(S)GD over Diagonal Linear Networks: Implicit Regularisation, Large Stepsizes and Edge of Stability

Mathieu Even[1,⋆], Scott Pesme [2,⋆], Suriya Gunasekar [3] and Nicolas Flammarion [2]

[1] Inria - ENS Paris

[2] Ecole Polytechnique Fédérale de Lausanne (EPFL)

[3] Microsoft Research

⋆ Equal contributions

Abstract. In this paper, we investigate the impact of stochasticity and large stepsizes on the implicit regularisation of gradient descent (GD) and stochastic gradient descent (SGD) over diagonal linear networks. We prove the convergence of GD and SGD with macroscopic stepsizes in an overparametrised regression setting and characterise their solutions through an implicit regularisation problem. Our crisp characterisation leads to qualitative insights about the impact of stochasticity and stepsizes on the recovered solution. Specifically, we show that large stepsizes consistently benefit SGD for sparse regression problems, while they can hinder the recovery of sparse solutions for GD. These effects are magnified for stepsizes in a tight window just below the divergence threshold, in the "edge of stability" regime. Our findings are supported by experimental results.

1. Introduction

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

To explain this second point, the concept of implicit regularisation has emerged: if overfitting is harmless in many real-world prediction tasks, it must be because the optimisation process is *implicitly favoring* solutions that have good generalisation properties for the task. The canonical example is overparametrised linear regression with more trainable parameters than number of samples: although there are infinitely many solutions that fit the samples, GD and SGD explore only a small subspace of all the possible parameters. As a result, they implicitly converge to the closest solution in terms of the ℓ_2 distance, and this without explicit regularisation [Zhang et al., 2017, Gunasekar et al., 2018a].

Currently, most theoretical works on implicit regularisation have primarily focused on continuous time approximations of (S)GD where the impact of crucial hyperparameters such as the stepsize and the minibatch size are ignored. One such common simplification is to analyse gradient flow, which is a continuous time limit of GD and minibatch SGD with an infinitesimal stepsize. By definition, this analysis cannot capture the effect of stepsize or stochasticity. Another approach is to approximate SGD by a stochastic gradient flow [Wojtowytsch, 2021, Pesme et al., 2021], which tries to capture the noise and the stepsize using an appropriate stochastic differential equation. However, there are no theoretical guarantees that these results can be transferred to minibatch SGD. This is problematic since the performances of most deep learning models are highly sensitive to the choice of stepsize and minibatch size—their importance is common knowledge in practice and has also been systematically established in controlled experiments [Keskar et al., 2017a, Masters and Luschi, 2018, Geiping et al., 2022].

In this work, we aim to address the gaps in our understanding of the impact of stochasticity and stepsizes by analysing the (S)GD trajectory in 2-layer diagonal networks (DLNs). We can already see in Fig. 1 the importance of these parameters in a noiseless sparse recovery problem which we detail later: the solutions recovered by SGD and GD have very different generalisation performances.

The 2-layer diagonal linear network which we consider is a simplified neural network that has received significant attention lately [Woodworth et al., 2020, Vaškevičius et al., 2019, HaoChen et al., 2021, Pillaud-Vivien et al., 2022]. Despite its simplicity, it surprisingly reveals training characteristics which are observed in much more complex architectures. It therefore serves as an

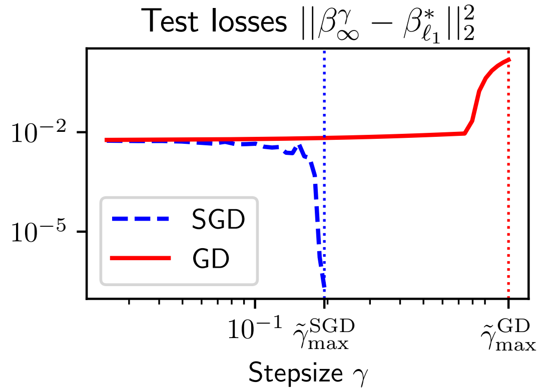

FIGURE 1. Noiseless sparse regression with a 2-layer diagonal linear network for stepsizes such that the iterates converge to a global solution: the solutions recovered by SGD and GD have very different generalisation properties. The dashed vertical lines correspond to the maximum stepsize that can be used before the iterates stop to converge. See the last paragraph of Section 2 for the precise experimental setting.

ideal proxy model for gaining a deeper understanding of complex phenomenons such as the roles of initialisation, stochasticity and stepsize.

1.1. Main results and paper organisation. Overparametrised regression and diagonal linear networks are introduced in Section 2. We formulate our main theorem (Theorem 1) in Section 3, and state it informally here.

Theorem (Informal)**.** *For macroscopic stepsizes, gradient descent and stochastic gradient descent over* 2*-layer diagonal linear networks converge to a specific zero-training loss solution* $\beta_\infty^\star$ *explicitly characterised through an implicit regularisation problem. Furthermore, the generalisation properties of* $\beta_\infty^\star$ *are fully characterised by the sum of the squared (stochastic) gradients along the iterates' trajectory.*

Contrary to previous works, we show the convergence of the iterates and establish the associated regularisation problem **without having to assume vanishing stepsizes**.

Then in Sections 4 and 5 we interpret our main result to understand the effect of the stepsize and of stochasticity on the recovered solution. As clearly illustrated in Fig. 1, in the sparse regression setting, there is a stark difference between the generalisation performances of GD and SGD. While using a large stepsize leads to a sparser solution in the case of SGD and is highly beneficial, the exact opposite holds for GD for which the use of a large stepsize is in fact detrimental. We explain this phenomenon by showing that GD tends to recover a low **weighted-ℓ_1**-norm solution which prevents the recovery of the sparse signal. For SGD, the story is different and thanks to noise the recovered solution enjoys a low ℓ_1-norm. We show that the previous observations are amplified as we push the stepsize towards a threshold value $\tilde{\gamma}_{\max}$, corresponding to the value above which the iterates do not converge towards a global solution anymore. The range of stepsizes just below this threshold corresponds to a brittle regime in which the iterates "oscillate" and where the loss converges very slowly. This regime is often denoted as the *Edge of Stability* regime and we explain why the sparse recovery performances of SGD are highly improved in this regime, while it is the opposite for GD.

1.2. Related works. Implicit bias. The concept of implicit bias in neural networks has been studied recently, starting with the seminal work of Soudry et al. [2018] on max-margin linear classification. This line of research has been further extended to multiplicative parametrisations [Gunasekar et al., 2018a], linear networks [Ji and Telgarsky, 2019], and homogeneous networks [Ji and Telgarsky, 2020, Chizat et al., 2019]. For diagonal linear networks, Woodworth et al. [2020] demonstrate that the scale of the initialisation determines the type of solution obtained, with large initialisations yielding minimum ℓ_2-norm solutions—the neural tangent kernel regime [Jacot et al., 2018] and small initialisation resulting in minimum ℓ_1-norm solutions—the *rich regime* [Chizat et al., 2019]. The analysis relies on the link between gradient descent and mirror descent established by Ghai et al. [2020] and further explored by Vaskevicius et al. [2020], Wu and Rebeschini [2020]. These works focus on full batch gradient, and most of them study the inifitesimal stepsize limit (gradient flow), leading to general insights and results that do not take into account the effect of stochasticity and large stepsizes.

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

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

2. Setup and preliminaries

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

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

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

$$\beta_w = u \odot v \text{ where } w = (u, v) \in \mathbb{R}^{2d} . \tag{1}$$

This parametrisation can be viewed as a simple neural network $x \mapsto \langle u, \sigma(\mathrm{diag}(v)x)\rangle$ where the output weights are represented by u, the inner weights is the diagonal matrix $\mathrm{diag}(v)$, 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 (1), the loss function F over parameters $w = (u, v) \in \mathbb{R}^{2d}$ is given by:

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

It is worth noting that despite the simplicity of the parametrisation, the corresponding optimisation problem is non-convex and is challenging to analyse.

Mini-batch Stochastic Gradient Descent. We minimise F using mini-batch SGD:

$$w_0 = (u_0, v_0)\,, \quad w_{k+1} = w_k - \gamma_k \nabla F_{\mathcal{B}_k}(w_k)\,, \tag{3}$$

where γ_k are stepsizes, $\mathcal{B}_k \subset [n]$ are mini-batches of $b \in [n]$ distinct samples sampled uniformly and independently, and $\nabla F_{\mathcal{B}_k}(w_k)$ are minibatch gradients of partial loss over $\mathcal{B}_k$

$$F_{\mathcal{B}_k}(w) := \mathcal{L}_{\mathcal{B}_k}(u \odot v) := \frac{1}{2b}\sum_{i\in\mathcal{B}_k}(y_i - \langle u \odot v, x_i\rangle)^2 \,.$$

We emphasise that our analysis holds for any batch-size $b \in [n]$ and stepsizes $\{\gamma_k\}_k$. Classical stochastic gradient descent and full-batch gradient descent are special cases with $b = 1$ and $b = n$, respectively. For $k \geqslant 0$, we consider the successive predicting parameters $\beta_k := u_k \odot v_k$ built from the neurons $w_k = (u_k, v_k)$.

Initialisation. We analyse SGD initialised at $u_0 = \sqrt{2}\boldsymbol{\alpha} \in \mathbb{R}^d_{>0}$ and $v_0 = \mathbf{0} \in \mathbb{R}^d$, resulting in $\beta_0 = \mathbf{0} \in \mathbb{R}^d$ independently of the chosen neuron initialisation $\boldsymbol{\alpha}$[1].

Notations. We denote by $y = \frac{1}{\sqrt{n}}(y_1, \dots, y_n) \in \mathbb{R}^n$ the normalised output vector and by $X \in \mathbb{R}^{n\times d}$ the input matrix whose i^{th} row is the normalised input $\frac{1}{\sqrt{n}}x_i \in \mathbb{R}^d$. $[n]$ denotes the set of all integers from 1 to n. Let $H := \frac{1}{n}\sum_i x_i x_i^\top$ denote the Hessian of $\mathcal{L}$. For a batch $\mathcal{B} \subset [n]$ we define the batch loss as $\mathcal{L}_{\mathcal{B}}(\beta) = \frac{1}{2|\mathcal{B}|}\sum_{i\in\mathcal{B}}(\langle x_i, \beta\rangle - y_i)^2$ and its Hessian as $H_{\mathcal{B}} := \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} x_i x_i^\top$. Let L denote the "smoothness" such that $\forall\beta$, $\|H_{\mathcal{B}}\beta\|_2 \leqslant L\|\beta\|_2$, $\|H_{\mathcal{B}}\beta\|_\infty \leqslant L\|\beta\|_\infty$ for all batches $\mathcal{B} \subset [n]$ of size b. All expectations in the paper are with respect to the uniform sampling of the batches $(\mathcal{B}_k)_k$. A real function (e.g, log, exp) applied to a vector must be understood as element-wise application, and for vectors $u, v \in \mathbb{R}^d$, $u^2 = (u_i^2)_{i\in[d]}$ and $u \odot v = (u_i v_i)_{i\in[d]}$. We write $\mathbf{1}$, $\mathbf{0}$ for the constant vectors with coordinates 1 and 0 respectively. For vectors $u, v \in \mathbb{R}^d$, we write $u \leqslant v$ for $\forall i \in [d]$, $u_i \leqslant v_i$ and for symmetric matrices $A, B \in \mathbb{R}^{d\times d}$ we write $A \succeq B$ for $B - A$ positive semi-definite.

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

3. Implicit bias of SGD and GD

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

$$\beta^*_{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\beta^\star\in\mathcal{S}} \psi_{\boldsymbol{\alpha}}(\beta^\star)\,, \tag{4}$$

where $\psi_{\boldsymbol{\alpha}}$ is the hyperbolic entropy function [Ghai et al., 2020] defined as:

$$\psi_{\boldsymbol{\alpha}}(\beta) = \frac{1}{2}\sum_{i=1}^{d}\Big(\beta_i \mathrm{arcsinh}(\frac{\beta_i}{\alpha_i^2}) - \sqrt{\beta_i^2 + \alpha_i^4} + \alpha_i^2\Big). \tag{5}$$

[1] In Appendix C, we show that the (S)GD trajectory with this initialisation exactly matches that of another common parametrisation $\beta_w = w_+^2 - w_-^2$ with initialisation $w_{+,0} = w_{-,0} = \boldsymbol{\alpha}$

The key characteristic of the hyperbolic entropy is its ability to interpolate between the ℓ_1 and ℓ_2 norms as the scale of the uniform initialisation $\boldsymbol{\alpha} = \alpha \mathbf{1} \in \mathbb{R}^d$ approaches zero and infinity respectively [Woodworth et al., 2020, Theorem 2].

The implicit regularisation characterisation in (4) highlights that the scale of the initial weights $\alpha > 0$ controls the shape of the solution recovered by the algorithm. Small initialisation results in solutions with low ℓ_1-norm known to induce sparse recovery guarantees [Candès et al., 2006]. This setting is often referred to as the "rich" regime [Woodworth et al., 2020]. In contrast, using large initial weights leads to solutions with low ℓ_2-norm, a setting known as the "kernel" or "lazy" regime. The weights of the neurons make only small adjustments to fit the data, resulting in dynamics similar to kernel linear regression. Overall, to retrieve the minimum ℓ_1-norm solution $\beta^\star_{\ell_1} := \operatorname{argmin}_{\beta^\star \in S} \|\beta^\star\|_1$, it is recommended to use the smallest possible initialisation scale α. However, with $\boldsymbol{\alpha} = 0$, $w_0 = (0, 0)$ is a saddle point of F, which makes the training longer as $\boldsymbol{\alpha} \to 0$.

In addition to the scale of $\boldsymbol{\alpha}$, a lesser studied aspect of initialisation is its "shape", which is a term we use to refer to the relative distribution of $\{\alpha_i\}$ along the d coordinates. For uniform initialisation $\boldsymbol{\alpha} = \alpha \mathbf{1}$, we know that the limit of $\alpha \to 0$ is $\psi_{\boldsymbol{\alpha}} \propto \|.\|_1$. However, if each entry of $\boldsymbol{\alpha}$ *does not go to zero "uniformly"*, the limit $\lim_{\boldsymbol{\alpha} \to \mathbf{0}} \psi_{\boldsymbol{\alpha}}$ becomes a **weighted** ℓ_1-norm instead of the standard ℓ_1-norm, which can negatively affect sparse recovery (Example 1).

3.1. Main result: convergence and implicit bias. In Theorem 1, we prove that for an initialisation $\boldsymbol{\alpha} \in \mathbb{R}^d$ and for macroscopic stepsizes, minibatch stochastic gradient descent on F converges almost surely to an interpolator, which we denote $\beta^\star_\infty$. Moreover, as for gradient flow, this interpolator minimises the hyperbolic entropy (5) but for a trajectory-dependent *effective* initialisation $\boldsymbol{\alpha}_\infty \in \mathbb{R}^d$ which is component-wise stricly smaller than $\boldsymbol{\alpha}$.

Theorem 1. *Let $(u_k, v_k)_{k \geqslant 0}$ follow the mini-batch SGD recursion (3) initialised at $u_0 = \sqrt{2}\boldsymbol{\alpha} \in \mathbb{R}^d_{>0}$ and $v_0 = \mathbf{0}$, and let $(\beta_k)_{k \geqslant 0} = (u_k \odot v_k)_{k \geqslant 0}$. There exists $B > 0$ and a numerical constant $c > 0$ such that for stepsizes satisfying $\gamma_k \leqslant \frac{c}{LB}$, the iterates satisfy $\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leqslant 1$ and $\|\beta_k\|_\infty \leqslant B$ for all k, and:*

(1) *$(\beta_k)_{k \geqslant 0}$ converges almost surely to some $\beta^\star_\infty \in \mathcal{S}$, that satisfies:*

$$\beta^\star_\infty = \operatorname*{argmin}_{\beta^\star \in \mathcal{S}} D_{\psi_{\boldsymbol{\alpha}_\infty}}(\beta^\star, \tilde{\beta}_0)\,, \tag{6}$$

where $\boldsymbol{\alpha}_\infty \in \mathbb{R}^d_{>0}$ and $\tilde{\beta}_0 \in \mathbb{R}^d$.

(2) *$\boldsymbol{\alpha}_\infty \in \mathbb{R}^d$ satisfies $\boldsymbol{\alpha}_\infty \leqslant \boldsymbol{\alpha}$ and is equal to:*

$$\boldsymbol{\alpha}_\infty^2 = \boldsymbol{\alpha}^2 \odot \exp\left(- \sum_{k=0}^{\infty} q\big(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\big) \right), \tag{7}$$

where $q(x) = -\frac{1}{2}\ln((1-x^2)^2) \geqslant 0$ for $|x| \leqslant \sqrt{2}$, and $\tilde{\beta}_0$ is a perturbation term equal to:

$$\tilde{\beta}_0 = \frac{1}{2}\big(\boldsymbol{\alpha}_+^2 - \boldsymbol{\alpha}_-^2\big),$$

where, $q_\pm(x) = \mp 2x - \ln((1 \mp x)^2)$, and $\boldsymbol{\alpha}_\pm^2 = \boldsymbol{\alpha}^2 \odot \exp\left(- \sum_{k=0}^{\infty} q_\pm(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))\right)$.

Full characterisation of the recovered solution. The regularisation problem (6) fully characterises the interpolator selected by the algorithm. It is worth noting that the assumption on the stepsize is only required to show the convergence of the iterates. Consequently, **regardless of stepsize sequence chosen**, as long as the iterates β_k converge to zero training error, the implicit regularisation characterisation in Eq. (6) holds true.

This remark even holds for adaptive stepsize schedules which keep the stepsize scalar such as AdaDelta [Zeiler, 2012]. To our knowledge, this is the first complete characterisation of the implicit bias of gradient methods with practical stepsizes. Thus our result extends beyond the classical continuous-time framework where all previous results were derived [Woodworth et al., 2020, Pesme et al., 2021]. Note that for $\gamma_k \to 0$ we have $\boldsymbol{\alpha}_\infty \to \boldsymbol{\alpha}$ and $\tilde{\beta}_0 \to \mathbf{0}$ (Proposition 13), recovering previous results for gradient flow (4).

$\tilde{\beta}_0$ **can be ignored.** We show in Proposition 12 that the magnitude of $\tilde{\beta}_0$ is negligable in front of the magnitudes of $\beta^\star \in S$. Hence, one can roughly ignore the term $\tilde{\beta}_0$ and the implicit regularisation Eq. (6) can be simplified as $\beta^\star_\infty \approx \operatorname{argmin}_{\beta^\star \in S} \psi_{\boldsymbol{\alpha}_\infty}(\beta^\star)$.

Effective initialisation $\boldsymbol{\alpha}_\infty$ (with a twist). Considering that $\tilde{\beta}_0 \approx 0$, the solution $\beta^\star_\infty$ obtained with minibatch SGD or GD minimises the same potential function that the solution of gradient flow Eq. (4) but with an effective initialisation of $\boldsymbol{\alpha}_\infty$, which is elementwise strictly smaller than $\boldsymbol{\alpha}$. Thus, based on the properties of hyper-entropy, we expect our effective implicit regulariser $\psi_{\boldsymbol{\alpha}_\infty}$ to be closer to the ℓ_1-norm than $\psi_{\boldsymbol{\alpha}}$. We thus expect the smaller scale of $\boldsymbol{\alpha}_\infty$ to always help in the recovery of low ℓ_1-norm solutions.

However, as clearly seen in Fig. 1, this is not always the case. This is because the recovery of low ℓ_1-norm solutions does not just depend on the scale $\|\boldsymbol{\alpha}_\infty\|_1$, but also on how uniform the "shape" of $\boldsymbol{\alpha}_\infty$ is. The full picture of how $\boldsymbol{\alpha}_\infty$ impacts recovery of minimum ℓ_1-norm solution is more nuanced and we discuss this in more details in the following sections. In particular, we show that the shape of our effective initialisation $\boldsymbol{\alpha}_\infty$ is affected by various factors including batchsize and input data X. We see that, unlike GD, the stochasticity in SGD leads to a more uniform $\boldsymbol{\alpha}_\infty$ and larger stepsizes (up to divergence) highly benefit the recovery of low ℓ_1-norm solutions, as seen Fig. 1.

Effect of the stepsizes. The impact of the macroscopic stepsizes is reflected in the effective initialisation $\boldsymbol{\alpha}_\infty$. The difference with gradient flow is directly associated with the quantity $\sum_k q(\gamma_k g_k)$: the larger this sum, the more the recovered solution differs from that of gradient flow. Also, as the (stochastic) gradients g_k converge to 0 and that $q(x) \overset{x\to 0}{\sim} x^2$, one should think of this sum as roughly being $\sum_k \gamma_k^2 g_k^2$. While it is clear that the stepsize influences our effective intialization $\boldsymbol{\alpha}_\infty$, it is not straightforward to predict the exact impact it has and whether it aids in recovering a low ℓ_1-norm solution. Our analysis in the next section shows that the improvement highly depends on the batch size and whether the inputs are centered or not.

3.2. Sketch of proof and time-varying mirror descent. Since the loss F is non-convex, it is non-trivial that the iterates (u_k, v_k) converge towards a global minimum. To prove this result, we consider the iterates $\beta_k = u_k \odot v_k$ and show that they follow a mirror descent recursion with time-varying potentials $(h_k)_{k \geqslant 0}$ [Orabona et al., 2015] on the convex loss $\mathcal{L}(\beta)$. The potentials are defined just below, and are closely related to the hyperbolic entropy (5).

Proposition 1. *$(\beta_k = u_k \odot v_k)_{k\geqslant 0}$ satisfy the Stochastic Mirror Descent recursion with varying potentials $(h_k)_k$:*

$$\nabla h_{k+1}(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,, \tag{8}$$

where $h_k : \mathbb{R}^d \to \mathbb{R}$ for $k \geqslant 0$ are strictly convex functions.

By suitably modifying classical convex optimization techniques to account for time-varying potentials (Proposition 8, that can also be of independent interest), we can prove the convergence of the iterates towards an interpolator $\beta^\star_\infty$ along with that of all the relevant quantities which appear in Theorem 1. The implicit regularisation problem then directly follows from: (1) the limit condition $\nabla h_\infty(\beta_\infty) \in \operatorname{Span}(x_1, \dots, x_n)$ as seen from Eq. (8) and (2) the interpolation condition $X\beta^\star_\infty = y$. Indeed, these two conditions exactly correspond to the KKT conditions of the convex problem Eq. (6). Furthermore, while we perform our analysis for minibatch SGD, Proposition 1 holds for any other stochastic gradient methods (*e.g.*, label-noise).

We now explicit the reference functions $(h_k)_{k\geqslant 0}$, that are related to the hyperbolic entropy of scale α_k with a translation ϕ_k, where α_k and ϕ_k are defined as follows. Let $q, q_\pm : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be defined as:

$$\begin{aligned} q_\pm(x) &= \mp 2x - \ln\left((1 \mp x)^2\right), \\ q(x) &= \frac{1}{2}(q_+(x) + q_-(x)) = -\frac{1}{2}\ln\left((1 - x^2)^2\right), \end{aligned}$$

with the convention that $q(1) = \infty$. Notice that $q(x) \geqslant 0$ for $|x| \leqslant \sqrt{2}$ and $q(x) < 0$ otherwise. For the iterates $\beta_k = u_k \odot v_k \in \mathbb{R}^d$, we define the following quantities:

$$\begin{aligned}
\alpha_{\pm,k}^2 &= \alpha^2 \exp\big(-\sum_{i=0}^{k-1} q_{\pm}(\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))\big) \in \mathbb{R}^d \,, \\
\alpha_k^2 &= \alpha_{+,k} \odot \alpha_{-,k} \,, \\
\phi_k &= \frac{1}{2} \operatorname{arcsinh}\big(\frac{\alpha_{+,k}^2 - \alpha_{-,k}^2}{2\alpha_k^2}\big) \in \mathbb{R}^d \,.
\end{aligned}$$

Finally for $k \geqslant 0$, the reference function $(h_k : \mathbb{R}^d \to \mathbb{R})_{k \geqslant 0}$ in Proposition 1 have the following expression:

$$h_k(\beta) = \psi_{\alpha_k}(\beta) - \langle \phi_k, \beta \rangle \,, \tag{9}$$

where ψ_{α_k} is the hyperbolic entropy (5) of scale α_k. A natural and crucial consequence of Proposition 1 is the following corollary, that characterises the limit, and that holds irrespectively of the structure of the stochastic gradients, as long as $g_k \in \mathrm{Span}(x_1, \ldots, x_n)$ holds for all k (as is the case for minibatch SGD, GD, or label noise SGD).

Corollary 1. *Assume that the iterates converge to some interpolator $\beta_\infty^\star$ and that there exist $\alpha_\infty, \phi_\infty \in \mathbb{R}^d$ such that $\alpha_k \to \alpha_\infty$ and $\phi_k \to \phi_\infty$. Then, $\beta_\infty^\star$ is uniquely defined by the following implicit regularization problem:*

$$\beta_\infty^\star = \underset{\beta^* \in S}{\operatorname{argmin}}\ D_{\psi_{\alpha_\infty}}(\beta^*, \tilde{\beta}_0) \,,$$

where $\tilde{\beta}_0 = 2\alpha_\infty^2 \sinh\big(2\phi_\infty\big)$.

This result is directly obtained by noticing that for all $k \geqslant 0$, we have $\nabla h_k(\beta_k) \in \mathrm{Span}(x_1, \ldots, x_n)$, the convergence of the iterates and of the reference functions being a consequence of the assumptions made in the corollary.

4. Analysis of the implicit bias and of the effective initialisation: stochasticity and stepsize

In this section, we analyse the effects of large stepsizes and stochasticity on the implicit bias of minibatch SGD, and in particular on the effective initialisation $\boldsymbol{\alpha}_\infty$. We explain how these two features influence the effective initialisation $\boldsymbol{\alpha}_\infty$ appearing in Theorem 1. Two major factors influence the recovered interpolator $\beta_\infty^\star$:

(1) **Scale** of $\boldsymbol{\alpha}_\infty$: for a homogeneous vector $\boldsymbol{\alpha} = \alpha\mathbf{1}$, as the scale α decreases, the function $\psi_{\boldsymbol{\alpha}}$ becomes increasingly similar to the ℓ_1-norm and the sparse recovery guarantees of $\beta_{\boldsymbol{\alpha}}^\star$ Eq. (4) improve. See Fig. 6 in Appendix D and Theorem 2 of Woodworth et al. [2020] for a precise characterisation.
(2) **Shape** of $\boldsymbol{\alpha}_\infty$: For $\boldsymbol{\alpha} \in \mathbb{R}^d_{\geqslant 0}$, we can show that (see Appendix D), $\psi_{\boldsymbol{\alpha}}(\beta) \overset{\boldsymbol{\alpha} \to \mathbf{0}}{\sim} \sum_{i=1}^d \ln(\frac{1}{\alpha_i})|\beta_i|$. Thus, a heterogeneous vector $\ln(1/\boldsymbol{\alpha}_\infty)$ with entries of differing magnitude results in minimising a **weighted** ℓ_1-norm at small scales. This phenomenon can lead to solutions with vastly different sparsity structure than the minimum ℓ_1-norm interpolator. See Fig. 6 in Appendix D for an intuitive illustration.

From Eq. (7), we see that the scale and the shape of $\boldsymbol{\alpha}_\infty$ are controlled by $\sum_k q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$. We henceforth call this quantity the *gain vector*. For simplicity, from now on, we consider constant stepsize $\gamma_k = \gamma$ for all $k \geqslant 0$ and a uniform initialisation of the neurons $\boldsymbol{\alpha} = \alpha\mathbf{1}$ where $\alpha > 0$ is the initialisation scale. We can then write the gain vector:

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

Following our discussion on the scale and the shape of $\boldsymbol{\alpha}_\infty$,

(1) The **magnitude** of $\|\mathrm{Gain}_\gamma\|_1$ indicates how much the implicit bias of (S)GD differs from that of gradient flow: $\|\mathrm{Gain}_\gamma\|_1 \sim 0$ implies that $\boldsymbol{\alpha}_\infty \sim \boldsymbol{\alpha}$ and therefore the recovered solution is close to that of gradient flow. On the contrary, $\|\mathrm{Gain}_\gamma\|_1 \gg \ln(1/\alpha)$ implies that

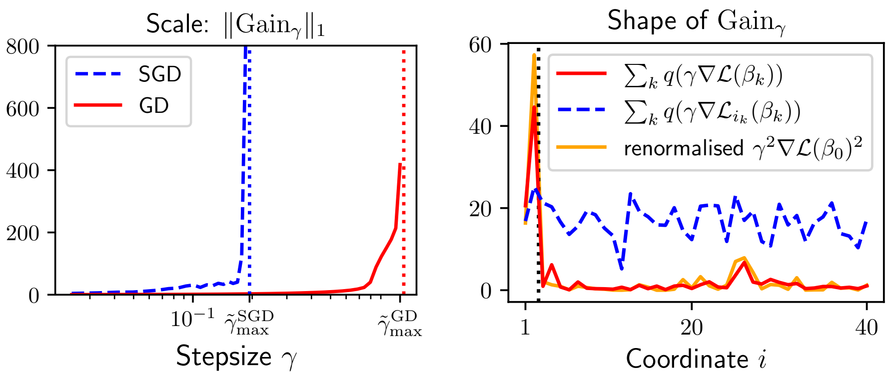

FIGURE 2. **Shape and scale of Gain_γ.** *Left:* the scale of Gain_γ explodes as $\gamma \to \tilde{\gamma}_{\max}$ for both GD and SGD. *Right:* The shape of $\mathrm{Gain}_\gamma^{\mathrm{SGD}}$ is homogeneous whereas that of GD is heterogeneous with much higher magnitude on the support of $\beta^\star_{\ell_1}$ (first two coordinates on the left of the dashed vertical line). The shape of $\mathrm{Gain}_\gamma^{\mathrm{GD}}$ matches that of the gradient at initialisation.

$\boldsymbol{\alpha}_\infty$ has effective scale much smaller than $\boldsymbol{\alpha}$ thereby changing the implicit regularisation Eq. (6).

(2) The **shape** of Gain_γ indicates which coordinates of β in the associated minimum weighted ℓ_1 problem are most penalised. Since

$$\psi_{\boldsymbol{\alpha}_\infty}(\beta) \sim \ln(\frac{1}{\alpha})\|\beta^\star\|_1 + \sum_{i=1}^{d} \mathrm{Gain}_\gamma(\mathrm{i})|\beta_\mathrm{i}|, \tag{10}$$

coordinates of β corresponding to the largest entries of Gain_γ are less likely to be recovered.

4.1. Scale of Gain_γ.

We start by introducing some data-dependent constants.

Definition 1. *Recall that for a batch $\mathcal{B} \subset [n]$ of size b, $H_{\mathcal{B}} = \frac{1}{b}\sum_{i\in\mathcal{B}} x_i x_i^\top$ and $H = \frac{1}{n}\sum_i x_i x_i^\top$. Let $\Lambda_b, \lambda_b > 0$ be the largest and smallest values, respectively, such that:*

$$\lambda_b H \preceq \mathbb{E}_{\mathcal{B}}\big[H_{\mathcal{B}}^2\big] \preceq \Lambda_b H.$$

For all b, (λ_b, Λ_b) only depend on X. For $b = n$, we have $(\lambda_n, \Lambda_n) = (\lambda^+_{\min}(H), \lambda_{\max}(H))$ where $\lambda^+_{\min}(H)$ is the smallest non-null eigenvalue of H. For $b = 1$, we have $\min_i \|x_i\|_2^2 \leqslant \lambda_1 \leqslant \Lambda_1 \leqslant \max_i \|x_i\|_2^2$. The following proposition highlights the dependencies of the scale of the gain $\|\mathrm{Gain}_\gamma\|_1$ in terms of various problem constants.

Proposition 2. *For any stepsize $\gamma > 0$, initialisation $\alpha\mathbf{1}$ and batch size $b \in [n]$, the magnitude of the gain satisfies:*

$$\lambda_b \gamma^2 \sum_k \mathcal{L}(\beta_k) \leqslant \mathbb{E}\left[\|\mathrm{Gain}_\gamma\|_1\right] \leqslant \Lambda_b \gamma^2 \sum_k \mathcal{L}(\beta_k)\,, \tag{11}$$

where the expectation is over uniform and independent sampling of the batches $(\mathcal{B}_k)_{k\geqslant 0}$ at each iteration. Furthermore, for stepsize $0 < \gamma \leqslant \gamma_{\max} = \frac{c}{BL}$, we have that:

$$\sum_k \gamma^2 \mathcal{L}(\beta_k) = \Theta\left(\gamma \ln\left(\frac{1}{\alpha}\right) \left\|\beta^\star_{\ell_1}\right\|_1\right). \tag{12}$$

The slower the training, the larger the gain. Eq. (11) shows that the slower the training loss converges to 0, the larger the sum of the loss, leading to a larger scale of Gain_γ. It extends observations previously made for stochastic gradient flow [Pesme et al., 2021] to SGD and GD.

Impact of the stepsize. The effect of the stepsize on the magnitude of the gain is not directly visible in Eq. (11) because larger stepsize γ tends to speed up the training. However, Eq. (12) clearly shows that increasing that stepsize **boosts** the magnitude $\|\mathrm{Gain}_\gamma\|_1$ up until the limit of

$\gamma_{\max}$. Therefore, the larger the stepsize the smaller is the effective scale of $\boldsymbol{\alpha}_\infty$, which in turn, if the gap is significant, leads to a large deviation of (S)GD from the gradient flow.

Impact of stochasticity. In the following Corollary, we demonstrate the effect of stochasticity (through the batch size b) on the magnitude of the gain. It requires some probabilistic assumptions[2] on the data distribution which enables the control of the values of λ_b and Λ_b.

Corollary 2. *Assume that the inputs are sampled from $\mathcal{N}(0, \sigma^2 I_d)$ for $\sigma^2 > 0$. Then, we have $\lambda_b = \Theta\left(\frac{\sigma^2 d}{b}\right)$ and $\Lambda_b = \Theta\left(\frac{\sigma^2 d}{b}\right)$ with probability $1 - Cne^{-cd}$ over the dataset and thus, with stepsize $0 < \gamma \leqslant \gamma_{\max} = \frac{c}{BL}$, we have that:*

$$\mathbb{E}\left[\|\mathrm{Gain}_\gamma\|_1\right] = \Theta\Big(\gamma \frac{\sigma^2 d}{b} \ln\big(\frac{1}{\alpha}\big) \|\beta^\star_{\ell_1}\|_1\Big)\,. \tag{13}$$

Corollary 2 directly shows that the scale of the Gain_γ decreases with the size of the batch and that there exists a factor n between that of SGD and that of GD. In Fig. 1, this explains why for $\gamma \leqslant \gamma_{\max}$, there is a difference of recovered solutions between SGD and gradient flow but not between GD and gradient flow.

Linear scaling rule. Notice from Corollary 2 that the magnitude of the Gain_γ depends on the value $\frac{\gamma}{b}$. This is reminiscent of the linear scaling rule, which is a standard practice in deep learning [Goyal et al., 2017]: SGD with $b = 1$ and stepsize γ is expected to behave similarly to SGD with batch-size b' but rescaled stepsize $\gamma' = \gamma \times b'$.

Loose analysis and Edge of Stability. Our results hold for stepsizes such that $\gamma \leqslant \gamma_{\max} = \frac{c}{LB}$ where c is some numerical constant. While this bound is accurate in terms of its dependencies on the problem constant, it tends to be conservative in terms of the value of c: empirically, the loss still converges for larger stepsizes. Let then $\tilde{\gamma}_{\max}$ be the largest stepsize one can use before the iterates do not converge, *i.e.*, $\tilde{\gamma}_{\max} = \sup_{\gamma \geqslant 0}\{\gamma \text{ s.t. } \forall \gamma' \leqslant \gamma,\ \sum_k \mathcal{L}(\beta_k^{\gamma'}) < \infty\}$. We directly have that $\gamma_{\max} \leqslant \tilde{\gamma}_{\max}$, and for $\gamma \to \tilde{\gamma}_{\max}^-$, (12) cannot hold and the sum $\sum_k \mathcal{L}(\beta_k)$ diverges as $\gamma \to \tilde{\gamma}_{\max}$,[3] which is clearly observed in Fig. 2 (left). Also, notice that the value $\tilde{\gamma}_{\max}$ depends on whether we consider GD or SGD and that as expected, one can use larger stepsizes for gradient descent, even though the stepsize regime $\gamma \in [\gamma_{\max}, \tilde{\gamma}_{\max}]$ is not captured in classical analyses.

For such very large stepsizes, the iterates of gradient descent tend to "bounce" and that of stochastic gradient descent to "fluctuate": this regime is commonly referred to as the *Edge of Stability*. As the convergence of the loss can be made arbitrarily slow due to these bouncing effects, the magnitude of Gain_γ can be made arbitrarily big, and the recovered solution heavily differs from that of gradient flow as seen in Fig. 1.

By analysing the magnitude $\|\mathrm{Gain}_\gamma\|_1$, we have explained how (S)GD with large stepsizes behave differently than gradient flow. However, our analysis so far does not show qualitatively a different behaviour between SGD and GD beyond the linear stepsize scaling rules. In contrast, Fig. 1 fundamentally shows different behaviours of SGD and GD: to explain this, we need to understand the shape of Gain_γ.

4.2. Shape of Gain_γ. In this section, we restrict our analysis to single batch SGD ($b = 1$) and full batch GD ($b = n$). We also focus on tractable sparse recovery settings, wherein the minimum ℓ_1-norm interpolator is also the sparsest interpolator [Candès et al., 2006]. Precise assumptions are stated below.

We first visualise in Fig. 2 (right) a typical shape of Gain_γ from SGD and GD trained with large stepsizes. In Fig. 2 (right), we see that GD and SGD indeed lead to different shapes of Gain_γ. Importantly, we see that for GD, the magnitude of Gain_γ is higher for coordinates in the support of $\beta^\star_{\ell_1}$. This is undesirable as based on our discussion above: the coordinates with high gain magnitude are adversely weighted in the asymptotic limit of $\psi_{\boldsymbol{\alpha}_\infty}$. This would explain the observation that GD in this regime has bad sparse recovery guarantees in spite of having small scale of $\boldsymbol{\alpha}_\infty$.

[2]The Gaussian assumption can be generalized to non-isotropic sub-Gaussian random variables using more refined concentration bounds. The zero-mean assumption can be relaxed at the cost of an additional $\|\mu\|^2$.

[3]note that we observe this experimentally, however we do not show it and leave it as future work.

The **shape** of Gain_γ is determined by the sum of the squared gradients $\sum_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)^2$, and in particular by the degree of heterogeneity among the coordinates of this sum. Precisely analysing the sum over the whole trajectory of the iterates $(\beta_k)_k$ is technically out of reach. However, we empirically observe for the trajectories shown in Fig. 2 that the shape is largely determined within the first few iterates and that the shape of the whole sum is close to that $\mathbb{E}[\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_0)]$. We formalise this observation below.

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

Having made the observation, to further understand the behaviour and the effects of the stochasticity and the stepsize on the shape of Gain_γ, we analyse a noiseless sparse recovery problem under the following standard assumption 1 [Candès et al., 2006] and as common in the sparse recovery literature, we make the following assumption 2 on the inputs.

Assumption 1. *There exists an s-sparse ground truth vector $\beta^\star_{\text{sparse}}$ where s verifies $n = \Omega(s \ln(d))$, such that $y_i = \langle \beta^\star_{\text{sparse}}, x_i \rangle$ for all $i \in [n]$.*

Assumption 2. *There exists $\delta, c_1, c_2 > 0$ such that for all s-sparse vectors β, there exists $\varepsilon \in \mathbb{R}^d$ such that $(X^\top X)\beta = \beta + \varepsilon$ where $\|\varepsilon\|_\infty \leqslant \delta \|\beta\|_2$ and $c_1 \|\beta\|_2^2 \mathbf{1} \leqslant \frac{1}{n}\sum_i x_i^2 \langle x_i, \beta \rangle^2 \leqslant c_2 \|\beta\|_2^2 \mathbf{1}$.*

The first part of Assumption 2 closely resembles the classical restricted isometry property (RIP) and is relevant for GD while the second part is relevant for SGD. Such an assumption is not restrictive and holds with high probability for Gaussian inputs $\mathcal{N}(0, \sigma^2 I_d)$ (see Lemma 9 in Appendix). Based on the claim above, we analyse the shape of the (stochastic) gradient at initialisation. For GD and SGD, it respectly writes, where $g_0 = \nabla \mathcal{L}_{i_0}(\beta_0)^2$, $i_0 \sim \mathrm{Unif}([\mathrm{n}])$:

$$\nabla \mathcal{L}(\beta_0)^2 = [X^\top X \beta^\star]^2\,, \quad \mathbb{E}_{i_0}[g_0] = \frac{1}{n}\sum_i x_i^2 \langle x_i, \beta^\star \rangle^2.$$

The following lemma then shows that while the initial stochastic gradients of SGD are homogeneous, it is not the case for that of GD.

Proposition 3. *Under Assumption 2, the squared full batch gradient and the expected stochastic gradient at initialisation satisfy, for some ε verifying $\|\varepsilon\|_\infty << \left\|\beta^\star_{\text{sparse}}\right\|_\infty^2$:*

$$\nabla \mathcal{L}(\beta_0)^2 = (\beta^\star_{\text{sparse}})^2 + \varepsilon\,, \tag{14}$$

$$\mathbb{E}_{i_0}[\nabla \mathcal{L}_{i_0}(\beta_0)^2] = \Theta\Big(\|\beta^\star\|_2^2 \mathbf{1}\Big)\,. \tag{15}$$

The gradient of GD is heterogeneous. Since $\beta^\star_{\text{sparse}}$ is sparse by definition, from Eq. (14) we deduce that $\nabla \mathcal{L}(\beta_0)$ is heterogeneous and that it takes larger values on the support of $\beta^\star_{\text{sparse}}$. Along with observation 1, this means that Gain_γ **has much larger values on the support of** $\beta^\star_{\text{sparse}}$. The corresponding weighted ℓ_1-norm therefore has bigger weights penalising the coordinates which belong to the support of $\beta^\star_{\text{sparse}}$, which is harmful for the recovery of $\beta^\star_{\text{sparse}}$ (as explained in Example 1, Appendix D).

The stochastic gradient of SGD is homogeneous. On the contrary, from Eq. (15), we have that the initial stochastic gradients are homogeneous, leading to a weighted ℓ_1-norm where the weights are roughly balanced. The corresponding weighted ℓ_1-norm is therefore close to the uniform ℓ_1-norm and the classical ℓ_1 recovery guarantees are expected.

4.3. Uncentered data. When the data is uncentered, the discussion and the conclusion for GD are somewhat different. This paragraph is motivated by the observation of Nacson et al. [2022] who notice that GD with large stepsizes helps to recover low ℓ_1 solutions for uncentered data (Fig. 4). We make the following assumptions on the uncentered inputs.

Assumption 3. *There exist $\mu \in \mathbb{R}^d$ and $\delta, c_0, c_1, c_2 > 0$ such that for all s-sparse vectors β verifying $\langle \mu, \beta \rangle \geqslant c_0 \|\beta\|_\infty \|\mu\|_\infty$, there exists $\varepsilon \in \mathbb{R}^d$ such that $(X^\top X)\beta = \langle \beta, \mu \rangle \mu + \varepsilon$ where $\|\varepsilon\|_2 \leqslant \delta \|\beta\|_2$ and $c_1 \langle \beta, \mu \rangle^2 \mu^2 \leqslant \frac{1}{n}\sum_i x_i^2 \langle x_i, \beta \rangle^2 \leqslant c_2 \langle \beta, \mu \rangle^2 \mu^2$.*

Assumption 3 is not restrictive and holds with high probability for $\mathcal{N}(\mu \mathbf{1}, \sigma^2 I_d)$ inputs when $\mu \gg \sigma \mathbf{1}$ (see Lemma 8 in Appendix). The following lemma characterises the initial shape of SGD and GD gradients for uncentered data.

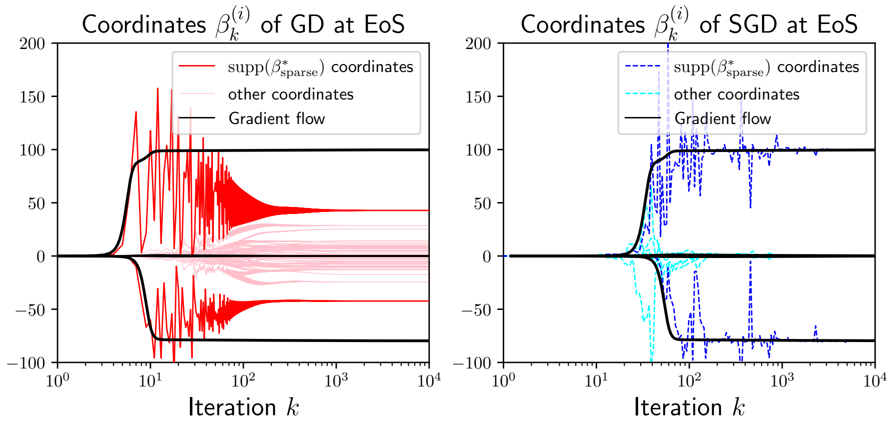

FIGURE 3. **(S)GD at the *EoS*.** *Left:* For GD, the coordinates on the support of $\beta^\star_{\text{sparse}}$ oscillate and drift towards 0. *Right:* For SGD, all the coordinates fluctuate and the iterates converge towards $\beta^\star_{\text{sparse}}$.

Proposition 4 (Shape of the (stochastic) gradient at initialisation)**.** *Under Assumption 3 and if $\langle \mu, \beta^\star_{\text{sparse}}\rangle \geqslant c_0 \|\beta\|_\infty \|\mu\|_\infty$, the squared full batch gradient and the expected stochastic gradient descent at initialisation satisfy, for some ε satisfying $\|\varepsilon\|_\infty \ll \|\beta_{\text{sparse}}\|_2$:*

$$\nabla \mathcal{L}(\beta_0) = \langle \beta^\star_{\text{sparse}}, \mu\rangle^2 \mu^2 + \varepsilon \,, \tag{16}$$

$$\mathbb{E}_{i\sim\text{Unif}([n])}[\nabla \mathcal{L}_i(\beta_0)^2] = \Theta\Big(\langle \beta^\star_{\text{sparse}}, \mu\rangle^2 \mu^2\Big)\,. \tag{17}$$

In this case the initial gradients of SGD and of GD **are both homogeneous**, explaining the behaviours of gradient descent in Fig. 4 (App. A): large stepsizes help in the recovery of the sparse solution in the presence of uncentered data, as opposed to centered data. Note that for decentered data with a $\mu \in \mathbb{R}^d$ orthogonal to $\beta^\star_{\text{sparse}}$, there is no effect of decentering on the recovered solution. If the support of μ is the same as that of $\beta^\star_{\text{sparse}}$, the effect is detrimental and the same discussion as in the centered data case applies.

5. Edge of Stability: the neural point of view

In recent years it has been noticed that when training neural networks with 'large' stepsizes at the limit of divergence, GD and SGD enter the *Edge of Stability (EoS)* regime. In this regime, as seen in Fig. 3, the iterates of GD 'oscillate' while the iterates of SGD 'fluctuate'. In this section we come back to the point of view of the neurons $w_k = (u_k, v_k) \in \mathbb{R}^{2d}$ and make the connection between our previous results and the common understanding of the *EoS* phenomenon for gradient descent. The question we seek to answer is: *in which case does GD enter the EoS regime, and if so, what are the consequences on the trajectory ?* We emphasise that this section aims to provide insights rather than formal statements.

We consider a small initialisation α such that gradient flow converges close to the sparse interpolator $\beta^\star_{\text{sparse}} = \beta_{w^\star_{\text{sparse}}}$. The trajectory of GD as seen in Fig. 3 (left) can be decomposed into up to 3 phases.

(1) **First phase: gradient flow.** The stepsize is appropriate for the local curvature and the iterates of GD remain close to the trajectory of gradient flow. If the stepsize is such that $\gamma < \frac{2}{\lambda_{\max}(\nabla^2 F(w^\star_{\text{sparse}}))}$, then the stepsize is compatible with the local curvature and the GD iterates converge: in this case GF and GD converge to the same point as seen in Fig. 1 for small stepsizes. For larger $\gamma > \frac{2}{\lambda_{\max}(\nabla^2 F(w^\star_{\text{sparse}}))}$, the iterates cannot converge and enter the oscillating phase.

(2) **Second phase: oscillations.** The iterates start oscillating. The gradient of F in the vicinity of $w^\star_{\text{sparse}}$ writes $\nabla_{(u,v)} F(w) \sim (\nabla \mathcal{L}(\beta) \odot v, \nabla \mathcal{L}(\beta) \odot u)$, therefore for $w \sim w^\star_{\text{sparse}}$ we have that $\nabla_u F(w)_i \sim \nabla_v F(w)_i \sim 0$ for $i \notin \mathrm{supp}(\beta^\star_{\text{sparse}})$ and the gradients roughly belong to $\mathrm{Span}(e_i, e_{i+d})_{i\in\mathrm{supp}(\beta^\star_{\text{sparse}})}$. This means that only the coordinates of the neurons (u_i, v_i) for $i \in \mathrm{supp}(\beta^\star_{\text{sparse}})$ can oscillate and similarly for $(\beta_i)_{i\in\mathrm{supp}(\beta^\star_{\text{sparse}})}$.

(3) **Last phase: convergence.** Due to the oscillations, the iterates gradually drift towards a region of lower curvature where they may (potentially) converge. Theorem 1 enables us to understand where they converge: the coordinates of β_k that have oscillated significantly along the trajectory belong to the support of $\beta^\star_{\text{sparse}}$, and therefore $\text{Gain}_\gamma(\text{i})$ becomes much larger for $i \in \text{supp}(\beta^\star_{\text{sparse}})$ than for the other coordinates. Therefore, the coordinates of the solution recovered in the *EoS* regime are heavily penalised on the support of the sparse solution. This is observed in Fig. 3 (left): the oscillations of $(\beta_i)_{i \in \text{supp}(\beta^\star_{\text{sparse}})}$ lead to a gradual shift of these coordinates towards 0, hindering an accurate recovery of the solution $\beta^\star_{\text{sparse}}$.

SGD in the *EoS* regime. In Fig. 3 (right), for stepsizes in the *EoS* regime, juste below the non-convergence threshold, the behavior of SGD is different to that of GD: the fluctuation of the coordinates occurs evenly over all coordinates, leading to a uniform $\boldsymbol{\alpha}_\infty$. These homogeneous fluctuations are reminiscent of label-noise SGD [Andriushchenko et al., 2022] and Pillaud-Vivien et al. [2022] showed that label-noise SGD can recover the sparse interpolator in DLNs.

Flat minima and generalisation. Common knowledges are that flatter minima are beneficial for generalisation, and that larger stepsizes lead to flatter minimas [Nacson et al., 2022, Hochreiter and Schmidhuber, 1997]. However in our DLN case, while larger stepsizes indeed drive the solution to a flatter minima (Fig. 5 (left), appendix A), as seen previously this leads to bad generalisation, illustrating the fact that generalisation is more complex than a story of flatness at the optimum.

Conclusion

We study the effect of stochasticity along with large stepsizes when training DLNs with (S)GD. We prove convergence of the iterates as well as explicitly characterise the recovered solution by exhibiting an implicit regularisation problem which depends on the iterates' trajectory. We show that large stepsizes highly changes the solution recovered by (S)GD with respect to gradient flow. However, surprisingly, the generalisation properties of GD with large stepsizes are highly different to those of SGD: without stochasticity, the use of large stepsizes can prevent the recovery of the sparse interpolator. We also provide insights on the link between the *Edge of Stability* regime and our results.

Aknowledgements. M. Even deeply thanks Laurent Massoulié for making it possible to visit Microsoft Research and the Washington state during an internship supervised by Suriya Gunasekar, the MSR Machine Learning Foundations group for hosting him, and Martin Jaggi for inviting him for a week in Lausanne at EPFL, making it possible to meet and discuss with Scott Pesme and Nicolas Flammarion.

References

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the "edge of stability". *arXiv preprint*, 2022.

M. Andriushchenko, A. Varre, L. Pillaud-Vivien, and N. Flammarion. SGD with large step sizes learns sparse features. *arXiv preprint*, 2022.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, January 2008.

H. H Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2): 330–348, 2017.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

G. Beugnot, J. Mairal, and A. Rudi. On the benefits of large learning rates for kernel methods. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 254–282. PMLR, 02–05 Jul 2022.

G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 483–513. PMLR, 09–12 Jul 2020.

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.

Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

Lei Chen and Joan Bruna. On gradient descent convergence beyond the edge of stability, 2022.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. *On Lazy Training in Differentiable Programming*. 2019.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Alex Damian, Tengyu Ma, and Jason D. Lee. Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, 2021.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023.

J. L. Doob. *Stochastic Processes*. John Wiley & Sons, 1990.

Radu Alexandru Dragomir, Mathieu Even, and Hadrien Hendrikx. Fast stochastic Bregman gradient methods: Sharp analysis and variance reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2815–2825. PMLR, 18–24 Jul 2021.

Mathieu Even and Laurent Massoulie. Concentration of non-isotropic random tensors with applications to learning and empirical risk minimization. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1847–1886. PMLR, 15–19 Aug 2021.

Nicolas Flammarion and Francis Bach. Stochastic composite least-squares regression with convergence rate $o(1/n)$. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 831–875. PMLR, 07–10 Jul 2017.

Jonas Geiping, Micah Goldblum, Phillip E Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. In *International Conference on Learning Representations*, 2022.

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

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841. PMLR, 10–15 Jul 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018b.

Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2315–2357. PMLR, 15–19 Aug 2021.

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

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, January 1997.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 1729–1739, 2017.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 8580–8589, 2018.

Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD, 2017.

Stanisław Jastrzebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amost Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *International Conference on Learning Representations*, 2019.

Stanislaw Jastrzkebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 392–402, 2018.

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

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186, 2020.

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

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

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707. PMLR, 10–15 Jul 2018.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, page 354–363, 2016.

Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In *Advances in Neural Information Processing Systems*, 2021.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *Proceedings of the 39th International*

Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pages 16270–16295. PMLR, 17–23 Jul 2022.

Ryan O'Donnell. Analysis of boolean functions, 2021.

Francesco Orabona, Koby Crammer, and Nicolò Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Mach. Learn.*, 99(3):411–435, jun 2015.

S. Pesme, L. Pillaud-Vivien, and N. Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, 2021.

L. Pillaud-Vivien, J. Reygner, and N. Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2127–2159. PMLR, 2022.

H. Robbins and S. Monro. A stochastic approxiation method. *Ann. Math. Statist*, 22(3):400–407, 1951.

Samuel L. Smith and Quoc V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19(1):2822–2878, jan 2018.

Terrence Tao. Concentration of measure. *254A, Notes 1, Blogpost*, 2010.

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

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022.

Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part II: Continuous time analysis. *arXiv preprint arXiv:2106.02588*, 2021.

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.

Fan Wu and Patrick Rebeschini. A continuous-time mirror descent approach to sparse phase retrieval. In *Advances in Neural Information Processing Systems*, volume 33, pages 20192–20203, 2020.

Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. In *International Conference on Learning Representations*, 2021.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

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

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. *International Conference on Learning Representations*, 2023.

Organization of the Appendix.

(1) In Appendix A, we provide additional experiments, for uncentered data, and on the behaviour of the Hessian (sharpness and trace) along the trajectory of the iterates.
(2) In Appendix B, we prove that (β_k) follows a Mirror descent recursion, with varying potentials. We explicit these potentials and discuss some consequences.
(3) In Appendix C we prove that (S)GD on the $\frac{1}{2}(w_+^2 - w_-^2)$ and $u \odot v$ parametrisation with suitable initialisations lead to the same sequence (β_k).
(4) In Appendix D, we discuss the limit of the hyper-entropy $\psi_{\boldsymbol{\alpha}}$ as a **weighted-ℓ_1**-norm for $\boldsymbol{\alpha} \to 0$ with heterogeneous speed amongst the coordinates. We then discuss the effects of this **weighted-ℓ_1**-norm for sparse recovery.
(5) In Appendix E, we provide our descent lemmas for mirror descent with varying potentials, and prove the boundedness of the iterates.
(6) In Appendix F, we prove our main result, Theorem 1.
(7) In Appendix G, we prove the lemmas and propositions given in the main text.
(8) In Appendix H, we provide technical lemmas used throughout the proof of Theorem 1.
(9) In Appendix I, we provide concentration results for random matrices and random vectors, used to estimate with high probability (wrt the dataset) quantities related to the data.

Appendix A. Additional experiments.

A.1. Uncentered data. Fig. 4: for uncentered data the solutions of GD and SGD have similar behaviours, corroborating Proposition 4.

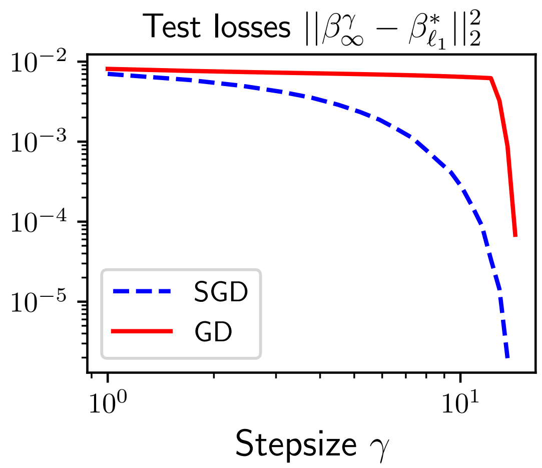

Figure 4. Noiseless sparse regression with a 2-layer DLN with uncentered data $x_i \sim \mathcal{N}(\mu \mathbf{1}, I_d)$ where $\mu = 5$. All the stepsizes lead to convergence to a global solution and the solutions of SGD and GD have similar behaviours, corroborating Proposition 4. The setup corresponds to $(n, d, s, \alpha) = (20, 30, 3, 0.1)$.

A.2. behaviour of the Hessian. Here in Fig. 5, we provide some additional experiments on the behaviour of: (1) the maximum eigenvalue of the hessian $\nabla^2 F(w^\gamma_\infty)$ at the convergence of the iterates of SGD and GD (2) the trace of hessian at the convergence of the iterates. As is clearly observed, increasing the stepsize for GD leads to a 'flatter' minimum in terms of the maximum eigenvalue of the hessian, while increasing the stepsize for SGD leads to a 'flatter' minimum in terms of its trace. These two solutions have very different structures. Indeed from the value of the hessian Eq. (22) at a global solution, and (very) roughly assuming that '$X^\top X = I_d$' and that '$\boldsymbol{\alpha} \sim 0$' (pushing the EoS phenomenon) , one can see that minimising $\lambda_{\max}(\nabla^2 F(w))$ under the constraints $X(w_+^2 - w_-^2) = 0$ and $w_+ \odot w_- = 0$ is equivalent to minimising $\|\beta\|_\infty$ under the constaint $X\beta = y$. On the other hand minimising the trace of the hessian is equivalent to minimising the ℓ_1-norm.

Appendix B. Main ingredients behind the proof of Theorem 1

In this section, we show that the iterates $(\beta_k)_{k\geqslant 0}$ follow a *stochastic mirror descent* with *varying potentials*. At the core of our analysis, this result enables us to *(i)* prove convergence of the iterates to an interpolator and *(ii)* completely characterise the inductive bias of the algorithm (SGD or GD). Unveiling a mirror-descent like structure to characterise the implicit bias of a gradient method is classical. For gradient flow over diagonal linear networks [Woodworth et al., 2020], the iterates follow a mirror flow with respect to the hyper-entropy (5) with parameter α the initialisation scale, while for stochastic gradient flow [Pesme et al., 2021] the mirror flow has a continuously evolving potential.

B.1. Mirror descent and varying potentials. We recall that for a strictly convex reference function $h : \mathbb{R}^d \to \mathbb{R}$, the (stochastic) mirror descent iterates algorithm write as [Beck and Teboulle, 2003, Bauschke et al., 2017, Flammarion and Bach, 2017, Dragomir et al., 2021], where the minimum is assumed to be attained over $\mathbb{R}^d$ and unique:

$$\beta_{k+1} = \mathrm{argmin}_{\beta \in \mathbb{R}^d} \left\{ \eta_k \langle g_k, \beta \rangle + D_h(\beta, \beta_k) \right\}, \tag{18}$$

for stochastic gradients g_k, stepsize $\gamma_k \geqslant 0$, and $D_h(\beta, \beta') = h(\beta) - h(\beta') - \langle \nabla h(\beta'), \beta - \beta' \rangle$ is the Bregman divergence associated to h. Iteration (18) can also be cast as

$$\nabla h(\beta_{k+1}) = \nabla h(\beta_k) - \gamma_k g_k \,. \tag{19}$$

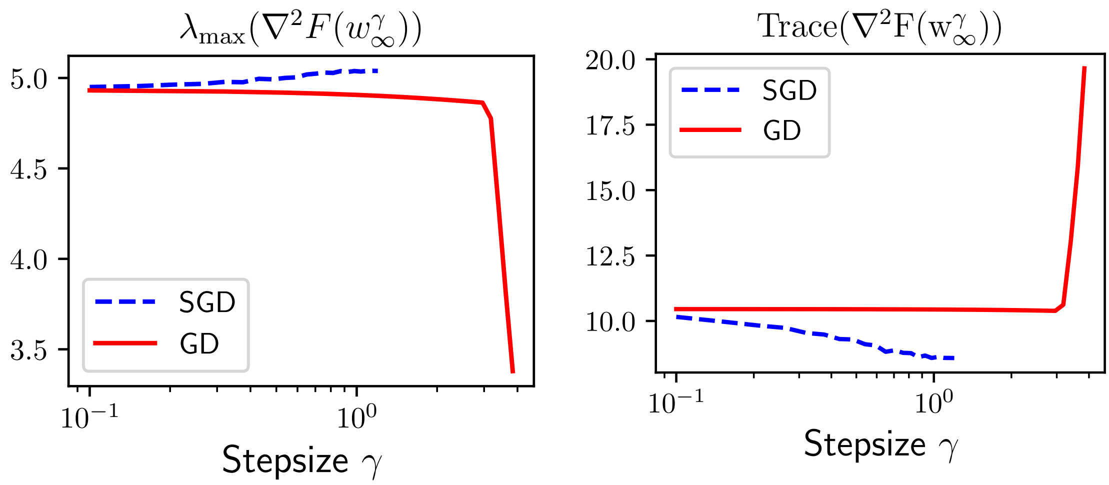

FIGURE 5. Noiseless sparse regression setting. Diagonal linear network. Centered data. Behaviour of 2 different types of flatness of the recovered solution by SGD and GD depending on the stepsize. The setup corresponds to $(n, d, s, \alpha) = (20, 30, 3, 0.1)$.

Now, let (h_k) be strictly convex reference functions $\mathbb{R}^d \to \mathbb{R}$. Whilst in continuous time, there is only one natural way to extend mirror flow to varying potentials, in discrete time the varying potentials can be incorporated in (18) (replacing h by h_k and leading to $\nabla h_k(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k g_k$), the mirror descent with varying potentials we study in this paper incorporates h_{k+1} and h_k in (19). The iterates are thus defined as through:

$$\beta_{k+1} = \mathrm{argmin}_{\beta \in \mathbb{R}^d} \left\{ \eta_k \langle g_k, \beta \rangle + D_{h_{k+1}, h_k}(\beta, \beta_k) \right\},$$

where $D_{h_{k+1}, h_k}(\beta, \beta') = h_{k+1}(\beta) - h_k(\beta') - \langle \nabla h_k(\beta'), \beta - \beta' \rangle$, a recursion that can also be cast as:

$$\nabla h_{k+1}(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k g_k .$$

To derive convergence of the iterates, we prove analogs to classical mirror descent lemmas, generalised to time-varying potentials.

B.2. Time-varying stochastic mirror descent: proof of Proposition 1. Using Proposition 5, we study the $\frac{1}{2}(w_+^2 - w_-^2)$ parametrisation, indeed this is the natural parametrisation to consider when doing the calculations as it "separates" the recursions on w_+ and w_-.

Proposition 1. *$(\beta_k = u_k \odot v_k)_{k \geqslant 0}$ satisfy the Stochastic Mirror Descent recursion with varying potentials $(h_k)_k$:*

$$\nabla h_{k+1}(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k), \tag{8}$$

where $h_k : \mathbb{R}^d \to \mathbb{R}$ for $k \geqslant 0$ are strictly convex functions.

Proof. Let us focus on the recursion of w_+:

$$w_{+,k+1} = (1 - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)) \cdot w_{+,k} .$$

We have:

$$\begin{aligned} w_{+,k+1}^2 &= (1 - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))^2 \cdot w_{+,k}^2 \\ &= \exp\left(\ln((1 - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))^2)\right) \cdot w_{+,k}^2 , \end{aligned}$$

with the convention that $\exp(\ln(0)) = 0$. This leads to:

$$\begin{aligned} w_{+,k+1}^2 &= \exp\left(-2\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(w_k) + 2\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k) + \ln((1 - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))^2)\right) \cdot w_{+,k}^2 \\ &= \exp\left(-2\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k) - q_+(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\right) \cdot w_{+,k}^2 , \end{aligned}$$

since $q_+(x) = -2x - \ln((1-x)^2)$. Expanding the recursion and using that $w_{+,k=0}$ is initialised at $w_{+,k=0} = \alpha$, we thus obtain:

$$\begin{aligned} w_{+,k}^2 &= \alpha^2 \exp(-\sum_{\ell=0}^{k-1} q_+(\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell))) \exp\big(-2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\big) \\ &= \alpha_{+,k}^2 \exp\big(-2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\big)\,, \end{aligned}$$

where we recall that $\alpha_{\pm,k}^2 = \alpha^2 \exp(-\sum_{\ell=0}^{k-1} q_\pm(\gamma_\ell g_\ell))$. One can easily check that we similarly get:

$$w_{-,k}^2 = \alpha_{-,k}^2 \exp\big(+2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\big)\,,$$

leading to:

$$\begin{aligned} \beta_k &= \frac{1}{2}(w_{+,k}^2 - w_{-,k}^2) \\ &= \frac{1}{2}\alpha_{+,k}^2 \exp\big(-2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\big) - \frac{1}{2}\alpha_{-,k}^2 \exp\big(+2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\big)\,. \end{aligned}$$

Using Lemma 3, the previous equation can be simplified into:

$$\beta_k = \alpha_{+,k}\alpha_{-,k} \sinh\Big(-2\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell) + \operatorname{arcsinh}\big(\frac{\alpha_{+,k}^2 - \alpha_{-,k}^2}{2\alpha_{+,k}\alpha_{-,k}}\big)\Big)\,,$$

which writes as:

$$\frac{1}{2}\operatorname{arcsinh}\big(\frac{\beta_k}{\alpha_k^2}\big) - \phi_k = -\sum_{\ell=0}^{k-1} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell) \in \operatorname{span}(x_1, \ldots, x_n)\,,$$

where $\phi_k = \frac{1}{2}\operatorname{arcsinh}\big(\frac{\alpha_{+,k}^2 - \alpha_{-,k}^2}{2\alpha_k^2}\big)$, $\alpha_k^2 = \alpha_{+,k} \odot \alpha_{-,k}$ and since the potentials h_k are defined in Eq. (9) as $h_k = \psi_{\alpha_k} - \langle \phi_k, \cdot \rangle$ with

$$\psi_\alpha(\beta) = \frac{1}{2}\sum_{i=1}^d \Big(\beta_i \operatorname{arcsinh}(\frac{\beta_i}{\alpha_i^2}) - \sqrt{\beta_i^2 + \alpha_i^4} + \alpha_i^2\Big) \tag{20}$$

specifically such that $\nabla h_k(\beta_k) = \frac{1}{2}\operatorname{arcsinh}\big(\frac{\beta_k}{\alpha_k^2}\big) - \phi_k$. Hence,

$$h_k(\beta_k) = \sum_{\ell<k} \gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_\ell}(\beta_\ell)\,,$$

so that:

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

which corresponds to a Mirror Descent with varying potentials $(h_k)_k$. $\square$

Further, we also have the identity:

$$|w_{+,k}||w_{-,k}| = \alpha_{-,k}\alpha_{+,k} = \alpha_k^2\,, \tag{21}$$

which means that α_k is decreasing coordinate wise as long as $|\gamma_\ell \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_\ell)| \leqslant 1$ for all $\ell \geqslant 0$, since $\alpha_{k+1}^2 = \alpha_k^2 e^{-q(\gamma_k g_k)}$, and $q(x) \geqslant 0$ for $|x| \leqslant \sqrt{2}$.

Appendix C. Equivalence of the parametrisations

We here prove the equivalence between the $\frac{1}{2}(w_+^2 - w_-^2)$ and $u \odot v$ parametrisations, that we use throughout the proofs in the Appendix.

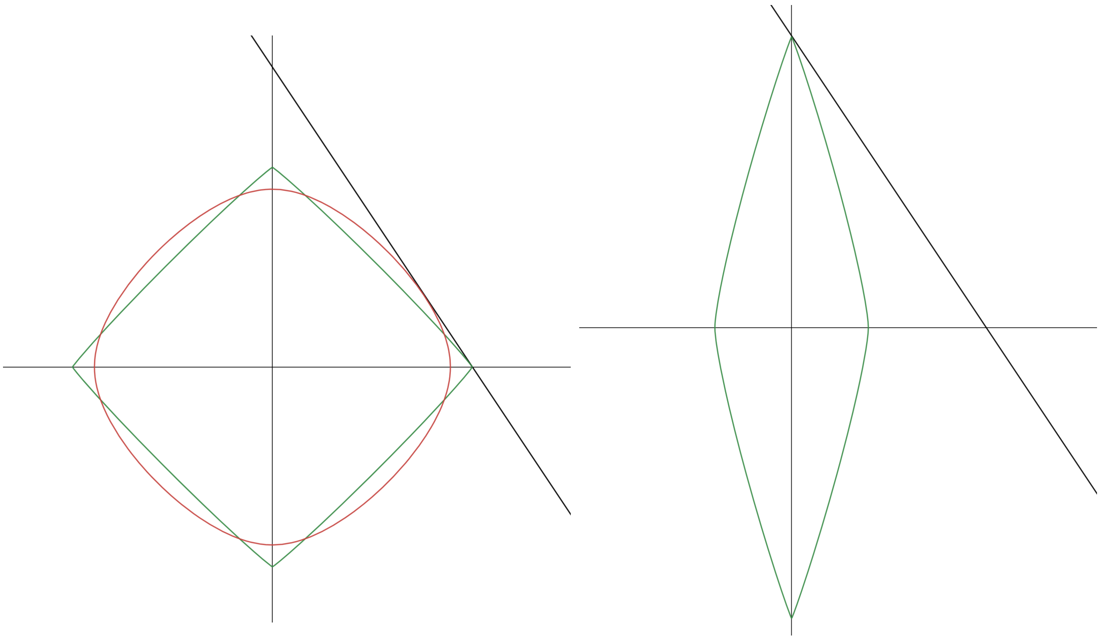

FIGURE 6. *Left*: Uniform $\boldsymbol{\alpha} = \alpha \mathbf{1}$: a smaller scale α leads to the potential ψ_α being closer to the ℓ_1-norm. *Right*: A non uniform $\boldsymbol{\alpha}$ can lead to the recovery of a solution which is very far from the minimum ℓ_1-norm solution. The affine line corresponds to the set of interpolators when $n = 1$, $d = 2$ and $s = 1$.

Proposition 5. *Let $(\beta_k)_{k\geqslant 0}$ and $(\beta'_k)_{k\geqslant 0}$ be respectively generated by stochastic gradient descent on the $u \odot v$ and $\frac{1}{2}(w_+^2 - w_-^2)$ parametrisations:*

$$(u_{k+1}, v_{k+1}) = (u_k, v_k) - \gamma_k \nabla_{u,v}\big(\mathcal{L}_{\mathcal{B}_k}(u \odot v)\big)(u_k, v_k)\,,$$

and

$$w_{\pm,k+1} = w_{\pm,k} - \gamma_k \nabla_{w_\pm}\big(\mathcal{L}_{\mathcal{B}_k}(\frac{1}{2}(w_+^2 - w_-^2))\big)(w_{+,k}, w_{-,k})\,,$$

initialised as $u_0 = \sqrt{2}\alpha, v_0 = 0$ and $w_{+,0} = w_{-,0} = \alpha$. Then for all $k \geqslant 0$, we have $\beta_k = \beta'_k$.

Proof. We have:

$$w_{\pm,0} = \alpha\,, \quad w_{\pm,k+1} = (1 \mp \gamma_k \nabla\mathcal{L}_{\mathcal{B}_k}(\beta'_k))w_{\pm,k}\,,$$

and

$$u_0 = \sqrt{2}\alpha\,, \quad v_0 = 0\,, \quad u_{k+1} = u_k - \gamma_k \nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)v_k\,, \quad v_{k+1} = v_k - \gamma_k \nabla\mathcal{L}(\beta_k)u_k\,.$$

Hence,

$$\beta_{k+1} = (1 + \gamma_k^2 \nabla\mathcal{L}(\beta_k)^2)\beta_k - \gamma_k(u_k^2 + v_k^2)\nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)\,,$$

and

$$\beta'_{k+1} = (1 + \gamma_k^2 \nabla\mathcal{L}_{\mathcal{B}_k}(\beta'_k)^2)\beta'_k - \gamma_k(w_{+,k}^2 + w_{-,k}^2)\nabla\mathcal{L}_{\mathcal{B}_k}(\beta'_k)\,.$$

Then, let $z_k = \frac{1}{2}(u_k^2 - v_k^2)$ and $z'_k = w_{+,k}w_{-,k}$. We have $z_0 = \alpha^2$, $z'_0 = \alpha^2$ and:

$$z_{k+1} = (1 - \gamma_k^2 \nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)^2)z_k\,, \quad z'_{k+1} = (1 - \gamma_k^2 \nabla\mathcal{L}_{\mathcal{B}_k}(\beta'_k)^2)z'_k\,.$$

Using $a^2 + b^2 = \sqrt{(2ab)^2 + (a^2 - b^2)^2}$ for $a, b \in \mathbb{R}$, we finally obtain that:

$$u_k^2 + v_k^2 = \sqrt{(2\beta_k)^2 + (2z_k)^2}\,, \quad w_{+,k}^2 + w_{-,k}^2 = \sqrt{(2\beta'_k)^2 + (2z'_k)^2}\,.$$

We conclude by observing that (β_k, z_k) and (β'_k, z'_k) follow the exact same recursions, initialised at the same value $(0, \alpha^2)$.

$\square$

Appendix D. Convergence of ψ_α to a weighted ℓ_1 norm and pathological behaviour

We show that when taking the scale of the initialisation to 0, one must be careful in the characterisation of the limiting norm, indeed if each entry does not go to zero "at the same speed", then the limit norm is a **weighted** ℓ_1-norm rather than the classical ℓ_1 norm.

Proposition 6. *For $\alpha \geqslant 0$ and a vector $h \in \mathbb{R}^d$, let $\tilde{\alpha} = \alpha \exp(-h \ln(1/\alpha)) \in \mathbb{R}^d$. Then we have that for all $\beta \in \mathbb{R}^d$*

$$\psi_{\tilde{\alpha}}(\beta) \underset{\alpha \to 0}{\sim} \ln(\frac{1}{\alpha}) \cdot \sum_{i=1}^{d} (1 + h_i)|\beta_i|.$$

Proof. Recall that

$$\psi_{\tilde{\alpha}}(\beta) = \frac{1}{2} \sum_{i=1}^{d} \Big(\beta_i \text{arcsinh}(\frac{\beta_i}{\tilde{\alpha}_i^2}) \; - \sqrt{\beta_i^2 + \tilde{\alpha_i}^4} + \tilde{\alpha}_i^2 \Big)$$

Using that $\text{arcsinh}(x) \underset{|x| \to \infty}{\sim} \text{sgn}(x) \ln(|x|)$, and that $\ln(\frac{1}{\tilde{\alpha}_i^2}) = (1 + h_i) \ln(\frac{1}{\alpha^2})$ we obtain that

$$\begin{aligned} \psi_{\tilde{\alpha}}(\beta) &\underset{\alpha \to 0}{\sim} \frac{1}{2} \sum_{i=1}^{d} \text{sgn}(\beta_i) \beta_i (1 + h_i) \ln(\frac{1}{\alpha^2}) \\ &= \frac{1}{2} \ln(\frac{1}{\alpha^2}) \sum_{i=1}^{d} (1 + h_i)|\beta_i|. \end{aligned}$$

□

More generally, for α such that $\alpha_i \to 0$ for all $i \in [d]$ at rates such that $\ln(1/\alpha_i) \sim q_i \ln(1/\max_i \alpha_i)$, we retrieve a weighted ℓ_1 norm:

$$\frac{\psi_\alpha(\beta)}{\ln(1/\alpha^2)} \to \sum_{i=1}^{d} q_i |\beta_i| \,.$$

Hence, even for arbitrary small $\max_i \alpha_i$, if the *shape* of α is 'bad', the interpolator β_α that minimizes ψ_α can be arbitrary far away from $\beta^\star_{\ell^1}$ the interpolator of minimal ℓ_1 norm.

We illustrate the importance of the previous proposition in the following example.

Example 1. *We illustrate how, even for arbitrary small $\max_i \alpha_i$, the interpolator $\beta^\star_\alpha$ that minimizes ψ_α can be far from the minimum ℓ_1 norm solution, due to the shape of $\boldsymbol{\alpha}$ that is not uniform. The message of this example is that for $\boldsymbol{\alpha} \to 0$ non-uniformly across coordinates, if the coordinates of α that go slowly to 0 coincide with the non-null coordinates of the sparse interpolator we want to retrieve, then $\beta^\star_\alpha$ will be far from the sparse solution.*

A simple counterexample can be built: let $\beta^\star_{\text{sparse}} = (1, \ldots, 1, 0, \ldots, 0)$ (with only the $s = o(d)$ first coordinates that are non-null), and let (x_i), (y_i) be generated as $y_i = \langle \beta^\star_{\ell^1}, x_i \rangle$ with $x_i \sim \mathcal{N}(0, 1)$. For n large enough (n of order $s \ln(d)$ where s is the sparsity), the design matrix X is RIP [Candès et al., 2006], so that the minimum ℓ_1 norm interpolator $\beta^\star_{\ell^1}$ is exactly equal to $\beta^\star_{\text{sparse}}$.

However, if α is such that $\max_i \alpha_i \to 0$ with $h_i >> 1$ for $j \leqslant s$ and $h_i = 1$ for $i \geqslant s + 1$ (h_i as in Proposition 6), $\beta^\star_\alpha$ will be forced to verify $\beta^\star_{\alpha,i} = 0$ for $i \leqslant s$ and hence $\|\beta^\star_{\alpha,1} - \beta^\star_{\ell^1}\|_1 \geqslant s$.

Appendix E. Main descent lemma and boundedness of the iterates

The goal of this section is to prove the following proposition, our main descent lemma: for well-chosen stepsizes, the Bregman divergences $(D_{h_k}(\beta^\star, \beta_k))_{k \geqslant 0}$ decrease. We then use this proposition to bound the iterates for both SGD and GD.

Proposition 7. *Let $k \geqslant 0$ and $B > 0$. Provided that $\|\beta_k\|_\infty, \|\beta_{k+1}\|_\infty, \|\beta^\star\|_\infty \leqslant B$ and $\gamma_k \leqslant \frac{c}{LB}$ where $c > 0$ is some numerical constant, we have:*

$$D_{h_{k+1}}(\beta^\star, \beta_{k+1}) \leqslant D_{h_k}(\beta^\star, \beta_k) - \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) \,.$$

E.1. Descent lemma for (stochastic) mirror descent with varying potentials. In the following we adapt a classical mirror descent equality but for time varying potentials, that differentiates from Orabona et al. [2015] in that it enables us to prove the decrease of the Bregman divergences of the iterates. Moreover, as for classical MD, it is an equality.

Proposition 8. *For $h, g : \mathbb{R}^d \to \mathbb{R}$ functions, let $D_{h,g}(\beta, \beta') = h(\beta) - g(\beta') - \langle \nabla g(\beta'), \beta - \beta' \rangle$[4] for $\beta, \beta' \in \mathbb{R}^d$. Let (h_k) strictly convex functions defined $\mathbb{R}^d$ $\mathcal{L}$ a convex function defined on $\mathbb{R}^d$. Let (β_k) defined recursively through $\beta_0 \in \mathbb{R}^d$, and*

$$\beta_{k+1} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \left\{ \gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta - \beta_k \rangle + D_{h_{k+1}, h_k}(\beta, \beta_k) \right\},$$

where we assume that the minimum is unique and attained in $\mathbb{R}^d$. Then, (β_k) satisfies

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

and for any $\beta \in \mathbb{R}^d$,

$$\begin{aligned} D_{h_{k+1}}(\beta, \beta_{k+1}) = D_{h_k}(\beta, \beta_k) - \gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta_k - \beta \rangle + D_{h_{k+1}}(\beta_k, \beta_{k+1}) \\ - \big(h_{k+1} - h_k\big)(\beta_k) + \big(h_{k+1} - h_k\big)(\beta). \end{aligned}$$

Proof. Let $\beta \in \mathbb{R}^d$. Since we assume that the minimum through which β_{k+1} is computed is attained in $\mathbb{R}^d$, the gradient of the function $V_k(\beta) = \gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta - \beta_k \rangle + D_{h_{k+1}, h_k}(\beta, \beta_k)$ evaluated at β_{k+1} is null, leading to $\nabla h_{k+1}(\beta_{k+1}) = \nabla h_k(\beta_k) - \gamma_k \nabla \mathcal{L}(\beta_k)$.

Then, since $\nabla V_k(\beta_{k+1}) = 0$, we have $D_{V_k}(\beta, \beta_{k+1}) = V_k(\beta) - V_k(\beta_{k+1})$. Using $\nabla^2 V_k = \nabla^2 h_{k+1}$, we also have $D_{V_k} = D_{h_{k+1}}$. Hence:

$$D_{h_{k+1}}(\beta, \beta_{k+1}) = \gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta - \beta_{k+1} \rangle + D_{h_{k+1}, h_k}(\beta, \beta_k) - D_{h_{k+1}, h_k}(\beta_{k+1}, \beta_k).$$

We write $\gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta - \beta_{k+1} \rangle = \gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta - \beta^k \rangle + \gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta_k - \beta_{k+1} \rangle$. We also have $\gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta_k - \beta_{k+1} \rangle = \langle \nabla h_k(\beta_k) - \nabla h_{k+1}(\beta_{k+1}), \beta_k - \beta_{k+1} \rangle = D_{h_k, h_{k+1}}(\beta_k, \beta_{k+1}) + D_{h_{k+1}, h_k}(\beta_{k+1}, \beta^k)$, so that $\gamma_k \langle \nabla \mathcal{L}(\beta_k), \beta_k - \beta_{k+1} \rangle - D_{h_{k+1}, h_k}(\beta_{k+1}, \beta^k) = D_{h_k, h_{k+1}}(\beta_k, \beta_{k+1})$. Thus,

$$D_{h_{k+1}}(\beta, \beta_{k+1}) = D_{h_{k+1}, h_k}(\beta, \beta_k) - \gamma_k \big(D_f(\beta, \beta_k) + D_f(\beta_k, \beta)\big) + D_{h_k, h_{k+1}}(\beta_k, \beta_{k+1}),$$

and writing $D_{h,g}(\beta, \beta') = D_g(\beta, \beta') + h(\beta) - g(\beta)$ concludes the proof. ☐

E.2. Proof of Proposition 7. In next proposition, we use Proposition 8 to prove our main descent lemma. To that end, we bound the error terms that appear in Proposition 8 as functions of $\mathcal{L}_{\mathcal{B}_k}(\beta_k)$ and norms of β_k, β_{k+1}, so that for explicit stepsizes, the error terms can be cancelled by half of the negative quantity $-2\mathcal{L}_{\mathcal{B}_k}(\beta_k)$.

Additional notation: let $L_2, L_\infty > 0$ such that $\forall \beta$, $\|H_{\mathcal{B}} \beta\|_2 \leqslant L \|\beta\|_2$, $\|H_{\mathcal{B}} \beta\|_\infty \leqslant L \|\beta\|_\infty$ for all batches $\mathcal{B} \subset [n]$ of size b.

Proposition 7. *Let $k \geqslant 0$ and $B > 0$. Provided that $\|\beta_k\|_\infty, \|\beta_{k+1}\|_\infty, \|\beta^\star\|_\infty \leqslant B$ and $\gamma_k \leqslant \frac{c}{LB}$ where $c > 0$ is some numerical constant, we have:*

$$D_{h_{k+1}}(\beta^\star, \beta_{k+1}) \leqslant D_{h_k}(\beta^\star, \beta_k) - \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k).$$

Proof. From Proposition 8:

$$D_{h_{k+1}}(\beta^*, \beta_{k+1}) = D_{h_k}(\beta^*, \beta_k) - 2\gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) + D_{h_{k+1}}(\beta_{k+1}, \beta_k) - (h_{k+1} - h_k)(\beta_k) + (h_{k+1} - h_k)(\beta^*).$$

We want to bound the last three terms of this equality. First, to bound the last two we apply Lemma 6 assuming that $\|\beta^*\|_\infty, \|\beta_{k+1}\|_\infty \leqslant B$:

$$-(h_{k+1} - h_k)(\beta_k) + (h_{k+1} - h_k)(\beta^*) \leqslant 24 B L_2 \gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)$$

We now bound $D_{h_{k+1}}(\beta_k, \beta_{k+1})$. Classical Bregman manipulations provide that

$$\begin{aligned} D_{h_{k+1}}(\beta_k, \beta_{k+1}) &= D_{h^*_{k+1}}(\nabla h_{k+1}(\beta_{k+1}), \nabla h_{k+1}(\beta_k)) \\ &= D_{h^*_{k+1}}(\nabla h_k(\beta^k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k), \nabla h_{k+1}(\beta_k)). \end{aligned}$$

[4]for $h = g$, we recover the classical Bregman divergence that we denote $D_h = D_{h,h}$

From Lemma 5 we have that h_{k+1} is $\min(1/(4\alpha_{k+1}^2), 1/(4B))$ strongly convex on the ℓ^∞-centered ball of radius B therefore h_{k+1}^* is $\max(4\alpha_{k+1}^2, 4B) = 4B$ (for α small enough or B big enough) smooth on this ball, leading to:

$$\begin{aligned}D_{h_{k+1}}(\beta_k, \beta_{k+1}) &\leqslant 2B\|\nabla h_k(\beta_k) - \gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k) - \nabla h_{k+1}(\beta_k)\|_2^2 \\ &\leqslant 4B\big(\|\nabla h_k(\beta_k) - \nabla h_{k+1}(\beta_k)\|_2^2 + \|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_2^2\big)\,.\end{aligned}$$

Using $|\nabla h_k(\beta) - \nabla h_{k+1}(\beta)| \leqslant 2\delta_k$ where $\delta_k = q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$, we get that:

$$D_{h_{k+1}}(\beta_k, \beta_{k+1}) \leqslant 8B\|\delta_k\|_2^2 + 4BL\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,.$$

Now, $\|\delta_k\|_2^2 \leqslant \|\delta_k\|_1 \|\delta_k\|_\infty$ and using Lemma 4, $\|\delta_k\|_1 \|\delta_k\|_\infty \leqslant 4\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_2^2 \|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty^2 \leqslant 2\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_2^2$ since $\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leqslant \gamma_k L_\infty \|\beta_k - \beta_\infty\| \leqslant \gamma_k \times 2LB \leqslant 1/2$ is verified for $\gamma_k \leqslant 1/(4LB)$. Thus,

$$D_{h_{k+1}}(\beta_k, \beta_{k+1}) \leqslant 40BL_2\gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,.$$

Hence, provided that $\|\beta_k\|_\infty \leqslant B$, $\|\beta_{k+1}\|_\infty \leqslant B$ and $\gamma_k \leqslant 1/(4LB)$, we have:

$$D_{h_{k+1}}(\beta^\star, \beta_{k+1}) \leqslant D_{h_k}(\beta^\star, \beta_k) - 2\gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) + 64L_2\gamma_k^2 B \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,,$$

and thus

$$D_{h_{k+1}}(\beta^\star, \beta_{k+1}) \leqslant D_{h_k}(\beta^\star, \beta_k) - \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k)\,.$$

if $\gamma_k \leqslant \frac{c}{BL}$, where $c = \frac{1}{64}$.

□

E.3. Bound on the iterates. We now bound the iterates (β_k) by an explicit constant B that depends on $\|\beta^\star\|_1$ (for any fixed $\beta^\star \in \mathcal{S}$).

The first bound we prove holds for both SGD and GD, and is of the form $\mathcal{O}(\|\beta^\star\|_1 \ln(1/\alpha^2)$ while the second bound, that holds only for GD ($b = n$) is of order $\mathcal{O}(\|\beta^\star\|_1)$ (independent of α). While a bound independent of α is only proved for GD, we believe that such a result also holds for SGD, and in both cases B should be thought of order $\mathcal{O}(\|\beta^\star\|_1)$.

E.3.1. *Bound that depends on α for GD and SGD.* A consequence of Proposition 7 is the boundedness of the iterates, as shown in next corollary. Hence, Proposition 7 can be applied using B a uniform bound on the iterates ℓ^∞ norm.

Corollary 3. *Let $B = \|\beta^\star\|_1 \ln\big(1 + \frac{\|\beta^\star\|_1}{\alpha^2}\big)$. For stepsizes $\gamma_k \leqslant \frac{c}{BL}$, we have $\|\beta_k\|_\infty \leqslant B$ for all $k \geqslant 0$.*

Proof. We proceed by induction. Let $k \geqslant 0$ such that $\|\beta_k\|_\infty \leqslant B$ for some $B > 0$ and $D_{h_k}(\beta^\star, \beta_k) \leqslant D_{h_0}(\beta^\star, \beta_0)$ (note that these two properties are verified for $k = 0$, since $\beta_0 = 0$). For γ_k sufficiently small (*i.e.*, that satisfies $\gamma_k \leqslant \frac{c}{B'L}$ where $B' \geqslant \|\beta_{k+1}\|_\infty, \|\beta_k\|_\infty, \|\beta^\star\|_\infty$), using Proposition 7, we have $D_{h_{k+1}}(\beta^\star, \beta_{k+1}) \leqslant D_{h_k}(\beta^\star, \beta_k)$ so that $D_{h_{k+1}}(\beta^\star, \beta_{k+1}) \leqslant D_{h_0}(\beta^\star, \beta_0)$, which can be rewritten as:

$$\sum_{i=1}^d \alpha_{k+1,i}^2 \Big(\sqrt{1 + (\frac{\beta_{k+1,i}}{\alpha_{k+1,i}^2})^2} - 1\Big) \leqslant \sum_{i=1}^d \beta_i^\star \operatorname{arcsinh}(\frac{\beta_{k+1,i}}{\alpha^2})\,.$$

Hence, $\|\beta_{k+1}\|_1 \leqslant \|\beta^\star\|_1 \ln(1 + \frac{\|\beta_{k+1}\|_1}{\alpha^2})$. We then notice that for $x, y > 0$, $x \leqslant y\ln(1+x) \implies x \leqslant 3y\ln(1+y)$: if $x > y\ln(1+y)$ and $x > y$, we have that $y\ln(1+y) < y\ln(1+x)$, so that $1 + y < 1 + x$, which contradicts our assumption. Hence, $x \leqslant \max(y, y\ln(1+y))$. In our case, $x = \left\|\beta^{k+1}\right\|_1/\alpha^2$, $y = \|\beta^\star\|_1/\alpha^2$ so that for small alpha, $\ln(1+y) \geqslant 1$.

Hence, we deduce that $\|\beta_{k+1}\|_1 \leqslant B$, where $B = \|\beta^\star\|_1 \ln(1 + \frac{\|\beta^\star\|_1}{\alpha^2})$.

This is true as long as γ_k is tuned using B' a bound on $\max(\|\beta_k\|_\infty, \|\beta_{k+1}\|_\infty)$. Using the continuity of β_{k+1} as a function of γ_k (β_k being fixed), we show that $\gamma_k \leqslant \frac{1}{2} \times \frac{c}{BL}$ can be used using this B. Indeed, let $\phi : \mathbb{R}^+ \to \mathbb{R}^d$ be the function that takes as entry $\gamma_k \geqslant 0$ and outputs the corresponding $\|\beta_{k+1}\|_\infty$: ϕ is continuous. Let $\gamma_r = \frac{1}{2} \times \frac{c}{rL}$ for $r > 0$ and $\bar{r} = \sup\{r \geqslant 0 : B < \phi(\gamma_r)\}$ (the set is upper-bounded; if is is empty, we do not need what follows since it means that any stepsize leads to $\|\beta_{k+1}\|_\infty \leqslant B$). By continuity of ϕ, $\phi(\gamma_{\bar{r}}) = B$. Furthermore, for all r that satisfies $r \geqslant \max(\phi(\gamma_r), B) \geqslant \max(\phi(\gamma_r), \|\beta_k\|_\infty, \|\beta^\star\|_\infty)$, we have, using what is proved just above, that $\|\beta_{k+1}\|_\infty \leqslant B$ and thus $\phi(\gamma_r) \leqslant B$ for such a r:

Lemma 1. *For $r > 0$ such that $r \geqslant \max(\phi(\gamma_r), B)$, we have $\phi(\gamma_r) \leqslant B$.*

Now, if $\bar{r} > B$, by definition of $\bar{r}$ and by continuity of ϕ, since $\phi(\bar{r}) = B$, there exists some $B < r < \bar{r}$ such that $\phi(\gamma_r) > B$ (definition of the supremum) and $\phi(\gamma_r) \leqslant 2B$ (continuity of ϕ). This particular choice of r thus satisfies $r > B$ and and $\phi(\gamma_r) \leqslant 2B \leqslant 2r$, leading to $\phi(\gamma_r) \leqslant B$, using Lemma 1, hence a contradiction: we thus have $\bar{r} \leqslant B$.

This concludes the induction: for all $r \geqslant B$, we have $r \geqslant \bar{r}$ so that $\phi(\gamma_r) \leqslant B$ and thus for all stepsizes $\gamma \leqslant \frac{c}{2LB}$, we have $\|\beta_{k+1}\|_\infty \leqslant B$.

□

E.3.2. *independent of* α. We here assume in this subsection that $b = n$. We prove that for gradient descent, the iterates are bounded by a constant that does not depend on α.

Proposition 9. *Assume that $b = n$ (full batch setting). There exists some $B = \mathcal{O}(\|\beta^\star\|_1)$ such that for stepsizes $\gamma_k \leqslant \frac{c}{BL}$, we have $\|\beta_k\|_\infty \leqslant B$ for all $k \geqslant 0$.*

Proof. We first begin by proving the following proposition: for sufficiently small stepsizes, the loss values decrease. In the following lemma we provide a bound on the gradient descent iterates $(w_{+,k}, w_{-,k})$ which will be useful to show that the loss is decreasing.

Proposition 10. *For $\gamma_k \leqslant \frac{c}{LB}$ where $B \geqslant \max(\|\beta_k\|_\infty, \|\beta_{k+1}\|_\infty)$, we have $\mathcal{L}(\beta_{k+1}) \leqslant \mathcal{L}(\beta_k)$*

Proof. Oddly, using the time-varying mirror descent recursion is not the easiest way to show the decrease of the loss, due to the error terms which come up. Therefore to show that the loss is decreasing we use the gradient descent recursion. Recall that the iterates $w_k = (w_{+,k}, w_{-,k}) \in \mathbb{R}^{2d}$ follow a gradient descent on the non convex loss $F(w) = \frac{1}{2}\|y - \frac{1}{2}X(w_+^2 - w_-^2)\|_2$.

For $k \geqslant 0$, using the Taylor formula we have that $F(w_{k+1}) \leqslant F(w_k) - \gamma_k(1 - \frac{\gamma_k L_k}{2})\|\nabla F(w_k)\|^2$ with the local smoothness $L_k = \sup_{w \in [w_k, w_{k+1}]} \lambda_{\max}(\nabla^2 F(w))$. Hence if $\gamma_k \leqslant 1/L_k$ for all k we get that the loss is non-increasing. We now bound L_k. Computing the hessian ot F, we obtain that:

$$\begin{aligned}\nabla^2 F(w_k) = &\begin{pmatrix} \operatorname{diag}(\nabla\mathcal{L}(\beta_k)) & 0 \\ 0 & -\operatorname{diag}(\nabla\mathcal{L}(\beta_k)) \end{pmatrix} \\ &+ \begin{pmatrix} \operatorname{diag}(w_{+,k})H\operatorname{diag}(w_{+,k}) & -\operatorname{diag}(w_{-,k})H\operatorname{diag}(w_{+,k}) \\ -\operatorname{diag}(w_{+,k})H\operatorname{diag}(w_{-,k}) & \operatorname{diag}(w_{-,k})H\operatorname{diag}(w_{-,k}) \end{pmatrix}.\end{aligned} \tag{22}$$

Let us denote by $M = \begin{pmatrix} M_+ & M_{+,-} \\ M_{+,-} & M_- \end{pmatrix} \in \mathbb{R}^{2d\times 2d}$ the second matrix in the previous equality. With this notation $\|\nabla^2 F(w_k)\| \leqslant \|\nabla\mathcal{L}(\beta_k)\|_\infty + 2\|M\|$ (where the norm corresponds to the Schatten 2-norm which is the largest eigenvalue for symmetric matrices). Now, notice that:

$$\begin{aligned}\|M\|^2 &= \sup_{u\in\mathbb{R}^{2d}, \|u\|=1} \|Mu\|^2 \\ &= \sup_{\substack{u_+\in\mathbb{R}^d, \|u_+\|=1 \\ u_-\in\mathbb{R}^d, \|u_-\|=1 \\ (a,b)\in\mathbb{R}^2, a^2+b^2=1}} \left\| M\begin{pmatrix} a\cdot u_+ \\ b\cdot u_- \end{pmatrix}\right\|^2.\end{aligned}$$

We have:

$$\begin{aligned}\left\| M\begin{pmatrix} a\cdot u_+ \\ b\cdot u_- \end{pmatrix}\right\|^2 &= \left\| \begin{pmatrix} aM_+u_+ + bM_{+-}u_- \\ aM_{+-}u_+ + bM_-u_- \end{pmatrix}\right\|^2 \\ &= \|aM_+u_+ + bM_{+-}u_-\|^2 + \|aM_{+-}u_+ + bM_-u_-\|^2 \\ &\leqslant 2\Big(a^2\|M_+u_+\|^2 + b^2\|M_{+-}u_-\|^2 + a^2\|M_{+-}u_+\|^2 + b^2\|M_-u_-\|^2\Big) \\ &\leqslant 2\Big(\|M_+\|^2 + \|M_{+-}\|^2 + \|M_-\|^2\Big).\end{aligned}$$

Since $\|M_\pm\| \leqslant \lambda_{max}\cdot\|w_\pm\|_\infty^2$ and $\|M_{+-}\| \leqslant \lambda_{max}\|w_+\|_\infty\|w_-\|_\infty$ we finally get that

$$\begin{aligned}\|M\|^2 &\leqslant 6\lambda_{max}^2\cdot\max(\|w_+\|_\infty^2, \|w_-\|_\infty^2)^2 \\ &\leqslant 6\lambda_{max}^2(\|w_+^2\|_\infty + \|w_-^2\|_\infty)^2 \\ &\leqslant 12\lambda_{max}^2\|w_+^2 + w_-^2\|_\infty^2.\end{aligned}$$

We now upper bound this quantity in the following lemma.

Lemma 2. *For all $k \geqslant 0$, the following inequality holds component-wise:*

$$w_{+,k}^2 + w_{-,k}^2 = \sqrt{4\alpha_k^4 + \beta_k^2}\,.$$

Proof. Recall that from Eq. (21) that $\alpha_k^2 = w_k^+ w_k^-$, with $\alpha_0 = \alpha^2$ and since α_k is decreasing (under our assumptions on the stepsizes, $\gamma_k^2 \nabla\mathcal{L}(\beta_k)^2 \leqslant (1/2)^2 < 1$), we get that.:

$$w_{+,k}^2 + w_{-,k}^2 = 2\sqrt{\alpha_k^4 + \beta_k^2} \leqslant 2\sqrt{\alpha^4 + \beta_k^2}$$

leading to $w_{+,k}^2 + w_{-,k}^2 \leqslant \sqrt{4\alpha^4 + B^2}$. □

From Lemma 2, $w_{+,k}^2 + w_{-,k}^2$ is bounded by $2\sqrt{\alpha^4 + B^2}$. Putting things together we finally get that $\|\nabla^2 F(w)\| \leqslant \|\nabla\mathcal{L}(\beta)\|_\infty + 8\lambda_{max}\sqrt{4\alpha^4 + B^2}$. Hence,

$$L_k \leqslant \sup_{\|\beta\|_\infty \leqslant B} \|\nabla\mathcal{L}(\beta)\|_\infty + 8\lambda_{\max}\sqrt{\alpha^4 + B^2} \leqslant LB + 8\lambda_{\max}\sqrt{\alpha^4 + B^2} \leqslant 10LB\,,$$

for $B \geqslant \alpha^2$. □

We finally prove the bound on $\|\beta_k\|_\infty$ independent of α, using the monotonic property of $\mathcal{L}$.

Proposition 11. *Assume that $b = n$ (full batch setting). There exists some $B = \mathcal{O}(\|\beta^\star\|_1)$ such that for stepsizes $\gamma_k \leqslant \frac{c}{BL}$, we have $\|\beta_k\|_\infty \leqslant B$ for all $k \geqslant 0$.*

Proof. In this proof, we first let B be a bound on the iterates. Tuning stepsizes using this bound, we prove that the iterates are bounded by a some $B' = \mathcal{O}(\|\beta^\star\|_1)$. Finally, we conclude by using the continuity of the iterates (at a finite horizon) that this explicit bound can be used to tune the stepsizes.

Writing the mirror descent with varying potentials, we have, since $\nabla h_0(\beta_0) = 0$,

$$\nabla h_k(\beta_k) = -\sum_{\ell<k} \gamma_\ell \nabla\mathcal{L}(\beta_\ell)\,,$$

leading to, by convexity of h_k:

$$h_k(\beta_k) - h_k(\beta^\star) \leqslant \langle \nabla h_k(\beta_k), \beta_k - \beta^\star\rangle = -\sum_{\ell<k}\langle \gamma_\ell \nabla\mathcal{L}(\beta_\ell), \beta_k - \beta^\star\rangle\,.$$

We then write, using $\nabla\mathcal{L}(\beta) = H(\beta - \beta^\star)$ for $H = XX^\top$, that $-\sum_{\ell<k}\langle \gamma_\ell\nabla\mathcal{L}(\beta_\ell), \beta_k - \beta^\star\rangle = -\sum_{\ell<k}\gamma_\ell\langle X^\top(\bar\beta_k - \beta^\star), X^\top(\beta_k - \beta^\star)\rangle \leqslant \sum_{\ell<k}\gamma_\ell\sqrt{\mathcal{L}(\bar\beta_k)\mathcal{L}(\beta_k)}$, leading to:

$$h_k(\beta_k) - h_k(\beta^\star) \leqslant 2\sqrt{\sum_{\ell<k}\gamma_\ell\mathcal{L}(\bar\beta_k)\sum_{\ell<k}\gamma_\ell\mathcal{L}(\beta_k)} \leqslant 2\sum_{\ell<k}\gamma_\ell\mathcal{L}(\bar\beta_k) \leqslant 2D_{h_0}(\beta^\star, \beta^0)\,,$$

where the last inequality holds provided that $\gamma_k \leqslant \frac{1}{CLB}$. Thus,

$$\psi_{\alpha_k}(\beta_k) \leqslant \psi_{\alpha_k}(\beta^\star) + 2\psi_{\alpha_0}(\beta^\star) + \langle \phi_k, \beta_k - \beta^\star\rangle\,.$$

Then, $\langle \phi_k, \beta_k - \beta^\star\rangle \leqslant \|\phi_k\|_1\|\beta_k - \beta^\star\|_\infty$ and $\|\phi_k\|_1 \leqslant C\lambda_{\max}\sum_{k<K}\gamma_k^2\mathcal{L}(\beta^k) \leqslant C\lambda_{\max}\gamma_{\max}h_0(\beta^\star)$. Then, using

$$\|\beta\|_\infty - \frac{1}{\ln(1/\alpha^2)} \leqslant \frac{\psi_\alpha(\beta)}{\ln(1/\alpha^2)} \leqslant \|\beta\|_1\big(1 + \frac{\ln(\|\beta\|_1 + \alpha^2)}{\ln(1/\alpha^2)}\big)\,,$$

we have:

$$\begin{aligned}\|\beta_k\|_\infty &\leqslant \frac{1}{\ln(1/\alpha_k^2)} + \|\beta^\star\|_1\big(1 + \frac{\ln(\|\beta^\star\|_1 + \alpha_k^2)}{\ln(1/\alpha_k^2)}\big) + \|\beta^\star\|_1\big(1 + \frac{\ln(\|\beta^\star\|_1 + \alpha^2)}{\ln(1/\alpha^2)}\big)\\ &\quad + B_0 C\lambda_{\max}\gamma_{\max}h_0(\beta^\star)/\ln(1/\alpha^2)\\ &\leqslant R + B_0 C\lambda_{\max}\gamma_{\max}h_0(\beta^\star)/\ln(1/\alpha^2)\,,\end{aligned}$$

where $R = \mathcal{O}(\|\beta^\star\|_1)$ is independent of α. Hence, since $B_0 = \sup_{k<\infty}\|\beta_k\|_\infty < \infty$, we have:

$$B_0(1 - C\lambda_{\max}\gamma_{\max}h_0(\beta^\star)/\ln(1/\alpha_k^2)) \leqslant R \implies B_0 \leqslant 2R\,,$$

provided that $\gamma_{\max} \leqslant 1/(2C\lambda_{\max}h_0(\beta^\star)/\ln(1/\alpha^2))$ (note that $h_0(\beta^\star)/\ln(1/\alpha^2)$ is independent of α^2).

Hence, if for all k we have $\gamma_k \leqslant \frac{1}{C'LB}$ where B bounds all $\|\beta_k\|_\infty$, we have $\|\beta_k\|_\infty \leqslant 2R$ for all k, where $R = \mathcal{O}(\|\beta^\star\|_1)$ is independent of α and stepsizes γ_k.

Let $K > 0$ be fixed, and

$$\bar{\gamma} = \inf \left\{ \gamma > 0 \quad \text{s.t.} \quad \sup_{k \leqslant K} \|\beta_k\|_\infty > 2R \right\}.$$

For $\gamma \geqslant 0$ a constant stepsize, let

$$\varphi(\gamma) = \sup_{k \leqslant K} \|\beta_k\|_\infty ,$$

which is a continuous function of γ. For $r > 0$, let $\gamma_r = \frac{1}{C'Lr}$.

An important feature to notice is that if $\gamma < \gamma_r$ and r bounds all $\|\beta_k\|_\infty, k \leqslant K$, then $\varphi(\gamma) \leqslant R$, as shown above. We will show that we have $\bar{\gamma} \geqslant \gamma_{2R}$. Reasoning by contradiction, if $\bar{\gamma} < \gamma_{2R}$: by continuity of φ, we have $\varphi(\bar{\gamma}) \leqslant R$ and thus, there exists some small $0 < \varepsilon < \gamma_{2R} - \bar{\gamma}$ such that for all $\gamma \in [\bar{\gamma}, \bar{\gamma} + \varepsilon]$, we have $\varphi(\bar{\gamma}) \leqslant 2R$.

However, such γ's verify both $\varphi(\gamma) \leqslant 2R$ (since $\gamma \in [\bar{\gamma}, \bar{\gamma}+\varepsilon]$ and by definition of ε) and $\gamma \leqslant \gamma_{2R}$ (by definition of ε), and hence $\varphi(\gamma) \leqslant R$. This contradicts the infimum of $\bar{\gamma}$, and hence $\bar{\gamma} \geqslant \gamma_{2R}$. Thus, for $\gamma \leqslant \gamma_{2R} = \frac{1}{2C'LR}$, we have $\|\beta_k\|_\infty \leqslant R$. ☐

☐

Appendix F. Proof of Theorem 1

We are now equipped to prove our main result.

Theorem 1. *Let $(u_k, v_k)_{k \geqslant 0}$ follow the mini-batch SGD recursion (3) initialised at $u_0 = \sqrt{2}\boldsymbol{\alpha} \in \mathbb{R}^d_{>0}$ and $v_0 = \mathbf{0}$, and let $(\beta_k)_{k \geqslant 0} = (u_k \odot v_k)_{k \geqslant 0}$. There exists $B > 0$ and a numerical constant $c > 0$ such that for stepsizes satisfying $\gamma_k \leqslant \frac{c}{LB}$, the iterates satisfy $\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leqslant 1$ and $\|\beta_k\|_\infty \leqslant B$ for all k, and:*

(1) *$(\beta_k)_{k \geqslant 0}$ converges almost surely to some $\beta^\star_\infty \in \mathcal{S}$, that satisfies:*

$$\beta^\star_\infty = \underset{\beta^\star \in \mathcal{S}}{\operatorname{argmin}}\ D_{\psi_{\boldsymbol{\alpha}_\infty}}(\beta^\star, \tilde{\beta}_0) , \tag{6}$$

where $\boldsymbol{\alpha}_\infty \in \mathbb{R}^d_{>0}$ and $\tilde{\beta}_0 \in \mathbb{R}^d$.

(2) *$\boldsymbol{\alpha}_\infty \in \mathbb{R}^d$ satisfies $\boldsymbol{\alpha}_\infty \leqslant \boldsymbol{\alpha}$ and is equal to:*

$$\boldsymbol{\alpha}^2_\infty = \boldsymbol{\alpha}^2 \odot \exp\left(- \sum_{k=0}^{\infty} q\big(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\big) \right) , \tag{7}$$

where $q(x) = -\frac{1}{2}\ln((1-x^2)^2) \geqslant 0$ for $|x| \leqslant \sqrt{2}$, and $\tilde{\beta}_0$ is a perturbation term equal to:

$$\tilde{\beta}_0 = \frac{1}{2}(\boldsymbol{\alpha}^2_+ - \boldsymbol{\alpha}^2_-),$$

where, $q_\pm(x) = \mp 2x - \ln((1 \mp x)^2)$, and $\boldsymbol{\alpha}^2_\pm = \boldsymbol{\alpha}^2 \odot \exp\left(- \sum_{k=0}^{\infty} q_\pm(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))\right)$.

Proof. The first point of the Theorem is a direct consequence of Corollary 3.

Then, for stepsizes $\gamma_k \leqslant \frac{c}{LB}$, using Proposition 7:

$$D_{h_{k+1}}(\beta^\star, \beta_{k+1}) \leqslant D_{h_k}(\beta^\star, \beta_k) - \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) . \tag{23}$$

Hence, summing:

$$\sum_k \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) \leqslant D_{h_0}(\beta^\star, \beta_0) ,$$

so that the series converges.

Under our stepsize rule, $\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leqslant \frac{1}{2}$, leading to $\|q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leqslant 3\|\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|^2_\infty$ by Lemma 4. Using $\|\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|^2 \leqslant 2L_2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)$, we have that $\ln(\alpha_{\pm,k})$, $\ln(\alpha_k)$ all converge.

We now show that $\sum_k \gamma_k \mathcal{L}(\beta_k) < \infty$. We have:

$$\sum_{\ell < k} \mathcal{L}(\beta_k) = \sum_{\ell < k} \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) + M_k ,$$

where $M_k = \sum_{\ell<k} \gamma_k(\mathcal{L}(\beta_k) - \mathcal{L}_{\mathcal{B}_k}(\beta_k))$. We have that (M_k) is a martingale with respect to the filtration $(\mathcal{F}_k)$ defined as $\mathcal{F}_k = \sigma(\beta_\ell, \ell \leqslant k)$. Using our upper-bound on $\sum_{\ell<k} \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k)$, we have:'

$$M_k \geqslant \sum_{\ell<k} \gamma_k \mathcal{L}(\beta_k) - \sum_{\ell<k} \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) \geqslant -D_{h_0}(\beta^\star, \beta_0)\,,$$

and hence (M_k) is a lower bounded martingale. Using Doob's first martingale convergence theorem (a lower bounded super-martingale converges almost surely, Doob [1990]), (M_k) converges almost surely. Consequently, since $\sum_{\ell<k} \gamma_k \mathcal{L}(\beta_k) = \sum_{\ell<k} \gamma_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) + M_k$, we have that $\sum_{\ell<k} \gamma_k \mathcal{L}(\beta_k)$ converges almost surely (the first term is upper bounded, the second converges almost surely).

We now prove the convergence of (β_k). Let β^* be a sublimit of (β_k) and $\sigma : \mathbb{N} \to \mathbb{N}$ increasing such that $\beta_{\sigma(k)} \to \beta^*$.

Almost surely, $\sum_k \gamma_k \mathcal{L}(\beta_k) < \infty$ and so $\gamma_k \mathcal{L}(\beta_k) \to 0$, leading to $\mathcal{L}(\beta_k) \to 0$ since stepsizes are lower bounded, so that $\mathcal{L}(\beta_{\sigma(k)}) \to 0$, and hence $\mathcal{L}(\beta^*) = 0$: β^* is an interpolator. Then, $\nabla h_\infty(\beta^*) = \big(\nabla h_\infty(\beta^*) - \nabla h_\infty(\beta_k)\big) + \big(\nabla h_\infty(\beta_k) - \nabla h_{\sigma(k)}(\beta_k)\big) + \nabla h_{\sigma(k)}(\beta_k)$. The first two terms converge to 0 (the first one is a direct consequence of the convergence to the sublimit, the second one is a consequence of the uniform convergence of $h_{\sigma(k)}$ to h_∞ on compact sets), and the last term is always in $\mathrm{Span}(x_1, \dots, x_n)$, leading to $\nabla h_\infty(\beta) \in \mathrm{Span}(x_1, \dots, x_n)$. Consequently, β^* is an interpolator satisfying $\nabla h_\infty(\beta^*) \in \mathrm{Span}(x_1, \dots, x_n)$: $\beta^* = \beta^\star_{\alpha^\infty}$ which is unique. (β_k) is bounded and only admits β_{α^∞} as a sublimit, hence $\beta_k \to \beta^\star_{\alpha^\infty}$ almost surely.

We have proved the almost sure convergence of (β_k), (α_k), $(\alpha_{\pm,k})$ and (ϕ_k). We now need to justify the expression of $\tilde\beta_0$. We have that $\beta^\star_\infty$ minimizes h_∞ over $\mathcal{S}$, where $h_\infty(\beta) = \psi_{\alpha_\infty}(\beta) - \langle \phi_\infty, \beta\rangle$. Thus, in order to have $h_\infty(\beta) = D_{\psi_{\alpha_\infty}}(\beta, \tilde\beta_0) + \text{constant}$ for all $\beta \in \mathbb{R}^d$, it is necessary and sufficient to have $\nabla\psi_{\alpha_\infty}(\beta) - \phi_\infty = \nabla\psi_{\alpha_\infty}(\beta) - \nabla\psi_{\alpha_\infty}(\tilde\beta_0)$ for all β, *i.e.*, $\nabla\psi_{\alpha_\infty}(\tilde\beta_0) = \phi_\infty$. Since $\phi_\infty = \frac{1}{2}\operatorname{arcsinh}(\frac{\alpha^2_{+,\infty} - \alpha^2_{-,\infty}}{2\alpha^2_\infty})$ and $\nabla\psi_{\alpha_\infty}(\tilde\beta_0) = \frac{1}{2}\operatorname{arcsinh}(\frac{\tilde\beta_0}{\alpha^2_\infty})$, we have $\tilde\beta_0 = \frac{1}{2}(\alpha^2_{+,\infty} - \alpha^2_{-,\infty})$. $\square$

Appendix G. Proof of miscellaneous results mentioned in the main text

In this section, we provide proofs for results mentioned in the main text and that are not directly directed to the proof of Theorem 1.

G.1. Proof of Proposition 2.

Proposition 2. *For any stepsize $\gamma > 0$, initialisation $\alpha\mathbf{1}$ and batch size $b \in [n]$, the magnitude of the gain satisfies:*

$$\lambda_b \gamma^2 \sum_k \mathcal{L}(\beta_k) \leqslant \mathbb{E}\left[\|\mathrm{Gain}_\gamma\|_1\right] \leqslant \Lambda_b \gamma^2 \sum_k \mathcal{L}(\beta_k)\,, \tag{11}$$

where the expectation is over uniform and independent sampling of the batches $(\mathcal{B}_k)_{k\geqslant 0}$ at each iteration. Furthermore, for stepsize $0 < \gamma \leqslant \gamma_{\max} = \frac{c}{BL}$, we have that:

$$\sum_k \gamma^2 \mathcal{L}(\beta_k) = \Theta\left(\gamma \ln\left(\frac{1}{\alpha}\right) \left\|\beta^\star_{\ell_1}\right\|_1\right)\,. \tag{12}$$

Proof. We want to lower bound $\sum_{k<\infty} \gamma_k^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)$. This will lead to an upper-bound on $\sum_i \ln\big(\frac{\alpha_{\infty,i}}{\alpha}\big)$. From Lemma 4, for all $-1/2 \leqslant x \leqslant 1/2$, it holds that $q(x) \geqslant x^2$. We have, using $\|\gamma_k \nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leqslant 1/2$ (the mean is taken with respect to the batches $\mathcal{B}_k$):

$$\begin{aligned}
\mathbb{E}\sum_i \ln\left(\frac{\alpha_{\infty,i}}{\alpha}\right) &= -\sum_{\ell<\infty}\sum_i \mathbb{E} q\big(\gamma_\ell \nabla_i \mathcal{L}_{\mathcal{B}_k}(\beta_\ell)\big) \\
&\leqslant -\sum_{\ell<\infty}\sum_i \mathbb{E}\big(\gamma_\ell \nabla_i \mathcal{L}_{\mathcal{B}_k}(\beta_\ell)\big)^2 \\
&= -\sum_{\ell<\infty} \gamma_\ell^2 \mathbb{E}\|\nabla\mathcal{L}_{\mathcal{B}_k}(\beta_\ell)\|_2^2 \\
&\leqslant -2\lambda_b \sum_{\ell<\infty} \gamma_\ell^2 \mathbb{E}\mathcal{L}_{\mathcal{B}_k}(\beta_\ell)\,.
\end{aligned}$$

To simplify notations, we now assume that stepsizes are constant: $\gamma_k \equiv \gamma$. We have the following equality, that holds for any k:

$$D_{h_{k+1}}(\beta^\star, \beta_{k+1}) = D_{h_k}(\beta^\star, \beta_k) - 2\gamma \mathcal{L}_{\mathcal{B}_k}(\beta_k) + D_{h_{k+1}}(\beta_k, \beta_{k+1}) + \big(h_k - h_{k+1}\big)(\beta_k) - \big(h_k - h_{k+1}\big)(\beta^\star)\,,$$

leading to, by summing for $k \in \mathbb{N}$:

$$\sum_{k<\infty} 2\gamma \mathcal{L}_{\mathcal{B}_k}(\beta_k) = D_{h_0}(\beta^\star, \beta_0) - \lim_{k\to\infty} D_{h_k}(\beta^\star, \beta_k) + \sum_{k<\infty} D_{h_{k+1}}(\beta_k, \beta_{k+1}) + \sum_{k<\infty} \big(h_k - h_{k+1}\big)(\beta_k) - \big(h_k - h_{k+1}\big)(\beta^\star)\,.$$

First, since $h_k \to h_\infty$, $\beta_k \to \beta_\infty$, we have $\lim_{k\to\infty} D_{h_k}(\beta^\star, \beta_k) = 0$. Then, $D_{h_{k+1}}(\beta_k, \beta_{k+1}) \geqslant 0$. Finally, $|\big(h_k - h_{k+1}\big)(\beta_k) - \big(h_k - h_{k+1}\big)(\beta^\star)| \leqslant 16BL_2\gamma^2 \mathcal{L}_{\mathcal{B}_k}(\beta_k)$. Hence :

$$\sum_{k<\infty} 2\gamma(1 + 16\gamma BL_2)\mathcal{L}_{\mathcal{B}_k}(\beta_k) \geqslant D_{h_0}(\beta^\star, \beta_0)\,,$$

and thus $\sum_{k<\infty} \gamma \mathcal{L}_{\mathcal{B}_k}(\beta_k) \geqslant D_{h_0}(\beta^\star, \beta_0)/4$ for $\gamma \leqslant c/(BL)$ (with $c \geqslant 16$). This gives the RHS inequality. The LHS is a direct consequence of bounds proved in previous subsections.

Hence, we have that

$$\gamma^2 \sum_k \mathcal{L}(\beta_k) = \Theta\left(\gamma D_{h_0}(\beta^\star, \beta_0)\right)\,.$$

Noting that $D_{h_0}(\beta^\star, \beta_0) = h_0(\beta^\star) = \Theta\big(\ln(1/\alpha)\|\beta^\star\|_1\big)$ concludes the proof. $\square$

In the following proposition we show that $\tilde\beta_0$ is close to $\mathbf{0}$ and therefore one should think of the implicit regularization problem as $\beta^*_\infty = \operatorname{argmin}_{\beta^* \in S} \psi_{\alpha_\infty}(\beta^*)$

Proposition 12. *Under the assumptions of Theorem 1,*

$$|\tilde\beta_0| \leqslant \alpha^2,$$

where the inequality must be understood coordinate-wise.

Proof.

$$\begin{aligned}
|\tilde\beta_0| &= \frac{1}{2}|\alpha_+^2 - \alpha_-^2| \\
&= \frac{1}{2}\alpha^2\big|\exp(-\sum_k q_+(\gamma_k \nabla\mathcal{L}(\beta_k)) - \exp(-\sum_k q_-(\gamma_k \nabla\mathcal{L}(\beta_k))\big| \\
&\leqslant \alpha^2,
\end{aligned}$$

where the inequality is because $q_+(\gamma_k \nabla\mathcal{L}(\beta_k)) \geqslant 0$, $q_-(\gamma_k \nabla\mathcal{L}(\beta_k)) \geqslant 0$ for all k.

$\square$

G.2. Proof of Corollary 2.

Proof. The bound on λ_b, Λ_b is proved using Lemma 12. Equation (13) is proved using Proposition 2 and the derived estimation of λ_b, Λ_b. $\square$

G.3. Proof of Proposition 3 and Proposition 4.

Proof of Proposition 3. Under Assumption 2, we have using:

$$\begin{aligned}
\nabla\mathcal{L}(\beta_0)^2 &= (X^\top X \beta^\star_{\text{sparse}}) \\
&= (\beta^\star_{\text{sparse}} + \varepsilon)^2 \\
&= {\beta^\star_{\text{sparse}}}^2 + \varepsilon^2 + 2\varepsilon\beta^\star_{\text{sparse}}\,.
\end{aligned}$$

We have $\left\|\varepsilon^2 + 2\varepsilon\beta^\star_{\text{sparse}}\right\|_\infty \leqslant \|\varepsilon\|_\infty^2 + 2\|\varepsilon\|_\infty\left\|\beta^\star_{\text{sparse}}\right\|_\infty$, and we conclude by using $\|\varepsilon\|_\infty \leqslant \delta\left\|\beta^\star_{\text{sparse}}\right\|_2$.

Then,

$$\mathbb{E}_{i\sim\text{Unif}([n])}[\nabla\mathcal{L}_i(\beta_0)^2] = \frac{1}{n}x_i^2\langle x_i, \beta^\star_{\text{sparse}}\rangle\,,$$

and we conclude using Assumption 2.

$\square$

Proof of Proposition 4. The proof proceeds as that of Proposition 3. ☐

G.4. Convergence of α_∞ and $\tilde\beta_0$ for $\gamma \to 0$.

Proposition 13. *Let $\tilde\beta_0(\gamma), \alpha_\infty(\gamma)$ be as defined in Theorem 1, for constant stepsizes $\gamma_k \equiv \gamma$. We have:*

$$\tilde\beta_0(\gamma) \to 0\,, \quad \boldsymbol{\alpha}_\infty \to \alpha\mathbf{1}\,,$$

when $\gamma \to 0$.

Proof. We have, as proved previoulsy, that

$$\begin{aligned}
\left\|\sum_k \gamma^2 \nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)^2\right\|_1 &\leqslant \sum_k \gamma^2 \left\|\nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)^2\right\|_1 \\
&= \sum_k \gamma^2 \|\nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_2^2 \\
&\leqslant 2L\gamma^2 \sum_k \mathcal{L}_{\mathcal{B}_k}(\beta_k) \\
&\leqslant 2L\gamma D_{h_0}(\beta^\star, \beta_0)\,,
\end{aligned}$$

for $\gamma \leqslant \frac{c}{BL}$. Thus, $\sum_k \gamma^2 \nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)^2 \to 0$ as $\gamma \to 0$ (note that β_k implicitly depends on γ, so that this result is not immediate).

Then, for $\gamma \leqslant \frac{c}{LB}$,

$$\left\|\ln(\boldsymbol{\alpha}_\infty^2/\alpha^2)\right\|_1 \leqslant \sum_k \|q(\gamma\mathcal{L}(\beta_k)\|_1 \leqslant 2\sum_k \gamma^2 \left\|\nabla\mathcal{L}_{\mathcal{B}_k}(\beta_k)^2\right\|_1\,,$$

which tends to 0 as $\gamma \to 0$. Similarly, $\left\|\ln(\boldsymbol{\alpha}_{+,\infty}^2/\alpha^2)\right\|_1 \to 0$ and $\left\|\ln(\boldsymbol{\alpha}_{-,\infty}^2/\alpha^2)\right\|_1 \to 0$ as $\gamma \to 0$, leading to $\tilde\beta_0(\gamma) \to 0$ as $\gamma \to 0$.

☐

Appendix H. Technical lemmas

In this section we present a few technical lemmas, used and referred to throughout the proof of Theorem 1.

Lemma 3. *Let $\alpha_+, \alpha_- > 0$ and $x \in \mathbb{R}$, and $\beta = \alpha_+^2 e^x - \alpha_-^2 e^{-x}$. We have:*

$$\operatorname{arcsinh}\left(\frac{\beta}{2\alpha_+\alpha_-}\right) = x + \ln\left(\frac{\alpha_+}{\alpha_-}\right) = x + \operatorname{arcsinh}\left(\frac{\alpha_+^2 - \alpha_-^2}{2\alpha_+\alpha_-}\right).$$

Proof. First,

$$\begin{aligned}
\frac{\beta}{2\alpha_+\alpha_-} &= \frac{1}{2}\left(\frac{\alpha_+}{\alpha^-}e^x - \left(\frac{\alpha_+}{\alpha^-}\right)^{-1} e^{-x}\right) \\
&= \frac{e^{x+\ln(\alpha_+/\alpha_-)} - e^{-x-\ln(\alpha_+/\alpha_-)}}{2} \\
&= \sinh(x + \ln(\alpha_+/\alpha_-))\,,
\end{aligned}$$

hence the result by taking the arcsinh of both sides. Note also that we have $\ln(\alpha_+/\alpha_-) = \operatorname{arcsinh}(\frac{\alpha_+^2-\alpha_-^2}{2\alpha_+\alpha_-})$. ☐

Lemma 4. *If $|x| \leqslant 1/2$ then $x^2 \leqslant q(x) \leqslant 2x^2$*

Lemma 5. *On the ℓ_∞ ball of radius B, the quadratic loss function $\beta \mapsto \mathcal{L}(\beta)$ is $4\lambda_{\max}\max(B,\alpha^2)$-relatively smooth w.r.t all the h_k's.*

Proof. We have:

$$\nabla^2 h_k(\beta) = \operatorname{diag}\left(\frac{1}{2\sqrt{\alpha_k^4 + \beta^2}}\right) \succeq \operatorname{diag}\left(\frac{1}{2\sqrt{\alpha^4 + \beta^2}}\right),$$

since $\alpha_k \leqslant \alpha$ component-wise. Thus, $\nabla^2 h_k(\beta) \succeq \frac{1}{2}\min\left(\min_{1\leqslant i\leqslant d}\frac{1}{2|\beta_i|}, \frac{1}{2\alpha^2}\right) I_d = \frac{1}{\max(4\|\beta\|_\infty, 4\alpha^2)} I_d$, and h_k is $\frac{1}{\max(4B,4\alpha^2)}$-strongly convex on the ℓ^∞ norm of radius B. Since $\mathcal{L}$ is $\lambda_{\max}$-smooth over $\mathbb{R}^d$, we have our result. ☐

Lemma 6. *For $k \geqslant 0$ and for all $\beta \in \mathbb{R}^d$:*

$$|h_{k+1}(\beta) - h_k(\beta)| \leqslant 8L_2\gamma_k^2\mathcal{L}_{\mathcal{B}_k}(\beta_k)\|\beta\|_\infty .$$

Proof. We have $\alpha_{+,k+1}^2 = \alpha_{+,k}^2 e^{-\delta_{+,k}}$ and $\alpha_{-,k+1}^2 = \alpha_{-,k}^2 e^{-\delta_{-,k}}$, for $\delta_{+,k} = \tilde{q}(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$ and $\delta_{-,k} = \tilde{q}(-\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$. And $\alpha_{k+1} = \alpha_k \exp(-\delta_k)$ where $\delta_k := \delta_{+,k} + \delta_{-,k} = q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k))$.

To prove the result we will use that for $\beta \in \mathbb{R}^d$, we have $|(h_{k+1} - h_k)(\beta)| \leqslant \sum_{i=1}^d \int_0^{|\beta_i|} |\nabla_i h_{k+1}(x) - \nabla_i h_k(x)| \mathrm{d}x$.

First, using that$|\operatorname{arcsinh}(a) - \operatorname{arcsinh}(b)| \leqslant |\ln(a/b)|$ for $ab > 0$. We have that

$$\begin{aligned}\left|\operatorname{arcsinh}\left(\frac{x}{\alpha_{k+1}^2}\right) - \operatorname{arcsinh}\left(\frac{x}{\alpha_k^2}\right)\right| &\leqslant \ln\left(\frac{\alpha_k^2}{\alpha_{k+1}^2}\right) \\ &= \delta_k ,\end{aligned}$$

since $\delta_k \geqslant 0$ due to our stepsize condition.

We now prove that $|\phi_{k+1} - \phi_k| \leqslant \frac{|\delta_{+,k} - \delta_{-,k}|}{2}$. We have $\phi_k = \operatorname{arcsinh}\left(\frac{\alpha_{+,k}^2 - \alpha_{-,k}^2}{2\alpha_{+,k}\alpha_{-,k}}\right)$ and hence,

$$|\phi_{k+1} - \phi_k| = \left|\operatorname{arcsinh}\left(\frac{\alpha_{+,k}^2 - \alpha_{-,k}^2}{2\alpha_{+,k}\alpha_{-,k}}\right) - \operatorname{arcsinh}\left(\frac{\alpha_{+,k+1}^2 - \alpha_{-,k+1}^2}{2\alpha_{+,k+1}\alpha_{-,k+1}}\right)\right| .$$

Then, assuming that $\alpha_{+,k,i} \geqslant \alpha_{-,k,i}$, we have:

$$\begin{aligned}\frac{\alpha_{+,k+1,i}^2 - \alpha_{-,k+1,i}^2}{2\alpha_{+,k+1,i}\alpha_{-,k+1,i}} &= e^{\delta_{k,i}/2}\frac{\alpha_{+,k,i}^2 e^{-\delta_{+,k,i}} - \alpha_{-,k,i}^2 e^{-\delta_{-,k,i}}}{2\alpha_{+,k,i}\alpha_{-,k,i}} \\ &\begin{cases} \leqslant \begin{cases} e^{\frac{\delta_{+,k,i} - \delta_{-,k,i}}{2}} \dfrac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}} & \text{if} \quad \delta_{+,k,i} \geqslant \delta_{-,k,i} \\ e^{\frac{\delta_{-,k,i} - \delta_{+,k,i}}{2}} \dfrac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}} & \text{if} \quad \delta_{-,k,i} \geqslant \delta_{+,k,i} \end{cases} \\ \geqslant \begin{cases} e^{-\frac{\delta_{+,k,i} - \delta_{-,k,i}}{2}} \dfrac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}} & \text{if} \quad \delta_{+,k,i} \geqslant \delta_{-,k,i} \\ e^{-\frac{\delta_{-,k,i} - \delta_{+,k,i}}{2}} \dfrac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}} & \text{if} \quad \delta_{-,k,i} \geqslant \delta_{+,k,i} \end{cases} \end{cases} .\end{aligned}$$

We thus have $\frac{\alpha_{+,k+1,i}^2 - \alpha_{-,k+1,i}^2}{2\alpha_{+,k+1,i}\alpha_{-,k+1,i}} \in \left[e^{-\frac{|\delta_{+,k,i} - \delta_{-,k,i}|}{2}}, e^{\frac{|\delta_{+,k,i} - \delta_{-,k,i}|}{2}}\right] \times \frac{\alpha_{+,k,i}^2 - \alpha_{-,k,i}^2}{2\alpha_{+,k,i}\alpha_{-,k,i}}$, and this holds similarly if $\alpha_{+,k,i} \leqslant \alpha_{-,k,i}$. Then, using $|\operatorname{arcsinh}(a) - \operatorname{arcsinh}(b)| \leqslant |\ln(a/b)|$ we obtain that:

$$\begin{aligned}|\phi_{k+1} - \phi_k| &= \left|\operatorname{arcsinh}\left(\frac{\alpha_{+,k}^2 - \alpha_{-,k}^2}{2\alpha_{+,k}\alpha_{-,k}}\right) - \operatorname{arcsinh}\left(\frac{\alpha_{+,k+1}^2 - \alpha_{-,k+1}^2}{2\alpha_{+,k+1}\alpha_{-,k+1}}\right)\right| \\ &\leqslant \frac{|\delta_{+,k} - \delta_{-,k}|}{2} .\end{aligned}$$

Wrapping things up, we have:

$$|\nabla h_k(\beta) - \nabla h_{k+1}(\beta)| \leqslant \delta_k + \frac{|\delta_{+,k} - \delta_{-,k}|}{2} \leqslant 2\delta_k ,$$

This leads to the following bound:

$$\begin{aligned}|h_{k+1}(\beta) - h_k(\beta)| &\leqslant \langle |2\delta_k|, |\beta| \rangle \\ &\leqslant 2\|\delta_k\|_1 \|\beta\|_\infty .\end{aligned}$$

Recall that $\delta_k = q(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)$, hence from Lemma 4 if $\gamma_k \|\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_\infty \leqslant 1/2$, we get that

$$\|\delta_k\|_1 \leqslant 2\gamma_k^2 \|\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\|_2^2 \leqslant 4L_2\gamma_k^2\mathcal{L}_{\mathcal{B}_k}(\beta_k).$$

Putting things together we obtain that

$$\begin{aligned}|h_{k+1}(\beta) - h_k(\beta)| &\leqslant \langle |2\delta_k|, |\beta| \rangle \\ &\leqslant 8L_2\gamma_k^2\mathcal{L}_{\mathcal{B}_k}(\beta_k)\|\beta\|_\infty .\end{aligned}$$

$\square$

Appendix I. Concentration inequalities for matrices

In this last section of the appendix, we provide and prove several concentration bounds for random vectors and matrices, with (possibly uncentered) isotropic gaussian inputs. These inequalities can easily be generalized to subgaussian random variables via more refined concentration bounds, and to non-isotropic subgaussian random variables [Even and Massoulie, 2021], leading to a dependence on an effective dimension and on the subgaussian matrix Σ. We present these lemmas before proving them in a row.

The next two lemmas closely ressemble the RIP assumption, for centered and then for uncentered gaussians.

Lemma 7. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be* i.i.d. *random variables of law $\mathcal{N}(0, I_d)$ and $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. Then, denoting by $\mathcal{C}$ the set of all s-sparse vector $\beta \in \mathbb{R}^d$ satisfying $\|\beta\|_2 \leqslant 1$, there exist $C_4, C_5 > 0$ such that for any $\varepsilon > 0$, if $n \geqslant C_4 s \ln(d) \varepsilon^{-2}$,*

$$\mathbb{P}\left(\sup_{\beta \in \mathcal{S}} \|H\beta - \beta\|_\infty \geqslant \varepsilon\right) \leqslant e^{-C_5 n}\,.$$

Lemma 8. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be* i.i.d. *random variables of law $\mathcal{N}(\mu, \sigma^2 I_d)$ and $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. Then, denoting by $\mathcal{C}$ the set of all s-sparse vector $\beta \in \mathbb{R}^d$ satisfying $\|\beta\|_2 \leqslant 1$, there exist $C_4, C_5 > 0$ such that for any $\varepsilon > 0$, if $n \geqslant C_4 s \ln(d) \varepsilon^{-2}$,*

$$\mathbb{P}\left(\sup_{\beta \in \mathcal{S}} \left\|H\beta - \mu\langle \mu, \beta\rangle - \sigma^2 \beta\right\|_\infty \geqslant \varepsilon\right) \leqslant e^{-C_5 n}\,.$$

We then provide two lemmas that estimate the mean Hessian of SGD.

Lemma 9. *Let $x_1, \ldots, x_n$ be i.i.d. random variables of law $\mathcal{N}(0, I_d)$. Then, there exist $c_1, c_2 > 0$ such that with probability $1 - \frac{1}{d^2}$ and if $n = \Omega(s^{5/4}\ln(d))$, we have for all s-sparse vectors β:*

$$c_1 \|\beta\|_2^2 \mathbf{1} \leqslant \frac{1}{n}\sum_{i=1}^n x_i^2 \langle x_i, \beta\rangle^2 \leqslant c_2 \|\beta\|_2^2 \mathbf{1}\,,$$

where the inequality is meant component-wise.

Lemma 10. *Let $x_1, \ldots, x_n$ be i.i.d. random variables of law $\mathcal{N}(\mu, \sigma^2 I_d)$. Then, there exist $c_0, c_1, c_2 > 0$ such that with probability $1 - \frac{c_0}{d^2} - \frac{1}{nd}$ and if $n = \Omega(s^{5/4}\ln(d))$ and $\mu \geqslant 4\sigma\sqrt{\ln(d)}\mathbf{1}$, we have for all s-sparse vectors β:*

$$\frac{\mu^2}{2}\left(\langle \mu, \beta\rangle^2 + \frac{1}{2}\sigma^2\|\beta\|_2^2\right) \leqslant \frac{1}{n}\sum_i x_i^2 \langle x_i, \beta\rangle^2 \leqslant 4\mu^2\left(\langle \mu, \beta\rangle^2 + 2\sigma^2 \|\beta\|_2^2\right)\,.$$

where the inequality is meant component-wise.

Finally, next two lemmas are used to estimate λ_b, Λ_b in our paper.

Lemma 11. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be* i.i.d. *random variables of law $\mathcal{N}(\mu\mathbf{1}, \sigma^2 I_d)$. Let $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and $\tilde{H} = \frac{1}{n}\sum_{i=1}^n \|x_i\|^2 x_i x_i^\top$. There exist numerical constants $C_2, C_3 > 0$ such that*

$$\mathbb{P}\Big(C_2(\mu^2 + \sigma^2) d H \preceq \tilde{H} \preceq C_3(\mu^2 + \sigma^2) d H\Big) \geqslant 1 - 2ne^{-d/16}\,.$$

Lemma 12. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be* i.i.d. *random variables of law $\mathcal{N}(\mu\mathbf{1}, \sigma^2 I_d)$ for some $\mu \in \mathbb{R}$. Let $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and for $1 \leqslant b \leqslant n$ let $\tilde{H}_b = \mathbb{E}_{\mathcal{B}}\left[\left(\frac{1}{b}\sum_{i \in \mathcal{B}} x_i x_i^\top\right)^2\right]$ where $\mathcal{B} \subset [n]$ is sampled uniformly at random in $\{\mathcal{B} \subset [n]\ s.t.\ |\mathcal{B}| = b\}$. With probability $1 - 2ne^{-d/16} - 3/n^2$, we have, for some numerical constants $c_1, c_2, c_3, C > 0$:*

$$\left(c_1 \frac{d(\mu^2 + \sigma^2)}{b} - c_2 \frac{(\sigma^2 + \mu^2)\ln(n)}{\sqrt{d}} - c_3 \frac{\mu^2 d}{n}\right) H \preceq \tilde{H}_b \preceq C\left(\frac{d(\mu^2 + \sigma^2)}{b} + \frac{(\sigma^2 + \mu^2)\ln(n)}{\sqrt{d}} + \mu^2 d\right)$$

Proof of Lemma 7. For $j \in [d]$, we have:

$$\begin{aligned}
(H\beta)_j &= \frac{1}{n}\sum_{i=1}^n x_{ij}\langle x_i, \beta\rangle \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{j'=1}^d x_{ij}x_{ij'}\beta_{j'} \\
&= \frac{1}{n}\sum_{i=1}^n x_{ij}^2\beta_j + \frac{1}{n}\sum_{i=1}^n \sum_{j'\neq j} x_{ij}x_{ij'}\beta_{j'} \\
&= \frac{\beta_j}{n}\sum_{i=1}^n x_{ij}^2 + \frac{1}{n}\sum_{i=1}^n x_{ij}\sum_{j'\neq j} x_{ij'}\beta_{j'}\,.
\end{aligned}$$

We thus notice that $\mathbb{E}\left[H\beta\right] = \beta$, and

$$(H\beta)_j = \beta_j + \frac{\beta_j}{n}\sum_{i=1}^n (x_{ij}^2 - 1) + \frac{1}{n}\sum_{i=1}^n z_i\,,$$

where $z_i = x_{ij}\sum_{j'\neq j} x_{ij'}\beta_{j'}$, and $\sum_{j'\neq j} x_{ij'}\beta_{j'} \sim \mathcal{N}(0, \|\beta\|^2 - \beta_j^2)$ and $\|\beta\|^2 - \beta_j^2 \leqslant 1$. Hence, $z_j + x_{ij}^2 - 1$ is a centered subexponential random variables (with a subexponential parameter of order 1). Thus, for $t \leqslant 1$:

$$\mathbb{P}\left(\left|\frac{\beta_j}{n}\sum_{i=1}^n (x_{ij}^2 - 1) + \frac{1}{n}\sum_{i=1}^n z_i\right| \geqslant t\right) \leqslant 2e^{-cnt^2}\,.$$

Hence, using an ε-net of $\mathcal{C} = \left\{\beta \in \mathbb{R}^d : \|\beta\|_2 \leqslant 1\,, \|\beta\|_0\right\}$ (of cardinality less than $d^s \times (C/\varepsilon)^s$, and for ε of order 1), we have, using the classical ε-net trick explained in [Chapt. 9, [Vershynin, 2018] or [App. C, Even and Massoulie [2021]]:

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{C},\, j\in[d]} |(H\beta)_j - \beta_j| \geqslant t\right) \leqslant d\times d^s (C/\varepsilon)^s \times 2e^{-cnt^2} = \exp\left(-c\ln(2)nt^2 + (s+1)\ln(d) + s\ln(C/\varepsilon)\right)\,.$$

Consequently, for $t = \varepsilon$ and if $n \geqslant C_4 s\ln(d)/\varepsilon^2$, we have:

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{C},\, j\in[d]} |(H\beta)_j - \beta_j| \geqslant t\right) \leqslant \exp\left(-C_5 nt^2\right)\,.$$

□

Proof of Lemma 8. We write $x_i = \sigma z_i + \mu$ where $z_i \sim \mathcal{N}(0, I_d)$. We have:

$$\begin{aligned}
X^\top X\beta &= \frac{1}{n}\sum_{i=1}^n (\mu + \sigma z_i)\langle \mu + \sigma z_i, \beta\rangle \\
&= \mu\langle\mu,\beta\rangle + \frac{\sigma^2}{n}\sum_{i=1}^n z_i\langle z_i,\beta\rangle + \frac{\sigma}{n}\sum_{i=1}^n \mu\langle z_i,\beta\rangle + \frac{\sigma}{n}\sum_{i=1}^n z_i\langle\mu,\beta\rangle \\
&= \mu\langle\mu,\beta\rangle + \frac{\sigma^2}{n}\sum_{i=1}^n z_i\langle z_i,\beta\rangle + \sigma\mu\langle\frac{1}{n}\sum_{i=1}^n z_i,\beta\rangle + \frac{\sigma\langle\mu,\beta\rangle}{n}\sum_{i=1}^n z_i\,.
\end{aligned}$$

The first term is deterministic and is to be kept. The second one is of order $\sigma^2\beta$ whp using Lemma 7. Then, $\frac{1}{n}\sum_{i=1}^n z_i \sim \mathcal{N}(0, I_d/n)$, so that

$$\mathbb{P}\left(\left|\langle\frac{1}{n}\sum_{i=1}^n z_i,\beta\rangle\right| \geqslant t\right) \leqslant 2e^{-nt^2/(2\|\beta\|_2^2)}\,,$$

and

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n z_{ij}\right| \geqslant t\right) \leqslant 2e^{-nt^2/2}\,.$$

Hence,

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{C}}\left\|\frac{1}{n}\sum_{i=1}^n z_{ij}\right\|_\infty \geqslant t\,,\ \sup_{\beta\in\mathcal{C}}\left|\langle\frac{1}{n}\sum_{i=1}^n z_i,\beta\rangle\right|\geqslant t\right)\leqslant 4e^{cs\ln(d)}e^{-nt^2/2}\,.$$

Thus, with probability $1-Ce^{-n\varepsilon^2}$ and under the assumptions of Lemma 7, we have $\left\|X^\top X\beta-\mu\langle\mu,\beta\rangle-\sigma^2\beta\right\|_\infty\leqslant\varepsilon$ ☐

Proof of Lemma 9. To ease notations, we assume that $\sigma=1$. We remind (O'Donnell [2021], Chapter 9 and Tao [2010]) that for *i.i.d.* real random variables $a_1,\dots,a_n$ that satisfy a tail inequality of the form

$$\mathbb{P}\big(|a_1-\mathbb{E}a_1|\geqslant t\big)\leqslant Ce^{-ct^p}\,, \tag{24}$$

for $p<1$, then for all $\varepsilon>0$ there exists C',c' such that for all t,

$$\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n a_i-\mathbb{E}a_1|\geqslant t\big)\leqslant C'e^{-c'nt^{p-\varepsilon}}\,.$$

We now expand $\frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i,\beta\rangle^2$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i,\beta\rangle^2&=\frac{1}{n}\sum_{i\in[n],k,\ell\in[d]}x_i^2x_{ik}x_{i\ell}\beta_k\beta_\ell\\&=\frac{1}{n}\sum_{i\in[n],k\in[d]}x_i^2x_{ik}^2\beta_k^2+\frac{1}{n}\sum_{i\in[n],k\neq\ell\in[d]}x_i^2x_{ik}x_{i\ell}\beta_k\beta_\ell\,.\end{aligned}$$

Thus, for $j\in[d]$,

$$\left(\frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i,\beta\rangle^2\right)_j=\sum_{k\in[d]}\frac{\beta_k^2}{n}\sum_{i\in[n]}x_{ij}^2x_{ik}^2+\sum_{k\neq\ell\in[d]}\frac{\beta_k\beta_\ell}{n}\sum_{i\in[n]}x_{ij}^2x_{ik}x_{i\ell}\,.$$

We notice that for all indices, all $x_{ij}^2x_{ik}x_{i\ell}$ and $x_{ij}^2x_{ik}^2$ satisfy the tail inequality Eq. (24) for $C=8$, $c=1/2$ and $p=1/2$, so that for $\varepsilon=1/4$:

$$\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n x_{ij}^2x_{ik}x_{i\ell}|\geqslant t\big)\leqslant C'e^{-c'nt^{1/4}}\quad,\quad\mathbb{P}\big(|\frac{1}{n}\sum_{i=1}^n x_{ij}^2x_{ik}^2-\mathbb{E}\left[x_{ij}^2x_{ik}^2\right]|\geqslant t\big)\leqslant C'e^{-c'nt^{1/4}}\,.$$

For $j\neq k$, we have $\mathbb{E}\left[x_{ij}^2x_{ik}^2\right]=1$ while for $j=k$, we have $\mathbb{E}\left[x_{ij}^2x_{ik}^2\right]=\mathbb{E}\left[x_{ij}^4\right]=3$. Hence,

$$\mathbb{P}\left(\exists j,k\neq\ell\,,\ |\frac{1}{n}\sum_{i=1}^n x_{ij}^2x_{ik}x_{i\ell}|\geqslant t\,,\ |\frac{1}{n}\sum_{i=1}^n x_{ij}^2x_{ik}^2-\mathbb{E}\left[x_{ij}^2x_{ik}^2\right]|\geqslant t\right)\leqslant C'd^2e^{-c'nt^{1/4}}\,.$$

Thus, with probability $1-C'd^2e^{-c'nt^{1/4}}$, for all $j\in[d]$,

$$\left|\left(\frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i,\beta\rangle^2\right)_j-2\beta_j^2-\|\beta\|_2^2\right|\leqslant t\sum_{k,\ell}|\beta_k||\beta_\ell|=t\|\beta\|_1^2\,.$$

Using the classical technique of Baraniuk et al. [2008], to make a union bound on all s-sparse vectors, we consider an ε-net of the set of s-sparse vectors of ℓ^2-norm smaller than 1. This ε-net is of cardinality less than $(C_0/\varepsilon)^sd^s$, and we only need to take ε of order 1 to obtain the result for all s-sparse vectors. This leads to:

$$\begin{aligned}\mathbb{P}\left(\exists\beta\in\mathbb{R}^d\ s\text{-sparse and }\|\beta\|_2\leqslant 1\,,\ \exists j\in\mathbb{R}^d\,,\quad\left|\left(\frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i,\beta\rangle^2\right)_j-2\beta_j^2-\|\beta\|_2^2\right|\geqslant t\|\beta\|_1^2\right)\\\leqslant C'd^2e^{c_1s+s\ln(d)}e^{-c'nt^{1/4}}\,.\end{aligned}$$

This probability is equal to C'/d^2 for $t=\left(\frac{(s+4)\ln(d)+c_1s}{c'n}\right)^4$. We conclude that with probability $1-C'/d^2$, all s-sparse vectors β satisfy:

$$\left|\left(\frac{1}{n}\sum_{i=1}^n x_i^2\langle x_i,\beta\rangle^2\right)_j-2\beta_j^2-\|\beta\|_2^2\right|\leqslant\left(\frac{(s+4)\ln(d)+c_1s}{c'n}\right)^4\|\beta\|_1^2\leqslant\left(\frac{(s+4)\ln(d)+c_1s}{c'n}\right)^4 s\|\beta\|_2^2\,,$$

and the RHS is smaller than $\|\beta\|_2^2/2$ for $n \geqslant \Omega(s^{5/4}\ln(d))$. □

Proof of Lemma 10. We write $x_i = \mu + \sigma z_i$ where $x_i \sim \mathcal{N}(0,1)$. We have:

$$\mathbb{P}\big(\forall i \in [n], \forall j \in [d],\, |z_{ij}| \geqslant t\big) \leqslant e^{\ln(nd) - t^2/2} = \frac{1}{nd}\,,$$

for $t = 2\sqrt{\ln(nd)}$. Thus, if $\mu \geqslant 4\sigma\sqrt{\ln(nd)}$ we have $\frac{\mu}{2} \leqslant x_i \leqslant 2\mu$, so that

$$\frac{\mu^2}{2n}\sum_i \langle x_i, \beta\rangle^2 \leqslant \frac{1}{n}\sum_i x_i^2 \langle x_i, \beta\rangle^2 \leqslant \frac{4\mu^2}{n}\sum_i \langle x_i, \beta\rangle^2\,.$$

Then, $\langle x_i, \beta\rangle \sim \mathcal{N}(\langle \mu, \beta\rangle, \sigma^2\|\beta\|_2^2)$. For now, we assume that $\|\beta\|_2 = 1$. We have $\mathbb{P}(|\langle x_i, \beta\rangle^2 - \langle \mu, \beta\rangle^2 - \sigma^2\|\beta\|_2^2| \geqslant t) \leqslant Ce^{-ct/\sigma^2}$, and for $t \leqslant 1$, using concentration of subexponential random variables [Vershynin, 2018]:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i \langle x_i, \beta\rangle^2 - \langle \mu, \beta\rangle^2 - \sigma^2\|\beta\|_2^2\right| \geqslant t\right) \leqslant C' e^{-nc't^2/\sigma^4}\,,$$

and using the ε-net trick of Baraniuk et al. [2008],

$$\mathbb{P}\left(\sup_{\beta \in \mathcal{C}}\left|\frac{1}{n}\sum_i \langle x_i, \beta\rangle^2 - \langle \mu, \beta\rangle^2 - \sigma^2\|\beta\|_2^2\right| \geqslant t\right) \leqslant C' e^{s\ln(d) - nc't^2/\sigma^4} = \frac{C'}{d^2}\,,$$

for $t = \sigma^2\|\beta\|_2^2\sqrt{\frac{2(cs+2)\ln(d)}{n}}$. Consequently, we have, with probability $1 - \frac{C'}{d^2} - \frac{1}{nd}$:

$$\frac{\mu^2}{2}\left(\langle \mu, \beta\rangle^2 + \frac{1}{2}\sigma^2\|\beta\|_2^2\right) \leqslant \frac{1}{n}\sum_i x_i^2\langle x_i, \beta\rangle^2 \leqslant 4\mu^2\left(\langle \mu, \beta\rangle^2 + 2\sigma^2\|\beta\|_2^2\right)\,.$$

□

Proof of Lemma 11. First, we write $x_i = \mu\mathbf{1} + \sigma z_i$, where $z_i \sim \mathcal{N}(0, I)$, leading to:

$$\frac{1}{n}\sum_{i \in [n]} \|x_i\|_2^2 x_i x_i^\top = \frac{1}{n}\sum_{i \in [n]} \big(\sigma^2\|z_i\|_2^2 + d\mu^2 + 2\sigma\mu\langle \mathbf{1}, z_i\rangle\big) x_i x_i^\top$$

We use concentration of χ_d^2 random variables around d:

$$\mathbb{P}(\chi_d^2 > d + 2t + 2\sqrt{dt}) \geqslant t) \leqslant e^{-t} \quad \text{and} \quad \mathbb{P}(\chi_d^2 > d - 2\sqrt{dt}) \leqslant t) \leqslant e^{-t}\,,$$

so that for all $i \in [n]$,

$$\mathbb{P}(\|z_i\|_2^2 \notin [d - 2\sqrt{dt}, d + 2t + 2\sqrt{dt}]) \leqslant 2e^{-t}\,.$$

Thus,

$$\mathbb{P}(\forall i \in [n],\, \|z_i\|_2^2 \in [d - 2\sqrt{dt}, d + 2t + 2\sqrt{dt}]) \geqslant 1 - 2ne^{-t}\,.$$

Taking $t = d/16$,

$$\mathbb{P}(\forall i \in [n],\, \|z_i\|_2^2 \in [\frac{d}{2}, 13d/8]) \geqslant 1 - 2ne^{-d/16}\,.$$

Then, for all i, $\langle \mathbf{1}, z_i\rangle$ is of law $\mathcal{N}(0, d)$, so that $\mathbb{P}(|\langle \mathbf{1}, z_i\rangle| \geqslant t) \leqslant 2e^{-t^2/(2d)}$ and

$$\mathbb{P}\big(\forall i \in [n],\, |\langle \mathbf{1}, z_i\rangle| \geqslant t\big) \leqslant 2ne^{-\frac{t^2}{2d}}\,.$$

Taking $t = \sqrt{2}d^{3/4}$,

$$\mathbb{P}\big(\forall i \in [n],\, |\langle \mathbf{1}, z_i\rangle| \geqslant d^{3/4}\big) \leqslant 2ne^{-d^{1/2}}\,.$$

Thus, with probability $1 - 2n(e^{-d/16} + e^{-\sqrt{d}}$, we have $\forall i \in [n]$, $|\langle \mathbf{1}, z_i\rangle| \geqslant d^{3/4}$ and $\|z_i\|_2^2 \in [\frac{d}{2}, 13d/8]$, so that

$$\big(\frac{d}{2}\sigma^2 + d\mu^2 - 2\mu\sigma d^{3/4}\big)H \preceq \tilde{H} \preceq \big(\frac{13d}{8}\sigma^2 + d\mu^2 + 2\mu\sigma d^{3/4}\big)H\,,$$

leading to the desired result. □

Proof of Lemma 12. We have:

$$
\begin{aligned}
\tilde{H}_b &= \mathbb{E}\left[\frac{1}{b^2}\sum_{i,j\in\mathcal{B}}\langle x_i, x_j\rangle x_i x_j^\top\right] \\
&= \mathbb{E}\left[\frac{1}{b^2}\sum_{i\in\mathcal{B}}\|x_i\|_2^2 x_i x_i^\top + \frac{1}{b^2}\sum_{i,j\in\mathcal{B},\, i\neq j}\langle x_i, x_j\rangle x_i x_j^\top\right] \\
&= \frac{1}{b^2}\sum_{i\in[n]}\mathbb{P}(i\in\mathcal{B})\|x_i\|_2^2 x_i x_i^\top + \frac{1}{b^2}\sum_{i\neq j}\mathbb{P}(i,j\in\mathcal{B})\langle x_i, x_j\rangle x_i x_j^\top \,.
\end{aligned}
$$

Then, since $\mathbb{P}(i\in\mathcal{B}) = \frac{b}{n}$ and $\mathbb{P}(i,j\in\mathcal{B}) = \frac{b(b-1)}{n(n-1)}$ for $i\neq j$, we get that:

$$
\tilde{H}_b = \frac{1}{bn}\sum_{i\in[n]}\|x_i\|_2^2 x_i x_i^\top + \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top \,.
$$

Using Lemma 11, the first term satisfies:

$$
\mathbb{P}\Big(\frac{d(\mu^2+\sigma^2)}{b}C_2 H \preceq \frac{1}{bn}\sum_{i\in[n]}\|x_i\|_2^2 x_i x_i^\top \preceq \frac{d(\mu^2+\sigma^2)}{b}C_3 H\Big) \geqslant 1 - 2ne^{-d/16}\,.
$$

We now show that the second term is of smaller order. Writing $x_i = \mu\mathbf{1} + \sigma z_i$ where $z_i\sim\mathcal{N}(0, I_d)$, we have:

$$
\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top = \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top
$$

For $i\neq j$, $\langle x_i, x_j\rangle = \sum_{k=1}^d x_{ik}x_{jk} = \sum_{k=1}^d a_k$ where $a_k = x_{ik}x_{jk}$ satisfies $\mathbb{E}a_k = 0$, $\mathbb{E}a_k^2 = 1$ and $\mathbb{P}(a_k\geqslant t)\leqslant 2\mathbb{P}(|x_{ik}|\geqslant\sqrt{t})\leqslant 4e^{-t/2}$. Hence, a_k is a centered subexponential random variables. Using concentration of subexponential random variables [Vershynin, 2018], for $t\leqslant 1$,

$$
\mathbb{P}\left(\frac{1}{d}|\langle x_i, x_j\rangle|\geqslant t\right)\leqslant 2e^{-cdt^2}\,.
$$

Thus,

$$
\mathbb{P}\left(\forall i\neq j,\ \frac{1}{d}|\langle x_i, x_j\rangle|\leqslant t\right)\geqslant 1 - n(n-1)e^{-cdt^2}\,.
$$

Then, taking $t = d^{-1/2}4\ln(n)/c$, we have:

$$
\mathbb{P}\left(\forall i\neq j,\ \frac{1}{d}|\langle x_i, x_j\rangle|\leqslant\frac{4\ln(n)}{c\sqrt{d}}\right)\geqslant 1 - \frac{1}{n^2}\,.
$$

Going back to our second term,

$$
\begin{aligned}
\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top &= \frac{(b-1)}{bn(n-1)}\sum_{i<j}\langle x_i, x_j\rangle\big(x_i x_j^\top + x_j x_i^\top\big) \\
&\preceq \frac{(b-1)}{bn(n-1)}\sum_{i<j}|\langle x_i, x_j\rangle|\big(x_i x_i^\top + x_j x_j^\top\big)\,,
\end{aligned}
$$

where we used $x_i x_j^\top + x_j x_i^\top \preceq x_i x_i^\top + x_j x_j^\top$. Thus,

$$
\begin{aligned}
\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle x_i x_j^\top &\preceq \sup_{i\neq j}|\langle x_i, x_j\rangle|\times\frac{(b-1)}{bn(n-1)}\sum_{i<j}\big(x_i x_i^\top + x_j x_j^\top\big) \\
&= \sup_{i\neq j}|\langle x_i, x_j\rangle|\times\frac{b-1}{b}\frac{1}{n-1}\sum_{i=1}^n x_i x_i^\top \\
&= \sup_{i\neq j}|\langle x_i, x_j\rangle|\times\frac{b-1}{b}\frac{n}{n-1}H\,.
\end{aligned}
$$

Similarly, we have

$$\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i,x_j\rangle x_i x_j^\top \succeq -\sup_{i\neq j}|\langle x_i,x_j\rangle| \times \frac{b-1}{b}\frac{n}{n-1}H\,.$$

Hence, with probability $1-1/n^2$,

$$-\frac{4\ln(n)}{c\sqrt{d}}\times\frac{b-1}{b}\frac{n}{n-1}H \preceq \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i,x_j\rangle x_i x_j^\top \preceq \frac{4\ln(n)}{c\sqrt{d}}\times\frac{b-1}{b}\frac{n}{n-1}H\,.$$

Wrapping things up, with probability $1-1/n^2-2ne^{-d/16}$,

$$\left(-\frac{4\ln(n)}{c\sqrt{d}}\frac{b-1}{b}\frac{n}{n-1}+C_2\frac{d}{b}\right)\times H\preceq \tilde H_b\preceq\left(\frac{4\ln(n)}{c\sqrt{d}}\frac{b-1}{b}\frac{n}{n-1}+C_3\frac{d}{b}\right)\times H\,.$$

Thus, provided that $\frac{4\ln(n)}{c\sqrt{d}}\leqslant\frac{C_2 d}{2b}$ and $d\geqslant 48\ln(n)$, we have with probability $1-3/n^2$:

$$C_2'\frac{d}{b}\times H\preceq\tilde H_b\preceq C_3'\frac{d}{b}\times H\,.$$

□

Proof of Lemma 12. We have:

$$\begin{aligned}
\tilde H_b &= \mathbb{E}\left[\frac{1}{b^2}\sum_{i,j\in\mathcal{B}}\langle x_i,x_j\rangle x_i x_j^\top\right]\\
&= \mathbb{E}\left[\frac{1}{b^2}\sum_{i\in\mathcal{B}}\|x_i\|_2^2 x_i x_i^\top+\frac{1}{b^2}\sum_{i,j\in\mathcal{B},\,i\neq j}\langle x_i,x_j\rangle x_i x_j^\top\right]\\
&= \frac{1}{b^2}\sum_{i\in[n]}\mathbb{P}(i\in\mathcal{B})\|x_i\|_2^2 x_i x_i^\top+\frac{1}{b^2}\sum_{i\neq j}\mathbb{P}(i,j\in\mathcal{B})\langle x_i,x_j\rangle x_i x_j^\top\,.
\end{aligned}$$

Then, since $\mathbb{P}(i\in\mathcal{B})=\frac{b}{n}$ and $\mathbb{P}(i,j\in\mathcal{B})=\frac{b(b-1)}{n(n-1)}$ for $i\neq j$, we get that:

$$\tilde H_b=\frac{1}{bn}\sum_{i\in[n]}\|x_i\|_2^2 x_i x_i^\top+\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i,x_j\rangle x_i x_j^\top\,.$$

Using Lemma 11, the first term satisfies:

$$\mathbb{P}\Big(\frac{d(\mu^2+\sigma^2)}{b}C_2H\preceq\frac{1}{bn}\sum_{i\in[n]}\|x_i\|_2^2x_ix_i^\top\preceq\frac{d(\mu^2+\sigma^2)}{b}C_3H\Big)\geqslant 1-2ne^{-d/16}\,.$$

We now show that the second term is of smaller order. Writing $x_i=\mu\mathbf{1}+\sigma z_i$ where $z_i\sim\mathcal{N}(0,I_d)$, we have:

$$\begin{aligned}
\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\langle x_i,x_j\rangle x_i x_j^\top &= \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i,z_j\rangle+\sigma\mu\langle\mathbf{1},z_i+z_j\rangle+\mu^2 d\big)x_i x_j^\top\\
&= \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i,z_j\rangle+\sigma\mu\langle\mathbf{1},z_i+z_j\rangle\big)x_i x_j^\top+\frac{(b-1)}{bn(n-1)}\mu^2 d\sum_{i\neq j}x_i x_j^\top
\end{aligned}$$

For $i\neq j$, $\langle z_i,z_j\rangle=\sum_{k=1}^d z_{ik}z_{jk}=\sum_{k=1}^d a_k$ where $a_k=z_{ik}z_{jk}$ satisfies $\mathbb{E}a_k=0$, $\mathbb{E}a_k^2=1$ and $\mathbb{P}(a_k\geqslant t)\leqslant 2\mathbb{P}(|z_{ik}|\geqslant\sqrt{t})\leqslant 4e^{-t/2}$. Hence, a_k is a centered subexponential random variables. Using concentration of subexponential random variables [Vershynin, 2018], for $t\leqslant 1$,

$$\mathbb{P}\left(\frac{1}{d}|\langle x_i,x_j\rangle|\geqslant t\right)\leqslant 2e^{-cdt^2}\,.$$

Thus,

$$\mathbb{P}\left(\forall i\neq j,\ \frac{1}{d}|\langle x_i,x_j\rangle|\leqslant t\right)\geqslant 1-n(n-1)e^{-cdt^2}\,.$$

Then, taking $t = d^{-1/2}4\ln(n)/c$, we have:

$$\mathbb{P}\left(\forall i \neq j,\ \frac{1}{d}|\langle x_i, x_j\rangle| \leqslant \frac{4\ln(n)}{c\sqrt{d}}\right) \geqslant 1 - \frac{1}{n^2}\,.$$

For $i \in [n]$, $\langle \mathbf{1}, z_i\rangle \sim \mathcal{N}(0,d)$ so that $\mathbb{P}(|\langle \mathbf{1}, z_i\rangle| \geqslant t) \leqslant 2e^{-t^2/(2d)}$, and

$$\mathbb{P}(\forall i \in [n],\ |\langle \mathbf{1}, z_i\rangle| \leqslant t) \geqslant 1 - 2ne^{-t^2/(2d)} = 1 - \frac{2}{n^2}\,,$$

for $t = 3\sqrt{d}\ln(n)$. Hence, with probability $1-3/n^2$, for all $i \neq j$ we have $|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i + z_j\rangle| \leqslant (\sigma^2 + \sigma\mu)C\ln(n)/\sqrt{d}$.

Now,

$$\begin{aligned}\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big)x_ix_j^\top &= \frac{(b-1)}{bn(n-1)}\sum_{i<j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big)(x_ix_j^\top + x_jx_i^\top)\\ &\preceq \frac{(b-1)}{bn(n-1)}\sum_{i<j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big)\big|\big(x_ix_i^\top + x_jx_j^\top\big)\,,\end{aligned}$$

where we used $x_ix_j^\top + x_jx_i^\top \preceq x_ix_i^\top + x_jx_j^\top$. Thus,

$$\begin{aligned}\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big)x_ix_j^\top &\preceq \sup_{i\neq j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big)\big| \times \frac{(b-1)}{bn(n-1)}\sum_{i<j}\big(x_ix_i^\top + x_jx_j^\top\big)\\ &= \sup_{i\neq j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big| \times \frac{b-1}{b}\frac{1}{n-1}\sum_{i=1}^n x_ix_i^\top\\ &= \sup_{i\neq j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big| \times \frac{b-1}{b}\frac{n}{n-1}H\,.\end{aligned}$$

Similarly, we have

$$\frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big)x_ix_j^\top \succeq -\sup_{i\neq j}\big|\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big)\big| \times \frac{b-1}{b}\frac{n}{n-1}H\,.$$

Hence, with probability $1 - 3/n^2$,

$$\begin{aligned}-\frac{(\sigma^2+\sigma\mu)C\ln(n)}{\sqrt{d}} \times \frac{b-1}{b}\frac{n}{n-1}H &\preceq \frac{(b-1)}{bn(n-1)}\sum_{i\neq j}\big(\sigma^2\langle z_i, z_j\rangle + \sigma\mu\langle \mathbf{1}, z_i+z_j\rangle\big)x_ix_j^\top\\ &\preceq \frac{(\sigma^2+\sigma\mu)C\ln(n)}{\sqrt{d}} \times \frac{b-1}{b}\frac{n}{n-1}H\,.\end{aligned}$$

We thus have shown that this term (the one in the middle of the above inequality) is of smaller order.

We are hence left with $\frac{(b-1)}{bn(n-1)}\mu^2 d\sum_{i\neq j}x_ix_j^\top$. Denoting $\bar{x} = \frac{1}{n}\sum_i x_i$, we have $\frac{1}{n^2}\sum_{i\neq j}x_ix_j^\top = \frac{1}{n^2}\sum_{i,j}x_ix_j^\top - \frac{1}{n^2}\sum_i x_ix_i^\top$, so that:

$$\frac{(b-1)}{bn(n-1)}\mu^2 d\sum_{i\neq j}x_ix_j^\top = \frac{(b-1)n}{b(n-1)}\mu^2 d\left(\bar{x}\bar{x}^\top - \frac{1}{n}H\right)\,.$$

We note that we have $H = \frac{1}{n}\sum_i x_ix_i^\top = \frac{1}{n^2}\sum_{i<j}x_ix_i^\top + x_jx_j^\top \succeq \frac{1}{n^2}\sum_{i<j}x_ix_j^\top + x_jx_i^\top = \bar{x}\bar{x}^\top$ using $x_ix_i^\top + x_jx_j^\top \succeq x_ix_j^\top + x_jx_i^\top$. Thus, $H \succeq \bar{x}\bar{x}^\top \succeq 0$, and:

$$-\frac{(b-1)n}{b(n-1)}\mu^2 d\frac{1}{n}H \preceq \frac{(b-1)}{bn(n-1)}\mu^2 d\sum_{i\neq j}x_ix_j^\top \preceq \frac{(b-1)n}{b(n-1)}\mu^2 d(1-1/n)H\,.$$

We are now able to wrap everything together. With probability $1 - 2ne^{-d/16} - 3/n^2$, we have, for some numerical constants $c_1, c_2, c_3, C > 0$:

$$\left(c_1\frac{d(\mu^2+\sigma^2)}{b} - c_2\frac{(\sigma^2+\mu^2)\ln(n)}{\sqrt{d}} - c_3\frac{\mu^2 d}{n}\right)H \preceq \tilde{H}_b \preceq C\left(\frac{d(\mu^2+\sigma^2)}{b} + \frac{(\sigma^2+\mu^2)\ln(n)}{\sqrt{d}} + \mu^2 d\right)$$

$\square$