Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity

Scott Pesme
EPFL
scott.pesme@epfl.ch

Loucas Pillaud-Vivien
EPFL
loucas.pillaud-vivien@epfl.ch

Nicolas Flammarion
EPFL
nicolas.flammarion@epfl.ch

Abstract

Understanding the implicit bias of training algorithms is of crucial importance in order to explain the success of overparametrised neural networks. In this paper, 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.

1 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 [3] and we steadily understand the convergence of training algorithms [29, 10] as well as the role of initialisation [20, 12]. 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 [39].

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 [31] 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.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

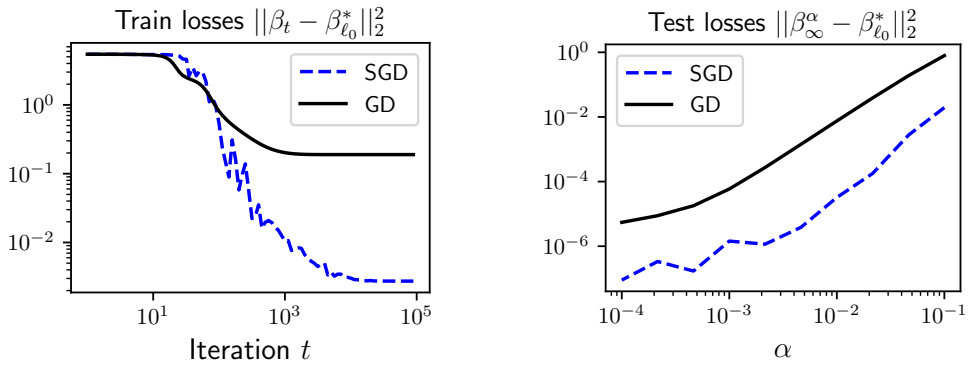

Figure 1: Sparse regression with $n = 40$, $d = 100$, $\|\beta^*_{\ell_0}\|_0 = 5$, $x_i \sim \mathcal{N}(0, I)$ $y_i = x_i^\top \beta^*_{\ell_0}$. ***Left***: for initialisation scale $\alpha = 0.05$, SGD converges towards a solution which generalises better than GD. *Right*: for different values of 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 5.1 for the precise experimental setup.

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 [23]. 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.

1.1 Main contributions and paper organisation.

In Section 2, we start by introducing the setup of our problem as well as the continuous modelisation of stochastic gradient descent. Then, in Section 3, we state our main result on the implicit bias of the stochastic gradient flow. We informally formulate it here and illustrate it in Figure 1:

Theorem 1 (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 [14, 36], 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 4 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 5.

1.2 Related work

As recalled, implicit bias has a recent history that has been initiated by the seminal work [31] 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 [14], linear networks [22] and more general homogeneous neural networks [27, 11]. In [36] the authors show that the scale of the initialisation leads to an interpolation between the neural tangent kernel regime [20, 12] (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 [28, 21, 16, 18, 24], 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 [23, 21, 16]. 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 [17, 23]. This idea has been further investigated by adopting a random walk on random landscape modelling [18], by suggesting that SGD's noise is smoothing the loss landscape, thus eliminating the sharp minima [24], by considering a dynamical stability perspective [38] or by interpreting SGD as a diffusion process [16, 21, 8]. Recently, label-noise has been shown to influence the implicit bias of SGD, by biasing the solution towards the origin for quadratically-parameterized models [15] or by implicitly regularising the expected squared norm of the gradient of the model with respect to the weights [5]. 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 [33] and stochastic gradient descent with label noise [15]. 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 [13] and further exploited by the same author in [37, 34] 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 [28, 21]. However, as highlighted by [1], 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 [26] that it can be done consistently. Yet, noise modelling still remains the principal issue [35] as it influences largely the behaviour of the dynamics [8, 9].

1.3 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 β^* 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 $\lim_{t \to \infty} z_t$.

2 Setup and preliminaries

2.1 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 $\beta^* \in \mathbb{R}^d$ which perfectly fits the training set, i.e. $y_i = \langle \beta^*, x_i \rangle$ 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:

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

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 [23]. 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 [36, 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. For the sake of completeness, the study of diagonal linear networks of arbitrary depth $p \geq 3$ is done in Appendix E.2. Also note that additionally to being a toy neural model, it has received recent attention for its practical ability to induce sparsity [33, 34, 15] or to solve phase retrieval problems [37].

Stochastic Gradient Descent. With this quadratic parametrisation, the loss now rewrites as: $L(w) = \frac{1}{4n} \sum_{i=1}^{n} \langle w_+^2 - w_-^2 - \beta^*, x_i \rangle^2$. 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 $\gamma > 0$:

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

It is convenient to rewrite this recursion as

$$w_{t+1,\pm} = w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \pm \gamma \, \mathrm{diag}(w_{t,\pm}) X^\top \xi_{i_t}(\beta_t), \tag{2}$$

where $\xi_{i_t}(\beta) = -\big(\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} - \mathbb{E}_{i_t}\big[\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t}\big]\big) \in \mathbb{R}^n$ 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 A.

2.2 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 (2), evaluated at $w_\pm$, the stochastic noise $\gamma \operatorname{diag}(w_\pm) X^\top \xi_{i_t}(w)$ has two main characteristics which we want to preserve:

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

It remains to understand the structure of the covariance of ξ_{i_t} which has the following closed form: $\operatorname{Cov}_{i_t}(\xi_{i_t}(\beta)) = \frac{1}{n}\operatorname{diag}(\langle \beta - \beta^*, x_i\rangle^2)_{1\le i\le n} - \frac{1}{n^2}\big(\langle \beta-\beta^*, x_i\rangle\langle \beta-\beta^*, x_j\rangle\big)_{1\le i,j\le n}$. 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 $\operatorname{Var}_{i_t}(\|\xi_{i_t}(\beta)\|_2) = \frac{4}{n}L(\beta) + O(\frac{1}{n^2})$. 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: $\Sigma_{\text{SGD}}(w_\pm) \cong \frac{4}{n}\gamma^2 L(w)[\operatorname{diag}(w_\pm)X^\top]^{\otimes 2}$.

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

$$\begin{aligned} \mathrm{d}w_{t,+} &= -\nabla_{w_+} L(w_t)\,\mathrm{d}t + 2\sqrt{\gamma n^{-1} L(w_t)}\; w_{t,+} \odot [X^\top \mathrm{d}B_t] \\ \mathrm{d}w_{t,-} &= -\nabla_{w_-} L(w_t)\,\mathrm{d}t - 2\sqrt{\gamma n^{-1} L(w_t)}\; w_{t,-} \odot [X^\top \mathrm{d}B_t], \end{aligned} \tag{3}$$

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 $\Sigma_{\text{SGD}}(w_\pm)$. We refer to [26] or [25] 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.

3 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 [36] 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 [13]:

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

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 $\phi_\alpha(\beta) \underset{\alpha\to+\infty}{=} -\frac{1}{2}\alpha^2 + \frac{1}{16\alpha^2}\|\beta\|_2^2 + o(\alpha^{-2})$. We refer to [36, Theorem 2] for more details on the asymptotic analysis. The implicit optimisation problem (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 [20, 12]. Overall, to retrieve a sparse solution, one should initialise with the smallest α possible. However, as is clearly explained in [36], 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.

[1] the general case is discussed in Appendix E.1

[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* (3) *with step size* $\gamma \leq O\big(\big[\ln(\frac{4}{p})\lambda_{\max} \max\{\|\beta^*_{\ell_1}\|_1 \ln\big(\frac{\|\beta^*_{\ell_1}\|_1}{\min_i \alpha_i^2}\big), \|\alpha\|_2^2\}\big]^{-1}\big)$ *where* $\beta^*_{\ell_1} = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta=y}{\arg\min} \|\beta\|_1$ *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*

$$\beta^\alpha_\infty = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta=y}{\arg\min} \phi_{\alpha_\infty}(\beta) \quad \textit{where} \quad \alpha_\infty = \alpha \odot \exp\left(-2\gamma \operatorname{diag}\left(\frac{X^\top X}{n}\right)\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s\right). \quad (5)$$

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 2. Now let us comment further the theorem.

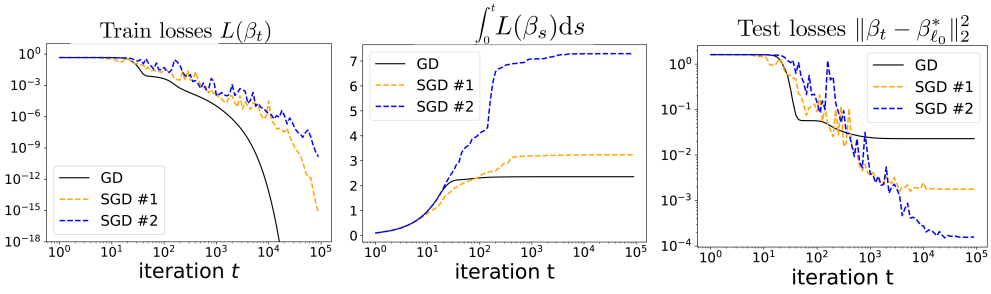

Figure 2: Sparse regression (see Section 5.1 for the detailed experimental setting). Both SGD and GD are initialised at $\alpha = 0.1$. 2 different runs of SGD over the training set are performed, they differ due to the inner stochasticity of the algorithm. *Left*: GD and SGD both converge towards a 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 [36], 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 2). Interestingly, Problem (5) tells us that the implicit bias of SGD initialised at α acts as if we run GD initialised at α_∞ (see Section 5.3). Note that the minimisation problem (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 3) and show under slightly stronger conditions that the relative scale α_∞/α decays as power of α (See Eq. (8) of the main text and Proposition 6 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 paper 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 [32] for a convergence result in the overparametrised least-squares setup). The explicit formula for the γ upper bound is $\gamma \leq \left(400 \ln\left(\frac{4}{p}\right) \lambda_{\max}(\frac{X^\top X}{n}) \max\left\{\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min_i \alpha_i^2}\right), \|\alpha\|_2^2\right\}\right)^{-1}$. It has a classical dependence on $\lambda_{\max}(X^\top X/n)$ which can be computed, but also on the unknown value of $\|\beta^*_{\ell_1}\|_1$. 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 [36, 14]. 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.

4 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 $\int_0^\infty L(\beta_s)\,\mathrm{d}s$.

4.1 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 [4, Theorem 2] for strongly convex potentials. 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 [2] 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:

$$\beta_\infty = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta = y}{\arg\min} D_\Psi(\beta, \beta_0), \tag{6}$$

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 [36] that the corresponding iterates $\beta_t = w_{t,+}^2 - w_{t,-}^2$ follow a mirror descent with potential ϕ_α defined in Eq.(4). Therefore we can apply the previous remarks to obtain the convergence towards an interpolator[3], as well as the associated implicit minimisation problem which in our case can be rewritten as $\beta^\alpha_\infty = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta = y}{\arg\min} \phi_\alpha(\beta)$ 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

[3] In our case, ϕ_α is not strongly convex so a bit more work is necessary to show the convergence of the iterates (see Appendix C).

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 1. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow in Eq.*(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:*

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma n^{-1} L(\beta_t)} X^\top \mathrm{d}B_t, \tag{7}$$

where $\alpha_t = \alpha \odot \exp\left(-2\gamma \operatorname{diag}\left(\frac{X^\top X}{n}\right) \int_0^t L(\beta_s)\,\mathrm{d}s\right)$ and ϕ_α is the hyperbolic entropy defined in (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) still yields $\nabla\phi_{\alpha_\infty}(\beta_\infty^\alpha) \in \operatorname{span}(X)$ which, along with $X\beta_\infty^\alpha = y$, are the KKT conditions of the implicit minimisation problem (5). We emphasise the fact that the structure of the noise, belonging to $\operatorname{span}(X)$, 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 [15].

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

Though it seems easy to derive the implicit minimisation problem (5) from the mirror-like structure of Eq.(7), it is necessary to ensure that the iterates converge towards an interpolator β_∞. This is the purpose of the following proposition.

Proposition 2 (Convergence of the iterates)**.** *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow* (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 β^* 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 3. *Under the same setting as in Proposition 2 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 7 (Appendix B.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: $\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s = \frac{1}{2} D_{\phi_\alpha}(\beta_\infty^\alpha, \beta_0)$ (see Proposition 7 in Appendix C). This matches the scale of the bounds given in Proposition 3, 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

$$\frac{\alpha_\infty}{\alpha} \underset{\alpha\to 0}{\leqslant} \left(\frac{\alpha^2}{\|\beta^*_{\ell_1}\|_1}\right)^\zeta, \tag{8}$$

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.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 [19].

5 Experiments

5.1 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 $\beta^*_{\ell_0}$ such that $\|\beta^*_{\ell_0}\|_0 = 5$. We generate the features as $x_i \sim \mathcal{N}(0, I)$ and the labels as $y_i = x_i^\top \beta^*_{\ell_0}$. 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 $\|\beta_t - \beta^*_{\ell_0}\|_2^2$ corresponds to the validation loss.

5.2 Validation of the SDE model

In this section, we present an experimental validation of the stochastic gradient flow model. In Figure 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 2.2.

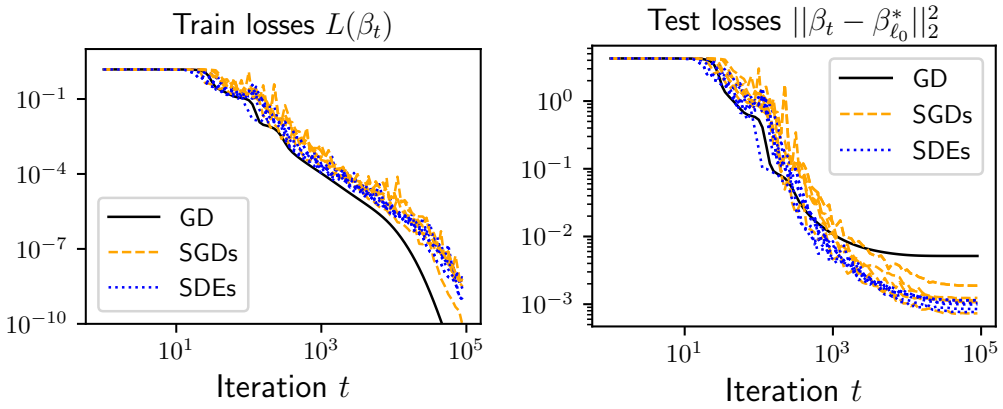

Figure 3: Sparse regression (see Section 5.1 for the detailed experimental setup). *Left and right*: the training and the validation losses behave very similarly, corroborating the continuous modelling.

5.3 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 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 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.

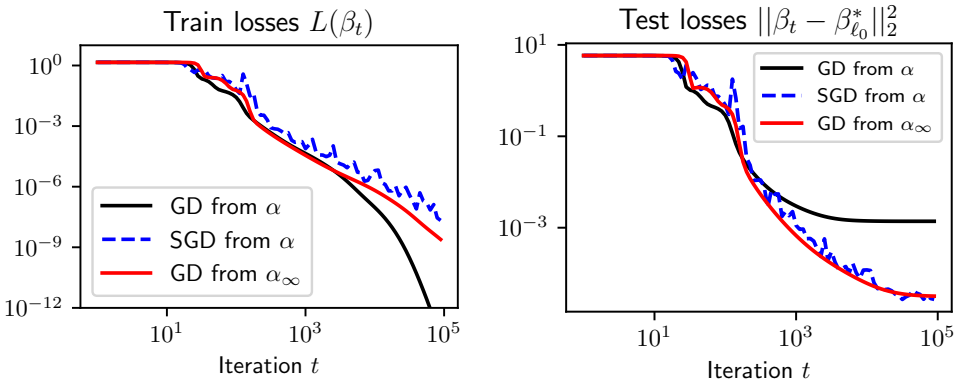

Figure 4: Sparse regression (see Section 5.1 for the detailed experimental setup). *Left and right*: SGD initialised at $\alpha\mathbf{1}$ converges towards the same point as GD initialised at $\alpha_\infty = \alpha \exp(-2\gamma \operatorname{diag}(X^\top X/n) \int_0^\infty L(\beta_s^{\mathrm{SGD}})\mathrm{d}s)$.

5.4 Doping the implicit bias with label noise

As largely discussed throughout the paper, 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:

$$w_{t+1,\pm} = w_{t,\pm} \mp \gamma \left(\langle \beta_w - \beta^*, x_{i_t}\rangle + \Delta_t\right) x_{i_t} \odot w_{t,+} \,, \qquad \text{where } i_t \sim \mathrm{Unif}(1,n). \tag{9}$$

As in Section 2.2, we can derive its related stochastic gradient flow (see Appendix D.1 for more details):

$$\mathrm{d}w_{t,\pm} = -\nabla_{w_\pm} L(w_t)\mathrm{d}t \pm 2\sqrt{\gamma n^{-1}(L(w_t) + \delta_t^2)}\, w_{t,+} \odot [X^\top \mathrm{d}B_t]. \tag{10}$$

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 5 in Appendix D.1.

6 Conclusion and Perspectives

In this paper, 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.

Acknowledgements. NF would like to thank Nathan Srebro for introducing him to the question of SGD's implicit bias as well as for the stimulating discussions they had during his visit at EPFL.

References

[1] Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pages 233–244. PMLR, 2020.

[2] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. *arXiv preprint arXiv:2102.09769*, 2021.

[3] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[4] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[5] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[7] Ioannis Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17(8):1505–1508, 2013.

[8] Pratik Chaudhari and Stefano 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.

[9] Xiang Cheng, Dong Yin, Peter Bartlett, and Michael 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.

[10] Lénaïc Chizat and Francis 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.

[11] Lenaic Chizat and Francis 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.

[12] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[13] Udaya Ghai, Elad Hazan, and Yoram 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.

[14] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[15] Jeff Z HaoChen, Colin Wei, Jason D Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*, 2020.

[16] Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Comput.*, 9(1):1–42, jan 1997.

[18] Elad Hoffer, Itay Hubara, and Daniel 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.

[19] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17(none):257 – 317, 2020. doi: 10.1214/18-PS321.

[20] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[21] Stanislaw Jastrzebski, Zac Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Amos Storkey, and Yoshua Bengio. Three factors influencing minima in SGD. In *International Conference on Learning Representations*, 2018.

[22] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

[23] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[24] Bobby Kleinberg, Yuanzhi Li, and Yang 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.

[25] Peter E Kloeden and Eckhard Platen. Stochastic differential equations. In *Numerical Solution of Stochastic Differential Equations*, pages 103–160. Springer, 1992.

[26] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[27] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

[28] Stephan Mandt, Matthew D. Hoffman, and David 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.

[29] Song Mei, Andrea Montanari, and Phan-Minh 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.

[30] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.

[31] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[32] Aditya Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. 2021.

[33] Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *arXiv preprint arXiv:1909.05122*, 2019.

[34] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. The statistical complexity of early-stopped mirror descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 253–264. Curran Associates, Inc., 2020.

[35] Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. Part I: Discrete time analysis, 2021.

[36] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[37] Fan Wu and Patrick 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.

[38] Lei Wu, Chao Ma, and Weinan 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.

[39] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Appendix

Organisation of the Appendix. The Appendix is structured as follows. In Section A, we give more precisions regarding the way we model stochastic gradient descent as a stochastic gradient flow. Section B.6 is the core of the Appendix as it provides the proof of the theorem in a self-contained fashion. For the sake of completeness, in Section C we gather the results on the link between mirror-descent and implicit bias as well as give convergence results in the deterministic case (gradient flow). In Section D.1, we provide more experiments supporting our results. In Section E.2, we discuss some extensions of our results ; (E.1) regarding a more general stochastic gradient flow model and in (E.2) we extend our results to depths $p \geq 3$. Finally, Section F provides the technical material needed for the proofs of our results.

A Details on the SDE modelling

We recall that the SGD recursion writes for $t \geqslant 1$ as:

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

Since the full gradient is $\nabla_{w_\pm} L(w) = \pm \big[\frac{1}{n} \sum_{k=1}^n \langle \beta_w - \beta^*, x_k \rangle \, x_k \big] \odot w_\pm \in \mathbb{R}^d$. We can rewrite the recursion as:

$$w_{t+1,\pm} = w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \mp \gamma \Big[\langle \beta_{w_t} - \beta^*, x_{i_t} \rangle \, x_{i_t} - \frac{1}{n} \sum_{k=1}^n \langle \beta_{w_t} - \beta^*, x_k \rangle \, x_k \Big] \odot w_{t,\pm}.$$

Now notice that

$$\langle \beta - \beta^*, x_{i_t} \rangle \, x_{i_t} - \frac{1}{n} \sum_{k=1}^n \langle \beta - \beta^*, x_k \rangle \, x_k = X^\top \Big(\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} - \mathbb{E}_{i_t} \big[\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} \big] \Big),$$

where $\mathbf{e}_i$ is the i^{th} element of the $\mathbb{R}^n$-canonical basis. Let us denote by $\xi_{i_t}(\beta) = -\big(\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} - \mathbb{E}_{i_t} \big[\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} \big] \big)$. 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 $\beta - \beta^*$, 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

$$\begin{aligned} \mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] &= \mathbb{E}_{i_t}[\xi_{i_t}(\beta)^{\otimes 2}] \\ &= \mathbb{E}_{i_t}[(\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t})^{\otimes 2}] - \mathbb{E}_{i_t}[\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t}]^{\otimes 2} \\ &= \frac{1}{n} \begin{pmatrix} \langle \beta - \beta^*, x_1 \rangle^2 & & 0 \\ & \ddots & \\ 0 & & \langle \beta - \beta^*, x_n \rangle^2 \end{pmatrix} - \frac{1}{n^2} \Big(\langle \beta - \beta^*, x_i \rangle \langle \beta - \beta^*, x_j \rangle \Big)_{1 \leq i,j \leq n} \\ &= \frac{4}{n} \begin{pmatrix} L_1(\beta) & & 0 \\ & \ddots & \\ 0 & & L_n(\beta) \end{pmatrix} - \frac{1}{n^2} \Big(\langle \beta - \beta^*, x_i \rangle \langle \beta - \beta^*, x_j \rangle \Big)_{1 \leq i,j \leq n} \end{aligned}$$

where $L_i(\beta) = \frac{1}{4} \langle \beta - \beta^*, x_i \rangle^2$ 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 E.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 in our paper therefore hold for mini-batch SGD by considering the effective step-size γ_{eff} instead of γ.

B 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.1, we derive the mirror-descent-like flow which the iterates follow as in Proposition 1 of the main text. Then, we upper bound the loss integral in subsection B.2. This leads us in proving the convergence of the iterates towards an interpolator in subsection B.3. Equipped with these results we prove the main result of the paper (Theorem 1) in subsection B.4. Finally, to complete the proof of Proposition 3 of the main text we derive a lower bound of the loss in subsection B.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.1 Proof of Proposition 1

In order to prove Proposition 1, we introduce the following lemma:

Lemma 1. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow in Eq.*(3) *with initialisation $w_{0,\pm} = \alpha \in (\mathbb{R}^*_+)^d$. Then we have the following implicit closed form expression for β_t:*

$$\beta_t = 2\alpha_t^2 \odot \sinh(2\bar{X}^\top \eta_t), \tag{11}$$

where $\eta_t = -\int_0^t \bar{X}(\beta_s - \beta^*)\,\mathrm{d}s + 2\sqrt{\gamma}\int_0^t \sqrt{L(\beta_s)}\mathrm{d}B_s \in \mathbb{R}^n$ and $\alpha_t = \alpha \odot \exp\big(-2\gamma \operatorname{diag}(\bar{X}^\top \bar{X}) \int_0^t L(\beta_s)\mathrm{d}s\big)$.

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 $r(w) = \bar{X}(w_+^2 - w_-^2 - \beta^*) = \bar{X}(\beta_w - \beta^*) \in \mathbb{R}^n$ 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:

$$w_{t,\pm} = w_{t=0,\pm} \odot \exp(\pm \bar{X}^\top \Big[- \int_0^t r(w_s)\,\mathrm{d}s + 2\sqrt{\gamma} \int_0^t \sqrt{L(w_s)}\,\mathrm{d}B_s \Big]) \odot \exp(-2\gamma\,\mathrm{diag}(\bar{X}^\top \bar{X}) \int_0^t L(w_s)\,\mathrm{d}s)$$
$$= \alpha_t \odot \exp(\pm \bar{X}^\top \eta_t).$$

Since $\beta = w_+^2 - w_-^2$, we get:

$$\beta_t = \alpha_t^2 \odot \big(\exp(+2\bar{X}^\top \eta_t) - \exp(-2\bar{X}^\top \eta_t)\big)$$
$$= 2\alpha_t^2 \odot \sinh(+2\bar{X}^\top \eta_t).$$

□

For clarity we recall the statement of Proposition 1.

Proposition 1. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow in Eq.(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:*

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma n^{-1} L(\beta_t)} X^\top \mathrm{d}B_t, \tag{7}$$

where $\alpha_t = \alpha \odot \exp\big(-2\gamma\,\mathrm{diag}\Big(\frac{X^\top X}{n}\Big) \int_0^t L(\beta_s)\,\mathrm{d}s\big)$ and ϕ_α is the hyperbolic entropy defined in (4)*.*

Proof. The results immediately follows from Lemma 1. Indeed, inverting the implicit equation on β_t, Eq. (11), we have,

$$\mathrm{arcsinh}\Big(\frac{\beta_t}{2\alpha_t^2}\Big) = 2X^\top \eta_t = -2\bar{X}^\top \int_0^t \bar{X}(\beta_s - \beta^*)\,\mathrm{d}s + 4\sqrt{\gamma}\bar{X}^\top \int_0^t \sqrt{L(\beta_s)}\mathrm{d}B_s.$$

Hence,

$$\mathrm{d}\,\mathrm{arcsinh}\Big(\frac{\beta_t}{2\alpha_t^2}\Big) = -2\bar{X}^\top \bar{X}(\beta_s - \beta^*)\,\mathrm{d}t + 4\sqrt{\gamma}\bar{X}^\top \sqrt{L(\beta_t)}\mathrm{d}B_t$$
$$= -4\nabla L(\beta_t)\,\mathrm{d}t + 4\sqrt{\gamma L(\beta_t)}\bar{X}^\top \mathrm{d}B_t.$$

Noticing that $\nabla\phi_\alpha(\beta) = \frac{1}{4}\mathrm{arcsinh}(\frac{\beta}{2\alpha^2})$ concludes the proof. □

B.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 3, 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. (12)]. Then, we fix a high-probability event on which we have a control of the Brownian diffusion term [Eq. (13)]. This gives an equation involving a weighted integral of the loss. After lower bounding this weight to access directly the loss integral [Lemma 4], we show that the iterates themselves are in fact bounded [Lemma 3]. We finally conclude the proof in Proposition 4.

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

$$\nabla_\beta \phi(\beta, \alpha^2) = \frac{1}{4}\operatorname{arcsinh}\left(\frac{\beta}{2\alpha^2}\right), \quad \nabla_z \phi(\beta, \alpha^2) = -\frac{1}{4\alpha^2}\sqrt{\beta^2 + 4\alpha^4} \in \mathbb{R}^d \qquad \text{as well as,}$$

$$\nabla^2_{\beta,\beta} \phi(\beta, \alpha^2) = \frac{1}{4}\operatorname{diag}\left[\frac{1}{\sqrt{\beta_i^2 + 4\alpha_i^2}}\right]_i \in \mathbb{R}^{d\times d}.$$

A first Lyapunov function. In this subsection we shall consider the following (stochastic) Lyapunov function:

$$V_t := -\phi_{\alpha_t}(\beta_t) + \langle \nabla \phi_{\alpha_t}(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle + \gamma \int_0^t L(\beta_s)\mathrm{d}s\, \langle |\beta^*_{\ell_1}|, \operatorname{diag}(\bar{X}^\top \bar{X})\rangle. \tag{12}$$

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 2. *For all $t > 0$, V_t verifies the following equation:*

$$V_t = V_0 - 2\int_0^t L(\beta_s)\Big(1 - \frac{1}{2}\gamma\langle \operatorname{diag}(\bar{X}^\top \bar{X}), \xi_s + |\beta^*_{\ell_1}|\rangle\Big)\mathrm{d}s + \int_0^t \sqrt{\gamma L(\beta_s)}\langle X^\top \mathrm{d}B_s, \beta_s - \beta^*_{\ell_1}\rangle.$$

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 $\alpha_t = \alpha \odot \exp\big(-2\gamma \operatorname{diag}(\bar{X}^\top \bar{X})\int_0^t L(w_s)\mathrm{d}s\big)$, we have $\mathrm{d}\alpha_t = -2\gamma \operatorname{diag}(\bar{X}^\top \bar{X})L(w_t)\alpha_t \mathrm{d}t$. 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 ${\beta_t}^4$, 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 $\mathrm{d}\langle \beta_t\rangle_{\mathrm{qv}} = \left[\mathrm{d}\langle \beta_t^i, \beta_t^j\rangle\right]_{ij} = 16\gamma L(\beta_t)(\bar{X}^\top \bar{X}) \odot (\xi_t \xi_t^\top) \in \mathbb{R}^{d\times d}$.

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,

$$\begin{aligned}
\mathrm{d}V_t = -\Big[&\langle \nabla_\beta \phi(\beta_t, \alpha_t^2), \mathrm{d}\beta_t\rangle + \langle \nabla_z \phi(\beta_t, \alpha_t^2), \mathrm{d}\left[\alpha_t^2\right]\rangle + \frac{1}{2}\mathrm{Tr}\left[\nabla^2_{\beta,\beta}\phi(\beta_t, \alpha_t^2)\mathrm{d}\langle \beta_t\rangle\right]\Big] \\
&+ \mathrm{d}\left[\langle \nabla_\beta \phi(\beta_t, \alpha_t^2), \beta_t - \beta^*_{\ell_1}\rangle\right] + \gamma L(\beta_t)\langle |\beta^*_{\ell_1}|, \operatorname{diag}(\bar{X}^\top \bar{X})\rangle \mathrm{d}t.
\end{aligned}$$

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:

$$\langle \nabla_z \phi(\beta_t, \alpha_t^2), \mathrm{d}\left[\alpha_t^2\right]\rangle = \left\langle -\frac{\xi_t}{4\alpha_t^2}, 2\alpha_t \odot \mathrm{d}\alpha_t \right\rangle = \gamma L(\beta_t)\left\langle \xi_t, \operatorname{diag}(\bar{X}^\top \bar{X})\right\rangle \mathrm{d}t.$$

Third term. We directly see that

$$\frac{1}{2}\mathrm{Tr}\left[\nabla^2_{\beta,\beta}\phi(\beta_t, \alpha_t^2)\mathrm{d}\langle \beta_t\rangle\right] = \frac{1}{2}\mathrm{Tr}\left[\frac{1}{4}\operatorname{diag}\left(\frac{1}{\xi_t}\right) \cdot 4\gamma L(\beta_t)\bar{X}^\top \bar{X} \odot (\xi_t \xi_t^\top)\right]\mathrm{d}t = 2\gamma L(\beta_t)\langle \xi_t, \operatorname{diag}(\bar{X}^\top \bar{X})\rangle \mathrm{d}t.$$

[4] It can be computed but its precise formula is not needed.

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), we have an expression for the first and last term, giving

$$\mathrm{d}\left[\langle \nabla_\beta \phi(\beta_t, \alpha_t^2), \beta_t - \beta^*_{\ell_1}\rangle\right] = -\langle \nabla L(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle \mathrm{d}t + 2\sqrt{\gamma L(\beta_t)}\langle \bar{X}^\top \mathrm{d}B_t, \beta_t - \beta^*_{\ell_1}\rangle + \langle \nabla_\beta \phi(\beta_t, \alpha_t^2), \mathrm{d}\beta_t\rangle \\ + 4\gamma L(\beta_t)\langle \xi_t, \mathrm{diag}(\bar{X}^\top \bar{X})\rangle \mathrm{d}t.$$

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 $\langle \nabla_\beta L(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle = 2L(\beta_t)$,

$$\mathrm{d}V_t = -\left[\gamma L(\beta_t)\left\langle \xi_t, \mathrm{diag}(\bar{X}^\top \bar{X})\right\rangle \mathrm{d}t + 2\gamma L(\beta_t)\langle \xi_t, \mathrm{diag}(\bar{X}^\top \bar{X})\rangle \mathrm{d}t\right] - 2L(\beta_t) \\ + \sqrt{\gamma L(\beta_t)}\langle \bar{X}^\top \mathrm{d}B_t, \beta_t - \beta^*_{\ell_1}\rangle + 4\gamma L(\beta_t)\langle \xi_t, \mathrm{diag}(\bar{X}^\top \bar{X})\rangle \mathrm{d}t + \gamma L(\beta_t)\langle |\beta^*_{\ell_1}|, \mathrm{diag}(\bar{X}^\top \bar{X})\rangle \mathrm{d}t.$$

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 2 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 [6]. To do so, let us fix $p < 1/2$ and we define two parameters: $a := \max\{\|\beta^*_{\ell_1}\|_1 \ln(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\}$ 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 3 and 4. These parameters being fixed, we can define the event:

$$\mathcal{A} = \{\forall t \geq 0, |S_t| \leq a + 2b\gamma\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\mathrm{d}s\}. \tag{13}$$

From Lemma 9, 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 2, 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:

$$V_t - V_0 \leq -2\int_0^t L(\beta_s)U_s\mathrm{d}s + a, \tag{14}$$

where $U_t := 1 - \frac{\gamma}{2}\left[\langle \mathrm{diag}(\bar{X}^\top \bar{X}), \xi_t + |\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}(\|\beta_t\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\right] \leq 1$. The following lemma show that as long as U_t stays positive, the iterates stay bounded.

Lemma 3. *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 $\|\xi_t\|_1$,*

$$\|\beta_t\|_1 \leq \|\xi_t\|_1 \leq 18\max\{\|\beta^*_{\ell_1}\|_1 \ln(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\}.$$

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 $-\phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle = \frac{1}{4}\|\xi_t\|_1 - \frac{1}{4}\langle \mathrm{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle$. Hence, we have:

$$\begin{aligned}
\|\xi_t\|_1 &= -4\phi_{\alpha_t}(\beta_t) + 4\langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta^*_{\ell_1}\rangle + \langle \mathrm{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle \\
&= 4V_t - 4\gamma\int_0^t L(\beta_s)\mathrm{d}s\,\langle |\beta^*_{\ell_1}|, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle + \langle \mathrm{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle \\
&\leq 2\|\alpha\|_2^2 + 4a + \langle \mathrm{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle - 4\gamma\int_0^t L(\beta_s)\mathrm{d}s\,\langle |\beta^*_{\ell_1}|, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle.
\end{aligned}$$

We now use the fact that $\mathrm{arcsinh}(x) \leq \ln(2(x+1))$ and that $|x| + |y| \leq \sqrt{2}\sqrt{x^2+y^2}$ for all $x, y \geq 0$.

$$\begin{aligned}
\|\xi_t\|_1 &\leq 2\|\alpha\|_2^2 + 4a + \sum_i |\beta_i^*| \ln\left(\frac{|\beta_i(t)| + 2\alpha_i(t)^2}{\alpha_i(t)^2}\right) - 4\gamma\int_0^t L(\beta_s)\mathrm{d}s\,\langle |\beta^*_{\ell_1}|, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle \\
&\leq 2\|\alpha\|_2^2 + 4a + \sum_i |\beta_i^*| \ln\left(\sqrt{2}\frac{\sqrt{|\beta_i(t)|^2 + 4\alpha_i(t)^4}}{\min \alpha_i^2}\right) - \sum_i |\beta_i^*| \ln\left(\exp\Big(-4\gamma\int_0^t L(\beta_s)\mathrm{d}s\,\mathrm{diag}(\bar{X}^\top\bar{X})\Big)\right) \\
&\qquad - 4\gamma\int_0^t L(\beta_s)\mathrm{d}s\,\langle |\beta^*_{\ell_1}|, \mathrm{diag}(\bar{X}^\top\bar{X})\rangle.
\end{aligned}$$

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 10 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,

$$\begin{aligned}
\|\beta_t\|_1 \leq \|\xi_t\|_1 &\leq \frac{5}{2}\left(2\|\alpha\|_2^2 + 4a + \|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right)\right) \\
&\leq 3\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right) + 5\|\alpha\|_2^2 + 10a \\
&\leq 18\max\{\|\beta^*_{\ell_1}\|_1 \ln(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\},
\end{aligned}$$

where in the last inequality we plug in the value of a. This concludes the proof of the lemma. □

Recall that we defined $U_t = 1 - \frac{\gamma}{2}\big[\langle \mathrm{diag}(\bar{X}^\top\bar{X}), \xi_t + |\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}(\|\beta_t\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\big]$. We now show that in fact $(U_t)_t$ is always lower bounded by a strictly positive constant. Hence, the result of Lemma 3 is valid at any time $t > 0$.

Lemma 4. *On $\mathcal{A}$, let us fix $\gamma \leq [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^*_{\ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\}]^{-1}$. Recall that $U_t = 1 - \frac{\gamma}{2}\big[\langle \mathrm{diag}(\bar{X}^\top\bar{X}), \xi_t + |\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}(\|\beta_t\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\big]$, 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

$$\begin{aligned}
U_0 &= 1 - \frac{\gamma}{2}\big[\langle \mathrm{diag}(\bar{X}^\top \bar{X}), 2\alpha^2 + |\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}\|\beta^*_{\ell_1}\|_1^2)\big] \\
&\geq 1 - \frac{\gamma}{2}\lambda_{\max}\big[2\|\alpha\|_2^2 + \|\beta^*_{\ell_1}\|_1 + 2b\|\beta^*_{\ell_1}\|_1^2) \\
&\geq 1 - 2\gamma\lambda_{\max} a \ln(\frac{4}{p}) \\
&> \frac{1}{2},
\end{aligned}$$

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 3 at time τ:

$$\|\beta_\tau\|_1 \leq \|\xi_\tau\|_1 \leq 18 \max\{\|\beta^*_{\ell_1}\|_1 \ln\Big(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\Big), \|\alpha\|_2^2\}.$$

Therefore:

$$\begin{aligned}
U_\tau &= 1 - \frac{\gamma}{2}\big[\langle \mathrm{diag}(\bar{X}^\top \bar{X}), \xi_\tau + |\beta^*_{\ell_1}|\rangle + 2b\lambda_{\max}(\|\beta_\tau\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\big] \\
&\geq 1 - \frac{\gamma}{2}\lambda_{\max}\big[\|\xi_\tau\|_1 + \|\beta^*_{\ell_1}\|_1 + 2b(\|\beta_\tau\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\big] \\
&\geq 1 - \frac{\gamma}{2}\lambda_{\max}\Big[18 \max\{\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\Big), \|\alpha\|_2^2\} \\
&\qquad\qquad + 2 \cdot 18^2 \cdot b \max\{\|\beta^*_{\ell_1}\|_1^2 \ln^2\Big(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\Big), \|\alpha\|_2^4\}\Big].
\end{aligned}$$

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:

$$\begin{aligned}
U_\tau &\geq 1 - \frac{\gamma}{2}\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\Big), \|\alpha\|_2^2\}[18 + 18^2] \\
&\geq 1 - 175\ln(\frac{4}{p})\gamma\lambda_{\max}\max\{\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\Big), \|\alpha\|_2^2\} \\
&> \frac{1}{2},
\end{aligned}$$

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 4, with Equation (14), we obtain:

$$\int_0^t L(\beta_s)\mathrm{d}s \leq V_0 - V_t + a \leq -V_t + 2\max\{\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\Big), \|\alpha\|_2^2\}. \tag{15}$$

Hence it remains to lower bound V_t in order to get the convergence of the integral of the loss.

Lemma 5. *On $\mathcal{A}$, let γ be set as in Lemma 3, for all $t > 0$, we have the following lower bound on V_t:*

$$V_t \geq -\frac{\|\beta^*_{\ell_1}\|_1}{4} \ln\Big(\frac{18\sqrt{2}}{\min \alpha_i^2} \max\Big\{\|\beta^*_{\ell_1}\|_1 \ln\Big(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\Big), \|\alpha\|_2^2\Big\}\Big).$$

Proof. We follow exactly the same proof as for upperbounding the iterates.

$$
\begin{aligned}
4V_t &= \sum_i \sqrt{\beta_i^2 + 4\alpha_i(t)^4} - \langle \operatorname{arcsinh}\frac{\beta_t}{2\alpha_t^2}, \beta^*_{\ell_1}\rangle + 4\gamma \int_0^t L(\beta_s)\mathrm{d}s \, \langle |\beta^*_{\ell_1}|, \operatorname{diag} H\rangle \\
&\geq \|\xi_t\|_1 - \sum_i |\beta_i^*| \ln\left(\frac{|\beta_i(t)| + 2\alpha_i(t)^2}{\alpha_i(t)^2}\right) + 4\gamma \int_0^t L(\beta_s)\mathrm{d}s \, \langle |\beta^*_{\ell_1}|, \operatorname{diag} H\rangle \\
&\geq \|\xi_t\|_1 - \|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\xi_t\|_1}{\min \alpha_i^2}\right) \\
&\geq -\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\xi_t\|_1}{\min \alpha_i^2}\right) \\
&\geq -\|\beta^*_{\ell_1}\|_1 \ln\Big(\frac{18\sqrt{2}}{\min \alpha_i^2} \max\left\{\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\right\}\Big).
\end{aligned}
$$

□

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 3. We recall it here in the following proposition.

Proposition 4. *On $\mathcal{A}$, let γ be set as in Lemma 3, we have the following upper bound on the loss integral:*

$$
\forall t > 0, \qquad \int_0^t L(\beta_s)\mathrm{d}s \leq \tilde{O}\left(\max\left\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\right\}\right).
$$

As a consequence, the integral $\int_0^\infty L(\beta_s)\mathrm{d}s$ converges.

Proof. From Equation (15), 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 5, it yields,

$$
\int_0^t L(\beta_s)\mathrm{d}s \leq \frac{\|\beta^*_{\ell_1}\|_1}{4} \ln\Big(\frac{18\sqrt{2}}{\min \alpha_i^2} \max\left\{\|\beta^*_{\ell_1}\|_1 \ln\left(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\right\}\Big) + 2\max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}\right), \|\alpha\|_2^2\},
$$

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.3 Proof of the convergence of the iterates: Proposition 2

In this subsection we prove the convergence of the iterates which corresponds to Proposition 2 of the main text. For the sake of completeness, we recall this fact in the following lemma.

Lemma 6. *On $\mathcal{A}$, let $\gamma \leq [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^*_{\ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\min \alpha_i^2}), \|\alpha\|_2^2\}]^{-1}$. 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 β^* :

$$
W_t = \phi_{\alpha_\infty}(\beta^*) - \phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta^*\rangle,
$$

where $\alpha_\infty = \alpha \exp\Big(-2\gamma\big(\int_0^\infty L(\beta_s)\mathrm{d}s\big)\operatorname{diag}(\bar{X}^\top\bar{X})\Big) > 0$ is well defined on $\mathcal{A}$ as a result of Proposition 4. The exact same computations as in Lemma 2 lead to:

$$
W_t = W_0 - 2\int_0^t L(\beta_s)\mathrm{d}s + \langle \operatorname{diag}(\bar{X}^\top\bar{X}), \gamma\int_0^t L(\beta_s)\xi_s\mathrm{d}s\rangle + \sqrt{\gamma}\int_0^t \sqrt{L(\beta_s)}\langle X^\top \mathrm{d}B_s, \beta_s - \beta^*\rangle.
$$

Note that:

- $\int_0^t L(\beta_s)\mathrm{d}s$ converges from Proposition 4.

- $\int_0^t \|L(\beta_s)\xi_s\|_1 \mathrm{d}s \leq \max_{s\geq 0}(\|\xi_s\|_1)\int_0^t L(\beta_s)\mathrm{d}s < \infty$ from Proposition 4. Hence $\int_0^t L(\beta_s)\xi_s \mathrm{d}s$ is absolutely convergent, hence converges.
- $\int_0^t \sqrt{L(\beta_s)}\langle X^\top \mathrm{d}B_s, \beta_s - \beta^*\rangle$ has a quadratic variation equal to $4\int_0^t L(\beta_s)^2\mathrm{d}s$ and $4\int_0^t L(\beta_s)^2\mathrm{d}s \leq 2\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_2^2 + \|\beta^*\|_1^2)\mathrm{d}s$. This implies that the quadratic variation converges. Hence we obtain the convergence [5] of the Brownian integral $\int_0^t \sqrt{L(\beta_s)}\langle X^\top \mathrm{d}B_s, \beta_s - \beta^*\rangle$.

Overall we get that W_t converges for all choice of interpolator β^*. 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 3 and 4), 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 $\beta^* = \beta_\infty$ 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:

$$\begin{aligned} W_t &= \phi_{\alpha_\infty}(\beta_\infty^\alpha) - \phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta_\infty^\alpha\rangle \\ &\geq \phi_{\alpha_t}(\beta_\infty^\alpha) - \phi_{\alpha_t}(\beta_t) + \langle \nabla\phi_{\alpha_t}(\beta_t), \beta_t - \beta_\infty^\alpha\rangle \\ &= D_{\phi_{\alpha_t}}(\beta_\infty^\alpha, \beta_t) \\ &\geq 0 \end{aligned}$$

where the first inequality is because $\alpha \mapsto \phi_\alpha(\beta)$ is decreasing and $\alpha_t \geq \alpha_\infty$. Therefore $D_{\phi_{\alpha_t}}(\beta_\infty^\alpha, \beta_t) \to 0$. Finally, since:

$$\begin{aligned} \nabla^2\phi_{\alpha_t}(\beta_t) &= \mathrm{diag}\Big(\frac{1}{\sqrt{\beta_i(t)^2 + 4\alpha_t^4(i)}}\Big)_i \\ &\geq \mathrm{diag}\Big(\frac{1}{\sqrt{\max_s\{\beta_i(s)^2\} + 4\alpha^4}}\Big)_i \\ &\geq \mathrm{diag}\Big(\frac{1}{\sqrt{\max_s\{\|\beta(s)\|_1^2\} + 4\alpha^4}}\Big)_i \\ &\geq \mu I_d, \end{aligned}$$

for some μ since the iterates are bounded. Therefore for all $t \geq 0$, ϕ_{α_t} is μ-strongly convex on some convex set in which the iterates β_s stay in. Which means that: $D_{\phi_{\alpha_t}}(\beta_\infty^\alpha, \beta_t) \geq \frac{\mu}{2}\|\beta_t - \beta_\infty^\alpha\|_2^2$. Hence $\beta_t \to \beta_\infty^\alpha$. □

Lemma 6 along with the fact that the event $\mathcal{A}$ has probability at least $1 - \frac{p}{2}$ (see Lemma 9 and paragraph around 13) concludes the proof of Proposition 2.

B.4 Proof of Theorem 1

We are now equipped to prove the main result of the paper. For clarity we recall the statement of the theorem here.

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* (3) *with step size $\gamma \leq O\big(\big[\ln(\frac{4}{p})\lambda_{\max}\max\{\|\beta_{\ell_1}^*\|_1 \ln\big(\frac{\|\beta_{\ell_1}^*\|_1}{\min_i \alpha_i^2}\big), \|\alpha\|_2^2\}\big]^{-1}\big)$ where $\beta_{\ell_1}^* = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta = y}{\arg\min} \|\beta\|_1$ 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*

[5] 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 [30].

- *the solution β^α_∞ satisfies*

$$\beta^\alpha_\infty = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta=y}{\arg\min}\ \phi_{\alpha_\infty}(\beta) \quad \textit{where} \quad \alpha_\infty = \alpha \odot \exp\left(-2\gamma \operatorname{diag}\left(\frac{X^\top X}{n}\right)\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s\right). \tag{5}$$

Proof. Recall first that on $\mathcal{A}$, Lemma 6 implies that the iterates converge towards a zero-training error we denote by β^α_∞. From Proposition 1 we also have that:

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma L(\beta_t)}\bar{X}^\top \mathrm{dB}_t, \tag{16}$$

where $\alpha_t = \alpha \odot \exp\left(-2\gamma \operatorname{diag}\left(\bar{X}^\top\bar{X}\right)\int_0^t L(\beta_s)\,\mathrm{d}s\right)$ and ϕ_α is the hyperbolic entropy defined in (4). Since the quantity $\int_0^\infty L(\beta_s)\,\mathrm{d}s$ is well defined on $\mathcal{A}$ (Proposition 4), we can integrate (16) 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 (5). From Lemma 9, the fact that the event $\mathcal{A}$ has probability at least $1-p$ concludes the proof. □

B.5 Lower bound on $\int L(\beta_s)\mathrm{d}s$ and proof of Proposition 3

Similarly to what has been done in subsection B.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 7 for the definition of W_0^α. Now that these parameters are fixed, consider the new event:

$$\mathcal{B} = \{\forall t \geq 0, |S_t| \leq \hat{a} + 2\hat{b}\gamma\lambda_{\max}\int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^*_{\ell_1}\|_1^2)\mathrm{d}s\}$$

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 7. *On $\mathcal{A}\cap\mathcal{B}$, let* $\gamma \leq [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^*_{\ell_1}\|_1\ln(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\alpha^2}), \alpha^2 d\}]^{-1}$,

$$\gamma\int_0^{+\infty} L(\beta_s)\mathrm{d}s \geq \frac{W_0^\alpha}{4}\frac{\gamma}{1+\gamma\frac{M}{W_0^\alpha}},$$

where $W_0^\alpha = \underset{\beta \text{ s.t } X\beta=Y}{\min}\phi_\alpha(\beta) - \phi_\alpha(0)$ *and* $M = \left[325\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^*_{\ell_1}\|_1^2\ln^2(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\alpha^2}), \alpha^4 d^2\}\right]$.

Proof. According to Lemma 6, 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 2:

$$\begin{aligned} 2\int_0^t L(\beta_s)\mathrm{d}s &= W_0 - W_t + \gamma\langle\operatorname{diag}(\bar{X}^\top\bar{X}), \int_0^t L(\beta_s)\xi_s\mathrm{d}s\rangle + S_t \\ &\geq W_0 - W_t + S_t, \end{aligned}$$

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 3 (which is still valid since we are on the event $\mathcal{A}$) and $M = \left[325\ln(4/p)\lambda_{\max}\max(\|\beta^*_{\ell_1}\|_1^2\ln^2(\sqrt{2}\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}), \alpha^4 d^2)\right]$. 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:

$$\begin{aligned}
W_0 &= \phi_{\alpha_\infty}(\beta_\infty) - \phi_\alpha(0) \\
&\geq \phi_\alpha(\beta_\infty) - \phi_\alpha(0) \\
&\geq \phi_\alpha(\beta^*_\alpha) - \phi_\alpha(0) := W_0^\alpha,
\end{aligned}$$

where $\beta^*_\alpha = \underset{\beta \text{ s.t } X\beta = Y}{\operatorname{argmin}} \phi(\beta, \alpha^2)$. 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 $\beta^*_{\ell_1}$ 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 8. *On $\mathcal{A}\cap\mathcal{B}$, let $\gamma \leq [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^*_{\ell_1}\|_1\ln(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\alpha^2}), \alpha^2 d\}]^{-1}$, then for α small enough:*

$$\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s \geqslant \frac{1}{8}\|\beta^*_{\ell_1}\|_1\ln\left(\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}\right).$$

Proof. Applying Lemma 11, 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,

$$W_0^\alpha \geq \frac{1}{4}\sum_i |\beta^*_{\alpha,i}|\ln\frac{|\beta^*_{\alpha,i}|}{2\alpha^2}.$$

Note that $\beta^*_\alpha = \underset{\beta \text{ s.t } X\beta = Y}{\operatorname{argmin}} \phi_\alpha(\beta)$ and $\beta^*_{\ell_1} = \underset{\beta \text{ s.t } X\beta = Y}{\operatorname{argmin}} \|\beta\|_1$. From Theorem 2 of [36]: $\|\beta^*_\alpha\|_1 \underset{\alpha\to 0}{\longrightarrow} \|\beta^*_{\ell_1}\|_1$ which leads to:

$$\sum_i |\beta^*_{\alpha,i}|\ln\frac{|\beta^*_{\alpha,i}|}{2\alpha^2} \underset{\alpha\to 0}{\sim} \|\beta^*_{\ell_1}\|_1\ln\frac{\|\beta^*_{\ell_1}\|_1}{\alpha^2}.$$

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 7 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 3. We recall it here to be self-contained.

Proposition 3. *Under the same setting as in Proposition 2 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 9. As this event is included in $\mathcal{A}$, the right inequality of the proof corresponds exactly to the Proposition 4 of Appendix B.2. The proof of left inequality of the proposition comes from Lemma 8. □

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 5. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow* (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. $\gamma = [400\lambda_{\max}\ln(\frac{4}{p})\max\{\|\beta^*_{\ell_1}\|_1 \ln(\frac{\sqrt{2}\|\beta^*_{\ell_1}\|_1}{\alpha^2 d}), \alpha^2 d\}]^{-1}$, 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 $\alpha_\infty = \alpha\exp\Big(-2\gamma\,\mathrm{diag}\Big(\frac{X^\top X}{n}\Big)\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s\Big)$ along with the lower bound from Lemma 8 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.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 3 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 $\mathcal{A}$, $\|\beta_t\|_1 \leq \|\xi_t\|_1 \leq \max\{\|\beta^*_{\ell_1}\|_1, \alpha^2 d\}$ 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 6. *Consider the iterates $(w_t)_{t\geq 0}$ issued from the stochastic gradient flow* (3)*, initialised at $w_{0,\pm} = \alpha\mathbf{1} \in (\mathbb{R}_+^*)^d$. Assume boundedness of the iterates and $\gamma = \Theta\big(\max\{\|\beta^*_{\ell_1}\|_1, \alpha^2 d\}^{-1}\big)$, 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:*

$$\frac{\alpha_\infty}{\alpha} \leq \left(\frac{\alpha^2}{\|\beta^*_{\ell_1}\|_1}\right)^\zeta.$$

Proof. As said earlier, we fix $p = 0.01$. Then, by following the proof of Lemma 4, and using the boundedness assumption instead of Lemma 3, one obtains that for $\gamma \leq O\big(\max\{\|\beta^*_{\ell_1}\|_1, \alpha^2 d\}^{-1}\big)$ (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 5, Proposition 4, Lemma 6 and therefore Theorem 1 then still hold with probability 0.99 but with the weaker condition that $\gamma \leq O\big(\max\{\|\beta^*_{\ell_1}\|_1, \alpha^2 d\}^{-1}\big)$.

For the upperbound on α_∞, we follow the exact same steps as in Appendix B.5. Indeed Lemma 7 now gives, for $\gamma \leq O\big((\lambda_{\max}\max\{\|\beta^*_{\ell_1}\|_1, \alpha^2 d\})^{-1}\big)$:

$$\gamma\int_0^{+\infty} L(\beta_s)\mathrm{d}s \geq \frac{W_0^\alpha}{4}\frac{\gamma}{1+\gamma\frac{M}{W_0^\alpha}},$$

where $M = \Theta\big(\lambda_{\max}\max\{\|\beta^*_{\ell_1}\|_1^2, \alpha^4 d^2\}\big)$. Plugging in the maximum value of γ, i.e. $\gamma = \Theta\big((\lambda_{\max}\max\{\|\beta^*_{\ell_1}\|_1, \alpha^2 d\})^{-1}\big)$: 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:

$$\alpha_\infty = \alpha\exp\left(-2\,\mathrm{diag}\left(\frac{X^\top X}{n}\right)\gamma\int_0^{+\infty} L(\beta_s)\,\mathrm{d}s\right) \leq \alpha\left(\frac{\alpha^2}{\|\beta^*_{\ell_1}\|_1}\right)^{\Omega\left(1\right)}.$$

This concludes the proof of the Proposition. □

This upperbound is significantly better than that of Proposition 5: the smaller the initialisation scale α and the greater the benefit of SGD over GD in terms of implicit bias. More precisely, Proposition 6 shows that the benefit scales as a power law with respect to the initialization α.

C Deterministic framework

In this section we recall some known results concerning the implicit bias of deterministic mirror descent as well as give convergence guarantees. In the previous section, the stochasticity of the flow made the analysis much more involved. In contrast, the analysis is straightforward in the deterministic setting and we believe this simple case can serve as a warmup to gain further intuition. Note that even though these results are known independently, we did not find a clear reference gathering them. See for example [4] for the convergence of the iterates towards an interpolator and [14] for the associated implicit minimisation problem.

Regression setting. Consider a loss function $\ell : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$, such that for any true label $y \in \mathbb{R}$, $\ell(\cdot, y)$ is differentiable, convex and has a unique minimum which we can assume to be y without loss of generality. Now for a training dataset $(x_i, y_i)_{1\leq i\leq n}$, consider the overall convex loss $L(\beta) = \frac{1}{n}\sum_{i=1}^n \ell(\langle\beta, x_i\rangle, y_i)$ and assume that there exists at least one zero-error interpolator β^* such that $L(\beta^*) = 0$, i.e $X\beta^* = y$. With these assumptions, the following property holds.

Proposition 7. *Consider the regression setting stated just above. Let $\Psi : \mathbb{R}^d \to \mathbb{R}$ be a strongly convex and twice differentiable function which we call potential. For any initialisation $\beta_0 \in \mathbb{R}^d$, consider the mirror descent flow $(\beta_t)_t$:*

$$\mathrm{d}\nabla\Psi(\beta_t) = -\nabla L(\beta_t)\mathrm{d}t. \tag{17}$$

Then the iterates $(\beta_t)_t$ converge to an interpolator β_∞ which satisfies:

$$\beta_\infty = \underset{\beta\in\mathbb{R}^d}{\arg\min}\ D_\Psi(\beta,\beta_0) \quad \textit{such that} \quad X\beta = y, \tag{18}$$

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.

Proof. We divide the proof into three steps.

First step: the loss goes to 0.

Note that:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}L(\beta_t) &= -\langle\nabla L(\beta_t), \dot\beta_t\rangle \\ &= -\langle[\nabla^2\Psi(\beta_t)]^{-1}\nabla L(\beta_t), \nabla L(\beta_t)\rangle \\ &\leq 0,\end{aligned}$$

where the inequality is by convexity of the potential Ψ. Hence the loss is decreasing. Now consider the Bregman divergence between an arbitrary interpolator β^* and β_t:

$$D_\Psi(\beta^*,\beta_t) = \Psi(\beta^*) - \Psi(\beta_t) - \langle\nabla\Psi(\beta_t), \beta^* - \beta_t\rangle \geq 0.$$

which is such that:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}D_\Psi(\beta^*,\beta_t) &= \langle\frac{\mathrm{d}}{\mathrm{d}t}\nabla\Psi(\beta_t), \beta_t - \beta^*\rangle \\ &= -\langle\nabla L(\beta_t), \beta_t - \beta^*\rangle \\ &\leq -L(\beta_t) \\ &\leq 0\end{aligned} \tag{19}$$

where the first inequality is by the convexity of the loss. Therefore:

$$\begin{aligned}L(\beta_t) &\leq \frac{1}{t}\int_0^t L(\beta_s)\,\mathrm{d}s \\ &\leq \frac{D_\Psi(\beta^*,\beta_0) - D_\Psi(\beta^*,\beta_t)}{t} \\ &\leq \frac{D_\Psi(\beta^*,\beta_0)}{t} \\ &\underset{t\to+\infty}{\longrightarrow} 0.\end{aligned}$$

Hence the loss converges to 0.

Second step: the iterates converge towards an interpolator β_∞.

Since $\frac{\mathrm{d}}{\mathrm{d}t}D_\Psi(\beta^*,\beta_t) \leq 0$, we have that whatever the interpolator β^*, $D_\Psi(\beta^*,\beta_t)$ is decreasing over the trajectory. Since it is a positive quantity we get that it converges. Moreover Ψ is μ-strongly convex which means that we also have $\|\beta_t - \beta^*\|_2^2 \leq \frac{2}{\mu}D_\Psi(\beta^*,\beta_t)$. The flow $(\beta_t)_t$ is therefore bounded and we can extract a convergent subsequence: let β_∞ be such that $\beta_{t_k} \underset{k\to\infty}{\longrightarrow} \beta_\infty$. Since from the first step $L(\beta_t) \to 0$, we have by unicity of the limit that $L(\beta_{t_k})$ also converges to 0, and we get by continuity of L that β_∞ is an interpolator. This means that (a) $D_\Psi(\beta_\infty,\beta_t)$ converges and (b) it converges towards the same limit as $D_\Psi(\beta_\infty,\beta_{t_k})$ which is 0. Finally:

$$0 \leq \|\beta_t - \beta_\infty\|_2^2 \leq \frac{2}{\mu}D_\Psi(\beta_\infty,\beta_t) \underset{t\to\infty}{\longrightarrow} 0,$$

and therefore β_t converges towards the interpolator β_∞.

Third step: implicit bias.

Note that for all β, $\nabla L(\beta) = \frac{1}{n}\sum_{i=1}^n \ell'(\langle \beta, x_i\rangle, y_i)\, x_i \in \mathrm{span}(X)$. Therefore,

$$\nabla\Psi(\beta_t) - \nabla\Psi(\beta_0) = -\int_0^t \nabla L(\beta_s)\,\mathrm{d}s \in \mathrm{span}(X).$$

Taking the limit, $\nabla\Psi(\beta_\infty) - \nabla\Psi(\beta_0) \in \mathrm{span}(X)$ and $X\beta_\infty = y$, which are exactly the KKT conditions of the following convex minimisation problem:

$$\min_{\beta\in\mathbb{R}^d} D_\Psi(\beta, \beta_0) \quad \text{such that} \quad X\beta = y.$$

Remark on the loss integral. We can also show that the integral of the loss converges. Indeed from inequality 19 with $\beta^* = \beta_\infty$, we immediately get that $\int_0^\infty L(\beta_s)\mathrm{d}s \leq D_\Psi(\beta_\infty, \beta_0)$. Furthermore, when L is the square loss $L(\beta) = \frac{1}{2}(\beta - \beta^*)^\top H(\beta - \beta^*)$, then inequality 19 becomes the equality $\frac{\mathrm{d}}{\mathrm{d}t} D_\Psi(\beta^*, \beta_t) = -2L(\beta_t)$ and hence $\int_0^\infty L(\beta_s)\mathrm{d}s = \frac{1}{2} D_\Psi(\beta_\infty, \beta_0)$.

□

In our framework, for the deterministic case, we cannot simply apply this result with ϕ_α, indeed it is not strongly convex over $\mathbb{R}^d$. However following the exact same proof as in Lemma 3 and Lemma 4 but in the deterministic case (which is easier since we do not need to use martingale concentration inequalities), we can show that the iterates β_t are bounded. Using Proposition 7 but on a convex set in which the iterates stay and over which ϕ_α is strongly convex (as done in Lemma 6) leads to the convergence of the iterates.

D Experiments

In the following section we consider the same experimental setup as in Section 5.1, which we recall here for clarity. We consider $n = 40$, $d = 100$ and randomly generate a sparse model $\beta^*_{\ell_0}$ such that $\|\beta^*_{\ell_0}\|_0 = 5$. We generate the features as $x_i \sim \mathcal{N}(0, I)$ and the labels as $y_i = x_i^\top \beta^*_{\ell_0}$. 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 $\|\beta_t - \beta^*_{\ell_0}\|_2^2$ corresponds to the validation loss.

D.1 Doping the implicit bias using label noise: experiments

We consider the label noise setting discussed in Section 5.4: 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 $\Delta_t \sim \mathrm{unif}\{2\delta_t, -2\delta_t\}$ 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:

$$w_{t+1,\pm} = w_{t,\pm} \mp \gamma\left(\langle \beta_w - \beta^*, x_{i_t}\rangle + \Delta_t\right)\, x_{i_t} \odot w_{t,+}\,, \qquad \text{where } i_t \sim \mathrm{unif}(1, n). \tag{20}$$

Using the same notations and following the same derivations as in Appendix A, we can rewrite the recursion as:

$$w_{t+1,\pm} = w_{t,\pm} - \gamma\nabla_{w_\pm} L(w_t) \pm \gamma\,\mathrm{diag}(w_{t,\pm})X^\top[\xi_{i_t}(\beta_t) + \Delta_t \mathbf{e}_{i_t}].$$

Since Δ_t is zero-mean and independent of i_t we get:

$$\begin{aligned}\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta) + \Delta_t\mathbf{e}_{i_t}] &= \mathbb{E}_{i_t}[\xi_{i_t}(\beta)^{\otimes 2}] + \mathbb{E}[\Delta_t^2 \mathbf{e}_{i_t}^{\otimes 2}] \\ &= \mathbb{E}_{i_t}[\xi_{i_t}(\beta)^{\otimes 2}] + \frac{4\delta_t^2}{n} I_n.\end{aligned}$$

Now following the same reasoning as in Appendix A, 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 1 we obtain that:

$$\beta_t = 2\tilde{\alpha}_t^2 \odot \sinh(2\bar{X}^\top \tilde{\eta}_t),$$

where $\tilde{\eta}_t = -\int_0^t \bar{X}(\beta_s - \beta^*)\,\mathrm{d}s + 2\sqrt{\gamma}\int_0^t \sqrt{\tilde{L}(\beta_s)}\mathrm{d}B_s \in \mathbb{R}^n$ and $\alpha_t = \alpha \odot \exp\big(-2\gamma\,\mathrm{diag}(\bar{X}^\top \bar{X})\int_0^t \tilde{L}(\beta_s)\mathrm{d}s\big)$. And following the proof of Proposition 1:

$$\mathrm{d}\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,\mathrm{d}t + \sqrt{\gamma n^{-1}\tilde{L}(\beta_t)}X^\top \mathrm{dB}_t. \tag{21}$$

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: $\tilde{\alpha}_\infty = \alpha \odot \exp\left(-2\gamma\,\mathrm{diag}(\frac{X^\top X}{n})\int_0^{+\infty}\tilde{L}(\beta_s)\,\mathrm{d}s\right)$ 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 5. 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.

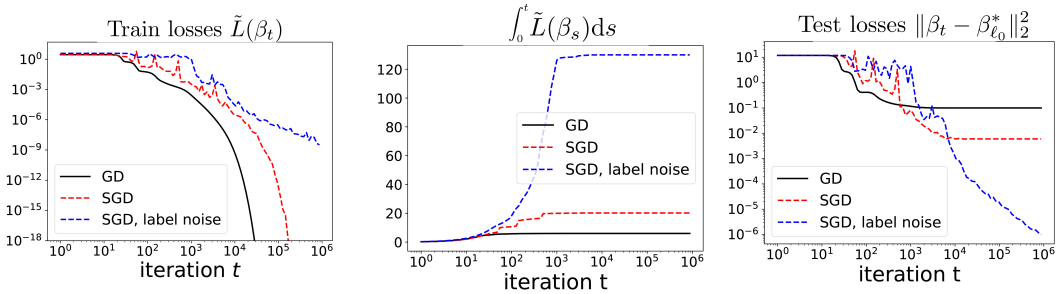

Figure 5: Sparse regression (see Section 5.1 for the detailed experimental setting), illustration of the benefits of using label noise. All experiments are initialised at $\alpha = 0.01$. *Left*: The use of label noise slows down the convergence of the effective training loss $\tilde{L}$. *Middle and 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.

E Extensions

We introduce two extensions of our results: subsection E.1 extends our results for a very general stochastic gradient flow model and subsection E.2 discuss them in the depth $p \geq 3$ case.

E.1 Towards a more general SDE modelling

Recall from the SDE modelling of Appendix A that $\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] = \frac{4}{n}\,\mathrm{diag}(L_i(\beta))_{1\leq i\leq n} + O(\frac{1}{n^2})$. If we assume n large enough we can neglect the second order term of order $1/n^2$:

$$\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] \cong \frac{4}{n}\,\mathrm{diag}(L_i(\beta))_{1\leq i\leq n}.$$

Assume we do not consider that $L_i(\beta) \sim L(\beta)$, then the overall SGD noise structure is captured by

$$\begin{aligned}\Sigma_{\mathrm{SGD}}(w_\pm) &:= \gamma^2\,\mathrm{diag}(w_\pm)X^\top \mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)]X\,\mathrm{diag}(w_\pm)\\ &\cong \frac{4}{n}\gamma^2[\mathrm{diag}(w_\pm)X^\top\,\mathrm{diag}(\sqrt{L_i(\beta)})]^{\otimes 2}.\end{aligned}$$

This leads us in considering the following SDE:

$$\begin{aligned}\mathrm{d}w_{t,+} &= -\nabla_{w_+}L(w_t)\,\mathrm{d}t + 2\sqrt{\gamma}\,w_{t,+}\odot[\bar{X}^\top\,\mathrm{diag}(\sqrt{L_i(\beta)})\mathrm{d}B_t]\\ \mathrm{d}w_{t,-} &= -\nabla_{w_-}L(w_t)\,\mathrm{d}t - 2\sqrt{\gamma}\,w_{t,-}\odot[\bar{X}^\top\,\mathrm{diag}(\sqrt{L_i(\beta)})\mathrm{d}B_t].\end{aligned} \tag{22}$$

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 $\alpha_t = \alpha \odot \exp(-2\gamma \operatorname{diag}(\bar{X}^\top \int_0^t \operatorname{diag}(L_i(w_s))\,\mathrm{d}s \bar{X}))$. 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), 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 paper 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 $\mathbb{E}_{i_t}[L_{i_t}(\beta_s)]\mathbb{E}_{i_t}[(x_{i_t}^{(k)})^2]$

E.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 2.2, stochastic gradient flow writes:

$$\mathrm{d}w_{t,\pm} = -\nabla_{w_\pm} L(w_t)\mathrm{d}t \pm 2\sqrt{\gamma n^{-1} L(\beta_t)}\operatorname{diag}(w_{t,\pm}^{p-1})X^\top \mathrm{d}B_t, \tag{23}$$

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 $\alpha_{\infty,\pm} = \alpha(1 + 2\gamma(p-2)(p-1)\alpha^{p-2}\operatorname{diag}(\frac{X^\top X}{n}) \odot \int_0^\infty L(\beta_s) w_{s,\pm}^{p-2}\mathrm{d}s)^{-\frac{1}{p-2}}$.

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[6] and that the convergence speed of the loss controls the biasing effect. As in [36], 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:

$$\mathrm{d}[w_{t,-}^{2-p}] = X^\top \mathrm{d}A_t + C_t^- \mathrm{d}t.$$

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 $\alpha_{\mathrm{eff},\pm}^{2-p} = \alpha^{2-p} + \int_0^\infty C_s^\pm \mathrm{d}s$ 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 $h_{\alpha,p,\pm}(z) = (\alpha_{\mathrm{eff},+}^{2-p} - z)^{\frac{p}{2-p}} - (\alpha_{\mathrm{eff},-}^{2-p} + z)^{\frac{p}{2-p}}$. Inverting this function and integrating gives the theorem with the standard KKT argument [see 36, under Theorem 1 page 4]. □

F Technical lemmas

In this section, we state and prove technical lemmas which we use to prove our main results.

Lemma 9. *For any interpolator β^*, $S_t = \int_0^t \sqrt{\gamma L(\beta_s)} \langle \bar{X}^\top \mathrm{d}B_s, \beta_s - \beta^* \rangle$ is a square-integrable martingale with a.s. continuous paths. And for any $a, b \geq 0$:*

$$\begin{aligned}
P(\forall t \geq 0, |S_t| \leq a + 2b\gamma\lambda_{\max} \int_0^t L(\beta_s)(\|\beta_s\|_1^2 + \|\beta^*\|_1^2)\mathrm{d}s) &\geq 1 - 2\exp(-2ab) \\
&= 1 - p,
\end{aligned}$$

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, [19, Corollary 11] gives that

$$P(\exists t \in (0,\infty) : S_t \geq a + b\langle S \rangle_t) \leq \exp\{-2ab\}).$$

[6] Note that, as the weights are initialized positively, they remain positive: $w_{t,\pm} > 0$, for all $t \geqslant 0$.

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 10. *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 [7]. 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 [7]). 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. □

[7] see `https://en.wikipedia.org/wiki/Lambert_W_function` for more details

Lemma 11. *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

$$f(\beta) := \phi_\alpha(\beta) - \phi_\alpha(0) = \frac{1}{4}\left[\beta \mathrm{arcsinh}\left(\frac{\beta}{2\alpha^2}\right) - \sqrt{\beta^2 + 4\alpha^4} + 2\alpha^2\right].$$

Trivially, $f'(\beta) = \frac{1}{4}\mathrm{arcsinh}\left(\frac{\beta}{2\alpha^2}\right) > 0$. 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

$$4[f'(\beta) - g'(\beta)] = \mathrm{arcsinh}\left(\frac{\beta}{2\alpha^2}\right) - \ln\left(\frac{\beta}{2\alpha^2}\right) + 1 = \ln\left(1 + \sqrt{1 + \frac{4\alpha^4}{\beta^2}}\right) + 1 > 0.$$

Hence, $f - g$ increases and as $f(0) - g(0) = 0$, we have that $f > g$ which concludes the proof. □