

Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation

Loucas Pillaud-Vivien
EPFL

LOUCAS.PILLAUD-VIVIEN@EPFL.CH

Julien Reygner
École des Ponts

JULIEN.REYGNER@ENPC.FR

Nicolas Flammarion
EPFL

NICOLAS.FLAMMARION@EPFL.CH

Editors: Po-Ling Loh and Maxim Raginsky

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 role of the label noise in the training dynamics of a quadratically parametrised model through its continuous time version. We explicitly characterise the solution chosen by the stochastic flow and prove that it implicitly solves a Lasso program. To fully complete our analysis, we provide nonasymptotic convergence guarantees for the dynamics as well as conditions for support recovery. We also give experimental results which support our theoretical claims. Our findings highlight the fact that structured noise can induce better generalisation and help explain the greater performances of stochastic dynamics as observed in practice.

Keywords: Label noise, Stochastic dynamics, Lasso, Sparse Regression.

1. Introduction

The many successes of deep learning are undoubtedly equal to the theoretical mysteries that surround it. However, while theoretical explanations were quite weak a decade ago, some recent progresses have refined our understanding of neural networks: by proving convergence in some cases (Mei et al., 2018; Chizat and Bach, 2018), or clarifying the role of initialisation (Jacot et al., 2018; Chizat et al., 2019). Still, one of their most surprising and unexplained aspect is their ability to generalise without explicit regularisation despite large overparametrisation (Zhang et al., 2017).

In fact, due to the high expressivity of overparametrised neural networks, they carry a large freedom while fitting a data set, yet without any hurt in the generalisation performance. That is to say that the way they are trained (initialisation, algorithm, specific architecture) specifies a large part of their generalisation abilities. This crucial aspect, often referred to as the *implicit bias* or *algorithmic regularisation*, has been a major line of research lately. For example, in the simple and prototypical least-square framework, it has been shown that both gradient descent and stochastic gradient descent converge towards the global solution which has the lowest squared distance from the initialisation (Zhang et al., 2017). For a linear parametrisation, Soudry et al. (2018) show in a seminal paper that gradient descent selects the max-margin classifier for logistic regression on separable data. Since then, many works have tried to characterise the implicit bias of specified settings: for classification with neural networks (Lyu and Li, 2020; Chizat and Bach, 2020), and

regression for a large variety of nonconvex models (Woodworth et al., 2020; Arora et al., 2019) for which initialisation always plays a central role (Maennel et al., 2018).

A common feature of these analyses is that they study (deterministic) gradient descents (GD), while it has been shown empirically that stochasticity may be of primary importance to match best generalisation guarantees (Keskar et al., 2017). Hence, it is natural to try to understand the role of stochasticity induced by the mini-batch training procedure of stochastic gradient descent (SGD). It is often shown that SGD tends to move towards *flat regions* of the training loss (Zhu et al., 2019; Chaudhari and Soatto, 2018). However, the flat minima selection phenomenon does not appear very clearly and noise models taken to rigorously prove these are disputable. In this perspective, specific noise models to understand the role of stochasticity are primordial: an example of that is the fact that minibatch stochasticity of SGD is state dependent and cancels itself at global optima (Wojtowysch, 2021; Pesme et al., 2021; Ali et al., 2020). Another important feature is that the noise has a specific geometry, e.g. in the least-square model it belongs to the span of the data inputs (Zhang et al., 2017).

To understand this without suffering from the degeneracy at global optima one may resort to *label noise* SGD, where some noise is systematically added to the output at each step of the descent. This injected noise has been shown to be a good surrogate model that exemplifies the geometry and the state dependence of the noise carried by SGD before reaching zero training loss (HaoChen et al., 2021). In this perspective, local implicit bias criteria have been sketched (Damian et al., 2021; Blanc et al., 2020) and notably, a limiting process has been introduced lately to formally explain how the label noise drives the dynamics (Li et al., 2022). However, all these correspond to local, nonexplicit and asymptotic results that might not be as satisfying as those proven in the deterministic case. We study such a label noise procedure in a nonconvex model, and provide an explicit, nonasymptotic description of the dynamics.

The quadratic parametrisations which we consider have become popular lately (Vaškevičius et al., 2019) since, despite their simplicity, 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 overparametrisation and of the initialisation (Woodworth et al., 2020). In this literature, it has been shown that a ℓ_1 -sparsifying regularisation rules the implicit biasing, together with initialisation: indeed, when the initialisation goes infinitely small, GD (Woodworth et al., 2020) and SGD (Pesme et al., 2021) select a sparse interpolator of the data. However, an important drawback is that as initialisation gets very small, optimisation time gets very large. Hence the following question

Does label noise help in recovering a sparse interpolator without infinitely small initialisation?

To tackle this question, we study the label noise (stochastic) gradient descent through its continuous version, namely the stochastic gradient flow (SGF). We stress that 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 contribution and paper organisation

In Section 2, we start by introducing the setup of our problem as well as the continuous stochastic model. Then, in Section 3, we state the main results on the dynamics convergence, deriving precise nonasymptotic statements, both in terms of time and noise. We informally formulate it here:

Theorem [Informal] *For any initialisation, the label noise stochastic gradient flow for quadratic parametrisation implicitly solves a weighted Lasso program. In consequence, under conditions on the design matrix, it recovers exactly the support of the ground-truth sparse estimator of the model.*

To reach this goal, we analyse thoroughly the dynamics, handling precisely its stochastic fluctuations. The proof sketch is depicted in Section 4. We support our results experimentally and validate our model in Section 5.

1.2. Additional related work

A large part of the related work has already been covered in the introduction. Let us complete it here. Our study leverages continuous time stochastic differential equation modelling of discrete time dynamics. We refer to Li et al. (2019) for a technical introduction to these techniques when related to machine learning problems.

Let us also compare the present work with the recent literature on label noise driven GD. Two different points of view are taken in the literature. The aim of the first one pioneered by Blanc et al. (2020) is to show that such stochastic dynamics are biased towards optimising a hidden objective related to the curvature of the loss. However, it seems hard to conclude as their calculations are essentially both local and asymptotic. In the same spirit, one of the most conclusive works related to this approach is certainly the recent work of Li et al. (2022) in which the authors exhibit a proper limiting dynamics upon the manifold of interpolators thanks to a time rescaling. Once again, the results shown are only asymptotic, nonquantitative and difficult to apprehend. In contrast, the aim of the present paper is to characterise quantitatively the convergence without resorting to any limiting argument. Finally, HaoChen et al. (2021) show a similar collapsing effect due to the label noise. However, their analysis relies on an extremely large noise (at least square of the dimension), so that our result on the large noise regime alone can be considered finer.

1.3. Notations

For $d \in \mathbb{N}^*$, \mathbb{R}_+^d is the cone of vectors of \mathbb{R}^d with nonnegative components. For vectors $\theta, \theta' \in \mathbb{R}^d$, $\langle \theta, \theta' \rangle$ denotes the standard scalar product of \mathbb{R}^d and $\|\cdot\|_2$ its associated Euclidean norm. For a matrix $X \in \mathbb{R}^{n \times d}$, $\|X\|$ denotes the operator norm associated with $\|\cdot\|_2$. In case of a square matrix, $H \in \mathbb{R}^{d \times d}$, $\text{diag}(H) \in \mathbb{R}^d$ denotes the vector constituted by its diagonal elements $(H_{kk})_{1 \leq k \leq d}$. Classically, $\|\theta\|_1 = \sum_{k=1}^d |\theta_k|$ stands for the ℓ_1 norm of θ . We denote by $\llbracket 1, d \rrbracket$ the set of integers between 1 and d . For vectors $\theta, \theta' \in \mathbb{R}^d$, $\theta \odot \theta'$ stands for the vector $(\theta_k \theta'_k)_{1 \leq k \leq d}$ and we define the square of a vector w.r.t this dot product: $\theta^2 = \theta \odot \theta$. Similarly $\log \theta$ and $\exp \theta$ (or e^θ) denote respectively the vectors $(\log \theta_k)_{1 \leq k \leq d}$ and $(\exp \theta_k)_{1 \leq k \leq d}$. For $\alpha, \alpha' \in \mathbb{R}$, the notation $\alpha \wedge \alpha'$ denotes the minimum between α and α' . $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian law of mean μ and variance σ^2 . For any subset S of $\llbracket 1, d \rrbracket$ of cardinal $|S|$, and vector $v \in \mathbb{R}^d$, we will denote by $v_S \in \mathbb{R}^{|S|}$, the vector $(v_k)_{k \in S}$. For any $v \in \mathbb{R}^d$, we will often write $v = [v_S, v_{S^c}]$, where $S^c = \llbracket 1, d \rrbracket \setminus S$. For the sake of clarity we also denote sometimes the k -th component of the vector v by $[v]_k$ instead of v_k as traditionally. Finally, $0_{\mathbb{R}^d}$ and $\mathbf{1} \in \mathbb{R}^d$ denote respectively the vectors of zeros and ones.

2. Setup and preliminaries

The aim of the study is to show that the geometry of the noise induced by SGD can bias the dynamics towards sparse data interpolators. To precisely support this claim, we introduce now the model and the algorithms we consider.

2.1. Overparametrised noiseless sparse regression

We consider a linear regression problem with inputs-outputs $(x_i, y_i)_{1 \leq i \leq n}$ in $\mathbb{R}^d \times \mathbb{R}$, and loss function

$$L(\beta) := \frac{1}{4n} \sum_{i=1}^n (\langle \beta, x_i \rangle - y_i)^2 = \frac{1}{4n} \|X\beta - y\|_2^2, \quad (1)$$

where $X \in \mathbb{R}^{n \times d}$ is the data matrix whose rows are the data vectors $(x_i^\top)_{1 \leq i \leq n}$ and $y \in \mathbb{R}^n$ is the vector of outputs $(y_i)_{1 \leq i \leq n}$. We study the overparametrised setting $d \geq n$ and assume that there exists at least one *nonnegative* interpolating parameter which perfectly fits the training set, namely:

(A.1) (Existence of nonnegative interpolators) *The set $\mathcal{J}_+ := \{\beta \in \mathbb{R}_+^d : L(\beta) = 0\}$ is nonempty.*

Since we work in the overparametrised setting, the set \mathcal{J}_+ may in principle be a high-dimensional polyhedron. We now provide assumptions on X and y ensuring the existence and characterisation of particular *sparse* interpolators. The first one is very mild and ensures that the vector $h := \text{diag}(X^\top X) \in \mathbb{R}_+^d$ has only positive coordinates.

(A.2) (No degenerate coordinate) *The data matrix X has no identically 0 column.*

For the following, let S be a nonempty subset of $\llbracket 1, d \rrbracket$ with cardinality s and define $X_S \in \mathbb{R}^{n \times s}$ as the matrix constituted by the coordinates of the inputs $(x_i)_{1 \leq i \leq n}$ solely in S . We use a similar notation for the vector $h_S \in \mathbb{R}_+^s$. Furthermore, the support of any $\beta \in \mathbb{R}_+^d$ is denoted and defined by $\text{supp}(\beta) := \{k \in \llbracket 1, d \rrbracket : \beta_k > 0\}$, and we introduce the set $\mathcal{S}_+ := \{\text{supp}(\beta), \beta \in \mathcal{J}_+\}$.

Lemma 1 (Domination condition) *Under Assumptions (A.1) and (A.2), there is at most one $S \in \mathcal{S}_+$ such that either $S = \emptyset$, or $X_S^\top X_S$ is invertible and the domination condition*

$$h_{S^c} > X_{S^c}^\top X_S (X_S^\top X_S)^{-1} h_S \quad (2)$$

holds, where the latter inequality is understood coordinatewise in \mathbb{R}^{d-s} .

Lemma 1 is proved in Appendix A. Let us briefly discuss some of its consequences. First, when $S \neq \emptyset$, the invertibility condition on $X_S^\top X_S$ implies that $|S| \leq n$, which gives that the interpolator associated with $S \in \mathcal{S}_+$ is at worst n -sparse. To understand the uniqueness property of S stated in the Lemma let us detail the case with one data point in dimension two.

Example 1 (Case $n = 1, d = 2$) *Let us set $x = (x^1, x^2) \in \mathbb{R}^2$ and e.g. $y = 1$. Clearly (A.1) and (A.2) are verified if and only if $x^1, x^2 \neq 0$ and either $x^1 > 0$ or $x^2 > 0$. If either one of the coordinate at least is negative, then the domination condition Eq. (2) is always fulfilled and uniqueness of S is obvious. Now suppose that $x^1, x^2 > 0$: if $S = \{1\}$, then Eq. (2) is equivalent to $x^2 > x^1$ and in the case $S = \{2\}$, Eq. (2) is true if and only if $x^1 > x^2$: hence, one possibility rejects the other one and uniqueness of S holds. Finally, in the particular case where $x^1 = x^2$, Eq. (2) is never satisfied; we exclude this pathological case thanks to the following assumption.*

(A.3) (Existence of the ground-truth) *There exists exactly one set $S^\circ \in \mathcal{S}_+$ which satisfies the conditions of Lemma 1.*

Notice that $S^\circ = \emptyset$ if and only if $y = 0$. Note additionally that under Assumption (A.3), we may define β° as the unique element of \mathcal{J}_+ with support S° and when $S^\circ \neq \emptyset$, we can write explicitly $\beta^\circ = [(\mathbf{X}_{S^\circ}^\top \mathbf{X}_{S^\circ})^{-1} \mathbf{X}_{S^\circ}^\top y, 0_{S^c}]$. We shall keep the notation β° , S° throughout the remainder of this article and call β° the *ground-truth estimator* of the model. In case the output y has been generated by a sparse nonnegative vector β^S with data matrix \mathbf{X} , i.e. $\mathbf{X}\beta^S = y$, and a support S that satisfy the conditions of Lemma 1, then this latter Lemma implies that $\beta^S = \beta^\circ$. The aim of Assumption (A.3) is to ensure the existence of such an interpolator. In the sparse recovery literature, the conditions asked in Lemma 1 are related respectively to the *subinvertibility* and *mutual incoherence* conditions. We refer to page 219 of Wainwright (2019) for further discussions on these.

2.2. Architecture and algorithm

A two-homogeneous reparametrisation. We reparametrise the linear prediction $x \mapsto \langle \beta, x \rangle$ using the nonlinear parametrisation $x \mapsto \langle \theta^2, x \rangle$, where the square of the vector θ stands for the coordinatewise square. This 2-positive homogeneous model can be viewed as a simple linear network with only pairwise connections and is often used as a first step towards understanding more general neural networks (Woodworth et al., 2020; Li et al., 2022). It is worth noting that the parametrisation $\beta = \theta_+^2 - \theta_-^2$ could be considered in order to attain negative values. It would only make the analysis more technical and therefore we prefer to restrict ourselves to the simpler setting. By a slight abuse of notation we rewrite the training loss (1) as a function of θ as

$$L(\theta) := \frac{1}{4n} \sum_{i=1}^n (\langle \theta^2, x_i \rangle - y_i)^2. \quad (3)$$

Even if the overall function space expressivity has not changed, the reparametrisation makes the least-square problem associated with Eq. (3) nonconvex. Thus minimising it with gradient based procedures is not guaranteed to converge to global optima anymore.

Label noise gradient descent. We minimise the training loss in Eq. (3) with GD and *Label Noise* (LNGD). Namely, at each gradient step $t > 0$, we deliberately add a random noise $\xi(t) \sim \mathcal{N}(0, \delta I_n)$ to the label y . The algorithm is started from θ_0 and used with a constant step size $\gamma > 0$. Noting explicitly the loss with input \mathbf{X} and output y as $L(\theta; \mathbf{X}, y)$, the update rule corresponds to

$$\begin{aligned} \theta(t+1) &= \theta(t) - \gamma \nabla_{\theta} L(\theta(t); \mathbf{X}, y + \xi(t)) \\ &= \theta(t) - \frac{\gamma}{n} \left[\mathbf{X}^\top (\mathbf{X}\theta(t)^2 - y) \right] \odot \theta(t) + \frac{\gamma}{n} \left[\mathbf{X}^\top \xi(t) \right] \odot \theta(t), \end{aligned} \quad (4)$$

where \odot stands for the coordinatewise product between two vectors. In the previous literature, LNGD was often studied together with SGD. However the stochasticity coming from the sampling procedure of SGD is rapidly negligible compared to the one triggered by the label noise. Indeed the label noise SGD update writes

$$\begin{aligned} \theta(t+1) &= \theta(t) - \gamma (\langle \theta(t)^2, x_{i(t)} \rangle - y_{i(t)}) x_{i(t)} \odot \theta(t) \\ &= \theta(t) - \frac{\gamma}{n} \left[\mathbf{X}^\top (\mathbf{X}\theta(t)^2 - y) \right] \odot \theta(t) + \frac{\gamma}{n} \left[\mathbf{X}^\top (\xi(t) + \varepsilon_{i(t)}) \right] \odot \theta(t), \end{aligned}$$

where for all $t > 0$, $i(t)$ is sampled from the uniform distribution over $\llbracket 1, n \rrbracket$, and we have defined $\varepsilon_{i(t)} := \mathbb{E}_{i(t)}[(\langle \theta(t)^2, x_{i(t)} \rangle - y_{i(t)})e_{i(t)}] - (\langle \theta(t)^2, x_{i(t)} \rangle - y_{i(t)})e_{i(t)}$ where $(e_i)_{1 \leq i \leq n}$ is the canonical basis of \mathbb{R}^n . With this notation, $\varepsilon_{i(t)}$ corresponds to the SGD noise: it is multiplicative, crucially vanishes at optimum and is rapidly negligible or comparable to $\sqrt{\delta}$. Thus there should be no qualitative difference between SGD and GD when both are used with label noise (see Figure 1 for an empirical validation of this fact). To understand the LNGD dynamics on the nonconvex objective Eq. (3), we resort to its continuous time model. This approach has the advantage of leading to clean calculations while comprehending the complexity of the model.

2.3. Label noise stochastic gradient flow

Continuous time stochastic dynamics modelling. Continuous time modelling of sequential processes provides a large set of tools, such as differential calculus, which are valuable when trying to understand the dynamics of a process. For this reason, many recent works have considered gradient flows with the aim of grasping the behaviour of gradient descent on complex nonconvex problems such as neural networks training. However the modelling of stochastic dynamics is more demanding. Indeed these dynamics are better modelled by stochastic processes which are solutions of stochastic differential equations (SDEs): $d\theta(t) = b(t, \theta(t))dt + \sigma(t, \theta(t))dB(t)$, where $(B(t))_{t \geq 0}$ is a standard Brownian motion. For a proper model, the drift term b and the noise σ need to be set in a particular manner:

- (i) The drift term b should match the negative gradient: $b = -\nabla L$.
- (ii) The noise covariance $\sigma\sigma^\top(t, \theta)$ should match $\text{Cov}[\frac{\gamma}{n} [X^\top \xi] \odot \theta(t) | \theta(t) = \theta]$, where we have set $\xi \sim \mathcal{N}(0, \delta I_n)$ independent of $\theta(t)$.
- (iii) The noise should belong to the correct space, i.e. to the manifold $\{\theta \odot \text{span}[X^\top], \text{ for } \theta \in \mathbb{R}^d\}$.

Stochastic process model. Following these rules, we propose the following SDE to model LNGD in continuous time. This gives the label noise gradient flow (LNGF)

$$d\theta(t) = -\frac{1}{n} [X^\top (X\theta(t)^2 - y)] \odot \theta(t) dt + \frac{\sqrt{\delta\gamma}}{n} \theta(t) \odot [X^\top dB(t)], \quad (5)$$

with $(B(t))_{t \geq 0}$ a standard Brownian motion in \mathbb{R}^n . Since LNGD is the Euler-Maruyama discretisation with step-size γ of the SDE (5), the SDE and the discrete models match perfectly for infinitesimal step-sizes (up to first order terms in γ). The model is said to be *consistent*. We also note that the same SDE is obtained for any label noise distribution ξ with zero-mean and δ times identity covariance. We simply assume that the injected label noise is Gaussian for clearness of exposition. In the same way, we could consider time-dependent covariance $\Sigma(t) \in \mathbb{R}^{n \times n}$ by replacing the noise term in the SDE by $\frac{\sqrt{\gamma}}{n} \theta(t) \odot [X^\top \Sigma(t)^{1/2} dB(t)]$.

Drift-Variance tradeoff. For the sake of clarity, we introduce the following renormalised variables: $X = X/\sqrt{n}$, $y = y/\sqrt{n}$, $h = h/n$, $\delta = \gamma\delta/n$. The SDE model we study reads

$$d\theta(t) = \underbrace{-[X^\top (X\theta(t)^2 - y)] \odot \theta(t)}_{\text{drift term}} dt + \underbrace{\sqrt{\delta} \theta(t) \odot [X^\top dB(t)]}_{\text{noise term}}. \quad (6)$$

As often with state-dependent noise SDEs, there is a competition between the noise and the drift components. On the one hand, considering the drift term alone amounts to only take into account the

gradient flow. As analysed by [Woodworth et al. \(2020\)](#); [Wu and Rebeschini \(2020\)](#), the dynamics is, in that case, driven to a certain interpolator related to the initialisation. On the other hand, the noise term acts as a multiplicative shrinking force akin to the one in the geometric Brownian motion $dS(t) = \mu S(t)dt + \sigma S(t)dW(t)$. For this dynamics, the noise can counter the repulsive force of the drift and drive $S(t)$ to 0 almost surely if its scale satisfies $\sigma^2 > 2\mu$. Thus, if the noise δ dominates the dynamics: $\delta \gg \sup_{t \geq 0} \|X^\top(X\theta(t)^2 - y)\|_\infty$, the process is similarly driven to 0 almost surely. We note that this argument is at the crux of the analysis of [HaoChen et al. \(2021\)](#). However, when the noise level δ is not infinitely large, the drift and the noise balance each other out and the dynamics becomes much more intricate to analyse.

2.4. Hidden mirror flow structure and Lasso

Itô calculus and hidden mirror flow. Let us recall that, for such reparametrised model, when θ follows a gradient flow $d\theta(t) = -\nabla L(\theta(t))dt$, then the corresponding iterate $\beta(t) = \theta(t)^2$ follows a mirror descent with potential defined through $\nabla\psi(\beta) = \log(\beta)$, where the log is taken componentwise ([Ghai et al., 2020](#)). This result is easily obtained using the chain rule on $\nabla\psi(\beta(t))$. This hidden mirror structure is then used to describe the implicit bias of such gradient flows. It turns out that the exact same procedure can be done based on Itô calculus (i.e. chain rule for stochastic processes) to exhibit a stochastic mirror flow. Indeed, setting $\beta(t) = \theta(t)^2$, where $(\theta(t))_{t \geq 0}$ is the solution to (6), we have

$$d\beta(t) = \beta(t) \odot \left[-2X^\top(X\beta(t) - y) + \delta h \right] dt + 2\sqrt{\delta}\beta(t) \odot [X^\top dB(t)], \quad (7)$$

and this entails the following stochastic differential equation for $\log(\beta)$

$$d\log \beta(t) = - \left(2X^\top(X\beta(t) - y) + \delta h \right) dt + 2\sqrt{\delta}X^\top dB(t). \quad (8)$$

This observation has already been used to understand the implicit bias due to the sampling noise of SGD ([Pesme et al., 2021](#)), where the authors resort to the unsigned parametrisation ($\beta = \theta_+^2 - \theta_-^2$) changing the mirror map to the argsh; without affecting qualitatively the following discussion.

Stochastic mirror on the weighted lasso. The equation (8) can be explicitly rewritten as a stochastic *mirror-like* flow

$$d\nabla\psi(\beta(t)) = -\nabla L_\delta(\beta(t))dt + 2\sqrt{\delta}X^\top dB(t), \quad (9)$$

with

$$L_\delta(\beta) := \|X\beta - y\|_2^2 + \delta \langle h, \beta \rangle, \quad \nabla\psi(\beta) = \log \beta. \quad (10)$$

The objective L_δ is importantly related to the celebrated *weighted* Lasso problem:

$$\min_{\beta \geq 0} L_\delta(\beta), \quad (\text{WL}_\delta)$$

where the quadratic loss is regularised by a weighted ℓ_1 -norm with weight h and regularisation parameter δ . Hence, the change of variable together with the mirror interpretation provides the following intuition:

The label noise gradient flow can be cast as a stochastic mirror flow on the weighted Lasso with weight h and regularisation parameter δ .

This picture serves as the main guideline for our study. However, the stochasticity present in the mirror prevents from directly applying mirror-based optimisation techniques. Instead, our approach is based on a direct analysis of the dynamics of $\beta(t)$. Our results are gathered in the next section.

3. Main results

3.1. The recovery problem for the weighted Lasso

As we have seen in the previous section, the stochastic dynamics Eq. (6) is intimately related to a mirror gradient optimisation on the weighted ℓ_1 least-square problem (WL_δ) . Hence, to understand the behaviour of the dynamical problem, it seems natural to understand first the properties of the weighted Lasso. This problem is referred to as the *variable selection consistency of the Lasso* in the literature (see e.g. Section 7.5.1 of [Wainwright \(2019\)](#)). Note that due to the weight h in the ℓ_1 norm and the positivity constraint in (WL_δ) , the characterisation of our problem is not immediately implied by standard theorems. We note that, since the function L_δ is convex, a vector $\beta \in \mathbb{R}_+^d$ is a solution to (WL_δ) if and only if there exists $\mu \in \mathbb{R}_+^d$ which satisfies the Karush–Kuhn–Tucker condition

$$2X^\top(X\beta - y) + \delta h = \mu, \quad \langle \mu, \beta \rangle = 0. \quad (\text{KKT}_\delta)$$

Theorem 2 *Let Assumptions (A.1), (A.2) and (A.3) hold, and let β°, S° be defined thereby.*

1. *If $S^\circ = \emptyset$, then for any $\delta > 0$, the pair $\beta^\text{L} := 0, \mu^\text{L} := \delta h$ satisfies the condition (KKT_δ) .*
2. *Otherwise, set*

$$\begin{aligned} \delta_- &:= \sup\{\delta > 0 : \forall k \in S^\circ, \beta_k^\circ > \delta[(2X_{S^\circ}^\top X_{S^\circ})^{-1}h_{S^\circ}]_k\}, \\ \delta_+ &:= \inf\{\delta > 0 : \forall k \in \llbracket 1, d \rrbracket, \delta h_k > 2[X^\top y]_k\}. \end{aligned}$$

- (i) *Standard noise regime: if $\delta < \delta_-$, then the pair*

$$\begin{aligned} \beta^\text{L} &:= [\beta_{S^\circ}^\circ - \delta(2X_{S^\circ}^\top X_{S^\circ})^{-1}h_{S^\circ}, 0_{S^\circ c}]^\top, \\ \mu^\text{L} &:= [0_{S^\circ}, \delta(h_{S^\circ c} - X_{S^\circ c}^\top X_{S^\circ}(X_{S^\circ}^\top X_{S^\circ})^{-1}h_{S^\circ})]^\top, \end{aligned}$$

satisfies the condition (KKT_δ) .

- (ii) *Large noise regime: if $\delta > \delta_+$, then the pair $\beta^\text{L} := 0, \mu^\text{L} := \delta h - 2X^\top y > 0$ satisfies the condition (KKT_δ) .*

In all cases, β^L is the unique solution to (WL_δ) .

Let us comment this theorem. First, uniqueness for (i) and (ii) is nontrivial because the (WL_δ) problem is not strongly convex on account of the degeneracy caused by the affine space $\beta^\circ + \text{Ker}X$. Second, (i) tells us that support recovery is perfectly achieved in the case of a standard noise level. The inequality $\delta < \delta_-$ should be interpreted as a high enough *signal to noise ratio* that allows the estimator to recover the support of the ground truth β° . To give an order of magnitude, in case of standard i.i.d. Gaussian data inputs, $X_{S^\circ}^\top X_{S^\circ} \sim I_{S^\circ}$, $h_{S^\circ} \sim \mathbb{1}_{S^\circ}$ and $X_{S^\circ c}^\top X_{S^\circ} \beta_S \sim 0_{S^\circ c}$. Hence, $\delta_- \sim 2\beta_{\min}^\circ$ and $\delta_+ \sim 2\beta_{\max}^\circ$ where β_{\min}° and β_{\max}° are respectively the minimum and the maximum value of $\beta_{S^\circ}^\circ$. The signal to noise ratios $\delta/\beta_{\min}^\circ$ and $\delta/\beta_{\max}^\circ$ rule the difference between the two regimes. Third, note that although our two regimes describe most of the cases we are interested in, a band of noise scale is not treated by Theorem 2: this is typically the case when $\delta \in [2\beta_{\min}^\circ, 2\beta_{\max}^\circ]$ in which case the solution could have support outside S° . The proof of this result is classical: we exhibit *ad-hoc* explicit solution to condition (KKT_δ) and show it is unique. This reasoning is referred to as the *Primal–dual witness construction* by [Wainwright \(2019, p.223\)](#) and is derived in Appendix A. Finally, note that the case $S^\circ = \emptyset$ can be seen as a subcase of the large noise regime, as when $y = 0$, we have $\delta_+ = 0$ and the solutions match. The distinction between standard and large noise regime finds an echo in the next section.

3.2. SDE convergence results

Before presenting the main results, we first recall the most important observations stated in the previous sections. First, as drift and noise parts are locally Lipschitz continuous, for any initial condition $\theta(0) \in \mathbb{R}^d$, Eq. (6) has a unique strong solution, which is defined up to some explosion time τ_∞ (Khasminskii, 2012). Throughout the sequel, we shall work with a fixed initial condition $\theta(0)$ such that $\beta(0) = \theta^2(0) > 0$. Our results will then entail that in both regimes introduced in the previous section, $\tau_\infty = +\infty$, almost surely. Remarkably, Eq. (9) of Section 2.4 shows that it can be cast as a Lasso stochastic mirror flow on the linear predictor $\beta = \theta^2$. Then, conditions are given in Theorem 2 under which support recovery is achieved by the minimiser β^L of the weighted Lasso program (WL_δ) . Here, we naturally distinguished between two different regimes: (i) the large noise regime where the best predictor is uniformly zero (Section 3.2.1) and (ii) the standard noise regime, when the signal to noise ratio is high enough to allow for support recovery (Section 3.2.2).

Our main results show that $(\beta(t))_{t \geq 0}$ recovers perfectly the support of β^L in these two regimes, namely that, first, $\beta(t) \rightarrow 0$ on S^c . Second, on the support S° , in the long run, $\beta(t)$ fluctuates in a neighbourhood of size $\sqrt{\delta}$ around β^L and hence of β° (by Theorem 2-(ii)). To quantify the noise that remains inherently in $\beta(t)$, we introduce comparison processes that allow to precisely specify the scale of the fluctuations. In the large noise regime, it is ruled by Brownian fluctuations, whereas in the standard noise regime, it is ruled by a rapidly mixing process which concentrates around β° at $\sqrt{\delta}$ scale.

3.2.1. THE LARGE NOISE REGIME

We first place ourselves in the large noise regime from Theorem 2. We recall that we take the convention that this regime contains the case $S^\circ = \emptyset$ (that is to say $y = 0$), for which we define $\delta_+ = 0$. In this regime, note that β^L is uniformly zero. In the following result, we show that the stochastic gradient flow we consider goes to zero almost surely at exponential speed.

Theorem 3 (Large noise regime convergence) *Let the assumptions of Theorem 2 hold, with $\delta > \delta_+$. Recall that $\mu^L = \delta h - 2X^\top y > 0$. Then $\tau_\infty = +\infty$, almost surely; besides, there exists $C \geq 0$ depending on the data X , y and on $\beta(0)$, and a one-dimensional process $(\zeta^\delta(t))_{t \geq 0}$, which has the same law as $(4\delta \|B(t)\|_2^2)_{t \geq 0}$, such that almost surely,*

$$\forall t \geq 0, \quad \beta(t) \leq C \exp\left(\|X^\top\| \sqrt{\zeta^\delta(t)} \mathbf{1} - \mu^L t\right). \quad (11)$$

In consequence,

$$\lim_{t \rightarrow +\infty} \beta(t) = \beta^L = 0, \quad \text{almost surely.} \quad (12)$$

The theorem shows quantitatively that all the coordinates of our predictor go to zero exponentially fast at some speed governed by δ (through the variable μ^L). Roughly speaking, as the fluctuations of the *Bessel process* $\sqrt{\zeta^\delta(t)}$ are of order \sqrt{t} (Revuz and Yor, 2013, Chapter XI), we see that they are negligible in front of the linear deterministic term $-\mu^L t$. These two terms in the exponential are quite reminiscent of the one dimensional geometric Brownian motion (GBW). Thus, we can see this regime as a multidimensional generalisation of the GBW. Note that the same idea has been expressed in HaoChen et al. (2021) but with an unnecessarily large noise $\delta \geq \mathcal{O}(\beta(0)d^2)$. To conclude on the almost sure convergence, a precise analysis is conducted in the proof using the law of the iterated logarithm (Revuz and Yor, 2013, Corollary 1.12, Chapter II). The proof as well as the explicit value of C can be found in Appendix B.1.

3.2.2. THE STANDARD NOISE REGIME

In this regime, recall that by Theorem 2, β^\perp has the same support as β° , denoted by $S^\circ \subset \llbracket 1, d \rrbracket$. For technical reasons we need to assume that the initial condition $\beta(0)$ satisfies the condition that $\beta_{S^\circ}(0) \neq \beta_{S^\circ}^\perp$, and moreover that $n \geq 2$. In the following result, we prove that the stochastic gradient flow we consider goes to zero almost surely at exponential speed on $S^{\circ c}$. We then show precisely that, on the support S° , the stochastic flow fluctuates in a region of size $\sqrt{\delta}$ around β^\perp . To quantitatively support this claim we introduce a comparison process $(\xi^\delta(t))_{t \geq 0}$ which quantifies the scale of the fluctuations.

Theorem 4 (Standard noise regime convergence) *Let the assumptions of Theorem 2 hold, with $S^\circ \neq \emptyset$ and $\delta < \delta_-$. Recall that $\mu_{S^{\circ c}}^\perp > 0$. Then $\tau_\infty = +\infty$, almost surely; besides, there exists a one dimensional stochastic process $(\xi^\delta(t))_{t \geq 0}$, such that the following assertions hold.*

(i) *On the support S° , almost surely,*

$$\forall t \geq 0, \quad \beta_{S^\circ}^\perp e^{-\|X_{S^\circ}^\top\| \sqrt{\xi^\delta(t)}} \leq \beta_{S^\circ}(t) \leq \beta_{S^\circ}^\perp e^{\|X_{S^\circ}^\top\| \sqrt{\xi^\delta(t)}}. \quad (13)$$

Furthermore, $\xi^\delta(t)$ converges in distribution towards a random variable ξ_∞^δ which satisfies the concentration property

$$\forall u > 0, \quad \mathbb{P}\left(\xi_\infty^\delta \geq u\delta\right) \leq 6 \exp\left(-\sqrt{\frac{u\delta}{\kappa^\delta}}\right), \quad (14)$$

for some $\kappa^\delta > 0$ such that $\kappa^\delta = \mathcal{O}(\delta)$ when δ is infinitesimally small.

(ii) *Outside the support S° , there exists $C \geq 0$, depending on X , y and $\beta(0)$, such that, almost surely,*

$$\forall t \geq 0, \quad \beta_{S^{\circ c}}(t) \leq C \exp\left(\|X_{S^{\circ c}}^\top\| \sqrt{\xi^\delta(t)} \mathbb{1}_{S^{\circ c}} - \mu_{S^{\circ c}}^\perp t\right). \quad (15)$$

This implies that

$$\lim_{t \rightarrow +\infty} \beta_{S^{\circ c}}(t) = 0, \quad \text{almost surely.}$$

We make the following remarks on the theorem. In the long term limit, for δ small, Equation (14) and the fact that $\kappa^\delta = \mathcal{O}(\delta)$ show that the random variable $\xi^\delta(t)$ is of order $\mathcal{O}(\delta)$. This concentration property follows from the fact that ξ_∞^δ can be expressed as the squared norm of a n -dimensional random variable, the law of which satisfies a Poincaré inequality with a constant which can be estimated thanks to the Laplace method. Therefore, on the support of β° , by Equation (13) we have, in the long term limit,

$$\beta_{S^\circ}(t) \simeq \beta_{S^\circ}^\perp + \mathcal{O}\left(\sqrt{\delta}\right) = \beta_{S^\circ}^\circ + \mathcal{O}\left(\sqrt{\delta}\right),$$

since Theorem 2 indicates that $\beta_{S^\circ}^\perp - \beta_{S^\circ}^\circ$ is of order δ .

On the other hand, from Equation (15) we see that the story outside the support is simpler: as in the large noise regime, the iterates go exponentially fast to 0 at rate $\mu_{S^{\circ c}}^\perp > 0$. Note however that the scale of the fluctuations is much smaller than in the large noise regime: it is given by the ergodic process $(\xi^\delta(t))_{t \geq 0}$, and no longer by the Bessel process $(\zeta^\delta(t))_{t \geq 0}$. The reason is that in the standard noise regime, there is an *effective* strong convexity effect that prevents the noise from truly fluctuating far from some deterministic trajectory.

Precise derivations on $(\xi^\delta(t))_{t \geq 0}$ and explicit constants can be found in Appendix B.2.

4. Dynamics description and proof sketch

4.1. Going beyond the mirror shape: a dual process

The time dependent mirror. Previous analysis (Gunasekar et al., 2018; Woodworth et al., 2020; Pesme et al., 2021) on the implicit bias of such models rely on the crucial observation that the dual iterates $\nabla\psi(\beta(t))$ belong to the linear span of the observations $\text{span}(X^\top)$. However, in our case, the additional drift term does not belong to such a space: i.e. $\delta h \notin \text{span}(X^\top)$. This is why we introduce a time dependent mirror map $\nabla_\beta\psi(t, \beta) = \log(\beta e^{\delta h t})$. By Eq. (8), the process $(\beta(t))_{0 \leq t < \tau_\infty}$ then satisfies

$$d\nabla_\beta\psi(t, \beta(t)) = -2X^\top(X\beta(t) - y)dt + 2\sqrt{\delta}X^\top dB(t),$$

which is a stochastic mirror-like descent with a geometry that depends on time. The vector $\nabla_\beta\psi(t, \beta(t))$ can then be decomposed in a unique way along $\text{span}(X^\top)$ and $(\text{span}(X^\top))^\perp = \text{Ker}X$. More precisely, there exist $u_* \in \text{Ker}X$ and $v(0) \in \mathbb{R}^n$ such that $\log \beta(0) = u_* + X^\top v(0)$. We next consider the unique strong solution to the SDE

$$dv(t) = -2 \left(X \exp(X^\top v(t) + u_* - \delta h t) - y \right) dt + 2\sqrt{\delta} dB(t),$$

and notice that $\exp(u_* + X^\top v(t) - \delta h t)$ then satisfies the SDE (7). Therefore, by pathwise uniqueness, $\beta(t) = \exp(u_* + X^\top v(t) - \delta h t)$ and both processes β and v share the same explosion time τ_∞ .

Defining a dual process. We next remove the constant term $2y$ in the drift of $v(t)$ by defining $\tilde{v}(t) := v(t) - 2yt \in \mathbb{R}^n$. Denoting by $F(t, \tilde{v}) := \|\exp(X^\top \tilde{v} + u_* - (\delta h - 2X^\top y)t)\|_1$, we remark that $(\tilde{v}(t))_{0 \leq t < \tau_\infty}$ satisfies the SDE

$$d\tilde{v}(t) = -2\nabla_v F(t, \tilde{v}(t))dt + 2\sqrt{\delta}dB(t). \quad (16)$$

We define as well the associated gradient flow $(\tilde{w}(t))_{0 \leq t < \tilde{\tau}_\infty}$:

$$d\tilde{w}(t) = -2\nabla_v F(t, \tilde{w}(t))dt, \quad (17)$$

initialised similarly, that is $\tilde{w}(0) = \tilde{v}(0) = v(0)$, and defined up to some (deterministic) explosion time $\tilde{\tau}_\infty$. For $t \geq 0$ we also introduce, with a slight abuse of notation, the map $\beta(t, \cdot)$ defined from \mathbb{R}^n to \mathbb{R}^d as $\beta(t, \tilde{v}) := \exp(X^\top \tilde{v} + u_* - (\delta h - 2X^\top y)t)$. This map represents the unique way to go from the process $\tilde{v}(t)$, solution of the SDE (16), to the associated process of linear predictors $\beta(t)$ that follows the SDE (7). Thus, we refer to $\tilde{v}(t)$ as *the dual process* associated to $\beta(t)$. It is also worth noting that $F(t, \tilde{v}) = \|\beta(t, \tilde{v})\|_1$ is a convex function of the second variable. This equality emphasises the sparse promoting effect of the dynamics on the linear predictor. Finally, an important quantity appearing in the exponential is the vector

$$c := \delta h - 2X^\top y.$$

In the two next sections, we show that, depending on its sign, the behaviour of Eq. (16) and (17) is very different.

4.2. Large noise regime and the lack of strong convexity

In the setting of Theorem 3, the vector c coincides with μ^\perp defined in Theorem 2, and thus satisfies $c > 0$. Hence the noise dominates the dynamics and entails that all coordinates of $\beta(t, \tilde{v}(t))$ are shrunk by a $e^{-\mu^\perp t}$ factor. Here, we proceed in two steps:

- (i) First, we show that the gradient flow (17) on the dynamics of the process $(\tilde{w}(t))_{0 \leq t < \tilde{\tau}_\infty}$ is bounded. Therefore, $\tilde{\tau}_\infty = +\infty$ and the associated linear estimator $\beta(t, \tilde{w}(t))$ goes to 0 at rate $e^{-\mu^\perp t}$.
- (ii) Then, we show that the stochastic process (16) on $\tilde{v}(t)$ fluctuates from the gradient flow $\tilde{w}(t)$ at a distance controlled by the norm of an n -dimensional Brownian motion, i.e. a Bessel process. Hence, as Brownian fluctuations are of order \sqrt{t} , these are negligible in front of the $-\mu^\perp t$ gradient flow decay rate, which yields in particular $\tau_\infty = +\infty$.

We put emphasis on the fact that this regime suffers from a lack of strong convexity. This is reminiscent of optimizing the Lasso outside of the support of the sparse ground-truth β° (see the discussion section 2.2 of Agarwal et al. (2012) on *Restricted strong convexity and smoothness*).

4.3. Standard noise regime and strong convexity

Failing of the previous reasoning. In the setting of Theorem 4, it is expected that the dual gradient flow (17) drives the primal iterate $\beta(t, \tilde{w}(t))$ towards the solution of the weighted Lasso β^\perp . The strategy of comparison with the gradient flow, outlined in the previous section, only works for the coordinates of $\beta(t, \tilde{v}(t))$ outside of the support S° of β^\perp . For these coordinates, the fluctuations are still dominated by the linear exponential shrinking. Unfortunately, when applying the same technique for coordinates in the support S° , the fluctuations totally blur the sparse recovery problem. To overcome this obstacle, we leverage the strong convexity of the function $\tilde{v} \mapsto F(t, \tilde{v})$ on the support S° , as explained in the following paragraph.

A linear classification problem structure. The reasoning rests on the following observation: the dual gradient flow (17) is similar to the one that solves a linear classification problem with the exponential loss *on the data points given by the coordinates of the $(x_i)_{1 \leq i \leq n}$* . More precisely, let us define the d data $(z_k)_{1 \leq k \leq d} \in (\mathbb{R}^n)^d$ such that $Z = X^\top$, i.e for all $k \in \llbracket 1, d \rrbracket$, $i \in \llbracket 1, n \rrbracket$ $[z_k]_i = [x_i]_k$. Then (17) is equivalent to

$$d\tilde{w}(t) = -2Z^\top \exp(Z\tilde{w}(t) + u_* - ct)dt = -2 \sum_{k=1}^d z_k \exp(\langle z_k, \tilde{w}(t) \rangle + [u_*]_k - c_k t) dt,$$

and the same goes for (16). Note that this observation is reminiscent of the primal-dual analysis presented in Ji and Telgarsky (2021). We leverage this equivalence and show that the selection of the support vectors (i.e, the samples with the smallest margin) of separable linear classification problem (Soudry et al., 2018) enables to recover the true support S° . In this aim, we show that the only terms that matter in the exponential sum are those in the set S' of indices k for which $\langle z_k, \tilde{w}(t) \rangle \sim c_k t$ when $t \rightarrow +\infty$, since the other terms go to zero exponentially fast. Hence, $\beta(t, \tilde{w}(t))$ is asymptotically supported by the coordinates of S' . As in the separable classification problem (Soudry et al., 2018), it turns out that we can totally identify S' as in fact $S' = S^\circ$. Finally, we show that the fluctuations due to the stochastic remaining terms are controlled by the strong convexity upon the support and are of order $\sqrt{\delta}$.

5. Experiments

We consider the following sparse regression setup for our experiments. We choose $n = 40$, $d = 100$ and randomly generate a sparse model β^S such that the cardinality of its support is $s = 4$. We generate Gaussian features as $x_i \sim \mathcal{N}(0, I_d)$, labels as $y_i = \langle \beta^S, x_i \rangle$ and check that this model satisfies the assumptions required in Theorem 2. We consider four different algorithms: the two first are the one we model i.e, Gradient descent + label noise and Stochastic gradient descent + label noise, and for comparison we also considered the Gradient descent and the Stochastic gradient descent. They are initialised at the same point and run with the same step size $\gamma = 0.1$ and noise $\delta = 10^{-3}$.

The experiment presented in Figure 1 perfectly illustrates Theorem 3 (right plot) and Theorem 4 (center plot). Overall, sparse recovery is achieved up to scale $\sqrt{\delta} = \sqrt{\delta\gamma/n} \sim 1.6 \cdot 10^{-3}$. Note that there is no qualitative difference between SGD + label noise (shadowed orange) and GD + label noise (orange). This validates our SDE model.

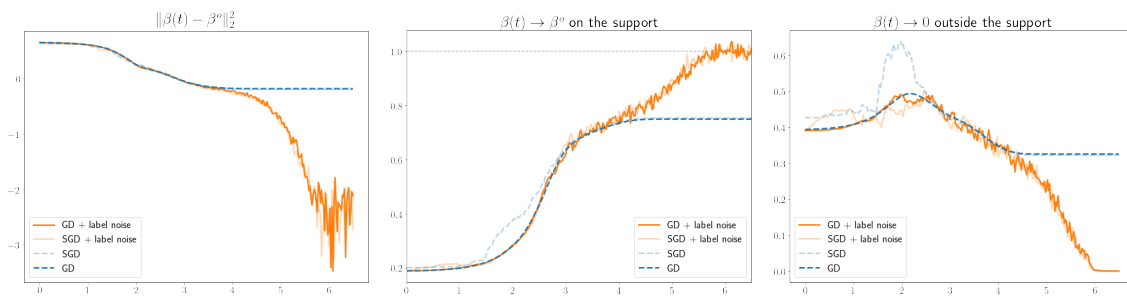


Figure 1: Sparse regression on synthetic data. *Left:* Square error on the parameters. GD and SGD do not recover the sparse truth whereas GD + label noise and SGD + label noise achieve recovery up to scale $\sqrt{\delta}$. *Center:* Recovery of the support in case of noisy label dynamics. A unique and prototypical coordinate is displayed here. *Right:* Convergence to 0 outside the support in case of noisy label dynamics for a prototypical coordinate.

6. Conclusion

In this paper, we have shown that, for quadratically parametrised predictors, the label noise gradient descent solves implicitly a weighted Lasso optimisation program. Hence, this stochastic descent is able to perfectly recover the support of the sparse ground-truth when the injected noise is not too large. In contrast with previous works, we derived precise nonasymptotic results, both in terms of time and noise. Surprisingly, the heart of the proof is based on an equivalence between the selection of support vectors for a classification problem and the one of the nonzero coordinates of the sparse ground-truth. Whether this is only a technical equivalence or a deep relationship could be of great interest. Also, we characterise the equivalence between our label noise dynamics and an optimisation problem enforcing sparsity. Whether it can be done for other architecture remains an interesting open question.

Acknowledgments

N.F. dedicates this work to his late dear friend O.E.. The work of J.R. is supported by the projects ANR EFI (ANR-17-CE40-0030) and ANR QuAMProcs (ANR-19-CE40-0010) from the French National Research Agency.

References

- Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. The Annals of Statistics, 40(5):2452 – 2482, 2012. doi: 10.1214/12-AOS1032.
- Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularisation of stochastic gradient flow for least squares. In International conference on machine learning, pages 233–244. PMLR, 2020.
- Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularisation in deep matrix factorization. Advances in Neural Information Processing Systems, 32, 2019.
- Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. Analysis and geometry of Markov diffusion operators, volume 103. Springer, 2014.
- Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularisation for deep neural networks driven by an ornstein-uhlenbeck like process. In Conference on learning theory, pages 483–513. PMLR, 2020.
- Sergey G Bobkov. Spectral gap and concentration for some spherically symmetric probability measures. In Geometric aspects of functional analysis, pages 37–43. Springer, 2003.
- 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.
- Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf>.
- Lenaïc 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.
- Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf>.
- Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. Advances in Neural Information Processing Systems, 34, 2021.
- Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In Algorithmic Learning Theory, pages 386–407. PMLR, 2020.

- Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 1832–1841. PMLR, 10–15 Jul 2018. URL <http://proceedings.mlr.press/v80/gunasekar18a.html>.
- Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In Conference on Learning Theory, pages 2315–2357. PMLR, 2021.
- Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf>.
- Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In Algorithmic Learning Theory, pages 772–804. PMLR, 2021.
- Ioannis Karatzas and Steven Shreve. Brownian motion and stochastic calculus, volume 113. Springer Science & Business Media, 2012.
- 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.
- Rafail Khasminskii. Stability of stochastic differential equations. In Stochastic Stability of Differential Equations, pages 145–176. Springer, 2012.
- 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. URL <http://jmlr.org/papers/v20/17-526.html>.
- Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss? –a mathematical framework. 2022.
- Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In International Conference on Learning Representations, 2020. URL <https://openreview.net/forum?id=SJeLIgBKPS>.
- Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. arXiv preprint arXiv:1803.08367, 2018.
- 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.
- Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. Advances in Neural Information Processing Systems, 34, 2021.

- Daniel Revuz and Marc Yor. Continuous martingales and Brownian motion, volume 293. Springer Science & Business Media, 2013.
- 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.
- Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularisation for optimal sparse recovery. arXiv preprint arXiv:1909.05122, 2019.
- Alexander Ju Veretennikov. On strong solutions and explicit formulas for solutions of stochastic integral equations. Mathematics of the USSR-Sbornik, 39(3):387, 1981.
- Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.
- Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. Part I: Discrete time analysis, 2021.
- Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrised models. In Conference on Learning Theory, pages 3635–3673. PMLR, 2020.
- Fan Wu and Patrick Rebeschini. A continuous-time mirror descent approach to sparse phase retrieval. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 20192–20203, 2020.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalisation. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. URL <https://openreview.net/forum?id=Sy8gdB9xx>.
- Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 7654–7663. PMLR, 09–15 Jun 2019. URL <http://proceedings.mlr.press/v97/zhu19e.html>.

Appendix

Appendix A. Proof of Lemma 1 and Theorem 2

We first recall the notation \mathcal{S}_+ introduced after Assumption (A.1), and set

$$\begin{aligned} \mathcal{S}'_+ &:= \left\{ S \in \mathcal{S}_+ : S \neq \emptyset, X_S^\top X \text{ is invertible, and } h_{S^c} > X_{S^c}^\top X_S (X_S^\top X_S)^{-1} h_S \right\} \\ &= \left\{ S \in \mathcal{S}_+ : S \neq \emptyset, X_S^\top X \text{ is invertible, and } h_{S^c} > X_{S^c}^\top X_S (X_S^\top X_S)^{-1} h_S \right\}. \end{aligned}$$

By construction, for any $S \in \mathcal{S}'_+$, there is a unique $\beta^{\circ, S} \in \mathbb{R}_+^d$ such that $X\beta^{\circ, S} = y$. Besides, the statement of Lemma 1 rewrites as: either $\emptyset \in \mathcal{S}_+$ and \mathcal{S}'_+ is empty, or $\emptyset \notin \mathcal{S}_+$ and \mathcal{S}'_+ contains at most one element. Therefore, the proof of Lemma 1 stems from the following result.

Lemma 5 (Weighted Lasso in the standard noise regime) *Let Assumptions (A.1) and (A.2) hold.*

1. *If $\emptyset \in \mathcal{S}_+$, then $\beta = 0$ is the unique solution to (WL_δ) , for any $\delta > 0$.*
2. *Assume that the set \mathcal{S}'_+ is nonempty. Define*

$$\delta_- := \sup\{\delta > 0 : \forall S \in \mathcal{S}'_+, \forall k \in S, \beta_k^{\circ, S} > \delta[(2X_S^\top X_S)^{-1} h_S]_k\} > 0,$$

and assume that $\delta < \delta_-$.

- (a) *For any $S \in \mathcal{S}'_+$, the pair*

$$\beta^S = [\beta^{\circ, S} - \delta(2X_S^\top X_S)^{-1} h_S, 0_{S^c}]^\top, \quad \mu^S = [0_S, \delta(h_{S^c} - X_{S^c}^\top X_S (X_S^\top X_S)^{-1} h_S)]^\top,$$

satisfies the condition (KKT_δ) ; in particular, β^S is a solution to (WL_δ) with support S .

- (b) *All solutions to (WL_δ) coincide.*

Proof We first assume that $\emptyset \in \mathcal{S}_+$. Then $y = 0$ and, for any $\delta > 0$, $L_\delta(\beta) = \|X\beta\|_2^2 + \delta\langle h, \beta \rangle$. It is obvious that $\beta = 0$ is a global minimiser of L_δ , and by Assumption (A.2), it is the only one.

We now assume that the set \mathcal{S}'_+ is nonempty, fix $\delta \in (0, \delta_-)$, let $S \in \mathcal{S}'_+$ and define (β^S, μ^S) accordingly. It is simple linear algebra to show that this pair satisfies the condition (KKT_δ) ; besides, it follows from the condition that $\delta \in (0, \delta_-)$ and the definition of \mathcal{S}'_+ that $\text{supp}(\beta^S) = S$ and $\mu_{S^c}^S > 0$. We therefore deduce that β^S solves (WL_δ) .

It remains to prove that all solutions to (WL_δ) coincide. To proceed, we let $\beta' \in \mathbb{R}_+^d$ be another minimiser of L_δ . Using the condition (KKT_δ) , which yields $\nabla L_0(\beta^S) + \delta h = \mu^S$ and $\langle \mu^S, \beta^S \rangle = 0$, we get

$$\begin{aligned} L_0(\beta') - L_0(\beta^S) &= L_\delta(\beta') - \delta\langle h, \beta' \rangle - L_\delta(\beta^S) + \delta\langle h, \beta^S \rangle \\ &= \delta\langle h, \beta^S - \beta' \rangle \\ &= \langle \mu^S - \nabla L_0(\beta^S), \beta^S - \beta' \rangle \\ &= -\langle \mu^S, \beta' \rangle + \langle \nabla L_0(\beta^S), \beta' - \beta^S \rangle. \end{aligned}$$

Since L_0 is convex, $L_0(\beta') - L_0(\beta^S) \geq \langle \nabla L_0(\beta^S), \beta' - \beta^S \rangle$ and therefore $\langle \mu^S, \beta' \rangle \leq 0$. Since both μ^S and β^S are nonnegative, $\langle \mu^S, \beta' \rangle = 0$, and since $\mu_{S^c}^S > 0$, we deduce that $S' := \text{supp}(\beta') \subset S$.

Finally, we let $\mu' \in \mathbb{R}_+^d$ be such that (β', μ') satisfies the condition (KKT_δ) . On the one hand, the same convexity argument for L_0 as above yields $\langle \mu', \beta^S \rangle = 0$, which implies that $\mu'_S = 0$. On the other hand, we deduce from the condition (KKT_δ) for both (β^S, μ^S) and (β', μ') that

$$2X^\top X(\beta^S - \beta') = \mu^S - \mu',$$

which entails $\beta'_S - \beta^S_S = -(2X_S^\top X_S)^{-1}(\mu^S_S - \mu'^S_S) = 0$ and completes the proof. \blacksquare

Lemma 5 implies in particular that S'_+ contains at most one element, and that if $\emptyset \in S_+$ then S'_+ is empty. This proves Lemma 1. We now let Assumption (A.3) hold, which defines β°, S° , and makes the statement of Theorem 2 in the case where $S^\circ = \emptyset$ a straightforward consequence of the first part of Lemma 5. Likewise, in the case where $S^\circ \neq \emptyset$, the statement of Theorem 2 in the standard noise regime immediately follows from Lemma 5. To complete the proof of the statement in the large noise regime, we resort to the following lemma.

Lemma 6 (Weighted Lasso in the large noise regime) *Under the assumptions of Theorem 2, let $S^\circ \neq \emptyset$ and assume that $\delta > \delta_+$. Then $\beta^L := 0, \mu^L := \delta h - 2X^\top y > 0$ is the unique pair of nonnegative vectors which satisfies the condition (KKT_δ) .*

Proof It is clear that (β^L, μ^L) satisfies the condition (KKT_δ) . For uniqueness, let (β', μ') be another solution. Then by the same convexity argument as in the proof of Lemma 5, we get $\langle \mu^L, \beta' \rangle = 0$, which yields $\beta' = 0$ and then $\mu' = \mu^L$. \blacksquare

Appendix B. Proof of Theorems 3 and 4

Recall the following facts that are developed in Section 4.1 of the main text. For $t \in [0, \tau_\infty)$, we introduced a dual variable $\tilde{v}(t) \in \mathbb{R}^n$ associated to $\beta(t)$ such that $\beta(t) = \beta(t, \tilde{v}(t)) = \exp(X^\top \tilde{v}(t) + u_* - ct) \in \mathbb{R}^d$, with $u_* \in \text{Ker} X$ and $c = \delta h - 2X^\top y \in \mathbb{R}^d$. Denoting $F(t, \tilde{v}) := \|\beta(t, \tilde{v})\|_1$, we see that $(\tilde{v}(t))_{0 \leq t < \tau_\infty}$ follows a nonautonomous overdamped Langevin dynamics with respect to the function $\tilde{v} \mapsto F(t, \tilde{v})$ (16), and we also introduce the related gradient flow $(\tilde{w}(t))_{0 \leq t < \tilde{\tau}_\infty}$ (17), which we recall here:

$$d\tilde{v}(t) = -2\nabla_v F(t, \tilde{v}(t))dt + 2\sqrt{\delta}dB_t, \quad d\tilde{w}(t) = -2\nabla_v F(t, \tilde{w}(t))dt, \quad \tilde{w}(0) = \tilde{v}(0).$$

The following statement, the verification of which is straightforward, shall play an important role in the study of (16) and (17).

Lemma 7 *For any $t \geq 0$, the function $\tilde{v} \mapsto F(t, \tilde{v})$ is convex on \mathbb{R}^n .*

B.1. The large noise regime: Theorem 3

Here we place ourselves in the setting of Theorem 3. Then the vector c coincides with the vector $\mu^L > 0$ from Theorem 2. Let us denote $\underline{c} > 0$ its smallest entry. As depicted in Section 4.2, we divide the proof in two steps.

- (i) First we show that the gradient flow $(\tilde{w}(t))_{0 \leq t < \tilde{\tau}_\infty}$ is bounded, which implies that $\tilde{\tau}_\infty = +\infty$. This is presented in Proposition 8 of Section B.1.1.
- (ii) Then we show that $(\tilde{v}(t))_{0 \leq t < \tau_\infty}$ remains in its vicinity with Brownian fluctuations of order \sqrt{t} , which implies that $\tau_\infty = +\infty$. This is presented in Proposition 10 of Section B.1.2.

We finally prove Theorem 3 in Section B.1.3.

B.1.1. BOUNDING THE GRADIENT FLOW

Proposition 8 *Let $(\tilde{w}(t))_{0 \leq t < \tilde{\tau}_\infty}$ be the solution of the gradient flow (17). For all $t \in [0, \tilde{\tau}_\infty)$,*

$$\forall w \in \mathbb{R}^n, \quad \|\tilde{w}(t) - w\|_2 \leq \|w(0) - w\|_2 + 2\|e^{X^\top w + u_*}\|_2 \frac{\|X\|}{\underline{c}}.$$

In particular, $\tilde{\tau}_\infty = +\infty$.

Proof Thanks to Lemma 7, we have, for all $w \in \mathbb{R}^n$, for all $t \in [0, \tilde{\tau}_\infty)$,

$$\begin{aligned} \frac{d}{dt} \frac{1}{2} \|\tilde{w}(t) - w\|_2^2 &= \left\langle \frac{d}{dt} \tilde{w}(t), \tilde{w}(t) - w \right\rangle \\ &= -2 \langle \nabla_v F(t, \tilde{w}(t)), \tilde{w}(t) - w \rangle \\ &= -2 \langle \nabla_v F(t, \tilde{w}(t)) - \nabla_v F(t, w), \tilde{w}(t) - w \rangle - 2 \langle \nabla_v F(t, w), \tilde{w}(t) - w \rangle \\ &\leq -2 \langle \nabla_v F(t, w), \tilde{w}(t) - w \rangle \\ &\leq 2 \|\nabla_v F(t, w)\|_2 \|\tilde{w}(t) - w\|_2 \\ &= 2 \|X e^{X^\top w + u_* - ct}\|_2 \|\tilde{w}(t) - w\|_2 \\ &\leq C e^{-ct} \|\tilde{w}(t) - w\|_2, \end{aligned}$$

with $C = 2 \|X\| \|e^{X^\top w + u_*}\|_2$. Then, denote $f(t) := \|\tilde{w}(t) - w\|_2^2$, we have,

$$\frac{1}{2} f'(t) \leq C e^{-ct} \sqrt{f(t)},$$

so that, for any $\varepsilon > 0$, $g^\varepsilon(t) := \sqrt{f(t)} + \varepsilon$ satisfies

$$[g^\varepsilon(t)]' = \frac{f'(t)}{2\sqrt{f(t)} + \varepsilon} \leq C e^{-ct} \frac{\sqrt{f(t)}}{\sqrt{f(t)} + \varepsilon} \leq C e^{-ct},$$

and by direct integration,

$$g^\varepsilon(t) \leq g^\varepsilon(0) + C \int_0^t e^{-cs} ds = g^\varepsilon(0) + \frac{C}{\underline{c}} (1 - e^{-ct}) \leq \sqrt{f(0)} + \varepsilon + \frac{C}{\underline{c}}.$$

By letting $\varepsilon \rightarrow 0$, we have $\sqrt{f(t)} \leq \sqrt{f(0)} + \frac{C}{\underline{c}}$, which proves the proposition. \blacksquare

B.1.2. CONTROL OF THE STOCHASTICITY: COMPARISON BETWEEN GRADIENT AND STOCHASTIC GRADIENT FLOW

The aim of this section is to bound the difference between the SDE solution of Eq. (16) and the gradient flow solution of Eq. (17). By Itô calculus, we have for $t \in [0, \tau_\infty)$,

$$\begin{aligned} d\|\tilde{v}(t) - \tilde{w}(t)\|_2^2 &= 2(\tilde{v}(t) - \tilde{w}(t), d\tilde{v}(t) - d\tilde{w}(t)) + 4n\delta dt \\ &= -4(\langle \tilde{v}(t) - \tilde{w}(t), \nabla_v F(t, \tilde{v}(t)) - \nabla_v F(t, \tilde{w}(t)) \rangle - n\delta) dt \\ &\quad + 4\sqrt{\delta}(\tilde{v}(t) - \tilde{w}(t), dB(t)). \end{aligned}$$

We begin by rewriting the previous equation with a one dimensional Brownian motion.

Lemma 9 *There exists a one dimensional Brownian motion $(W(t))_{t \geq 0}$ such that, almost surely, for $t \in [0, \tau_\infty)$,*

$$\begin{aligned} d\|\tilde{v}(t) - \tilde{w}(t)\|_2^2 &= -4(\langle \tilde{v}(t) - \tilde{w}(t), \nabla_v F(t, \tilde{v}(t)) - \nabla_v F(t, \tilde{w}(t)) \rangle - n\delta) dt \\ &\quad + 4\sqrt{\delta}\|\tilde{v}(t) - \tilde{w}(t)\|_2 dW(t). \end{aligned}$$

Proof From regularity theory of elliptic equations, for any $t > 0$ the law of $\tilde{v}(t)$ on the event $\{t < \tau_\infty\}$ is absolutely continuous with respect to the Lebesgue measure and hence $\mathbb{P}(t < \tau_\infty, \tilde{v}(t) = \tilde{w}(t)) = 0$. As a consequence, $\mathbb{E}[\int_0^t \mathbb{1}_{\{s < \tau_\infty, \tilde{v}(s) = \tilde{w}(s)\}} ds] = \int_0^t \mathbb{P}(s < \tau_\infty, \tilde{v}(s) = \tilde{w}(s)) ds = 0$, hence $\int_0^t \mathbb{1}_{\{s < \tau_\infty, \tilde{v}(s) = \tilde{w}(s)\}} ds = 0$, almost surely. Therefore, by Levy's characterisation, the local martingale $(W(t))_{t \geq 0}$ defined by

$$\forall t \geq 0, \quad W(t) = \int_0^t \left(\mathbb{1}_{\{s < \tau_\infty\}} \frac{\langle \tilde{v}(s) - \tilde{w}(s), dB(s) \rangle}{\|\tilde{v}(s) - \tilde{w}(s)\|_2} + \mathbb{1}_{\{s \geq \tau_\infty\}} dB^1(s) \right),$$

where $(B^1(t))_{t \geq 0}$ is the first coordinate of $(B(t))_{t \geq 0}$, is a Brownian motion. This completes the proof. \blacksquare

From this, using Lemma 7, we have

$$d\|\tilde{v}(t) - \tilde{w}(t)\|_2^2 \leq 4n\delta dt + 4\sqrt{\delta}\|\tilde{v}(t) - \tilde{w}(t)\|_2 dW(t).$$

Then, define $Y(t) := \|\tilde{v}(t) - \tilde{w}(t)\|_2^2$, we have $Y(0) = 0$, and for $t \in [0, \tau_\infty)$,

$$dY(t) \leq 4n\delta dt + 4\sqrt{\delta Y(t)} dW(t).$$

To control $Y(t)$, it is interesting to introduce the process that saturates the constraint in the inequality above. It turns out that we can provide an exact representation of the distribution of this process, which is the squared norm of an n -dimensional Brownian motion.

Proposition 10 *The stochastic differential equation*

$$d\zeta^\delta(t) = 4n\delta dt + 4\sqrt{\delta\zeta^\delta(t)} dW(t), \quad \zeta^\delta(0) = 0, \quad (18)$$

has a unique strong solution $(\zeta^\delta(t))_{t \geq 0}$, which is defined globally in time. Besides:

- (i) *the process $(\zeta^\delta(t))_{t \geq 0}$ has the same law as $(4\delta\|B(t)\|_2^2)_{t \geq 0}$;*
- (ii) *almost surely, for any $t \in [0, \tau_\infty)$, $Y(t) \leq \zeta^\delta(t)$.*

Proof First, we define $\tilde{\zeta}^\delta(t) := 4\delta\|B(t)\|_2^2$ and deduce from Itô's formula that

$$d\tilde{\zeta}^\delta(t) = 4n\delta dt + 8\delta\langle B(t), dB(t) \rangle.$$

By the same arguments as in the proof of Lemma 9, there exists a one-dimensional Brownian motion $(\tilde{W}(t))_{t \geq 0}$ such that $\langle B(t), dB(t) \rangle = \|B(t)\|_2 d\tilde{W}(t)$, which yields

$$8\delta\langle B(t), dB(t) \rangle = 4\sqrt{\delta\tilde{\zeta}^\delta(t)} d\tilde{W}(t)$$

and shows that $(\widetilde{\zeta}^\delta(t), \widetilde{W}(t))_{t \geq 0}$ is a weak solution to the SDE (18). On the other hand, by Theorem 3.5-(ii) of (Revuz and Yor, 2013, Chapter IX), pathwise uniqueness is known to hold for this SDE. Therefore, by the Yamada–Watanabe theorem (Karatzas and Shreve, 2012, Chapter 5, Corollary 3.23), strong existence also holds. Besides, by uniqueness in law (Karatzas and Shreve, 2012, Chapter 5, Proposition 3.20), the strong solution $(\zeta^\delta(t))_{t \geq 0}$ driven by $(W(t))_{t \geq 0}$ has the same law as $(\widetilde{\zeta}^\delta(t))_{t \geq 0}$.

To prove the last statement of the proposition, we follow the lines of (Revuz and Yor, 2013, Theorem 3.7, Chapter IX). First, for any $M \geq 0$, we set $\tau_M := \inf\{t \geq 0 : \|\beta(t)\|_2 \geq M\}$, so that $\tau_\infty = \limsup_{M \rightarrow +\infty} \tau_M$. We may then write, for any $t \geq 0$,

$$(Y(t \wedge \tau_M) - \zeta^\delta(t \wedge \tau_M))_+ \leq 4\sqrt{\delta} \int_0^{t \wedge \tau_M} \mathbb{1}_{\{Y(s) \geq \zeta^\delta(s)\}} \left(\sqrt{Y(s)} - \sqrt{\zeta^\delta(s)} \right) dW(s),$$

which then implies that $\mathbb{E}[(Y(t \wedge \tau_M) - \zeta^\delta(t \wedge \tau_M))_+] = 0$ and therefore, by continuity of the trajectories of $Y(t \wedge \tau_M) - \zeta^\delta(t \wedge \tau_M)$, we have almost surely

$$\forall t \geq 0, \quad \forall M \geq 0, \quad Y(t \wedge \tau_M) \leq \zeta^\delta(t \wedge \tau_M).$$

The final claim easily follows. ■

B.1.3. CONVERGENCE OF THE INITIAL STOCHASTIC FLOW TO THE ORIGIN

We are now in place to give a bound on the iterates $\beta(t, \tilde{v}(t))$ for all $t \geq 0$.

Theorem 11 *Let $(\zeta^\delta(t))_{t \geq 0}$ be defined by Proposition 10. Almost surely,*

$$\forall t \in [0, \tau_\infty), \quad \beta(t) \leq e^{\|X^\top\|(\sqrt{\zeta^\delta(t)} + 2\|X\|\|\beta(0)\|_2/c)} \beta(0) \odot e^{-ct}.$$

In consequence, $\tau_\infty = +\infty$ and $\lim_{t \rightarrow +\infty} \beta(t) = 0$ almost surely.

Proof Thanks to Propositions 8 and 10, we have the following almost sure inequality

$$\begin{aligned} \|\tilde{v}(t) - \tilde{v}(0)\|_2 &\leq \|\tilde{v}(t) - \tilde{w}(t)\|_2 + \|\tilde{w}(t) - \tilde{w}(0)\|_2 \\ &\leq \sqrt{\zeta^\delta(t)} + 2\|\beta_0\|_2 \frac{\|X\|}{c}. \end{aligned}$$

Transferring this estimation to the iterates of the initial flow, we get,

$$\begin{aligned} \beta(t) &= \exp\left(X^\top \tilde{v}(t) + u_* - ct\right) \\ &= \beta(0) \odot \exp\left(X^\top (\tilde{v}(t) - \tilde{v}(0)) - ct\right) \\ &\leq e^{\|X^\top\| \|\tilde{v}(t) - \tilde{v}(0)\|_2} \beta(0) \odot e^{-ct} \\ &\leq e^{\|X^\top\|(\sqrt{\zeta^\delta(t)} + 2\|\beta_0\|_2\|X\|/c)} \beta(0) \odot e^{-ct}, \end{aligned}$$

which proves the first inequality in the theorem and implies that $\tau_\infty = +\infty$, almost surely. To prove that $\beta(t)$ goes to 0, we first note that, by the law of the iterated logarithm (Revuz and Yor, 2013, Corollary 1.12, Chapter II),

$$\limsup_{t \rightarrow +\infty} \frac{\|B(t)\|_2^2}{2t \log \log(t)} \leq n, \quad \text{almost surely,}$$

which by Proposition 10 then implies that

$$\limsup_{t \rightarrow +\infty} \|X^\top\| \sqrt{\zeta^\delta(t)} \mathbb{1} - ct = -\infty, \quad \text{almost surely,}$$

and completes the proof. \blacksquare

B.2. The standard noise regime: Theorem 4

B.2.1. INTRODUCTION OF THE DUAL VARIABLES

We now work in the setting of Theorem 4, where in particular $S^\circ \neq \emptyset$ and $\delta < \delta_-$. We recall the explicit form of $\beta^\mathbb{L}$ and its conic dual variable $\mu^\mathbb{L}$ as derived in Theorem 2,

$$\beta^\mathbb{L} = \underbrace{[\beta_{S^\circ}^\circ - \delta(2X_{S^\circ}^\top X_{S^\circ})^{-1}h_{S^\circ}, 0_{S^{\circ c}}]}_{\beta_{S^\circ}^\mathbb{L} > 0}, \quad \mu^\mathbb{L} = [0_{S^\circ}, \underbrace{\delta(h_{S^{\circ c}} - X_{S^{\circ c}}^\top X_{S^\circ}(X_{S^\circ}^\top X_{S^\circ})^{-1}h_{S^\circ})}_{\mu_{S^{\circ c}}^\mathbb{L} > 0}].$$

As said in Section 4.3, the conceptual crux is to see the problem like a maximum margin selection of a linearly separable classification problem *on the data points given by the coordinates of the* $(x_i)_{1 \leq i \leq n}$. Let us recall the definition of this transposed dataset more precisely: define the d data inputs $(z_k)_{1 \leq k \leq d} \in (\mathbb{R}^n)^d$ such that $Z = X^\top$, i.e for all $k \in \llbracket 1, d \rrbracket$, $i \in \llbracket 1, n \rrbracket$ $[z_k]_i = [x_i]_k$. Let us define, the following dual variables of \mathbb{R}^n :

$$\begin{aligned} v^\mathbb{L} &:= X_{S^\circ}(X_{S^\circ}^\top X_{S^\circ})^{-1} \log(\beta_{S^\circ}^\mathbb{L}), \\ v^\infty &:= -2X\beta^\mathbb{L} = -2X_{S^\circ}\beta_{S^\circ}^\mathbb{L}. \end{aligned}$$

The two variables have the following property that will be crucial to derive the behaviour of the dual process (16), and easily follows from the definitions of $v^\mathbb{L}$, v^∞ together with the condition (KKT $_\delta$). We recall here that $c = \delta h - 2X^\top y$.

Lemma 12 *We have the following properties on $(v^\mathbb{L}, v^\infty)$:*

- (i) $v^\infty = -2X_{S^\circ}e^{X_{S^\circ}^\top v^\mathbb{L}}$, i.e translated in terms of the z 's variables as $v^\infty = -2 \sum_{k \in S^\circ} z_k e^{\langle z_k, v^\mathbb{L} \rangle}$.
- (ii) $-X^\top v^\infty + c = \mu^\mathbb{L}$. This translates in terms of z 's variables as for $k \in S^\circ$, $-\langle z_k, v^\infty \rangle + c_k = 0$ and for $k \in S^{\circ c}$, $-\langle z_k, v^\infty \rangle + c_k = \mu_k^\mathbb{L} > 0$.

B.2.2. THE RESIDUAL PROCESS

We have introduced such vectors for a concrete purpose. Indeed, we are going to show that at first order, the process $(\tilde{v}(t))_{0 \leq t < \tau_\infty}$ defined in Eq. (16) will approximately diverge as t along the ray $\{v^\infty t + v^\mathbb{L}, t \geq 0\}$. Define, for $t \in [0, \tau_\infty)$, what can be called the residual process,

$$r(t) := \tilde{v}(t) - v^\infty t - v^\mathbb{L} + X_{S^\circ}(X_{S^\circ}^\top X_{S^\circ})^{-1}[u_*]_{S^\circ},$$

where $[u_*]_{S^\circ} \in \mathbb{R}^s$ is the restriction of $u_* \in \mathbb{R}^d$ on the support S° . We also define $u'_* := X_{S^\circ}^\top X_{S^\circ} (X_{S^\circ}^\top X_{S^\circ})^{-1} [u_*]_{S^\circ} \in \mathbb{R}^{d-s}$, and deduce from Itô calculus and Lemma 12 that

$$\begin{aligned}
 d\|r(t)\|_2^2 &= 2\langle dr(t), r(t) \rangle + 4n\delta dt \\
 &= -4\langle X e^{X^\top \tilde{v}(t) + u_* - ct} - v^\infty, r(t) \rangle dt + 4n\delta dt + 4\sqrt{\delta} \langle r(t), dB(t) \rangle \\
 &= -4\langle X e^{X^\top r(t) + X^\top v^\perp + u_* - u'_* + (X^\top v^\infty - c)t} + v^\infty, r(t) \rangle dt + 4n\delta dt + 4\sqrt{\delta} \langle r(t), dB(t) \rangle \\
 &= -4\langle X \left(e^{X^\top r(t) + X^\top v^\perp + u_* - u'_* - \mu^\perp t} - \beta^\perp \right), r(t) \rangle dt + 4n\delta dt + 4\sqrt{\delta} \langle r(t), dB(t) \rangle \\
 &= -4 \sum_{k \in S^\circ} e^{\langle z_k, v^\perp \rangle} \left(e^{\langle z_k, r(t) \rangle} - 1 \right) \langle z_k, r(t) \rangle dt \\
 &\quad - 4 \sum_{k \in S^{\circ c}} e^{\langle z_k, v^\perp \rangle + [u_*]_k - [u'_*]_k - \mu_k^\perp t} e^{\langle z_k, r(t) \rangle} \langle z_k, r(t) \rangle dt + 4n\delta dt + 4\sqrt{\delta} \langle r(t), dB(t) \rangle.
 \end{aligned}$$

By the same argument as in Lemma 9, there exists a one dimensional Brownian motion $(W(t))_{t \geq 0}$ such that for all $t \in [0, \tau_\infty)$,

$$\begin{aligned}
 d\|r(t)\|_2^2 &= -4 \sum_{k \in S^\circ} \beta_k^\perp \left(e^{\langle z_k, r(t) \rangle} - 1 \right) \langle z_k, r(t) \rangle dt \\
 &\quad - 4 \sum_{k \in S^{\circ c}} e^{\langle z_k, v^\perp \rangle + [u_*]_k - [u'_*]_k - \mu_k^\perp t} e^{\langle z_k, r(t) \rangle} \langle z_k, r(t) \rangle dt + 4n\delta dt + 4\sqrt{\delta} \|r(t)\| dW(t).
 \end{aligned}$$

Using that for any $x \in \mathbb{R}$, $-xe^x \leq |x|$, defining $\underline{\mu} = \min_{k \in S^{\circ c}} \mu_k^\perp$, and finally noting the sum $b = \sum_{k \in S^{\circ c}} e^{\langle z_k, v^\perp \rangle + [u_*]_k - [u'_*]_k} \|z_k\|_2$, we get that the second term is upper bounded by $be^{-\underline{\mu}t} \|r(t)\|_2$.

The first term is a bit more involved and crucially rests on the strong convexity on the support given by the invertibility of $X_{S^\circ}^\top X_{S^\circ}$. Indeed, first note that for any $x \in \mathbb{R}$, $x(e^x - 1) \geq x^2/(1 + |x|)$. Moreover if we denote $a = \min_{k \in S^\circ} \beta_k^\perp$, $\Omega = \sup_{k \in S^\circ} \|z_k\|_2$, and let $\rho_{S^\circ} > 0$ such that $X_{S^\circ}^\top X_{S^\circ} \geq \rho_{S^\circ} I_{S^\circ}$,

$$\begin{aligned}
 -4 \sum_{k \in S^\circ} \beta_k^\perp \left(e^{\langle z_k, r(t) \rangle} - 1 \right) \langle z_k, r(t) \rangle &\leq -4a \sum_{k \in S^\circ} \frac{\langle z_k, r(t) \rangle^2}{1 + |\langle z_k, r(t) \rangle|} \\
 &\leq -4a \sum_{k \in S^\circ} \frac{\langle z_k, r(t) \rangle^2}{1 + \Omega \|r(t)\|_2} \\
 &\leq -\frac{4a}{1 + \Omega \|r(t)\|_2} \sum_{k \in S^\circ} \langle z_k, r(t) \rangle^2 \\
 &\leq -4a\rho_{S^\circ} \frac{\|r(t)\|_2^2}{1 + \Omega \|r(t)\|_2}.
 \end{aligned}$$

Finally,

$$d\|r(t)\|_2^2 \leq \left(4n\delta - 4a\rho_{S^\circ} \frac{\|r(t)\|_2^2}{1 + \Omega \|r(t)\|_2} + 4be^{-\underline{\mu}t} \|r(t)\|_2 \right) dt + 4\sqrt{\delta} \|r(t)\|_2 dW(t). \quad (19)$$

B.2.3. THE COMPARISON PROCESS $(\xi^\delta(t))_{t \geq 0}$ AND THE FIRST PART OF THE THEOREM

Everything is now in order to apply the same SDE comparison argument we detailed in the non-strongly convex case in Subsection B.1.2.

Proposition 13 *Assume that $n \geq 2$ and $r(0) \neq 0$. The stochastic differential equation*

$$d\xi^\delta(t) = \left(4n\delta - 4a\rho_{S^0} \frac{\xi^\delta(t)}{1 + \Omega\sqrt{\xi^\delta(t)}} + 4be^{-\mu t} \sqrt{\xi^\delta(t)} \right) dt + 4\sqrt{\delta\xi^\delta(t)} dW(t), \quad (20)$$

initialised at $\xi^\delta(0) = \|r(0)\|_2^2$, has a unique strong solution $(\xi^\delta(t))_{t \geq 0}$, which is defined globally in time. Besides:

(i) *the process $(\xi^\delta(t))_{t \geq 0}$ has the same law as $(\|R^\delta(t)\|_2^2)_{t \geq 0}$, where $(R^\delta(t))_{t \geq 0}$ is the unique strong solution to the n -dimensional SDE*

$$dR^\delta(t) = \left(-2a\rho_{S^0} \frac{R^\delta(t)}{1 + \Omega\|R^\delta(t)\|_2} + 2be^{-\mu t} \frac{R^\delta(t)}{\|R^\delta(t)\|_2} \right) dt + 2\sqrt{\delta} dB(t), \quad (21)$$

initialised at $R^\delta(0) = r(0)$;

(ii) *almost surely, for any $t \in [0, \tau_\infty)$, $\|r(t)\|_2^2 \leq \xi^\delta(t)$.*

Proof The well-posedness of (21) follows from Veretennikov (1981), and Itô's formula together with Lévy's characterisation show that the process $(\|R^\delta(t)\|_2^2)_{t \geq 0}$ is a weak solution to (20). However, in contrast with the proof of Proposition 10, the presence of the square root in the drift of the SDE (20) prevents us from using standard results to claim pathwise uniqueness. To recover this property, we note that since the drift remains Lipschitz continuous, uniformly in time, on all sets of the form $[\varepsilon, +\infty)$, $\varepsilon > 0$, any two strong solutions to (20) coincide, and have the same law as the process $\|R^\delta(t)\|_2^2$, until they hit 0. As a consequence, to obtain pathwise uniqueness for (20) it suffices to check that, almost surely, $R^\delta(t)$ never hits 0. But since the drift in (21) is bounded, uniformly in time, the Girsanov theorem shows that for any $T > 0$, the laws of $(R^\delta(t))_{t \in [0, T]}$ and $(r(0) + 2\sqrt{\delta}B(t))_{t \in [0, T]}$ are equivalent, and therefore

$$\mathbb{P} \left(\exists t \in [0, T] : R^\delta(t) = 0 \right) = \mathbb{P} \left(\exists t \in [0, T] : r(0) + 2\sqrt{\delta}B(t) = 0 \right) = 0,$$

where the second equality is well-known in dimension $n \geq 2$. This implies that, almost surely, $R^\delta(t) \neq 0$, for any $t \geq 0$, and therefore completes the proof of pathwise uniqueness for (20).

We now detail the comparison between $\|r(t)\|_2^2$ and $\xi^\delta(t)$. For any $M > M_0 := 1 \vee \|r(0)\|_2^{-2}$, we set $\tau'_M := \inf\{t \in [0, \tau_\infty) : \|r(t)\|_2^2 \wedge \xi^\delta(t) \leq 1/M \text{ or } \|\beta(t)\|_2 \geq M\}$, so that $\limsup_{M \rightarrow +\infty} \tau'_M = \tau_0 \wedge \tau_\infty$, with $\tau_0 := \inf\{t \in [0, \tau_\infty) : \|r(t)\|_2 = 0\}$ (we recall that, almost surely, $\xi^\delta(t)$ never hits 0). With similar arguments to (Revuz and Yor, 2013, Theorem 3.7, Chapter IX), we may write

$$\mathbb{E} \left[\left(\|r(t \wedge \tau'_M)\|_2^2 - \xi^\delta(t \wedge \tau'_M) \right)_+ \right] \leq \mathbb{E} \left[\int_0^{t \wedge \tau'_M} \mathbb{1}_{\{\|r(s)\|_2^2 \geq \xi^\delta(s)\}} \left| g(s, \|r(s)\|_2^2) - g(s, \xi^\delta(s)) \right| ds \right],$$

where $g(t, \xi) = 4n\delta - 4a\rho_{S^0}\xi/(1 + \Omega\sqrt{\xi}) + 4be^{-\mu t}\sqrt{\xi}$ is the drift of (20). Denoting by C_M the Lipschitz constant of $g(s, \cdot)$ on $[1/M, +\infty)$, which is uniform in s , we get

$$\begin{aligned} \mathbb{E} \left[\left(\|r(t \wedge \tau'_M)\|_2^2 - \xi^\delta(t \wedge \tau'_M) \right)_+ \right] &\leq C_M \mathbb{E} \left[\int_0^{t \wedge \tau'_M} \mathbf{1}_{\{\|r(s)\|_2^2 \geq \xi^\delta(s)\}} \left| \|r(s)\|_2^2 - \xi^\delta(s) \right| ds \right] \\ &= C_M \mathbb{E} \left[\int_0^{t \wedge \tau'_M} \left(\|r(s)\|_2^2 - \xi^\delta(s) \right)_+ ds \right] \\ &\leq C_M \int_0^t \mathbb{E} \left[\left(\|r(s \wedge \tau'_M)\|_2^2 - \xi^\delta(s \wedge \tau'_M) \right)_+ \right] ds, \end{aligned}$$

which by Gronwall's lemma and the continuity of the trajectories of $\|r(t \wedge \tau'_M)\|_2^2 - \xi^\delta(t \wedge \tau'_M)$ yields, almost surely,

$$\forall t \geq 0, \quad \forall M > M_0, \quad \|r(t \wedge \tau'_M)\|_2^2 \leq \xi^\delta(t \wedge \tau'_M),$$

and therefore, almost surely,

$$\forall t \in [0, \tau_\infty \wedge \tau_0), \quad \|r(t)\|_2^2 \leq \xi^\delta(t).$$

To complete the proof, let us fix $T \in [0, \tau_\infty)$ and call $z := \inf_{t \in [0, T]} \xi^\delta(t) > 0$. The previous argument shows that $\|r(t)\|_2^2 \leq \xi^\delta(t)$ for all $t \leq T \wedge \tau_0$, and if $\tau_0 < T$, this inequality remains trivially true as long as $t \leq \tau_z^{(1)} := \inf\{t \in [\tau_0, T] : \|r(t)\|_2^2 \geq z\}$. But if $\tau_z^{(1)} < T$, the argument above can be repeated to show that the inequality holds up to $\tau_0^{(1)} := \inf\{t \in [\tau_z^{(1)}, T] : \|r(t)\|_2 = 0\}$, and then, if $\tau_0^{(1)} < T$, up to $\tau_z^{(2)} := \inf\{t \in [\tau_0^{(1)}, T] : \|r(t)\|_2^2 \geq z\}$. Iterating the argument, we thus construct two sequences $\tau_0^{(l)}, \tau_z^{(l)} \leq T$ such that $\|r(\tau_0^{(l)})\|_2 = 0$ if $\tau_0^{(l)} < T$, and $\|r(\tau_z^{(l)})\|_2^2 \geq z$ if $\tau_z^{(l)} < T$, and such that the inequality $\|r(t)\|_2^2 \leq \xi^\delta(t)$ holds on $[0, \tau_0^{(l)}]$. By continuity of the trajectory of $r(t)$ on $[0, T]$, there are only finitely many $\tau_0^{(l)}, \tau_z^{(l)}$ which are strictly below T , and therefore the inequality $\|r(t)\|_2^2 \leq \xi^\delta(t)$ finally holds on $[0, T]$. \blacksquare

We are now in place to give a bound on the iterates $\beta(t, \tilde{v}(t))$ for all $t \in [0, \tau_\infty)$.

Theorem 14 *Let $(\xi^\delta(t))_{t \geq 0}$ be defined in Proposition 13.*

(i) *On the support: almost surely,*

$$\forall t \in [0, \tau_\infty), \quad \beta_{S^0}^L e^{-\|X_{S^0}^\top\| \sqrt{\xi^\delta(t)}} \leq \beta_{S^0}(t) \leq \beta_{S^0}^L e^{\|X_{S^0}^\top\| \sqrt{\xi^\delta(t)}}.$$

(ii) *Outside the support: there exists $C \geq 0$, depending on X , y and $\beta(0)$, such that, almost surely,*

$$\forall t \in [0, \tau_\infty), \quad \beta_{S^{0c}}(t) \leq C \exp \left(\|X_{S^{0c}}^\top\| \sqrt{\xi^\delta(t)} \mathbf{1}_{S^{0c}} - \mu_{S^{0c}}^L t \right).$$

In particular, $\tau_\infty = +\infty$, almost surely.

Proof We transfer the estimates on $\tilde{v}(t)$ to the iterates of the initial flow on the linear predictor $\beta(t)$. Indeed as $X^\top \tilde{v}(t) = X^\top r(t) + X^\top v^\infty t + X^\top v^L - [[u_*]_{S^o}, u'_*]^\top$, we get for $t \in [0, \tau_\infty)$,

$$\begin{aligned} \beta(t) &= \exp\left(X^\top \tilde{v}(t) + u_* - ct\right) \\ &= \exp\left(X^\top r(t) + X^\top v^\infty t + X^\top v^L - [[u_*]_{S^o}, u'_*]^\top + u_* - ct\right) \\ &= \exp\left(X^\top r(t) + X^\top v^L + [0_{S^o}, [u_*]_{S^o c} - u'_*]^\top + (X^\top v^\infty - c)t\right) \\ &= \exp\left(X^\top r(t) + X^\top v^L + [0_{S^o}, [u_*]_{S^o c} - u'_*]^\top - \mu^L t\right), \end{aligned}$$

thanks to Lemma 12. On the support, we have $[e^{X^\top v^L}]_{S^o} = \beta_{S^o}^L$ and $\mu_{S^o}^L = 0$. Hence, $\beta_{S^o}(t) = \beta_{S^o}^L \odot \exp(X_{S^o}^\top r(t))$, and the first part of the Theorem follows from Proposition 13. The second part of the Theorem follows similarly, with $C = \max_{k \in S^o c} \exp([X^\top v^L]_k + [u_*]_k - [u'_*]_k)$. Finally, the fact that $\xi^\delta(t)$ does not explode yields $\tau_\infty = +\infty$, which completes the proof. \blacksquare

B.2.4. STUDY OF THE PROCESS $(R^\delta(t))_{t \geq 0}$ AND THE SECOND PART OF THE THEOREM

This section is dedicated to a detailed study of the long-time behaviour of the process $(R^\delta(t))_{t \geq 0}$, and therefore of $(\xi^\delta(t))_{t \geq 0}$. To proceed, we first rewrite (20) under the form

$$dR^\delta(t) = G(t, R^\delta(t))dt + 2\sqrt{\delta}dB(t), \quad G(t, r) := -\nabla V(r) + 2be^{-\mu t} \frac{r}{\|r\|_2},$$

where $V : \mathbb{R}^n \rightarrow \mathbb{R}$ is the C^1 function with globally bounded gradient defined by

$$V(r) := \frac{2a\rho_{S^o}}{\Omega^2} (\Omega\|r\|_2 - \log(1 + \Omega\|r\|_2)). \quad (22)$$

We also introduce the time-homogeneous diffusion process

$$d\bar{R}^\delta(t) = -\nabla V(\bar{R}^\delta(t))dt + 2\sqrt{\delta}dB(t), \quad \bar{R}^\delta(0) = r(0).$$

We first state a trajectorial comparison result between $R^\delta(t)$ and $\bar{R}^\delta(t)$.

Proposition 15 *Almost surely,*

$$\sup_{t \geq 0} \|\bar{R}^\delta(t) - R^\delta(t)\|_2 \leq \frac{b}{\mu}, \quad \text{and} \quad \lim_{t \rightarrow +\infty} \|\bar{R}^\delta(t) - R^\delta(t)\|_2 = 0.$$

Proof Step 1. Convexity of V . We start the proof by showing that V is convex on \mathbb{R}^n and bounding $\nabla^2 V(r)$ from below. To proceed, we compute

$$\nabla^2 V(r) = \frac{2a\rho_{S^o}}{(1 + \Omega\|r\|_2)^2} \left((1 + \Omega\|r\|_2)I_n - \Omega \frac{rr^\top}{\|r\|_2} \right), \quad (23)$$

so that, for any $u \in \mathbb{R}^n$,

$$\begin{aligned} \langle u, \nabla^2 V(r)u \rangle &= \frac{2a\rho_{S^\circ}}{(1 + \Omega\|r\|_2)^2} \left((1 + \Omega\|r\|_2)\|u\|_2^2 - \Omega \frac{\langle r, u \rangle^2}{\|r\|_2} \right) \\ &\geq \frac{2a\rho_{S^\circ}}{(1 + \Omega\|r\|_2)^2} ((1 + \Omega\|r\|_2)\|u\|_2^2 - \Omega\|r\|_2\|u\|_2^2) \\ &= \frac{2a\rho_{S^\circ}}{1 + \Omega\|r\|_2} \|u\|_2^2. \end{aligned}$$

As a consequence, for any $r, \bar{r} \in \mathbb{R}^n$,

$$\begin{aligned} \langle \bar{r} - r, \nabla V(\bar{r}) - \nabla V(r) \rangle &= \left\langle \bar{r} - r, \int_0^1 \frac{d}{dt} \nabla V(t\bar{r} + (1-t)r) dt \right\rangle \\ &= \int_0^1 \langle \bar{r} - r, \nabla^2 V(t\bar{r} + (1-t)r)(\bar{r} - r) \rangle dt \\ &\geq 2a\rho_{S^\circ} \int_0^1 \frac{\|\bar{r} - r\|_2^2}{1 + \Omega\|t\bar{r} + (1-t)r\|_2} dt \\ &\geq 2a\rho_{S^\circ} \frac{\|\bar{r} - r\|_2^2}{1 + \Omega(\|\bar{r}\|_2 \vee \|r\|_2)}. \end{aligned}$$

Step 2. Global estimate. Using only the nonnegativity of $\langle \bar{R}^\delta(t) - R^\delta(t), -\nabla V(\bar{R}^\delta(t)) + \nabla V(R^\delta(t)) \rangle$, we get, for any $t \geq 0$,

$$\begin{aligned} \frac{d}{dt} \|\bar{R}^\delta(t) - R^\delta(t)\|_2^2 &= 2 \left\langle \bar{R}^\delta(t) - R^\delta(t), -\nabla V(\bar{R}^\delta(t)) + 2be^{-\mu t} \frac{R^\delta(t)}{\|\bar{R}^\delta(t)\|_2} + \nabla V(R^\delta(t)) \right\rangle \\ &\leq 2be^{-\mu t} \|\bar{R}^\delta(t) - R^\delta(t)\|_2, \end{aligned}$$

and therefore the global estimate on $\|\bar{R}^\delta(t) - R^\delta(t)\|_2$ follows from Lemma 16.

Step 3. Long time convergence. We shall make use of this global estimate to prove the long time convergence of $\|\bar{R}^\delta(t) - R^\delta(t)\|_2$ to 0. We first note that, for any $r, \bar{r} \in \mathbb{R}^n$, if $\|\bar{r} - r\|_2 \leq b/\underline{\mu}$, then $\|\bar{r}\|_2 \vee \|r\|_2 \leq \|\bar{r}\|_2 + b/\underline{\mu}$ and thus

$$\langle \bar{r} - r, \nabla V(\bar{r}) - \nabla V(r) \rangle \geq 2a\rho_{S^\circ} \frac{\|\bar{r} - r\|_2^2}{1 + \Omega(\|\bar{r}\|_2 + b/\underline{\mu})}.$$

Let us now fix $M > 1$ large enough for the inequality

$$\mu' := 2a\rho_{S^\circ}/(1 + \Omega(M + b/\underline{\mu})) < \underline{\mu}$$

to hold, and define

$$\begin{aligned} \tau_0^- &:= \inf\{t \geq 0 : \|\bar{R}^\delta(t)\|_2 \leq M - 1\}, \\ \tau_\ell^+ &:= \inf\{t \geq \tau_\ell^- : \|\bar{R}^\delta(t)\|_2 \geq M\}, \quad \ell \geq 0, \\ \tau_\ell^- &:= \inf\{t \geq \tau_{\ell-1}^+ : \|\bar{R}^\delta(t)\|_2 \leq M - 1\}, \quad \ell \geq 1. \end{aligned}$$

Since $\|\bar{R}^\delta(t)\|_2^2$ satisfies a time-homogeneous SDE and is ergodic, the sequences of positive random variables $(\tau_\ell^+ - \tau_\ell^-)_{\ell \geq 0}$ and $(\tau_{\ell+1}^- - \tau_\ell^+)_{\ell \geq 0}$ are well-defined, and by the strong Markov property they are iid and therefore $\tau_\ell^-, \tau_\ell^+ \rightarrow +\infty$ with ℓ . Besides, on account on the previous discussion, we have

$$\frac{d}{dt} \|\bar{R}^\delta(t) - R^\delta(t)\|_2^2 \leq -2\mu' \|\bar{R}^\delta(t) - R^\delta(t)\|_2^2 + 2be^{-\mu t} \|\bar{R}^\delta(t) - R^\delta(t)\|_2, \quad \text{on } [\tau_\ell^-, \tau_\ell^+],$$

and

$$\frac{d}{dt} \|\bar{R}^\delta(t) - R^\delta(t)\|_2^2 \leq 2be^{-\mu t} \|\bar{R}^\delta(t) - R^\delta(t)\|_2, \quad \text{on } [\tau_\ell^+, \tau_{\ell+1}^-].$$

We deduce from Lemma 16 that

$$\|\bar{R}^\delta(\tau_0^-) - R^\delta(\tau_0^-)\|_2 \leq \frac{b}{\underline{\mu}} \left(1 - e^{-\underline{\mu}\tau_0^-}\right),$$

and for any $t \in [\tau_\ell^-, \tau_\ell^+]$,

$$\|\bar{R}^\delta(t) - R^\delta(t)\|_2 \leq e^{-\mu'(t-\tau_\ell^-)} \|\bar{R}^\delta(\tau_\ell^-) - R^\delta(\tau_\ell^-)\|_2 + \frac{be^{-\mu't}}{\underline{\mu} - \mu'} \left(e^{-(\underline{\mu}-\mu')\tau_\ell^-} - e^{-(\underline{\mu}-\mu')t}\right),$$

and for any $t \in [\tau_\ell^+, \tau_{\ell+1}^-]$,

$$\|\bar{R}^\delta(t) - R^\delta(t)\|_2 \leq \|\bar{R}^\delta(\tau_\ell^+) - R^\delta(\tau_\ell^+)\|_2 + \frac{b}{\underline{\mu}} \left(e^{-\underline{\mu}\tau_\ell^+} - e^{-\underline{\mu}t}\right).$$

As a consequence, the sequence $Q_\ell := \|\bar{R}^\delta(\tau_\ell^-) - R^\delta(\tau_\ell^-)\|_2$ satisfies the recursive inequation $Q_{\ell+1} \leq \alpha_\ell Q_\ell + \beta_\ell$, where

$$\alpha_\ell := e^{-\mu'(\tau_\ell^+ - \tau_\ell^-)}, \quad \beta_\ell := \frac{be^{-\mu'\tau_\ell^+}}{\underline{\mu} - \mu'} \left(e^{-(\underline{\mu}-\mu')\tau_\ell^-} - e^{-(\underline{\mu}-\mu')\tau_\ell^+}\right) + \frac{b}{\underline{\mu}} \left(e^{-\underline{\mu}\tau_\ell^+} - e^{-\underline{\mu}\tau_{\ell+1}^-}\right),$$

and moreover we have the intermediate (rough) control

$$\sup_{\tau_\ell^- \leq t \leq \tau_{\ell+1}^-} \|\bar{R}^\delta(t) - R^\delta(t)\|_2 \leq Q_\ell + \frac{b}{\underline{\mu} - \mu'} e^{-(\underline{\mu}-\mu')\tau_\ell^-} + \frac{b}{\underline{\mu}} e^{-\underline{\mu}\tau_\ell^+},$$

so that, since $\tau_\ell^-, \tau_\ell^+ \rightarrow +\infty$, to show that $\|\bar{R}^\delta(t) - R^\delta(t)\|_2 \rightarrow 0$ is suffices to show that $Q_\ell \rightarrow 0$. The recursive inequation yields, for any $\ell \geq 1$,

$$Q_\ell \leq \left(\prod_{m=0}^{\ell-1} \alpha_m\right) Q_0 + \sum_{m=0}^{\ell-1} \left(\prod_{k=m+1}^{\ell-1} \alpha_k\right) \beta_m.$$

On the one hand,

$$\prod_{m=0}^{\ell-1} \alpha_m = \exp\left(-\mu' \sum_{m=0}^{\ell-1} (\tau_m^+ - \tau_m^-)\right),$$

and since the sequence $(\tau_m^+ - \tau_m^-)_{m \geq 0}$ is iid with $\mathbb{P}(\tau_0^+ - \tau_0^- > 0) = 1$, we have $\sum_{m=0}^{\ell-1} (\tau_m^+ - \tau_m^-) \rightarrow +\infty$, almost surely. On the other hand, we first note that

$$\begin{aligned} \beta_m &= \frac{b}{\underline{\mu} - \mu'} \left(e^{-\mu'(\tau_m^+ - \tau_m^-)} e^{-\underline{\mu}\tau_m^-} - e^{-\underline{\mu}\tau_m^+} \right) + \frac{b}{\underline{\mu}} \left(e^{-\underline{\mu}\tau_m^+} - e^{-\underline{\mu}\tau_{m+1}^-} \right) \\ &\leq \frac{b}{\underline{\mu} - \mu'} \left(e^{-\underline{\mu}\tau_m^-} - e^{-\underline{\mu}\tau_{m+1}^-} \right), \end{aligned}$$

so that

$$\begin{aligned} \sum_{m=0}^{\ell-1} \left(\prod_{k=m+1}^{\ell-1} \alpha_k \right) \beta_m &\leq \frac{b}{\underline{\mu} - \mu'} \sum_{m=0}^{\ell-1} \left(\prod_{k=m+1}^{\ell-1} \alpha_k \right) \left(e^{-\underline{\mu}\tau_m^-} - e^{-\underline{\mu}\tau_{m+1}^-} \right) \\ &= \frac{b}{\underline{\mu} - \mu'} \sum_{m=0}^{\ell-1} e^{-\underline{\mu}\tau_m^-} (1 - \alpha_m) \left(\prod_{k=m+1}^{\ell-1} \alpha_k \right) - e^{-\underline{\mu}\tau_{\ell}^-}. \end{aligned}$$

Since $\tau_{\ell}^- \rightarrow +\infty$ when $\ell \rightarrow +\infty$, the remainder $e^{-\underline{\mu}\tau_{\ell}^-}$ vanishes when $\ell \rightarrow +\infty$. Besides, for any $\varepsilon > 0$, there exists m_0 such that for all $m \geq m_0$, $e^{-\underline{\mu}\tau_m^-} \leq \varepsilon$. We may then write, for $\ell - 1 \geq m_0$,

$$\sum_{m=0}^{\ell-1} e^{-\underline{\mu}\tau_m^-} (1 - \alpha_m) \left(\prod_{k=m+1}^{\ell-1} \alpha_k \right) \leq \sum_{m=0}^{m_0-1} e^{-\underline{\mu}\tau_m^-} (1 - \alpha_m) \left(\prod_{k=m+1}^{\ell-1} \alpha_k \right) + \varepsilon \sum_{m=m_0}^{\ell-1} (1 - \alpha_m) \left(\prod_{k=m+1}^{\ell-1} \alpha_k \right).$$

Each one of the m_0 terms of the sum in the first term of the right-hand side goes to 0 when $\ell \rightarrow +\infty$, while the second term is telescopic and rewrites

$$\varepsilon \sum_{m=m_0}^{\ell-1} (1 - \alpha_m) \left(\prod_{k=m+1}^{\ell-1} \alpha_k \right) = \varepsilon \sum_{m=m_0}^{\ell-1} \left(\prod_{k=m+1}^{\ell-1} \alpha_k - \prod_{k=m}^{\ell-1} \alpha_k \right) \leq \varepsilon.$$

Therefore, for any $\varepsilon > 0$, we get $\limsup_{\ell \rightarrow +\infty} Q_{\ell} \leq \varepsilon$, which proves that $Q_{\ell} \rightarrow 0$, almost surely, and finally $\|\bar{R}^{\delta}(t) - R^{\delta}(t)\|_2 \rightarrow 0$ as well. \blacksquare

Lemma 16 (A Gronwall-type estimate) *Let $0 \leq \mu' < \mu$ and $b \geq 0$. Assume that $u(t) \geq 0$ satisfies*

$$u'(t) \leq -2\mu' u(t) + 2be^{-\mu t} \sqrt{u(t)}$$

on some interval $[t_1, t_2]$. Then

$$\forall t \in [t_1, t_2], \quad \sqrt{u(t)} \leq e^{-\mu'(t-t_1)} \sqrt{u(t_1)} + \frac{be^{-\mu' t}}{\mu - \mu'} \left(e^{-(\mu - \mu')t_1} - e^{-(\mu - \mu')t} \right).$$

Proof We first set $\hat{u}(t) = e^{2\mu' t} u(t)$, so that

$$\hat{u}'(t) \leq 2be^{(2\mu' - \mu)t} \sqrt{u(t)} = 2be^{(\mu' - \mu)t} \sqrt{\hat{u}(t)}.$$

For any $\varepsilon > 0$, we therefore deduce that the function \hat{u}_{ε} defined by $\hat{u}_{\varepsilon}(t) = \hat{u}(t) + \varepsilon$ satisfies $\hat{u}_{\varepsilon}(t) > 0$ and, for any $t \in [t_1, t_2]$,

$$\hat{u}'_{\varepsilon}(t) = \hat{u}'(t) \leq 2be^{(\mu' - \mu)t} \sqrt{\hat{u}(t)} \leq 2be^{(\mu' - \mu)t} \sqrt{\hat{u}_{\varepsilon}(t)}.$$

Therefore

$$\sqrt{\widehat{u}_\varepsilon(t)} - \sqrt{\widehat{u}_\varepsilon(t_1)} = \int_{t_1}^t \frac{\widehat{u}'_\varepsilon(s)}{2\sqrt{\widehat{u}_\varepsilon(s)}} ds \leq b \int_{t_1}^t e^{(\mu' - \mu)s} ds.$$

We deduce that

$$\sqrt{e^{2\mu't}u(t) + \varepsilon} \leq \sqrt{e^{2\mu't_1}u(t_1) + \varepsilon} + b \int_{t_1}^t e^{(\mu' - \mu)s} ds,$$

in which we take the $\varepsilon \rightarrow 0$ limit to rewrite

$$\sqrt{u(t)} \leq e^{-\mu'(t-t_1)} \sqrt{u(t_1)} + be^{-\mu't} \int_{t_1}^t e^{(\mu' - \mu)s} ds,$$

and obtain the claimed estimate. ■

We are now ready to complete the proof of Theorem 4 (see next subsection for the concentration property). To proceed, we first note that it is easily checked that

$$Z^\delta := \int_{\mathbb{R}^n} \exp\left(-\frac{V(r)}{2\delta}\right) dr < +\infty,$$

which allows to define the probability measure

$$\mu_\infty^\delta(dr) := \frac{1}{Z^\delta} \exp\left(-\frac{V(r)}{2\delta}\right) dr$$

on \mathbb{R}^n . By standard arguments, the process $(\overline{R}^\delta(t))_{t \geq 0}$ is ergodic with respect to the probability measure μ_∞^δ . In particular, $\overline{R}^\delta(t)$ converges in distribution towards $R_\infty^\delta \sim \mu_\infty^\delta$, which combined with Proposition 15 entails that $R^\delta(t)$ converges in distribution to R_∞^δ , and therefore $\xi^\delta(t)$ converges in distribution to $\xi_\infty^\delta = \|R_\infty^\delta\|_2^2$.

Outside the support, it remains to show that

$$\limsup_{t \rightarrow +\infty} \|X_{S^c}^\top\| \sqrt{\xi^\delta(t) \mathbb{1}_{S^c} - \mu_{S^c}^\delta t} = -\infty, \quad \text{almost surely.}$$

We start by noting that, for any $t \geq 0$,

$$\frac{\overline{R}^\delta(t)}{t} = \frac{r(0)}{t} - \frac{1}{t} \int_0^t \nabla V(\overline{R}^\delta(s)) ds + 2\sqrt{\delta} \frac{B(t)}{t}.$$

Since V is even, by the ergodic theorem,

$$\lim_{t \rightarrow +\infty} \frac{1}{t} \int_0^t \nabla V(\overline{R}^\delta(s)) ds = \int_{\mathbb{R}^n} \nabla V(r) \mu_\infty^\delta(dr) = 0, \quad \text{almost surely,}$$

while by the law of the iterated logarithm,

$$\lim_{t \rightarrow +\infty} \frac{r(0)}{t} + 2\sqrt{\delta} \frac{B(t)}{t} = 0, \quad \text{almost surely.}$$

Therefore,

$$\lim_{t \rightarrow +\infty} \frac{\overline{R}^\delta(t)}{t} = 0, \quad \text{almost surely,}$$

and by Proposition 15 we deduce that

$$\lim_{t \rightarrow +\infty} \frac{R^\delta(t)}{t} = 0, \quad \text{almost surely.}$$

Since the processes $(\|R^\delta(t)\|)_{t \geq 0}$ and $(\sqrt{\xi^\delta(t)})_{t \geq 0}$ have the same law, we deduce that

$$\lim_{t \rightarrow +\infty} \frac{\sqrt{\xi^\delta(t)}}{t} = 0, \quad \text{almost surely,}$$

which proves the claim.

B.2.5. POINCARÉ INEQUALITY AND CONCENTRATION PROPERTIES

We first show, in Lemma 17, that the probability measure μ_∞^δ satisfies a Poincaré inequality with a constant κ^δ on which we provide an explicit bound. We recall that it means that, for any smooth function f on \mathbb{R}^n ,

$$\int_{\mathbb{R}^n} \left(f(r) - \int_{\mathbb{R}^n} f(r') \mu_\infty^\delta(dr') \right)^2 \mu_\infty^\delta(dr) \leq \kappa^\delta \int_{\mathbb{R}^n} \|\nabla f(r)\|^2 \mu_\infty^\delta(dr).$$

From this result, the concentration inequality (14) stated in Theorem 4 follows from Eq. (4.4.6), p. 192 in Bakry et al. (2014). Last, the fact that $\kappa^\delta = \mathcal{O}(\delta)$ in the $\delta \rightarrow 0$ regime follows from a basic application of the Laplace method, which is detailed in Remark 18.

Lemma 17 (Poincaré constant) *The probability measure μ_∞^δ satisfies a Poincaré inequality with constant $\kappa^\delta \leq 13(\sigma^\delta)^2/n$, where*

$$\sigma^\delta := \sqrt{\int_{\mathbb{R}^n} \|r\|_2^2 \mu_\infty^\delta(dr)} = \sqrt{\mathbb{E}[\|R_\infty^\delta\|_2^2]}.$$

Proof Let us denote by $\mu_\infty^\delta(r)$ the density of the probability measure $\mu_\infty^\delta(dr)$ with respect to the Lebesgue measure on \mathbb{R}^n . It is an immediate computation to show that the probability density $\tilde{\mu}_\infty^\delta$ defined by

$$\tilde{\mu}_\infty^\delta(\tilde{r}) := \left(\frac{\sigma^\delta}{\sqrt{n}} \right)^n \mu_\infty^\delta \left(\frac{\sigma^\delta}{\sqrt{n}} \tilde{r} \right)$$

satisfies

$$\int_{\mathbb{R}^n} \|\tilde{r}\|_2^2 \tilde{\mu}_\infty^\delta(\tilde{r}) d\tilde{r} = n,$$

and that this density writes as a log-concave function of $\|\tilde{r}\|_2$. Therefore, by Theorem 1 in Bobkov (2003), it satisfies a Poincaré inequality with constant $\tilde{\kappa}^\delta \leq 13$, and the final statement follows from an elementary rescaling argument for Poincaré inequalities. \blacksquare

Remark 18 (Estimate of the Poincaré constant) *Finally let us give some estimate on the Poincaré constant when δ is small. To give an equivalent of $(\sigma^\delta)^2 = \int_{\mathbb{R}^n} \|r\|_2^2 \mu_\infty^\delta(\mathrm{d}r)$ with the Laplace method, we approximate V by a quadratic near its minimum in $0_{\mathbb{R}^n}$, $V(r) \simeq V(0) + \frac{1}{2} \langle r, \nabla^2 V(0)r \rangle = a\rho_{S^\circ} \|r\|_2^2$ by the equation (22). Therefore by the Laplace method, in the small δ regime, the distribution with density μ_∞^δ should approximately behave as a centered Gaussian random variable, with covariance matrix $\delta/(a\rho_{S^\circ})I_n$. In particular, this gives an upperbound of the Poincaré constant:*

$$\kappa^\delta \leq \frac{13(\sigma^\delta)^2}{n} \simeq \frac{13\delta}{a\rho_{S^\circ}} = \mathcal{O}(\delta).$$

Finally remark that in the small δ limit, this estimate is tight up to the numerical constant as a lower bound with numerical constant 1 is also given in [Bobkov \(2003\)](#).