École Normale Supérieure

Département d'informatique

Rapport de stage

Non-linear parameterizations of least-squares gradient flow: mirror flow, implicit bias and convergence speed

David A. R. Robin

david.robin@ens.fr

Oct 2020 — Mar 2021

Advisor

Lénaïc Chizat
(Laboratoire de Mathématiques d'Orsay)

Contents

1 Context & motivation

Artificial neural networks have demonstrated puzzling efficacy in a wide range of seemingly very difficult classification tasks, despite little theorical guarantees to back these sucesses. In particular, simple architectures trained only with a basic gradient descent and no explicit regularization reach impressive prediction accuracies, even on non-convex problems, and in regimes where the number of parameters of the model far exceeds the number of available samples (also known as "over-parameterized" regime).

1.1 Implicit bias and reparameterizations

The main idea to explain the quality of the obtained predictor in the over-parameterized regime is to describe not the growing number of parameters and their evolution, but rather the prediction function that is learned by the network. If the prediction function converges over time to a well-behaved limit predictor, then the quality of the network's predictions can be explained by the regularities of the limit predictor. This is known as the "implicit bias" of the algorithm, for it does not appear explicitely as a regularization term in the objective. On the contrary, it is a bias naturally enforced by the choice of gradient descent as a learning algorithm. Gradient descent does not converge to *any* minimizer of the objective, but rather to a very specific and possibly very well-behaved one. Using this approach, [Chizat and Bach, 2020] show that two-layer networks trained with gradient descent on a logistic loss (a.k.a. cross-entropy) learn a form of max-margin classifier of the data. The analysis is not quantitative, but the limit predictor exhibits interesting properties that guarantee strong generalization bounds, for instance in the presence of low-dimensional structure in the data, the maximal margin does not depend on the ambient dimension.

In the setting of linear networks (i.e. without non-linearities) for the squared loss, [Woodworth et al., 2020] show that the implicit bias depends heavily on the parameterization and choice of initialization. In particular, the linear predictor $\beta \mapsto X\beta$ will learn the interpolator minimizing the ℓ_2-difference to initialization if gradient descent is performed on the parameters β, but will learn an interpolator minimizing an ℓ_1-like disparity to initialization if the gradient descent is performed on the reparameterization w satisfying $w \odot w = \beta$. The latter option, akin to a two-layer linear network, induces an implicit bias that promotes sparsity of the linear predictor learned.

1.2 Choice of focus for this internship

The goal of this internship was to extend these results on the implicit bias of two-layer networks to make them quantitative, ideally with convergence speed guarantees that accurately describe the behavior of networks in settings similar to the widespread use of neural networks.

From the max-margin convergence analysis, we retain the continuous gradient flow point of view, and the separation of the first layer weights into a direction (on the sphere) and a magnitude. [Chizat and Bach, 2020] showed that with the logistic loss, if the directions converge, then they must converge to the max-margin directions, however there is no guarantee that they will converge. It therefore seems unlikely that we could obtain guarantees on convergence speed on the directions. To simplify, we consider the directions fixed during training. This fixed-direction choice is questionnable, but the results could still be interesting. Because of the logistic loss, the magnitudes of weights diverged to infinity over time. In contrast, the squared loss is coercive and will have weights that do not diverge over time. For these reasons, we choose to focus on two-layer networks with fixed directions for the first layer, trained by gradient flow on a squared loss objective.

2 Mirror descent for over-parameterized least-squares

Following the approach of [Woodworth et al., 2020], we will show that it is possible to reduce the setting we are interested in to linear predictor following a mirror flow. We start by analyzing this setting in depth, and show the reduction to this form in the following section.

We consider an over-parameterized linear regression problem with the square loss. Given a design matrix $X \in \mathbb{R}^{n\times d}$ with $\operatorname{rank}(X) = n < d$ and a response vector $y \in \mathbb{R}^n$, the objective function $F : \mathbb{R}^d \to \mathbb{R}$ is the square loss

$$F(\beta) \coloneqq \frac{1}{2}\|X\beta - y\|_2^2.$$

Let $\phi : \mathcal{D} \to \mathbb{R}$ be a convex function of *Legendre type*[1] defined on a nonempty convex open set $\mathcal{D} \subset \mathbb{R}^d$ such that $\nabla\phi(\mathcal{D}) = \mathbb{R}^d$[2]. We study the continuous time mirror descent dynamics starting from $\beta_0 \in \mathcal{D}$, which we will refer to as the *mirror flow*. It is the unique solution $\beta : \mathbb{R}_+ \to \mathcal{D}$ to the following system

$$\begin{cases} \beta(0) = \beta_0, \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(\beta(t)) = -\nabla F(\beta(t)). \end{cases}$$

See Lemma (A.1) in appendix for proof of existence and uniqueness of this solution.

2.1 Implicit bias of the mirror flow

For clarity in the statement of the following two theorems, we distinguish two types of minimizers of the objective.

Definition 2.1. *(Following Bauschke and Borwein [1997, Sec. 3.3])*

$$\arg\inf_{\mathcal{D}} F = \left\{ \bar\beta \in \bar{\mathcal{D}} \;\middle|\; \forall\beta \in \mathcal{D},\, F(\bar\beta) \le F(\beta) \right\}$$

$$\arg\min_{\mathcal{D}} F = (\arg\inf_{\mathcal{D}} F) \cap \mathcal{D}$$

The implicit bias of this dynamics was characterized in [Gunasekar et al., 2018, Thm. 1] in terms of the Bregman divergence $D_\phi(\beta_1, \beta_0) \coloneqq \phi(\beta_1) - \phi(\beta_0) - \langle\nabla\phi(\beta_0), \beta_1 - \beta_0\rangle$. Here we state an improved version of their result where we remove a convergence assumption[3].

Theorem 2.2 (Implicit bias, qualitative)**.** *Assume that* $\arg\min_{\mathcal{D}} F \neq \emptyset$*. Then the mirror flow converges to the unique Bregman projection of* β_0 *onto the set of minimizers of* F *i.e.* $\lim_{t\to\infty}\beta(t) = \beta^*$ *where*

$$\{\beta^*\} = \arg\min\left\{D_\phi(\beta, \beta_0) \;:\; \beta \in \arg\min_{\mathcal{D}} F\right\}. \tag{1}$$

The existence and uniqueness of solutions to the problem in Eq. (1) under our assumptions is shown in [Bauschke and Borwein, 1997, Theorem 3.12]. The unique solution $\beta^* \in \mathcal{D}$ is characterized by

$$\begin{cases} \beta^* \in \arg\min_{\mathcal{D}} F \\ \nabla\phi(\beta^*) - \nabla\phi(\beta_0) \in \operatorname{Im}(X^\top) \end{cases}$$

[1] That is ϕ is strictly convex, differentiable and $\lim_{\beta\to\partial\mathcal{D}} \|\nabla\phi(\beta)\| = +\infty$, see [Rockafellar, 1970, Sec. 26] or [Bauschke and Borwein, 1997, Sec. 2].

[2] Needed for coercivity of $D_\phi(\beta, \cdot)$ [Bauschke and Borwein, 1997, Cor. 3.11].

[3] In Gunasekar et al. [2018], it is assumed that $\beta(t)$ admits a limit β^*, and that $X\beta^* = y$.

Proof. We first start by exploiting the classical mirror descent argument, see e.g. [Bubeck, 2015, Chap. 4]. Let $\bar{\beta} \in \arg\min_{\mathcal{D}} F \subseteq \mathcal{D}$. By the definition of D_ϕ and convexity of F we have for $t > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} D_\phi\left(\bar{\beta}, \beta(t)\right) = \left\langle \nabla F(\beta(t)), \bar{\beta} - \beta(t) \right\rangle \leq -\left(F\left(\beta(t)\right) - F\left(\bar{\beta}\right)\right) \leq 0. \tag{2}$$

As we have assumed that $\nabla\phi(\mathcal{D}) = \mathbb{R}^d$, the sublevel sets of $\beta \mapsto D_\phi(\bar{\beta}, \beta)$ are compact [Bauschke and Borwein, 1997, Cor. 3.11] so there exists a limit point $\beta^* \in \bar{\mathcal{D}}$. Integrating Inequality (2) and using that $t \mapsto F(\beta(t))$ is decreasing and $D_\phi(\bar{\beta}, \beta(t)) \geq 0$, it follows for $t > 0$,

$$F(\beta(t)) - F(\bar{\beta}) \leq \frac{1}{t} \int_0^t \left(F(\beta(s)) - F(\bar{\beta})\right) \mathrm{d}s \leq \frac{1}{t} D_\phi\left(\bar{\beta}, \beta(0)\right). \tag{3}$$

As a consequence, any limit point β^* must satisfy $F(\beta^*) = F(\bar{\beta})$.

The second part of the proof now follows Gunasekar *et al.*'s argument. Since $\nabla F(\beta) \in \mathrm{Im}(X^\top) = \ker(X)^\perp$ it follows that $\nabla\phi(\beta(t)) - \nabla\phi(\beta(0)) = -\int_0^t \nabla F(\beta(s))\mathrm{d}s \in \mathrm{Im}(X^\top)$. Given the optimality conditions of Equation (2.1) and since $\nabla\phi$ is continuous on $\mathcal{D}$ it only remains to show that any limit point $\beta^* \notin \partial\mathcal{D}$.

Indeed, if $\beta(t) \to \beta^* \in \partial\mathcal{D}$, then in particular $\|\nabla\phi(\beta(t))\| \to \infty$ by hypothesis, thus $D_\phi(\bar{\beta}, \beta(t)) = D_{\phi^*}(\nabla\phi(\beta(t)), \nabla\phi(\bar{\beta})) \to \infty$ by coercivity (by [Bauschke and Borwein, 1997, Fact. 2.11], since ϕ Legendre implies ϕ^* closed convex proper, and $\bar{\beta} \in \arg\min_{\mathcal{D}} F \subseteq \mathcal{D}$), which contradicts the decrease proven in Equation (2). Hence $\beta^* \notin \partial\mathcal{D}$ and $F(\beta^*) = F(\bar{\beta})$, therefore $\beta^* \in \arg\min_{\mathcal{D}} F$. Since $g : t \mapsto D_\phi(\beta^*, \beta(t))$ is decreasing and β^* is a limit point of β, g must tend to zero, and so it follows that $\beta(t) \underset{t\to\infty}{\longrightarrow} \beta^*$. □

At the expense of a slightly more intricate proof, the assumption that the minimum in attained inside the domain can be removed in some cases. In the event that no such minimum exists in the domain, i.e. $\arg\min_{\mathcal{D}} F = \emptyset$, the minimizers of F are all on the border $\arg\inf_{\mathcal{D}} F \subseteq \partial\mathcal{D}$, and the mirror flow will converge to a limit point on the border $\partial\mathcal{D}$ as well.

Theorem 2.3 (Implicit bias, qualitative, extended)**.** *If the mirror flow potential admits a continuous extension $\phi : \bar{\mathcal{D}} \to \mathbb{R}$, then the mirror flow $\beta : \mathbb{R}_+ \to \mathcal{D}$ converges to the unique Bregman projection of β_0 onto the set of minimizers of F i.e. $\lim_{t\to\infty} \beta(t) = \beta^*$ where*

$$\{\beta^*\} = \arg\min\left\{D_\phi(\beta, \beta_0) \; : \; \beta \in \arg\inf_{\mathcal{D}} F \subseteq \bar{\mathcal{D}}\right\}. \tag{4}$$

Moreover, $(\arg\min_{\mathcal{D}} F \neq \emptyset) \Leftrightarrow (\beta^* \in \arg\min_{\mathcal{D}} F \subseteq \mathcal{D})$*.*

Proof. The Bregman divergence is extended continuously to $D_\phi : \bar{\mathcal{D}} \to \mathbb{R}_+$. The existence and uniqueness of solutions again follows from [Bauschke and Borwein, 1997, Theorem 3.12]. The classical mirror descent argument, see e.g. [Bubeck, 2015, Chap. 4], extends naturally. Let $\bar{\beta} \in \arg\inf_{\mathcal{D}} F \subseteq \bar{\mathcal{D}}$. By definition of D_ϕ and convexity of F, we have for $t > 0$

$$\frac{\mathrm{d}}{\mathrm{d}t} D_\phi\left(\bar{\beta}, \beta(t)\right) = \left\langle \nabla F(\beta(t)), \bar{\beta} - \beta(t) \right\rangle \leq -\left(F\left(\beta(t)\right) - F\left(\bar{\beta}\right)\right) \leq 0$$

Since [Bauschke and Borwein, 1997, Cor. 3.11] guarantee compactness of the sublevel sets $\beta \mapsto D_\phi(\bar{\beta}, \beta)$ only if $\bar{\beta} \in \mathcal{D}$, instead we note that $t \mapsto F(\beta(t))$ is also decreasing since $\frac{\mathrm{d}}{\mathrm{d}t} F(\beta) = \nabla F(\beta) \cdot \frac{\mathrm{d}}{\mathrm{d}t}\beta = -\nabla F(\beta) \cdot \nabla^2\phi(\beta)^{-1} \cdot \nabla F(\beta) \leq 0$, and F has compact sublevel sets, so there exists a limit point $\beta^* \in \bar{\mathcal{D}}$. Integrating the previous inequality and using this decrease as previously, together with $D_\phi(\bar{\beta}, \beta(t)) \geq 0$, it follows for $t > 0$,

$$F(\beta(t)) - F(\bar{\beta}) \leq \frac{1}{t} \int_0^t \left(F(\beta(s)) - F(\bar{\beta})\right) \mathrm{d}s \leq \frac{1}{t} D_\phi\left(\bar{\beta}, \beta(0)\right). \tag{5}$$

As a consequence, any limit point β^* must satisfy $F(\beta^*) = F(\bar{\beta})$, thus $\beta^* \in \arg\inf_{\mathfrak{D}} F$, and as previously $\beta(t) \to \beta^*$. Moreover, it holds $X\beta^* = X\bar{\beta}$, otherwise $\frac{1}{2}(\beta^* + \bar{\beta}) \in \bar{\mathfrak{D}}$ would achieve strictly lower loss by strict convexity of the ℓ_2 loss $u \mapsto \|u - y\|_2^2$. It then only remains to show that $D_\phi(\beta^*, \beta_0) \leq D_\phi(\bar{\beta}, \beta_0)$. For this purpose, let $f : t \mapsto D_\phi(\beta^*, \beta(t)) - D_\phi(\bar{\beta}, \beta(t))$. Observe that $\frac{\mathrm{d}f}{\mathrm{d}t}(t) = \langle \nabla F(\beta(t)), \beta^* - \bar{\beta}\rangle = (X\beta(t) - y)^T(X\beta^* - X\bar{\beta}) = 0$. Hence $f(0) = \lim_{t\to\infty} f(t) \leq 0$, which concludes the proof of the first claim.

For the second claim, if $\arg\min_{\mathfrak{D}} F \neq \emptyset$, then let $\bar{\beta} \in \arg\min_{\mathfrak{D}} F$, and we obtain as previously that $D_\phi(\bar{\beta}, \beta(t)) \to \infty$ if $\beta^* \in \partial\mathfrak{D}$, thus $\beta^* \in \mathfrak{D}$ because $D_\phi(\bar{\beta}, \beta(t))$ is decreasing and finite at $t = 0$. The other implication is immediate. □

2.2 Local convergence speed

In the next theorem, we give assumptions under which convergence to a minimizer is guaranteed and also gives an asymptotic convergence speed. This control will be obtained by showing that the system satisfies a Łojasiewicz condition, then integrating with respect to time to bound the remaining path length.

Lemma 2.4 (Least-squares-induced Bregman divergence is Mahalanobis distance)**.** *The least-squares objective $F : \beta \mapsto \frac{1}{2}\|X\beta - y\|_2^2$ induces the symmetric Bregman divergence $D_F : (\beta_0, \beta_1) \mapsto \frac{1}{2}\|X\beta_0 - X\beta_1\|_2^2$, also known as Mahalanobis squared distance*[4] $\|\beta_0 - \beta_1\|^2_{X^\top X}$

Proof. Since the objective F is a convex function, it induces a well-defined Bregman divergence $D_F : (\beta_0, \beta_1) \mapsto F(\beta_0) - F(\beta_1) - \langle \nabla F(\beta_1), \beta_0 - \beta_1\rangle$. Moreover, since $F(\beta) = \frac{1}{2}\|X\beta - y\|_2^2$, it has gradient $\nabla F(\beta) = X^\top(X\beta - y)$. Thus for any $(\beta_0, \beta_1) \in \mathfrak{D}^2$, we can rewrite the divergence using bilinearity of the inner product in the following way

$$\begin{aligned}
D_F(\beta_0, \beta_1) &= \frac{1}{2}\|X\beta_0 - y\|_2^2 - \frac{1}{2}\|X\beta_1 - y\|_2^2 - (X\beta_1 - y)^T X(\beta_0 - \beta_1) \\
&= \frac{1}{2}\|X\beta_0 - y\|_2^2 - \frac{1}{2}\|X\beta_1 - y\|_2^2 - (X\beta_1 - y)^T(X\beta_0 - y) + (X\beta_1 - y)^T(X\beta_1 - y) \\
&= \frac{1}{2}\left(\|X\beta_0 - y\|_2^2 + \|X\beta_1 - y\|_2^2 - 2(X\beta_0 - y)^T(X\beta_1 - y)\right) \\
&= \frac{1}{2}\|X\beta_0 - X\beta_1\|_2^2
\end{aligned}$$

□

Lemma 2.5 (Convergence under θ-Łojasiewicz condition)**.** *Assume that there exists a radius $r > 0$, a multiplicative constant $\kappa > 0$ and an exponent $\theta \geq 0$ such that on the set $B_{\beta^*}(r) := \{\beta \in \mathfrak{D} \ : \ D_\phi(\beta^*, \beta) \leq r, \nabla\phi(\beta) \in \nabla\phi(\beta^*) + \mathrm{Im}(X^\top)\}$, it holds $D_F(\beta^*, \beta) \geq \frac{\kappa}{2} D_\phi(\beta^*, \beta)^\theta$. Then $T_0 = \inf\{t \in \mathbb{R}_+ | D_\phi(\beta^*, \beta(t)) \leq r\} < \infty$. Moreover for all $t \geq T_0$, the following differential inequality is satisfied: $\frac{\mathrm{d}}{\mathrm{d}t} D_\phi(\beta^*, \beta(t)) \leq -\kappa D_\phi(\beta^*, \beta(t))^\theta$.*

The $\frac{\kappa}{2}$ constant instead of just κ in the Łojasiewicz condition is due to the symmetry of the Bregman divergence D_F. Introducing the $\frac{1}{2}$ scaling only slightly complicates the proof and yields tighter bounds in the following sections.

Proof. By Theorem (2.2), $\beta(t) \to \beta^*$, therefore by continuity of the Bregman divergence, $D_\phi(\beta^*, \beta(t)) \to D_\phi(\beta^*, \beta^*) = 0$. Hence, define $T_0 = \inf\{t \in \mathbb{R}_+ \ : \ D_\phi(\beta^*, \beta(t)) \leq r\}$. Let us show now that the set $B_{\beta^*}(r)$ is stable by the training dynamics. As previously, since $\nabla F(\beta) \in \mathrm{Im}(X^\top)$ for all β, it holds at all times $\nabla\phi(\beta(t)) - \nabla\phi(\beta^*) = \int_t^\infty \nabla F(\beta(s))ds \in \mathrm{Im}(X^\top)$. Moreover, $D_\phi(\beta^*, \beta(T_0)) \leq r$, and by Inequality (2), for all $t \geq T_0$, it holds $D_\phi(\beta^*, \beta(t)) \leq$

[4] It is not always a distance, in the sense that it does not separate points if X does not have full rank

$D_\phi(\beta^*, \beta(T_0)) \leq r$, thus for all $t \geq T_0$, $\beta(t) \in B_{\beta^*}(r)$. Then, by definition of the Bregman divergence, and symmetry of the previously computed divergence D_F, $\langle \nabla F(\beta^*) - \nabla F(\beta), \beta^* - \beta \rangle = D_F(\beta^*, \beta) + D_F(\beta, \beta^*) = 2D_F(\beta^*, \beta)$. However, since β^* is a minimum of F over the convex set $\mathcal{D}$, it satisfies the optimality condition $\forall \beta \in \mathcal{D}$, $\langle \nabla F(\beta^*), \beta - \beta^* \rangle \geq 0$ [Rockafellar, 1970, Theorem 25.6]. Therefore, for any $\beta \in \mathcal{D}$, it holds $\langle -\nabla F(\beta), \beta^* - \beta \rangle \geq 2D_F(\beta^*, \beta)$. The result follows from the derivation argument of Inequality (2), the symmetry, and the hypothesis.

$$\frac{\mathrm{d}}{\mathrm{d}t} D_\phi(\beta^*, \beta(t)) = \langle \nabla F(\beta(t)), \beta^* - \beta(t) \rangle \leq -2D_F(\beta^*, \beta(t)) \leq -\kappa D_\phi(\beta^*, \beta(t))^\theta \tag{6}$$

□

2.2.1 Local linear convergence speed

Theorem 2.6 (Implicit bias, local linear convergence)**.** *Assume that there exists two norms $\|\cdot\|$ on $\mathcal{D}$ and $\|\cdot\|_*$ on $\mathbb{R}^n$, and a radius $r > 0$ such that $\beta^* \in \mathcal{D}$, and on the set $B_{\beta^*}(r) := \{\beta \in \mathcal{D} \,:\, D_\phi(\beta^*, \beta) \leq r,\, \nabla\phi(\beta) \in \nabla\phi(\beta^*) + \mathrm{Im}(X^\top)\}$, it holds $\nabla\phi : (\mathcal{D}, \|\cdot\|) \to (\mathbb{R}^n, \|\cdot\|_*)$ is L_ϕ-Lipschitz continuous. Then $T_0 = \inf\{t \in \mathbb{R}_+ | D_\phi(\beta^*, \beta(t)) \leq r\} < \infty$. Moreover, for all $t \geq T_0$, it holds*

$$D_\phi(\beta^*, \beta(t)) \leq D_\phi(\beta^*, \beta(T_0)) \exp\left(-\frac{1}{L_\phi \left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2} (t - T_0) \right)$$

Where $\left\|X^{\top\dagger}\right\|_{\mathrm{op}} = \sup_{\|v\|_* \leq 1} \left\| (X^\top)^\dagger v \right\|_2$ is the operator norm of the pseudo-inverse of $X^\top$.

While this shows that L_ϕ-Lipschitz continuity of $\nabla\phi$ is sufficient to get linear convergence, this bound risks becoming loose when L_ϕ grows. If this happens only far from the optimum, then decreasing the radius r will grant a local linear convergence bound with a more reasonable constant. If even near the optimum L_ϕ is too large, we show in the next section a variation of this technique that can give a more interesting bound.

Proof. Let $\kappa_1 = \left(L_\phi \left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2 \right)^{-1}$. We will start by showing that the system satisfies a 1-Łojasiewicz condition $D_F(\beta^*, \beta) \geq \frac{\kappa_1}{2} D_\phi(\beta^*, \beta)$ for all $\beta \in B_{\beta^*}(r)$. The claim shall then be easily obtained by applying Lemma (2.5) and integrating with respect to time.

The L_ϕ-Lipschitz continuity implies that for $\beta \in B_{\beta^*}(r)$, it holds $\langle \nabla\phi(\beta) - \nabla\phi(\beta^*), \beta - \beta^* \rangle \leq L_\phi^{-1} \|\nabla\phi(\beta) - \nabla\phi(\beta^*)\|_*^2$ by co-coercivity of the gradient. Let $\beta \in B_{\beta^*}(r)$. By definition of $B_{\beta^*}(r)$, there exists $u \in \mathbb{R}^n$ such that $\nabla\phi(\beta) - \nabla\phi(\beta^*) = X^\top u$. Then, leveraging the Cauchy-Schwarz inequality (a), projection smoothness (b) and potential smoothness (c), we obtain

$$\begin{aligned}
\langle \nabla\phi(\beta) - \nabla\phi(\beta^*), \beta - \beta^* \rangle^2 &= \langle X^\top u, \beta - \beta^* \rangle^2 = \langle X\beta - X\beta^*, u \rangle^2 \\
&\leq \|X\beta - X\beta^*\|_2^2 \|u\|_2^2 && (a) \\
&\leq \left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2 \|X\beta - X\beta^*\|_2^2 \|X^\top u\|_*^2 && (b) \\
&\leq L_\phi \left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2 \|X\beta - X\beta^*\|_2^2 \langle \nabla\phi(\beta) - \nabla\phi(\beta^*), \beta - \beta^* \rangle && (c)
\end{aligned}$$

Since by convexity $\langle \nabla\phi(\beta) - \nabla\phi(\beta^*), \beta - \beta^* \rangle \geq 0$ [Rockafellar, 1970, Sec. 24], this implies

$$\langle \nabla\phi(\beta) - \nabla\phi(\beta^*), \beta - \beta^* \rangle \leq L_\phi \left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2 \|X\beta - X\beta^*\|_2^2 = \frac{2}{\kappa_1} D_F(\beta^*, \beta)$$

By definition of the Bregman divergence $D_\phi(\beta^*,\beta)+D_\phi(\beta,\beta^*)=\langle\nabla\phi(\beta)-\nabla\phi(\beta^*),\beta-\beta^*\rangle$. By positivity of the Bregman divergence $D_\phi(\beta,\beta^*)\geq 0$, this implies the previously announced 1-Łojasiewicz condition $D_F(\beta^*,\beta)\geq\frac{\kappa_1}{2}D_\phi(\beta^*,\beta)$.

Applying Lemma (2.5), it follows that $T_0=\inf\{t\in\mathbb{R}_+|D_\phi(\beta^*,\beta(t))\leq r\}<\infty$. Moreover, for all $t\geq T_0$, $\frac{\mathrm{d}}{\mathrm{d}t}D_\phi(\beta^*,\beta(t))=-\kappa_1 D_\phi(\beta^*,\beta(t))$. Integrating with respect to time yields the result $D_\phi(\beta^*,\beta(t))\leq D_\phi(\beta^*,\beta(T_0))\exp(-\kappa_1(t-T_0))$ by Grönwall's lemma.

□

2.2.2 Local sublinear convergence speed

The local linear convergence speed was obtained by leveraging a 1-Łojasiewicz condition $D_F(\beta^*,\beta)\geq\frac{\kappa_1}{2}D_\phi(\beta^*,\beta)$, obtained by a Lipschitz-smoothness assumption on $\nabla\phi$. Since it was already established that $F(\beta(t))-F(\beta^*)\leq C_0/t$ and $\kappa_1 D_F(\beta^*,\beta)\leq F(\beta)-F(\beta^*)$, this immediately gives the sublinear convergence speed $D_\phi(\beta^*,\beta(t))\leq C_0/(\kappa_1 t)$. However the constant κ_1 originates from a very strong 1-Łojasiewicz condition. In the event that this forces a choice of κ_1 too small to be informative, as will be the case in our application below, it can be useful to resort to a proof of sublinear convergence under milder assumptions, through a 2-Łojasiewicz condition, resulting in a possibly much better constant.

Theorem 2.7 (Implicit bias, sublinear convergence)**.** *Assume that there exists a norm $\|\cdot\|_*$ on $\mathbb{R}^n$, and a radius $r>0$ and a constant $C_\phi\geq 0$ such that $\beta^*\in\mathcal{D}$, and on the set $B_{\beta^*}(r):=\{\beta\in\mathcal{D}\ :\ D_\phi(\beta^*,\beta)\leq r,\ \nabla\phi(\beta)\in\nabla\phi(\beta^*)+\mathrm{Im}(X^\top)\}$, it holds $\|\nabla\phi(\beta)-\nabla\phi(\beta^*)\|_*\leq C_\phi$. Then $T_0=\inf\{t\in\mathbb{R}_+|D_\phi(\beta^*,\beta(t))\leq r\}<\infty$. Moreover, for all $t\geq T_0$, it holds*

$$D_\phi(\beta^*,\beta(t))\leq\frac{1}{\kappa_2(t-T_0)+c}$$

With $\kappa_2=\left(C_\phi^2\left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2\right)^{-1}$ and $c=D_\phi(\beta^*,\beta(T_0))^{-1}$, where $\left\|X^{\top\dagger}\right\|_{\mathrm{op}}=\sup_{\|v\|_*\leq 1}\left\|(X^\top)^\dagger v\right\|_2$ is the operator norm of the pseudo-inverse of $X^\top$.

Proof. Let $\kappa_2=\left(C_\phi^2\left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2\right)^{-1}$. The idea for the proof is the same as the previous, we start by showing a 2-Łojasiewicz condition $D_F(\beta^*,\beta)\geq\frac{\kappa_2}{2}D_\phi(\beta^*,\beta)^2$, then apply Lemma (2.5) and integrate the obtained inequality.

Let $\beta\in B_{\beta^*}(r)$. By definition of $B_{\beta^*}(r)$, there exists $u\in\mathbb{R}^n$ such that $\nabla\phi(\beta)-\nabla\phi(\beta^*)=X^\top u$. Then leveraging as previously the Cauchy-Schwartz inequality (a), projection smoothness (b), and dual boundedness (c), we obtain

$$\begin{aligned}\langle\nabla\phi(\beta)-\nabla\phi(\beta^*),\beta-\beta^*\rangle^2&=\langle X^\top u,\beta-\beta^*\rangle^2=\langle X\beta-X\beta^*,u\rangle^2\\&\leq\|X\beta-X\beta^*\|_2^2\|u\|_2^2 &&(a)\\&\leq\left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2\|X\beta-X\beta^*\|_2^2\|X^\top u\|_*^2 &&(b)\\&=\left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2\|X\beta-X\beta^*\|_2^2\|\nabla\phi(\beta)-\nabla\phi(\beta^*)\|_*^2\\&\leq C_\phi^2\left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2\|X\beta-X\beta^*\|_2^2 &&(c)\end{aligned}$$

By definition of the Bregman divergence $D_\phi(\beta^*,\beta)+D_\phi(\beta,\beta^*)=\langle\nabla\phi(\beta)-\nabla\phi(\beta^*),\beta-\beta^*\rangle$. Hence by positivity of the Bregman divergence $D_\phi(\beta,\beta^*)\geq 0$, the previous inequality implies

$$D_\phi(\beta^*,\beta)^2\leq\langle\nabla\phi(\beta)-\nabla\phi(\beta^*),\beta-\beta^*\rangle^2\leq C_\phi^2\left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2\|X\beta-X\beta^*\|_2^2=\frac{2}{\kappa_2}D_F(\beta^*,\beta)$$

Thus the previously announced 2-Łojasiewicz condition $D_F(\beta^*,\beta) \geq \frac{\kappa_2}{2} D_\phi(\beta^*,\beta)^2$. Applying Lemma (2.5), it follows that $T_0 = \inf\{t \in \mathbb{R}_+ | D_\phi(\beta^*,\beta(t)) \leq r\} < \infty$. Moreover, for all $t \geq T_0$, $\frac{\mathrm{d}}{\mathrm{d}t} D_\phi(\beta^*,\beta(t)) = -\kappa_2 D_\phi(\beta^*,\beta(t))^2$. Integrating with respect to time yields the result $D_\phi(\beta^*,\beta(t)) \leq \left(\kappa_2(t-T_0) + D_\phi(\beta^*,\beta(T_0))^{-1}\right)^{-1}$. □

3 Application: two-layer network for least-squares

We consider a regression problem with training set $(x_i, y_i^*)_{i\in[n]}$ of n pairs of observations with $x_i \in \mathbb{R}^k$ and $y_i^* \in \mathbb{R}$. The prediction functions we are interested in are two-layer neural networks with ReLU non-linearities $(x \mapsto (x)_+ = \max(0,x))$ with $m \in \mathbb{N} \setminus \{0\}$ hidden nodes. We assume that the directions of the first layer $(\theta_i)_{i\in[n]}$ are fixed to $\theta_i \in \mathbb{S}^{k-1} = \{u \in \mathbb{R}^k \,|\, \|u\|_2 = 1\}$. As such, the predictions of a network with trainable parameters $(a,b) \in \mathbb{R}^m \times \mathbb{R}^m$ is obtained with $N^\theta : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^n$ defined as

$$N_i^\theta : (a,b) \mapsto \sum_{j\in[m]} b_j \left(\langle a_j \theta_j, x_i\rangle\right)_+$$

We use $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ to denote the square loss $y \mapsto \frac{1}{2}\|y - y^*\|_2^2$, and consider minimization using a gradient flow on the trainable parameters $w : \mathbb{R}_+ \to \mathbb{R}^{2m}$.

$$\frac{\mathrm{d}w}{\mathrm{d}t}(t) = -\nabla(\mathcal{L} \circ N^\theta)(w(t))$$

Under the assumption that for all $i \in [m]$, $a_i(0)^2 \geq b_i(0)^2$, this gradient flow is well defined at all times, see Lemma (A.2) in appendix.

3.1 Reparameterization to mirror flow

We start by showing that there exists a reparameterization $\psi : \mathbb{R}^m \times \mathbb{R}^m \to \mathcal{D}$ with $\mathcal{D} \subseteq \mathbb{R}^d$ non-empty open convex, a Legendre potential $\phi : \mathcal{D} \to \mathbb{R}$ with $\nabla\phi(\mathcal{D}) = \mathbb{R}^d$, and a design matrix $X \in \mathbb{R}^{n\times d}$, such that $N^\theta(a,b) = X \cdot \psi(a,b)$, and $\psi(a,b)$ follows a mirror flow with potential ϕ. From there, applying the results of the previous section will yield a caracterization of the limit point obtained by gradient flow, together with bounds on the convergence speed.

Definition 3.1. *For $(\alpha, z) \in (\mathbb{R}_+ \times \mathbb{R}) \setminus \{(0,0)\}$, define $\mathcal{D}_{\alpha,z} := \mathbb{R}$ if $\alpha \neq 0$, or $\mathcal{D}_{0,z} := \mathbb{R}_+^*$ if $z > 0$, or $\mathcal{D}_{0,z} := \mathbb{R}_-^*$ if $z < 0$. The one-dimensional α-hypentropy compatible with z is the function $\phi_\alpha : \mathcal{D}_{\alpha,z} \to \mathbb{R}$, defined for $\alpha \neq 0$ as*

$$\phi_\alpha : x \mapsto x \operatorname{arcsinh}\left(\frac{x}{\alpha}\right) - \sqrt{x^2 + \alpha^2} + \alpha$$

And extended to $\alpha = 0$ as

$$\phi_0 : x \mapsto |x| \log(|x|) - |x| + 1$$

Definition 3.2. *For $(\alpha, z) \in \mathbb{R}_+^k \times \mathbb{R}^k$ such that $\forall i \in [k]$, $(\alpha_i, z_i) \neq (0,0)$, define by cartesian product the domain $\mathcal{D}_{\alpha,z} := \prod_{i\in[k]} \mathcal{D}_{\alpha_i,z_i} \subseteq \mathbb{R}^k$. The k-dimensional α-hypentropy compatible with z is the function $\Phi_\alpha : \mathcal{D}_{\alpha,z} \to \mathbb{R}$ defined as $\Phi_\alpha : x \mapsto \sum_{i\in[k]} \phi_{\alpha_i}(x_i)$.*

Although perhaps not directly obvious from the definition, the multi-dimensional hypentropy of Definition (3.2) can take crucially different shapes depending on the parameter $\alpha \in \mathbb{R}_+^k$. Some examples in dimension one and two are depicted in Figure (1). For α strictly positive, $\phi_\alpha(0) = 0$ and $\nabla\phi_\alpha(0) = 0$. Therefore, in the setting of Theorem (2.2), if $\beta_0 = 0$, then the limit point β^* is the minimizer of $D_{\Phi_\alpha}(\cdot, \beta_0) = \Phi_\alpha(\cdot)$ among minimizers of the objective. The classifier learned by gradient descent and its properties will vary greatly depending on this parameter. As observed by [Woodworth et al., 2020, p6], a large such parameter will yield an ℓ_2-like implicit regularization, while a choice closer to zero will behave more like an ℓ_1-like regularization.

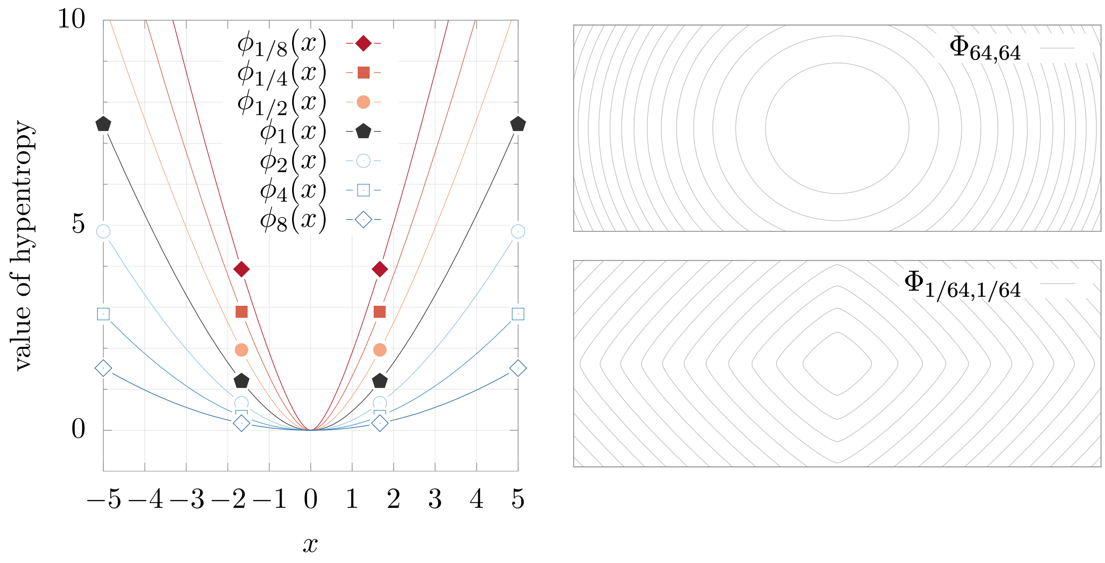

Figure 1: Comparison of 1D hypentropies (left), and level sets of 2D hypentropies (right)

We now show that the so-defined hypentropy is a valid choice of potential for mirror flow.

Lemma 3.3. *For any $(\alpha, z) \in \mathbb{R}_+^k \times \mathbb{R}^k$ such that $\forall i \in [k]$, $(\alpha_i, z_i) \neq (0,0)$, the domain $\mathcal{D}_{\alpha,z} \subseteq \mathbb{R}^k$ is a convex open subset of $\mathbb{R}^k$ containing z, and $\Phi_\alpha : \mathcal{D}_{\alpha,z} \to \mathbb{R}$ the α-hypentropy compatible with z is of Legendre type, and satisfies $\nabla\Phi_\alpha(\mathcal{D}_{\alpha,z}) = \mathbb{R}^k$.*

Proof. For all $i \in [k]$, by hypothesis $(\alpha_i, z_i) \neq (0,0)$, thus $\mathcal{D}_{\alpha_i,z_i} \in \{\mathbb{R}, \mathbb{R}_+^*, \mathbb{R}_-^*\}$, hence $\mathcal{D}_{\alpha_i,z_i}$ is a convex open subset of $\mathbb{R}$. Moreover $z_i \in \mathcal{D}_{\alpha_i,z_i}$ by definition. Then $\mathcal{D}_{\alpha,z}$ is a convex set as a cartesian product of convex sets by Rockafellar [1970, Thm. 3.5], and $z \in \mathcal{D}_{\alpha,z}$. For every $i \in [k]$, the one-dimensional α_i-hypentropy is twice differentiable on $\mathcal{D}_{\alpha_i,z_i}$ and $\frac{\mathrm{d}^2\phi_{\alpha_i}}{\mathrm{d}x^2}(x) = (x^2 + \alpha_i^2)^{-1/2} > 0$ (see Lemma (A.3) in appendix). Thus all $\phi_{\alpha_i} : \mathcal{D}_{\alpha_i,z_i} \to \mathbb{R}$ are strictly convex on their domain. Furthermore, they are all essentially smooth, since $\partial\mathcal{D}_{\alpha_i,z_i} = \emptyset$ if $\alpha_i \neq 0$, $\partial\mathcal{D}_{0,z_i} = \{0\}$ and $\langle \nabla\phi_0(tx), x\rangle = |x|\log(t|x|) \to -\infty$ when $t \to 0$ for any $x \in \mathcal{D}_{0,z_i}$ (see [Bauschke and Borwein, 1997, Def 2.1]). Thus all ϕ_i are Legendre (see [Bauschke and Borwein, 1997, Def 2.8]). Then $\Phi_\alpha : \prod_{i\in[k]} \mathcal{D}_{\alpha_i,z_i} \to \mathbb{R}$ is Legendre, for it is strictly convex differentiable by separability, and $\langle \nabla\Phi_\alpha(x + t(y-x)), y - x\rangle \to -\infty$ when $t \to 0$ for any $(x, y) \in \partial\mathcal{D}_{\alpha,z} \times \mathcal{D}_{\alpha,z}$ by separating each dimension and applying the previous argument. For the last claim, observe that since each $\phi_{\alpha_i} : \mathcal{D}_{\alpha_i,z_i} \to \mathbb{R}$ is Legendre, it holds $\nabla\phi_{\alpha_i}(\mathcal{D}_{\alpha_i,z_i}) = \mathbb{R}$. Thus by separability, $\nabla\Phi_\alpha(\mathcal{D}_{\alpha,z}) = (\nabla\phi_{\alpha_i,z_i}(\mathcal{D}_{\alpha_i,z_i}))_{i\in[k]} = \mathbb{R}^k$. □

Theorem 3.4. *Let $(x_i, y_i^*)_{i\in[n]} \in (\mathbb{R}^d \times \mathbb{R})^n$, and $\theta_j \in \mathbb{S}^{d-1}$ for $j \in [m]$. Let $(a_0, b_0) \in (\mathbb{R}_+^*)^m \times \mathbb{R}^m$ be such that $r = \frac{1}{2}\left(a_0^2 - b_0^2\right) \in \mathbb{R}_+^m$. Let $(a, b) : \mathbb{R}_+ \to \mathbb{R}^m \times \mathbb{R}^m$ be the weights of the two-layer fixed-direction network trained by gradient flow on the least-squares objective with data $(x_i, y_i^*)_{i\in[n]}$ and initialization (a_0, b_0). Then $\beta : \mathbb{R}_+ \to \mathbb{R}^m$, defined by $\beta_j : t \mapsto a_j(t)b_j(t)$ follows a mirror flow with potential $\frac{1}{2}\Phi_r$, where Φ_r is the r-hypentropy compatible with $\beta(0)$, and objective $\beta \mapsto \frac{1}{2}\|X\beta - y^*\|_2^2$, where $X_{i,j} = (\langle \theta_j, x_i\rangle)_+$. Moreover, $\forall t \in \mathbb{R}_+, \forall x \in \mathbb{R}^d, \forall j \in [m]$, $b_j(t)(\langle a_j(t)\theta_j, x\rangle)_+ = \beta_j(t)(\langle \theta_j, x\rangle)_+$.*

The idea for this proof is that since the objective depends only on the product $a_j b_j$ and not on a_j and b_j individually, then at all times and for all $j \in [m]$, training by gradient flow will preserve the quantity $a_j^2 - b_j^2$. The pair (a_j, b_j) thus remains constrained on a hyperbole[5], and can be safely reparameterized to a single parameter β_j. Similarly, the hypentropy potential appears in [Woodworth et al., 2020, Theorem 1] because their objective depends only on a difference of squares, therefore the product is conserved, and the change of coordinates $(a, b) \mapsto (a^2 - b^2, ab)$ is identical. Their reparameterization is not recovered directy as a particular case of this theorem, for it merges two nodes together, but it will be recovered as a particular case of the following lemma which extends this theorem.

Proof. Let $r : t \mapsto \frac{1}{2}\left(a(t)^2 - b(t)^2\right)$. By Lemma (A.2), the condition $\frac{1}{2}\left(a_0^2 - b_0^2\right) \in \mathbb{R}_+^m$ guarantees that the gradient flow is well defined, that r is constant, and that $\forall t \in \mathbb{R}_+, \forall j \in [m], a_j(t) \geq 0$. It follows that for any $t \in \mathbb{R}_+$, $x \in \mathbb{R}^d$, and $j \in [m]$, it holds $b_j(t)(\langle a_j(t)\theta_j, x\rangle)_+ = b_j(t)a_j(t)(\langle \theta_j, x\rangle)_+ = \beta_j(t)(\langle \theta_j, x\rangle)_+$, where we let $\beta : t \mapsto (a_j(t)b_j(t))_{j\in[m]}$. This proves the last claim of equivalence between the two-layer network in (a, b) and the linear model in β. It remains to show that β follows a mirror flow for objective $F : \beta \mapsto \frac{1}{2}\|X\beta - y^*\|_2^2$.

Let us show that β satisfies the differential equation $\frac{\mathrm{d}}{\mathrm{d}t}\beta = -\left(\frac{1}{2}\nabla^2\Phi_r(\beta)\right)^{-1} \cdot \nabla F(\beta)$. Using the definition of gradient flow for (a, b), then the chain rule with the intermediate variable $y_i : t \mapsto \sum_{j\in[m]} \beta_j(t)\ (\langle \theta_j, x_i\rangle)_+$, and definition of the objective F,

$$\begin{aligned}
\frac{\mathrm{d}\beta_j}{\mathrm{d}t}(t) &= \frac{\mathrm{d}a_j}{\mathrm{d}t}(t)\ b_j(t) + a_j(t)\ \frac{\mathrm{d}b_j}{\mathrm{d}t}(t) \\
&= -\frac{\partial \mathcal{L} \circ N^\theta}{\partial a_j}\Big(a(t), b(t)\Big)\ b_j(t) - a_j(t)\ \frac{\partial \mathcal{L} \circ N^\theta}{\partial b_j}\Big(a(t), b(t)\Big) \\
&= -\left(a_j(t)^2 + b_j(t)^2\right) \sum_{i\in[n]} (\langle \theta_j, x_i\rangle)_+ \frac{\partial \mathcal{L}}{\partial y_i}(y_i(t)) \\
&= -\left(a_j(t)^2 + b_j(t)^2\right) \Big[\nabla F(\beta(t))\Big]_j
\end{aligned}$$

Using the identity $(u + v) = \sqrt{4uv + (u - v)^2}$, which holds for all $(u, v) \in \mathbb{R}_+^2$, and since the Hessian $\nabla^2\Phi_r$ is diagonal by separability of Φ_r, it follows

$$\begin{aligned}
\frac{\mathrm{d}\beta_j}{\mathrm{d}t}(t) &= -\left(\sqrt{4\beta_j(t)^2 + \left(a_j(t)^2 - b_j(t)^2\right)^2}\right) \Big[\nabla F(\beta(t))\Big]_j \\
&= -\left(2\sqrt{\beta_j(t)^2 + {r_j}^2}\right) \Big[\nabla F(\beta(t))\Big]_j \\
&= -\left[\frac{1}{2}\nabla^2\Phi_r(\beta(t))\right]_{j,j}^{-1} \Big[\nabla F(\beta(t))\Big]_j \\
&= -\left[\left(\frac{1}{2}\nabla^2\Phi_r(\beta(t))\right)^{-1} \nabla F(\beta(t))\right]_j
\end{aligned}$$

Hence β satisfies the claimed differential equation $\frac{\mathrm{d}}{\mathrm{d}t}\beta = -\left(\frac{1}{2}\nabla^2\Phi_r(\beta)\right)^{-1} \cdot \nabla F(\beta)$, which by chain rule is equivalent to $\frac{\mathrm{d}}{\mathrm{d}t}\nabla\left(\frac{1}{2}\Phi_r\right)(\beta) = -\nabla F(\beta)$, thus β follows a mirror flow with objective F and potential $\frac{1}{2}\Phi_r$, which concludes the proof. □

[5] unless the difference of squares is null, in which case it is constrained to a open half-line

Lemma 3.5 (reparameterization, merging identical directions)**.** *Let* $(x_i, y_i^*) \in (\mathbb{R}^d \times \mathbb{R})$ *for* $i \in [n]$ *be the training data, and* $\theta_j \in \mathbb{S}^{d-1}$ *for* $j \in [m]$ *the fixed directions. Let* $I : [m] \to [p]$ *be a surjective map such that* $\forall (i,j) \in [m]^2$, $I(i) = I(j) \Rightarrow \theta_i = \theta_j$. *Let* $(a_0, b_0) \in (\mathbb{R}_+^*)^m \times \mathbb{R}^m$ *be such that* $(a_0^2 - b_0^2) \in \mathbb{R}_+^m$. *Let* $r \in \mathbb{R}_+^p$ *defined for* $k \in [p]$ *by* $r_k^2 = \frac{1}{4} \sum_{i,j \in I^{-1}(k)} \Big([a_0]_i[a_0]_j - [b_0]_i[b_0]_j\Big)^2 + \Big([a_0]_i[b_0]_j - [b_0]_i[a_0]_j\Big)^2$. *Let* $(a, b) : \mathbb{R}_+ \to \mathbb{R}^m \times \mathbb{R}^m$ *be the weights of the two-layer fixed-direction network trained by gradient flow on the least-squares objective with data* $(x_i, y_i^*)_{i \in [n]}$ *and initialization* (a_0, b_0). *Then* $\beta : \mathbb{R}_+ \to \mathbb{R}^p$ *defined by* $\beta_k : t \mapsto \sum_{j \in I^{-1}(k)} a_j(t) b_j(t)$ *follows a mirror flow with potential* $\frac{1}{2}\Phi_r$, *where* Φ_r *is the* r*-hypentropy compatible with* $\beta(0)$, *and objective* $\beta \mapsto \frac{1}{2}\|X\beta - y^*\|_2^2$, *for* $X_{k,i} = (\langle \theta_{I^{-1}(k)}, x_i \rangle)_+$. *Moreover,* $\forall t \in \mathbb{R}_+, \forall x \in \mathbb{R}^d, \forall k \in [p]$, $\sum_{j \in I^{-1}(k)} b_j(t)(\langle a_j(t)\theta_j, x \rangle)_+ = \beta_k(t)(\langle \theta_{I^{-1}(k)}, x \rangle)_+$.

Proof. We start by noting that X is well defined, despite the slight abuse of notation, by the assumption on I, and that $I = \mathrm{id} : [m] \to [m]$ the identity recovers the previous theorem. The proof is identical, so we only show the change in the computation of the differential equation.

$$\begin{aligned}
\frac{\mathrm{d}\beta_k}{\mathrm{d}t} &= - \sum_{j \in I^{-1}(k)} \Big(a_j(t)^2 + b_j(t)^2\Big) \sum_{i \in [n]} (\langle \theta_j, x_i \rangle)_+ \frac{\partial \mathcal{L}}{\partial y_i}(y_i(t)) \\
&= - \sum_{j \in I^{-1}(k)} \Big(a_j(t)^2 + b_j(t)^2\Big) \Big[\nabla F(\beta(t))\Big]_k
\end{aligned}$$

To shorten notations a little, for $i \in [m]$ let $q_i : t \mapsto a_i(t)^2 + b_i(t)^2$, and $\alpha_i : t \mapsto a_i(t)b_i(t)$. Let $k \in [p]$ and perform summations over $I^{-1}(k)$ by default when ommitted, such that $\beta_k = \sum_j \alpha_j$. In this form, the previous equality can be written $\frac{\mathrm{d}\beta_k}{\mathrm{d}t} = -(\sum_j q_j)[\nabla F(\beta)]_k$, and the potential's inverse hessian is $\left[\frac{1}{2}\nabla^2\Phi(r)(\beta)\right]_{k,k}^{-1} = 2\sqrt{\beta_k^2 + r_k^2}$. To show that β follows a mirror flow with potential $\frac{1}{2}\Phi_r$, it is thus sufficient to show that $\sum_j q_j = \left[\frac{1}{2}\nabla^2\Phi(r)(\beta)\right]_{k,k}^{-1}$, or equivalently (by positivity) that their squares are equal. Since $(\sum_j q_j)^2 - 4\beta_k^2 = (\sum_j q_j)^2 - 4(\sum_j \alpha_j)^2 = \sum_{i,j} q_i q_j - 4\alpha_i\alpha_j$, it is sufficient to show that $4r_k^2 = \sum_{i,j} q_i q_j - 4\alpha_i\alpha_j$. By simply expanding each term of both sums, $q_i q_j - 4\alpha_i\alpha_j = (a_i^2 + b_i^2)(a_j^2 + b_j^2) - 4a_ia_jb_ib_j = (a_i^2a_j^2 - 2a_ia_jb_ib_j + b_i^2b_j^2) + (a_i^2b_j^2 - 2a_ib_ja_jb_i + a_j^2b_i^2) = (a_ia_j - b_ib_j)^2 + (a_ib_j - a_jb_i)^2$, which concludes the proof. We could also show as before that each term $q_iq_j - 4\alpha_i\alpha_j$ is constant. □

3.2 Analysis of reparameterized mirror flow

In the event that the hypentropy parameter is sufficiently far from zero, a global linear convergence rate of the Bregman divergence to the optimum can be guaranteed.

Lemma 3.6. *Let* $X \in \mathbb{R}^{n \times m}$ *be a design matrix of rank* n, $y \in \mathbb{R}^n$ *a response vector,* $\beta_0 \in \mathbb{R}^m$ *an initial point, and* $r \in \mathbb{R}_+^m$ *such that* $\inf_i r_i > 0$. *If* $\beta : \mathbb{R}_+ \to (\mathbb{R}^m, \|\cdot\|_2)$ *follows a mirror flow with objective* $\beta \mapsto \frac{1}{2}\|X\beta - y\|_2^2$ *and potential* $\frac{1}{2}\Phi_r$ *the* r*-hypentropy compatible with* β_0, *from* $\beta(0) = \beta_0$, *then* $\beta(t) \to \beta^* = \arg\min\{D_{\Phi_r}(\beta, \beta_0) : \beta \in \arg\min F\}$. *Moreover for all* $t \in \mathbb{R}_+$, *it holds* $D_{\Phi_r}(\beta^*, \beta(t)) \leq D_{\Phi_r}(\beta^*, \beta_0)e^{-\kappa t}$, *where* $\kappa = (\inf_i r_i)/\left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2$.

Proof. Let $R = \inf_{i \in [m]} r_i$. Let us show that $\nabla\Phi_r : (\mathbb{R}^m, \|\cdot\|_2) \to (\mathbb{R}^m, \|\cdot\|_2)$ is $(1/R)$-Lipschitz.

$$\sup_{(x,y) \in \mathbb{R}^m \times \mathbb{R}^m} \frac{\|\nabla\Phi_r(x) - \nabla\Phi_r(y)\|_2}{\|x - y\|_2} \leq \sup_{x \in \mathbb{R}^m} \sup_{i \in [m]} \Big[\nabla^2\Phi_r(x)\Big]_{i,i} = \sup_{x \in \mathbb{R}^m} \sup_{i \in [m]} \sqrt{\frac{1}{x^2 + r_i^2}} \leq \frac{1}{R}$$

The result follows by Theorem (2.6) for $T_0 = 0$. □

Similarly if the limit β^* is sufficiently far from zero, then a local linear convergence rate would be obtained in the same manner by upper-bounding the highest eigenvalue of $\nabla^2\Phi_r$ near β^*. However in the event that the optimum is sparse, and the hypentropy parameter r is close to zero, then these bounds would quickly become uninformative since the constant κ would tend to zero like $(\inf_{i\in[m]} r_i)$. While it remains that $F(\beta(t)) \leq C_0/t$, we have no such guarantee for the Bregman divergence, which merely satisfies $D_\phi(\beta^*, \beta(t)) \leq C_0/(\kappa t)$ for a vanishing κ. Since that sparsity-inducing regime appears to be interesting, we turn to a 2-Łojasiewicz condition to attempt to salvage the sublinear convergence rate's constant.

Lemma 3.7. *Let $X \in \mathbb{R}^{n\times m}$ be a design matrix of rank n, $y \in \mathbb{R}^n$ a response vector, $\beta_0 \in \mathbb{R}^m$ an initial point, and $r \in \mathbb{R}_+^m$ such that $\inf_i r_i > 0$. If $\beta : \mathbb{R}_+ \to (\mathbb{R}^m, \|\cdot\|_2)$ follows a mirror flow with objective $\beta \mapsto \frac{1}{2}\|X\beta - y\|_2^2$ and potential $\frac{1}{2}\Phi_r$ the r-hypentropy compatible with β_0, from $\beta(0) = \beta_0$, then $\beta(t) \to \beta^* = \arg\min\{D_{\Phi_r}(\beta, \beta_0) : \beta \in \arg\min F\}$. Moreover if for all $t \in \mathbb{R}_+$, $\|\beta(t)\|_1 \leq M \in \mathbb{R}_+$, then for all $t \in \mathbb{R}_+$, it holds $D_{\Phi_r}(\beta^*, \beta(t)) \leq 1/(\kappa t + c)$, where $\kappa = 1/(C^2 \left\|X^{\top\dagger}\right\|_{\text{op}}^2)$, and $C = 2m\log\left(1 + \frac{2M}{m(\inf_i r_i)}\right)$ and $c = 1/D_{\Phi_r}(\beta^*, \beta_0)$.*

Proof. Let $R = \inf_{i\in[m]} r_i$. Let us show that for all $t \in \mathbb{R}_+$, $\|\nabla\Phi_r(\beta(t)) - \nabla\Phi_r(\beta^*)\|_2 \leq C$. First, observe that for $x \in \mathbb{R}^m$, it holds in each dimension

$$|\nabla\phi_{r_i}(x_i)| = \operatorname{arcsinh}\left(\left|\frac{x_i}{r_i}\right|\right) = \log\left(\left|\frac{x_i}{r_i}\right| + \sqrt{\left(\frac{x_i}{r_i}\right)^2 + 1}\right) \leq \log\left(1 + 2\left|\frac{x_i}{r_i}\right|\right)$$

Thus with the additional assumption $\|x\|_1 \leq M$, and arithmetic-geometric inequality

$$\|\nabla\Phi_r(x)\|_1 = \sum_{i\in[m]} |\nabla\phi_{r_i}(x_i)| \leq \log\left(\prod_{i\in[m]}\left(1 + 2\left|\frac{x_i}{R}\right|\right)\right) \leq m\log\left(\frac{1}{m}\sum_{i\in[m]} 1 + 2\left|\frac{x_i}{R}\right|\right)$$
$$= m\log\left(1 + \frac{2}{mR}\sum_{i\in[m]}|x_i|\right) \leq m\log\left(1 + \frac{2M}{mR}\right) = \frac{C}{2}$$

Then since $\|\beta(t)\|_1 \leq M$ for all t, and thus $\|\beta^*\|_1 \leq M$ by continuity, it follows

$$\|\nabla\Phi_r(\beta(t)) - \nabla\Phi_r(\beta^*)\|_2 \leq \|\nabla\Phi_r(\beta(t))\|_2 + \|\nabla\Phi_r(\beta^*)\|_2 \leq \|\nabla\Phi_r(\beta(t))\|_1 + \|\nabla\Phi_r(\beta^*)\|_1 \leq C$$

The result follows by Theorem (2.7) with $T_0 = 0$. □

Note that although this results may be interesting for a fixed number of neurons m and vanishing constant R, the constant C quickly degrades with m, as $2m\log\left(1 + \frac{2M}{mR}\right) \underset{m\to\infty}{\longrightarrow} \frac{4M}{R}$.

4 Numerical experiments

Unfortunately, neither of these bounds appear to accurately capture the behavior observed in practice. Still, to get a sense of the influence of some hyperparameters, we perform experiments on randomly-generated data. We generate $n \in \mathbb{N}$ points uniformly at random on the d-dimensional hypercube of side 2, with labels taken uniformly at random in $\{\pm 1\}$. We then initialize the weights in $\mathbb{R}^{d\times m} \times \mathbb{R}^m$ independently uniformly at random on the segment $[-\alpha/\sqrt{m}, +\alpha/\sqrt{m}]$, for some parameter $\alpha \in \mathbb{R}_+$, and split the directions θ and trainable parameters $w \in \mathbb{R}^m \times \mathbb{R}^m$. The scaling by $\sqrt{m}$ is chosen such that the response remains approximately constant even when the number of neurons tends to infinity, and decreasing the parameter α will allow us to artifically let the hypentropy parameter r tend to zero.

We resample if the initial weights do not satisfy our condition for a well-defined flow, then train the network by gradient descent $w_{k+1} = w_k - \mu \cdot \nabla(\mathcal{L} \circ N^\theta)(w_k)$. Our bounds being constrained to the continuous time domain, we use small step sizes ($\mu = 10^{-4}$) and the surrogate $t_k = \mu \times k$ for the time variable at iteration k. The training is continued until the objective reaches zero (up to machine precision) and the corresponding value of β is used as approximation of the limit point β^* for the computation of the Bregman divergence to optimum. The resulting figures serve only as vague illustration of the result since the proofs do not extend as-is to discrete time steps.

Figure (2) depicts the evolution of the Bregman divergence $D_\phi(\beta^*, \beta(t))$ over time, for $n = 20$ samples in dimension $d = 50$ and $m = 50$ neurons with scaling $\alpha = 1$. This is by far the most representative situation we encounter in experiments, very fast convergence of the network, with a linear convergence phase near the optimum. The linear bound's rate ($\kappa_1 \approx 6.2 \times 10^{-2}$) does not quite match the observed rate ($\kappa \approx 3 \times 10^{-1}$, measured on the figure), which results in a large gap with the observed behavior, and the sublinear bound's constant ($\kappa_2 \approx 1.7 \times 10^{-5}$) is too small to show visible improvements before the network has converged. In this example, $(\inf_i r_i) \approx 1.3 \times 10^{-1}$ and $\left\|X^{\top\dagger}\right\|_{\mathrm{op}} \approx 1.5$ are computed at initialization, and $M \approx 1.4 \times 10^1$ is measured after training.

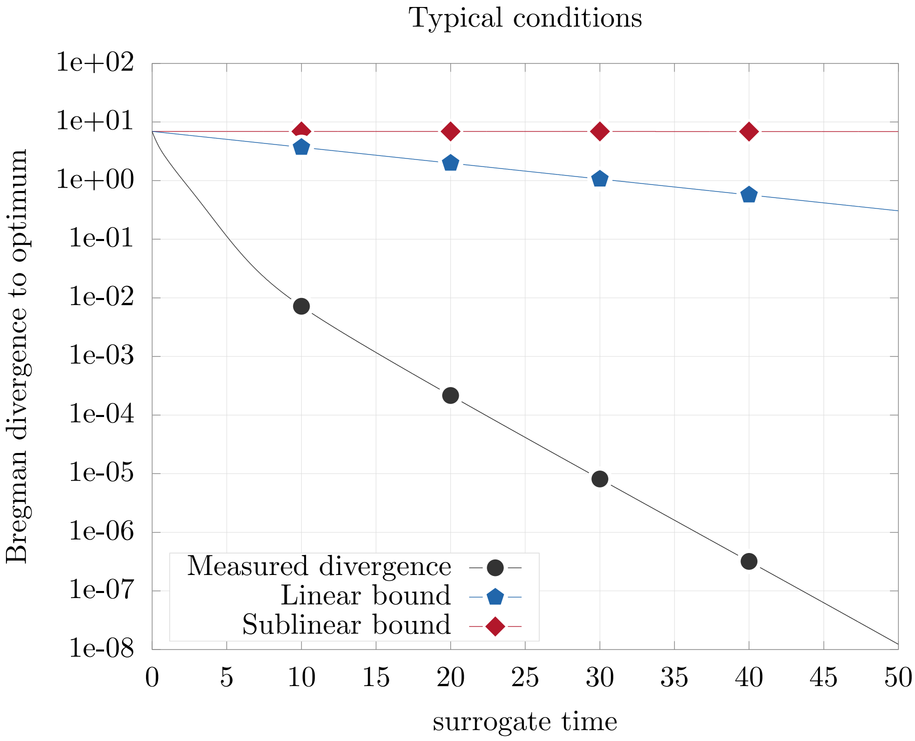

Figure 2: Convergence speed measurements $[\, n = 20, d = 50, m = 50, \alpha = 1\,]$

Given the expressions for the constants κ_1 and κ_2, we can craft a setting in which we expect the sublinear bound to become more informative. Setting $d = 100$ to lower the variance of first layer's norms, and choosing few neurons ($m = 20$), with $n = 10$ samples and a scaling of $\alpha = 10^{-4}$, we get the results depicted in Figure (3). As expected, the linear bound's rate ($\kappa_1 \approx 1.8 \times 10^{-9}$) decreases so much it becomes completely uninformative since $(\inf_i r_i) \approx 6.9 \times 10^{-9}$, while the sublinear bound's constant ($\kappa_2 \approx 4.7 \times 10^{-7}$) is only affected logarithmically and retains some use. In this example, $\left\|X^{\top\dagger}\right\|_{\mathrm{op}} \approx 1.9$ and $M \approx 6.2$. Although the convergence is again very fast once a neighborhood of the optimum is reached, what makes this regime particulary interesting is that there is a much slower (by several orders of magnitude) first phase where the sublinear bound is relatively accurate. When the hypentropy parameter $r \in \mathbb{R}_+^m$ varies, the limit point reached can be very different, and the time required to reach it varies accordingly. Here r is close to zero, leading to a more sparsity-inducing task, with slower convergence speed.

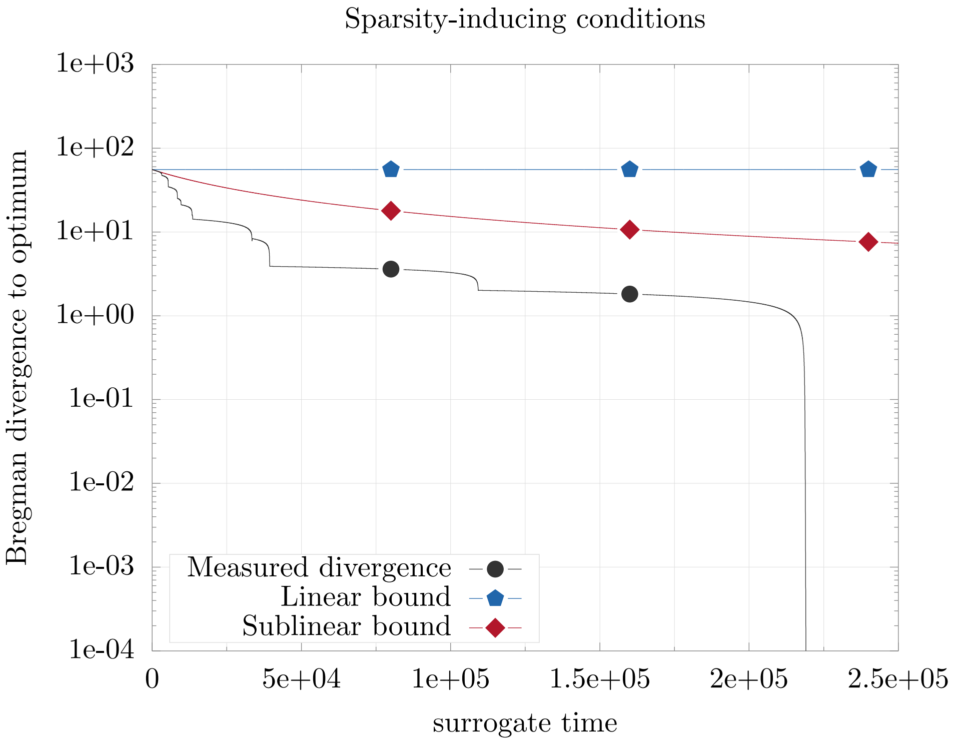

Figure 3: Convergence speed measurements $[\, n = 10, d = 100, m = 20, \alpha = 10^{-4} \,]$

An interesting surprise when performing these experiments was the accidental consequences of altering hyperparameter choices. Our analysis revolved around the split between the linear predictor $\beta \in \mathbb{R}^m$ and the hypentropy parameter $r \in \mathbb{R}_+^m$, with different inductive biases on the linear predictor characterized by different choices of hypentropy parameters. With a good prior on the desired linear predictor, one could choose the appropriate parameter to align the inductive bias with the prior. But that is not how the choice of initialization is typically made. Instead, weights are initialized in a different parameterization ($\mathbb{R}^{d \times m} \times \mathbb{R}^m$), with very different objectives in mind, for instance the concern that some intermediate variables might blow up past what the machine precision can handle can lead to a factor $1/\sqrt{m}$ in the choice of initialization scale. In such a setting, increasing the number of neurons will have unintended consequences, such as altering the hypentropy parameter, and thus the inductive bias, which is crucial in overparameterized regimes.

5 Ablations and consistency

5.1 Convergence to the Bregman projection of initialization

We have proven in the first section that the mirror flow with least-squares objective converges to the Bregman projection of the initialization, for all Legendre potentials alike, provided the mirror flow is well defined. However, this need not be true for all convex objectives, it appears as property of the least-squares objective specifically.

Intuitively, the least-squares objective is non-biasing, in the sense that the gradient is orthogonal to all segments between minima, such that the gradient never biases the flow towards one particular minimum. This guarantees that the Bregman projection is constant. However the Bregman projection to the set of minimizers of the limit point is itself, since the limit point is a minimizer as well by convexity of the objective, thus the limit point is the Bregman projection of the initialization.

If β^* is the limit point, $\bar{\beta}$ a minimizer of F, then for a mirror flow with potential ϕ,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(D_\phi(\beta^*, \beta_t) - D_\phi(\bar{\beta}, \beta_t)\Big) = -\left\langle \nabla F(\beta_t), \beta^* - \bar{\beta} \right\rangle$$

If F is a least squares objective, then this quantity is always null and thus the projection is constant. A similar approach is used in Gunasekar et al. [2020] to generalize to linear models (with respect to the weights w) with a more complicated loss $\mathcal{L}$, that is $w \mapsto \mathcal{L}(\langle w, x \rangle)$, where the "linear model" assumption grants the same property under mild assumptions on the loss.

5.2 Necessity of a link between objective and potential

Convexity of the objective F is sufficient to obtain a bound of $F(\beta(t)) \leq C_0/t$, regardless of the potential chosen for the mirror flow. However, in general no such guarantee can be obtained for the Bregman divergence D_ϕ. The reason is that the convergence measured by F and the one measured by ϕ can be very different. We give a simple example with a different objective in dimension one below. Note that the minimum is unique in this example, to avoid the discussion of the previous section.

Lemma 5.1. *Choose the domain $\mathcal{D} = \mathbb{R}_+^*$, let $\phi : \mathcal{D} \to \mathbb{R}$ be the entropy $\phi : x \mapsto x \log x - x$, and $F : x \mapsto \frac{\mu}{n(n+1)} x^{n+1}$ for some $n \in \mathbb{N}^*$ and $\mu \in \mathbb{R}_+^*$, with (unique) optimum $\beta^* = 0 \in \partial\mathcal{D}$. The solution $\beta : \mathbb{R}_+^* \to \mathcal{D}$ to any mirror flow defined by F and ϕ satisfies for some $c \in \mathbb{R}_+^*$*

$$\begin{cases} F(\beta(t)) = \dfrac{\mu}{n(n+1)}(c + \mu\, t)^{-\frac{n+1}{n}} \\ D_\phi(\beta^*, \beta(t)) = (c + \mu\, t)^{-\frac{1}{n}} \end{cases}$$

Letting $n \to \infty$ gives an arbitrarily slow convergence, as measured by ϕ, while the convergence measured by F remains relatively similar to C_0/t.

Proof. We have already proved that ϕ is Legendre with $\nabla\phi(\mathcal{D}) = \mathbb{R}$. The objective $F : \mathbb{R} \to \mathbb{R}$ is clearly convex, therefore for any $\beta_0 \in \mathcal{D}$, the mirror flow $\beta : \mathbb{R}_+ \to \mathcal{D}$ with objective F, potential ϕ and starting from $\beta(0) = \beta_0$ is well defined, unique, and converges to $\beta^* = 0$. Moreover, we can expand the mirror flow equation $\frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(\beta) = -\nabla F(\beta)$ to obtain $\frac{\mathrm{d}}{\mathrm{d}t}\log(\beta) = -\frac{\mu}{n}\beta^n$, i.e. $\frac{\mathrm{d}}{\mathrm{d}t}\beta = -\frac{\mu}{n}\beta^{n+1}$, which yields by integration $\beta(t) = (c + \mu t)^{-1/n}$ for $c = \beta_0^{-n}$. Then, observe that for $\beta^* = 0$, it holds $D_\phi(\beta^*, \beta) = \phi(\beta^*) - \phi(\beta) - \nabla\phi(\beta)(\beta^* - \beta) = \beta$. The result follows by replacing $\beta(t)$ by its expression. □

5.3 Reduction to standard gradient flow bound

When the domain is $\mathcal{D} = \mathbb{R}^m$ and $\phi : \beta \mapsto \frac{1}{2}\|\beta\|_2^2$, the mirror flow with potential ϕ is a gradient flow. If the objective F is μ-strongly convex, i.e. if $D_F(\beta_0, \beta_1) = \frac{1}{2}\|X\beta_0 - X\beta_1\|_2^2 \geq \frac{\mu}{2}\|\beta_0 - \beta_1\|_2^2$, then choosing the ℓ_2 norm as reference, $\left\|X^{\top\dagger}\right\|_{\mathrm{op}}^2 = \sup_{\|v\|_2 \leq 1}\|(X^\top)^\dagger v\|_2^2 = \sup_{\|X^\top u\|_2 \leq 1}\|u\|_2^2 = \frac{1}{\mu}$, and $\nabla\phi$ is 1-Lipschitz with respect to ℓ_2. The bound obtained by Theorem (2.6) in this case is $D_\phi(\beta^*, \beta(t)) = \frac{1}{2}\|\beta^* - \beta(t)\|_2^2 \leq \frac{1}{2}\|\beta^* - \beta_0\|_2^2 \exp(-\mu t)$, which corresponds to $\|\beta^* - \beta(t)\|_2 \leq \|\beta^* - \beta_0\|_2 \exp(-\frac{\mu}{2}t)$. In contrast, the standard bound obtained in this setting by analyzing the gradient flow is $\|\beta^* - \beta(t)\|_2 \leq \|\beta^* - \beta_0\|_2 \exp(-\mu t)$. Our bound is off only by a factor two in the constant, which is expected since our proof uses $D_\phi(\beta^*, \beta) \leq D_\phi(\beta^*, \beta) + D_\phi(\beta, \beta^*) = \langle \nabla\phi(\beta^*) - \nabla\phi(\beta), \beta^* - \beta \rangle$ and proceeds to bound the latter. If the Bregman divergence D_ϕ is symmetric, as is the case here, then it holds $2D_\phi(\beta^*, \beta) = \langle \nabla\phi(\beta^*) - \nabla(\beta), \beta^* - \beta \rangle$, which can be used to gain the missing factor two. Note that while we can't improve our bound by this method, for we need it to cover the hypentropy which has asymmetric Bregman divergence, we did properly incorporate the factor two gained from the symmetry of the least-squares-induced Bregman divergence D_F.

5.4 A closer look at the operator norm involved

Let m be the number of neurons, $(x_i \in \mathbb{R}^d)_{i\in[n]}$ the training samples, and $(\theta_j)_{j\in[m]} \in (\mathbb{S}^{d-1})^m$ the directions of the first layer neurons. The design matrix $X \in \mathbb{R}^{n\times m}$ is defined as $X_{i,j} = (\langle \theta_j, x_i \rangle)_+$. We wish to understand the operator norm L of the left inverse of $X^\top$, with respect to the norm $\|\cdot\|_*$ on $\mathbb{R}^m$,

$$L = \sup \left\{ \|u\|_2 \mid u \in \mathbb{R}^n, \|X^\top u\|_* \leq 1 \right\}$$

In the underspecified case, if the rank of X is less than n, then $L = +\infty$. Indeed for $u \in \mathrm{Ker}(X^\top) \setminus \{0\}$, and $\lambda \in \mathbb{R}_+$, it holds $X^\top(\lambda u) = 0$, however $\|\lambda u\|_2 \to_\lambda +\infty$. This happens for instance if there are two identical samples $x_i = x_j$ for $i \neq j$.

While it seems reasonable that the bound should become uninformative if there are two samples very close with wildly different responses y_i and y_j, it appears that if the responses are identical, the problem is not made much more difficult if the samples are similar. This absence of assumption on the response could be a good candidate explanation for the discrepancies observed between our bound and empirical behavior.

6 Conclusion

We have shown some bounds on the convergence speed of a two-layer network with fixed first layer directions and a least-squares objective, by reparameterizing this gradient flow as a mirror flow that we could analyze by showing that it satisfies some Łojasiewicz conditions. The choices of parameterization played a central role in this work, and the restricted setting chosen allowed us to obtain several consistent pictures of the training dynamic given by different choices of parameterizations. A choice of gradient flow in one parameterization induces a different metric in another parameterization, with far from obvious consequences on the inductive bias that it introduces, although these can be somewhat disentangled by changing points of view. Another important point that came up was the obscure consequences on the inductive bias triggered by some hyperparameter choices. What looks like a harmless engineering trick from one point of view can be interpreted as a dramatic change of hyperparameters in another. Different parameterizations, even when the dynamic they describe is identical, could have hyperparameters entangled differently, and it remains unclear how to perform meaningful comparisons to understand the intrinsic influence of a single hyperparameter.

However, the bounds obtained in this way do not seem to give a very accurate description of the network's weights' evolution in our numerical experiments, and the conditions in which our bounds are most accurate are far from the settings in which neural networks are practically used. The analysis also relies heavily on there being two layers with fixed directions for the first layer, which is a rather unusual choice limiting the applicability of these results. Moreover, the crucial steps of reparameterizing to obtain a mirror flow, and the convergence to the Bregman projection of the initialization seem to be specific to our assumptions, leaving little space to generalize these results to broader settings.

References

Heinz H. Bauschke and Jonathan M. Borwein. Legendre functions and the method of random bregman projections. *Journal of convex analysis*, 4(1):27–67, 1997.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

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

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry, 2020.

R Tyrrell Rockafellar. *Convex analysis*, volume 36. Princeton university press, 1970.

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. *arXiv preprint arXiv:2002.09277*, 2020.

A Minor technical lemmae

A.1 Existence and uniqueness of least-squares mirror flow

Lemma A.1. *Let* $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, *and define* $F : \mathbb{R}^d \to \mathbb{R}$ *the loss* $\beta \mapsto \frac{1}{2}\|X\beta - y\|_2^2$. *If* $\mathcal{D}$ *is a nonempty convex open subset of* $\mathbb{R}^d$, *and* $\phi : \mathcal{D} \to \mathbb{R}$ *is a convex function of Legendre type satisfying* $\nabla\phi(\mathcal{D}) = \mathbb{R}^d$, *then for every* $\beta_0 \in \mathbb{R}^d$, *there exists a unique solution* $\beta : \mathbb{R}_+ \to \mathcal{D}$ *to the differential equation* $\frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi(\beta) = -\nabla F(\beta)$ *such that* $\beta(0) = \beta_0$.

Proof. Let ϕ^* be the convex conjugate of ϕ. Since ϕ is Legendre and $\nabla\phi : \mathcal{D} \to \mathbb{R}^d$ is surjective, it is also a bijection of inverse $\nabla\phi^*$. Let $G : \mathbb{R}^d \to \mathbb{R}^d$ be the function $\gamma \mapsto \nabla F(\nabla\phi^*(\gamma))$, differentiable as a composition of differentiable functions. By strict convexity of ϕ, the inverse mirror map $\nabla\phi^*$ is locally Lipschitz, thus so is G. The gradient flow $\frac{\mathrm{d}}{\mathrm{d}t}\gamma = -G(\gamma)$ hence has a unique maximal differentiable solution $\gamma : I \to \mathbb{R}^d$, with I an open interval of $\mathbb{R}$ containing 0, by Picard-Lindelöf theorem. Taking $\beta = \nabla\phi^*(\gamma) : I \mapsto \mathcal{D}$ thus gives the existence of a solution. Similarly, if β and β' are two solutions, then $\nabla\phi(\beta)$ and $\nabla\phi(\beta')$ are two solutions of the previous gradient flow. By uniqueness of such solutions, $\beta = \beta'$. It then only remains to show that $\sup(I) = +\infty$. By maximality of the open interval I, it is sufficient to show that if $\sup(I) < +\infty$, then the path $\gamma = \nabla\phi(\beta)$ has finite length, so that it can be extended to contradict maximality of I. To do so, note that by a classical mirror descent argument (see Thm 2.3, or [Bubeck, 2015, Chap. 4]), the objective value $t \mapsto F(\beta(t))$ is decreasing over time, hence since the sublevel sets of F are compact, $t \mapsto \|\beta(t)\|_2$ is bounded. Observing that $\|\frac{\mathrm{d}}{\mathrm{d}t}\gamma\|_2 = \|\nabla F(\beta)\|_2$ is bounded, because ∇F is Lipschitz, concludes the proof. □

A.2 Lorentz cone condition for well-defined gradient flow

Lemma A.2. *Let $(x_i, y_i)_{i\in[n]}$ be a set of n observations with $x_i \in \mathbb{R}^k$ and $y_i \in \mathbb{R}$. Let $(\theta_i)_{j\in[m]}$ be the first layer directions $\theta_j \in \mathbb{S}^{k-1}$. Let $f : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ be the objective*

$$f : (a, b) \mapsto \frac{1}{2} \sum_{i\in[n]} \left(y_i - \sum_{j\in[m]} b_j (a_j \theta_j \cdot x_i)_+ \right)^2$$

If $(a_0, b_0) \in \mathbb{R}^m \times \mathbb{R}^m$ is such that $\forall i \in [m]$, $[a_0]_i^2 - [b_0]_i^2 \geq 0$, then there exists a unique solution $w : \mathbb{R}_+ \to \mathbb{R}^m \times \mathbb{R}^m$ to the differential equation $\frac{\mathrm{d}w}{\mathrm{d}t} = -\nabla f(w)$ satisfying $w(0) = (a_0, b_0)$.

The problem is that the objective f is not differentiable on its whole domain. Indeed, the ReLU non-linearity $x \mapsto (x)_+ = \max(0, x)$ is not differentiable at zero. Under the stated condition, we can rule out the singularities introduced by the ReLU non-linearity such that the gradient flow remains well defined. See Figure (4) for a visual depiction of this issue. Assume for simplicity that for all $(i, j) \in [n] \times [m]$, it holds $\theta_j \cdot x_i \neq 0$ (otherwise ommit this term in the loss to remove the singularity consistently, similarly to the following proof).

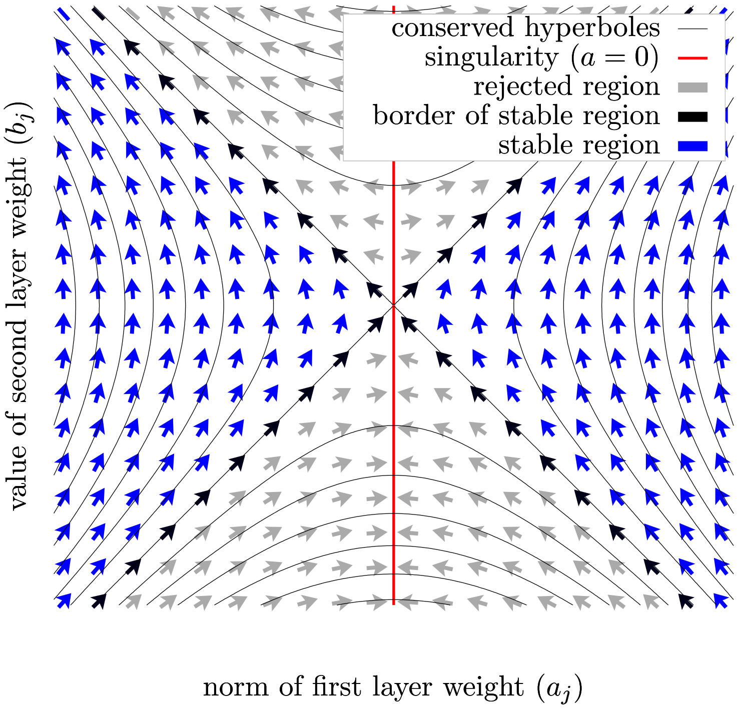

Figure 4: Directions of a gradient vector field with ReLU (colored by stability of region)

Proof. Let $K = \{(a, b) \in \mathbb{R}^m \times \mathbb{R}^m \mid \forall i \in [m], a_i^2 - b_i^2 \geq 0\}$, and for all $j \in [m]$, let $S_j = \{(a, b) \in K \mid (a_j, b_j) = (0, 0)\}$ be the j-th slice singularities. The objective function f is differentiable on $K \setminus (\cup_j S_j)$, and we can extend the definition of ∇f to all K by continuity, by setting $[\nabla f(a, b)]_j = (0, 0)$ if $(a, b) \in S_j$ as a convention in the definition of the gradient flow. Let $\delta_i : K \mapsto \mathbb{R}$ be the error in the response of the network, $\delta_i : (a, b) \mapsto \sum_{j\in[m]} b_j (a_j \theta_j \cdot x_i)_+ - y_i$. The gradient of the objective is as follows (well-defined since $a_j = 0 \Rightarrow b_j = 0$ in K).

$$\nabla f|_K (a, b) = \left(\sum_{i\in[n]} \operatorname{sign}(a_j) b_j (\theta_j \cdot x_i)_+ \delta_i(a, b),\ \sum_{i\in[n]} |a_j| (\theta_j \cdot x_i)_+ \delta_i(a, b) \right)_{j\in[m]}$$

It is then sufficient to check that any solution to the gradient flow equation remains in K. To do so, we will simply observe that the quantity $a_j^2 - b_j^2$ is constant, hence on every slice the hyperboles from Figure (4) that are conserved by gradient flow. Let $(a,b) : \mathbb{R}_+ \to \mathbb{R}^m \times \mathbb{R}^m$ be a solution to the gradient flow equation with $(a,b)(0) \in K$. Let $j \in [m]$. For succintness, write $\mu_j(a,b) = \sum_{i\in[n]} (\theta_j \cdot x_i)_+ \delta_i(a,b)$, and observe that $a_j^2 - b_j^2$ is constant over time

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(a_j^2 - b_j^2\right) = 2a_j\frac{\mathrm{d}a_j}{\mathrm{d}t} - 2b_j\frac{\mathrm{d}b_j}{\mathrm{d}t} = -2a_j\frac{\partial f}{\partial a_j} + 2b_j\frac{\partial f}{\partial b_j} = 2\mu_j(a,b)\left(-\operatorname{sign}(a_j)a_j b_j + b_j|a_j|\right) = 0$$

Hence for all $t \in \mathbb{R}_+$, $a_j(t)^2 - b_j(t)^2 = a_j(0)^2 - b_j(0)^2 \geq 0$, thus $(a(t), b(t)) \in K$. Moreover, a_j does not change signs, for if $a_j(t) = 0$ for some t, then $b_j(t) = 0$ because $a_j(t)^2 - b_j(t)^2 \geq 0$, and thus (a_j, b_j) is stationnary. □

In high dimensions[6], and with independently identically distributed weights for each neuron, the above condition is satisfied with very high probability. Intuitively, the gaussian distribution in high dimensions is similar to the uniform distribution on a sphere, that is to say that the ℓ_2-norm of a gaussian variable in high dimensions is very concentrated around a single value that increases with the dimension. In this setting, the condition that states that the first layer's norm must be larger than the second layer's corresponding weight is almost always satisfied. Figure (5) depicts the probability that all m neurons satisfy the condition as a function of the input dimension d, for gaussian and uniform initializations.

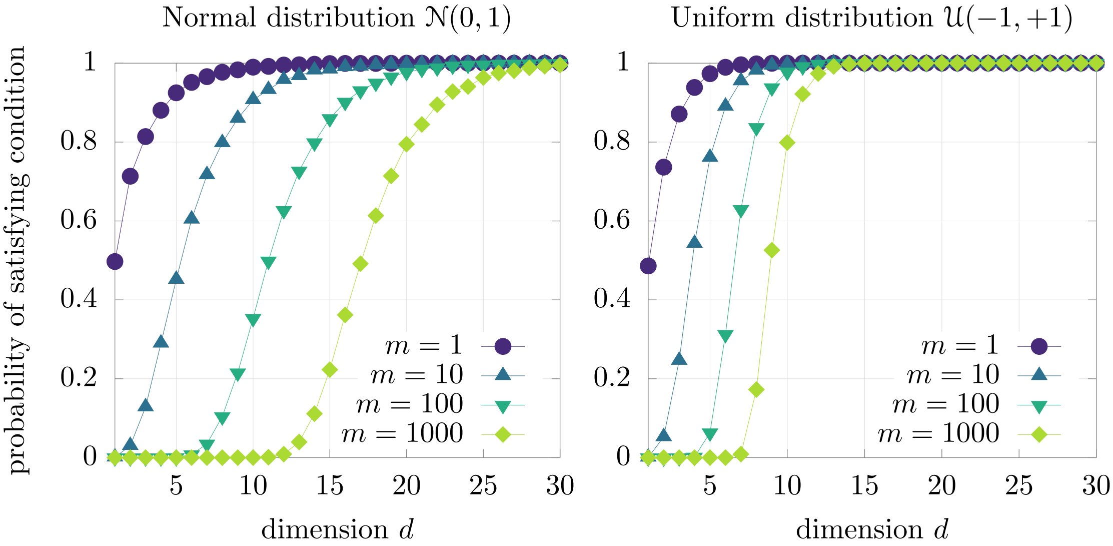

Figure 5: Probability of satisfying the condition for iid weights and m neurons.

A.3 Hypentropy properties

Lemma A.3. *For $\alpha \in \mathbb{R}_+^*$, the α-hypentropy $\phi_\alpha : x \mapsto x \operatorname{arcsinh}(x/\alpha) - \sqrt{x^2+\alpha^2} + \alpha$ is twice differentiable over $\mathbb{R}$ with first and second derivatives $\phi'_\alpha : x \mapsto \operatorname{arcsinh}(x/\alpha)$ and $\phi''_\alpha : x \mapsto 1/\sqrt{x^2+\alpha^2}$. For the case $\alpha = 0$, for any $z \in \mathbb{R} \setminus \{0\}$, the 0-hypentropy compatible with z, $\phi_0 : x \mapsto |x| \log |x| - |x| + 1$ is twice differentiable over its domain with first and second derivatives $\phi'_0 : x \mapsto (x/|x|) \log |x|$ and $\phi''_0 : x \mapsto 1/|x|$.*

Note that in both cases, the second derivative of ϕ_α can be written $x \mapsto 1/\sqrt{x^2+\alpha^2}$.

[6] that is input dimension d, not number of neurons m, with the previous notation

Proof. For the first part, recall that $\operatorname{arcsinh} : x \mapsto \log(x + \sqrt{x^2+1})$ is differentiable over $\mathbb{R}$ with derivative $x \mapsto 1/\sqrt{x^2+1}$. Then for any $\alpha > 0$, ϕ_α is differentiable as a composition of differentiable functions, and a quick computation shows the result. For the second part, if $z > 0$ then the domain is $\mathbb{R}_+^*$, and $\phi_0(x) = x \log x - x + 1$, which is twice differentiable as before with first and second derivatives $\phi_0'(x) = \log x$ and $\phi_0''(x) = 1/x$. If on the other hand $z < 0$, then the domain is $\mathbb{R}_-^*$, and $\phi_0(x) = -x\log(-x) + x + 1$ is also twice differentiable with first and second derivatives $\phi_0'(x) = -\log(-x)$ and $\phi_0''(x) = -1/x$, hence the result. □

A.4 Some more numerical experiments

For the bound $D_\phi(\beta^*, \beta(t)) \leq \frac{1}{\kappa_2 t + c}$ from Lemma (3.7), our constant $\kappa_2 = \left(C^2 \left\|X^{\top\dagger}\right\|_{\text{op}}^2\right)^{-1}$ uses a constant that remains finite when the number of neurons m tends to infinity, $C = 2m\log\left(1 + \frac{2M}{m(\inf_i r_i)}\right) \underset{m\to\infty}{\longrightarrow} \frac{4M}{\inf_i r_i}$. However, as the number of neurons grows, its dependence on the vanishing parameter $(\inf_{i\in[m]} r_i)$ becomes worse, rendering this second bound almost as uninformative as the linear bound. Figure (6) depicts some results for $m = 1000$ neurons.

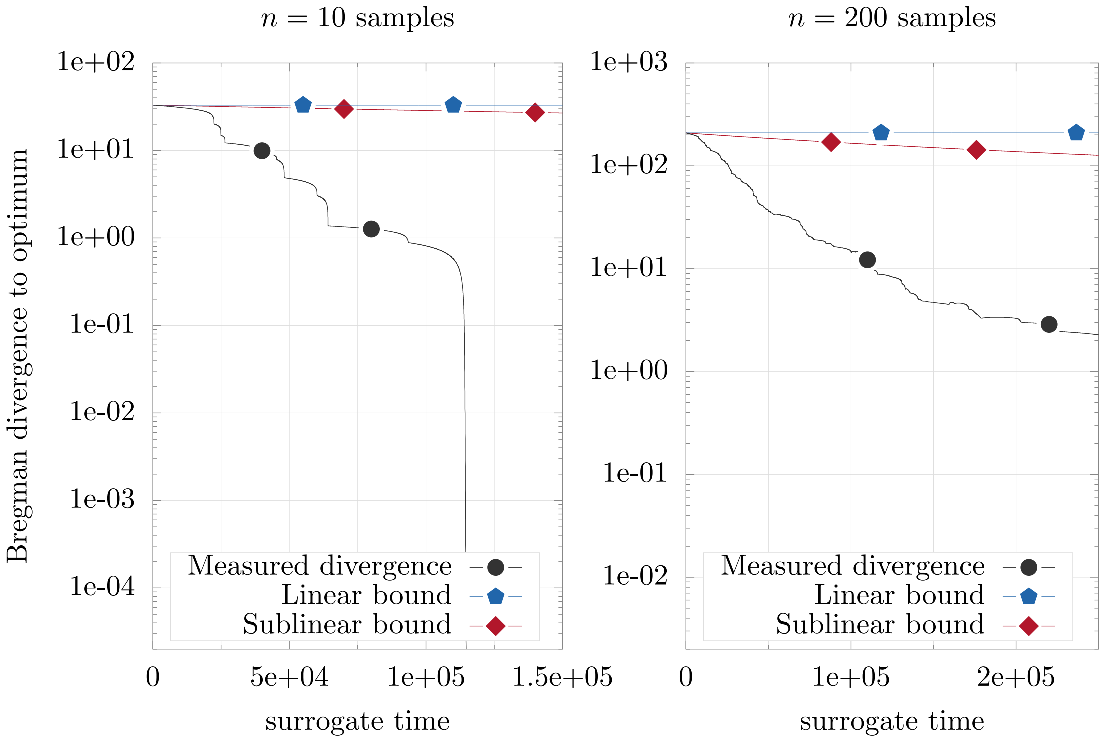

Figure 6: Convergence speed measurements $[\,d = 100, m = 1000, \alpha = 10^{-4}\,]$

In the first example, both constants ($\kappa_1 \approx 6.9\times10^{-9}$ and $\kappa_2 \approx 4.7\times10^{-8}$) are too small to indicate meaningful improvements. In this example, $(\inf_i r_i) \approx 1.1 \times 10^{-10}$, $\left\|X^{\top\dagger}\right\|_{\text{op}} \approx 0.13$, and $M \approx 3.2$. The divergence is seen dropping on the figure and reaches 10^{-15}, after which it becomes numerically unstable.

In the second example, the constants have relatively similar values ($\kappa_1 \approx 2.2 \times 10^{-9}$ and $\kappa_2 \approx 1.3 \times 10^{-8}$). The maximum ℓ_1 norm of β on the trajectory rises to $M \approx 20$, likely due to the higher number of samples leading to a less sparse optimum, and $(\inf_i r_i) \approx 1.2 \times 10^{-10}$, $\left\|X^{\top\dagger}\right\|_{\text{op}} \approx 2.3 \times 10^{-1}$. The training is stopped early for this latter experiment (objective value at 10^{-5} at $t = 4.0 \times 10^5$) due to the particularly slow convergence speed and increased computational cost of each iteration.