Saddle-to-Saddle Dynamics in Diagonal Linear Networks

Scott Pesme and Nicolas Flammarion
EPFL

ABSTRACT. In this paper we fully describe the trajectory of gradient flow over diagonal linear networks in the limit of vanishing initialisation. We show that the limiting flow successively jumps from a saddle of the training loss to another until reaching the minimum ℓ_1-norm solution. This saddle-to-saddle dynamics translates to an incremental learning process as each saddle corresponds to the minimiser of the loss constrained to an active set outside of which the coordinates must be zero. We explicitly characterise the visited saddles as well as the jumping times through a recursive algorithm reminiscent of the Homotopy algorithm used for computing the Lasso path. Our proof leverages a convenient arc-length time-reparametrisation which enables to keep track of the heteroclinic transitions between the jumps. Our analysis requires negligible assumptions on the data, applies to both under and overparametrised settings and covers complex cases where there is no monotonicity of the number of active coordinates. We provide numerical experiments to support our findings.

1. INTRODUCTION

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

To explain this second point, a major line of work has focused on the concept of implicit regularisation: amongst the infinite space of zero-loss solutions, the optimisation process must be implicitly biased towards solutions which have good generalisation properties for the considered real-world prediction task. Many papers have therefore shown that gradient methods have the fortunate property of asymptotically leading to solutions which have a well-behaving structure [Neyshabur, 2017, Gunasekar et al., 2017]. For example, convergence towards max-margin classifiers have been shown in various classification settings [Soudry et al., 2018, Lyu and Li, 2020, Chizat and Bach, 2020] and convergence towards low-norm solutions in regression settings [Woodworth et al., 2020, Boursier et al., 2022].

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

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

1.1. Main contribution and paper organisation.

In this paper, we provide a full description of the trajectory of gradient flow over diagonal linear networks in the limit of vanishing initialisation. **We show that the iterates successively jump from a saddle of the training loss to another and we fully characterise each visited saddle as well as the jumping times.** This result is informally presented here and illustrated in Figure 1.

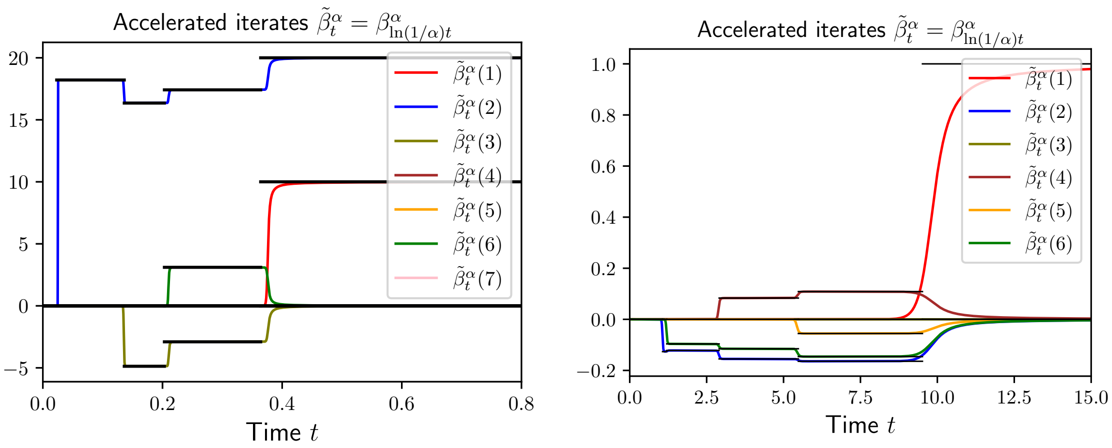

FIGURE 1. Regression problem (for the precise experimental setting, see Appendix A). *Left and right:* The magnitudes of the coordinates of the accelerated iterates $\tilde\beta^\alpha$ are plotted across time. For small initialisation scales α, a saddle-to-saddle process appears. Coordinates are not monotonic and the number of active coordinates neither as several coordinates can deactivate at the same time. The piecewise constant process plotted in black is the limiting process $\tilde\beta^\circ$ predicted by our theory.

Theorem 1. [Main result, informal.] *In the regression setting and in the limit of vanishing initialisation, the trajectory of gradient flow over diagonal linear networks converges towards a limiting process $(\tilde\beta_t^\circ)_t$ which is piecewise constant and defined as follows:*

$$\text{(“Saddles”)} \qquad \tilde\beta_t^\circ \equiv \beta_k \qquad \text{for } t \in (t_k, t_{k+1}) \text{ and } 0 \le k \le p,$$

where the “saddles” $(\beta_0 = \mathbf{0}, \beta_1, \dots, \beta_{p-1}, \beta_p)$ and jump times $(t_0 = 0, t_1, \dots, t_p, t_{p+1} = +\infty)$ can recursively and explicitly be computed by an algorithm (see Algorithm 1) reminiscent of the Homotopy algorithm for the Lasso. The final point β_p corresponds to the minimum ℓ_1-norm solution.

The learning is said to be “incremental" as each saddle corresponds to the minimiser of the loss constrained to a set of coordinates which can be non-zero, the size of this set is typically (but not necessarily) increasing.

We make minimal assumptions on the data and our analysis holds for complex datasets where there is no monotonicity of the number of active coordinates and where the successive active sets are highly non-trivial as depicted in Figure 1.

In Section 2 we introduce the regression setting and the diagonal network architecture. Our main result exhibiting the limiting saddle-to-saddle dynamics is provided Section 3 and the sketch of proof based on the arc-length parametrisation is given Section 4.

1.2. Related works.

Diagonal linear networks. Diagonal linear networks are simplified neural networks which have seen a surge of interest recently [Woodworth et al., 2020, Vaskevicius et al., 2019, HaoChen et al., 2021, Pesme et al., 2021, Even et al., 2023]. For these networks, the scale of the initialisation determines the structure of the recovered solution. Large initialisations yield low ℓ_2-norm solutions (referred to as either the neural tangent kernel regime [Jacot et al., 2018], or the lazy regime [Chizat et al., 2019]). Small initialisations result in low ℓ_1-norm solutions (known as the rich regime [Woodworth et al., 2020]). Despite their simplicity, diagonal linear networks reveal training characteristics observed in much more complex architectures. We point out that an even more drastic simplification is to consider a quadratic parametrisation where the regression predictor β is parametrised $\beta = u^2$, with element-wise multiplication [Amid and Warmuth, 2020, Vivien et al., 2022, Berthier, 2022].

Incremental learning / saddle-to-saddle dynamics. In all the papers mentioned in the introduction, the analysed incremental learning phenomenon and the saddle-to-saddle process are complementary facets of the same idea. Indeed for gradient flows $\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t$, fixed points of the dynamics correspond to critical points of the loss. Stages with little progress in learning and minimal movement of the iterates necessarily correspond to the iterates being in the vicinity of a critical point of the loss. It turns out that in many settings (linear networks [Kawaguchi, 2016], matrix sensing [Bhojanapalli et al., 2016, Park et al., 2017]), critical points are necessarily saddle points of the loss (if not global minima) and they have a very particular structure (high sparsity, low rank etc.). The concepts of incremental learning and of saddle-to-saddle process refer therefore to the same phenomenon in many settings. In the dynamical systems literature, saddle-to-saddle dynamics are referred as heteroclinic networks [Krupa, 1997, Ashwin and Field, 1999] and are characterised by the connection of multiple fixed points through orbits of the flow. We note that an alternative approach to realising these dynamics is through the perturbation of the gradient flow by a vanishing noise as studied by Bakhtin [2011].

Rate-independent systems. Rate-independent systems refer to time-dependent processes which are invariant under time rescaling. Examples of such systems are ubiquitous in mechanics and appear in problems related to friction, crack propagation, elastoplasticity, ferromagnetism to name a few [Mielke, 2005, Ch. 6 for a survey]. Such systems emerge from differential problems of the form $\partial_q E(t, q_t) \in \partial h(\dot{q}_t)$ for a time dependent *energy* functional $E(t, \cdot) : \mathbb{R}^d \to \mathbb{R}$ and where the *dissipation* potential $h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is 1-homogeneous. A main difficulty with rate independent processes is the possible appearance of jumps when the energy is non-convex. To deal with these jumps, a popular approach is to add a small viscosity regularisation, and to consider a convenient arc-length parametrisation which enables to keep track of the transition path between each jump and to analyse the solutions which arise when taking the viscosity term to zero [Efendiev and Mielke, 2006, Mielke et al., 2009, 2012]. We follow this approach to prove our main result.

2. Setup and preliminaries

2.1. Setup.

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

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

We do not make any assumptions concerning the number of samples n or the dimension d, however we recall the two classical settings which are covered by our results. The *underparametrised* setting, where $d \leq n$ and for which there exists a unique solution $\beta^\star = \arg\min_\beta L(\beta)$. The *overparametrised* setting, where $d > n$ and where there exists an infinite number of minimisers $\beta^\star \in \mathbb{R}^d$ which attain zero training loss and which correspond to linear predictors that perfectly fit the training data, *i.e.* $y_i = \langle \beta^\star, x_i \rangle$ for all $1 \leq i \leq n$. These predictors are referred to as *interpolators*. We assume throughout the papier that the inputs $(x_1, \dots, x_n)$ are in *general position*. In order to state the assumption, let $X \in \mathbb{R}^{n \times d}$ be the feature matrix whose i^{th} row is x_i and let us denote by $\tilde{x}_j \in \mathbb{R}^n$ it's j^{th} column for $j \in [d]$.

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

This assumption is standard in the Lasso framework as it ensures that the Lasso solution is unique for any regularisation parameter, in our case it similarly ensures that our saddle-to-saddle algorithm is well-defined as shown Proposition 2. Similarly, in the overparametrised setting, this assumption ensures that the minimum ℓ_1-norm interpolator is unique [Dossal, 2012, Theorem 2.2], we can therefore define $\beta^\star_{\ell_1} := \arg\min_{y_i = \langle x_i, \beta^\star \rangle, \forall i} \|\beta^\star\|_1$. Note that this assumption is not restrictive as it holds almost surely when the data is drawn from a continuous probability distribution [Tibshirani, 2013, Lemma 4].

2-layer diagonal linear network. We represent the regression vector β as a function β_w of a trainable parameter $w \in \mathbb{R}^p$. Despite the linearity of the final prediction function $x \mapsto \langle \beta_w, x \rangle$, the parametrisation significantly affects the training dynamics. In an effort to understand the training dynamics of neural networks, we consider a 2-layer diagonal linear neural network given by:

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

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

$$F(w) := L(u \odot v) = \frac{1}{2n} \sum_{i=1}^{n} (\langle u \odot v, x_i \rangle - y_i)^2 . \tag{3}$$

The optimisation problem is non-convex and highly non-trivial training dynamics occur.

Gradient Flow. We minimise the loss F using gradient flow:

$$\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t , \tag{4}$$

initialised at $u_0 = \sqrt{2}\alpha\mathbf{1} \in \mathbb{R}^d_{>0}$ with $\alpha > 0$, and $v_0 = \mathbf{0} \in \mathbb{R}^d$. This initialisation results in $\beta_0 = \mathbf{0} \in \mathbb{R}^d$ independently of the chosen neuron initialisation scale α. We denote $\beta^\alpha_t := u^\alpha_t \odot v^\alpha_t$ the prediction iterates generated from the gradient flow to highlight its dependency on the initialisation scale α[1].

Saddle points. As seen in Figure 1 for gradient flow with small initialisation scale, the iterates jump from a critical point of the loss, where the iterates barely make any progress, to another. The following proposition shows that these critical points are saddle points (*i.e.* not local extrema) and that they correspond to points which have a very particular structure as highlighted in Eq. (5). The proof is deferred to Appendix B.

Proposition 1. *All the critical points $w_c = (u_c, v_c)$ of F which are not global minima, i.e. $\nabla F(w_c) = \mathbf{0}$ and $F(w_c) > \min_w F(w)$, are necessarily saddle points (*i.e. *not local extrema). Each critical point maps to a parameter $\beta_c = u_c \odot v_c$ which satisfies $|\beta_c| \odot \nabla L(\beta_c) = \mathbf{0}$ and verifies:*

$$\beta_c \in \underset{\beta_i = 0 \text{ for } i \notin \mathrm{supp}(\beta_c)}{\arg\min} L(\beta), \tag{5}$$

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

Necessity of "accelerating" time. As the origin $\mathbf{0} \in \mathbb{R}^{2d}$ is a critical point of the function F, taking the initialisation $\alpha \to 0$ arbitrarily slows down the dynamics. In fact it can easily be shown for any fixed time t, that $(u^\alpha_t, v^\alpha_t) \to \mathbf{0}$ as $\alpha \to 0$. Therefore if we restrict ourselves to a finite time analysis, there is no hope of exhibiting the observed saddle-to-saddle behaviour. To do so, we must find an appropriate bijection $\tilde{t}_\alpha$ in $\mathbb{R}_{\geq 0}$ which "accelerates" time (*i.e.* $\tilde{t}_\alpha(t) \underset{\alpha \to 0}{\longrightarrow} +\infty$ for all t) and consider the accelerated iterates $\beta^\alpha_{\tilde{t}_\alpha(t)}$.

2.2. Leveraging the mirror flow structure.

It is shown in Azulay et al. [2021] that the iterates β^α_t follow a mirror flow with potential ϕ_α initialised at $\beta^\alpha_{t=0} = \mathbf{0}$:

$$\mathrm{d}\nabla\phi_\alpha(\beta^\alpha_t) = -\nabla L(\beta^\alpha_t)\mathrm{d}t, \tag{6}$$

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

$$\phi_\alpha(\beta) = \frac{1}{2} \sum_{i=1}^{d} \Big(\beta_i \mathrm{arcsinh}(\frac{\beta_i}{\alpha_i^2}) - \sqrt{\beta_i^2 + \alpha_i^4} + \alpha_i^2 \Big). \tag{7}$$

Unveiling the mirror flow structure enables to leverage convex optimisation tools to prove convergence of iterates to a global minimiser $\beta^\star_\alpha$. For overparametrised problems, the mirror formulation provides a simple proof of the associated implicit regularisation problem. As shown by Woodworth et al. [2020], the limit $\beta^\star_\alpha$ of the *gradient flow* is the solution of the following minimisation problem:

$$\beta^\star_\alpha = \underset{y_i = \langle x_i, \beta \rangle, \forall i}{\arg\min} \phi_\alpha(\beta) \tag{8}$$

Given the fact that ϕ_α behaves as the ℓ_1-norm as α goes to 0, it is shown in [Woodworth et al., 2020, Theorem 2] that $\beta^\star_\alpha$ converges to the minimum ℓ_1-norm interpolator as $\alpha \to 0$. We use the nice structure from Eq. (6) to bring to light the "saddle-to-saddle" dynamics which occurs as we take the initialisation to 0.

Taking the initialisation scale to 0, mirror point of view. As mentioned before, the iterates β^α_t stay stuck at $\mathbf{0}$ when $\alpha \to 0$. The mirror flow point of view sheds a new light on this observation. Differentiating the left side of equation 6 and computing the Hessian of ϕ_α, the iterates β^α_t are shown to follow $\dot{\beta}^\alpha_t =$

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

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

Data: $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$
$(t, \beta, s, r) \leftarrow (0, \mathbf{0}_d, \mathbf{0}_d, -\nabla L(\mathbf{0_d}))$;
while $r \neq \mathbf{0}$ **do**
    $\Delta \leftarrow \inf\{\delta > 0 \text{ s.t. } \exists i \in \{j \in [d], r(j) \neq 0\}, s(i) + \delta r(i) = \pm 1\}$
    $s \leftarrow s + \Delta \cdot r$
    $t \leftarrow t + \Delta$
    $I_\pm \leftarrow \{i \in [d] \text{ s.t. } s(i) = \pm 1\}$
    $\beta \leftarrow \underset{\substack{\beta_i \geq 0, i \in I_+ \\ \beta_i \leq 0, i \in I_- \\ \beta_i = 0, i \notin I_+ \cup I_-}}{\arg\min} \; L(\beta)$
    $r \leftarrow -\nabla L(\beta)$
end
Output: Successive values of β and t

$-\sqrt{(\beta_t^\alpha)^2 + \alpha^4} \odot \nabla L(\beta_t^\alpha)$. Taking $\alpha \to 0$, this flow is informally equivalent to $\dot{\beta}_t = -|\beta_t| \odot \nabla L(\beta_t)$ for which $\mathbf{0}$ is a fixed point as well as all vectors $\beta_c = \beta_{w_c}$ where w_c is a saddle point of F (following Proposition 1).

Time-rescaled iterates. Since for $\beta \neq \mathbf{0}$, $\|\nabla\phi_\alpha(\beta)\| \to \infty$ when $\alpha \to 0$, the potential ϕ_α becomes degenerate for small α. The formulation from Eq. (6) is therefore not well defined in the limit $\alpha \to 0$. We can nonetheless obtain a meaningful limit by considering the appropriate time acceleration $\tilde{t}_\alpha(t) = \ln(1/\alpha) \cdot t$ and looking at the accelerated iterates

$$\tilde{\beta}_t^\alpha := \beta_{\tilde{t}_\alpha(t)}^\alpha = \beta_{\ln(1/\alpha)t}^\alpha. \tag{9}$$

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

$$\mathrm{d}\nabla\tilde{\phi}_\alpha(\tilde{\beta}_t^\alpha) = -\nabla L(\tilde{\beta}_t^\alpha)\mathrm{d}t, \qquad \text{where} \qquad \tilde{\phi}_\alpha := \frac{1}{\ln(1/\alpha)} \cdot \phi_\alpha, \tag{10}$$

with $\tilde{\beta}_{t=0} = \mathbf{0}$ and where ϕ_α is defined Eq. (7). Also, in contrast with ϕ_α, the rescaled potential $\tilde{\phi}_\alpha$ is non-degenerate as α approaches 0. Indeed it satisfies $\tilde{\phi}_\alpha(\beta) \to \|\beta\|_1$ as $\alpha \to 0$ and its gradients satisfy $[\nabla\tilde{\phi}_\alpha(\beta)]_i \to \pm 1$ if $\beta_i \gtrless 0$ (see Lemma 2 in Appendix E). As a result, it is tempting to examine the limiting equation obtained when taking the limit $\alpha \to 0$ in Eq. (10).

Notations and terminology. In all generality we denote by $\beta_{\ell_1}^\star := \arg\min_{\beta^\star \in \arg\min_\beta L(\beta)} \|\beta^\star\|_1$ the minimum ℓ_1-norm solution, we point out that it simplifies to $\beta_{\ell_1}^\star := \arg\min_\beta L(\beta)$ in the underparametrised setting, and to $\beta_{\ell_1}^\star := \arg\min_{y_i = \langle x_i, \beta^\star\rangle, \forall i} \|\beta^\star\|_1$ in the overparametrised setting. For an integer d, we denote $[d]$ the set $\{1, \ldots, d\}$ and for $i \in [d]$ we denote $\{i\}^C$ the complementary set of $\{i\}$ in $[d]$. We abuse of the terminology of "saddle" when speaking of a point $\beta_c \in \mathbb{R}^d$ which **is not** an actual saddle of the loss L which is convex. We use it because it corresponds to a point $w_c \in \mathbb{R}^{2d}$ which is a saddle of the non-convex loss F.

3. Main result: limiting process

Our main result formalises the informal theorem given in the introduction. The limiting flow progresses by jumping from one saddle to another, where each saddle corresponds to the minimiser of the loss function under constraints on the set of active coordinates and their sign. The procedure for identifying and characterising these saddles as well as the jump times is given by a standalone algorithm outlined in Algorithm 1. We first verify that all the steps are well defined and that the algorithm terminates in a finite number of iterations in the following proposition. The proof is deferred to Appendix D.1.

Proposition 2. *Algorithm 1 is well defined: at each iteration (i) the attribution of Δ is well defined as $\Delta < +\infty$, (ii) the constrained minimisation problem has a unique solution and the attribution of the value of β is therefore well-founded. Furthermore, along the loops: the iterates β have at most n non-zero coordinates, the loss is strictly decreasing and the algorithm terminates in at most $\min(2^d, \sum_{k=0}^n \binom{d}{k})$ steps by outputting the minimum ℓ_1-norm solution $\beta_{\ell_1}^\star$.*

Proposition 2 highlights that Algorithm 1 is on its own an algorithm of interest for finding the minimum ℓ_1-norm solution in an overparametrised regression setting. We point out that the provided upperbound on the number of iterations is very crude and could certainly be improved. However, analysing Algorithm 1 is not the focus of our paper.

We can now state our main result. We show that at vanishing initialisation, the limiting flow progresses by jumping between different saddles, where each saddle corresponds to a minimiser of the loss function subject to specific sign constraints.

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

$$\textit{(“Saddles”)} \qquad \tilde{\beta}^\circ_t \equiv \beta_k \qquad \text{for } t \in (t_k, t_{k+1}) \text{ and } 0 \leq k \leq p, \;\; t_{p+1} = +\infty.$$

The accelerated flow $(\tilde{\beta}^\alpha_t)_t$ defined in Eq. (9) uniformly converges towards the limiting process $(\tilde{\beta}^\circ_t)_t$ on any compact subset of $\mathbb{R}_{\geq 0} \backslash \{t_1, \ldots, t_p\}$.

Behaviour of the saddle-to-saddle process. The visited "saddles" are entirely provided in a recursive manner by Algorithm 1. Unlike all previous results on incremental learning, complex behaviours can occur when the feature matrix is ill designed and the RIP property [Candès et al., 2006] is not satisfied. Several coordinates can activate and deactivate at the same time (in Figure 1 (Right), 4 coordinates deactivate at the same time, in Figure 2 (Right), two coordinates activate at the same time). Moreover, there is no monotonicity of the total number of non-zero coordinates and each coordinate can increase, decrease and change sign many times.

Convergence result. First we point out that we could not expect uniform convergence of $(\tilde{\beta}^\alpha_t)$ on intervals of the form $[0, T]$ given the fact that the limit process is discontinuous. Hence the uniform convergence outside of the discontinuity times is one of the strongest we could expect. In Corollary 2 in the following section we give an even stronger result by showing a graph convergence of the iterates. We also highlight that showing the convergence to a limiting process is in fact the toughest challenge from a theoretical point of view and is done Section 4. However, in Section 3.1, assuming its existence, we show that constructing $\tilde{\beta}^\circ$ is rather intuitive and naturally leads to Algorithm 1.

Estimate for the iterates β^α_t. We point out that our result provides no speed of convergence of $\tilde{\beta}^\alpha$ towards $\tilde{\beta}^\circ$. We believe that a non-asymptotic result is challenging and leave it as future work. We experimentally notice that the convergence rate quickly degrades after each saddle. Nonetheless, we can still write for the "non-accelerated" iterates that $\beta^\alpha_t = \tilde{\beta}^\alpha_{t/\ln(1/\alpha)} \sim \tilde{\beta}^\circ_{t/\ln(1/\alpha)}$ when $\alpha \to 0$. This approximation tells us for α small enough that the iterates β^α_t are roughly equal to 0 until time $t_1 \cdot \ln(1/\alpha)$ and that the minimum ℓ_1-norm interpolator is reached at time $t_p \cdot \ln(1/\alpha)$. Such a precise estimate of the global convergence time is remarkable and goes beyond classical Lyapunov analyses which only leads to $L(\beta^\alpha_t) \lesssim \ln(1/\alpha)/t$ (see Proposition 4 in Appendix C).

"Neural" point of view. We can map back our result to the accelerated flow $\tilde{w}^\alpha_t := w^\alpha_{\tilde{t}^\alpha(t)}$. Indeed there is a bijective mapping between the mirror flow $\tilde{\beta}^\alpha_t$ and the gradient flow as shown in Lemma 1 in the appendix. From there we can show that $(\tilde{u}^\alpha_t, \tilde{v}^\alpha_t) \underset{\alpha \to 0}{\longrightarrow} (\sqrt{|\tilde{\beta}^\circ_t|}, \mathrm{sign}(\tilde{\beta}^\circ_t)\sqrt{|\tilde{\beta}^\circ_t|})$ uniformly on any compact subset of $\mathbb{R}_{\geq 0} \backslash \{t_1, \ldots, t_p\}$.

Justification of the saddle-to-saddle terminology. We emphasise that the β_k's which we refer to as "saddles" **are not** saddles of the convex loss L. We use this terminology as the mapping $(u_k, v_k) = (\sqrt{|\beta_k|}, \mathrm{sign}(\beta_k)\sqrt{|\beta_k|})$ results in actual saddle points of the non-convex loss function F.

3.1. Intuitive construction of the limiting process.

In this section we provide some intuition on how the limiting flow $(\tilde{\beta}^\circ_t)_t$ is constructed. To do so, the subdifferential of the ℓ_1-norm turns out very useful. We recall its definition:

$$\partial\|\tilde{\beta}\|_1 = \bullet\ \{1\} \text{ if } \tilde{\beta} > 0 \quad \bullet\ \{-1\} \text{ if } \tilde{\beta} < 0 \quad \bullet\ [-1, 1] \text{ if } \tilde{\beta} = 0. \tag{11}$$

Recall the accelerated mirror flow Eq. (9) satisfies Eq. (10) which can be integrated as:

$$-\int_0^t \nabla L(\tilde{\beta}_s^\alpha)\mathrm{d}s = \nabla\tilde{\phi}_\alpha(\tilde{\beta}_t^\alpha). \tag{12}$$

From this equation, the complicated part is to show that the iterates $\tilde{\beta}_t^\alpha$ indeed converge to some piecewise constant process. We show this convergence in the following section. The goal here is to provide some intuition on why Algorithm 1 describes the jumps times and the visited saddles. Assuming that a limiting process indeed exists, *i.e.* $\lim_{\alpha\to 0}\tilde{\beta}_t^\alpha =: \tilde{\beta}_t$ exists for all t, the dominated convergence theorem then yields that the process $(\tilde{\beta}_t)_t$ must satisfy:

$$-\int_0^t \nabla L(\tilde{\beta}_s)\mathrm{d}s \in \partial\|\tilde{\beta}_t\|_1, \tag{13}$$

Indeed, by the definition of $\tilde{\phi}_\alpha$, $\nabla\tilde{\phi}_\alpha(\tilde{\beta}_t^\alpha)$ must converge to an element of $\partial\|\tilde{\beta}_t\|_1$ (see Lemma 2 in the appendix). We start by providing a few comments concerning Eq. (13).

Links with Lasso. Notice that Eq. (13) closely resembles to the Lasso optimality condition:

$$\beta_\lambda^\star = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d}\ L(\beta) + \lambda\|\beta\|_1. \tag{14}$$

Indeed, the optimality condition of Eq. (14) writes $\nabla L(\beta_\lambda) \in \lambda\partial\|\beta_\lambda\|_1$, and is almost the same as Eq. (13) with $\lambda = 1/t$. For the quadratic loss, the trajectory $(\beta_\lambda^\star)_{\lambda\geq 0}$ is a piecewise linear path with a discrete number of vertices between $\beta_{\lambda\to\infty}^\star = 0$ and $\beta_{\lambda\to 0}^\star = \beta_{\ell_1}^\star$. This path can be computed using the Homotopy method [see, e.g., Tibshirani, 2013, and references therein] which operates in an iterative fashion and finds each vertex starting from $\mathbf{0}$.

Subdifferential equations and rate-independant systems. Similar subdifferential inclusions of the form $\nabla L(\beta_t) \in \frac{\mathrm{d}}{\mathrm{d}t}\partial h(\beta_t)$ for non-differential functions h have been studied by Attouch et al. [2004] but for strongly convex functions h. In this case, the solutions are continuous, do not exhibit jumps and we cannot leverage their tools. On another hand, Efendiev and Mielke [2006], Mielke et al. [2009, 2012] consider rate-independent systems of the form $\partial_q E(t, q_t) \in \partial h(\dot{q}_t)$ for 1-homogeneous functions h. Though this formulation is quite different from Eq. (12), its solutions can have jumps and we leverage their tools to show our result. This is detailed in Section 4.

3.2. Intuitive proof of Theorem 2.

We provide here an intuitive sketch of the proof of Theorem 2 assuming that the limiting process exists. Note that this **is not** a rigorous proof but has the advantage of preserving intuitive comprehension (the full proof is a consequence of Theorem 3 and can be found Appendix D.3). The idea is similar to the Homotopy algorithm which is used to find the Lasso path: we start from $\mathbf{0}$ and successively look at the breaking times of Eq. (13) and determine the unique update of $\tilde{\beta}_t$ that keeps the conditions satisfied. Let us denote $s_t := -\int_0^t \nabla L(\tilde{\beta}_s)\mathrm{d}s$. The function s_t is therefore continuous and, as noted in Eq. (13), satisfies $s_t \in \partial\|\tilde{\beta}_t\|_1$.

First saddle: $[0, t_1)$**,** $\tilde{\beta}_t = \beta_0 \equiv \mathbf{0}$**.** While $\|s_t\|_\infty < 1$, we must have $\tilde{\beta}_t = \mathbf{0}$ from Eq. (13) and therefore $s_t = -t\cdot\nabla L(\beta_0 = \mathbf{0})$. Now notice that t_1 defined in Algorithm 1 corresponds to the time such that $\|s_{t_1}\|_\infty = \pm 1$. The iterates must then move: indeed assume that $\tilde{\beta}_t$ is still $\mathbf{0}$ for $t \in [t_1, t_1+\varepsilon]$, then we would have $\|s_t\|_\infty > 1$ which contradicts Eq. (13). For simplicity, assume that there is a unique coordinate i_1 such that $|s_{t_1}(i_1)| = 1$ and without loss of generality assume that $s_{t_1}(i_1) = +1$.

Jump at time t_1**.** We show that there must be a discontinuous jump at time t_1. To understand this claim, we consider $\tilde{\beta}_t^\alpha$ for very small α. Differentiating the left side from Eq. (10) and computing the Hessian of $\tilde{\phi}_\alpha$, we obtain that $\mathrm{d}\tilde{\beta}_t^\alpha = -\ln(1/\alpha)\sqrt{\tilde{\beta}_t^\alpha + \alpha^4} \odot \nabla L(\tilde{\beta}_t^\alpha) \sim -\ln(1/\alpha)|\tilde{\beta}_t^\alpha| \odot \nabla L(\tilde{\beta}_t^\alpha)$ $(*)$. Since t_1 corresponds to the time the iterates leave the first saddle, $\tilde{\beta}_{t_1}^\alpha$ cannot converge to a fixed point of $\dot{\beta}_t = -|\beta_t| \odot \nabla L(\beta_t)$. Therefore $(*)$ entails $\|\dot{\tilde{\beta}}_t^\alpha\| \to +\infty$ as $\alpha \to 0$, which translates into a jump. Following the jump, the velocity should return to a bounded state and $(*)$ implies that the iterates can only stick to another saddle where $|\beta| \odot \nabla L(\beta) = \mathbf{0}$.

Second saddle: (t_1, t_2)**.** The iterates have jumped to another saddle point which we denote β_1, the question is now to determine which precise saddle it is. First note that by continuity of s_t at time t_1 and because $\|s_{t_1}(\{i_1\}^C)\|_\infty < 1$ we must have $\beta_1(\{i_1\}^C) = \mathbf{0}$ from Eq. (13). Furthermore, since we assumed that $s_{t_1}(i_1) = 1$, we must have that $\beta_1(i_1) \geq 0$. Therefore, we necessary have

$$\beta_1 = \underset{\substack{\beta(\{i_1\}^C)=\mathbf{0} \\ \beta(i_1)\geq 0}}{\arg\min} \; L(\beta), \tag{15}$$

which matches the first loop of Algorithm 1. Then $\tilde{\beta}_t$ stays constant at β_1 until time t_2 when an inactive coordinate i_2 is such that $s_{t_2}(i_2) = \pm 1$.

Following saddles. We can recursively follow the previous logic to compute the jump times and the following saddles. However one must be careful in not forgetting the sign constraints. Indeed a coordinate cannot jump from a strictly positive value to a strictly negative value since crossing 0 always corresponds to a fixed point of the dynamics. This feature also appears in Eq. (13): the function s_t, being continuous at the jump times, cannot pass from a value of $+1$ to -1 instantly.

3.3. Additional comments.

Comparisons to the Homotopy and OMP algorithms. The jump times as well as the visited saddles depend on all the previously visited saddles, and not just on the last one. This behaviour is in stark contrast with the Homotopy algorithm for the Lasso. On a separate note, a common belief is that in the vicinity of a saddle point, the following activated coordinate should correspond to the direction of most negative curvature (i.e. eigenvector corresponding to the most negative eigenvalue). However, this statement cannot be accurate as it is inconsistent with our algorithm (though it holds true for the first coordinate). In fact, it can be shown that by selecting this particular active coordinate the resulting algorithm aligns with the orthogonal matching pursuit (OMP) algorithm [Pati et al., 1993, Davis et al., 1997] which does not necessarily lead to the minimum ℓ_1-norm interpolator.

Natural extensions of our setting. More general initialisations than $u_{t=0} = \sqrt{2}\alpha\mathbf{1}$ can easily be dealt with. For instance, initialisations of the form $u_{t=0} = \alpha\mathbf{u_0} \in \mathbb{R}^d$ lead to the exact same result as it is shown in Woodworth et al. [2020] (Discussion after Theorem 1) that the associated mirror still converges to the ℓ_1-norm. Initialisations of the form $[u_{t=0}]_i = \alpha^{k_i}$, where $k_i > 0$, lead to the associated potential converging towards a weighted ℓ_1-norm and one should modify Algorithm 1 by accordingly weighting $\nabla L(\beta)$ in the algorithm. Similarly, deeper linear architectures of the form $\beta_w = w_+^D - w_-^D$ as in Woodworth et al. [2020] do not change our result as the associated mirror converges towards the ℓ_1-norm too. Only the square loss is considered in the paper, however we believe that all our results should hold for any loss of the type $L(\beta) = \sum_{i=1}^n \ell(y_i, \langle x_i, \beta\rangle)$ where for all $y \in \mathbb{R}$, $\ell(y, \cdot)$ is strictly convex with a unique minimiser at y. In fact, the only property which cannot directly be adapted from our results is showing the uniform boundedness of the iterates (see discussion before Proposition 5 in Appendix C).

4. Sketch of proof and arc-length parametrisation

In this section, we consider a new time reparametrisation which circumvents the apparition of discontinuous jumps and leads to the proof of Theorem 2. The main difficulty stems from the non-continuity of the limit process $\tilde{\beta}^\circ$. Therefore we cannot expect uniform convergence of $\tilde{\beta}^\alpha$ towards $\tilde{\beta}$ as $\alpha \to 0$. In addition, $\tilde{\beta}^\circ$ does not provide any insights into the path followed between the jumps.

Arc-length parametrisation. The high-level idea is to "slow-down" time when the jumps occur. To do so we follow the approach from Efendiev and Mielke [2006], Mielke et al. [2009] and we consider an arc-length parametrisation of the path, i.e., we consider τ^α equal to:

$$\tau^\alpha(t) = t + \int_0^t \|\dot{\tilde{\beta}}_s^\alpha\| \mathrm{d}s.$$

In Proposition 6 in the appendix, we show that the full path length $\int_0^{+\infty} \|\dot{\tilde{\beta}}_s^\alpha\| \mathrm{d}s$ is finite and bounded independently of α. Therefore τ^α is a bijection in $\mathbb{R}_{\geq 0}$. We can then define the following quantities:

$$\hat{t}_\tau^\alpha = (\tau^\alpha)^{-1}(\tau) \quad \text{and} \quad \hat{\beta}_\tau^\alpha = \tilde{\beta}_{\hat{t}^\alpha(\tau)}^\alpha.$$

By construction, a simple chain rule leads to $\dot{\hat{t}}^\alpha(\tau) + \|\dot{\hat{\beta}}_\tau^\alpha\| = 1$, which means that the speed of $(\hat{\beta}_\tau^\alpha)_\tau$ is always upperbounded by 1, independently of α. This behaviour is in stark contrast with the process $(\tilde{\beta}_t^\alpha)_t$ which has a speed which explodes at the jumps. It presents a major advantage as we can now use

Arzelà-Ascoli's theorem to extract a converging subsequent. A simple change of variable shows that the new process satisfies the following equations:

$$-\int_0^\tau \dot{\hat{t}}_s^\alpha \nabla L(\hat{\beta}_s^\alpha)\mathrm{d}s = \nabla\tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha) \quad \text{and} \quad \dot{\hat{t}}_\tau^\alpha + \|\dot{\hat{\beta}}_\tau^\alpha\| = 1 \tag{16}$$

started from $\hat{\beta}_\tau^\alpha = 0$ and $\hat{t}_0 = 0$. The next proposition states the convergence of the rescaled process, up to a subsequence. The proof is deferred to Appendix D.2

Proposition 3. *Let $T \geq 0$. For every $\alpha > 0$, let $(\hat{t}^\alpha, \hat{\beta}^\alpha)$ be the solution of Eq. (16). Then, there exists a subsequence $(\hat{t}^{\alpha_k}, \hat{\beta}^{\alpha_k})_{k\in\mathbb{N}}$ and $(\hat{t}, \hat{\beta})$ such that as $\alpha_k \to 0$:*

$$(\hat{t}^{\alpha_k}, \hat{\beta}^{\alpha_k}) \to (\hat{t}, \hat{\beta}) \qquad \text{in } (C^0([0,T], \mathbb{R}\times\mathbb{R}^d), \|\cdot\|_\infty) \tag{17}$$

$$(\dot{\hat{t}}^{\alpha_k}, \dot{\hat{\beta}}^{\alpha_k}) \rightharpoonup (\dot{\hat{t}}, \dot{\hat{\beta}}) \qquad \text{in } L_1[0,T] \tag{18}$$

Limiting dynamics. *The limits $(\hat{t}, \hat{\beta})$ satisfy:*

$$-\int_0^\tau \dot{\hat{t}}_s \nabla L(\hat{\beta}_s)\mathrm{d}s \in \partial\|\hat{\beta}_\tau\|_1 \quad \text{and} \quad \dot{\hat{t}}_\tau + \|\dot{\hat{\beta}}_\tau\| \leq 1 \tag{19}$$

Heteroclinic orbit. *In addition, when $\hat{\beta}_\tau$ is such that $|\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau) \neq 0$, we have*

$$\dot{\hat{\beta}}_\tau = -\frac{|\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)}{\||\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)\|} \quad \text{and} \quad \dot{\hat{t}}_\tau = 0. \tag{20}$$

Furthermore, the loss strictly decreases along the heteroclinic orbits and the path length $\int_0^T \|\dot{\hat{\beta}}_\tau\|\mathrm{d}\tau$ is upperbounded independently of T.

The proof can be found in Appendix D and relies on the Arzelà-Ascoli theorem. Borrowing terminologies from Efendiev and Mielke [2006], we can distinguish two regimes: when $\dot{\hat{\beta}}_\tau = 0$, the system is *sticked* to the saddle point. When $\dot{\hat{t}}_\tau = 0$ and $\|\dot{\hat{\beta}}_\tau\| = 1$ the system switches to a *viscous slip* which follows the normalised flow Eq. (20). We use the term of *heteroclinic orbit* as in the dynamical systems literature since in the neuron space (u, v) it corresponds to a path with links two distinct critical points of the loss F. Since $\dot{\hat{t}}_\tau = 0$, this regime happens instantly for the original t time scale (*i.e.* a jump occurs).

From Proposition 3, following the same reasoning as in Section 3.1, we can show that the rescaled process converges uniformly to a continuous saddle-to-saddle process where the saddles are linked by normalized flows.

Theorem 3. *Let $T > 0$. For all subsequences defined in Proposition 3, there exist times $0 = \tau_0' < \tau_1 < \tau_1' < \cdots < \tau_p < \tau_p' < \tau_{p+1} = +\infty$ such that the the iterates $(\hat{\beta}_\tau^{\alpha_k})_\tau$ converge uniformly on $[0, T]$ to the following limit trajectory :*

("Saddle") $\qquad \hat{\beta}_\tau = \beta_k \qquad$ *for $\tau \in [\tau_k', \tau_{k+1}]$ where $0 \leq k \leq p$*

(Orbit) $\qquad \dot{\hat{\beta}}_\tau = -\dfrac{|\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)}{\||\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)\|} \qquad$ *for $\tau \in [\tau_{k+1}, \tau_{k+1}']$ where $0 \leq k \leq p-1$*

where the saddles $(\beta_0 = 0, \beta_1, \ldots, \beta_p = \beta_{\ell_1}^\star)$ are constructed in Algorithm 1. Also, the loss $(L(\hat{\beta}_\tau))_\tau$ is constant on the saddles and strictly decreasing on the orbits. Finally, independently of the chosen subsequence, for $k \in [p]$ we have $\hat{t}_{\tau_k} = \hat{t}_{\tau_k'} = t_k$ where the times $(t_k)_{k\in[p]}$ are defined through Algorithm 1.

Proof. Some parts of the proof are slightly technical. To simplify the understanding, we make use of auxiliary lemmas which are stated in Appendix E. The overall spirit follows the sketch of proof given after Theorem 2 and relies on showing that Eq. (19) can only be satisfied if the iterates visit the saddles from Algorithm 1.

We let $\hat{s}_\tau := -\int_0^\tau \dot{\hat{t}}_s \nabla L(\hat{\beta}_s)\mathrm{d}s$, which is continuous and satisfies $\hat{s}_\tau \in \partial\|\hat{\beta}_\tau\|_1$ from Eq. (19). Let $S = \{\beta \in \mathbb{R}^d, |\beta|\odot\nabla L(\beta) = \mathbf{0}\}$ denote the set of critical points and let (β_k, t_k, s_k) be the successive values of (β, t, s) which appear in the loops of Algorithm 1.

We do a proof by induction: we start by assuming that the iterates are stuck at the saddle β_{k-1} at time $\tau \geq \tau_{k-1}'$ where $\hat{t}_{\tau_{k-1}'} = t_{k-1}$ and $\hat{s}_{\tau_{k-1}'} = s_{k-1}$ (recurrence hypothesis), we then show that they can only move at a time τ_k and follow the normalised flow Eq. (20). We finally show that they must end up "stuck" at the new critical point β_k, validating the recurrence hypothesis.

Proof of the jump time τ_k such that $\hat{t}_{\tau_k} = t_k$: we set ourselves at time $\tau \geq \tau'_{k-1}$, stuck at the saddle β_{k-1}. Let $\tau_k := \sup\{\tau, \hat{t}_\tau \leq t_k\}$, we have that $\tau_k < \infty$ from Lemma 3 in the Appendix. Note that by continuity of $\hat{t}_\tau$ it holds that $\hat{t}_{\tau_k} = t_k$. Now notice that $\hat{s}_\tau = \hat{s}_{\tau'_{k-1}} - (\hat{t}_\tau - \hat{t}_{\tau'_{k-1}})\nabla L(\beta_{k-1}) = s_{k-1} - (\hat{t}_\tau - t_{k-1})\nabla L(\beta_{k-1})$. We argue that for any $\varepsilon > 0$, we cannot have $\hat{\beta}_\tau = \beta_{k-1}$ on $(\tau_k, \tau_k + \varepsilon)$. Indeed by the definition of τ_k and from the algorithmic construction of time t_k, it would lead to $|\hat{s}_\tau(i)| > 1$ for some coordinate $i \in [d]$, which contradicts Eq. (19). Therefore the iterates must move at the time τ_k.

Heterocline leaving β_{k-1} for $\tau \in [\tau_k, \tau'_k]$: contrary to before, our time rescaling enables to capture what happens during the "jump". We have shown that for any ε, there exists $\tau_\varepsilon \in (\tau_k, \tau_k + \varepsilon)$, such that $\hat{\beta}_{\tau_\varepsilon} \neq \beta_{k-1}$. From Lemma 4, since the saddles are distinct along the flow, we must have that $\hat{\beta}_{\tau_\varepsilon} \notin S$ for ε small enough. The iterates therefore follow a heterocline flow leaving β_{k-1} with a speed of 1 given by Eq. (20). We now define $\tau'_k := \inf\{\tau > \tau_k, \exists \varepsilon_0 > 0, \forall \varepsilon \in [0, \varepsilon_0],\ \hat{\beta}_{\tau+\varepsilon} \in S\}$ which corresponds to the time at which the iterates reach a new critical point and stay there for at least a small time ε_0. We have just shown that $\tau'_k > \tau_k$. Now from Proposition 3, the path length of $\hat{\beta}$ is finite, and from Lemma 4 the flow visits a finite number of distinct saddles at a speed of 1. These two arguments put together, we get that $\tau'_k < +\infty$ and also $\hat{\beta}_{\tau'_k+\varepsilon} = \hat{\beta}_{\tau'_k}, \forall \varepsilon \in [0, \varepsilon_0]$. On another note, since $\dot{\hat{t}}_\tau = 0$ for $\tau \in [\tau_k, \tau'_k]$ we have $\hat{t}_{\tau'_k} = \hat{t}_{\tau_k} (= t_k)$ as well as $\hat{s}_{\tau_k} = \hat{s}_{\tau'_k} (= s_k)$.

Proof of the landing point β_k : we now want to find to which saddle $\hat{\beta}_{\tau'_k} \in S$ the iterates have moved to. To that end, we consider the following sets which also appear in Algorithm 1:

$$I_{\pm,k} := \{i \in \{1, \dots, d\}, \text{ s.t. } \hat{s}_{\tau'_k}(i) = \pm 1\} \quad \text{and} \quad I_k = I_{+,k} \cup I_{-,k}. \tag{21}$$

The set I_k corresponds to the coordinates of $\hat{\beta}_{\tau'_k}$ which "are allowed" (but not obliged) to be activated (*i.e.* non-zero). For $\tau \in [\tau'_k, \tau'_k + \varepsilon_0]$ we have that $\hat{s}_\tau = \hat{s}_{\tau'_k} - (\hat{t}_\tau - t_k)\nabla L(\hat{\beta}_{\tau'_k})$. By continuity of $\hat{s}$ and the fact that $\hat{s}_\tau \in \partial \|\hat{\beta}_{\tau'_k}\|_1$, the equality translates into: • if $i \notin I_k$, $\hat{\beta}_{\tau'_k}(i) = 0$, • if $i \in I_{+,k}$, then $[\nabla L(\hat{\beta}_{\tau'_k})]_i \geq 0$ and $\hat{\beta}_{\tau'_k}(i) \geq 0$, • if $i \in I_{-,k}$, then $[\nabla L(\hat{\beta}_{\tau'_k})]_i \leq 0$ and $\hat{\beta}_{\tau'_k}(i) \leq 0$ and finally • for $i \in I_k$, if $\hat{\beta}_{\tau'_k}(i) \neq 0$, then $[\nabla L(\hat{\beta}_{\tau'_k})]_i = 0$. One can then notice that these conditions exactly correspond to the optimality conditions of the following constrained minimisation problem:

$$\underset{\substack{\beta_i \geq 0, i \in I_{k,+} \\ \beta_i \leq 0, i \in I_{k,-} \\ \beta_i = 0, i \notin I_k}}{\arg\min} \quad L(\beta). \tag{22}$$

We showed in Proposition 2 that the solution to this problem is unique and equal to β_k from Algorithm 1. Therefore $\hat{\beta}_\tau = \beta_k$ for $\tau \in [\tau'_k, \tau'_k + \varepsilon_0]$. It finally remains to show that $\hat{\beta}_\tau = \beta_k$ while $\tau \leq \tau_{k+1}$, where $\tau_{k+1} := \sup\{\tau, \hat{t}_\tau = t_{k+1}\}$. For this let $\tau \in [\tau'_k, \tau_{k+1}]$, notice that for $i \notin I_k$, we necessarily have that $\hat{\beta}_\tau(i) = \beta_k(i) = 0$, otherwise we break the continuity of $\hat{s}_\tau$. Similarly, for $i \in I_{k,+}$, we necessarily have that $\hat{\beta}_\tau(i) \geq 0$ and for $i \in I_{k,-}$, $\hat{\beta}_\tau(i) \leq 0$ for the same continuity reasons. Now assume that $\hat{\beta}_\tau(I_k) \neq \beta_k(I_k)$. Then from Lemma 4 and continuity of the flow, $\exists \tau' \in (\tau'_k, \tau)$ such that $\hat{\beta}_{\tau'} \notin S$ and there must exist a heterocline flow Eq. (20) starting from β_k which passes through $\beta_{\tau'}$. This is absurd since along this flow the loss strictly decreases, which is in contradiction with the definition of β_k which minimises the problem Eq. (22). □

Theorem 3 enables to prove without difficulty Theorem 2. Indeed we can show that any extracted limit $\hat{\beta}$ maps back to the unique discontinuous process $\tilde{\beta}^\circ$. We refer to Appendix D.3 for the full proof of Theorem 2.

Graph convergence. A nice and pictorial consequence of Theorem 3 is that the graph of the iterates $(\tilde{\beta}^\alpha_t)_t$ converges towards that of $(\hat{\beta}_\tau)_\tau$.

Corollary 1. *For all $T > t_p$, the graph of the iterates $(\tilde{\beta}^\alpha_t)_{t \leq T}$ converges to that of $(\hat{\beta}_\tau)_\tau$:*

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

where $\mathrm{dist}(\cdot, \cdot)$ *corresponds to the Hausdorff distance between 2 sets.*

Unlike Theorem 2, the convergence result presented in Corollary 2 allows to track the path followed between the saddles.

Conclusion

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

Acknowledgments

S.P. would like to thank Quentin Rebjock, Loucas Pillaud-Vivien and Etienne Boursier for the helpful and enlightening discussions. The authors also thank Jérôme Bolte for the discussions concerning subdifferential equations, Aris Daniilidis for the reference of Kurdyka [1998], as well as Aditya Varre for proofreading the paper.

References

Ehsan Amid and Manfred K. Warmuth. Winnowing with gradient descent. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 163–182. PMLR, 09–12 Jul 2020.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter Ashwin and Michael Field. Heteroclinic networks in coupled cell systems. *Arch. Ration. Mech. Anal.*, 148(2):107–143, 1999.

H. Attouch, J. Bolte, P. Redont, and M. Teboulle. Singular Riemannian barrier methods and gradient-projection dynamical systems for constrained optimization. *Optimization*, 53(5-6):435–454, 2004.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 468–477. PMLR, 18–24 Jul 2021.

Yuri Bakhtin. Noisy heteroclinic networks. *Probab. Theory Related Fields*, 150(1-2):1–42, 2011.

Raphaël Berthier. Incremental learning in diagonal linear networks. *arXiv preprint arXiv:2208.14673*, 2022.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow reLU networks for square loss and orthogonal inputs. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1305–1338. PMLR, 09–12 Jul 2020.

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

Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13:57–98, 1997.

Charles Dossal. A necessary and sufficient condition for exact sparse recovery by ℓ_1 minimization. *C. R. Math. Acad. Sci. Paris*, 350(1-2), 2012.

Messoud A. Efendiev and Alexander Mielke. On the rate-independent limit of systems with dry friction and small viscosity. *J. Convex Anal.*, 13(1):151–167, 2006.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s)gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*, 2023.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 386–407. PMLR, 08 Feb–11 Feb 2020.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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 *Advances in Neural Information Processing Systems*, volume 31, 2018.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Liwei Jiang, Yudong Chen, and Lijun Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *arXiv preprint arXiv:2203.02839*, 2022.

Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon S Du, and Jason D Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. *arXiv preprint arXiv:2301.11500*, 2023.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

M. Krupa. Robust heteroclinic cycles. *J. Nonlinear Sci.*, 7(2):129–176, 1997.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)*, 48(3):769–783, 1998.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

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

Alexander Mielke. Evolution of rate-independent systems. *Evolutionary equations*, 2:461–559, 2005.

Alexander Mielke, Riccarda Rossi, and Giuseppe Savaré. Modeling solutions with jumps for rate-independent systems on metric spaces. *Discrete Contin. Dyn. Syst.*, 25(2):585–615, 2009.

Alexander Mielke, Riccarda Rossi, and Giuseppe Savaré. Variational convergence of gradient flows and rate-independent evolutions in metric spaces. *Milan Journal of Mathematics*, 80:381–410, 2012.

Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 65–74. PMLR, 20–22 Apr 2017.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, 2021.

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pages 8913–8924. PMLR, 2021.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Stat.*, 7:1456–1490, 2013.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Conference on Learning Theory*, pages 2127–2159. PMLR, 2022.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Organisation of the Appendix.

(1) In Appendix A, we provide additional experiments as well as the experimental setup
(2) In Appendix B, we prove Proposition 1 and provide additional comments concerning the unicity of the minimisation problem which appears in the proposition
(3) In Appendix C, we provide some general results on the flow
(4) In Appendix D, we prove Proposition 2, Proposition 3, Theorem 2 as well as Corollary 2
(5) In Appendix E, we provide technical lemmas which are useful to prove the main results

Appendix A. Additional experiments and experimental setup

Experimental setup. For each experiment we generate our dataset as $y_i = \langle x_i, \beta^\star \rangle$ where $x_i = \mathcal{N}(\mathbf{0}, H)$ for a a diagonal covariance matrix H and $\beta^\star$ is a vector of $\mathbb{R}^d$. Gradient descent is run with a small step size and from initialisation $u_{t=0} = \sqrt{2}\alpha \mathbf{1} \in \mathbb{R}^d$ and $v_{t=0} = \mathbf{0}$ for some initialisation scale $\alpha > 0$.

- Figure 1 (Left): $(n, d, \alpha) = (5, 7, 10^{-120})$, $H = I_d$, $\beta^\star = (10, 20, 0, 0, 0, 0, 0) \in \mathbb{R}^7$.
- Figure 1 (Right): $(n, d, \alpha) = (6, 6, 10^{-10})$, $H = \mathrm{diag}(1, 10, 10, 10, 10, 10) \in \mathbb{R}^{6\times 6}$, $\beta^\star = (10, 20, 0, 0, 0, 0, 0) \in \mathbb{R}^6$.
- Figure 2 (Left): $(n, d, \alpha_1, \alpha_2) = (7, 2, 10^{-100}, 10^{-10})$, $H = I_d$, $\beta^\star = (10, 20) \in \mathbb{R}^7$.
- Figure 2 (Right): $(n, d, \alpha) = (3, 3, 10^{-100})$, X is the square root matrix of the matrix $((20, 6, -1.4), (6, 2, -0.4), (-1.4, -0.4, 0.12)) \in \mathbb{R}^{3\times 3}$, $\beta^\star = (1, 9, 10)$.

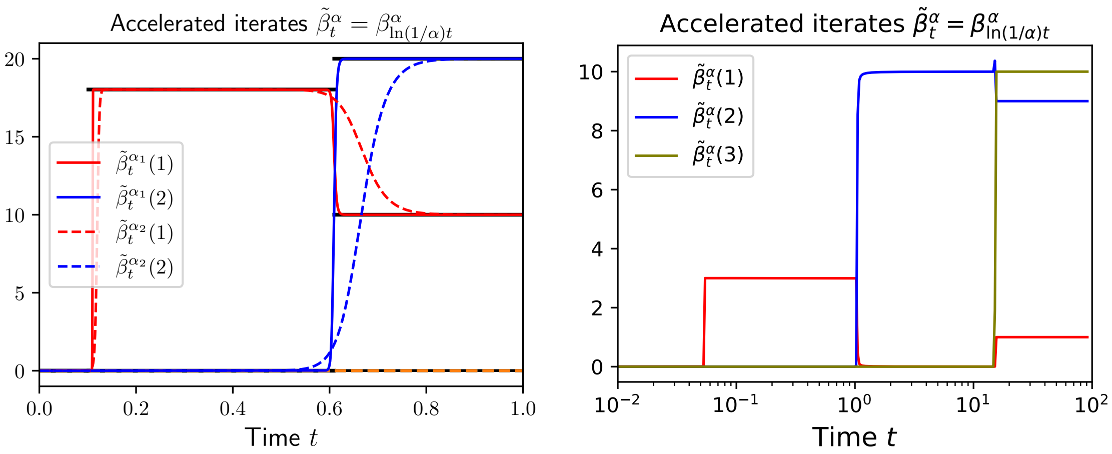

FIGURE 2. *Left:* Visualisation of the uniform convergence of $\tilde\beta^\alpha$ towards $\tilde\beta^\circ$ as $\alpha \to 0$. $\alpha_1 = 10^{-100} \ll \alpha_2 = 10^{-10}$ *Right:* In some cases, 2 coordinates can activate at the same time. Note that the time axis is in log-scale for better visualisation.

Appendix B. Proof of Proposition 1

Proposition 1. *All the critical points $w_c = (u_c, v_c)$ of F which are not global minima, i.e. $\nabla F(w_c) = \mathbf{0}$ and $F(w_c) > \min_w F(w)$, are necessarily saddle points (*i.e. *not local extrema). Each critical point maps to a parameter $\beta_c = u_c \odot v_c$ which satisfies $|\beta_c| \odot \nabla L(\beta_c) = \mathbf{0}$ and verifies:*

$$\beta_c \in \underset{\beta_i = 0 \text{ for } i \notin \mathrm{supp}(\beta_c)}{\arg\min} L(\beta), \tag{5}$$

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

Proof. **Non-existence of maxima / non-global minima.** This is a simple version of results in Kawaguchi [2016], for the sake of completeness we provide a simple proof. The intuition follows the fact that if there existed a local maximum / non-global minimum for F then this would translate to the existence of a local maximum / non-global minimum for the convex loss L, which is absurd.

Assume that there exists a local maximum $w^\star = (u^\star, v^\star)$, i.e. assume that there exists $\varepsilon > 0$ such that for all $w = (u, v)$ such that $\|w - w^\star\|_2^2 \leq \varepsilon$, $F(w) \leq F(w^\star)$. We show that this would imply that $\beta^\star = u^\star \odot v^\star$ is a local maximum of L, which is absurd.

The mapping $g : (u, v) \mapsto (u \odot v, \sqrt{(u^2 - v^2)/2}$ from $\mathbb{R}^d_{\geq 0} \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d_{\geq 0}$ is a bijection with inverse

$$g^{-1} : (\beta, \alpha) \mapsto (\sqrt{\alpha^2 + \sqrt{\beta^2 + \alpha^4}}, \mathrm{sign}(\beta) \odot \sqrt{-\alpha^2 + \sqrt{\beta^2 + \alpha^4}}). \tag{23}$$

Let $\tilde{\varepsilon} > 0$ and let $\beta \in \mathbb{R}^d$ such that $\|\beta - \beta^\star\|_2^2 \leq \tilde{\varepsilon}$, then for $(u, v) = g^{-1}(\beta, \alpha_*)$ where $\alpha_* = \sqrt{((u^\star)^2 - (v^\star)^2)/2}$ we have that:

$$\begin{aligned}
\|(u, v) - (u^\star, v^\star)\|_2^2 &= 2\left\|\left(\sqrt{\alpha_*^4 + \beta^2} - \sqrt{\alpha_*^4 + {\beta^\star}^2}\right)^2\right\|_1 \\
&\leq 2\|\beta^2 - {\beta^\star}^2\|_1 \\
&= 2\|(\beta - \beta^\star)^2 + 2(\beta - \beta^\star)\beta^\star\|_1 \\
&\leq 2\|(\beta - \beta^\star)^2\|_1 + 2\|\beta^\star\|_\infty \|\beta - \beta^\star\|_1 \\
&\leq 2(1 + \sqrt{d}\|\beta^\star\|_\infty)\tilde{\varepsilon} \\
&\leq \varepsilon
\end{aligned}$$

where the last inequality is for $\tilde{\varepsilon}$ small enough. Therefore if $w^\star$ were a local maximum of F, then $\beta^\star$ would be a local maximum of L. The exact same proof holds to show that there are no local minima of F which are not global minima.

Critical points. The gradient of the loss function F writes:

$$\nabla_w F(w) = \begin{pmatrix} \nabla_u F(w) \\ \nabla_v F(w) \end{pmatrix} = \begin{pmatrix} \nabla L(\beta) \odot v \\ \nabla L(\beta) \odot u \end{pmatrix} \in \mathbb{R}^{2d}.$$

Therefore $\nabla F(w_c) = \mathbf{0} \in \mathbb{R}^{2d}$ implies that $\nabla L(\beta_c) \odot \beta_c = \mathbf{0} \in \mathbb{R}^d$. Now consider such a β_c and let $\mathrm{supp}(\beta_c) = \{i \in [d] \text{ such that } \beta_c(i) \neq 0\}$ denote the support of β_c. Since $[\nabla L(\beta_c)]_i = 0$ for $i \notin \mathrm{supp}(\beta_c)$, we can therefore write that

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

Furthermore we point out that since $\mathrm{supp}(\beta_c) \subset [d]$, there are at most 2^d distinct sets $\mathrm{supp}(\beta_c)$, and therefore at most 2^d values $F(w_c) = L(\beta_c)$, where w_c is a critical point of F. □

Additional comment concerning the uniqueness of $\arg\min_{\beta_i=0, i\notin \mathrm{supp}(\beta_c)} L(\beta)$**.**

We point out that the constrained minimisation problem (5) does not necessarily have a unique solution, even when β_c is not a global solution. Though not required for any of our results, for the sake of completeness, we show here that under an additional mild assumption on the data, we can ensure that the minimisation problem (5) which appears in Proposition 1 has a unique minimum when $L(\beta_c) > 0$. Under this additional assumption, there is therefore a finite number of saddles β_c. Recall that we let $X \in \mathbb{R}^{n\times d}$ be the feature matrix and $(\tilde{x}_1, \ldots, \tilde{x}_d)$ be its columns. Now assume *temporarily* that the following assumption holds.

Assumption 2 (Assumption used just in this short section)**.** *Any subset of* $(\tilde{x}_1, \ldots, \tilde{x}_d)$ *of size smaller than* $\min(n, d)$ *is linearly independent.*

One can easily check that this assumption holds with probability 1 as soon as the data is drawn from a continuous probability distribution, similarly to [Tibshirani, 2013, Lemma 4]). In the following, for a subset $\xi = \{i_1, \ldots, i_k\} \subset [d]$, we write $X_\xi = (\tilde{x}_{i_1}, \ldots, \tilde{x}_{i_k}) \in \mathbb{R}^{n\times k}$ (we extract the columns from X). For a vector $\beta \in \mathbb{R}^d$ we write $\beta[\xi] = (\beta_{i_1}, \ldots, \beta_{i_k})$ and $\beta[\xi^C] = (\beta_i)_{i\notin\xi}$. We distinguish two different settings:

- Underparametrised setting $(n \geq d)$: in this case, for any $\xi = \{i_1, \ldots, i_k\} \subset [d]$, then $\beta^\star := \underset{\beta_i=0, i\notin\xi}{\mathrm{argmin}} L(\beta)$ is unique. Indeed we simply set the gradient to 0 and notice that due to Assumption 2, there exists a unique solution, indeed it is $\beta^\star$ such that $\beta^\star[\xi] = (X_\xi^\top X_\xi)^{-1} X_\xi^\top y$ and $\beta^\star[\xi^C] = 0$.

- Overparametrised setting $(d > n)$: **Global solutions:** $\arg\min_{\beta\in\mathbb{R}^d} L(\beta)$ is an affine space spanned by the orthogonal of $(x_1, \ldots, x_n)$ in $\mathbb{R}^d$. Since $\mathrm{span}(\tilde{x}_1, \ldots, \tilde{x}_d) = \mathbb{R}^n$ from Assumption 2, any $\beta^\star \in \arg\min_{\beta\in\mathbb{R}^d} L(\beta)$ satisfies $X\beta^\star = y$ and $L(\beta^\star) = 0$. **"Saddle points":** now let $\beta_c \in \mathbb{R}^d$ be such that we can write $\beta_c \in \arg\min_{\beta_i=0, i\notin\mathrm{supp}(\beta_c)} L(\beta)$ and assume that $L(\beta_c) > 0$ (i.e., not a global solution), then: (1) β_c has at most n non-zero entries, indeed if it were not the case, then y would necessarily belong to $\mathrm{span}(\tilde{x}_i)_{i\in\mathrm{supp}(\beta_c)}$ due to the assumption on the data, and this would lead to $L(\beta_c) = 0$, (2) therefore, similar to the underparametrised case,

$\arg\min_{\beta_i=0, i\notin \mathrm{supp}(\beta_c)} L(\beta)$ is unique, equal to β_c, and we have that $\beta_c[\xi] = (X_\xi^\top X_\xi)^{-1} X_\xi^\top y$ and $\beta_c[\xi^C] = 0$ where $\xi = \mathrm{supp}(\beta_c)$.

Thus, in both the underparametrised and overparametrised settings, the minimisation problem (5) appearing in Proposition 1 has a unique minimum when $L(\beta_c) > 0$ and Assumption 2 holds.

Appendix C. General results on the iterates

In the following lemma we recall a few results concerning the gradient flow Eq. (4):

$$\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t\,, \tag{24}$$

where F is defined in Eq. (3) as:

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

Lemma 1. *For an initialisation $u_0 = \sqrt{2}\alpha$, $v_0 = \mathbf{0}$, the flow $w_t^\alpha = (u_t^\alpha, v_t^\alpha)$ from Eq. (24) is such that the quantity $(u_t^\alpha)^2 - (v_t^\alpha)^2$ is constant and equal to $2\alpha^2\mathbf{1}$. Furthermore $u_t^\alpha > |v_t^\alpha| \geq 0$ and therefore from the bijection Eq. (23) we have that:*

$$u_t^\alpha = \sqrt{\alpha^2 + \sqrt{(\beta_t^\alpha)^2 + \alpha^4}}, \quad v_t^\alpha = \mathrm{sign}(\beta_t^\alpha) \odot \sqrt{-\alpha^2 + \sqrt{(\beta_t^\alpha)^2 + \alpha^4}}.$$

Proof. From the expression of $\nabla F(w)$, notice that the derivative of $(u_t^\alpha)^2 - (v_t^\alpha)^2$ is equal to $\mathbf{0}$ and therefore equal to its initial value.

Since $(u_t^\alpha)^2 - (v_t^\alpha)^2 = (u_t^\alpha + v_t^\alpha)(u_t^\alpha - v_t^\alpha) > 0$, by continuity we get that $u_t^\alpha + v_t^\alpha > 0$ and $u_t^\alpha - v_t^\alpha > 0$ and therefore $u_t^\alpha > |v_t^\alpha|$. □

In this section we consider the accelerated iterates Eq. (9) which follow:

$$\mathrm{d}\nabla\tilde{\phi}_\alpha(\tilde{\beta}_t^\alpha) = -\nabla L(\tilde{\beta}_t^\alpha)\mathrm{d}t, \qquad \text{where} \qquad \tilde{\phi}_\alpha := \frac{1}{\ln(1/\alpha)} \cdot \tilde{\phi}_\alpha \tag{25}$$

with $\tilde{\beta}_{t=0} = \mathbf{0}$ and where ϕ_α is defined Eq. (7).

Proposition 4. *For all $\alpha > 0$ and minimum $\beta^\star \in \arg\min_\beta L(\beta)$, the loss values $L(\tilde{\beta}_t^\alpha)$ and the Bregman divergence $\mathrm{D}_{\tilde{\phi}_\alpha}(\beta^\star, \tilde{\beta}_t^\alpha)$ are decreasing. Moreover*

$$L(\tilde{\beta}_t^\alpha) - L(\beta^\star) \leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}, \tag{26}$$

$$L\Big(\frac{1}{t}\int_0^t \tilde{\beta}_s^\alpha \mathrm{d}s\Big) - L(\beta^\star) \leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}. \tag{27}$$

Proof. The loss is decreasing since: $\frac{\mathrm{d}}{\mathrm{d}t}L(\tilde{\beta}_t^\alpha) = \nabla L(\tilde{\beta}_t^\alpha)^\top \dot{\tilde{\beta}}_t^\alpha = -\dot{\tilde{\beta}}_t^{\alpha\top} \nabla^2\tilde{\phi}_\alpha(\tilde{\beta}_t^\alpha)\dot{\tilde{\beta}}_t^\alpha \leq 0$.

$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\tilde{\phi}_\alpha}(\beta^\star, \tilde{\beta}_t^\alpha) = -\nabla L(\tilde{\beta}_t^\alpha)^\top(\tilde{\beta}_t^\alpha - \beta^\star) = -2(L(\tilde{\beta}_t^\alpha) - L(\beta^\star))$ (since L is the quadratic loss), therefore the Bregman distance is decreasing. We can also integrate this last equality from 0 to t, and divide by $-2t$:

$$\begin{aligned}\frac{1}{t}\int_0^t L(\tilde{\beta}_s^\alpha)\mathrm{d}s - L(\beta^\star) &= \frac{\mathrm{D}_{\tilde{\phi}_\alpha}(\beta^\star, \beta_0^\alpha = \mathbf{0}) - \mathrm{D}_{\tilde{\phi}_\alpha}(\beta^\star, \beta_t^\alpha)}{2t} \\ &\leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}.\end{aligned}$$

Since the loss is decreasing we get that $L(\tilde{\beta}_t^\alpha) - L(\beta^\star) \leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}$ and from the convexity of L we get that $L\Big(\frac{1}{t}\int_0^t \tilde{\beta}_s^\alpha \mathrm{d}s\Big) - L(\beta^\star) \leq \frac{\tilde{\phi}_\alpha(\beta^\star)}{2t}$. □

In the following proposition, we show that for α small enough, the iterates are bounded independently of α. Note that this result unfortunately only holds for the quadratic loss, we expect it to hold for other convex losses of the type $L(\beta) = \frac{1}{n}\sum_i \ell(y_i, \langle x_i, \beta\rangle)$ where $\ell(y, \cdot)$ is strictly convex has a unique root at y but we don't know how to show it. Also note that bounding the accelerated iterates $\tilde{\beta}^\alpha$ is equivalent to bounding the iterates β^α since $\tilde{\beta}_t^\alpha = \beta_{\ln(1/\alpha)t}^\alpha$.

Proposition 5. *For $\alpha < \alpha_0$, where α_0 depends on $\beta^\star_{\ell_1}$, the iterates $\tilde{\beta}^\alpha_t$ are bounded independently of α:*

$$\|\tilde{\beta}^\alpha_t\|_\infty \leq 3\|\beta^\star_{\ell_1}\|_1 + 1$$

Proof. From Eq. (25), integrating and using that L is the quadratic loss, we get:

$$\nabla\tilde{\phi}_\alpha(\tilde{\beta}^\alpha_t) = \frac{t}{n}X^\top(y - X\bar{\beta}^\alpha_t) = -\frac{t}{n}X^\top X(\bar{\beta}^\alpha_t - \beta^\star),$$

where we recall that $X \in \mathbb{R}^{n\times d}$ is the input data represented as a matrix and where we denote the averaged iterate by $\bar{\beta}^\alpha_t = \frac{1}{t}\int_0^t \tilde{\beta}^\alpha_s \mathrm{d}s$. Thus we get

$$\nabla\tilde{\phi}_\alpha(\tilde{\beta}^\alpha_t)^\top(\tilde{\beta}^\alpha_t - \beta^\star) = -\frac{t}{n}(\bar{\beta}^\alpha_t - \beta^\star)^\top X^\top X(\tilde{\beta}^\alpha_t - \beta^\star). \tag{28}$$

By convexity of $\tilde{\phi}_\alpha$ we have $\tilde{\phi}_\alpha(\beta^\alpha_t) - \tilde{\phi}_\alpha(\beta^\star) \leq \nabla\tilde{\phi}_\alpha(\beta^\alpha_t)^\top(\beta^\alpha_t - \beta^\star)$. By the Cauchy-Schwarz inequality, we also have $(\bar{\beta}^\alpha_t - \beta^\star)^\top X^\top X(\beta^\alpha_t - \beta^\star) \leq \|X(\beta^\alpha_t - \beta^\star)\|\|X(\bar{\beta}^\alpha_t - \beta^\star)\|$. Using Proposition 4: $\|X(\beta^\alpha_t - \beta^\star)\|^2 \leq n\tilde{\phi}_\alpha(\beta^\star)/t$ and $\|X(\bar{\beta}^\alpha_t - \beta^\star)\|^2 \leq n\tilde{\phi}_\alpha(\beta^\star)/t$ we can further bound the right hand side of Eq. (28) as

$$-\frac{t}{n}(\bar{\beta}^\alpha_t - \beta^\star)^\top X^\top X(\beta^\alpha_t - \beta^\star) \leq \tilde{\phi}_\alpha(\beta^\star).$$

Thus it yields

$$\tilde{\phi}_\alpha(\beta^\alpha_t) - \tilde{\phi}_\alpha(\beta^\star) \leq \tilde{\phi}_\alpha(\beta^\star).$$

From Woodworth et al. [2020] (proof of Lemma 1 in the appendix) we get that for

$$\alpha < \min\left\{1, \sqrt{\|\beta\|_1}, (2\|\beta\|_1)^{-1}\right\}$$

then:

$$\tilde{\phi}_\alpha(\beta) \leq \frac{3}{2}\|\beta\|_1,$$

and for all $\alpha < \exp(-d/2)$:

$$\begin{aligned}\tilde{\phi}_\alpha(\beta) &\geq \|\beta\|_1 - \frac{d}{\ln(1/\alpha^2)}\\ &\geq \|\beta\|_1 - 1,\end{aligned}$$

which finally leads for

$$\alpha < \alpha_0 := \min\left\{1, \sqrt{\|\beta^\star_{\ell_1}\|_1}, (2\|\beta^\star_{\ell_1}\|_1)^{-1}, \exp(-d/2)\right\}$$

to the result. □

The following proposition shows that we can bound the path length of the flow $\tilde{\beta}^\alpha$ independently of α. Keep in mind that the path length of $\tilde{\beta}^\alpha$ is equivalent to that of β^α as the first is just an acceleration of the second: $\tilde{\beta}^\alpha_t = \beta^\alpha_{\ln(1/\alpha)t}$.

Proposition 6. *For $\alpha < \alpha_0$ where α_0 is the same as in Proposition 5, the path length of the iterates $(\beta^\alpha_t)_{t\geq 0}$ is bounded independently of $\alpha > 0$:*

$$\int_0^{+\infty}\|\dot{\beta}^\alpha_t\|\mathrm{d}t < C,$$

where C does not depend on α. Hence the path length of the accelerated flow $\tilde{\beta}^\alpha$ is also bounded independently of α.

Proof. Having shown that the iterates β^α_t are bounded independently of α, it also implies that the iterates $w_t = (u_t, v_t)$ are bounded following Lemma 1. Since the loss $w \mapsto F(w)$ is a multivariate polynomial function, it is a semialgebraic function and we can consequently apply the result of Kurdyka [1998, Theorem 2] which grants that

$$\int_0^{+\infty}\|\dot{w}_t\|\mathrm{d}t < C,$$

where the constant C only depends on the loss and on the bound on the iterates. We further use that $\dot{\beta} = \dot{u}\odot v + u\odot\dot{v}$ and $\|\dot{u}\odot v + u\odot\dot{v}\| \leq C_1(\|\dot{u}\| + \|\dot{v}\|)$ using that u and v are bounded and $\|\dot{u}\| + \|\dot{v}\| \leq C_2\|\dot{w}\|$

using the equivalence of norms. Therefore $\int_0^{+\infty}\|\dot{\beta}_t^\alpha\|\mathrm{d}t < C$ for some C which is independent of the initialisation scale α.

□

Appendix D. Proof of Proposition 2, Proposition 3, Theorem 2 and Corollary 2

D.1. Proof of Proposition 2.

We start by proving Proposition 2 which we recall below.

Proposition 2. *Algorithm 1 is well defined: at each iteration (i) the attribution of Δ is well defined as $\Delta < +\infty$, (ii) the constrained minimisation problem has a unique solution and the attribution of the value of β is therefore well-founded. Furthermore, along the loops: the iterates β have at most n non-zero coordinates, the loss is strictly decreasing and the algorithm terminates in at most $\min\left(2^d, \sum_{k=0}^n \binom{d}{k}\right)$ steps by outputting the minimum ℓ_1-norm solution $\beta^\star_{\ell_1}$.*

Proof. In the following, for the matrix X and for a subset $I = \{i_1, \dots, i_k\} \subset [d]$, we write $X_I = (\tilde{x}_{i_1}, \dots, \tilde{x}_{i_k}) \in \mathbb{R}^{n\times k}$ (we extract the columns from X). For a vector $\beta \in \mathbb{R}^d$ we write $\beta_I = (\beta_{i_1}, \dots, \beta_{i_k})$.

(1) The constrained minimisation problem has a unique solution: we follow the proof of [Tibshirani, 2013, Lemma 2]. Following the notations in Algorithm 1, we define $I = \{i \in [d],\ |s_i| = 1\}$ and we point out that after k loops of the algorithm, the value of s is equal to $s = -(\Delta_1\nabla L(\beta_0) + \dots + \Delta_k \nabla L(\beta_{k-1})) \in \mathrm{span}(x_1, \dots, x_n)$. We can therefore write $s = X^\top r$ for some $r \in \mathbb{R}^n$.

Now assume that $\ker(X_I) \neq \{0\}$. Then, for some $i \in I$, we have $\tilde{x}_i = \sum_{j\in I\setminus\{i\}} c_j \tilde{x}_j$ where $c_j \in \mathbb{R}$. Without loss of generality, we can assume that $I \setminus \{i\}$ has at most n elements. Indeed, we can otherwise always find n elements $\tilde{I} \subset I \setminus \{i\}$ such that $\tilde{x}_i = \sum_{j\in\tilde{I}} c_j \tilde{x}_j$. Rewriting the previous equality, we get

$$s_i\tilde{x}_i = \sum_{j\in I\setminus\{i\}} (s_i s_j c_j)(s_j \tilde{x}_j). \tag{29}$$

Now by definitions of the set I and of r, we have that $\langle \tilde{x}_j, r\rangle = s_j \in \{+1, -1\}$ for any $j \in I$. Taking the inner product of Eq. (29) with r, we obtain that $1 = \sum_{j\in I\setminus\{i\}}(s_i s_j c_j)$. Consequently, we have shown that if $\ker(X_I) \neq \{0\}$, then we necessarily have for some $i \in I$,

$$s_i\tilde{x}_i = \sum_{j\in I\setminus\{i\}} a_j (s_j\tilde{x}_j),$$

with $\sum_{j\in I\setminus\{i\}} a_j = 1$, which means that $s_i\tilde{x}_i$ lies in the affine space generated by $(s_j\tilde{x}_j)_{j\in I\setminus\{i\}}$. This fact is however impossible due to Assumption 1 (recall that without loss of generality we have that $I \setminus \{i\}$ has at most n elements, and trivially less that d elements). **Therefore X_I is full rank**, and $\mathrm{Card}(I) \le n$. Now notice that the constrained minimisation problem corresponds to $\arg\min_{\substack{\beta_i\ge 0, i\in I_+\\ \beta_i\le 0, i\in I_-}} \|y - X_I\beta_I\|_2^2$. Since X_I is full rank, this restricted loss is strictly convex and the constrained minimisation problem **has a unique minimum.**

(2) $\Delta < +\infty$: Notice that the optimality conditions of

$$\beta = \arg\min_{\substack{\beta_i\ge 0, i\in I_+\\ \beta_i\le 0, i\in I_-\\ \beta_i = 0, i\notin I}} \|y - X_I\beta_I\|_2^2,$$

are (i) β satisfies the constraints, (ii) if $i \in I_+$ (resp $i \in I_-$) then $[-\nabla L(\beta)]_i \le 0$ (resp $[-\nabla L(\beta)]_i \ge 0$) and (iii) if $\beta_i \neq 0$ then $[\nabla L(\beta)]_i = 0$. One can notice that condition (ii) ensures that at each iteration, for $\delta \le \Delta_k$, $s_{k-1} - \delta\nabla L(\beta_{k-1}) \in [-1, 1]$ coordinate wise. Also, if $L(\beta_{k-1}) \neq \mathbf{0}$, then a coordinate of the vector $|s_{k-1} - \delta\nabla L(\beta_{k-1})|$ must necessarily hit 1, this value of δ corresponds to Δ_k.

(3) The loss is strictly decreasing: Let $I_{k-1,\pm}$ and $I_{k,\pm}$ be the equicorrelation sets defined in the algorithm at step $k-1$ and k, and β_{k-1} and β_k the solutions of the constrained minimisation problems. Also, let i_k be the newly added coordinate which breaks the constraint at step k (which we assume to be unique for simplicity). Without loss of generality, assume that $s_k(i_k) = +1$. Since the sets $I_{k-1,+} \setminus (I_{k,+} \setminus \{i_k\})$ and $I_{k-1,-} \setminus I_{k,-}$ are (if not empty) only composed of indexes of coordinates of β_{k-1} which are equal to 0, one can notice that β_{k-1} also satisfies the new constraints at step k. Therefore $L(\beta_k) \le L(\beta_{k-1})$. Now since $[-\nabla L(\beta_{k-1})]_{i_k} > 0$, from the strict convexity of the restricted loss on I_k, this means that

$\beta_k(i_k) > 0$ (which also means that newly activated coordinate i_k **must activate**), and therefore $\beta_{k-1} \neq \beta_k$ and $L(\beta_k) < L(\beta_{k-1})$.

(4) The algorithm terminates in at most $\min\left(2^d, \sum_{k=0}^{n} \binom{d}{k}\right)$ **steps:** Recall that we showed in part (1) of the proof that at each iteration k of the algorithm, I_k as at most $\min(n, d)$ elements. Since $\mathrm{supp}(\beta_k) \subset I_k$, we have that β_k has at most $\min(n, d)$ non-zero elements, also recall that we always have $\beta_k = \arg\min_{\beta_i = 0, i \notin \mathrm{supp}(\beta_k)} L(\beta)$ (we here have unicity of this minimisation problem following part (1) of the proof). There are hence at most

$$\sum_{k=0}^{\min(n,d)} \binom{d}{k} = \min\left(2^d, \sum_{k=0}^{n} \binom{d}{k}\right)$$

such minimisation problems. The loss being strictly decreasing, the algorithm cannot output the same solution β at two different loops, and the algorithm must terminate in at most $\min\left(2^d, \sum_{k=0}^{n} \binom{d}{k}\right)$ iterations by outputting a vector $\beta^\star$ such that $\nabla L(\beta^\star) = 0$, *i.e.* $\beta^\star \in \arg\min L(\beta)$.

(5) The algorithm outputs the minimum ℓ_1-norm solution. Let $\beta^\star$ be the output of the algorithm after p iterations. Notice that by the definition of the successive sets $I_{k,\pm}$ and of the constraints on the minimisation problem, we have that at each iteration $s_k \in \partial \|\beta_k\|_1$. Therefore $s_p \in \partial \|\beta^\star\|_1$. Also, recall from part (1) of the proof that $s_p \in \mathrm{span}(x_1, \ldots, x_n)$ which means that there exists $r \in \mathbb{R}^n$ such that $s_p = X^\top r$. Putting the two together we get that $X^\top r \in \partial \|\beta^\star\|_1$, this condition along with the fact that $L(\beta^\star) = \min L(\beta)$ are exactly the KKT conditions of $\underset{\beta \in \arg\min L}{\arg\min} \|\beta\|_1$.

□

D.2. Proof of Proposition 3.

We recall the definition of the considered quantities as well as restate Proposition 3:

$$-\int_0^\tau \dot{\hat{t}}_s^\alpha \nabla L(\hat{\beta}_s^\alpha) \mathrm{d}s = \nabla \phi_\alpha(\hat{\beta}_\tau^\alpha) \tag{30}$$

$$\dot{\hat{t}}_\tau^\alpha + \|\dot{\hat{\beta}}_\tau^\alpha\| = 1. \tag{31}$$

Proposition 3. *Let $T \geq 0$. For every $\alpha > 0$, let $(\hat{t}^\alpha, \hat{\beta}^\alpha)$ be the solution of Eq. (16). Then, there exists a subsequence $(\hat{t}^{\alpha_k}, \hat{\beta}^{\alpha_k})_{k \in \mathbb{N}}$ and $(\hat{t}, \hat{\beta})$ such that as $\alpha_k \to 0$:*

$$(\hat{t}^{\alpha_k}, \hat{\beta}^{\alpha_k}) \to (\hat{t}, \hat{\beta}) \qquad \textit{in } (C^0([0,T], \mathbb{R} \times \mathbb{R}^d), \|\cdot\|_\infty) \tag{17}$$

$$(\dot{\hat{t}}^{\alpha_k}, \dot{\hat{\beta}}^{\alpha_k}) \rightharpoonup (\dot{\hat{t}}, \dot{\hat{\beta}}) \qquad \textit{in } L_1[0,T] \tag{18}$$

Limiting dynamics. *The limits $(\hat{t}, \hat{\beta})$ satisfy:*

$$-\int_0^\tau \dot{\hat{t}}_s \nabla L(\hat{\beta}_s) \mathrm{d}s \in \partial \|\hat{\beta}_\tau\|_1 \quad \textit{and} \quad \dot{\hat{t}}_\tau + \|\dot{\hat{\beta}}_\tau\| \leq 1 \tag{19}$$

Heteroclinic orbit. *In addition, when $\hat{\beta}_\tau$ is such that $|\hat{\beta}_\tau| \odot \nabla L(\hat{\beta}_\tau) \neq 0$, we have*

$$\dot{\hat{\beta}}_\tau = -\frac{|\hat{\beta}_\tau| \odot \nabla L(\hat{\beta}_\tau)}{\||\hat{\beta}_\tau| \odot \nabla L(\hat{\beta}_\tau)\|} \quad \textit{and} \quad \dot{\hat{t}}_\tau = 0. \tag{20}$$

Furthermore, the loss strictly decreases along the heteroclinic orbits and the path length $\int_0^T \|\dot{\hat{\beta}}_\tau\| \mathrm{d}\tau$ is upperbounded independently of T.

Proof. Differentiating Eq. (30) and from the Hessian of $\tilde{\phi}_\alpha$ we get:

$$\begin{aligned} \dot{\hat{\beta}}_\tau^\alpha &= -\dot{\hat{t}}_\tau^\alpha (\nabla^2 \tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1} \nabla L(\hat{\beta}_\tau^\alpha) \\ &= -(1 - \|\dot{\hat{\beta}}_\tau^\alpha\|)(\nabla^2 \tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1} \nabla L(\hat{\beta}_\tau^\alpha). \end{aligned}$$

Therefore taking the norm on the right hand side we obtain that

$$\|\dot{\hat{\beta}}_\tau^\alpha\| = \frac{\|(\nabla^2 \tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1} \nabla L(\hat{\beta}_\tau^\alpha)\|}{1 + \|(\nabla^2 \tilde{\phi}_\alpha(\hat{\beta}_\tau^\alpha))^{-1} \nabla L(\hat{\beta}_\tau^\alpha)\|},$$

and therefore

$$\dot{\hat{\beta}}^\alpha_\tau = -\frac{(\nabla^2\tilde{\phi}_\alpha(\hat{\beta}^\alpha_\tau))^{-1}\nabla L(\hat{\beta}^\alpha_\tau)}{1+\|(\nabla^2\tilde{\phi}_\alpha(\hat{\beta}^\alpha_\tau))^{-1}\nabla L(\hat{\beta}^\alpha_\tau)\|}. \tag{32}$$

Subsequence extraction. By construction we have $\dot{\hat{t}}^\alpha_\tau + \|\dot{\hat{\beta}}^\alpha_\tau\| = 1$ (Eq. (31)), therefore the sequences $(\dot{\hat{t}}^\alpha)_\alpha$, $(\dot{\hat{\beta}}^\alpha)_\alpha$ as well as $(\hat{t}^\alpha)_\alpha$, $(\hat{\beta}^\alpha)_\alpha$ are uniformly bounded on $[0,T]$. The Arzelà-Ascoli theorem yields that, up to a subsequence, there exists $(\hat{t},\hat{\beta})$ such that $(\hat{t}^{\alpha_k},\hat{\beta}^{\alpha_k}) \to (\hat{t},\hat{\beta})$ in $(C^0([0,T],\mathbb{R}\times\mathbb{R}^d),\|\cdot\|_\infty)$. Since $\|\dot{\hat{\beta}}^\alpha_\tau\|, \|\dot{\hat{t}}^\alpha_\tau\| \le 1$ we have, applying the Banach–Alaoglu theorem, that up to a new subsequence

$$(\dot{\hat{t}}^{\alpha_k},\dot{\hat{\beta}}^{\alpha_k}) \overset{*}{\rightharpoonup} (\dot{\hat{t}},\dot{\hat{\beta}}) \text{ in } L_\infty(0,T) \tag{33}$$

and $\|\dot{\hat{\beta}}_\tau\| \le \liminf_{\alpha_k}\|\dot{\hat{\beta}}^{\alpha_k}_\tau\| \le 1$ and thus $\dot{\hat{t}}_\tau + \|\dot{\hat{\beta}}_\tau\| \le 1$:

$$\int_0^T \|\dot{\hat{\beta}}_\tau\|\mathrm{d}\tau \le \int_0^T \liminf_{\alpha_k}\|\dot{\hat{\beta}}^{\alpha_k}_\tau\|\mathrm{d}\tau \le \int_0^{+\infty} \liminf_{\alpha_k}\|\dot{\hat{\beta}}^{\alpha_k}_\tau\|\mathrm{d}\tau \le \liminf_{\alpha_k}\int_0^{+\infty}\|\dot{\hat{\beta}}^{\alpha_k}_\tau\|\mathrm{d}\tau < C,$$

where the third inequality is by Fatou's lemma. Note that since $[0,T]$ is bounded then it also implies the weak convergence in any $L_p(0,T)$, $1 \le p < \infty$. Since $(\hat{\beta}^\alpha)$ converges uniformly on $[0,T]$, and ∇L is continuous, we have that $\nabla L(\hat{\beta}^\alpha)$ converges uniformly to $\nabla L(\hat{\beta})$. Since $\dot{\hat{t}}^{\alpha_k} \rightharpoonup \dot{\hat{t}}$ in $L_1(0,T)$, passing to the limit in the equation $\nabla\tilde{\phi}_\alpha(\hat{\beta}^\alpha_\tau) = -\int_0^\tau \dot{\hat{t}}^\alpha_s \nabla L(\hat{\beta}^\alpha_s)\mathrm{d}s$ leads to

$$-\int_0^\tau \dot{\hat{t}}_s\nabla L(\hat{\beta}_s)\mathrm{d}s \in \partial\|\hat{\beta}_\tau\|_1,$$

due to Lemma 2.

Recall from Eq. (32) and the definition of $\tilde{\phi}_\alpha$ that:

$$\dot{\hat{\beta}}^\alpha_\tau = -\frac{\sqrt{\hat{\beta}^\alpha_\tau + \alpha^4}\odot\nabla L(\hat{\beta}^\alpha_\tau)}{1/\ln(1/\alpha) + \|\sqrt{\hat{\beta}^\alpha_\tau+\alpha^4}\odot\nabla L(\hat{\beta}^\alpha_\tau)\|}. \tag{34}$$

Hence assuming that $\hat{\beta}_\tau$ is such that $\||\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)\| \ne 0$, we can ensure that $\||\hat{\beta}_{\tau'}|\odot\nabla L(\hat{\beta}_{\tau'})\| \ne 0$ for $\tau' \in [\tau,\tau+\varepsilon]$ and ε small enough. We have then $\frac{\sqrt{\hat{\beta}^\alpha_{\tau'}+\alpha^4}\odot\nabla L(\hat{\beta}^\alpha_{\tau'})}{1/\ln(1/\alpha)+\|\sqrt{\hat{\beta}^\alpha_{\tau'}+\alpha^4}\odot\nabla L(\hat{\beta}^\alpha_{\tau'})\|}$ converges uniformly toward $-\frac{|\hat{\beta}_{\tau'}|\odot\nabla L(\hat{\beta}_{\tau'})}{\||\hat{\beta}_{\tau'}|\odot\nabla L(\hat{\beta}_{\tau'})\|}$ on $[\tau,\tau+\varepsilon]$. Using the dominated convergence theorem, we have $\int_\tau^{\tau+\varepsilon}\frac{\sqrt{\hat{\beta}^\alpha_{\tau'}+\alpha^4}\odot\nabla L(\hat{\beta}^\alpha_{\tau'})}{1/\log(1/\alpha)+\|\sqrt{\hat{\beta}^\alpha_{\tau'}+\alpha^4}\odot\nabla L(\hat{\beta}^\alpha_{\tau'})\|}\mathrm{d}\tau' \to \int_\tau^{\tau+\varepsilon}\frac{|\hat{\beta}_{\tau'}|\odot\nabla L(\hat{\beta}_{\tau'})}{\||\hat{\beta}_{\tau'}|\odot\nabla L(\hat{\beta}_{\tau'})\|}\mathrm{d}\tau'$. We therefore obtain $\dot{\hat{\beta}}_\tau = -\frac{|\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)}{\||\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)\|}$ in $L_1[0,T]$. Consequently $\|\dot{\hat{\beta}}_\tau\| = 1$ and $\dot{\hat{t}}_\tau = 0$.

Proof that the loss stricly decreases along the heteroclinic orbits.
Assume $\hat{\beta}_\tau$ is such that $|\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau) \ne 0$, then the flow follows

$$\dot{\hat{\beta}}_\tau = -\frac{|\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)}{\||\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)\|}$$

Letting $\gamma(\tau) = \frac{1}{\||\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)\|}$ we get:

$$\mathrm{d}L(\hat{\beta}_\tau) = -\gamma(\tau)\sum_i |\hat{\beta}_\tau(i)|\odot[\nabla L(\hat{\beta}_\tau)]_i^2\mathrm{d}\tau < 0,$$

because $|\hat{\beta}_\tau|\odot\nabla L(\hat{\beta}_\tau)^2 \ne 0$. $\square$

D.3. Proof of Theorem 2.

We can now prove Theorem 2.

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

$$\text{(“Saddles”)} \qquad \tilde\beta^\circ_t \equiv \beta_k \qquad \text{for } t \in (t_k, t_{k+1}) \text{ and } 0 \leq k \leq p, \;\; t_{p+1} = +\infty.$$

The accelerated flow $(\tilde\beta^\alpha_t)_t$ defined in Eq. (9) uniformly converges towards the limiting process $(\tilde\beta^\circ_t)_t$ on any compact subset of $\mathbb{R}_{\geq 0} \backslash \{t_1, \dots, t_p\}$.

Proof. We directly apply Theorem 3, let α_k be the subsequence from the theorem. Let $\varepsilon > 0$, for simplicity we prove the result on $[t_1 + \varepsilon, t_2 - \varepsilon]$, all the other compacts easily follow the same line of proof. Note that since $\hat t^{\alpha_k}(\tau_1') \to t_1$ and $\hat t^{\alpha_k}(\tau_2) \to t_2$, for α_k small enough $\hat t^{\alpha_k}(\tau_1') \leq t_1 + \varepsilon$ and $\hat t^{\alpha_k}(\tau_2) \geq t_2 - \varepsilon$, by the monotonicity of τ^{α_k}, this means that for α_k small enough, $\tau_1' \leq \tau^{\alpha_k}(t_1 + \varepsilon)$ and $\tau_2 \geq \tau^{\alpha_k}(t_2 - \varepsilon)$. Therefore

$$\begin{aligned}
\sup_{t \in [t_1+\varepsilon, t_2-\varepsilon]} \|\tilde\beta^{\alpha_k}_t - \beta_1\| &= \sup_{t \in [t_1+\varepsilon, t_2-\varepsilon]} \|\hat\beta^{\alpha_k}(\tau_{\alpha_k}(t)) - \beta_1\| \\
&= \sup_{\tau \in [\tau^{\alpha_k}(t_1+\varepsilon), \tau^{\alpha_k}(t_2-\varepsilon)]} \|\hat\beta^{\alpha_k}(\tau) - \beta_1\| \\
&\leq \sup_{\tau \in [\tau_1', \tau_2]} \|\hat\beta^{\alpha_k}(\tau) - \beta_1\|,
\end{aligned}$$

which goes uniformly to 0 following Theorem 3. Since this result is independent of the subsequence α_k, we get the result of Theorem 2. $\square$

D.4. Proof of Corollary 2.

We restate and prove Corollary 2 below.

Corollary 2. *For all $T > t_p$, the graph of the iterates $(\tilde\beta^\alpha_t)_{t \leq T}$ converges to that of $(\hat\beta_\tau)_\tau$:*

$$\operatorname{dist}(\{\tilde\beta^\alpha_t\}_{t \leq T}, \{\hat\beta_\tau\}_{\tau \geq 0}) \underset{\alpha \to 0}{\longrightarrow} 0,$$

where $\operatorname{dist}(\cdot, \cdot)$ corresponds to the Hausdorff distance between 2 sets.

Proof. For α small enough, we have that $\hat t^\alpha_{\tau_p'} \leq t_p + \varepsilon \leq T$

$$\begin{aligned}
\sup_{\tau \geq 0} d(\hat\beta_\tau, \{\tilde\beta^\alpha_t\}_{t \leq T}) &= \sup_{\tau \leq \tau_p'} d(\hat\beta_\tau, \{\tilde\beta^\alpha_t\}_{t \leq T}) \\
&\leq \sup_{\tau \leq \tau_p'} \|\hat\beta_\tau - \tilde\beta^\alpha_{\hat t^\alpha_\tau}\| \\
&= \sup_{\tau \leq \tau_p'} \|\hat\beta_\tau - \hat\beta^\alpha_\tau\| \underset{\alpha \to 0}{\longrightarrow} 0,
\end{aligned}$$

according to Theorem 3.

Similarly:

$$\begin{aligned}
\sup_{t \leq T} d(\tilde\beta^\alpha_t, \{\hat\beta_{\tau'}\}_{\tau'}) &= \sup_{\tau \leq \tau^\alpha_T} d(\hat\beta^\alpha_\tau, \{\hat\beta_{\tau'}\}_{\tau'}) \\
&\leq \sup_{\tau \leq \tau^\alpha_T} \|\hat\beta^\alpha_\tau - \hat\beta_\tau\| \underset{\alpha \to 0}{\longrightarrow} 0,
\end{aligned}$$

according to Theorem 3, which concludes the proof. $\square$

Appendix E. Technical lemmas

The following lemma describes the behaviour of $\nabla\tilde\phi_\alpha(\beta^\alpha)$ as $\alpha \to 0$ in function of the subdifferential $\partial\|\cdot\|_1$.

Lemma 2. *Let $(\beta^\alpha)_{\alpha > 0}$ such that $\beta^\alpha \underset{\alpha \to 0}{\longrightarrow} \beta \in \mathbb{R}^d$.*

- *if $\beta_i > 0$ then $[\nabla\tilde\phi_\alpha(\beta^\alpha)]_i$ converges to 1*
- *if $\beta_i < 0$ then $[\nabla\tilde\phi_\alpha(\beta^\alpha)]_i$ converges to -1.*

Moreover if we assume that $\nabla\tilde\phi_\alpha(\beta^\alpha)$ converges to $\eta \in \mathbb{R}^d$, we have that:

- $\eta_i \in (-1, 1) \Rightarrow \beta_i = 0$

- $\beta_i = 0 \Rightarrow \eta_i \in [-1, 1]$.

Overall, assuming that $(\beta^\alpha, \nabla\tilde{\phi}_\alpha(\beta^\alpha)) \underset{\alpha\to 0}{\longrightarrow} (\beta, \eta)$*, we can write:*

$$\eta \in \partial\|\beta\|_1.$$

Proof. We have that

$$\begin{aligned}[\nabla\tilde{\phi}_\alpha(\beta^\alpha)]_i &= \frac{1}{2\ln(1/\alpha)}\operatorname{arcsinh}(\frac{\beta_i^\alpha}{\alpha^2}) \\ &= \frac{1}{2\ln(1/\alpha)}\ln\Big(\frac{\beta_i^\alpha}{\alpha^2} + \sqrt{\frac{(\beta_i^\alpha)^2}{\alpha^4} + 1}\Big).\end{aligned}$$

Now assume that $\beta_i^\alpha \to \beta_i > 0$, then $[\nabla\tilde{\phi}_\alpha(\beta^\alpha)]_i \to 1$, if $\beta_i < 0$ we conclude using that $\operatorname{arcsinh}$ is an odd function. All the claims are simple consequences of this. $\square$

The following lemma shows that the extracted limits $\hat{t}$ as defined in Proposition 3 diverge to ∞. This divergence is crucial as it implies that the rescaled iterates $(\hat{\beta}_\tau)_\tau$ explore the whole trajectory..

Lemma 3. *For any extracted limit $\hat{t}$ as defined in Proposition 3, we have that $\tau - C \leq \hat{t}_\tau$ where C is the upperbound on the length of the curves defined in proposition 6.*

Proof. Recall that

$$\tau^\alpha(t) = t + \int_0^t \|\dot{\hat{\beta}}_s^\alpha\| \mathrm{d}s.$$

From Proposition 6, the full path length $\int_0^{+\infty} \|\dot{\beta}_s^\alpha\| \mathrm{d}s$ is finite and bounded by some constant C independently of α. Therefore τ^α is a bijection in $\mathbb{R}_{\geq 0}$ and we defined $\hat{t}_\tau^\alpha = (\tau^\alpha)^{-1}(\tau)$. Furthermore $\tau^\alpha(t) \leq t + C$ leads to $t \leq \hat{t}^\alpha(t + C)$ and therefore $\tau - C \leq \hat{t}^\alpha(\tau)$ for all $\tau \geq 0$. This inequality still holds for any converging subsequence, which proves the result. $\square$

Under a mild additional assumption on the data (see Assumption 2), we showed after the proof of Proposition 1 in Appendix B that the number of saddles of F is finite. Without this assumption, the number of saddles is *a priori* not finite. However the following lemma shows that along the flow of $\hat{\beta}$ the number of saddles which can potentially be visited is indeed finite.

Lemma 4. *The limiting flow $\hat{\beta}$ as defined in Proposition 3 can only visit a finite number of critical points $\beta \in S \coloneqq \{\beta \in \mathbb{R}^d, \beta \odot \nabla L(\beta) = \mathbf{0}\}$ and can visit each one of them at most once.*

Proof. Let $\tau \geq 0$, and assume that $\hat{\beta}_\tau \in S$, i.e., we are at a critical point at time τ. From Proposition 1, we have that

$$\hat{\beta}_\tau \in \underset{\beta_i = 0 \text{ for } i \notin \operatorname{supp}(\hat{\beta}_\tau)}{\arg\min} L(\beta), \tag{35}$$

Let us define the sets

$$I_\pm \coloneqq \{i \in \{1, \dots, d\}, \text{ s.t. } \hat{s}_\tau(i) = \pm 1\} \quad \text{and} \quad I = I_+ \cup I_-.$$

The set I corresponds to the coordinate of $\hat{\beta}_\tau$ which "are allowed" (but not obliged) to be non-zero since from Eq. (19), $\operatorname{supp}(\hat{\beta}_\tau) \subset I$. Now given the fact that the sub-matrix $X_I = (\tilde{x}_i)_{i\in I} \in \mathbb{R}^{n\times \operatorname{card}(I)}$ is full rank (see part (1) of the proof of Proposition 2 for the explanation), the solution of the minimisation problem (35) is unique and equal to $\beta[\xi] = (X_\xi^\top X_\xi)^{-1} X_\xi^\top y$ and $\beta[\xi^C] = 0$ where $\xi = \operatorname{supp}(\hat{\beta}_\tau)$. There are $2^d = \operatorname{Card}(P([d]))$ (where $P([d])$ contains all the subsets of $[d]$) number of constraints of the form $\{\beta_i = 0, i \notin \mathcal{A}\}$, where $\mathcal{A} \subset [d]$, and $\hat{\beta}_\tau$ is the unique solution of one of them. $\hat{\beta}_\tau$ can therefore take at most 2^d values (very crude upperbound). There is therefore a finite number of critical points which can be reached by the flow $\hat{\beta}$. Furthermore, from Proposition 3, the loss is strictly decreasing along the heteroclinic orbits, each of these critical points can therefore be visited at most once.

$\square$