}_t^\circ)_t$ *on any compact subset of* $\mathbb{R}_{\ge 0} \backslash \{t_1, \dots, t_p\}$.

Convergence result. We recall that from a technical point of view, showing the existence of a limiting process $\lim_{\alpha \to 0} \tilde{\beta}^\alpha$ is the toughest part. Theorem 1 provides this existence as well as the uniform convergence of the accelerated iterates towards $\tilde{\beta}^\circ$ over all closed intervals of $\mathbb{R}$ which do not contain the jump times. We highlight that this is the strongest type of convergence we could expect and a uniform convergence over all intervals of the form $[0, T]$ is impossible given that the limiting process $\tilde{\beta}^\circ$ is discontinuous. In Proposition 13, we give an even stronger result by showing a graph convergence of the iterates which takes into account the path followed between the jumps. We also point out that we can easily show the same type of convergence for the accelerated weights $\tilde{w}_t^\alpha := w_{\tilde{t}^\alpha(t)}^\alpha$. Indeed, using the bijective mapping which links the weights w_t and the predictors β_t (see Lemma 6 in Appendix A.3), we immediately get that the accelerated weights $(\tilde{u}^\alpha, \tilde{v}^\alpha)$ uniformly converge towards the limiting process $(\sqrt{|\tilde{\beta}^\circ|}, \mathrm{sign}(\tilde{\beta}^\circ)\sqrt{|\tilde{\beta}^\circ|})$ on any compact subset of $\mathbb{R}_{\ge 0} \backslash \{t_1, \dots, t_p\}$.

Estimates for the non-accelerated iterates β_t^α. We point out that our result provides no speed of convergence of $\tilde{\beta}^\alpha$ towards $\tilde{\beta}^\circ$. We believe that a non-asymptotic result is challenging and leave it as future work. Note that we experimentally notice that the convergence rate quickly degrades after each saddle. Nonetheless, we can still write for the non-accelerated iterates that $\beta_t^\alpha = \tilde{\beta}_{t/\ln(1/\alpha)}^\alpha \sim \tilde{\beta}_{t/\ln(1/\alpha)}^\circ$ as $\alpha \to 0$. Hence, for α small enough the iterates β_t^α are roughly equal to 0 until time $t_1 \cdot \ln(1/\alpha)$ and the minimum ℓ_1-norm interpolator is reached at time $t_p \cdot \ln(1/\alpha)$. **Such a precise estimate of the global convergence time is rather remarkable** and goes beyond classical Lyapunov analysises which only leads to $L(\beta_t^\alpha) \lesssim \ln(1/\alpha)/t$ (see Proposition 23 in Appendix A.3).

Natural extensions of our setting. More general initialisations can easily be dealt with. For instance, initialisations of the form $u_{t=0} = \alpha \mathbf{u_0} \in \mathbb{R}^d$ lead to the exact same result as it is shown in Woodworth et al. [2020b] (Discussion after Theorem 1) that the associated mirror still converges to the ℓ_1-norm. Initialisations of the form $[u_{t=0}]_i = \alpha^{k_i}$, where $k_i > 0$, lead to the associated potential converging towards a weighted ℓ_1-norm and one should modify Algorithm 1

by accordingly weighting $\nabla L(\beta)$ in the algorithm. Also, deeper linear architectures of the form $\beta_w = w_+^D - w_-^D$ as in Woodworth et al. [2020b] do not change our result as the associated mirror still converges towards the ℓ_1-norm. Though we only consider the square loss in the chapter, we believe that all our results should hold for any loss of the type $L(\beta) = \sum_{i=1}^n \ell(y_i, \langle x_i, \beta \rangle)$ where for all $y \in \mathbb{R}$, $\ell(y, \cdot)$ is strictly convex with a unique minimiser at y. In fact, the only property which cannot directly be adapted from our results is showing the uniform boundedness of the iterates (see discussion before Proposition 24 in Appendix A.3).

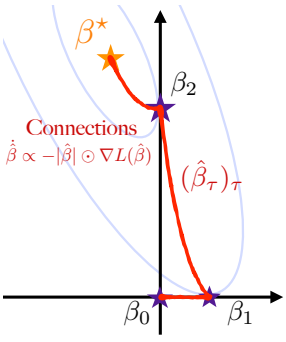
6.5.1 High level sketch of proof of $\tilde{\beta}^\alpha \to \tilde{\beta}^\circ$ which leverages an arc-length parametrisation

In this section, we give the high level ideas concerning the proof of the convergence $\tilde{\beta}^\alpha \to \tilde{\beta}^\circ$ given in Theorem 1. A full and detailed proof can be found in Appendix A.5. The main difficulty stems from the non-continuity of the limit process $\tilde{\beta}^\circ$. To circumvent this difficulty, a clever trick which we borrow to Efendiev and Mielke [2006], Mielke et al. [2009] is to "slow-down" time when the jumps occur by considering **an arc-length parametrisation of the path**. We consider the $\mathbb{R}_{\geq 0}$ arclength bijection τ^α and leverage it to define the 'appropriately slowed down' iterates $\hat{\beta}_\tau^\alpha$ as:

$$\hat{\beta}_\tau^\alpha = \tilde{\beta}^\alpha_{\hat{t}^\alpha(\tau)} \qquad \text{where} \qquad \hat{t}_\tau^\alpha = (\tau^\alpha)^{-1}(\tau) \quad \text{and} \quad \tau^\alpha(t) = t + \int_0^t \|\dot{\tilde{\beta}}_s^\alpha\| \mathrm{d}s.$$

This time reparametrisation has the fortunate but crucial property of leading to $\dot{\hat{t}}^\alpha(\tau) + \|\dot{\hat{\beta}}_\tau^\alpha\| = 1$ by a simple chain rule, which means that the speed of $(\hat{\beta}_\tau^\alpha)_\tau$ **is uniformly upperbounded by** 1 **independently of** α. This behaviour is in stark contrast with the process $(\tilde{\beta}_t^\alpha)_t$ which has a speed which explodes at the jumps. This change of time now allows us to use Arzelà-Ascoli's theorem to extract a subsequence which uniformly converges to a limiting process which we denote $\hat{\beta}$. Importantly, $\hat{\beta}$ enables to keep track of the path followed between the jumps as we show that its trajectory has two regimes:

$$\textbf{Saddles: } \hat{\beta}_\tau = \beta_k \qquad \textbf{Connections: } \dot{\hat{\beta}}_\tau = -\frac{|\hat{\beta}_\tau| \odot \nabla L(\hat{\beta}_\tau)}{\||\hat{\beta}_\tau| \odot \nabla L(\hat{\beta}_\tau)\|}.$$



The process $\hat{\beta}$ is illustrated on the right: the red curves correspond to the paths which the iterates follow during the jumps. These paths are called *heteroclinic orbits* in the dynamical systems literature [Krupa, 1997, Ashwin and Field, 1999]. To prove Theorem 1, we can map back the convergence of $\hat{\beta}^\alpha$ to show that of $\tilde{\beta}^\alpha$. Moreover from the convergence $\hat{\beta}^\alpha \to \hat{\beta}$ we get a more complete picture of the limiting dynamics of $\tilde{\beta}^\alpha$ as it naturally implies the convergence of the graph of the iterates $(\tilde{\beta}_t^\alpha)_t$ converges towards that of $(\hat{\beta}_\tau)_\tau$. The graph convergence result is formalised in this last proposition.

Proposition 13. *For all $T > t_p$, the graph of the iterates $(\tilde{\beta}_t^\alpha)_{t \leq T}$ converges to that of $(\hat{\beta}_\tau)_\tau$:*

$$\mathrm{dist}(\{\tilde{\beta}_t^\alpha\}_{t \leq T}, \{\hat{\beta}_\tau\}_{\tau \geq 0}) \underset{\alpha \to 0}{\longrightarrow} 0 \qquad \textit{(Hausdorff distance)}$$

6.6 Further discussion and conclusion

Link between incremental learning and saddle-to-saddle dynamics. The incremental learning phenomenon and the saddle-to-saddle process are often complementary facets of the same idea and refer to the same phenomenon. Indeed for gradient flows $\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t$, fixed points of the dynamics correspond to critical points of the loss. Stages with little progress in learning and minimal movement of the iterates necessarily correspond to the iterates being in the vicinity of a critical point of the loss. It turns out that in many settings (linear networks [Kawaguchi, 2016], matrix sensing [Bhojanapalli et al., 2016, Park et al., 2017]), critical points are necessarily saddle points of the loss (if not global minima) and that they have a very particular structure (high sparsity, low rank, etc.). We finally note that an alternative approach to realising saddle-to-saddle dynamics is through the perturbation of the gradient flow by a vanishing noise as studied in [Bakhtin, 2011].

Characterisation of the visited saddles. A common belief is that the saddle-to-saddle trajectory can be found by successively computing the direction of most negative curvature of the loss (i.e. the eigenvector corresponding to the most negative eigenvalue) and following this direction until reaching the next saddle [Jacot et al., 2021]. However this statement cannot be accurate as it is inconsistent with our algorithm in our setting. In fact, it can be shown that this algorithm would match the orthogonal matching pursuit (OMP) algorithm [Pati et al., 1993, Davis et al., 1997] which does not necessarily lead to the minimum ℓ_1-norm interpolator. In [Berthier, 2022], which is the closest to our work and the first to prove convergence of the iterates towards a piece-wise constant process, the successive saddles are entirely characterised and connected to the Lasso regularisation path in the underparameterised setting. Recently, Boix-Adsera et al. [2023] extended the diagonal linear network setting to diagonal parametrisations of the form $f_{u\odot v}$, but at the cost of stronger assumptions on the trajectory.

Adaptive Inverse Scale Space Method. Following the submission of the paper this chapter is based on, we were informed that Algorithm 1 had already been proposed and analysed in the compressed sensing literature. Indeed it exactly corresponds to the Adaptive Inverse Scale Space Method (aISS) proposed in Burger et al. [2013]. The motivations behind its study are extremely different from ours and originate from the study of Bregman iteration [Cai et al., 2010, Osher et al., 2005, Yin et al., 2008] which is an efficient method for solving ℓ_1 related minimisation problems. The so-called inverse scale space flow which corresponds to eq. (6.12) in our chapter can be seen as the continuous version of Bregman iteration. As in our chapter, Burger et al. [2013] show that this equation can be solved through an iterative algorithm. We refer to Yang et al. [2013, Section 2] for further details. However we did not find any results in this literature concerning the uniqueness of the constrained minimisation problem due to Assumption 6, nor on the maximum number of iterations, the behaviour under RIP assumptions and the maximum number of active coordinates.

Subdifferential equations and rate-independent systems. As in eq. (6.12), subdifferential inclusions of the form $\nabla L(\beta_t) \in \frac{\mathrm{d}}{\mathrm{d}t}\partial h(\beta_t)$ for non-differential functions h have been studied by Attouch et al. [2004] but for strongly convex functions h. In this case, the solutions are continuous and do not exhibit jumps. On another hand, Efendiev and Mielke [2006], Mielke et al. [2009, 2012] consider so-called *rate-independent systems* of the form $\partial_q E(t, q_t) \in \partial h(\dot{q}_t)$ for 1-homogeneous *dissipation* potentials h. Examples of such systems are ubiquitous in mechanics and appear in problems related to friction, crack propagation, elastoplasticity and ferromagnetism to name a few [Mielke, 2005, Ch. 6 for a survey]. As in our case, the main difficulty with such processes is the possible appearance of jumps when the energy E is non-convex.

6.7 Conclusion.

Our study examines the behaviour of gradient flow with vanishing initialisation over diagonal linear networks. We prove that it leads to the flow jumping from a saddle point of the loss to another. Our analysis characterises each visited saddle point as well as the jumping times through an algorithm which is reminiscent of the LARS method used in the Lasso framework. There are several avenues for further exploration. The most compelling one is the extension of these techniques to broader contexts for which the implicit bias of gradient flow has not yet fully been understood.

Part III

Effect of hyperparameters

Chapter 7

Effect of noise

7.1 Preface

This chapter follows Pesme et al. [2021].

Summary We study the dynamics of stochastic gradient descent over diagonal linear networks through its continuous time version, namely stochastic gradient flow. We explicitly characterise the solution chosen by the stochastic flow and prove that it always enjoys better generalisation properties than that of gradient flow. Quite surprisingly, we show that the convergence speed of the training loss controls the magnitude of the biasing effect: the slower the convergence, the better the bias. To fully complete our analysis, we provide convergence guarantees for the dynamics. We also give experimental results which support our theoretical claims. Our findings highlight the fact that structured noise can induce better generalisation and they help explain the greater performances of stochastic gradient descent over gradient descent observed in practice.

Co-authors Loucas Pillaud-Vivien and Nicolas Flammarion.

Contributions Scott and Loucas worked together on the project.

7.2 Introduction

Understanding the performance of neural networks is certainly one of the most thrilling challenges for the current machine learning community. From the theoretical point of view, progress has been made in several directions: we have a better functional analysis description of neural networks [Bach, 2017] and we steadily understand the convergence of training algorithms [Mei et al., 2018, Chizat and Bach, 2018] as well as the role of initialisation [Jacot et al., 2018, Chizat et al., 2019]. Yet there remain many unanswered questions. One of which is why do the currently used training algorithms converge to solutions which generalise well, and this with very little use of explicit regularisation [Zhang et al., 2017].

To understand this phenomenon, the concept of *implicit bias* has emerged: if over-fitting is benign, it must be because the optimisation procedure converges towards some particular global minimum which enjoys good generalisation properties. Though no explicit regularisation is added, the algorithm is implicitly selecting a particular solution: this is referred to as the implicit bias of the training procedure. The implicit regularisation of several algorithms has been studied, the simplest and most emblematic being that of gradient descent and stochastic gradient descent in the least-squares framework: they both converge towards the global solution which has the lowest squared distance from the initialisation. For logistic regression on separable data, Soudry et al. show in the seminal paper [Soudry et al., 2018] that gradient descent selects the max-margin classifier. This type of result has then been extended to neural networks and

to other frameworks. Overall, characterising the implicit bias of gradient methods has almost always come down to unveiling mirror-descent like structures which underlie the algorithms.

While mostly all of the results focus on gradient descent, it must be pointed out that this full batch algorithm is not used in practice for neural networks since it does not lead to solutions which generalise well [Keskar et al., 2017]. Instead, results on stochastic gradient descent, which is widely used and shows impressive results, are still missing or unsatisfactory. This has certainly to do with the fact that grasping the nature of the noise induced by the stochasticity of the algorithm is particularly hard: it mixes properties from the model's architecture, the data's distribution and the loss. In our work, by focusing on simplified neural networks, we answer to the following fundamental questions: do SGD's and GD's implicit bias differ? What is the role of SGD's noise over the algorithm's implicit bias?

The simplified neural networks which we consider are diagonal linear neural networks; despite their simplicity they have become popular since they already enable to grasp the complexity of more general networks. Indeed, they highlight important aspects of the theoretical concerns of modern machine learning: the neural tangent kernel regime, the roles of over-parametrisation, of the initialisation and of the step size. For a regression problem where we assume the existence of an interpolating solution, we study stochastic gradient descent through its continuous version, namely stochastic gradient flow (SGF). Though the continuous modelling of SGD has not yet led to many fruitful results compared to the well studied gradient flow, we believe it is because capturing the essence of the stochastic noise is particularly difficult. It has generally been done in a non realistic and over simplified manner, such as considering constant and isotropic noise. In our work, we attach peculiar attention to the adequate modelling of the noise. Tools from Itô calculus are then leveraged in order to derive exact formulas, quantitative bounds and interesting interpretations for our problem.

7.2.1 Main contributions and chapter organisation.

In Section 8.3, we start by introducing the setup of our problem as well as the continuous modelisation of stochastic gradient descent. Then, in Section 7.4, we state our main result on the implicit bias of the stochastic gradient flow. We informally formulate it here and illustrate it in Figure 7.1:

Theorem 3 (Informal)**.** *Stochastic gradient flow over diagonal linear networks converges with high probability to a zero-loss solution which enjoys better generalisation properties than the one obtained by gradient flow. Furthermore, the speed of convergence of the training loss controls the magnitude of the biasing effect: the slower the convergence, the better the bias.*

Unlike previous works [Gunasekar et al., 2018a, Woodworth et al., 2020b], in addition to characterising the implicit bias effect of SGF, we also prove the convergence of the iterates towards a zero-loss solution with high-probability. To accomplish this, we leverage in Section 7.5 the fact that the iterates follow a stochastic continuous mirror descent with a time-varying potential. We support our results experimentally and validate our model in Section 7.6.

7.2.2 Related work

As recalled, implicit bias has a recent history that has been initiated by the seminal work Soudry et al. [2018] on max-margin classification with log-loss for a linear setup and separable data. This work has been extended to other architectures, *e.g.* multiplicative parametrisations [Gunasekar et al., 2018a], linear networks [Ji and Telgarsky, 2019] and more general homogeneous neural networks [Lyu and Li, 2020, Chizat and Bach, 2020]. In Woodworth et al. [2020b] the authors show that the scale of the initialisation leads to an interpolation between the neural tangent

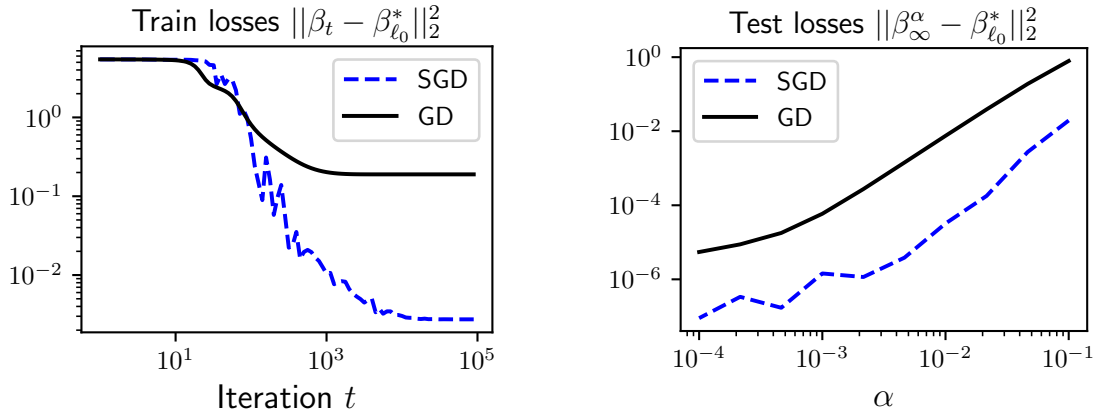


Figure 7.1: Sparse regression with $n = 40$, $d = 100$, $\|\beta^\star_{\ell_0}\|_0 = 5$, $x_i \sim \mathcal{N}(0, I)$ $y_i = x_i^\top \beta^\star_{\ell_0}$. *Left*: for initialisation scale $\alpha = 0.05$, SGD converges towards a solution which generalises better than GD. *Right*: for different values of the initialisation scale α, the solution recovered by SGD has better validation loss than that of GD. The sparsifying effect due to their implicit biases differ by more than an order of magnitude. See Section 7.6.1 for the precise experimental setup.

kernel regime [Jacot et al., 2018, Chizat et al., 2019] (which is a linear regression on fixed features) leading to ℓ_2 minimum norm solutions and the rich regimes leading to ℓ_1 minimum norm solutions. Note that these works focus on full batch gradient descent (or flow) and are deeply linked to mirror descent.

While the links between SGD's stochasticity and generalisation have been looked into in numerous works [Mandt et al., 2016, Jastrzebski et al., 2018, He et al., 2019, Hoffer et al., 2017, Kleinberg et al., 2018], no such explicit characterisation of implicit regularisation have ever been given. It has been empirically observed that SGD often outputs models which generalise better than GD [Keskar et al., 2017, Jastrzebski et al., 2018, He et al., 2019]. One suggested explanation is that SGD is prone to pick flatter solutions than GD and that bad generalisation solutions are correlated with sharp minima, i.e., with strong curvature, while good generalisation solutions are correlated with flat minima, i.e., with low curvature [Hochreiter and Schmidhuber, 1997, Keskar et al., 2017]. This idea has been further investigated by adopting a random walk on random landscape modelling [Hoffer et al., 2017], by suggesting that SGD's noise is smoothing the loss landscape, thus eliminating the sharp minima [Kleinberg et al., 2018], by considering a dynamical stability perspective [Wu et al., 2018] or by interpreting SGD as a diffusion process [He et al., 2019, Jastrzebski et al., 2018, Chaudhari and Soatto, 2018]. Recently, label-noise has been shown to influence the implicit bias of SGD, by biasing the solution towards the origin for quadratically-parameterized models [HaoChen et al., 2021] or by implicitly regularising the expected squared norm of the gradient of the model with respect to the weights [Blanc et al., 2020]. Thus, if the notion of implicit bias of GD is fairly well understood both in the cases of regression and classification, it remains unclear for SGD, and its explicit characterisation is missing.

The linear diagonal neural networks we consider have been studied in the case of gradient descent [Vaskevicius et al., 2019] and stochastic gradient descent with label noise [HaoChen et al., 2021]. In both cases the authors show that this model has the ability to implicitly bias the training procedure to help retrieve a sparse predictor. The link between gradient descent and mirror descent for this model has been initiated by Ghai et al. [2020] and further exploited by the same author in Wu and Rebeschini [2020], Vaskevicius et al. [2020a] for its sparse inducing property.

Contrary to the deterministic case, the modelling of stochastic gradient descent as a stochastic differential equation is quite recent, see Mandt et al. [2016], Jastrzebski et al. [2018]. However,

as highlighted by Ali et al. [2020], early attempts often suffer from the drawback that they model the noise using a constant covariance matrix. On the contrary, state dependant noise has now become the legitimate manner for modelling SGD as a stochastic gradient flow and it is shown in Li et al. [2019a] that it can be done consistently. Yet, noise modelling still remains the principal issue [Wojtowytsch, 2021] as it influences largely the behaviour of the dynamics Chaudhari and Soatto [2018], Cheng et al. [2020].

7.2.3 Notations

For input data $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ and output $(y_1, \dots, y_n) \in \mathbb{R}^n$, we denote respectively $X \in \mathbb{R}^{n \times d}$ the design matrix whose i-th row is feature $x_i \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$ the vector of outputs. $\mathbb{R}_+^*$ denotes the set of strictly positive real numbers. For $p = 1, 2$, the ℓ_p-norm of $x \in \mathbb{R}^d$ is $\|x\|_p^p = \sum_i^d |x_i|^p$. The operations $\odot$ will stand for coordinate-wise product between vector: $[u \odot v]_i = u_i v_i$ and $u^2 = u \odot u$. For $p \in \mathbb{N}^*$, we also define $u^p := u \odot \dots \odot u$, the p times product of u with itself. All inequalities between vectors should be understood value by value. For $f, g \in \mathbb{R}$, the existence of $C > 0$ such that $f \leq Cg$ and $Cg \leq f$ will be denoted $f \leq O(g)$ and $\Omega(g) \leq f$ respectively. We shall use the symbole $\widetilde{O}$ when this is true up to log factors. For a vector $u \in \mathbb{R}^d$, $\mathrm{diag}(u)$ denotes the $d \times d$ diagonal matrix which has its diagonal equal to u. For a matrix $M \in \mathbb{R}^{d \times d}$, $\mathrm{diag}(M)$ denotes the vector $(M_{11}, \dots, M_{dd}) \in \mathbb{R}^d$. The indexed vector $\beta^\star$ will stand for any β interpolating the data, i.e. any vector in the affine space $\{\beta \in \mathbb{R}^d \; s.t. \; X\beta = Y\}$ of dimension at least $d - n$. Out of all these, let $\beta_{\ell_1}^* = \underset{\beta \in \mathbb{R}^d \text{ s.t. } X\beta = y}{\arg\min} \|\beta\|_1$. For z any vector, z_∞ or z^∞ will always designate of $\lim_{t \to \infty} z_t$.

7.3 Setup and preliminaries

7.3.1 Architecture and algorithm.

Overparametrised noiseless regression. We consider a linear regression problem with outputs $(y_1, \dots, y_n) \in \mathbb{R}^n$ and inputs $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$. We study an overparametrised setting $(n < d)$ and assume that there exists at least one interpolating parameter $\beta^\star \in \mathbb{R}^d$ which perfectly fits the training set, i.e. $y_i = \langle \beta^\star, x_i \rangle$ for all $1 \leq i \leq n$. We parametrise the regression vector β as β_w with $w \in \mathbb{R}^p$. We will see that though in the end our final models $x \mapsto \langle \beta_w, x \rangle$ are classical linear models whatever the parametrisation $w \mapsto \beta_w$, the choice of this parametrisation has crucial consequences on the solution recovered by the learning algorithms. We study the quadratic loss and the overall loss is written as:

$$L(w) = L(\beta_w) := \frac{1}{4n} \sum_{i=1}^n (\langle \beta_w, x_i \rangle - y_i)^2 = \frac{1}{4n} \sum_{i=1}^n \langle \beta_w - \beta^\star, x_i \rangle^2,$$

where by abuse of notation we use $L(w) = L(\beta_w)$.

2-layer diagonal linear network. The simplest parametrisation of β_w is to consider $\beta_w = w$ which corresponds to the classical least-squares framework. It is well known that in this case, many first order methods (GD, SGD, with and without momentum) will converge towards the same solution: we say that they have the same implicit bias. This is experimentally not the case for neural networks where SGD has been shown to lead to solutions which have better generalisation properties compared to GD [Keskar et al., 2017]. To theoretically confirm this observation, we study a simple non-linear parametrisation: $\beta_w = w_+^2 - w_-^2$ with $w = [w_+, w_-]^\top \in \mathbb{R}^{2d}$. We point out that it is 2-positive homogeneous and that it is equivalent to the parametrisation $\beta_{u,v} = u \odot v$ with $u, v \in \mathbb{R}^d$. It should be thought of a simplified linear network of depth 2

(see [Woodworth et al., 2020b, Section 4] for more details). We consider two weight vectors w_+ and w_- (and not only $\beta_w = w^2$) in order to ensure that our final linear predictor parameter β_w can take negative values. Also note that additionally to being a toy neural model, it has received recent attention for its practical ability to induce sparsity [Vaskevicius et al., 2019, 2020a, HaoChen et al., 2021] or to solve phase retrieval problems [Wu and Rebeschini, 2020].

Stochastic Gradient Descent. With this quadratic parametrisation, the loss now rewrites as: $L(w) = \frac{1}{4n}\sum_{i=1}^n \langle w_+^2 - w_-^2 - \beta^\star, x_i\rangle^2$. Note that despite its simplicity, this loss is non convex and its minimisation is non trivial. The algorithm we shall consider is the well known SGD algorithm, where for a step size $\gamma > 0$:

$$\begin{aligned} w_{t+1,+} &= w_{t,+} - \gamma\langle \beta_w - \beta^\star, x_{i_t}\rangle \; x_{i_t} \odot w_{t,+} \\ w_{t+1,-} &= w_{t,-} + \gamma\langle \beta_w - \beta^\star, x_{i_t}\rangle \; x_{i_t} \odot w_{t,-} \end{aligned} \qquad \text{where } i_t \sim \mathrm{Unif}(1,n). \tag{7.1}$$

It is convenient to rewrite this recursion as

$$w_{t+1,\pm} = w_{t,\pm} - \gamma\nabla_{w_\pm} L(w_t) \pm \gamma \,\mathrm{diag}(w_{t,\pm})X^\top \xi_{i_t}(\beta_t), \tag{7.2}$$

where $\xi_{i_t}(\beta) = -\big(\langle \beta - \beta^\star, x_{i_t}\rangle \mathbf{e}_{i_t} - \mathbb{E}_{i_t}\big[\langle \beta - \beta^\star, x_{i_t}\rangle \mathbf{e}_{i_t}\big]\big) \in \mathbb{R}^n$ is a zero-mean *multiplicative* noise which vanishes at any global optimum ($\mathbf{e}_i$ denotes the i^{th} element of the canonical basis). We point out that all the results we shall give hold for any initialisation such that $w_{t=0,+} = w_{t=0,-} \in \mathbb{R}^d$, under which we have that $\beta_{w_{t=0}} = 0$. To understand under what conditions the SGD procedure converges and towards which point it does, we shall consider its continuous counterpart which has the advantage of leading to clean and intuitive calculations. We highlight the fact that we consider a bath-size equal to 1 for clarity, however all our analysis holds for mini-batch SGD (with and without replacement) simply by considering an effective step-size γ_{eff} instead of γ, this is clearly explained in Appendix B.1.

7.3.2 Stochastic gradient flow

Continuous time modelling of sequential processes offer a large set of tools, such as derivation, which come in helpful to understand the dynamics of the processes. This has led to a large part of the recent literature to consider continuous gradient flow in order and understand the behaviour of gradient descent on complicated architectures such as neural nets. However, the continuous time modelling of stochastic gradient descent is more challenging: it requires to add on top of the gradient flow a diffusion term whose covariance matches the one of SGD. Hence, it is fundamental to understand its structure and scale.

Understanding the noise's structure. As seen in equation (7.2), evaluated at $w_\pm$, the stochastic noise $\gamma\,\mathrm{diag}(w_\pm)X^\top\xi_{i_t}(w)$ has two main characteristics which we want to preserve:

- It belongs to $\mathrm{span}(w_\pm \odot x_1, \dots, w_\pm \odot x_n)$
- It has covariance $\Sigma_{\text{SGD}}(w_\pm) := \gamma^2\,\mathrm{diag}(w_\pm)X^\top \mathrm{Cov}_{i_t}(\xi_{i_t}(\beta))X\,\mathrm{diag}(w_\pm) \in \mathbb{R}^{d\times d}$

It remains to understand the structure of the covariance of ξ_{i_t} which has the following closed form: $\mathrm{Cov}_{i_t}(\xi_{i_t}(\beta)) = \frac{1}{n}\,\mathrm{diag}(\langle\beta-\beta^\star, x_i\rangle^2)_{1\le i\le n} - \frac{1}{n^2}\big(\langle\beta-\beta^\star, x_i\rangle\langle\beta-\beta^\star, x_j\rangle\big)_{1\le i,j\le n}$. We identify the two key facts: (i) it is diagonal at the leading n^{-1} order and (ii) its trace is linked to the loss as $\mathrm{Var}_{i_t}(\|\xi_{i_t}(\beta)\|_2) = \frac{4}{n}L(\beta) + O(\frac{1}{n^2})$. This leads us in modelling $\xi_{i_t}(\beta)$'s covariance matrix as $\frac{4}{n}L(\beta)I_n$ as it preserves these two characteristics [1]. Finally this brings us to consider the following modelling of the overall noise's structure: $\Sigma_{\text{SGD}}(w_\pm) \cong \frac{4}{n}\gamma^2 L(w)[\mathrm{diag}(w_\pm)X^\top]^{\otimes 2}$.

[1] the general case is discussed in Appendix B.4.1

Stochastic differentiable equation modelling. Guided by the previous considerations, we study the following stochastic gradient flow:

$$\begin{aligned} \mathrm{d}w_{t,+} &= -\nabla_{w_+} L(w_t)\,\mathrm{d}t + 2\sqrt{\gamma n^{-1} L(w_t)}\; w_{t,+} \odot [X^\top \mathrm{d}B_t] \\ \mathrm{d}w_{t,-} &= -\nabla_{w_-} L(w_t)\,\mathrm{d}t - 2\sqrt{\gamma n^{-1} L(w_t)}\; w_{t,-} \odot [X^\top \mathrm{d}B_t], \end{aligned} \tag{7.3}$$

where $\mathrm{d}B_t$ is a standard $\mathbb{R}^n$ Brownian motion. The SDE is a perturbed gradient flow with a diffusion term that is defined such that its Euler discretisation with step size γ leads to a Markov Chain whose covariance exactly matches SGD's noise covariance $\Sigma_{\mathrm{SGD}}(w_\pm)$. We refer to Li et al. [2019a] or Kloeden and Platen [1992] for the technical details regarding consistency of such a procedure in the limit of small step sizes. This stochastic differential equation is the starting point of the analysis.

7.4 The implicit bias of the stochastic gradient flow

Implicit bias and hyperbolic entropy. To understand the relevance of the main result and how stochasticity induces a preferable bias, we start by recalling some known results for gradient flow. In Woodworth et al. [2020b] it is shown, assuming global convergence, that the solution selected by the gradient flow initialised at $\alpha \in \mathbb{R}^d$ and denoted β_∞^α solves a constrained optimisation problem involving the *hyperbolic entropy* introduced by Ghai et al. [2020]:

$$\beta_\infty^\alpha = \underset{\beta \in \mathbb{R}^d \text{ s.t. } X\beta = y}{\arg\min}\; \phi_\alpha(\beta) := \frac{1}{4}\Big[\sum_{i=1}^d \beta_i \operatorname{arcsinh}\Big(\frac{\beta_i}{2\alpha_i^2}\Big) - \sqrt{\beta_i^2 + 4\alpha_i^4}\Big], \tag{7.4}$$

Though the hyperbolic entropy function has a non-trivial expression, its principal characteristic is that it interpolates between the ℓ_1 and the ℓ_2 norms according to the scale of α. More precisely for $\alpha \in \mathbb{R}$ [2]: $\phi_\alpha(\beta) \underset{\alpha \to 0}{\sim} \frac{1}{2}\ln\big(\frac{1}{\alpha}\big)\|\beta\|_1$ and $\phi_\alpha(\beta) \underset{\alpha \to +\infty}{=} 2\alpha^2 + \frac{1}{4\alpha^2}\|\beta\|_2^2 + o(\alpha^{-2})$. We refer to [Woodworth et al., 2020b, Theorem 2] for more details on the asymptotic analysis. The implicit optimisation problem (7.4) therefore highlights the fact that the initialisation scale of the weights controls the shape of the recovered solution. Small initialisations lead to low ℓ_1-norm solutions which are known to induce good generalisation properties: this is what is often referred to as the *rich regime*. Large initialisations lead to low ℓ_2-norm solutions: this is referred to as the *kernel regime* or *lazy regime* in which the weights move only very slightly. The dynamics of the gradient flow are then very similar to the one of kernel linear regression with the kernel depending on the initialisation [Jacot et al., 2018, Chizat et al., 2019]. Overall, to retrieve a sparse solution, one should initialise with the smallest α possible. However, as is clearly explained in Woodworth et al. [2020b], it is important to stress out that there is a generalisation / optimisation tradeoff: the point $w = 0$ happens to be a saddle point for the loss and a smaller α will lead to a longer training time.

Main result. In the main theorem we show that, for an initialisation scale α, the stochasticity of SGF biases the flow towards solutions which still minimise the hyperbolic entropy. However, what is remarkable is that it does so with an effective parameter α_∞ which is strictly smaller than α. The recovered solution therefore minimises an optimisation problem which has better sparsity inducing properties than that of gradient flow.

[2] If $\alpha \in \mathbb{R}$ we consider the abuse of notation $\phi_\alpha := \phi_{\alpha \mathbf{1}}$.

Theorem 1. *For* $p \leq \frac{1}{2}$ *and* $w_{0,\pm} = \alpha \in (\mathbb{R}_+^*)^d$*, let* $(w_t)_{t\geq 0}$ *follow the stochastic gradient flow* (7.3) *with step size* $\gamma \leq O\big(\big[\ln(\frac{4}{p})\lambda_{\max} \max\{\|\beta^\star_{\ell_1}\|_1 \ln\big(\frac{\|\beta^\star_{\ell_1}\|_1}{\min_i \alpha_i^2}\big), \|\alpha\|_2^2\}\big]^{-1}\big)$ *where* $\beta^\star_{\ell_1} = \underset{\beta\in\mathbb{R}^d \, s.t. \, X\beta=y}{\arg\min} \|\beta\|_1$ *and* $\lambda_{\max}$ *is the largest eigenvalue of* $X^\top X/n$*. Then, with probability at least* $1-p$*:*

- $(\beta_t)_{t\geq 0}$ *converges towards a zero-training error solution* β^α_∞
- *the solution* β^α_∞ *satisfies*

$$\beta^\alpha_\infty = \underset{\beta\in\mathbb{R}^d \ s.t. \ X\beta=y}{\arg\min} \phi_{\alpha_\infty}(\beta) \quad \text{where} \quad \alpha_\infty = \alpha \odot \exp\left(-2\gamma \operatorname{diag}\left(\frac{X^\top X}{n}\right)\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s\right). \tag{7.5}$$

The theorem is three-fold: with high probability and for an explicit choice of constant step size γ, (i) the flow $(\beta_t)_{t\geq 0}$ converges, (ii) its limit β^α_∞ is an interpolating solution, i.e. $X\beta^\alpha_\infty = y$, (iii) this solution minimises the hyperbolic entropy problem with a parameter that depends on the dynamics. We illustrate these results in Figure 8.1. Now let us comment further the theorem.

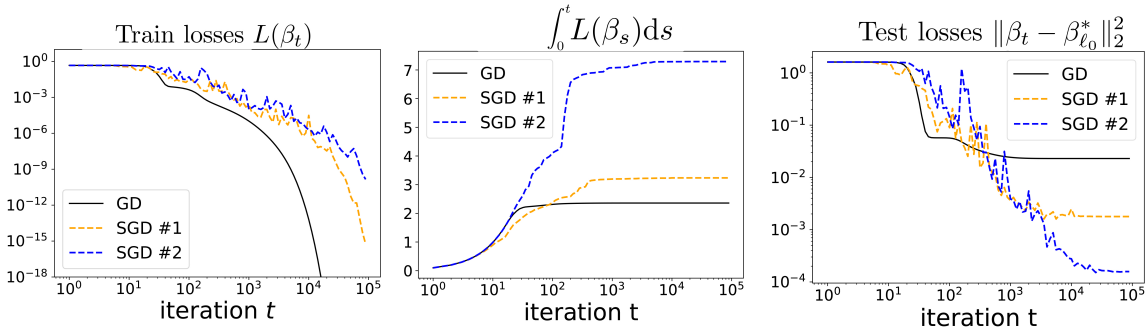


Figure 7.2: Sparse regression (see Section 7.6.1 for the detailed experimental setting). Both SGD and GD are initialised at $\alpha = 0.1$. 2 different runs of SGD over the training set are performed, they differ due to the inner stochasticity of the algorithm. *Left*: GD and SGD both converge towards a global minimum. *Middle and right*: for two different trajectories of SGD, the higher the value of the loss integral at convergence, the better the validation loss. In both cases SGD converges towards a solution which generalises better than GD. This figure illustrates Theorem 1.

Beneficial implicit bias through effective initialisation. The most remarkable aspect of the result is that the recovered solution β^α_∞ minimises the same potential as for gradient flow but with an *effective parameter* α_∞ which is strictly smaller than α. Hence, the hyperbolic entropy is closer to the ℓ_1 norm compared to the deterministic case, proving a systematic benefit of stochasticity. Note that this effective parameter is random and controlled by the loss integral $\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s$: the higher the integral, the smaller the effective initialisation scale. In other words and quite surprisingly, the slower the loss converges to 0, the "richer" the implicit bias. However, it must be kept in mind that, as explained in Woodworth et al. [2020b], there is a tension between generalisation and optimisation: a longer training time might improve generalisation but comes at the cost of... a longer training time. Yet it is clear experimentally that SGD systematically largely wins the trade-off over GD (see Figure 8.1). Interestingly, Problem (7.5) tells us that the implicit bias of SGD initialised at α acts as if we run GD initialised at α_∞ (see Section 7.6.3). Note that the minimisation problem (7.5) only makes sense *a posteriori* since the quantity α_∞

depends on the whole stochastic trajectory. Finally, an interesting question is whether one can quantify the scale of this beneficial phenomenon, i.e. how small α_∞ is compared to α. To answer this, we quantify the scale of the loss integral w.r.t. γ and α (see Proposition 16) and show under slightly stronger conditions that the relative scale α_∞/α decays as power of α (See Eq. (7.8) of the main text and Proposition 30 of the appendix for a proof).

Kernel regime. Though it is less our focus, our result still holds as $\alpha \to +\infty$ which corresponds to the kernel regime. In this regime, we believe that $\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s \underset{\alpha\to\infty}{\to} 0$ (not shown in the chapter but experimentally observed) and hence SGF and GF converge towards the same solution. This is expected since in the NTK regime, the iterates follow a kernel linear regression for which the bias of SGF and GF are the same.

Step size. Note that the convergence of the iterates holds for a constant step size. This is not illogical since in the overparametrised setting, the noise vanishes at the optimum (see Varre et al. [2021] for a convergence result in the overparametrised least-squares setup). The explicit formula for the γ upper bound is $\gamma \leq \left(400 \ln\left(\frac{4}{p}\right)\lambda_{\max}(\frac{X^\top X}{n}) \max\left\{\|\beta^\star_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^\star_{\ell_1}\|_1}{\min_i \alpha_i^2}\right), \|\alpha\|_2^2\right\}\right)^{-1}$. It has a classical dependence on $\lambda_{\max}(X^\top X/n)$ which can be computed, but also on the unknown value of $\|\beta^\star_{\ell_1}\|_1$. However in practice we choose the highest value of γ for which the iterates converge. Note that in practice the weights are often initialised such that $\|\alpha\|_2^2$ is roughly equal to 1 and hence it is sensible to consider $\|\alpha\|_2^2 < \|\beta^*_{\ell_1}\|_1$. In the explicit bound, there is a $\ln\left(\|\beta^*_{\ell_1}\|_1/\min_i \alpha_i^2\right)^{-1}$ factor, we believe that it is an artefact of our analysis and could be removed. It is hence best to think of the upperbound on γ to simply be $\gamma \leq O(\frac{1}{\lambda_{\max}\|\beta^*_{\ell_1}\|_1})$.

Convergence and proof sketch. Let us put emphasis on the fact that since we deal with a non-convex problem, neither convergence nor convergence towards a global minimum are obvious. In most of similar works, convergence of the iterates is assumed [Woodworth et al., 2020b, Gunasekar et al., 2018a]. In fact, the hardest and most technical part of our result is to show the convergence of the flow with high probability: once the convergence is shown, describing the minimisation problem β^α_∞ verifies is straightforward. In the following section we give several properties which constitute the major keys of the theorem's proof.

7.5 Links with mirror descent

The aim of this section is to show that the sequence $(\beta_t)_{t\geq 0}$ follows a stochastic version of continuous mirror descent with a time dependent mirror. From this crucial property, we show how the convergence and implicit bias characterisation follow. Finally, as it is one of the central objects of our main theorem, we give an estimation of $\int_0^\infty L(\beta_s)\,\mathrm{d}s$.

7.5.1 Stochastic continuous mirror descent with time-varying potential

We start by recalling known results on the link between implicit bias and mirror descent. We recall also convergence guarantees for mirror descent dynamics.

Mirror descent: convergence and implicit bias. For any $\beta_0 \in \mathbb{R}^d$ and convex potential function Ψ, consider the mirror descent flow $(\beta_t)_t$ which corresponds to $\mathrm{d}\nabla\Psi(\beta_t) = -\nabla L(\beta_t)\mathrm{d}t$. Though the convergence of the loss to 0 is straightforward, showing the convergence of the iterates requires more work and is shown in [Bauschke et al., 2017, Theorem 2]. Yet, once the convergence of the iterates is shown, deriving the implicit minimisation problem is straightforward.

We recall the reasoning here (see Section 3 of Azulay et al. [2021] for more details): integrating the flow yields $\nabla\Psi(\beta_\infty) - \nabla\Psi(\beta_0) = -\int_0^\infty \nabla L(\beta_s)\,\mathrm{d}s = -4X^\top \int_0^\infty X(\beta_s - \beta_\infty)\,\mathrm{d}s \in \mathrm{span}(X)$. This condition, along with the fact that $X\beta_\infty = y$ exactly corresponds to the KKT conditions of the problem:

$$\beta_\infty = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta=y}{\arg\min} D_\Psi(\beta, \beta_0), \tag{7.6}$$

where $D_\Psi(\beta, \beta_0) = \Psi(\beta) - \Psi(\beta_0) - \langle \nabla\Psi(\beta_0), \beta - \beta_0\rangle$ is the Bregman divergence w.r.t. Ψ.

Link with our model. It turns out that these general observations on mirror descent apply to our framework when $(w_t)_t$ follows the gradient flow $\mathrm{d}w_{t,\pm} = -\nabla_{w_\pm} L(w_t)\,\mathrm{d}t$. Indeed it has been shown in Woodworth et al. [2020b] that the corresponding iterates $\beta_t = w_{t,+}^2 - w_{t,-}^2$ follow a mirror descent with potential ϕ_α defined in Eq.(7.4). Therefore we can apply the previous remarks to obtain the convergence towards an interpolator, as well as the associated implicit minimisation problem which in our case can be rewritten as $\beta_\infty^\alpha = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta=y}{\arg\min} \phi_\alpha(\beta)$ since $\nabla\phi_\alpha(\beta_0 = 0) = 0$.

Stochastic Mirror descent with a time varying potential. To address the problem where $(w_t)_t$ follows a stochastic gradient flow instead of a gradient flow, it is natural, as in the deterministic framework, to see what type of flow $(\beta_t)_t$ follows. Because of the noise, we cannot hope to simply recover a classical mirror descent. However interestingly the next property shows that it follows a stochastic mirror-like descent with a geometry that depends on time.

Proposition 14. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow in Eq.(7.3) with initialisation $w_{0,\pm} = \alpha \in (\mathbb{R}_+^*)^d$. Then the corresponding flow $(\beta_t)_{t\geq 0}$ follows a "stochastic continuous mirror descent with time varying potential" defined by:*

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma n^{-1} L(\beta_t)}X^\top \mathrm{d}B_t, \tag{7.7}$$

where $\alpha_t = \alpha \odot \exp\left(-2\gamma\,\mathrm{diag}\left(\frac{X^\top X}{n}\right)\int_0^t L(\beta_s)\,\mathrm{d}s\right)$ and ϕ_α is the hyperbolic entropy defined in (7.4).

Under this form we clearly see that the iterates $(\beta_t)_t$ follow a flow which closely resembles that of mirror descent but with two major differences: (i) the potential ϕ_{α_t} changes over time according to the random quantity $\int_0^t L(\beta_s)\,\mathrm{d}s$, (ii) the flow is perturbed by noise. We highlight the fact that viewing the dynamics this way has the major advantage of giving a clear roadmap for the proof of Theorem 1: (i) we can adapt classical mirror-descent results to our framework and construct appropriate Lyapunov functions to prove the convergence of the flow with high probability to some interpolator β_∞^α, (ii) we immediately recover the corresponding minimisation problem as in the deterministic case. Indeed, integrating Eq.(7.7) still yields $\nabla\phi_{\alpha_\infty}(\beta_\infty^\alpha) \in \mathrm{span}(X)$ which, along with $X\beta_\infty^\alpha = y$, are the KKT conditions of the implicit minimisation problem (7.5). We emphasise the fact that the structure of the noise, belonging to $\mathrm{span}(X)$, is crucial in order to obtain this minimisation problem. This would for instance clearly not be true if we considered isotropic noise in the SDE modelling. This highlights the fact that not every form of noise improves the implicit bias: the shape of the intrinsic SGD noise is of primal importance [HaoChen et al., 2021].

7.5.2 Convergence and control of $\int_0^\infty L(\beta_s)\,\mathrm{d}s$

Though it seems easy to derive the implicit minimisation problem (7.5) from the mirror-like structure of Eq.(7.7), it is necessary to ensure that the iterates converge towards an interpolator β_∞. This is the purpose of the following proposition.

Proposition 15 (Convergence of the iterates)**.** *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow (7.3), initialised at $w_{0,\pm} = \alpha \in (\mathbb{R}_+^*)^d$. For $p \leq \frac{1}{2}$ and γ such as in Theorem 1, then with probability at least $1-p$, the flow $(\beta_t)_t$ converges to an interpolating solution β_∞^α.*

The convergence of the iterates is technical and requires several intermediate results. We start by considering an appropriate Bregman-type stochastic function with a time-varying potential and show that it converges with high probability. Leveraging the fact that we are able to bound the iterates β_t, we are able to show that the limit of the function is in fact 0. Owing to the fact that the function we consider also controls the distance of β_t to a particular $\beta^\star$ we finally get that the iterates converge.

However for the objects (such as α_∞) and functions we introduce to be well defined, we need to guarantee the convergence of $\int_0^\infty L(\beta_s)\mathrm{d}s$. Besides, it is crucial to grasp the scale of this quantity since it gives the overall scale of α_∞. This is done in the following proposition where we lower and upper bound its value.

Proposition 16. *Under the same setting as in Proposition 15 with initialisation $w_{0,\pm} = \alpha\mathbf{1}$, we have with probability at least $1-p$:*

$$\Omega\Big(\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\Big)\Big) \underset{\alpha\to 0}{\leqslant} \int_0^{+\infty} L(\beta_s)\,\mathrm{d}s \leqslant O\Big(\max\big\{\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\Big), \alpha^2 d\big\}\Big).$$

We point out that the lower bound is given for small α's for simplicity but we provide in Lemma 16 (Appendix B.2.5) a lower bound which holds for all α's. Note that when $\gamma = 0$, which corresponds to deterministic gradient flow, we can give the exact value for the integral: $\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s = \frac{1}{2} D_{\phi_\alpha}(\beta_\infty^\alpha, \beta_0)$. This matches the scale of the bounds given in Proposition 16, hence showing the tightness of the result. We focus now on how this translates to the scale of the effective initialisation w.r.t. α when this latter is small enough. In fact, this lower bound on the integral of the loss along with a stronger assumption on the boundedness of the iterates lead to

$$\frac{\alpha_\infty}{\alpha} \underset{\alpha\to 0}{\leqslant} \left(\frac{\alpha^2}{\|\beta^\star_{\ell_1}\|_1}\right)^\zeta, \tag{7.8}$$

for some $\zeta > 0$. Hence the smaller the initialisation scale α and the greater the benefit of SGD over GD in terms of implicit bias (see Appendix B.2.6 for more details).

Again, the proof of this proposition is technical and relies on considering appropriate Lyapunov functions which highly resemble to Bregman divergences, but which take into account the fact that the geometry changes over time. These overall decreasing Lyapunov's enable to bound the iterates as well as lower and upper bound the integral of the loss. The stochastic integrals which naturally appear are controlled with high probability using time-uniform concentration of martingales [Howard et al., 2020].

7.6 Experiments

7.6.1 Experimental setup for sparse regression

We consider the following sparse regression setup for our experiments. We choose $n = 40$, $d = 100$ and randomly generate a sparse model $\beta^\star_{\ell_0}$ such that $\|\beta^\star_{\ell_0}\|_0 = 5$. We generate the features as $x_i \sim \mathcal{N}(0, I)$ and the labels as $y_i = x_i^\top \beta^\star_{\ell_0}$. SGD, GD and the SGF are always initialised using the same scale $\alpha > 0$ and it is specified each time. We use the same step size for GD and SGD and choose it to be the biggest as possible why still ensuring convergence. Note that since the true population covariance $\mathbb{E}[xx^\top]$ is equal to identity, the quantity $\|\beta_t - \beta^\star_{\ell_0}\|_2^2$ corresponds to the validation loss.

7.6.2 Validation of the SDE model

In this section, we present an experimental validation of the stochastic gradient flow model. In Figure 7.3, for the same step size, we run: (i) the trajectory of gradient descent, (ii) 5 trajectories of stochastic gradient descent that correspond to different realisations of the uniform sampling over the data, (iii) 5 trajectories of the stochastic gradient flow (its Euler discretisation with $\mathrm{d}t = \gamma/10$)) corresponding to different realisations of the Brownian. We clearly see (left) that the loss behaves similarly for SGD and SGF across time. We also see that the validation losses (right) of the iterates of SGD and SGF have very similar behaviours. This tends to validate our continuous modelling from Section 7.3.2.

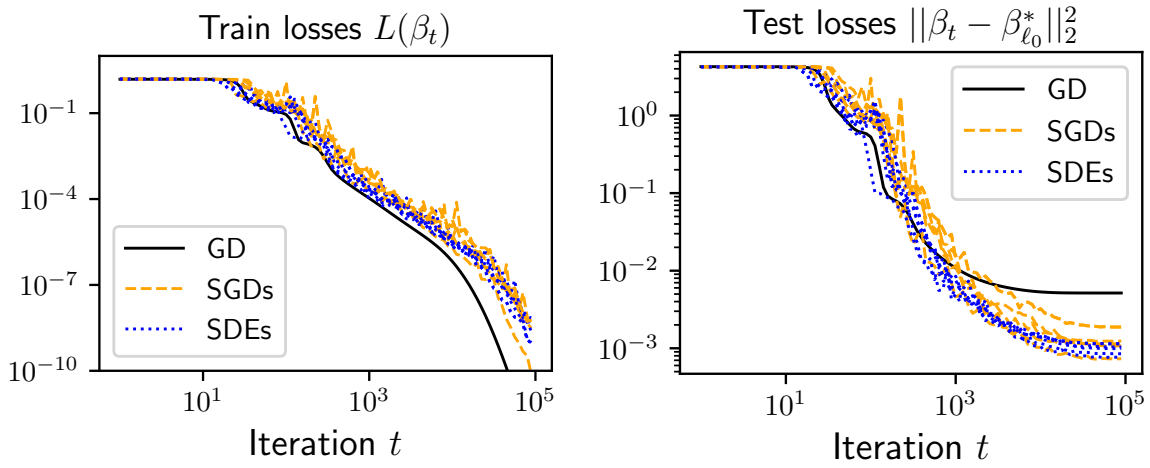


Figure 7.3: Sparse regression (see Section 7.6.1 for the detailed experimental setup). *Left and right*: the training and the validation losses behave very similarly, corroborating the continuous modelling.

7.6.3 GD and SGD have the same implicit bias, but from different initialisations

In order to confirm and illustrate the main Theorem 1, we provide the following experiment which is illustrated Figure 7.4. We first run GD and SGD with the same step-size and initialise them both at $\alpha\mathbf{1}$ with $\alpha = 0.01$. As expected, the solution recovered by SGD generalises better. Then, using the iterates β_t^{SGD} from the first SGD run, we compute the value $\alpha_\infty = \alpha \exp(-2\gamma \operatorname{diag}(X^\top X/n) \int_0^\infty L(\beta_s^{\mathrm{SGD}})\mathrm{d}s) \in \mathbb{R}^d$ (the integral is approximated by its discrete time approximation with $\mathrm{d}t = \gamma$). We then run gradient descent but this time initialised at $w_{0,\pm} = \alpha_\infty$. According to our main result from Theorem 1, it should approximately (it would be exact

if we ran SGF and GF) converge to the same solution as SGD initialised at $\alpha\mathbf{1}$. This is clearly observed Figure 7.4 (right). Also note that SGD and GD (initialised at α_∞) seem to have overall very similar dynamics, this is not shown by our results and we leave this as future work. However keep in mind that though the validation losses converge at the same iteration rate, in terms of computation time, SGD is n times faster.

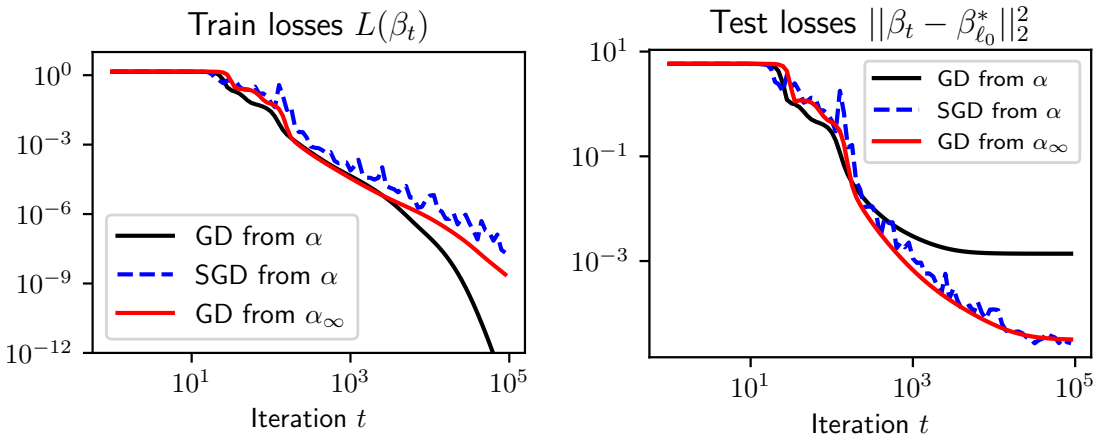


Figure 7.4: Sparse regression (see Section 7.6.1 for the detailed experimental setup). *Left and right*: SGD initialised at $\alpha\mathbf{1}$ converges towards the same point as GD initialised at $\alpha_\infty = \alpha \exp(-2\gamma \operatorname{diag}(X^\top X/n) \int_0^\infty L(\beta_s^{\mathrm{SGD}})\mathrm{d}s)$.

7.6.4 Doping the implicit bias with label noise

As largely discussed throughout the chapter, the effect of the implicit bias is controlled by the convergence speed of the loss: the slower it converges, the sparser the selected solution will be. Hence the following question: can we leverage this knowledge to dope the implicit bias? We argue in this Section that the answer to this question is affirmative. Indeed, consider a sequence $(\delta_t)_{t\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ and assume that we artificially inject some label noise Δ_t at time t, say for example $\Delta_t \sim \mathrm{Unif}\{2\delta_t, -2\delta_t\}$ (independently from i_t). This injected label noise perturbs the SGD recursion as follows:

$$w_{t+1,\pm} = w_{t,\pm} \mp \gamma \left(\langle \beta_w - \beta^\star, x_{i_t}\rangle + \Delta_t\right) \; x_{i_t} \odot w_{t,+} \; , \qquad \text{where} \;\; i_t \sim \mathrm{Unif}(1,n). \tag{7.9}$$

As in Section 7.3.2, we can derive its related stochastic gradient flow (see Appendix B.3.1 for more details):

$$\mathrm{d}w_{t,\pm} = -\nabla_{w_\pm} L(w_t)\mathrm{d}t \pm 2\sqrt{\gamma n^{-1}(L(w_t) + \delta_t^2)} \; w_{t,+} \odot [X^\top \mathrm{d}B_t]. \tag{7.10}$$

Assuming that $(\delta_t)_{t\geq 0} \in (\mathbb{R}_+)^{\mathbb{R}}$ and γ are such that the iterates converge, the corresponding implicit regularisation minimisation problem is preserved but with a "slowed down" loss: $\tilde{L}(\beta_t) := L(\beta_t) + \delta_t^2$ and the effective initialisation writes: $\tilde{\alpha}_\infty = \alpha \odot \exp\left(-2\gamma \operatorname{diag}(\frac{X^\top X}{n}) \int_0^{+\infty} \tilde{L}(\beta_s)\,\mathrm{d}s\right)$. The label noise therefore helps recovering a solution which has better sparsity properties. However, it must be kept in mind that adding too much label noise can significantly slow down the convergence of the validation loss or even prevent the iterates from converging. Yet, experimental results showing the impressive effect of label noise are provided Figure B.1 in Appendix B.3.1.

7.7 Conclusion

In this chapter, we have shown the benefit of using stochastic gradient descent over gradient descent for diagonal linear networks in terms of their implicit bias. Indeed, we prove that stochastic gradient flow acts as gradient flow but initialised at a smaller scale: this induces a sparser finale iterate. This effect is controlled by the speed of convergence of the loss. Moreover, we prove the convergence of the flow and exhibit an interesting link with mirror descent. Fully understanding this novel type of dynamics could help to grasp the implicit biasing properties of stochastic gradient descent in other frameworks. It is also natural to ask whether the integral of the loss also controls the difference of implicit regularisation for more general architectures. It would also be interesting to analyse how this property adapts to log losses known to lead to max-margin solutions in classification.

Chapter 8

Effect of the step size

8.1 Preface

This chapter follows Even et al. [2023].

Summary We investigate the impact of stochasticity and large stepsizes on the implicit regularisation of gradient descent (GD) and stochastic gradient descent (SGD) over 2-layer diagonal linear networks. We prove the convergence of GD and SGD with macroscopic stepsizes in an overparametrised regression setting and provide a characterisation of their solution through an implicit regularisation problem. Our characterisation provides insights on how the choice of minibatch sizes and stepsizes lead to qualitatively distinct behaviors in the solutions. Specifically, we show that for sparse regression learned with 2-layer diagonal linear networks, large stepsizes consistently benefit SGD, whereas they can hinder the recovery of sparse solutions for GD. These effects are amplified for stepsizes in a tight window just below the divergence threshold, known as the "edge of stability" regime.

Co-authors Mathieu Even, Suriya Gunasekar and Nicolas Flammarion.

Contributions Mathieu and Scott worked together on the project.

8.2 Introduction

The stochastic gradient descent algorithm (SGD) [Robbins and Monro, 1951] is the foundational algorithm for almost all neural network training. Though a remarkably simple algorithm, it has led to many impressive empirical results and is a key driver of deep learning. However the performances of SGD are quite puzzling from a theoretical point of view as (1) its convergence is highly non-trivial and (2) there exist many global minimums for the training objective which generalise very poorly [Zhang et al., 2017].

To explain this second point, the concept of implicit regularisation has emerged: if overfitting is harmless in many real-world prediction tasks, it must be because the optimisation process is *implicitly favoring* solutions that have good generalisation properties for the task. The canonical example is overparametrised linear regression with more trainable parameters than number of samples: although there are infinitely many solutions that fit the samples, GD and SGD explore only a small subspace of all the possible parameters. As a result, it can be shown that they implicitly converge to the closest solution in terms of the ℓ_2 distance, and this without explicit regularisation [Lemaire, 1996, Gunasekar et al., 2018a].

Currently, most theoretical works on implicit regularisation have primarily focused on continuous time approximations of (S)GD where the impact of crucial hyperparameters such as the stepsize and the minibatch size are ignored. One such common simplification is to analyse gradient flow, which is a continuous time limit of GD and minibatch SGD with an infinitesimal

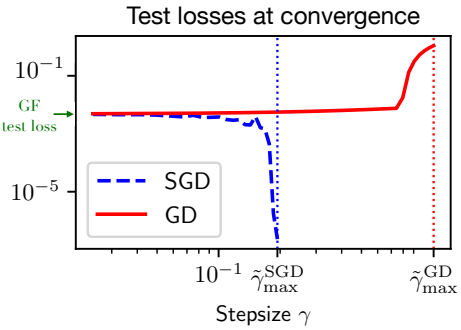


Figure 8.1: Noiseless sparse regression with a diagonal linear network using SGD and GD, with parameters initialized at the scale of $\alpha = 0.1$ (Section 8.3). The test losses at convergence for various stepsizes are plotted for GD and SGD. Small stepsizes correspond to gradient flow (GF) performance. We see that increasing the stepsize improves the generalisation properties of SGD, but deteriorates that of GD. The dashed vertical lines at stepsizes $\tilde{\gamma}^{\mathrm{SGD}}_{\mathrm{max}}$ and $\tilde{\gamma}^{\mathrm{GD}}_{\mathrm{max}}$ denote the largest stepsizes for which SGD and GD, respectively, converge. See Section 8.3 for the precise experimental setting.

stepsize. By definition, this analysis does not capture the effect of stepsize or stochasticity. Another approach is to approximate SGD by a stochastic gradient flow [Wojtowytsch, 2021, Pesme et al., 2021], which tries to capture the noise and the stepsize using an appropriate stochastic differential equation. However, there are no theoretical guarantees that these results can be transferred to minibatch SGD as used in practice. This is a limitation in our understanding since the performances of most deep learning models are often sensitive to the choice of stepsize and minibatch size. The importance of stepsize and SGD minibatch size is common knowledge in practice and has also been systematically established in controlled experiments [Keskar et al., 2017, Masters and Luschi, 2018, Geiping et al., 2022].

In this work, we aim to expand our understanding of the impact of stochasticity and stepsizes by analysing the (S)GD trajectory in 2-layer diagonal networks (DLNs). In Figure 8.1, we show that even in our simple network, there are significant differences between the nature of the solutions recovered by SGD and GD at macroscopic stepsizes. We discuss this behavior further in the later sections.

The 2-layer diagonal linear network which we consider is a simplified neural network that has received significant attention lately [Woodworth et al., 2020a, Vaskevicius et al., 2019, HaoChen et al., 2021, Pillaud-Vivien et al., 2022]. Despite its simplicity, it surprisingly reveals training characteristics which are observed in much more complex architectures, such as the role of the initialisation [Woodworth et al., 2020a], the role of noise [Pesme et al., 2021, Pillaud-Vivien et al., 2022], or the emergence of saddle-to-saddle dynamics [Berthier, 2022, Pesme and Flammarion, 2023]. It therefore serves as an ideal proxy model for gaining a deeper understanding of complex phenomenons such as the roles of stepsizes and of stochasticity as highlighted in this chapter. We also point out that implicit bias and convergence for more complex architectures such as 2-layer ReLU networks, matrix multiplication are not yet fully understood, even for the simple gradient flow. Therefore studying the subtler effects of large stepsizes and stochasticity in these settings is currently out of reach.

8.2.1 Main results and chapter organisation

The overparametrised regression setting and diagonal linear networks are introduced in Section 8.3. We formulate our theoretical results (Theorems 1 and 2) in Section 8.4: we prove that for **macroscopic stepsizes**, gradient descent and stochastic gradient descent over 2-layer diagonal linear networks converge to a zero-training loss solution $\beta^\star_\infty$. We further provide a refined characterization of $\beta^\star_\infty$ through a trajectory-dependent implicit regularisation problem, that captures the effects of hyperparameters of the algorithm, such as stepsizes and batchsizes, in useful and analysable ways. In Section 8.5 we then leverage this crisp characterisation to

explain the influence of crucial parameters such as the stepsize and batch-size on the recovered solution. Importantly **our analysis shows a stark difference between the generalisation performances of GD and SGD for large stepsizes**, hence explaining the numerical results seen in Figure 8.1 for the sparse regression setting. Finally, in Section 8.6, we use our results to shed new light on the *Edge of Stability* (*EoS*) phenomenon [Cohen et al., 2021].

8.2.2 Related works

Implicit bias. The concept of implicit bias from optimization algorithm in neural networks has been studied extensively in the past few years, starting with early works of [Telgarsky, 2013, Neyshabur et al., 2014, Keskar et al., 2017, Soudry et al., 2018]. The theoretical results on implicit regularisation have been extended to multiplicative parametrisations [Gunasekar et al., 2017, 2018b], linear networks [Ji and Telgarsky, 2019], and homogeneous networks [Lyu and Li, 2020, Ji and Telgarsky, 2020, Chizat et al., 2019]. For regression loss on diagonal linear networks studied in this work, Woodworth et al. [2020a] demonstrate that the scale of the initialisation determines the type of solution obtained, with large initialisations yielding minimum ℓ_2 norm solutions—the neural tangent kernel regime [Jacot et al., 2018] and small initialisation resulting in minimum ℓ_1 norm solutions—the *rich regime* [Chizat et al., 2019]. The analysis relies on the link between gradient descent and mirror descent established by Ghai et al. [2020] and further explored by Vaskevicius et al. [2020a], Wu and Rebeschini [2020]. These works focus on full batch gradient, and often in the inifitesimal stepsize limit (gradient flow), leading to general insights and results that do not take into account the effects of stochasticity and large stepsizes.

The effect of stochasticity in SGD on generalisation. The relationship between stochasticity in SGD and generalisation has been studied in various works [Mandt et al., 2016, Hoffer et al., 2017, Chaudhari and Soatto, 2018, Kleinberg et al., 2018, Wu et al., 2018]. Empirically, models generated by SGD exhibit better generalisation performance than those generated by GD [Keskar et al., 2017, Jastrzebski et al., 2019, He et al., 2019]. Explanations related to the flatness of the minima picked by SGD have been proposed [Hochreiter and Schmidhuber, 1997]. Label noise has been shown to influence the implicit bias of SGD [HaoChen et al., 2021, Blanc et al., 2020, Damian et al., 2021, Pillaud-Vivien et al., 2022] by implicitly regularising the sharp minimisers. Recently, studying a *stochastic gradient flow* that models the noise of SGD in continuous time with Brownian diffusion, Pesme et al. [2021] characterised for diagonal linear networks the limit of their stochastic process as the solution of an implicit regularisation problem. However similar explicit characterisation of the implicit bias remains unclear for SGD with large stepsizes.

The effect of stepsizes in GD and SGD. Recent efforts to understand how the choice of stepsizes affects the learning process and the properties of the recovered solution suggest that larger stepsizes lead to the minimisation of some notion of flatness of the loss function [Smith and Le, 2018, Keskar et al., 2017, Nacson et al., 2022, Jastrzkebski et al., 2018, Wu et al., 2018, Mulayoff et al., 2021], backed by empirical evidences or stability analyses. Larger stepsizes have also been proven to be beneficial for specific architectures or problems: two-layer network [Li et al., 2019b], regression [Wu et al., 2021], kernel regression [Beugnot et al., 2022] or matrix factorisation [Wang et al., 2022b]. For large stepsizes, it has been observed that GD enters an *Edge of Stability (EoS)* regime [Jastrzebski et al., 2019, Cohen et al., 2021], in which the iterates and the train loss oscillate before converging to a zero-training error solution; this phenomenon has then been studied on simple toy models [Ahn et al., 2022, Zhu et al., 2023, Chen and Bruna, 2022, Damian et al., 2023] for GD. Recently, Andriushchenko et al. [2022] presented empirical evidence that large stepsizes can lead to loss stabilisation and towards simpler predictors.

8.3 Setup and preliminaries

Overparametrised linear regression. We consider a linear regression over inputs $X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ and outputs $y = (y_1, \dots, y_n) \in \mathbb{R}^n$. We consider *overparametrised* problems where input dimension d is (much) larger than the number of samples n. In this case, there exists infinitely many linear predictors $\beta^\star \in \mathbb{R}^d$ which perfectly fit the training set, *i.e.*, $y_i = \langle \beta^\star, x_i \rangle$ for all $1 \leq i \leq n$. We call such vectors *interpolating predictors* or *interpolators* and we denote by $\mathcal{S}$ the set of all interpolators $\mathcal{S} = \{\beta^\star \in \mathbb{R}^d \text{ s.t. } \langle \beta^\star, x_i \rangle = y_i, \forall i \in [n]\}$. Note that $\mathcal{S}$ is an affine space of dimension greater than $d-n$ and equal to $\beta^\star + \operatorname{span}(x_1, \dots, x_n)^\perp$ for any $\beta^\star \in \mathcal{S}$. We consider the following quadratic loss: $\mathcal{L}(\beta) = \frac{1}{2n} \sum_{i=1}^n (\langle \beta, x_i \rangle - y_i)^2$, for $\beta \in \mathbb{R}^d$.

2-layer linear diagonal network. We parametrise regression vectors β as functions β_w of trainable parameters $w \in \mathbb{R}^p$. Although the final prediction function $x \mapsto \langle \beta_w, x \rangle$ is linear in the input x, the choice of the parametrisation drastically changes the solution recovered by the optimisation algorithm [Gunasekar et al., 2018b]. In the case of the linear parametrisation $\beta_w = w$ many first-order methods (SGD, GD, with or without momentum) converge towards the same solution and the choice of stepsize does not impact the recovered solution beyond convergence. In an effort to better understand the effects of stochasticity and large stepsize, we consider the next simple parametrisation, that of a 2-layer diagonal linear neural network given by:

$$\beta_w = u \odot v \text{ where } w = (u, v) \in \mathbb{R}^{2d} . \tag{8.1}$$

This parametrisation can be viewed as a simple neural network $x \mapsto \langle u, \sigma(\operatorname{diag}(v)x) \rangle$ where the output weights are represented by u, the inner weights is the diagonal matrix $\operatorname{diag}(v)$, and the activation σ is the identity function. In this spirit, we refer to the entries of $w = (u, v) \in \mathbb{R}^{2d}$ as the *weights* and to $\beta \coloneqq u \odot v \in \mathbb{R}^d$ as the *prediction parameter*. Despite the simplicity of the parametrisation (8.1), the loss function F over parameters $w = (u, v) \in \mathbb{R}^{2d}$ is **non-convex** (and thus the corresponding optimization problem is challenging to analyse), and is given by:

$$F(w) \coloneqq \mathcal{L}(u \odot v) = \frac{1}{2n} \sum_{i=1}^n (y_i - \langle u \odot v, x_i \rangle)^2 . \tag{8.2}$$

Mini-batch SGD. We minimise F using mini-batch SGD: let $w_0 = (u_0, v_0)$ and for $k \geq 0$,

$$w_{k+1} = w_k - \gamma_k \nabla F_{\mathcal{B}_k}(w_k) , \quad \text{where} \quad F_{\mathcal{B}_k}(w) \coloneqq \frac{1}{2b} \sum_{i \in \mathcal{B}_k} (y_i - \langle u \odot v, x_i \rangle)^2 , \tag{8.3}$$

where γ_k are stepsizes, $\mathcal{B}_k \subset [n]$ are mini-batches of $b \in [n]$ distinct samples sampled uniformly and independently, and $\nabla F_{\mathcal{B}_k}(w_k)$ are minibatch gradients of partial loss over $\mathcal{B}_k$, $F_{\mathcal{B}_k}(w) \coloneqq \mathcal{L}_{\mathcal{B}_k}(u \odot v)$ defined above. Classical SGD and full-batch GD are special cases with $b = 1$ and $b = n$, respectively. For $k \geq 0$, we consider the successive prediction parameters $\beta_k \coloneqq u_k \odot v_k$ built from the weights $w_k = (u_k, v_k)$. We analyse SGD initialised at $u_0 = \sqrt{2}\boldsymbol{\alpha} \in \mathbb{R}^d_{>0}$ and $v_0 = \mathbf{0} \in \mathbb{R}^d$, resulting in $\beta_0 = \mathbf{0} \in \mathbb{R}^d$ independently of the chosen weight initialisation $\boldsymbol{\alpha}$[1].

[1] In Appendix C.3, we show that the (S)GD trajectory with this initialisation exactly matches that of another common parametrisation $\beta_w = w_+^2 - w_-^2$ with initialisation $w_{+,0} = w_{-,0} = \boldsymbol{\alpha}$. The second layer of our diagonal linear network is set to 0 in order to obtain results that are easier to interpret. However, our proof techniques can be applied directly to a general initialisation, at the cost of additional notations in our Theorems.

Experimental details. We consider the noiseless sparse regression setting where $(x_i)_{i\in[n]} \sim \mathcal{N}(0, I_d)$ and $y_i = \langle \beta^\star_{\ell_1}, x_i \rangle$ for some s-sparse vector $\beta^\star_{\ell_1}$. We perform (S)GD over the DLN with a uniform initialisation $\boldsymbol{\alpha} = \alpha \mathbf{1} \in \mathbb{R}^d$ where $\alpha > 0$. Figure 8.1 and Figure 8.2 (left) correspond to the setup $(n, d, s, \alpha) = (20, 30, 3, 0.1)$, Figure 8.2 (right) to $(n, d, s, \alpha) = (50, 100, 4, 0.1)$ and Figure 8.3 to $(n, d, s, \alpha) = (50, 100, 2, 0.1)$.

Notations. Let $H \coloneqq \nabla^2 \mathcal{L} = \frac{1}{n}\sum_i x_i x_i^\top$ denote the Hessian of $\mathcal{L}$, and for a batch $\mathcal{B} \subset [n]$ let $H_\mathcal{B} \coloneqq \nabla^2 \mathcal{L}_\mathcal{B} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} x_i x_i^\top$ denote the Hessian of the partial loss over the batch $\mathcal{B}$. Let L denote the "smoothness" such that $\forall \beta$, $\|H_\mathcal{B}\beta\|_2 \leq L\|\beta\|_2$, $\|H_\mathcal{B}\beta\|_\infty \leq L\|\beta\|_\infty$ for all batches $\mathcal{B} \subset [n]$ of size b. A real function (e.g, $\log, \exp$) applied to a vector must be understood as element-wise application, and for vectors $u, v \in \mathbb{R}^d$, $u^2 = (u_i^2)_{i\in[d]}$, $u \odot v = (u_i v_i)_{i\in[d]}$ and $u/v = (u_i/v_i)_{i\in[d]}$. We write $\mathbf{1}$, $\mathbf{0}$ for the constant vectors with coordinates 1 and 0 respectively. The Bregman divergence [Bregman, 1967] of a differentiable convex function $h : \mathbb{R}^d \to \mathbb{R}$ is defined as $D_h(\beta_1, \beta_2) = h(\beta_1) - (h(\beta_2) + \langle \nabla h(\beta_2), \beta_1 - \beta_2 \rangle)$.

8.4 Implicit bias of SGD and GD

We start by recalling some known results on the implicit bias of gradient flow on diagonal linear networks before presenting our main theorems on characterising the (stochastic) gradient descent solutions (theorem 1) as well as proving the convergence of the iterates (theorem 2).

8.4.1 Warmup: gradient flow

We first review prior findings on gradient flow on diagonal linear neural networks. Woodworth et al. [2020a] show that the limit $\beta^\star_{\boldsymbol{\alpha}}$ of the *gradient flow* $\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t$ initialised at $(u_0, v_0) = (\sqrt{2}\boldsymbol{\alpha}, \mathbf{0})$ is the solution of the minimal interpolation problem:

$$\beta^\star_{\boldsymbol{\alpha}} = \operatorname*{arg\,min}_{\beta^\star \in \mathcal{S}} \ \psi_{\boldsymbol{\alpha}}(\beta^\star)\,, \quad \text{where} \quad \psi_{\boldsymbol{\alpha}}(\beta) = \frac{1}{2}\sum_{i=1}^{d}\left(\beta_i \operatorname{arcsinh}\left(\frac{\beta_i}{\alpha_i^2}\right) - \sqrt{\beta_i^2 + \alpha_i^4} + \alpha_i^2\right). \tag{8.4}$$

The convex potential $\psi_{\boldsymbol{\alpha}}$ is the **hyperbolic entropy function** (or **hypentropy**) [Ghai et al., 2020]. Depending on the structure of the vector $\boldsymbol{\alpha}$, the generalisation properties of $\beta^\star_{\boldsymbol{\alpha}}$ highly vary. We point out the two main characteristics of $\boldsymbol{\alpha}$ that affect the behaviour of $\psi_{\boldsymbol{\alpha}}$ and therefore also the solution $\beta^\star_{\boldsymbol{\alpha}}$.

1. The **Scale** of $\boldsymbol{\alpha}$. For an initialisation vector $\boldsymbol{\alpha}$ we call the ℓ_1-norm $\|\boldsymbol{\alpha}\|_1$ the **scale** of the initialisation. It is an important quantity affecting the properties of the recovered solution $\beta^\star_{\boldsymbol{\alpha}}$. To see this let us consider a uniform initialisation of the form $\boldsymbol{\alpha} = \alpha\mathbf{1}$ for a scalar value $\alpha > 0$. In this case the potential $\psi_{\boldsymbol{\alpha}}$ has the property of resembling the ℓ_1-norm as the scale α vanishes: $\psi_{\boldsymbol{\alpha}} \sim \ln(1/\alpha)\|.\|_1$ as $\alpha \to 0$. Hence, a small initialisation results in a low ℓ_1-norm solution which is known to induce sparse recovery guarantees [Candès et al., 2006]. This setting is often referred to as the "rich" regime [Woodworth et al., 2020a]. In contrast, using a large initialisation scale leads to solutions with low ℓ_2-norm: $\psi_{\boldsymbol{\alpha}} \sim \|.\|_2^2/(2\alpha^2)$ as $\alpha \to \infty$, a setting known as the "kernel" or "lazy" regime. Overall, to retrieve the minimum ℓ_1-norm solution, one should use a uniform initialisation with small scale α, see Figure C.4 in Appendix C.4 for an illustration and [Woodworth et al., 2020a, Theorem 2] for a precise characterisation.

2. The **Shape** of $\boldsymbol{\alpha}$. In addition to the scale of the initialisation $\boldsymbol{\alpha}$, a lesser studied aspect is its "shape", which is a term we use to refer to the relative distribution of $\{\alpha_i\}_i$ along the d coordinates [Azulay et al., 2021]. It is a crucial property because having $\boldsymbol{\alpha} \to \mathbf{0}$ **does not** necessarily lead to the potential $\psi_{\boldsymbol{\alpha}}$ being close to the ℓ_1-norm. Indeed, we have that $\psi_{\boldsymbol{\alpha}}(\beta) \overset{\boldsymbol{\alpha}\to\mathbf{0}}{\sim}$

$\sum_{i=1}^{d} \ln(\frac{1}{\alpha_i})|\beta_i|$ (see Appendix C.4), therefore if the vector $\ln(1/\boldsymbol{\alpha})$ has entries changing at different rates, then $\psi_{\boldsymbol{\alpha}}(\beta)$ is a **weighted** ℓ_1-norm. In words, if the entries of $\boldsymbol{\alpha}$ *do not go to zero "uniformly"*, then the resulting implicit bias minimizes a weighed ℓ_1-norm. This phenomenon can lead to solutions with vastly different sparsity structure than the minimum ℓ_1-norm interpolator. See Figure C.4 and Example 1 in Appendix C.4.

8.4.2 Implicit bias of (stochastic) gradient descent

In theorem 1, we prove that for an initialisation $\sqrt{2}\boldsymbol{\alpha} \in \mathbb{R}^d$ and for **arbitrary** stepsize sequences $(\gamma_k)_{k\geq 0}$ **if the iterates converge to an interpolator**, then this interpolator is the solution of a constrained minimisation problem which involves the hyperbolic entropy $\psi_{\boldsymbol{\alpha}_\infty}$ defined in (8.4), where $\boldsymbol{\alpha}_\infty \in \mathbb{R}^d$ is an effective initialisation which depends on the trajectory and on the stepsize sequence. Later, **we prove the convergence of iterates for macroscopic step sizes** in theorem 2.

Theorem 1 (Implicit bias of (S)GD)**.** *Let $(u_k, v_k)_{k\geq 0}$ follow the mini-batch SGD recursion (8.3) initialised at $(u_0, v_0) = (\sqrt{2}\boldsymbol{\alpha}, \mathbf{0})$ and with stepsizes $(\gamma_k)_{k\geq 0}$. Let $(\beta_k)_{k\geq 0} = (u_k \odot v_k)_{k\geq 0}$ and assume that they converge to some interpolator $\beta^\star_\infty \in \mathcal{S}$. Then, $\beta^\star_\infty$ satisfies:*

$$\beta^\star_\infty = \underset{\beta^\star \in \mathcal{S}}{\arg\min}\ D_{\psi_{\boldsymbol{\alpha}_\infty}}(\beta^\star, \tilde{\beta}_0)\,, \tag{8.5}$$

*where $D_{\psi_{\boldsymbol{\alpha}_\infty}}$ is the Bregman divergence with hyperentropy potential $\psi_{\boldsymbol{\alpha}_\infty}$ of the **effective initialisation** $\boldsymbol{\alpha}_\infty$, and $\tilde{\beta}_0$ is a small **perturbation term**. The **effective initialisation** $\boldsymbol{\alpha}_\infty$ is given by,*

$$\boldsymbol{\alpha}^2_\infty = \boldsymbol{\alpha}^2 \odot \exp\left(-\sum_{k=0}^{\infty} q\big(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\big)\right)\,, \tag{8.6}$$

*where $q(x) = -\frac{1}{2}\ln((1-x^2)^2)$ satisfies $q(x) \geq 0$ for $|x| \leq \sqrt{2}$, with the convention $q(1) = +\infty$. The **perturbation term** $\tilde{\beta}_0 \in \mathbb{R}^d$ is explicitly given by $\tilde{\beta}_0 = \frac{1}{2}\big(\boldsymbol{\alpha}^2_+ - \boldsymbol{\alpha}^2_-\big)$, where $q_\pm(x) = \mp 2x - \ln((1\mp x)^2)$, and $\boldsymbol{\alpha}^2_\pm = \boldsymbol{\alpha}^2 \odot \exp\big(-\sum_{k=0}^{\infty} q_\pm(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))\big)$.*

Trajectory-dependent characterisation. The characterisation of $\beta^\star_\infty$ in Theorem 1 holds for any stepsize schedule such that the iterates converge and goes beyond the continuous-time frameworks previously studied [Woodworth et al., 2020a, Pesme et al., 2021]. The result even holds for adaptive stepsize schedules which keep the stepsize scalar such as AdaDelta [Zeiler, 2012]. An important aspect of our result is that $\boldsymbol{\alpha}_\infty$ and $\tilde{\beta}_0$ depend on the iterates' trajectory. Nevertheless, we argue that our formulation provides useful ingredients for understanding the implicit regularisation effects of (S)GD for this problem compared to trivial characterisations (such as *e.g.*, $\min_\beta \|\beta - \beta^\star_\infty\|$). Importantly, **the key parameters $\boldsymbol{\alpha}_\infty, \tilde{\beta}_0$ depend on crucial parameters such as the stepsize and noise in a useful and analysable manner**: understanding how they affect $\boldsymbol{\alpha}_\infty$ and $\tilde{\beta}_0$ coincides with understanding how they affect the recovered solution $\beta^\star_\infty$ and its generalisation properties. This is precisely the object of Sections 8.5 and 8.6 where we discuss the qualitative and quantitative insights from Theorem 1 in greater detail.

The perturbation $\tilde{\beta}_0$ can be ignored. We show in Proposition 43, under reasonable assumptions on the stepsizes, that $|\tilde{\beta}_0| \leq \boldsymbol{\alpha}^2$ and $\boldsymbol{\alpha}_\infty \leq \boldsymbol{\alpha}$ (component-wise). The magnitude of $\tilde{\beta}_0$ is therefore negligible in front of the magnitudes of $\beta^\star \in S$ and one can roughly ignore the term $\tilde{\beta}_0$. Hence, the implicit regularisation eq. (8.5) can be thought of as $\beta^\star_\infty \approx \arg\min_{\beta^\star \in S} D_{\psi_{\boldsymbol{\alpha}_\infty}}(\beta^\star, 0) = \psi_{\boldsymbol{\alpha}_\infty}(\beta^\star)$, and thus *the solution $\beta^\star_\infty$ minimises the same potential*

function that the solution of gradient flow (see Equation (8.4)), but with an effective initialisation $\boldsymbol{\alpha}_\infty$. Also note that for $\gamma_k \equiv \gamma \to 0$ we have $\boldsymbol{\alpha}_\infty \to \boldsymbol{\alpha}$ and $\tilde{\beta}_0 \to \mathbf{0}$ (proposition 46), recovering the previously known result for gradient flow (8.4).

Deviation from gradient flow. The difference with gradient flow is directly associated with the quantity $\sum_k q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$. Also, as the (stochastic) gradients converge to 0 and $q(x) \overset{x \to 0}{\sim} x^2$, one should think of this sum as roughly being $\sum_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)^2$: the larger this sum, the more the recovered solution differs from that of gradient flow. The full picture of how large stepsizes and stochasticity impact the generalisation properties of $\beta^\star_\infty$ and the recovery of minimum ℓ_1-norm solution is nuanced as clearly seen in fig. 8.1.

8.4.3 Convergence of the iterates

Theorem 1 provides the implicit minimisation problem but says nothing about the convergence of the iterates. Here we show under very reasonable assumptions on the stepsizes that the iterates indeed converge towards a global optimum. Note that since the loss F is non-convex, such a convergence result is non-trivial and requires an involved analysis.

Theorem 2 (Convergence of the iterates)**.** *Let* $(u_k, v_k)_{k\geq 0}$ *follow the mini-batch SGD recursion (8.3) initialised at* $u_0 = \sqrt{2}\boldsymbol{\alpha} \in \mathbb{R}^d_{>0}$ *and* $v_0 = \mathbf{0}$*, and let* $(\beta_k)_{k\geq 0} = (u_k \odot v_k)_{k\geq 0}$*. Recall the "smoothness" parameter* L *on the minibatch loss defined in the notations. There exist* $B > 0$ *verifying* $B = \tilde{\mathcal{O}}(\min_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_\infty)$ *and a numerical constant* $c > 0$ *such that for stepsizes satisfying* $\gamma_k \leq \frac{c}{LB}$*, the iterates* $(\beta_k)_{k\geq 0}$ *converge almost surely to the interpolator* $\beta^\star_\infty$ *solution of Equation (8.5).*

In fact, we can be more precise by showing an exponential rate of convergence of the losses as well as characterise the rate of convergence of the iterates as follows.

Proposition 17 (Quantitative convergence rates)**.** *For a uniform initialisation* $\boldsymbol{\alpha} = \alpha \mathbf{1}$ *and under the assumptions of Theorem 2, we have:*

$$\mathbb{E}\left[\mathcal{L}(\beta_k)\right] \leq \left(1 - \frac{1}{2}\gamma\alpha^2\lambda_b\right)^k \mathcal{L}(\beta_0) \quad \text{and} \quad \mathbb{E}\left[\|\beta_k - \beta^\star_{\alpha_k}\|^2\right] \leq C\left(1 - \frac{1}{2}\gamma\alpha^2\lambda_b\right)^k,$$

where $\lambda_b > 0$ *is the largest value such that* $\lambda_b H \preceq \mathbb{E}_{\mathcal{B}}[H_{\mathcal{B}}]$, $C = 2B(\alpha^2\lambda^+_{\min})^{-1}\left(1 + (4B\lambda_{\max})(\alpha^2\lambda^+_{\min})^{-1}\right)\mathcal{L}(\beta_0)$ *and* $\lambda^+_{\min}, \lambda_{\max} > 0$ *are respectively the smallest non-null and the largest eigenevalues of* H*, and* $\beta^\star_{\alpha_k}$ *is the interpolator that minimises the perturbed hypentropy* h_k *of parameter* α_k*, as defined in Equation (8.7) in the next subsection.*

The convergence of the losses is proved directly using the time-varying mirror structure that we exhibit in the next subsection, the convergence of the iterates is proved by studying the curvature of the mirror maps on a small neighborhood around the affine interpolation space.

8.4.4 Sketch of proof through a time varying mirror descent

As in the continuous-time framework, our results heavily rely on showing that the iterates $(\beta_k)_k$ follow a mirror descent recursion with time-varying potentials on the convex loss $\mathcal{L}(\beta)$. To show this, we first define the following quantities:

$$\boldsymbol{\alpha}_k^2 := \boldsymbol{\alpha}_{+,k} \odot \boldsymbol{\alpha}_{-,k} \qquad \text{and} \qquad \phi_k := \frac{1}{2}\operatorname{arcsinh}\left(\frac{\boldsymbol{\alpha}^2_{+,k} - \boldsymbol{\alpha}^2_{-,k}}{2\boldsymbol{\alpha}_k^2}\right) \in \mathbb{R}^d,$$

where $\boldsymbol{\alpha}_{\pm,k} \coloneqq \boldsymbol{\alpha} \exp\left(-\frac{1}{2}\sum_{i=0}^{k-1} q_{\pm}\big(\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\big)\right) \in \mathbb{R}^d$. Finally for $k \geq 0$, we define the potentials $(h_k : \mathbb{R}^d \to \mathbb{R})_{k\geq 0}$ as:

$$h_k(\beta) = \psi_{\boldsymbol{\alpha}_k}(\beta) - \langle \phi_k, \beta \rangle. \tag{8.7}$$

Where $\psi_{\boldsymbol{\alpha}_k}$ is the hyperbolic entropy function defined Equation (8.4). Now that all the relevant quantities are defined, we can state the following proposition which explicits the time-varying stochastic mirror descent.

Proposition 18. *The iterates $(\beta_k = u_k \odot v_k)_{k\geq 0}$ from Equation (8.3) satisfy the Stochastic Mirror Descent recursion with varying potentials $(h_k)_k$:*

$$\nabla h_{k+1}(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,,$$

where $h_k : \mathbb{R}^d \to \mathbb{R}$ for $k \geq 0$ are defined Equation (8.7). Since $\nabla h_0(\beta_0) = 0$ we have:

$$\nabla h_k(\beta_k) \in \operatorname{span}(x_1, \dots, x_n). \tag{8.8}$$

Theorems 1 and 2 and proposition 17 follow from this key proposition: by suitably modifying classical convex optimization techniques to account for the time-varying potentials, we can prove the convergence of the iterates towards an interpolator $\beta^\star_\infty$ along with that of the relevant quantities $\boldsymbol{\alpha}_{\pm,k}$, $\boldsymbol{\alpha}_k$ and ϕ_k. The implicit regularisation problem then directly follows from: (1) the limit condition $\nabla h_\infty(\beta_\infty) \in \operatorname{span}(x_1, \dots, x_n)$ as seen from eq. (8.8) and (2) the interpolation condition $X\beta^\star_\infty = y$. Indeed, these two conditions exactly correspond to the KKT conditions of the convex problem eq. (8.5).

8.5 Analysis of the impact of the stepsize and stochasticity on α_∞

In this section, we analyse the effects of large stepsizes and stochasticity on the implicit bias of (S)GD. We focus on how these factors influence the effective initialisation $\boldsymbol{\alpha}_\infty$, which plays a key role as shown in Theorem 1. From its definition in eq. (8.6), we see that $\boldsymbol{\alpha}_\infty$ is a function of the vector $\sum_k q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$. We henceforth call this quantity the *gain vector*. For simplicity of the discussions, from now on, we consider constant stepsizes $\gamma_k = \gamma$ for all $k \geq 0$ and a uniform initialisation of the weights $\boldsymbol{\alpha} = \alpha \mathbf{1}$ with $\alpha > 0$. We can then write the gain vector as:

$$\mathrm{Gain}_\gamma \coloneqq \ln\left(\frac{\boldsymbol{\alpha}^2}{\boldsymbol{\alpha}_\infty^2}\right) = \sum_k q(\gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)) \in \mathbb{R}^d\,.$$

Following our discussion in section 8.4.1 on the scale and the shape of $\boldsymbol{\alpha}_\infty$, we recall the link between the scale and shape of Gain_γ and the recovered solution:

1. The **scale** of Gain_γ, i.e. the magnitude of $\|\mathrm{Gain}_\gamma\|_1$ indicates how much the implicit bias of (S)GD differs from that of gradient flow: $\|\mathrm{Gain}_\gamma\|_1 \sim 0$ implies that $\boldsymbol{\alpha}_\infty \sim \boldsymbol{\alpha}$ and therefore the recovered solution is close to that of gradient flow. On the contrary, $\|\mathrm{Gain}_\gamma\|_1 \gg \ln(1/\alpha)$ implies that $\boldsymbol{\alpha}_\infty$ has effective scale much smaller than $\boldsymbol{\alpha}$ thereby changing the implicit regularisation eq. (8.5).

2. The **shape** of Gain_γ indicates which coordinates of β in the associated minimum weighted ℓ_1 problem are most penalised. First recall from Section 8.4.1 that a uniformly large Gain_γ leads to $\psi_{\boldsymbol{\alpha}_\infty}$ being closer to the ℓ_1-norm. However, with small weight initialisation $\alpha \to 0$, we have,

$$\psi_{\boldsymbol{\alpha}_\infty}(\beta) \sim \ln(\frac{1}{\alpha})\|\beta\|_1 + \sum_{i=1}^{d} \mathrm{Gain}_\gamma(\mathrm{i})|\beta_\mathrm{i}|\,, \tag{8.9}$$

In this case, having a heterogeneously large vector Gain_γ leads to a weighted ℓ_1 norm as the effective implicit regularisation, where the coordinates of β corresponding to the largest entries of Gain_γ are less likely to be recovered.

8.5.1 The scale of Gain_γ is increasing with the stepsize

The following proposition highlights the dependencies of the scale of the gain $\|\mathrm{Gain}_\gamma\|_1$ in terms of various problem constants. Let $\Lambda_b, \lambda_b > 0$ [2] be the largest and smallest values, respectively, such that $\lambda_b H \preceq \mathbb{E}_{\mathcal{B}}\left[H_{\mathcal{B}}^2\right] \preceq \Lambda_b H$. For any stepsize $\gamma > 0$ satisfying $\gamma \leq \frac{c}{BL}$ (as in theorem 2), initialisation $\alpha\mathbf{1}$ and batch size $b \in [n]$, the magnitude of the gain satisfies:

$$\lambda_b \gamma^2 \sum_k \mathbb{E}\mathcal{L}(\beta_k) \leq \mathbb{E}\left[\|\mathrm{Gain}_\gamma\|_1\right] \leq 2\Lambda_b \gamma^2 \sum_k \mathbb{E}\mathcal{L}(\beta_k)\,, \tag{8.10}$$

where the expectation is over a uniform and independent sampling of the batches $(\mathcal{B}_k)_{k\geq 0}$.

The slower the training, the larger the gain. eq. (8.10) shows that the slower the training loss converges to 0, the larger the sum of the loss and therefore the larger the scale of Gain_γ. This means that the (S)GD trajectory deviates from that of gradient flow if the stepsize and/or noise slows down the training. This supports observations previously made from stochastic gradient flow [Pesme et al., 2021] analysis.

The bigger the stepsize, the larger the gain. The effect of the stepsize on the magnitude of the gain is not directly visible in eq. (8.10) because a larger stepsize tends to speed up the training. For stepsize $0 < \gamma \leq \gamma_{\max} = \frac{c}{BL}$ as in Theorem 2 we have that (see Appendix C.7.1):

$$\sum_k \gamma^2 \mathcal{L}(\beta_k) = \Theta\left(\gamma \ln\left(\frac{1}{\alpha}\right) \|\beta^\star_{\ell_1}\|_1\right)\,. \tag{8.11}$$

eq. (8.11) clearly shows that increasing the stepsize **boosts** the magnitude $\|\mathrm{Gain}_\gamma\|_1$ up until the limit of $\gamma_{\max}$. Therefore, the larger the stepsize the smaller is the effective scale of $\boldsymbol{\alpha}_\infty$. In turn, larger gap between $\boldsymbol{\alpha}_\infty$ and $\boldsymbol{\alpha}$ leads to a larger deviation of (S)GD from the gradient flow.

Large stepsizes and Edge of Stability. The previous paragraph holds for stepsizes smaller than $\gamma_{\max}$ for which we can theoretically prove convergence. But what if we use even bigger stepsizes? Let $(\beta_k^\gamma)_k$ denote the iterates generated with stepsize γ and let us define $\tilde{\gamma}_{\max} \coloneqq \sup_{\gamma\geq 0}\{\gamma \text{ s.t. } \forall \gamma' \in (0,\gamma),\ \sum_k \mathcal{L}(\beta_k^{\gamma'}) < \infty\}$, which corresponds to the largest stepsize such that the iterates still converge for a given problem (even if not provably so). From Section 8.5.1 we have that $\gamma_{\max} \leq \tilde{\gamma}_{\max}$. As we approach this upper bound on convergence $\gamma \to \tilde{\gamma}_{\max}$, the sum $\sum_k \mathcal{L}(\beta_k^\gamma)$ diverges. For such large stepsizes, the iterates of gradient descent tend to "bounce" and this regime is commonly referred to as the *Edge of Stability*. In this regime, the convergence of the loss can be made arbitrarily slow due to these bouncing effects. As a consequence, as seen through Equation (8.10), the magnitude of Gain_γ can be become arbitrarily big as observed in fig. 8.2 (left). In this regime, the recovered solution tends to dramatically differ from the gradient flow solution, as seen in fig. 8.1.

Impact of stochasticity and linear scaling rule. Assuming inputs x_i sampled from $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma^2 > 0$, we obtain $\mathbb{E}\left[\|\mathrm{Gain}_\gamma\|_1\right] = \Theta\Big(\gamma\frac{\sigma^2 d}{b} \ln\left(\frac{1}{\alpha}\right)\|\beta^\star_{\ell_1}\|_1\Big)$, w.h.p. over the dataset (see Appendix C.7.3, proposition 44). The scale of Gain_γ decreases with batch size and there exists a factor n between that of SGD and that of GD. Additionally, the magnitude of Gain_γ depends on $\frac{\gamma}{b}$, resembling the **linear scaling rule** commonly used in deep learning [Goyal et al., 2017].

By analysing the magnitude $\|\mathrm{Gain}_\gamma\|_1$, we have explained **the distinct behavior of (S)GD with large stepsizes compared to gradient flow**. However, our current analysis does not

[2] $\Lambda_b, \lambda_b > 0$ are data-dependent constants; for $b = n$, we have $(\lambda_n, \Lambda_n) = (\lambda^+_{\min}(H), \lambda_{\max}(H))$ where $\lambda^+_{\min}(H)$ is the smallest non-null eigenvalue of H; for $b = 1$, we have $\min_i \|x_i\|_2^2 \leq \lambda_1 \leq \Lambda_1 \leq \max_i \|x_i\|_2^2$.

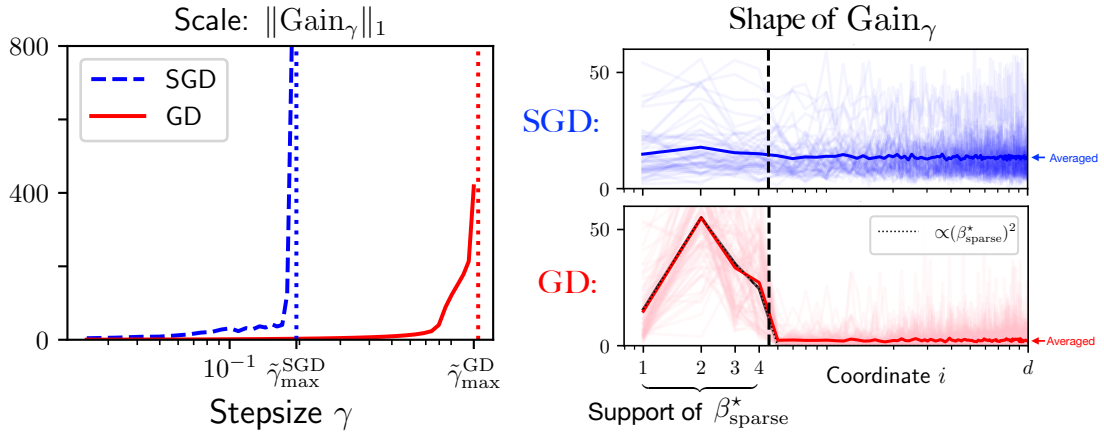


Figure 8.2: *Left:* the scale of Gain_γ explodes as $\gamma \to \tilde{\gamma}_{\max}$ for both GD and SGD. *Right:* $\beta^\star_{\text{sparse}}$ is fixed, we perform 100 runs of GD and SGD with different feature matrices, and we plot the d coordinates of Gain_γ (for GD and SGD) on the x-axis (which is in log scale for better visualisation). The shape of $\text{Gain}^{\text{SGD}}_\gamma$ is homogeneous whereas that of GD is heterogeneous with much higher magnitude on the support of $\beta^\star_{\text{sparse}}$. The shape of $\text{Gain}^{\text{GD}}_\gamma$ is proportional to the expected gradient at initialisation which is $(\beta^\star_{\text{sparse}})^2$.

qualitatively distinguish the behavior between SGD and GD beyond the linear stepsize scaling rules, in contrast with fig. 8.1. A deeper understanding of the shape of Gainγ is needed to explain this disparity.

8.5.2 The shape of Gain_γ explains the differences between GD and SGD

In this section, we restrict our presentation to single batch SGD ($b = 1$) and full batch GD ($b = n$). When visualising the typical shape of Gain_γ for large stepsizes (see Figure 8.2 - right), we note that GD and SGD behave very differently. For GD, the magnitude of Gain_γ is higher for coordinates in the support of $\beta^\star_{\ell_1}$ and thus these coordinates are adversely weighted in the asymptotic limit of $\psi_{\boldsymbol{\alpha}_\infty}$ (per (8.9)). This explains the distinction seed in fig. 8.1, where GD in this regime has poor sparse recovery despite having a small scale of $\boldsymbol{\alpha}_\infty$, as opposed to SGD that behaves well.

The **shape** of Gain_γ is determined by the sum of the squared gradients $\sum_k \nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)^2$, and in particular by the degree of heterogeneity among the coordinates of this sum. Precisely analysing the sum over the whole trajectory of the iterates $(\beta_k)_k$ is technically out of reach. However, we empirically observe for the trajectories shown in Figure 8.2 that the shape is largely determined within the first few iterates as formalized in the observation below.

Observation 1. $\sum_k \nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)^2 \eqsim \mathbb{E}[\nabla\mathcal{L}_{\mathcal{B}_k}(\beta_0)^2]$.

In the simple case of a Gaussian noiseless sparse recovery problem (where $y_i = \langle\beta^\star_{\text{sparse}}, x_i\rangle$ for some sparse vector $\beta^\star_{\text{sparse}}$), we can control these gradients for GD and SGD (Appendix C.7.4) as:

$$\nabla\mathcal{L}(\beta_0)^2 = (\beta^\star_{\text{sparse}})^2 + \varepsilon\,, \text{ for some } \varepsilon \text{ verifying } \|\varepsilon\|_\infty << \|\beta^\star_{\text{sparse}}\|^2_\infty\,, \tag{8.12}$$

$$\mathbb{E}_{i_0}[\nabla\mathcal{L}_{i_0}(\beta_0)^2] = \Theta\Big(\|\beta^\star_{\text{sparse}}\|_2^2\mathbf{1}\Big)\,. \tag{8.13}$$

The gradient of GD is heterogeneous. Since $\beta^\star_{\text{sparse}}$ is sparse by definition, we deduce from eq. (C.12) that $\nabla\mathcal{L}(\beta_0)$ is heterogeneous with larger values corresponding to the support

of $\beta^\star_{\text{sparse}}$. Along with observation 1, this means that Gain_γ **has much larger values on the support of** $\beta^\star_{\text{sparse}}$. The corresponding weighted ℓ_1-norm therefore penalises the coordinates belonging to the support of $\beta^\star_{\text{sparse}}$, which hinders the recovery of $\beta^\star_{\text{sparse}}$ (as explained in Example 1, Appendix C.4).

The stochastic gradient of SGD is homogeneous. On the contrary, from eq. (C.13), we have that the initial stochastic gradients are homogeneous, leading to a weighted ℓ_1-norm where the weights are roughly balanced. The corresponding weighted ℓ_1-norm is therefore close to the uniform ℓ_1-norm and the classical ℓ_1 recovery guarantees are expected.

Overall summary of the joint effects of the scale and shape. In summary we have the following trichotomy which fully explains Figure 8.1:

1. for small stepsizes, the scale is small, and (S)GD solutions are close to that of gradient flow;
2. for large stepsizes the scale is significant and the recovered solutions differ from GF:
   - for SGD the shape of $\boldsymbol{\alpha}_\infty$ is uniform, the associated norm is closer to the ℓ_1-norm and the recovered solution is closer to the sparse solution;
   - for GD, the shape is heterogeneous, the associated norm is weighted such that it hinders the recovery of the sparse solution.

In this last section, we relate heuristically these findings to the *Edge of Stability* phenomenon.

8.6 Edge of stability: the neural point of view

In recent years it has been noticed that when training neural networks with 'large' stepsizes at the limit of divergence, GD enters the *Edge of Stability (EoS)* regime. In this regime, as seen in Figure 8.3, the iterates of GD 'bounce' / 'oscillate'. In this section, we come back to the point of view of the weights $w_k = (u_k, v_k) \in \mathbb{R}^{2d}$ and make the connection between our previous results and the common understanding of the *EoS* phenomenon. The question we seek to answer is: in which case does GD enter the *EoS* regime, and if so, what are the consequences on the trajectory? *Keep in mind that this section aims to provide insights rather than formal statements.* We study the GD trajectory starting from a small initialisation $\boldsymbol{\alpha} = \alpha\mathbf{1}$ where $\alpha \ll 1$ such that we can consider that gradient flow converges close to the sparse interpolator $\beta^\star_{\text{sparse}} = \beta_{w^\star_{\text{sparse}}}$ corresponding to the weights $w^\star_{\text{sparse}} = (\sqrt{|\beta^\star_{\text{sparse}}|}, \text{sign}(\beta^\star_{\text{sparse}})\sqrt{|\beta^\star_{\text{sparse}}|})$ (see Lemma 1 in Pesme and Flammarion [2023] for the mapping from the predictors to weights for gradient flow). The trajectory of GD as seen in fig. 8.3 (left) can be decomposed into up to 3 phases.

First phase: gradient flow. The stepsize is appropriate for the local curvature (as seen in Figure 8.3, lower right) around initialisation and the iterates of GD remain close to the trajectory of gradient flow (in black in fig. 8.3). If the stepsize is such that $\gamma < \frac{2}{\lambda_{\max}(\nabla^2 F(w^\star_{\text{sparse}}))}$, then it is compatible with the local curvature and the iterates can converge: in this case GF and GD converge to the same point (as seen in fig. 8.1 for small stepsizes). For larger $\gamma > \frac{2}{\lambda_{\max}(\nabla^2 F(w^\star_{\text{sparse}}))}$ (as is the case for γ_{GD} in fig. 8.3, lower right), the iterates cannot converge to $\beta^\star_{\text{sparse}}$ and we enter the oscillating phase.

Second phase: oscillations. The iterates start oscillating. The gradient of F writes $\nabla_{(u,v)} F(w) \sim (\nabla\mathcal{L}(\beta) \odot v, \nabla\mathcal{L}(\beta) \odot u)$ and for w in the vicinity of $w^\star_{\text{sparse}}$ we have that $u_i \approx v_i \approx 0$ for

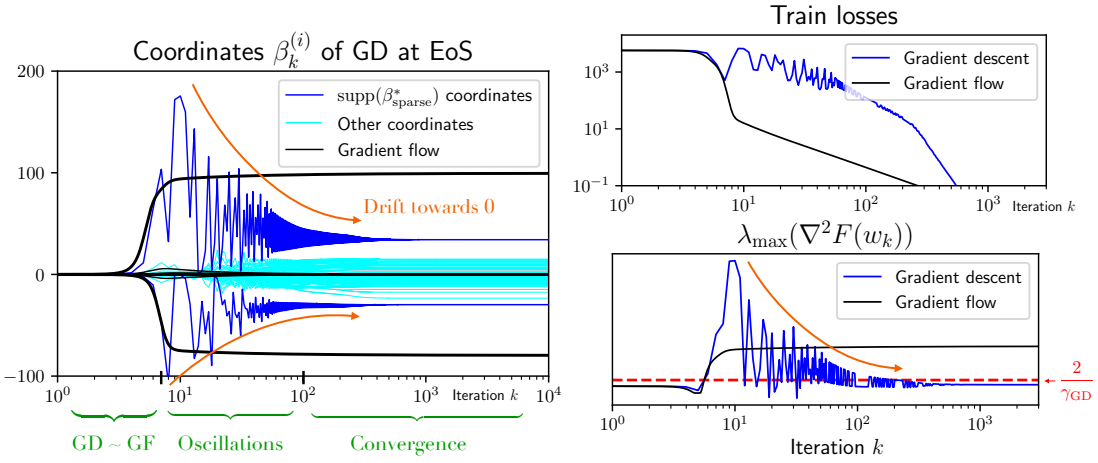


Figure 8.3: GD at the *EoS*. *Left:* For GD, the coordinates on the support of $\beta^\star_{\text{sparse}}$ oscillate and drift towards 0. *Right, top:* The GD train losses saturate before eventually converging. *Bottom:* GF converges towards a solution that has a high hessian maximum eigenvalue. GD cannot converge towards this solution because of its large stepsize: it therefore drifts towards a solution that has a curvature just below $2/\gamma$.

.

$i \notin \text{supp}(\beta^\star_{\text{sparse}})$. Therefore for $w \sim w^\star_{\text{sparse}}$ we have that $\nabla_u F(w)_i \approx \nabla_v F(w)_i \approx 0$ for $i \notin \text{supp}(\beta^\star_{\text{sparse}})$ and the gradients roughly belong to $\text{span}(e_i, e_{i+d})_{i \in \text{supp}(\beta^\star_{\text{sparse}})}$. This means that only the coordinates of the weights (u_i, v_i) for $i \in \text{supp}(\beta^\star_{\text{sparse}})$ can oscillate and similarly for $(\beta_i)_{i \in \text{supp}(\beta^\star_{\text{sparse}})}$ (as seen Figure 8.3 left).

Last phase: convergence. Due to the oscillations, the iterates gradually drift towards a region of lower curvature (fig. 8.3, lower right, the sharpness decreases) where they may (potentially) converge. theorem 1 enables us to understand where they converge: the coordinates of β_k that have oscillated significantly along the trajectory belong to the support of $\beta^\star_{\text{sparse}}$, and therefore $\text{Gain}_\gamma(\text{i})$ becomes much larger for $i \in \text{supp}(\beta^\star_{\text{sparse}})$ than for the other coordinates. Thus, the coordinates of the solution recovered in the *EoS* regime are heavily penalised on the support of the sparse solution. This is observed in Figure 8.3 (left): the oscillations of $(\beta_i)_{i \in \text{supp}(\beta^\star_{\text{sparse}})}$ lead to a gradual shift of these coordinates towards 0, hindering an accurate recovery of the solution $\beta^\star_{\text{sparse}}$.

SGD in the *EoS* regime. In contrast to the behavior of GD where the oscillations primarily occur on the non-sparse coordinates of ground truth sparse model, for SGD we see a different behavior in Figure C.3 (Appendix C.1). For stepsizes in the *EoS* regime, just below the non-convergence threshold: the fluctuation of the coordinates occurs evenly over all coordinates, leading to a uniform $\boldsymbol{\alpha}_\infty$. These fluctuations are reminiscent of label-noise SGD [Andriushchenko et al., 2022], that have been shown to recover the sparse interpolator in diagonal linear networks [Pillaud-Vivien et al., 2022].

8.7 Conclusion

We study the effect of stochasticity along with large stepsizes when training DLNs with (S)GD. We prove convergence of the iterates as well as explicitly characterise the recovered solution by exhibiting an implicit regularisation problem which depends on the iterates' trajectory. In essence the impact of stepsize and minibatch size are captured by the effective initialisation

parameter $\boldsymbol{\alpha}_\infty$ that depends on these choices in an informative way. We then use our characterisation to explain key empirical differences between SGD and GD and provide further insights on the role of stepsize and stochasticity. In particular, our characterisation explains the fundamentally different generalisation properties of SGD and GD solutions at large stepsizes as seen in Figure 8.1: without stochasticity, the use of large stepsizes can prevent the recovery of the sparse interpolator, even though the effective scale of the initialization decreases with larger stepsize for both SGD and GD. We also provide insights on the link between the *Edge of Stability* regime and our results.

Chapter 9

Effect of momentum

9.1 Preface

This chapter follows Papazov et al. [2024].

Summary We investigate the effect of momentum on the optimisation trajectory of gradient descent. We leverage a continuous-time approach in the analysis of momentum gradient descent with step size γ and momentum parameter β that allows us to identify an intrinsic quantity $\lambda = \frac{\gamma}{(1-\beta)^2}$ which uniquely defines the optimisation path and provides a simple acceleration rule. When training a 2-layer diagonal linear network in an overparametrised regression setting, we characterise the recovered solution through an implicit regularisation problem. We then prove that small values of λ help to recover sparse solutions. Finally, we give similar but weaker results for stochastic momentum gradient descent. We provide numerical experiments which support our claims.

Co-authors Hristo Papazov and Nicolas Flammarion.

Contributions Hristo and Scott worked together on the project.

9.2 Introduction

Momentum methods [Sutskever et al., 2013] have now become a staple of optimal neural network training due to the provided gains in both optimisation efficiency and generalisation performance. This pivotal role is underscored by the widespread use of momentum in the successful training of most state-of-the-art deep networks, including CLIP [Radford et al., 2021], Chinchilla [Hoffmann et al., 2022], GPT-3 [Brown et al., 2020], and PaLM [Chowdhery et al., 2022].

Originating in the work of Polyak [1964], momentum first featured in the heavy-ball method devised to accelerate convergence in convex optimisation. However, when applied to neural network training, momentum exhibits a distinct and complementary characteristic: a steering towards models with superior generalisation performance compared to networks trained with gradient descent. We note that while the effect of momentum on optimisation has been researched extensively [Defazio, 2020, Sun et al., 2019], the generalisation aspect of momentum has been left relatively underexplored.

The performance of gradient descent methods presents intriguing challenges from a theoretical perspective. First, establishing convergence is highly non-trivial. Second, the existence of numerous global minima for the training objective, some of which generalise poorly, adds to the puzzle [Zhang et al., 2017]. To elucidate this second point, the notion of implicit regularisation has come to the forefront. It posits that the optimisation process implicitly favors solutions with strong generalisation properties, even in the absence of explicit regularisation. The canonical example is overparametrised linear regression with more trainable parameters than the number

of samples. While there exist infinitely many solutions that fit the data, gradient methods navigate in a restricted parameter subspace and converge towards the solution closest in terms of the ℓ_2 distance [Lemaire, 1996].

In this work, we aim to expand our understanding of the implicit bias of momentum by analysing its impact on the optimisation trajectory in 2-layer diagonal linear networks. The 2-layer diagonal linear network has garnered significant attention recently [Woodworth et al., 2020b, Vaskevicius et al., 2019, HaoChen et al., 2021, Pesme et al., 2021, Pillaud-Vivien et al., 2022]. Despite its apparent simplicity, this network has surprisingly shed light on training behaviours typically associated with much more complex architectures. Some of these insights include the influence of initialisation [Woodworth et al., 2020b], the impact of noise [Pesme et al., 2021], and the role of the step size [Even et al., 2023]. Consequently, this architecture serves as an excellent surrogate model for gaining a deeper understanding of intricate phenomena such as the role of momentum in the generalisation performance.

9.2.1 Main contributions

In this chapter, we investigate the influence of momentum on the optimisation trajectory of neural networks trained with momentum gradient descent (MGD). Leveraging the continuous-time approximation of MGD – momentum gradient flow (MGF), we show that the optimisation trajectory strongly depends on the key quantity $\lambda = \frac{\gamma}{(1-\beta)^2}$, where γ and β denote the step size and momentum parameter of MGD, respectively. Surprisingly, this continuous-time framework experimentally proves to be a good approximation of the discrete trajectory even for large values of γ.

We proceed to list our main contributions.

- First, using the key quantity λ, we derive a straightforward acceleration rule that maintains the optimisation path while accelerating the optimisation speed.
- Then, focusing on MGF on 2-layer diagonal linear networks, we precisely characterise the recovered solution and prove that for suitably small values of λ, MGF recovers solutions which generalise better than the ones selected by gradient flow (GF) in a sparse regression setting.
- Finally, we provide similar but slightly weaker results for stochastic MGD.

9.2.2 Related works

Momentum and acceleration. Momentum algorithms have their roots in acceleration methods, and many studies have investigated their convergence speed when optimising both convex and non-convex functions: [Ghadimi et al., 2015, Flammarion and Bach, 2015, Kidambi et al., 2018, Can et al., 2019, Sebbouh et al., 2021, Mai and Johansson, 2020, Liu et al., 2020b, Cutkosky and Mehta, 2020, Defazio, 2020, Orvieto et al., 2020, Sebbouh et al., 2021]. Moreover, apart from accelerating training, heavy-ball methods come with the additional advantage of always escaping saddle points [Jin et al., 2018, Sun et al., 2019].

Momentum and continuous-time models. Building upon the foundational work of Alvarez [2000], Attouch et al. [2000], researchers have analysed accelerated gradient methods using second-order differential equations. Su et al. [2014] extended the previous ODE to encompass the Nesterov accelerated method, demonstrating convergence rates similar to the discrete case. Wibisono et al. [2016] adopted a variational perspective to scrutinise the mechanics of

acceleration. A significant advancement emerged with the introduction of Lyapunov analysis, undertaken by Wilson et al. [2021], Sanz Serna and Zygalakis [2021], Moucer et al. [2023]. This analytical approach sheds light on the stability and convergence properties of these methods. Further refinement has been achieved by Shi et al. [2021], who developed high-resolution ODEs tailored to various momentum-based acceleration techniques and able to distinguish between Nesterov's Accelerated Gradient and Polyak's Heavy Ball methods. Finally, error bounds for the discretisation of MGF have been developed by Kovachki and Stuart [2021].

Momentum and Implicit Bias. Sutskever et al. [2013], Leclerc and Madry [2020] have empirically shown significant generalisation improvements in architectures trained with momentum on common vision tasks. Building on these empirical observations, Jelassi and Li [2022] designed a synthetic binary classification problem where a 2-layer convolutional neural network trained with MGD provably generalises better than gradient descent (GD). Recently, Ghosh et al. [2023] reveal that the MGD trajectory closely resembles the gradient flow trajectory of a regularised loss. Through the specific regularisation, the authors argue that the MGD trajectory favors flatter minima than the GD trajectory. The study's findings apply to any reasonable loss, but due to the finite time horizon restriction, cannot characterise the solution to which MGD converges. Additionally, Wang et al. [2023] show that in deep diagonal linear networks with identical weights across layers, increasing the depth biases the optimisation towards sparse solutions.

9.3 From discrete to continuous

Momentum Gradient Descent. We consider minimising a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ using *momentum gradient descent* (MGD) with step size $\gamma > 0$ and momentum parameter $\beta \in [0, 1)$. Initialised at two points $(w_0, w_1) \in \mathbb{R}^{2d}$, the iterates follow the discrete recursion for $k \geq 1$:

$$w_{k+1} = w_k - \gamma \nabla F(w_k) + \beta(w_k - w_{k-1}). \tag{MGD(γ, β)}$$

Momentum Gradient Flow. Directly analysing the discrete recursion MGD(γ, β) appears intractable in many settings. To overcome this difficulty, we follow the classical approach of considering a second order differential equation of the form

$$a\ddot{w}_t + b\dot{w}_t + \nabla F(w_t) = 0 \tag{9.1}$$

with leading coefficient $a \geq 0$ and damping coefficient $b > 0$. In fact, without loss of generality, the previous differential equation can be reduced to a new one which depends on a single parameter λ. Indeed, assume that w_t follows ODE (9.1) with initialisation $(w_{t=0}, \dot{w}_{t=0}) = (w_0, \dot{w}_0)$, then a simple chain rule shows that $\tilde{w}_t = w_{bt}$ follows

$$\frac{a}{b^2}\ddot{\tilde{w}}_t + \dot{\tilde{w}}_t + \nabla F(\tilde{w}_t) = 0,$$

with initialisation $(\tilde{w}_{t=0}, \dot{\tilde{w}}_{t=0}) = (w_0, b\dot{w}_0)$. Hence, up to a time reparametrisation, it is sufficient to consider the following differential equation which depends on a unique parameter $\lambda \geq 0$:

$$\lambda\ddot{w}_t + \dot{w}_t + \nabla F(w_t) = 0. \tag{MGF(λ)}$$

We call the differential equation MGF(λ) *momentum gradient flow* (MGF) with parameter λ. To show the link with the MGD(γ, β) recursion, we discretise MGF(λ) with a second-order

central difference, first-order backward difference, and discretisation step $\varepsilon > 0$ as carried out by Kovachki and Stuart [2021]:

$$\lambda \, \tfrac{w_{k+1}-2w_k+w_{k-1}}{\varepsilon^2} + \tfrac{w_k-w_{k-1}}{\varepsilon} + \nabla F(w_k) = 0. \tag{9.2}$$

Rewriting, we obtain

$$w_{k+1} = w_k - \frac{\varepsilon^2}{\lambda}\nabla F(w_k) + (1 - \frac{\varepsilon}{\lambda})(w_k - w_{k-1}),$$

which corresponds to momentum gradient descent with parameters $\gamma = \frac{\varepsilon^2}{\lambda}$ and $\beta = 1 - \frac{\varepsilon}{\lambda}$. Solving for ε and λ leads to the following proposition:

Proposition 19. *For $(w_0, w_1) \in \mathbb{R}^{2d}$, consider momentum gradient flow MGF(λ) with*

$$\lambda = \frac{\gamma}{(1-\beta)^2}$$

and initialisation $w_{t=0} = w_0$, $\dot{w}_{t=0} = (w_1 - w_0)/\sqrt{\lambda\gamma}$. Then, discretising as (9.2) with discretisation step $\varepsilon = \sqrt{\lambda\gamma} = \gamma/(1-\beta)$ leads to the momentum gradient descent recursion MGD(γ, β) with step size γ, momentum parameter β, and initialisation (w_0, w_1).

Proposition 19 motivates studying MGF(λ) as a continuous proxy for MGD(γ, β) assuming that the discretisation (9.2) closely approximates the continuous path.

Discretisation Error Bounds. Unfortunately, applying known discretisation error bounds to our setting leads to very pessimistic bounds. Indeed, for step size γ and momentum parameter β, consider the iterates w_k from MGD(γ, β) initialised at (w_0, w_1). Now, let $w(t)$ be the solution of MGF(λ) with $\lambda = \gamma/(1-\beta)^2$ and the appropriate initialisation from Proposition 19. Then, for a finite horizon $K > 0$, classical discretisation error bounds (see Kovachki and Stuart [2021], Theorem 4) lead to a catastrophic

$$\sup_{k \le K} \|w_k - w(k\varepsilon)\| \le \exp(CK)\varepsilon,$$

where the constant C depends on λ and F. Such an exponential dependence in the time horizon K questions the suitability of momentum gradient flow as a good proxy for momentum gradient descent. However, empirically, the above bound appears excessively pessimistic (see Figure 9.1: Left and Middle). The MGF and MGD trajectories behave similarly in various settings, even with non-convex losses F and relatively large step sizes γ (see Appendix D.6 for additional experiments).

Intertwined Roles of γ and β. When the discretisation accurately follows the continuous path, Proposition 19 implies that the trajectory of MGD(γ, β) is solely determined by a single parameter $\lambda = \gamma/(1-\beta)^2$, intertwining step size and momentum as observed in Figures 9.1 and 9.2. **Consequently, γ and β serve interchangeable roles in influencing the trajectory of MGD(γ, β).** Note that this single-parameter dependence aligns with empirical results from Leclerc and Madry [2020] where generalisation performance with large step sizes can be replicated with momentum and smaller step sizes. Though the quantity $\gamma/(1-\beta)^2$ spontaneously appears in works studying MGD [Ghosh et al., 2023], to the best of our knowledge, its natural presence was never clearly explained and motivated.

MGD Acceleration Rule. Though all couples (γ, β) with the same same value of λ yield the same trajectory, the iterates do not follow this path at the same speed.

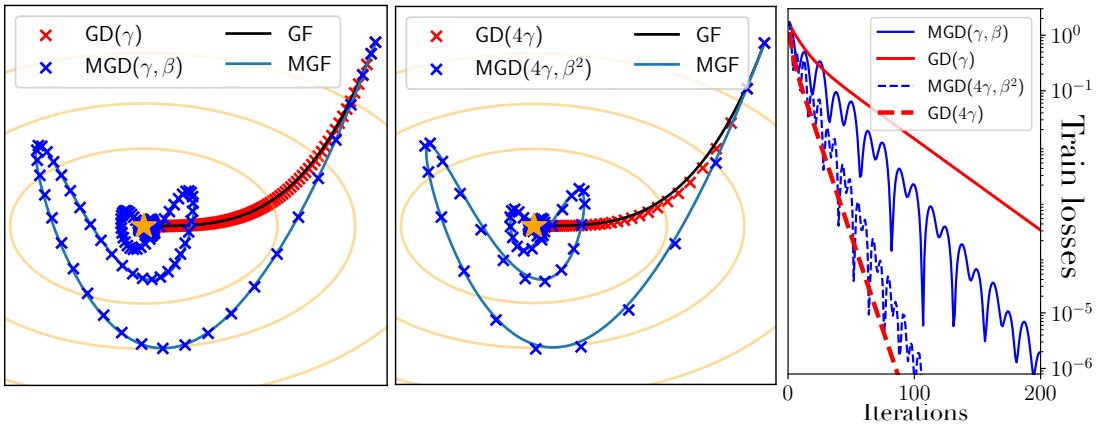


Figure 9.1: (M)GD over a 2D quadratic. *Left and Middle*: The (M)GD trajectories closely follow the continuous trajectories of (M)GF as suggested by Proposition 19. *Right*: MGD$(4\gamma, \beta^2)$ follows the same trajectory as MGD(γ, β) but twice as fast as suggested by Corollary 1. In contrast, GD(4γ) runs four times faster than GD(γ).

Corollary 1 (Acceleration rule)**.** *Let* MGD(γ, β) *initialised at* $w_0 = w_1 \in \mathbb{R}^d$ *correspond to the discretisation of* MGF(λ) *with discretisation step* ε*. Now, for* $\rho \in \mathbb{R}_{>0}$*, consider the different parameter couple*

$$\hat{\gamma} = \rho^2 \gamma \quad \textit{and} \quad \hat{\beta} = 1 - \rho(1-\beta) =_{\beta \to 1} \beta^\rho + O((1-\beta)^2).$$

Then, since $\hat{\gamma}/(1-\hat{\beta})^2 = \lambda$, MGD$(\hat{\gamma}, \hat{\beta})$ *initialised at* $w_0 = w_1$ *becomes the discretisation of the same* MGF(λ) *but with discretisation step* $\hat{\varepsilon} = \rho \cdot \varepsilon$*.*

Following the notations of the previous corollary for an integer $\rho \geq 2$ and letting w_k and $\hat{w}_k$ denote the iterates of MGD(γ, β) and MGD$(\hat{\gamma}, \hat{\beta})$, respectively, Corollary 1 implies that we expect $w_{\rho \cdot k}$ and $\hat{w}_k$ to be close. This is in contrast with gradient descent, where scaling the step size by a factor ρ^2 leads to a speedup of ρ^2. This acceleration rule is illustrated in Figure 9.1 with $\rho = 2$.

Optimisation Regimes. The link between λ, γ, and β highlights several regimes:

- β **large – the iterates converge arbitrarily slow.** Taking β close to 1 while keeping γ constant leads to $\lambda \gg 1$. As explained previously, a chain rule shows that $\tilde{w}_t = w_{\sqrt{\lambda} t}$ follows the ODE $\ddot{\tilde{w}}_t + \lambda^{-1/2} \cdot \dot{\tilde{w}}_t + \nabla F(\tilde{w}_t) = 0$. Consequently, the damping parameter $\lambda^{-1/2}$ goes to 0, and we expect the iterates to heavily oscillate and converge arbitrarily slowly.

- γ **small – the iterates follow GF.** Taking $\gamma \to 0$ while keeping β fixed leads to $\lambda \ll 1$, and MGF(λ) boils down to gradient flow. We expect the MGD(γ, β) iterates to be close to the discretisation of GF with discretisation step $\varepsilon = \gamma/(1-\beta)$. That is, MGD$(\gamma, \beta)$ will approximate GD with step size $\gamma/(1-\beta)$. Hence, MGD gains a speed-up of $1/(1-\beta)$ over GD without a change of trajectory.

- **The "momentum" regime.** In this regime, γ and β are such that λ is non-degenerate, and gradient flow cannot capture the trajectory of MGD(γ, β). Hence, β has an impact on the optimisation path, and the iterates can still converge in reasonable time.

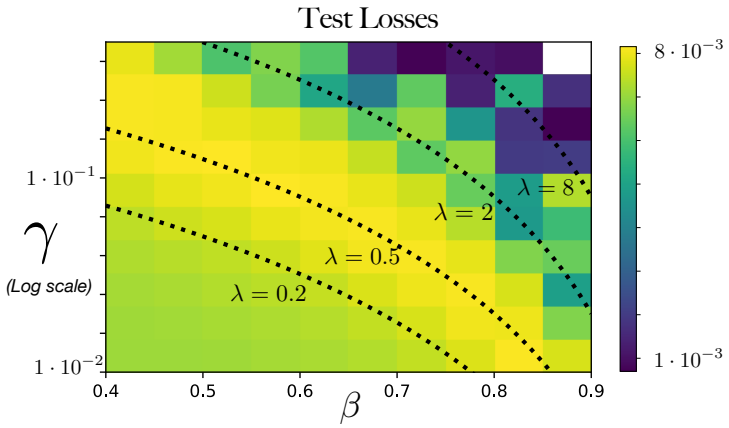


Figure 9.2: Teacher-student framework with a fully-connected 1-hidden layer ReLU network. The level lines of the test loss after training with MGD(γ, β) correspond to values of γ, β which have a fixed value $\lambda = \gamma/(1-\beta)^2$, as predicted by Proposition 19.

9.4 Momentum gradient flow over diagonal linear networks

Overparametrised Linear Regression. We consider a linear regression over n samples $(x_i, y_i)_{i=1}^n$ with inputs x_i living in $\mathbb{R}^d$ and scalar outputs $y_i \in \mathbb{R}$. We assume the dimension d to be larger than the number of samples n, in which case there exists an infinite number of vectors $\theta^\star$ which perfectly fit the dataset with $y_i = \langle \theta^\star, x_i \rangle$ for all $1 \leq i \leq n$. We call these vectors *interpolators* and we denote by $\mathcal{S}$ the set of such vectors: $\mathcal{S} = \{\theta^\star \in \mathbb{R}^d : y_i = \langle \theta^\star, x_i \rangle, \ \forall i \in [n]\}$. Note that $\mathcal{S}$ is an affine space of dimension at least $(d-n)$ equal to $\theta^\star + \text{span}(x_1, \dots, x_n)^\perp$ for any interpolator $\theta^\star$. We consider the quadratic loss:

$$\mathcal{L}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2. \tag{9.3}$$

MGF over Least Squares. A classical result found in Lemaire [1996] and Gunasekar et al. [2018c] shows that when initialised at θ_0, gradient flow over the quadratic loss (9.3) converges to the orthogonal projection of the initialisation on $\mathcal{S}$: $\arg\min_{\theta^\star \in \mathcal{S}} \|\theta^\star - \theta_0\|_2$. This next proposition from Alvarez [2000] shows that momentum does not fundamentally change the implicit bias.

Proposition 20 (Alvarez [2000])**.** *Initialised at θ_0 with initial speed $\dot{\theta}_0$, momentum gradient flow* MGF(λ) *over the least squares loss (9.3) converges towards*

$$\arg\min_{\theta^\star \in \mathcal{S}} \|\theta^\star - (\theta_0 + \lambda \dot{\theta}_0)\|_2.$$

MGF(λ) recovers the same solution as gradient flow but with an effective initialisation $\theta_0 + \lambda\dot{\theta}_0$ which takes into account the drift along $\text{span}(x_1, \cdots, x_n)^\perp$ due to the initial speed $\dot{\theta}_0$. Note that in practice, $\dot{\theta}_0$ is chosen equal to 0, in which case the presence of momentum has no effect on the recovered solution.

To better understand momentum's effect on neural networks, we move beyond simple linear parametrization.

2-Layer Diagonal Linear Network. We consider a toy neural network, which corresponds to reparametrising the regression vector as $\theta = u \odot v$ for weights $(u, v) \in \mathbb{R}^{2d}$. This parametrisation

can be viewed as a simple neural network $x \mapsto \langle u, \sigma(\mathrm{diag}(v)x)\rangle$, where the output weights are u, the inner weights are the diagonal matrix $\mathrm{diag}(v)$, and where the activation function σ is the identity. The loss function over the trainable weights $w = (u, v) \in \mathbb{R}^{2d}$ now writes

$$F(w) = \mathcal{L}(u \odot v) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle x_i, u \odot v\rangle)^2,$$

where $\odot$ denotes the Hadamard product. Despite the simplicity of this reparametrisation, the loss function F is non-convex and challenging to analyse.

Momentum Gradient Flow. We consider momentum gradient flow MGF(λ) with parameter $\lambda \geq 0$ over the diagonal-linear-network loss F:

$$\begin{aligned} \lambda\ddot{u}_t + \dot{u}_t + \nabla\mathcal{L}(\theta_t) \odot v_t = 0 \\ \lambda\ddot{v}_t + \dot{v}_t + \nabla\mathcal{L}(\theta_t) \odot u_t = 0. \end{aligned} \tag{9.4}$$

We initialise the flow with zero speed $\dot{u}_0 = \dot{v}_0 = 0$, and apart from requiring the quantity $|u_0^2 - v_0^2|$ to have non-zero coordinates[1], we impose no further constraints on the weight initialisations (u_0, v_0). In what follows, we often rely on the reparametrisation $(w_{+,t}, w_{-,t}) \coloneqq (u_t + v_t, u_t - v_t)$ which makes our formulas more succinct. We will also make use of the *initialisation scale* α, which we define as $\alpha \coloneqq \max(\|u_0\|_\infty, \|v_0\|_\infty)$ and consider as a small quantity.

Balancedness. In our results, the *balancedness* of the weights plays a key role. We recall its definition here.

Definition (Balancedness)**.** The *balancedness*[2] of the weights of the diagonal linear network corresponds to the quantity $\Delta_t \coloneqq |u_t^2 - v_t^2| \in \mathbb{R}^d_{\geq 0}$. We define $\Delta_\infty \coloneqq \lim_{t\to\infty}\Delta_t$ as the *asymptotic balancedness*.

Notice that with the above definition we simply adapted the classical notion of balancedness for general linear neural networks [see Du et al., 2019, Arora et al., 2019] to our toy setting. In the case of gradient flow, a simple derivation shows that balancedness is a conserved quantity: i.e., $\Delta_t = \Delta_0$ for all $t \geq 0$. However, the evolution of Δ_t becomes more complicated as soon as $\lambda > 0$, and our findings emphasise that the *asymptotic balancedness* Δ_∞ plays a crucial role in the generalisation properties of the recovered solution.

Experimental Details. In our numerical experiments, we explore the effects of momentum in the noiseless sparse regression setting with **uncentered data** as in Nacson et al. [2022]. Specifically, we choose $(x_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu\mathbf{1}, \sigma^2 I_d)$ and $y_i = \langle x_i, \theta_s^\star\rangle$ for $i \in [n]$, where $\theta_s^\star$ is s-sparse with nonzero entries equal to $1/\sqrt{s}$. The use of uncentered data is necessary in order to experimentally observe a clear impact of momentum over the training trajectory (see Figure D.5 for experiments with centered data). We train a 2-layer diagonal linear network with (M)GD and (M)GF with a uniform initialisation $u_0 = \alpha \cdot \mathbf{1}$, $v_0 = 0$, where $\alpha = 0.01$. For the plots presented in the main part of our chapter, we fixed $(n, d, s) = (20, 30, 5)$, $(\mu, \sigma) = (1, 1)$. We show results averaged over 5 replications. We refer the reader to Appendix D.6 for additional experiments where we vary the parameters of the data distribution (e.g., centered data), change the architecture of the trained model, and give further details on the implementation of the (M)GF simulation.

[1] If initially $u_{i,0} = \pm v_{i,0}$ for some coordinate $i \in [d]$, then $u_{i,t} = \pm v_{i,t}$, $\forall t \geq 0$. Hence, imposing $|u_0^2 - v_0^2| \neq 0$ becomes equivalent to working with $2d$ distinct weights. See Appendix D.3.3 for the full argument from uniqueness.

[2] The absolute value in the definition must be understood coordinate-wise.

Notations. We let $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times d}$ denote the feature matrix and $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ – the output vector. For a vector $z \in \mathbb{R}^d$ and a scalar function $f : \mathbb{R} \to \mathbb{R}$, the action of f on z must be understood element-wise: $f(z) \in \mathbb{R}^d$ represents the vector $(f(z_k))_{k=1}^d$. Inequalities between vectors will also be interpreted as holding coordinate-wise. Additionally, when we write $q_\pm$ for some place-holder quantity q, we mean that we refer to both q_+ and q_-. For example: $w_{\pm,t} = (u_t \pm v_t)$. Finally, for a strictly convex function $\Phi : \mathbb{R}^d \to \mathbb{R}$, which we call a *potential*, the Bregman divergence is defined as the nonnegative quantity $D_\Phi(\theta_1, \theta_2) = \Phi(\theta_1) - \Phi(\theta_2) - \langle \nabla\Phi(\theta_2), \theta_1 - \theta_2 \rangle, \ \forall \theta_1, \theta_2 \in \mathbb{R}^d$.

9.4.1 Implicit bias of gradient flow

Before analysing the effect of momentum, we start by recalling the known results for gradient flow on diagonal linear networks, which corresponds to taking $\lambda = 0$ in eq. (9.4). Woodworth et al. [2020b] show that the predictors $\theta_t = u_t \odot v_t$ converge towards an interpolator θ^{GF} uniquely defined by the following constrained minimisation problem:

$$\theta^{\mathrm{GF}} = \underset{\theta^\star \in \mathcal{S}}{\arg\min}\ D_{\psi_{\Delta_0}}(\theta^\star, \theta_0), \tag{9.5}$$

where for $\Delta \in \mathbb{R}^d_{>0}$, $\psi_\Delta : \mathbb{R}^d \to \mathbb{R}$ denotes the hyperbolic entropy function [Ghai et al., 2020] at scale Δ:

$$\psi_\Delta(\theta) = \frac{1}{4} \sum_{i=1}^d \left(2\theta_i \mathrm{arcsinh}\left(\frac{2\theta_i}{\Delta_i} \right) - \sqrt{4\theta_i^2 + \Delta_i^2} + \Delta_i \right), \tag{9.6}$$

and D_{ψ_Δ} is the Bregman divergence. Note that through eq. (9.5), θ^{GF} corresponds to the Bregman-projection of the initialisation on the set of interpolators.

Effect of the Initialisation Scale. For a small initialisation scale α, $\theta_0 = O(\alpha^2)$ becomes much smaller than any interpolator $\theta^\star \in \mathcal{S}$. Hence, $D_{\psi_{\Delta_0}}(\theta^\star, \theta_0)$ roughly equals $D_{\psi_{\Delta_0}}(\theta^\star, 0)$, and eq. (9.5) should be thought of as

$$\theta^{\mathrm{GF}} \approx \underset{\theta^\star \in \mathcal{S}}{\arg\min}\ \psi_{\Delta_0}(\theta^\star). \tag{9.7}$$

This last equation highlights the fact that the recovered solution simply depends on the initial balancedness Δ_0, making it a key quantity. Importantly, the hyperbolic entropy is a convex function which interpolates between the ℓ_1 and ℓ_2 norms as the magnitude of Δ_0 goes from 0 to $+\infty$ (see Woodworth et al. [2020b], Theorem 2). So, as $\Delta_0 = O(\alpha^2)$ goes to 0, ψ_{Δ_0} becomes asymptotically identical to the ℓ_1-norm (see Appendix D.5). Hence, as seen through eq. (9.7), a small initialisation scale α leads to the recovery of a solution with a small ℓ_1-norm, which facilitates sparse recovery and explains why this setting is referred to as the "rich" or "feature-learning" regime. On the other hand, larger initialisation scales lead to the so-called "kernel" or "lazy" regime, where gradient flow selects small-ℓ_2-norm solutions. **Overall, the smaller the initialisation scale, the closer the retrieved solution will be to the minimum-ℓ_1-norm solution.** We refer the reader to the work of Wind et al. [2023] for precise recovery bounds. However, as noted in Even et al. [2023], the picture remains incomplete if we do not take into account the homogeneity of Δ_0. Indeed, initialisations with entries of different magnitudes can hinder the recovery of a sparse vector. However, in our case, our experiments (for uncentered data) verify that the overall magnitudes of Δ_0 and Δ_∞ are sufficient to explain the effects of momentum. We therefore put aside potential homogeneity considerations.

9.4.2 Implicit bias of momentum gradient flow

We now move on to describe the impact of momentum on the solution recovered by MGF(λ). Our work proceeds under the following assumption.

Assumption 7 (Boundedness)**.** *The optimisation trajectory $(u_t, v_t)_{t\geq 0}$ of MGF (9.4) is bounded.*

Unfortunately, even though Assumption 7 holds true in all our experiments, the boundedness of the trajectory of a second-order gradient flow has only been established under stronger assumption on the loss function [Alvarez, 2000, Goudou and Munier, 2009, Apidopoulos et al., 2022]. We defer further details to Appendix D.3.1. Crucially, the boundedness assumption allows us to prove the convergence of the iterates, and we let $(u_\infty, v_\infty) \coloneqq \lim_{t\to\infty}(u_t, v_t)$. Our goal now becomes to characterise the recovered predictor which we denote with $\theta^{\text{MGF}} \coloneqq u_\infty \odot v_\infty$. For our proofs, we make the following additional assumption.

Assumption 8 (Balancedness)**.** *The asymptotic balancedness Δ_∞ has non-zero coordinates: $\Delta_{\infty,i} > 0$ for all $i \in [d]$.*

Again, Assumption 8 holds true empirically in all our experiments, and in Section 9.4.3, we prove that small values of λ lead to nonzero asymptotic balancedness. Positing Assumption 8 allows us to prove that the recovered solution θ^{MGF} interpolates the dataset.

General Characterisation of MGF Bias

In our main result for MGF, we prove that the iterates converge towards an interpolator characterised as the solution of a constrained minimisation problem which involves the hyperbolic entropy (9.6) scaled at the asymptotic balancedness Δ_∞. Moreover, we derive an insightful description of the asymptotic balancedness in terms of the full optimisation trajectory which allows us to compare the generalisation properties of MGF and GF for small values of λ. Before stating our main continuous-time theorem, we define two integral quantities which appear in our results.

Lemma 2. *The following integral quantities Ω_+ and Ω_- are well-defined and finite:*

$$\Omega_\pm \coloneqq \int_0^\infty \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_{\pm,t} w_{\pm,s}) \mathrm{d}s \; \mathrm{d}t$$

where $\operatorname{sgn}(\cdot)$ *denotes the sign function,* $w_{\pm,t} = u_t \pm v_t$*, and* m.p.v. *denotes a modified Cauchy principal value defined in Appendix D.1.*

The fact that the weights $w_{\pm,t}$ can cross zero necessitates the use of the modified Cauchy principal value since otherwise the integrals would diverge. Now, for succinctness, let us introduce the integral quantities

$$I_\pm \coloneqq \Omega_\pm + \Lambda_\pm,$$

where the terms $\Lambda_\pm$ vanish whenever the balancedness Δ_t remains strictly positive for all $t \in [0, \infty]$. The precise form of $\Lambda_\pm$ is uninformative and can be found in Equation (D.9), Appendix D.3.3. We now proceed to characterise the recovered solution θ^{MGF}.

Theorem 3. *The solution θ^{MGF} of MGF (9.4) interpolates the dataset and satisfies the following implicit regularisation:*

$$\theta^{MGF} = \operatorname*{arg\,min}_{\theta^\star \in \mathcal{S}} \; D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

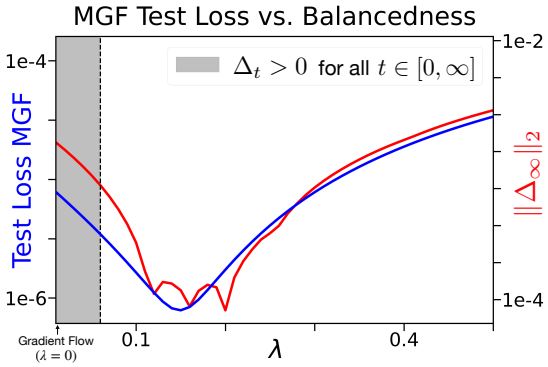


Figure 9.3: Test loss (in blue) and magnitude of balancedness (in red) at convergence of MGF(λ) over a diagonal linear network in a sparse regression setting with uncentered data. As predicted by Theorem 3, a more balanced solution generalises better. The shaded zone corresponds to values of λ for which the balancedness never hits zero during training and for which Corollary 2 therefore holds.

In the above expression, $D_{\psi_{\Delta_\infty}}$ denotes the Bregman divergence with potential ψ_{Δ_∞}, where the asymptotic balancedness equals

$$\Delta_\infty = \Delta_0 \odot \exp\big(-(I_+ + I_-)\big)$$

and $\tilde{\theta}_0 = \frac{1}{4}(w_{+,0}^2 \odot \exp(-2I_+) - w_{-,0}^2 \odot \exp(-2I_-))$ denotes a perturbed initialisation term.

The proof of Theorem 3 appears in Appendix D.3.3 as well as explicit formulas for Δ_∞ and $\tilde{\theta}_0$. We explain the significance and shed more light on the different parts of Theorem 3 below.

Perturbed Initialisation $\tilde{\theta}_0$. In all our experiments, we observe that the perturbed initialisation $\tilde{\theta}_0$ remains negligible in the sense that for any interpolator $\theta^\star \in \mathcal{S}$, $\|\tilde{\theta}_0\|_2 \ll \|\theta^\star\|_2$. Moreover, in the next section, we prove that whenever the balancedness remains nonzero during training, $\tilde{\theta}_0$ becomes smaller than α^2, where α stands for the initialisation scale. Hence, exactly for the same reasons as for gradient flow, the implicit regularisation problem from Theorem 3 should be though of as

$$\theta^{\text{MGF}} \approx \arg\min_{\theta^\star \in S} \psi_{\Delta_\infty}(\theta^\star). \tag{9.8}$$

Appendix D.3.3 provides more details. Thus, the asymptotic balancedness Δ_∞ becomes the key quantity governing the properties of the recovered solution.

Key Role of Δ_∞. If during optimisation the weights become more balanced, i.e. $\Delta_\infty < \Delta_0$, then as discussed previously, based on the properties of ψ_{Δ_∞}, the recovered solution will enjoy better sparsity guarantees than the solution of gradient flow. Figure 9.3 illustrates this point: the smaller the magnitude of Δ_∞, the better the generalisation. Finally note that by eqs. (9.5) and (9.8), θ^{MGF} approximately equals the solution recovered from gradient flow initialised at $u_0 = \sqrt{\Delta_\infty}, v_0 = 0$, which we denote by $\theta^{\text{GF}}_{\Delta_\infty}$. We observe $||\theta^{\text{MGF}} - \theta^{\text{GF}}_{\Delta_\infty}||_2 / ||\theta^{\text{GF}}_{\Delta_\infty}||_2 < 0.01$ in all our experiments, which validates the approximation in eq. (9.8).

Path-Dependent Quantity. Unfortunately, the asymptotic balancedness depends on the whole optimisation trajectory in a very intricate way, and we cannot compare $\|\Delta_\infty\|$ and $\|\Delta_0\|$. Thus, in general, we cannot meaningfully compare the recovered interpolators θ^{MGF} and θ^{GF}. However, in the following section we prove that with the additional assumption that the balancedness remains nonzero, we have $\Delta_\infty < \Delta_0$.

9.4.3 Provable benefits of momentum for small values of λ

In this subsection, we prove that for small values of the momentum flow parameter λ, the recovered solution becomes more balanced (and therefore sparser) than the solution of gradient flow. As a starting point for our argument, notice that if the balancedness $\Delta_t = |u_t^2 - v_t^2| = |w_{+,t} w_{-,t}|$ remains strictly positive throughout training, then the weights $w_{\pm,t}$ never change sign. Hence, the integral quantities $\Lambda_\pm$ become 0, and $\Omega_\pm > 0$. Thus, $I_\pm > 0$, which combined with Theorem 3 implies the following corollary.

Corollary 2. *For $\lambda > 0$, if the balancedness Δ_t remains strictly positive during training (i.e. $\Delta_t \neq 0$ for $t \in [0, +\infty]$), then the perturbed initialisation satisfies $|\tilde{\theta}_0| < \alpha^2$ and*

$$\Delta_\infty = \Delta_0 \odot \exp\Big(- \lambda \int_0^\infty \Big(\frac{\dot{w}_{+,t}}{w_{+,t}}\Big)^2 + \Big(\frac{\dot{w}_{-,t}}{w_{-,t}}\Big)^2 \mathrm{d}t\Big).$$

Importantly, $\Delta_\infty < \Delta_0$.

In words, the above corollary (proved in Appendix D.3.4) implies that if the balancedness Δ_t does not hit zero during training, then (i) the perturbation term $\tilde{\theta}_0$ is provably negligible, (ii) the asymptotic balancedness is coordinate-wise smaller than initial balancedness Δ_0 which translates into a solution with better sparsity properties than the gradient flow interpolator. This regime corresponds to the gray zone in Figure 9.3. The following proposition proved in Appendix D.5 demonstrates that for small values of λ, the balancedness remains strictly positive.

For $\lambda \leq \frac{n}{\|y\|_2^2} \cdot (\min_{i \leq d} \Delta_{0,i})$, the balancedness Δ_t never vanishes: $\Delta_t \neq 0$, $\forall t \in [0, +\infty]$. Hence, through Section 9.4.3 and Corollary 2, we show that small values of λ lead to solutions with better sparse recovery guarantees.

Limitations of Our Analysis. In Appendix D.3.3, we prove that Δ_t can vanish at most a finite number of times. Experimentally, Δ_t never hits 0 for much larger values of λ than $\frac{n}{\|y\|_2^2} \cdot (\min_{i \leq d} \Delta_{0,i})$, making the bound from Section 9.4.3 relatively loose. In Figure 9.3, we empirically observe an interval $(0, \lambda_{max})$ in which MGF(λ) outperforms GF in terms of generalisation. Moreover, there exists an optimal value $\lambda^\star$ (roughly corresponding to the smallest Δ_∞) which brings about the most improvement compared to gradient flow. Unfortunately, as observed Figure 9.3, the balancedness vanishes for $\lambda = \lambda^\star$, and therefore Corollary 2 does not cover the optimal value. Also note that $(0, \lambda_{max})$ and $\lambda^\star$ depend on the data.

Behaviour of Δ_∞ for Small Values of λ. Unfortunately, determining the precise effect of λ on Δ_∞ is challenging. Nonetheless, for small λ, we informally show in Appendix D.3.5 that

$$\Delta_\infty^2 \underset{\lambda \to 0}{\approx} \Delta_0^2 \odot \exp\big(- 2\lambda \int_0^\infty \nabla \mathcal{L}(\theta_s)^2 \mathrm{d}s\big).$$

This approximate equivalence for small λ echoes the implicit bias of SGD [Even et al., 2023, Pesme et al., 2021], which involves a similar formulation for the effective initialisation where the step size γ appears instead of λ. Note that the above approximation suggests that for small values of λ, Δ_∞ monotonically decreases with λ as experimentally confirmed by Figure 9.3.

9.4.4 Sketch of proof

Implicit bias through a second-order time-varying mirror flow. A natural way of showing the implicit regularisation (9.5) of gradient flow on a 2-layer diagonal linear network goes through proving that the predictors θ_t^{GF} follow the mirror flow $\mathrm{d}\nabla\psi_{\Delta_0}(\theta_t^{\mathrm{GF}}) = -\nabla\mathcal{L}(\theta_t^{\mathrm{GF}})\mathrm{d}t$. In

our setting, we prove that the predictors θ_t^{MGF} follow a second-order time-varying mirror flow. Specifically, we define a family of potentials $(\Phi_t)_{t\geq 0}$ with $\Phi_t(\theta) := \psi_{\Delta_t}(\theta) - \langle \phi_t, \theta \rangle$ where ψ_{Δ_t} corresponds to the hyperbolic entropy (9.6) depending on the balancedness Δ_t and a perturbation function ϕ_t. We then prove the following proposition.

Proposition 21. *The predictors θ_t^{MGF} follow a momentum mirror flow with time-varying potentials Φ_t:*

$$\lambda \frac{\mathrm{d}^2 \nabla \Phi_t(\theta_t^{MGF})}{\mathrm{d}t^2} + \frac{\mathrm{d} \nabla \Phi_t(\theta_t^{MGF})}{\mathrm{d}t} + \nabla \mathcal{L}(\theta_t^{MGF}) = 0.$$

The implicit regularisation follows from integrating the ODE: $\nabla \Phi_\infty(\theta^{\text{MGF}}) = -\int_0^\infty \nabla \mathcal{L}(\theta_t^{\text{MGF}}) \mathrm{d}t \in \mathtt{span}(x_1, \dots, x_n)$ which exactly corresponds to the KKT conditions of the constrained minimisation from Theorem 3. Assuming that $w_{\pm,t}$ do not change sign, the proof of Proposition 21 comes naturally and relies on the writing $w_{\pm,t} = \mathrm{sgn}(w_{\pm,0}) \exp(\rho_{\pm,t})$. When the iterates cross 0, this reparametrisation does not hold anymore. The analysis can still be carried out by decomposing $\mathbb{R}_{\geq 0}$ into intervals on which the iterates have constant sign and appropriately sticking the intervals using a modified Cauchy principal value.

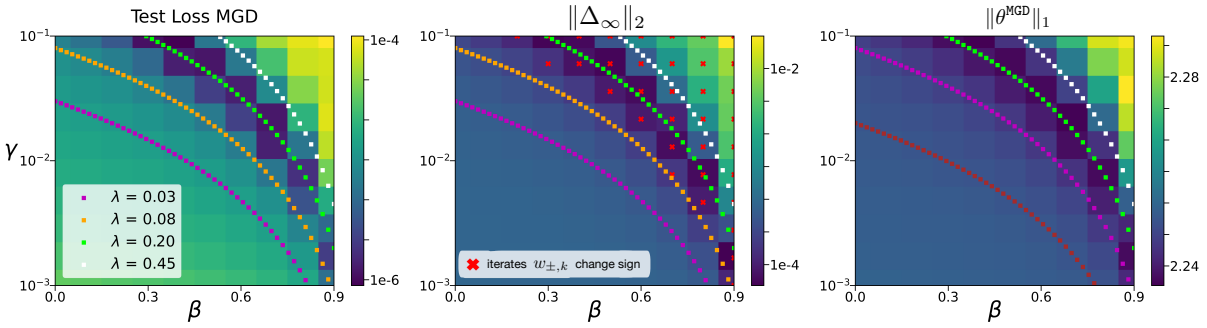


Figure 9.4: (Non-stochastic) MGD over a diagonal linear network in a sparse regression setting with uncentered data. As predicted by Proposition 19, the three quantities at convergence only depend on the single parameter $\lambda := \gamma/(1-\beta)^2$. As predicted by Theorem 4, a more balanced solution (*center plot*) leads to a solution with a smaller ℓ_1-norm (*right plot*), which in turn translates into better generalisation (*left plot*). Finally, as predicted by Corollary 3, the trajectories for which the iterates do not cross zero satisfy $\Delta_\infty < \Delta_0$, where Δ_0 (approximately) corresponds to the asymptotic balancedness for $\beta = 0$ and $\gamma = 10^{-3}$.

9.5 Momentum SGD over diagonal linear networks

In this section, we move from continuous to discrete time and focus on the original MGD(γ, β) recursion for which we can prove similar but slightly weaker results than the ones for MGF. In fact, our results hold for stochastic momentum gradient descent (SMGD) with any batch size $B \in [n]$. For step size $\gamma > 0$ and momentum parameter $\beta \in [0, 1)$, the SMGD recursion writes as follows:

$$\begin{aligned} u_{k+1} &= u_k - \gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot v_k + \beta(u_k - u_{k-1}) \\ v_{k+1} &= v_k - \gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot u_k + \beta(v_k - v_{k-1}), \end{aligned} \tag{9.9}$$

where $\mathcal{L}_{\mathcal{B}_k}(\theta) := \frac{1}{2B} \sum_{i \in \mathcal{B}_k} (y_i - \langle u \odot v, x_i \rangle)^2$ corresponds to the partial loss over the batch $\mathcal{B}_k \subset [n]$ of size B. The batches could be sampled with or without replacement. As for continuous time,

we let $\theta_k = u_k \odot v_k$ correspond to the regression predictor. We initialise at $u_1 = u_0$ and $v_1 = v_0$, and we again consider the balancedness of the weights $\Delta_k \coloneqq |u_k^2 - v_k^2|$ for $k \geq 0$, the reparametrised iterates $w_{\pm,k} \coloneqq u_k \pm v_k$, and the *initialisation scale* $\alpha \coloneqq \max(\|u_0\|_\infty, \|v_0\|_\infty)$. In contrast to our continuous-time prerequisites where we only assumed boundedness of the optimisation trajectory, here we assume that the iterates converge:

Assumption 9 (Convergence)**.** *The iterates (u_k, v_k) converge towards the limiting weights (u_∞, v_∞). We denote by $\theta^{SMGD} \coloneqq u_\infty \odot v_\infty$ the recovered predictor.*

As in continuous time, we again assume that the asymptotic balancedness is nonzero.

Assumption 10 (Balancedness)**.** *The asymptotic balancedness $\Delta_\infty \coloneqq |u_\infty^2 - v_\infty^2|$ has non-zero coordinates.*

Similar to Lemma 2, we define two discrete infinite sums which depend on the entire trajectory and appear in our discrete-time result.

Lemma 3. *The following two sums S_+ and S_- converge to finite vectors:*

$$S_\pm = \frac{1}{1-\beta} \sum_{k=1}^{\infty} \left[r\Big(\frac{w_{\pm,k+1}}{w_{\pm,k}}\Big) + \beta r\Big(\frac{w_{\pm,k}}{w_{\pm,k+1}}\Big) \right],$$

where $r(z) = (z-1) - \ln(|z|)$ for $z \neq 0$.

Importantly, the function $r(z)$ from Lemma 3 is positive for $z > 0$. Contrary to the continuous-time case, in discrete time, the iterates $w_{\pm,k}$ never exactly equal zero. Indeed, since ∇L is linear, we have that for all $k \geq 0$, $w_{\pm,k}(\gamma, \beta)$ is a polynomial in (γ, β). Therefore, the set of pairs (γ, β) for which there exists $k \geq 0$ such that $w_{\pm,k}(\gamma, \beta) = 0$ is a negligible set in $\mathbb{R}^2$. The iterates therefore 'jump' over zero, making the sums from Lemma 3 well-defined.

9.5.1 General characterisation of SMGD bias

The following theorem represents the discrete counterpart of Theorem 3 and generalises [Even et al., 2023, Theorem 1] which considers SGD without momentum.

Theorem 4. *The solution θ^{SMGD} of SMGD (9.9) interpolates the dataset and satisfies the following implicit regularisation:*

$$\theta^{SMGD} = \operatorname*{arg\,min}_{\theta^\star \in \mathcal{S}} \; D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

In the above expression, $D_{\psi_{\Delta_\infty}}$ denotes the Bregman divergence with potential ψ_{Δ_∞}, where the asymptotic balancedness equals

$$\Delta_\infty = \Delta_0 \odot \exp\big(-(S_+ + S_-)\big)$$

and $\tilde{\theta}_0 = \frac{1}{4}(w_{+,0}^2 \odot \exp(-2S_+)) - w_{-,0}^2 \odot \exp(-2S_-))$ denotes a perturbed initialisation term.

Due to the strong similarities with Theorem 3, we proceed by making similar comments. In our experiments, the norm of the perturbed initialisation $\tilde{\theta}_0$ remains much smaller than that of any interpolator $\theta^\star$. Hence, arguing as before, the implicit regularisation problem from Theorem 4 should be though of as

$$\theta^{\mathrm{SMGD}} \approx \operatorname*{arg\,min}_{\theta^\star \in S} \psi_{\Delta_\infty}(\theta^\star). \tag{9.10}$$

Again, the asymptotic balancedness Δ_∞ controls the generalisation properties of the recovered solution. Thus, if $\|\Delta_\infty(\gamma, \beta)\|_2 < \|\Delta_\infty(\gamma', \beta')\|_2$, we expect the interpolator $\theta^{\text{SMGD}}(\gamma, \beta)$ to be sparser than $\theta^{\text{SMGD}}(\gamma', \beta')$. Figure 9.4 illustrates this point: the smaller the magnitude of Δ_∞ (center plot), the better the sparsity of the interpolator (right plot), which translates into better generalisation (left plot). Unfortunately, as for MGF, the asymptotic balancedness Δ_∞ depends on the whole optimisation trajectory in an intricate way, which prevents us from extracting an insightful formula for Δ_∞ in terms of γ and β. However, Figure 9.4 indicates that Δ_∞ effectively depends on the single parameter $\lambda = \gamma/(1-\beta)^2$. As in Figure 9.2, λ again clearly appears to be the relevant quantity which governs the performance of MGD, and not γ and β considered individually. These empirical observations support the idea that even for 'practical' step sizes γ and momentum parameters β, MGD(γ, β) closely follows MGF(λ).

Figure 9.4 also clearly shows that the asymptotic balancedness decreases as the key quantity λ increases over an interval $[0, \lambda^\star]$ where $\lambda^\star$ denotes the parameter inducing the best generalisation performances. Then, for λ above $\lambda^\star$, the magnitude of Δ_∞ starts to grow and the sparsity of the solutions deteriorates. We expect proving this phenomenon to be very challenging. Such a proof would require a fine-grained analysis of the sums $S_\pm$, which becomes already quite involved when $\beta = 0$ as performed by Even et al. [2023].

Now, similar to the continuous-time result, the following corollary shows that if the iterates do not change sign, then the asymptotic balancedness becomes smaller than the initial balancedness.

Corollary 3. *For $\gamma, \beta > 0$, if the iterates $w_{\pm,k} = (u_k \pm v_k)$ do not change sign during training, then $|\tilde{\theta}_0| < \alpha^2$ and $\Delta_\infty < \Delta_0$.*

The above corollary implies that the recovered solution θ^{SMGD} must perform at least as well as the gradient flow interpolator θ^{GF}. However, in contrast to the continuous case and even though we believe it to be true, we were unable to prove that the SMGD iterates do not change sign for small values of λ.

9.6 Conclusion

Considering an appropriate second-order differential equation which discretises into MGD, we highlight the existence of a single key quantity $\lambda = \gamma/(1-\beta)^2$ which fully determines the trajectory of MGF. This continuous-time perspective also provides a simple acceleration rule and insight into several relevant optimisation regimes. Then, focusing on 2-layer diagonal linear networks, we prove that the asymptotic balancedness Δ_∞ solely governs the generalisation performances of MGF and SMGD. We additionally prove that small values of λ aid the recovery of sparse MGF solutions. Future work should consider MGF/MGD optimisation on more complex architectures and understand precisely the non-trivial effect of λ on the asymptotic balancedness Δ_∞.

Part IV

Mirror flow over classification tasks

Chapter 10

Implicit bias of mirror flow on separable data

10.1 Preface

This chapter follows a paper which is currently under review.

Summary We examine the continuous-time counterpart of mirror descent, namely mirror flow, on classification problems which are linearly separable. Such problems are minimised 'at infinity' and have many possible solutions; we study which solution is preferred by the algorithm depending on the mirror potential. For exponential tailed losses and under mild assumptions on the potential, we show that the iterates converge in direction towards a ϕ_∞-maximum margin classifier. The function ϕ_∞ is the *horizon function* of the mirror potential and characterises its shape 'at infinity'. When the potential is separable, a simple formula allows to compute this function. We analyse several examples of potentials and provide numerical experiments highlighting our results.

Co-authors Radu-Alexandru Dragomir and Nicolas Flammarion.

Contributions Scott and Radu-Alexandru worked together on the project.

10.2 Introduction

Heavily over-parametrised yet barely regularised neural networks can easily perfectly fit a noisy training set while still performing very well on unseen data [Zhang et al., 2017]. This statistical phenomenon is surprising since it is known that there exists interpolating solutions which have terrible generalisation performances [Liu et al., 2020a]. To understand this benign overfitting, it is essential to take into account the training algorithm. If overfitting is indeed harmless, it must be because the optimisation process has steered us towards a solution with favorable generalisation properties.

From this simple observation, a major line of work studying the *implicit regularisation* of gradient methods has emerged. These results show that the recovered solution enjoys some type of low norm property in the infinite space of interpolating solution. Gradient descent (and its variations) has therefore been analysed in various settings, the simplest and most emblematic being that of gradient descent for least-squares regression: it converges towards the solution which has the lowest ℓ_2 distance from the initialisation [Lemaire, 1996]. In the classification setting with linearly separable data, iterates of gradient methods must diverge to infinity to minimise the loss. Therefore, the directional convergence of the iterates is considered and Soudry et al. [2018] show in their seminal paper that gradient descent selects the ℓ_2-max-margin solution amongst all classifiers.

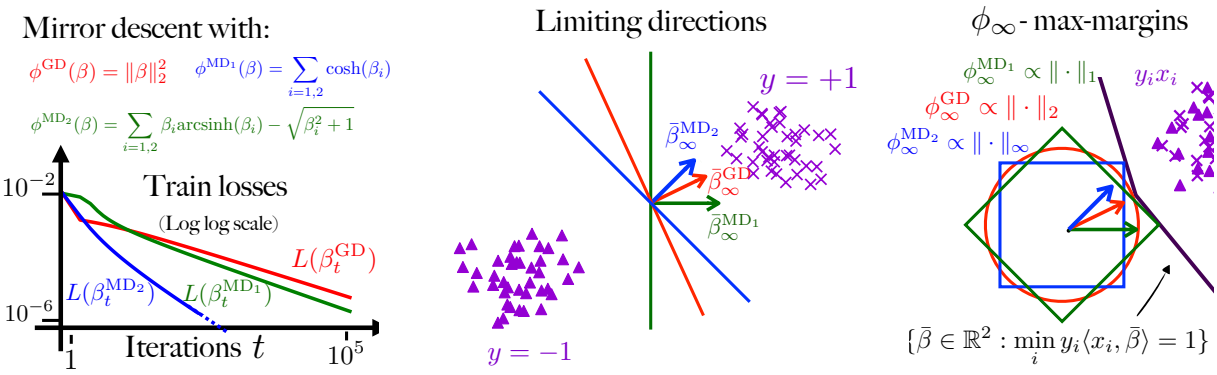


Figure 10.1: Mirror descent is performed using 3 different potentials on the same toy 2d dataset. *Left:* the losses converge to zero. *Center:* the iterates converge in direction towards 3 different vectors, the 3 lines passing through the origin correspond to the 3 different separating hyperplanes. *Right:* these directions are each proportional to $\arg\min \phi_\infty(\bar{\beta})$ under the constraint $\min_i y_i\langle x_i, \bar{\beta}\rangle \geq 1$ for their respective ϕ_∞'s, as predicted by our theory (Theorem 5). See Section 10.6 for more details on the experimental setup.

Going beyond linear settings, it has been observed that **an underlying mirror-descent structure very recurrently emerges** when analysing gradient descent in a wide range of non-linear parametrisations [Woodworth et al., 2020b, Azulay et al., 2021]. Providing convergence and implicit regularisation results for mirror descent has therefore gained significant importance.

In this context, for linear regression, Gunasekar et al. [2018a] show that the iterates converge to the solution that has minimal Bregman distance to the initial point. Turning towards the classification setting, an apparent gap emerges as there is still no clear understanding of what happens: Can directional convergence be characterised in terms of a max-margin problem? And if so, what is the associated norm? Quite surprisingly, this question remains unanswered. This chapter bridges this gap by formally characterising the implicit bias of mirror descent for separable classification problems.

10.2.1 Informal statement of the main result

For a separable dataset $(x_i, y_i)_{i\in[n]}$, we study the mirror flow $\mathrm{d}\nabla\phi(\beta_t) = -\nabla L(\beta_t)\mathrm{d}t$ with potential $\phi : \mathbb{R}^d \to \mathbb{R}$ and an exponential tailed classification loss L. We prove that β_t converges in direction towards the solution of the ϕ_∞-maximum margin solution where the (asymmetric) norm ϕ_∞ captures the shape of the potential ϕ 'at infinity' (see Figure 10.2 for an intuitive illustration).

Theorem 5 (Main result, Informal)**.** *There exists a* ***horizon function*** ϕ_∞ *such for any separable dataset, the normalised mirror flow iterates* $\bar{\beta}_t := \beta_t/\|\beta_t\|$ *converge and satisfy:*

$$\lim_{t\to\infty} \bar{\beta}_t \quad \textit{is proportional to} \quad \arg\max_{\phi_\infty(\bar{\beta})\leq 1} \ \min_{i\in[n]} \ y_i\langle x_i, \bar{\beta}\rangle.$$

This also means that $\lim_{t\to\infty}\bar{\beta}_t$ *is proportional to* $\arg\min_{\min_i y_i\langle x_i,\bar{\beta}\rangle\geq 1} \phi_\infty(\bar{\beta})$ *because* ϕ_∞ *is positively 1-homogeneous.*

Our result holds for (nearly) all reasonable potentials ϕ and it recovers previous results which were obtained for $\phi = \|\cdot\|_p^p$ [Sun et al., 2022] and for L-homogeneous potentials [Sun et al., 2023]. For general potentials, showing convergence towards a maximum margin classifier is much harder

because, in stark contrast with homogeneous potentials, ϕ's geometry changes as the iterates diverge. To capture the behaviour of ϕ at infinity, we geometrically construct its horizon function ϕ_∞. By considering ϕ's successive level sets (and re-normalising them to prevent them blowing-up), we show that under mild assumptions, these sets asymptotically converge towards a limiting horizon set S_∞. The horizon function ϕ_∞ is then simply the asymmetric norm which has S_∞ as its unit ball (see Figure 10.3 for an illustration). In addition, when the function ϕ is 'separable' and can be written $\phi(\beta) = \sum_i \varphi(\beta_i)$ for a real valued function φ, then a very simple and practical formula enables to calculate ϕ_∞ (Theorem 7).

The chapter is organised as follows. The classification setting as well as the assumptions on the loss and the potential are provided in Section 10.3. The proof sketch and an intuitive construction of the horizon function are given in Section 10.4. In Section 10.5, we state the formal definition and results. Simple examples of horizon potentials and numerical experiments supporting our claims are finally given Section 10.6.

10.2.2 Relevance of mirror descent and related works.

We first provide motivations for why understanding the implicit regularisaton of mirror descent is relevant to the machine learning community, as well as related works that places our contribution in context.

Relevance of studying mirror descent in the context of machine learning. Though mirror descent is not *per se* an algorithm used by machine learning practitioners, it proves to be a very useful tool for theoreticians in the field. Indeed, when analysing gradient descent (and its stochastic and accelerated variants) on neural-network architectures, an underlying mirror-descent-like structure somehow very recurrently emerges. General results for mirror descent then enable to prove convergence as well as characterise the implicit regularisation of gradient descent for these architectures. Diagonal linear networks, which are ideal proxy models for gaining insights on complex deep-learning phenomenons, is the most notable example of such an architecture. The hyperbolic entropy potential natural appears and enables to prove countless results: implicit bias of gradient descent in regression [Woodworth et al., 2020b, Vaskevicius et al., 2019] and in classification [Moroshko et al., 2020], effect of stochasticity [Pesme et al., 2021], convergence of gradient descent and effect of the step-size [Even et al., 2023], saddle-to-saddle dynamics [Pesme and Flammarion, 2023]. Unveiling an underlying mirror-like structure goes beyond these simple networks as they also appear in: matrix factorisation with commuting observations [Gunasekar et al., 2017, Wu and Rebeschini, 2021], fully connected linear networks [Azulay et al., 2021, Varre et al., 2023] and 2-layer ReLU networks [Chizat and Bach, 2020]. Building on these examples, Li et al. [2022] further broaden the scope of mirror descent for implicit bias by investigating the formal conditions that ensure the existence of a mirror flow reformulation for general parameterizations, extending previous results by Amid and Warmuth [2020a,b].

Gradient descent in classification. Numerous works have studied gradient descent in the classification setting. For linear parametrisations, separable data and exponentially tailed losses, Soudry et al. [2018] prove that GD converges in direction towards the ℓ_2-maximum margin classifier and provides convergence rates. A very fine description of this divergence trajectory is conducted by Ji and Telgarsky [2018] and a different primal-dual analysis leading to tighter rates is given by Ji and Telgarsky [2021]. Similar results are proven for stochastic gradient descent by Nacson et al. [2019a]. In the case of general loss tails, Ji et al. [2020] prove that gradient descent asymptotically follows the ℓ_2-norm regularisation path. A whole 'astral theory' is developed by

Dudík et al. [2022] who provide a framework which enables to handle 'minimisation at infinity'. Beyond the linear case, Lyu and Li [2020] proves for homogeneous neural networks that any directional limit point of gradient descent is along a KKT point of the ℓ_2-max margin problem. A weaker version of this result was previously obtained by Nacson et al. [2019b]. Furthermore, convergence results for linear networks are provided by Yun et al. [2021]. Finally, for 2-layer networks in the infinite width limit, assuming directional convergence, Chizat and Bach [2020] proves that the limit can be characterised as a max-margin classifier in a certain space of functions.

10.2.3 Notations

We provide here a few notations which will be useful throughout the chapter. We let $[n]$ be the integers from 1 to n. We denote by $Z \in \mathbb{R}^{n\times d}$ the feature matrix whose i^{th} line corresponds to datapoint $y_i x_i$. When not specified, $\|\cdot\|$ corresponds to any (definable) norm on $\mathbb{R}^d$. For a convex function h, $\partial h(\beta)$ denotes its subdifferential at β: $\partial h(\beta) = \{g \in \mathbb{R}^d : h(\beta') \geq h(\beta) + \langle g, \beta' - \beta\rangle, \forall \beta' \in \mathbb{R}^d\}$. For any scalar function $f : \mathbb{R} \to \mathbb{R}$ and vector $u \in \mathbb{R}^p$, the vector $f(u) \in \mathbb{R}^p$ corresponds to the component-wise application of f over u. We denote by $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ the softmax function equal to $\sigma(z) = \exp(z)/\sum_{i=1}^n \exp(z_i) \in \Delta_n$ where Δ_n is the unit simplex. For a convex potential ϕ, we denote $D_\phi(\beta, \beta_0)$ the Bregman divergence equal to $\phi(\beta) - (\phi(\beta_0) + \langle \nabla\phi(\beta_0), \beta - \beta_0\rangle) \geq 0$.

10.3 Problem set-up

We consider a dataset $(x_i, y_i)_{1\leq i\leq n}$ with points $x_i \in \mathbb{R}^d$ and binary labels $y_i \in \{-1, 1\}$. We choose a loss function $\ell : \mathbb{R} \to \mathbb{R}$ and seek to minimize the empirical risk

$$L(\beta) = \sum_{i=1}^{n} \ell(y_i \langle x_i, \beta\rangle).$$

We propose to study the dynamics of mirror flow, which is the continuous-time limit of the *mirror descent* algorithm [Beck and Teboulle, 2003]. Mirror descent is a generalization of gradient descent to non-Euclidean geometries induced by a given convex potential function $\phi : \mathbb{R}^d \to \mathbb{R}$. The method generates a sequence $(\hat{\beta}_k)_{k\geq 0}$ with $\hat{\beta}_0 = \beta_0 \in \mathbb{R}^d$ and

$$\nabla\phi(\hat{\beta}_{k+1}) = \nabla\phi(\hat{\beta}_k) - \gamma \nabla L(\hat{\beta}_k).$$

When the step size γ goes to 0, the mirror descent iterates approach the solution $(\beta_t)_{t\geq 0}$ to the following first-order differential equation (ODE):

$$\mathrm{d}\nabla\phi(\beta_t) = -\nabla L(\beta_t)\mathrm{d}t, \tag{MF}$$

initialised at β_0. Studying the mirror flow (MF) allows for simpler computations than its discrete counterpart, and still allows to obtain rich insights about the algorithm's behaviour.

We now state our standing assumptions on the loss function ℓ and potential ϕ.

Assumption 11. *The loss ℓ satisfies:*

1. *ℓ is convex, twice continuously differentiable, decreasing and $\lim_{z\to+\infty} \ell(z) = 0$.*
2. *ℓ has an exponential tail, in the sense that $-\ell'(z) \underset{z\to\infty}{\sim} \exp(-z)$.*

The first part of the assumptions is very general and insures that the empirical loss L can be minimised 'at infinity'. The exponential tail is crucial: it enables to identify a unique maximum margin solution towards which the iterates converge in direction, independently of the considered loss. Both the exponential $\ell(z) = \exp(-z)$ and the logistic loss $\ell(z) = \ln(1+\exp(-z))$ satisfy the conditions. On the other hand, losses with polynomial tails do not satisfy the second criterion. Similar assumptions on the tail appear when investigating the implicit bias of gradient descent for separable data [Soudry et al., 2018, Nacson et al., 2019c, Ji et al., 2020, Ji and Telgarsky, 2021, Chizat and Bach, 2020].

Assumption 12. *The potential $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfies:*

1. *ϕ is twice continuously differentiable, strictly convex and coercive.*
2. *$\nabla\phi$ diverges at infinity: $\lim_{\|\beta\|\to\infty} \|\nabla\phi(\beta)\| = +\infty$.*
3. *$\nabla^2\phi(\beta)$ is positive-definite for all $\beta \in \mathbb{R}^d$.*
4. *for every $c \in \mathbb{R}_{\geq 0}$ and $\beta_2 \in \mathbb{R}^d$, the sub-level set $\{\beta_1 \in \mathbb{R}^d, D_\phi(\beta_2, \beta_1) \leq c\}$ is bounded.*

These assumptions are common when considering mirror descent [Bauschke et al., 2017]. We provide several examples of potentials in Section 10.6. Crucially, these assumptions ensure the existence and uniqueness of (MF) as shown in the following lemma.

Lemma 4. *For any initialisation $\beta_0 \in \mathbb{R}^d$, there exists a unique solution defined over $\mathbb{R}_{\geq 0}$ which satisfies (MF) for all $t \geq 0$ and with initial condition $\beta_{t=0} = \beta_0$.*

The proof is standard and relies on ensuring that the iterates do not diverge in finite time, we defer it to Appendix E.1. Finally, we assume that $\inf_\beta L(\beta) = 0$, meaning that there exists a hyperplane that perfectly separates the data.

Assumption 13. *The dataset is linearly separable: there exists $\beta^\star \in \mathbb{R}^d$ such that $y_i\langle\beta^\star, x_i\rangle > 0$ for every $i \in [n]$.*

Notice that such $\beta^\star$'s correspond to minimisation directions: $L(\lambda\beta^\star) \overset{\lambda\to\infty}{\longrightarrow} 0$. Under the three previous assumptions, we can show that the mirror flow iterates $(\beta_t)_{t\geq 0}$ minimise the loss while diverging to infinity.

Proposition 22. *Considering the mirror flow $(\beta_t)_{t\geq 0}$, the loss converges towards 0 and the iterates diverge: $\lim_{t\to\infty} L(\beta_t) = 0$ and $\lim_{t\to\infty} \|\beta_t\| = +\infty$.*

The proof relies on classical techniques used to analyse gradient methods in continuous time and we defer the proof to Appendix E.1. We now turn to the main question addressed in this chapter:

Among all minimising directions $\beta^\star$, towards which does the mirror flow converge?

We initially offer a heuristic and intuitive answer to this question, setting the stage for the formal construction of the implicit regularisation problem.

10.4 Intuitive construction of the implicit regularisation problem

In this section, we give an informal presentation and proof sketch of our main result. A fully rigorous exposition is then provided in Section 10.5.

Preliminaries. Assume here for simplicity that $\ell(z) = \exp(-z)$. The mirror flow then writes

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(\beta_t) = L(\beta_t) \cdot Z^T q(\beta_t),$$

with $q(\beta_t) = \sigma(-Z\beta_t)$. We recall that σ is the softmax function and Z the matrix with rows $(y_i x_i)_{i\in[n]}$. Note that $q(\beta_t)$ belongs to the unit simplex Δ_n.

We simplify the differential equation by performing a time rescaling, which does not change the asymptotical behaviour. As $\theta : t \mapsto \int_0^t L(\beta_s)\mathrm{d}s$ is a bijection in $\mathbb{R}_{\geq 0}$ (see Lemma 40) which typically slowly grows as $\ln(t)$, we can speed up time and consider the accelerated iterates $\tilde{\beta}_t = \beta_{\theta^{-1}(t)}$ [1]. By the chain rule, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(\tilde{\beta}_t) = Z^\top q(\tilde{\beta}_t),$$

and therefore

$$\frac{1}{t}\nabla\phi(\tilde{\beta}_t) = \frac{1}{t}\nabla\phi(\beta_0) - Z^\top\Big(\frac{1}{t}\int_0^t q(\tilde{\beta}_s)\mathrm{d}s\Big). \tag{10.1}$$

From now on, we drop the tilde notation and assume that change of time scale has been done. We want to characterize the directional limit of the diverging iterates β_t. To do so, we study their normalisation $\bar{\beta}_t \coloneqq \frac{\beta_t}{\|\beta_t\|}$. As they form a bounded sequence, and $q(\beta_t) \in \Delta_n$ is also bounded, we can extract a subsequence $(\bar{\beta}_{t_s}, q(\beta_{t_s}))_{s>0}$, with $\lim_{s\to\infty} t_s = \infty$ converging to some limit $(\bar{\beta}_\infty, q_\infty)$. By the Césaro average property, $\frac{1}{t_s}\int_0^{t_s} q(\beta_s)\mathrm{d}s$ also converges towards q_∞. Equation (10.1) then yields

$$\frac{1}{t_s}\nabla\phi(\beta_{t_s}) \underset{s\to\infty}{\longrightarrow} Z^\top q_\infty \tag{10.2}$$

Observe that $q(\beta) = \sigma(-Z\beta_t)$ and the softmax function σ approaches the argmax operator at infinity. Hence, as β_t diverges, we expect that $q(\beta_t)_k \to 0$ for coordinates k for which $(-Z\beta_t)_k$ is not maximal, *i.e.* $(Z\beta_t)_k$ not minimal. This observation is made formal in the following lemma. Its proof is straightforward and is given in Appendix E.1.

Lemma 5. *Assume that* $(\bar{\beta}_{t_s}, q(\beta_{t_s})) \overset{s\to\infty}{\longrightarrow} (\bar{\beta}_\infty, q_\infty)$. *It holds that:*

$$(q_\infty)_k = 0 \quad \textit{if} \quad y_k\langle x_k, \bar{\beta}_\infty\rangle > \min_{1\leq i\leq n} y_i\langle x_i, \bar{\beta}_\infty\rangle.$$

In words, coordinates of q_∞ which do not correspond to support vectors of $\bar{\beta}_\infty$ must be zero. Our goal is now to uniquely characterise $\bar{\beta}_\infty$ as the solution of a maximum margin problem.

10.4.1 Warm-up: gradient flow

As a warm-up, let us consider standard gradient flow, which corresponds to mirror flow with potential $\phi = \|\cdot\|_2^2/2$. In this case, Equation (10.2) becomes $\beta_{t_s}/t_s \to Z^\top q_\infty$. Since the normalized iterates satisfy $\bar{\beta}_{t_s} \to \bar{\beta}_\infty$, we get

$$\bar{\beta}_\infty = \frac{Z^\top q_\infty}{\|Z^\top q_\infty\|_2}.$$

[1] $\tilde{\beta}_t$ can also be seen as the mirror flow trajectory but on the log-sum-exp function instead of the sum-exp function

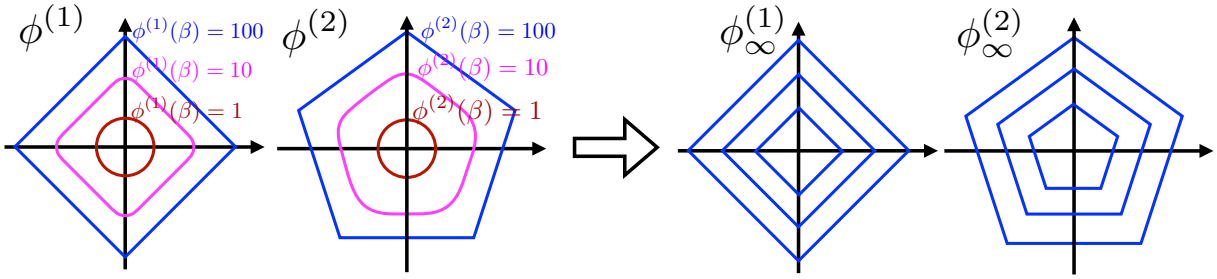


Figure 10.2: *Left two:* Sketch of the level lines of two different potentials $\phi^{(1)}, \phi^{(2)} : \mathbb{R}^2 \to \mathbb{R}$. *Right two:* Their corresponding asymptotic functions $\phi^{(1)}_\infty$, $\phi^{(1)}_\infty$ which are positive 1-homogeneous.

Now notice that this equation along with the slackness conditions from Lemma 5 exactly correspond to the optimality conditions of the following convex minimisation problem:

$$\max_{\bar{\beta}} \ \min_{i\in[n]} \ y_i\langle x_i, \bar{\beta}\rangle \quad \text{under the constraint} \quad \|\bar{\beta}\|_2 \leq 1. \tag{10.3}$$

Furthermore, the ℓ_2-unit ball being strictly convex, Problem (10.3) has a unique solution to which $\bar{\beta}_\infty$ must therefore be equal. Importantly notice that Problem (10.3) uniquely defines the limit of **any extraction** on the normalised iterates $\bar{\beta}_t$: the normalised iterates $\bar{\beta}_t$ must therefore converge towards the ℓ_2-maximum margin and we recover the implicit regularization result from Soudry et al. [2018]:

$$\bar{\beta}_\infty = \arg\max_{\|\bar{\beta}\|_2\leq 1} \ \min_{i\in[n]} \ y_i\langle x_i, \bar{\beta}\rangle.$$

10.4.2 General potential: introducing the horizon function ϕ_∞

We now tackle general potentials ϕ. We first need to introduce the definition of an asymmetric norm. Simply said, an asymmetric norm has all the properties of a norm except for being centrally symmetric.

Definition. [Asymmetric norm.] A function $p : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is said to be an asymmetric norm if we have $p(\beta_1 + \beta_2) \leq p(\beta_1) + p(\beta_2)$ (triangle inequality), $p(\beta) > 0$ for $\beta \neq 0$ (positivity), and $p(r\beta) = rp(\beta)$ for $r \geq 0$ (positive homogeneity).

The challenge of identifying the max-margin problem, to which the iterates converge in direction, stems from the fact that if the potential ϕ is not homogeneous[2] then its geometry changes as the iterates norm increases. More formally, its sub-level sets $S_c := \{\beta \in \mathbb{R}^d, \phi(\beta) \leq c\}$ change of shape as c increase, as illustrated by Figure 10.2 (Left).

However, we can hope that these sets have a **limiting shape** at infinity, meaning that the normalised sub-level sets $\bar{S}_c := S_c/R_c$ where $R_c := \max_{\beta\in S_c} \|\beta\|$ converge to some limiting convex set S_∞ as $c \to \infty$. We can then construct an asymmetric norm ϕ_∞ which has S_∞ as its unit ball. **In words, ϕ_∞ captures the shape of ϕ at infinity.** This informal construction is made rigorous in Section 10.5.1. We state here the crucial consequence of this construction.

Corollary 4. *The horizon function ϕ_∞ is such such that for any sequence β_t for which $\bar{\beta}_t := \frac{\beta_t}{\|\beta_t\|}$ and $\frac{\nabla\phi(\beta_t)}{t}$ both converge, then:*

$$\lim_{t\to\infty} \frac{1}{t}\nabla\phi(\beta_t) \in \lambda \cdot \partial\phi_\infty(\bar{\beta}_\infty), \quad \textit{where} \quad \bar{\beta}_\infty = \lim_{t\to\infty} \bar{\beta}_t,$$

[2] a function is said to be homogeneous if there exists $L > 0$ such that $\phi(c\beta) = c^L\phi(\beta)$ for all β and $c > 0$

for some positive factor λ.

Using this construction, we are able to derive the optimality conditions satisfied by $\bar{\beta}_\infty$. From the convergence in Equation (10.2) and that of $\bar{\beta}_t \to \bar{\beta}_\infty$, applying Corollary 4, we get that:

$$Z^\top q_\infty \in \lambda \cdot \partial\phi_\infty(\bar{\beta}_\infty).$$

Up to a positive multiplicative factor (which is irrelevant due to the positive homogeneity the quantities involved), this condition along with Lemma 5 are exactly the optimality conditions of the convex problem

$$\max_{\bar{\beta}\in\mathbb{R}^d} \min_{i\in[n]} y_i\langle x_i, \bar{\beta}\rangle \quad \text{under the constraint} \quad \phi_\infty(\bar{\beta}) \leq 1.$$

$\bar{\beta}_\infty$ must therefore belong to the set of its solutions. Assuming that this set contains a single element of norm 1 (we refer to the next section for comments concerning the unicity), we obtain that the iterates $\bar{\beta}_t$ must converge towards it:

$$\lim_{t\to\infty} \frac{\beta_t}{\|\beta_t\|} \propto \arg\max_{\phi_\infty(\bar{\beta})\leq 1} \min_{i\in[n]} y_i\langle x_i, \bar{\beta}\rangle.$$

10.5 Main result: directional convergence of the iterates towards the ϕ_∞-max margin

We now state our formal results, begining with the precise construction of the horizon function ϕ_∞, and following with the theorem showing convergence of the iterates towards the ϕ_∞-max margin.

10.5.1 Construction of the horizon function ϕ_∞

We first define the **horizon shape** of a convex potential, and provide sufficient conditions for its existence. Then, we use this shape to construct a **horizon function** ϕ_∞, which allows the interpretation of the directional limits of gradients of ϕ at infinity. The proofs require technical elements from variational analysis to ensure that the limits are well-defined; these are deferred to the appendix.

Definition and sufficient conditions for existence. Assume without loss of generality that $\phi(0) = 0$. For $c \geq 0$, consider the sublevel set

$$S_c(\phi) = \{\beta \in \mathbb{R}^d \,:\, \phi(\beta) \leq c\},$$

which is nonempty and compact by coercivity of ϕ. We can then define the normalised sublevel set:

$$\bar{S}_c = \frac{1}{R_c} S_c, \quad R_c = \max\{\|\beta\| \,:\, \beta \in S_c\}. \tag{10.4}$$

By construction, the set $\bar{S}_c$ belongs to the unit ball. We are interested in the limit of $\bar{S}_c$ as $c \to \infty$.

Definition. We say that ϕ admits a horizon shape if the family of normalized sublevel sets $(\bar{S}_c)_{c>0}$ defined in Equation (10.4) converges to some set S_∞ as $c \to \infty$ for the Hausdorff distance. In addition, we say that this shape is non-degenerate if the origin belongs to the interior of S_∞.

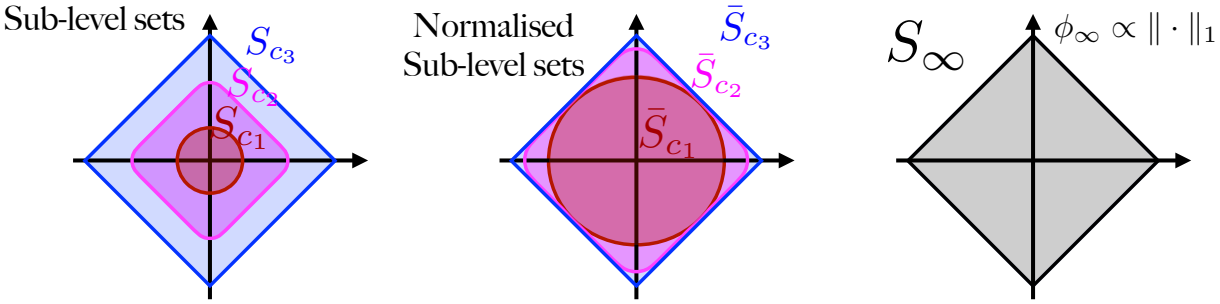


Figure 10.3: Illustration of the construction of the horizon shape S_∞. *Left:* the sub-level sets S_c change of shape and are increasing. *Middle:* in order to avoid the shapes blowing up, we normalized them to keep them in the unit ball (here we chose the arbitray constraining norm to be the ℓ_1-norm). *Right*: the normalised sub-level sets $\bar{S}_c$ converge to a limiting set S_∞ for the Hausdorff distance.

Hausdorff distance is a natural distance on compact sets [see Rockafellar and Wets, 1998, Section 4.C., for a definition]. In Proposition 10.5.3, we prove the existence of the horizon shape for a large class of functions, containing all potentials encoutered in practice.

Although the horizon shape is guaranteed to exist for most functions, we cannot a priori prove that it is non-degenerate, as the normalized sub-levels $\bar{S}_c$ can become 'flat' as $c \to \infty$[3]. As this case would be much more technical and involved, we now restrict to non-degenerate horizon shapes.

Horizon function. If h admits a non-degenerate horizon shape S_∞, we define its horizon function as the *Minkowski gauge* [Rockafellar and Wets, 1998, Section 11.E] of S_∞:

$$\phi_\infty(\bar{\beta}) \coloneqq \inf\{r > 0 \,:\, \frac{\bar{\beta}}{r} \in S_\infty\}$$

for $\bar{\beta} \in \mathbb{R}^d$. The horizon function ϕ_∞ is an asymmetric norm as defined in Definition . Importantly, it is defined such that its sub-level sets correspond to scaled versions of S_∞ [see Rockafellar and Wets, 1998, Section 11.C, for more properties]. For example, in the case of the horizon shape S_∞ illustrated in Figure 10.3, the corresponding horizon function ϕ_∞ is proportional to the ℓ_1-norm. Although the construction of ϕ_∞ presented here is rather abstract, we show in Theorem 7 that for separable potentials, as commonly encountered in practice, it can be computed with a simple formula. Though different, our definition of the horizon function shares many similarities with the classical concept of horizon function from convex analysis [Rockafellar and Wets, 1998]. We discuss the links between the two notions at the end of Section 10.5.3.

10.5.2 ϕ_∞-max margin problem and main result

Now that we have properly constructed the horizon function ϕ_∞, we can define its corresponding maximum-margin problem.

Definition. The ϕ_∞-max margin problem is defined as:

$$\max_{\bar{\beta}\in\mathbb{R}^d} \min_{i\in[n]} y_i\langle x_i, \bar{\beta}\rangle \quad \text{under the constraint} \quad \phi_\infty(\bar{\beta}) \leq 1.$$

We can now state our main result which fully characterises the directional convergence of the mirror flow iterates.

[3]Consider for instance $\phi(x, y) = x^2 + y^4$ on $\mathbb{R}^2$, for which the horizon shape is $[-1, 1] \times \{0\}$.

Theorem 6. *Let ϕ_∞ be the horizon function of ϕ. Assuming that the ϕ_∞-max margin problem has a unique solution, the mirror flow normalised iterates $\bar{\beta}_t = \frac{\beta_t}{\|\beta_t\|}$ converge towards a vector $\bar{\beta}_\infty$ and*

$$\bar{\beta}_\infty \propto \arg\max_{\phi_\infty(\bar{\beta})\leq 1} \min_{i\in[n]} y_i\langle x_i, \bar{\beta}\rangle,$$

where the symbol $\propto$ denotes positive proportionality.

Remark on the unicity of the margin problem. If the unit ball of ϕ_∞ is strictly convex, then the max-margin problem has a unique maximiser. However, in the general case, there may exist an infinity of solutions and weak but ad hoc assumptions on the dataset are required to guarantee uniqueness. For instance, if ϕ_∞ is proportional to the ℓ_1-norm, a common assumption which ensures uniqueness is assuming that the data is in *general position*. It is not restrictive as it is almost surely satisfied for data drawn from a continuous probability distribution, we refer to Rosset et al. [2004], Appendix B, for more details.

10.5.3 Assumptions guaranteeing the existence of ϕ_∞ and computable formula

Our main result, presented in Theorem 6 relies on the existence of a horizon shape, S_∞, as described in Definition . From this shape, the asymmetric norm ϕ_∞ is constructed.

We show here that the existence of S_∞ is ensured for a large class of 'nice' functions, specifically those *definable in o-minimal structures* [Dries, 1998]. For the reader unfamiliar with this notion, this class contains all 'reasonable' functions used in practice, such as polynomials, logarithms, exponentials, and 'reasonable' combinations of those. This is a typical assumption used for instance to prove the convergence of optimisation methods through the Kurdyka–Łojasiewicz property [Attouch et al., 2011].

If any of the three following conditions hold: *(i)* ϕ is a finite composition of polynomials, exponentials and logarithms, *(ii)* ϕ is globally sub-analytic, *(iii)* ϕ is definable in a o-minimal structure on $\mathbb{R}$; then ϕ admits a horizon shape S_∞. The proof is technical and we defer it to Appendix E.2. Although the previous proposition ensures the existence ϕ_∞ for a wide range of potentials, it does not offer a direct method for computing it. In the following, we show that for potentials that are both separable and even, a simple formula exists, allowing for the direct calculation of ϕ_∞.

Assumption 14. *The potential ϕ is separable in the sense that there exists $\varphi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that $\phi(\beta) = \sum_{i=1}^d \varphi(\beta_i)$. We assume that φ satisfies Assumption 12, that it is definable in a o-minimal structure on $\mathbb{R}$ and that it is an even function. W.l.o.g. we assume that $\varphi(0) = 0$.*

We note that φ is bijective over $\mathbb{R}_{\geq 0}$ from Assumption 12, and we denote φ^{-1} its inverse. We consider the function $\varphi^{-1} \circ \phi$, which has the suitable properties of: *(i)* maintaining the same sub-level sets as ϕ and *(ii)* not growing 'too fast', ensuring that $\lim_{\eta\to 0} \eta\varphi^{-1}(\phi(\bar{\beta}/\eta))$ exists in $\mathbb{R}_{>0}$ for all $\bar{\beta}$. These two observations lead to the following theorem.

Theorem 7. *Under Assumption 14, there exists $\lambda > 0$ such that the horizon function ϕ_∞ of ϕ as defined in the previous section satisfies:*

$$\phi_\infty(\bar{\beta}) = \lambda \lim_{\eta\to 0} \eta \cdot \varphi^{-1}\left(\phi\left(\frac{\bar{\beta}}{\eta}\right)\right)$$

for every $\bar{\beta} \in \mathbb{R}^d$.

We use this simple formula when computing ϕ_∞ for various potentials in the next section.

Remark on previous notions of horizon function. In the convex analysis literature, the horizon function is typically defined as $\lim_{\eta\to 0}\eta\phi(\bar{\beta}/\eta)$ [Rockafellar and Wets, 1998, Laghdir and Volle, 1999]. A direct application of this concept in our context would yield a function valued at $+\infty$ everywhere except at the origin, a consequence of Assumption 12. In contrast, our definition ensures that ϕ_∞ attains finite values over at least a portion of $\mathbb{R}^d$. This distinction stems from our way of normalising the level sets by R_c in the general definition, or alternatively, the composition by φ^{-1} in the separable case.

The two constructions would coincide only if ϕ was Lipschitz continuous, which is at odds with Assumption 12. To align more closely with the standard terminology in the literature, we could refer to our notion as the *normalised* horizon function.

10.6 Applications and experiments

In this section, we illustrate our main result using three different potentials. We implement mirror descent with each one of these potentials, and present the results in Figure 10.1.

Homogeneous potentials. We first consider potentials ϕ which are L-homogeneous, i.e., there exists $L > 0$ such that for all $c > 0$ and $\beta \in \mathbb{R}^d$, $\phi(c\beta) = c^L\phi(\beta)$. In this case, since $\bar{S}_c$ is equal for all $c > 0$, it follows that there exists λ such that $\phi_\infty \propto \phi^{1/L}$. An important example is the case of $\phi = \|\cdot\|_p^p$ where $\|\cdot\|_p$ corresponds to the ℓ_p-norm with $p > 1$, for which we get that $\phi_\infty \propto \|\cdot\|_p$ and we recover the result from Sun et al. [2022, 2023].

Hyperbolic entropy potential. The hyperbolic entropy potential $\phi(\beta) = \sum_{i=1}^d(\beta_i\text{arcsinh}(\beta_i) - \sqrt{\beta_i^2+1} - 1)$ plays a central role in works considering diagonal linear networks [Woodworth et al., 2020b, Pesme and Flammarion, 2023]. Applying Theorem 7, we obtain that $\phi_\infty \propto \|\cdot\|_1$ and we recover the result from Moroshko et al. [2020]. We note that the geometry induced by this potential changes across different scales, interpolating between the ℓ_2 and ℓ_1 geometries.

Hyperbolic-cosine entropy potential. We finally consider the following potential $\phi(\beta) = \sum_{i=1}^d(\cosh(\beta_i) - 1)$. Applying Theorem 7, we get that $\phi_\infty \propto \|\cdot\|_\infty$.

Experimental details concerning Figure 10.1. As shown in Figure 10.1 (Middle), we generate 40 points with positive labels and 40 points with negative labels. Starting from $\beta_0 = 0$, we run mirror descent with the three following potentials:

$$(i)\ \ \phi^{\text{GD}} = \|\cdot\|_2^2, \qquad (ii)\ \ \phi^{\text{MD}_1} = \text{Hyperbolic entropy}, \qquad (iii)\ \ \phi^{\text{MD}_2} = \text{cosh-entropy}.$$

We first observe in Figure 10.1 (Left) that the training loss converges to zero, as predicted by Proposition 22, with a linear convergence rate that varies across different potentials. Moreover, as illustrated in Figure 10.1 (Middle and Right), the iterates converge in direction towards their respective unique ϕ_∞-max margin solutions associated with the following geometries:

$$(i)\ \ \phi_\infty^{\text{GD}} \propto \|\cdot\|_2, \qquad (ii)\ \ \phi_\infty^{\text{MD}_1} = \|\cdot\|_1, \qquad (iii)\ \ \phi_\infty^{\text{MD}_2} \propto \|\cdot\|_\infty.$$

Therefore, by using different Therefore, by employing various potentials, we can induce different implicit biases, leading to distinct generalization properties based on the data distribution.

10.7 Conclusion and limitations

In this chapter, we offer a comprehensive characterisation of the implicit bias of mirror flow for separable classification problems. This characterisation is framed in terms of the horizon function associated with the mirror descent potential, leveraging the asymptotic geometry induced by the potential.

Our results being purely asymptotic, characterising the rate at which the normalised iterates converge towards the maximum-margin solution is an open direction for future research. Furthermore, we note that our analysis does not cover potentials that are defined only on a strict subset of $\mathbb{R}^d$ (such as the log-barrier and the negative entropy), and with possibly non-coercive gradients. This class of potentials is of interest as it arises when investigating deep architectures, such as diagonal linear networks of depth $D > 2$. In this setting, it is known that gradient flow on the weights lead to a mirror flow on the predictors with a certain potential ϕ_D [Woodworth et al., 2020b]. Interestingly, the potentials ϕ_D have non-coercive gradients and their horizon functions do not depend on the depth D as they are all proportional to the ℓ_1-norm. The predictors are, however, known to converge in direction towards a KKT point of the non-convex $\ell_{2/D}$-max-margin problem [**?**] which can be different from the ℓ_1-max-margin problem [Moroshko et al., 2020]. This observation highlights that our coercive gradient assumption is necessary for our result to hold. However, extending our analysis beyond this assumption is a promising direction for understanding gradient dynamics in deep architectures.

Conclusion and future directions

10.7.1 Conclusion

In this PhD thesis, we investigated optimisation phenomena which occur in deep learning by focusing solely on the 2-layer diagonal linear network architecture.

In the first part, we gave an introduction to the notion of implicit regularisation alongside an overview of the 2-layer diagonal linear network architecture and the mirror flow algorithm. We recalled key results upon which the rest of the manuscript is built. This section lays the groundwork for the subsequent results presented in the thesis.

In the following part, in a regression setting, we showed that we can describe the entire gradient flow trajectory for a vanishing initialisation: the iterates jump from a saddle of the training loss to another and coordinates are learnt one at a time, highlighting an incremental learning.

In the third section, we examined the influence of training hyperparameters on the recovered solution. Initially, we employed a stochastic differential equation to model the trajectory of Stochastic Gradient Descent (SGD), demonstrating that noise aids in recovering a sparser solution compared to gradient flow. While we didn't emphasise the role of the step-size extensively here, in the ensuing chapter, we directly tackle the discrete SGD and GD recursions and investigate the impact of the step-size. Our analysis revealed that while large step-sizes are beneficial for sparse recovery when using SGD, they can be counterproductive for GD. In the third and last chapter, we explored the effect of momentum, employing a continuous-time approach to model momentum gradient descent. We identified an intrinsic quantity, $\lambda = \frac{\gamma}{(1-\beta)^2}$, which uniquely characterises the optimisation path and provides a straightforward acceleration rule applicable beyond the scope of the diagonal network architecture. Additionally, we demonstrated that moderate values of λ help in the recovery of sparse solutions.

In the fourth and final part, we adopted a slightly different perspective, diverging from a strict focus on diagonal linear networks. Instead we focused our attention towards understanding mirror flow's implicit regularisation in linear classification tasks with separable data. We proved that the iterates converge directionally towards a ϕ_∞ max margin classifier, where the function ϕ_∞ captures the shape of the potential at infinity.

10.7.2 Future directions

Beyond diagonal linear networks. While the study of diagonal linear networks is rich in insights, it is undoubtedly a limited framework. A clear avenue for future research would involve exploring more complex architectures that bear closer resemblance to real-world models, such as linear networks, ReLU networks, ResNets, and Transformers. We would not be surprised if many of the techniques and methodologies developed within the context of diagonal linear networks can be extended and adapted to these broader frameworks. Indeed, the 2-layer diagonal network already captures a crucial characteristic found in various architectures: weight multiplication. Investigating the non-convexity introduced by other types of weight multiplication, such as

matrix-matrix products, as seen in Transformers and ReLU networks, is a direction for future research.

Exploring the 'data | architecture | gradient method' alignment. As emphasised in the introduction, understanding the generalisation capabilities of deep learning models requires considering the structure of the data: neural networks with gradient training excel in real-world tasks because they align well with the underlying data structures. Drawing a parallel with linear regression: there exists a class of data distributions for which the minimum ℓ_2-norm interpolator generalises well [Bartlett et al., 2020], and another class where it is the minimum ℓ_1-norm interpolator [Wang et al., 2022a]. The question is then: for which class of data-distributions do neural networks with gradient training perform well?

Beyond the classical supervised learning setting. The success of large language models relies on the next-token prediction task they are trained on. Though this task can be cast as supervised learning, it is more commonly designated as 'self-supervised'. However there exists no appropriate statistical framework which enables to formalise and analyse these learning problems. What space of inputs should we consider? What complexity measure can we associate to functions acting on this space? Exploring these questions is an exciting avenue for future research.

From continuous to discrete. The continuous-time framework constitutes a fantastic framework which allows to painlessly prove convergence of the iterates, give convergence speeds as well as expose the implicit regularisation problem. However all known discretisation bounds are overly pessimistic and fail to account for two important observations: *(i)* experimental evidence shows that the discrete algorithms tend to closely follow their continuous counterpart, *(ii)* once the continuous-time proof is obtained, the proof techniques can often be adapted and transferred to the discrete case by using the same ideas and Lyapunov functions. An interesting research avenue would be to understand: *(i)* for which class of functions do favorable discretisation bounds hold? *(ii)* could we build a framework which enables to immediately transfer proofs?

Appendix A

Appendix for Chapter 6

Organisation of the Appendix.

1. In Appendix A.1, we give the experimental setup and provide additional experiments.
2. In Appendix A.2, we prove Proposition 11 and provide additional comments concerning the unicity of the minimisation problem which appears in the proposition.
3. In Appendix A.3, we provide some general results on the flow.
4. In Appendix A.4, we prove Proposition 12 and give standalone properties of Algorithm 1.
5. In Appendix A.5, we explain in more detail the arc-length parametrisation explained in the main text as well as prove Theorem 1 and Proposition 13.
6. In Appendix A.6, we provide technical lemmas which are useful to prove the main results.

A.1 Experimental setup and additional: experiments, extension, related works.

Experimental setup and additional experiments. For each experiment we generate our dataset as $y_i = \langle x_i, \beta^\star \rangle$ where $x_i = \mathcal{N}(\mathbf{0}, H)$ for a a diagonal covariance matrix H and $\beta^\star$ is a vector of $\mathbb{R}^d$. Gradient descent is run with a small step size and from initialisation $u_{t=0} = \sqrt{2}\alpha\mathbf{1} \in \mathbb{R}^d$ and $v_{t=0} = \mathbf{0}$ for some initialisation scale $\alpha > 0$.

- Figure 6.1 and Figure A.2 (Left): $(n, d, \alpha) = (5, 7, 10^{-120})$, $H = I_d$, $\beta^\star = (10, 20, 0, 0, 0, 0, 0) \in \mathbb{R}^7$.
- Figure A.2 (Right): $(n, d, \alpha) = (6, 6, 10^{-10})$, $H = \mathrm{diag}(1, 10, 10, 10, 10, 10) \in \mathbb{R}^{6\times 6}$, $\beta^\star = (1, 0, 0, 0, 0, 0, 0) \in \mathbb{R}^6$.
- Figure A.1 (Left): $(n, d, \alpha_1, \alpha_2) = (7, 2, 10^{-100}, 10^{-10})$, $H = I_d$, $\beta^\star = (10, 20) \in \mathbb{R}^7$.
- Figure A.1 (Right): $(n, d, \alpha) = (3, 3, 10^{-100})$, X is the square root matrix of the matrix $((20, 6, -1.4), (6, 2, -0.4), (-1.4, -0.4, 0.12)) \in \mathbb{R}^{3\times 3}$, $\beta^\star = (1, 9, 10)$.

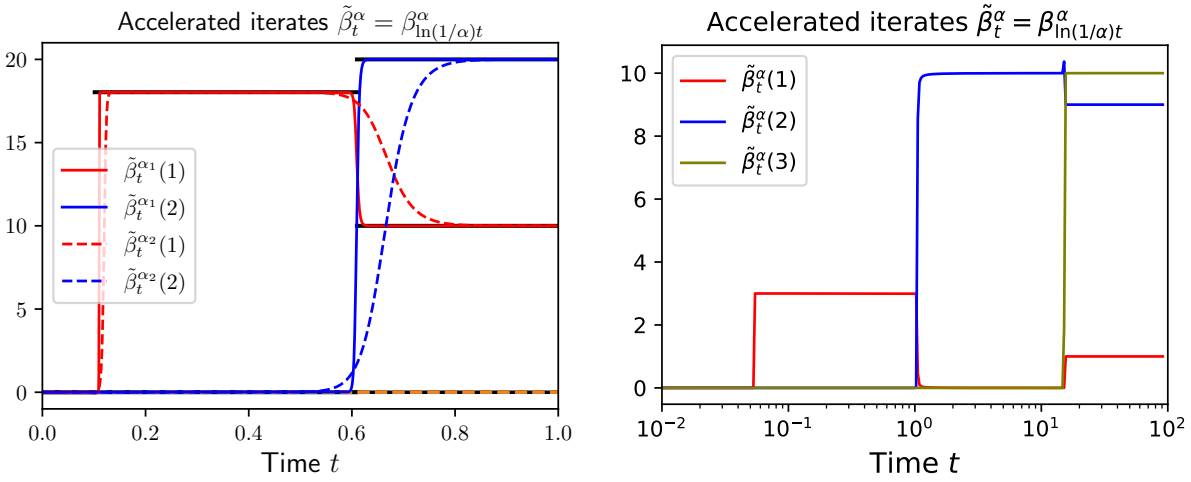


Figure A.1: *Left:* Visualisation of the uniform convergence of $\tilde{\beta}^\alpha$ towards $\tilde{\beta}^\circ$ as $\alpha \to 0$. $\alpha_1 = 10^{-100} \ll \alpha_2 = 10^{-10}$ *Right:* In some cases, 2 coordinates can activate at the same time. Note that the time axis is in log-scale for better visualisation.

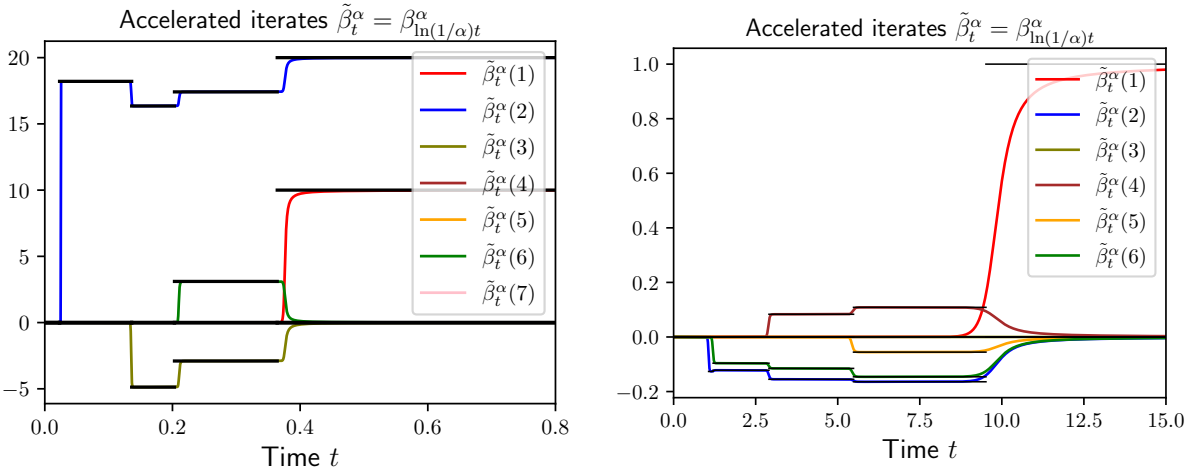


Figure A.2: Complex dynamics can occur. *Left and right:* Coordinates are not monotonic and the number of active coordinates neither as several coordinates can deactivate at the same time. The piecewise constant process plotted in black is the limiting process $\tilde{\beta}^\circ$ predicted by our theory.

A.2 Proof of Proposition 11

Proposition 11. *All the critical points w_c of F which are not global minima, i.e., $\nabla F(w_c) = \mathbf{0}$ and $F(w_c) > \min_w F(w)$, are necessarily saddle points* (i.e., *not local extrema). They map to parameters $\beta_c = u_c \odot v_c$ which satisfy $|\beta_c| \odot \nabla L(\beta_c) = \mathbf{0}$ and:*

$$\beta_c \in \underset{\beta[i]=0 \text{ for } i \notin \mathrm{supp}(\beta_c)}{\arg\min} L(\beta) \tag{6.4}$$

where $\mathrm{supp}(\beta_c) = \{i \in [d], \beta_c[i] \neq 0\}$ *corresponds to the support of* β_c.

Proof. **Non-existence of maxima / non-global minima.** This is a simpler version of results which appear in Kawaguchi [2016], for the sake of completeness we provide here a simple proof adapted to our setting. The intuition follows the fact that if there existed a local maximum / non-global minimum for F then this would translate to the existence of a local maximum / non-global minimum for the convex loss L, which is absurd.

Assume that there exists a local maximum $w^\star = (u^\star, v^\star)$, i.e. assume that there exists $\varepsilon > 0$ such that for all $w = (u, v)$ such that $\|w - w^\star\|_2^2 \leq \varepsilon$, $F(w) \leq F(w^\star)$. We show that this would imply that $\beta^\star = u^\star \odot v^\star$ is a local maximum of L, which is absurd.

The mapping $g : (u, v) \mapsto (u \odot v, \sqrt{(u^2 - v^2)/2})$ from $\mathbb{R}^d_{\geq 0} \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d_{\geq 0}$ is a bijection with inverse

$$g^{-1} : (\beta, \alpha) \mapsto (\sqrt{\alpha^2 + \sqrt{\beta^2 + \alpha^4}}, \mathrm{sign}(\beta) \odot \sqrt{-\alpha^2 + \sqrt{\beta^2 + \alpha^4}}). \tag{A.1}$$

Also notice that $F(g^{-1}(\beta, \alpha)) = L(\beta)$ for all β and α. Now let $\tilde{\varepsilon} > 0$ and let $\beta \in \mathbb{R}^d$ such that

$\|\beta - \beta^\star\|_2^2 \leq \tilde{\varepsilon}$, then for $(u, v) = g^{-1}(\beta, \alpha_*)$ where $\alpha_* = \sqrt{((u^\star)^2 - (v^\star)^2)/2}$ we have that:

$$\begin{aligned}
\|(u, v) - (u^\star, v^\star)\|_2^2 &= 2\left\| \left(\sqrt{\alpha_*^4 + \beta^2} - \sqrt{\alpha_*^4 + \beta^{\star 2}} \right)^2 \right\|_1 \\
&\leq 2\|\beta^2 - \beta^{\star 2}\|_1 \\
&= 2\|(\beta - \beta^\star)^2 + 2(\beta - \beta^\star)\beta^\star\|_1 \\
&\leq 2\|(\beta - \beta^\star)^2\|_1 + 2\|\beta^\star\|_\infty \|\beta - \beta^\star\|_1 \\
&\leq 2(1 + \sqrt{d}\|\beta^\star\|_\infty)\tilde{\varepsilon} \\
&\leq \varepsilon
\end{aligned}$$

where the last inequality is for $\tilde{\varepsilon}$ small enough. This means that $L(\beta) = F(w) \leq F(w^\star) = L(\beta^\star)$ and $\beta^\star$ is a local maximum of L, which is absurd.

The exact same proof holds to show that there are no local minima of F which are not global minima.

Critical points. The gradient of the loss function F writes:

$$\nabla_w F(w) = \begin{pmatrix} \nabla_u F(w) \\ \nabla_v F(w) \end{pmatrix} = \begin{pmatrix} \nabla L(\beta) \odot v \\ \nabla L(\beta) \odot u \end{pmatrix} \in \mathbb{R}^{2d}.$$

Therefore $\nabla F(w_c) = \mathbf{0} \in \mathbb{R}^{2d}$ implies that $\nabla L(\beta_c) \odot \beta_c = \mathbf{0} \in \mathbb{R}^d$. Now consider such a β_c and let $\text{supp}(\beta_c) = \{i \in [d] \text{ such that } \beta_c(i) \neq 0\}$ denote the support of β_c. Since $[\nabla L(\beta_c)]_i = 0$ for $i \notin \text{supp}(\beta_c)$, we can therefore write that

$$\beta_c \in \underset{\beta_i = 0 \text{ for } i \notin \text{supp}(\beta_c)}{\arg\min} L(\beta).$$

Furthermore we point out that since $\text{supp}(\beta_c) \subset [d]$, there are at most 2^d distinct sets $\text{supp}(\beta_c)$, and therefore at most 2^d values $F(w_c) = L(\beta_c)$, where w_c is a critical point of F. □

Additional comment concerning the uniqueness of $\arg\min_{\beta_i = 0, i \notin \text{supp}(\beta_c)} L(\beta)$**.**

We point out that the constrained minimisation problem (6.4) does not necessarily have a unique solution, even when β_c is not a global solution. Though not required for any of our results, for the sake of completeness, we show here that under an additional mild assumption on the data, we can ensure that the minimisation problem (6.4) which appears in Proposition 11 has a unique minimum when $L(\beta_c) > 0$. Under this additional assumption, there is therefore a finite number of saddles β_c. Recall that we let $X \in \mathbb{R}^{n \times d}$ be the feature matrix and $(\tilde{x}_1, \ldots, \tilde{x}_d)$ be its columns. Now assume *temporarily* that the following assumption holds.

Assumption 15 (Assumption used just in this short section)**.** *Any subset of* $(\tilde{x}_1, \ldots, \tilde{x}_d)$ *of size smaller than* $\min(n, d)$ *is linearly independent.*

One can easily check that this assumption holds with probability 1 as soon as the data is drawn from a continuous probability distribution, similarly to Tibshirani [2013, Lemma 4]). In the following, for a subset $\xi = \{i_1, \ldots, i_k\} \subset [d]$, we write $X_\xi = (\tilde{x}_{i_1}, \ldots, \tilde{x}_{i_k}) \in \mathbb{R}^{n \times k}$ (we extract the columns from X). For a vector $\beta \in \mathbb{R}^d$ we write $\beta[\xi] = (\beta_{i_1}, \ldots, \beta_{i_k})$ and $\beta[\xi^C] = (\beta_i)_{i \notin \xi}$. We distinguish two different settings:

- Underparametrised setting ($n \geq d$) : in this case, for any $\xi = \{i_1, \dots, i_k\} \subset [d]$, then $\beta^\star := \underset{\beta_i=0, i\notin\xi}{\operatorname{argmin}} L(\beta)$ is unique. Indeed we simply set the gradient to 0 and notice that due to Assumption 15, there exists a unique solution, indeed it is $\beta^\star$ such that $\beta^\star[\xi] = (X_\xi^\top X_\xi)^{-1} X_\xi^\top y$ and $\beta^\star[\xi^C] = 0$.

- Overparametrised setting ($d > n$) : **Global solutions:** $\arg\min_{\beta\in\mathbb{R}^d} L(\beta)$ is an affine space spanned by the orthogonal of $(x_1, \dots, x_n)$ in $\mathbb{R}^d$. Since $\operatorname{span}(\tilde{x}_1, \dots, \tilde{x}_d) = \mathbb{R}^n$ from Assumption 15, any $\beta^\star \in \arg\min_{\beta\in\mathbb{R}^d} L(\beta)$ satisfies $X\beta^\star = y$ and $L(\beta^\star) = 0$. **"Saddle points":** now let $\beta_c \in \mathbb{R}^d$ be such that we can write $\beta_c \in \arg\min_{\beta_i=0, i\notin\operatorname{supp}(\beta_c)} L(\beta)$ and assume that $L(\beta_c) > 0$ (i.e., not a global solution), then: (1) β_c has at most n non-zero entries, indeed if it were not the case, then y would necessarily belong to $\operatorname{span}(\tilde{x}_i)_{i\in\operatorname{supp}(\beta_c)}$ due to the assumption on the data, and this would lead to $L(\beta_c) = 0$, (2) therefore, similar to the underparametrised case, $\arg\min_{\beta_i=0, i\notin\operatorname{supp}(\beta_c)} L(\beta)$ is unique, equal to β_c, and we have that $\beta_c[\xi] = (X_\xi^\top X_\xi)^{-1} X_\xi^\top y$ and $\beta_c[\xi^C] = 0$ where $\xi = \operatorname{supp}(\beta_c)$.

Thus, in both the underparametrised and overparametrised settings, the minimisation problem (6.4) appearing in Proposition 11 has a unique minimum when $L(\beta_c) > 0$ and Assumption 15 holds.

A.3 General results on the iterates

In the following lemma we recall a few results concerning the gradient flow eq. (8.3):

$$\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t\,, \tag{A.2}$$

where F is defined in eq. (8.2) as:

$$F(w) \coloneqq L(u \odot v) = \frac{1}{2n}\sum_{i=1}^{n}(\langle u \odot v, x_i\rangle - y_i)^2\,.$$

Lemma 6. *For an initialisation $u_0 = \sqrt{2}\alpha$, $v_0 = \mathbf{0}$, the flow $w_t^\alpha = (u_t^\alpha, v_t^\alpha)$ from eq. (A.2) is such that the quantity $(u_t^\alpha)^2 - (v_t^\alpha)^2$ is constant and equal to $2\alpha^2\mathbf{1}$. Furthermore $u_t^\alpha > |v_t^\alpha| \geq 0$ and therefore from the bijection eq. (A.1) we have that:*

$$u_t^\alpha = \sqrt{\alpha^2 + \sqrt{(\beta_t^\alpha)^2 + \alpha^4}}, \quad v_t^\alpha = \operatorname{sign}(\beta_t^\alpha) \odot \sqrt{-\alpha^2 + \sqrt{(\beta_t^\alpha)^2 + \alpha^4}}.$$

Proof. From the expression of $\nabla F(w)$, notice that the derivative of $(u_t^\alpha)^2 - (v_t^\alpha)^2$ is equal to $\mathbf{0}$ and therefore equal to its initial value.

Since $(u_t^\alpha)^2 - (v_t^\alpha)^2 = (u_t^\alpha + v_t^\alpha)(u_t^\alpha - v_t^\alpha) > 0$, by continuity we get that $u_t^\alpha + v_t^\alpha > 0$ and $u_t^\alpha - v_t^\alpha > 0$ and therefore $u_t^\alpha > |v_t^\alpha|$. □

In this section we consider the accelerated iterates eq. (6.9) which follow:

$$\mathrm{d}\nabla\tilde{\phi}_\alpha(\tilde{\beta}_t^\alpha) = -\nabla L(\tilde{\beta}_t^\alpha)\mathrm{d}t, \qquad \text{where} \qquad \tilde{\phi}_\alpha \coloneqq \frac{1}{\ln(1/\alpha)}\cdot\tilde{\phi}_\alpha \tag{A.3}$$

with $\tilde{\beta}_{t=0} = \mathbf{0}$ and where ϕ_α is defined eq. (6.7).

Proposition 23. *For all $\alpha > 0$ and minimum $\beta^\star \in \arg\min_\beta L(\beta)$, the loss values $L(\tilde{\beta}_t^\alpha)$ and the Bregman divergence $\mathrm{D}_{\tilde{\phi}_\alpha}(\beta^\star, \tilde{\beta}_t^\alpha)$ are decreasing. Moreover*

$$L(\tilde{\beta}_t^\alpha) - L(\beta^\star) \leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}, \tag{A.4}$$

$$L\Big(\frac{1}{t}\int_0^t \tilde{\beta}_s^\alpha \mathrm{d}s\Big) - L(\beta^\star) \leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}. \tag{A.5}$$

Proof. The loss is decreasing since: $\frac{\mathrm{d}}{\mathrm{d}t}L(\tilde{\beta}_t^\alpha) = \nabla L(\tilde{\beta}_t^\alpha)^\top \dot{\beta}_t^\alpha = -\dot{\tilde{\beta}}_t^{\alpha\top}\nabla^2\tilde{\phi}_\alpha(\tilde{\beta}_t^\alpha)\dot{\tilde{\beta}}_t^\alpha \leq 0$.

$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\tilde{\phi}_\alpha}(\beta^\star, \tilde{\beta}_t^\alpha) = -\nabla L(\tilde{\beta}_t^\alpha)^\top(\tilde{\beta}_t^\alpha - \beta^\star) = -2(L(\tilde{\beta}_t^\alpha) - L(\beta^\star))$ (since L is the quadratic loss), therefore the Bregman distance is decreasing. We can also integrate this last equality from 0 to t, and divide by $-2t$:

$$\begin{aligned}\frac{1}{t}\int_0^t L(\tilde{\beta}_s^\alpha)\mathrm{d}s - L(\beta^\star) &= \frac{\mathrm{D}_{\tilde{\phi}_\alpha}(\beta^\star, \beta_0^\alpha = \mathbf{0}) - \mathrm{D}_{\tilde{\phi}_\alpha}(\beta^\star, \beta_t^\alpha)}{2t} \\ &\leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}.\end{aligned}$$

Since the loss is decreasing we get that $L(\tilde{\beta}_t^\alpha) - L(\beta^\star) \leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}$ and from the convexity of L we get that $L\Big(\frac{1}{t}\int_0^t \tilde{\beta}_s^\alpha \mathrm{d}s\Big) - L(\beta^\star) \leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}$. □

In the following proposition, we show that for α small enough, the iterates are bounded independently of α. Note that this result unfortunately only holds for the quadratic loss, we expect it to hold for other convex losses of the type $L(\beta) = \frac{1}{n}\sum_i \ell(y_i, \langle x_i, \beta\rangle)$ where $\ell(y, \cdot)$ is strictly convex has a unique root at y but we don't know how to show it. Also note that bounding the accelerated iterates $\tilde{\beta}^\alpha$ is equivalent to bounding the iterates β^α since $\tilde{\beta}^\alpha_t = \beta^\alpha_{\ln(1/\alpha)t}$.

Proposition 24. *For $\alpha < \alpha_0$, where α_0 depends on $\beta^\star_{\ell_1}$, the iterates $\tilde{\beta}^\alpha_t$ are bounded independently of α:*

$$\|\tilde{\beta}^\alpha_t\|_\infty \leq 3\|\beta^\star_{\ell_1}\|_1 + 1$$

Proof. From eq. (A.3), integrating and using that L is the quadratic loss, we get:

$$\nabla\tilde{\phi}_\alpha(\tilde{\beta}^\alpha_t) = \frac{t}{n}X^\top(y - X\bar{\beta}^\alpha_t) = -\frac{t}{n}X^\top X(\bar{\beta}^\alpha_t - \beta^\star),$$

where we recall that $X \in \mathbb{R}^{n\times d}$ is the input data represented as a matrix and where we denote the averaged iterate by $\bar{\beta}^\alpha_t = \frac{1}{t}\int_0^t \tilde{\beta}^\alpha_s \mathrm{d}s$. Thus we get

$$\nabla\tilde{\phi}_\alpha(\tilde{\beta}^\alpha_t)^\top(\tilde{\beta}^\alpha_t - \beta^\star) = -\frac{t}{n}(\bar{\beta}^\alpha_t - \beta^\star)^\top X^\top X(\tilde{\beta}^\alpha_t - \beta^\star). \tag{A.6}$$

By convexity of $\tilde{\phi}_\alpha$ we have $\tilde{\phi}_\alpha(\beta^\alpha_t) - \tilde{\phi}_\alpha(\beta^\star) \leq \nabla\tilde{\phi}_\alpha(\beta^\alpha_t)^\top(\beta^\alpha_t - \beta^\star)$. By the Cauchy-Schwarz inequality, we also have $(\bar{\beta}^\alpha_t - \beta^\star)^\top X^\top X(\beta^\alpha_t - \beta^\star) \leq \|X(\beta^\alpha_t - \beta^\star)\|\|X(\bar{\beta}^\alpha_t - \beta^\star)\|$. Using Proposition 23: $\|X(\beta^\alpha_t - \beta^\star)\|^2 \leq n\tilde{\phi}_\alpha(\beta^\star)/t$ and $\|X(\bar{\beta}^\alpha_t - \beta^\star)\|^2 \leq n\tilde{\phi}_\alpha(\beta^\star)/t$ we can further bound the right hand side of eq. (A.6) as

$$-\frac{t}{n}(\bar{\beta}^\alpha_t - \beta^\star)^\top X^\top X(\beta^\alpha_t - \beta^\star) \leq \tilde{\phi}_\alpha(\beta^\star).$$

Thus it yields

$$\tilde{\phi}_\alpha(\beta^\alpha_t) - \tilde{\phi}_\alpha(\beta^\star) \leq \tilde{\phi}_\alpha(\beta^\star).$$

From Woodworth et al. [2020b] (proof of Lemma 1 in the appendix) we get that for

$$\alpha < \min\left\{1, \sqrt{\|\beta\|_1}, (2\|\beta\|_1)^{-1}\right\}$$

then:

$$\tilde{\phi}_\alpha(\beta) \leq \frac{3}{2}\|\beta\|_1,$$

and for all $\alpha < \exp(-d/2)$:

$$\begin{aligned}\tilde{\phi}_\alpha(\beta) &\geq \|\beta\|_1 - \frac{d}{\ln(1/\alpha^2)} \\ &\geq \|\beta\|_1 - 1,\end{aligned}$$

which finally leads for

$$\alpha < \alpha_0 \coloneqq \min\left\{1, \sqrt{\|\beta^\star_{\ell_1}\|_1}, \left(2\|\beta^\star_{\ell_1}\|_1\right)^{-1}, \exp(-d/2)\right\}$$

to the result. □

APPENDIX A. APPENDIX FOR CHAPTER 6

The following proposition shows that we can bound the path length of the flow $\tilde{\beta}^\alpha$ independently of α. Keep in mind that the path length of $\tilde{\beta}^\alpha$ is equivalent to that of β^α as the first is just an acceleration of the second: $\tilde{\beta}^\alpha_t = \beta^\alpha_{\ln(1/\alpha)t}$.

Proposition 25. *For $\alpha < \alpha_0$ where α_0 is the same as in Proposition 24, the path length of the iterates $(\beta^\alpha_t)_{t\geq 0}$ is bounded independently of $\alpha > 0$:*

$$\int_0^{+\infty} \|\dot{\beta}^\alpha_t\| \mathrm{d}t < C,$$

where C does not depend on α. Hence the path length of the accelerated flow $\tilde{\beta}^\alpha$ is also bounded independently of α.

Proof. Having shown that the iterates β^α_t are bounded independently of α, it also implies that the iterates $w_t = (u_t, v_t)$ are bounded following Lemma 6. Since the loss $w \mapsto F(w)$ is a multivariate polynomial function, it is a semialgebraic function and we can consequently apply the result of Kurdyka [1998, Theorem 2] which grants that

$$\int_0^{+\infty} \|\dot{w}_t\| \mathrm{d}t < C,$$

where the constant C only depends on the loss and on the bound on the iterates. We further use that $\dot{\beta} = \dot{u} \odot v + u \odot \dot{v}$ and $\|\dot{u} \odot v + u \odot \dot{v}\| \leq C_1(\|\dot{u}\| + \|\dot{v}\|)$ using that u and v are bounded and $\|\dot{u}\| + \|\dot{v}\| \leq C_2\|\dot{w}\|$ using the equivalence of norms. Therefore $\int_0^{+\infty} \|\dot{\beta}^\alpha_t\| \mathrm{d}t < C$ for some C which is independent of the initialisation scale α. □

A.4 Standalone properties of Algorithm 1

A.4.1 "Well-definedness" of Algorithm 1 and upperbound on its number of loops

Notice that this proposition highlights the fact that Algorithm 1 is on its own an algorithm of interest for finding the minimum ℓ_1-norm solution in an overparametrised regression setting. We point out that the provided upperbound on the number of iterations is very crude and could certainly be improved.

Proposition 26. *Algorithm 1 is well defined: at each iteration (i) the attribution of Δ is well defined as $\Delta < +\infty$, (ii) the constrained minimisation problem has a unique solution and the attribution of the value of β is therefore well-founded. Furthermore, along the loops: the iterates β have at most n non-zero coordinates, the loss is strictly decreasing and the algorithm terminates in at most $\min\left(2^d, \sum_{k=0}^{n}\binom{d}{k}\right)$ steps by outputting the minimum ℓ_1-norm solution $\beta^\star_{\ell_1} \coloneqq \underset{\beta \in \arg\min L}{\operatorname{arg\,min}} \|\beta\|_1$.*

Proof. In the following, for the matrix X and for a subset $I = \{i_1, \dots, i_k\} \subset [d]$, we write $X_I = (\tilde{x}_{i_1}, \dots, \tilde{x}_{i_k}) \in \mathbb{R}^{n \times k}$ (we extract the columns from X). For a vector $\beta \in \mathbb{R}^d$ we write $\beta_I = (\beta_{i_1}, \dots, \beta_{i_k})$.

(1) The constrained minimisation problem has a unique solution: we follow the proof of Tibshirani [2013, Lemma 2]. Following the notations in Algorithm 1, we define $I = \{i \in [d],\ |s_i| = 1\}$ and we point out that after k loops of the algorithm, the value of s is equal to $s = -(\Delta_1 \nabla L(\beta_0) + \dots + \Delta_k \nabla L(\beta_{k-1})) \in \operatorname{span}(x_1, \dots, x_n)$. We can therefore write $s = X^\top r$ for some $r \in \mathbb{R}^n$.

Now assume that $\ker(X_I) \neq \{0\}$. Then, for some $i \in I$, we have $\tilde{x}_i = \sum_{j \in I \setminus \{i\}} c_j \tilde{x}_j$ where $c_j \in \mathbb{R}$. Without loss of generality, we can assume that $I \setminus \{i\}$ has at most n elements. Indeed, we can otherwise always find n elements $\tilde{I} \subset I \setminus \{i\}$ such that $\tilde{x}_i = \sum_{j \in \tilde{I}} c_j \tilde{x}_j$. Rewriting the previous equality, we get

$$s_i \tilde{x}_i = \sum_{j \in I \setminus \{i\}} (s_i s_j c_j)(s_j \tilde{x}_j). \tag{A.7}$$

Now by definitions of the set I and of r, we have that $\langle \tilde{x}_j, r \rangle = s_j \in \{+1, -1\}$ for any $j \in I$. Taking the inner product of eq. (A.7) with r, we obtain that $1 = \sum_{j \in I \setminus \{i\}} (s_i s_j c_j)$. Consequently, we have shown that if $\ker(X_I) \neq \{0\}$, then we necessarily have for some $i \in I$,

$$s_i \tilde{x}_i = \sum_{j \in I \setminus \{i\}} a_j (s_j \tilde{x}_j),$$

with $\sum_{j \in I \setminus \{i\}} a_j = 1$, which means that $s_i \tilde{x}_i$ lies in the affine space generated by $(s_j \tilde{x}_j)_{j \in I \setminus \{i\}}$. This fact is however impossible due to Assumption 6 (recall that without loss of generality we have that $I \setminus \{i\}$ has at most n elements, and trivially less that d elements). **Therefore X_I is full rank**, and $\operatorname{Card}(I) \leq n$. Now notice that the constrained minimisation problem corresponds to $\arg\min_{\substack{\beta_i \geq 0, i \in I_+ \\ \beta_i \leq 0, i \in I_-}} \|y - X_I \beta_I\|_2^2$. Since X_I is full rank, this restricted loss is strictly convex and the constrained minimisation problem **has a unique minimum.**

(2) $\Delta < +\infty$: Notice that the optimality conditions of

$$\beta = \underset{\substack{\beta_i \geq 0, i \in I_+ \\ \beta_i \leq 0, i \in I_- \\ \beta_i = 0, i \notin I}}{\arg\min} \|y - X_I \beta_I\|_2^2,$$

are (i) β satisfies the constraints, (ii) if $i \in I_+$ (resp $i \in I_-$) then $[-\nabla L(\beta)]_i \leq 0$ (resp $[-\nabla L(\beta)]_i \geq 0$) and (iii) if $\beta_i \neq 0$ then $[\nabla L(\beta)]_i = 0$. One can notice that condition (ii) ensures that at each iteration, for $\delta \leq \Delta_k$, $s_{k-1} - \delta \nabla L(\beta_{k-1}) \in [-1, 1]$ coordinate wise. Also, if $L(\beta_{k-1}) \neq \mathbf{0}$, then a coordinate of the vector $|s_{k-1} - \delta \nabla L(\beta_{k-1})|$ must necessarily hit 1, this value of δ corresponds to Δ_k.

(3) The loss is strictly decreasing: Let $I_{k-1,\pm}$ and $I_{k,\pm}$ be the equicorrelation sets defined in the algorithm at step $k-1$ and k, and β_{k-1} and β_k the solutions of the constrained minimisation problems. Also, let i_k be the newly added coordinate which breaks the constraint at step k (which we assume to be unique for simplicity). Without loss of generality, assume that $s_k(i_k) = +1$. Since the sets $I_{k-1,+} \backslash \big(I_{k,+} \backslash \{i_k\}\big)$ and $I_{k-1,-} \backslash I_{k,-}$ are (if not empty) only composed of indexes of coordinates of β_{k-1} which are equal to 0, one can notice that β_{k-1} also satisfies the new constraints at step k. Therefore $L(\beta_k) \leq L(\beta_{k-1})$. Now since $[-\nabla L(\beta_{k-1})]_{i_k} > 0$, from the strict convexity of the restricted loss on I_k, this means that $\beta_k(i_k) > 0$ (which also means that newly activated coordinate i_k **must activate**), and therefore $\beta_{k-1} \neq \beta_k$ and $L(\beta_k) < L(\beta_{k-1})$.

(4) The algorithm terminates in at most $\min\Big(2^d, \sum_{k=0}^n \binom{d}{k}\Big)$ steps: Recall that we showed in part (1) of the proof that at each iteration k of the algorithm, I_k as at most $\min(n, d)$ elements. Since $\mathrm{supp}(\beta_k) \subset I_k$, we have that β_k has at most $\min(n, d)$ non-zero elements, also recall that we always have $\beta_k = \arg\min_{\beta_i = 0, i \notin \mathrm{supp}(\beta_k)} L(\beta)$ (we here have unicity of this minimisation problem following part (1) of the proof). There are hence at most

$$\sum_{k=0}^{\min(n,d)} \binom{d}{k} = \min\Big(2^d, \sum_{k=0}^{n} \binom{d}{k}\Big)$$

such minimisation problems. The loss being strictly decreasing, the algorithm cannot output the same solution β at two different loops, and the algorithm must terminate in at most $\min\Big(2^d, \sum_{k=0}^n \binom{d}{k}\Big)$ iterations by outputting a vector $\beta^\star$ such that $\nabla L(\beta^\star) = 0$, *i.e.* $\beta^\star \in \arg\min L(\beta)$.

(5) The algorithm outputs the minimum ℓ_1-norm solution. Let $\beta^\star$ be the output of the algorithm after p iterations. Notice that by the definition of the successive sets $I_{k,\pm}$ and of the constraints on the minimisation problem, we have that at each iteration $s_k \in \partial \|\beta_k\|_1$. Therefore $s_p \in \partial \|\beta^\star\|_1$. Also, recall from part (1) of the proof that $s_p \in \mathrm{span}(x_1, \ldots, x_n)$ which means that there exists $r \in \mathbb{R}^n$ such that $s_p = X^\top r$. Putting the two together we get that $X^\top r \in \partial \|\beta^\star\|_1$, this condition along with the fact that $L(\beta^\star) = \min L(\beta)$ are exactly the KKT conditions of $\underset{\beta \in \arg\min L}{\arg\min} \|\beta\|_1$. □

To put our upperbound on the number of iterations into perspective, the worst-case number of iterations for the LARS algorithm is $(3^d + 1)/2$ [Mairal and Yu, 2012]. Hence Algorithm 1 has fewer iterations in the worst-case setting. Whether an exponential dependency in the dimension is inevitable for Algorithm 1 is unknown and we leave this as future work.

However, when the number of samples is much smaller than the dimension we lose the exponential dependency. Indeed, for $\varepsilon := n/d \leq 1/2$, we have the upperbound $\sum_{k=0}^n \binom{d}{k} \leq 2^{H(\varepsilon)d}$ where $H(\varepsilon) = -\varepsilon \log_2(\varepsilon) - (1-\varepsilon)\log_2(1-\varepsilon)$ is the binary entropy. Since for $\varepsilon \leq 1/2$, $H(\varepsilon) \leq -2\varepsilon \log_2(\varepsilon)$, we get the upperbound $\sum_{k=0}^n \binom{d}{k} \leq 2^{H(\varepsilon)d} \leq (\frac{d}{n})^{2n}$, which is much better than 2^d.

A.4.2 Proof of Proposition 12

As mentioned several times, for general feature matrices X complex behaviours can occur with coordinates deactivating and changing sign several times. Here we show that for simple datasets

which have a feature matrix X that satisfy the restricted isometry property (RIP) [Candès et al., 2006], we can simply determine the jump times and the saddles as a function of the sparse predictor which we seek to recover.

The non-realistic but enlightening extreme case of the RIP assumption is to consider that the feature matrix is such that $X^\top X/n = I_d$. In this case, by letting $\beta^\star$ be the unique vector such that $y = \langle x, \beta^\star \rangle$ and assuming that $\beta^\star = (\beta_1^\star, \dots, \beta_r^\star, 0, \dots, 0)$ with $|\beta_1^\star| > \cdots > |\beta_r^\star| > 0$, then the loss writes $L(\beta) = \|\beta - \beta^\star\|_2^2/2$ and one can easily check that Algorithm 1 would terminate in r loops and output exactly $t_i = \frac{1}{|\beta_i^\star|}$ and $\beta_i = (\beta_1^\star, \dots, \beta_i^\star, 0, \dots, 0)$ for $i \leq r$ (the case where several coordinates of $\beta^\star$ are stricly equal can also be treated: for example if $\beta_1^\star = \beta_2^\star$ then the first output of the algorithm is directly $\beta_1 = (\beta_1^\star, \beta_2^\star, 0, \dots, 0)$).

We now recall the more realistic RIP setting which is an adaptation of the previous observation.

Sparse regression with RIP and gap assumption. ***(RIP)*** Assume that there exists an r-sparse vector $\beta^\star$ such that $y_i = \langle x_i, \beta^\star \rangle$. Furthermore we assume that the feature matrix $X \in \mathbb{R}^{n,d}$ satisfies the $2r$-restricted isometry property with constant $\tilde{\varepsilon} < \sqrt{2} - 1 < 1/2$: i.e. for all submatrix X_s where we extract any $s \leq 2r$ columns of X, the matrix $X_s^\top X_s/n$ of size $s \times s$ has all its eigenvalues in the interval $[1 - \tilde{\varepsilon}, 1 + \tilde{\varepsilon}]$. ***(Gap assumption)*** Furthermore we assume that the r-sparse vector $\beta^\star$ has coordinates which have a "sufficient gap'. W.l.o.g we write $\beta^\star = (\beta_1^\star, \dots, \beta_r^\star, 0, \dots, 0)$ with $|\beta_1^\star| \geq \cdots \geq |\beta_r^\star| > 0$ and we define $\lambda := \min_{i \in [r]}(|\beta_i^\star| - |\beta_{i+1}^\star|) \geq 0$ which corresponds to the smallest gap between the entries of $|\beta^\star|$. We assume that $5\tilde{\varepsilon}\|\beta^\star\|_2 < \lambda/2$ and we let $\varepsilon := 5\tilde{\varepsilon}$.

A classic result from compressed sensing (see Candes [2008, Theorem 1.2]) is that the $2r$-restricted isometry property with constant $\sqrt{2} - 1$ ensures that the minimum ℓ_0-minimisation problem has a unique r-sparse solution which is $\beta^\star$. Furthermore it ensures that the minimum ℓ_1-norm solution is unique and is equal to $\beta^\star$. This means that Algorithm 1 will have $\beta^\star$ as a final output.

We now recall the result which characterises the outputs of Algorithm 1 when the data satisfies the previous assumptions.

Proposition 12. *Under the restricted isometry property and the gap assumption stated right above, Algorithm 1 terminates in r-loops and outputs:*

$$
\begin{aligned}
&\beta_1 = (\beta_1[1], 0, \dots, 0) && \text{with} \quad \beta_1[1] \in \left[\beta_1^\star - \varepsilon\|\beta^\star\|, \beta_2^\star + \varepsilon\|\beta^\star\|\right] \\
&\beta_2 = (\beta_2[1], \beta_2[2], 0, \dots, 0) && \text{with} \quad \begin{cases} \beta_2[1] \in \left[\beta_1^\star - \varepsilon\|\beta^\star\|, \beta_1^\star + \varepsilon\|\beta^\star\|\right] \\ \beta_2[2] \in [\beta_2^\star - \varepsilon\|\beta^\star\|, \beta_2^\star + \varepsilon\|\beta^\star\|] \end{cases} \\
&\vdots \\
&\beta_{r-1} = (\beta_{r-1}[1], \dots, \beta_{r-1}[r-1], 0, \dots, 0) && \text{with} \quad \beta_{r-1}[i] \in \left[\beta_i^\star - \varepsilon\|\beta^\star\|, \beta_i^\star + \varepsilon\|\beta^\star\|\right] \\
&\beta_r = \beta^\star = (\beta_1^\star, \dots, \beta_r^\star, 0, \dots, 0),
\end{aligned}
$$

at times $t_1, \dots, t_r$ such that $t_i \in \left[\frac{1}{|\beta_i^\star| + \varepsilon\|\beta^\star\|}, \frac{1}{|\beta_i^\star| - \varepsilon\|\beta^\star\|}\right]$ and where $\|\cdot\|$ denotes the ℓ_2 norm.

Proof. In all the proof $\|\cdot\|$ denotes the ℓ_2 norm $\|\cdot\|_2$. For simplicity we assume that $\beta_i^\star > 0$ for all $i \in [r]$, the proof can easily be adapted to the general case. We first define $\xi := X^\top X/n - I_d$. By the restricted isometry property, for any $k \leq 2r$, we have that any $k \times k$ square matrix extracted from ξ which we denote ξ_{kk} has its eigenvalues in $[-\tilde{\varepsilon}, \tilde{\varepsilon}]$. It also means that the eigenvalues of $(I_k + \xi_{kk})^{-1} - I_k$ are in $[\frac{1}{1+\tilde{\varepsilon}} - 1, \frac{1}{1-\tilde{\varepsilon}} - 1] \subset [-2\tilde{\varepsilon}, 2\tilde{\varepsilon}]$.

We now proceed by induction with the following induction hypothesis:

- β_{k-1} has its support on its $(k-1)$ first coordinates with $|\beta_{k-1}[i] - \beta_i^\star| \leq 5\tilde{\varepsilon}\|\beta^\star\|$ for $i < k$
- $t_k \in \left[\frac{1}{\beta_k^\star + 5\tilde{\varepsilon}\|\beta^\star\|}, \frac{1}{\beta_k^\star - 5\tilde{\varepsilon}\|\beta^\star\|}\right]$ and $s_{t_k}[k] = 1$
- $s_{t_k}[i] \in [t_k(\beta_i^\star - 5\tilde{\varepsilon}\|\beta^\star\|), t_k(\beta_i^\star + 5\tilde{\varepsilon}\|\beta^\star\|)] \subset (-1, 1)$ for $i > k$

From the recurrence hypothesis, the output of the algorithm at step k is hence $\beta_k = \arg\min L(\beta)$ under the constraint $\beta[i] \geq 0$ for $i \leq k$ and $\beta[i] = 0$ otherwise. We first search for the solution of the minimisation problem without the sign constraint and still (abusively) denote it β_k: we will show that it turns out to satisfy the sign constraint and that it is therefore indeed β_k.

In the following, for a vector v, we denote by $v[:k]$ its k first coordinates. Setting the k first coordinates of the gradient to 0, we get that $[X^\top X(\beta_k - \beta^\star)][:k] = \mathbf{0}$, which leads to $(I_k + \xi_{kk})\beta_k[:k] = \beta^\star[:k] + [\xi\beta^\star][:k]$, which gives:

$$\begin{aligned}\beta_k[:k] &= (I_k + \xi_{kk})^{-1}(\beta^\star[:k] + [\xi\beta^\star][:k]) \\ &= \beta^\star[:k] + [\xi\beta^\star][:k] + v_1\end{aligned}$$

where from the bound on the eigenvalues of $(I_k + \xi_{kk})^{-1} - I_k$ and $\|\xi\beta^\star\| \leq \tilde{\varepsilon}\|\beta^\star\|$:

$$\begin{aligned}\|v_1\| &\leq 2\tilde{\varepsilon}\|\beta^\star[:k] + [\xi\beta^\star][:k])\| \\ &\leq 2\tilde{\varepsilon}(\|\beta^\star\| + \|\xi\beta^\star\|) \\ &\leq 2\tilde{\varepsilon}(\|\beta^\star\| + \tilde{\varepsilon}\|\beta^\star\|) \\ &\leq 4\tilde{\varepsilon}\|\beta^\star\|.\end{aligned}$$

Therefore

$$\beta_k[:k] = \beta^\star[:k] + v_2$$

where $v_2 = [\xi\beta^\star][:k] + v_1$ hence $\|v_2\|_\infty \leq \|v_2\| \leq 5\tilde{\varepsilon}\|\beta^\star\|$. Notice that from the definition of λ and the fact that $5\tilde{\varepsilon}\|\beta^\star\| < \lambda/2$ we have that $\beta_k[:k] \geq 0$ coordinate-wise, hence verifying the sign constraint. Also note that $\|\beta_k\| \leq \|\beta^\star\| + 5\tilde{\varepsilon}\|\beta^\star\| \leq 4\|\beta^\star\|$.

For $t \geq t_k$, $s_t = s_{t_k} - (t - t_k)\nabla L(\beta_k)$, and $[\nabla L(\beta_k)][:k] = 0$ therefore $s_t[:k] = s_{t_k}[:k]$. Now for $i > k$, $[-\nabla L(\beta_k)]_i = n^{-1}[X^\top X(\beta^\star - \beta_k)]_i = \beta_i^\star + [\xi(\beta_k - \beta^\star)]_i$. Now since $(\beta_k - \beta^\star)$ is r-sparse we have that:

$$\begin{aligned}\|\xi(\beta_k - \beta^\star)\|_\infty &\leq \|\xi(\beta_k - \beta^\star)\| \\ &\leq \tilde{\varepsilon}\|\beta_k - \beta^\star\| \\ &\leq \tilde{\varepsilon}(\|\beta_k\| + \|\beta^\star\|) \\ &\leq 5\tilde{\varepsilon}\|\beta^\star\| < \lambda/2, \qquad \text{(A.8)}\end{aligned}$$

Now from the fact that $s_t[i] = s_{t_k}[i] + (t - t_k)\beta_i^\star + (t - t_k)[\xi(\beta_k - \beta^\star)]_i$ and using the recurrence hypothesis: $s_{t_k}[i] \in [t_k(\beta_i^\star - 5\tilde{\varepsilon}\|\beta^\star\|), t_k(\beta_i^\star + 5\tilde{\varepsilon}\|\beta^\star\|)]$, we get (using the bound eq. (A.8)) that $s_t[i] \in [t(\beta_i^\star - 5\tilde{\varepsilon}\|\beta^\star\|), t(\beta_i^\star + 5\tilde{\varepsilon}\|\beta^\star\|)]$. From the "separation assumption" we have that $5\tilde{\varepsilon}\|\beta^\star\| < \lambda/2$ and therefore the next coordinate to activate is necessarily the $(k+1)^{th}$ at time t_{k+1} with $s_{t_{k+1}}[k+1] = 1$ and:

$$t_{k+1} \in \left[\frac{1}{\beta_{k+1}^\star + 5\tilde{\varepsilon}\|\beta^\star\|}, \frac{1}{\beta_{k+1}^\star - 5\tilde{\varepsilon}\|\beta^\star\|}\right].$$

This proves the recursion. The algorithm cannot stop before iteration r as $\beta^\star$ is the unique minimiser of L that has at most r non-zero coordinates. But it stops at iteration r as $\beta^\star$ is the unique minimiser of $L(\beta)$ under the constraints $\beta_i \geq 0$ for $i \leq r$ and $\beta_i = 0$ otherwise. □

A.5 Proof of Theorem 1 and Proposition 13 through the arc-length parametrisation

In this section, we explain in more details the arc-length reparametrisation which circumvents the apparition of discontinuous jumps and leads to the proof of Theorem 1. The main difficulty to show the convergence stems from the non-continuity of the limit process $\tilde{\beta}^\circ$. Therefore we cannot expect uniform convergence of $\tilde{\beta}^\alpha$ towards $\tilde{\beta}$ as $\alpha \to 0$. In addition, $\tilde{\beta}^\circ$ does not provide any insights into the path followed between the jumps.

Arc-length parametrisation. The high-level idea is to "slow-down" time when the jumps occur. To do so we follow the approach from Efendiev and Mielke [2006], Mielke et al. [2009] and we consider an arc-length parametrisation of the path, i.e., we consider τ^α equal to:

$$\tau^\alpha(t) = t + \int_0^t \|\dot{\tilde{\beta}}^\alpha_s\| \mathrm{d}s.$$

In Proposition 25, we showed that the full path length $\int_0^{+\infty} \|\dot{\beta}^\alpha_s\| \mathrm{d}s$ is finite and bounded independently of α. Therefore τ^α is a bijection in $\mathbb{R}_{\geq 0}$. We can then define the following quantities:

$$\hat{t}^\alpha_\tau = (\tau^\alpha)^{-1}(\tau) \quad \text{and} \quad \hat{\beta}^\alpha_\tau = \tilde{\beta}^\alpha_{\hat{t}^\alpha(\tau)}.$$

By construction, a simple chain rule leads to $\dot{\hat{t}}^\alpha(\tau) + \|\dot{\hat{\beta}}^\alpha_\tau\| = 1$, which means that the speed of $(\hat{\beta}^\alpha_\tau)_\tau$ is always upperbounded by 1, independently of α. This behaviour is in stark contrast with the process $(\tilde{\beta}^\alpha_t)_t$ which has a speed which explodes at the jumps. It presents a major advantage as we can now use Arzelà-Ascoli's theorem to extract a converging subsequent. A simple change of variable shows that the new process satisfies the following equations:

$$-\int_0^\tau \dot{\hat{t}}^\alpha_s \nabla L(\hat{\beta}^\alpha_s) \mathrm{d}s = \nabla \tilde{\phi}_\alpha(\hat{\beta}^\alpha_\tau) \quad \text{and} \quad \dot{\hat{t}}^\alpha_\tau + \|\dot{\hat{\beta}}^\alpha_\tau\| = 1 \tag{A.9}$$

started from $\hat{\beta}^\alpha_\tau = 0$ and $\hat{t}_0 = 0$. The next proposition states the convergence of the rescaled process, up to a subsequence.

Proposition 27. *Let $T \geq 0$. For every $\alpha > 0$, let $(\hat{t}^\alpha, \hat{\beta}^\alpha)$ be the solution of eq. (A.9). Then, there exists a subsequence $(\hat{t}^{\alpha_k}, \hat{\beta}^{\alpha_k})_{k\in\mathbb{N}}$ and $(\hat{t}, \hat{\beta})$ such that as $\alpha_k \to 0$:*

$$(\hat{t}^{\alpha_k}, \hat{\beta}^{\alpha_k}) \to (\hat{t}, \hat{\beta}) \qquad \text{in } (C^0([0,T], \mathbb{R}\times\mathbb{R}^d), \|\cdot\|_\infty) \tag{A.10}$$

$$(\dot{\hat{t}}^{\alpha_k}, \dot{\hat{\beta}}^{\alpha_k}) \rightharpoonup (\dot{\hat{t}}, \dot{\hat{\beta}}) \qquad \text{in } L_1[0,T] \tag{A.11}$$

Limiting dynamics. *The limits $(\hat{t}, \hat{\beta})$ satisfy:*

$$-\int_0^\tau \dot{\hat{t}}_s \nabla L(\hat{\beta}_s) \mathrm{d}s \in \partial\|\hat{\beta}_\tau\|_1 \quad \text{and} \quad \dot{\hat{t}}_\tau + \|\dot{\hat{\beta}}_\tau\| \leq 1 \tag{A.12}$$

Heteroclinic orbit. *In addition, when $\hat{\beta}_\tau$ is such that $|\hat{\beta}_\tau| \odot \nabla L(\hat{\beta}_\tau) \neq 0$, we have*

$$\dot{\hat{\beta}}_\tau = -\frac{|\hat{\beta}_\tau| \odot \nabla L(\hat{\beta}_\tau)}{\||\hat{\beta}_\tau| \odot \nabla L(\hat{\beta}_\tau)\|} \quad \text{and} \quad \dot{\hat{t}}_\tau = 0. \tag{A.13}$$

Furthermore, the loss strictly decreases along the heteroclinic orbits and the path length $\int_0^T \|\dot{\hat{\beta}}_\tau\| \mathrm{d}\tau$ is upperbounded independently of T.

Proof. Differentiating eq. (A.9) and from the Hessian of $\tilde{\phi}_\alpha$ we get:

$$\begin{aligned}\dot{\hat{\beta}}_\tau^\alpha &= -\dot{\hat{t}}_\tau^\alpha(\nabla^2\tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1}\nabla L(\hat{\beta}_\tau^\alpha)\\ &= -(1-\|\dot{\hat{\beta}}_\tau^\alpha\|)(\nabla^2\tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1}\nabla L(\hat{\beta}_\tau^\alpha).\end{aligned}$$

Therefore taking the norm on the right hand side we obtain that

$$\|\dot{\hat{\beta}}_\tau^\alpha\| = \frac{\|(\nabla^2\tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1}\nabla L(\hat{\beta}_\tau^\alpha)\|}{1+\|(\nabla^2\tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1}\nabla L(\hat{\beta}_\tau^\alpha)\|},$$

and therefore

$$\dot{\hat{\beta}}_\tau^\alpha = -\frac{(\nabla^2\tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1}\nabla L(\hat{\beta}_\tau^\alpha)}{1+\|(\nabla^2\tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1}\nabla L(\hat{\beta}_\tau^\alpha)\|}. \tag{A.14}$$

Subsequence extraction. By construction eq. (A.9) we have $\dot{\hat{t}}_\tau^\alpha + \|\dot{\hat{\beta}}_\tau^\alpha\| = 1$, therefore the sequences $(\dot{\hat{t}}^\alpha)_\alpha$, $(\dot{\hat{\beta}}^\alpha)_\alpha$ as well as $(\hat{t}^\alpha)_\alpha$, $(\hat{\beta}^\alpha)_\alpha$ are uniformly bounded on $[0,T]$. The Arzelà-Ascoli theorem yields that, up to a subsequence, there exists $(\hat{t},\hat{\beta})$ such that $(\hat{t}^{\alpha_k},\hat{\beta}^{\alpha_k}) \to (\hat{t},\hat{\beta})$ in $(C^0([0,T],\mathbb{R}\times\mathbb{R}^d),\|\cdot\|_\infty)$. Since $\|\dot{\hat{\beta}}_\tau^\alpha\|, \|\dot{\hat{t}}_\tau^\alpha\| \leq 1$ we have, applying the Banach–Alaoglu theorem, that up to a new subsequence

$$(\dot{\hat{t}}^{\alpha_k},\dot{\hat{\beta}}^{\alpha_k}) \overset{*}{\rightharpoonup} (\dot{\hat{t}},\dot{\hat{\beta}}) \text{ in } L_\infty(0,T) \tag{A.15}$$

and $\|\dot{\hat{\beta}}_\tau\| \leq \liminf_{\alpha_k}\|\dot{\hat{\beta}}_\tau^{\alpha_k}\| \leq 1$ and thus $\dot{\hat{t}}_\tau + \|\dot{\hat{\beta}}_\tau\| \leq 1$:

$$\int_0^T \|\dot{\hat{\beta}}_\tau\|\mathrm{d}\tau \leq \int_0^T \liminf_{\alpha_k}\|\dot{\hat{\beta}}_\tau^{\alpha_k}\|\mathrm{d}\tau \leq \int_0^{+\infty}\liminf_{\alpha_k}\|\dot{\hat{\beta}}_\tau^{\alpha_k}\|\mathrm{d}\tau \leq \liminf_{\alpha_k}\int_0^{+\infty}\|\dot{\hat{\beta}}_\tau^{\alpha_k}\|\mathrm{d}\tau < C,$$

where the third inequality is by Fatou's lemma. Note that since $[0,T]$ is bounded then it also implies the weak convergence in any $L_p(0,T)$, $1 \leq p < \infty$. Since $(\hat{\beta}^\alpha)$ converges uniformly on $[0,T]$, and ∇L is continuous, we have that $\nabla L(\hat{\beta}^\alpha)$ converges uniformly to $\nabla L(\hat{\beta})$. Since $\dot{\hat{t}}^{\alpha_k} \rightharpoonup \dot{\hat{t}}$ in $L_1(0,T)$, passing to the limit in the equation $\nabla\tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha) = -\int_0^\tau \dot{\hat{t}}_s^\alpha \nabla L(\hat{\beta}_s^\alpha)\mathrm{d}s$ leads to

$$-\int_0^\tau \dot{\hat{t}}_s\nabla L(\hat{\beta}_s)\mathrm{d}s \in \partial\|\hat{\beta}_\tau\|_1,$$

due to Lemma 7.

Recall from eq. (A.14) and the definition of $\tilde{\phi}_\alpha$ that:

$$\dot{\hat{\beta}}_\tau^\alpha = -\frac{\sqrt{\hat{\beta}_\tau^\alpha+\alpha^4}\odot\nabla L(\hat{\beta}_\tau^\alpha)}{1/\ln(1/\alpha)+\|\sqrt{\hat{\beta}_\tau^\alpha+\alpha^4}\odot\nabla L(\hat{\beta}_\tau^\alpha)\|}. \tag{A.16}$$

Hence assuming that $\hat{\beta}_\tau$ is such that $\||\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)\| \neq 0$, we can ensure that $\||\hat{\beta}_{\tau'}|\odot\nabla L(\hat{\beta}_{\tau'})\| \neq 0$ for $\tau' \in [\tau,\tau+\varepsilon]$ and ε small enough. We have then $\frac{\sqrt{\hat{\beta}_{\tau'}^\alpha+\alpha^4}\odot\nabla L(\hat{\beta}_{\tau'}^\alpha)}{1/\ln(1/\alpha)+\|\sqrt{\hat{\beta}_{\tau'}^\alpha+\alpha^4}\odot\nabla L(\hat{\beta}_{\tau'}^\alpha)\|}$ converges uniformly toward $-\frac{|\hat{\beta}_{\tau'}|\odot\nabla L(\hat{\beta}_{\tau'})}{\||\hat{\beta}_{\tau'}|\odot\nabla L(\hat{\beta}_{\tau'})\|}$ on $[\tau,\tau+\varepsilon]$. Using the dominated convergence theorem,

we have $\int_\tau^{\tau+\varepsilon} \frac{\sqrt{\hat\beta^\alpha_{\tau'}+\alpha^4}\odot\nabla L(\hat\beta^\alpha_{\tau'})}{1/\log(1/\alpha)+\|\sqrt{\hat\beta^\alpha_{\tau'}+\alpha^4}\odot\nabla L(\hat\beta^\alpha_{\tau'})\|}\mathrm{d}\tau' \to \int_\tau^{\tau+\varepsilon} \frac{|\hat\beta_{\tau'}|\odot\nabla L(\hat\beta_{\tau'})}{\||\hat\beta_{\tau'}|\odot\nabla L(\hat\beta_{\tau'})\|}\mathrm{d}\tau'$. We therefore obtain $\dot{\hat\beta}_\tau = -\frac{|\hat\beta_\tau|\odot\nabla L(\hat\beta_\tau)}{\||\hat\beta_\tau|\odot\nabla L(\hat\beta_\tau)\|}$ in $L_1[0,T]$. Consequently $\|\dot{\hat\beta}_\tau\| = 1$ and $\dot{\hat t}_\tau = 0$.

Proof that the loss stricly decreases along the heteroclinic orbits.

Assume $\hat\beta_\tau$ is such that $|\hat\beta_\tau|\odot\nabla L(\hat\beta_\tau) \neq 0$, then the flow follows

$$\dot{\hat\beta}_\tau = -\frac{|\hat\beta_\tau|\odot\nabla L(\hat\beta_\tau)}{\||\hat\beta_\tau|\odot\nabla L(\hat\beta_\tau)\|}$$

Letting $\gamma(\tau) = \frac{1}{\||\hat\beta_\tau|\odot\nabla L(\hat\beta_\tau)\|}$ we get:

$$\mathrm{d}L(\hat\beta_\tau) = -\gamma(\tau)\sum_i |\hat\beta_\tau(i)|\odot[\nabla L(\hat\beta_\tau)]_i^2\mathrm{d}\tau < 0,$$

because $|\hat\beta_\tau|\odot\nabla L(\hat\beta_\tau)^2 \neq 0$. □

Borrowing terminologies from Efendiev and Mielke [2006], we can distinguish two regimes: when $\dot{\hat\beta}_\tau = 0$, the system is *sticked* to the saddle point. When $\dot{\hat t}_\tau = 0$ and $\|\dot{\hat\beta}_\tau\| = 1$ the system switches to a *viscous slip* which follows the normalised flow eq. (A.13). We use the term of *heteroclinic orbit* as in the dynamical systems literature since in the weight space (u,v) it corresponds to a path with links two distinct critical points of the loss F. Since $\dot{\hat t}_\tau = 0$, this regime happens instantly for the original t time scale (*i.e.* a jump occurs).

From Proposition 27, following the same reasoning as in Section 6.4, we can show that the rescaled process converges uniformly to a continuous saddle-to-saddle process where the saddles are linked by normalized flows.

Theorem 1. *Let $T > 0$. For all subsequences defined in Proposition 27, there exist times $0 = \tau'_0 < \tau_1 < \tau'_1 < \cdots < \tau_p < \tau'_p < \tau_{p+1} = +\infty$ such that the the iterates $(\hat\beta^{\alpha_k}_\tau)_\tau$ converge uniformly on $[0,T]$ to the following limit trajectory :*

("Saddle")	$\hat\beta_\tau = \beta_k$	*for $\tau \in [\tau'_k, \tau_{k+1}]$ where $0 \le k \le p$*
(Orbit)	$\dot{\hat\beta}_\tau = -\frac{\lvert\hat\beta_\tau\rvert\odot\nabla L(\hat\beta_\tau)}{\||\hat\beta_\tau|\odot\nabla L(\hat\beta_\tau)\|}$	*for $\tau \in [\tau_{k+1}, \tau'_{k+1}]$ where $0 \le k \le p-1$*

where the saddles $(\beta_0 = 0, \beta_1, \dots, \beta_p = \beta^\star_{\ell_1})$ are constructed in Algorithm 1. Also, the loss $(L(\hat\beta_\tau))_\tau$ is constant on the saddles and strictly decreasing on the orbits. Finally, independently of the chosen subsequence, for $k \in [p]$ we have $\hat t_{\tau_k} = \hat t_{\tau'_k} = t_k$ where the times $(t_k)_{k\in[p]}$ are defined through Algorithm 1.

Proof. Some parts of the proof are slightly technical. To simplify the understanding, we make use of auxiliary lemmas which are stated in Appendix A.6. The overall spirit follows the intuitive ideas given in Section 6.4 and relies on showing that eq. (A.12) can only be satisfied if the iterates visit the saddles from Algorithm 1.

We let $\hat s_\tau \coloneqq -\int_0^\tau \dot{\hat t}_s \nabla L(\hat\beta_s)\mathrm{d}s$, which is continuous and satisfies $\hat s_\tau \in \partial\|\hat\beta_\tau\|_1$ from eq. (A.12). Let $S = \{\beta \in \mathbb{R}^d, |\beta|\odot\nabla L(\beta) = \mathbf{0}\}$ denote the set of critical points and let (β_k, t_k, s_k) be the successive values of (β, t, s) which appear in the loops of Algorithm 1.

We do a proof by induction: we start by assuming that the iterates are stuck at the saddle β_{k-1} at time $\tau \ge \tau'_{k-1}$ where $\hat t_{\tau'_{k-1}} = t_{k-1}$ and $\hat s_{\tau'_{k-1}} = s_{k-1}$ (recurrence hypothesis), we then show that they can only move at a time τ_k and follow the normalised flow eq. (A.13). We

finally show that they must end up "stuck" at the new critical point β_k, validating the recurrence hypothesis.

Proof of the jump time τ_k such that $\hat{t}_{\tau_k} = t_k$: we set ourselves at time $\tau \geq \tau'_{k-1}$, stuck at the saddle β_{k-1}. Let $\tau_k \coloneqq \sup\{\tau, \hat{t}_\tau \leq t_k\}$, we have that $\tau_k < \infty$ from Lemma 8. Note that by continuity of $\hat{t}_\tau$ it holds that $\hat{t}_{\tau_k} = t_k$. Now notice that $\hat{s}_\tau = \hat{s}_{\tau'_{k-1}} - (\hat{t}_\tau - \hat{t}_{\tau'_{k-1}})\nabla L(\beta_{k-1}) = s_{k-1} - (\hat{t}_\tau - t_{k-1})\nabla L(\beta_{k-1})$. We argue that for any $\varepsilon > 0$, we cannot have $\hat{\beta}_\tau = \beta_{k-1}$ on $(\tau_k, \tau_k + \varepsilon)$. Indeed by the definition of τ_k and from the algorithmic construction of time t_k, it would lead to $|\hat{s}_\tau(i)| > 1$ for some coordinate $i \in [d]$, which contradicts eq. (A.12). Therefore the iterates must move at the time τ_k.

Heterocline leaving β_{k-1} for $\tau \in [\tau_k, \tau'_k]$: contrary to before, our time rescaling enables to capture what happens during the "jump". We have shown that for any ε, there exists $\tau_\varepsilon \in (\tau_k, \tau_k + \varepsilon)$, such that $\hat{\beta}_{\tau_\varepsilon} \neq \beta_{k-1}$. From Lemma 9, since the saddles are distinct along the flow, we must have that $\hat{\beta}_{\tau_\varepsilon} \notin S$ for ε small enough. The iterates therefore follow a heterocline flow leaving β_{k-1} with a speed of 1 given by eq. (A.13). We now define $\tau'_k \coloneqq \inf\{\tau > \tau_k, \exists \varepsilon_0 > 0, \forall \varepsilon \in [0, \varepsilon_0],\ \hat{\beta}_{\tau+\varepsilon} \in S\}$ which corresponds to the time at which the iterates reach a new critical point and stay there for at least a small time ε_0. We have just shown that $\tau'_k > \tau_k$. Now from Proposition 27, the path length of $\hat{\beta}$ is finite, and from Lemma 9 the flow visits a finite number of distinct saddles at a speed of 1. These two arguments put together, we get that $\tau'_k < +\infty$ and also $\hat{\beta}_{\tau'_k+\varepsilon} = \hat{\beta}_{\tau'_k}$, $\forall \varepsilon \in [0, \varepsilon_0]$. On another note, since $\dot{\hat{t}}_\tau = 0$ for $\tau \in [\tau_k, \tau'_k]$ we have $\hat{t}_{\tau'_k} = \hat{t}_{\tau_k} (= t_k)$ as well as $\hat{s}_{\tau_k} = \hat{s}_{\tau'_k} (= s_k)$.

Proof of the landing point β_k : we now want to find to which saddle $\hat{\beta}_{\tau'_k} \in S$ the iterates have moved to. To that end, we consider the following sets which also appear in Algorithm 1:

$$I_{\pm,k} \coloneqq \{i \in \{1, \dots, d\},\ \text{s.t. } \hat{s}_{\tau'_k}(i) = \pm 1\} \quad \text{and} \quad I_k = I_{+,k} \cup I_{-,k}. \tag{A.17}$$

The set I_k corresponds to the coordinates of $\hat{\beta}_{\tau'_k}$ which "are allowed" (but not obliged) to be activated (*i.e.* non-zero). For $\tau \in [\tau'_k, \tau'_k + \varepsilon_0]$ we have that $\hat{s}_\tau = \hat{s}_{\tau'_k} - (\hat{t}_\tau - t_k)\nabla L(\hat{\beta}_{\tau'_k})$. By continuity of $\hat{s}$ and the fact that $\hat{s}_\tau \in \partial \|\hat{\beta}_{\tau'_k}\|_1$, the equality translates into:

- if $i \notin I_k$, $\hat{\beta}_{\tau'_k}(i) = 0$
- if $i \in I_{+,k}$, then $[\nabla L(\hat{\beta}_{\tau'_k})]_i \geq 0$ and $\hat{\beta}_{\tau'_k}(i) \geq 0$
- if $i \in I_{-,k}$, then $[\nabla L(\hat{\beta}_{\tau'_k})]_i \leq 0$ and $\hat{\beta}_{\tau'_k}(i) \leq 0$
- for $i \in I_k$, if $\hat{\beta}_{\tau'_k}(i) \neq 0$, then $[\nabla L(\hat{\beta}_{\tau'_k})]_i = 0$

One can then notice that these conditions exactly correspond to the optimality conditions of the following constrained minimisation problem:

$$\underset{\substack{\beta_i \geq 0, i \in I_{k,+} \\ \beta_i \leq 0, i \in I_{k,-} \\ \beta_i = 0, i \notin I_k}}{\arg\min} \quad L(\beta). \tag{A.18}$$

We showed in Proposition 26 that the solution to this problem is unique and equal to β_k from Algorithm 1. Therefore $\hat{\beta}_\tau = \beta_k$ for $\tau \in [\tau'_k, \tau'_k + \varepsilon_0]$. It finally remains to show that $\hat{\beta}_\tau = \beta_k$ while $\tau \leq \tau_{k+1}$, where $\tau_{k+1} \coloneqq \sup\{\tau, \hat{t}_\tau = t_{k+1}\}$. For this let $\tau \in [\tau'_k, \tau_{k+1}]$, notice that for $i \notin I_k$, we necessarily have that $\hat{\beta}_\tau(i) = \beta_k(i) = 0$, otherwise we break the continuity of $\hat{s}_\tau$. Similarly, for $i \in I_{k,+}$, we necessarily have that $\hat{\beta}_\tau(i) \geq 0$ and for $i \in I_{k,-}$, $\hat{\beta}_\tau(i) \leq 0$ for the same continuity reasons. Now assume that $\hat{\beta}_\tau(I_k) \neq \beta_k(I_k)$. Then from Lemma 9 and continuity of the flow,

$\exists \tau' \in (\tau'_k, \tau)$ such that $\hat{\beta}_{\tau'} \notin S$ and there must exist a heteroclinic flow eq. (A.13) starting from β_k which passes through $\beta_{\tau'}$. This is absurd since along this flow the loss strictly decreases, which is in contradiction with the definition of β_k which minimises the problem eq. (A.18). □

A.5.1 Proof of Theorem 1

Theorem 1 enables to prove without difficulty Theorem 1 which we recall below. Indeed we can show that any extracted limit $\hat{\beta}$ maps back to the unique discontinuous process $\tilde{\beta}^\circ$.

Theorem 2. *Let the saddles $(\beta_0 = \mathbf{0}, \beta_1, \dots, \beta_{p-1}, \beta_p = \beta^\star_{\ell_1})$ and jump times $(t_0 = 0, t_1, \dots, t_p)$ be the outputs of Algorithm 1 and let $(\tilde{\beta}^\circ_t)_t$ be the piecewise constant process defined as follows:*

$$\textit{(Saddles)} \qquad \tilde{\beta}^\circ_t = \beta_k \qquad \text{for } t \in (t_k, t_{k+1}) \text{ and } 0 \le k \le p, \quad t_{p+1} = +\infty.$$

The accelerated flow $(\tilde{\beta}^\alpha_t)_t$ defined in eq. (6.9) uniformly converges towards the limiting process $(\tilde{\beta}^\circ_t)_t$ on any compact subset of $\mathbb{R}_{\ge 0} \backslash \{t_1, \dots, t_p\}$.

Proof. We directly apply Theorem 1, let α_k be the subsequence from the theorem. Let $\varepsilon > 0$, for simplicity we prove the result on $[t_1 + \varepsilon, t_2 - \varepsilon]$, all the other compacts easily follow the same line of proof. Note that since $\hat{t}^{\alpha_k}(\tau'_1) \to t_1$ and $\hat{t}^{\alpha_k}(\tau_2) \to t_2$, for α_k small enough $\hat{t}^{\alpha_k}(\tau'_1) \le t_1 + \varepsilon$ and $\hat{t}^{\alpha_k}(\tau_2) \ge t_2 - \varepsilon$, by the monotonicity of τ^{α_k}, this means that for α_k small enough, $\tau'_1 \le \tau^{\alpha_k}(t_1 + \varepsilon)$ and $\tau_2 \ge \tau^{\alpha_k}(t_2 - \varepsilon)$. Therefore

$$\begin{aligned}
\sup_{t \in [t_1+\varepsilon, t_2-\varepsilon]} \|\tilde{\beta}^{\alpha_k}_t - \beta_1\| &= \sup_{t \in [t_1+\varepsilon, t_2-\varepsilon]} \|\hat{\beta}^{\alpha_k}(\tau_{\alpha_k}(t)) - \beta_1\| \\
&= \sup_{\tau \in [\tau^{\alpha_k}(t_1+\varepsilon), \tau^{\alpha_k}(t_2-\varepsilon)]} \|\hat{\beta}^{\alpha_k}(\tau) - \beta_1\| \\
&\le \sup_{\tau \in [\tau'_1, \tau_2]} \|\hat{\beta}^{\alpha_k}(\tau) - \beta_1\|,
\end{aligned}$$

which goes uniformly to 0 following Theorem 1. Since this result is independent of the subsequence α_k, we get the result of Theorem 1. □

A.5.2 Proof of Proposition 13

We restate and prove Proposition 13 below.

Proposition 13. *For all $T > t_p$, the graph of the iterates $(\tilde{\beta}^\alpha_t)_{t \le T}$ converges to that of $(\hat{\beta}_\tau)_\tau$:*

$$\operatorname{dist}(\{\tilde{\beta}^\alpha_t\}_{t \le T}, \{\hat{\beta}_\tau\}_{\tau \ge 0}) \underset{\alpha \to 0}{\longrightarrow} 0 \qquad \textit{(Hausdorff distance)}$$

Proof. For α small enough, we have that $\hat{t}^\alpha_{\tau'_p} \le t_p + \varepsilon \le T$

$$\begin{aligned}
\sup_{\tau \ge 0} d(\hat{\beta}_\tau, \{\tilde{\beta}^\alpha_t\}_{t \le T}) &= \sup_{\tau \le \tau'_p} d(\hat{\beta}_\tau, \{\tilde{\beta}^\alpha_t\}_{t \le T}) \\
&\le \sup_{\tau \le \tau'_p} \|\hat{\beta}_\tau - \tilde{\beta}^\alpha_{\hat{t}^\alpha_\tau}\| \\
&= \sup_{\tau \le \tau'_p} \|\hat{\beta}_\tau - \hat{\beta}^\alpha_\tau\| \quad \underset{\alpha \to 0}{\longrightarrow} \quad 0,
\end{aligned}$$

according to Theorem 1.

Similarly:

$$\begin{aligned}\sup_{t\leq T} d(\tilde{\beta}_t^\alpha, \{\hat{\beta}_{\tau'}\}_{\tau'}) &= \sup_{\tau\leq\tau_T^\alpha} d(\hat{\beta}_\tau^\alpha, \{\hat{\beta}_{\tau'}\}_{\tau'}) \\ &\leq \sup_{\tau\leq\tau_T^\alpha} \|\hat{\beta}_\tau^\alpha - \hat{\beta}_\tau\| \underset{\alpha\to 0}{\longrightarrow} \quad 0,\end{aligned}$$

according to Theorem 1, which concludes the proof. □

A.6 Technical lemmas

The following lemma describes the behaviour of $\nabla\tilde{\phi}_\alpha(\beta^\alpha)$ as $\alpha \to 0$ in function of the subdifferential $\partial\|\cdot\|_1$.

Lemma 7. *Let $(\beta^\alpha)_{\alpha>0}$ such that $\beta^\alpha \xrightarrow[\alpha\to 0]{} \beta \in \mathbb{R}^d$.*

- *if $\beta_i > 0$ then $[\nabla\tilde{\phi}_\alpha(\beta^\alpha)]_i$ converges to 1*
- *if $\beta_i < 0$ then $[\nabla\tilde{\phi}_\alpha(\beta^\alpha)]_i$ converges to -1.*

Moreover if we assume that $\nabla\tilde{\phi}_\alpha(\beta^\alpha)$ converges to $\eta \in \mathbb{R}^d$, we have that:

- $\eta_i \in (-1,1) \Rightarrow \beta_i = 0$
- $\beta_i = 0 \Rightarrow \eta_i \in [-1,1]$.

Overall, assuming that $(\beta^\alpha, \nabla\tilde{\phi}_\alpha(\beta^\alpha)) \xrightarrow[\alpha\to 0]{} (\beta,\eta)$, we can write:

$$\eta \in \partial\|\beta\|_1.$$

Proof. We have that

$$\begin{aligned}[\nabla\tilde{\phi}_\alpha(\beta^\alpha)]_i &= \frac{1}{2\ln(1/\alpha)}\operatorname{arcsinh}\Big(\frac{\beta_i^\alpha}{\alpha^2}\Big)\\ &= \frac{1}{2\ln(1/\alpha)}\ln\Big(\frac{\beta_i^\alpha}{\alpha^2} + \sqrt{\frac{(\beta_i^\alpha)^2}{\alpha^4}+1}\Big).\end{aligned}$$

Now assume that $\beta_i^\alpha \to \beta_i > 0$, then $[\nabla\tilde{\phi}_\alpha(\beta^\alpha)]_i \to 1$, if $\beta_i < 0$ we conclude using that arcsinh is an odd function. All the claims are simple consequences of this. □

The following lemma shows that the extracted limits $\hat{t}$ as defined in Proposition 27 diverge to ∞. This divergence is crucial as it implies that the rescaled iterates $(\hat{\beta}_\tau)_\tau$ explore the whole trajectory..

Lemma 8. *For any extracted limit $\hat{t}$ as defined in Proposition 27, we have that $\tau - C \le \hat{t}_\tau$ where C is the upperbound on the length of the curves defined in Proposition 25.*

Proof. Recall that

$$\tau^\alpha(t) = t + \int_0^t \|\dot{\tilde{\beta}}_s^\alpha\|\mathrm{d}s.$$

From Proposition 25, the full path length $\int_0^{+\infty}\|\dot{\beta}_s^\alpha\|\mathrm{d}s$ is finite and bounded by some constant C independently of α. Therefore τ^α is a bijection in $\mathbb{R}_{\ge 0}$ and we defined $\hat{t}_\tau^\alpha = (\tau^\alpha)^{-1}(\tau)$. Furthermore $\tau^\alpha(t) \le t + C$ leads to $t \le \hat{t}^\alpha(t+C)$ and therefore $\tau - C \le \hat{t}^\alpha(\tau)$ for all $\tau \ge 0$. This inequality still holds for any converging subsequence, which proves the result. □

Under a mild additional assumption on the data (see Assumption 15), we showed after the proof of Proposition 11 in Appendix A.2 that the number of saddles of F is finite. Without this assumption, the number of saddles is *a priori* not finite. However the following lemma shows that along the flow of $\hat{\beta}$ the number of saddles which can potentially be visited is indeed finite.

Lemma 9. *The limiting flow $\hat{\beta}$ as defined in Proposition 27 can only visit a finite number of critical points $\beta \in S := \{\beta \in \mathbb{R}^d, \beta \odot \nabla L(\beta) = \mathbf{0}\}$ and can visit each one of them at most once.*

Proof. Let $\tau \geq 0$, and assume that $\hat{\beta}_\tau \in S$, i.e., we are at a critical point at time τ. From Proposition 11, we have that

$$\hat{\beta}_\tau \in \underset{\beta_i=0 \text{ for } i \notin \text{supp}(\hat{\beta}_\tau)}{\arg\min} L(\beta), \tag{A.19}$$

Let us define the sets

$$I_\pm \coloneqq \{i \in \{1, \dots, d\}, \text{ s.t. } \hat{s}_\tau(i) = \pm 1\} \quad \text{and} \quad I = I_+ \cup I_-.$$

The set I corresponds to the coordinate of $\hat{\beta}_\tau$ which "are allowed" (but not obliged) to be non-zero since from eq. (A.12), $\text{supp}(\hat{\beta}_\tau) \subset I$. Now given the fact that the sub-matrix $X_I = (\tilde{x}_i)_{i \in I} \in \mathbb{R}^{n \times \text{card}(I)}$ is full rank (see part (1) of the proof of Proposition 26 for the explanation), the solution of the minimisation problem (A.19) is unique and equal to $\beta[\xi] = (X_\xi^\top X_\xi)^{-1} X_\xi^\top y$ and $\beta[\xi^C] = 0$ where $\xi = \text{supp}(\hat{\beta}_\tau)$. There are $2^d = \text{Card}(P([d]))$ (where $P([d])$ contains all the subsets of $[d]$) number of constraints of the form $\{\beta_i = 0, i \notin \mathcal{A}\}$, where $\mathcal{A} \subset [d]$, and $\hat{\beta}_\tau$ is the unique solution of one of them. $\hat{\beta}_\tau$ can therefore take at most 2^d values (very crude upperbound). There is therefore a finite number of critical points which can be reached by the flow $\hat{\beta}$. Furthermore, from Proposition 27, the loss is strictly decreasing along the heteroclinic orbits, each of these critical points can therefore be visited at most once. □

Appendix B

Appendix for Chapter 7

Organisation of the Appendix. The Appendix is structured as follows.

- In Section B.1, we give more precisions regarding the way we model stochastic gradient descent as a stochastic gradient flow.
- Section B.2 is the core of the Appendix as it provides the proof of the theorem in a self-contained fashion.
- In Section B.3, we provide more experiments supporting our results.
- Finally, Section B.5 provides the technical material needed for the proofs of our results.

B.1 Details on the SDE modelling

We recall that the SGD recursion writes for $t \geqslant 1$ as:

$$\begin{aligned} w_{t+1,+} &= w_{t,+} - \gamma \langle \beta_w - \beta^\star, x_{i_t} \rangle \; x_{i_t} \odot w_{t,+} \\ w_{t+1,-} &= w_{t,-} + \gamma \langle \beta_w - \beta^\star, x_{i_t} \rangle \; x_{i_t} \odot w_{t,-} \end{aligned} \qquad \text{where } i_t \sim \mathrm{Unif}(1,n).$$

Since the full gradient is $\nabla_{w_\pm} L(w) = \pm \big[\frac{1}{n} \sum_{k=1}^n \langle \beta_w - \beta^\star, x_k \rangle \; x_k \big] \odot w_\pm \in \mathbb{R}^d$. We can rewrite the recursion as:

$$w_{t+1,\pm} = w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \mp \gamma \Big[\langle \beta_{w_t} - \beta^\star, x_{i_t} \rangle \; x_{i_t} - \frac{1}{n} \sum_{k=1}^n \langle \beta_{w_t} - \beta^\star, x_k \rangle \; x_k \Big] \odot w_{t,\pm}.$$

Now notice that

$$\langle \beta - \beta^\star, x_{i_t} \rangle \; x_{i_t} - \frac{1}{n} \sum_{k=1}^n \langle \beta - \beta^\star, x_k \rangle \; x_k = X^\top \Big(\langle \beta - \beta^\star, x_{i_t} \rangle \mathbf{e}_{i_t} - \mathbb{E}_{i_t} \big[\langle \beta - \beta^\star, x_{i_t} \rangle \mathbf{e}_{i_t} \big] \Big),$$

where $\mathbf{e}_i$ is the i^{th} element of the $\mathbb{R}^n$-canonical basis. Let us denote by $\xi_{i_t}(\beta) = -\big(\langle \beta - \beta^\star, x_{i_t} \rangle \mathbf{e}_{i_t} - \mathbb{E}_{i_t} \big[\langle \beta - \beta^\star, x_{i_t} \rangle \mathbf{e}_{i_t} \big] \big)$. It is a zero-mean random variable with values in $\mathbb{R}^n$ and it can be seen as a multiplicative noise, i.e., proportional to $\beta - \beta^\star$, which vanishes at the optimum. The SGD recursion then writes as:

$$\begin{aligned} w_{t+1,\pm} &= w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \pm \gamma \big[X^\top \xi_{i_t}(\beta_t) \big] \odot w_{t,\pm} \\ &= w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \pm \gamma \, \mathrm{diag}(w_{t,\pm}) X^\top \xi_{i_t}(\beta_t). \end{aligned}$$

As we are interested in the stochastic differential model of the SGD recursion, let us now compute the covariance of the SGD noise. We first notice that

$$\begin{aligned} \mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] &= \mathbb{E}_{i_t}[\xi_{i_t}(\beta)^{\otimes 2}] \\ &= \mathbb{E}_{i_t}[(\langle \beta - \beta^\star, x_{i_t} \rangle \mathbf{e}_{i_t})^{\otimes 2}] - \mathbb{E}_{i_t}[\langle \beta - \beta^\star, x_{i_t} \rangle \mathbf{e}_{i_t}]^{\otimes 2} \\ &= \frac{1}{n} \begin{pmatrix} \langle \beta - \beta^\star, x_1 \rangle^2 & & 0 \\ & \ddots & 0 \\ 0 & & \langle \beta - \beta^\star, x_n \rangle^2 \end{pmatrix} - \frac{1}{n^2} \Big(\langle \beta - \beta^\star, x_i \rangle \langle \beta - \beta^\star, x_j \rangle \Big)_{1 \leq i,j \leq n} \\ &= \frac{4}{n} \begin{pmatrix} L_1(\beta) & & 0 \\ & \ddots & 0 \\ 0 & & L_n(\beta) \end{pmatrix} - \frac{1}{n^2} \Big(\langle \beta - \beta^\star, x_i \rangle \langle \beta - \beta^\star, x_j \rangle \Big)_{1 \leq i,j \leq n} \end{aligned}$$

where $L_i(\beta) = \frac{1}{4} \langle \beta - \beta^\star, x_i \rangle^2$ is the individual loss of the observation x_i, such that $L(\beta) = \frac{1}{n} \sum_{i=1}^n L_i(\beta)$.

Thus, the covariance satisfies the relation $\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] = \frac{4}{n} \mathrm{diag}(L_i(\beta))_{1 \leq i \leq n} + O(\frac{1}{n^2})$. From this expression we can obtain a good model for $\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)]$. First, we neglect the second term of order $1/n^2$. Then, we assume that all partial losses are approximately uniformly equal to their mean: i.e. for any i, $L_i(\beta) \cong \mathbb{E}_{i_t}[L_{i_t}(\beta)]$ (the general case is discussed Appendix B.4.1). Hence,

$$\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] \cong \frac{4}{n} \mathrm{diag} \Big(\frac{1}{n} \sum_i L_i(\beta) \Big) = \frac{4}{n} L(\beta) I_n.$$

The overall SGD's noise structure is then captured by

$$\begin{aligned}\Sigma_{\text{SGD}}(w_\pm) &:= \gamma^2 \operatorname{diag}(w_\pm) X^\top \operatorname{Cov}_{i_t}[\xi_{i_t}(\beta)] X \operatorname{diag}(w_\pm) \\ &\cong \frac{4}{n}\gamma^2 L(\beta)[\operatorname{diag}(w_\pm)X^\top]^{\otimes 2}.\end{aligned}$$

This leads us in considering the following SDE:

$$\begin{aligned}\mathrm{d}w_{t,+} &= -\nabla_{w_+} L(w_t)\,\mathrm{d}t + 2\sqrt{\gamma n^{-1} L(w_t)}\; w_{t,+} \odot [X^\top \mathrm{d}B_t] \\ \mathrm{d}w_{t,-} &= -\nabla_{w_-} L(w_t)\,\mathrm{d}t - 2\sqrt{\gamma n^{-1} L(w_t)}\; w_{t,-} \odot [X^\top \mathrm{d}B_t],\end{aligned}$$

since its Euler discretisation with step size γ is :

$$w_{t+1,\pm} = w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \pm 2\sqrt{\gamma n^{-1} L(w_t)}\; w_{t,\pm} \odot [X^\top \varepsilon_t],$$

where $\varepsilon_t \sim \mathcal{N}(0, \sqrt{\gamma} I_n)$. This corresponds to a Markov-Chain whose noise covariance is equal to Σ_{SGD}.

Remark on mini-batch SGD. This analysis can easily be extended to a batch size larger than 1. Indeed, using a mini-batch sampled with replacement of size b only changes the noise covariance up to a multiplicative constant as: $\operatorname{Cov}_{i_t}[\xi^b_{i_t}(\beta)] = \frac{1}{b}\operatorname{Cov}_{i_t}[\xi^{b'=1}_{i_t}(\beta)]$. The associated SDE, for a step size γ, is therefore $\mathrm{d}w_{t,\pm} = -\nabla_{w_\pm} L(w_t)\,\mathrm{d}t \pm 2\sqrt{\gamma b^{-1} n^{-1} L(w_t)}\; w_{t,\pm} \odot [X^\top \mathrm{d}B_t]$. Hence, it the same SDE as for a batch-size equal to 1 but with an effective step-size $\gamma_{\text{eff}} = \gamma/b$ (hence larger step-sizes can be used, as expected). The exact same reasoning can be done for mini-batch without replacement and our analysis would hold this time with: $\gamma_{\text{eff}} = \gamma(n-b)/((n-1)b)$. Note that all the results therefore hold for mini-batch SGD by considering the effective step-size γ_{eff} instead of γ.

B.2 Proofs of the main results

This section contains all the proofs of the main results. It is self contained as we recall each time the propositions we prove. In subsection B.2.1, we derive the mirror-descent-like flow which the iterates follow as in Proposition 14 of the main text. Then, we upper bound the loss integral in subsection B.2.2. This leads us in proving the convergence of the iterates towards an interpolator in subsection B.2.3. Equipped with these results we prove the main result (Theorem 1) in subsection B.2.4. Finally, to complete the proof of Proposition 16 of the main text we derive a lower bound of the loss in subsection B.2.5.

For the sake of easy reading, we adopt the following notations in this section: we denote by $\bar{X} := X/\sqrt{n}$, and $\lambda_{\max} := \lambda_{\max}(\bar{X}^\top \bar{X})$.

B.2.1 Proof of Proposition 14

In order to prove Proposition 14, we introduce the following lemma:

Lemma 10. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow in Eq.(7.3) with initialisation $w_{0,\pm} = \alpha \in (\mathbb{R}_+^*)^d$. Then we have the following implicit closed form expression for β_t:*

$$\beta_t = 2\alpha_t^2 \odot \sinh(2\bar{X}^\top \eta_t), \tag{B.1}$$

where $\eta_t = -\int_0^t \bar{X}(\beta_s - \beta^\star)\,\mathrm{d}s + 2\sqrt{\gamma}\int_0^t \sqrt{L(\beta_s)}\mathrm{d}B_s \in \mathbb{R}^n$ and $\alpha_t = \alpha \odot \exp\left(-2\gamma\,\mathrm{diag}(\bar{X}^\top\bar{X})\int_0^t L(\beta_s)\mathrm{d}s\right)$.

Note that this **is not** an explicit closed form for β_t since the right hand side depends on $(\beta_s)_{0\leq s\leq t}$.

Proof. Recall that the SDE we consider writes as:

$$\begin{aligned}\mathrm{d}w_{t,\pm} &= -\nabla_{w_\pm} L(w_t)\,\mathrm{d}t \pm 2\sqrt{\gamma n^{-1} L(w_t)}\; w_{t,\pm} \odot [X^\top \mathrm{d}B_t] \\ &= \pm\big(-[\bar{X}^\top r(w_t)] \odot w_{t,\pm}\,\mathrm{d}t + 2\sqrt{\gamma L(w_t)}\; w_{t,\pm} \odot [\bar{X}^\top \mathrm{d}B_t]\big),\end{aligned}$$

where $r(w) = \bar{X}(w_+^2 - w_-^2 - \beta^\star) = \bar{X}(\beta_w - \beta^\star) \in \mathbb{R}^n$ are the (normalised) rests.

It turns out that there is an implicit closed form solution to this SDE. Indeed deriving the Itô formula on $\ln(w_{t,\pm})$ gives the following integral expression:

$$\begin{aligned}w_{t,\pm} &= w_{t=0,\pm} \odot \exp(\pm\bar{X}^\top\Big[-\int_0^t r(w_s)\,\mathrm{d}s + 2\sqrt{\gamma}\int_0^t \sqrt{L(w_s)}\,\mathrm{d}B_s\Big]) \odot \exp(-2\gamma\,\mathrm{diag}(\bar{X}^\top\bar{X})\int_0^t L(w_s)\,\mathrm{d}s) \\ &= \alpha_t \odot \exp(\pm\bar{X}^\top\eta_t).\end{aligned}$$

Since $\beta = w_+^2 - w_-^2$, we get:

$$\begin{aligned}\beta_t &= \alpha_t^2 \odot \big(\exp(+2\bar{X}^\top\eta_t) - \exp(-2\bar{X}^\top\eta_t)\big) \\ &= 2\alpha_t^2 \odot \sinh(+2\bar{X}^\top\eta_t).\end{aligned}$$

□

For clarity we recall the statement of Proposition 14.

Proposition 14. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow in Eq.(7.3) with initialisation $w_{0,\pm} = \alpha \in (\mathbb{R}_+^*)^d$. Then the corresponding flow $(\beta_t)_{t\geq 0}$ follows a "stochastic continuous mirror descent with time varying potential" defined by:*

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma n^{-1} L(\beta_t)}X^\top \mathrm{d}B_t, \tag{7.7}$$

where $\alpha_t = \alpha \odot \exp\big(-2\gamma \operatorname{diag}\left(\frac{X^\top X}{n}\right) \int_0^t L(\beta_s)\,\mathrm{d}s\big)$ *and* ϕ_α *is the hyperbolic entropy defined in* (7.4).

Proof. The results immediately follows from Lemma 10. Indeed, inverting the implicit equation on β_t, Eq. (B.1), we have,

$$\operatorname{arcsinh}\Big(\frac{\beta_t}{2\alpha_t^2}\Big) = 2X^\top \eta_t = -2\bar{X}^\top \int_0^t \bar{X}(\beta_s - \beta^\star)\,\mathrm{d}s + 4\sqrt{\gamma}\bar{X}^\top \int_0^t \sqrt{L(\beta_s)}\mathrm{d}B_s.$$

Hence,

$$\begin{aligned}\mathrm{d}\operatorname{arcsinh}\Big(\frac{\beta_t}{2\alpha_t^2}\Big) &= -2\bar{X}^\top\bar{X}(\beta_s - \beta^\star)\,\mathrm{d}t + 4\sqrt{\gamma}\bar{X}^\top\sqrt{L(\beta_t)}\mathrm{d}B_t \\ &= -4\nabla L(\beta_t)\,\mathrm{d}t + 4\sqrt{\gamma L(\beta_t)}\bar{X}^\top \mathrm{d}B_t.\end{aligned}$$

Noticing that $\nabla\phi_\alpha(\beta) = \frac{1}{4}\operatorname{arcsinh}(\frac{\beta}{2\alpha^2})$ concludes the proof. □

B.2.2 Upperbound of the integral of the loss

This section contains several technical arguments that permit us to derive the upperbound of the integral of the loss [Proposition 16, right side]. Let us try to highlight the key features of this proof. First, as for classical mirror descent, we define a Lyapunov function that resembles a Bregman divergence plus a necessary control term [Eq. (B.2)]. Then, we fix a high-probability event on which we have a control of the Brownian diffusion term [Eq. (B.3)]. This gives an equation involving a weighted integral of the loss. After lower bounding this weight to access directly the loss integral [Lemma 13], we show that the iterates themselves are in fact bounded [Lemma 12]. We finally conclude the proof in Proposition 28.

Notations and standard calculations. Let us introduce some notations that are important throughout the proofs. We consider the hyperbolic entropy $\phi_\alpha(\beta)$ as a function of two variables $(y,z) \mapsto \phi(y,z)$ evaluated at the point $(\beta,\alpha^2) \in \mathbb{R}^d \times \mathbb{R}^d$. With a slight abuse of notation, we denote by $\nabla_\beta\phi(\beta,\alpha^2) \in \mathbb{R}^d$, the gradient with respect to the first vector evaluated in (β,α^2), and $\nabla_z\phi(\beta,\alpha^2) \in \mathbb{R}^d$, the gradient with respect to the second variable evaluated in (β,α^2). Let us also define the process $(\xi_t)_{t\geqslant 0}$, as the vector $\xi_t := \sqrt{\beta_t^2 + 4\alpha_t^4} \in \mathbb{R}^d$, for all $t \geqslant 0$. For the sake of clarity, we recall here the expression of the hyperbolic entropy as well as its derivatives: we have $\phi(\beta,\alpha^2) = \frac{1}{4}\sum_{i=1}^d \beta_i \operatorname{arcsinh}(\frac{\beta_i}{2\alpha_i^2}) - \sqrt{\beta_i^2 + 4\alpha_i^4}$, and

$$\nabla_\beta\phi(\beta,\alpha^2) = \frac{1}{4}\operatorname{arcsinh}\left(\frac{\beta}{2\alpha^2}\right), \quad \nabla_z\phi(\beta,\alpha^2) = -\frac{1}{4\alpha^2}\sqrt{\beta^2+4\alpha^4} \in \mathbb{R}^d \quad \text{as well as,}$$

$$\nabla^2_{\beta,\beta}\phi(\beta,\alpha^2) = \frac{1}{4}\operatorname{diag}\left[\frac{1}{\sqrt{\beta_i^2+4\alpha_i^2}}\right]_i \in \mathbb{R}^{d\times d}.$$

A first Lyapunov function. In this subsection we shall consider the following (stochastic) Lyapunov function:

$$V_t := -\phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle + \gamma\int_0^t L(\beta_s)\mathrm{d}s\ \langle|\beta^*_{\ell_1}|, \operatorname{diag}(\bar{X}^\top\bar{X})\rangle. \tag{B.2}$$

This Lyapunov resembles to a Bregman divergence with respect to the hyperbolic entropy. The added term is however required to have a proper control on its decrease. Just as in the deterministic framework, we want to show that the Lyapunov is decreasing, i.e. it has a negative derivative. With this aim, we compute its Îto derivative $\mathrm{d}V_t$ in the following lemma.

Lemma 11. *For all $t > 0$, V_t verifies the following equation:*

$$V_t = V_0 - 2\int_0^t L(\beta_s)\Big(1 - \frac{1}{2}\gamma\langle \mathrm{diag}(\bar{X}^\top \bar{X}), \xi_s + |\beta^*_{\ell_1}|\rangle\Big)\mathrm{d}s + \int_0^t \sqrt{\gamma L(\beta_s)}\langle X^\top \mathrm{d}B_s, \beta_s - \beta^*_{\ell_1}\rangle.$$

Proof. To derive the formula for the Lyapunov V_t, we compute its derivatives $\mathrm{d}V_t$ thanks to Itô formula and then integrate it with respect to the time. Let us stress that as V_t is a function of β_t and α_t we need both their full Itô decomposition. For α_t, as we know that $\alpha_t = \alpha \odot \exp\big(-2\gamma\,\mathrm{diag}(\bar{X}^\top\bar{X})\int_0^t L(w_s)\mathrm{d}s\big)$, we have $\mathrm{d}\alpha_t = -2\gamma\,\mathrm{diag}(\bar{X}^\top\bar{X})L(w_t)\alpha_t\mathrm{d}t$. For β_t, we only need the noise compound of the Itô decomposition. Let us denote by $b(\beta_{w_t})$ the drift in the Itô decomposition of β_t[1], we have,

$$\begin{aligned}
\mathrm{d}\beta_t &= \mathrm{d}w^2_{t,+} - \mathrm{d}w^2_{t,-}\\
&= b(\beta_{w_t})\mathrm{d}t + 4\sqrt{\gamma L(\beta_t)}(w_{t,+}\odot w_{t,+}\odot\left[\bar{X}^\top \mathrm{d}B_t\right] + w_{t,-}\odot w_{t,-}\odot\left[\bar{X}^\top \mathrm{d}B_t\right])\\
&= b(\beta_{w_t})\mathrm{d}t + 4\sqrt{\gamma L(\beta_t)}\;\xi_t\odot\left[\bar{X}^\top \mathrm{d}B_t\right].
\end{aligned}$$

From this expression, we deduce the matrix of its quadratic variations $\mathrm{d}\langle\beta_t\rangle_{\mathrm{qv}} = \left[\mathrm{d}\langle\beta^i_t, \beta^j_t\rangle\right]_{ij} = 16\gamma L(\beta_t)(\bar{X}^\top\bar{X})\odot(\xi_t\xi_t^\top) \in \mathbb{R}^{d\times d}$.

We are now equipped to apply the Itô formula on V_t. Indeed, it is clear that ϕ is a C^2 function of (β,α), hence,

$$\begin{aligned}
\mathrm{d}V_t = -\Big[&\langle\nabla_\beta\phi(\beta_t,\alpha_t^2), \mathrm{d}\beta_t\rangle + \langle\nabla_z\phi(\beta_t,\alpha_t^2), \mathrm{d}\left[\alpha_t^2\right]\rangle + \frac{1}{2}\mathrm{Tr}\left[\nabla^2_{\beta,\beta}\phi(\beta_t,\alpha_t^2)\mathrm{d}\langle\beta_t\rangle\right]\Big]\\
&+ \mathrm{d}\left[\langle\nabla_\beta\phi(\beta_t,\alpha_t^2), \beta_t - \beta^*_{\ell_1}\rangle\right] + \gamma L(\beta_t)\langle|\beta^*_{\ell_1}|, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle\mathrm{d}t.
\end{aligned}$$

The fifth term is explicit. Let us treat the first four terms separately:

First term. This term cancels with a compound of the fourth term.

Second term. We apply simply the chain rule for this term as α_t does not have any quadratic variation:

$$\langle\nabla_z\phi(\beta_t,\alpha_t^2), \mathrm{d}\left[\alpha_t^2\right]\rangle = \left\langle -\frac{\xi_t}{4\alpha_t^2}, 2\alpha_t\odot\mathrm{d}\alpha_t\right\rangle = \gamma L(\beta_t)\left\langle\xi_t, \mathrm{diag}(\bar{X}^\top\bar{X})\right\rangle\mathrm{d}t.$$

Third term. We directly see that

$$\frac{1}{2}\mathrm{Tr}\left[\nabla^2_{\beta,\beta}\phi(\beta_t,\alpha_t^2)\mathrm{d}\langle\beta_t\rangle\right] = \frac{1}{2}\mathrm{Tr}\left[\frac{1}{4}\,\mathrm{diag}\left(\frac{1}{\xi_t}\right)\cdot 4\gamma L(\beta_t)\bar{X}^\top\bar{X}\odot(\xi_t\xi_t^\top)\right]\mathrm{d}t = 2\gamma L(\beta_t)\langle\xi_t, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle\mathrm{d}t.$$

Fourth term. We apply Itô formula once again to get:

$$\mathrm{d}\left[\langle\nabla_\beta\phi(\beta_t,\alpha_t^2), \beta_t - \beta^*_{\ell_1}\rangle\right] = \langle\mathrm{d}\left[\nabla_\beta\phi(\beta_t,\alpha_t^2)\right], \beta_t - \beta^*_{\ell_1}\rangle + \langle\nabla_\beta\phi(\beta_t,\alpha_t^2), \mathrm{d}\beta_t\rangle + \mathrm{Tr}\left[\mathrm{d}\langle\nabla_\beta\phi(\beta_t,\alpha_t^2), \beta_t\rangle_{\mathrm{vq}}\right],$$

and thanks to Eq. (7.7), we have an expression for the first and last term, giving

$$\begin{aligned}
\mathrm{d}\left[\langle\nabla_\beta\phi(\beta_t,\alpha_t^2), \beta_t - \beta^*_{\ell_1}\rangle\right] = &-\langle\nabla L(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle\mathrm{d}t + 2\sqrt{\gamma L(\beta_t)}\langle\bar{X}^\top\mathrm{d}B_t, \beta_t - \beta^*_{\ell_1}\rangle + \langle\nabla_\beta\phi(\beta_t,\alpha_t^2), \mathrm{d}\beta_t\rangle\\
&+ 4\gamma L(\beta_t)\langle\xi_t, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle\mathrm{d}t.
\end{aligned}$$

Final expression. Let us gather the four expressions to get $\mathrm{d}V_t$. We remark that the terms $\langle\nabla_\beta\phi(\beta_t,\alpha_t^2), \mathrm{d}\beta_t\rangle$ cancels (from first and fourth terms) and since $\langle\nabla_\beta L(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle = 2L(\beta_t)$,

$$\begin{aligned}
\mathrm{d}V_t = &-\left[\gamma L(\beta_t)\left\langle\xi_t, \mathrm{diag}(\bar{X}^\top\bar{X})\right\rangle\mathrm{d}t + 2\gamma L(\beta_t)\langle\xi_t, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle\mathrm{d}t\right] - 2L(\beta_t)\\
&+ \sqrt{\gamma L(\beta_t)}\langle\bar{X}^\top\mathrm{d}B_t, \beta_t - \beta^*_{\ell_1}\rangle + 4\gamma L(\beta_t)\langle\xi_t, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle\mathrm{d}t + \gamma L(\beta_t)\langle|\beta^*_{\ell_1}|, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle\mathrm{d}t.
\end{aligned}$$

[1] It can be computed but its precise formula is not needed.

And finally, we have the expression:

$$\mathrm{d}V_t = -2L(\beta_t) + \gamma L(\beta_t)\left\langle \xi_t, \mathrm{diag}(\bar{X}^\top \bar{X})\right\rangle \mathrm{d}t + \gamma L(\beta_t)\langle |\beta^*_{\ell_1}|, \mathrm{diag}(\bar{X}^\top \bar{X})\rangle \mathrm{d}t \\ + \sqrt{\gamma L(\beta_t)}\langle \bar{X}^\top \mathrm{d}B_t, \beta_t - \beta^*_{\ell_1}\rangle.$$

Integrating this equation between 0 and t concludes the proof. □

Control of the martingale term and definition of $\mathcal{A}$. Lemma 11 shows that in order to control V_t, we need to control the local martingale $S_t = \sqrt{\gamma}\int_0^t \sqrt{L(\beta_s)}\langle \bar{X}^\top \mathrm{d}B_s, \beta_s - \beta^*_{\ell_1}\rangle$. In fact, it is expected that the deviation of S_t from its quadratic variation is very small: this is a concentration property of local martingales similar to the Bernstein inequality for discrete ones [Boucheron et al., 2013]. To do so, let us fix $p < 1/2$ and we define two parameters: $a := \max\{\|\beta^\star_{ell_1}\|_1 \ln(\sqrt{2}\frac{\|\beta^\star_{\ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\}$ and $b := \frac{1}{2}\ln(4/p)a^{-1}$. The reason behind the precise value of a will appear clearly in the proof of Lemmas 12 and 13. These parameters being fixed, we can define the event:

$$\mathcal{A} = \{\forall t \geq 0, |S_t| \leq a + 2b\gamma\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^\star_{\ell_1}\|_1^2)\mathrm{d}s\}. \tag{B.3}$$

From Lemma 18, we know that $\mathbb{P}(\mathcal{A}) \geq 1 - 2\exp(-2ab) = 1 - \frac{p}{2}$. Note that p is a free parameter that can be chosen as small as we want.

From now on and until the end of the Section, we place ourselves *on* the event $\mathcal{A}$, that is, all (in)equalities between random variables should be considered pointwise for any $\omega \in \mathcal{A}$. To make it clear, we will recall from time to time laconically this fact by writing, "on $\mathcal{A}$".

From Lemma 11, we deduce the following inequalities,

$$V_t - V_0 \leq -2\int_0^t L(\beta_s)(1 - \frac{1}{2}\gamma\langle \mathrm{diag}(\bar{X}^\top \bar{X}), \xi_s + |\beta^*_{\ell_1}|\rangle)\mathrm{d}s + 2b\gamma\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\mathrm{d}s + a \\ \leq -2\int_0^t L(\beta_s)(1 - \frac{1}{2}\gamma\langle \mathrm{diag}(\bar{X}^\top \bar{X}), \xi_s + |\beta^*_{\ell_1}|\rangle - b\gamma\lambda_{\max}(\|\beta_s\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\mathrm{d}s + a.$$

Hence, we have the following control on V_t with respect to a weighted loss integral:

$$V_t - V_0 \leq -2\int_0^t L(\beta_s)U_s\mathrm{d}s + a, \tag{B.4}$$

where $U_t := 1 - \frac{\gamma}{2}\left[\langle \mathrm{diag}(\bar{X}^\top \bar{X}), \xi_t + C|\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}(\|\beta_t\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\right] \leq 1$. The following lemma show that as long as U_t stays positive, the iterates stay bounded.

Lemma 12. *Let us place ourselves on the event $\mathcal{A}$. Let $\tau > 0$. Assume $(U_t)_{0\leq t\leq \tau}$ is positive. Then for all $t \leqslant \tau$ we have the following explicit upper bound on both $\|\beta_t\|_1$ and $\|\xi_t\|_1$,*

$$\|\beta_t\|_1 \leq \|\xi_t\|_1 \leq 18\max\{\|\beta^\star_{\ell_1}\|_1 \ln(\sqrt{2}\frac{\|\beta^\star_{\ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\}.$$

Proof. Let $t \leq \tau$. Remember that $\alpha(t) = \alpha \odot \exp\Big(-2\gamma\big(\int_0^t L(w_s)\mathrm{d}s\big)\,\mathrm{diag}(\bar{X}^\top \bar{X})\Big) \in \mathbb{R}^d$. Since $V_t \leq V_0 - 2\int_0^t L(\beta_s)U(s)\mathrm{d}s + a$ and since by assumption $U(s) \geq 0$ for all $s \leq t$, we immediately get

that $V_t \leq V_0 + a = -\phi_\alpha(0) + a = \frac{1}{2}\|\alpha\|_2^2 + a$. Notice furthermore that $-\phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle = \frac{1}{4}\|\xi_t\|_1 - \frac{1}{4}\langle \operatorname{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle$. Hence, we have:

$$\begin{aligned}
\|\xi_t\|_1 &= -4\phi_{\alpha_t}(\beta_t) + 4\langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle + \langle \operatorname{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle \\
&= 4V_t - 4\gamma\int_0^t L(\beta_s)\mathrm{d}s\ \langle |\beta^*_{\ell_1}|, \operatorname{diag}(\bar{X}^\top\bar{X})\rangle + \langle \operatorname{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle \\
&\leq 2\|\alpha\|_2^2 + 4a + \langle \operatorname{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle - 4\gamma\int_0^t L(\beta_s)\mathrm{d}s\ \langle |\beta^*_{\ell_1}|, \operatorname{diag}(\bar{X}^\top\bar{X})\rangle.
\end{aligned}$$

We now use the fact that $\operatorname{arcsinh}(x) \leq \ln(2(x+1))$ and that $|x| + |y| \leq \sqrt{2}\sqrt{x^2+y^2}$ for all $x, y \geq 0$.

$$\begin{aligned}
\|\xi_t\|_1 &\leq 2\|\alpha\|_2^2 + 4a + \sum_i |\beta_i^\star| \ln\left(\frac{|\beta_i(t)| + 2\alpha_i(t)^2}{\alpha_i(t)^2}\right) - 4\gamma\int_0^t L(\beta_s)\mathrm{d}s\ \langle |\beta^*_{\ell_1}|, \operatorname{diag}(\bar{X}^\top\bar{X})\rangle \\
&\leq 2\|\alpha\|_2^2 + 4a + \sum_i |\beta_i^\star| \ln\left(\sqrt{2}\frac{\sqrt{|\beta_i(t)|^2 + 4\alpha_i(t)^4}}{\min\alpha_i^2}\right) - \sum_i |\beta_i^\star| \ln\left(\exp\Big(-4\gamma\int_0^t L(\beta_s)\mathrm{d}s\,\operatorname{diag}(\bar{X}^\top\bar{X})\Big)\right) \\
&\qquad - 4\gamma\int_0^t L(\beta_s)\mathrm{d}s\ \langle |\beta^*_{\ell_1}|, \operatorname{diag}(\bar{X}^\top\bar{X})\rangle.
\end{aligned}$$

Since the last two terms cancel and for all i, $\sqrt{|\beta_i(t)|^2 + 4\alpha_i(t)^4} \leqslant \|\xi\|_1$, we have

$$\|\xi_t\|_1 \leq 2\|\alpha\|_2^2 + 4a + \|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\xi_t\|_1}{\min\alpha_i^2}\right).$$

To obtain the explicit upperbound we use Lemma 19 with $A = \frac{2\sqrt{2}\|\alpha\|^2}{\min\alpha_i^2} + \frac{4a\sqrt{2}}{\min\alpha_i^2}$ and $B = \frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min\alpha_i^2}$ since the condition on A, B are satisfied as $\frac{A}{B} + \ln(B) \geq \frac{2\|\alpha\|_2^2}{\|\beta^*_{\ell_1}\|_1} + \ln(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min\alpha_i^2}) \geq 1 + \ln(\sqrt{8}d) \geq 2$, as soon as $d \geq 3$. Hence,

$$\begin{aligned}
\|\beta_t\|_1 \leq \|\xi_t\|_1 &\leq \frac{5}{2}\left(2\|\alpha\|_2^2 + 4a + \|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min\alpha_i^2}\right)\right) \\
&\leq 3\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min\alpha_i^2}\right) + 5\|\alpha\|_2^2 + 10a \\
&\leq 18\max\{\|\beta^\star_{\ell_1}\|_1 \ln(\sqrt{2}\frac{\|\beta^\star_{\ell_1}\|_1}{\min\alpha_i^2}), \|\alpha\|_2^2\},
\end{aligned}$$

where in the last inequality we plug in the value of a. This concludes the proof of the lemma. □

Recall that we defined $U_t = 1 - \frac{\gamma}{2}\big[\langle \operatorname{diag}(\bar{X}^\top\bar{X}), \xi_t + |\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}(\|\beta_t\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\big]$. We now show that in fact $(U_t)_t$ is always lower bounded by a strictly positive constant. Hence, the result of Lemma 12 is valid at any time $t > 0$.

Lemma 13. *On $\mathcal{A}$, let us fix $\gamma \leq [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^\star_{\ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^\star_{ell_1}\|_1}{\min\alpha_i^2}), \|\alpha\|_2^2\}]^{-1}$. Recall that $U_t = 1 - \frac{\gamma}{2}\big[\langle \operatorname{diag}(\bar{X}^\top\bar{X}), \xi_t + |\beta^\star_{\ell_1}|\rangle + 2b\lambda_{\max}(\|\beta_t\|_1^2 + \|\beta^\star_{\ell_1}\|_1^2)\big]$, then for all $t \geq 0$,*

$$U_t \geq \frac{1}{2}.$$

Proof. Let us define the stopping time $\tau = \inf\{t \geq 0 \text{ such that } U(t) \leq \frac{1}{2}\}$. Note that

$$\begin{aligned} U_0 &= 1 - \frac{\gamma}{2}\left[\langle \operatorname{diag}(\bar{X}^\top \bar{X}), 2\alpha^2 + |\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}\|\beta^*_{\ell_1}\|_1^2)\right] \\ &\geq 1 - \frac{\gamma}{2}\lambda_{\max}\left[2\|\alpha\|_2^2 + \|\beta^*_{\ell_1}\|_1 + 2b\|\beta^*_{\ell_1}\|_1^2\right) \\ &\geq 1 - 2\gamma\lambda_{\max} a \ln(\frac{4}{p}) \\ &> \frac{1}{2}, \end{aligned}$$

where the last inequality comes from the upperbound on γ. Since U_t is continuous we have that $\tau > 0$. Assume that $\tau < +\infty$, by definition of the stopping time, for $t \leq \tau$: $U(t) \geq 0$ and we can apply Lemma 12 at time τ:

$$\|\beta_\tau\|_1 \leq \|\xi_\tau\|_1 \leq 18 \max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\}.$$

Therefore:

$$\begin{aligned} U_\tau &= 1 - \frac{\gamma}{2}\left[\langle \operatorname{diag}(\bar{X}^\top \bar{X}), \xi_\tau + |\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}(\|\beta_\tau\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\right] \\ &\geq 1 - \frac{\gamma}{2}\lambda_{\max}\left[\|\xi_\tau\|_1 + \|\beta^*_{\ell_1}\|_1 + 2b(\|\beta_\tau\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\right] \\ &\geq 1 - \frac{\gamma}{2}\lambda_{\max}\Big[18 \max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\} \\ &\qquad + 2 \cdot 18^2 \cdot b \max\{\|\beta^*_{\ell_1}\|_1^2 \ln^2\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^4\}\Big]. \end{aligned}$$

Since $b = \frac{1}{2}\ln(\frac{4}{p}) \max\{\|\beta^*_{\ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\}^{-1}$ we get that:

$$\begin{aligned} U_\tau &\geq 1 - \frac{\gamma}{2}\lambda_{\max}\ln(\frac{4}{p}) \max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\}[18 + 18^2] \\ &\geq 1 - 175\ln(\frac{4}{p})\gamma\lambda_{\max} \max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\} \\ &> \frac{1}{2}, \end{aligned}$$

where the last inequality comes from the choice of γ.

This is inconsistent since $U_\tau = \frac{1}{2}$. Hence $\tau = +\infty$ and thus $U_t \geq 1/2$ for all t. □

From the result of Lemma 13, with Equation (B.4), we obtain:

$$\int_0^t L(\beta_s)\mathrm{d}s \leq V_0 - V_t + a \leq -V_t + 2\max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\}. \tag{B.5}$$

Hence it remains to lower bound V_t in order to get the convergence of the integral of the loss.

Lemma 14. *On $\mathcal{A}$, let γ be set as in Lemma 12, for all $t > 0$, we have the following lower bound on V_t:*

$$V_t \geq -\frac{\|\beta^\star_{\ell_1}\|_1}{4} \ln\left(\frac{18\sqrt{2}}{\min \alpha_i^2} \max\left\{\|\beta^\star_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^\star_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\right\}\right).$$

Proof. We follow exactly the same proof as for upperbounding the iterates.

$$\begin{aligned}
4V_t &= \sum_i \sqrt{\beta_i^2 + 4\alpha_i(t)^4} - \langle \operatorname{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle + 4\gamma \int_0^t L(\beta_s)\mathrm{d}s \ \langle |\beta^*_{\ell_1}|, \operatorname{diag} H\rangle \\
&\geq \|\xi_t\|_1 - \sum_i |\beta_i^\star| \ln\left(\frac{|\beta_i(t)| + 2\alpha_i(t)^2}{\alpha_i(t)^2}\right) + 4\gamma \int_0^t L(\beta_s)\mathrm{d}s \ \langle |\beta^*_{\ell_1}|, \operatorname{diag} H\rangle \\
&\geq \|\xi_t\|_1 - \|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\xi_t\|_1}{\min \alpha_i^2}\right) \\
&\geq -\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\xi_t\|_1}{\min \alpha_i^2}\right) \\
&\geq -\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{18\sqrt{2}}{\min \alpha_i^2} \max\left\{\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\right\}\right).
\end{aligned}$$

□

Hence $(V_t)_{t\geq 0}$ is lowerbounded and we can derive an upper bound on the loss integral to show the right part of Proposition 16. We recall it here in the following proposition.

Proposition 28. *On $\mathcal{A}$, let γ be set as in Lemma 12, we have the following upper bound on the loss integral:*

$$\forall t > 0, \qquad \int_0^t L(\beta_s)\mathrm{d}s \leq \tilde{O}\left(\max\left\{\|\beta^\star_{ell_1}\|_1 \ln\left(\frac{\|\beta^\star_{ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\right\}\right).$$

As a consequence, the integral $\int_0^\infty L(\beta_s)\mathrm{d}s$ converges.

Proof. From Equation (B.5), we have that

$$\int_0^t L(\beta_s)\mathrm{d}s \leq -V_t + 2\max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\},$$

and thanks to the lower bound on V_t from Lemma 14, it yields,

$$\int_0^t L(\beta_s)\mathrm{d}s \leq \frac{\|\beta^*_{\ell_1}\|_1}{4} \ln\left(\frac{18\sqrt{2}}{\min \alpha_i^2} \max\left\{\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\right\}\right) + 2\max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\},$$

hence the integral $\int_0^\infty L(\beta_s)\mathrm{d}s$ converges and we have furthermore the $\tilde{O}$ bound of the proposition. □

B.2.3 Proof of the convergence of the iterates: Proposition 15

In this subsection we prove the convergence of the iterates which corresponds to Proposition 15 of the main text. For the sake of completeness, we recall this fact in the following lemma.

Lemma 15. *On $\mathcal{A}$, let $\gamma \leq [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^\star_{ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^\star_{ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\}]^{-1}$. The iterates $(\beta_t)_{t\geq 0}$ converge to an interpolator β^α_∞, i.e. such that $L(\beta^\alpha_\infty) = 0$.*

Proof. Consider the following Bregman divergence style function for any interpolator $\beta^\star$:

$$W_t = \phi_{\alpha_\infty}(\beta^\star) - \phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta^\star\rangle,$$

where $\alpha_\infty = \alpha \exp\Big(-2\gamma\big(\int_0^\infty L(\beta_s)\mathrm{d}s\big)\,\mathrm{diag}(\bar{X}^\top\bar{X})\Big) > 0$ is well defined on $\mathcal{A}$ as a result of Proposition 28. The exact same computations as in Lemma 11 lead to:

$$W_t = W_0 - 2\int_0^t L(\beta_s)\mathrm{d}s + \langle \mathrm{diag}(\bar{X}^\top\bar{X}), \gamma\int_0^t L(\beta_s)\xi_s\mathrm{d}s\rangle + \sqrt{\gamma}\int_0^t \sqrt{L(\beta_s)}\langle X^\top \mathrm{d}B_s, \beta_s - \beta^\star\rangle.$$

Note that:

- $\int_0^t L(\beta_s)\mathrm{d}s$ converges from Proposition 28.
- $\int_0^t \|L(\beta_s)\xi_s\|_1\mathrm{d}s \le \max_{s\ge 0}(\|\xi_s\|_1)\int_0^t L(\beta_s)\mathrm{d}s < \infty$ from Proposition 28. Hence $\int_0^t L(\beta_s)\xi_s\mathrm{d}s$ is absolutely convergent, hence converges.
- $\int_0^t \sqrt{L(\beta_s)}\langle X^\top \mathrm{d}B_s, \beta_s - \beta^\star\rangle$ has a quadratic variation equal to $4\int_0^t L(\beta_s)^2\mathrm{d}s$ and $4\int_0^t L(\beta_s)^2\mathrm{d}s \le 2\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_2^2 + \|\beta^\star\|_1^2)\mathrm{d}s$. This implies that the quadratic variation converges. Hence we obtain the convergence [2] of the Brownian integral $\int_0^t \sqrt{L(\beta_s)}\langle X^\top \mathrm{d}B_s, \beta_s - \beta^\star\rangle$.

Overall we get that W_t converges for all choice of interpolator $\beta^\star$. Now note that since $\int_0^\infty L(\beta_s)\mathrm{d}s < +\infty$ we can extract a subsequence such that $L(\beta_{\phi(t)}) \underset{t\to\infty}{\to} 0$. Since $(\beta_t)_t$ is bounded (Lemmas 12 and 13), so is $(\beta_{\phi(t)})_t$ and we can extract a new subsequence which converges. Let β_∞^α denote the limit: $\beta_{\phi_2(t)} \underset{t\to\infty}{\longrightarrow} \beta_\infty^\alpha$ where ϕ_2 is the double extraction. Since $L(\beta_{\phi(t)}) \underset{t\to\infty}{\to} 0$ so does $L(\beta_{\phi_2(t)}) \underset{t\to\infty}{\to} 0$. By continuity of the loss we have that β_∞^α is an interpolator. Now notice that since the Lyapunov W_t with the choice $\beta^\star = \beta_\infty$ converges and that $W_{\phi_2(t)} \underset{t\to\infty}{\to} 0$ we get that $W_t \underset{t\to\infty}{\to} 0$.

Furthermore:

$$\begin{aligned} W_t &= \phi_{\alpha_\infty}(\beta_\infty^\alpha) - \phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta_\infty^\alpha\rangle \\ &\ge \phi_{\alpha_t}(\beta_\infty^\alpha) - \phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta_\infty^\alpha\rangle \\ &= D_{\phi_{\alpha_t}}(\beta_\infty^\alpha, \beta_t) \\ &\ge 0 \end{aligned}$$

where the first inequality is because $\alpha \mapsto \phi_\alpha(\beta)$ is decreasing and $\alpha_t \ge \alpha_\infty$. Therefore $D_{\phi_{\alpha_t}}(\beta_\infty^\alpha, \beta_t) \to 0$. Finally, since:

$$\begin{aligned} \nabla^2\phi_{\alpha_t}(\beta_t) &= \mathrm{diag}\Big(\frac{1}{\sqrt{\beta_i(t)^2 + 4\alpha_t^4(i)}}\Big)_i \\ &\ge \mathrm{diag}\Big(\frac{1}{\sqrt{\max_s\{\beta_i(s)^2\} + 4\alpha^4}}\Big)_i \\ &\ge \mathrm{diag}\Big(\frac{1}{\sqrt{\max_s\{\|\beta(s)\|_1^2\} + 4\alpha^4}}\Big)_i \\ &\ge \mu I_d, \end{aligned}$$

for some μ since the iterates are bounded. Therefore for all $t \ge 0$, ϕ_{α_t} is μ-strongly convex on some convex set in which the iterates β_s stay in. Which means that: $D_{\phi_{\alpha_t}}(\beta_\infty^\alpha, \beta_t) \ge \frac{\mu}{2}\|\beta_t - \beta_\infty^\alpha\|_2^2$. Hence $\beta_t \to \beta_\infty^\alpha$. □

Lemma 15 along with the fact that the event $\mathcal{A}$ has probability at least $1 - \frac{p}{2}$ (see Lemma 18 and paragraph around B.3) concludes the proof of Proposition 15.

[2] See for example Theorem 5 of https://almostsuremath.com/2010/04/01/continuous-local-martingales/ for a proof of this fact. For the moment we did not find a precise reference of this standard fact in the classical Revuz and Yor [2013].

B.2.4 Proof of Theorem 1

We are now equipped to prove the main result of the chapter. For clarity we recall the statement of the theorem here.

Theorem 8. *For $p \leq \frac{1}{2}$ and $w_{0,\pm} = \alpha \in (\mathbb{R}_+^*)^d$, let $(w_t)_{t\geq 0}$ follow the stochastic gradient flow (7.3) with step size $\gamma \leq O\big(\big[\ln(\frac{4}{p})\lambda_{\max}\max\{\|\beta^\star_{\ell_1}\|_1 \ln\big(\frac{\|\beta^\star_{\ell_1}\|_1}{\min_i \alpha_i^2}\big), \|\alpha\|_2^2\}\big]^{-1}\big)$ where $\beta^\star_{\ell_1} = \underset{\beta\in\mathbb{R}^d \, s.t. \, X\beta=y}{\arg\min} \|\beta\|_1$ and $\lambda_{\max}$ is the largest eigenvalue of $X^\top X/n$. Then, with probability at least $1-p$:*

- *$(\beta_t)_{t\geq 0}$ converges towards a zero-training error solution β^α_∞*
- *the solution β^α_∞ satisfies*

$$\beta^\alpha_\infty = \underset{\beta\in\mathbb{R}^d \ s.t. \ X\beta=y}{\arg\min} \phi_{\alpha_\infty}(\beta) \quad \text{where} \quad \alpha_\infty = \alpha \odot \exp\left(-2\gamma \operatorname{diag}\left(\frac{X^\top X}{n}\right)\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s\right). \tag{C.11}$$

Proof. Recall first that on $\mathcal{A}$, Lemma 15 implies that the iterates converge towards a zero-training error we denote by β^α_∞. From Proposition 14 we also have that:

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma L(\beta_t)}\bar{X}^\top \mathrm{dB}_t, \tag{B.6}$$

where $\alpha_t = \alpha \odot \exp\big(-2\gamma \operatorname{diag}\big(\bar{X}^\top\bar{X}\big)\int_0^t L(\beta_s)\,\mathrm{d}s\big)$ and ϕ_α is the hyperbolic entropy defined in (7.4). Since the quantity $\int_0^\infty L(\beta_s)\,\mathrm{d}s$ is well defined on $\mathcal{A}$ (Proposition 28), we can integrate (B.6) from $t=0$ to $t=\infty$ which leads to $\nabla\phi_{\alpha_\infty}(\beta^\alpha_\infty) \in \operatorname{span}(X)$. This condition, along with the fact that $X\beta^\alpha_\infty = y$, exactly corresponds to the KKT conditions of the implicit minimisation problem (7.5). From Lemma 18, the fact that the event $\mathcal{A}$ has probability at least $1-p$ concludes the proof. □

B.2.5 Lower bound on $\int L(\beta_s)\mathrm{d}s$ and proof of Proposition 16

Similarly to what has been done in subsection B.2.2, in order to lower bound the loss integral, we need a (different) control on the deviation of the local martingale S_t. We choose $\hat{a} := W_0^\alpha/2$ and $\hat{b} := \frac{1}{2}\ln(4/p)\hat{a}^{-1}$ so that once again $\hat{a}\hat{b} = \frac{1}{2}\ln(4/p)$. We refer to Lemma 16 for the definition of W_0^α. Now that these parameters are fixed, consider the new event:

$$\mathcal{B} = \{\forall t \geq 0, |S_t| \leq \hat{a} + 2\hat{b}\gamma\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^\star_{ell_1}\|_1^2)\mathrm{d}s\}$$

In this entire subsection we shall put ourselves on the intersection $\mathcal{A}\cap\mathcal{B}$ which occurs with probability $\mathbb{P}(\mathcal{A}\cap\mathcal{B}) \geq 1 - (\mathbb{P}(\mathcal{A}^C) + \mathbb{P}(\mathcal{B}^C)) \geq 1-p$. Furthermore since the goal of this section is to obtain an idea of the dependency on α of the integral of the loss as α goes to 0, we shall consider the initialisations $\alpha = \alpha\mathbf{1}$, therefore for now on α is a positive scalar. Note that with this convention $\|\alpha\|_2^2 = \alpha^2 d$.

Notice that the quantity $\gamma\int_0^{+\infty} L(\beta_s)\mathrm{d}s$, through α_∞, controls the magnitude of the sparse-inducing effect. In the following lemma we show that this quantity is lower bounded by a quantity which is strictly increasing with γ. **This recommends to pick the largest γ (as long as the iterates converge). This fact is also observed in practice.**

Lemma 16. *On $\mathcal{A} \cap \mathcal{B}$, let $\gamma \leq [400\lambda_{\max} \ln(\frac{4}{p}) \max\{\|\beta^\star_{ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^\star_{ell_1}\|_1}{\alpha^2}), \alpha^2 d\}]^{-1}$,*

$$\gamma \int_0^{+\infty} L(\beta_s) \mathrm{d}s \geq \frac{W_0^\alpha}{4} \frac{\gamma}{1 + \gamma \frac{M}{W_0^\alpha}},$$

where $W_0^\alpha = \min_{\beta \ s.t \ X\beta = Y} \phi_\alpha(\beta) - \phi_\alpha(0)$ and $M = \big[325\lambda_{\max} \ln(\frac{4}{p}) \max\{\|\beta^\star_{ell_1}\|_1^2 \ln^2(\frac{\sqrt{2}\|\beta^\star_{ell_1}\|_1}{\alpha^2}), \alpha^4 d^2\}\big]$.

Proof. According to Lemma 15, the flow converges to an interpolator β^α_∞. We consider the same Lyapunov as before:

$$W_t = \phi_{\alpha_\infty}(\beta^\alpha_\infty) - \phi_{\alpha_t}(\beta_t) + \langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta^\alpha_\infty \rangle,$$

which is such that, following the same computations as in Lemma 11:

$$\begin{aligned} 2\int_0^t L(\beta_s)\mathrm{d}s &= W_0 - W_t + \gamma \langle \mathrm{diag}(\bar{X}^\top \bar{X}), \int_0^t L(\beta_s)\xi_s \mathrm{d}s \rangle + S_t \\ &\geq W_0 - W_t + S_t, \end{aligned}$$

where $S_t = \int_0^t \sqrt{\gamma L(\beta_s)} \langle X^\top \mathrm{d}B_s, \beta_s - \beta^*_{\ell_1} \rangle$.

Now since we put ourselves on $\mathcal{B}$:

$$\begin{aligned} 2\int_0^\infty L(\beta_s)\mathrm{d}s &\geq W_0 - \hat{a} - 2\hat{b}\gamma\lambda_{\max} \int_0^{+\infty} L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\mathrm{d}s \\ &\geq W_0 - \hat{a} - 2\hat{b}\gamma\lambda_{\max}(18^2 + 1) \max\left(\|\beta^*_{\ell_1}\|_1^2 \ln^2\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\right), \alpha^4 d^2\right) \int_0^{+\infty} L(\beta_s)\mathrm{d}s \\ &\geq W_0 - \hat{a} - 2\gamma\hat{b}M \ln(4/p)^{-1} \int_0^{+\infty} L(\beta_s)\mathrm{d}s, \end{aligned}$$

where the second inequality comes from Lemma 12 (which is still valid since we are on the event $\mathcal{A}$) and $M = \big[325 \ln(4/p)\lambda_{\max} \max(\|\beta^*_{\ell_1}\|_1^2 \ln^2(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}), \alpha^4 d^2)\big]$. Hence, we can lowerbound the integral as

$$\int_0^{+\infty} L(\beta_s)\mathrm{d}s \geq \frac{W_0 - \hat{a}}{2 + 2\gamma\hat{b}M \ln(\frac{4}{p})^{-1}}.$$

Importantly $W_0 = \phi_{\alpha_\infty}(\beta_\infty) - \phi_\alpha(0)$ depends on β_∞ and is therefore stochastic. However, since for all $\beta \in \mathbb{R}^d$, $\alpha \mapsto \phi(\beta, \alpha^2)$ is decreasing and $\alpha_\infty \leq \alpha$, we obtain:

$$\begin{aligned} W_0 &= \phi_{\alpha_\infty}(\beta_\infty) - \phi_\alpha(0) \\ &\geq \phi_\alpha(\beta_\infty) - \phi_\alpha(0) \\ &\geq \phi_\alpha(\beta^\star_\alpha) - \phi_\alpha(0) := W_0^\alpha, \end{aligned}$$

where $\beta^\star_\alpha = \operatorname*{argmin}_{\beta \ \mathrm{s.t} \ X\beta = Y} \phi(\beta, \alpha^2)$. Therefore, we control the integral of the loss as

$$\int_0^{+\infty} L(\beta_s)\mathrm{d}s \geq \frac{W_0^\alpha - \hat{a}}{2 + 2\gamma\hat{b}M \ln(\frac{4}{p})^{-1}}$$

We now plug in the values $\hat{a} = \frac{W_0^\alpha}{2}$ and $\hat{b} = \frac{1}{W_0^\alpha}\ln(\frac{4}{p})$:

$$\gamma \int_0^{+\infty} L(\beta_s)\mathrm{d}s \geq \frac{W_0^\alpha}{4}\frac{\gamma}{1+\gamma\frac{M}{W_0^\alpha}}.$$

□

To complete our understanding of the dependency of the integral of the loss in terms of α and $\beta^\star_{\ell_1}$ we need to know the dependency of W_0^α in α. The following lemma does so. We consider the limit $\alpha \to 0$ which corresponds to the rich regime we are interested in.

Lemma 17. *On $\mathcal{A}\cap\mathcal{B}$, let $\gamma \leq [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^\star_{ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^\star_{ell_1}\|_1}{\alpha^2}), \alpha^2 d\}]^{-1}$, then for α small enough:*

$$\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s \geqslant \frac{1}{8}\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\Big).$$

Proof. Applying Lemma 20, for all $\beta \in \mathbb{R}^d$, $\phi_\alpha(\beta) - \phi_\alpha(0) \geq \frac{1}{4}\sum_i \max\left\{0, |\beta_i| \ln\frac{|\beta_i|}{2\alpha^2}\right\}$. Therefore,

$$W_0^\alpha \geq \frac{1}{4}\sum_i |\beta^\star_{\alpha,i}| \ln\frac{|\beta^\star_{\alpha,i}|}{2\alpha^2}.$$

Note that $\beta^\star_\alpha = \underset{\beta \text{ s.t } X\beta=Y}{\operatorname{argmin}} \phi_\alpha(\beta)$ and $\beta^\star_{\ell_1} = \underset{\beta \text{ s.t } X\beta=Y}{\operatorname{argmin}} \|\beta\|_1$. From Theorem 2 of Woodworth et al. [2020b]: $\|\beta^\star_\alpha\|_1 \underset{\alpha\to 0}{\longrightarrow} \|\beta^\star_{\ell_1}\|_1$ which leads to:

$$\sum_i |\beta^\star_{\alpha,i}| \ln\frac{|\beta^\star_{\alpha,i}|}{2\alpha^2} \underset{\alpha\to 0}{\sim} \|\beta^\star_{\ell_1}\|_1 \ln\frac{\|\beta^\star_{\ell_1}\|_1}{\alpha^2}.$$

and $W_0^\alpha \underset{\alpha\to 0}{\geqslant} \frac{1}{4}\|\beta^*_{\ell_1}\|_1 \ln\big(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\big)$. Finally, for α small enough, from the upperbound on γ, the value of M and the lower bound on W_0^α:

$$\gamma\frac{M}{W_0^\alpha} \underset{\alpha\to 0}{\leq} 1,$$

which along with Lemma 16 concludes the proof. □

Therefore through this lemma we see that by picking the biggest step-size which ensures convergence, we have a dependency of the integral of the loss as $\ln\frac{1}{\alpha}$.

Now we are equipped to prove Proposition 16. We recall it here to be self-contained.

Proposition 16. *Under the same setting as in Proposition 15 with initialisation $w_{0,\pm} = \alpha\mathbf{1}$, we have with probability at least $1-p$:*

$$\Omega\Big(\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\Big)\Big) \underset{\alpha\to 0}{\leqslant} \int_0^{+\infty} L(\beta_s)\,\mathrm{d}s \leqslant O\Big(\max\big\{\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\Big), \alpha^2 d\big\}\Big).$$

Proof. Let us place ourselves on the event $\mathcal{A}\cap\mathcal{B}$. Let us recall that $\mathbb{P}(\mathcal{A}\cap\mathcal{B}) \geq 1 - (\mathbb{P}(\mathcal{A}^C) + \mathbb{P}(\mathcal{B}^C)) \geq 1-p$, where the last inequality results from the definitions of $\mathcal{A}$ and $\mathcal{B}$ and Lemma 18. As this event is included in $\mathcal{A}$, the right inequality of the proof corresponds exactly to the Proposition 28 of Appendix B.2.2. The proof of left inequality of the proposition comes from Lemma 17. □

In the final proposition of this subsection, we give the scale of α_∞ we obtain thanks to our analysis. Indeed though we know that in all case $\alpha_\infty < \alpha$, we would like to quantitatively know **how much** smaller the effective initialisation is in order to have an idea of the gain of SGD over GD (in terms of implicit bias).

Proposition 29. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow (7.3), initialised at $w_{0,\pm} = \alpha\mathbf{1} \in (\mathbb{R}_+^*)^d$. Let $p \leq \frac{1}{2}$ and γ matching the upperbound in Theorem 1, i.e. $\gamma = [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^\star_{ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^\star_{ell_1}\|_1}{\alpha^2 d}), \alpha^2 d\}]^{-1}$, then with probability at least $1-p$ and for α small enough:*

$$\frac{\alpha_\infty}{\alpha} \leq \exp\Big(-\frac{1}{1600\ln(\frac{4}{p})}\frac{\mathrm{diag}(\frac{X^\top X}{n})}{\lambda_{\max}}\Big).$$

Proof. The fact that $\alpha_\infty = \alpha\exp\left(-2\gamma\,\mathrm{diag}\left(\frac{X^\top X}{n}\right)\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s\right)$ along with the lower bound from Lemma 17 and the value of γ gives the result. □

This result tends to show that the overall gain of SGD over GD is only by a constant factor $\exp(-\frac{1}{1600\ln(\frac{4}{p})}\frac{\mathrm{diag}(\frac{X^\top X}{n})}{\lambda_{\max}}) < 1$. We believe that our analysis is not tight and that the gain is in fact more consequent, this is explained in the following subsection.

B.2.6 Scale of α_∞ when assuming that the iterates are bounded independently of α.

In this subsection we explain why we believe that our analysis lacks of tightness. In Lemma 12 there is a dependency in $\ln(\frac{1}{\alpha})$ in the upperbound of the ℓ_1 norm of the iterates. We believe that this dependency is an artifact of our analysis and that the true bound is independent of α, this is also what is observed in practice. This is the reason why we formulate the following assumption:

Boundedness assumption. *On $\mathcal{A}$, $\|\beta_t\|_1 \leq \|\xi_t\|_1 \leq \max\{\|\beta^\star_{\ell_1}\|_1, \alpha^2 d\}$ for all $t \geq 0$.*

Under this assumption, we obtain convergence of the iterates towards an interpolating solution under a weaker constraint on γ (bigger step-sizes can be used while still ensuring convergence) as well as a much better upperbound on the scale of α_∞. The aim of the following result is to give the relevant scale of how small is α_∞ w.r.t. α. Hence, for the sake of clarity, we will assume that $\mathrm{diag}(X^\top X/n) \sim \lambda_{\max}\mathbf{1}$ (which is true for sub-gaussian inputs with high probability). We also fix $p = 0.01$ and drop all the numerical constants under some universal constant $\zeta > 0$.

Proposition 30. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow (7.3), initialised at $w_{0,\pm} = \alpha\mathbf{1} \in (\mathbb{R}_+^*)^d$. Assume boundedness of the iterates and $\gamma = \Theta\big(\max\{\|\beta^\star_{\ell_1}\|_1, \alpha^2 d\}^{-1}\big)$, then with probability at least 0.99, the iterates $(\beta_t)_{t\geq 0}$ converge towards an interpolating solution $\beta^\alpha_\infty = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta = y}{\arg\min}\ \phi_{\alpha_\infty}(\beta)$. Furthermore, for α small enough, there exists $\zeta > 0$ such that:*

$$\frac{\alpha_\infty}{\alpha} \leq \left(\frac{\alpha^2}{\|\beta^\star_{\ell_1}\|_1}\right)^\zeta.$$

Proof. As said earlier, we fix $p = 0.01$. Then, by following the proof of Lemma 13, and using the boundedness assumption instead of Lemma 12, one obtains that for $\gamma \leq O\big(\max\{\|\beta^\star_{\ell_1}\|_1, \alpha^2 d\}^{-1}\big)$

(as mentioned the precise numerical constants are dropped for simplicity) then $U_t \geq \frac{1}{2}$ for all $t \geq 0$. The results of Lemma 14, Proposition 28, Lemma 15 and therefore Theorem 1 then still hold with probability 0.99 but with the weaker condition that $\gamma \leq O\big(\max\{\|\beta^\star_{\ell_1}\|_1, \alpha^2 d\}^{-1}\big)$.

For the upperbound on α_∞, we follow the exact same steps as in Appendix B.2.5. Indeed Lemma 16 now gives, for $\gamma \leq O\big((\lambda_{\max} \max\{\|\beta^\star_{\ell_1}\|_1, \alpha^2 d\})^{-1}\big)$:

$$\gamma \int_0^{+\infty} L(\beta_s) \mathrm{d}s \geq \frac{W_0^\alpha}{4} \frac{\gamma}{1 + \gamma \frac{M}{W_0^\alpha}},$$

where $M = \Theta\big(\lambda_{\max} \max\{\|\beta^\star_{\ell_1}\|_1^2, \alpha^4 d^2\}\big)$. Plugging in the maximum value of γ, i.e. $\gamma = \Theta\big((\lambda_{\max} \max\{\|\beta^\star_{\ell_1}\|_1, \alpha^2 d\})^{-1}\big)$: we have that $\gamma \frac{M}{W_0^\alpha} \underset{\alpha \to 0}{\longrightarrow} 0$ and for α small enough $\gamma W_0^\alpha \geq \Omega\left(\lambda_{\max}^{-1} \ln\left(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\right)\right)$. Therefore for α small enough:

$$\gamma \int_0^{+\infty} L(\beta_s) \mathrm{d}s \geq \Omega\left(\lambda_{\max}^{-1} \ln\left(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\right)\right)$$

Plugging this inequality into the definition of α_∞ and assuming that $\mathrm{diag}(X^\top X/n) \sim \lambda_{\max} \mathbf{1}$ leads to:

$$\alpha_\infty = \alpha \exp\left(-2\,\mathrm{diag}\left(\frac{X^\top X}{n}\right) \gamma \int_0^{+\infty} L(\beta_s)\,\mathrm{d}s\right) \leq \alpha \left(\frac{\alpha^2}{\|\beta^\star_{\ell_1}\|_1}\right)^{\Omega(1)}.$$

This concludes the proof of the Proposition. □

This upperbound is significantly better than that of Proposition 29: the smaller the initialisation scale α and the greater the benefit of SGD over GD in terms of implicit bias. More precisely, Proposition 30 shows that the benefit scales as a power law with respect to the initialization α.

B.3 Experiments

In the following section we consider the same experimental setup as in Section 7.6.1, which we recall here for clarity. We consider $n = 40$, $d = 100$ and randomly generate a sparse model $\beta^\star_{\ell_0}$ such that $\|\beta^\star_{\ell_0}\|_0 = 5$. We generate the features as $x_i \sim \mathcal{N}(0, I)$ and the labels as $y_i = x_i^\top \beta^\star_{\ell_0}$. We use the same step size for GD and SGD and choose it to be the biggest as possible while still ensuring convergence. Note that since the true population covariance $\mathbb{E}[xx^\top]$ is equal to identity, the quantity $\|\beta_t - \beta^\star_{\ell_0}\|_2^2$ corresponds to the validation loss.

B.3.1 Doping the implicit bias using label noise: experiments

We consider the label noise setting discussed in Section 7.6.4: for a sequence $(\delta_t)_{t\in\mathbb{N}} \in \mathbb{R}_+$, assume that we artificially inject some label noise Δ_t at time t, say for example $\Delta_t \sim \text{unif}\{2\delta_t, -2\delta_t\}$ and independently from i_t (other type of label noise can of course be considered, but we consider here this one for simplicity). This injected label noise perturbs the SGD recursion as follows:

$$w_{t+1,\pm} = w_{t,\pm} \mp \gamma \left(\langle \beta_w - \beta^\star, x_{i_t}\rangle + \Delta_t\right) \; x_{i_t} \odot w_{t,+} \;, \qquad \text{where} \;\; i_t \sim \text{unif}(1, n). \tag{B.7}$$

Using the same notations and following the same derivations as in Appendix B.1, we can rewrite the recursion as:

$$w_{t+1,\pm} = w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \pm \gamma \,\text{diag}(w_{t,\pm}) X^\top [\xi_{i_t}(\beta_t) + \Delta_t \mathbf{e}_{i_t}].$$

Since Δ_t is zero-mean and independent of i_t we get:

$$\begin{aligned}\text{Cov}_{i_t}[\xi_{i_t}(\beta) + \Delta_t \mathbf{e}_{i_t}] &= \mathbb{E}_{i_t}[\xi_{i_t}(\beta)^{\otimes 2}] + \mathbb{E}[\Delta_t^2 \mathbf{e}_{i_t}^{\otimes 2}] \\ &= \mathbb{E}_{i_t}[\xi_{i_t}(\beta)^{\otimes 2}] + \frac{4\delta_t^2}{n} I_n.\end{aligned}$$

Now following the same reasoning as in Appendix B.1, it is natural to consider the following SDE:

$$\mathrm{d}w_{t,\pm} = -\nabla_{w_\pm} L(w_t)\mathrm{d}t \pm 2\sqrt{\gamma n^{-1}(L(w_t) + \delta_t^2)}\; w_{t,+} \odot [X^\top \mathrm{d}B_t].$$

Let $\tilde{L}(\beta_t) = L(\beta_t) + \delta_t^2$ be the "slowed down" loss. Following the same computations as for Lemma 10 we obtain that:

$$\beta_t = 2\tilde{\alpha}_t^2 \odot \sinh(2\bar{X}^\top \tilde{\eta}_t),$$

where $\tilde{\eta}_t = -\int_0^t \bar{X}(\beta_s - \beta^\star)\,\mathrm{d}s + 2\sqrt{\gamma}\int_0^t \sqrt{\tilde{L}(\beta_s)}\mathrm{d}B_s \in \mathbb{R}^n$ and $\alpha_t = \alpha \odot \exp\left(-2\gamma\,\text{diag}(\bar{X}^\top \bar{X})\int_0^t \tilde{L}(\beta_s)\mathrm{d}s\right)$. And following the proof of Proposition 14:

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma n^{-1}\tilde{L}(\beta_t)}X^\top \mathrm{dB}_t. \tag{B.8}$$

Assuming that $(\delta_t)_{t\geq 0} \in (\mathbb{R}_+)^{\mathbb{R}}$ and γ are such that the iterates converge (here we do not show under which conditions we have convergence and leave this as future work), the corresponding implicit regularisation minimisation problem is preserved but with an effective initialisation: $\tilde{\alpha}_\infty = \alpha \odot \exp\left(-2\gamma\,\text{diag}(\frac{X^\top X}{n})\int_0^{+\infty}\tilde{L}(\beta_s)\,\mathrm{d}s\right)$ which takes into account the slowed down loss $\tilde{L}(\beta_t) = L(\beta_t) + \delta_t^2$. Since it is reasonable to consider that $\tilde{\alpha}_\infty < \alpha_\infty$, the label noise therefore helps to recover a solution which has better sparsity properties.

We experimentally validate the advantage of adding label noise by choosing the sequence $\delta_t = 1$ if $t \leq 10^3$ and $\delta_t = 0$ if $t > 10^3$. The results are illustrated Figure B.1. Note that the training loss is heavily slowed down, however the recovered solution at iteration $t = 10^6$ is much better than that of SGD, and it has not even converged yet. However, it must be kept in mind that adding too much label noise can significantly slow down the convergence of the validation loss or even prevent the iterates from converging.

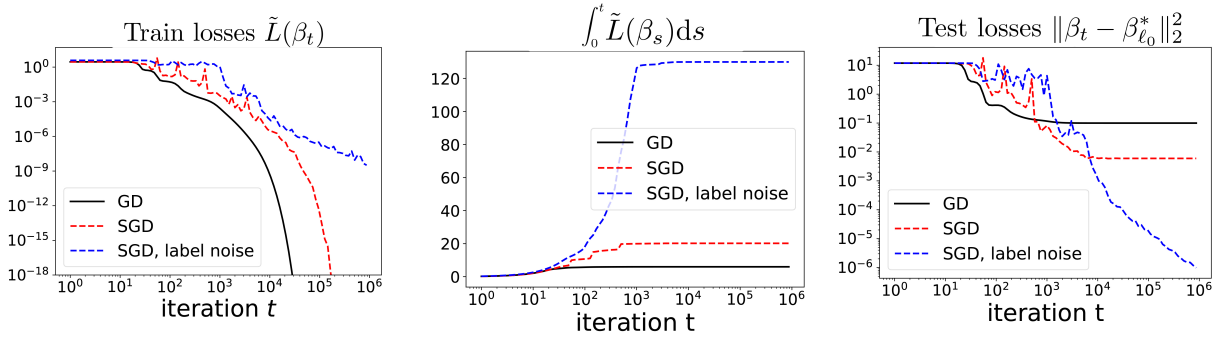


Figure B.1: Sparse regression (see Section 7.6.1 for the detailed experimental setting), illustration of the benefits of using label noise. All experiments are initialised at $\alpha = 0.01$. *Left*: The use of label noise slows down the convergence of the effective training loss $\tilde{L}$. *Middle and right*: the value of the integral of the slowed down loss $\tilde{L}$ is much higher for the recursion with label noise, leading to a solution which generalises much better.

B.4 Extensions

We introduce two extensions of our results: subsection B.4.1 extends our results for a very general stochastic gradient flow model and subsection B.4.2 discuss them in the depth $p \geq 3$ case.

B.4.1 Towards a more general SDE modelling

Recall from the SDE modelling of Appendix B.1 that $\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] = \frac{4}{n}\operatorname{diag}(L_i(\beta))_{1\leq i\leq n} + O(\frac{1}{n^2})$. If we assume n large enough we can neglect the second order term of order $1/n^2$:

$$\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] \cong \frac{4}{n}\operatorname{diag}(L_i(\beta))_{1\leq i\leq n}.$$

Assume we do not consider that $L_i(\beta) \sim L(\beta)$, then the overall SGD noise structure is captured by

$$\begin{aligned}\Sigma_{\mathrm{SGD}}(w_\pm) &:= \gamma^2 \operatorname{diag}(w_\pm) X^\top \mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] X \operatorname{diag}(w_\pm) \\ &\cong \frac{4}{n}\gamma^2 [\operatorname{diag}(w_\pm) X^\top \operatorname{diag}(\sqrt{L_i(\beta)})]^{\otimes 2}.\end{aligned}$$

This leads us in considering the following SDE:

$$\begin{aligned}\mathrm{d}w_{t,+} &= -\nabla_{w_+} L(w_t)\,\mathrm{d}t + 2\sqrt{\gamma}\; w_{t,+} \odot [\bar{X}^\top \operatorname{diag}(\sqrt{L_i(\beta)})\mathrm{d}B_t] \\ \mathrm{d}w_{t,-} &= -\nabla_{w_-} L(w_t)\,\mathrm{d}t - 2\sqrt{\gamma}\; w_{t,-} \odot [\bar{X}^\top \operatorname{diag}(\sqrt{L_i(\beta)})\mathrm{d}B_t].\end{aligned} \tag{B.9}$$

As previously, this SDE admits an implicit integral formulation (multiplication must be understood component-wise):

$$\begin{aligned}w_{t,\pm} &= w_{t=0,\pm} \odot \exp(\pm\bar{X}^\top\big[- \int_0^t r(w_s)\,\mathrm{d}s + 2\sqrt{\gamma}\int_0^t \operatorname{diag}(\sqrt{L_i(w_s)})\,\mathrm{d}B_s\big]) \\ &\quad \odot \exp(-2\gamma \operatorname{diag}(\bar{X}^\top \int_0^t \operatorname{diag}(L_i(w_s))\,\mathrm{d}s\bar{X})) \\ &= \alpha_t \odot \exp(\pm\bar{X}^\top\eta_t),\end{aligned}$$

where $\eta_t = -\int_0^t \bar{X}(\beta_s - \beta^*)\,\mathrm{d}s + 2\sqrt{\gamma}\int_0^t \operatorname{diag}(\sqrt{L_i(w_s)})\,\mathrm{d}B_s \in \mathbb{R}^n$ and $\alpha_t = \alpha \odot \exp(-2\gamma \operatorname{diag}(\bar{X}^\top \int_0^t \operatorname{diag}(L_i(w_s))$ Since $\beta = w_+^2 - w_-^2$, we get:

$$\begin{aligned}\beta_t &= \alpha_t^2 \odot \big(\exp(+2\bar{X}^\top\eta_t) - \exp(-2\bar{X}^\top\eta_t)\big) \\ &= 2\alpha_t^2 \odot \sinh(+2\bar{X}^\top\eta_t).\end{aligned}$$

And we obtain the following mirror-type descent flow:

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma}\bar{X}^\top \operatorname{diag}(\sqrt{L_i(\beta_t)})\,\mathrm{d}B_t.$$

Assuming convergence of the iterates and of α_t (we do not show the convergence, though we think the proof could straightforwardly be adapted following Appendix B.2), the corresponding minimisation problem is:

$$\beta_\infty^\alpha = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta = y}{\arg\min} \phi_{\alpha_\infty}(\beta) \quad \text{where} \quad \alpha_\infty = \alpha \odot \exp(-2\gamma \operatorname{diag}(\bar{X}^\top \int_0^\infty \operatorname{diag}(L_i(\beta_s))\,\mathrm{d}s\; \bar{X})).$$

Note that the main result of the chapter is very similar, the difference relies in:

- the k^{th} coordinate of $\operatorname{diag}(\bar{X}^\top \operatorname{diag}(L_i(\beta_s))\; \bar{X})$ is $\mathbb{E}_{i_t}[L_{i_t}(\beta_s)(x_{i_t}^{(k)})^2]$
- the k^{th} coordinate of $L(\beta_s)\operatorname{diag}(\bar{X}^\top\bar{X})$ is $\mathbb{E}_{i_t}[L_{i_t}(\beta_s)]\mathbb{E}_{i_t}[(x_{i_t}^{(k)})^2]$

B.4.2 Higher order models: the cases of depth $p > 2$

Until now, we have focused on a 2-homogeneous parametrisation of the estimator. A legitimate question is how the implicit bias changes as we go to a higher degree of homogeneity. In terms of networks architecture, this corresponds to increasing the depth of the neural networks. Let us fix $p \geqslant 3$ with the new parametrisation $\beta_w = w_+^p - w_-^p$, the loss of our new model writes: $L(w) = \frac{1}{4n}\sum_{i=1}^n \langle w_+^p - w_-^p - \beta^*, x_i\rangle^2$. As previously, we want to consider the stochastic differential equation related to stochastic gradient descent on the above loss. With the same modelling as in Section 7.3.2, stochastic gradient flow writes:

$$\mathrm{d}w_{t,\pm} = -\nabla_{w_\pm} L(w_t)\mathrm{d}t \pm 2\sqrt{\gamma n^{-1} L(\beta_t)}\,\mathrm{diag}(w_{t,\pm}^{p-1})X^\top \mathrm{d}B_t, \tag{B.10}$$

where B_t is a standard Brownian motion in $\mathbb{R}^n$. We would like to put emphasis that, unlike the 2-depth model, we do not provide a dynamical analysis enabling convergence proof and control of interesting quantities. Here, the aim is to show how our framework naturally extends to general depth and how the convergence speed of the loss still seems to controls the effect of the stochastic flow biasing. Contrary to the 2-depth case, the potential cannot be defined in close form, but we still have the following explicit expression, $\phi_{\alpha,\pm}^p(\beta) = \sum_{i=1}^d \psi_{\alpha,\pm}^p(\beta_i)$, where $\psi_{\alpha,\pm}^p = \int [h_{\alpha,\pm}^p]^{-1}$ is a primitive of the unique inverse of $h_{\alpha,\pm}^p(z) := (\alpha_+^{2-p} - z)^{-\frac{p}{p-2}} - (\alpha_-^{2-p} + z)^{-\frac{p}{p-2}}$ in $(-\alpha_-^{2-p}, \alpha_+^{2-p})$. In the following theorem we characterize the implicit bias of the stochastic gradient flow when applied with higher order models.

Theorem. *Initialise the stochastic gradient flow with $w_0 = \alpha\mathbf{1} \in \mathbb{R}^{2d}$. If we assume that the flow $(\beta_t)_{t\geq 0}$ converges almost surely towards a zero-training error solution $\beta_\infty^{\alpha,p}$, and that the quantities $\int_0^\infty L(\beta_s) w_{s,\pm}^{p-2}\mathrm{d}s$ and $\int_0^\infty L(\beta_s)\mathrm{d}s$ exist a.s., then the limit satisfies*

$$\beta_{\infty,p} = \underset{\beta \ s.t \ X\beta = y}{\arg\min}\ \phi_{\alpha_\infty,\pm}^p(\beta),$$

with $\alpha_{\infty,\pm} = \alpha\big(1 + 2\gamma(p-2)(p-1)\alpha^{p-2}\,\mathrm{diag}(\frac{X^\top X}{n}) \odot \int_0^\infty L(\beta_s) w_{s,\pm}^{p-2}\mathrm{d}s\big)^{-\frac{1}{p-2}}$.

First let us stress that without a close form expression of ϕ_α^d and proper control of $\int_0^\infty L(\beta_s) w_{s,\pm}^{p-2}\mathrm{d}s$ with respect to p or α, it is difficult to conclude directly on the magnitude of the stochastic bias. Yet, the main aspect we can comment on is that, as in the depth-2 case, $\alpha_{\infty,\pm} \leqslant \alpha$ almost surely[3] and that the convergence speed of the loss controls the biasing effect. As in Woodworth et al. [2020b], it can be shown empirically that $\phi_{\alpha,\pm}^p$ interpolate between the ℓ_1 and the ℓ_2 norm as $\alpha_\pm \to 0$ and $\alpha_\pm \to +\infty$ respectively and that the transition is faster than for the depth-2 case.

We directly prove this theorem here.

Proof. We apply the Itô formula on $w_{t,+}^{2-p}$ and $w_{t,-}^{2-p}$ to get the following:

$$\begin{aligned}
\mathrm{d}[w_{t,+}^{2-p}] &= (2-p)w_{t,+}^{1-p} \odot \mathrm{d}w_{t,+} + 2(2-p)(1-p)\gamma L(\beta_t) w_{t,+}^{-p} \odot w_{t,+}^{2p-2} \odot \mathrm{diag}(H) \\
&= -p(2-p)X^\top r(\beta_t)\mathrm{d}t + 2(2-p)(1-p)\gamma L(\beta_t) w_{t,+}^{p-2} \odot \mathrm{diag}(H)\mathrm{d}t + (2-p)\sqrt{\gamma L(\beta_t)}X^\top \mathrm{d}B_t \\
&= -X^\top \mathrm{d}A_t + C_t^+ \mathrm{d}t,
\end{aligned}$$

where $\mathrm{d}A_t := -p(p-2)r(\beta_t)\mathrm{d}t + 2(p-2)\sqrt{\gamma L(\beta_t)}\mathrm{d}B_t$ and $C_t^+ := 2(p-2)(p-1)\gamma L(\beta_t) w_{t,+}^{p-2} \odot \mathrm{diag}(H)$. Similarly, with explicit notations, we have that:

$$\mathrm{d}[w_{t,-}^{2-p}] = X^\top \mathrm{d}A_t + C_t^- \mathrm{d}t.$$

[3] Note that, as the weights are initialized positively, they remain positive: $w_{t,\pm} > 0$, for all $t \geqslant 0$.

Hence,

$$w_{t,+}^p = \left[\alpha^{2-p} - X^\top \int_0^t \mathrm{d}A_s + \int_0^t C_s^+ \mathrm{d}s\right]^{\frac{p}{2-p}} \quad \text{and} \quad w_{t,-}^p = \left[\alpha^{2-p} + X^\top \int_0^t \mathrm{d}A_s + \int_0^t C_s^- \mathrm{d}s\right]^{\frac{p}{2-p}}.$$

And finally,

$$\beta_t = w_{t,+}^p - w_{t,-}^p = \left[\alpha^{2-p} + \int_0^t C_s^+ \mathrm{d}s - X^\top \int_0^t \mathrm{d}A_s\right]^{\frac{p}{2-p}} - \left[\alpha^{2-p} + \int_0^t C_s^- \mathrm{d}s + X^\top \int_0^t \mathrm{d}A_s\right]^{\frac{p}{2-p}}.$$

Defining $\alpha_{\text{eff},\pm}^{2-p} = \alpha^{2-p} + \int_0^\infty C_s^\pm \mathrm{d}s$ and $\nu_\infty = \int_0^\infty \mathrm{d}A_s$, if all quantities have limits when $t \to \infty$ we have that $\beta_\infty = h_{\alpha,p,\pm}(X^\top \nu_\infty)$, where $h_{\alpha,p,\pm}(z) = (\alpha_{\text{eff},+}^{2-p} - z)^{\frac{p}{2-p}} - (\alpha_{\text{eff},-}^{2-p} + z)^{\frac{p}{2-p}}$. Inverting this function and integrating gives the theorem with the standard KKT argument [see Woodworth et al., 2020b, under Theorem 1 page 4]. □

B.5 Technical lemmas

In this section, we state and prove technical lemmas which we use to prove our main results.

Lemma 18. *For any interpolator $\beta^\star$, $S_t = \int_0^t \sqrt{\gamma L(\beta_s)}\langle \bar{X}^\top \mathrm{d}B_s, \beta_s - \beta^\star\rangle$ is a square-integrable martingale with a.s. continuous paths. And for any $a, b \geq 0$:*

$$P(\forall t \geq 0, |S_t| \leq a + 2b\gamma\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^*\|_1^2)\mathrm{d}s) \geq 1 - 2\exp(-2ab)$$
$$= 1 - p,$$

where $p = 2\exp(-2ab)$.

Proof. Since $(S_t)_{t\geq 0}$ is a is a locally square-integrable martingale with a.s. continuous paths, [Howard et al., 2020, Corollary 11] gives that

$$P(\exists t \in (0,\infty) : S_t \geq a + b\langle S\rangle_t) \leq \exp\{-2ab\}).$$

We now compute the quadratic variation $\langle S\rangle_t$. Notice that $\langle \bar{X}^\top \mathrm{d}B_t, \beta_t - \beta^*\rangle = \sum_{k=1}^n [\bar{X}(\beta_t - \beta^*)]_k \mathrm{d}B_t^k$, hence the quadratic variation of S_t equals:

$$\begin{aligned}\langle S\rangle_t &= \gamma\int_0^t L(\beta_s)\sum_{k=1}^n [\bar{X}(\beta_t - \beta^*)]_k^2 \mathrm{d}s \\ &= \gamma\int_0^t L(\beta_s)\|\bar{X}(\beta_s - \beta^*)\|^2 \mathrm{d}s \\ &= 4\gamma\int_0^t L(\beta_s)^2 \mathrm{d}s.\end{aligned}$$

Furthermore, since:

$$\begin{aligned}4\int_0^t L(\beta_s)^2\mathrm{d}s &= \int_0^t L(\beta_s)(\beta_s - \beta^*)^T\bar{X}^\top\bar{X}(\beta_s - \beta^*)\mathrm{d}s \\ &\leq \lambda_{\max}\int_0^t L(\beta_s)\|\beta_s - \beta^*\|_2^2\mathrm{d}s \\ &\leq 2\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_2^2 + \|\beta^*\|_2^2)\mathrm{d}s \\ &\leq 2\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^*\|_1^2)\mathrm{d}s,\end{aligned}$$

we obtain that:

$$\langle S\rangle_t \leq 2\gamma\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^*\|_1^2)\mathrm{d}s,$$

and:

$$\begin{aligned}P(\exists t \geq 0, |S_t| \geq a + 2b\gamma\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_2^2 + \|\beta^*\|_1^2)\mathrm{d}s) \\ \leq P(\exists t \geq 0, |S_t| \geq a + b\langle S\rangle_t) \\ \leq 2\exp(-2ab).\end{aligned}$$

□

Lemma 19. *Let $A, B > 0$ such that $\frac{A}{B} + \ln(B) \geq 2$. Assume that $x \leq A + B \ln x$, then*

$$x \leq \frac{5}{2}(A + B \ln(B)).$$

Proof. $x \leq A + B \ln x$ is equivalent to $x \leq \exp(-\frac{A}{B}) \exp(\frac{x}{B})$. Standard analysis on the Lambert W function shows that this leads to $x \leq -B\, W_{-1}(-\frac{1}{B}\exp(-\frac{A}{B}))$, where W_{-1} is the lower branch [4]. For $-\frac{1}{e} \leq z \leq 0$, the branch W_{-1} can be lower bounded as: $W_{-1}(z) \geq -\sqrt{-2(1+\ln(-z))} + \ln(-z)$ (see Theorem 1 of Chatzigeorgiou [2013]). Since $\ln(-z) = \ln(\frac{1}{B}\exp(-\frac{A}{B})) = -(\frac{A}{B} + \ln(B))$:

$$\begin{aligned} x &\leq B(\sqrt{2(-1 + \frac{A}{B} + \ln(B))} + \frac{A}{B} + \ln(B)) \\ &\leq B(\sqrt{2}(-1 + \frac{A}{B} + \ln(B)) + \frac{A}{B} + \ln(B)) \\ &\leq (\sqrt{2}+1)B(\frac{A}{B} + \ln(B)) \\ &\leq \frac{5}{2}(A + B\ln(B)). \end{aligned}$$

This concludes the proof of the Lemma. □

Lemma 20. *For any $\alpha > 0$ and $\beta \in \mathbb{R}$, we have the following inequality:*

$$\phi_\alpha(\beta) - \phi_\alpha(0) \geq \frac{1}{4}\max\left\{0, |\beta| \ln \frac{|\beta|}{2\alpha^2}\right\}.$$

Proof. Let us fix $\alpha \in \mathbb{R}$. First notice that by parity in β of the functions involved, and as the inequality holds in $\beta = 0$, we can suppose that $\beta > 0$ and define

$$f(\beta) := \phi_\alpha(\beta) - \phi_\alpha(0) = \frac{1}{4}\left[\beta \text{arcsinh}\left(\frac{\beta}{2\alpha^2}\right) - \sqrt{\beta^2 + 4\alpha^4} + 2\alpha^2\right].$$

Trivially, $f'(\beta) = \frac{1}{4}\text{arcsinh}\left(\frac{\beta}{2\alpha^2}\right) > 0$. Hence, it increases on $\mathbb{R}_+$ and as $f(0) = 0$, f is always positive. This show the inequality for the left term of the max.

For the other term of the max, let us define $g(\beta) := \frac{1}{4}\beta \ln \frac{\beta}{2\alpha^2}$, we have that

$$4[f'(\beta) - g'(\beta)] = \text{arcsinh}\left(\frac{\beta}{2\alpha^2}\right) - \ln\left(\frac{\beta}{2\alpha^2}\right) + 1 = \ln\left(1 + \sqrt{1 + \frac{4\alpha^4}{\beta^2}}\right) + 1 > 0.$$

Hence, $f - g$ increases and as $f(0) - g(0) = 0$, we have that $f > g$ which concludes the proof. □

[4] see https://en.wikipedia.org/wiki/Lambert_W_function for more details

Appendix C

Appendix for Chapter 8

Organisation of the Appendix.

1. In Appendix C.1, we provide additional experiments for uncentered data as well as on the behaviour of the sharpness and trace of the Hessian along the trajectory of the iterates. We finally provide an experiment highlighting the EoS regime for SGD.

2. In Appendix C.2, we prove that (β_k) follows a Mirror descent recursion with varying potentials. We explicit these potentials and discuss some consequences.

3. In Appendix C.3 we prove that (S)GD on the $\frac{1}{2}(w_+^2 - w_-^2)$ and $u \odot v$ parametrisations with suitable initialisations lead to the same sequence (β_k).

4. In Appendix C.4, we show that the hypentropy $\psi_{\boldsymbol{\alpha}}$ converges to a **weighted**-ℓ_1-norm when $\boldsymbol{\alpha}$ converges to 0 non-uniformly. We then discuss the effects of this **weighted** ℓ_1-norm for sparse recovery.

5. In Appendix C.5, we provide our descent lemmas for mirror descent with varying potentials and prove the boundedness of the iterates.

6. In Appendix C.6, we prove our main results: Theorem 1 and Theorem 2, as well as quantitative convergence (Proposition 17).

7. In Appendix C.7, we prove the lemmas and propositions given in the main text.

8. In Appendix C.8, we provide technical lemmas used throughout the proof of Theorem 1 and Theorem 2.

9. In Appendix C.9, we provide concentration results for random matrices and random vectors, used to estimate with high probability (w.r.t. the dataset) quantities related to the data.

C.1 Additional experiments and results

C.1.1 Uncentered data

When the data is uncentered, the discussion and the conclusion for GD are somewhat different. This paragraph is motivated by the observation of Nacson et al. [2022] who notice that GD with large stepsizes helps to recover low ℓ_1 solutions for uncentered data (Figure C.1). We make the following assumptions on the uncentered inputs.

Assumption 16. *There exist $\mu \in \mathbb{R}^d$ and $\delta, c_0, c_1, c_2 > 0$ such that for all s-sparse vectors β verifying $\langle \mu, \beta \rangle \geq c_0 \|\beta\|_\infty \|\mu\|_\infty$, there exists $\varepsilon \in \mathbb{R}^d$ such that $(X^\top X)\beta = \langle \beta, \mu \rangle \mu + \varepsilon$ where $\|\varepsilon\|_2 \leq \delta \|\beta\|_2$ and $c_1 \langle \beta, \mu \rangle^2 \mu^2 \leq \frac{1}{n} \sum_i x_i^2 \langle x_i, \beta \rangle^2 \leq c_2 \langle \beta, \mu \rangle^2 \mu^2$.*

Assumption 16 is not restrictive and holds with high probability for $\mathcal{N}(\mu \mathbf{1}, \sigma^2 I_d)$ inputs when $\mu \gg \sigma \mathbf{1}$ (see Lemma 29 in Appendix). The following lemma characterises the initial shape of SGD and GD gradients for uncentered data.

Proposition 31 (Shape of the (stochastic) gradient at initialisation)**.** *Under Assumption 16 and if $\langle \mu, \beta^\star_{\text{sparse}} \rangle \geq c_0 \|\beta\|_\infty \|\mu\|_\infty$, the squared full batch gradient and the expected stochastic gradient descent at initialisation satisfy, for some ε satisfying $\|\varepsilon\|_\infty \ll \|\beta_{\text{sparse}}\|_2$:*

$$\nabla \mathcal{L}(\beta_0) = \langle \beta^\star_{\text{sparse}}, \mu \rangle^2 \mu^2 + \varepsilon \,, \tag{C.1}$$

$$\mathbb{E}_{i \sim \text{Unif}([n])} [\nabla \mathcal{L}_i(\beta_0)^2] = \Theta\Big(\langle \beta^\star_{\text{sparse}}, \mu \rangle^2 \mu^2 \Big) \,. \tag{C.2}$$

In this case the initial gradients of SGD and of GD **are both homogeneous**, explaining the behaviours of gradient descent in Figure C.1 (App. C.1): large stepsizes help in the recovery of the sparse solution in the presence of uncentered data, as opposed to centered data. Note that for decentered data with a $\mu \in \mathbb{R}^d$ orthogonal to $\beta^\star_{\text{sparse}}$, there is no effect of decentering on the recovered solution. If the support of μ is the same as that of $\beta^\star_{\text{sparse}}$, the effect is detrimental and the same discussion as in the centered data case applies.

Figure C.1: for uncentered data the solutions of GD and SGD have similar behaviours, corroborating Proposition 31.

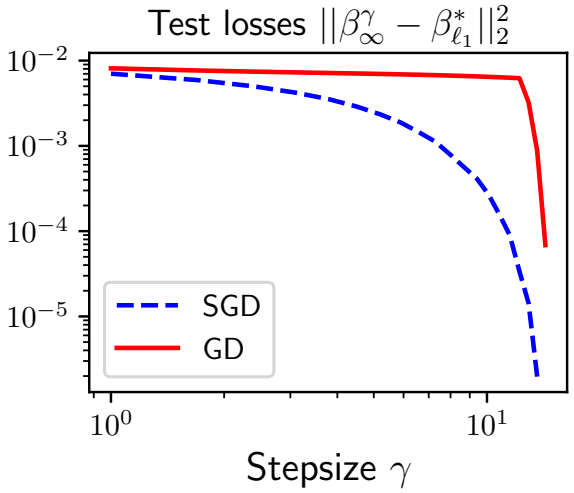


Figure C.1: Noiseless sparse regression with a 2-layer DLN with uncentered data $x_i \sim \mathcal{N}(\mu \mathbf{1}, I_d)$ where $\mu = 5$. All the stepsizes lead to convergence to a global solution and the solutions of SGD and GD have similar behaviours, corroborating Proposition 31. The setup corresponds to $(n, d, s, \alpha) = (20, 30, 3, 0.1)$.

C.1.2 Behaviour of the maximal value and trace of the hessian

Here in Figure C.2, we provide some additional experiments on the behaviour of: (1) the maximum eigenvalue of the hessian $\nabla^2 F(w_\infty^\gamma)$ at the convergence of the iterates of SGD and GD (2) the trace of hessian at the convergence of the iterates. As is clearly observed, increasing the stepsize for GD leads to a 'flatter' minimum in terms of the maximum eigenvalue of the hessian, while increasing the stepsize for SGD leads to a 'flatter' minimum in terms of its trace. These two solutions have very different structures. Indeed from the value of the hessian Equation (C.9) at a global solution, and (very) roughly assuming that '$X^\top X = I_d$' and that '$\boldsymbol{\alpha} \sim 0$' (pushing the EoS phenomenon), one can see that minimising $\lambda_{\max}(\nabla^2 F(w))$ under the constraints $X(w_+^2 - w_-^2) = y$ and $w_+ \odot w_- = 0$ is equivalent to minimising $\|\beta\|_\infty$ under the constaint $X\beta = y$. On the other hand minimising the trace of the hessian is equivalent to minimising the ℓ_1-norm.

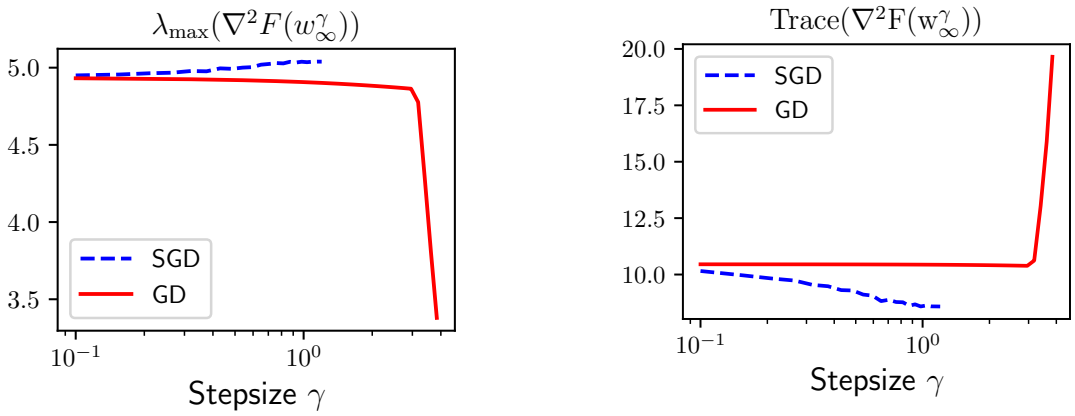


Figure C.2: Noiseless sparse regression setting. Diagonal linear network. Centered data. Behaviour of 2 different types of flatness of the recovered solution by SGD and GD depending on the stepsize. The setup corresponds to $(n, d, s, \alpha) = (20, 30, 3, 0.1)$.

C.1.3 Edge of Stability for SGD

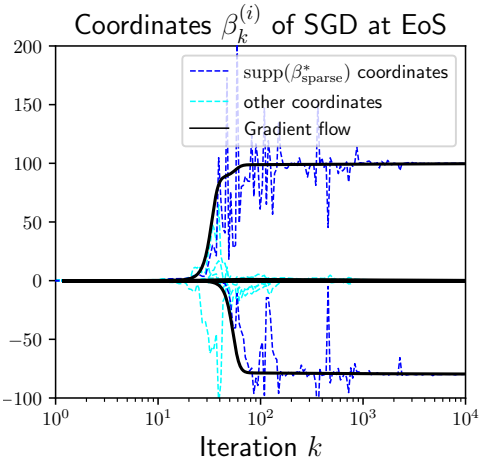


Figure C.3: SGD at the edge of stability: all coordinates fluctuate, and the sparse solution is recovered. As opposed to GD at the EoS, since all coordinates fluctuate, the coordinates to recover are not more penalised than the others.

C.2 Main ingredients behind the proof of Theorem 1 and Theorem 2

In this section, we show that the iterates $(\beta_k)_{k\geq 0}$ follow a *stochastic mirror descent* with *varying potentials*. At the core of our analysis, this result enables us to *(i)* prove convergence of the iterates to an interpolator and *(ii)* completely characterise the inductive bias of the algorithm (SGD or GD). Unveiling a mirror-descent like structure to characterise the implicit bias of a gradient method is classical. For gradient flow over diagonal linear networks [Woodworth et al., 2020a], the iterates follow a mirror flow with respect to the hypentropy (8.4) with parameter α the initialisation scale, while for stochastic gradient flow [Pesme et al., 2021] the mirror flow has a continuously evolving potential.

C.2.1 Mirror descent and varying potentials

We recall that for a strictly convex reference function $h : \mathbb{R}^d \to \mathbb{R}$, the (stochastic) mirror descent iterates algorithm write as [Bauschke et al., 2017, Dragomir et al., 2021], where the minimum is assumed to be attained over $\mathbb{R}^d$ and unique:

$$\beta_{k+1} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \left\{\eta_k \langle g_k, \beta\rangle + D_h(\beta, \beta_k)\right\}, \tag{C.3}$$

for stochastic gradients g_k, stepsize $\gamma_k \geq 0$, and $D_h(\beta, \beta') = h(\beta) - h(\beta') - \langle \nabla h(\beta'), \beta - \beta'\rangle$ is the Bregman divergence associated to h. Iteration (C.3) can also be cast as

$$\nabla h(\beta_{k+1}) = \nabla h(\beta_k) - \gamma_k g_k \,. \tag{C.4}$$

Now, let (h_k) be strictly convex reference functions $\mathbb{R}^d \to \mathbb{R}$. Whilst in continuous time, there is only one natural way to extend mirror flow to varying potentials, in discrete time the varying potentials can be incorporated in eq. (C.3) (replacing h by h_k and leading to $\nabla h_k(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k g_k$), the mirror descent with varying potentials we study incorporates h_{k+1} and h_k in eq. (C.4). The iterates are thus defined as through:

$$\beta_{k+1} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \left\{\eta_k \langle g_k, \beta\rangle + D_{h_{k+1},h_k}(\beta, \beta_k)\right\},$$

where $D_{h_{k+1},h_k}(\beta,\beta') = h_{k+1}(\beta) - h_k(\beta') - \langle \nabla h_k(\beta'), \beta - \beta' \rangle$, a recursion that can also be cast as:

$$\nabla h_{k+1}(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k g_k \,.$$

To derive convergence of the iterates, we prove analogs to classical mirror descent lemmas, generalised to time-varying potentials.

C.2.2 The iterates (β_k) follow a stochastic mirror descent with varying potential recursion

In this section we show and prove that the iterates $(\beta_k)_k$ follow a stochastic mirror descent with varying potentials. Before stating the proposition, we recall the definition of the potentials. To do so we introduce several quantities.

Let $q, q_\pm : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be defined as:

$$\begin{aligned} q_\pm(x) &= \mp 2x - \ln\left((1 \mp x)^2\right), \\ q(x) &= \frac{1}{2}(q_+(x) + q_-(x)) = -\frac{1}{2}\ln\left((1-x^2)^2\right), \end{aligned}$$

with the convention that $q(1) = \infty$. Notice that $q(x) \geq 0$ for $|x| \leq \sqrt{2}$ and $q(x) < 0$ otherwise. For the iterates $\beta_k = u_k \odot v_k \in \mathbb{R}^d$, we recall the definition of the following quantities:

$$\begin{aligned} \boldsymbol{\alpha}_{\pm,k} &= \boldsymbol{\alpha} \exp(-\frac{1}{2}\sum_{i=0}^{k-1} q_\pm(\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))) \in \mathbb{R}^d_{>0}\,, \\ \boldsymbol{\alpha}_k^2 &= \boldsymbol{\alpha}_{+,k} \odot \boldsymbol{\alpha}_{-,k}\,, \\ \phi_k &= \frac{1}{2}\operatorname{arcsinh}\left(\frac{\boldsymbol{\alpha}_{+,k}^2 - \boldsymbol{\alpha}_{-,k}^2}{2\boldsymbol{\alpha}_k^2}\right) \in \mathbb{R}^d\,. \end{aligned}$$

Finally for $k \geq 0$, we define the potentials $(h_k : \mathbb{R}^d \to \mathbb{R})_{k\geq 0}$ as:

$$h_k(\beta) = \psi_{\boldsymbol{\alpha}_k}(\beta) - \langle \phi_k, \beta \rangle\,, \tag{C.5}$$

where $\psi_{\boldsymbol{\alpha}_k}$ is the hyperbolic entropy defined in (8.4) of scale $\boldsymbol{\alpha}_k$:

$$\psi_{\boldsymbol{\alpha}_k}(\beta) = \frac{1}{2}\sum_{i=1}^{d}\left(\beta_i \operatorname{arcsinh}(\frac{\beta_i}{\alpha_{k,i}^2}) - \sqrt{\beta_i^2 + \alpha_{k,i}^4} + \alpha_{k,i}^2\right)$$

where $\alpha_{k,i}$ corresponds to the i^{th} coordinate of the vector $\boldsymbol{\alpha}_k$.

Now that all the relevant quantities are define, we can state the following proposition which explicits the time-varying stochastic mirror descent followed by $(\beta_k)_k$

Proposition 32. *The iterates $(\beta_k = u_k \odot v_k)_{k\geq 0}$ from eq. (8.3) satisfy the Stochastic Mirror Descent recursion with varying potentials $(h_k)_k$:*

$$\nabla h_{k+1}(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,, \tag{C.6}$$

where $h_k : \mathbb{R}^d \to \mathbb{R}$ for $k \geq 0$ are defined Equation (C.5). Since $\nabla h_0(\beta_0) = 0$ we have:

$$\nabla h_k(\beta_k) \in \operatorname{span}(x_1, \ldots, x_n)$$

Proof. Using Proposition 33, we study the $\frac{1}{2}(w_+^2 - w_-^2)$ parametrisation instead of the $u \odot v$, indeed this is the natural parametrisation to consider when doing the calculations as it "separates" the recursions on w_+ and w_-.

Let us focus on the recursion of w_+:

$$w_{+,k+1} = (1 - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)) \cdot w_{+,k} \,.$$

We have:

$$\begin{aligned} w_{+,k+1}^2 &= (1 - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))^2 \cdot w_{+,k}^2 \\ &= \exp\left(\ln((1 - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))^2)\right) \cdot w_{+,k}^2 \,, \end{aligned}$$

with the convention that $\exp(\ln(0)) = 0$. This leads to:

$$\begin{aligned} w_{+,k+1}^2 &= \exp\big(- 2\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(w_k) + 2\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k) + \ln((1 - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))^2)\big) \cdot w_{+,k}^2 \\ &= \exp\big(- 2\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k) - q_+(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))\big) \cdot w_{+,k}^2 \,, \end{aligned}$$

since $q_+(x) = -2x - \ln((1-x)^2)$. Expanding the recursion and using that $w_{+,k=0}$ is initialised at $w_{+,k=0} = \boldsymbol{\alpha}$, we thus obtain:

$$\begin{aligned} w_{+,k}^2 &= \boldsymbol{\alpha}^2 \exp(-\sum_{\ell=0}^{k-1} q_+(\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))) \exp\left(-2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\right) \\ &= \boldsymbol{\alpha}_{+,k}^2 \exp\left(-2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\right) , \end{aligned}$$

where we recall that $\boldsymbol{\alpha}_{\pm,k}^2 = \boldsymbol{\alpha}^2 \exp(-\sum_{\ell=0}^{k-1} q_\pm(\gamma_\ell g_\ell))$. One can easily check that we similarly get:

$$w_{-,k}^2 = \boldsymbol{\alpha}_{-,k}^2 \exp\left(+2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\right) ,$$

leading to:

$$\begin{aligned} \beta_k &= \frac{1}{2}(w_{+,k}^2 - w_{-,k}^2) \\ &= \frac{1}{2}\boldsymbol{\alpha}_{+,k}^2 \exp\left(-2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\right) - \frac{1}{2}\boldsymbol{\alpha}_{-,k}^2 \exp\left(+2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\right) . \end{aligned}$$

Using Lemma 24, the previous equation can be simplified into:

$$\beta_k = \boldsymbol{\alpha}_{+,k}\boldsymbol{\alpha}_{-,k} \sinh\Big(- 2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell) + \operatorname{arcsinh}\Big(\frac{\boldsymbol{\alpha}_{+,k}^2 - \boldsymbol{\alpha}_{-,k}^2}{2\boldsymbol{\alpha}_{+,k}\boldsymbol{\alpha}_{-,k}}\Big)\Big) ,$$

which writes as:

$$\frac{1}{2}\operatorname{arcsinh}\Big(\frac{\beta_k}{\boldsymbol{\alpha}_k^2}\Big) - \phi_k = -\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell) \in \operatorname{span}(x_1, \ldots, x_n) \,,$$

where $\phi_k = \frac{1}{2}\operatorname{arcsinh}\left(\frac{\boldsymbol{\alpha}_{+,k}^2 - \boldsymbol{\alpha}_{-,k}^2}{2\boldsymbol{\alpha}_k^2}\right)$, $\boldsymbol{\alpha}_k^2 = \boldsymbol{\alpha}_{+,k} \odot \boldsymbol{\alpha}_{-,k}$ and since the potentials h_k are defined in Equation (C.5) as $h_k = \psi_{\boldsymbol{\alpha}_k} - \langle \phi_k, \cdot \rangle$ with

$$\psi_{\boldsymbol{\alpha}}(\beta) = \frac{1}{2}\sum_{i=1}^{d}\left(\beta_i \operatorname{arcsinh}\left(\frac{\beta_i}{\boldsymbol{\alpha}_i^2}\right) - \sqrt{\beta_i^2 + \boldsymbol{\alpha}_i^4} + \boldsymbol{\alpha}_i^2\right) \tag{C.7}$$

specifically such that $\nabla h_k(\beta_k) = \frac{1}{2}\operatorname{arcsinh}\left(\frac{\beta_k}{\boldsymbol{\alpha}_k^2}\right) - \phi_k$. Hence,

$$\nabla h_k(\beta_k) = \sum_{\ell<k} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\,,$$

so that:

$$\nabla h_{k+1}(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,,$$

which corresponds to a Mirror Descent with varying potentials $(h_k)_k$. □

C.3 Equivalence of the $u \odot v$ and $\frac{1}{2}(w_+^2 - w_-^2)$ parametrisations

We here prove the equivalence between the $\frac{1}{2}(w_+^2 - w_-^2)$ and $u \odot v$ parametrisations, **that we use throughout the proofs in the Appendix.**

Proposition 33. *Let $(\beta_k)_{k\geq 0}$ and $(\beta'_k)_{k\geq 0}$ be respectively generated by stochastic gradient descent on the $u \odot v$ and $\frac{1}{2}(w_+^2 - w_-^2)$ parametrisations:*

$$(u_{k+1}, v_{k+1}) = (u_k, v_k) - \gamma_k \nabla_{u,v}\big(\mathcal{L}_{\mathcal{B}_k}(u \odot v)\big)(u_k, v_k)\,,$$

and

$$w_{\pm,k+1} = w_{\pm,k} - \gamma_k \nabla_{w_\pm}\big(\mathcal{L}_{\mathcal{B}_k}(\frac{1}{2}(w_+^2 - w_-^2))\big)(w_{+,k}, w_{-,k})\,,$$

initialised as $u_0 = \sqrt{2}\boldsymbol{\alpha}, v_0 = 0$ and $w_{+,0} = w_{-,0} = \boldsymbol{\alpha}$. Then for all $k \geq 0$, we have $\beta_k = \beta'_k$.

Proof. We have:

$$w_{\pm,0} = \boldsymbol{\alpha}\,, \quad w_{\pm,k+1} = (1 \mp \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta'_k))w_{\pm,k}\,,$$

and

$$u_0 = \sqrt{2}\boldsymbol{\alpha}\,, \quad v_0 = 0\,, \quad u_{k+1} = u_k - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)v_k\,, \quad v_{k+1} = v_k - \gamma_k \nabla \mathcal{L}(\beta_k)u_k\,.$$

Hence,

$$\beta_{k+1} = (1 + \gamma_k^2 \nabla \mathcal{L}(\beta_k)^2)\beta_k - \gamma_k(u_k^2 + v_k^2)\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,,$$

and

$$\beta'_{k+1} = (1 + \gamma_k^2 \nabla \mathcal{L}_{\mathcal{B}_k}(\beta'_k)^2)\beta'_k - \gamma_k(w_{+,k}^2 + w_{-,k}^2)\nabla \mathcal{L}_{\mathcal{B}_k}(\beta'_k)\,.$$

Then, let $z_k = \frac{1}{2}(u_k^2 - v_k^2)$ and $z'_k = w_{+,k}w_{-k}$. We have $z_0 = \boldsymbol{\alpha}^2$, $z'_0 = \boldsymbol{\alpha}^2$ and:

$$z_{k+1} = (1 - \gamma_k^2 \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)^2)z_k\,, \quad z'_{k+1} = (1 - \gamma_k^2 \nabla \mathcal{L}_{\mathcal{B}_k}(\beta'_k)^2)z'_k\,.$$

Using $a^2 + b^2 = \sqrt{(2ab)^2 + (a^2 - b^2)^2}$ for $a, b \in \mathbb{R}$, we finally obtain that:

$$u_k^2 + v_k^2 = \sqrt{(2\beta_k)^2 + (2z_k)^2}\,, \quad w_{+,k}^2 + w_{-,k}^2 = \sqrt{(2\beta'_k)^2 + (2z'_k)^2}\,.$$

We conclude by observing that (β_k, z_k) and (β'_k, z'_k) follow the exact same recursions, initialised at the same value $(0, \boldsymbol{\alpha}^2)$.

□

C.4 Convergence of ψ_α to a weighted ℓ_1 norm and harmful behaviour

We show that when taking the scale of the initialisation to 0, one must be careful in the characterisation of the limiting norm, indeed if each entry does not go to zero "at the same speed", then the limit norm is a **weighted** ℓ_1-norm rather than the classical ℓ_1 norm.

Proposition 34. *For $\alpha \geq 0$ and a vector $h \in \mathbb{R}^d$, let $\tilde{\alpha} = \alpha \exp(-h \ln(1/\alpha)) \in \mathbb{R}^d$. Then we have that for all $\beta \in \mathbb{R}^d$*

$$\psi_{\tilde{\alpha}}(\beta) \underset{\alpha \to 0}{\sim} \ln(\frac{1}{\alpha}) \cdot \sum_{i=1}^{d} (1 + h_i)|\beta_i|.$$

Proof. Recall that

$$\psi_{\tilde{\alpha}}(\beta) = \frac{1}{2} \sum_{i=1}^{d} \left(\beta_i \text{arcsinh}(\frac{\beta_i}{\tilde{\alpha}_i^2}) \ - \sqrt{\beta_i^2 + {\tilde{\alpha}_i}^4} + \tilde{\alpha}_i^2 \right)$$

Using that $\text{arcsinh}(x) \underset{|x| \to \infty}{\sim} \text{sgn}(x) \ln(|x|)$, and that $\ln(\frac{1}{\tilde{\alpha}_i^2}) = (1 + h_i) \ln(\frac{1}{\alpha^2})$ we obtain that

$$\begin{aligned} \psi_{\tilde{\alpha}}(\beta) &\underset{\alpha \to 0}{\sim} \frac{1}{2} \sum_{i=1}^{d} \text{sgn}(\beta_i) \beta_i (1 + h_i) \ln(\frac{1}{\alpha^2}) \\ &= \frac{1}{2} \ln(\frac{1}{\alpha^2}) \sum_{i=1}^{d} (1 + h_i)|\beta_i|. \end{aligned}$$

□

The following Figure C.4 illustrates the effect of the non-uniform shape $\boldsymbol{\alpha}$ on the corresponding potential $\psi_{\boldsymbol{\alpha}}$.

More generally, for α such that $\alpha_i \to 0$ for all $i \in [d]$ at rates such that $\ln(1/\alpha_i) \sim q_i \ln(1/\max_i \alpha_i)$, we retrieve a weighted ℓ_1 norm:

$$\frac{\psi_\alpha(\beta)}{\ln(1/\alpha^2)} \to \sum_{i=1}^{d} q_i |\beta_i| \,.$$

Hence, even for arbitrary small $\max_i \alpha_i$, if the *shape* of α is 'bad', the interpolator β_α that minimizes ψ_α can be arbitrary far away from $\beta^\star_{\ell^1}$ the interpolator of minimal ℓ_1 norm.

We illustrate the importance of the previous proposition in the following example.

Example 1. *We illustrate how, even for arbitrary small $\max_i \alpha_i$, the interpolator $\beta^\star_\alpha$ that minimizes ψ_α can be far from the minimum ℓ_1 norm solution, due to the shape of $\boldsymbol{\alpha}$ that is not uniform. The message of this example is that for $\boldsymbol{\alpha} \to 0$ non-uniformly across coordinates, if the coordinates of α that go slowly to 0 coincide with the non-null coordinates of the sparse interpolator we want to retrieve, then $\beta^\star_\alpha$ will be far from the sparse solution.*

A simple counterexample can be built: let $\beta^\star_{\text{sparse}} = (1, \dots, 1, 0, \dots, 0)$ (with only the $s = o(d)$ first coordinates that are non-null), and let (x_i), (y_i) be generated as $y_i = \langle \beta^\star_{\text{sparse}}, x_i \rangle$ with $x_i \sim \mathcal{N}(0, 1)$. For n large enough (n of order $s \ln(d)$ where s is the sparsity), the design matrix X is RIP [Candès et al., 2006], so that the minimum ℓ_1 norm interpolator $\beta^\star_{\ell^1}$ is exactly equal to $\beta^\star_{\text{sparse}}$.

However, if α is such that $\max_i \alpha_i \to 0$ with $h_i >> 1$ for $j \leq s$ and $h_i = 1$ for $i \geq s + 1$ (h_i as in Proposition 34), $\beta^\star_\alpha$ will be forced to verify $\beta^\star_{\alpha,i} = 0$ for $i \leq s$ and hence $\|\beta^\star_{\alpha,1} - \beta^\star_{\ell^1}\|_1 \geq s$.

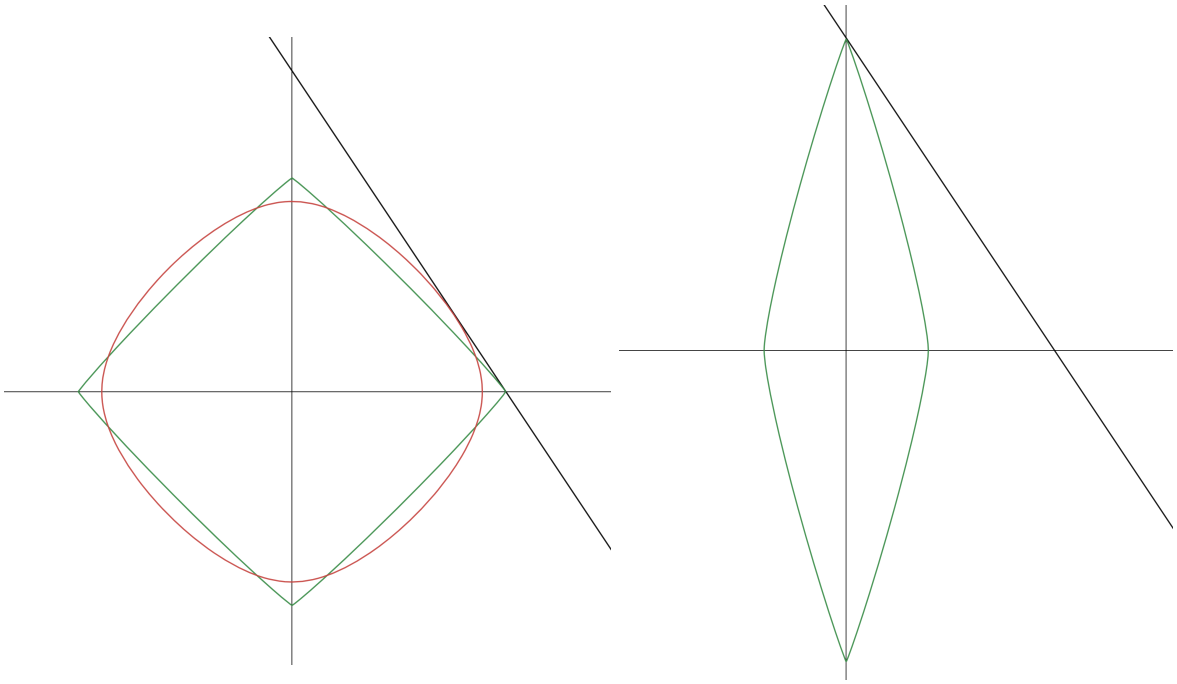


Figure C.4: *Left*: Uniform $\boldsymbol{\alpha} = \alpha \mathbf{1}$: a smaller scale α leads to the potential ψ_α being closer to the ℓ_1-norm. *Right*: A non uniform $\boldsymbol{\alpha}$ can lead to the recovery of a solution which is very far from the minimum ℓ_1-norm solution. The affine line corresponds to the set of interpolators when $n = 1$, $d = 2$ and $s = 1$.

C.5 Main descent lemma and boundedness of the iterates

The goal of this section is to prove the following proposition, our main descent lemma: for well-chosen stepsizes, the Bregman divergences $(D_{h_k}(\beta^\star, \beta_k))_{k \geq 0}$ decrease. We then use this proposition to bound the iterates for both SGD and GD.

Proposition 35. *There exist a constant $c > 0$ and $B > 0$ such that $B = \mathcal{O}(\inf_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_\infty)$ for GD and $B = \mathcal{O}(\ln(1/\alpha) \inf_{\beta^\star \in \mathcal{S}} \|\beta^\star\|_\infty)$ for SGD, such that if $\gamma_k \leq \frac{c}{LB}$ for all k, then we have, for all $k \geq 0$ and any interpolator $\beta^\star \in \mathcal{S}$:*

$$D_{h_{k+1}}(\mathsf{w}^\star, \mathsf{w}_{k+1}) \leq D_{h_k}(\mathsf{w}^\star, \mathsf{w}_k) - \gamma_k \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\,.$$

To prove this result, we first provide a general descent lemma for time-varying mirror descent (Proposition 36, Appendix C.5.1), before proving the proposition for fixed iteration k and bound $B > 0$ on the iterates infinity norm in Appendix C.5.2 (Proposition 37). We finally use this to prove a bound on the iterates infinity norm in Appendix C.5.3.

C.5.1 Descent lemma for (stochastic) mirror descent with varying potentials

In the following we adapt a classical mirror descent equality but for time varying potentials, that differentiates from Orabona et al. [2015] in that it enables us to prove the decrease of the Bregman divergences of the iterates. Moreover, as for classical MD, it is an equality.

Proposition 36. *For $h, g : \mathbb{R}^d \to \mathbb{R}$ functions, let $D_{h,g}(\mathsf{w}, \mathsf{w}') = h(\mathsf{w}) - g(\mathsf{w}') - \langle \nabla g(\mathsf{w}'), \mathsf{w} - \mathsf{w}' \rangle$[1] for $\beta, \beta' \in \mathbb{R}^d$. Let (h_k) strictly convex functions defined $\mathbb{R}^d$ $\mathcal{L}$ a convex function defined on $\mathbb{R}^d$.*

[1] for $h = g$, we recover the classical Bregman divergence that we denote $D_h = D_{h,h}$

Let (w_k) *defined recursively through* $\mathsf{w}_0 \in \mathbb{R}^d$*, and*

$$\mathsf{w}_{k+1} \in \operatorname*{arg\,min}_{\mathsf{w}\in\mathbb{R}^d} \left\{ \gamma_k \langle \nabla \mathcal{L}(\mathsf{w}_k), \mathsf{w} - \mathsf{w}_k \rangle + D_{h_{k+1},h_k}(\mathsf{w}, \mathsf{w}_k) \right\},$$

where we assume that the minimum is unique and attained in $\mathbb{R}^d$*. Then,* (w_k) *satisfies*

$$\nabla h_{k+1}(\mathsf{w}_{k+1}) = \nabla h_k(\mathsf{w}_k) - \gamma_k \nabla \mathcal{L}(\mathsf{w}_k)\,,$$

and for any $\beta \in \mathbb{R}^d$*,*

$$\begin{aligned} D_{h_{k+1}}(\mathsf{w}, \mathsf{w}_{k+1}) = D_{h_k}(\mathsf{w}, \mathsf{w}_k) - \gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta_k - \beta \rangle + D_{h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1}) \\ - \big(h_{k+1} - h_k\big)(\mathsf{w}_k) + \big(h_{k+1} - h_k\big)(\mathsf{w})\,. \end{aligned}$$

Proof. Let $\beta \in \mathbb{R}^d$. Since we assume that the minimum through which w_{k+1} is computed is attained in $\mathbb{R}^d$, the gradient of the function $V_k(\mathsf{w}) = \gamma_k \langle \nabla \mathcal{L}(\mathsf{w}_k), \mathsf{w} - \mathsf{w}_k \rangle + D_{h_{k+1},h_k}(\mathsf{w}, \mathsf{w}_k)$ evaluated at w_{k+1} is null, leading to $\nabla h_{k+1}(\mathsf{w}_{k+1}) = \nabla h_k(\mathsf{w}_k) - \gamma_k \nabla \mathcal{L}(\mathsf{w}_k)$.

Then, since $\nabla V_k(\mathsf{w}_{k+1}) = 0$, we have $D_{V_k}(\mathsf{w}, \mathsf{w}_{k+1}) = V_k(\mathsf{w}) - V_k(\mathsf{w}_{k+1})$. Using $\nabla^2 V_k = \nabla^2 h_{k+1}$, we also have $D_{V_k} = D_{h_{k+1}}$. Hence:

$$D_{h_{k+1}}(\mathsf{w}, \mathsf{w}_{k+1}) = \gamma_k \langle \nabla \mathcal{L}(\mathsf{w}_k), \mathsf{w} - \mathsf{w}_{k+1} \rangle + D_{h_{k+1},h_k}(\mathsf{w}, \mathsf{w}_k) - D_{h_{k+1},h_k}(\mathsf{w}_{k+1}, \mathsf{w}_k)\,.$$

We write $\gamma_k \langle \nabla \mathcal{L}(\mathsf{w}_k), \mathsf{w} - \mathsf{w}_{k+1} \rangle = \gamma_k \langle \nabla \mathcal{L}(\mathsf{w}_k), \mathsf{w} - \mathsf{w}^k \rangle + \gamma_k \langle \nabla \mathcal{L}(\mathsf{w}_k), \mathsf{w}_k - \mathsf{w}_{k+1} \rangle$. We also have $\gamma_k \langle \nabla \mathcal{L}(\mathsf{w}_k), \mathsf{w}_k - \mathsf{w}_{k+1} \rangle = \langle \nabla h_k(\mathsf{w}_k) - \nabla h_{k+1}(\mathsf{w}_{k+1}), \mathsf{w}_k - \mathsf{w}_{k+1} \rangle = D_{h_k,h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1}) + D_{h_{k+1},h_k}(\mathsf{w}_{k+1}, \mathsf{w}^k)$, so that $\gamma_k \langle \nabla \mathcal{L}(\mathsf{w}_k), \mathsf{w}_k - \mathsf{w}_{k+1} \rangle - D_{h_{k+1},h_k}(\mathsf{w}_{k+1}, \mathsf{w}^k) = D_{h_k,h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1})$. Thus,

$$D_{h_{k+1}}(\mathsf{w}, \mathsf{w}_{k+1}) = D_{h_{k+1},h_k}(\mathsf{w}, \mathsf{w}_k) - \gamma_k \big(D_f(\mathsf{w}, \mathsf{w}_k) + D_f(\mathsf{w}_k, \mathsf{w})\big) + D_{h_k,h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1})\,,$$

and writing $D_{h,g}(\mathsf{w}, \mathsf{w}') = D_g(\mathsf{w}, \mathsf{w}') + h(\mathsf{w}) - g(\mathsf{w})$ concludes the proof. □

C.5.2 Proof of Proposition 37

In next proposition, we use Proposition 36 to prove our main descent lemma. To that end, we bound the error terms that appear in Proposition 36 as functions of $\mathcal{L}_{\mathcal{B}_k}(\beta_k)$ and norms of β_k, β_{k+1}, so that for explicit stepsizes, the error terms can be cancelled by half of the negative quantity $-2\mathcal{L}_{\mathcal{B}_k}(\beta_k)$.

Additional notation: let $L_2, L_\infty > 0$ such that $\forall \beta$, $\|H_{\mathcal{B}}\beta\|_2 \leq L\|\beta\|_2$, $\|H_{\mathcal{B}}\beta\|_\infty \leq L\|\beta\|_\infty$ for all batches $\mathcal{B} \subset [n]$ of size b.

Proposition 37. *Let* $k \geq 0$ *and* $B > 0$*. Provided that* $\|\beta_k\|_\infty, \|\beta_{k+1}\|_\infty, \|\beta^\star\|_\infty \leq B$ *and* $\gamma_k \leq \frac{c}{LB}$ *where* $c > 0$ *is some numerical constant, we have:*

$$D_{h_{k+1}}(\mathsf{w}^\star, \mathsf{w}_{k+1}) \leq D_{h_k}(\mathsf{w}^\star, \mathsf{w}_k) - \gamma_k \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\,. \tag{C.8}$$

Proof. Let $\beta^\star \in \mathcal{S}$ be any interpolator. From Proposition 36:

$$D_{h_{k+1}}(\beta^\star, \beta_{k+1}) = D_{h_k}(\beta^\star, \beta_k) - 2\gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) + D_{h_{k+1}}(\beta_{k+1}, \beta_k) - (h_{k+1} - h_k)(\beta_k) + (h_{k+1} - h_k)(\beta^\star).$$

We want to bound the last three terms of this equality. First, to bound the last two we apply Lemma 27 assuming that $\|\beta^\star\|_\infty, \|\beta_{k+1}\|_\infty \leq B$:

$$-(h_{k+1} - h_k)(\beta_k) + (h_{k+1} - h_k)(\beta^\star) \leq 24BL_2\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)$$

We now bound $D_{h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1})$. Classical Bregman manipulations provide that

$$\begin{aligned} D_{h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1}) &= D_{h^*_{k+1}}(\nabla h_{k+1}(\mathsf{w}_{k+1}), \nabla h_{k+1}(\mathsf{w}_k)) \\ &= D_{h^*_{k+1}}(\nabla h_k(\mathsf{w}^k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k), \nabla h_{k+1}(\mathsf{w}_k)) \,. \end{aligned}$$

From Lemma 26 we have that h_{k+1} is $\min(1/(4\alpha^2_{k+1}), 1/(4B))$ strongly convex on the ℓ^∞-centered ball of radius B therefore h^*_{k+1} is $\max(4\alpha^2_{k+1}, 4B) = 4B$ (for α small enough or B big enough) smooth on this ball, leading to:

$$\begin{aligned} D_{h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1}) &\leq 2B \|\nabla h_k(\mathsf{w}_k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k) - \nabla h_{k+1}(\mathsf{w}_k)\|_2^2 \\ &\leq 4B \big(\|\nabla h_k(\mathsf{w}_k) - \nabla h_{k+1}(\mathsf{w}_k)\|_2^2 + \|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\|_2^2 \big) \,. \end{aligned}$$

Using $|\nabla h_k(\mathsf{w}) - \nabla h_{k+1}(\mathsf{w})| \leq 2\delta_k$ where $\delta_k = q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$, we get that:

$$D_{h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1}) \leq 8B \|\delta_k\|_2^2 + 4BL\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k) \,.$$

Now, $\|\delta_k\|_2^2 \leq \|\delta_k\|_1 \|\delta_k\|_\infty$ and using Lemma 25, $\|\delta_k\|_1 \|\delta_k\|_\infty \leq 4\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\|_2^2 \|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\|_\infty^2 \leq 2\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\|_2^2$ since $\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\|_\infty \leq \gamma_k L_\infty \|\beta_k - \beta_\infty\| \leq \gamma_k \times 2LB \leq 1/2$ is verified for $\gamma_k \leq 1/(4LB)$. Thus,

$$D_{h_{k+1}}(\mathsf{w}_k, \mathsf{w}_{k+1}) \leq 40BL_2\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k) \,.$$

Hence, provided that $\|\beta_k\|_\infty \leq B$, $\|\beta_{k+1}\|_\infty \leq B$ and $\gamma_k \leq 1/(4LB)$, we have:

$$D_{h_{k+1}}(\mathsf{w}^\star, \mathsf{w}_{k+1}) \leq D_{h_k}(\mathsf{w}^\star, \mathsf{w}_k) - 2\gamma_k \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k) + 64L_2\gamma_k^2 B \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k) \,,$$

and thus

$$D_{h_{k+1}}(\mathsf{w}^\star, \mathsf{w}_{k+1}) \leq D_{h_k}(\mathsf{w}^\star, \mathsf{w}_k) - \gamma_k \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k) \,.$$

if $\gamma_k \leq \frac{c}{BL}$, where $c = \frac{1}{64}$.

□

C.5.3 Bound on the iterates

We now bound the iterates (β_k) by an explicit constant B that depends on $\|\beta^\star\|_1$ (for any fixed $\beta^\star \in \mathcal{S}$).

The first bound we prove holds for both SGD and GD, and is of the form $\mathcal{O}(\|\beta^\star\|_1 \ln(1/\alpha^2)$ while the second bound, that holds only for GD ($b = n$) is of order $\mathcal{O}(\|\beta^\star\|_1)$ (independent of α). While a bound independent of α is only proved for GD, we believe that such a result also holds for SGD, and in both cases B should be thought of order $\mathcal{O}(\|\beta^\star\|_1)$.

Bound that depends on α for GD and SGD

A consequence of Proposition 37 is the boundedness of the iterates, as shown in next corollary. Hence, Proposition 37 can be applied using B a uniform bound on the iterates ℓ^∞ norm.

Corollary 5. *Let $B = 3\|\beta^\star\|_1 \ln\big(1 + \frac{\|\beta^\star\|_1}{\alpha^2}\big)$. For stepsizes $\gamma_k \leq \frac{c}{BL}$, we have $\|\beta_k\|_\infty \leq B$ for all $k \geq 0$.*

Proof. We proceed by induction. Let $k \geq 0$ such that $\|\beta_k\|_\infty \leq B$ for some $B > 0$ and $D_{h_k}(\mathsf{w}^\star, \mathsf{w}_k) \leq D_{h_0}(\mathsf{w}^\star, \mathsf{w}_0)$ (note that these two properties are verified for $k = 0$, since $\beta_0 = 0$). For γ_k sufficiently small (*i.e.*, that satisfies $\gamma_k \leq \frac{c}{B'L}$ where $B' \geq \|\beta_{k+1}\|_\infty, \|\beta_k\|_\infty, \|\beta^\star\|_\infty$), using

Proposition 37, we have $D_{h_{k+1}}(\mathsf{w}^\star, \mathsf{w}_{k+1}) \leq D_{h_k}(\mathsf{w}^\star, \mathsf{w}_k)$ so that $D_{h_{k+1}}(\mathsf{w}^\star, \mathsf{w}_{k+1}) \leq D_{h_0}(\mathsf{w}^\star, \mathsf{w}_0)$, which can be rewritten as:

$$\sum_{i=1}^{d} \alpha_{k+1,i}^2 \Big(\sqrt{1 + \big(\frac{\mathsf{w}_{k+1,i}}{\alpha_{k+1,i}^2}\big)^2} - 1\Big) \leq \sum_{i=1}^{d} \mathsf{w}_i^\star \operatorname{arcsinh}\big(\frac{\mathsf{w}_{k+1,i}}{\alpha^2}\big)\,.$$

Hence, $\|\mathsf{w}_{k+1}\|_1 \leq \|\mathsf{w}^\star\|_1 \ln(1 + \frac{\|\mathsf{w}_{k+1}\|_1}{\alpha^2})$. We then notice that for $x, y > 0$, $x \leq y\ln(1+x) \implies x \leq 3y\ln(1+y)$: if $x > y\ln(1+y)$ and $x > y$, we have that $y\ln(1+y) < y\ln(1+x)$, so that $1 + y < 1 + x$, which contradicts our assumption. Hence, $x \leq \max(y, y\ln(1+y))$. In our case, $x = \|\beta^{k+1}\|_1/\alpha^2$, $y = \|\beta^\star\|_1/\alpha^2$ so that for small alpha, $\ln(1+y) \geq 1$.

Hence, we deduce that $\|\mathsf{w}_{k+1}\|_1 \leq B$, where $B = \|\mathsf{w}^\star\|_1 \ln(1 + \frac{\|\mathsf{w}^\star\|_1}{\alpha^2})$.

This is true as long as γ_k is tuned using B' a bound on $\max(\|\beta_k\|_\infty, \|\beta_{k+1}\|_\infty)$. Using the continuity of β_{k+1} as a function of γ_k (β_k being fixed), we show that $\gamma_k \leq \frac{1}{2} \times \frac{c}{BL}$ can be used using this B. Indeed, let $\phi : \mathbb{R}^+ \to \mathbb{R}^d$ be the function that takes as entry $\gamma_k \geq 0$ and outputs the corresponding $\|\beta_{k+1}\|_\infty$: ϕ is continuous. Let $\gamma_r = \frac{1}{2} \times \frac{c}{rL}$ for $r > 0$ and $\bar{r} = \sup\{r \geq 0 : B < \phi(\gamma_r)\}$ (the set is upper-bounded; if is is empty, we do not need what follows since it means that any stepsize leads to $\|\beta_{k+1}\|_\infty \leq B$). By continuity of ϕ, $\phi(\gamma_{\bar{r}}) = B$. Furthermore, for all r that satisfies $r \geq \max(\phi(\gamma_r), B) \geq \max(\phi(\gamma_r), \|\beta_k\|_\infty, \|\beta^\star\|_\infty)$, we have, using what is proved just above, that $\|\beta_{k+1}\|_\infty \leq B$ and thus $\phi(\gamma_r) \leq B$ for such a r:

Lemma 21. *For $r > 0$ such that $r \geq \max(\phi(\gamma_r), B)$, we have $\phi(\gamma_r) \leq B$.*

Now, if $\bar{r} > B$, by definition of $\bar{r}$ and by continuity of ϕ, since $\phi(\bar{r}) = B$, there exists some $B < r < \bar{r}$ such that $\phi(\gamma_r) > B$ (definition of the supremum) and $\phi(\gamma_r) \leq 2B$ (continuity of ϕ). This particular choice of r thus satisfies $r > B$ and and $\phi(\gamma_r) \leq 2B \leq 2r$, leading to $\phi(\gamma_r) \leq B$, using Lemma 21, hence a contradiction: we thus have $\bar{r} \leq B$.

This concludes the induction: for all $r \geq B$, we have $r \geq \bar{r}$ so that $\phi(\gamma_r) \leq B$ and thus for all stepsizes $\gamma \leq \frac{c}{2LB}$, we have $\|\beta_{k+1}\|_\infty \leq B$.

□

Bound independent of α

We here assume in this subsection that $b = n$. We prove that for gradient descent, the iterates are bounded by a constant that does not depend on α.

Proposition 38. *Assume that $b = n$ (full batch setting). There exists some $B = \mathcal{O}(\|\beta^\star\|_1)$ such that for stepsizes $\gamma_k \leq \frac{c}{BL}$, we have $\|\beta_k\|_\infty \leq B$ for all $k \geq 0$.*

Proof. We first begin by proving the following proposition: for sufficiently small stepsizes, the loss values decrease. In the following lemma we provide a bound on the gradient descent iterates $(w_{+,k}, w_{-,k})$ which will be useful to show that the loss is decreasing.

Proposition 39. *For $\gamma_k \leq \frac{c}{LB}$ where $B \geq \max(\|\beta_k\|_\infty, \|\beta_{k+1}\|_\infty)$, we have $\mathcal{L}(\beta_{k+1}) \leq \mathcal{L}(\beta_k)$*

Proof. Oddly, using the time-varying mirror descent recursion is not the easiest way to show the decrease of the loss, due to the error terms which come up. Therefore to show that the loss is decreasing we use the gradient descent recursion. Recall that the iterates $w_k = (w_{+,k}, w_{-,k}) \in \mathbb{R}^{2d}$ follow a gradient descent on the non convex loss $F(w) = \frac{1}{2}\|y - \frac{1}{2}X(w_+^2 - w_-^2)\|_2$.

For $k \geq 0$, using the Taylor formula we have that $F(w_{k+1}) \leq F(w_k) - \gamma_k(1 - \frac{\gamma_k L_k}{2})\|\nabla F(w_k)\|^2$ with the local smoothness $L_k = \sup_{w \in [w_k, w_{k+1}]} \lambda_{\max}(\nabla^2 F(w))$. Hence if $\gamma_k \leq 1/L_k$ for all k we

get that the loss is non-increasing. We now bound L_k. Computing the hessian ot F, we obtain that:

$$\begin{aligned}\nabla^2 F(w_k) = &\begin{pmatrix}\operatorname{diag}(\nabla\mathcal{L}(\beta_k)) & 0\\ 0 & -\operatorname{diag}(\nabla\mathcal{L}(\beta_k))\end{pmatrix}\\ &+\begin{pmatrix}\operatorname{diag}(w_{+,k})H\operatorname{diag}(w_{+,k}) & -\operatorname{diag}(w_{-,k})H\operatorname{diag}(w_{+,k})\\ -\operatorname{diag}(w_{+,k})H\operatorname{diag}(w_{-,k}) & \operatorname{diag}(w_{-,k})H\operatorname{diag}(w_{-,k})\end{pmatrix}.\end{aligned} \tag{C.9}$$

Let us denote by $M = \begin{pmatrix}M_+ & M_{+,-}\\ M_{+,-} & M_-\end{pmatrix} \in \mathbb{R}^{2d\times 2d}$ the second matrix in the previous equality. With this notation $\|\nabla^2 F(w_k)\| \leq \|\nabla\mathcal{L}(\beta_k)\|_\infty + 2\|M\|$ (where the norm corresponds to the Schatten 2-norm which is the largest eigenvalue for symmetric matrices). Now, notice that:

$$\begin{aligned}\|M\|^2 &= \sup_{u\in\mathbb{R}^{2d},\|u\|=1}\|Mu\|^2\\ &= \sup_{\substack{u_+\in\mathbb{R}^d,\|u_+\|=1\\ u_-\in\mathbb{R}^d,\|u_-\|=1\\ (a,b)\in\mathbb{R}^2,a^2+b^2=1}}\left\|M\begin{pmatrix}a\cdot u_+\\ b\cdot u_-\end{pmatrix}\right\|^2.\end{aligned}$$

We have:

$$\begin{aligned}\left\|M\begin{pmatrix}a\cdot u_+\\ b\cdot u_-\end{pmatrix}\right\|^2 &= \left\|\begin{pmatrix}aM_+u_+ + bM_{+-}u_-\\ aM_{+-}u_+ + bM_-u_-\end{pmatrix}\right\|^2\\ &= \|aM_+u_+ + bM_{+-}u_-\|^2 + \|aM_{+-}u_+ + bM_-u_-\|^2\\ &\leq 2\Big(a^2\|M_+u_+\|^2 + b^2\|M_{+-}u_-\|^2 + a^2\|M_{+-}u_+\|^2 + b^2\|M_-u_-\|^2\Big)\\ &\leq 2\Big(\|M_+\|^2 + \|M_{+-}\|^2 + \|M_-\|^2\Big).\end{aligned}$$

Since $\|M_\pm\| \leq \lambda_{max}\cdot\|w_\pm\|_\infty^2$ and $\|M_{+-}\| \leq \lambda_{max}\|w_+\|_\infty\|w_-\|_\infty$ we finally get that

$$\begin{aligned}\|M\|^2 &\leq 6\lambda_{max}^2\cdot\max(\|w_+\|_\infty^2, \|w_-\|_\infty^2)^2\\ &\leq 6\lambda_{max}^2(\|w_+^2\|_\infty + \|w_-^2\|_\infty)^2\\ &\leq 12\lambda_{max}^2\|w_+^2 + w_-^2\|_\infty^2.\end{aligned}$$

We now upper bound this quantity in the following lemma.

Lemma 22. *For all $k \geq 0$, the following inequality holds component-wise:*

$$w_{+,k}^2 + w_{-,k}^2 = \sqrt{4\boldsymbol{\alpha}_k^4 + \beta_k^2}.$$

Proof. Notice from the definition of $w_{+,k}$ and $w_{-,k}$ given in the proof of Proposition 32 that:

$$|w_{+,k}||w_{-,k}| = \boldsymbol{\alpha}_{-,k}\boldsymbol{\alpha}_{+,k} = \boldsymbol{\alpha}_k^2. \tag{C.10}$$

And $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}^2$. Now since $\boldsymbol{\alpha}_k$ is decreasing coordinate-wise (under our assumptions on the stepsizes, $\gamma_k^2\nabla\mathcal{L}(\beta_k)^2 \leq (1/2)^2 < 1$), we get that.:

$$w_{+,k}^2 + w_{-,k}^2 = 2\sqrt{\boldsymbol{\alpha}_k^4 + \beta_k^2} \leq 2\sqrt{\boldsymbol{\alpha}^4 + \beta_k^2}$$

leading to $w_{+,k}^2 + w_{-,k}^2 \leq \sqrt{4\boldsymbol{\alpha}^4 + B^2}$. □

From Lemma 22, $w_{+,k}^2 + w_{-,k}^2$ is bounded by $2\sqrt{\boldsymbol{\alpha}^4 + B^2}$. Putting things together we finally get that $\|\nabla^2 F(w)\| \leq \|\nabla \mathcal{L}(\beta)\|_\infty + 8\lambda_{max}\sqrt{4\|\boldsymbol{\alpha}\|_\infty^4 + B^2}$. Hence,

$$L_k \leq \sup_{\|\beta\|_\infty \leq B} \|\nabla \mathcal{L}(\beta)\|_\infty + 8\lambda_{\max}\sqrt{\|\boldsymbol{\alpha}\|_\infty^4 + B^2} \leq LB + 8\lambda_{\max}\sqrt{\|\boldsymbol{\alpha}\|_\infty^4 + B^2} \leq 10LB\,,$$

for $B \geq \|\boldsymbol{\alpha}\|_\infty^2$. □

We finally prove the bound on $\|\beta_k\|_\infty$ independent of α for a uniform initialisation $\boldsymbol{\alpha} = \alpha\mathbf{1}$, using the monotonic property of $\mathcal{L}$.

Proposition 40. *Assume that $b = n$ (full batch setting). There exists some $B = \mathcal{O}(\|\beta^\star\|_1)$ such that for stepsizes $\gamma_k \leq \frac{c}{BL}$, we have $\|\beta_k\|_\infty \leq B$ for all $k \geq 0$.*

Proof. In this proof, we first let B be a bound on the iterates. Tuning stepsizes using this bound, we prove that the iterates are bounded by a some $B' = \mathcal{O}(\|\beta^\star\|_1)$. Finally, we conclude by using the continuity of the iterates (at a finite horizon) that this explicit bound can be used to tune the stepsizes.

Writing the mirror descent with varying potentials, we have, since $\nabla h_0(\beta_0) = 0$,

$$\nabla h_k(\beta_k) = -\sum_{\ell<k} \gamma_\ell \nabla \mathcal{L}(\beta_\ell)\,,$$

leading to, by convexity of h_k:

$$h_k(\beta_k) - h_k(\beta^\star) \leq \langle \nabla h_k(\beta_k), \beta_k - \beta^\star\rangle = -\sum_{\ell<k}\langle \gamma_\ell \nabla \mathcal{L}(\beta_\ell), \beta_k - \beta^\star\rangle\,.$$

We then write, using $\nabla\mathcal{L}(\beta) = H(\beta - \beta^\star)$ for $H = XX^\top$, that $-\sum_{\ell<k}\langle \gamma_\ell \nabla \mathcal{L}(\beta_\ell), \beta_k - \beta^\star\rangle = -\sum_{\ell<k}\gamma_\ell \langle X^\top(\bar\beta_k - \beta^\star), X^\top(\beta_k - \beta^\star)\rangle \leq \sum_{\ell<k}\gamma_\ell\sqrt{\mathcal{L}(\bar\beta_k)\mathcal{L}(\beta_k)}$, leading to:

$$h_k(\beta_k) - h_k(\beta^\star) \leq 2\sqrt{\sum_{\ell<k}\gamma_\ell \mathcal{L}(\bar\beta_k)\sum_{\ell<k}\gamma_\ell\mathcal{L}(\beta_k)} \leq 2\sum_{\ell<k}\gamma_\ell\mathcal{L}(\bar\beta_k) \leq 2D_{h_0}(\mathsf{w}^\star, \mathsf{w}^0)\,,$$

where the last inequality holds provided that $\gamma_k \leq \frac{1}{CLB}$. Thus,

$$\psi_{\boldsymbol{\alpha}_k}(\beta_k) \leq \psi_{\boldsymbol{\alpha}_k}(\beta^\star) + 2\psi_{\boldsymbol{\alpha}_0}(\beta^\star) + \langle \phi_k, \beta_k - \beta^\star\rangle\,.$$

Then, $\langle \phi_k, \beta_k - \beta^\star\rangle \leq \|\phi_k\|_1\|\beta_k - \beta^\star\|_\infty$ and $\|\phi_k\|_1 \leq C\lambda_{\max}\sum_{k<K}\gamma_k^2\mathcal{L}(\beta^k) \leq C\lambda_{\max}\gamma_{\max}h_0(\beta^\star)$. Then, using

$$\|\beta\|_\infty - \frac{1}{\ln(1/\alpha^2)} \leq \frac{\psi_\alpha(\beta)}{\ln(1/\alpha^2)} \leq \|\beta\|_1\big(1 + \frac{\ln(\|\beta\|_1 + \alpha^2)}{\ln(1/\alpha^2)}\big)\,,$$

we have:

$$\begin{aligned}\|\beta_k\|_\infty &\leq \frac{1}{\ln(1/\alpha^2)} + \|\beta^\star\|_1\big(1 + \frac{\ln(\|\beta^\star\|_1 + \alpha^2)}{\ln(1/\alpha^2)}\big) + \|\beta^\star\|_1\big(1 + \frac{\ln(\|\beta^\star\|_1 + \alpha^2)}{\ln(1/\alpha^2)}\big)\\ &\quad + B_0 C\lambda_{\max}\gamma_{\max}h_0(\beta^\star)/\ln(1/\alpha^2)\\ &\leq R + B_0 C\lambda_{\max}\gamma_{\max}h_0(\beta^\star)/\ln(1/\alpha^2)\,,\end{aligned}$$

where $R = \mathcal{O}(\|\beta^\star\|_1)$ is independent of α. Hence, since $B_0 = \sup_{k<\infty}\|\beta_k\|_\infty < \infty$, we have:

$$B_0(1 - C\lambda_{\max}\gamma_{\max}h_0(\beta^\star)/\ln(1/\alpha^2)) \leq R \implies B_0 \leq 2R\,,$$

provided that $\gamma_{\max} \leq 1/(2C\lambda_{\max}h_0(\beta^\star)/\ln(1/\alpha^2))$ (note that $h_0(\beta^\star)/\ln(1/\alpha^2)$ is independent of α^2).

Hence, if for all k we have $\gamma_k \leq \frac{1}{C'LB}$ where B bounds all $\|\beta_k\|_\infty$, we have $\|\beta_k\|_\infty \leq 2R$ for all k, where $R = \mathcal{O}(\|\beta^\star\|_1)$ is independent of α and stepsizes γ_k.

Let $K > 0$ be fixed, and

$$\bar{\gamma} = \inf \left\{ \gamma > 0 \quad \text{s.t.} \quad \sup_{k \leq K} \|\beta_k\|_\infty > 2R \right\}.$$

For $\gamma \geq 0$ a constant stepsize, let

$$\varphi(\gamma) = \sup_{k \leq K} \|\beta_k\|_\infty \,,$$

which is a continuous function of γ. For $r > 0$, let $\gamma_r = \frac{1}{C'Lr}$.

An important feature to notice is that if $\gamma < \gamma_r$ and r bounds all $\|\beta_k\|_\infty, k \leq K$, then $\varphi(\gamma) \leq R$, as shown above. We will show that we have $\bar{\gamma} \geq \gamma_{2R}$. Reasoning by contradiction, if $\bar{\gamma} < \gamma_{2R}$: by continuity of φ, we have $\varphi(\bar{\gamma}) \leq R$ and thus, there exists some small $0 < \varepsilon < \gamma_{2R} - \bar{\gamma}$ such that for all $\gamma \in [\bar{\gamma}, \bar{\gamma} + \varepsilon]$, we have $\varphi(\bar{\gamma}) \leq 2R$.

However, such γ's verify both $\varphi(\gamma) \leq 2R$ (since $\gamma \in [\bar{\gamma}, \bar{\gamma} + \varepsilon]$ and by definition of ε) and $\gamma \leq \gamma_{2R}$ (by definition of ε), and hence $\varphi(\gamma) \leq R$. This contradicts the infimum of $\bar{\gamma}$, and hence $\bar{\gamma} \geq \gamma_{2R}$. Thus, for $\gamma \leq \gamma_{2R} = \frac{1}{2C'LR}$, we have $\|\beta_k\|_\infty \leq R$. □

□

C.6 Proof of Theorems 1 and 2, and of Proposition 17

C.6.1 Proof of Theorems 1 and 2

We are now equipped to prove Theorem 1 and Theorem 2, condensed in the following Theorem.

Theorem 8. *Let $(u_k, v_k)_{k\geq 0}$ follow the mini-batch SGD recursion (8.3) initialised at $u_0 = \sqrt{2}\boldsymbol{\alpha} \in \mathbb{R}^d_{>0}$ and $v_0 = \mathbf{0}$, and let $(\beta_k)_{k\geq 0} = (u_k \odot v_k)_{k\geq 0}$. There exists and explicit $B > 0$ and a numerical constant $c > 0$ such that:*

1. *For stepsizes satisfying $\gamma_k \leq \frac{c}{LB}$, the iterates satisfy $\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leq 1$ and $\|\beta_k\|_\infty \leq B$ for all k;*
2. *For stepsizes satisfying $\gamma_k \leq \frac{c}{LB}$, $(\beta_k)_{k\geq 0}$ converges almost surely to some $\beta^\star_\infty \in \mathcal{S}$,*
3. *If $(\beta_k)_k$ and the neurons $(u_k, v_k)_k$ respectively converge to a model $\beta^\star_\infty$ and neurons (u_∞, v_∞) satisfying $\beta^\star_\infty \in \mathcal{S}$ (and $\beta^\star_\infty = u_\infty \odot v_\infty$), then for almost all stepsizes (with respect to the Lebesgue measure), the limit $\beta^\star_\infty$ satisfies:*

$$\beta^\star_\infty = \underset{\beta^\star \in \mathcal{S}}{\arg\min} \; D_{\psi_{\boldsymbol{\alpha}_\infty}}(\beta^\star, \tilde{\beta}_0) \,,$$

for $\boldsymbol{\alpha}_\infty \in \mathbb{R}^d_{>0}$ and $\tilde{\beta}_0 \in \mathbb{R}^d$ satisfying

$$\boldsymbol{\alpha}^2_\infty = \boldsymbol{\alpha}^2 \odot \exp\left(-\sum_{k=0}^{\infty} q\big(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\big) \right) ,$$

where $q(x) = -\frac{1}{2}\ln((1-x^2)^2) \geq 0$ for $|x| \leq \sqrt{2}$, and $\tilde{\beta}_0$ is a perturbation term equal to:

$$\tilde{\beta}_0 = \frac{1}{2}\big(\boldsymbol{\alpha}^2_+ - \boldsymbol{\alpha}^2_-\big),$$

where, $q_\pm(x) = \mp 2x - \ln((1 \mp x)^2)$, and $\boldsymbol{\alpha}^2_\pm = \boldsymbol{\alpha}^2 \odot \exp\left(-\sum_{k=0}^{\infty} q_\pm(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))\right)$.

Proof. **Point 1.** The first point of the Theorem is a direct consequence of Corollary 5 and the bounds proved in Appendix C.5.3.

Point 2. Then, for stepsizes $\gamma_k \leq \frac{c}{LB}$, using Proposition 35 for any interpolator $\beta^\star \in \mathcal{S}$:

$$D_{h_{k+1}}(\mathsf{w}^\star, \mathsf{w}_{k+1}) \leq D_{h_k}(\mathsf{w}^\star, \mathsf{w}_k) - \gamma_k \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\,. \tag{C.11}$$

Hence, summing:

$$\sum_k \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) \leq D_{h_0}(\beta^\star, \beta_0)\,,$$

so that the series converges.

Under our stepsize rule, $\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leq \frac{1}{2}$, leading to $\|q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leq 3\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty^2$ by Lemma 25. Using $\|\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|^2 \leq 2L_2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)$, we have that $\ln(\boldsymbol{\alpha}_{\pm,k})$, $\ln(\boldsymbol{\alpha}_k)$ all converge.

We now show that $\sum_k \gamma_k \mathcal{L}(\beta_k) < \infty$. We have:

$$\sum_{\ell<k} \mathcal{L}(\beta_k) = \sum_{\ell<k} \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) + M_k\,,$$

where $M_k = \sum_{\ell<k} \gamma_k(\mathcal{L}(\beta_k) - \mathcal{L}_{\mathcal{B}_k}(\beta_k))$. We have that (M_k) is a martingale with respect to the filtration $(\mathcal{F}_k)$ defined as $\mathcal{F}_k = \sigma(\beta_\ell, \ell \leq k)$. Using our upper-bound on $\sum_{\ell<k} \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k)$, we have:

$$M_k \geq \sum_{\ell<k} \gamma_k \mathcal{L}(\beta_k) - \sum_{\ell<k} \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) \geq -D_{h_0}(\beta^\star, \beta_0)\,,$$

and hence (M_k) is a lower bounded martingale. Using Doob's first martingale convergence theorem (a lower bounded super-martingale converges almost surely [Doob, 1990]), (M_k) converges almost surely. Consequently, since $\sum_{\ell<k} \gamma_k \mathcal{L}(\beta_k) = \sum_{\ell<k} \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) + M_k$, we have that $\sum_{\ell<k} \gamma_k \mathcal{L}(\beta_k)$ converges almost surely (the first term is upper bounded, the second converges almost surely).

We now prove the convergence of (w_k). Since it is a bounded sequence, let $\mathsf{w}_{\sigma(k)}$ be a convergent sub-sequence and let $\mathsf{w}^\star_\infty$ denote its limit: $\mathsf{w}_{\sigma(k)} \to \beta^\star_\infty$.

Almost surely, $\sum_k \gamma_k \mathcal{L}(\beta_k) < \infty$ and so $\gamma_k \mathcal{L}(\beta_k) \to 0$, leading to $\mathcal{L}(\beta_k) \to 0$ since the stepsizes are lower bounded, so that $\mathcal{L}(\beta_{\sigma(k)}) \to 0$, and hence $\mathcal{L}(\beta^\star_\infty) = 0$: this means that $\beta^\star_\infty$ is an interpolator.

Since the quantities $(\boldsymbol{\alpha}_k)_k$, $(\boldsymbol{\alpha}_{\pm,k})_k$ and $(\phi_k)_k$ converge almost surely to $\boldsymbol{\alpha}_\infty$, $\boldsymbol{\alpha}_\pm$ and ϕ_∞, we get that the potentials h_k uniformly converge to $h_\infty = \psi_{\boldsymbol{\alpha}_\infty} - \langle \phi_\infty, \cdot \rangle$ on all compact sets. Now notice that we can decompose $\nabla h_\infty(\beta^\star_\infty)$ as:

$$\nabla h_\infty(\beta^\star_\infty) = \big(\nabla h_\infty(\beta^\star_\infty) - \nabla h_\infty(\mathsf{w}_{\sigma(k)})\big) + \big(\nabla h_\infty(\mathsf{w}_{\sigma(k)}) - \nabla h_{\sigma(k)}(\mathsf{w}_{\sigma(k)})\big) + \nabla h_{\sigma(k)}(\mathsf{w}_{\sigma(k)}).$$

The first two terms converge to 0: the first is a direct consequence of the convergence of the extracted subsequence, the second is a consequence of the uniform convergence of $h_{\sigma(k)}$ to h_∞ on compact sets. Finally the last term is always in $\mathrm{span}(x_1, \ldots, x_n)$ due to Proposition 32, leading to $\nabla h_\infty(\beta^\star_\infty) \in \mathrm{span}(x_1, \ldots, x_n)$. Consequently, $\nabla h_\infty(\beta^\star_\infty) \in \mathrm{span}(x_1, \ldots, x_n)$. Notice that from the definition of h_∞, we have that $\nabla h_\infty(\beta^\star_\infty) = \nabla \psi_{\boldsymbol{\alpha}_\infty}(\beta^\star_\infty) - \phi_\infty$. Now since $\phi_\infty = \frac{1}{2}\,\mathrm{arcsinh}(\frac{\boldsymbol{\alpha}_+^2 - \alpha_-^2}{2\boldsymbol{\alpha}_\infty^2})$, one can notice that $\tilde{\beta}_0$ is precisely defined such that $\nabla \psi_{\alpha_\infty}(\tilde{\beta}_0) = \phi_\infty$. Therefore $\nabla \psi_{\boldsymbol{\alpha}_\infty}(\beta^\star_\infty) - \nabla \psi_{\boldsymbol{\alpha}_\infty}(\tilde{\beta}_0) \in \mathrm{span}(x_1, \ldots, x_n)$. This condition along with the fact that $\mathsf{w}^\star_\infty$ is an interpolator are exactly the optimality conditions of the convex minimisation problem:

$$\min_{\beta^\star \in \mathcal{S}} D_{\psi_{\boldsymbol{\alpha}_\infty}}(\beta^\star, \tilde{\beta}_0)$$

Therefore $\beta^\star_\infty$ must be equal to the unique minimiser of this problem. Since this is true for any sub-sequence we get that w_k converges almost surely to:

$$\beta^\star_\infty = \arg\min_{\beta \in \mathcal{S}} \quad D_{\psi_{\boldsymbol{\alpha}_\infty}}(\beta^\star, \tilde{\beta}_0).$$

Point 3. From what we just proved, note that it is sufficient to prove that $\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{\pm,k}, \phi_k$ converge to limits $\boldsymbol{\alpha}_\infty, \boldsymbol{\alpha}_{\pm,\infty}, \phi_\infty$ satisfying $\boldsymbol{\alpha}_\infty, \boldsymbol{\alpha}_{\pm,\infty} \in \mathbb{R}^d_{>0}$ (with positive and non-null coordinates) and $\phi_\infty \in \mathbb{R}^d$. Indeed, if this holds and since we assume that the iterates converge to some interpolator, we proved just above that this interpolator is uniquely defined through the desired implicit regularization problem. We thus prove the convergence of $\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{\pm,k}, \phi_k$.

Note that the convergence of u_k, v_k is equivalent to the convergence of $w_{\pm,k}$ in the $w_+^2 - w_-^2$ parameterisation used in our proofs, that we use there too. We have:

$$w_{\pm,k+1} = (1 \mp \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)) \odot w_{\pm,k}\,,$$

so that

$$\ln(w^2_{\pm,k}) = \sum_{\ell<k} \ln((1 \mp \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))^2)\,.$$

We now assume that stepsizes are such that for all $\ell \geq 0$ and $i \in [d]$, stepsizes are such that we have $|\gamma_\ell \nabla_i \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)| \neq 1$: this is true for all stepsizes except a countable number of stepsizes, and so this is true for almost all stepsizes. Since we assume that the iterates β_k converge to some interpolator, this leads to $\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell) \to 0$ if we assume that stepsizes do not diverge.

Taking the limit, we have

$$\ln(w^2_{\pm,\infty}) = \sum_{\ell<\infty} \ln((1 \mp \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))^2)\,.$$

This limit is in $(\{-\infty\} \cup \mathbb{R})^d$ (since $w_{\pm,\infty} \in \mathbb{R}^d$), and a coordinate of the limit is equal to $-\infty$ if and only if the sum on the RHS diverges to $-\infty$ (note that from our assumption just above, no term of the sum can be equal to $-\infty$).

We have $\ln((1\mp\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))^2) \sim \mp 2\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)$ as $\ell \to \infty$, so that if for some coordinate i we have $\sum_\ell \gamma_\ell \nabla_i \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell) = \mp\infty$, then the coordinate i of the limit satisfies $\ln(w^2_{i,\pm,\infty}) = +\infty$, which is impossible. Hence, the sum $\sum_\ell \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)$ is in $\mathbb{R}^d$ (and is thus converging); consequently, $\sum_\ell \gamma_\ell^2 \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)^2$ converges and thus $\sum_\ell q(\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))$ and $\sum_\ell q_\pm(\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))$ all converge: the sequences $\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{\pm,k}$ thus converge to limits in $\mathbb{R}^d_{>0}$, and ϕ_k converges, concluding our proof. □

C.6.2 Proof of Proposition 17

We begin with the following Lemma, that explicits the curvature of D_h around the set of interpolators.

Lemma 23. *For all $k \geq 0$, if $\mathcal{L}(\beta_k) \leq \frac{1}{2\lambda_{\max}}(\alpha^2 \lambda^+_{\min})^2$, we have $\|\beta_k - \beta^\star_{\alpha_k}\|^2 \leq 2B(\alpha^2\lambda^+_{\min})^{-1}\mathcal{L}(\beta_k)$.*

Proof. Recall that the sequence $\mathbf{z}^k = \nabla h_k(\mathsf{w}^k)$ satisfies $\mathbf{z}^0 = 0$ and $\mathbf{z}^{k+1} = \mathbf{z}^k - \gamma_k \mathcal{L}(\mathsf{w}^k)$, so that we have that $\mathbf{z}^k \in V = \mathrm{Im}(\mathbf{X}\mathbf{X}^\top)$ for all $k \geq 0$. Then, let w^α_k be the unique minimizer of h_k over $\mathcal{S}$ the space of interpolators: w^α_k is exactly characterized by $\mathbf{X}^\top \mathsf{w}^\alpha_k = \mathbf{Y}$ and $\nabla h_k(\mathsf{w}^\alpha_k) \in V$. We define $\mathbf{z}^\alpha_k \in V$ as $\mathbf{z}^\alpha_k = \nabla h_k(\mathsf{w}^\alpha_k)$.

Now, fix $\mathbf{z}^\alpha = \mathbf{z}^\alpha_k$ and $h = h_k$, and let us define $\psi : \mathbf{z} \in V \to D_{h^*}(\mathbf{z}, \mathbf{z}^\alpha)$ and $\phi : \mathbf{z} \in V \to \mathcal{L}(\nabla h^*(\mathbf{z}))$. We next show that for all $\mathbf{z} \in V$, there exists μ_z such that $\nabla^2\phi(\mathbf{z}) \geq \mu_z \nabla^2\psi(\mathbf{z})$,

and that $\mu_z \geq \mu$ for $\mathbf{z}$ in an open convex set of V around $\mathbf{z}^\alpha$, for some $\mu > 0$. For $A \in \mathbb{R}^{d\times d}$ an operator/matrix on $\mathbb{R}^d$, let us denote A_V its restriction/co-restriction to V.

First, for $\mathbf{z} \in V$, we have $\nabla^2\psi(\mathbf{z}) = \nabla^2(h^*(\mathbf{z}) - h^*(\mathbf{z}) - \langle \nabla h^*(\mathbf{z}^\alpha), z - z^\alpha\rangle)(\mathbf{z}) = \nabla^2 h^*(\mathbf{z})_V$. Then, $\nabla\phi(\mathbf{z}) = \nabla^2 h^*(\mathbf{z})\nabla\mathcal{L}(\nabla h^*(\mathbf{z}))$, so that $\nabla^2\phi(\mathbf{z}) = \big(\nabla^2 h^*(\mathbf{z})\nabla^2\mathcal{L}(\nabla h^*(\mathbf{z}))\nabla^2 h^*(\mathbf{z})\big)_V + \nabla^3 h^*(\mathbf{z})(\nabla\mathcal{L}(\nabla h^*(\mathbf{z})), \cdot, \cdot)_V$.

Since h is $1/(2\alpha^2)$ smooth (on $\mathbb{R}^d$ and thus on V), h^* is $2\alpha^2$ strongly convex (on V and on $\mathbb{R}^d$). Using $V = \mathrm{Im}(\mathbf{X}\mathbf{X}^\top)$ and $\nabla^2\mathcal{L} \equiv \mathbf{X}\mathbf{X}^\top$, we have $\big(\nabla^2 h^*(\mathbf{z})\nabla^2\mathcal{L}(\nabla h^*(\mathbf{z}))\nabla^2 h^*(\mathbf{z})\big)_V = \nabla^2 h^*(\mathbf{z})_V\nabla^2\mathcal{L}(\nabla h^*(\mathbf{z}))_V\nabla^2 h^*(\mathbf{z})_V$, and thus $\big(\nabla^2 h^*(\mathbf{z})\nabla^2\mathcal{L}(\nabla h^*(\mathbf{z}))\nabla^2 h^*(\mathbf{z})\big)_V \succeq 2\alpha^2\lambda^+_{\min}\nabla^2 h^*(\mathbf{z})_V$.

For the other term of $\nabla^2\phi$, namely $\nabla^3 h^*(\mathbf{z})(\nabla\mathcal{L}(\nabla h^*(\mathbf{z})), \cdot, \cdot)_V$, we compute $\nabla^3_{ijk} h^*(\mathbf{z}) = \mathbf{1}_{i=j=k} 2\alpha^2_{i,k}\sinh(\mathbf{z}_i)$, leading to: $\nabla^3 h^*(\mathbf{z})(\nabla\mathcal{L}(\nabla h^*(\mathbf{z})), \cdot, \cdot)_V = \mathrm{diag}(2\alpha^2\sinh(\mathbf{z})\odot(\mathbf{X}\mathbf{X}^\top(2\alpha^2\sinh(\mathbf{z}) - \mathsf{w}^\alpha)))_V$. Thus, writing $\mathsf{w}_\mathbf{z} = 2\alpha^2_{i,k}\sinh(\mathbf{z}) = \nabla h^*(\mathbf{z})$ the primal surrogate of $\mathbf{z}$, we have:

$$\begin{aligned}
\nabla^3 h^*(\mathbf{z})(\nabla\mathcal{L}(\nabla h^*(\mathbf{z})), \cdot, \cdot)_V &= \mathrm{diag}(2\alpha^2_{i,k}\sinh(\mathbf{z}) \odot (\mathbf{X}\mathbf{X}^\top(\mathsf{w}_\mathbf{z} - \mathsf{w}^\alpha_k)))_V \\
&\succeq -\|\mathbf{X}\mathbf{X}^\top(\mathsf{w}_\mathbf{z} - \mathsf{w}^\alpha_k)\|_\infty \,\mathrm{diag}(2\alpha^2_k \odot |\sinh(\mathbf{z})|)_V \\
&\succeq -\|\mathbf{X}\mathbf{X}^\top(\mathsf{w}_\mathbf{z} - \mathsf{w}^\alpha_k)\|_\infty \,\mathrm{diag}(2\alpha^2_k \odot \cosh(\mathbf{z}))_V \\
&= -\|\mathbf{X}\mathbf{X}^\top(\mathsf{w}_\mathbf{z} - \mathsf{w}^\alpha_k)\|_\infty \nabla^2\psi(\mathbf{z})\,.
\end{aligned}$$

Wrapping things together,

$$\nabla^2\phi(\mathbf{z}) \succeq \big(2\alpha^2\lambda^+_{\min} - \|\mathbf{X}\mathbf{X}^\top(\mathsf{w}_\mathbf{z} - \mathsf{w}^\alpha)\|_\infty\big)\nabla^2\psi(\mathbf{z})\,.$$

Let $\mathcal{Z} = \big\{\mathbf{z} \in V : \|\mathbf{X}\mathbf{X}^\top(\mathsf{w}_\mathbf{z} - \mathsf{w}^\alpha_k)\|_\infty < \alpha^2\lambda^+_{\min}\big\}$ that satisfies $\Big\{\mathsf{w} \in V : \mathcal{L}(\mathsf{w}_\mathbf{z}) < \frac{1}{2\lambda_{\max}}(\alpha^2\lambda^+_{\min})^2\Big\} \subset \mathcal{Z}$. $\mathcal{Z}$ is an open convex set of V containing $\mathbf{z}^\alpha$. On $\mathcal{Z}$, $\nabla^2\phi \succeq \alpha^2\lambda^+_{\min}\nabla^2\psi$, and $\psi(\mathbf{z}^\alpha) = \phi(\mathbf{z}^\alpha) = 0$, so that for all $\mathbf{z} \in \mathcal{Z}$, we have $\phi(\mathbf{z}) \geq \alpha^2\lambda^+_{\min}\psi(\mathbf{z})$. Hence, for all $\mathbf{z} \in \mathcal{Z}$, we have $D_{h_k}(\beta^\alpha_k, \beta_\mathbf{z}) \leq D_{h^\star}(\mathbf{z}, \mathbf{z}^\alpha) \leq (\alpha^2\lambda^+_{\min})^{-1}\mathcal{L}(\beta_\mathbf{z})$, and using the fact that D_{h_k} is $\frac{1}{4B}$ strongly convex, we obtain, for $\beta_\mathbf{z} = \beta_k$ (since $\mathbf{z}^k \in V$): if $\mathcal{L}(\beta_k) \leq \frac{1}{2\lambda_{\max}}(\alpha^2\lambda^+_{\min})^2$, we have $\|\beta^\alpha_k - \beta_k\|_2^2 \leq (\alpha^2\lambda^+_{\min})^{-1}\mathcal{L}(\beta_k)$. □

Proposition 41. *As assume $\mathcal{L}$ is L_r-relatively smooth with respect to all the h_k's. Then for all β we have the following inequality.*

$$\begin{aligned}
\gamma_k(\mathcal{L}(\beta_{k+1}) - \mathcal{L}(\beta)) \leq\ & D_{h_k}(\beta, \beta_k) - D_{h_{k+1}}(\beta, \beta_{k+1}) - (1 - \gamma_k L_r)D_{h_k}(\beta_{k+1}, \beta_k) \\
& + (h_{k+1} - h_k)(\beta) - (h_{k+1} - h_k)(\beta_{k+1})\,.
\end{aligned}$$

Proof. For any $\beta, \beta_k, \beta_{k+1}$, the following holds (three points identity for time varying potentials, Proposition 36):

$$\begin{aligned}
D_{h_k}(\beta, \beta_k) - D_{h_{k+1}}(\beta, \beta_{k+1}) &= \big[h_k(\beta) - (h_k(\beta_k) + \langle \nabla h_k(\beta_k), \beta - \beta_k\rangle)\big] \\
&\quad - \big[h_{k+1}(\beta) - (h_{k+1}(\beta_{k+1}) + \langle \nabla h_{k+1}(\beta_{k+1}), \beta - \beta_{k+1}\rangle)\big] \\
&= h_k(\beta) - h_{k+1}(\beta) + \langle \nabla h_{k+1}(\beta_{k+1}) - \nabla h_k(\beta_k), \beta - \beta_{k+1}\rangle \\
&\quad + h_{k+1}(\beta_{k+1}) - \big[h_k(\beta_k) + \langle \nabla h_k(\beta_k), \beta_{k+1} - \beta_k\rangle\big] \\
&= h_k(\beta) - h_{k+1}(\beta) + \langle \nabla h_{k+1}(\beta_{k+1}) - \nabla h_k(\beta_k), \beta - \beta_{k+1}\rangle \\
&\quad + h_{k+1}(\beta_{k+1}) - h_k(\beta_{k+1}) + D_{h_k}(\beta_{k+1}, \beta_k).
\end{aligned}$$

Rearranging and plugging in our mirror update we obtain that for all β:

$$\begin{aligned}
\gamma_k\langle \nabla\mathcal{L}(\beta_k), \beta_{k+1} - \beta\rangle = & D_{h_k}(\beta, \beta_k) - D_{h_{k+1}}(\beta, \beta_{k+1}) \\
& - D_{h_k}(\beta_{k+1}, \beta_k) - (h_{k+1} - h_k)(\beta_{k+1}) + (h_{k+1} - h_k)(\beta).
\end{aligned}$$

From the convexity of $\mathcal{L}$ and its L_r-relative smoothness we also have that:

$$\mathcal{L}(\beta_{k+1}) \leq \mathcal{L}(\beta) + \langle \nabla \mathcal{L}(\beta_k), \beta_{k+1} - \beta \rangle + L_r D_{h_k}(\beta_{k+1}, \beta_k),$$

Finally:

$$\begin{aligned}\gamma_k(\mathcal{L}(\beta_{k+1}) - \mathcal{L}(\beta)) \leq D_{h_k}(\beta, \beta_k) - D_{h_{k+1}}(\beta, \beta_{k+1}) - (1 - \gamma_k L_r) D_{h_k}(\beta_{k+1}, \beta_k) \\ + (h_{k+1} - h_k)(\beta) - (h_{k+1} - h_k)(\beta_{k+1}).\end{aligned}$$

Note that in our setting, for any β, $k \mapsto h_k(\beta)$ is **increasing**. We can therefore write that:

$$\gamma_k(\mathcal{L}(\beta_{k+1}) - \mathcal{L}(\beta)) \leq D_{h_k}(\beta, \beta_k) - D_{h_{k+1}}(\beta, \beta_{k+1}) - (1 - \gamma_k L_r) D_{h_k}(\beta_{k+1}, \beta_k) + (h_{k+1} - h_k)(\beta).$$

In particular, for $\beta = \beta^*$:

$$\begin{aligned}\gamma_k \mathcal{L}(\beta_{k+1}) &\leq D_{h_k}(\beta^*, \beta_k) - D_{h_{k+1}}(\beta^*, \beta_{k+1}) - (1 - \gamma_k L) D_{h_k}(\beta_{k+1}, \beta_k) + (h_{k+1} - h_k)(\beta^*) \\ &\quad - (h_{k+1} - h_k)(\beta_{k+1}) \\ &\leq D_{h_k}(\beta^*, \beta_k) - D_{h_{k+1}}(\beta^*, \beta_{k+1}) - (1 - \gamma_k L_r) D_{h_k}(\beta_{k+1}, \beta_k) + (h_{k+1} - h_k)(\beta^*)\end{aligned}$$

and in $\beta = \beta_k$:

$$\begin{aligned}\gamma_k \mathcal{L}(\beta_{k+1}) &\leq \gamma_k \mathcal{L}(\beta_k) - D_{h_{k+1}}(\beta_k, \beta_{k+1}) - (1 - \gamma_k L_r) D_{h_k}(\beta_{k+1}, \beta_k) + (h_{k+1} - h_k)(\beta_k) \\ &\quad - (h_{k+1} - h_k)(\beta_{k+1}) \\ &\leq \gamma_k \mathcal{L}(\beta_k) - D_{h_{k+1}}(\beta_k, \beta_{k+1}) - (1 - \gamma_k L_r) D_{h_k}(\beta_{k+1}, \beta_k) + (h_{k+1} - h_k)(\beta_k)\end{aligned}$$

□

Proof of Proposition 17. We apply Proposition 41 for $\beta = \beta_k$, with $L_r = 4BL$ (using Lemma 26) and replacing $\mathcal{L}$ by $\mathcal{L}_{\mathcal{B}_k}$, to obtain:

$$\begin{aligned}\gamma_k(\mathcal{L}_{\mathcal{B}_k}(\beta_{k+1}) - \mathcal{L}_{\mathcal{B}_k}(\beta_k)) \leq -D_{h_{k+1}}(\beta_k, \beta_{k+1}) - (1 - \gamma_k L_r) D_{h_k}(\beta_{k+1}, \beta_k) \\ + (h_{k+1} - h_k)(\beta_k) - (h_{k+1} - h_k)(\beta_{k+1}),\end{aligned}$$

and thus, taking the mean wrt $\mathcal{B}_k$,

$$\begin{aligned}\gamma_k(\mathbb{E}_{\mathcal{B}_k}\mathcal{L}(\beta_{k+1}) - \mathcal{L}(\beta_k)) &\leq -\mathbb{E}_{\mathcal{B}_k} D_{h_{k+1}}(\beta_k, \beta_{k+1}) - (1 - \gamma_k L_r)\mathbb{E}_{\mathcal{B}_k} D_{h_k}(\beta_{k+1}, \beta_k) \\ &\quad + \mathbb{E}_{\mathcal{B}_k}(h_{k+1} - h_k)(\beta_k) - \mathbb{E}_{\mathcal{B}_k}(h_{k+1} - h_k)(\beta_{k+1}) \\ &\leq -(1 - \gamma_k L_r)\mathbb{E}_{\mathcal{B}_k} D_{h_k}(\beta_{k+1}, \beta_k) \\ &\quad + \mathbb{E}_{\mathcal{B}_k}(h_{k+1} - h_k)(\beta_k) - \mathbb{E}_{\mathcal{B}_k}(h_{k+1} - h_k)(\beta_{k+1}).\end{aligned}$$

First, as in the proof of Proposition 37, using the fact that h_k is $\ln(1/\alpha_k)$ smooth,

$$\begin{aligned}D_{h_k}(\mathsf{w}_{k+1}, \beta_k) &\geq \frac{1}{2\ln(1/\alpha_k)} \|\nabla h_k(\mathsf{w}_k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k) - \nabla h_k(\mathsf{w}_k) + \nabla h_{k+1}(\mathsf{w}_{k+1}) - \nabla h_k(\mathsf{w}_{k+1})\|_2^2 \\ &\geq -\frac{1}{2\ln(1/\alpha_k)} \|\nabla h_k(\mathsf{w}_k) - \nabla h_{k+1}(\mathsf{w}_k)\|_2^2 + \frac{1}{4\ln(1/\alpha_k)} \|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k)\|_2^2,\end{aligned}$$

and thus

$$\mathbb{E} D_{h_k}(\mathsf{w}_{k+1}, \beta_k) \geq \mathbb{E}\left[-\frac{1}{2\ln(1/\alpha_k)} \|\nabla h_k(\mathsf{w}_k) - \nabla h_{k+1}(\mathsf{w}_k)\|_2^2 + \frac{\lambda_b}{2\ln(1/\alpha_k)} \gamma_k^2 \mathcal{L}_{\mathcal{B}}(\mathsf{w}_k) \right].$$

Now, we apply Lemma 27 assuming that $\|\beta^\star\|_\infty, \|\beta_{k+1}\|_\infty \leq B$ (which is satisfied since we are under the assumption of Theorem 2):

$$(h_{k+1} - h_k)(\beta_k) - (h_{k+1} - h_k)(\beta^\star) \leq 24BL\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,.$$

Using $|\nabla h_k(\mathsf{w}) - \nabla h_{k+1}(\mathsf{w})| \leq 2\delta_k$ where $\delta_k = q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$ as in Proposition 37, we have:

$$\mathbb{E}\|\nabla h_k(\mathsf{w}_k) - \nabla h_{k+1}(\mathsf{w}_k)\|_2^2 \leq 16B\gamma_k^2 \mathbb{E}\|\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|^2 \leq 32BL\gamma_k^2 \mathbb{E}\mathcal{L}(\beta_k)\,.$$

Wrapping everything together,

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\beta_{k+1}) - \mathcal{L}(\beta_k)\right] \leq &-(1 - \gamma_k 4BL)\frac{\lambda_b}{2\ln(1/\alpha_k)}\gamma_k \mathbb{E}\mathcal{L}(\beta_k) \\ &+ \big(\gamma_k^2(1 - 4\gamma_k BL)24BL + \frac{32BL}{\ln(1/\alpha_k)}\big)\gamma_k^2 \mathbb{E}\mathcal{L}(\beta_k)\,.\end{aligned}$$

Thus, for $\gamma_k \leq \frac{c'}{LB\ln(1/(\min_i \alpha_{k,i}))}$, we have the first part of Proposition 17.

Using Lemma 23, we then have:

$$\begin{aligned}\mathbb{E}\left[\|\beta_k - \beta^\star_{\alpha_k}\|^2\right] &= \mathbb{E}\left[\mathbf{1}_{\left\{\mathcal{L}(\beta_k) \leq \frac{1}{2\lambda_{\max}}(\alpha^2\lambda^+_{\min})^2\right\}}\|\beta_k - \beta^\star_{\alpha_k}\|^2\right] \\ &\quad + \mathbb{E}\left[\mathbf{1}_{\left\{\mathcal{L}(\beta_k) > \frac{1}{2\lambda_{\max}}(\alpha^2\lambda^+_{\min})^2\right\}}\|\beta_k - \beta^\star_{\alpha_k}\|^2\right] \\ &\leq \mathbb{E}\left[\mathbf{1}_{\left\{\mathcal{L}(\beta_k) \leq \frac{1}{2\lambda_{\max}}(\alpha^2\lambda^+_{\min})^2\right\}} 2B(\alpha^2\lambda^+_{\min})^{-1}\mathcal{L}(\beta_k)\right] \\ &\quad + \mathbb{P}\left(\mathcal{L}(\beta_k) > \frac{1}{2\lambda_{\max}}(\alpha^2\lambda^+_{\min})^2\right) \times 4B^2 \\ &\leq 2B(\alpha^2\lambda^+_{\min})^{-1}\mathbb{E}\left[\mathcal{L}(\beta_k)\right] \\ &\quad + \frac{\mathbb{E}\left[\mathcal{L}(\beta_k)\right]}{\frac{1}{2\lambda_{\max}}(\alpha^2\lambda^+_{\min})^2} \times 4B^2 \\ &= 2B(\alpha^2\lambda^+_{\min})^{-1}\Big(1 + \frac{4B\lambda_{\max}}{\alpha^2\lambda^+_{\min}}\Big)\mathbb{E}\left[\mathcal{L}(\beta_k)\right]\,.\end{aligned}$$

□

C.7 Proof of miscellaneous results mentioned in the main text

In this section, we provide proofs for results mentioned in the main text and that are not directly directed to the proof of Theorem 8.

C.7.1 Proof of Proposition 8.5.1 and the sum of the losses

We start by proving the following proposition. We then continue with upper and lower bounds (of similar magnitude) on the sum of the losses.

Let $\Lambda_b, \lambda_b > 0$ [2] be the largest and smallest values, respectively, such that $\lambda_b H \preceq \mathbb{E}_{\mathcal{B}}\big[H^2_{\mathcal{B}}\big] \preceq \Lambda_b H$. For any stepsize $\gamma > 0$ satisfying $\gamma \leq \frac{c}{BL}$ (as in theorem 2), initialisation $\alpha\mathbf{1}$ and batch size $b \in [n]$, the magnitude of the gain satisfies:

$$\lambda_b \gamma^2 \sum_k \mathbb{E}\mathcal{L}(\beta_k) \leq \mathbb{E}\left[\|\mathrm{Gain}_\gamma\|_1\right] \leq 2\Lambda_b \gamma^2 \sum_k \mathbb{E}\mathcal{L}(\beta_k)\,, \tag{8.10}$$

where the expectation is over a uniform and independent sampling of the batches $(\mathcal{B}_k)_{k\geq 0}$.

Proof. From Lemma 25, for all $-1/2 \leq x \leq 1/2$, it holds that $x^2 \leq q(x) \leq 2x^2$. We have, using $\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leq 1/2$ (which holds under the stepsize assumption):

$$\begin{aligned}
\mathbb{E}\|\mathrm{Gain}_\gamma\|_1 &= -\mathbb{E}\sum_i \ln\left(\frac{\alpha_{\infty,i}}{\alpha}\right) \\
&= \sum_{\ell<\infty}\sum_i \mathbb{E}q\big(\gamma_\ell \nabla_i \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\big) \\
&\leq 2\sum_{\ell<\infty}\sum_i \mathbb{E}\big(\gamma_\ell \nabla_i \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\big)^2 \\
&= \sum_{\ell<\infty} \gamma_\ell^2 \mathbb{E}\|\nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\|_2^2 \\
&\leq 4\Lambda_b \sum_{\ell<\infty} \gamma_\ell^2 \mathbb{E}\mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\,,
\end{aligned}$$

since $\mathbb{E}\|\nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\|_2^2 \leq 2\Lambda_b \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)$. For the left handside we use $q(x) \geq x^2$ for $|x| \leq 1/2$ and $\mathbb{E}\|\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_\ell)\|_2^2 \geq 2\lambda_b \mathcal{L}_{\mathcal{B}_k}(\beta_\ell)$. Finally, since $\mathcal{B}_\ell$ independent freom β_ℓ, we have $\mathbb{E}\mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell) = \mathbb{E}\mathcal{L}(\beta_\ell)$. □

Proposition 42. *For stepsizes $\gamma_k \equiv \gamma \leq \frac{c}{LB}$ (as in Theorem 2), we have:*

$$\sum_{k\geq 0} \gamma^2 \mathbb{E}\mathcal{L}(\beta_k) = \Theta\left(\gamma \|\beta^\star\|_1 \ln(1/\alpha)\right)\,.$$

Proof. We first lower bound $\sum_{k<\infty} \gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)$. We have the following equality, that holds for any k:

$$\begin{aligned}
D_{h_{k+1}}(\beta^\star, \beta_{k+1}) = D_{h_k}(\beta^\star, \beta_k) - 2\gamma \mathcal{L}_{\mathcal{B}_k}(\beta_k) + D_{h_{k+1}}(\beta_k, \beta_{k+1}) \\
+ \big(h_k - h_{k+1}\big)(\beta_k) - \big(h_k - h_{k+1}\big)(\beta^\star)\,,
\end{aligned}$$

leading to, by summing for $k \in \mathbb{N}$:

$$\sum_{k<\infty} 2\gamma \mathcal{L}_{\mathcal{B}_k}(\beta_k) = D_{h_0}(\beta^\star, \beta_0) - \lim_{k\to\infty} D_{h_k}(\beta^\star, \beta_k) + \sum_{k<\infty} D_{h_{k+1}}(\beta_k, \beta_{k+1}) + \sum_{k<\infty} \big(h_k - h_{k+1}\big)(\beta_k) - \big(h_k - h_{k+1}\big)(\beta^\star)\,.$$

[2] $\Lambda_b, \lambda_b > 0$ are data-dependent constants; for $b = n$, we have $(\lambda_n, \Lambda_n) = (\lambda^+_{\min}(H), \lambda_{\max}(H))$ where $\lambda^+_{\min}(H)$ is the smallest non-null eigenvalue of H; for $b = 1$, we have $\min_i \|x_i\|_2^2 \leq \lambda_1 \leq \Lambda_1 \leq \max_i \|x_i\|_2^2$.

First, since $h_k \to h_\infty$, $\beta_k \to \beta_\infty$, we have $\lim_{k\to\infty} D_{h_k}(\beta^\star, \beta_k) = 0$. Then, $D_{h_{k+1}}(\beta_k, \beta_{k+1}) \geq 0$. Finally, $|\big(h_k - h_{k+1}\big)(\beta_k) - \big(h_k - h_{k+1}\big)(\beta^\star)| \leq 16BL_2\gamma^2\mathcal{L}_{\mathcal{B}_k}(\beta_k)$. Hence :

$$\sum_{k<\infty} 2\gamma(1 + 16\gamma BL_2)\mathcal{L}_{\mathcal{B}_k}(\beta_k) \geq D_{h_0}(\beta^\star, \beta_0)\,,$$

and thus $\sum_{k<\infty}\gamma\mathcal{L}_{\mathcal{B}_k}(\beta_k) \geq D_{h_0}(\beta^\star, \beta_0)/4$ for $\gamma \leq c/(BL)$ (with $c \geq 16$). This gives the RHS inequality. The LHS is a direct consequence of bounds proved in previous subsections.

Hence, we have that

$$\gamma^2 \sum_k \mathcal{L}(\beta_k) = \Theta\left(\gamma D_{h_0}(\beta^\star, \beta_0)\right)\,.$$

Noting that $D_{h_0}(\beta^\star, \beta_0) = h_0(\beta^\star) = \Theta\big(\ln(1/\alpha)\|\beta^\star\|_1\big)$ concludes the proof. □

C.7.2 $\tilde{\beta}_0$ is negligible

In the following proposition we show that $\tilde{\beta}_0$ is close to $\mathbf{0}$ and therefore one should think of the implicit regularization problem as $\beta^\star_\infty = \arg\min_{\beta^\star \in S} \psi_{\alpha_\infty}(\beta^\star)$

Proposition 43. *Under the assumptions of Theorem 2,*

$$|\tilde{\beta}_0| \leq \alpha^2,$$

where the inequality must be understood coordinate-wise.

Proof.

$$\begin{aligned}
|\tilde{\beta}_0| &= \frac{1}{2}|\alpha_+^2 - \alpha_-^2| \\
&= \frac{1}{2}\alpha^2\big|\exp(-\sum_k q_+(\gamma_k\nabla\mathcal{L}(\beta_k)) - \exp(-\sum_k q_-(\gamma_k\nabla\mathcal{L}(\beta_k))\big| \\
&\leq \alpha^2,
\end{aligned}$$

where the inequality is because $q_+(\gamma_k\nabla\mathcal{L}(\beta_k)) \geq 0$, $q_-(\gamma_k\nabla\mathcal{L}(\beta_k)) \geq 0$ for all k.

□

C.7.3 Impact of stochasticity and linear scaling rule

Proposition 44. *With probability* $1 - 2ne^{-d/16} - 3/n^2$ *over the* $x_i \sim_{\text{iid}} \mathcal{N}(0, \sigma^2 I_d)$, $c_1\frac{d\sigma^2}{b}(1 + o(1)) \leq \lambda_b \leq \Lambda_b \leq c_2\frac{d\sigma^2}{b}(1 + o(1))\,,$

so that under these assumptions,

$$\sum_k \gamma_k \mathbb{E}\mathcal{L}(\beta_k) = \Theta\left(\frac{\gamma}{b}\sigma^2\|\beta^\star\|_1 \ln(1/\alpha)\right)\,.$$

Proof. The bound on λ_b, Λ_b is a direct consequence of the concentration bound provided in Lemma 33. □

C.7.4 (Stochastic) gradients at the initialisation

To understand the behaviour and the effects of the stochasticity and the stepsize on the shape of Gain_γ, we analyse a noiseless sparse recovery problem under the following standard Assumption 17 (see Candès et al. [2006]). As as common in the sparse recovery literature, we also make the following Assumption 18 on the inputs.

Assumption 17. *There exists an s-sparse ground truth vector $\beta^\star_{\text{sparse}}$ where s verifies $n = \Omega(s\ln(d))$, such that $y_i = \langle \beta^\star_{\text{sparse}}, x_i\rangle$ for all $i \in [n]$.*

Assumption 18. *There exists $\delta, c_1, c_2 > 0$ such that for all s-sparse vectors β, there exists $\varepsilon \in \mathbb{R}^d$ such that $(X^\top X)\beta = \beta + \varepsilon$ where $\|\varepsilon\|_\infty \leq \delta\|\beta\|_2$ and $c_1\|\beta\|_2^2\mathbf{1} \leq \frac{1}{n}\sum_i x_i^2\langle x_i, \beta\rangle^2 \leq c_2\|\beta\|_2^2\mathbf{1}$.*

The first part of Assumption 18 closely resembles the classical restricted isometry property (RIP) and is relevant for GD while the second part is relevant for SGD. Such an assumption is not restrictive and holds with high probability for Gaussian inputs $\mathcal{N}(0, \sigma^2 I_d)$ (see Lemma 30 in Appendix).

Based on the claim above, we analyse the shape of the (stochastic) gradient at initialisation. For GD and SGD, it respectly writes, where $g_0 = \nabla\mathcal{L}_{i_0}(\beta_0)^2$, $i_0 \sim \text{Unif}([\text{n}])$:

$$\nabla\mathcal{L}(\beta_0)^2 = [X^\top X\beta^\star]^2\,, \quad \mathbb{E}_{i_0}[g_0] = \frac{1}{n}\sum_i x_i^2\langle x_i, \beta^\star\rangle^2.$$

The following lemma then shows that while the initial stochastic gradients of SGD are homogeneous, it is not the case for that of GD.

Proposition 45. *Under Assumption 18, the squared full batch gradient and the expected stochastic gradient at initialisation satisfy, for some ε verifying $\|\varepsilon\|_\infty << \|\beta^\star_{\text{sparse}}\|_\infty^2$:*

$$\nabla\mathcal{L}(\beta_0)^2 = (\beta^\star_{\text{sparse}})^2 + \varepsilon\,, \tag{C.12}$$

$$\mathbb{E}_{i_0}[\nabla\mathcal{L}_{i_0}(\beta_0)^2] = \Theta\Big(\|\beta^\star\|_2^2\mathbf{1}\Big)\,. \tag{C.13}$$

Proof of Proposition 45. Under Assumption 18, we have using:

$$\begin{aligned}\nabla\mathcal{L}(\beta_0)^2 &= (X^\top X\beta^\star_{\text{sparse}}) \\ &= (\beta^\star_{\text{sparse}} + \varepsilon)^2 \\ &= {\beta^\star_{\text{sparse}}}^2 + \varepsilon^2 + 2\varepsilon\beta^\star_{\text{sparse}}\,.\end{aligned}$$

We have $\|\varepsilon^2 + 2\varepsilon\beta^\star_{\text{sparse}}\|_\infty \leq \|\varepsilon\|_\infty^2 + 2\|\varepsilon\|_\infty\|\beta^\star_{\text{sparse}}\|_\infty$, and we conclude by using $\|\varepsilon\|_\infty \leq \delta\|\beta^\star_{\text{sparse}}\|_2$.

Then,

$$\mathbb{E}_{i\sim\text{Unif}([\text{n}])}[\nabla\mathcal{L}_i(\beta_0)^2] = \frac{1}{n}x_i^2\langle x_i, \beta^\star_{\text{sparse}}\rangle\,,$$

and we conclude using Assumption 18.

□

Proof of Proposition 31. The proof proceeds as that of Proposition 45. □

C.7.5 Convergence of α_∞ and $\tilde{\beta}_0$ for $\gamma \to 0$

Proposition 46. *Let $\tilde{\beta}_0(\gamma), \alpha_\infty(\gamma)$ be as defined in Theorem 1, for constant stepsizes $\gamma_k \equiv \gamma$. We have:*

$$\tilde{\beta}_0(\gamma) \to 0\,, \quad \boldsymbol{\alpha}_\infty \to \alpha \mathbf{1}\,,$$

when $\gamma \to 0$.

Proof. We have, as proved previoulsy, that

$$\begin{aligned}
\| \sum_k \gamma^2 \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)^2 \|_1 &\leq \sum_k \gamma^2 \| \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)^2 \|_1 \\
&= \sum_k \gamma^2 \| \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k) \|_2^2 \\
&\leq 2L\gamma^2 \sum_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) \\
&\leq 2L\gamma D_{h_0}(\beta^\star, \beta_0)\,,
\end{aligned}$$

for $\gamma \leq \frac{c}{BL}$. Thus, $\sum_k \gamma^2 \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)^2 \to 0$ as $\gamma \to 0$ (note that β_k implicitly depends on γ, so that this result is not immediate).

Then, for $\gamma \leq \frac{c}{LB}$,

$$\| \ln(\boldsymbol{\alpha}_\infty^2 / \alpha^2) \|_1 \leq \sum_k \| q(\gamma \mathcal{L}(\beta_k) \|_1 \leq 2 \sum_k \gamma^2 \| \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)^2 \|_1\,,$$

which tends to 0 as $\gamma \to 0$. Similarly, $\| \ln(\boldsymbol{\alpha}_{+,\infty}^2 / \alpha^2) \|_1 \to 0$ and $\| \ln(\boldsymbol{\alpha}_{-,\infty}^2 / \alpha^2) \|_1 \to 0$ as $\gamma \to 0$, leading to $\tilde{\beta}_0(\gamma) \to 0$ as $\gamma \to 0$.

□

C.8 Technical lemmas

In this section we present a few technical lemmas, used and referred to throughout the previous proofs.

Lemma 24. *Let $\alpha_+, \alpha_- > 0$ and $x \in \mathbb{R}$, and $\beta = \alpha_+^2 e^x - \alpha_-^2 e^{-x}$. We have:*

$$\operatorname{arcsinh}\left(\frac{\beta}{2\alpha_+\alpha_-}\right) = x + \ln\left(\frac{\alpha_+}{\alpha_-}\right) = x + \operatorname{arcsinh}\left(\frac{\alpha_+^2 - \alpha_-^2}{2\alpha_+\alpha_-}\right).$$

Proof. First,

$$\begin{aligned}
\frac{\beta}{2\alpha_+\alpha_-} &= \frac{1}{2}\left(\frac{\alpha_+}{\alpha^-} e^x - \left(\frac{\alpha_+}{\alpha^-}\right)^{-1} e^{-x}\right) \\
&= \frac{e^{x + \ln(\alpha_+/\alpha_-)} - e^{-x - \ln(\alpha_+/\alpha_-)}}{2} \\
&= \sinh(x + \ln(\alpha_+/\alpha_-))\,,
\end{aligned}$$

hence the result by taking the arcsinh of both sides. Note also that we have $\ln(\alpha_+/\alpha_-) = \operatorname{arcsinh}(\frac{\alpha_+^2 - \alpha_-^2}{2\alpha_+\alpha_-})$. □

Lemma 25. *If $|x| \leq 1/2$ then $x^2 \leq q(x) \leq 2x^2$*

Lemma 26. *On the ℓ_∞ ball of radius B, the quadratic loss function $\beta \mapsto \mathcal{L}(\beta)$ is $4\lambda_{\max}\max(B, \alpha^2)$-relatively smooth w.r.t all the h_k's.*

Proof. We have:

$$\nabla^2 h_k(\beta) = \operatorname{diag}\left(\frac{1}{2\sqrt{\alpha_k^4 + \beta^2}}\right) \succeq \operatorname{diag}\left(\frac{1}{2\sqrt{\alpha^4 + \beta^2}}\right),$$

since $\alpha_k \leq \alpha$ component-wise. Thus, $\nabla^2 h_k(\beta) \succeq \frac{1}{2}\min\left(\min_{1\leq i\leq d}\frac{1}{2|\beta_i|}, \frac{1}{2\alpha^2}\right) I_d = \frac{1}{\max(4\|\beta\|_\infty, 4\alpha^2)} I_d$, and h_k is $\frac{1}{\max(4B, 4\alpha^2)}$-strongly convex on the ℓ^∞ norm of radius B. Since $\mathcal{L}$ is $\lambda_{\max}$-smooth over $\mathbb{R}^d$, we have our result. □

Lemma 27. *For $k \geq 0$ and for all $\beta \in \mathbb{R}^d$:*

$$|h_{k+1}(\mathsf{w}) - h_k(\mathsf{w})| \leq 8L_2\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|\mathsf{w}\|_\infty.$$

Proof. We have $\alpha_{+,k+1}^2 = \alpha_{+,k}^2 e^{-\delta_{+,k}}$ and $\alpha_{-,k+1}^2 = \alpha_{-,k}^2 e^{-\delta_{-,k}}$, for $\delta_{+,k} = \tilde{q}(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k))$ and $\delta_{-,k} = \tilde{q}(-\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\mathsf{w}_k))$. And $\alpha_{k+1} = \alpha_k \exp(-\delta_k)$ where $\delta_k \coloneqq \delta_{+,k} + \delta_{-,k} = q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$.

To prove the result we will use that for $\mathsf{w} \in \mathbb{R}^d$, we have $|(h_{k+1} - h_k)(\mathsf{w})| \leq \sum_{i=1}^d \int_0^{|\mathsf{w}_i|} |\nabla_i h_{k+1}(x) - \nabla_i h_k(x)| \mathrm{d}x$.

First, using that $|\operatorname{arcsinh}(a) - \operatorname{arcsinh}(b)| \leq |\ln(a/b)|$ for $ab > 0$. We have that

$$\begin{aligned}\left|\operatorname{arcsinh}\left(\frac{x}{\alpha_{k+1}^2}\right) - \operatorname{arcsinh}\left(\frac{x}{\alpha_k^2}\right)\right| &\leq \ln\left(\frac{\alpha_k^2}{\alpha_{k+1}^2}\right) \\ &= \delta_k\,,\end{aligned}$$

since $\delta_k \geq 0$ due to our stepsize condition.

We now prove that $|\phi_{k+1} - \phi_k| \leq \frac{|\delta_{+,k} - \delta_{-,k}|}{2}$. We have $\phi_k = \operatorname{arcsinh}\left(\frac{\alpha_{+,k}^2 - \alpha_{-,k}^2}{2\alpha_{+,k}\alpha_{-,k}}\right)$ and hence,

$$|\phi_{k+1} - \phi_k| = \left|\operatorname{arcsinh}\left(\frac{\alpha_{+,k}^2 - \alpha_{-,k}^2}{2\alpha_{+,k}\alpha_{-,k}}\right) - \operatorname{arcsinh}\left(\frac{\alpha_{+,k+1}^2 - \alpha_{-,k+1}^2}{2\alpha_{+,k+1}\alpha_{-,k+1}}\right)\right|.$$

Then, assuming that $\alpha_{+,k,i} \geq \alpha_{-,k,i}$, we have:

$$\begin{aligned}\frac{\alpha_{+,k+1,i}^2 - \alpha_{-,k+1,i}^2}{2\alpha_{+,k+1,i}\alpha_{-,k+1,i}} &= e^{\delta_{k,i}/2}\frac{\alpha_{+,k,i}^2 e^{-\delta_{+,k,i}} - \alpha_{-,k,i}^2 e^{-\delta_{-,k,i}}}{2\alpha_{+,k,i}\alpha_{-,k,i}} \\ &\begin{cases} \leq \begin{cases} e^{\frac{\delta_{+,k,i} - \delta_{-,k,i}}{2}}\dfrac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}} & \text{if} \quad \delta_{+,k,i} \geq \delta_{-,k,i} \\ e^{\frac{\delta_{-,k,i} - \delta_{+,k,i}}{2}}\dfrac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}} & \text{if} \quad \delta_{-,k,i} \geq \delta_{+,k,i} \end{cases} \\ \geq \begin{cases} e^{-\frac{\delta_{+,k,i} - \delta_{-,k,i}}{2}}\dfrac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}} & \text{if} \quad \delta_{+,k,i} \geq \delta_{-,k,i} \\ e^{-\frac{\delta_{-,k,i} - \delta_{+,k,i}}{2}}\dfrac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}} & \text{if} \quad \delta_{-,k,i} \geq \delta_{+,k,i} \end{cases} \end{cases}.\end{aligned}$$

We thus have $\frac{\alpha^2_{+,k+1,i}-\alpha^2_{-,k+1,i}}{2\alpha_{+,k+1,i}\alpha_{-,k+1,i}} \in \left[e^{-\frac{|\delta_{+,k,i}-\delta_{-,k,i}|}{2}}, e^{\frac{|\delta_{+,k,i}-\delta_{-,k,i}|}{2}}\right] \times \frac{\alpha^2_{+,k,i}-\alpha^2_{-,k,i}}{2\alpha_{+,k,i}\alpha_{-,k,i}}$, and this holds similarly if $\alpha_{+,k,i} \leq \alpha_{-,k,i}$. Then, using $|\operatorname{arcsinh}(a) - \operatorname{arcsinh}(b)| \leq |\ln(a/b)|$ we obtain that:

$$\begin{aligned}|\phi_{k+1} - \phi_k| &= \left|\operatorname{arcsinh}\left(\frac{\alpha^2_{+,k} - \alpha^2_{-,k}}{2\alpha_{+,k}\alpha_{-,k}}\right) - \operatorname{arcsinh}\left(\frac{\alpha^2_{+,k+1} - \alpha^2_{-,k+1}}{2\alpha_{+,k+1}\alpha_{-,k+1}}\right)\right| \\ &\leq \frac{|\delta_{+,k} - \delta_{-,k}|}{2}\,.\end{aligned}$$

Wrapping things up, we have:

$$|\nabla h_k(\mathsf{w}) - \nabla h_{k+1}(\mathsf{w})| \leq \delta_k + \frac{|\delta_{+,k} - \delta_{-,k}|}{2} \leq 2\delta_k\,,$$

This leads to the following bound:

$$\begin{aligned}|h_{k+1}(\mathsf{w}) - h_k(\mathsf{w})| &\leq \langle |2\delta_k|, |\mathsf{w}| \rangle \\ &\leq 2\|\delta_k\|_1 \|\mathsf{w}\|_\infty.\end{aligned}$$

Recall that $\delta_k = q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)$, hence from Lemma 25 if $\gamma_k \|\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leq 1/2$, we get that

$$\|\delta_k\|_1 \leq 2\gamma_k^2 \|\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_2^2 \leq 4L_2\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k).$$

Putting things together we obtain that

$$\begin{aligned}|h_{k+1}(\mathsf{w}) - h_k(\mathsf{w})| &\leq \langle |2\delta_k|, |\mathsf{w}| \rangle \\ &\leq 8L_2\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k) \|\mathsf{w}\|_\infty.\end{aligned}$$

□

C.9 Concentration inequalities for matrices

In this last section of the appendix, we provide and prove several concentration bounds for random vectors and matrices, with (possibly uncentered) isotropic gaussian inputs. These inequalities can easily be generalized to subgaussian random variables via more refined concentration bounds, and to non-isotropic subgaussian random variables [Even and Massoulie, 2021], leading to a dependence on an effective dimension and on the subgaussian matrix Σ. We present these lemmas before proving them in a row.

The next two lemmas closely ressemble the RIP assumption, for centered and then for uncentered gaussians.

Lemma 28. *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be* i.i.d. *random variables of law $\mathcal{N}(0, I_d)$ and $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. Then, denoting by $\mathcal{C}$ the set of all s-sparse vector $\beta \in \mathbb{R}^d$ satisfying $\|\beta\|_2 \leq 1$, there exist $C_4, C_5 > 0$ such that for any $\varepsilon > 0$, if $n \geq C_4 s \ln(d)\varepsilon^{-2}$,*

$$\mathbb{P}\left(\sup_{\beta \in \mathcal{S}} \|H\beta - \beta\|_\infty \geq \varepsilon\right) \leq e^{-C_5 n}\,.$$

Lemma 29. *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be* i.i.d. *random variables of law $\mathcal{N}(\mu, \sigma^2 I_d)$ and $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. Then, denoting by $\mathcal{C}$ the set of all s-sparse vector $\beta \in \mathbb{R}^d$ satisfying $\|\beta\|_2 \leq 1$, there exist $C_4, C_5 > 0$ such that for any $\varepsilon > 0$, if $n \geq C_4 s \ln(d)\varepsilon^{-2}$,*

$$\mathbb{P}\left(\sup_{\beta \in \mathcal{S}} \|H\beta - \mu\langle \mu, \beta\rangle - \sigma^2\beta\|_\infty \geq \varepsilon\right) \leq e^{-C_5 n}\,.$$

We then provide two lemmas that estimate the mean Hessian of SGD.

Lemma 30. *Let $x_1, \dots, x_n$ be i.i.d. random variables of law $\mathcal{N}(0, I_d)$. Then, there exist $c_1, c_2 > 0$ such that with probability $1 - \frac{1}{d^2}$ and if $n = \Omega(s^{5/4}\ln(d))$, we have for all s-sparse vectors β:*

$$c_1\|\beta\|_2^2 \mathbf{1} \leq \frac{1}{n}\sum_{i=1}^n x_i^2 \langle x_i, \beta\rangle^2 \leq c_2\|\beta\|_2^2\mathbf{1}\,,$$

where the inequality is meant component-wise.

Lemma 31. *Let $x_1, \dots, x_n$ be i.i.d. random variables of law $\mathcal{N}(\mu, \sigma^2 I_d)$. Then, there exist $c_0, c_1, c_2 > 0$ such that with probability $1 - \frac{c_0}{d^2} - \frac{1}{nd}$ and if $n = \Omega(s^{5/4}\ln(d))$ and $\mu \geq 4\sigma\sqrt{\ln(d)}\mathbf{1}$, we have for all s-sparse vectors β:*

$$\frac{\mu^2}{2}\left(\langle \mu, \beta\rangle^2 + \frac{1}{2}\sigma^2\|\beta\|_2^2\right) \leq \frac{1}{n}\sum_i x_i^2\langle x_i, \beta\rangle^2 \leq 4\mu^2\left(\langle \mu, \beta\rangle^2 + 2\sigma^2\|\beta\|_2^2\right)\,.$$

where the inequality is meant component-wise.

Finally, next two lemmas are used to estimate λ_b, Λ_b.

Lemma 32. *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be* i.i.d. *random variables of law $\mathcal{N}(\mu\mathbf{1}, \sigma^2 I_d)$. Let $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and $\tilde{H} = \frac{1}{n}\sum_{i=1}^n \|x_i\|^2 x_i x_i^\top$. There exist numerical constants $C_2, C_3 > 0$ such that*

$$\mathbb{P}\Big(C_2(\mu^2 + \sigma^2)dH \preceq \tilde{H} \preceq C_3(\mu^2 + \sigma^2)dH\Big) \geq 1 - 2ne^{-d/16}\,.$$

Lemma 33. *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be* i.i.d. *random variables of law $\mathcal{N}(\mu \mathbf{1}, \sigma^2 I_d)$ for some $\mu \in \mathbb{R}$. Let $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and for $1 \le b \le n$ let $\tilde{H}_b = \mathbb{E}_{\mathcal{B}}\left[\left(\frac{1}{b}\sum_{i\in\mathcal{B}} x_i x_i^\top\right)^2\right]$ where $\mathcal{B} \subset [n]$ is sampled uniformly at random in $\{\mathcal{B} \subset [n] \, s.t. \, |\mathcal{B}| = b\}$. With probability $1 - 2ne^{-d/16} - 3/n^2$, we have, for some numerical constants $c_1, c_2, c_3, C > 0$:*

$$\left(c_1 \frac{d(\mu^2+\sigma^2)}{b} - c_2 \frac{(\sigma^2+\mu^2)\ln(n)}{\sqrt{d}} - c_3 \frac{\mu^2 d}{n}\right) H \preceq \tilde{H}_b \preceq C\left(\frac{d(\mu^2+\sigma^2)}{b} + \frac{(\sigma^2+\mu^2)\ln(n)}{\sqrt{d}} + \mu^2 d\right)$$

Proof of Lemma 28. For $j \in [d]$, we have:

$$\begin{aligned}
(H\beta)_j &= \frac{1}{n}\sum_{i=1}^n x_{ij}\langle x_i, \beta\rangle \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{j'=1}^d x_{ij}x_{ij'}\beta_{j'} \\
&= \frac{1}{n}\sum_{i=1}^n x_{ij}^2\beta_j + \frac{1}{n}\sum_{i=1}^n \sum_{j'\neq j} x_{ij}x_{ij'}\beta_{j'} \\
&= \frac{\beta_j}{n}\sum_{i=1}^n x_{ij}^2 + \frac{1}{n}\sum_{i=1}^n x_{ij}\sum_{j'\neq j} x_{ij'}\beta_{j'} \, .
\end{aligned}$$

We thus notice that $\mathbb{E}\,[H\beta] = \beta$, and

$$(H\beta)_j = \beta_j + \frac{\beta_j}{n}\sum_{i=1}^n (x_{ij}^2 - 1) + \frac{1}{n}\sum_{i=1}^n z_i \, ,$$

where $z_i = x_{ij}\sum_{j'\neq j} x_{ij'}\beta_{j'}$, and $\sum_{j'\neq j} x_{ij'}\beta_{j'} \sim \mathcal{N}(0, \|\beta\|^2 - \beta_j^2)$ and $\|\beta\|^2 - \beta_j^2 \le 1$. Hence, $z_j + x_{ij}^2 - 1$ is a centered subexponential random variables (with a subexponential parameter of order 1). Thus, for $t \le 1$:

$$\mathbb{P}\left(\left|\frac{\beta_j}{n}\sum_{i=1}^n (x_{ij}^2 - 1) + \frac{1}{n}\sum_{i=1}^n z_i\right| \ge t\right) \le 2e^{-cnt^2} \, .$$

Hence, using an ε-net of $\mathcal{C} = \left\{\beta \in \mathbb{R}^d : \|\beta\|_2 \le 1 \, , \|\beta\|_0\right\}$ (of cardinality less than $d^s \times (C/\varepsilon)^s$, and for ε of order 1), we have, using the classical ε-net trick explained in [Chapt. 9, Vershynin [2018] or [App. C, Even and Massoulie [2021]]:

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{C},\, j\in[d]} |(H\beta)_j - \beta_j| \ge t\right) \le d \times d^s (C/\varepsilon)^s \times 2e^{-cnt^2} = \exp\left(-c\ln(2)nt^2 + (s+1)\ln(d) + s\ln(C/\varepsilon)\right) \, .$$

Consequently, for $t = \varepsilon$ and if $n \ge C_4 s\ln(d)/\varepsilon^2$, we have:

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{C},\, j\in[d]} |(H\beta)_j - \beta_j| \ge t\right) \le \exp\left(-C_5 nt^2\right) \, .$$

□

Proof of Lemma 29. We write $x_i = \sigma z_i + \mu$ where $z_i \sim \mathcal{N}(0, I_d)$. We have:

$$\begin{aligned} X^\top X\beta &= \frac{1}{n}\sum_{i=1}^n (\mu + \sigma z_i)\langle \mu + \sigma z_i, \beta\rangle \\ &= \mu\langle \mu, \beta\rangle + \frac{\sigma^2}{n}\sum_{i=1}^n z_i\langle z_i, \beta\rangle + \frac{\sigma}{n}\sum_{i=1}^n \mu\langle z_i, \beta\rangle + \frac{\sigma}{n}\sum_{i=1}^n z_i\langle \mu, \beta\rangle \\ &= \mu\langle \mu, \beta\rangle + \frac{\sigma^2}{n}\sum_{i=1}^n z_i\langle z_i, \beta\rangle + \sigma\mu\langle \frac{1}{n}\sum_{i=1}^n z_i, \beta\rangle + \frac{\sigma\langle \mu, \beta\rangle}{n}\sum_{i=1}^n z_i\,. \end{aligned}$$

The first term is deterministic and is to be kept. The second one is of order $\sigma^2\beta$ whp using Lemma 28. Then, $\frac{1}{n}\sum_{i=1}^n z_i \sim \mathcal{N}(0, I_d/n)$, so that

$$\mathbb{P}\left(\left|\langle \frac{1}{n}\sum_{i=1}^n z_i, \beta\rangle\right| \geq t\right) \leq 2e^{-nt^2/(2\|\beta\|_2^2)}\,,$$

and

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n z_{ij}\right| \geq t\right) \leq 2e^{-nt^2/2}\,.$$

Hence,

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{C}}\left\|\frac{1}{n}\sum_{i=1}^n z_{ij}\right\|_\infty \geq t\,,\, \sup_{\beta\in\mathcal{C}}\left|\langle \frac{1}{n}\sum_{i=1}^n z_i, \beta\rangle\right| \geq t\right) \leq 4e^{cs\ln(d)}e^{-nt^2/2}\,.$$

Thus, with probability $1 - Ce^{-n\varepsilon^2}$ and under the assumptions of Lemma 28, we have $\|X^\top X\beta - \mu\langle \mu, \beta\rangle - \sigma^2\beta\|_\infty \leq \varepsilon$ □

Proof of Lemma 30. To ease notations, we assume that $\sigma = 1$. We remind (O'Donnell [2021], Chapter 9 and Tao [2010]) that for *i.i.d.* real random variables $a_1, \ldots, a_n$ that satisfy a tail inequality of the form

$$\mathbb{P}\big(|a_1 - \mathbb{E}a_1| \geq t\big) \leq Ce^{-ct^p}\,, \tag{C.14}$$

for $p < 1$, then for all $\varepsilon > 0$ there exists C', c' such that for all t,

$$\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n a_i - \mathbb{E}a_1| \geq t\big) \leq C'e^{-c'nt^{p-\varepsilon}}\,.$$

We now expand $\frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i, \beta\rangle^2$:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i, \beta\rangle^2 &= \frac{1}{n}\sum_{i\in[n],k,\ell\in[d]} x_i^2 x_{ik}x_{i\ell}\beta_k\beta_\ell \\ &= \frac{1}{n}\sum_{i\in[n],k\in[d]} x_i^2 x_{ik}^2\beta_k^2 + \frac{1}{n}\sum_{i\in[n],k\neq\ell\in[d]} x_i^2 x_{ik}x_{i\ell}\beta_k\beta_\ell\,. \end{aligned}$$

Thus, for $j \in [d]$,

$$\left(\frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i, \beta\rangle^2\right)_j = \sum_{k\in[d]}\frac{\beta_k^2}{n}\sum_{i\in[n]} x_{ij}^2 x_{ik}^2 + \sum_{k\neq\ell\in[d]}\frac{\beta_k\beta_\ell}{n}\sum_{i\in[n]} x_{ij}^2 x_{ik}x_{i\ell}\,.$$

We notice that for all indices, all $x_{ij}^2 x_{ik} x_{i\ell}$ and $x_{ij}^2 x_{ik}^2$ satisfy the tail inequality Equation (C.14) for $C = 8$, $c = 1/2$ and $p = 1/2$, so that for $\varepsilon = 1/4$:

$$\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n x_{ij}^2 x_{ik} x_{i\ell}| \geq t\big) \leq C' e^{-c' n t^{1/4}} \quad , \quad \mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n x_{ij}^2 x_{ik}^2 - \mathbb{E}\left[x_{ij}^2 x_{ik}^2\right]| \geq t\big) \leq C' e^{-c' n t^{1/4}} .$$

For $j \neq k$, we have $\mathbb{E}\left[x_{ij}^2 x_{ik}^2\right] = 1$ while for $j = k$, we have $\mathbb{E}\left[x_{ij}^2 x_{ik}^2\right] = \mathbb{E}\left[x_{ij}^4\right] = 3$. Hence,

$$\mathbb{P}\left(\exists j, k \neq \ell\,,\, |\frac{1}{n}\sum_{i=1}^n x_{ij}^2 x_{ik} x_{i\ell}| \geq t\,,\, |\frac{1}{n}\sum_{i=1}^n x_{ij}^2 x_{ik}^2 - \mathbb{E}\left[x_{ij}^2 x_{ik}^2\right]| \geq t\right) \leq C' d^2 e^{-c' n t^{1/4}} .$$

Thus, with probability $1 - C' d^2 e^{-c' n t^{1/4}}$, for all $j \in [d]$,

$$\left| \left(\frac{1}{n}\sum_{i=1}^n x_i^2 \langle x_i, \beta\rangle^2\right)_j - 2\beta_j^2 - \|\beta\|_2^2 \right| \leq t \sum_{k,\ell} |\beta_k||\beta_\ell| = t\|\beta\|_1^2 .$$

Using the classical technique of Baraniuk et al. [2008], to make a union bound on all s-sparse vectors, we consider an ε-net of the set of s-sparse vectors of ℓ^2-norm smaller than 1. This ε-net is of cardinality less than $(C_0/\varepsilon)^s d^s$, and we only need to take ε of order 1 to obtain the result for all s-sparse vectors. This leads to:

$$\mathbb{P}\left(\exists \beta \in \mathbb{R}^d \text{ } s\text{-sparse and } \|\beta\|_2 \leq 1\,,\, \exists j \in \mathbb{R}^d\,, \quad \left| \left(\frac{1}{n}\sum_{i=1}^n x_i^2 \langle x_i, \beta\rangle^2\right)_j - 2\beta_j^2 - \|\beta\|_2^2 \right| \geq t\|\beta\|_1^2 \right) \leq C' d^2 e^{c_1 s + s\ln(d)} e^{-c' n t^{1/4}} .$$

This probability is equal to C'/d^2 for $t = \left(\frac{(s+4)\ln(d) + c_1 s}{c' n}\right)^4$. We conclude that with probability $1 - C'/d^2$, all s-sparse vectors β satisfy:

$$\left| \left(\frac{1}{n}\sum_{i=1}^n x_i^2 \langle x_i, \beta\rangle^2\right)_j - 2\beta_j^2 - \|\beta\|_2^2 \right| \leq \left(\frac{(s+4)\ln(d) + c_1 s}{c' n}\right)^4 \|\beta\|_1^2 \leq \left(\frac{(s+4)\ln(d) + c_1 s}{c' n}\right)^4 s\|\beta\|_2^2 ,$$

and the RHS is smaller than $\|\beta\|_2^2/2$ for $n \geq \Omega(s^{5/4}\ln(d))$. □

Proof of Lemma 31. We write $x_i = \mu + \sigma z_i$ where $x_i \sim \mathcal{N}(0, 1)$. We have:

$$\mathbb{P}\big(\forall i \in [n], \forall j \in [d],\, |z_{ij}| \geq t\big) \leq e^{\ln(nd) - t^2/2} = \frac{1}{nd} ,$$

for $t = 2\sqrt{\ln(nd)}$. Thus, if $\mu \geq 4\sigma\sqrt{\ln(nd)}$ we have $\frac{\mu}{2} \leq x_i \leq 2\mu$, so that

$$\frac{\mu^2}{2n}\sum_i \langle x_i, \beta\rangle^2 \leq \frac{1}{n}\sum_i x_i^2 \langle x_i, \beta\rangle^2 \leq \frac{4\mu^2}{n}\sum_i \langle x_i, \beta\rangle^2 .$$

Then, $\langle x_i, \beta\rangle \sim \mathcal{N}(\langle \mu, \beta\rangle, \sigma^2\|\beta\|_2^2)$. For now, we assume that $\|\beta\|_2 = 1$. We have $\mathbb{P}(|\langle x_i, \beta\rangle^2 - \langle \mu, \beta\rangle^2 - \sigma^2\|\beta\|_2^2| \geq t) \leq Ce^{-ct/\sigma^2}$, and for $t \leq 1$, using concentration of subexponential random variables [Vershynin, 2018]:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i \langle x_i, \beta\rangle^2 - \langle \mu, \beta\rangle^2 - \sigma^2\|\beta\|_2^2\right| \geq t\right) \leq C' e^{-nc' t^2/\sigma^4} ,$$

and using the ε-net trick of Baraniuk et al. [2008],

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{C}}\left|\frac{1}{n}\sum_i\langle x_i,\beta\rangle^2-\langle\mu,\beta\rangle^2-\sigma^2\|\beta\|_2^2\right|\geq t\right)\leq C'e^{s\ln(d)-nc't^2/\sigma^4}=\frac{C'}{d^2}\,,$$

for $t=\sigma^2\|\beta\|_2^2\sqrt{\frac{2(cs+2)\ln(d)}{n}}$. Consequently, we have, with probability $1-\frac{C'}{d^2}-\frac{1}{nd}$:

$$\frac{\mu^2}{2}\left(\langle\mu,\beta\rangle^2+\frac{1}{2}\sigma^2\|\beta\|_2^2\right)\leq\frac{1}{n}\sum_i x_i^2\langle x_i,\beta\rangle^2\leq 4\mu^2\left(\langle\mu,\beta\rangle^2+2\sigma^2\|\beta\|_2^2\right).$$

□

Proof of Lemma 32. First, we write $x_i=\mu\mathbf{1}+\sigma z_i$, where $z_i\sim\mathcal{N}(0,I)$, leading to:

$$\frac{1}{n}\sum_{i\in[n]}\|x_i\|_2^2x_ix_i^\top=\frac{1}{n}\sum_{i\in[n]}\big(\sigma^2\|z_i\|_2^2+d\mu^2+2\sigma\mu\langle\mathbf{1},z_i\rangle\big)x_ix_i^\top$$

We use concentration of χ^2_d random variables around d:

$$\mathbb{P}(\chi^2_d>d+2t+2\sqrt{dt})\geq t)\leq e^{-t}\quad\text{and}\quad\mathbb{P}(\chi^2_d>d-2\sqrt{dt})\leq t)\leq e^{-t}\,,$$

so that for all $i\in[n]$,

$$\mathbb{P}(\|z_i\|_2^2\notin[d-2\sqrt{dt},d+2t+2\sqrt{dt}])\leq 2e^{-t}\,.$$

Thus,

$$\mathbb{P}(\forall i\in[n],\ \|z_i\|_2^2\in[d-2\sqrt{dt},d+2t+2\sqrt{dt}])\geq 1-2ne^{-t}\,.$$

Taking $t=d/16$,

$$\mathbb{P}(\forall i\in[n],\ \|z_i\|_2^2\in[\frac{d}{2},13d/8])\geq 1-2ne^{-d/16}\,.$$

Then, for all i, $\langle\mathbf{1},z_i\rangle$ is of law $\mathcal{N}(0,d)$, so that $\mathbb{P}(|\langle\mathbf{1},z_i\rangle|\geq t)\leq 2e^{-t^2/(2d)}$ and

$$\mathbb{P}\big(\forall i\in[n],\ |\langle\mathbf{1},z_i\rangle|\geq t\big)\leq 2ne^{-\frac{t^2}{2d}}\,.$$

Taking $t=\sqrt{2}d^{3/4}$,

$$\mathbb{P}\big(\forall i\in[n],\ |\langle\mathbf{1},z_i\rangle|\geq d^{3/4}\big)\leq 2ne^{-d^{1/2}}\,.$$

Thus, with probability $1-2n(e^{-d/16}+e^{-\sqrt{d}}$, we have $\forall i\in[n]$, $|\langle\mathbf{1},z_i\rangle|\geq d^{3/4}$ and $\|z_i\|_2^2\in[\frac{d}{2},13d/8]$, so that

$$\big(\frac{d}{2}\sigma^2+d\mu^2-2\mu\sigma d^{3/4}\big)H\preceq\tilde{H}\preceq\big(\frac{13d}{8}\sigma^2+d\mu^2+2\mu\sigma d^{3/4}\big)H\,,$$

leading to the desired result. □

Proof of Lemma 33. We have:

$$\begin{aligned}\tilde{H}_b&=\mathbb{E}\left[\frac{1}{b^2}\sum_{i,j\in\mathcal{B}}\langle x_i,x_j\rangle x_ix_j^\top\right]\\&=\mathbb{E}\left[\frac{1}{b^2}\sum_{i\in\mathcal{B}}\|x_i\|_2^2x_ix_i^\top+\frac{1}{b^2}\sum_{i,j\in\mathcal{B},\,i\neq j}\langle x_i,x_j\rangle x_ix_j^\top\right]\\&=\frac{1}{b^2}\sum_{i\in[n]}\mathbb{P}(i\in\mathcal{B})\|x_i\|_2^2x_ix_i^\top+\frac{1}{b^2}\sum_{i\neq j}\mathbb{P}(i,j\in\mathcal{B})\langle x_i,x_j\rangle x_ix_j^\top\,.\end{aligned}$$

Then, since $\mathbb{P}(i \in \mathcal{B}) = \frac{b}{n}$ and $\mathbb{P}(i,j \in \mathcal{B}) = \frac{b(b-1)}{n(n-1)}$ for $i \neq j$, we get that:

$$\tilde{H}_b = \frac{1}{bn}\sum_{i\in[n]} \|x_i\|_2^2 x_i x_i^\top + \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top .$$

Using Lemma 32, the first term satisfies:

$$\mathbb{P}\Big(\frac{d(\mu^2+\sigma^2)}{b}C_2 H \preceq \frac{1}{bn}\sum_{i\in[n]} \|x_i\|_2^2 x_i x_i^\top \preceq \frac{d(\mu^2+\sigma^2)}{b}C_3 H\Big) \geq 1 - 2ne^{-d/16}.$$

We now show that the second term is of smaller order. Writing $x_i = \mu\mathbf{1} + \sigma z_i$ where $z_i \sim \mathcal{N}(0, I_d)$, we have:

$$\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top = \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top$$

For $i \neq j$, $\langle x_i, x_j\rangle = \sum_{k=1}^d x_{ik}x_{jk} = \sum_{k=1}^d a_k$ where $a_k = x_{ik}x_{jk}$ satisfies $\mathbb{E}a_k = 0$, $\mathbb{E}a_k^2 = 1$ and $\mathbb{P}(a_k \geq t) \leq 2\mathbb{P}(|x_{ik}| \geq \sqrt{t}) \leq 4e^{-t/2}$. Hence, a_k is a centered subexponential random variables. Using concentration of subexponential random variables Vershynin [2018], for $t \leq 1$,

$$\mathbb{P}\left(\frac{1}{d}|\langle x_i, x_j\rangle| \geq t\right) \leq 2e^{-cdt^2}.$$

Thus,

$$\mathbb{P}\left(\forall i \neq j,\ \frac{1}{d}|\langle x_i, x_j\rangle| \leq t\right) \geq 1 - n(n-1)e^{-cdt^2}.$$

Then, taking $t = d^{-1/2}4\ln(n)/c$, we have:

$$\mathbb{P}\left(\forall i \neq j,\ \frac{1}{d}|\langle x_i, x_j\rangle| \leq \frac{4\ln(n)}{c\sqrt{d}}\right) \geq 1 - \frac{1}{n^2}.$$

Going back to our second term,

$$\begin{aligned}\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top &= \frac{(b-1)}{bn(n-1)}\sum_{i<j}\langle x_i, x_j\rangle\big(x_i x_j^\top + x_j x_i^\top\big)\\ &\preceq \frac{(b-1)}{bn(n-1)}\sum_{i<j}\big|\langle x_i, x_j\rangle\big|\big(x_i x_i^\top + x_j x_j^\top\big),\end{aligned}$$

where we used $x_i x_j^\top + x_j x_i^\top \preceq x_i x_i^\top + x_j x_j^\top$. Thus,

$$\begin{aligned}\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top &\preceq \sup_{i\neq j}|\langle x_i, x_j\rangle| \times \frac{(b-1)}{bn(n-1)}\sum_{i<j}\big(x_i x_i^\top + x_j x_j^\top\big)\\ &= \sup_{i\neq j}|\langle x_i, x_j\rangle| \times \frac{b-1}{b}\frac{1}{n-1}\sum_{i=1}^n x_i x_i^\top\\ &= \sup_{i\neq j}|\langle x_i, x_j\rangle| \times \frac{b-1}{b}\frac{n}{n-1}H.\end{aligned}$$

Similarly, we have

$$\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top \succeq -\sup_{i\neq j}|\langle x_i, x_j\rangle| \times \frac{b-1}{b}\frac{n}{n-1}H.$$

Hence, with probability $1 - 1/n^2$,

$$-\frac{4\ln(n)}{c\sqrt{d}} \times \frac{b-1}{b}\frac{n}{n-1}H \preceq \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top \preceq \frac{4\ln(n)}{c\sqrt{d}} \times \frac{b-1}{b}\frac{n}{n-1}H\,.$$

Wrapping things up, with probability $1 - 1/n^2 - 2ne^{-d/16}$,

$$\left(-\frac{4\ln(n)}{c\sqrt{d}}\frac{b-1}{b}\frac{n}{n-1} + C_2\frac{d}{b}\right) \times H \preceq \tilde{H}_b \preceq \left(\frac{4\ln(n)}{c\sqrt{d}}\frac{b-1}{b}\frac{n}{n-1} + C_3\frac{d}{b}\right) \times H\,.$$

Thus, provided that $\frac{4\ln(n)}{c\sqrt{d}} \leq \frac{C_2 d}{2b}$ and $d \geq 48\ln(n)$, we have with probability $1 - 3/n^2$:

$$C_2'\frac{d}{b} \times H \preceq \tilde{H}_b \preceq C_3'\frac{d}{b} \times H\,.$$

□

Proof of Lemma 33. We have:

$$\begin{aligned}
\tilde{H}_b &= \mathbb{E}\left[\frac{1}{b^2}\sum_{i,j\in\mathcal{B}}\langle x_i, x_j\rangle x_i x_j^\top\right] \\
&= \mathbb{E}\left[\frac{1}{b^2}\sum_{i\in\mathcal{B}}\|x_i\|_2^2 x_i x_i^\top + \frac{1}{b^2}\sum_{i,j\in\mathcal{B}, i\neq j}\langle x_i, x_j\rangle x_i x_j^\top\right] \\
&= \frac{1}{b^2}\sum_{i\in[n]}\mathbb{P}(i\in\mathcal{B})\|x_i\|_2^2 x_i x_i^\top + \frac{1}{b^2}\sum_{i\neq j}\mathbb{P}(i,j\in\mathcal{B})\langle x_i, x_j\rangle x_i x_j^\top\,.
\end{aligned}$$

Then, since $\mathbb{P}(i\in\mathcal{B}) = \frac{b}{n}$ and $\mathbb{P}(i,j\in\mathcal{B}) = \frac{b(b-1)}{n(n-1)}$ for $i\neq j$, we get that:

$$\tilde{H}_b = \frac{1}{bn}\sum_{i\in[n]}\|x_i\|_2^2 x_i x_i^\top + \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top\,.$$

Using Lemma 32, the first term satisfies:

$$\mathbb{P}\Big(\frac{d(\mu^2+\sigma^2)}{b}C_2 H \preceq \frac{1}{bn}\sum_{i\in[n]}\|x_i\|_2^2 x_i x_i^\top \preceq \frac{d(\mu^2+\sigma^2)}{b}C_3 H\Big) \geq 1 - 2ne^{-d/16}\,.$$

We now show that the second term is of smaller order. Writing $x_i = \mu\mathbf{1} + \sigma z_i$ where $z_i \sim \mathcal{N}(0, I_d)$, we have:

$$\begin{aligned}
\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top &= \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle\mathbf{1}, z_i + z_j\rangle + \mu^2 d\big)x_i x_j^\top \\
&= \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle\mathbf{1}, z_i + z_j\rangle\big)x_i x_j^\top + \frac{(b-1)}{bn(n-1)}\mu^2 d\sum_{i\neq j}x_i x_j^\top
\end{aligned}$$

For $i\neq j$, $\langle z_i, z_j\rangle = \sum_{k=1}^d z_{ik}z_{jk} = \sum_{k=1}^d a_k$ where $a_k = z_{ik}z_{jk}$ satisfies $\mathbb{E}a_k = 0$, $\mathbb{E}a_k^2 = 1$ and $\mathbb{P}(a_k \geq t) \leq 2\mathbb{P}(|z_{ik}| \geq \sqrt{t}) \leq 4e^{-t/2}$. Hence, a_k is a centered subexponential random variables. Using concentration of subexponential random variables [Vershynin, 2018], for $t \leq 1$,

$$\mathbb{P}\left(\frac{1}{d}|\langle x_i, x_j\rangle| \geq t\right) \leq 2e^{-cdt^2}\,.$$

Thus,

$$\mathbb{P}\left(\forall i \neq j,\ \frac{1}{d}|\langle x_i, x_j\rangle| \leq t\right) \geq 1 - n(n-1)e^{-cdt^2}\,.$$

Then, taking $t = d^{-1/2}4\ln(n)/c$, we have:

$$\mathbb{P}\left(\forall i \neq j,\ \frac{1}{d}|\langle x_i, x_j\rangle| \leq \frac{4\ln(n)}{c\sqrt{d}}\right) \geq 1 - \frac{1}{n^2}\,.$$

For $i \in [n]$, $\langle \mathbf{1}, z_i\rangle \sim \mathcal{N}(0,d)$ so that $\mathbb{P}(|\langle \mathbf{1}, z_i\rangle| \geq t) \leq 2e^{-t^2/(2d)}$, and

$$\mathbb{P}(\forall i \in [n],\ |\langle \mathbf{1}, z_i\rangle| \leq t) \geq 1 - 2ne^{-t^2/(2d)} = 1 - \frac{2}{n^2}\,,$$

for $t = 3\sqrt{d}\ln(n)$. Hence, with probability $1 - 3/n^2$, for all $i \neq j$ we have $|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle| \leq (\sigma^2 + \sigma\mu)C\ln(n)/\sqrt{d}$.

Now,

$$\begin{aligned}\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big)x_i x_j^\top &= \frac{(b-1)}{bn(n-1)}\sum_{i<j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big)(x_i x_j^\top + x_j x_i^\top) \\ &\preceq \frac{(b-1)}{bn(n-1)}\sum_{i<j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big|\big(x_i x_i^\top + x_j x_j^\top\big)\,,\end{aligned}$$

where we used $x_i x_j^\top + x_j x_i^\top \preceq x_i x_i^\top + x_j x_j^\top$. Thus,

$$\begin{aligned}\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big)x_i x_j^\top &\preceq \sup_{i\neq j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big)\big| \times \frac{(b-1)}{bn(n-1)}\sum_{i<j}\big(x_i x_i^\top + x_j x_j^\top \\ &= \sup_{i\neq j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big| \times \frac{b-1}{b}\frac{1}{n-1}\sum_{i=1}^n x_i x_i^\top \\ &= \sup_{i\neq j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big| \times \frac{b-1}{b}\frac{n}{n-1}H\,.\end{aligned}$$

Similarly, we have

$$\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big)x_i x_j^\top \succeq -\sup_{i\neq j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big)\big| \times \frac{b-1}{b}\frac{n}{n-1}H\,.$$

Hence, with probability $1 - 3/n^2$,

$$\begin{aligned}-\frac{(\sigma^2 + \sigma\mu)C\ln(n)}{\sqrt{d}} \times \frac{b-1}{b}\frac{n}{n-1}H &\preceq \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle\big)x_i x_j^\top \\ &\preceq \frac{(\sigma^2 + \sigma\mu)C\ln(n)}{\sqrt{d}} \times \frac{b-1}{b}\frac{n}{n-1}H\,.\end{aligned}$$

We thus have shown that this term (the one in the middle of the above inequality) is of smaller order.

We are hence left with $\frac{(b-1)}{bn(n-1)}\mu^2 d\sum_{i\neq j} x_i x_j^\top$. Denoting $\bar{x} = \frac{1}{n}\sum_i x_i$, we have $\frac{1}{n^2}\sum_{i\neq j} x_i x_j^\top = \frac{1}{n^2}\sum_{i,j} x_i x_j^\top - \frac{1}{n^2}\sum_i x_i x_i^\top$, so that:

$$\frac{(b-1)}{bn(n-1)}\mu^2 d\sum_{i\neq j} x_i x_j^\top = \frac{(b-1)n}{b(n-1)}\mu^2 d\left(\bar{x}\bar{x}^\top - \frac{1}{n}H\right)\,.$$

We note that we have $H = \frac{1}{n}\sum_i x_i x_i^\top = \frac{1}{n^2}\sum_{i<j} x_i x_i^\top + x_j x_j^\top \succeq \frac{1}{n^2}\sum_{i<j} x_i x_j^\top + x_j x_i^\top = \bar{x}\bar{x}^\top$ using $x_i x_i^\top + x_j x_j^\top \succeq x_i x_j^\top + x_j x_i^\top$. Thus, $H \succeq \bar{x}\bar{x}^\top \succeq 0$, and:

$$-\frac{(b-1)n}{b(n-1)}\mu^2 d\frac{1}{n}H \preceq \frac{(b-1)}{bn(n-1)}\mu^2 d\sum_{i\neq j} x_i x_j^\top \preceq \frac{(b-1)n}{b(n-1)}\mu^2 d(1-1/n)H\,.$$

We are now able to wrap everything together. With probability $1 - 2ne^{-d/16} - 3/n^2$, we have, for some numerical constants $c_1, c_2, c_3, C > 0$:

$$\left(c_1\frac{d(\mu^2+\sigma^2)}{b} - c_2\frac{(\sigma^2+\mu^2)\ln(n)}{\sqrt{d}} - c_3\frac{\mu^2 d}{n}\right)H \preceq \tilde{H}_b \preceq C\left(\frac{d(\mu^2+\sigma^2)}{b} + \frac{(\sigma^2+\mu^2)\ln(n)}{\sqrt{d}} + \mu^2 d\right)$$

□

Appendix D

Appendix for Chapter 9

D.1 Additional notations and comments on discretisation methods

Vector Operations. Moving forward, all arithmetic operations and real-valued functions will be considered as being applied coordinate-wise. In other words, if a and b are vectors in $\mathbb{R}^d$ and $p, q \in \mathbb{Q}$, then $a^p b^q \in \mathbb{R}^d$ will be used as a shorthand for the vector with entries $\{a_i^p b_i^q\}_{i=1}^d$. And for any $f : \mathbb{R} \to \mathbb{R}$, $f(a)$ will represent the vector with entries $\{f(a_i)\}_{i=1}^d$. Inequalities between vectors will also be interpreted as holding coordinate-wise.

Mirror Maps. Various definitions of a *mirror map* $\Phi : \mathbb{R}^d \to (-\infty, +\infty]$ exist in the optimization literature [see Nemirovski, 1979], and a common one coincides with the concept of a Legendre function [see Bauschke et al., 2017, Bauschke and Borwein, 1997]. In our proofs, we do not deal with extended real-valued functions, and the term mirror map is applied to C^∞-smooth strictly convex functions with coercive gradients. In particular, our mirror maps are of Legendre type.

For such a mirror map $\Phi : \mathbb{R}^d \to \mathbb{R}$, we define the Bregman divergence $D_\Phi(\theta_1, \theta_2)$ for $\theta_1, \theta_2 \in \mathbb{R}^d$ as

$$D_\Phi(\theta_1, \theta_2) = \Phi(\theta_1) - \Phi(\theta_2) - \langle \nabla\Phi(\theta_2), \theta_1 - \theta_2 \rangle.$$

Notice that due to the strict convexity of Φ, $D_\Phi(\theta_1, \theta_2) > 0$ whenever $\theta_1 \neq \theta_2$.

Modified Cauchy Principal Value. Let $f : \mathbb{R}_{\geq 0} \to [-\infty, +\infty]$ be an extended real-valued function with a finite set of poles $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$ (*i.e.* points $t \in \mathbb{R}_{\geq 0}$ at which $f(t) = \pm\infty$) such that f is continuous on $\mathbb{R}_{\geq 0} \setminus \mathcal{T}$. Let $0 < T_1 < \cdots < T_N$. Let $T \in \mathcal{T}$ and let $\varepsilon > 0$ be small enough such that $(T - \varepsilon, T + \varepsilon) \cap \mathcal{T} = \{T\}$. Recall that, provided the limit below exists, the Cauchy principal value p.v. $\int_{T-\varepsilon}^{T+\varepsilon} f(t)\mathrm{d}t$ is defined as

$$\text{p.v.} \int_{T-\varepsilon}^{T+\varepsilon} f(t)\mathrm{d}t \coloneqq \lim_{\delta \to 0} \left[\int_{T-\varepsilon}^{T-\delta} f(t)\mathrm{d}t + \int_{T+\delta}^{T+\varepsilon} f(t)\mathrm{d}t \right].$$

Now, let $\varepsilon_m > 0$ be such that $(T_m - \varepsilon_m, T_m + \varepsilon_m) \cap \mathcal{T} = \{T_m\}$ for $m \in [N]$. Moreover, let $T_0 = \varepsilon_0 = 0$ and $T_{N+1} = +\infty$. Suppose f has finite Cauchy principal values at all poles. Then, for any $\tau \geq 0$ such that $\tau \notin \mathcal{T}$, we could define p.v. $\int_0^\tau f(t)\mathrm{d}t$ as

$$\text{p.v.} \int_0^\tau f(t)\mathrm{d}t \coloneqq \sum_{m : T_{m+1} < \tau} \left[\text{p.v.} \int_{T_m - \varepsilon_m}^{T_m + \varepsilon_m} f(t)\mathrm{d}t + \int_{T_m + \varepsilon_m}^{T_{m+1} - \varepsilon_{m+1}} f(t)\mathrm{d}t \right] + \text{p.v.} \int_{T_k - \varepsilon_k}^{T_k + \varepsilon_k} f(t)\mathrm{d}t + \int_{T_k + \varepsilon_k}^{\tau} f(t)\mathrm{d}t,$$

where $T_k < \tau < T_{k+1}$.

For our proofs of Lemma 2 and Theorem 3, we require a modification to the Cauchy principal value. For the aforementioned function f with the described properties and for $T \in \mathcal{T}$, $\varepsilon > 0$

such that $(T-\varepsilon, T+\varepsilon) \cap \mathcal{T} = \{T\}$, we define the modified principal value m.p.v. $\int_{T-\varepsilon}^{T+\varepsilon} f(t)\mathrm{d}t$ as

$$\text{m.p.v.} \int_{T-\varepsilon}^{T+\varepsilon} f(t)\mathrm{d}t \coloneqq \lim_{\delta \to 0} \left[\int_{T-\varepsilon}^{T-\delta} f(t)\mathrm{d}t \cdot e^{\frac{\delta}{\lambda}} + \int_{T+\delta}^{T+\varepsilon} f(t)\mathrm{d}t \cdot e^{-\frac{\delta}{\lambda}} \right], \tag{D.1}$$

where λ denotes our familiar MGF parameter. We also extend the m.p.v. definition to integrals $\int_0^\tau f(t)\mathrm{d}t$ for arbitrary $\tau \geq 0$ by mimicking the Cauchy-principal-value construction:

$$\begin{aligned}\text{m.p.v.} \int_0^\tau f(t)\mathrm{d}t \coloneqq \sum_{m: T_{m+1} < \tau} \left[\text{m.p.v.} \int_{T_m - \varepsilon_m}^{T_m + \varepsilon_m} f(t)\mathrm{d}t + \int_{T_m + \varepsilon_m}^{T_{m+1} - \varepsilon_{m+1}} f(t)\mathrm{d}t \right] \\ + \text{m.p.v.} \int_{T_k - \varepsilon_k}^{T_k + \varepsilon_k} f(t)\mathrm{d}t + \int_{T_k + \varepsilon_k}^{\tau} f(t)\mathrm{d}t,\end{aligned}$$

where $T_k < \tau < T_{k+1}$. Note that the above definition implies that whenever f has no poles on an interval $(a, b) \subset \mathbb{R}_{\geq 0}$, then

$$\text{m.p.v.} \int_a^b f(t)\mathrm{d}t = \int_a^b f(t)\mathrm{d}t.$$

Additional Comments on the Discretisation of MGF(λ). Following our discussion from Section 9.3, we want to point out that that there are other ways of discretising

$$\lambda \ddot{w}_t + \dot{w}_t + \nabla F(w_t) = 0.$$

Indeed, instead of discretising as (9.2) in the chapter:

$$\lambda \, \frac{w_{k+1} - 2w_k + w_{k-1}}{\varepsilon^2} + \frac{w_k - w_{k-1}}{\varepsilon} + \nabla F(w_k) = 0,$$

one could also consider a central first-order difference:

$$\lambda \, \frac{w_{k+1} - 2w_k + w_{k-1}}{\varepsilon^2} + \frac{w_{k+1} - w_{k-1}}{2\varepsilon} + \nabla F(w_k) = 0.$$

Rearranging, this leads to

$$w_{k+1} = w_k - \frac{\varepsilon^2}{\lambda(1 + \frac{\varepsilon}{2\lambda})} \nabla F(w_k) + \frac{1 - \frac{\varepsilon}{2\lambda}}{1 + \frac{\varepsilon}{2\lambda}} (w_k - w_{k-1}),$$

which corresponds to momentum with $\gamma = \frac{\varepsilon^2}{\lambda(1+\frac{\varepsilon}{2\lambda})}$ and $\beta = \frac{1-\frac{\varepsilon}{2\lambda}}{1+\frac{\varepsilon}{2\lambda}}$. Solving for ε and λ, we get

$$\lambda = \frac{(1+\beta)\gamma}{2(1-\beta)^2} \qquad \text{and} \qquad \varepsilon = \frac{\gamma}{1-\beta}.$$

Hence, we obtain the same discretisation step ε as in Proposition 19 and a slightly different expression for λ. However, note that the two versions of λ become indistinguishable for large values of β since $\frac{1+\beta}{2} \to_{\beta \to 1} 1$. Experimentally, running MGF(λ) with the two different values for λ leads to similar results. Thus, the discretisation scheme was chosen due to the more concise definition of λ in this case.

D.2 (w_+, w_-)-reparametrisation

MGF Reparametrisation. We recall that we consider momentum gradient flow MGF(λ) with parameter $\lambda > 0$ over the diagonal-linear-network loss $F((u, v))) = \mathcal{L}(u \odot v)$:

$$\lambda \ddot{u}_t + \dot{u}_t + \nabla L(\theta_t) \odot v_t = 0;$$
$$\lambda \ddot{v}_t + \dot{v}_t + \nabla L(\theta_t) \odot u_t = 0.$$

For proof-writing convenience, we consider the simple reparametrisation outlined below.

In order to eliminate the cross-dependencies in (u, v) in the above equations, it is natural to consider the quantities $(w_{+,t}, w_{-,t})$ where $w_{\pm,t} = u_t \pm v_t$ for $t \geq 0$. Hence, we get the following reparametrised ODE:

$$\begin{cases} \lambda \ddot{w}_{\pm,t} + \dot{w}_{\pm,t} \pm \nabla \mathcal{L}(\theta_t) \odot w_{\pm,t} = 0; \\ w_{\pm,0} = u_0 \pm v_0, \quad \dot{w}_{\pm,0} = 0. \end{cases} \tag{D.2}$$

Notice that with these new quantities, we have

$$\theta_t = \frac{w_{+,t}^2 - w_{-,t}^2}{4} \quad \text{and} \quad \Delta_t = |w_{+,t} w_{-,t}|.$$

MGD Reparametrisation. For the discrete-time setting, we follow the same reparametrisation from the MGD recursion:

$$u_{k+1} = u_k - \gamma \nabla L(\theta_k) \odot v_k + \beta(u_k - u_{k-1});$$
$$v_{k+1} = v_k - \gamma \nabla L(\theta_k) \odot u_k + \beta(v_k - v_{k-1}).$$

We let $w_{\pm,k} = u_k \pm v_k$ for $k \geq 0$. Then, for $k \geq 1$, the equations above transform into

$$\begin{cases} w_{\pm,k+1} = w_{\pm,k} \mp \gamma \nabla \mathcal{L}(\theta_k) \odot w_{\pm,k} + \beta(w_{\pm,k} - w_{\pm,k-1}); \\ w_{\pm,1} = w_{\pm,0} = u_0 \pm v_0. \end{cases} \tag{D.3}$$

Again, with the newly defined quantities, we have

$$\theta_k = \frac{w_{+,k}^2 - w_{-,k}^2}{4} \quad \text{and} \quad \Delta_k = |w_{+,k} w_{-,k}|.$$

D.3 Continuous-time theorems

D.3.1 Convergence of momentum gradient flow

Momentum gradient flow (with $\lambda > 0$),

$$\lambda \ddot{w}_t + \dot{w}_t + \nabla F(w_t) = 0,$$

also known in the optimisation literature as the heavy-ball with friction ODE or the heavy-ball dynamical system with constant damping coefficient, has been the object of extensive mathematical study over the years [Haraux and Jendoubi, 1998, Attouch et al., 2000, Alvarez, 2000, Goudou and Munier, 2009, Polyak and Shcherbakov, 2017, Apidopoulos et al., 2022]. If we abstract away from the diagonal linear network setting and consider an unspecified loss $F \in C^1(\mathbb{R}^D, \mathbb{R}_{\geq 0})$ with

locally Lipschitz gradient, we can still identify a useful Lyapunov function, which perhaps motivated the study of the ODE in the first place. The function in question happens to be the energy of the system

$$E_t = F(w_t) + \frac{\lambda}{2}\|\dot{w}_t\|_2^2, \tag{D.4}$$

whose nonpositive time-derivative $\dot{E}_t = -\|\dot{w}_t\|_2^2$ allows us to prove the global existence and uniqueness of a solution to MGF [Attouch et al. [2000], Theorem 3.1] in this more general setting. We note that by an easy inductive argument, when the function F is C^k-smooth, the MGF solution w_t is C^{k+1}-smooth. Hence, in our setting where the diagonal-neural-network loss F is C^∞-smooth, the learning trajectory w_t is also C^∞-smooth.

Convergence under Assumption 7.

Under the assumption of a bounded trajectory – $w_t \in L^\infty(0,\infty)$, one can prove the following convergences [Attouch et al., 2000]:

$$\lim_{t\to\infty} \dot{w}_t = \lim_{t\to\infty} \nabla F(w_t) = 0.$$

However, even when bounded, the iterates w_t need not converge as demonstrated by the coercive function from Section 4.3 in [Attouch et al., 2000]. Nevertheless, when the loss F is also analytic, as in the case of diagonal linear networks, assuming boundedness, one can further prove iterate convergence $\lim_{t\to\infty} w_t = w_\infty$ [Haraux and Jendoubi [1998]].

Unfortunately, without assuming boundedness, iterate convergence has been established only in the cases of convex loss [Alvarez [2000]], quasiconvex loss [Goudou and Munier [2009]], and loss satisfying the Polyak-Lojasiewicz inequality [Apidopoulos et al. [2022]]. Thus, the square loss for a diagonal linear network (and neural networks in general) falls out of the scope of these few favorable cases due to non-convexity and an abundance of local and global minima. For that reason, we posit Assumption 7, which holds true empirically in all our experiments on diagonal linear networks.

Convergence to 0 Loss under Assumption 8.

Let us now go back to the specific case of diagonal linear networks where the loss is given by $F(w) = \mathcal{L}(u \odot v)$ for $w = (u, v)$. Notice that from the discussion above, if we assume boundedness of the trajectory, we have

$$\lim_{t\to\infty} \nabla F(w_t) = (\nabla\mathcal{L}(\theta_\infty) \odot v_\infty, \nabla\mathcal{L}(\theta_\infty) \odot u_\infty) = 0.$$

Therefore, since $\nabla\mathcal{L}(\theta_\infty) \odot \Delta_\infty = 0$, if the balancedness at infinity Δ_∞ has nonzero coordinates, we can conclude that $\nabla\mathcal{L}(\theta_\infty) = 0$. Recalling that $\mathcal{L}$ is convex, we get that $\mathcal{L}(\theta_\infty) = 0$. Hence, θ_∞ interpolates the dataset.

D.3.2 Proof of time-varying momentum mirror flow

In our discussion in Appendix D.3.1, we saw that assuming

1) iterate boundedness: $u_t, v_t \in L^\infty(0,\infty)$, and

2) nonzero balancedness at infinity: $\Delta_{\infty,i} \neq 0, \ \ \forall i \in [d]$,

we can prove that MGF over a diagonal linear network (9.4) converges to an interpolator θ_∞.[1] Before we jump into the proof of Proposition 21, we need to establish the following lemma.

[1] Note that we also refer to θ_∞ as θ^{MGF}.

Lemma 34. *Assuming that $u_t, v_t \in L^\infty(0,\infty)$ and $\Delta_{\infty,i} \neq 0,\ \forall i \in [d]$, the following integral limit exists:*

$$\lim_{t\to\infty} \int_0^t \nabla\mathcal{L}(\theta_s)ds = \int_0^\infty \nabla\mathcal{L}(\theta_t)\mathrm{d}t.$$

Consequently,

$$\lim_{t\to\infty} \int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}ds = 0.$$

Proof. Let us consider the (w_+, w_-)-reparametrisation of MGF (9.4) given by Equation (D.2):

$$\lambda\ddot{w}_{\pm,t} + \dot{w}_{\pm,t} \pm \nabla\mathcal{L}(\theta_t) \odot w_{\pm,t} = 0.$$

Since we assumed that Δ_∞ has nonzero coordinates, there exists $T \geq 0$ such that for all $t \geq T$, $w_{\pm,t}$ have nonzero coordinates. Hence, for $t \geq T$, we can safely divide by $w_{\pm,t}$ to obtain

$$\lambda\frac{\mathrm{d}^2 \ln|w_{\pm,t}|}{\mathrm{d}t^2} + \frac{\mathrm{d}\ln|w_{\pm,t}|}{\mathrm{d}t} \pm \nabla\mathcal{L}(\theta_t) + \lambda\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 = 0.$$

Let us notice a couple of things. First, as we discussed in Appendix D.3.1, the boundedness of the iterates forces $w_{\pm,t}$ to converge to some vectors $w_{\pm,\infty}$ with nonzero coordinates since we assumed the coordinates of $\Delta_\infty = w_{+,\infty}w_{-,\infty}$ are nonzero. Hence,

$$\|\frac{\dot{w}^2_{\pm,t}}{w^2_{\pm,t}}\|_\infty \leq \texttt{const}\cdot(\|\dot{u}_t^2\|_\infty + \|\dot{v}_t^2\|_\infty),$$

where the RHS is integrable as we saw in the proof of Section 9.4.3. Second, from the discussion in Appendix D.3.1, we know that $\lim_{t\to\infty}\dot{w}_{\pm,t} = 0$, so $\lim_{t\to\infty}\frac{\mathrm{d}\ln|w_{\pm,t}|}{\mathrm{d}t} = 0$.

Now, for $t \geq T$,

$$\begin{aligned}\int_T^t \nabla\mathcal{L}(\theta_s)ds &= \mp\left(\lambda\int_T^t \frac{\mathrm{d}^2\ln|w_{\pm,s}|}{\mathrm{d}t^2}ds + \int_T^t \frac{\mathrm{d}\ln|w_{\pm,s}|}{\mathrm{d}t}ds + \int_T^t\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 ds\right)\\ &= \mp\left(\lambda\frac{\mathrm{d}\ln|w_{\pm,s}|}{\mathrm{d}t}\Big|_T^t + \ln|w_{\pm,s}|\Big|_T^t + \int_T^t\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 ds\right).\end{aligned}$$

So, using the above observations and letting $t \to \infty$ yields

$$\lim_{t\to\infty}\int_T^t \nabla\mathcal{L}(\theta_s)ds = \mp\left(-\lambda\frac{\mathrm{d}\ln|w_{\pm,T}|}{\mathrm{d}t} - \ln|w_{\pm,T}| + \ln|w_{\pm,\infty}| - \int_T^\infty\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2\mathrm{d}t\right).$$

Thus, we conclude that $\lim_{t\to\infty}\int_0^t \nabla\mathcal{L}(\theta_s)ds$ exists, and therefore, $\lim_{t\to\infty}\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}ds = 0$. □

We are now well-equipped to prove Proposition 21. We note that we phrased Proposition 21 rather succinctly in the chapter due to space considerations. In what follows, we restate Proposition 21 by precisely specifying the underlying assumptions.

Proposition. *Assume the solution (u_t, v_t) of MGF (9.4) is bounded. If we also assume that the balancedness at infinity Δ_∞ has nonzero coordinates, then there exists a time $T \geq 0$, after which the predictors $\theta_t = u_t \odot v_t$ follow a momentum mirror flow with time-varying potentials Φ_t:*

$$\lambda\frac{\mathrm{d}^2\nabla\Phi_t(\theta_t)}{\mathrm{d}t^2} + \frac{\mathrm{d}\nabla\Phi_t(\theta_t)}{\mathrm{d}t} + \nabla\mathcal{L}(\theta_t) = 0.$$

Furthermore, if we assume that the balancedness Δ_t remains nonzero for $t \in [0, +\infty]$, then the momentum mirror flow holds for every $t \geq 0$.

Proof. We will consider the (w_+, w_-)-reparametrisation of momentum gradient flow (9.4) introduced in Appendix D.2. For convenience of the reader, we recall this reparametrisation here:

$$\lambda \ddot{w}_{\pm,t} + \dot{w}_{\pm,t} \pm \nabla \mathcal{L}(\theta_t) \odot w_{\pm,t} = 0.$$

Now, let $\xi : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be the $C^\infty(0, \infty)$ solution of the following ODE:

$$\lambda \ddot{\xi}_t + \dot{\xi}_t + \nabla \mathcal{L}(\theta_t) = 0,$$

with the constraint $\xi_0 = \dot{\xi}_0 = 0$. Hence, by Lemma 36,

$$\xi_t = -\int_0^t \nabla \mathcal{L}(\theta_s)(1 - e^{-\frac{t-s}{\lambda}}) \mathrm{d}s,$$

and by Lemma 34,

$$\xi_\infty = -\int_0^\infty \nabla \mathcal{L}(\theta_t) \mathrm{d}t.$$

Thus, $\xi_t \in \mathtt{span}(x_1, \ldots, x_n)$, $\forall t \in [0, +\infty]$.

Having fixed ξ_t, we define the quantities $\alpha_{\pm,t}$ for every $t \in [0, +\infty]$ through the following relation:

$$\alpha_{\pm,t} = w_{\pm,t} \exp(\mp \xi_t).$$

So, $\Delta_t = |w_{+,t} w_{-,t}| = |\alpha_{+,t} \alpha_{-,t}|$. Furthermore,

$$\begin{aligned} \theta_t &= \frac{1}{4}(w_{+,t}^2 - w_{-,t}^2) \\ &= \frac{1}{4}(\alpha_{+,t}^2 \exp(2\xi_t) - \alpha_{-,t}^2 \exp(-2\xi_t)) \\ &= \frac{1}{2}\Delta_t \sinh\left(2\xi_t + \ln \frac{|\alpha_{+,t}|}{|\alpha_{-,t}|}\right) \end{aligned}$$

Since we assumed that Δ_∞ has nonzero coordinates, there exists $T \geq 0$ such that for all $t \geq T$, $w_{\pm,t}$ have nonzero coordinates. Hence, for $t \geq T$, the logarithm $\ln \frac{|\alpha_{+,t}|}{|\alpha_{-,t}|}$ is well-defined. If we assume positive balancedness for $t \in [0, +\infty]$, then we can choose $T = 0$. From now until the end of the proof, whenever a time-dependent quantity features division by Δ_t, we will tacitly assume that $t \geq T$.

Let us now introduce the helper quantity ϕ_t through the following identity:

$$\phi_t = \frac{1}{2} \ln \frac{|\alpha_{+,t}|}{|\alpha_{-,t}|} = \frac{1}{2} \mathrm{arcsinh}\left(\frac{\alpha_{+,t}^2 - \alpha_{-,t}^2}{2\Delta_t}\right).$$

Then,

$$\frac{1}{2} \mathrm{arcsinh}\left(\frac{2\theta_t}{\Delta_t}\right) - \phi_t = \xi_t \in \mathtt{span}(x_1, \ldots, x_n).$$

So, if we consider the time-varying potential

$$\begin{aligned} \Phi_t(\theta) &= \frac{1}{4} \sum_{i=1}^d \left(2\theta_i \mathrm{arcsinh}\left(\frac{2\theta_i}{\Delta_{t,i}}\right) - \sqrt{4\theta_i^2 + \Delta_{t,i}^2} + \Delta_{t,i}\right) - \langle \phi_t, \theta \rangle \\ &= \psi_{\Delta_t}(\theta) - \langle \phi_t, \theta \rangle, \end{aligned} \tag{D.5}$$

where ψ_{Δ_t} is the hyperbolic entropy defined in Equation (9.6), then

$$\nabla \Phi_t(\theta) = \frac{1}{2} \mathrm{arcsinh}\left(\frac{2\theta}{\Delta_t}\right) - \phi_t.$$

Notice that $\nabla^2\Phi_t = \texttt{diag}\left(1/\sqrt{4\theta^2+\Delta_t^2}\right) \succ 0$. Hence, Φ_t is a mirror map. Furthermore, $\nabla\Phi_t(\theta_t) = \xi_t$ for $t \geq T$, so

$$\lambda\frac{\mathrm{d}^2\nabla\Phi_t(\theta_t)}{\mathrm{d}t^2} + \frac{\mathrm{d}\nabla\Phi_t(\theta_t)}{\mathrm{d}t} + \nabla\mathcal{L}(\theta_t) = 0.$$

□

D.3.3 Proof of Theorem 3

We are now ready to prove our main result for the implicit bias of momentum gradient flow on diagonal linear networks.

Theorem 3. *The solution θ^{MGF} of MGF (9.4) interpolates the dataset and satisfies the following implicit regularisation:*

$$\theta^{MGF} = \underset{\theta^\star\in\mathcal{S}}{\arg\min}\ \ D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

In the above expression, $D_{\psi_{\Delta_\infty}}$ denotes the Bregman divergence with potential ψ_{Δ_∞}, where the asymptotic balancedness equals

$$\Delta_\infty = \Delta_0 \odot \exp\big(-(I_+ + I_-)\big)$$

and $\tilde{\theta}_0 = \frac{1}{4}(w_{+,0}^2 \odot \exp(-2I_+) - w_{-,0}^2 \odot \exp(-2I_-))$ denotes a perturbed initialisation term.

We split the proof into two parts for conceptual clarity. In the first part, we utilise the time-varying mirror flow from Proposition 21 to derive the implicit regularisation $\theta^{\text{MGF}} = \arg\min_{\theta^\star\in\mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$. Then, in the second part, we prove that the integral quantities $I_\pm$ from Lemma 2 are well-defined, and we give the trajectory-dependent characterisations of the asymptotic balancedness Δ_∞ and the perturbed initialisation $\tilde{\theta}_0$.

Proof of Implicit Regularisation.

In Proposition 21, we proved that whenever the MGF trajectory is bounded and the coordinates of Δ_∞ are nonzero, there exists a time $T \geq 0$, after which the predictors θ_t follow a momentum mirror flow with potentials given by Equation (D.5). Recall that for $t \geq T$,

$$\nabla\Phi_t(\theta_t) = \frac{1}{2}\text{arcsinh}\left(\frac{2\theta}{\Delta_t}\right) - \phi_t = -\xi_t \in \texttt{span}(x_1, \ldots, x_n).$$

where $\xi_t = -\int_0^t \nabla\mathcal{L}(\theta_s)(1 - e^{-\frac{t-s}{\lambda}})ds$, $\alpha_{\pm,t} = w_{\pm,t}\exp(\mp\xi_t)$, and $\phi_t = \frac{1}{2}\text{arcsinh}\left(\frac{\alpha_{+,t}^2 - \alpha_{-,t}^2}{2\Delta_t}\right)$.

Now, as $t \to \infty$, ξ_t and the MGF iterates converge, so we know that $\nabla\Phi_\infty(\theta_\infty) \in \texttt{span}(x_1, \ldots, x_n)$, where $\Phi_\infty(\theta) = \psi_{\Delta_\infty}(\theta) - \langle\phi_\infty, \theta\rangle$. Thus, we can use the familiar Bregman-Cosine-Theorem trick to characterise the interpolator θ_∞. We proceed with this characterisation.

Let $\tilde{\theta}_0$ be a perturbation term such that $\nabla\Phi_\infty(\tilde{\theta}_0) = 0$. Equivalently,

$$\frac{1}{2}\text{arcsinh}\left(\frac{2\tilde{\theta}_0}{\Delta_\infty^2}\right) - \phi_\infty = 0 \iff$$

$$\text{arcsinh}\left(\frac{2\tilde{\theta}_0}{\Delta_\infty^2}\right) - \text{arcsinh}\left(\frac{\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2}{2\Delta_\infty^2}\right) = 0 \iff$$

$$\tilde{\theta}_0 = \frac{\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2}{4}.$$

Note that $\alpha_{\pm,\infty} = w_{\pm,\infty}\exp(\pm\int_0^\infty \nabla\mathcal{L}(\theta_t)\mathrm{d}t)$ by Lemma 34 and $\Delta_\infty = |\alpha_{+,\infty}\alpha_{-,\infty}|$.

Now, let $\theta^\star \in \mathcal{S}$ be an arbitrary interpolator of the dataset. Then, $\theta^\star - \theta_\infty \in \ker(X) = \mathtt{span}(x_1,\dots,x_n)^\perp$. Hence, the Bregman Cosine Theorem yields

$$\begin{aligned} D_{\Phi_\infty}(\theta^\star,\tilde\theta_0) &= D_{\Phi_\infty}(\theta^\star,\theta_\infty) + D_{\Phi_\infty}(\theta_\infty,\tilde\theta_0) + \langle\theta^\star-\theta_\infty, \nabla\Phi(\theta_\infty)-\nabla\Phi(\tilde\theta_0)\rangle \\ &= D_{\Phi_\infty}(\theta^\star,\theta_\infty) + D_{\Phi_\infty}(\theta_\infty,\tilde\theta_0), \end{aligned}$$

where we used that $\nabla\Phi(\theta_\infty)-\nabla\Phi(\tilde\theta_0) \in \mathtt{span}(x_1,\dots,x_n)$. Thus,

$$\begin{aligned} \theta_\infty &= \arg\min_{\theta^\star\in\mathcal{S}} D_{\Phi_\infty}(\theta^\star,\tilde\theta_0) \\ &= \arg\min_{\theta^\star\in\mathcal{S}} \Phi_\infty(\theta^\star). \end{aligned}$$

Finally, notice that $\nabla\psi_{\Delta_\infty}(\tilde\theta_0) = \frac{1}{2}\mathrm{arcsinh}\left(\frac{2\tilde\theta_0}{\Delta_\infty}\right) = \phi_\infty$ as we showed above. Hence,

$$D_{\psi_{\Delta_\infty}}(\theta,\tilde\theta_0) = \Phi_\infty(\theta) - \psi_{\Delta_\infty}(\tilde\theta_0) + \langle\nabla\psi_{\Delta_\infty}(\tilde\theta_0),\tilde\theta_0\rangle.$$

Thus, we conclude that

$$\theta_\infty = \arg\min_{\theta^\star\in\mathcal{S}} \Phi_\infty(\theta^\star) = \arg\min_{\theta^\star\in\mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star,\tilde\theta_0).$$

□

Proof of Trajectory-Dependent Characterisation.

We just showed that the recovered interpolator by MGF solves the constrained minimisation problem $\theta_\infty = \arg\min_{\theta^\star\in\mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star,\tilde\theta_0)$, where $\Delta_\infty = |\alpha_{+,\infty}\alpha_{-,\infty}|$, $\tilde\theta_0 = (\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2)/4$, and $\alpha_{\pm,\infty} = w_{\pm,\infty}\exp(\pm\int_0^\infty \nabla\mathcal{L}(\theta_t)\mathrm{d}t)$. Clearly, these opaque characterisations of Δ_∞ and $\tilde\theta_0$ prevent us from describing how the magnitude of these quantities compares to the magnitude of the initial balancedness Δ_0 and the initialisation scale $\alpha = \max(|u_0|,|v_0|)$. Ideally, we would like to find formulas for Δ_∞ and $\tilde\theta_0$ which show that $\tilde\theta_0 \ll \theta^\star$, $\forall\theta^\star\in\mathcal{S}$ and $\Delta_\infty < \Delta_0$ so that we can conclude that $\theta^{\mathrm{MGF}} \approx \arg\min_{\theta^\star\in\mathcal{S}}\psi_{\Delta_\infty}(\theta^\star)$ enjoys better sparsity guarantees than $\theta^{\mathrm{GF}} \approx \arg\min_{\theta^\star\in\mathcal{S}}\psi_{\Delta_0}(\theta^\star)$. In what follows, we derive such formulas.

In our subsequent arguments, for a vector $z\in\mathbb{R}^d$ and a coordinate $i\in[d]$, we will denote with $z(i)$ the i^{th} coordinate of z in order to reduce the index bloat. Let us consider again the (w_+,w_-)-reparametrisation of MGF discussed in Appendix D.2:

$$\lambda\ddot w_{\pm,t} + \dot w_{\pm,t} \pm \nabla\mathcal{L}(\theta_t)\odot w_{\pm,t} = 0. \tag{D.6}$$

Notice that if for some $T>0$ and $i\in[d]$, $w_{+,T}(i)=0$, then $\dot w_{+,T}(i)$ must be nonzero. Indeed, as we argued in Appendix D.3.1, MGF (9.4) admits a unique global solution. And if $w_{+,T}(i) = \dot w_{+,T}(i) = 0$, then we could construct another solution $(w'_{+,t}, w'_{-,t})$ of MGF such that $w'_{+,t}(i) = \dot w'_{+,t}(i) = 0$, $\forall t\geq 0$, and $w_{\pm,T} = w'_{\pm,T}$. By uniqueness, we get that $w_{\pm,t} = w'_{\pm,t}$, $\forall t\geq 0$ However, the newly constructed solution will not be consistent with the imposed initialisation $\Delta_0 \neq 0$,. Hence, $\Delta_0\neq 0$ prevents $w_{+,t}(i)$ and $\dot w_+(i)$ from hitting 0 simultaneously. Similarly, this situation cannot occur for $w_{-,t}$.

Until further notice, we fix a coordinate $i\in[d]$ and consider eq. (D.6) only in the i^{th} coordinate without explicit mention. If $w_{\pm,T}=0$ for some $T>0$, then $\dot w_{\pm,T}\neq 0$. Hence, for some small ¿0, $\dot w_{\pm,t}$ does not change sign on $[T-\dot{,}T+\dot{]}$, so $w_{\pm,t}$ either strictly increases or decreases on $[T-\dot{,}T+\dot{]}$. Therefore, $w_{\pm,t}\neq 0$ on $[T-\dot{,}T+\dot{]}\setminus\{T\}$ implying that $w_{\pm,t}$ equals 0 at

most a countable number of times. Recall that by Assumption 8, there exists a time T_∞ after which $w_\pm$ does not change sign. Therefore, if we assume that $w_\pm$ vanishes on infinitely many points $T_1 < T_2 < \cdots < T_\infty$, then by compactness, the limit $\tau = \lim_{m\to\infty} T_m$ exists. Since $w_\pm$ is continuous, we infer that $w_{\pm,\tau} = 0$. Moreover, by the Mean Value Theorem, for every $m \geq 1$, there exists $T'_m \in (T_m, T_{m+1})$ such that $\dot{w}_{T'_m} = 0$. Notice that $\lim_{m\to\infty} T'_m = \tau$ as well. Hence, by continuity, $w_{\pm,\tau} = \dot{w}_{\pm,\tau} = 0$ – a contradiction.

Hence, $w_\pm$ vanishes on a finite set of points. Let us order these vanishing times as $0 < T_1 < \cdots < T_N$ and let $T_0 = 0$ and $T_{N+1} = +\infty$. Observe that for $t \notin \mathcal{T} = \{T_i : i \in [N]\}$, we can safely divide both sides of eq. (D.6) by $w_{\pm,t}$ to obtain

$$\lambda \frac{\ddot{w}_{\pm,t}}{w_{\pm,t}} + \frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \pm \nabla\mathcal{L}(\theta_t) = 0.$$

The last expression is equivalent to

$$\lambda\left(\frac{\ddot{w}_{\pm,t}}{w_{\pm,t}} - \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2\right) + \frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \pm \nabla\mathcal{L}(\theta_t) + \lambda\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 = 0,$$

which can be rewritten as

$$\lambda \frac{\mathrm{d}^2 \ln(\mathrm{sgn}(w_{\pm,t})w_{\pm,t})}{\mathrm{d}t^2} + \frac{\mathrm{d}\ln(\mathrm{sgn}(w_{\pm,t})w_{\pm,t})}{\mathrm{d}t} \pm \nabla\mathcal{L}(\theta_t) + \lambda\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 = 0.$$

Let us define a new function $g_\pm : \mathbb{R}_{\geq 0}\backslash\mathcal{T} \to \mathbb{R}^d$ through the relation $g_{\pm,t} = \ln(\mathrm{sgn}(w_{\pm,t})w_{\pm,t})$. Then, $g_\pm$ is C^∞-smooth on $\mathbb{R}_{\geq 0} \setminus \mathcal{T}$ and satisfies the following ODE:

$$\lambda\ddot{g}_{\pm,t} + \dot{g}_{\pm,t} \pm \nabla\mathcal{L}(\theta_t) + \lambda\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 = 0. \tag{D.7}$$

Induction on Vanishing Times. Now, we proceed to prove by induction on $N - 1 \geq m \geq 0$ that for $\tau \in (T_m, T_{m+1})$ the following 3 things hold:

- The following integral quantities[2] exist and are finite:

$$\text{m.p.v.} \int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 e^{-\frac{\tau-s}{\lambda}} \mathrm{sgn}(w_{\pm,\tau}w_{\pm,t})\mathrm{d}t \quad \text{and} \quad \int_0^\tau \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \mathrm{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s\,\mathrm{d}t.$$

- The following identity holds:

$$\dot{g}_{\pm,\tau} = -\text{m.p.v.} \int_0^\tau \left[\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)\right] e^{-\frac{\tau-t}{\lambda}} \mathrm{sgn}(w_{\pm,\tau}w_{\pm,t})\mathrm{d}t - \frac{1}{\lambda}\sum_{k=1}^m (-1)^{m-k} e^{-\frac{\tau-T_k}{\lambda}}.$$

- The following identity holds:

$$g_{\pm,\tau} = g_{\pm,0} - \int_0^\tau \text{m.p.v.} \int_0^t \left[\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}} \mathrm{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s\,\mathrm{d}t$$
$$- \sum_{k=1}^m (-1)^{m-k}\left(1 - e^{-\frac{\tau-T_k}{\lambda}}\right) + 2\sum_{1\leq i<j\leq m} (-1)^{j-i}\left(1 - e^{-\frac{T_j-T_i}{\lambda}}\right).$$

[2] See Equation (D.1) for the definition of m.p.v.

Recall that $\nabla\mathcal{L}(\theta_t)$ is a bounded function, so if the modified principal value from the first bullet point exists, then the modified principal values in the above identities are also well-defined.

Base case: $m = 0$**.** Recall from the proof of Proposition 21 in Appendix D.3.2 that $\dot{w}_\pm \in L^2(0,\infty)$. Now, since $w_{\pm,t}$ does not change signs on the interval (T_0, τ), we know that $1/w_{\pm,t} = \Omega(1)$. Hence,

$$\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \in L^1(0,\infty).$$

Similarly, $\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2$ is integrable on all intervals $[T_i+\varepsilon, T_{i+1}-\varepsilon]$ for any small $\varepsilon > 0$. Consequently, the integral quantities

$$\text{m.p.v.}\int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_{\pm,\tau}w_{\pm,t})\mathrm{d}t = \int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 e^{-\frac{t-s}{\lambda}}\mathrm{d}t$$

and

$$\begin{aligned}\int_0^\tau \text{m.p.v.}\int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s\ \mathrm{d}t &= \int_0^\tau\int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}}\mathrm{d}s\ \mathrm{d}t \\ &= \int_0^\tau \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 (1-e^{-\frac{\tau-t}{\lambda}})\mathrm{d}t\end{aligned}$$

are well-defined. Moreover, after applying Lemma 36 to eq. (D.7), we get

$$\begin{aligned}\dot{g}_{\pm,\tau} &= -\int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 e^{-\frac{\tau-t}{\lambda}}\mathrm{d}t \ \mp \frac{1}{\lambda}\int_0^\tau \nabla\mathcal{L}(\theta_t)e^{-\frac{\tau-t}{\lambda}}\mathrm{d}t \\ g_{\pm,\tau} &= g_{\pm,0} - \lambda\int_0^\tau \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 (1-e^{-\frac{\tau-t}{\lambda}})\mathrm{d}t \ \mp \int_0^\tau \nabla\mathcal{L}(\theta_t)(1-e^{-\frac{\tau-t}{\lambda}})\mathrm{d}t,\end{aligned}$$

which concludes the proof of the base case.

Induction step: $m \to m+1$**.** For $m \geq 0$, assume that for every $\tau \in [0, T_{m+1}) \setminus \mathcal{T}$ the expressions

$$\dot{g}_{\pm,\tau} = -\text{m.p.v.}\int_0^\tau \left[\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)\right] e^{-\frac{\tau-t}{\lambda}}\operatorname{sgn}(w_{\pm,\tau}w_{\pm,t})\mathrm{d}t - \frac{1}{\lambda}\sum_{k=1}^m (-1)^{m-k}e^{-\frac{\tau-T_k}{\lambda}}$$

and

$$\begin{aligned}g_{\pm,\tau} = g_{\pm,0} &- \int_0^\tau \text{m.p.v.}\int_0^t \left[\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_{\pm,t}w_{\pm,s})\mathrm{d}s\ \mathrm{d}t \\ &- \sum_{k=1}^m (-1)^{m-k}\left(1-e^{-\frac{\tau-T_k}{\lambda}}\right) + 2\sum_{1\leq i<j\leq m}(-1)^{j-i}\left(1-e^{-\frac{T_j-T_i}{\lambda}}\right).\end{aligned}$$

are true and well-defined. We now want to extend the validity of these identities to $\tau \in (T_{m+1}, T_{m+2})$. For ease of notation during the induction step, let $T_{m+1} = T$, $w_\pm = w$, and $g_\pm = g$. Let $\varepsilon > 0$ and let $T_\pm = T \pm \varepsilon$.

Now, applying Lemma 36 to eq. (D.7) yields

$$\begin{aligned}\dot{g}_\tau &= \dot{g}_{T_+}e^{-\frac{\tau-T_+}{\lambda}} - \int_{T_+}^\tau \left[\left(\frac{\dot{w}_t}{w_t}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)\right] e^{-\frac{\tau-t}{\lambda}}\mathrm{d}t \\ g_\tau &= g_{T_+} + \dot{g}_{T_+}\int_{T_+}^\tau e^{-\frac{t-T_+}{\lambda}}\mathrm{d}t - \int_{T_+}^\tau\int_{T_+}^t \left[\left(\frac{\dot{w}_s}{w_s}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}}\mathrm{d}s\ \mathrm{d}t.\end{aligned}$$

For further ease of notation and with some abuse of notation, let $f_t = \left(\frac{\dot{w}_t}{w_t}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)$ on $\mathbb{R}_{\geq 0} \setminus \mathcal{T}$. We will shortly prove that $g_{T_+} - g_{T_-} = O(\varepsilon)$ and $\dot{g}_{T_+} + \dot{g}_{T_-} + \frac{1}{\lambda} = O(\varepsilon)$.[3] Hence, the following limits will hold:

$$\dot{g}_\tau = \lim_{\varepsilon\to 0}\left[-\frac{1}{\lambda}e^{-\frac{\tau-T_+}{\lambda}} - \dot{g}_{T_-}e^{-\frac{\tau-T_+}{\lambda}} - \int_{T_+}^{\tau} f_t e^{-\frac{\tau-t}{\lambda}}\mathrm{d}t\right]$$
$$g_\tau = \lim_{\varepsilon\to 0}\left[g_{T_-} - \frac{1}{\lambda}\int_{T_+}^{\tau} e^{-\frac{t-T_+}{\lambda}}\mathrm{d}t - \dot{g}_{T_-}\int_{T_+}^{\tau} e^{-\frac{t-T_+}{\lambda}}\mathrm{d}t - \int_{T_+}^{\tau}\int_{T_+}^{t} f_s e^{-\frac{t-s}{\lambda}}\mathrm{d}s\ \mathrm{d}t\right].$$

Induction step for $\dot{g}_\tau$. Let us begin to untangle the first limit by substituting $\dot{g}_{T_-}$ with its integral formula given by the induction hypothesis. Notice that

$$\begin{aligned}\dot{g}_{T_-}e^{-\frac{\tau-T_+}{\lambda}} &= -\text{m.p.v.}\int_0^{T_-} f_t e^{-\frac{T_- - t}{\lambda}}\mathrm{sgn}(w_{T_-}w_t)\mathrm{d}t\cdot e^{-\frac{\tau-T_+}{\lambda}} - \frac{e^{-\frac{\tau-T_+}{\lambda}}}{\lambda}\sum_{k=1}^{m}(-1)^{m-k}e^{-\frac{T_- - T_k}{\lambda}}\\ &= \text{m.p.v.}\int_0^{T_-} f_t e^{-\frac{\tau-t}{\lambda}}\mathrm{sgn}(w_\tau w_t)\mathrm{d}t\cdot e^{\frac{2\varepsilon}{\lambda}} + \frac{e^{\frac{2\varepsilon}{\lambda}}}{\lambda}\sum_{k=1}^{m}(-1)^{(m+1)-k}e^{-\frac{\tau-T_k}{\lambda}},\end{aligned}$$

where we used that $\mathrm{sgn}(\tau) = -\mathrm{sgn}(T_-)$ since w changes signs at T. Hence, we have that

$$\begin{aligned}\dot{g}_\tau &= -\lim_{\varepsilon\to 0}\Big[\frac{1}{\lambda}e^{-\frac{\tau-T_+}{\lambda}} + \frac{e^{\frac{2\varepsilon}{\lambda}}}{\lambda}\sum_{k=1}^{m}(-1)^{(m+1)-k}e^{-\frac{\tau-T_k}{\lambda}}\\ &\quad + \text{m.p.v.}\int_0^{T_-} f_t e^{-\frac{\tau-t}{\lambda}}\mathrm{sgn}(w_\tau w_t)\mathrm{d}t\cdot e^{\frac{2\varepsilon}{\lambda}} + \int_{T_+}^{\tau} f_t e^{-\frac{\tau-t}{\lambda}}\mathrm{sgn}(w_\tau w_t)\mathrm{d}t\Big]\\ &= \mp\int_0^{\tau}\frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)\mathrm{sgn}(w_\tau w_t) - \frac{1}{\lambda}\sum_{k=1}^{m+1}(-1)^{(m+1)-k}e^{-\frac{\tau-T_k}{\lambda}}\\ &\quad - \lim_{\varepsilon\to 0}\left[\text{m.p.v.}\int_0^{T_-}\left(\frac{\dot{w}_t}{w_t}\right)^2 e^{-\frac{\tau-t}{\lambda}}\mathrm{sgn}(w_\tau w_t)\mathrm{d}t\cdot e^{\frac{2\varepsilon}{\lambda}} + \int_{T_+}^{\tau}\left(\frac{\dot{w}_t}{w_t}\right)^2 e^{-\frac{\tau-t}{\lambda}}\mathrm{sgn}(w_\tau w_t)\mathrm{d}t\right],\end{aligned}$$

where the limit on the last line formally equals the modified principal value m.p.v. $\int_0^\tau \left(\frac{\dot{w}_t}{w_t}\right)^2 e^{-\frac{\tau-s}{\lambda}}\mathrm{sgn}(w_\tau w_t)\mathrm{d}t$ whose existence we want to prove as part of the induction step. In fact, notice that we just proved the existence of m.p.v. $\int_0^\tau \left(\frac{\dot{w}_t}{w_t}\right)^2 e^{-\frac{\tau-s}{\lambda}}\mathrm{sgn}(w_\tau w_t)\mathrm{d}t$ since both $\dot{g}_\tau$ and $\mp\int_0^\tau \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)\mathrm{sgn}(w_\tau w_t)$ are finite quantities. Hence, for $\tau \in (T_{m+1}, T_{m+2})$,

$$\dot{g}_\tau = -\text{m.p.v.}\int_0^{\tau}\left[\left(\frac{\dot{w}_t}{w_t}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_t)\right]e^{-\frac{\tau-t}{\lambda}}\mathrm{sgn}(w_\tau w_t)\mathrm{d}t - \frac{1}{\lambda}\sum_{k=1}^{m+1}(-1)^{(m+1)-k}e^{-\frac{\tau-T_k}{\lambda}}.$$

Induction step for g_τ. We move on to untangle the limit which equals g_τ. By the induction

[3] Whenever we write an equation of the form $A = B + O(\varepsilon^r)$ for some $r > 0$, we mean that $A = B + C$, where $|C| = O(\varepsilon^r)$.

hypothesis,

$$\dot{g}_{T_-} = -\text{m.p.v.}\int_0^{T_-} f_t e^{-\frac{T_- - t}{\lambda}}\operatorname{sgn}(w_{T_-}w_t)\mathrm{d}t - \frac{1}{\lambda}\sum_{k=1}^{m}(-1)^{m-k}e^{-\frac{T_- - T_k}{\lambda}}$$
$$g_{T_-} = g_0 - \int_0^{T_-}\text{m.p.v.}\int_0^t f_s e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t$$
$$-\sum_{k=1}^{m}(-1)^{m-k}\left(1-e^{-\frac{T_- - T_k}{\lambda}}\right) + 2\sum_{1\le i<j\le m}(-1)^{j-i}\left(1-e^{-\frac{T_j - T_i}{\lambda}}\right).$$

Again, we can substitute $\operatorname{sgn}(w_{T_-})$ with $-\operatorname{sgn}(w_\tau)$, and after performing the familiar integral and limit manipulations, we obtain

$$g_\tau = g_0 \mp \int_0^\tau\int_0^t \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t - \lim_{\varepsilon\to 0}\left[A_\varepsilon + B_\varepsilon + C_\varepsilon\right]$$
$$-\sum_{k=1}^{m+1}(-1)^{(m+1)-k}\left(1-e^{-\frac{\tau - T_k}{\lambda}}\right) + 2\sum_{1\le i<j\le m+1}(-1)^{j-i}\left(1-e^{-\frac{T_j - T_i}{\lambda}}\right),$$

where

$$A_\varepsilon = \int_0^{T_-}\text{m.p.v.}\int_0^t\left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t$$
$$B_\varepsilon = \int_{T_+}^{\tau}\text{m.p.v.}\int_0^{T_-}\left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t\cdot e^{\frac{2\varepsilon}{\lambda}}$$
$$C_\varepsilon = \int_{T_+}^{\tau}\int_{T_+}^{t}\left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t.$$

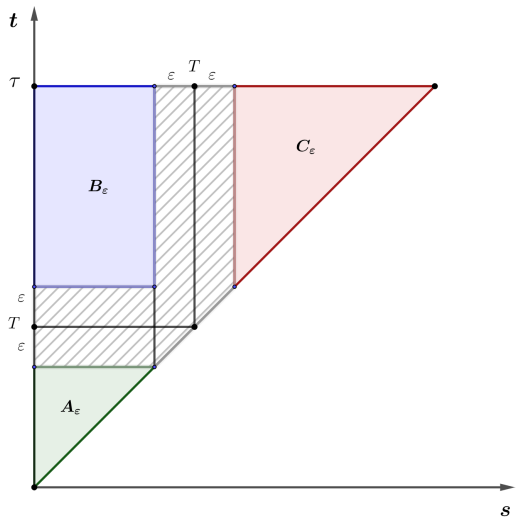


Figure D.1: A visualisation of the areas over which we integrate $\left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)$ in the above limit.

Notice that formally the limit $\lim_{\varepsilon\to 0}[A_\varepsilon + B_\varepsilon + C_\varepsilon]$ equals the integral quantity

$$\int_0^\tau\text{m.p.v.}\int_0^t\left(\frac{\dot{w}_s}{w_s}\right)^2 e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t,$$

whose existence we just proved as a consequence of the fact that

$$g_0 \mp \int_0^\tau \int_0^t \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t - \sum_{k=1}^{m+1}(-1)^{(m+1)-k}\left(1-e^{-\frac{\tau-T_k}{\lambda}}\right) + 2\sum_{1\le i<j\le m+1}(-1)^{j-i}\left(1-e^{-\frac{T_j-T_i}{\lambda}}\right.$$

is well-defined and finite. Thus, for $\tau \in (T_{m+1}, T_{m+2})$,

$$g_\tau = g_0 - \int_0^\tau \text{m.p.v.}\int_0^t \left[\left(\frac{\dot{w}_s}{w_s}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right]e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t - \sum_{k=1}^{m+1}(-1)^{(m+1)-k}\left(1-e^{-\frac{\tau-T_k}{\lambda}}\right) + 2\sum_{1\le i<j\le m+1}(-1)^{j-i}\left(1-e^{-\frac{T_j-T_i}{\lambda}}\right).$$

Proof of bounds. In order to conclude the induction step, we still have to prove the following bounds:

$$g_{T_+} - g_{T_-} = O(\varepsilon) \quad \text{and} \quad \dot{g}_{T_+} + \dot{g}_{T_-} + \frac{1}{\lambda} = O(\varepsilon).$$

Recall that $g_{T\pm\varepsilon} = \log|w_{T\pm\varepsilon}|$ and that $w_T = 0,\ \dot{w_T} \neq 0$. From the Taylor expansion of w_t, we know that

$$w_{T\pm\varepsilon} = \pm\varepsilon\dot{w}_T + O(\varepsilon^2).$$

Hence, $|w_{T+\varepsilon}/w_{T-\varepsilon}| = 1 + O(\varepsilon)$. Therefore, using the Taylor expansion of the logarithm around 1, we get that

$$|g_{T_+} - g_{T_-}| = |\log(1+O(\varepsilon))| = O(\varepsilon).$$

Now, recall that $\dot{g}_{T\pm\varepsilon} = \dot{w}_{T\pm\varepsilon}/w_{T\pm\varepsilon}$ and observe that

$$\begin{aligned} w_{T\pm\varepsilon} &= \pm\varepsilon\dot{w}_T + \frac{1}{2}\varepsilon^2\ddot{w}_T + O(\varepsilon^3) \\ \dot{w}_{T\pm\varepsilon} &= \dot{w}_T \pm \varepsilon\ddot{w}_T + O(\varepsilon^2). \end{aligned}$$

Hence,

$$\frac{w_{T+\varepsilon}\dot{w}_{T-\varepsilon} + w_{T-\varepsilon}\dot{w}_{T+\varepsilon}}{w_{T+\varepsilon}w_{T-\varepsilon}} = \frac{-\varepsilon^2\dot{w}_T\ddot{w}_T + O(\varepsilon^3)}{-\varepsilon^2\dot{w}_T^2 + O(\varepsilon^3)} = \frac{\ddot{w}_T}{\dot{w}_T} + O(\varepsilon).$$

Since, $\lambda\ddot{w}_T + \dot{w}_T \pm \nabla\mathcal{L}(\theta_T)\odot w_T = 0$ and $w_T = 0$, we get that $\frac{\ddot{w}_T}{\dot{w}_T} = -\frac{1}{\lambda}$, which concludes the induction step.

Proof of Lemma 2. Thus, we proved that for $\tau \in (T_m, T_{m+1})$, $m \in \{0, 1, \ldots, N\}$,

$$\ln|w_\tau| = \ln|w_0| - \int_0^\tau \text{m.p.v.}\int_0^t \left[\left(\frac{\dot{w}_s}{w_s}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right]e^{-\frac{t-s}{\lambda}}\operatorname{sgn}(w_t w_s)\mathrm{d}s\ \mathrm{d}t - \sum_{k=1}^{m}(-1)^{m-k}\left(1-e^{-\frac{\tau-T_k}{\lambda}}\right) + 2\sum_{1\le i<j\le m}(-1)^{j-i}\left(1-e^{-\frac{T_j-T_i}{\lambda}}\right).$$

Recall that throughout our inductive proof we worked with a fixed coordinate $i \in [d]$ of $w_\pm$. Different coordinates of $w_\pm$ vanish at different points in time, so writing the sum the last line in a coordinate-agnostic way becomes impossible. Thus, deriving a simple expression for the

full d-dimensional vector $w_{\pm,\tau}$ for any $\tau \in \mathbb{R}_{\geq 0}$ also becomes impossible. However, remembering that the finite nonzero limits $\lim_{\tau\to\infty} |w_{\pm,\tau}| = |w_{\pm,\infty}|$ exist and letting $\tau \to \infty$ yields an interesting result for the weights at infinity. Indeed, notice that for every vanishing time T, $\lim_{\tau\to\infty} \left(1 - e^{-\frac{\tau - T_k}{\lambda}}\right) = 1$. Hence,

$$\frac{1}{\lambda}\sum_{k=1}^{N}(-1)^{N-k}\left(1 - e^{-\frac{\tau - T_k}{\lambda}}\right) = \mathbf{1}_{\{N- \text{ odd}\}}.$$

For every $i \in [d]$, let $N_{\pm}(i)$ denote the number of vanishing points for the coordinate $w_{\pm}(i)$. Let us define the d-dimensional parity vectors $P_{\pm} \in \{0,1\}^d$ such that $P_{\pm}(i) \equiv N_{\pm}(i) \mod 2$. Let us also define the d-dimensional vectors $Q_{\pm} \in \mathbb{R}^d$ such that for each coordinate $k \in [d]$,

$$Q_{\pm}(k) \coloneqq -2 \sum_{1\leq i<j\leq N_{\pm}(k)} (-1)^{j-i}\left(1 - e^{-\frac{T_{\pm,k}(j) - T_{\pm,k}(i)}{\lambda}}\right),$$

where $0 < T_{\pm,k}(1) < \cdots < T_{\pm,k}(N_{\pm}(k))$ denote the vanishing times of the weight $w_{\pm}(k)$. Hence, we obtain the formula

$$|w_{\pm,\infty}| = |w_{\pm,0}|e^{-(P_{\pm}+Q_{\pm})} \exp\left(-\int_0^{\infty} \text{m.p.v.} \int_0^t \left[\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 \pm \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right] e^{-\frac{t-s}{\lambda}} \text{sgn}(w_{\pm,t}w_{\pm,s}) \text{d}s \ \text{d}t\right). \tag{D.8}$$

Recall that in Lemma 34, we proved that the limit

$$\lim_{t\to\infty} \int_0^t \nabla\mathcal{L}(\theta_s)ds = \int_0^{\infty} \nabla\mathcal{L}(\theta_t)\text{d}t = \frac{1}{\lambda}\int_0^{\infty}\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\text{d}s \ \text{d}t$$

exists and is finite. Therefore, we can decouple $\left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2$ and $\nabla\mathcal{L}(\theta_s)$ from the above integral and show that the following integral limits exist and are finite:

$$\int_0^{\infty} \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \text{sgn}(w_{\pm,t}w_{\pm,s}) \text{d}s \ \text{d}t \quad \text{and} \quad \int_0^{\infty}\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}} \text{sgn}(w_{\pm,t}w_{\pm,s}) \text{d}s \ \text{d}t.$$

Hence, the integral quantities $\emptyset_{\pm}$ from Lemma 2 are well-defined and finite. Thus, we finally proved Lemma 2.

Trajectory-Dependent Characterisation. We started this section with a promise for more insightful representations of $\Delta_{\infty} = |\alpha_{+,\infty}\alpha_{-,\infty}|$ and $\tilde{\theta}_0 = (\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2)/4$. We now deliver on that promise.

Recall that $\alpha_{\pm,\infty} = w_{\pm,\infty} \exp\left(\pm \int_0^{\infty} \nabla\mathcal{L}(\theta_t)\text{d}t\right)$ and notice that

$$\int_0^{\infty}\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\text{d}s \ \text{d}t - \int_0^{\infty}\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\text{sgn}(w_{\pm,t}w_{\pm,s})\text{d}s \ \text{d}t = 2\int_0^{\infty}\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\mathbf{1}_{\{w_{\pm,t}w_{\pm,s}<0\}}\text{d}s$$

Hence, using the formula for $w_{\pm}$ from Equation (D.8), we derive the following:

$$\begin{aligned}|\alpha_{\pm,\infty}| = |w_{\pm,0}|e^{-(P_{\pm}+Q_{\pm})} &\odot \exp\left(-\int_0^{\infty} \text{m.p.v.} \int_0^t \left(\frac{\dot{w}_{\pm,s}}{w_{\pm,s}}\right)^2 e^{-\frac{t-s}{\lambda}} \text{sgn}(w_{\pm,t}w_{\pm,s})\text{d}s \ \text{d}t\right) \\ &\odot \exp\left(\pm\frac{2}{\lambda}\int_0^{\infty}\int_0^t \nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\mathbf{1}_{\{w_{\pm,t}w_{\pm,s}<0\}}\text{d}s \ \text{d}t\right).\end{aligned}$$

Now, let

$$\Lambda_{\pm} \coloneqq \mp\frac{2}{\lambda}\int_0^{\infty}\int_0^{t}\nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\mathbf{1}_{\{w_{\pm,t}w_{\pm,s}<0\}}\mathrm{d}s\ \mathrm{d}t + P_{\pm} + Q_{\pm}, \tag{D.9}$$

where the quantities $P_{\pm}$ and $Q_{\pm}$ were defined in the previous paragraph. Notice that as we promised underneath Lemma 2, $\Lambda_{\pm}$ vanish whenever the balancedness Δ_t remains strictly positive. Using the abbreviation $I_{\pm} = \text{Ø}_{\pm} + \Lambda_{\pm}$, we get that

$$|\alpha_{\pm,\infty}| = |w_{\pm,0}| \odot \exp\left(-I_{\pm}\right).$$

Multiplying $|\alpha_{+,\infty}|$ by $|\alpha_{-,\infty}|$, we derive a formula for the asymptotic balancedness:

$$\begin{aligned}\Delta_{\infty} = \Delta_0 e^{-(P_+ + P_- + Q_+ + Q_-)} &\odot \exp\left(-\int_0^{\infty}\text{m.p.v.}\int_0^{t}\left[\left(\frac{\dot{w}_{+,s}}{w_{+,s}}\right)^2 + \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right]e^{-\frac{t-s}{\lambda}}\mathrm{sgn}(w_{+,t}w_{+,s})\mathrm{d}s\ \mathrm{d}t\right)\\ &\odot \exp\left(-\int_0^{\tau}\text{m.p.v.}\int_0^{t}\left[\left(\frac{\dot{w}_{-,s}}{w_{-,s}}\right)^2 - \frac{1}{\lambda}\nabla\mathcal{L}(\theta_s)\right]e^{-\frac{t-s}{\lambda}}\mathrm{sgn}(w_{-,t}w_{-,s})\mathrm{d}s\ \mathrm{d}t\right)\\ &\odot \exp\left(\frac{2}{\lambda}\int_0^{\infty}\int_0^{t}\nabla\mathcal{L}(\theta_s)e^{-\frac{t-s}{\lambda}}\left[\mathbf{1}_{\{w_{+,t}w_{+,s}<0\}} - \mathbf{1}_{\{w_{-,t}w_{-,s}<0\}}\right]\mathrm{d}s\ \mathrm{d}t\right).\end{aligned} \tag{D.10}$$

Now, we can write Δ_{∞} and $\tilde{\theta}_0$ more succinctly as

$$\Delta_{\infty} = \Delta_0 \odot \exp\Big(-(I_+ + I_-)\Big)$$

and

$$\tilde{\theta}_0 = \frac{1}{4}\Big(w_{+,0}^2 \odot \exp\left(-2I_+\right) - w_{-,0}^2 \odot \exp\left(-2I_-\right)\Big),$$

which concludes the proof of Theorem 3. □

Consequences for Generalisation.

We just proved that whenever MGF on a diagonal linear network converges and the balancedness at infinity is nonzero, we can characterize the recovered interpolator through the implicit regularization problem

$$\begin{aligned}\theta^{\text{MGF}} &= \arg\min_{\theta^\star\in\mathcal{S}}\ D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)\\ &= \arg\min_{\theta^\star\in\mathcal{S}}\ \left[\psi_{\Delta_\infty}(\theta^\star) - \langle\nabla\psi_{\Delta_\infty}(\tilde{\theta}_0), \theta^\star\rangle\right].\end{aligned}$$

Since

$$\psi_{\Delta_\infty}(\theta) = \frac{1}{4}\sum_{i=1}^{d}\left(2\theta_i\mathrm{arcsinh}\left(\frac{2\theta_i}{\Delta_{\infty,i}}\right) - \sqrt{4\theta_i^2 + \Delta_{\infty,i}^2} + \Delta_{\infty,i}\right)$$

and

$$\nabla\psi_{\Delta_\infty}(\theta) = \frac{1}{2}\mathrm{arcsinh}\left(\frac{2\theta}{\Delta_\infty}\right),$$

for a small asymptotic balancedness $\Delta_\infty = O(\Delta_0) = O(\alpha^2)$ and small perturbed initialisation $|\tilde{\theta}_0| = O(\alpha^2) \ll |\theta^\star|$, we would expect $\psi_{\Delta_\infty}(\theta^\star)$ to dominate $\langle\nabla\psi_{\Delta_\infty}(\tilde{\theta}_0), \theta^\star\rangle$. More formally, for a fixed $\theta^\star \in \mathcal{S}$ and small Δ_∞ and $\tilde{\theta}_0$, we have the following asymptotic equivalence:

$$\psi_{\Delta_\infty}(\theta^\star) \underset{\alpha\to 0}{\sim} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

Hence, $\theta^{\text{MGF}} \approx \arg\min_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star) = \theta^{\text{GF}}_{\Delta_\infty}$ as we discussed in Section 9.4.1. So, if $\Delta_\infty < \Delta_0$, Lemma 35 implies that the MGF predictor will benefit from better sparsity guarantees than the GF solution.

Therefore, to recap, for a small initialisation scale α and provided that the bounds $\Delta_\infty = O(\alpha^2)$ and $\tilde{\theta}_0 = O(\alpha^2)$ hold, we conclude that the asymptotic balancedness at infinity Δ_∞ roughly controls the sparsity of the recovered interpolator. And when $\Delta_\infty < \Delta_0$, θ^{MGF} will be sparser than $\theta^{\text{GF}}_\alpha$. Unfortunately, without the assumption that the balancedness Δ_t remains strictly positive for all $t \in [0, +\infty]$, we cannot formally compare Δ_∞ and $\tilde{\theta}_0$ with α.

Note that even without the bounds $\Delta_\infty = O(\alpha^2)$ and $\tilde{\theta}_0 = O(\alpha^2)$, if $|\tilde{\theta}_0| \ll |\theta^\star|$, then $\psi_{\Delta_\infty}(\theta^\star)$ still dominates $\langle \nabla \psi_{\Delta_\infty}(\tilde{\theta}_0), \theta^\star \rangle$. Indeed, our experiments clearly show that the perturbation term $\tilde{\theta}_0$ can safely be ignored since $\theta^{\text{MGF}} \approx \arg\min_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$ (see the discussion around Figure D.4.)

D.3.4 Non-vanishing balancedness

If we work under the assumption that the balancedness $\Delta_t = |w_{+,t} w_{-,t}|$ never vanishes, then much of the analysis from Appendix D.3.3 greatly simplifies. First, the integral quantities $P_\pm$ and $Q_\pm$ from the previous subsection become 0. Second, the multipliers $\operatorname{sgn}(w_{\pm,t} w_{\pm,s})$ become equal to 1 for all $t, s \in \mathbb{R}_{\geq 0}$. Hence, using Fubini's Theorem as in the proof of Lemma 36, we get that

$$\begin{aligned}
\int_0^\tau \text{m.p.v.} \int_0^t \left[\left(\frac{\dot{w}_s}{w_s} \right)^2 \pm \frac{1}{\lambda} \nabla \mathcal{L}(\theta_s) \right] e^{-\frac{t-s}{\lambda}} \operatorname{sgn}(w_t w_s) \mathrm{d}s \, \mathrm{d}t \\
&= \int_0^\tau \int_0^t \left[\left(\frac{\dot{w}_s}{w_s} \right)^2 \pm \frac{1}{\lambda} \nabla \mathcal{L}(\theta_s) \right] e^{-\frac{t-s}{\lambda}} \mathrm{d}s \, \mathrm{d}t \\
&= \lambda \int_0^\tau \left(\frac{\dot{w}_t}{w_t} \right)^2 \left(1 - e^{-\frac{\tau - t}{\lambda}} \right) \mathrm{d}t \pm \int_0^\tau \nabla \mathcal{L}(\theta_t) \left(1 - e^{-\frac{\tau - t}{\lambda}} \right) \mathrm{d}t.
\end{aligned}$$

Therefore, referencing eq. (D.8), we can express the evolution of the iterates as follows:

$$w_{\pm,\tau} = w_{\pm,0} \exp\left(-\lambda \int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \right)^2 \left(1 - e^{-\frac{\tau - t}{\lambda}} \right) \mathrm{d}t \right) \exp\left(\mp \int_0^\tau \nabla \mathcal{L}(\theta_s) \left(1 - e^{-\frac{\tau - t}{\lambda}} \right) \mathrm{d}t \right). \tag{D.11}$$

Thus, the balancedness evolves as

$$\Delta_t = \Delta_0 \exp\left(-\lambda \int_0^\tau \left[\left(\frac{\dot{w}_{+,t}}{w_{+,t}} \right)^2 + \left(\frac{\dot{w}_{-,t}}{w_{-,t}} \right)^2 \right] \left(1 - e^{-\frac{\tau - t}{\lambda}} \right) \mathrm{d}t \right). \tag{D.12}$$

Now, from Lemma 34, we know that $\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \right)^2$ is integrable and that

$$\lim_{\tau \to \infty} \int_0^\tau \nabla \mathcal{L}(\theta_s) \left(1 - e^{-\frac{\tau - t}{\lambda}} \right) \mathrm{d}t = \int_0^\infty \nabla \mathcal{L}(\theta_s) \mathrm{d}t$$

exists. Furthermore, from Lemma 37, we know that that

$$\lim_{\tau \to \infty} \int_0^\tau \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \right)^2 \left(1 - e^{-\frac{\tau - t}{\lambda}} \right) \mathrm{d}t = \int_0^\infty \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \right)^2) \mathrm{d}t.$$

Therefore, letting $\tau \to \infty$, we obtain the formulas

$$w_{\pm,\tau} = w_{\pm,0} \exp\left(-\lambda \int_0^\infty \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}} \right)^2 \mathrm{d}t \right) \exp\left(\mp \int_0^\infty \nabla \mathcal{L}(\theta_s) \mathrm{d}t \right) \tag{D.13}$$

$$\Delta_\infty = \Delta_0 \exp\left(-\lambda \int_0^\infty \left[\left(\frac{\dot{w}_{+,t}}{w_{+,t}} \right)^2 + \left(\frac{\dot{w}_{-,t}}{w_{-,t}} \right)^2 \right] \mathrm{d}t \right). \tag{D.14}$$

Hence, clearly, $\Delta_\infty < \Delta_0$.

Finally, let us consider how the perturbed initialisation $\tilde{\theta}_0$ looks like when Δ_t remains nonzero. Recall that $\tilde{\theta}_0 = (\alpha_+^2 - \alpha_-^2)/4$, where $\alpha_{\pm,\infty} = w_{\pm,\infty} \exp\left(\pm \int_0^\infty \nabla\mathcal{L}(\theta_t)\mathrm{d}t\right)$. Thus,

$$\alpha_{\pm,\infty} = w_{\pm,0} \exp\left(-\lambda \int_0^\infty \left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \mathrm{d}t\right)$$

and

$$\tilde{\theta}_0 = \frac{1}{4}\left[w_{+,0}^2 \exp\left(-2\lambda \int_0^\infty \left(\frac{\dot{w}_{+,t}}{w_{+,t}}\right)^2 \mathrm{d}t\right) - w_{,0}^2 \exp\left(-2\lambda \int_0^\infty \left(\frac{\dot{w}_{-,t}}{w_{-,t}}\right)^2 \mathrm{d}t\right)\right]. \tag{D.15}$$

Now, $\alpha_{\pm,\infty} < w_{\pm,0} \leq 2\alpha$, where $\alpha = \max(\|u_0\|_\infty, \|v_0\|_\infty)$ stood for the initialisation scale. Hence, $|\tilde{\theta}_0| < \alpha^2$.

Therefore, we just proved

Corollary 2. *For $\lambda > 0$, if the balancedness Δ_t remains strictly positive during training (i.e. $\Delta_t \neq 0$ for $t \in [0, +\infty]$), then the perturbed initialisation satisfies $|\tilde{\theta}_0| < \alpha^2$ and*

$$\Delta_\infty = \Delta_0 \odot \exp\Big(-\lambda \int_0^\infty \left(\frac{\dot{w}_{+,t}}{w_{+,t}}\right)^2 + \left(\frac{\dot{w}_{-,t}}{w_{-,t}}\right)^2 \mathrm{d}t\Big).$$

Importantly, $\Delta_\infty < \Delta_0$.

D.3.5 Behaviour of Δ_∞ for small values of λ

Since a precise asymptotic result for small λ is technically difficult, in this section we focus on giving some qualitative results. For $\lambda > 0$, recall that our iterates follow

$$\lambda \ddot{w}_{\pm,t}^{(\lambda)} + \dot{w}_{\pm,t}^{(\lambda)} \pm \nabla\mathcal{L}(\theta_t^{(\lambda)}) \odot w_{\pm,t}^{(\lambda)} = 0,$$

where we explicitly highlight the dependency on λ. Therefore, we have

$$\frac{\dot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}^{(\lambda)}} = \mp\nabla\mathcal{L}(\theta_t^{(\lambda)}) - \lambda \frac{\ddot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}}$$

and

$$\Big(\frac{\dot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}^{(\lambda)}}\Big)^2 = \nabla\mathcal{L}(\theta_t^{(\lambda)})^2 + \lambda^2\Big(\frac{\ddot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}}\Big)^2 \pm 2\lambda\mathcal{L}(\theta_t^{(\lambda)})\Big(\frac{\ddot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}}\Big).$$

Informally, we expect $(t \mapsto \nabla L(\theta_t^{(\lambda)}))_{0<\lambda\leq 1} \in L^2(0, +\infty)$ and $(t \mapsto \lambda\frac{\ddot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}})_\lambda \underset{\lambda\to 0}{\longrightarrow} 0$ in L^2-norm (see Theorem 5.1 in Attouch et al. [2000]). Hence, we get

$$\int_0^\infty \Big(\frac{\dot{w}_{\pm,t}^{(\lambda)}}{w_{\pm,t}^{(\lambda)}}\Big)^2 \underset{\lambda\to 0}{\sim} \int_0^\infty \nabla\mathcal{L}(\theta_t^{(\lambda)})^2 \mathrm{d}t$$

and

$$\Delta_\infty \underset{\lambda\to 0}{\approx} \Delta_0 \exp\big(-2\lambda \int_0^\infty \nabla\mathcal{L}(\theta_s^{(\lambda)})^2 \mathrm{d}s\big).$$

D.4 Discrete time results

In this section, we cover the proofs of our discrete-time results from Section 9.5. We first recall the SMGD recursion (D.3) with the $w_\pm$-parametrisation from Appendix D.2. Initialised at $w_{\pm,0} = w_{\pm,1} \in \mathbb{R}^d$, for $k \geq 1$, the iterates follow

$$w_{\pm,k+1} = w_{\pm,k} \mp \gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot w_{\pm,k} + \beta(w_{\pm,k} - w_{\pm,k-1}). \tag{D.16}$$

In what follows, we will adapt our continuous-time proof technique to the discrete case and identify a quantity which follows a momentum mirror descent with time-varying potentials. Our proofs closely follow the proof techniques from Even et al. [2023] which considers SGD without momentum.

D.4.1 Proof of Lemma 3, Theorem 4 and Corollary 3

We start by recalling the chapter's results. The first lemma introduces two convergent series which will appear in our main result.

Lemma 3. *The following two sums S_+ and S_- converge to finite vectors:*

$$S_\pm = \frac{1}{1-\beta} \sum_{k=1}^{\infty} \left[r\Big(\frac{w_{\pm,k+1}}{w_{\pm,k}}\Big) + \beta r\Big(\frac{w_{\pm,k}}{w_{\pm,k+1}}\Big) \right],$$

where $r(z) = (z-1) - \ln(|z|)$ for $z \neq 0$.

The proof of the lemma can be found in the proof of the following main theorem.

Theorem 4. *The solution θ^{SMGD} of SMGD (9.9) interpolates the dataset and satisfies the following implicit regularisation:*

$$\theta^{SMGD} = \arg\min_{\theta^\star \in \mathcal{S}} \; D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0).$$

In the above expression, $D_{\psi_{\Delta_\infty}}$ denotes the Bregman divergence with potential ψ_{Δ_∞}, where the asymptotic balancedness equals

$$\Delta_\infty = \Delta_0 \odot \exp\big(-(S_+ + S_-)\big)$$

and $\tilde{\theta}_0 = \frac{1}{4}(w_{+,0}^2 \odot \exp(-2S_+)) - w_{-,0}^2 \odot \exp(-2S_-))$ denotes a perturbed initialisation term.

Proving Convergence towards an Interpolator. By Assumption 9, we have that the iterates $w_{\pm,k}$ converge towards limiting weights $w_{\pm,\infty}$ and that the predictors converge towards a vector θ^{MGF}. Taking the limit in Equation (D.16), we get that $\lim_{k\to\infty} \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) \odot w_{\pm,k} = 0$. By Assumption 10, $w_{\pm,\infty}$ have non-zero coordinates. Therefore, $\lim_{k\to\infty} \nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k) = 0$. For any fixed batch $\mathcal{B} \subset \{1, \cdots, n\}$, the sampling with or without replacement is such that (almost surely) the set $M_k := \{k \geq 0, \mathcal{B}_k = \mathcal{B}\}$ is infinite. Hence, by continuity of $\nabla L_\mathcal{B}$, $\lim_{k\to\infty, k\in M_k} \nabla L_\mathcal{B}(\theta_k) = \nabla L_\mathcal{B}(\theta^{\text{SMGD}})$. Therefore, for all fixed batches $\mathcal{B}$, $\nabla L_\mathcal{B}(\theta^{\text{SMGD}}) = 0$ and hence θ^{SMGD} interpolates the dataset.

From here on now, for ease of notation, we do the proof for deterministic MGD. The proof for stochastic MGD is exactly the same after replacing $\nabla \mathcal{L}(\theta_k)$ with $\nabla \mathcal{L}_{\mathcal{B}_k}(\theta_k)$.

Deriving the Momentum Mirror Descent. Recall that the set of pairs (γ, β) such that there exists k where $w_{\pm,k} = 0$ is negligible in $\mathbb{R}^2$. We can hence assume that the iterates are never exactly zero, and we consider the logarithmic reparametrisation of the iterates $w_{\pm,k}$ as

$$g_{\pm,k} = \begin{cases} \ln(w_{\pm,k}), & \text{if } w_{\pm,k} > 0, \\ \ln(|w_{\pm,k}|) + i\pi, & \text{if } w_{\pm,k} < 0. \end{cases}$$

This way we have that that $w_{\pm,k} = \exp(g_{\pm,k})$ for all k. Equation (D.16) then becomes

$$\exp(g_{\pm,k+1}) = \exp(g_{\pm,k}) \mp \gamma \nabla \mathcal{L}(\theta_k) \odot \exp(g_{\pm,k}) + \beta(\exp(g_{\pm,k}) - \exp(g_{\pm,k-1})).$$

Dividing by $\exp(g_{\pm,k})$ yields

$$\exp(g_{\pm,k+1} - g_{\pm,k}) = 1 \mp \gamma \nabla \mathcal{L}(\theta_k) + \beta(1 - \exp(-(g_{\pm,k} - g_{\pm,k-1})).$$

Now, for $k \geq 1$, let $\delta_{\pm,k} = g_{\pm,k} - g_{\pm,k-1}$ so that we can more compactly write the above recurrence as

$$\exp(\delta_{\pm,k+1}) = 1 \mp \gamma \nabla \mathcal{L}(\theta_k) + \beta(1 - \exp(-\delta_{\pm,k})).$$

The trick, inspired by Even et al. [2023], is to consider the function $q(z) = \exp(z) - (1 + z)$ for $z \in \mathbb{C}$. Importantly, note that $q(z) \geq 0$ for $z \in \mathbb{R}$. Using this function, we can now rewrite the recurrence as

$$\delta_{\pm,k+1} + q(\delta_{\pm,k+1}) = \mp\gamma \nabla \mathcal{L}(\theta_k) + \beta(\delta_{\pm,k} - q(-\delta_{\pm,k})).$$

Setting the residues $Q_{\pm,k} \coloneqq q(\delta_{\pm,k+1}) + \beta q(-\delta_{\pm,k})$ leads to

$$\delta_{\pm,k+1} = \beta \delta_{\pm,k} \mp \gamma \nabla \mathcal{L}(\theta_k) - Q_{\pm,k}.$$

This can be seen as a first-order recurrence relation with variable coefficients. For $\beta = 0$ we exactly recover the analysis from Even et al. [2023]. For $\beta > 0$, since $\delta_{\pm,1} = 0$, for $m \geq 1$, we can expand the relation as

$$\delta_{\pm,m+1} = -\sum_{k=1}^{m} \beta^{m-k} \left[\pm\gamma \nabla \mathcal{L}(\theta_k) + Q_{\pm,k}\right].$$

Summing over m, we now get for $N \geq 1$ the following expression:

$$\begin{aligned} g_{\pm,N+1} - g_{\pm,1} &= \sum_{m=1}^{N} \delta_{\pm,m+1} \\ &= -\sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} \left[\pm\gamma \nabla \mathcal{L}(\theta_k) + Q_{\pm,k}\right] \end{aligned}$$

Finally, taking the exponential for $N \geq 1$, we obtain

$$\begin{aligned} w_{\pm,N+1} &= w_{\pm,0} \exp\left(-\sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} \left[\pm\gamma \nabla \mathcal{L}(\theta_k) + Q_{\pm,k}\right]\right) \\ &= w_{\pm,0} \exp\left(\pm \sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} Q_{\pm,k}\right) \exp\left(\mp\gamma \sum_{m=1}^{N} \sum_{k=1}^{m} \beta^{m-k} \nabla \mathcal{L}(\theta_k)\right) \\ &= w_{\pm,0} \exp\left(-\frac{1}{1-\beta} \sum_{m=1}^{N} (1 - \beta^{N+1-m}) Q_{\pm,m}\right) \exp\left(\mp\frac{\gamma}{1-\beta} \sum_{m=1}^{N} (1 - \beta^{N+1-m}) \nabla \mathcal{L}(\theta_m)\right), \end{aligned}$$

where the last equality is obtained by changing the order of summation. Following our continuous-time approach, for $N \geq 2$, we define $\alpha_{\pm,N+1}$ as

$$\begin{aligned}\alpha_{\pm,N+1} &\coloneqq w_{\pm,0} \exp\left(\pm \sum_{m=1}^{N}\sum_{k=1}^{m} \beta^{m-k} Q_{\pm,k}\right) \\ &= w_{\pm,0} \exp\left(-\frac{1}{1-\beta}\sum_{m=1}^{N}(1-\beta^{N+1-m})Q_{\pm,m}\right).\end{aligned} \tag{D.17}$$

We can now write the iterates $w_{\pm,k}$ as

$$w_{\pm,N+1} = \alpha_{\pm,N+1} \exp\big(\mp\gamma \sum_{m=1}^{N}\sum_{k=1}^{m} \beta^{m-k}\nabla\mathcal{L}(\theta_k)\big).$$

Thus, the regression parameter θ_N becomes

$$\begin{aligned}\theta_{N+1} &= \frac{1}{4}(w_{+,N+1}^2 - w_{-,N+1}^2) \\ &= \frac{1}{4}\alpha_{+,N+1}^2 \exp\big(-2\gamma \sum_{m=1}^{N}\sum_{k=1}^{m} \beta^{m-k}\nabla\mathcal{L}(\theta_k)\big) - \frac{1}{4}\alpha_{-,N+1}^2 \exp\big(2\gamma \sum_{m=1}^{N}\sum_{k=1}^{m} \beta^{m-k}\nabla\mathcal{L}(\theta_k)\big) \\ &= \frac{1}{2}\Delta_{N+1}\ \sinh\left(-2\gamma \sum_{m=1}^{N}\sum_{k=1}^{m} \beta^{m-k}\nabla\mathcal{L}(\theta_k) + \operatorname{arcsinh}\left(\frac{\alpha_{+,N+1}^2 - \alpha_{-,N+1}^2}{2\Delta_{N+1}}\right)\right),\end{aligned}$$

where we recall that $\Delta_N = |w_{+,N}w_{-,N}| = |\alpha_{+,N}\alpha_{-,N}|$. Hence, similar to the continuous case,

$$\frac{1}{2}\operatorname{arcsinh}\left(\frac{2\theta_{N+1}}{\Delta_{N+1}}\right) - \frac{1}{2}\operatorname{arcsinh}\left(\frac{\alpha_{+,N+1}^2 - \alpha_{-,N+1}^2}{2\Delta_{N+1}}\right) = -\gamma \sum_{m=1}^{N}\sum_{k=1}^{m} \beta^{m-k}\nabla\mathcal{L}(\theta_k).$$

For $N \geq 1$, the above identity becomes exactly

$$\nabla\Phi_{N+1}(\theta_{N+1}) = -\gamma \sum_{m=1}^{N}\sum_{k=1}^{m} \beta^{m-k}\nabla\mathcal{L}(\theta_k), \tag{D.18}$$

where the time-varying potential $\Phi_N : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\begin{aligned}\Phi_N(\theta) &= \frac{1}{4}\sum_{i=1}^{d}\left(2\theta_i \operatorname{arcsinh}\left(\frac{2\theta_i}{\Delta_{N,i}}\right) - \sqrt{4\theta_i^2 + \Delta_{N,i}^2} + \Delta_{N,i}\right) + \langle\phi_N, \theta\rangle \\ &= \psi_{\Delta_N}(\theta) + \langle\phi_N, \theta\rangle,\end{aligned}$$

where $\phi_N = \frac{1}{2}\operatorname{arcsinh}\left(\frac{\alpha_{+,N}^2 - \alpha_{-,N}^2}{2\Delta_N}\right)$ and ψ_{Δ_N} is the hyperbolic entropy defined in Equation (9.6). Notice that with this definition we arrive at the following time-varying momentum mirror descent for $N \geq 1$:

$$\nabla\Phi_{N+1}(\theta_{N+1}) = \nabla\Phi_N(\theta_N) - \gamma\nabla\mathcal{L}(\theta_N) + \underline{(}\nabla\Phi_N(\theta_N) - \nabla\Phi_{N-1}(\theta_{N-1})). \tag{D.19}$$

Convergent Quantities. From Lemma 38, we have that $\alpha_{\pm,N}$ must converge and that the limiting vectors $\alpha_{\pm,\infty}$ have non-zero coordinates. Therefore, the series $\sum_{m=1}^{\infty}\sum_{k=1}^{m}\beta^{m-k}Q_{\pm,k}$ are convergent and their terms must hence converge to zero: $\sum_{k=1}^{m}\beta^{m-k}Q_{\pm,k} \underset{m\to\infty}{\longrightarrow} 0$. Therefore,

$$\alpha_{\pm,N} \to \alpha_{\pm,\infty} = w_{\pm,0}\exp\left(-\frac{1}{1-\beta}\sum_{m=1}^{\infty}Q_{\pm,m}\right).$$

We now develop the formulas for $Q_{\pm,m}$ in order to arrive at the sums $S_\pm$ from Lemma 3. Recall that for $m \geq 1$, $Q_{\pm,m} = q(\delta_{\pm,m+1}) + \beta q(-\delta_{\pm,m})$ and $\delta_{\pm,1} = q(\delta_{\pm,1}) = 0$. Therefore,

$$\begin{aligned}\sum_{m=1}^{\infty} Q_{\pm,m} &= \sum_{m=1}^{\infty} q(\delta_{\pm,m+1}) + \beta q(-\delta_{\pm,m}) \\ &= \sum_{m=1}^{\infty} q(\delta_{\pm,m+1}) + \beta q(-\delta_{\pm,m+1}).\end{aligned}$$

Since $\delta_{\pm,m+1} = g_{\pm,m+1} - g_{\pm,m}$, we have

$$\delta_{\pm,m+1} = \begin{cases} \ln\left(\frac{w_{\pm,m+1}}{w_{\pm,m}}\right) & \text{if } w_{\pm,m+1} \text{ and } w_{\pm,m} \text{ have the same sign,} \\ \ln\left(\left|\frac{w_{\pm,m+1}}{w_{\pm,m}}\right|\right) + \operatorname{sgn}(w_{\pm,m})i\pi & \text{if they have different signs.} \end{cases}$$

It remains to notice that since $q(z) = \exp(z) - (1+z)$, we get that

$$\begin{aligned} q(\ln(z)) &= (z-1) - \ln(z) && \text{for } z \in \mathbb{R}_{>0}, \\ q(\ln(|z|) \pm i\pi) &= (z-1) - (\ln(|z|) \pm i\pi) && \text{for } z \in \mathbb{R}_{<0}. \end{aligned}$$

Therefore letting $r(z) = (z-1) - \ln(|z|)$ as in Lemma 3, we get

$$\begin{aligned} q(\delta_{\pm,m+1}) &= r\Big(\frac{w_{\pm,m+1}}{w_{\pm,m}}\Big) - \xi_{\pm,m}\operatorname{sgn}(w_{\pm,m})i\pi \\ q(-\delta_{\pm,m+1}) &= r\Big(\frac{w_{\pm,m}}{w_{\pm,m+1}}\Big) + \xi_{\pm,m}\operatorname{sgn}(w_{\pm,m})i\pi, \end{aligned}$$

where $\xi_{\pm,m} = 0$ if $\operatorname{sgn}(w_{\pm,m+1}) = \operatorname{sgn}(w_{\pm,m})$ and 1 otherwise. This leads to

$$\begin{aligned}\frac{1}{1-\beta}\sum_{m=1}^{\infty} Q_{\pm,m} &= \frac{1}{1-\beta}\sum_{m=1}^{\infty}\Big[r\Big(\frac{w_{\pm,m+1}}{w_{\pm,m}}\Big) + \beta r\Big(\frac{w_{\pm,m}}{w_{\pm,m+1}}\Big)\Big] - \sum_{m=1}^{\infty}\xi_{\pm,m}\operatorname{sgn}(w_{\pm,m})i\pi \\ &= S_\pm - \sum_{m=1}^{\infty}\xi_{\pm,m}\operatorname{sgn}(w_{\pm,m})i\pi < \infty.\end{aligned}$$

The last equality is due to the definition of $S_\pm$ from Lemma 3, and the last inequality is due to the summability of $(Q_{\pm,m})_m$. This therefore proves lemma Lemma 3. Now notice that

$$\alpha_{\pm,\infty}^2 = w_{\pm,0}^2 \exp\left(-2S_\pm\right).$$

Since $\Delta_\infty = |\alpha_{+,\infty}\alpha_{-,\infty}|$, we finally get that

$$\Delta_\infty = \Delta_0 \odot \exp\Big(-(S_+ + S_-)\Big).$$

Implicit Regularisation Problem. Notice that

$$\nabla\Phi_{N+1}(\theta_{N+1}) = -\gamma\sum_{m=1}^{N}\sum_{k=1}^{m}\beta^{m-k}\nabla\mathcal{L}(\theta_k) \in \mathtt{span}(x_1, \cdots, x_n).$$

Let $\Phi_\infty(\theta) \coloneqq \psi_{\Delta_\infty}(\theta) + \langle\phi_\infty, \theta\rangle$ and consider

$$\nabla\Phi_\infty(\theta^{\mathrm{MGD}}) = (\nabla\Phi_\infty(\theta^{\mathrm{MGD}}) - \nabla\Phi_\infty(\theta_N)) + (\nabla\Phi_\infty(\theta_N) - \nabla\Phi_N(\theta_N)) + \nabla\Phi_N(\theta_N).$$

The first two terms converge to 0: the first due to the convergence $\theta_N \to \theta^{\text{MGD}}$ and the second due to the uniform convergence of $\nabla\Phi_N$ to $\nabla\Phi_\infty$ on compact sets. The last term is in $\texttt{span}(x_1, \cdots, x_n)$ for all N. Therefore, we get that $\nabla\Phi_\infty(\theta_\infty) \in \texttt{span}(x_1, \cdots, x_n)$, and following the exact same proof as in the continuous-time framework, we finally get that

$$\theta^{\text{MGD}} = \underset{\theta^\star \in \mathcal{S}}{\arg\min}\ D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$$

where

$$\begin{aligned}\tilde{\theta}_0 &= \frac{\alpha_{+,\infty}^2 - \alpha_{-,\infty}^2}{4} \\ &= \frac{1}{4}\Big(w_{+,0}^2 \odot \exp\left(-2S_+\right) - w_{-,0}^2 \odot \exp\left(-2S_-\right)\Big).\end{aligned}$$

□

We recall and prove the following corollary.

Corollary 3. *For $\gamma, \beta > 0$, if the iterates $w_{\pm,k} = (u_k \pm v_k)$ do not change sign during training, then $|\tilde{\theta}_0| < \alpha^2$ and $\Delta_\infty < \Delta_0$.*

Proof. The corollary follows from the fact that if the iterates $w_{\pm,k}$ do not change sign, then since $r(z) \geq 0$ for $z > 0$, we get that $S_\pm > 0$ and $\Delta_\infty < \Delta_0$. Furthermore, $|\tilde{\theta}_0| < \max(w_{+,0}^2, w_{-,0}^2)/4 \leq \alpha^2$ □

D.4.2 Link to the continuous-time result.

In this subsection we link our continuous results with the discrete when the iterates do not cross zero. Indeed, at first sight, the discrete-time expression for Δ_∞ might seem quite different from its continuous-time counterpart:

$$\Delta_\infty^{\text{MGD}} = \Delta_0 \exp\left(-\frac{1}{1-\beta}\sum_{k=1}^{\infty}\left[r\left(\frac{w_{+,k+1}}{w_{+,k}}\right) + r\left(\frac{w_{-,k+1}}{w_{-,k}}\right)\right] + \beta\left[r\left(\frac{w_{+,k}}{w_{+,k+1}}\right) + r\left(\frac{w_{-,k}}{w_{-,k+1}}\right)\right]\right)$$

$$\Delta_\infty^{\text{MGF}} = \Delta_0 \exp\left(-\lambda\int_0^\infty \left(\frac{\dot{w}_{+,t}}{w_{+,t}}\right)^2 + \left(\frac{\dot{w}_{-,t}}{w_{-,t}}\right)^2 \mathrm{d}t\right).$$

However, upon closer inspection, by letting the discretisation step $\varepsilon = \sqrt{l\gamma} = \frac{\gamma}{(1-\beta)}$ from Proposition 19 go to 0, we can recover the continuous-time result. Indeed, as $\varepsilon \to 0$, we expect successive iterates $w_{\pm,k}$ to be close and hence $w_{\pm,k+1}/w_{\pm,k} \approx 1$. Now, since $r(z) \sim_{z\to 1} (z-1)^2/2$, we roughly have

$$r\Big(\frac{w_{\pm,k+1}}{w_{\pm,k}}\Big) \approx \frac{1}{2}\big(\frac{w_{\pm,k+1} - w_{\pm,k}}{w_{\pm,k}}\big)^2$$

and

$$r\Big(\frac{w_{\pm,k}}{w_{\pm,k+1}}\Big) \approx \frac{1}{2}\big(\frac{w_{\pm,k+1} - w_{\pm,k}}{w_{\pm,k+1}}\big)^2 \approx \frac{1}{2}\big(\frac{w_{\pm,k+1} - w_{\pm,k}}{w_{\pm,k}}\big)^2$$

Putting the approximations together:

$$\begin{aligned}\frac{1}{1-\beta}\sum_k\Big[r\Big(\frac{w_{\pm,k+1}}{w_{\pm,k}}\Big) + \beta r\Big(\frac{w_{\pm,k}}{w_{\pm,k+1}}\Big)\Big] &\approx \frac{1}{2}\frac{\varepsilon(1+\beta)}{1-\beta}\sum_k\left(\frac{w_{\pm,k+1} - w_{\pm,k}}{\varepsilon}\right)^2 \frac{1}{(w_{\pm,k})^2}\cdot\varepsilon \\ &\approx \frac{1+\beta}{2}\frac{\gamma}{(1-\beta)^2}\int_0^\infty\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \mathrm{d}t \\ &= \frac{1+\beta}{2}\lambda\int_0^\infty\left(\frac{\dot{w}_{\pm,t}}{w_{\pm,t}}\right)^2 \mathrm{d}t.\end{aligned}$$

Notice that in order for λ to remain constant and ε to go to 0, we must both have $\gamma \to 0$ and $\beta \to 1$. Hence, $(1+\beta)/2 \to 1$, and we recover the continuous-time expression for the balancedness.

However, note that when the iterates cross zero it is unclear to the authors how the continuous formula and its discrete counterpart compare.

Another Safe-Check Computation. Recall that MGD with stepsize γ and momentum parameter β corresponds to the discretisation of MGF with $\lambda = \gamma/(1-\beta)^2$ and discretisation step $\varepsilon = \sqrt{\lambda\gamma}$. To check the consistency between the discrete time equations and continuous time equations, we look at the value of $\exp(-\frac{t-s}{\lambda})$ and times '$t = m\varepsilon$' and '$s = k\varepsilon$':

$$\begin{aligned}\exp(-\frac{t-s}{\lambda}) &= \exp(-\frac{(m-k)\varepsilon}{\lambda}) \\ &= \exp(-(m-k)(1-\beta)) \\ &= [\exp(\beta-1)]^{m-k} \\ &\sim_{\beta\to 1} \beta^{m-k}.\end{aligned}$$

This small computation serves as a safe-check, affirming the correspondence between the continuous-time analysis expression $\exp(-\frac{t-s}{\lambda})$ and its discrete-time counterpart β^{m-k}.

D.5 Technical lemmas

In this section we present various technical lemmas which allow us to prove our main results. For $\Delta \in \mathbb{R}^d_{>0}$, we recall the definition of the hyperbolic entropy function $\psi_\Delta : \mathbb{R}^d \to \mathbb{R}$ at scale Δ:

$$\psi_\Delta(\theta) = \frac{1}{4}\sum_{i=1}^{d}\left(2\theta_i \operatorname{arcsinh}\left(\frac{2\theta_i}{\Delta_i}\right) - \sqrt{4\theta_i^2+\Delta_i^2} + \Delta_i\right).$$

The following lemma shows that the potential behaves as the ℓ_1-norm as Δ approaches 0.

Lemma 35. *For $\theta \in \mathbb{R}^d$ the following asymptotic equivalence holds:*

$$\psi_\Delta(\theta) \underset{\Delta\to 0}{\sim} \frac{1}{4}\sum_{i=1}^{d}\ln\left(\frac{1}{\Delta_i}\right)|\theta_i|.$$

Proof. The lemma easily follows from the asymptotic convergence

$$\operatorname{arcsinh}(x) \underset{|x|\to\infty}{\sim} \operatorname{sgn}(x)\ln|x|.$$

□

The following lemma is a classical result which gives a closed-form expression to the solution of a first order ODE.

Lemma 36. *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a differentiable function and let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a continuous function such that for some $\lambda \neq 0$,*

$$\lambda\dot{f} + f + g = 0, \quad \forall t \in \mathbb{R}_{\geq 0}.$$

Then,

$$f(t) = f(0)e^{-\frac{t}{\lambda}} - \frac{1}{\lambda}\int_0^t g(s)e^{-\frac{(t-s)}{\lambda}}ds.$$

Moreover, we have the following formula for the integral of $f(t)$:

$$\int_0^T f(t)dt = \lambda f(0)(1 - e^{-\frac{T}{\lambda}}) - \int_0^T g(t)(1 - e^{-\frac{(T-t)}{\lambda}})dt.$$

Proof. If we integrate the identity $\frac{d}{dt}\left[f(t)e^{t/\lambda}\right] = -\frac{1}{\lambda}g(t)e^{t/\lambda}$, we get that

$$f(t) = f(0)e^{-\frac{t}{\lambda}} - \frac{1}{\lambda}\int_0^t g(s)e^{-\frac{(t-s)}{\lambda}}ds.$$

As for the second part of the lemma, notice that

$$\int_0^T f(t)dt = \int_0^T \left[f(0)e^{-\frac{t}{\lambda}} - \frac{1}{\lambda}\int_0^t g(s)e^{-\frac{(t-s)}{\lambda}}ds\right]dt.$$

Hence, using Fubini, we get

$$\begin{aligned}
\int_0^T \int_0^t g(s)e^{-\frac{(t-s)}{\lambda}}dsdt &= \int_0^T \int_0^T g(s)\mathbf{1}_{s\leq t}(s,t)e^{-\frac{(t-s)}{\lambda}}dsdt \\
&= \int_0^T g(s)\int_0^T \mathbf{1}_{s\leq t}(s,t)e^{-\frac{(t-s)}{\lambda}}dtds \\
&= \int_0^T g(s)\int_s^T e^{-\frac{(t-s)}{\lambda}}dtds \\
&= \int_0^T g(s)l(1 - e^{-\frac{(T-s)}{\lambda}})ds,
\end{aligned}$$

which concludes the proof of the lemma. □

The following lemma gives various properties on integrability and convergence of the solution f of the aforementioned ODE.

Lemma 37. *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a differentiable function such that $f(0) = 0$ and let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}^d$ be a continuous function such that for some $\lambda \neq 0$,*

$$\lambda\dot{f} + f + g = 0, \quad \forall t \in \mathbb{R}_{\geq 0}.$$

If $g \in L^\infty(0, +\infty)$, then $f \in L^\infty(0, +\infty)$ and $\|f\|_\infty \leq \|g\|_\infty$. Moreover, if $g \in L^1(0, +\infty)$, then the following hold:

- *$f \in L^1(0, +\infty)$ and $\int_0^t |f(s)|ds \leq \int_0^t |g(s)|ds, \quad \forall t \in [0, +\infty]$;*
- *$\lim_{t\to\infty} f(t) = 0$;*
- *$\int_0^\infty f = -\int_0^\infty g$.*

Proof. First, assume $g \in L^\infty(0, \infty)$. From Lemma 36, we have that $f(t) = -\frac{1}{\lambda}\int_0^t g(s)e^{-\frac{(t-s)}{\lambda}}ds$. Hence,

$$\begin{aligned}
|f(t)| &\leq \frac{\|g\|_\infty}{\lambda}\int_0^t e^{-\frac{(t-s)}{\lambda}}ds \\
&= \|g\|_\infty(1 - e^{-t/\lambda}) \leq \|g\|_\infty,
\end{aligned}$$

which proves the first assertion.

Second, assume $g \in L^1(0,\infty)$. Then, $|f(t)| \le \frac{1}{\lambda}\int_0^t |g(s)| e^{-\frac{(t-s)}{\lambda}} ds$. Therefore,

$$\begin{aligned}\int_0^t |f(s)| ds &\le \int_0^t |g(s)|(1 - e^{-\frac{(t-s)}{\lambda}}) ds \\ &\le \int_0^t |g(s)| ds \le \|g\|_{L^1}.\end{aligned}$$

Moving on, we will show that $\lim_{t\to\infty} f(t) = 0$. Recall that $f(t) = -\frac{1}{\lambda}\int_0^t g(s) e^{-\frac{(t-s)}{\lambda}} ds$. Then,

$$\begin{aligned}\left|\int_0^t g(s) e^{-\frac{(t-s)}{\lambda}} ds\right| &= \left|\int_0^{t/2} g(s) e^{-\frac{(t-s)}{\lambda}} ds + \int_{t/2}^t g(s) e^{-\frac{(t-s)}{\lambda}} ds\right| \\ &\le e^{-\frac{t}{2\lambda}} \int_0^{t/2} |g(s)| ds + \int_{t/2}^{\infty} |g(s)| ds \\ &\xrightarrow{t\to\infty} 0.\end{aligned}$$

Finally, notice that

$$\begin{aligned}&\lim_{t\to\infty}\left[\lambda\int_0^t \dot{f} + \int_0^t (f+g)\right] = 0 \iff \\ &\lambda \lim_{t\to\infty} f(t) + \int_0^{\infty} (f+g) = 0 \iff \\ &\int_0^{\infty} f + \int_0^{\infty} g = 0,\end{aligned}$$

where we used that $\lim_{t\to\infty} f(t) = 0$ and the linearity of the Lebesgue integral. □

With the help of Lemma 36 and Lemma 37, we can finally prove Section 9.4.3, which considers ODE (9.4) and establishes the positivity of the balancedness for small λ.

For $\lambda \le \frac{n}{\|y\|_2^2} \cdot (\min_{i\le d} \Delta_{0,i})$, the balancedness Δ_t never vanishes: $\Delta_t \neq 0,\ \forall t \in [0, +\infty]$.

Proof. We consider MGF(λ) with the diagonal-linear-network loss $F(w) = \mathcal{L}(u \odot v)$, where $w = (u, v)$. From the energy of the system, defined in Equation (D.4) as $E_t = F(w_t) + \frac{\lambda}{2}\|\dot{w}_t\|_2^2$ with derivative $\dot{E}_t = -\|\dot{w}_t\|_2^2$, we get that

$$\mathcal{L}(\theta_t) = \frac{\|y\|_2^2}{2n} - \frac{\lambda}{2}\|\dot{w}_t\|_2^2 - \int_0^t \|\dot{w}_s\|_2^2 ds.$$

Hence, since the LHS of the above equation is nonnegative, we get

$$\int_0^{\infty} \|\dot{w}_t\|^2 dt \le \frac{\|y\|^2}{2n}.$$

Therefore,

$$\int_0^{\infty} |\dot{u}_t^2 - \dot{v}_t^2| dt < \frac{\|y\|^2}{2n}\mathbf{1}.$$

Consequently, $\dot{u}_t^2 - \dot{v}_t^2 \in L^1(0,\infty)$. Now, notice that from ODE (9.4), we obtain

$$\begin{aligned}&l(\ddot{u}_t u_t - \ddot{v}_t v_t) + (\dot{u}_t u_t - \dot{v}_t v_t) = 0 \iff \\ &\lambda\frac{d}{dt}(\dot{u}_t u_t - \dot{v}_t v_t) + (\dot{u}_t u_t - \dot{v}_t v_t) - l(\dot{u}_t^2 - \dot{v}_t^2) = 0.\end{aligned}$$

Applying Lemma 36 yields

$$\dot{u}_t u_t - \dot{v}_t v_t = \int_0^t (\dot{u}_s^2 - \dot{v}_s^2) e^{-\frac{(t-s)}{\lambda}} ds$$

and

$$u_t^2 - v_t^2 = \Delta_0 + 2\lambda \int_0^t (\dot{u}_s^2 - \dot{v}_s^2)(1 - e^{-\frac{(t-s)}{\lambda}}) ds. \tag{D.20}$$

Applying Lemma 37 allows us to conclude that for every $t \in [0, +\infty]$,

$$\begin{aligned} \Delta_t &\geq \Delta_0 - 2\lambda \int_0^t |\dot{u}_s^2 - \dot{v}_s^2| ds \\ &> \Delta_0 - \frac{\lambda \|y\|_2^2}{n} \mathbf{1} \geq 0, \end{aligned}$$

where the last inequality is due to the inequality assumption over λ. □

Our final technical lemma helps with the proof of Theorem 4. The definition of the quantities $Q_{\pm,m}$ can be found in the proof of this theorem.

Lemma 38. *The quantities $\alpha_{\pm,N}$ defined in eq. (D.17):*

$$\alpha_{\pm,N+1} = \alpha \exp\left(-\frac{1}{1-\beta} \sum_{m=1}^{N} (1 - \beta^{N+1-m}) Q_{\pm,m} \right),$$

converge as $N \to \infty$ to vectors $\alpha_{\pm,\infty}$ with non-zero coordinates.

Proof. From Assumption 9 and Assumption 10, we have that the iterates $w_{\pm,N}$ converge towards vectors $w_{\pm,\infty}$ such that $\Delta_\infty = |w_{+,\infty} \odot w_{-,\infty}|$ has non-zero coordinates. This means that there exists $N_0 > 0$ such that $w_{\pm,N}$ do not change sign for $N \geq N_0$. Consequently, the imaginary parts of $g_{\pm,N}$ are constant (equal to 0 or π depending on the sign of $w_{\pm,\infty}$) for $N \geq N_0$, and $\delta_{\pm,N} \in \mathbb{R}$ for $N \geq N_0$. This finally means that $Q_{\pm,N} \geq 0$ for $N \geq N_0$ and

$$\sum_{m=1}^{N} (1 - \beta^{N+1-m}) Q_{\pm,m} = \sum_{m=1}^{N_0} (1 - \beta^{N+1-m}) Q_{\pm,m} + \sum_{m=N_0+1}^{N} (1 - \beta^{N+1-m}) Q_{\pm,m}$$

The first term converges to $\sum_{m=1}^{N_0} Q_{\pm,m}$ as $N \to \infty$. The second term is increasing because $Q_{\pm,N}$ are positive for $N \geq N_0$ and $(1 - \beta^{N+1-m})$ is increasing. Therefore, the second term also converges to a finite value since otherwise $\alpha_{\pm,\infty} = 0$, which contradicts $\Delta_\infty = |\alpha_{+,\infty} \alpha_{-,\infty}| \neq 0$. □

D.6 Additional experiments

In this section of the appendix, we clarify experimental details and discuss additional experiments.

D.6.1 MGF: a good continuous surrogate

Most of our experiments deal with 2-layer diagonal networks, but before we constrain ourselves to that tractable setting, we present a couple of experiments on more general architectures. These experiments highlight our observation from Section 9.3 that MGF(λ) serves as a good

continuous proxy for MGD(γ, β) even for complicated non-convex losses F and large step sizes γ. We provide evidence for that conclusion by showing that the single parameter $\lambda = \gamma/(1-\beta)^2$ controls the generalisation performance of models trained with MGD(γ, β).

Teacher-Student Fully Connected Network. We detail the experimental setting which leads to Figure 9.2. We consider a teacher-student setup where the teacher is a one-hidden-layer fully-connected ReLU network with 5 hidden neurons and the student is a one-hidden-layer fully-connected ReLU network with 20 hidden neurons. We randomly generate 15 inputs $x_i \in \mathbb{R}^2$ according to a standard multivariate normal distribution. Each y_i corresponds to the output by the teacher network on input x_i. The student is trained using momentum gradient descent with a square loss. Figure 9.2 corresponds to the test loss after the student reaches 10^{-5} training error. Each grid point corresponds to the same data set and initialisation of the student network. We observe that the quantity $\lambda = \frac{\gamma}{(1-\beta)^2}$ aligns well with the level lines of the test loss as expected from Proposition 19.

Deep Linear Network. The network used for Figure D.2 contains 5 layers with widths (30, 60, 120, 60, 1) and was trained for 1000 epochs for each pair of momentum parameter β and step size γ. Each network weight was randomly initialised according to $\mathcal{N}(0, 0.1^2)$ with fixed randomness for each (γ, β)-trial. The training data was chosen as follows: $(x_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu \mathbf{1}, \sigma^2 I_d)$ and $y_i = \langle x_i, \theta_s^\star \rangle$ for $i \in [n]$ where $\theta_s^\star$ is s-sparse with nonzero entries equal to $1/\sqrt{s}$, where $(n, d, s) = (20, 30, 5)$ and $(\mu, \sigma) = (1, 1)$. We show results averaged over 5 replications.

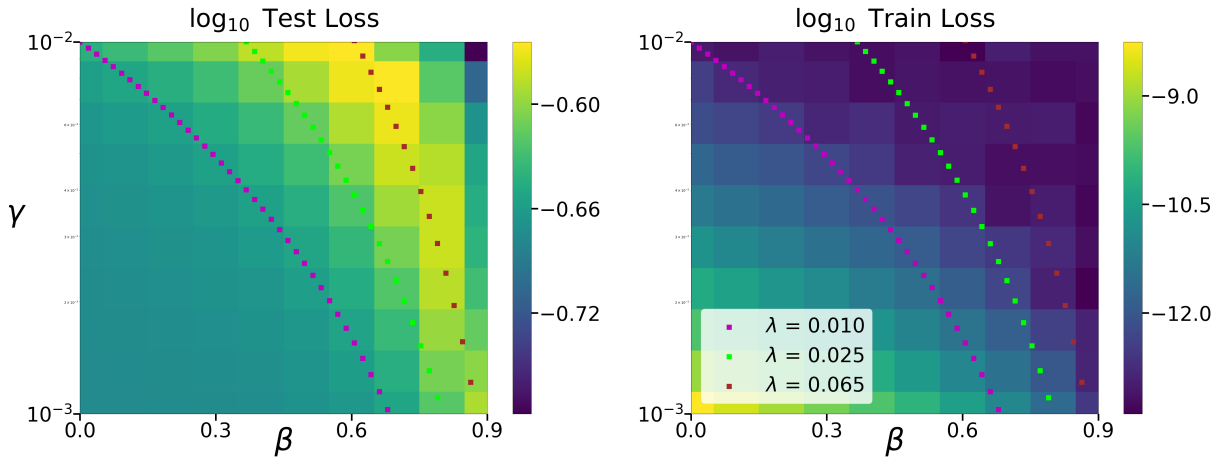


Figure D.2: Test and train loss of a fully connected deep linear network trained with MGD(γ, β) in a noiseless sparse overparametrised regression setting. The test loss appears considerably correlated with the intrinsic parameter $\lambda = \gamma/(1-\beta)^2$, evincing that MGF$(\lambda)$ approximates MGD(γ, β) sufficiently well even on complex architectures.

2-Layer Diagonal Linear Network. The plots from Figure D.3 were obtained for a 2-layer diagonal linear network trained in the noiseless sparse overparametrised regression setting described above. The first network layer was initialised with the uniform initialisation $\alpha \mathbf{1}$, where $\alpha = 0.01$, and the weights of the second layer were set to 0. The momentum gradient flow evolution of the weights was simulated with the default version of the ODE solver `scipy.integrate.odeint`.

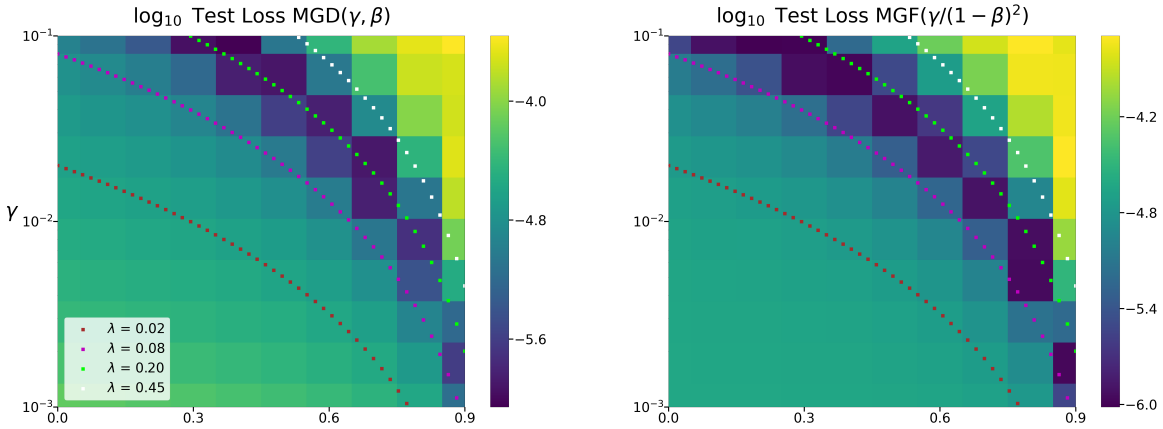


Figure D.3: *Left*: Decimal logarithm of the test loss of a 2-layer diagonal linear network trained with MGD(γ, β) for 1 million epochs. *Right*: Decimal logarithm of the test loss of a 2-layer diagonal linear whose weights evolved according to MGF(λ) – where $\lambda = \gamma/(1-\beta)^2$ – and converged to an interpolator of the training dataset. We observe an almost one-to-one correspondence in terms of generalisation capacity, which demonstrates that MGF(λ) serves as a suitable continuous surrogate for MGD(γ, β) in the diagonal linear setting.

D.6.2 Experiments with diagonal linear networks

Having seen empirical proof that MGF(λ) approximates well the optimisation trajectory of MGD(γ, β) on complicated models, we proceed with experiments that illustrate the conclusions of our results for 2-layer diagonal linear networks. In particular, we provide experimental evidence that both in the continuous and discrete-time cases, the recovered interpolators by MGD and MGF satisfy

$$\theta^{\mathrm{MGF}\,/\,\mathrm{MGD}} = \operatorname*{arg\,min}_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0) \approx \operatorname*{arg\,min}_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star),$$

as we explain underneath Theorem 3, Theorem 4, and in Appendix D.3.3. Indeed, we observe that the perturbation term $\tilde{\theta}_0$ can be safely ignored even without the assumption of strictly positive balancedness. The asymptotic balancedness Δ_∞ then uniquely controls the properties of the recovered solution. We now specify our experimental setting.

Experimental Details. We work in the noiseless sparse overparametrised regression setting with uncentered data. More precisely, we let $(x_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu\mathbf{1}, \sigma^2 I_d)$ and $y_i = \langle x_i, \theta_s^\star \rangle$ for $i \in [n]$ where $\theta_s^\star$ is s-sparse with nonzero entries equal to $1/\sqrt{s}$. We train a 2-layer diagonal linear network with (M)GD and (M)GF with the uniform initialisation $u_0 = \alpha\mathbf{1}$, where $\alpha = 0.01$ and $v_0 = 0$. In order to simulate gradient flow or momentum gradient flow on the network weights, we use the vanilla version of the ODE solver `scipy.integrate.odeint`. For most of the incoming plots, we have fixed $(n, d, s, \sigma) = (20, 30, 5, 1)$ and we let $\mu \in \{0, 0.5, 1, 1.5\}$. In what follows, all plots show results averaged over 5 replications.

Continuous-Time Plots

We first present a set of 3 continuous-time plots (Figure D.4) for the setting where the input data follows a Gaussian distribution $\mathcal{N}(\mu\mathbf{1}, I_d)$ with $\mu = 1$.

Experimental Setup. For a sampled dataset (X, y), we train our diagonal network with MGF(λ), $\lambda \in [0, 1]$, and initialisation $(u_0, v_0) = (\alpha \cdot \mathbf{1}, 0)$ until convergence to an interpolator [4] θ^{MGF}. During the training of MGF(λ), we also take note of whether the balancedness Δ_t remains strictly positive at all times, thereby checking the explanatory range of Section 9.4.3. Having completed the MGF training, we plot the **Test Loss** of θ^{MGF}, the **ℓ_2-Norm of Δ_∞**, and the **ℓ_1-Norm of θ^{MGF}** in order to visualise the gain in generalisation performance.

Insignificance of $\tilde{\theta}_0$. Now, recall from Theorem 3 that $\theta^{\text{MGF}} = \arg\min_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$ and that for $\|\tilde{\theta}_0\|_\infty \ll \|\theta^{\text{MGF}}\|_\infty$, $D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0) \approx \psi_{\Delta_\infty}(\theta^\star)$. We proved that for small values of λ, the balancedness remains strictly positive at all times, which allowed us to show that $\|\tilde{\theta}_0\|_\infty < \alpha^2$. We conjecture that $\theta^{\text{MGF}} \approx \arg\min_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$ continues to hold for larger values of λ. We experimentally test this claim by measuring the precise distance between θ^{MGF} and $\theta^{\text{GF}}_{\Delta_\infty} = \arg\min_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$. Indeed, we initialise a gradient flow with initial balancedness equal to Δ_∞ and such that $\theta_0 = 0$, which converges to the predictor $\theta^{\text{GF}}_{\Delta_\infty}$ as discussed in Section 9.4.1. Hence, we can calculate the **Normalised Distance between θ^{MGF} and $\theta^{\text{GF}}_{\Delta_\infty}$** equal to $\|\theta^{\text{MGF}} - \theta^{\text{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\text{GF}}_{\Delta_\infty}\|_2$, and we obtain that $\|\theta^{\text{MGF}} - \theta^{\text{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\text{GF}}_{\Delta_\infty}\|_2 < 0.01$ for $\lambda \in (0, 1)$.

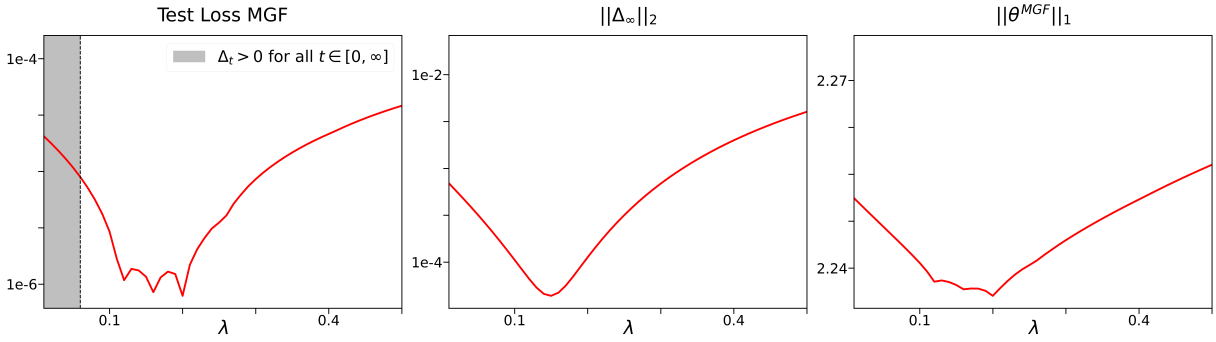


Figure D.4: Continuous-time experiments on uncentered data with mean $\mu = 1$. Here, θ^{MGF} denotes the interpolator recovered by MGF(λ) and Δ_∞ stands for the balancedness at infinity for MGF(λ). We observe that the test loss and sparsity of θ^{MGF} correlate with the magnitude of Δ_∞ as predicted by Theorem 3.

Insights from Continuous-Time Experiments. First, we observe that no matter the mean of the data distribution[5] or the size of $\lambda \in (0, 1)$, the normalised distance between θ^{MGF} and $\theta^{\text{GF}}_{\Delta_\infty}$ is always upper-bounded by 0.01. Hence, we can empirically confirm our conjecture from Theorem 3 that $\theta^{\text{MGF}} \approx \theta^{\text{GF}}_{\Delta_\infty}$ for larger λ when the balancedness changes sign. Second, we see that regardless of the mean of the dataset, the balancedness at infinity (i.e., the effective initialisation Δ_∞) controls the generalisation behavior of the recovered interpolator. We can explain this observation again through the approximate equivalence $\theta^{\text{MGF}} \approx \arg\min_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$.

The Effect of the Data Mean. In Figure D.5, we summarise our empirical results for data with various means. Notice that there exists a difference between the generalisation behavior for centered and uncentered data. Indeed, for centered data (top left), the key quantity λ has little impact on the sparsity of the recovered solution. This circumstance is reminiscent of the observations from Nacson et al. [2022] and Even et al. [2023]. However, for uncentered data, we

[4] We know that θ^{MGF} interpolates the dataset (X, y) because we also record the **Train Loss** (θ^{MGF}), which falls under 10^{-20}.

[5] We performed the continuous-time experiments depicted in Figure D.5 for data with mean $\mu = 0, 0.5, 1, 1.5$.

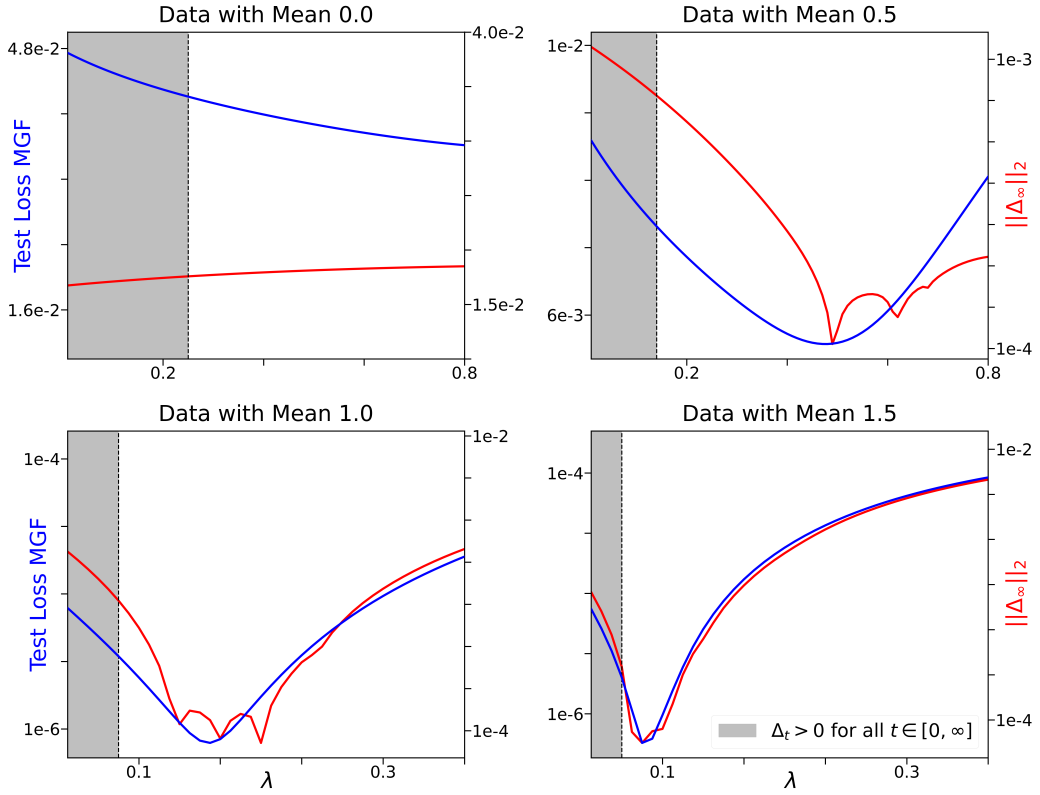


Figure D.5: We observe that for uncentered data the magnitude of the balancedness at infinity Δ_∞ correlates with the test loss of the interpolator selected by MGF(λ). However, this relationship breaks for centered data.

observe an interval $\mathcal{I}_{\mathcal{D}_x} = (0, \lambda_{\max})$ (which depends on the data distribution $\mathcal{D}_x$) for which MGF with $\lambda \in \mathcal{I}_{\mathcal{D}_x}$ outperforms GF in terms of generalisation. Furthermore, there appears to exist a constant $\lambda^\star_{\mathcal{D}_x} \in \mathcal{I}_{\mathcal{D}_x}$ (roughly corresponding to the minimum magnitude of Δ_∞) which brings about the most improvement compared to gradient flow. We note that the following tendency seems to hold empirically:

$$\lim_{|\mu| \to +\infty} \lambda^\star_{\mathcal{D}_x} = 0.$$

Discrete-Time Plots

For the sake of brevity[6], we only present a single set of plots for the discrete-time noiseless sparse recovery given in Figure 9.4. Our input data follows a unit-mean Gaussian distribution $\mathcal{N}(\mathbf{1}, I_d)$.

Experimental Setup. For a sampled dataset (X, y) and hyperparameter pair $(,\gamma)$, we train our 2-layer diagonal linear network with MGD(γ, β) initialised at $(u_0, v_0) = (\alpha\mathbf{1}, 0)$ for 1 million epochs (which suffices for convergence[7]). During the MGD(γ, β) training, we also take note of whether the iterates $w_{\pm,k}$ change sign or not thereby checking the explanatory range of Corollary 3. Having completed the MGD training, we plot the **Test Loss** of θ^{MGD}, the **ℓ_2-Norm of Δ_∞**, and the **ℓ_1-Norm of θ^{MGD}** in order to visualise the gain in generalisation performance.

Insignificance of $\tilde{\theta}_0$. Recall from Theorem 4 that $\theta^{\text{MGD}} = \arg\min_{\theta^\star \in \mathcal{S}} D_{\psi_{\Delta_\infty}}(\theta^\star, \tilde{\theta}_0)$. Again, we

[6] We performed discrete-time experiments for data with means $\mu = 0, 0.5, 1, 1.5$.
[7] Again, we record the **Train Loss** ($\theta^{\text{MGD}}_{\gamma, \beta\alpha}$), which falls under 10^{-8}.

want to characterise the recovered interpolator as $\theta^{\text{MGD}} \approx \arg\min_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$. In order to verify empirically that the effect of the perturbation term is negligible, we follow the same strategy as in the continuous-time case. We initialise a gradient flow with initial balancedness equal to Δ_∞ and $\theta_0 = 0$, which converges to the predictor $\theta^{\text{GF}}_{\Delta_\infty}$ as discussed in Section 9.4.1. Hence, we can calculate the **Normalised Distance between θ^{MGD} and $\theta^{\text{GF}}_{\Delta_\infty}$** equal to $\|\theta^{\text{MGD}}_{\gamma,\underline{,}\alpha} - \theta^{\text{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\text{GF}}_{\Delta_\infty}\|_2$, and we find that $\|\theta^{\text{MGD}}_{\gamma,\underline{,}\alpha} - \theta^{\text{GF}}_{\Delta_\infty}\|_2 / \|\theta^{\text{GF}}_{\Delta_\infty}\|_2 < 0.01$ for all pairs (γ, β) in Figure 9.4. This experimentally shows that $\theta^{\text{MGD}} \approx \arg\min_{\theta^\star \in \mathcal{S}} \psi_{\Delta_\infty}(\theta^\star)$ and that the asymptotic balancedess is the key quantity which predicts the recovered solution.

Insights from Discrete-Time Experiments. As predicted by Theorem 4, a more balanced solution (center plot) leads to a solution with a lower ℓ_1-norm (right plot), which in turn translates to better generalisation (left plot). Finally, as proven in Corollary 3, the trajectories for which the iterates do not cross zero satisfy $\Delta_\infty < \Delta_0$, where Δ_0 (approximately) corresponds to the asymptotic balancedness for the pair $(\underline{,}\gamma) = (0, 10^{-3})$ in the bottom left corner of the center plot. Clearly, the pairs $(\underline{,}\gamma)$ for which $w_{\pm,k}$ do not change sign lead to better generalisation than the pair $(0, 10^{-3})$. Again, we note that for centered data the story changes, and we lose the clear correspondence between small $\|\Delta_\infty\|_2$ and small $\|\theta^{\text{MGD}}\|_1$.

Appendix E

Appendix for Chapter 10

Organisation of the Appendix.

1. In Appendix E.1, we provide the proofs of the existence and uniqueness of (MF), of the convergence of the loss, the divergence of the iterates and the proof of Lemma 5.

2. In Appendix E.2, we provide all the proofs concerning the construction of the horizon shape and that of our main Theorems 6 and 7.

E.1 Proofs of properties of the mirror flow in the classification setting

We start by proving Lemma 4 which ensures the existence and unicity of (MF).

Lemma 4. *For any initialisation $\beta_0 \in \mathbb{R}^d$, there exists a unique solution defined over $\mathbb{R}_{\geq 0}$ which satisfies (MF) for all $t \geq 0$ and with initial condition $\beta_{t=0} = \beta_0$.*

Proof. First note that the two first points of Assumption 12 correspond to the definition of a *Legendre function* (see Rockafellar [1970], Chapter 26). It follows that the Fenchel conjugate ϕ^* is also a Legendre function and that the gradient $\nabla\phi$ is a bijection over $\mathbb{R}^d$ with $(\nabla\phi)^{-1} = \nabla\phi^*$.

Note that the existence of a global solution of (MF) is *a priori* not obvious. To prove it, we first consider the following differential equation:

$$\mathrm{d}u_t = -\nabla L(\nabla\phi^*(u_t))\mathrm{d}t, \tag{E.1}$$

with initial condition $u_{t=0} = \nabla\phi^*(\beta_0)$.

Since L is $\mathcal{C}^2$, ∇L is Lipschitz on all compact sets. Furthermore, since $\nabla^2\phi$ is p.s.d., $\nabla\phi^* = (\nabla\phi)^{-1}$ is $\mathcal{C}^1$ and therefore Lipschitz on all compact sets. Hence $\nabla L \circ \nabla\phi^*$ is Lipschitz on all compact sets and from the Picard-Lindelöf theorem, there exists a unique maximal (*i.e.* which cannot be extended) solution (u_t) satisfying eq. (E.1) such that $u_{t=0} = \nabla\phi^*(\beta_0)$. We denote $[0, T_{\max})$ the intersection of this maximal interval of definition (which must be open) and $\mathbb{R}_{\geq 0}$. Our goal is now to prove that $T_{\max} = +\infty$. To do so, we assume that $T_{\max}$ is finite and we will show that this leads to a contradiction due to the fact that the iterates β_t cannot diverge in finite time. Let $\beta_t \coloneqq \nabla\phi^*(u_t)$ and notice that β_t is therefore the unique solution satisfying (MF) over $[0, T_{\max})$ with $\beta_{t=0} = \beta_0$.

Bounding the trajectory of β_t over $[0, T_{\max})$. Pick any $\beta \in \mathbb{R}^d$ and notice that by convexity of L:

$$\frac{\mathrm{d}}{\mathrm{d}t} D_\phi(\beta, \beta_t) = -\langle \nabla L(\beta_t), \beta_t - \beta\rangle \leq -(L(\beta_t) - L(\beta)) \leq L(\beta) - L_{\min}.$$

Where $L_{\min}$ is a lower bound on the loss. Integrating from 0 to $t < T_{\max}$ we get:

$$\begin{aligned} D_\phi(\beta, \beta_t) &\leq t \cdot (L(\beta) - L_{\min}) + D_\phi(\beta, \beta_0) \\ &\leq T_{\max} \cdot (L(\beta) - L_{\min}) + D_\phi(\beta, \beta_0) \end{aligned}$$

Therefore, due to Assumption 12, the iterates β_t are bounded over $[0, T_{\max})$. The proof from here is standard (see e.g. Attouch et al. [2000], Theorem 3.1): from eq. (E.1) we get that $\dot{u}_t$ is bounded over $[0, T_{\max})$ and $\sup_{t\in[0,T_{\max})} \|\dot{u}_t\| =: C < +\infty$ which means that $\|u_t - u_{t'}\| \leq C|t - t'|$. Hence $\lim_{t\to T_{\max}} u_t =: u_\infty$ must exist. Applying the Picard-Lindelöf again at time $T_{\max}$ with initial condition u_∞ violates the initial maximal interval assumption. Therefore $T_{\max} = +\infty$ which concludes the proof. □

We now recall and prove classical results on the mirror flow in the classification setting.

Proposition 22. *Considering the mirror flow $(\beta_t)_{t\geq 0}$, the loss converges towards 0 and the iterates diverge: $\lim_{t\to\infty} L(\beta_t) = 0$ and $\lim_{t\to\infty} \|\beta_t\| = +\infty$.*

Proof. **The loss is decreasing.** $\frac{\mathrm{d}}{\mathrm{d}t} L(\beta_t) = -\langle \nabla L(\beta_t), \dot{\beta}_t\rangle = -\langle \nabla^2\phi(\beta_t)^{-1}\nabla L(\beta_t), \nabla L(\beta_t)\rangle \leq 0$, where the inequality is due to the convexity of the potential ϕ.

Convergence of the loss towards 0. Now consider the Bregman divergence between an arbitrary point β and β_t:

$$D_\phi(\beta, \beta_t) = \phi(\beta) - \phi(\beta_t) - \langle \nabla\phi(\beta_t), \beta - \beta_t \rangle \geq 0.$$

which is such that:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} D_\phi(\beta, \beta_t) &= \langle \frac{\mathrm{d}}{\mathrm{d}t} \nabla\phi(\beta_t), \beta_t - \beta \rangle \\ &= -\langle \nabla L(\beta_t), \beta_t - \beta \rangle \\ &\leq -(L(\beta_t) - L(\beta)) \end{aligned} \tag{E.2}$$

where the inequality is by convexity of the loss. Integrating and due to the decrease of the loss, we get that:

$$\begin{aligned} L(\beta_t) &\leq \frac{1}{t} \int_0^t L(\beta_s)\, \mathrm{d}s \\ &\leq L(\beta) + \frac{D_\phi(\beta, \beta_0) - D_\phi(\beta, \beta_t)}{t} \\ &\leq L(\beta) + \frac{D_\phi(\beta, \beta_0)}{t} \end{aligned}$$

Since this is true for all point β, we get that $L(\beta_t) \leq \inf_{\beta \in \mathbb{R}^d} L(\beta) + \frac{D_\phi(\beta, \beta_0)}{t}$. It remains to show that the right hand term goes to 0 as t goes to infinity. To show this, let $\varepsilon > 0$, by the separability assumption we get that there exists $\beta^\star$ such that $\min y_i \langle x_i, \beta^\star \rangle > 0$. Since $L(\lambda \beta_\star) \underset{\lambda \to \infty}{\longrightarrow} 0$, we can choose λ big enough such that $L(\lambda \beta^\star) < \varepsilon$ and then t_λ large enough such that $\frac{1}{t_\lambda} D_\phi(\lambda \beta^\star, \beta_0) < \varepsilon$. The loss therefore converges to 0.

Divergence of the iterates. For all $i \in [n]$, $\ell(y_i \langle x_i, \beta_t \rangle) \leq L(\beta_t) \underset{t \to \infty}{\longrightarrow} 0$. Due to the assumptions on the loss, this translates into $y_i \langle x_i, \beta_t \rangle \underset{t \to \infty}{\longrightarrow} \infty$, hence $\|\beta_t\| \underset{t \to \infty}{\longrightarrow} +\infty$. □

In the following lemma we recall and prove that a coordinate $q_\infty[k]$ must be equal to 0 if datapoint x_k is not a support vector of $\bar{\beta}_\infty$.

Lemma 39. *For some function $C_t \to \infty$, if the iterates $\bar{\beta}_t = \frac{\beta_t}{C_t}$ converge towards a vector which we denote $\bar{\beta}_\infty$ and q_t converges towards a vector $q_\infty \in [0, 1]^n$. Then it holds that:*

$$q_\infty[k] = 0 \qquad \text{if} \qquad y_k \langle x_k, \bar{\beta}_\infty \rangle > \min_{1 \leq i \leq n} y_i \langle x_i, \bar{\beta}_\infty \rangle.$$

In words, $q_\infty[k] = 0$ if x_k is not a support vector.

Proof. Recall that

$$q(\beta_t) = \frac{\ell'(Z\beta_t)}{\ell'(\ell^{-1}(\sum_i \ell(y_i \langle x_i, \beta_t \rangle)))}. \tag{E.3}$$

From Proposition 22, we have that $\min_{i \in [n]} y_i \langle x_i, \bar{\beta}_\infty \rangle > 0$ and we denote this margin as γ. Now consider $k \in [n]$ which is not a support vector, i.e, $y_k \langle x_k, \bar{\beta}_\infty \rangle > \min_{i \in [n]} y_i \langle x_i, \bar{\beta}_\infty \rangle$ and without loss of generality assume that $y_1 \langle x_1, \bar{\beta}_\infty \rangle = \min_{1 \leq i \leq n} y_i \langle x_i, \bar{\beta}_\infty \rangle$. We denote by $\delta = \langle y_k x_k - y_1 x_1, \bar{\beta}_\infty \rangle > 0$ the gap. Then

$$\begin{aligned} q(\beta_t)_k &= \frac{\ell'(C_t \langle x_k, \bar{\beta}_t \rangle)}{\ell'(\ell^{-1}(\sum_i \ell(C_t y_i \langle x_i, \bar{\beta}_t \rangle)))} \\ &\leq \frac{\ell'(C_t y_k \langle x_k, \bar{\beta}_t \rangle)}{\ell'(C_t y_1 \langle x_1, \bar{\beta}_t \rangle)} \end{aligned}$$

We write $\bar{\beta}_t = \bar{\beta}_\infty + r_t$ where $(r_t)_{t\geq 0} \in \mathbb{R}^d$ converges to 0. For t big enough, we have that $y_k\langle x_k, \bar{\beta}_t\rangle \geq y_k\langle x_k, \bar{\beta}_\infty\rangle - \frac{\delta}{4}$ and $y_1\langle x_1, \bar{\beta}_t\rangle \leq y_1\langle x_1, \bar{\beta}_\infty\rangle + \frac{\delta}{4}$. Therefore for t large enough, since ℓ' is negative and increasing:

$$\begin{aligned} q(\beta_t)[k] &\leq \frac{\ell'(C_t(y_k\langle x_k, \bar{\beta}_\infty\rangle - \delta/4))}{\ell'(C_t(y_1\langle x_1, \bar{\beta}_\infty\rangle + \delta/4))} \\ &\leq \frac{\ell'(C_t(\gamma + \delta/2 + \delta/4))}{\ell'(C_t(\gamma + \delta/2))} \underset{t\to\infty}{\longrightarrow} 0, \end{aligned}$$

where the last term converge to 0 due to the exponential tail of $-\ell'$ and that $C_t \to \infty$. □

We here reformulate and prove Lemma 5.

Lemma 40 (Reformulation of Lemma 5)**.** *Denoting $a(\beta_t) \coloneqq -\ell'(\ell^{-1}(\sum_i \ell(y_i\langle x_i, \beta_t\rangle))) > 0$, we have that $\int_0^t a(\beta_s)\mathrm{d}s \underset{t\to\infty}{\longrightarrow} +\infty$. For $\ell(z) = \exp(-z)$, this translates to $\int_0^t L(\beta_s)\mathrm{d}s \to \infty$.*

Proof. Recall that $\nabla\phi(\beta_t) = \nabla\phi(\beta_0) + Z^\top \int_0^t a(\beta_s)q(\beta_s)\mathrm{d}s$, therefore

$$\begin{aligned} \|\nabla\phi(\beta_t)\| &\leq \|\nabla\phi(\beta_0)\| + \sum_{i=1}^n \|x_i\| \int_0^t a(\beta_s)q(\beta_s)[i]\mathrm{d}s \\ &\leq \Big(\sum_{i=1}^n \|x_i\|\Big) \int_0^t a(\beta_s)\mathrm{d}s. \end{aligned}$$

Where the first inequality is due to the triangle inequality and the second to the fact $q(\beta) \in (0,1]^n$. Since the iterates diverge, we have from assumption 12 that $\|\nabla\phi(\beta_t)\| \underset{t\to\infty}{\longrightarrow} \infty$ and therefore that $\int_0^t a(\beta_s)\mathrm{d}s \underset{t\to\infty}{\longrightarrow} +\infty$. □

E.2 Differed proofs on the construction of ϕ_∞

As mentioned in the main text, the following property highlights the fact that all 'reasonable' potentials have a horizon shape.

If any of the three following conditions hold: *(i)* ϕ is a finite composition of polynomials, exponentials and logarithms, *(ii)* ϕ is globally sub-analytic, *(iii)* ϕ is definable in a o-minimal structure on $\mathbb{R}$; then ϕ admits a horizon shape S_∞.

Proof. Note that points (i) and (ii) are particular cases of (iii) [Dries, 1998, Bolte et al., 2007]. If h is definable in a o-minimal structure, then so is the sublevel set S_c for $c > 0$, and so is the normalization factor R_c since it can be defined in first-order logic as

$$R_c = \{r \in \mathbb{R} \,:\, \exists \beta^* \in S_c, \|\beta^*\| = r \text{ and } \forall \beta \in S_c, \|\beta\| \leq r\}.$$

Therefore, $(\bar{S}_c)_{c>0}$ if a definable family of definable and compact sets. Then so is the family $(\bar{S}_{t^{-1}})_{t\in(0,1]}$. Since all the sets belong to the unit ball of $\mathbb{R}^d$, they lie in the sets of compact subsets of $B(0,1)$. This set is compact for the Hausdorff metric [Aliprantis and Border, 2006, Thm 3.85]; therefore, there exists a sequence $(t_k)_{k\in\mathbb{N}}$ such that $t_k \to 0$ and $(\bar{S}_{t_k^{-1}})_{k\in\mathbb{N}}$ converges to some set $\bar{S}$.

We can then apply Corollary 2 of Kocel-Cynk et al. [2014], which states that there exists a definable arc $\gamma : (0,1] \to (0,1]$ such that $\lim_{\tau\to 0} \gamma(\tau) = 0$ and $\bar{S} = \lim_{\tau\to 0} \bar{S}_{\gamma(\tau)^{-1}}$. This implies that the limit $\bar{S}$ is uniquely defined and therefore that $\lim_{t\to 0} \bar{S}_{t^{-1}} = \bar{S}$. □

The next corollary is a more general restatement of Corollary 4. It shows that the construction of ϕ_∞ enables to take the limit $\lim_t \frac{\nabla\phi(\beta_t)}{t} \propto \partial\phi_\infty(\bar{\beta}_\infty)$.

Corollary 6. *Assume that ϕ admits a non-degenerate horizon shape S_∞. Then its horizon function ϕ_∞ satisfies the following properties.*

1. *ϕ_∞ is convex and finite-valued on $\mathbb{R}^d$,*
2. *Let $(\beta_s)_{s>0}$ be a continous sequence such that when $s \to \infty$:*

$$(a)\ \ \|\beta_s\| \to \infty, \quad (b)\ \ \frac{\beta_s}{\|\beta_s\|} \to \bar{\beta} \ \text{ for some } \ \bar{\beta} \in \mathbb{R}^d, \quad (c)\ \ \frac{\nabla\phi(\beta_s)}{\|\nabla\phi(\beta_s)\|} \to \bar{g} \ \text{ for some } \ \bar{g} \in \mathbb{R}^d.$$

Then $\bar{g}$ is proportional to a subgradient of ϕ_∞ at $\bar{\beta}$:

$$\bar{g} \in \lambda\partial\phi_\infty(\bar{\beta}) \quad \text{for some } \lambda > 0.$$

Proof. The sequence of sets $(\bar{S}_c)$ is contained in the compact ball $B(0,1)$; therefore, Hausdorff convergence is equivalent to Painlevé-Kuratowski convergence [Rockafellar and Wets, 1998, Section 4.C]. Hence, as $(\bar{S}_c)$ are convex, so is their limit S_∞ [Rockafellar and Wets, 1998, Prop 4.15]. It follows that h_∞ is convex [Rockafellar and Wets, 1998, Ex 3.50].

Since S_∞ is non-degenerate, there exists a radius r_0 such that $B(0, r_0) \subset S_\infty$, which implies that $h_\infty(\beta)$ is finite-valued for every β.

To prove point (ii), consider the sequence of functions $(\eta_c)_{c>0}$ formed by the indicators of convex sets $\bar{S}_c$:

$$\eta_c(\beta) = I_{\bar{S}_c}(\beta) = \begin{cases} 0 & \text{if } \beta \in \bar{S}_c, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that the epigraph of η_c is $\bar{S}_c \times \mathbb{R}_+$; these sets also converge to $S_\infty \times \mathbb{R}_+$ [Rockafellar and Wets, 1998, Ex 4.29], from which we conclude that function η_c converge *epigraphically* to the

indicator function η_∞ of S_∞ ($\eta_\infty = I_{S_\infty}$). We can then apply Attouch's theorem [Attouch and Beer, 1993, Combari and Thibault, 1998] ensuring that the graph of the subdifferentials of η_c

$$\mathcal{G}(\partial\eta_c) = \{(\beta, g) \,:\, g \in \partial\eta_c(\beta)\}$$

converge in Painlevé-Kuratowski sense to the graph $\mathcal{G}(\partial\eta_\infty)$ of subdifferential of η_∞. This means that if a sequence $(\beta_c, g_c)_{c>0}$ such that $(\beta_c, \beta_c) \in \mathcal{G}(\partial\eta_c)$ for every $c > 0$ converges, then its limit belongs to $\mathcal{G}(\partial\eta_\infty)$.

Consider now a sequence $(\beta_s)_{s>0}$ satisfying the conditions described in (ii). Since it diverges to infinity and h is coercive, we have $h(\beta_s) \to \infty$, and we may assume w.l.o.g that $h(\beta_s) > 0$ for all s. We have by definition of sublevel sets $\beta_s \in S_{h(\beta_s)}$, and therefore

$$\nabla h(\beta_s) \in \partial I_{S_{h(\beta_s)}}(\beta_s), \tag{E.4}$$

which can be derived easily from convexity of h (geometrically, this means that the gradients of h are normal to the sublevel sets). Consider now the normalized levels sets as defined in (10.4). Denoting

$$\bar{\beta}_s = \frac{\beta_s}{R_{h(\beta_s)}},$$

we have $\bar{\beta}_s \in \bar{S}_{h(\beta_s)}$ and thus by simple rescaling (E.4) becomes

$$\nabla h(\beta_s) \in \partial I_{\bar{S}_{h(\beta_s)}}\left(\bar{\beta}_s\right).$$

Since $\partial I_{\bar{S}_c}$ is a cone (the normal cone to $\bar{S}_c$), this also holds for any positive multiple of $\nabla h(\beta_s)$. We deduce that for every $s > 0$

$$\left(\bar{\beta}_s, \frac{\nabla h(\beta_s)}{\|\nabla h(\beta_s)\|}\right) \in \mathcal{G}(\partial\eta_{h(\beta_s)}).$$

Note that since $\bar{\beta}_s$ belongs to the normalized level sets, this sequence is bounded. We can extract a subsequence $(\bar{\beta}_{s_k}, \frac{\nabla h(\beta_{s_k})}{\|\nabla h(\beta_{s_k})\|})_{k\geq 0}$ which converges to a limit point $(\hat{\beta}, \hat{g})$. By the previous remark on graphical convergence of subdifferentials, we have $(\hat{\beta}, \hat{g}) \in \mathcal{G}(\partial I_{S_\infty})$, i.e.,

$$\hat{g} \in \partial I_{S_\infty}(\hat{\beta}). \tag{E.5}$$

We need to prove that $\hat{\beta}$ is not 0. Since h is strictly convex, the level set $\{\phi(\beta) = c\}$ is exactly the boundary of the sublevel set $\{\phi(\beta) \leq c\}$. Therefore, β_s lies on the boundary of $S_{h(\beta_s)}$, and hence so does $\bar{\beta}_s$ lie on the boundary of $\bar{S}_{h(\beta_s)}$. Since 0 is in the interior of S_∞, it also belongs to the interior of $\bar{S}_{\phi(\beta_s)}$ for s larger than some s_0. Then, there exists $r_0 > 0$ such that $B(0, r_0) \subset \bar{S}_{\phi(\beta_s)}$ for $s \geq s_0$. By definition of boundary, we then have for $s \geq s_0$ $\|\bar{\beta}_s\| > r_0$, which leads to $\|\hat{\beta}\| > 0$.

To achieve the desired result; we need to relate $(\hat{\beta}, \hat{g})$ to $(\bar{\beta}, \bar{g})$. First, notice that by construction we have necessarily $\hat{g} = \bar{g} = \lim_{s\to\infty} \nabla h(\beta_s)/\|h(\beta_s)\|$. Then, note that

$$\bar{\beta} = \lim_{k\to\infty} \frac{\beta_{s_k}}{\|\beta_{s_k}\|}, \quad \hat{\beta} = \lim_{k\to\infty} \frac{\beta_{s_k}}{R_{h(\beta_{s_k})}}.$$

Taking the norm of the second limit, we have $\|\hat{\beta}\| = \lim_{k\to\infty} \frac{\|\beta_{s_k}\|}{R_{h(\beta_{s_k})}}$. Injecting back in the first limit yields

$$\bar{\beta} = \lim_{k\to\infty} \frac{\beta_{s_k}}{R_{h(\beta_{s_k})}} \frac{R_{h(\beta_{s_k})}}{\|\beta_{s_k}\|} = \frac{\hat{\beta}}{\|\hat{\beta}\|}.$$

Therefore, (E.5) becomes

$$\bar{g} \in \partial I_{S_\infty}(\|\hat{\beta}\|\bar{\beta}).$$

This means that $\bar{g}$ belongs to the *normal cone* of S_∞ at $\|\hat{\beta}\|\bar{\beta}$ [Rockafellar, 1970, Sec. 23]. We note the level set $\{\beta : \phi_\infty(\beta) \leq \phi_\infty\left(\|\hat{\beta}\|\bar{\beta}\right)\}$ is exactly τS_∞ for some $\tau > 0$. We use Corollary 23.7.1 from Rockafellar [1970] which states that if a vector is in the normal cone of the level set of ϕ_∞, then it must be a positive multiple of a subgradient. This implies that that there exists $\lambda \geq 0$ such that

$$\bar{g} \in \lambda \partial h_\infty(\|\hat{\beta}\|\bar{\beta}).$$

Finally, $\lambda > 0$ since $\|\bar{g}\| = 1$, and $\partial h_\infty(\|\hat{\beta}\|\bar{\beta}) = \partial h_\infty(\bar{\beta})$ by positive homogenity of h_∞. □

We can now prove our main result, which we restate here.

Theorem 6. *Let ϕ_∞ be the horizon function of ϕ. Assuming that the ϕ_∞-max margin problem has a unique solution, the mirror flow normalised iterates $\bar{\beta}_t = \frac{\beta_t}{\|\beta_t\|}$ converge towards a vector $\bar{\beta}_\infty$ and*

$$\bar{\beta}_\infty \propto \arg\max_{\phi_\infty(\bar{\beta})\leq 1} \min_{i\in[n]} y_i\langle x_i, \bar{\beta}\rangle,$$

where the symbol $\propto$ denotes positive proportionality.

The proof essentially follows exactly the same lines as in Section 10.4 but taking into account the fact that the loss is not exactly the exponential one.

Proof. Recall that Z is the data matrix of size $n \times d$ whose i^{th} row is $y_i x_i$, we then have that $\nabla L(\beta) = Z^\top \ell'(Z\beta)$, where ℓ' is applied component wise. We now denote by $q(\beta)$ the vector in $\mathbb{R}^n$ equal to:

$$q(\beta) = \frac{\ell'(Z\beta)}{\ell'(\ell^{-1}(\sum_i \ell(y_i\langle x_i, \beta\rangle)))} \tag{E.6}$$

Notice that the facts that $\ell > 0$, $\ell' < 0$, ℓ^{-1} is increasing and ℓ' is decreasing, we have that $q(\beta) > 0$ and that for all $i_0 \in [n]$,

$$q(\beta)_{i_0} = \frac{\ell'(y_{i_0}\langle x_{i_0}, \beta\rangle)}{\ell'(\ell^{-1}(\sum_i \ell(y_i\langle x_i, \beta\rangle)))} \leq \frac{\ell'(y_{i_0}\langle x_{i_0}, \beta\rangle)}{\ell'(\ell^{-1}(\ell(y_{i_0}\langle x_{i_0}, \beta\rangle)))} \leq 1.$$

Therefore $q(\beta) \in (0, 1]^n$.

We further denote $a_t := -\ell'(\ell^{-1}(\sum_i \ell(y_i\langle x_i, \beta_t\rangle))) > 0$. This way we can write $\nabla L(\beta_t) = -a_t Z^\top q_t$ with $q_t := q(\beta_t)$

Integrating the flow we have that

$$\begin{aligned}\nabla\phi(\beta_t) &= \nabla\phi(\beta_0) - \int_0^t \nabla L(\beta_s)\mathrm{d}s \\ &= \nabla\phi(\beta_0) + Z^\top \int_0^t a_s q_s \mathrm{d}s.\end{aligned}$$

Similar to the time change we performed in Section 10.4, we consider $\theta(t) = \int_0^t a_s \mathrm{d}s$. From Lemma 40, θ is a bijection over $\mathbb{R}_{\geq 0}$ and perform the time change $\tilde{\beta}_t = \beta_{\theta^{-1}(t)}$. Due to the chain rule, after the time change and dropping the tilde notation we obtain:

$$\nabla\phi(\beta_t) = \nabla\phi(\beta_0) + Z^\top \int_0^t q_s \mathrm{d}s.$$

Dividing by t we get:

$$\frac{1}{t}\nabla\phi(\beta_t) = \frac{1}{t}\nabla\phi(\beta_0) + Z^\top \bar{q}_t, \tag{E.7}$$

where $\bar{q}_t \coloneqq \frac{1}{t}\int_0^t q_s \mathrm{d}s$ corresponds to the average of $(q_s)_{s\leq t}$.

Extracting a convergent subsequence: We now consider the normalised iterates $\bar{\beta}_t = \frac{\beta_t}{\|\beta_t\|}$ and up to an extraction we get that $\bar{\beta}_t \to \bar{\beta}_\infty$. Since q_t is a bounded function, up to a second extraction, we have that $q_t \to q_\infty$, and the same holds for its average: $\bar{q}_t \to q_\infty$. Taking the limit in Equation (E.8) we immediately obtain that:

$$\lim_t \frac{1}{t}\nabla\phi(\beta_t) = Z^\top q_\infty, \tag{E.8}$$

which also means that

$$\frac{\nabla\phi(\beta_t)}{\|\nabla\phi(\beta_t)\|} \xrightarrow[t\to\infty]{} \frac{Z^\top q_\infty}{\|Z^\top \bar{q}_\infty\|}$$

We can now directly apply Corollary 6 and there exists $\lambda > 0$ such that:

$$Z^\top q_\infty \in \lambda \partial\phi_\infty(\bar{\beta}_\infty)$$

The end of the proof is then as explained in Section 10.4.

□

Finally we recall and prove Theorem 7 which provides a simple formula for the horizon function in the case of separable potentials.

Theorem 7. *Under Assumption 14, there exists $\lambda > 0$ such that the horizon function ϕ_∞ of ϕ as defined in the previous section satisfies:*

$$\phi_\infty(\bar{\beta}) = \lambda \lim_{\eta\to 0} \eta \cdot \varphi^{-1}\left(\phi\left(\frac{\bar{\beta}}{\eta}\right)\right)$$

for every $\bar{\beta} \in \mathbb{R}^d$.

Proof. **Lipschitzness, upper and lower boundedness.** For $\eta > 0$, let us denote by $h_\eta : \beta \mapsto \eta \cdot \varphi^{-1}(\phi(\beta/\eta))$ and notice that $\nabla h_\eta(\beta) = (\frac{\varphi'(\beta_k/\eta)}{\varphi'(\varphi^{-1}(\sum_i \varphi(\beta_i/\eta)))})_{k\in[d]} \geq 0$ Since $\varphi \geq 0$ and that φ^{-1} and φ' are increasing we get that that $\nabla h_\eta(\beta) \in [0,1]^d$. Therefore $(h_\eta)_{\eta>0}$ are uniformly Lipschitz-continuous. Consequently, for all β, $h_\eta(\beta)$ is upper-bounded independently of η. Lastly, since $\varphi \geq 0$, notice that $h_\eta(\beta) \geq \min_i |\beta_i| > 0$ for all $\beta \neq 0$.

Point-wise and epi-convergence of h_η. For all $\bar{\beta}$, by composition, $\eta \mapsto \eta \cdot \varphi^{-1}(\phi(\bar{\beta}/\eta))$ is a definable function, the monotonicity Lemma [Van den Dries and Miller, 1996] (Theorem 4.1) ensures that it has a unique limit in $\mathbb{R}$ which we denote $h_0(\beta)$. From the uniform Lipschitzness of h_η, we get that $(\eta, \beta) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d \mapsto h_\eta(\beta)$ is continuous. Hence for all sequence $\eta_k \to 0$, we get that h_{η_k} epi-converges to h_0. Therefore (epi $h_{\eta_k})_k$ converges in the Painlevé–Kuratowski sense towards epi h_{η_0}.

Link between the level sets of h_η and those of ϕ. To conclude the proof it remains to notice that for all $c \geq 0$:

$$\{\beta \in \mathbb{R}^d, \phi(\beta) \leq c\} = \frac{1}{\eta}\{\bar{\beta} \in \mathbb{R}^d, h_\eta(\bar{\beta}) \leq \eta\varphi^{-1}(c)\}.$$

Therefore letting $\eta_c = 1/\varphi^{-1}(c)$ we get that

$$\eta_c \cdot S_c = \{\bar{\beta} \in \mathbb{R}^d, h_{\eta_c}(\bar{\beta}) \leq 1\}.$$

This simply means that η_c is an appropriate normalising quantity, it replaces the normalisation by the radius of S_c. Since $\{\bar{\beta} \in \mathbb{R}^d, h_{\eta_c}(\bar{\beta}) \leq 1\}$ converges in the Painlevé–Kuratowski sense towards $\{\bar{\beta} \in \mathbb{R}^d, h_0(\bar{\beta}) \leq 1\}$, we get that $R_c\eta_c \cdot \bar{S}_c$ converges towards the same set. However, with our previous construction, we also have that $\bar{S}_c$ converges towards $\bar{S}_\infty$. The sets $\bar{S}_\infty$ and $\{\bar{\beta} \in \mathbb{R}^d, h_0(\bar{\beta}) \leq 1\}$ are therefore proportional and $h_0 \propto \phi_\infty$ which concludes the proof.

□

Bibliography

E. Abbe, E. B. Adsera, and T. Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

K. Ahn, S. Bubeck, S. Chewi, Y. T. Lee, F. Suarez, and Y. Zhang. Learning threshold neurons via the "edge of stability". *arXiv preprint*, 2022.

A. Ali, E. Dobriban, and R. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pages 233–244. PMLR, 2020.

C. Aliprantis and K. Border. *Infinite Dimensional Analysis*. Springer Berlin, Heidelberg, 2006.

F. Alvarez. On the minimizing property of a second order dissipative system in hilbert spaces. *SIAM Journal on Control and Optimization*, 38(4):1102–1119, 2000.

E. Amid and M. K. Warmuth. Reparameterizing mirror descent as gradient descent. *Advances in Neural Information Processing Systems*, 33:8430–8439, 2020a.

E. Amid and M. K. Warmuth. Winnowing with gradient descent. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 163–182. PMLR, 09–12 Jul 2020b.

M. Andriushchenko, A. Varre, L. Pillaud-Vivien, and N. Flammarion. SGD with large step sizes learns sparse features. *arXiv preprint*, 2022.

V. Apidopoulos, N. Ginatta, and S. Villa. Convergence rates for the heavy-ball continuous dynamics for non-convex optimization, under polyak–lojasiewicz condition. *Journal of Global Optimization*, 84(3):563–589, 2022. ISSN 1573-2916.

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

P. Ashwin and M. Field. Heteroclinic networks in coupled cell systems. *Arch. Ration. Mech. Anal.*, 148(2):107–143, 1999.

H. Attouch and G. Beer. On the convergence of subdifferentials of convex functions. *Archiv der Mathematik*, 1993.

H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, i. the continuous dynamical system: Global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipitive dynamical system. *Communications in Contemporary Mathematics*, 2(1):1–34, 2000.

H. Attouch, J. Bolte, P. Redont, and M. Teboulle. Singular Riemannian barrier methods and gradient-projection dynamical systems for constrained optimization. *Optimization*, 53(5-6): 435–454, 2004.

H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 2011.

S. Azulay, E. Moroshko, M. S. Nacson, B. Woodworth, N. Srebro, A. Globerson, and D. Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. *arXiv preprint arXiv:2102.09769*, 2021.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Y. Bakhtin. Noisy heteroclinic networks. *Probab. Theory Related Fields*, 150(1-2):1–42, 2011.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, Jan. 2008.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

H. G. Bauschke and J. M. Borwein. Legendre functions and the method of random bregman projections. *Journal of Convex Analysis*, 4:27–67, 1997.

H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42 (2):330–348, 2017.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

R. Berthier. Incremental learning in diagonal linear networks. *arXiv preprint arXiv:2208.14673*, 2022.

G. Beugnot, J. Mairal, and A. Rudi. On the benefits of large learning rates for kernel methods. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 254–282. PMLR, 02–05 Jul 2022.

S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

E. Boix-Adsera, E. Littwin, E. Abbe, S. Bengio, and J. Susskind. Transformers learn through gradual rank increase. *arXiv preprint arXiv:2306.07042*, 2023.

J. Bolte, A. Daniilidis, and A. Lewis. Tame functions are semismooth. *Mathematical Programming*, 2007.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

E. Boursier, L. Pillaud-Vivien, and N. Flammarion. Gradient flow dynamics of shallow reLU networks for square loss and orthogonal inputs. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. ISSN 0041-5553.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

M. Burger, M. Möller, M. Benning, and S. Osher. An adaptive inverse scale space method for compressed sensing. *Mathematics of Computation*, 82(281):269–299, 2013.

J.-F. Cai, S. Osher, and Z. Shen. Split bregman methods and frame based image restoration. *Multiscale modeling & simulation*, 8(2):337–369, 2010.

B. Can, M. Gurbuzbalaban, and L. Zhu. Accelerated linear convergence of stochastic momentum methods in wasserstein distances. In *International Conference on Machine Learning*, pages 891–901. PMLR, 2019.

E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

I. Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17(8):1505–1508, 2013.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

L. Chen and J. Bruna. On gradient descent convergence beyond the edge of stability, 2022.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

X. Cheng, D. Yin, P. Bartlett, and M. Jordan. Stochastic gradient and Langevin processes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1810–1819. PMLR, 13–18 Jul 2020.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

L. Chizat and F. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

C. Combari and L. Thibault. On the graph convergence of subdifferentials of convex functions. *Proceedings of the American Mathematical Society*, 1998.

C. Cortes and V. Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

A. Cutkosky and H. Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

A. Damian, T. Ma, and J. D. Lee. Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, 2021.

A. Damian, E. Nichani, and J. D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023.

G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13:57–98, 1997.

A. Defazio. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis. *ArXiv*, 2020.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. L. Doob. *Stochastic Processes*. John Wiley & Sons, 1990.

C. Dossal. A necessary and sufficient condition for exact sparse recovery by ℓ_1 minimization. *C. R. Math. Acad. Sci. Paris*, 350(1-2), 2012.

R.-A. Dragomir. *Bregman gradient methods for relatively-smooth optimization*. PhD thesis, UT1 Capitole, 2021.

R. A. Dragomir, M. Even, and H. Hendrikx. Fast stochastic Bregman gradient methods: Sharp analysis and variance reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2815–2825. PMLR, 18–24 Jul 2021.

L. P. D. v. d. Dries. *Tame Topology and O-minimal Structures*. Cambridge University Press, 1998.

S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

M. Dudík, R. E. Schapire, and M. Telgarsky. Convex analysis at infinity: An introduction to astral space. *arXiv preprint arXiv:2205.03260*, 2022.

A. Eckert. OpenAI's Gigantic Salaries Are Making Apple, Google, Microsoft Pay Look... Meh. *Bezinga*, November 2023.

M. A. Efendiev and A. Mielke. On the rate-independent limit of systems with dry friction and small viscosity. *J. Convex Anal.*, 13(1):151–167, 2006.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. 2004.

A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1): 24–29, 2019.

M. Even. *Towards Decentralization, Asynchrony, Privacy and Personalization in Federated Learning*. PhD thesis, PSL Research University; Ecole normale supérieure, 2024. PhD manuscript currently under review.

M. Even and L. Massoulie. Concentration of non-isotropic random tensors with applications to learning and empirical risk minimization. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1847–1886. PMLR, 15–19 Aug 2021.

M. Even, S. Pesme, S. Gunasekar, and N. Flammarion. (s) gd over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695. PMLR, 2015.

S. Fort, G. K. Dziugaite, M. Paul, S. Kharaghani, D. M. Roy, and S. Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.

J. Geiping, M. Goldblum, P. E. Pope, M. Moeller, and T. Goldstein. Stochastic training is not necessary for generalization. In *International Conference on Learning Representations*, 2022.

E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

U. Ghai, E. Hazan, and Y. Singer. Exponentiated gradient meets gradient descent. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 386–407, San Diego, California, USA, 08 Feb–11 Feb 2020. PMLR.

A. Ghosh, H. Lyu, X. Zhang, and R. Wang. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

G. Gidel, F. Bach, and S. Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

D. Gissin, S. Shalev-Shwartz, and A. Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020.

X. Goudou and J. Munier. The gradient and heavy ball with friction dynamical systems: the quasiconvex case. *Mathematical Programming*, 116(1):173–191, 2009.

P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch SGD: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018a.

S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018b.

S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1827–1836. PMLR, 2018c.

S. Gunasekar, B. Woodworth, and N. Srebro. Mirrorless mirror descent: A natural derivation of mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pages 2305–2313. PMLR, 2021.

J. Z. HaoChen, C. Wei, J. Lee, and T. Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR, 2021.

A. Haraux and M. Jendoubi. Convergence of solutions of second-order gradient-like systems with analytic nonlinearities. *Journal of Differential Equations*, 144(2):313–320, 1998.

F. He, T. Liu, and D. Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, Jan. 1997.

P. D. Hoff. Lasso, fractional norm and structured sparse estimation using a hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.

E. Hoffer, I. Hubara, and D. Soudry. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1729–1739, 2017.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17(none):257 – 317, 2020. doi: 10.1214/18-PS321.

T. Hsu. Fake and Explicit Images of Taylor Swift Started on 4chan, Study Says. *The New York Times*, 2024.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

A. Jacot, F. Ged, B. Şimşek, C. Hongler, and F. Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, A. Storkey, and Y. Bengio. Three factors influencing minima in SGD. In *International Conference on Learning Representations*, 2018.

S. Jastrzebski, Z. Kenton, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *International Conference on Learning Representations*, 2019.

S. Jastrzkebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 392–402, 2018.

S. Jelassi and Y. Li. Towards understanding how momentum improves generalization in deep learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9965–10040. PMLR, 2022.

Z. Ji and M. Telgarsky. Risk and parameter convergence of logistic regression. *CoRR*, 2018.

Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186, 2020.

Z. Ji and M. Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

Z. Ji, M. Dudík, R. E. Schapire, and M. Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

L. Jiang, Y. Chen, and L. Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *arXiv preprint arXiv:2203.02839*, 2022.

C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

J. Jin, Z. Li, K. Lyu, S. S. Du, and J. D. Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. *arXiv preprint arXiv:2301.11500*, 2023.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

D. Kalimeris, G. Kaplun, P. Nakkiran, B. Edelman, T. Yang, B. Barak, and H. Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

K. Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

R. Kidambi, P. Netrapalli, P. Jain, and S. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

B. Kleinberg, Y. Li, and Y. Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707. PMLR, 10–15 Jul 2018.

P. E. Kloeden and E. Platen. Stochastic differential equations. In *Numerical Solution of Stochastic Differential Equations*, pages 103–160. Springer, 1992.

B. Kocel-Cynk, W. Pawłucki, and A. Valette. A short geometric proof that hausdorff limits are definable in any o-minimal structure. *advg*, 2014.

N. B. Kovachki and A. M. Stuart. Continuous time analysis of momentum methods. *J. Mach. Learn. Res.*, 22(1), 2021.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

M. Krupa. Robust heteroclinic cycles. *J. Nonlinear Sci.*, 7(2):129–176, 1997.

K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)*, 48(3):769–783, 1998.

M. Laghdir and M. Volle. A general formula for the horizon function of a convex composite function. *Archiv der Mathematik*, 1999.

G. Leclerc and A. Madry. The two regimes of deep network training. *arXiv preprint arXiv:2002.10376*, 2020.

Y. Lecun. The Epistemology of Deep Learning , February 2019. YouTube. Videospecific content referenced around time 48:00.

Y. Lecun, March 2023. LinkedIn post.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

B. Lemaire. An asymptotical variational principle associated with the steepest descent method for a convex function. *Journal of Convex Analysis*, 3:63–70, 1996.

Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019a.

Y. Li, C. Wei, and T. Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019b.

Z. Li, Y. Luo, and K. Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Z. Li, T. Wang, J. D. Lee, and S. Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. *Advances in Neural Information Processing Systems*, 35:34626–34640, 2022.

S. Liu, D. Papailiopoulos, and D. Achlioptas. Bad global minima exist and SGD can reach them. In *Advances in Neural Information Processing Systems*, volume 33, pages 8543–8552, 2020a.

Y. Liu, Y. Gao, and W. Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020b.

K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

V. Mai and M. Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International conference on machine learning*, pages 6630–6639. PMLR, 2020.

J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*, 2012.

S. Mandt, M. D. Hoffman, and D. M. Blei. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 354–363, 2016.

S. Marcotte, R. Gribonval, and G. Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. *Advances in Neural Information Processing Systems*, 36, 2024.

D. Masters and C. Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

C. Metz. Google's AI Wins Pivotal Second Game in Match With Go Grandmaster. *Wired*, March 2016.

A. Mielke. Evolution of rate-independent systems. *Evolutionary equations*, 2:461–559, 2005.

A. Mielke, R. Rossi, and G. Savaré. Modeling solutions with jumps for rate-independent systems on metric spaces. *Discrete Contin. Dyn. Syst.*, 25(2):585–615, 2009.

A. Mielke, R. Rossi, and G. Savaré. Variational convergence of gradient flows and rate-independent evolutions in metric spaces. *Milan Journal of Mathematics*, 80:381–410, 2012.

E. Moroshko, B. E. Woodworth, S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. In *Advances in Neural Information Processing Systems*, volume 33, pages 22182–22193, 2020.

C. Moucer, A. Taylor, and F. Bach. A systematic approach to lyapunov analyses of continuous-time models in convex optimization. *SIAM Journal on Optimization*, 33(3):1558–1586, 2023.

R. Mulayoff, T. Michaeli, and D. Soudry. The implicit bias of minima stability: A view from function space. In *Advances in Neural Information Processing Systems*, 2021.

M. Nacson, N. Srebro, and D. Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3051–3059. PMLR, 2019a.

M. S. Nacson, S. Gunasekar, J. Lee, N. Srebro, and D. Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4683–4692. PMLR, 09–15 Jun 2019b.

M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019c.

M. S. Nacson, K. Ravichandran, N. Srebro, and D. Soudry. Implicit bias of the step size in linear diagonal neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16270–16295. PMLR, 17–23 Jul 2022.

A. Nemirovski. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15, 1979.

A. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley–Blackwell, 1983.

B. Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

B. Neyshabur, R. R. Salakhutdinov, and N. Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *Advances in neural information processing systems*, 28, 2015.

R. O'Donnell. Analysis of boolean functions, 2021.

OpenAI. Sora: creating video from text, 2024.

F. Orabona, K. Crammer, and N. Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Mach. Learn.*, 99(3):411–435, jun 2015.

A. Orvieto, J. Kohler, and A. Lucchi. The role of memory in stochastic optimization. In *Uncertainty in Artificial Intelligence*, pages 356–366. PMLR, 2020.

M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

H. G. Papazov, S. Pesme, and N. Flammarion. Leveraging continuous time to understand momentum when training diagonal linear networks. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AIS-TATS) 2024,*, 2024.

D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 65–74. PMLR, 20–22 Apr 2017.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.

S. Pesme and N. Flammarion. Online robust regression via sgd on the l1 loss. *Advances in Neural Information Processing Systems*, 33:2540–2552, 2020.

S. Pesme and N. Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. *Advances in Neural Information Processing Systems*, 2023.

S. Pesme, A. Dieuleveut, and N. Flammarion. On convergence-diagnostic based step sizes for stochastic gradient descent. In *International Conference on Machine Learning*, pages 7641–7651. PMLR, 2020.

S. Pesme, L. Pillaud-Vivien, and N. Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, 2021.

L. Pillaud-Vivien, J. Reygner, and N. Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2127–2159. PMLR, 2022.

B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

B. Polyak and P. Shcherbakov. Lyapunov functions: An optimization theory perspective. *IFAC-PapersOnLine*, 50(1):7456–7461, 2017.

C. Poon and G. Peyré. Smooth bilevel programming for sparse regularization. *Advances in Neural Information Processing Systems*, 34:1543–1555, 2021.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

N. Razin, A. Maman, and N. Cohen. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pages 8913–8924. PMLR, 2021.

D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.

H. Robbins and S. Monro. A stochastic approxiation method. *Ann. Math. Statist*, 22(3):400–407, 1951.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer Berlin Heidelberg, 1998.

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *The Journal of Machine Learning Research*, 2004.

J. M. Sanz Serna and K. C. Zygalakis. The connections between lyapunov functions for some optimization algorithms and differential equations. *SIAM Journal on Numerical Analysis*, 59 (3):1542–1565, 2021.

A. M. Saxe, J. L. McClelland, and S. Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

O. Sebbouh, R. M. Gower, and A. Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971. PMLR, 2021.

B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.

S. L. Smith and Q. V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19(1):2822–2878, jan 2018.

W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

H. Sun, K. Ahn, C. Thrampoulidis, and N. Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. *Advances in Neural Information Processing Systems*, 35, 2022.

H. Sun, K. Gatmiry, K. Ahn, and N. Azizan. A unified approach to controlling implicit regularization via mirror descent. *arXiv preprint arXiv:2306.13853*, 2023.

T. Sun, D. Li, Z. Quan, H. Jiang, S. Li, and Y. Dou. Heavy-ball algorithms always escape saddle points. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 3520–3526. AAAI Press, 2019.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 2013. PMLR.

T. Tao. Concentration of measure. *254A, Notes 1, Blogpost*, 2010.

M. Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

R. J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Stat.*, 7:1456–1490, 2013.

L. Van den Dries and C. Miller. Geometric categories and o-minimal structures. 1996.

V. N. Vapnik. *The Nature of Statistical learning theory*. Springer New York, NY, 2nd edition, 1999.

G. Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66 (6):86–93, 2023.

A. Varre, L. Pillaud-Vivien, and N. Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. 2021.

A. V. Varre, M.-L. Vladarean, L. Pillaud-Vivien, and N. Flammarion. On the spectral bias of two-layer linear networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

T. Vaskevicius, V. Kanade, and P. Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

T. Vaskevicius, V. Kanade, and P. Rebeschini. The statistical complexity of early-stopped mirror descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 253–264. Curran Associates, Inc., 2020a.

T. Vaskevicius, V. Kanade, and P. Rebeschini. The statistical complexity of early-stopped mirror descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 253–264, 2020b.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

G. Wang, K. Donhauser, and F. Yang. Tight bounds for minimum *ell*_1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pages 10572–10602. PMLR, 2022a.

L. Wang, Z. Fu, Y. Zhou, and Z. Yan. The implicit regularization of momentum gradient descent in overparametrized models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):10149–10156, 2023.

Y. Wang, M. Chen, T. Zhao, and M. Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022b.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

A. C. Wilson, B. Recht, and M. I. Jordan. A lyapunov analysis of accelerated methods in optimization. *The Journal of Machine Learning Research*, 22(1):5040–5073, 2021.

J. S. Wind, V. Antun, and A. C. Hansen. Implicit regularization in ai meets generalized hardness of approximation in optimization–sharp results for diagonal linear networks. *arXiv preprint arXiv:2307.07410*, 2023.

S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type. Part I: Discrete time analysis, 2021.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020a.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020b.

F. Wu and P. Rebeschini. A continuous-time mirror descent approach to sparse phase retrieval. In *Advances in Neural Information Processing Systems*, volume 33, pages 20192–20203. Curran Associates, Inc., 2020.

F. Wu and P. Rebeschini. Implicit regularization in matrix sensing via mirror descent. *Advances in Neural Information Processing Systems*, 34:20558–20570, 2021.

J. Wu, D. Zou, V. Braverman, and G. Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. In *International Conference on Learning Representations*, 2021.

L. Wu, C. Ma, and W. E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Y. Yang, M. Möller, and S. Osher. A dual split bregman method for fast l1 minimization. *Mathematics of computation*, 82(284):2061–2085, 2013.

W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for l1-minimization with applications to compressed sensing: Siam journal on imaging sciences, 1, 143–168. *LIST OF FIGURES*, 2008.

C. Yun, S. Krishnan, and H. Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

P. Zhao, Y. Yang, and Q.-C. He. High-dimensional linear regression via implicit regularization. *arXiv preprint arXiv:1903.09367*, 2019.

X. Zhu, Z. Wang, X. Wang, M. Zhou, and R. Ge. Understanding edge-of-stability training dynamics with a minimalist example. *International Conference on Learning Representations*, 2023.

Scott Pesme

Curriculum Vitae

May 2024

scott.pesme@epfl.ch
scottpesme
Lausanne, Switzerland

Education

2019-2024 | **École Polytechnique Fédérale de Lausanne**
Ph.D. in the Theory of Machine Learning laboratory supervised by Prof. Nicolas Flammarion.

2018–2019 | **École Normale Supérieure Paris-Saclay**
Master Mathématiques Vision Apprentissage (MVA).

2015–2018 | **École Polytechnique, Palaiseau**
B.Sc. and M.Sc. in applied mathematics.

2013–2015 | **Lycée Henri IV, Paris**
Preparatory classses in mathematics and physics for the french grandes écoles.

Experience

2023 | **RIKEN AIP** · Research exchange (*5 months*) · Tokyo
Research stay in Taiji Suzuki's laboratory at the University of Tokyo and RIKEN AIP.

2019 | **EPFL** · Master thesis (*5 months*) · Lausanne
On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent.

2018 | **McGill University** · Intern at the Montreal Neurological Institute (*4 months*) · Montréal
Implementation of new methods for extracting event related brain potentials using neural networks.

2017 | **General Electric** · R&D intern (*3 months*) · Grenoble
Development of a software program in Python language that predicts the damage of water turbines.

2015 | **Paris Fire Brigade** · Military service (*7 months*) · Paris
École Polytechnique's mandatory military service as a paramedic team-leader in the fire brigade.

Publications

Leveraging Continuous Time to Understand Momentum When Training Diagonal Linear Networks.
H. Papazov, S. Pesme, N. Flammarion, AISTATS 2024.

Saddle-to-Saddle Dynamics in Diagonal Linear Networks.
S. Pesme, N. Flammarion, Neurips 2023.

(S)GD over Diagonal Linear Networks: Implicit Regularisation, Large Stepsizes and Edge of Stability.
M. Even, S. Pesme, S. Gunasekar, N. Flammarion, Neurips 2023.

Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity.
S. Pesme, L. Pillaud-Vivien, N. Flammarion, Neurips 2021.

Online Robust Regression via SGD on the ℓ_1 loss. S. Pesme, N. Flammarion, Neurips 2020.

On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent.
S. Pesme, A. Dieuleveut, N. Flammarion, ICML 2020.

Teaching Assistant

Machine Learning and Optimisation for ML courses at EPFL.
Mathematics and physics examiner at Lycée Henry IV preparatory classes.

Skills

Languages | Python, Matlab, Java, C++, R.

Languages

French (mother-tongue) · English (C2) · Spanish (B1)