Journal of Machine Learning Research 24 (2023) 1-26 Submitted 12/22; Revised 4/23; Published 5/23

Incremental Learning in Diagonal Linear Networks

Raphaël Berthier RAPHAEL.BERTHIER@EPFL.CH
EPFL
Lausanne, Switzerland

Editor: Lorenzo Rosasco

Abstract

Diagonal linear networks (DLNs) are a toy simplification of artificial neural networks; they consist in a quadratic reparametrization of linear regression inducing a sparse implicit regularization. In this paper, we describe the trajectory of the gradient flow of DLNs in the limit of small initialization. We show that incremental learning is effectively performed in the limit: coordinates are successively activated, while the iterate is the minimizer of the loss constrained to have support on the active coordinates only. This shows that the sparse implicit regularization of DLNs decreases with time. This work is restricted to the underparametrized regime with anti-correlated features for technical reasons.

Keywords: diagonal linear networks, incremental learning, saddle-to-saddle dynamics, implicit bias, Lotka-Volterra

1. Introduction

Artificial neural networks are the state of the art for many machine learning tasks (Le Cun et al., 2015); however, we lack theoretical understanding of this success (Zhang et al., 2021). Indeed, the parametrization of neural networks induces a non-convex loss, and consequently it is challenging to analyze the optimization error of gradient descent methods. Moreover, neural networks can be successful even without any (explicit) regularizer; this challenges the statistical wisdom in overparametrized settings.

Recent research suggests that these two problems are intertwined: through its non-convex parametrization, the gradient descent dynamics of neural networks induce an implicit regularization that controls the statistical performance (Bartlett et al., 2021). However, this phenomenon is difficult to describe because it is a joint effect of the parametrization, the gradient descent dynamics and the initialization.

As a consequence, theoretical research has focused on studying implicit regularization in toy simplifications of neural networks (Soudry et al., 2018; Gunasekar et al., 2017; Li et al., 2018; Chizat and Bach, 2020; Li et al., 2020). We are interested in an extreme simplification, called *diagonal linear networks* (DLNs) (Vaskevicius et al., 2019; Zhao et al., 2019; Woodworth et al., 2020; HaoChen et al., 2021; Li et al., 2021; Azulay et al., 2021; Pesme et al., 2021; Pillaud-Vivien et al., 2022; Nacson et al., 2022; Chou et al., 2023). In fact, it is only a linear regression where regressors θ_i are parametrized quadratically; specifically, in this paper, we parametrize $\theta_i = u_i^2/4$. We then perform a gradient descent in terms of u_i and not θ_i (see Section 2.1 for more details). This quadratic reparametrization is loosely argued to have an effect similar to the composition of two layers in a neural network. When started from a small initialization, DLNs were rigorously shown to enforce an implicit sparse regularization; in an overparametrized setting, DLNs converge to a sparse interpolator.

Previous works have thus focused on describing the limit point of the dynamics of DLNs. Instead, in this work, we study the full trajectory of the continuous-time gradient flow dynamics. In the limit of small initialization, we show that *incremental learning* (see, e.g., Saxe et al. (2019); Gidel et al. (2019); Gissin et al. (2019); Li et al. (2020)) is effectively performed: as time increases, coordinates are successively activated and the iterate is the minimizer of the loss constrained to have support on the active coordinates only. The main contribution of this paper is the description of the time-dependent set of active coordinates, and the rigorous proof of the convergence to a regressor whose sparsity depends on the stopping time. The take-home message is that DLNs enforce a sparse implicit regularization that decreases as the stopping time increases.

As a corollary of our description of the dynamics, we obtain an asymptotic equivalent of the convergence time to the minimizer of the loss in the limit of small initialization. It is quite remarkable that such a precise estimate of the global convergence time can be obtained for a non-convex optimization problem.

For technical reasons, our work is restricted to the special case of anti-correlated feature; consequently, we study only underparametrized problems (see Section 2.2). However, we do not expect this restriction to be necessary for incremental learning to occur in DLNs. We leave the proof of this for future work.

The rest of this paper is organized as follows. In Section 2, we present our setting, our assumptions and our results. In Section 3, we articulate the related work in more detail. Sections 4 and 5 prove our results.

Notations. We use bold notations for vectors and matrices: for instance, if $\boldsymbol{\theta} \in \mathbb{R}^d$ is a multi-dimensional vector, we denote θ_i its coordinates. Similarly, if $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, we denote M_{ij} its entries. If I is a subset of $\{1, \dots, d\}$, we denote I^c its complement and $|I|$ its cardinality. We denote $\boldsymbol{\theta}_I \in \mathbb{R}^{|I|}$ the subvector obtained from $\boldsymbol{\theta} \in \mathbb{R}^d$ by keeping only the coordinates indexed by $i \in I$. Similarly, if I, J are subsets of $\{1, \dots, d\}$, we denote $\boldsymbol{M}_{IJ}$ the submatrix obtained from $\boldsymbol{M}$ by keeping only the rows indexed by $i \in I$ and columns indexed by $j \in J$.

If $\boldsymbol{\theta}, \boldsymbol{\nu} \in \mathbb{R}^d$, we write $\boldsymbol{\theta} \geqslant \boldsymbol{\nu}$ (resp. $\boldsymbol{\theta} > \boldsymbol{\nu}$) to denote that for all $i \in \{1, \dots, d\}$, $\theta_i \geqslant \nu_i$ (resp. $\theta_i > \nu_i$). In particular, $\boldsymbol{\theta} \geqslant \mathbf{0}$ (resp. $\boldsymbol{\theta} > \mathbf{0}$) denotes that all coordinates of $\boldsymbol{\theta}$ are non-negative (resp. positive).

We use the notations $\langle ., . \rangle$ and $\|.\|$ to denote the Euclidean dot product and norm respectively.

2. Main Results

In this section, we first introduce the parametrization of DLNs and the induced gradient flow dynamics (Section 2.1). Then, we state the assumption that features are anti-correlated and discuss the consequences (Section 2.2). Finally, we state our results (Section 2.3).

2.1 Setting

We perform a linear regression of n output variables $y_1, \dots, y_n \in \mathbb{R}$ from n corresponding input variables $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$. The traditional approach is to minimize the quadratic loss:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ f(\boldsymbol{\theta}) = \frac{1}{2} \sum_{k=1}^{n} (y_k - \langle \boldsymbol{\theta}, \boldsymbol{x}_k \rangle)^2 \right\} . \tag{1}$$

Denote $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ the matrix whose rows are $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$, and $\boldsymbol{X}_1, \dots, \boldsymbol{X}_d \in \mathbb{R}^n$ the columns of the matrix $\boldsymbol{X}$, i.e., the features of the regression problem. The quadratic loss f can be expressed as a function of the covariance $\boldsymbol{M} = \boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{d \times d}$ of the features and of the covariance $\boldsymbol{r} = \boldsymbol{X}^\top \boldsymbol{y}$ between the features and the output:

$$f(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{y}\|^2 - \langle \boldsymbol{r}, \boldsymbol{\theta} \rangle + \frac{1}{2} \langle \boldsymbol{\theta}, \boldsymbol{M} \boldsymbol{\theta} \rangle .$$

A strategy to minimize this convex function is to perform a gradient descent, i.e., an Euler discretization of the gradient flow

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = -\frac{\mathrm{d}f}{\mathrm{d}\theta_i} = r_i - \sum_{j=1}^{d} M_{ij} \theta_j , \tag{2}$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}(t)$ is a function from $\mathbb{R}_{\geqslant 0}$ to $\mathbb{R}^d$. It is widely known that for any initial point, this gradient flow converges exponentially fast to a minimizer of f.

In this paper, we are interested in the effect of reparametrizing

$$\theta_i = \frac{1}{4} u_i^2 .$$

This reparametrization of linear regression is called a diagonal linear network (DLN). We perform the gradient flow in terms of $\boldsymbol{u} \in \mathbb{R}^d$ instead of $\boldsymbol{\theta} \in \mathbb{R}^d$: $\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\frac{\mathrm{d}f}{\mathrm{d}u_i}$. Using that $\mathrm{d}\theta_i = \frac{1}{2} u_i \mathrm{d}u_i$, we compute the resulting equation in θ_i:

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \frac{1}{2} u_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = -\frac{1}{2} u_i \frac{\mathrm{d}f}{\mathrm{d}u_i} = -\frac{1}{4} u_i^2 \frac{\mathrm{d}f}{\mathrm{d}\theta_i} ,$$

and thus

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = -\theta_i \frac{\mathrm{d}f}{\mathrm{d}\theta_i} = \theta_i \left(r_i - \sum_{j=1}^{d} M_{ij} \theta_j \right) . \tag{3}$$

Compare (3) with (2). The reparametrization has added a factor θ_i in the derivative of θ_i. This implies that if θ_i is initialized at 0, then it remains at 0 in the DLN dynamics (3). In particular, $\boldsymbol{\theta} = \boldsymbol{0} \in \mathbb{R}^d$ is a stable point of the dynamics.

In this paper, we are interested in the DLN when initialized close to this stable point. More precisely, for $\varepsilon > 0$, define $\boldsymbol{\theta}^{(\varepsilon)} = \boldsymbol{\theta}^{(\varepsilon)}(t)$ as the solution of the DLN dynamics (3) initialized from $\boldsymbol{\theta}^{(\varepsilon)}(0) = (C_1 \varepsilon^{k_1}, \dots, C_d \varepsilon^{k_d})$, where $\boldsymbol{C} = (C_1, \dots, C_d) > \boldsymbol{0}$ and $\boldsymbol{k} = (k_1, \dots, k_d) > \boldsymbol{0}$ are constants.

2.2 Assumptions

Before we get to a rigorous statement of our results, let us state our assumptions.

(A1) $\boldsymbol{r} = \boldsymbol{X}^\top \boldsymbol{y} > \boldsymbol{0}$, i.e., the covariance $r_i = \langle \boldsymbol{X}_i, \boldsymbol{y} \rangle$ between the output $\boldsymbol{y}$ and the feature $\boldsymbol{X}_i$ is positive for all $i \in \{1, \dots, d\}$.

The reparametrization $\theta_i = \frac{1}{4} u_i^2$ constrains the linear regression to have non-negative weights. In this situation, it is natural to preprocess the data by potentially changing the signs of the features $\boldsymbol{X}_1, \dots, \boldsymbol{X}_d$ so that the output is positively correlated with the features. Assumption (A1) assumes that this pre-processing has been done, and—for technical reasons—that the correlations are non-zero.

(A2) For all $i \neq j$, $M_{ij} = \langle \boldsymbol{X}_i, \boldsymbol{X}_j \rangle \leqslant 0$, i.e., the features are anti-correlated.

We assume that once the features have been positively correlated with the output, they are anti-correlated. This assumption is a strong restriction to the class of studied problems and weakening it is left as an open problem. A major motivation for this assumption is that it implies that the trajectories of the DLN dynamics are nondecreasing.

Lemma 1 *Assume (A1)-(A2). There exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0]$, for all $i \in \{1, \dots, d\}$, $\theta_i^{(\varepsilon)}(t)$ is nondecreasing in t.*

The proof of this result is postponed to Appendix B.

As a side comment, note that Assumptions (A1) and (A2) jointly constrain the problem to be in the underparametrized regime $n \geqslant d$.

Proposition 1 *Assume (A1)-(A2). Then $\boldsymbol{M} = \boldsymbol{X}^\top \boldsymbol{X}$ is positive definite. In particular, as $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, we have $n \geqslant d$.*

The proof of this result is postponed to Appendix C.

2.3 Statement of the Results

Our main result (Theorem 1 below) states that the DLN spends long periods of time in the vicinity of fixed points of (3), and describes the times at which transitions occur. To start with, we describe this family of fixed points, using the notations introduced in Section 1.

Proposition 2 *Assume (A1)-(A2). For all $I \subset \{1, \dots, d\}$, there exists a unique $\boldsymbol{\theta} \geqslant 0$ fixed point of (3) with support $\{i \in \{1, \dots, d\} \,|\, \theta_i > 0\}$ equal to I. We denote this fixed point as $\boldsymbol{\theta}_*^{(I)}$. Its non-zero coordinates are $(\boldsymbol{\theta}_*^{(I)})_I = (\boldsymbol{M}_{II})^{-1} \boldsymbol{r}_I$. There are thus 2^d fixed points of (3).*

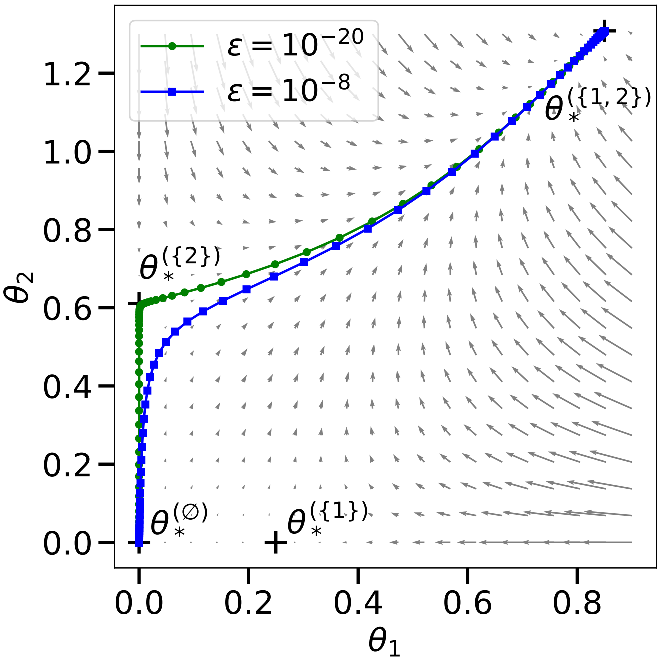


Figure 1: In dimension $d = 2$, we show the vector field $V(\theta_1, \theta_2) = (\theta_1(r_1 - A_{11}\theta_1 - A_{12}\theta_2), \theta_2(r_2 - A_{21}\theta_1 - A_{22}\theta_2))$ associated to the DLN dynamics (3) (gray arrows), its fixed points $\theta_*^{(I)}$ for $I \subset \{1, 2\}$ (black crosses) and the trajectories of $\theta^{(\varepsilon)}(t)$ for $\varepsilon = 10^{-8}$ (blue) and $\varepsilon = 10^{-20}$ (green). In this simulation, $n = 3$ and the data $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{y} \in \mathbb{R}^n$ is generated randomly with i.i.d. standard Gaussian entries, conditionally on the event that Assumptions (A1) and (A2) hold. The initialization is $\boldsymbol{\theta}^{(\varepsilon)}(0) = (\varepsilon, \varepsilon)$.

The proof that $(\boldsymbol{M}_{II})^{-1}$ exists and the proof of the proposition are postponed to Appendix D. We give here a high-level intuition. For each $i \in \{1, \ldots, d\}$, there are two ways of canceling out the right hand side of (3): either $\theta_i = 0$ or $r_i - \sum_j M_{ij}\theta_j = 0$. For the fixed point $\boldsymbol{\theta}_*^{(I)}$, the set $I \subset \{1, \ldots, d\}$ is the set of coordinates i such that $\theta_{*,i}^{(I)} \neq 0$ and $r_i - \sum_j M_{ij}\theta_{*,j}^{(I)} = 0$; conversely for $i \notin I$, $\theta_{*,i}^{(I)} = 0$. We say that the coordinates in I are the *active* coordinates of $\boldsymbol{\theta}_*^{(I)}$.

If no coordinate is active, we obtain the fixed point $\boldsymbol{\theta}_*^{(\emptyset)} = \boldsymbol{0}$ of (3). If all coordinates are active, $\boldsymbol{\theta}_*^{(\{1,\ldots,d\})} = \boldsymbol{M}^{-1}\boldsymbol{r}$ is the minimum of f, thus a fixed point of both gradient flows (2) and (3). In Figure 1, we provide an illustration in dimension $d = 2$. We show the vector field defined by the DLN dynamics (3), its $2^d = 4$ fixed points enumerated above and the trajectories $\theta^\varepsilon(t)$ for different values of ε.

We are now in position to state our main theorem. Recall that for $\varepsilon > 0$, $\boldsymbol{\theta}^{(\varepsilon)} = \boldsymbol{\theta}^{(\varepsilon)}(t)$ is the solution of the DLN dynamics (3) initialized from $\boldsymbol{\theta}^{(\varepsilon)}(0) = (C_1\varepsilon^{k_1}, \dots, C_d\varepsilon^{k_d})$, where $\boldsymbol{C} = (C_1, \dots, C_d) > \mathbf{0}$ and $\boldsymbol{k} = (k_1, \dots, k_d) > \mathbf{0}$ are constants.

Theorem 1 *Assume (A1)-(A2). For $s > 0$, define $\boldsymbol{\mu}(s)$ as the unique minimizer of the regularized and constrained minimization problem*

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d,\, \boldsymbol{\theta}\geqslant\mathbf{0}}. \left\{ f(\boldsymbol{\theta}) + \frac{1}{s}\langle \boldsymbol{k}, \boldsymbol{\theta}\rangle \right\} \tag{4}$$

and

$$I(s) = \{i \in \{1, \dots, d\} \,|\, \mu_i(s) > 0\} \, .$$

Then we have the following:

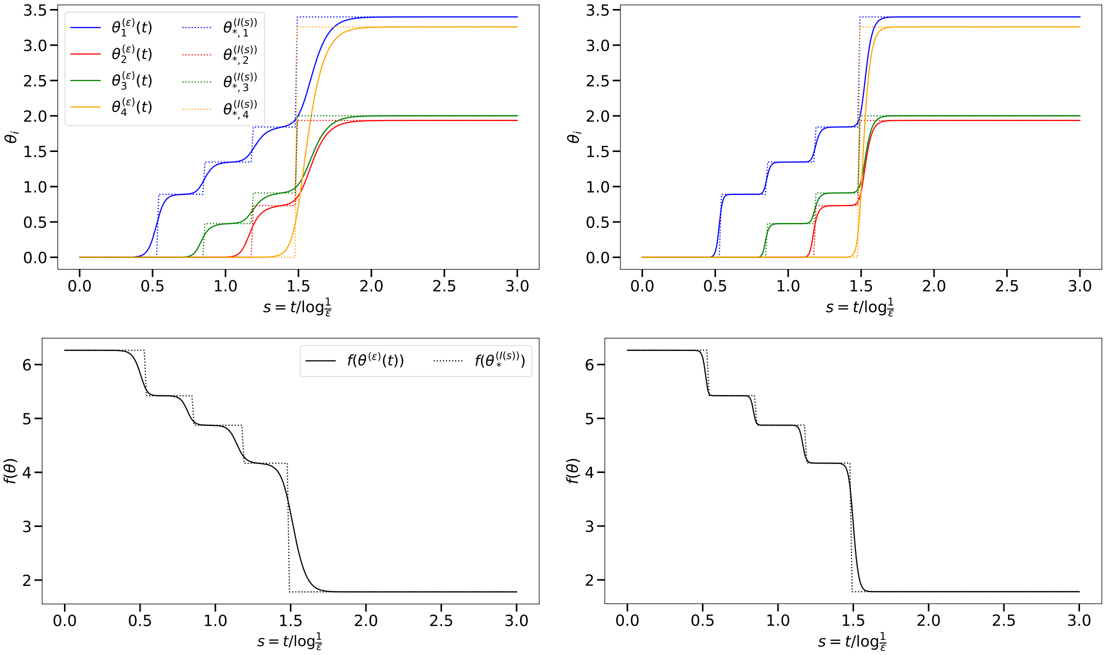
1. *The minimizer $\boldsymbol{\mu}(s)$ is nondecreasing in s, i.e., for all $i \in \{1, \dots, d\}$, $\mu_i(s)$ is nondecreasing in s. Consequently, $I(s)$ is nondecreasing in s for the inclusion.*
2. *Denote $s_1, \dots, s_q$ the points of discontinuity of the function $s \mapsto I(s)$. For all $s > 0$, $s \neq s_1, \dots, s_q$,*
$$\boldsymbol{\theta}^{(\varepsilon)}\left(s \log \frac{1}{\varepsilon}\right) \xrightarrow[\varepsilon\to 0]{} \boldsymbol{\theta}_*^{(I(s))} \, .$$
Moreover, the convergence is uniform for s in compact subsets of $\mathbb{R}_{>0}\backslash\{s_1, \dots, s_q\}$.
3. *For all $s > 0$,*
$$\frac{1}{s\log\frac{1}{\varepsilon}} \int_0^{s\log\frac{1}{\varepsilon}} \mathrm{d}t\, \boldsymbol{\theta}^{(\varepsilon)}(t) \xrightarrow[\varepsilon\to 0]{} \boldsymbol{\mu}(s) \, .$$
Moreover, the convergence is uniform for s in compact subsets of $\mathbb{R}_{>0}$.

This theorem is proved in Section 4. It states that $\boldsymbol{\theta}^{(\varepsilon)}\left(s\log\frac{1}{\varepsilon}\right)$ converges to a piecewise constant function, taking values at the fixed points of the DLN dynamics (3). Moreover, the set $I(s)$ of active coordinates of the limit is nondecreasing, showing that there are successive coordinate activations.

We provide an illustration of these successive coordinate activations in Figure 2. Note that when a new coordinate is activated, all other coordinates are perturbed. Moreover, as ε decreases from 10^{-8} to 10^{-20}, one can observe that $\boldsymbol{\theta}_*^{(I(s))}$ is becoming a sharper approximation of $\boldsymbol{\theta}^{(\varepsilon)}\left(s\log\frac{1}{\varepsilon}\right)$ and $f(\boldsymbol{\theta}_*^{(I(s))})$ is becoming a sharper approximation of $f\left(\boldsymbol{\theta}^{(\varepsilon)}\left(s\log\frac{1}{\varepsilon}\right)\right)$.

The set $I(s)$ of active coordinates of the limit is obtained by solving a regularized and constrained version (4) of the original optimization problem (1). The non-active constraints at the optimum $\boldsymbol{\mu}(s)$ correspond to the active coordinates of $\boldsymbol{\theta}_*^{(I(s))}$.

The regularization term $+\frac{1}{s}\langle\boldsymbol{k}, \boldsymbol{\theta}\rangle$ in (4) has a decreasing sparse regularizing effect. The author did not find a finer high-level motivation to explain why $I(s)$ should be defined through (4); his insights come only from the proof of the theorem. However, Theorem 1.(3) states a second relation between the DLN dynamics (3) and the optimization problem (4): the average $\frac{1}{s\log\frac{1}{\varepsilon}}\int_0^{s\log\frac{1}{\varepsilon}}\mathrm{d}t\,\boldsymbol{\theta}^{(\varepsilon)}(t)$ of the trajectory converges to the minimizer $\boldsymbol{\mu}(s)$ as $\varepsilon \to 0$; said differently, the average of the trajectory computes the regularization path (Hastie et al., 2009, Section 3) of the regularized optimization problem (4).



(a) $\varepsilon = 10^{-8}$

(b) $\varepsilon = 10^{-20}$

Figure 2: Comparison between the coordinates $\theta_i^{(\varepsilon)}\left(s \log \frac{1}{\varepsilon}\right)$ and their asymptotic approximation $\theta_{*,i}^{(I(s))}$ (upper plots) and between the losses $f\left(\boldsymbol{\theta}^{(\varepsilon)}\left(s \log \frac{1}{\varepsilon}\right)\right)$ and $f(\boldsymbol{\theta}_*^{(I(s))})$ (lower plots). The simulations are run with $\varepsilon = 10^{-8}$ (left plots) and $\varepsilon = 10^{-20}$ (right plots). In this simulation, $n = 5$, $d = 4$ and the data $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{y} \in \mathbb{R}^n$ is generated randomly with i.i.d. standard Gaussian entries, conditionally on the event that Assumptions (A1) and (A2) hold. The initialization is $\boldsymbol{\theta}^{(\varepsilon)}(0) = (\varepsilon, \ldots, \varepsilon)$ and thus $\boldsymbol{k} = (1, \ldots, 1)$.

Remark 3 *In Theorem 1.(2), it is not possible to have uniform convergence in neighborhoods of $s_1, \dots, s_q$ as the functions $\boldsymbol{\theta}^{(\varepsilon)}\left(s \log \frac{1}{\varepsilon}\right)$ are continuous while $\boldsymbol{\theta}_*^{(I(s))}$ is discontinuous at $s_1, \dots, s_q$; a uniform convergence would contradict the Arzelà theorem (Dunford and Schwartz, 1988, Section IV.6).*

The definition of the asymptotic process $\boldsymbol{\theta}_*^{(I(s))}$ is rather complex as one has to solve an optimization problem. Nevertheless, in the following corollary, we show that it is still possible to deduce a simple expression for the convergence time of $\boldsymbol{\theta}^{(\varepsilon)}(t)$ to the minimizer $\boldsymbol{\theta}_*^{(\{1,\dots,d\})}$ of f.

Corollary 1 (convergence time to the minimizer) *Assume (A1)-(A2). For all $\eta > 0$, denote*

$$\tau_\eta^{(\varepsilon)} = \inf\left\{t \geqslant 0 \,\middle|\, \left\|\boldsymbol{\theta}^{(\varepsilon)}(t) - \boldsymbol{\theta}_*^{(\{1,\dots,d\})}\right\| \leqslant \eta\right\}$$

the hitting time of the ball centered around the minimizer $\boldsymbol{\theta}_*^{(\{1,\dots,d\})}$ of f and of radius η. Then for η small enough,

$$\frac{\tau_\eta^{(\varepsilon)}}{\log \frac{1}{\varepsilon}} \xrightarrow[\varepsilon \to 0]{} \max_{i \in \{1,\dots,d\}} \frac{(\boldsymbol{M}^{-1}\boldsymbol{k})_i}{(\boldsymbol{M}^{-1}\boldsymbol{r})_i}\,.$$

This corollary is proved in Section 5. Note that we describe the hitting time $\tau_\eta^{(\varepsilon)}$ only in the asymptotic limit $\varepsilon \to 0$, and in this limit, $\tau_\eta^{(\varepsilon)}/\log\frac{1}{\varepsilon}$ is independent of η (for η small enough). This surprising property is due to the fact that the limit trajectory reaches the global minimizer $\boldsymbol{\theta}_*^{(\{1,\dots,d\})}$ through a last jump of the iterates (see Figure 2) and the duration of this jump is negligible before $\log\frac{1}{\varepsilon}$.

3. Related Work

Diagonal linear networks (DLNs). Previous studies of DLNs show that their dynamics select sparse estimators in an overparametrized setting (Vaskevicius et al., 2019; Zhao et al., 2019; Woodworth et al., 2020; HaoChen et al., 2021; Li et al., 2021; Azulay et al., 2021; Pesme et al., 2021; Pillaud-Vivien et al., 2022; Nacson et al., 2022; Chou et al., 2023). Our work differs from previous studies in two ways: first, we describe the limit of the full trajectory of the dynamics, but second, we are technically restricted to the underparametrized setting.

Many studies consider the more general quadratic reparametrization $\theta_i = (u_i^2 - v_i^2)/4$ or equivalently $\theta_i = u_i v_i$, which do not constrain θ_i to be non-negative, while our reparametrization $\theta_i = u_i^2/4$ does. However, under our Assumptions (A1) and (A2), the restriction to non-negative regressors is benign. Heuristically, the regressors θ_i are nondecreasing (Lemma 1), thus they do not "try" to become negative. We thus claim that under the more general parametrization $\theta_i = (u_i^2 - v_i^2)/4$, the variables v_i would remain negligible and the results would be the same.

When $\boldsymbol{M} = \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_d$, the DLNs dynamics (3) are separable across coordinates and can be solved using the logistic equation. In this special case, the activation of a coordinate does not affect the other coordinates. Further, one can check that the coordinates are activated in the decreasing order of the loss decrease that they induce. In (Vaskevicius et al., 2019; Zhao et al., 2019; Li et al., 2021), a restricted isometry property or incoherence property controls the deviation from this special case. On the contrary, in this paper, we do not make such an assumption and observe richer phenomena. In Figure 2, we observe each coordinate activation has a large influence on other active coordinates; moreover, the coordinate introduced last is the second largest coordinate of the optimum $\boldsymbol{\theta}_*^{(\{1,\dots,4\})}$ and induces the largest loss decrease.

To the best of our knowledge, previous analyses of DLNs have focused on the case where the initializations $\theta_i^{(\varepsilon)}(0) = C_i\varepsilon^{k_i}$ of all coordinates have the same order of magnitude, i.e., $\boldsymbol{k} = (1, \dots, 1)$. In this paper, we generalize to $\boldsymbol{k} \neq (1, \dots, 1)$: this has the effect of weighting the sparse regularizing term of (4).

We believe that the techniques of this paper can be adapted to deeper DLNs, i.e., when $\theta_i \propto u_i^l$, $l > 2$. One would only need to assume additionally that $k_1 = \cdots = k_d$. In this case, the time rescaling to have a limiting trajectory would change from $\log 1/\varepsilon$ for $l = 2$ to $\varepsilon^{2/l-1}$ for $l > 2$. Moreover, in this latter case, one observes that the effective regularization in Theorem 1 depends on the constants $C_1, \dots, C_d$ (but is still linear in $\boldsymbol{\theta}$). This is observed by repeating the proof of Section 4, redefining $w_i^{(\varepsilon)}$ as $(\theta_i/\varepsilon)^{2/l-1}$. We have omitted this adaptation for simplicity.

Finally, we note that when a ℓ^2 penalization on $\boldsymbol{u}$ (or on $\boldsymbol{u}$ and $\boldsymbol{v}$) is added to f, DLNs are related to iterative reweighted least-squares, a reparametrization of the Lasso problem appreciated for computational purposes, see (Bach et al., 2012, Section 5) or (Poon and Peyré, 2021). However, in this paper, there is no explicit ℓ^2 penalty on $\boldsymbol{u}$ and thus no explicit ℓ^1 penality on $\boldsymbol{\theta}$.

Incremental learning. Incremental learning describes some learning curves observed in human and machine learning that are almost piecewise constant: they consist of stages where little progress is made, separated by sharp transitions. For instance, this phenomenon occurs in non-diagonal linear networks (Saxe et al., 2019; Gidel et al., 2019; Gissin et al., 2019; Arora et al., 2019; Chou et al., 2020; Li et al., 2020), in tensor decomposition (Ge et al., 2021; Razin et al., 2021, 2022; Hariz et al., 2022) and in shallow ReLU networks (Boursier et al., 2022). In general, obtaining a mathematical description of the process—of the times of the transitions and the progress made—is mathematically challenging. To the best of our knowledge, existing works obtain a rigorous and complete mathematical description only in "separable" cases where the learning dynamics can be separated into several one-dimensional learning dynamics. For instance, Gissin et al. (2019) study DLNs but only in the special case $\boldsymbol{M} = \boldsymbol{I}_d$. As a consequence, a major contribution of our work is to describe precisely some *non-separable* incremental learning dynamics.

Heteroclinic networks. From a dynamical systems perspective, the dynamics (3) form a heteroclinic network (Bakhtin, 2011): it has several fixed points (also called *saddle points*) $(\boldsymbol{\theta}_*^{(I)})_{I\subset\{1,\dots,d\}}$ connected by geodesics of the flow. Such a dynamical system spends large amounts of time in the vicinity of fixed points, with sharp transitions between them. In our case, this is closely related to incremental learning. For our dynamical system, we describe the sequence of visited fixed points and the transition times. The paper of Jacot et al. (2021) attempted a similar study for linear networks; we prove rigorously such results in the special case of *diagonal* linear networks.

Lotka–Volterra equations. To finish, we note that the quadratic system (3) of ordinary differential equations are Lotka–Volterra (LV) equations (Hofbauer and Sigmund, 1998; Baigent, 2017). Traditionally, in mathematical biology, these equations represent the evolution the populations sizes $\theta_1, \dots, \theta_d$ of d interacting species. The parameter r_i represents the intrinsic growth of population i while the parameter M_{ij} represents the interaction between populations i and j.

This point of view, and in particular the paper of Goh (1979), inspired the author to use the function (10) in the proof of Theorem 1. In general, our paper can be interpreted as a study of LV equations for cooperative and symmetric interactions from infinitesimal initial population sizes. To the best of our knowledge, such a study did not exist in the literature on LV equations; its implications will be the subject of a forthcoming paper.

4. Proof of Theorem 1

In this proof, we use both time variables t and s, with $t = s \log \frac{1}{\varepsilon}$. As it is frequent in the literature on ordinary differential equations (ODEs), we abusively use the same notation for functions of t and s. For instance, by convention, $\boldsymbol{\theta}^{(\varepsilon)}(s) := \boldsymbol{\theta}^{(\varepsilon)}(t)$ with $t = s \log \frac{1}{\varepsilon}$. In fact, we often drop the dependence on time. For instance, $\boldsymbol{\theta}^{(\varepsilon)} := \boldsymbol{\theta}^{(\varepsilon)}(s) = \boldsymbol{\theta}^{(\varepsilon)}(t)$.

We start with a crude estimate of the trajectories $\boldsymbol{\theta}^{(\varepsilon)}(t)$ that is useful several times later in the proof.

Lemma 2 *The trajectory $\boldsymbol{\theta}^{(\varepsilon)}(t)$ is bounded uniformly for $\varepsilon \in (0, 1]$ and $t \in \mathbb{R}_{\geqslant 0}$, i.e., there exists a constant $B > 0$ such that $\forall \varepsilon \in (0,1], \forall t \in \mathbb{R}_{\geqslant 0}, \|\boldsymbol{\theta}^{(\varepsilon)}(t)\| \leqslant B$.*

Proof As Equation (3) is a (reparametrized) gradient flow of f, f is a Lyapunov function, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{\theta}) = \sum_{i=1}^{d} \frac{\mathrm{d}f}{\mathrm{d}\theta_i} \frac{\mathrm{d}\theta_i}{\mathrm{d}t} \underset{(3)}{=} -\sum_{i=1}^{d} \theta_i \left(\frac{\mathrm{d}f}{\mathrm{d}\theta_i} \right)^2 \leqslant 0 \, .$$

Thus for all $\varepsilon \in (0, 1]$, for all $t \geqslant 0$,

$$f(\boldsymbol{\theta}^{(\varepsilon)}(t)) \leqslant f(\boldsymbol{\theta}^{(\varepsilon)}(0)) \leqslant \sup_{\varepsilon \in (0,1]} f(\boldsymbol{\theta}^{(\varepsilon)}(0)) \, . \tag{5}$$

This supremum is finite as f is continuous and $\boldsymbol{\theta}^{(\varepsilon)}(0)$ is uniformly bounded for $\varepsilon \in (0, 1]$. Further, as $\boldsymbol{M}$ is positive definite (Proposition 1),

$$f(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y}\|^2 - \langle \boldsymbol{r}, \boldsymbol{\theta} \rangle + \frac{1}{2} \langle \boldsymbol{\theta}, \boldsymbol{M}\boldsymbol{\theta} \rangle \to \infty \qquad \text{as } \|\boldsymbol{\theta}\| \to \infty \, .$$

Thus the uniform bound (5) implies a uniform bound on $\|\boldsymbol{\theta}^{(\varepsilon)}(t)\|$. ■

The central idea of the proof of Theorem 1 is to keep track of the size of the coordinates of $\boldsymbol{\theta}^{(\varepsilon)}(t)$, in order to be able to determine which coordinates of $\boldsymbol{\theta}^{(\varepsilon)}(t)$ are activated depending on time t. More precisely, define

$$w_i^{(\varepsilon)} = \frac{\log \theta_i^{(\varepsilon)}}{\log \varepsilon} \, . \tag{6}$$

Equivalently, this gives $\theta_i^{(\varepsilon)} = \varepsilon^{w_i^{(\varepsilon)}}$. This logarithmic transformation of the coordinates is particularly convenient because its time derivative is affine in $\boldsymbol{\theta}^{(\varepsilon)}$:

$$\frac{\mathrm{d}w_i^{(\varepsilon)}}{\mathrm{d}s} = \frac{\mathrm{d}t}{\mathrm{d}s}\frac{\mathrm{d}w_i^{(\varepsilon)}}{\mathrm{d}t} = \left(\log\frac{1}{\varepsilon}\right)\frac{1}{\theta_i^{(\varepsilon)}\log\varepsilon}\frac{\mathrm{d}\theta_i^{(\varepsilon)}}{\mathrm{d}t} \underset{(3)}{=} \sum_{j=1}^{d} M_{ij}\theta_j^{(\varepsilon)} - r_i\,,$$

or, using the vector notation $\boldsymbol{w}^{(\varepsilon)} = (w_1^{(\varepsilon)}, \dots, w_d^{(\varepsilon)})$,

$$\frac{\mathrm{d}\boldsymbol{w}^{(\varepsilon)}}{\mathrm{d}s} = \boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{r}\,.$$

Our proof technique determines the limit of $\boldsymbol{w}^{(\varepsilon)}$ as $\varepsilon \to 0$. The limit is described as the Lagrange multiplier of an optimization problem closely related to (4). We start with a brief reminder on duality in optimization.

Proposition 4 *Let $\boldsymbol{q}, \boldsymbol{z} \in \mathbb{R}^d$. The two following statements are equivalent:*

1. *$\boldsymbol{z}$ is the unique minimizer of the constrained optimization problem*

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d,\,\boldsymbol{\theta}\geqslant\boldsymbol{0}}. \left\{\langle\boldsymbol{q},\boldsymbol{\theta}\rangle + \frac{1}{2}\langle\boldsymbol{\theta},\boldsymbol{M}\boldsymbol{\theta}\rangle\right\}. \tag{7}$$

2. *There exists $\boldsymbol{w} \in \mathbb{R}^d$ such that $(\boldsymbol{w}, \boldsymbol{z})$ is the unique solution of*

$$\begin{aligned} &\boldsymbol{w} = \boldsymbol{q} + \boldsymbol{M}\boldsymbol{z}\,,\\ &\boldsymbol{w} \geqslant \boldsymbol{0}\,, \qquad \boldsymbol{z} \geqslant \boldsymbol{0}\,, \qquad \boldsymbol{w}^\top\boldsymbol{z} = 0\,. \end{aligned}$$

The four conditions above form a so-called linear complementarity problem (LCP)*, where $\boldsymbol{q}$ and $\boldsymbol{M}$ are the parameters and $\boldsymbol{w}$ and $\boldsymbol{z}$ are the variables.*

The linear complementarity problem is widely studied; see for instance the monograph of Cottle et al. (2009) or Appendix A. In this connection with the quadratic programming problem (7), the variable $\boldsymbol{w}$ should be seen as the Lagrange multiplier associated to the constraint $\boldsymbol{\theta} \geqslant \boldsymbol{0}$. The LCP expresses the Karush–Kuhn–Tucker (KKT) conditions for optimality to hold. More precisely, $\boldsymbol{w} = \boldsymbol{q} + \boldsymbol{M}\boldsymbol{z}$ is a condition of stationarity of the Lagrangian; $\boldsymbol{w} \geqslant \boldsymbol{0}$ and $\boldsymbol{z} \geqslant \boldsymbol{0}$ are respectively dual and primal feasibility conditions; and $\boldsymbol{w}^\top\boldsymbol{z} = 0$ is a complementarity slackness condition. Put together, the conditions $\boldsymbol{w} \geqslant \boldsymbol{0}$, $\boldsymbol{z} \geqslant \boldsymbol{0}$ and $\boldsymbol{w}^\top\boldsymbol{z} = 0$ impose that for all $i \in \{1, \dots, d\}$, either $w_i = 0$ or $z_i = 0$.

Proposition 4 is classical; nevertheless we detail the appropriate references in Appendix E.

We are now in position to describe the asymptotic behavior of $w_i^{(\varepsilon)} = \frac{\log\theta_i^{(\varepsilon)}}{\log\varepsilon}$. Define

$$\boldsymbol{z}^{(\varepsilon)}(s) = \int_0^s \mathrm{d}u\,\boldsymbol{\theta}^{(\varepsilon)}(u)\,.$$

Proposition 5 *Let $(\boldsymbol{w}(s), \boldsymbol{z}(s))$ be the unique solution of the linear complementarity problem*

$$\begin{aligned} &\boldsymbol{w} = \boldsymbol{k} - s\boldsymbol{r} + \boldsymbol{M}\boldsymbol{z}\,, \\ &\boldsymbol{w} \geqslant \boldsymbol{0}\,, \qquad \boldsymbol{z} \geqslant \boldsymbol{0}\,, \qquad \boldsymbol{w}^\top \boldsymbol{z} = 0\,. \end{aligned} \tag{8}$$

Then $\boldsymbol{w}^{(\varepsilon)}(s)$ and $\boldsymbol{z}^{(\varepsilon)}(s)$ converge respectively to $\boldsymbol{w}(s)$ and $\boldsymbol{z}(s)$ as $\varepsilon \to 0$, uniformly on compact subsets of $\mathbb{R}_{\geqslant 0}$.

Proof In this proof, we define $\varepsilon_0 > 0$ as in Lemma 1 and $B > 0$ as in Lemma 2. Further we take $\varepsilon_1 = \min\left(\varepsilon_0, \frac{1}{2}\right)$ and assume that $\varepsilon \leqslant \varepsilon_1$.

Fix $S > 0$ and define the continuous functions

$$\begin{aligned} &\boldsymbol{\varphi}^{(\varepsilon)} : s \in [0, S] \mapsto \left(\boldsymbol{w}^{(\varepsilon)}(s), \boldsymbol{z}^{(\varepsilon)}(s)\right) \in (\mathbb{R}^d)^2\,, \\ &\boldsymbol{\varphi} : s \in [0, S] \mapsto (\boldsymbol{w}(s), \boldsymbol{z}(s)) \in (\mathbb{R}^d)^2\,. \end{aligned}$$

We want to show that $\boldsymbol{\varphi}^{(\varepsilon)} \to \boldsymbol{\varphi}$ uniformly as $\varepsilon \to 0$. First, we use the Arzelà–Ascoli theorem to check that the set $\{\boldsymbol{\varphi}^{(\varepsilon)}, \varepsilon \in (0, \varepsilon_1)\}$ is relatively compact in the space of continuous functions from $[0, S]$ to $(\mathbb{R}^d)^2$. The reader can consult the monograph of Dunford and Schwartz (1988, Section IV.6) for a reference on the Arzelà–Ascoli theorem for real-valued functions; the multidimensional extension is straightforward.

- For $\varepsilon \in (0, \varepsilon_1)$, $s \in [0, S]$, we bound $\|\boldsymbol{\varphi}^{(\varepsilon)}(s)\|^2 = \|\boldsymbol{w}^{(\varepsilon)}(s)\|^2 + \|\boldsymbol{z}^{(\varepsilon)}(s)\|^2$. We bound the two terms separately.

  – From Lemmas 1 and 2,

  $$C_i \varepsilon^{k_i} = \theta_i^{(\varepsilon)}(0) \leqslant \theta_i^{(\varepsilon)}(s) \leqslant B\,,$$

  thus ($\log \varepsilon < 0$),

  $$\frac{\log B}{\log \varepsilon} \leqslant w_i^{(\varepsilon)} = \frac{\log \theta_i^{(\varepsilon)}}{\log \varepsilon} \leqslant \frac{\log C_i}{\log \varepsilon} + k_i\,.$$

  As $\varepsilon \leqslant 1/2$, $\log \varepsilon$ is bounded away from 0. This shows that $w_i^{(\varepsilon)}$ is bounded uniformly for $\varepsilon \in (0, \varepsilon_1]$ and $s \in [0, S]$.

  – From Lemma 2,

  $$\|\boldsymbol{z}^{(\varepsilon)}(s)\| \leqslant \int_0^s \mathrm{d}u\, \|\boldsymbol{\theta}^{(\varepsilon)}(u)\| \leqslant sB \leqslant SB\,.$$

  The two points above show that $\|\boldsymbol{\varphi}^{(\varepsilon)}(s)\|^2$ is bounded uniformly for $\varepsilon \in (0, \varepsilon_1]$ and $s \in [0, S]$.

- The square norm of the derivative

$$\begin{aligned} \left\|\frac{\mathrm{d}\boldsymbol{\varphi}^{(\varepsilon)}}{\mathrm{d}s}(s)\right\|^2 &= \left\|\frac{\mathrm{d}\boldsymbol{w}^{(\varepsilon)}}{\mathrm{d}s}(s)\right\|^2 + \left\|\frac{\mathrm{d}\boldsymbol{z}^{(\varepsilon)}}{\mathrm{d}s}(s)\right\|^2 \\ &= \left\|\boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)}(s) - \boldsymbol{r}\right\|^2 + \left\|\boldsymbol{\theta}^{(\varepsilon)}(s)\right\|^2 \end{aligned}$$

is bounded uniformly for $\varepsilon \in (0, \varepsilon_1]$ and $s \in [0, S]$ by Lemma 2. Thus the set $\{\boldsymbol{\varphi}^{(\varepsilon)}, \varepsilon \in (0, \varepsilon_1)\}$ is equicontinuous.

The two points above show that we can apply the Arzelà–Ascoli theorem: $\{\boldsymbol{\varphi}^{(\varepsilon)}, \varepsilon \in (0, \varepsilon_1)\}$ is relatively compact in the space of continuous functions from $[0, S]$ to $(\mathbb{R}^d)^2$. To conclude on Proposition 5, it is then sufficient to show that the only subsequential uniform limit of $\boldsymbol{\varphi}^{(\varepsilon)}$ as $\varepsilon \to 0$ is $\boldsymbol{\varphi}$.

Let $\boldsymbol{\varphi}' = (\boldsymbol{w}', \boldsymbol{z}')$ be a subsequential uniform limit of $\boldsymbol{\varphi}^{(\varepsilon)} = (\boldsymbol{w}^{(\varepsilon)}, \boldsymbol{z}^{(\varepsilon)})$ as $\varepsilon \to 0$. There exists $\varepsilon(n) \in (0, \varepsilon_1]$ such that $\varepsilon(n) \to 0$ and $\boldsymbol{\varphi}^{(\varepsilon(n))} \to \boldsymbol{\varphi}'$ uniformly as $n \to \infty$. Then $\boldsymbol{w}^{(\varepsilon(n))} \to \boldsymbol{w}'$ and $\boldsymbol{z}^{(\varepsilon(n))} \to \boldsymbol{z}'$ uniformly as $n \to \infty$. We check that $(\boldsymbol{w}', \boldsymbol{z}')$ is a solution of the LCP (8).

- First,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\left(\boldsymbol{w}^{(\varepsilon(n))}(s) - \boldsymbol{k} + s\boldsymbol{r} - \boldsymbol{M}\boldsymbol{z}^{(\varepsilon(n))}(s)\right) &= \frac{\mathrm{d}\boldsymbol{w}^{(\varepsilon(n))}}{\mathrm{d}s}(s) + \boldsymbol{r} - \boldsymbol{M}\frac{\mathrm{d}\boldsymbol{z}^{(\varepsilon(n))}}{\mathrm{d}s}(s) \\ &= \boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon(n))}(s) - \boldsymbol{r} + \boldsymbol{r} - \boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon(n))}(s) \\ &= \boldsymbol{0}\,.\end{aligned}$$

  Thus $\boldsymbol{w}^{(\varepsilon(n))}(s) - \boldsymbol{k} + s\boldsymbol{r} - \boldsymbol{M}\boldsymbol{z}^{(\varepsilon(n))}(s)$ is constant in s, equal to its initial value $\boldsymbol{w}^{(\varepsilon(n))}(0) - \boldsymbol{k}$. Moreover, for all i,

$$w_i^{(\varepsilon(n))}(0) - k_i = \frac{\log C_i \varepsilon(n)^{k_i}}{\log \varepsilon(n)} - k_i = \frac{\log C_i}{\log \varepsilon(n)} \xrightarrow[n\to\infty]{} 0\,.$$

  Thus $\boldsymbol{w}^{(\varepsilon(n))}(0) - \boldsymbol{k} = \boldsymbol{w}^{(\varepsilon(n))}(s) - \boldsymbol{k} + s\boldsymbol{r} - \boldsymbol{M}\boldsymbol{z}^{(\varepsilon(n))}(s) \to \boldsymbol{0}$ as $n \to \infty$. By identification of the limit, we have $\boldsymbol{w}'(s) - \boldsymbol{k} + s\boldsymbol{r} - \boldsymbol{M}\boldsymbol{z}'(s) = \boldsymbol{0}$.

- Second, using Lemma 2 and that $\log \varepsilon(n) < 0$,

$$w_i^{(\varepsilon(n))}(s) = \frac{\log \theta_i^{(\varepsilon(n))}(s)}{\log \varepsilon(n)} \geqslant \frac{\log B}{\log \varepsilon(n)}\,,$$

  thus taking $n \to \infty$, we obtain $\boldsymbol{w}' \geqslant \boldsymbol{0}$.

- Third, we have $\boldsymbol{z}^{(\varepsilon(n))}(s) \geqslant \boldsymbol{0}$ trivially from the definition, and thus taking $n \to \infty$, we obtain $\boldsymbol{z}' \geqslant \boldsymbol{0}$.

- Finally,

$$\begin{aligned}\left|\boldsymbol{w}^{(\varepsilon(n))}(s)^\top \boldsymbol{z}^{(\varepsilon(n))}(s)\right| &\leqslant \sum_i \left|w_i^{(\varepsilon(n))}(s) z_i^{(\varepsilon(n))}(s)\right| \\ &= \sum_i \frac{|\log \theta_i^{(\varepsilon(n))}(s)|}{|\log \varepsilon(n)|} \int_0^s \mathrm{d}u\, \theta_i^{(\varepsilon(n))}(u) \\ &\leqslant \frac{s}{|\log \varepsilon(n)|} \sum_i |\log \theta_i^{(\varepsilon(n))}(s)| \theta_i^{(\varepsilon(n))}(s)\,,\end{aligned}$$

where in this last inequality we use Lemma 1. The function $x \mapsto |\log x| x$ can be continuously extended in 0; it is thus bounded on $[0, B]$. Thus

$$\left| \boldsymbol{w}^{(\varepsilon(n))}(s)^\top \boldsymbol{z}^{(\varepsilon(n))}(s) \right| \leqslant \frac{sd}{|\log \varepsilon(n)|} \max_{x \in [0,B]} |\log x| x \,.$$

Taking $n \to \infty$, we obtain $\boldsymbol{w}'(s)^\top \boldsymbol{z}'(s) = 0$.

The four points above show that for all $s \in [0, S]$, $(\boldsymbol{w}'(s), \boldsymbol{z}'(s))$ is a solution of the LCP (8). As the solution is unique (Proposition 6), $\boldsymbol{\varphi}' = (\boldsymbol{w}', \boldsymbol{z}') = (\boldsymbol{w}, \boldsymbol{z}) = \boldsymbol{\varphi}$. Thus $\boldsymbol{\varphi}$ is the unique subsequent limit of $\boldsymbol{\varphi}^{(\varepsilon)}$ as $\varepsilon \to 0$. Thus $\boldsymbol{\varphi}^{(\varepsilon)} \xrightarrow[\varepsilon \to 0]{} \boldsymbol{\varphi}$ uniformly on $[0, S]$. ■

We now show how the proof of Theorem 1 essentially follows from Proposition 5.

Proof [Proof of Theorem 1] First, note that it is shown in Proposition 4 that there is a unique minimizer to (4). We continue by proving successively the three points of the theorem.

(1) The fact that $\boldsymbol{\mu}(s)$ is nondecreasing follows from the connection with the LCP (Proposition 4) and the antitonicity property of the solution of the LCP (Proposition 7). We detail this argument.

Recall that $\boldsymbol{\mu}(s)$ is the unique minimizer of

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d,\, \boldsymbol{\theta} \geqslant \boldsymbol{0}}. \left\{ f(\boldsymbol{\theta}) + \frac{1}{s}\langle \boldsymbol{k}, \boldsymbol{\theta} \rangle = \frac{1}{2}\|\boldsymbol{y}\|^2 + \left\langle \frac{1}{s}\boldsymbol{k} - \boldsymbol{r}, \boldsymbol{\theta} \right\rangle + \frac{1}{2}\langle \boldsymbol{\theta}, \boldsymbol{M}\boldsymbol{\theta} \rangle \right\}$$

Thus by Proposition 4, there exists $\boldsymbol{v}(s) \in \mathbb{R}^d$ such that $(\boldsymbol{v}(s), \boldsymbol{\mu}(s))$ is the unique solution of the LCP

$$\begin{aligned} &\boldsymbol{v} = \frac{1}{s}\boldsymbol{k} - \boldsymbol{r} + \boldsymbol{M}\boldsymbol{\mu} \,, \\ &\boldsymbol{v} \geqslant \boldsymbol{0} \,, \qquad \boldsymbol{\mu} \geqslant \boldsymbol{0} \,, \qquad \boldsymbol{v}^\top \boldsymbol{\mu} = 0 \,. \end{aligned}$$

Let us make a side remark for later purposes. We rewrite the LCP above as

$$\begin{aligned} &s\boldsymbol{v} = \boldsymbol{k} - s\boldsymbol{r} + \boldsymbol{M}(s\boldsymbol{\mu}) \,, \\ &s\boldsymbol{v} \geqslant \boldsymbol{0} \,, \qquad s\boldsymbol{\mu} \geqslant \boldsymbol{0} \,, \qquad (s\boldsymbol{v})^\top (s\boldsymbol{\mu}) = 0 \,. \end{aligned}$$

Thus $(s\boldsymbol{v}(s), s\boldsymbol{\mu}(s))$ is a solution of the LCP (8), of which $(\boldsymbol{w}(s), \boldsymbol{z}(s))$ is also the solution. By unicity of the solution of the LCP (Proposition 6),

$$\boldsymbol{z}(s) = s\boldsymbol{\mu}(s) \,. \tag{9}$$

We now return to the proof of point (1) of the theorem. Consider $s_1 \leqslant s_2$. Then $(\boldsymbol{v}(s_1), \boldsymbol{\mu}(s_1))$ (resp. $(\boldsymbol{v}(s_2), \boldsymbol{\mu}(s_2))$) is the unique solution of the LCP with parameters $\boldsymbol{q}^{(1)} = \frac{1}{s_1}\boldsymbol{k} - \boldsymbol{r}$ (resp. $\boldsymbol{q}^{(2)} = \frac{1}{s_2}\boldsymbol{k} - \boldsymbol{r}$) and $\boldsymbol{M}$. Note that as $\boldsymbol{k} \geqslant \boldsymbol{0}$, $\boldsymbol{q}^{(1)} \geqslant \boldsymbol{q}^{(2)}$. As $\boldsymbol{M}$ is symmetric positive definite (Proposition 1) with non-positive off-diagonal entries (Assumption (A2)), we apply the antitonicity property (Proposition 7) and obtain that $\boldsymbol{\mu}(s_1) \leqslant \boldsymbol{\mu}(s_2)$.

(3) Abusing again on the notation of the time indexations, we have

$$\frac{1}{s\log\frac{1}{\varepsilon}}\int_0^{s\log\frac{1}{\varepsilon}} \mathrm{d}t\,\boldsymbol{\theta}^{(\varepsilon)}(t) = \frac{1}{s}\int_0^s \mathrm{d}u\,\boldsymbol{\theta}^{(\varepsilon)}(u) = \frac{1}{s}\boldsymbol{z}^{(\varepsilon)}(s)\,.$$

By Proposition 5, $\boldsymbol{z}^{(\varepsilon)}(s)$ converges to $\boldsymbol{z}(s)$ uniformly on compact subsets of $\mathbb{R}_{\geqslant 0}$. Thus $\frac{1}{s}\boldsymbol{z}^{(\varepsilon)}(s)$ converges uniformly to $\frac{1}{s}\boldsymbol{z}(s)$ on compact subsets of $\mathbb{R}_{>0}$. (Note that as the factor $1/s$ diverges at 0, we need to consider compact subsets bounded away from 0.) As $\frac{1}{s}\boldsymbol{z}(s) = \boldsymbol{\mu}(s)$ (Equation (9)), this concludes point (3) of the theorem.

(2) Fix $p \in \{1,\dots,q,q+1\}$ and $u, v \in \mathbb{R}_{>0}$ such that $s_{p-1} < u < v < s_p$ (with the conventions $s_0 = 0$, $s_{q+1} = \infty$). Let I be the constant value of $I(s)$ for $s \in (s_{p-1}, s_p)$. To prove point (2) of the theorem, it is sufficient to show $\boldsymbol{\theta}^{(\varepsilon)}\left(s\log\frac{1}{\varepsilon}\right) \xrightarrow[\varepsilon\to 0]{} \boldsymbol{\theta}_*^{(I)}$ uniformly for $s \in [u, v]$.

For $\boldsymbol{\theta} \in \mathbb{R}^d$, define

$$V(\boldsymbol{\theta}) = \sum_{i\in I}\left(\theta_i - \theta_{*,i}^{(I)}\log\theta_i\right)\,. \tag{10}$$

The function V is inspired from a Lyapunov function used in the study of Lotka–Volterra equations (Goh, 1979). Here, we show that V is "almost" decreasing on the interval (s_{p-1}, s_p). We first compute

$$\frac{\mathrm{d}}{\mathrm{d}t}V(\boldsymbol{\theta}^{(\varepsilon)}) \underset{(3)}{=} \sum_{i\in I}\left(\theta_i^{(\varepsilon)} - \theta_{*,i}^{(I)}\right)\left(r_i - \left(\boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)}\right)_i\right)\,. \tag{11}$$

Using the definition of $\boldsymbol{\theta}_*^{(I)}$ in Proposition 2, we have

$$(\boldsymbol{M}\boldsymbol{\theta}_*^{(I)})_I = \boldsymbol{M}_{II}\boldsymbol{\theta}_{*,I}^{(I)} + \boldsymbol{M}_{II^c}\boldsymbol{\theta}_{*,I^c}^{(I)} = \boldsymbol{M}_{II}\boldsymbol{M}_{II}^{-1}\boldsymbol{r}_I + \boldsymbol{M}_{II^c}\boldsymbol{0} = \boldsymbol{r}_I\,,$$

thus we can rewrite (11) as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V(\boldsymbol{\theta}^{(\varepsilon)}) &= \sum_{i\in I}\left(\theta_i^{(\varepsilon)} - \theta_{*,i}^{(I)}\right)\left(\left(\boldsymbol{M}\boldsymbol{\theta}_*^{(I)}\right)_i - \left(\boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)}\right)_i\right)\\ &= -\left\langle\boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{\theta}_*^{(I)}, \boldsymbol{M}(\boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{\theta}_*^{(I)})\right\rangle - \sum_{i\notin I}\theta_i^{(\varepsilon)}\left(\left(\boldsymbol{M}\boldsymbol{\theta}_*^{(I)}\right)_i - \left(\boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)}\right)_i\right)\,.\end{aligned}$$

Fix u', v' such that $s_{p-1} < u' < u < v < v' < s_p$. We integrate the equality above for $s \in [u', v']$:

$$\begin{aligned}V(\boldsymbol{\theta}^{(\varepsilon)}(v')) - V(\boldsymbol{\theta}^{(\varepsilon)}(u')) &= \int_{u'}^{v'}\mathrm{d}s\frac{\mathrm{d}}{\mathrm{d}s}V(\boldsymbol{\theta}^{(\varepsilon)})\\ &= \left(\log\frac{1}{\varepsilon}\right)\int_{u'}^{v'}\mathrm{d}s\frac{\mathrm{d}}{\mathrm{d}t}V(\boldsymbol{\theta}^{(\varepsilon)})\\ &= \left(\log\frac{1}{\varepsilon}\right)\Big[-\int_{u'}^{v'}\mathrm{d}s\left\langle\boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{\theta}_*^{(I)}, \boldsymbol{M}(\boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{\theta}_*^{(I)})\right\rangle\\ &\qquad - \sum_{i\notin I}\int_{u'}^{v'}\mathrm{d}s\,\theta_i^{(\varepsilon)}\left(\left(\boldsymbol{M}\boldsymbol{\theta}_*^{(I)}\right)_i - \left(\boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)}\right)_i\right)\Big]\,.\end{aligned}$$

Denote $\lambda_{\min}(\boldsymbol{M})$ the minimal eigenvalue of $\boldsymbol{M}$. By Proposition 1, $\lambda_{\min}(\boldsymbol{M}) > 0$.

$$\begin{aligned}\int_{u'}^{v'} \mathrm{d}s \left\|\boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{\theta}_*^{(I)}\right\|^2 &\leqslant \frac{1}{\lambda_{\min}(\boldsymbol{M})} \int_{u'}^{v'} \mathrm{d}s \left\langle \boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{\theta}_*^{(I)}, \boldsymbol{M}(\boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{\theta}_*^{(I)})\right\rangle \\ &= \frac{1}{\lambda_{\min}(\boldsymbol{M})} \Big[\frac{1}{\log \frac{1}{\varepsilon}} \left(V\left(\boldsymbol{\theta}^{(\varepsilon)}(u')\right) - V\left(\boldsymbol{\theta}^{(\varepsilon)}(v')\right)\right) \\ &\qquad - \sum_{i \notin I} \int_{u'}^{v'} \mathrm{d}s\, \theta_i^{(\varepsilon)} \left(\left(\boldsymbol{M}\boldsymbol{\theta}_*^{(I)}\right)_i - \left(\boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)}\right)_i \right) \Big] .\end{aligned} \tag{12}$$

Take $\varepsilon_0 > 0$ as defined in Lemma 1 and now assume $\varepsilon \leqslant \min(1, \varepsilon_0)$ so that both Lemmas 1 and 2 apply. Then we have the following estimates:

- Fix $i \notin I$. From Lemma 2, there exists a constant C independent of ε such that

$$\begin{aligned}\left| \int_{u'}^{v'} \mathrm{d}s\, \theta_i^{(\varepsilon)} \left(\left(\boldsymbol{M}\boldsymbol{\theta}_*^{(I)}\right)_i - \left(\boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)}\right)_i \right) \right| &\leqslant C \int_{u'}^{v'} \mathrm{d}s\, \theta_i^{(\varepsilon)} \\ &= C \left(v' \frac{1}{v'} \int_0^{v'} \mathrm{d}s\, \theta_i^{(\varepsilon)} - u' \frac{1}{u'} \int_0^{u'} \mathrm{d}s\, \theta_i^{(\varepsilon)} \right)\end{aligned}$$

  Using Theorem 1.(3), this last quantity converges as $\varepsilon \to 0$ to $v'\mu_i(v') - u'\mu_i(u')$, which is equal to 0 as $i \notin I$. We thus have

$$\int_{u'}^{v'} \mathrm{d}s\, \theta_i^{(\varepsilon)} \left(\left(\boldsymbol{M}\boldsymbol{\theta}_*^{(I)}\right)_i - \left(\boldsymbol{M}\boldsymbol{\theta}^{(\varepsilon)}\right)_i \right) \xrightarrow[\varepsilon \to 0]{} 0 . \tag{13}$$

- Further, if $s = u'$ or $s = v'$,

$$\begin{aligned} V(\boldsymbol{\theta}^{(\varepsilon)}(s)) &= \sum_{i \in I} \left(\theta_i^{(\varepsilon)}(s) - \theta_{*,i}^{(I)} \log \theta_i^{(\varepsilon)}(s) \right) \\ &= \sum_{i \in I} \theta_i^{(\varepsilon)}(s) + \left(\log \frac{1}{\varepsilon} \right) \sum_{i \in I} \theta_{*,i}^{(I)} w_i^{(\varepsilon)}(s) \end{aligned} \tag{14}$$

  by the definition of $w_i^{(\varepsilon)}$ in Equation (6). The first term $\sum_{i \in I} \theta_i^{(\varepsilon)}(s)$ is bounded independently of ε by Lemma 2. Further, for all $i \in I$, $\mu_i(s) > 0$ and thus by Equation (9), $z_i(s) > 0$. By the complementary slackness of $\boldsymbol{z}(s)$ and $\boldsymbol{w}(s)$, we must have $w_i(s) = 0$. As a consequence, using Proposition 5,

$$\sum_{i \in I} \theta_{*,i}^{(I)} w_i^{(\varepsilon)}(s) \xrightarrow[\varepsilon \to 0]{} \sum_{i \in I} \theta_{*,i}^{(I)} w_i(s) = 0 .$$

  Returning to Equation (14), we obtain that

$$V(\boldsymbol{\theta}^{(\varepsilon)}(u')) = o\left(\log \frac{1}{\varepsilon}\right) , \qquad V(\boldsymbol{\theta}^{(\varepsilon)}(v')) = o\left(\log \frac{1}{\varepsilon}\right) . \tag{15}$$

Putting together Equations (12), (13) and (15), we obtain

$$\int_{u'}^{v'} \mathrm{d}s \|\boldsymbol{\theta}^{(\varepsilon)} - \boldsymbol{\theta}_*^{(I)}\|^2 \xrightarrow[\varepsilon \to 0]{} 0\,. \tag{16}$$

To conclude on uniform convergence, we use an elementary argument based on monotonicity:

$$\max_{s\in[u,v]} \left|\theta_i^{(\varepsilon)}(s) - \theta_{*,i}^{(I)}\right| = \max_{s\in[u,v]} \max\left(\theta_i^{(\varepsilon)}(s) - \theta_{*,i}^{(I)}, \theta_{*,i}^{(I)} - \theta_i^{(\varepsilon)}(s)\right)\,. \tag{17}$$

By Lemma 1, for all $s \in [u,v]$ and $s' \in [v,v']$, $\theta_i^{(\varepsilon)}(s) - \theta_{*,i}^{(I)} \leqslant \theta_i^{(\varepsilon)}(s') - \theta_{*,i}^{(I)}$, thus

$$\theta_i^{(\varepsilon)}(s) - \theta_{*,i}^{(I)} \leqslant \frac{1}{v'-v}\int_v^{v'} \mathrm{d}s' \left(\theta_i^{(\varepsilon)}(s') - \theta_{*,i}^{(I)}\right)\,.$$

Using Hölder's inequality, we obtain

$$\begin{aligned}\theta_i^{(\varepsilon)}(s) - \theta_{*,i}^{(I)} &\leqslant \frac{1}{\sqrt{v'-v}}\left(\int_v^{v'} \mathrm{d}s' \left(\theta_i^{(\varepsilon)}(s') - \theta_{*,i}^{(I)}\right)^2\right)^{1/2} \\ &\leqslant \frac{1}{\sqrt{v'-v}}\left(\int_v^{v'} \mathrm{d}s' \left\|\boldsymbol{\theta}^{(\varepsilon)}(s') - \boldsymbol{\theta}_*^{(I)}\right\|^2\right)^{1/2}\,.\end{aligned}$$

Similarly,

$$\theta_{*,i}^{(I)} - \theta_i^{(\varepsilon)}(s) \leqslant \frac{1}{\sqrt{u-u'}}\left(\int_{u'}^{u} \mathrm{d}s' \left\|\boldsymbol{\theta}^{(\varepsilon)}(s') - \boldsymbol{\theta}_*^{(I)}\right\|^2\right)^{1/2}\,.$$

Finally, plugging these estimates back in Equation (17) and using Equation (16), we obtain

$$\begin{aligned}\max_{s\in[u,v]} \left|\theta_i^{(\varepsilon)}(s) - \theta_{*,i}^{(I)}\right| &\leqslant \max\Bigg(\frac{1}{\sqrt{v'-v}}\left(\int_v^{v'} \mathrm{d}s' \left\|\boldsymbol{\theta}^{(\varepsilon)}(s') - \boldsymbol{\theta}_*^{(I)}\right\|^2\right)^{1/2}, \\ &\qquad\qquad \frac{1}{\sqrt{u-u'}}\left(\int_{u'}^{u} \mathrm{d}s' \left\|\boldsymbol{\theta}^{(\varepsilon)}(s') - \boldsymbol{\theta}_*^{(I)}\right\|^2\right)^{1/2}\Bigg) \\ &\xrightarrow[\varepsilon\to 0]{} 0\,.\end{aligned}$$

This being true for all $i \in \{1,\dots,d\}$, we conclude that $\boldsymbol{\theta}^{(\varepsilon)}$ converges to $\boldsymbol{\theta}_*^{(I)}$ uniformly on $[u,v]$ and thus point (2) of the theorem holds.

■

5. Proof of Corollary 1

We first study under which condition on s we have $I(s) = \{1, \dots, d\}$. By definition of $I(s)$, this is equivalent to having $\boldsymbol{\mu}(s) > \mathbf{0}$ and thus $I(s) = \{1, \dots, d\}$ if and only if $f(\boldsymbol{\theta}) + \frac{1}{s}\langle \boldsymbol{k}, \boldsymbol{\theta}\rangle$ is minimized at a positive point. Note that

$$f(\boldsymbol{\theta}) + \frac{1}{s}\langle \boldsymbol{k}, \boldsymbol{\theta}\rangle = \frac{1}{2}\|\boldsymbol{y}\|^2 - \langle \boldsymbol{r} - \frac{1}{s}\boldsymbol{k}, \boldsymbol{\theta}\rangle + \frac{1}{2}\langle \boldsymbol{\theta}, \boldsymbol{M}\boldsymbol{\theta}\rangle$$

is minimized at $\boldsymbol{M}^{-1}\left(\boldsymbol{r} - \frac{1}{s}\boldsymbol{k}\right)$ thus

$$\begin{aligned} I(s) = \{1, \dots, d\} \quad &\Leftrightarrow \quad \boldsymbol{M}^{-1}\boldsymbol{r} - \frac{1}{s}\boldsymbol{M}^{-1}\boldsymbol{k} > \mathbf{0} \\ &\Leftrightarrow \quad s > s_*\,, \qquad s_* := \max_i \frac{(\boldsymbol{M}^{-1}\boldsymbol{k})_i}{(\boldsymbol{M}^{-1}\boldsymbol{r})_i}\,. \end{aligned} \tag{18}$$

Consider $s > s_*$. From Theorem 1.(2),

$$\boldsymbol{\theta}^{(\varepsilon)}\left(s \log\frac{1}{\varepsilon}\right) \xrightarrow[\varepsilon \to 0]{} \boldsymbol{\theta}_*^{(\{1,\dots,d\})}\,.$$

Take $\eta > 0$. Then there exists $\varepsilon_1 > 0$ such that for all $\varepsilon \leqslant \varepsilon_1$,

$$\left\|\boldsymbol{\theta}^{(\varepsilon)}\left(s \log\frac{1}{\varepsilon}\right) - \boldsymbol{\theta}_*^{(\{1,\dots,d\})}\right\| \leqslant \eta\,.$$

Thus $\tau_\eta^{(\varepsilon)} \leqslant s \log\frac{1}{\varepsilon}$. This being true for all $\varepsilon \leqslant \varepsilon_1$, we have $\limsup_{\varepsilon\to 0} \frac{\tau_\eta^{(\varepsilon)}}{\log 1/\varepsilon} \leqslant s$. This being true for all $s > s_*$, we have $\limsup_{\varepsilon\to 0} \frac{\tau_\eta^{(\varepsilon)}}{\log 1/\varepsilon} \leqslant s_*$.

We are left with showing that $\liminf_{\varepsilon\to 0} \frac{\tau_\eta^{(\varepsilon)}}{\log 1/\varepsilon} \geqslant s_*$. We assume that $\eta < \min_j \theta_{*,j}^{(\{1,\dots,d\})}$, which is possible as from Proposition 2, $\boldsymbol{\theta}_*^{(\{1,\dots,d\})} > \mathbf{0}$. Consider $s < s_*$. Then from (18), $I(s) \neq \{1, \dots, d\}$ thus we can consider $i \notin I(s)$. Assume further that $s \neq s_1, \dots, s_q$. Then by Theorem 1.(2),

$$\theta_i^{(\varepsilon)}\left(s \log\frac{1}{\varepsilon}\right) \xrightarrow[\varepsilon \to 0]{} \theta_{*,i}^{(I(s))} = 0\,.$$

Denote $\nu = \min_j \theta_{*,j}^{(\{1,\dots,d\})} - \eta > 0$. There exists $\varepsilon_2 > 0$ such that for all $\varepsilon \leqslant \varepsilon_2$,

$$\theta_i^{(\varepsilon)}\left(s \log\frac{1}{\varepsilon}\right) < \nu\,. \tag{19}$$

Define ε_0 as in Lemma 1 and assume $\varepsilon \leqslant \min(\varepsilon_0, \varepsilon_2)$ so that both Lemma 1 and Equation (19) apply. Then for all $t \leqslant s\log\frac{1}{\varepsilon}$,

$$\theta_i^{(\varepsilon)}(t) \underset{\text{(Lemma 1)}}{\leqslant} \theta_i^{(\varepsilon)}\left(s \log\frac{1}{\varepsilon}\right) \underset{\text{(Equation (19))}}{<} \nu\,.$$

Thus

$$\begin{aligned}\left\|\boldsymbol{\theta}^{(\varepsilon)}(t)-\boldsymbol{\theta}_*^{(\{1,\dots,d\})}\right\| &\geqslant \left|\theta_i^{(\varepsilon)}(t)-\theta_{*,i}^{(\{1,\dots,d\})}\right| \geqslant \theta_{*,i}^{(\{1,\dots,d\})}-\theta_i^{(\varepsilon)}(t)\\ &> \theta_{*,i}^{(\{1,\dots,d\})}-\nu \geqslant \min_j \theta_{*,j}^{(\{1,\dots,d\})}-\nu=\eta\,.\end{aligned}$$

This being true for all $t \leqslant s\log\frac{1}{\varepsilon}$, this gives $\tau_\eta^{(\varepsilon)} \geqslant s\log\frac{1}{\varepsilon}$. This being true for all $\varepsilon \leqslant \min(\varepsilon_0,\varepsilon_2)$, this gives $\liminf_{\varepsilon\to 0}\frac{\tau_\eta^{(\varepsilon)}}{\log 1/\varepsilon} \geqslant s$. This being true for all $s < s_*$, $s \neq s_1,\dots,s_q$, this gives $\liminf_{\varepsilon\to 0}\frac{\tau_\eta^{(\varepsilon)}}{\log 1/\varepsilon} \geqslant s_*$.

We thus conclude that $\lim_{\varepsilon\to 0}\frac{\tau_\eta^{(\varepsilon)}}{\log 1/\varepsilon} = s_*$.

6. Conclusion

In this paper, we have shown how the implicit regularization of DLNs is generated by incremental learning with successive coordinate activations. We obtain a sharp description of the incremental learning process using an associated regularized optimization problem with decreasing regularization.

An immediate open question is to obtain a similar description without the anti-correlation assumption (A2). This would cover the overparametrized setting. In this case, it should be necessary to parametrize $\theta_i = (u_i^2 - v_i^2)/4$ (as in the article of Vaskevicius et al. (2019), for instance) so that the sign of θ_i is not constrained.

Further, we leave open whether our strategy can be adapted to study incremental learning in matrix factorization problems and more general neural networks, as well as the statistical benefits of the induced implicit regularization.

Acknowledgements

The author thanks Loucas Pillaud-Vivien for many discussions motivating this project.

Appendix A. Properties of the Linear Complementarity Problem

This section gathers a few properties of the linear complementarity problem (LCP) from the monograph of Cottle et al. (2009). Let $\boldsymbol{q} \in \mathbb{R}^d$ and $\boldsymbol{M} \in \mathbb{R}^{d \times d}$. We recall that the LCP associated to the parameters $\boldsymbol{q}$ and $\boldsymbol{M}$ is the problem of finding $(\boldsymbol{w}, \boldsymbol{z}) \in \left(\mathbb{R}^d\right)^2$ such that

$$\begin{aligned} &\boldsymbol{w} = \boldsymbol{q} + \boldsymbol{M}\boldsymbol{z}\,, \\ &\boldsymbol{w} \geqslant \boldsymbol{0}\,, \qquad \boldsymbol{z} \geqslant \boldsymbol{0}\,, \qquad \boldsymbol{w}^\top \boldsymbol{z} = 0\,. \end{aligned} \tag{20}$$

Proposition 6 *Assume that $\boldsymbol{M}$ is symmetric positive definite. Then the LCP* (20) *has a unique solution.*

This result is provided in the monograph of Cottle et al. (2009, Theorem 3.1.6) (actually without the symmetry requirement).

Proposition 7 (antitonicity property) *Assume that $\boldsymbol{M}$ is symmetric positive definite, with non-positive off-diagonal entries. Consider $\boldsymbol{q}^{(1)} \leqslant \boldsymbol{q}^{(2)}$ and let $(\boldsymbol{w}_1, \boldsymbol{z}_1)$, $(\boldsymbol{w}_2, \boldsymbol{z}_2)$ be the unique solutions of* (20) *with $\boldsymbol{q} = \boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}$ respectively. Then $\boldsymbol{z}_1 \geqslant \boldsymbol{z}_2$.*

Proof This result is provided in the monograph of Cottle et al. (2009, Theorem 3.11.9).

Indeed, the fact that $\boldsymbol{M}$ has non-positive off-diagonal entries means that $\boldsymbol{M}$ is a $\boldsymbol{Z}$-matrix in the sense of Cottle et al. (2009, Definition 3.11.1). Further, $\boldsymbol{M}$ is a symmetric positive definite matrix, thus $\boldsymbol{M}$ is a $\boldsymbol{P}$-matrix in the sense of Cottle et al. (2009, Section 3.3). Thus $\boldsymbol{M}$ is a $\boldsymbol{K}$-matrix in the sense of Cottle et al. (2009, Definition 3.11.1). Thus Theorem 3.11.9 of Cottle et al. (2009) applies. ■

Appendix B. Proof of Lemma 1

The DLN dynamics (3) form an autonomous ordinary differential equation (ODE)

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \Psi(\boldsymbol{\theta})\,, \qquad\qquad (\Psi(\boldsymbol{\theta}))_i = \theta_i \left(r_i - \sum_{j=1}^d M_{ij}\theta_j \right)\,. \tag{21}$$

We first show that the set

$$Q = \left\{ \boldsymbol{\theta} \geqslant \boldsymbol{0} \;\middle|\; \forall i \in \{1, \dots, d\}, r_i - \sum_{j=1}^d M_{ij}\theta_j \geqslant 0 \right\}$$

is positively invariant for this ODE, i.e., if $\boldsymbol{\theta}(0) \in Q$, then $\boldsymbol{\theta}(t) \in Q$ for all $t \geqslant 0$. The proof is based on Nagumo's theorem, see the original result of Nagumo (1942) or the recent introduction of Blanchini and Miani (2008, Section 4.2) for instance. Heuristically, Nagumo's theorem states that Q is positively invariant if the vector field $\Psi(\boldsymbol{\theta})$ points "in" the set Q if $\boldsymbol{\theta}$ is on the boundary of Q.

More precisely, for $\boldsymbol{\theta} \in Q$, denote

$$\mathrm{Act}(\boldsymbol{\theta}) = \left\{ i \in \{1, \dots, d\} \;\middle|\; r_i - \sum_j M_{ij}\theta_j = 0 \right\}$$

the set of active constraints at $\boldsymbol{\theta}$ and

$$T_Q(\boldsymbol{\theta}) = \left\{ \nu \in \mathbb{R}^d \;\middle|\; \forall i \in \mathrm{Act}(\boldsymbol{\theta}), -\sum_j M_{ij}\nu_j \geqslant 0 \right\}$$

the tangent cone to Q at $\boldsymbol{\theta}$ (Blanchini and Miani, 2008, Eq. (4.6)). Then Nagumo's theorem (Blanchini and Miani, 2008, Corollary 4.8) states that Q is positively invariant for the dynamics (21) if for all $\boldsymbol{\theta} \in Q$, $\Psi(\boldsymbol{\theta}) \in T_Q(\boldsymbol{\theta})$.

We now check that this latter condition is satisfied. Let $\boldsymbol{\theta} \in Q$ and $i \in \mathrm{Act}(\boldsymbol{\theta})$. We need to show that $-\sum_j M_{ij} \left(\Psi(\boldsymbol{\theta})\right)_j \geqslant 0$. We have

$$\begin{aligned} -\sum_j M_{ij} \left(\Psi(\boldsymbol{\theta})\right)_j &= -\sum_j M_{ij}\theta_j \left(r_j - \sum_k M_{jk}\theta_k \right) \\ &= -M_{ii}\theta_i \left(r_i - \sum_k M_{ik}\theta_k \right) - \sum_{j \neq i} M_{ij}\theta_j \left(r_j - \sum_k M_{jk}\theta_k \right) . \end{aligned}$$

For the first term, we have $i \in \mathrm{Act}(\boldsymbol{\theta})$ and thus $r_i - \sum_k M_{ik}\theta_k = 0$. For the sum, we have $M_{ij} \leqslant 0$ (Assumption (A2)), $\theta_j \geqslant 0$ and $r_j - \sum_k M_{jk}\theta_k \geqslant 0$ (because $\boldsymbol{\theta} \in Q$). Thus we indeed have $-\sum_j M_{ij} \left(\Psi(\boldsymbol{\theta})\right)_j \geqslant 0$ and thus $\Psi(\boldsymbol{\theta}) \in T_Q(\boldsymbol{\theta})$. We conclude that Q is positively invariant.

As $\boldsymbol{\theta}^{(\varepsilon)}(0) = (C_1\varepsilon^{k_1}, \dots, C_d\varepsilon^{k_d}) \to \mathbf{0}$ as $\varepsilon \to 0$ and $\boldsymbol{r} > 0$, there exists $\varepsilon_0 > 0$ such that $\forall \varepsilon \in (0, \varepsilon_0], \boldsymbol{\theta}^{(\varepsilon)}(0) \in Q$. Thus $\forall \varepsilon \in (0, \varepsilon_0], \forall t \geqslant 0, \boldsymbol{\theta}^{(\varepsilon)}(t) \in Q$. Thus $\forall \varepsilon \in (0, \varepsilon_0], \forall t \geqslant 0, \forall i \in \{1, \dots, d\}$,

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \theta_i \left(r_i - \sum_{j=1}^{d} M_{ij}\theta_j \right) \geqslant 0 .$$

Appendix C. Proof of Proposition 1

Consider the block matrix

$$\widetilde{\boldsymbol{M}} = \left(\begin{array}{c|c} \boldsymbol{X} & -\boldsymbol{y} \end{array} \right)^\top \left(\begin{array}{c|c} \boldsymbol{X} & -\boldsymbol{y} \end{array} \right) = \left(\begin{array}{c|c} \boldsymbol{M} & -\boldsymbol{r} \\ \hline -\boldsymbol{r}^\top & \|\boldsymbol{y}\|^2 \end{array} \right) .$$

From Assumptions (A1)-(A2), $\widetilde{\boldsymbol{M}}$ is a matrix with non-positive off-diagonal entries. Thus there exists $\mu \in \mathbb{R}$ such that $\boldsymbol{A} = \mu \boldsymbol{I}_{d+1} - \widetilde{\boldsymbol{M}}$ is a matrix with non-negative entries. Moreover, from Assumption (A1), for $i \in \{1, \dots, d\}$, $A_{i,d+1} = A_{d+1,i} = r_i > 0$. This implies that $\boldsymbol{A}$ is irreducible (see Cottle et al., 2009, Section 2.2 for a definition). By the Perron-Frobenius theorem (Cottle et al., 2009, Theorem 2.2.21), the largest eigenvalue of $\boldsymbol{A}$ is simple and there exists an eigenvector $\widetilde{\boldsymbol{v}}$ with positive entries associated to this eigenvalue. As a consequence, the smallest eigenvalue λ of $\widetilde{\boldsymbol{M}} = \mu \boldsymbol{I}_{d+1} - \boldsymbol{A}$ is simple and associated to $\widetilde{\boldsymbol{v}}$.

We now have two cases:

- If the smallest eigenvalue λ is positive, then $\widetilde{\boldsymbol{M}}$ is positive definite and thus so is the principal submatrix $\boldsymbol{M}$.
- If $\lambda = 0$, then $\ker \widetilde{\boldsymbol{M}} = \{\alpha \widetilde{\boldsymbol{v}}, \alpha \in \mathbb{R}\}$. We want to show that $\ker \boldsymbol{M} = \{\boldsymbol{0}\}$. Consider $\boldsymbol{v} \in \mathbb{R}^d$ such that $\boldsymbol{M}\boldsymbol{v} = \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v} = \boldsymbol{0}$. This implies that $\boldsymbol{X}\boldsymbol{v} = \boldsymbol{0}$. (This can be seen, for instance, using a singular value decomposition of $\boldsymbol{X}$.) Then we perform the block computation

$$\widetilde{\boldsymbol{M}} \left(\begin{array}{c} \boldsymbol{v} \\ \hline 0 \end{array} \right) = \left(\begin{array}{c|c} \boldsymbol{M} & -\boldsymbol{r} \\ \hline -\boldsymbol{r}^\top & \|\boldsymbol{y}\|^2 \end{array} \right) \left(\begin{array}{c} \boldsymbol{v} \\ \hline 0 \end{array} \right) = \left(\begin{array}{c} \boldsymbol{M}\boldsymbol{v} \\ \hline -\boldsymbol{r}^\top \boldsymbol{v} \end{array} \right)$$
$$= \left(\begin{array}{c} \boldsymbol{M}\boldsymbol{v} \\ \hline -\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{v} \end{array} \right) = \boldsymbol{0} \,.$$

Thus $\left(\begin{array}{c} \boldsymbol{v} \\ \hline 0 \end{array} \right) \in \ker \widetilde{\boldsymbol{M}} = \{\alpha \widetilde{\boldsymbol{v}}, \alpha \in \mathbb{R}\}$. As $\widetilde{\boldsymbol{v}} > \boldsymbol{0}$, elements of $\ker \widetilde{\boldsymbol{M}}$ have all entries non-zero or all entries equal to 0. Thus it must be that $\boldsymbol{v} = \boldsymbol{0}$. This concludes that $\ker \boldsymbol{M} = \{\boldsymbol{0}\}$ and thus that $\boldsymbol{M}$ is positive definite.

Appendix D. Proof of Proposition 2

Let $I \subset \{1, \dots, d\}$.

We first prove that $(\boldsymbol{M}_{II})^{-1}$ exists. The matrix $\boldsymbol{M}_{II}$ has non-positive off-diagonal entries thus $\boldsymbol{M}_{II}$ is a $\boldsymbol{Z}$-matrix in the sense of Cottle et al. (2009, Definition 3.11.1). Further, $\boldsymbol{M}_{II}$ is a symmetric positive definite matrix as a principal submatrix of a positive definite matrix (Proposition 1). Thus $\boldsymbol{M}_{II}$ is a $\boldsymbol{P}$-matrix in the sense of Cottle et al. (2009, Section 3.3). Thus $\boldsymbol{M}_{II}$ is a $\boldsymbol{K}$-matrix in the sense of Cottle et al. (2009, Definition 3.11.1). From Theorem 3.11.10 of Cottle et al. (2009), $(\boldsymbol{M}_{II})^{-1}$ exists and has non-negative entries.

Thus, we can define $\boldsymbol{\theta}_*^{(I)} \in \mathbb{R}^d$ by the equations $(\boldsymbol{\theta}_*^{(I)})_I = (\boldsymbol{M}_{II})^{-1}\boldsymbol{r}_I$ and $(\boldsymbol{\theta}_*^{(I)})_{I^c} = \boldsymbol{0}$.

We check that $(\boldsymbol{\theta}_*^{(I)})_I > 0$. Fix $i \in I$. Then $\theta_{*,i}^{(I)} = \sum_{j \in I} \left((\boldsymbol{M}_{II})^{-1}\right)_{ij} r_j \geqslant 0$ from Assumption (A1) and the fact that $(\boldsymbol{M}_{II})^{-1}$ has non-negative entries. Moreover, assume by contradiction that $\theta_{*,i}^{(I)} = 0$. As from Assumption (A1), $\boldsymbol{r} > \boldsymbol{0}$, we have for all $j \in I$, $\left((\boldsymbol{M}_{II})^{-1}\right)_{ij} = 0$. Thus a full row of $(\boldsymbol{M}_{II})^{-1}$ is $\boldsymbol{0}$, which contradicts the fact that $(\boldsymbol{M}_{II})^{-1}$ is invertible. Thus for all $i \in I$, $\theta_{*,i}^{(I)} > 0$.

We now check that $\boldsymbol{\theta}_*^{(I)}$ is a fixed point of (3). Fix $i \in \{1, \dots, d\}$. If $i \in I$,

$$r_i - \sum_{j=1}^{d} M_{ij} \theta_{*,j}^{(I)} = r_i - \sum_{j \in I} M_{ij} \left((\boldsymbol{M}_{II})^{-1}\boldsymbol{r}_I\right)_j = \left(\boldsymbol{r}_I - \boldsymbol{M}_{II}(\boldsymbol{M}_{II})^{-1}\boldsymbol{r}_I\right)_i = 0 \,.$$

If $i \notin I$, by definition, $\theta_{*,i}^{(I)} = 0$. In both cases, $\theta_{*,i}^{(I)} \left(r_i - \sum_{j=1}^d M_{ij} \theta_{*,j}^{(I)} \right) = 0$. As this is true for all $i \in \{1, \dots, d\}$, this proves that $\boldsymbol{\theta}_*^{(I)}$ is a fixed point of (3).

We now take a fixed point $\boldsymbol{\theta}$ of (3) with support I, and show that $\boldsymbol{\theta} = \boldsymbol{\theta}_*^{(I)}$. It is sufficient to show the equality on the support of the vectors, namely that $\boldsymbol{\theta}_I = \left(\boldsymbol{\theta}_*^{(I)}\right)_I = (\boldsymbol{M}_{II})^{-1} \boldsymbol{r}_I$. Consider $i \in I$. As $\boldsymbol{\theta}$ is a fixed point, $\theta_i \left(r_i - \sum_{j=1}^d M_{ij} \theta_j \right) = 0$. But as $i \in I$, the first factor is non-zero. Thus $r_i - \sum_{j=1}^d M_{ij} \theta_j = r_i - \sum_{j \in I} M_{ij} \theta_j = 0$. With vector notation, we proved $\boldsymbol{r}_I - \boldsymbol{M}_{II} \boldsymbol{\theta}_I = 0$, which gives the claimed equality.

Appendix E. Proof of Proposition 4

In this proof, we use convex duality (Boyd and Vandenberghe, 2004, Section 5.5). The Lagrangian associated to (7) is

$$L(\boldsymbol{\theta}, \boldsymbol{w}) = \langle \boldsymbol{q}, \boldsymbol{\theta} \rangle + \frac{1}{2} \langle \boldsymbol{\theta}, \boldsymbol{M} \boldsymbol{\theta} \rangle - \langle \boldsymbol{w}, \boldsymbol{\theta} \rangle$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is the Lagrange multiplier associated to the constraint $\boldsymbol{\theta} \geqslant \boldsymbol{0}$. As the optimization problem (7) is convex, the KKT conditions are necessary and sufficient for optimality. The stationarity condition is

$$\boldsymbol{0} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{w}) = \boldsymbol{q} + \boldsymbol{M} \boldsymbol{\theta} - \boldsymbol{w} \,,$$

the feasibility conditions are $\boldsymbol{\theta} \geqslant \boldsymbol{0}$ and $\boldsymbol{w} \geqslant \boldsymbol{0}$, and the complementary slackness condition is $\boldsymbol{w}^\top \boldsymbol{\theta} = 0$. At this point, we have proven the equivalence between:

1. $\boldsymbol{z}$ is a minimizer of the constrained optimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d, \, \boldsymbol{\theta} \geqslant \boldsymbol{0}} \left\{ \langle \boldsymbol{q}, \boldsymbol{\theta} \rangle + \frac{1}{2} \langle \boldsymbol{\theta}, \boldsymbol{M} \boldsymbol{\theta} \rangle \right\} .$$

2. There exists $\boldsymbol{w} \in \mathbb{R}^d$ such that $(\boldsymbol{w}, \boldsymbol{z})$ is a solution of

$$\begin{aligned} &\boldsymbol{w} = \boldsymbol{q} + \boldsymbol{M} \boldsymbol{z} \,, \\ &\boldsymbol{w} \geqslant \boldsymbol{0} \,, \qquad \boldsymbol{z} \geqslant \boldsymbol{0} \,, \qquad \boldsymbol{w}^\top \boldsymbol{z} = 0 \,. \end{aligned}$$

We are left with proving that the solutions of both problems are indeed unique. For the LCP, this is given by Proposition 6 as $\boldsymbol{M}$ is positive definite (Proposition 1). We can then use the equivalence shown above to prove that the constrained optimization problem (7) has a unique solution.

References

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

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

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1): 1–106, 2012.

Stephen Baigent. Lotka–Volterra dynamical systems. In *Dynamical and Complex Systems*, pages 157–188. World Scientific, 2017.

Yuri Bakhtin. Noisy heteroclinic networks. *Probability Theory and Related Fields*, 150(1): 1–42, 2011.

Peter Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

Franco Blanchini and Stefano Miani. *Set-Theoretic Methods in Control*. Springer, 2008.

Etienne Boursier, Loucas Pillaud-Viven, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pages 20105–20118, 2022.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *arXiv preprint arXiv:2011.13772*, 2020.

Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: inducing sparsity via overparameterization. *Information and Inference: A Journal of the IMA*, 12(3), 2023.

Richard Cottle, Jong-Shi Pang, and Richard Stone. *The Linear Complementarity Problem*. SIAM, 2009.

Nelson Dunford and Jacob Schwartz. *Linear Operators, Part 1: General Theory*, volume 10. John Wiley & Sons, 1988.

Rong Ge, Yunwei Ren, Xiang Wang, and Mo Zhou. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34:1299–1311, 2021.

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*, 2019.

Bean-San Goh. Stability in models of mutualism. *The American Naturalist*, 113(2):261–275, 1979.

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

Jeff 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.

Kais Hariz, Hachem Kadri, Stéphane Ayache, Maher Moakher, and Thierry Artières. Implicit regularization with polynomial growth in deep tensor factorization. In *International Conference on Machine Learning*, pages 8484–8501. PMLR, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics.* Cambridge University Press, 1998.

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.

Yann Le Cun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

Jiangyuan Li, Thanh Nguyen, Chinmay Hegde, and Ka Wai Wong. Implicit sparse regularization: The impact of depth and early stopping. *Advances in Neural Information Processing Systems*, 34:28298–28309, 2021.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference on Learning Theory*, pages 2–47. PMLR, 2018.

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*, 2020.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pages 16270–16295. PMLR, 2022.

Mitio Nagumo. Über die lage der integralkurven gewöhnlicher differentialgleichungen. *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, 24:551–559, 1942.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

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.

Clarice Poon and Gabriel Peyré. Smooth bilevel programming for sparse regularization. *Advances in Neural Information Processing Systems*, 34:1543–1555, 2021.

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

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. In *International Conference on Machine Learning*, pages 18422–18462. PMLR, 2022.

Andrew Saxe, James 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. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

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

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit regularization via Hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*, 2019.