# Implicit Bias of Mirror Flow in Homogeneous Neural Networks: Sparse and Dense Feature Learning

**Tom Jacobs**
CISPA Helmholtz Center
tom.jacobs@cispa.de

**Guido Montúfar**
UCLA and MPI MiS
montufar@math.ucla.edu

## Abstract

We study the max-margin solutions reached by mirror flow in deep neural networks with homogeneous activation functions. Extending classical results on gradient flow, we derive a novel balance equation for mirror flow from convex duality, enabling a characterization of the horizon function governing the induced margin. We further establish max-margin characterizations together with convergence rates and norm growth estimates. Finally, we support our theory through experiments on synthetic datasets and standard vision tasks. Concretely, we show that: (1) distinct non-homogeneous mirror maps can induce the same max-margin solution; (2) convergence can be extremely slow, including exponentially slow regimes; and (3) although all considered mirror maps exhibit feature learning, they can produce markedly different representations, ranging from sparse to dense neuron activations. Together, these results provide a unified perspective on sparse and dense feature learning in homogeneous neural networks, highlighting how mirror maps shape both optimization dynamics and the geometry of the learned classifiers.

## 1 Introduction

Understanding why gradient-based optimization can generalize well is one of the central questions in deep learning theory. A key insight from the implicit regularization literature is that gradient descent training does not simply return a minimizer of the loss function but instead is implicitly biased towards particular types of minimizers. One of the most general results presently available in this context states that for homogeneous neural network classifiers trained on data that is separable by the network, gradient flow approaches an $L_2$ max-margin classifier [33]. This result connects the optimization dynamics induced by the algorithm and the network parameterization to the geometric properties of the resulting decision boundary, which directly influence the generalization behavior.

However, gradient flow is only one algorithm within a much broader family of optimization algorithms. Mirror flows generalize gradient flows by replacing the Euclidean geometry of the search space with the geometry induced by a strictly convex mirror potential $R$. The resulting parameter updates take place in the dual space induced by the mirror map $\nabla R$. Mirror flows arise naturally in several contexts. In particular, they arise as reparameterized gradient flows of overparameterized models [29, 15] and they are used to implement structural parameter constraints such as sparsity or non-negativity. In spite of the importance of these methods, a unified theory of implicit regularization for mirror flows on neural networks is still missing.

Prior work on implicit regularization of mirror flows has largely focused on linear models. In this context, [45] characterized the max-margin for homogeneous mirror maps, and more recently, [40] showed how to deal with non-homogeneous separable mirror maps by introducing a horizon function that captures the limiting direction of the iterates. This is particularly useful for mirror maps like the hyperbolic entropy used for sparse training [15, 17]. To extend this line of results to the case of non-linear homogeneous networks, we identify and resolve two key difficulties: (i) defining an appropriate notion of margin, and (ii) establishing convergence of the iterates in direction to a KKT point of

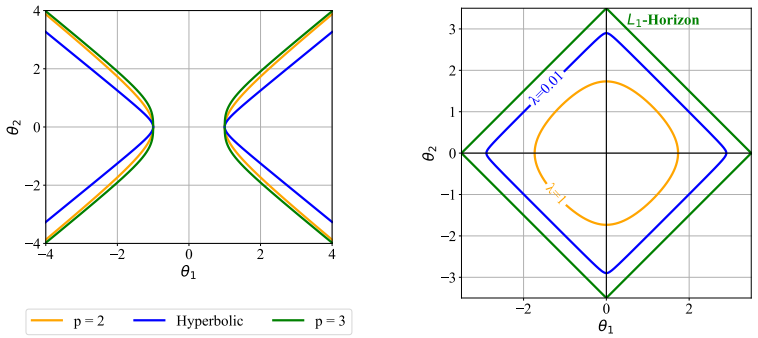


Figure 1: Left: Balance equations. Shown are the solution sets of the balance equations for distinct mirror maps, characterized by $Q(\nabla R)$. Right: Horizon functions. Shown are the level sets $Q(\nabla R) = c$ for the hyperbolic entropy with different values of the hyperparameter $\lambda$. Although all choices of $\lambda$ share the same horizon function, given by the $L_1$-norm, for larger $\lambda$ substantially larger parameter magnitudes $\|\theta\|$ are required for these level sets to approximate the limiting shape.

the corresponding max-margin. Note that for homogeneous networks with an exponential-tail loss, driving the loss to zero requires the parameters to diverge. One therefore cannot study convergence of the parameters directly, but rather convergence in direction.

Our starting point is the identification of a new *balance equation* for mirror flow, illustrated in Figure 1. This provides an analog of the layer-wise norm balance equation well known for gradient flow. Building on this,

we introduce the *$Q$-margin* as a suitable notion of margin for mirror flows, defined via the dual potential $Q(\nabla R(\theta))$, where $Q := R^*$ denotes the convex conjugate of $R$. This is the direct mirror analog of the $L_2$-margin for gradient descent. Using these ingredients, we show that mirror flow converges in direction to a KKT point of a constrained optimization problem whose objective is the square of a *horizon function* of the form $\phi_\alpha(\theta) := \lim_{\eta\to 0} \eta(\alpha Q(\nabla R(\theta/\eta)))^{1/\alpha}$, corresponding to a max-margin solution in the induced geometry. The horizon function captures the asymptotic geometry of the mirror potential at large parameter scales and governs the limiting direction of the dynamics. Here, $\alpha$ is the asymptotic homogeneity degree of the mirror potential. In the homogeneous case $R(\theta) = \frac{1}{p}\|\theta\|_p^p$, this reduces to $\|\theta\|_p$, with $\alpha = p$.

Next, to gain further insight into the role of non-homogeneity in the potential function, we consider a family of potentials with a hyperparameter $\lambda \geq 0$ which controls the level of non-homogeneity. We show that although distinct non-homogeneous mirror potentials can induce the same horizon function, and thus converge to the same max-margin solution, they can have vastly different convergence rates, in some cases requiring exponentially longer time. Figure 1 illustrates that larger values of $\lambda$ require disproportionately larger parameter magnitudes for the horizon function to closely approximate its asymptotic limit. These results provide concrete guidance for selecting the degree of inhomogeneity to balance convergence rate and other factors such as training stability.

Finally, we show that mirror descent induces different feature representations through its geometry. In Theorem 4.9, we show for a two-layer network, hyperbolic entropy leads to fewer active neurons, whereas a smoothed homogeneous mirror potential activates more, resulting in sparse versus dense feature learning (Figure 2).

**Main contributions.** In this work, we characterize the implicit bias of mirror flow in training homogeneous neural network classifiers, with the following contributions:

- **Balance equation.** For layerwise separable mirror maps, we derive a novel balance equation, a conserved quantity of the mirror flow (Lemma 3.1), using the Fenchel-Young identity (Lemma 2.5).
- **Max-margin.** Based on the balance equation, we introduce a novel $Q$-margin for mirror flow in Theorem 4.5 and derive the corresponding constrained optimization problem in Theorem 4.7.
- **Reaching the max-margin solution.** We derive tight growth rates for $Q$ and decrease rates for the loss in Theorem 4.6. These highlight for non-homogeneous mirrors the difficulty of reaching the

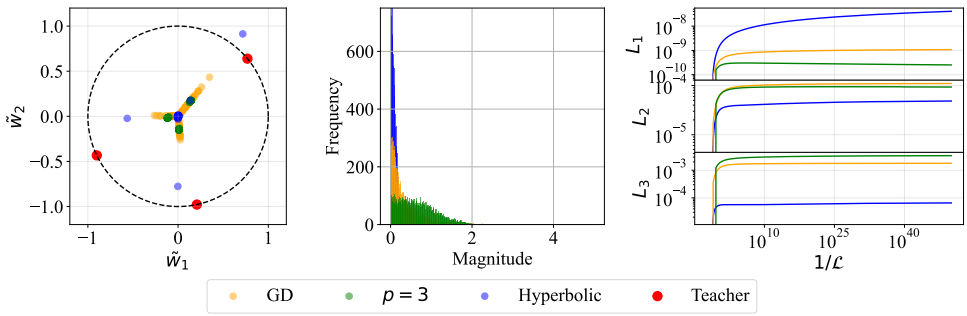


Figure 2: **Different types of feature learning under mirror flows.** The hyperbolic entropy induces sparse feature learning with fewer active neurons, whereas the smoothed homogeneous potential ($p = 3$) induces dense feature learning with more active neurons. **(Left)** Input weight representations $\tilde{w} = |a_j|w_j$ of a two-layer student network in a student-teacher setup: GD (orange) produces diffuse weights near the origin, $p = 3$ (green) learns teacher directions with all active neurons, and the hyperbolic map (blue) concentrates mass onto only a few neurons aligned with the teacher (red), reflecting sparsity. **(Middle)** Weight magnitude distribution of the hidden layer in a three-layer network: GD and $p = 3$ yield spread out distributions with many active weights, while the hyperbolic entropy produces a sharply peaked distribution near zero with few large-magnitude weights. **(Right)** The $L_k$-margin, for $k = 1, 2, 3$ as a function of $1/\mathcal{L}$ for a three-layer network: each mirror flow converges in direction to their corresponding max-margin.

corresponding max-margin solution. We leverage this analysis to provide guidance for selecting hyperparameters for mirror maps in Lemma 4.10.

- **Sparse and dense feature learning.** We show that mirror descent can learn both sparse and dense representations, as established in Theorem 4.9, and we empirically verify the different margins reached and the difficulty of reaching them depending on the smoothening parameter $\lambda$.
- **NTK and modularity.** In Appendix I, we show that the constrained optimization problem admits an SVM formulation that cannot be captured within an RKHS framework, but instead requires a Reproducing Kernel Banach Space (RKBS). We further formulate the constrained optimization problem for layer-wise distinct mirror maps, highlighting its modular structure.

**Related work.** We highlight key related works here and defer broader discussion to Appendix A, including connections to general implicit bias literature and hyperparameter transfer.

- **Max-margin results.** A substantial body of work characterizes the implicit bias of gradient-based optimization in terms of margin maximization. Classical results establish that gradient descent on (linearly) separable data converges in direction to the $L_2$ max-margin classifier [44, 33, 19]. These results have recently been extended to steepest-descent [46], inhomogeneous neural networks [4], simplicity bias [47], and the edge-of-stability regime [51]. Moreover, benign overfitting has been studied under approximate max-margin convergence [21]. However, a max-margin characterization of mirror flow for neural networks beyond the linear setting has remained open [45, 40].
- **Mirror flow.** Mirror descent and its continuous-time limit, mirror flow, are classical tools in optimization that enforce constraints on iterates via a Bregman divergence [50, 37, 22, 34]. Mirror flow also arises naturally as the dynamics of reparametrized gradient descent [29, 15, 17, 6, 49, 5], and has been used to promote sparsity [15, 31]. Its implicit bias has been characterized in restricted settings such as linear classification [45, 40], matrix factorization [13], univariate regression with two-layer neural networks [30], and single layer attention [20]. We provide the first max-margin result for mirror flow in the broader setting of homogeneous neural networks.
- **Balance equation.** The balance equation describes a conserved quantity of gradient flow that appears in deep linear networks [9, 35], Riemannian gradient flow [37], and structured architectures such as transformers and ResNets [36]. Beyond standard gradient descent, analogous balance equations are known only for diagonal linear networks [18]. In contrast to prior derivations relying on architectural homogeneity or Riemannian structure, we derive a new balance equation via convex duality, which is central to our max-margin analysis for homogeneous networks.

## 2 Preliminaries

We consider homogeneous neural networks trained for binary classification. In this section, we first recall the balance equation for gradient flow and then introduce the definitions and assumptions from convex analysis needed to define mirror flow. Informally, given an objective function $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$, the mirror flow associated with a strictly convex differentiable potential $R : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$d\nabla R(\theta_t) = -\nabla \mathcal{L}(\theta_t)dt, \qquad \theta_0 = \theta_{\text{init}}.$$

**Neural network and loss.** Let $f$ be a neural network that, given an input $x \in \mathbb{R}^d$ and parameters $\theta$, outputs a real value $f(\theta, x)$. The sign of $f(\theta, x)$ determines the predicted class. We denote the training dataset by $Z := \{(x_i, y_i) : i \in [K]\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$ denote the input and label, respectively. For a loss function $\ell : \mathbb{R} \to \mathbb{R}$, the empirical training loss is given by $\mathcal{L}(\theta) := \sum_{i=1}^{K} \ell(y_i f(\theta, x_i))$.

**Assumption 2.1.** We impose the following assumption on the classifier $f$ and loss function $\ell$, following [33]:

- (Regularity). For any fixed $x$, the map $\theta \mapsto f(\theta, x)$ is locally Lipschitz and satisfies the chain rule.
- (Homogeneity). There exists $L > 0$ such that for all $c > 0$, $f(c\theta, x) = c^L f(\theta, x)$.
- (Exponential Loss). $\ell(q) = \exp(-q)$.
- (Separability). There exists $t_0$ such that $\mathcal{L}(\theta(t_0)) < 1$.

Since $f$ is only assumed to be locally Lipschitz, we work with Clarke subdifferentials. The subdifferential of $f$ at a point $\theta \in A$, where $A$ is a convex set, is defined as

$$\partial^\circ f(\theta) := \text{conv}\left\{ \lim_{k\to\infty} \nabla f(\theta_k) : \theta_k \to \theta, f \text{ is differentiable at } \theta_k \right\}.$$

Following [33, 8], we say that a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ admits the chain rule if, for any arc $z : [0, \infty) \to \mathbb{R}^n$, $\frac{d}{dt}(f \circ z)(t) = \langle h, \dot{z}(t) \rangle$ for all $h \in \partial^\circ f(z(t))$, for a.e. $t > 0$. Concretely, we consider neural networks of the form

$$f(\theta, x) := W_1 \sigma\left(W_2 \cdots \sigma\left(W_L x\right)\right), \tag{1}$$

where $\theta = [W_1, \ldots, W_L]$ and $\sigma$ is a homogeneous activation function, such as the ReLU or linear.

**Balance equation and margin for gradient flow.** For homogeneous neural networks as in Eq. (1), standard gradient flow with $R(\theta) = \frac{1}{2}\|\theta\|_2^2$ satisfies the following *balance equation* for all $t \geq 0$ and $i \in [L-1]$:

$$\frac{1}{2}\|W_{i,t}\|_2^2 - \frac{1}{2}\|W_{i+1,t}\|_2^2 = \frac{1}{2}\lambda_{0,i}, \tag{2}$$

where $\lambda_{0,i} = \|W_{i,0}\|_2^2 - \|W_{i+1,0}\|_2^2$. Eq. (2) shows that the squared Euclidean norms of the weight matrices across adjacent layers grow proportionally over time. This property was exploited in [33] to derive the $L_2$ max-margin characterization of the implicit bias of gradient flow.

We recall the definition of a generalized margin with respect to the $L_2$-norm, following [33], and later extend it to more general settings. This notion captures the minimal signed distance of the training input data to the decision boundary, measured in the geometry induced by the parameter norm.

**Definition 2.2** ($L_2$-margin)**.** Given a dataset $\{(x_i, y_i) : i \in [K]\}$ and a function $f$ that is $L$-homogeneous in its parameters, the generalized $L_2$-margin is a function of $\theta$ defined as

$$\gamma := \min_i y_i f\left(\frac{\theta}{\|\theta\|_2}, x_i\right) = \min_i \frac{y_i f(\theta, x_i)}{\|\theta\|_2^L}.$$

Moreover, we denote $q_i := y_i f(\theta, x_i)$ and $q_{\min} := \min_i q_i$.

**Convex analysis.** To define mirror flow, we first introduce the mirror potential $R : \mathbb{R}^n \to \mathbb{R}$.

**Assumption 2.3.** For our analysis, we impose the following assumptions on $R$: (1) strict convexity (2) twice differentiable (3) coercivity, i.e., $\|\theta\| \to \infty$ implies $\|\nabla R\| \to \infty$, so that $\nabla R$ is surjective onto $\mathbb{R}^n$, and (4) $\nabla^2 R$ is locally Lipschitz.

We recall the convex conjugate and Fenchel-Young duality in Definition 2.4 and Lemma 2.5.

**Definition 2.4.** The convex conjugate of $R$ is the function $Q : \mathbb{R}^n \to \mathbb{R}$ defined by $Q(\theta) := \sup_{\xi \in \mathbb{R}^n} \{\langle \theta, \xi \rangle - R(\xi)\}$.

**Lemma 2.5** (Fenchel-Young identity)**.** *Let $R$ satisfy Assumption 2.3 and let $Q$ denote its convex conjugate. Then for all $\theta \in \mathbb{R}^n$,*

$$\langle \theta, \nabla R(\theta) \rangle = R(\theta) + Q(\nabla R(\theta)).$$

Applying Lemma 2.5 to the Euclidean case $R(\theta) := \frac{1}{2}\|\theta\|_2^2$, we obtain $Q(\nabla R(\theta)) = \frac{1}{2}\|\theta\|_2^2$. This matches the layer-wise quantities appearing in the gradient flow balance equation in Eq. (2). As we will show in Section 3, this identity allows us to define an analogous balance equation for mirror flow, where $R(\theta)$ and $Q(\nabla R(\theta))$ need not coincide.

**Mirror flow.** In homogeneous neural networks, mirror flow is defined by the differential inclusion

$$\frac{d\nabla R(\theta_t)}{dt} \in -\partial^\circ \mathcal{L}(\theta_t), \quad \text{or equivalently} \quad \frac{d\theta_t}{dt} \in -(\nabla^2 R)^{-1}(\theta_t)\, \partial^\circ \mathcal{L}(\theta_t).$$

When $f$ is differentiable, this reduces to $\frac{d\theta_t}{dt} = -(\nabla^2 R)^{-1}(\theta_t)\nabla \mathcal{L}(\theta_t)$. Under our regularity assumptions, the training loss satisfies the chain rule, yielding

$$\frac{d}{dt}\mathcal{L}(\theta_t) = -\left\langle h_t, \frac{d\theta_t}{dt} \right\rangle = -\left\| \frac{d\theta_t}{dt} \right\|^2_{\nabla^2 R(\theta_t)}, \quad \text{for some } h_t \in \partial^\circ \mathcal{L}(\theta_t).$$

Here $\|\cdot\|_{\nabla^2 R(\theta)} := \sqrt{\langle \cdot, \nabla^2 R(\theta) \cdot \rangle}$ denotes the local norm induced by the mirror potential.

## 3 Balance equation for mirror flow

For homogeneous networks, the direction of convergence is governed by the growth of the parameter norm. We derive a balance equation for the mirror flow of a homogeneous neural network $f$ of the form given in Eq. (1). To this end, we assume that the mirror potential is layer-wise separable, i.e., $R(\theta) = \sum_{i=1}^{L} R_i(W_i)$. Then the mirror flow then takes the form

$$d\nabla R_i(W_{i,t}) = -\nabla_i \mathcal{L}(\theta_t)dt, \qquad W_{i,0} = W_{i,\text{init}}, \tag{3}$$

for $i \in [L]$, where $\nabla_i$ denotes the gradient with respect to the parameters of the $i$th layer. Since the network is not assumed to be $C^1$, this should be interpreted as a differential inclusion. In particular, solutions need not be unique and are understood in a set-valued sense.

**Lemma 3.1.** *The mirror flow in Eq. (3) satisfies the following balance equation for all $t \geq 0$:*

$$Q_i(\nabla_i R_i(W_{i,t})) - Q_{i+1}(\nabla_{i+1} R_{i+1}(W_{i+1,t})) = \lambda_{0,i},$$

*where $\lambda_{0,i} = Q_i(\nabla_i R_i(W_{i,0})) - Q_{i+1}(\nabla_{i+1} R_{i+1}(W_{i+1,0}))$, for all $i \in [L-1]$.*

*Proof sketch.* It follows from homogeneity of $f$ and convex duality, with details in Appendix B. □

Lemma 3.1 recovers the gradient flow balance equation in the special case $R(\theta) = \frac{1}{2}\|\theta\|_2^2$. Moreover, it identifies $Q(\nabla R(\theta))$ as a natural normalization for the margin, as both the balance equation and the growth of the margin are governed by the homogeneity of $f$. To relate the dynamics to the margin, we invoke Euler's identity: if $f$ is homogeneous of degree $L$, then $\langle \theta_t, h \rangle_2 = L f(\theta_t; x),$ for all $h \in \partial^\circ f(\theta_t; x)$. This, together with Lemma 2.5, implies

$$\frac{d}{dt} Q(\nabla R(\theta_t)) = \left\langle \theta_t, \frac{d}{dt} \nabla R(\theta_t) \right\rangle_2 = -\langle \theta_t, g_t \rangle_2 \geq L\mathcal{L}(\theta_t)\, \log\big(1/\mathcal{L}(\theta_t)\big), \tag{4}$$

for some $g_t \in \partial^\circ \mathcal{L}(\theta_t)$. Here the first equation follows from Lemma 2.5, the second from mirror flow dynamics, and the third from Euler's identity (using the homogeneity of $f$) and the structure of the exponential loss, the full calculation is in Eq. (7) in the appendix. This shows that the evolution of $Q(\nabla R(\theta))$ is controlled by the loss. This extends the analysis of the gradient flow setting in [33], corresponding to $R(\theta) := \frac{1}{2}\|\theta\|_2^2$, to the mirror flow setting with general potentials.

**Mirror potentials.** We focus on two classes of mirror potentials: the hyperbolic entropy and smoothened homogeneous potentials, summarized in Table 1. The hyperbolic entropy is known to promote sparsity [50, 16]. The horizon function $\phi(\theta)$ characterizes the asymptotic behavior of the potential, capturing the shape of the level sets of $R(\theta)$ as $\|\theta\| \to \infty$, as illustrated in Figure 1. The corresponding balance equation is illustrated in Figure 1.

Table 1: Examples of separable mirror potentials $R$ and their corresponding dual potentials $Q(\nabla R(\theta))$ and horizon functions $\phi(\theta)$. Moreover we have that $\frac{1}{p} + \frac{1}{q} = 1$.

| **Mirror Potential** | $R(\theta)$ | $Q(\nabla R(\theta))$ | $\phi(\theta)$ |
|---|---|---|---|
| Hyperbolic Entropy | $\sum_{i=1}^{n} \theta_i \operatorname{arcsinh}\left(\frac{\theta_i}{\sqrt{\lambda}}\right) - \sqrt{\theta_i^2 + \lambda}$ | $\sum_{i=1}^{n} \sqrt{\theta_i^2 + \lambda}$ | $\|\theta\|_1$ |
| Smoothened Hom. | $\frac{1}{p}\|\theta\|_p^p + \frac{\lambda}{2}\|\theta\|_2^2,\ p \geq 2$ | $\frac{1}{q}\|\theta\|_p^p + \frac{\lambda}{2}\|\theta\|_2^2$ | $(p-1)^{1/p}\|\theta\|_p$ |

## 4 Late-phase max-margin characterization

We are now ready to characterize the implicit bias of mirror flow towards max-margin solutions. First we introduce the $Q$-margin and the corresponding $Q$-soft margin inspired by the balance equation presented in Section 3 which allows us to relate the dynamics to the loss.

**The $Q$-margin and a soft approximation.** Analogous to the $L_2$-margin we define the $Q$-margin for asymptotically homogeneous mirror maps.

**Definition 4.1** ($Q$-margin)**.** For an $L$-homogeneous function $f$ and asymptotically $\alpha$-homogeneous mirror map $R$, the $Q$-margin is a function of $\theta$ and defined as

$$\gamma_Q := \min_i y_i f\left(\frac{\theta}{(\alpha Q(\nabla(R(\theta)))^{1/\alpha}}, x_i\right) = \min_i \frac{y_i f(\theta, x_i)}{(\alpha Q(\nabla(R(\theta)))^{L/\alpha}}.$$

Definition 4.1 generalizes the $L_2$-margin for GD which is recovered by setting $R(\theta) = \frac{1}{2}\|\theta\|_2^2$. Next, we introduce the *$Q$-soft margin* as a tractable surrogate to track the $Q$-margin during training. Similar to [33] and [46], we define the soft margin using the log-sum-exp (LSE). The term soft refers to replacing the exact minimum with a smooth approximation, in which the margin is directly controlled by the dynamics via Eq. (4).

**Definition 4.2** ($Q$-soft margin)**.** The $Q$-soft margin is defined as

$$\tilde{\gamma}_Q := \frac{\log(\frac{1}{\mathcal{L}(\theta)})}{(\alpha Q(\nabla(R(\theta)))^{L/\alpha}},$$

where $\mathcal{L}(\theta)$ is the loss.

Note that when the training set is correctly classified we have that $\mathcal{L}(t_0) < 1$ and $\tilde{\gamma}_Q > 0$ by Definition 4.2. This is a key property we will use to show the $Q$-soft margin continues to grow once the data is correctly classified. Lemma 4.3 shows that the $Q$-soft margin approximates the $Q$-margin.

**Lemma 4.3.** *The $Q$-soft margin is a $\mathcal{O}((\alpha Q \nabla(R(\theta)))^{-L/\alpha})$ additive approximation of the $Q$-margin.*

*Proof sketch.* This follows from using log-sum-exp, with details in Appendix D. □

**Assumption 4.4.** The mirror potential $R : \mathbb{R}^n \to \mathbb{R}$ is asymptotically $\alpha$-homogeneous with $\alpha \geq 1$ and satisfies:

$$\alpha Q(\nabla R(\theta)) \geq \|\theta\|^2_{\nabla^2 R}.$$

Assumption 4.4 is satisfied by all mirror potentials listed in Table 1, as shown in Appendix C.

**Alignment of the $Q$-margin.** We now show that mirror flow naturally controls the growth of the $Q$-soft margin, which in turn leads to an increase in the $Q$-margin.

**Theorem 4.5.** *Under Assumptions 2.1 and 2.3 and Assumption 4.4, the $Q$-soft margin is an $\mathcal{O}((\alpha Q \nabla(R(\theta)))^{-L/\alpha})$ additive approximation for the $Q$-margin and the following hold:*

- *For a.e. $t > t_0$, $\frac{d}{dt}\tilde{\gamma}_Q(\theta_t)) \geq 0$.*
- *$\mathcal{L}(\theta_t) \to 0$ and $Q(\nabla R(\theta_t)) \to \infty$ as $t \to \infty$, therefore $|\gamma_Q - \tilde{\gamma}_Q| \to 0$.*

*Proof sketch.* We use Eq. (4) to control the $Q$-soft margin growth and use Lemma 4.3, for details see Appendix D. □

Theorem 4.5 shows that mirror flow naturally drives the $Q$-soft margin to grow, and that as the loss converges to zero, the $Q$-soft margin closely tracks the $Q$-margin.

**Convergence rate** We obtain the following convergence rates for mirror flow, which depend on the degree of homogeneity $\alpha \geq 1$ of the mirror map.

**Theorem 4.6.** *Under the same assumptions as in Theorem 4.5, we have the following.*
*For $\alpha \in [1, 2]$: $\mathcal{L}(\theta_t) = \Theta\left(\frac{1}{t\ \log(t)^{2-\frac{\alpha}{L}}}\right)$ and $(\alpha Q(\nabla R(\theta_t)))^{L/\alpha} = \Theta(\log\ t)$.*
*For $\alpha > 2$: $\mathcal{L}(\theta_t) = O\left(\frac{1}{t\ \log(t)^{2-\frac{\alpha}{L}}}\right)$ and $(\alpha Q(\nabla R(\theta_t)))^{L/\alpha} = \Omega(\log\ t)$.*

*Proof.* See Appendix F. □

Theorem 4.6 shows that a larger $\alpha$ may slow down convergence up to a multiplicative logarithmic factor. However, for large depth $L$ this may become negligible.

**Implicit bias description.** We obtain an implicit bias description of mirror flow in terms of a corresponding constrained optimization problem and its approximate KKT conditions. For simplicity of presentation, the following Theorem 4.7 specializes to the smoothened homogeneous mirror potentials listed in Table 1. We provide a more general result in Theorem G.2.

**Theorem 4.7.** *(Smoothened homogeneous) Assume $R$ is a smoothened homogeneous potential with $\alpha = p \geq 2$. Then, under the same assumptions as in Theorem 4.5, the parameter direction $\bar{\theta}_t := \frac{\theta_t}{\|\theta_t\|_2}$ converges to a KKT point of the following optimization problem:*

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\phi_\alpha^2(\theta) \qquad \text{such that} \qquad y_i f(\theta, x_i) \geq 1, \text{ for all } i \in [K],$$

*where $\phi_\alpha := \lim_{\eta \to 0} \eta(\alpha Q(\nabla R(\theta/\eta)))^{1/\alpha}$ is the horizon function.*

*Proof sketch.* The proof follows a similar strategy as used in [33] and [46]. The main differences are: (1) Approximating the horizon function by the iterates of $\nabla(\alpha Q(\nabla R(\theta)))^{2/\alpha}$, which requires $\alpha \geq 2$. (2) Showing that $\nabla(\alpha Q(\nabla R(\theta)))^{1/\alpha}/q_{\min}^{1/L}$ tends to an approximate KKT point. (3) Showing convergence of the normalized dual iterates to a limit point, this implies that the normalized primal iterates converge as well. The full proof is given in Appendix G.2. □

Theorem 4.7 relies on $\alpha \geq 2$. For the hyperbolic entropy, which does not satisfy that assumption (see details in Appendix G.4), we can exploit a connection to homogeneous reparameterizations trained with gradient flow from [29, 15, 17]. This allows us to apply results for homogeneous neural networks trained with gradient flow and show the following.

**Corollary 4.8.** *(Hyperbolic entropy) For the hyperbolic entropy $R_\lambda(\theta)$ with hyperparameter $\lambda > 0$ under same assumptions as Theorem 4.5, the parameter direction $\bar{\theta}_t := \frac{\theta_t}{\|\theta_t\|_1}$ converges to a KKT point of the following optimization problem:*

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\phi_1^2(\theta) \qquad \text{such that} \qquad y_i f(\theta, x_i) \geq 1, \text{ for all } i \in [K],$$

*where $\phi_1 := \|\theta\|_1$ is the corresponding horizon function.*

*Proof sketch.* We use the relationship between 2-homogeneous reparameterization and the hyperbolic mirror map, which allows us to apply Theorem 4.4 in [33]. See Appendix G.3 for details. □

**From sparse to dense feature learning.** Consider now a two layer neural network $f((a,w),x) := \sum_{j=1}^{N} a_j \sigma(w_j^T x)$. The previous two results can be reformulated as follows:

**Theorem 4.9.** *Let $R$ be the hyperbolic entropy ($\alpha = 1$) or a smoothened homogeneous mirror potential ($\alpha = p \geq 2$). Then the constrained optimization problems from Theorems 4.7 and 4.8 can reformulated as:*

$$\min_{\tilde{a}\in\mathbb{R}^N, \tilde{w}\in\mathbb{R}^{Nd}} \sum_{j=1} |\tilde{a}_j|^{\alpha/2} \qquad \textit{such that} \qquad \|\tilde{w}_j\|_{L_\alpha} = 1, \textit{ for all } j \in [N], \tag{5}$$

*where $\tilde{a}_j = a_j \|w_j\|_{L_\alpha}$ and $\tilde{w}_j = w_j / \|w_j\|_{L_\alpha}$, and such that $y_i f(\theta, x_i) \geq 1$, for all $i \in [K]$.*

*Proof sketch.* It follows from the balance equation for the parameter iterates and the rescale invariance of the two-layer neural network. The full proof is presented in Appendix G.5. □

Theorem 4.9 reveals that different values of the homogeneity degree $\alpha$ lead to different forms of feature learning. Specifically, for the hyperbolic entropy ($\alpha = 1$) mirror flow is biased towards networks with fewer active neurons, whereas for smoothened homogeneous potentials ($\alpha \geq 2$) it is biased towards networks with more active neurons. This is illustrated numerically in Figure 2. Note in particular that all the considered mirrors exhibit feature learning in the sense that the input weights of the neurons move significantly from their random initial positions.

**Reaching the margin.** The results presented so far characterize the implicit bias of mirror flow through the introduction of a $Q$-margin and its maximization in the late phase of training. Reaching the max-margin solution, determined by the (homogeneous) horizon function, critically depends on the rate of parameter growth. In particular, convergence to the max-margin direction can be extremely slow, in some cases even exponentially slow, as illustrated in Figure 1. As established in Theorems 4.7 and 4.8, the limiting direction is characterized by the horizon function $\phi_\alpha$. However, along the dynamics, it is the $Q$-margin that governs parameter growth. Thus it is of interest how quickly the $Q$-margin approximates the $\phi_\alpha$-margin. We characterize this next.

**Lemma 4.10.** *For the hyperbolic entropy and smoothened homogeneous mirror flow $R_\lambda : \mathbb{R}^n \to \mathbb{R}$, we have that once the time surpasses $\Omega(\exp((\sqrt{\lambda}n)^L))$ and $\Omega(\exp(\left(\frac{1}{2}p\lambda n\right)^{L/p}))$, respectively, the relative difference between $Q_\lambda(\nabla R_\lambda(\theta))$ and $\phi_1(\theta)$ is $O(1)$.*

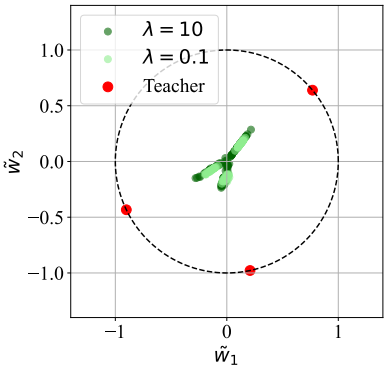


Figure 3: Training with smoothened homogeneous potential ($p = 3$). This shows that increasing $\lambda > 0$ slows convergence toward the max-margin solution, so that for finite training time the learned representation remains closer to that of gradient descent.

*Proof.* See Appendix H. □

Lemma 4.10 can be used to select training hyperparameters mitigating slow margin alignment. The takeaway is that the hyperparameter $\lambda$ should to be relatively small compared to $n$ if we want to reach the max-margin solution in sub-exponential time. For example, it is sufficient to choose $\lambda = 1/n^2$ for the hyperbolic entropy and $\lambda = 1/pn$ for the smoothened potentials. Note that otherwise it can take an exponential time in $n$ to reach the max-margin solution, effectively making it unreachable. This is shown in Figure 3, for smoothened homogeneous potential in a two layer student-teacher setting.

## 5 Experimental validation

We validate here our theoretical results. First we confirm Theorem 4.9 by considering a two-layer network trained in a student-teacher setup. Next, we test the mirror descent algorithms on a standard vision task, CIFAR10. The full training details and additional ablations are provided in Appendix J.

**Sparse and dense feature learning.** To illustrate Theorem 4.9, we consider a teacher network with 3 neurons together with two- and three-layer student networks of hidden dimension 100. The final parameters are shown in Figure 2. We train using (i) hyperbolic entropy, (ii) gradient descent, and (iii) a smoothened homogeneous mirror potential ($p = 3$). Observe that all 3 optimization procedures exhibit feature learning. However, as predicted by the theorem, smaller homogeneity degrees ($\alpha = 1$) produce solutions with fewer active neurons or weights, whereas larger values ($\alpha = 3$) produce solutions with more active neurons or weights. We refer to these regimes as sparse and dense feature learning. Furthermore, Figure 3 validates the dependence on $\lambda > 0$, confirming Lemma 4.10.

**Shaping representation structure in a vision task.** We train a VGG-16 [43] on CIFAR-10 [25], initialized in the lazy regime, using a hyperparameter sweep over $\lambda$[1] and the learning rate (details in Appendix J). Hyperbolic entropy achieves the best validation accuracy (Table 9 in Appendix J). The weight distributions in Figure 4 confirm the sparse and dense feature learning patterns observed in the student-teacher setting, and show that $\lambda$ must be small to alter the weight representation, consistent with Lemma 4.10. Furthermore, as a consequence of the changed weight distribution, pruning the network leads to reduced performance degradation for the hyperbolic entropy and increased degradation for the smoothened homogeneous potential. Moreover, we find that the weight distribution of the last layer remains unchanged (Figure 14 in Appendix J), consistent with training under standard parameterization (SP) [53].

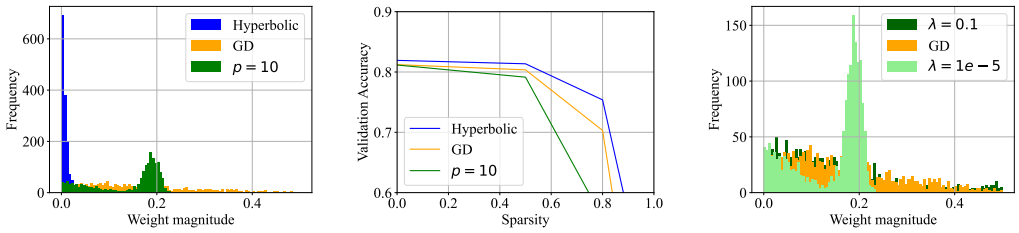


Figure 4: Weight distributions for the first layer of a VGG-16 and weight pruning. (Left) We show the distribution for 3 mirror maps leading to dense and sparse representations. (Middle) Validation accuracy versus sparsity, we can prune using layer-wise magnitude pruning more weights when training with the hyperbolic entropy and less with the homogeneous potential ($p = 10$) due to the change in weight distribution. (Right) Homogeneous mirror potential with $p = 10$ both for large and small $\lambda$, large $\lambda$ leads to a similar representation as gradient descent (GD), confirming Lemma 4.10.

## 6 Conclusion and future work

We established directional late-phase implicit bias results for a class of mirror descent algorithms in homogeneous neural networks. Our analysis is based on a novel balance equation derived via the Fenchel-Young identity. In addition, we show that the hyperparameter $\lambda > 0$ plays a critical role in whether the implicit bias induced by mirror flow is realized in practice, as large values of $\lambda$ can substantially slow convergence to the corresponding max-margin solution. Together, these results show how mirror geometry steers optimization toward distinct max-margin solutions and induces either sparse or dense feature learning depending on the chosen geometry.

In particular, the $L_1$ max-margin promotes sparsity by selecting a small subset of active parameters. In contrast, the $L_p$ max-margin for $p \geq 2$ encourages dense but homogeneous weight distributions, where weights concentrate around similar magnitudes. While the former aligns naturally with sparse feature selection, the latter may benefit quantization, as more uniform weight scales facilitate mapping parameters to discrete levels. Overall, our results suggest that mirror geometry provides a unified mechanism for steering models toward either sparse or quantization-friendly representations.

Finally, this raises several questions beyond the scope of the present work, including extending Theorem 4.7 to the $\alpha < 2$ regime (discussed in Appendix G.4), as well as understanding the effects of finite learning rate, early stopping, and the generalization properties of these max-margin solutions.

[1]Here $\lambda$ is layerwise rescaled correcting for the width of each layer as detailed in Appendix J.

**Acknowledgments**

This work was supported in part by the DFG project 464109215 within the Priority Programme SPP 2298 "Theoretical Foundations of Deep Learning". TJ has been supported by funding from the European Research Council (ERC) under the Horizon Europe Framework Programme (HORIZON) for proposal number 101116395 SPARSE-ML. GM has been supported in part by the DARPA AIQ grant HR00112520014, NSF grants DMS-2522495, DMS-2145630, CCF-2212520, and the BMFTR in DAAD project 57616814 (SECAI).

## References

[1] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

[3] Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *Forty-second International Conference on Machine Learning*, 2025.

[4] Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. In *Forty-second International Conference on Machine Learning*, 2025.

[5] Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: inducing sparsity via overparameterization. *Information and Inference*, 12(3):1437–1460, 2023.

[6] Hung-Hsu Chou, Holger Rauhut, and Rachel Ward. Robust implicit regularization via weight normalization. *Information and Inference: A Journal of the IMA*, 13(3):iaae022, 09 2024.

[7] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley-Interscience, 1983.

[8] Damek Davis, Dmitriy Drusvyatskiy, Sham M. Kakade, and J. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20:119–154, 2018.

[9] Clémentine Carla Juliette Dominé, Nicolas Anguita, Alexandra Maria Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

[10] A. F. Filippov. *Differential Equations with Discontinuous Right-Hand Sides*. Springer Dordrecht, 1988.

[11] Advait Gadhikar, Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Sign-in to the lottery: Reparameterizing sparse training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[12] Advait Harshal Gadhikar and Rebekka Burkholz. Masks, signs, and learning rate rewinding. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[14] Tjeerd Jan Heeringa, Len Spek, and Christoph Brune. Deep networks are reproducing kernel chains. *ArXiv*, abs/2501.03697, 2025.

[15] Tom Jacobs and Rebekka Burkholz. Mask in the mirror: Implicit sparsification. In *The Thirteenth International Conference on Learning Representations*, 2025.

[16] Tom Jacobs, Advait Gadhikar, Celia Rubio-Madrigal, and Rebekka Burkholz. Hyperbolic aware minimization: Implicit bias for sparsity. In *The Fourteenth International Conference on Learning Representations*, 2026.

[17] Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Mirror, mirror of the flow: How does regularization shape implicit bias? In *Forty-second International Conference on Machine Learning*, 2025.

[18] Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Never saddle: Reparameterized steepest descent as mirror flow. In *The Fourteenth International Conference on Learning Representations*, 2026.

[19] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, 2017.

[20] Aaron Alvarado Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *ArXiv*, abs/2410.14581, 2024.

[21] Kedar Karhadkar, Erin George, Michael Murray, Guido Montúfar, and Deanna Needell. Benign overfitting in leaky ReLU networks with moderate input dimension. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[22] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

[23] Chris Kolb, Laetitia Frost, Bernd Bischl, and David Rügamer. Differentiable sparsity via $d$-gating: Simple and versatile structured penalization, 2025.

[24] Chris Kolb, Tobias Weber, Bernd Bischl, and David Rügamer. Deep weight factorization: Sparse learning through the lens of artificial symmetries. In *The Thirteenth International Conference on Learning Representations*, 2025.

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[26] Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. *CoRR*, abs/2405.14813, 2024.

[27] Jiangyuan Li, Thanh V. Nguyen, Chinmay Hegde, and Raymond K. W. Wong. Implicit sparse regularization: The impact of depth and early stopping, 2021.

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

[29] Zhiyuan Li, Tianhao Wang, Jason D. Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. In *Advances in Neural Information Processing Systems*, 2022.

[30] Shuang Liang and Guido Montúfar. Implicit bias of mirror flow for shallow neural networks in univariate regression. In *The Thirteenth International Conference on Learning Representations*, 2025.

[31] Yannick Lunk, Sebastian James Scott, and Leon Bungert. Sparse training of neural networks based on multilevel mirror descent, 2026.

[32] Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *ICLR*, 2024.

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

[34] Negin Majidi, Ehsan Amid, Hossein Talebi, and Manfred K. Warmuth. Exponentiated gradient reweighting for robust training under label noise and beyond. *ArXiv*, abs/2104.01493, 2021.

[35] Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: conservation laws for gradient flows. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[36] Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Transformative or conservative? conservation laws for resnets and transformers. In *Forty-second International Conference on Machine Learning*, 2025.

[37] Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Keep the momentum: Conservation laws beyond Euclidean gradient flows, 2024.

[38] Pierre Marion and Lénaïc Chizat. Deep linear networks for regression are implicitly regularized towards flat minima. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[39] Chaewon Moon, Dongkuk Si, and Chulhee Yun. Minor first, major last: A depth-induced implicit bias of sharpness-aware minimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

[40] Scott Pesme, Radu-Alexandru Dragomir, and Nicolas Flammarion. Implicit bias of mirror flow on separable data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

[42] Pedro H. P. Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Annual Conference Computational Learning Theory*, 2019.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[44] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data, 2017.

[45] Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *ArXiv*, abs/2306.13853, 2023.

[46] Nikolaos Tsilivis, Gal Vardi, and Julia Kempe. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*, 2024.

[47] Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 48728–48767. PMLR, 21–27 Jul 2024.

[48] Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery, 2019.

[49] 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.

[50] Fan Wu and Patrick Rebeschini. Implicit regularization in matrix sensing via mirror descent. In *Neural Information Processing Systems*, 2021.

[51] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pages 74229–74256. Curran Associates, Inc., 2023.

[52] Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: $\ell_\infty$-norm constrained optimization. In *Forty-first International Conference on Machine Learning*, 2024.

[53] Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021.

[54] Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

## Appendix Contents

# A Extended related work

**Implicit regularization results.** Implicit bias of gradient-based optimization has been widely studied as an explanation for generalization in overparameterized models. Prior work shows that gradient descent induces structured solutions such as low-rank or sparse representations even without explicit regularization in regression settings [13, 1, 48, 27, 49]. Extensions analyze the role of depth, initialization, gradient noise, and training dynamics, including phase transitions such a grokking and alternative optimizers [28, 38, 2, 32, 52, 54, 39, 41]. More specifically for $\alpha = 2$, the optimization problem in Theorem 4.9 resembles the cost minimization problem for infinite width two layer neural networks studied in [42].

From a theoretical perspective, nonsmooth analysis provides the natural framework for studying subgradient and gradient flow dynamics in modern machine learning. Classical results on generalized gradients and differential inclusions [7, 10] formalize the behavior of optimization algorithms in the presence of nonsmooth objectives, and are particularly relevant for architectures involving nonsmooth components such as ReLU activations.

**Sparse training.** Mirror flow has been connected to reparameterized gradient flow [29] and sparse training using a time-varying mirror flow [15]. Sparsity has also been accomplished using deeper reparameterizations [24] and for group sparsity [23, 28]. These structures however do not correspond anymore to a mirror flow as they would violate conditions in [29]. Moreover, mirror flow can also be used to promote sign flips which have been shown to be important in sparse training [12, 11].

**Hyperparameter transfer.** Mirror descent introduces hyperparameters beyond the learning rate, such as the scaling of the mirror map [16, 50]. Principled transfer of hyperparameters across model scales has been studied in the tensor programs framework [53], which provides conditions under which learning rates and weight decay remain stable as the width grows. We extend this perspective to mirror descent, deriving hyperparameter transfer rules that ensure the corresponding max-margin solution remains reachable, with scaling that depends on the network width.

## B Proof of Lemma 3.1

*Proof of Lemma 3.1.* First note that the homogeneity of $f$ implies

$$\langle W_i, \nabla_i \mathcal{L}(\theta) \rangle_F = \langle W_{i+1}, \nabla_{i+1} \mathcal{L}(\theta) \rangle_F, \tag{6}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. Indeed, homogeneity of $f$ implies that the reparameterization

$$(W_i, W_{i+1}) \mapsto (e^s W_i, e^{-s} W_{i+1})$$

leaves $f_\theta$ and hence $\mathcal{L}(\theta)$ invariant. Thus differentiating at $s = 0$ yields

$$0 = \frac{d}{ds} \mathcal{L}(\theta(t))\Big|_{s=0} = \langle \nabla_i \mathcal{L}(\theta), W_i \rangle_F - \langle \nabla_{i+1} \mathcal{L}(\theta), W_{i+1} \rangle_F,$$

which implies

$$\langle W_i, \nabla_i \mathcal{L}(\theta) \rangle_F = \langle W_{i+1}, \nabla_{i+1} \mathcal{L}(\theta) \rangle_F.$$

Now let us consider the evolution of the proposed conserved quantity. Since $R$ is layer-wise separable, so is $Q$. We have

$$\begin{aligned} d\Big( Q_i(\nabla_i R_i(W_i)) - Q_{i+1}(\nabla_{i+1} R_{i+1}(W_{i+1})) \Big) &= \big\langle \nabla Q_i(\nabla_i R_i(W_i)),\, d\big(\nabla_i R_i(W_i)\big) \big\rangle \\ &\quad - \big\langle \nabla Q_{i+1}(\nabla_{i+1} R_{i+1}(W_{i+1})),\, d\big(\nabla_{i+1} R_{i+1}(W_{i+1})\big) \big\rangle . \end{aligned}$$

By the definition of $Q_i$ (as the Legendre dual of $R_i$), we have

$$\nabla Q_i(\nabla_i R_i(W_i)) = W_i, \qquad \nabla Q_{i+1}(\nabla_{i+1} R_{i+1}(W_{i+1})) = W_{i+1}.$$

Moreover, along gradient flow,

$$d\big(\nabla_i R_i(W_i)\big) = -\nabla_i \mathcal{L}(\theta)\, dt, \qquad d\big(\nabla_{i+1} R_{i+1}(W_{i+1})\big) = -\nabla_{i+1} \mathcal{L}(\theta)\, dt.$$

Substituting, we obtain

$$\begin{aligned} d\Big( Q_i(\nabla_i R_i(W_i)) - Q_{i+1}(\nabla_{i+1} R_{i+1}(W_{i+1})) \Big) &= -\langle W_i, \nabla_i \mathcal{L}(\theta) \rangle_F\, dt + \langle W_{i+1}, \nabla_{i+1} \mathcal{L}(\theta) \rangle_F\, dt \\ &= \Big( - \langle W_i, \nabla_i \mathcal{L}(\theta) \rangle_F + \langle W_{i+1}, \nabla_{i+1} \mathcal{L}(\theta) \rangle_F \Big) dt \\ &= 0, \end{aligned}$$

where the last equality follows from Eq. (6). This concludes the proof. □

## C Verification of Assumption 4.4

We provide explicit calculations for the considered mirror maps presented in Table 1 to verify that they satisfy Assumption 4.4. Note that since all mirror potentials are separable, we can verify the assumption pointwise. Note that both types of potentials are asymptotically homogeneous, with $\alpha = 1$ for the Hyperbolic entropy, i.e., $R(\theta) \sim \|\theta\|_1$ for $\|\theta\| \to \infty$, and $\alpha = p$ for the smoothened homogeneous potentials, as then $R(\theta) \sim \|\theta\|_p^p$ for $\|\theta\| \to \infty$.

**Hyperbolic entropy.** Recall the hyperbolic entropy with parameter $\lambda$:

$$R(\theta) = \theta \operatorname{arcsinh}\Big(\frac{\theta}{\sqrt{\lambda}}\Big) - \sqrt{\theta^2 + \lambda}.$$

Then we have to show that for all $\lambda \geq 0$:

$$Q(\nabla R(\theta)) \geq \|\theta\|^2_{\nabla^2 R}.$$

Plugging in the exact quantities:

$$\sqrt{\theta^2 + \lambda} \geq \frac{\theta^2}{\sqrt{\theta^2 + \lambda}},$$

where the denominator is always positive. Rearranging gives:

$$\theta^2 + \lambda \geq \theta^2,$$

which is true for all $\lambda \geq 0$, verifying the assumption.

**Smoothened homogeneous potentials.** Recall the smoothened homogeneous potentials with parameter $\lambda \geq 0$:

$$R(\theta) = \frac{1}{p}|\theta|^p + \frac{\lambda}{2}\theta^2.$$

Then we have to show that for all $\lambda \geq 0$:

$$pQ(\nabla R(\theta)) \geq \|\theta\|^2_{\nabla^2 R}.$$

Plugging in the exact quantities:

$$(p-1)|\theta|^p + \frac{p\lambda}{2}\theta^2 \geq (p-1)|\theta|^p + \lambda\theta^2.$$

Rearranging gives:

$$\frac{p\lambda}{2}\theta^2 \geq \lambda\theta^2,$$

which is true for all $\lambda \geq 0$ and $p \geq 2$, verifying the assumption. This also includes homogeneous potentials with $\lambda = 0$ and therefore gradient flow.

## D Proof of Theorem 4.5

The proof of Theorem 4.5 is divided into three lemmas. First we show that the soft-margin approximates the $Q$-margin as stated in the main text in Lemma 4.3.

*Proof of Lemma 4.3.* Observe that $e^{a_{\max}} \leq \sum_{i=1}^{K} e^{a_i} \leq K e^{a_{\max}}$ holds for $a_{\max} = \max\{a_1, \ldots, a_K\}$. Applying the log we have $a_{\max} \leq \text{LSE}(a_1, \ldots, a_K) \leq a_{\max} + \log K$. Now using the definitions of both $\tilde{\gamma}$ and $\gamma_Q$ gives $\gamma_Q(\theta) - \log K \ \left(\alpha Q(\nabla(R(\theta)))^{-L/\alpha}\right) \leq \tilde{\gamma}_Q(\theta) \leq \gamma_Q(\theta)$. Note that this proof does not rely on the particular definition of $Q$. □

Next, for Lemma D.1 below we want to utilize the following inequality as in [33]:

$$-\langle \theta_t, g_t \rangle_2 = \langle \theta_t, \sum_{i=1}^{K} e^{-q_i} \partial^\circ q_i \rangle_2 = L \sum_{i=1}^{K} e^{-q_i} q_i \geq L \sum_{i=1}^{K} e^{-q_i} q_{\min} \geq L \mathcal{L}(\theta_t) \log\big(1/\mathcal{L}(\theta_t)\big), \quad (7)$$

where the last inequality follows from $e^{-q_{\min}} \leq \mathcal{L}$.

**Lemma D.1.** *If the mirror potential $R$ satisfies Assumption 4.4 and $\mathcal{L}(\theta_{t_0}) < 1$, then the $Q$-soft margin satisfies:*

$$\frac{d}{dt} \log Q(\nabla R(\theta_t)) > 0 \qquad \textit{and} \qquad \frac{d}{dt} \log \tilde{\gamma} \geq L \left( \frac{d}{dt} \log Q(\nabla R(\theta_t)) \right)^{-1} E_{tan} \geq 0,$$

*where* $E_{tan} := \frac{1}{Q} \left( \| \frac{d}{dt} \theta_t \|^2_{\nabla^2 R} - \frac{\langle \theta, \nabla^2 R(\theta_t) \frac{d}{dt} \theta_t \rangle^2}{\|\theta\|^2_{\nabla^2 R}} \right)$.

*Proof.* We have

$$\begin{aligned}
\frac{d}{dt} \log \tilde{\gamma}_t &= \frac{d}{dt} \log \frac{\log(\frac{1}{\mathcal{L}(\theta_t)})}{(\alpha Q(\nabla(R(\theta_t)))^{L/\alpha}} \\
&= \frac{d}{dt} \left( \log \log(1/\mathcal{L}(\theta_t)) - \frac{L}{\alpha} \log(\alpha Q(\nabla(R(\theta_t))) \right) \\
&= \frac{-1}{\mathcal{L}(\theta_t) \log(1/\mathcal{L}(\theta_t))} \frac{d}{dt} \mathcal{L}(\theta_t) - \frac{L}{\alpha} \frac{1}{Q(\nabla(R(\theta_t))} \langle \theta_t, \frac{d \nabla R(\theta_t)}{dt} \rangle
\end{aligned}$$

Now define $\nu_t := \mathcal{L}(\theta_t) \log(1/\mathcal{L}(\theta_t))$, and note that $\nu_t \to 0$ if $\mathcal{L}(\theta_t) \to 0$. We use the same idea as in [33, Lemma 5.1] where the energy can be split into radial and tangential parts. We can use that $\frac{d}{dt} \mathcal{L}(\theta_t) = -\| \frac{d}{dt} \theta_t \|^2_{\nabla^2 R}$ and $\dot{Q} = \langle \theta_t, \frac{d \nabla R(\theta_t)}{dt} \rangle \geq L \nu_t > 0$ to get:

$$\begin{aligned}
\frac{1}{\nu_t} \| \frac{d}{dt} \theta_t \|^2_{\nabla^2 R} - \frac{L}{\alpha} \frac{1}{Q} \dot{Q} &= \frac{Q}{\nu_t}(E_{tan}) + \frac{1}{\nu_t} \frac{\langle \theta, \frac{d}{dt} \nabla R(\theta) \rangle^2}{\|\theta\|^2_{\nabla^2 R}} - \frac{L}{\alpha} \frac{\dot{Q}}{Q} \\
&= \frac{Q}{\nu_t}(E_{tan}) + \frac{1}{\nu_t} \frac{\dot{Q}^2}{\|\theta\|^2_{\nabla^2 R}} - \frac{L}{\alpha} \frac{\dot{Q}}{Q} \\
&= \frac{Q}{\nu_t}(E_{tan}) + \frac{\dot{Q}}{\alpha Q} \left( \frac{\alpha Q \dot{Q}}{\nu_t \|\theta\|^2_{\nabla^2 R}} - L \right) \\
&\geq \frac{Q}{\nu_t}(E_{tan}) + \frac{L \dot{Q}}{\alpha Q} \left( \frac{\alpha Q}{\|\theta\|^2_{\nabla^2 R}} - 1 \right) \\
&\geq \frac{Q}{\nu_t}(E_{tan}) \\
&\geq \frac{LQ}{\dot{Q}}(E_{tan}) \\
&= L \left( \frac{d}{dt} \log Q(\nabla R(\theta_t)) \right)^{-1} E_{\text{tan}} \\
&\geq 0.
\end{aligned}$$

Here the main step was to use Assumption 4.4. Note that the tangential part $E_{tan}$ is always non-negative. This can be seen from the fact that $Q$ is non-negative and non-decreasing and using the identity $\frac{d}{dt}\nabla R(\theta_t) = \nabla^2 R(\theta_t)\frac{d}{dt}\theta_t$ and the Cauchy-Schwarz inequality to bound:

$$\begin{aligned} QE_{tan} &:= \|\tfrac{d}{dt}\theta_t\|^2_{\nabla^2 R} - \frac{\langle \theta, \frac{d}{dt}\nabla R(\theta_t)\rangle^2}{\|\theta\|^2_{\nabla^2 R}} \\ &= \|\tfrac{d}{dt}\theta_t\|^2_{\nabla^2 R} - \frac{\langle \theta, \nabla^2 R(\theta_t)\frac{d}{dt}\theta_t\rangle^2}{\|\theta\|^2_{\nabla^2 R}} \\ &\geq \|\tfrac{d}{dt}\theta_t\|^2_{\nabla^2 R} - \|\tfrac{d}{dt}\theta_t\|^2_{\nabla^2 R} \\ &= 0. \end{aligned}$$

Moreover, the fact that $\frac{d}{dt}Q > 0$ implies that $\frac{d}{dt}\log Q > 0$, which concludes the result. □

It remains to show that the loss goes to zero and consequently that the iterates diverge, as established in the next Lemma D.2.

**Lemma D.2.** *For all* $t \geq 0$

$$G(\frac{1}{\mathcal{L}(\theta_t)}) \geq \frac{L^2}{\alpha}\tilde{\gamma}(t_0)^{\alpha/L}(t - t_0) \qquad \textit{for } G(x) := \int_{1/\mathcal{L}(t_0)}^{x} (\log(u))^{\frac{\alpha}{L}-2}\, du.$$

*The loss* $\mathcal{L}(\theta_t) \to 0$ *and* $Q(\nabla R(\theta_t)) \to \infty$ *as* $t \to \infty$.

*Proof.* We have

$$\begin{aligned} -\frac{d\mathcal{L}(\theta_t)}{dt} &= \|\frac{d}{dt}\theta_t\|^2_{\nabla^2 R} \geq \langle \frac{\theta_t}{\|\theta_t\|_{\nabla^2 R}}, \frac{d}{dt}\theta_t\rangle^2_{\nabla^2 R} \\ &\geq \frac{1}{\alpha Q(\nabla R(\theta_t))}\langle \theta_t, \frac{d}{dt}\theta_t\rangle^2_{\nabla^2 R} \\ &\geq \frac{L^2}{\alpha}\frac{\nu_t^2}{Q(\nabla R(\theta_t))}, \end{aligned}$$

where we use the CS inequality, Assumption 4.4 and the definition of $L\nu$. Note this exactly matches the lower bound in Lemma B.6 of [33] in the case of gradient descent. We replace $\alpha Q(\nabla R(\theta_t))$ with $(\log(\frac{1}{L(\theta_t)})/\tilde{\gamma}_Q)^{\alpha/L}$ to get an expression in terms of the $Q$-soft margin and the loss:

$$\begin{aligned} -\frac{d\mathcal{L}(\theta_t)}{dt} &\geq \frac{L^2}{\alpha}\left(\mathcal{L}(\theta_t)\log(\frac{1}{\mathcal{L}(\theta_t)})\right)^2 \left(\tilde{\gamma}_Q(t)/\log(\frac{1}{\mathcal{L}(\theta_t)})\right)^{\alpha/L} \\ &\geq \frac{L^2}{\alpha}\mathcal{L}(\theta_t)^2\left(\log(\frac{1}{\mathcal{L}(\theta_t)})\right)^{2-\frac{\alpha}{L}} \tilde{\gamma}_Q(t_0)^{\alpha/L}, \end{aligned}$$

where the last inequality follows from monotonicity of the $Q-$soft margin. So the following holds for a.e. $t \geq t_0$:

$$\left(\log(\frac{1}{\mathcal{L}(\theta_t)})\right)^{\frac{\alpha}{L}-2} \cdot \frac{d}{dt}\frac{1}{\mathcal{L}(\theta_t)} \geq \frac{L^2}{\alpha}\tilde{\gamma}(t_0)^{\alpha/L}.$$

Integrating on both sides from $t_0$ to $t$, we can conclude that

$$G(\frac{1}{\mathcal{L}(\theta_t)}) \geq L^2\tilde{\gamma}(t_0)^{\alpha/L}(t - t_0).$$

Note that $1/\mathcal{L}(\theta_t)$ is non-decreasing. If $1/\mathcal{L}(\theta_t)$ does not grow to $+\infty$, then neither does $G(1/\mathcal{L}(\theta_t))$. But the RHS grows to $+\infty$, which leads to a contradiction. So $\mathcal{L}(\theta_t) \to 0$. To make $\mathcal{L}(\theta_t) \to 0$, $q_{\min} := \min_i y_i f(\theta, x_i)$ must converge to $+\infty$. So $Q(\nabla R(\theta_t)) \to \infty$. □

*Proof of Theorem 4.5.* The result now follows directly from Lemmas D.1 and D.2 combined with Lemma 4.3. □

## E Bounds on the normalized dual iterates

The following two lemmas help us control the dual iterates, which in turn allows us to show convergence to KKT points in Appendix G.2. For Lemma E.1, we introduce the constants $B_0$ and $B_1$. As in Section C.3 of [33], the function $q_n$ is locally Lipschitz and therefore Lipschitz on a compact set such that

$$B_0 := \sup\{\frac{q_i}{(\alpha Q(\nabla R(\theta)))^{L/\alpha}} : \theta \in \mathbb{R}^n \setminus \{0\}, h \in \partial^\circ q_i, i \in [K]\} < \infty$$

and

$$B_1 := \sup\{\frac{\|h\|_2}{(\alpha Q(\nabla R(\theta)))^{(L-1)/\alpha}} : \theta \in \mathbb{R}^n \setminus \{0\}, h \in \partial^\circ q_i, i \in [K]\} < \infty.$$

**Lemma E.1.** *The normalized dual derivative is bounded such that we have for a.e. $t > t_0$:*

$$\|\frac{d}{dt}\frac{\nabla R(\theta_t)}{\|\nabla R(\theta_t)\|_2}\|_2 \leq \frac{2B_1\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t))\rangle}{L\tilde{\gamma}_Q C_2 \|\theta\|_2 \|\nabla R(\theta_t)\|_2},$$

*where $C_2 > 0$ is the constant such that $(\alpha Q(\nabla R(\theta)))^{1/\alpha} \geq C_2\|\theta\|_2$ for all $\theta \in \mathbb{R}^n$.*

*Proof.* By Cauchy-Schwarz:

$$\begin{aligned}\|\frac{d}{dt}\frac{\nabla R(\theta_t)}{\|\nabla R(\theta_t)\|_2}\|_2 &= \|\frac{\nabla^2 R(\theta_t)\frac{d}{dt}\theta_t}{\|\nabla R(\theta_t)\|} - \frac{\nabla R(\theta_t)\langle \nabla R(\theta_t), \nabla^2 R(\theta_t)\frac{d}{dt}\theta_t\rangle}{\|\nabla R(\theta_t)\|^3}\|_2 \\ &\leq 2\frac{\|\nabla^2 R(\theta_t)\frac{d}{dt}\theta_t\|}{\|\nabla R(\theta_t)\|_2}.\end{aligned}$$

We now use the definition of the mirror flow and a bound for the gradient. By the chain rule there exists $h_1, \ldots, h_K : [0, \infty) \to \mathbb{R}^n$ satisfying that for a.e. $t > 0$ $h_{i,t} \in \partial^\circ q_n$ and $\nabla^2 R(\theta_t)\frac{d\theta_t}{dt} = \sum_{i=1}^K e^{-q_i} h_{i,t}$. Now using $B_1$:

$$\begin{aligned}\|\nabla^2 R(\theta_t)\frac{d}{dt}\theta_t\|_2 &\leq \sum_{i=1}^K e^{-q_n}\|h_i\|_2 \\ &\leq \sum_{i=1}^K e^{-q_i} q_i \frac{1}{q_i} B_1 (\alpha Q)^{\frac{L-1}{\alpha}}.\end{aligned}$$

We can use that $q_i \geq q_{\min} \geq \log\frac{1}{\mathcal{L}}$ and that $\sum_{i=1}^K e^{-q_i} q_i = \frac{1}{L}\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t))\rangle$:

$$\begin{aligned}\|\nabla^2 R(\theta_t)\frac{d}{dt}\theta_t\|_2 &\leq \frac{B_1\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t))\rangle}{L\log 1/\mathcal{L}}(\alpha Q)^{\frac{L-1}{\alpha}} \\ &\leq \frac{B_1\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t))\rangle}{L\tilde{\gamma}_Q(\alpha Q)^{1/\alpha}}.\end{aligned}$$

Next we can put this together giving and using that there exists a constant $C_2$ such that $(\alpha Q)^{1/\alpha} \geq C_2\|\theta\|_2$ for all $\theta \in \mathbb{R}^n$:

$$\|\frac{d}{dt}\frac{\nabla R(\theta_t)}{\|\nabla R(\theta_t)\|_2}\|_2 \leq \frac{2B_1\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t))\rangle}{L\tilde{\gamma}_Q(\alpha Q)^{1/\alpha}\|\nabla R(\theta_t)\|_2} \leq \frac{2B_1\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t))\rangle}{L\tilde{\gamma}_Q C_2\|\theta\|_2\|\nabla R(\theta_t)\|_2}$$

This concludes the proof. □

**Lemma E.2.** *There exists a time $t_1$ such that for a.e. $t > t_1$:*

$$\frac{d}{dt}\log(Q(\nabla R(\theta_t))) \geq \frac{\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t)\rangle}{\|\theta_t\|\|\nabla R(\theta_t)\|}.$$

*Proof.* Since $Q \to \infty$ we have that $R \to \infty$ by definition of the mirror map. Therefore, there exists a time $t_1$ such that for all $t > t_1$ $R > 0$. Therefore, we have for all $t > t_1$:

$$\frac{d}{dt}\log(Q(\nabla R(\theta_t))) = \frac{\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t)\rangle}{Q} \geq \frac{\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t)\rangle}{\langle \theta_t, \nabla R(\theta_t)\rangle} \geq \frac{\langle \theta_t, \frac{d}{dt}\nabla R(\theta_t)\rangle}{\|\theta_t\|\|\nabla R(\theta_t)\|}.$$

Note from Lemma E.1 we know that the right hand side is greater than zero. □

## F Proof of Theorem 4.6

This proof relies on the characterization of $G$ as in Lemma D.2.

**Lemma F.1.** *For $G$ defined in Lemma D.2 and its inverse, we have the following bounds:*

$$G(x) = \Theta\left((log\ x)^{\alpha/L-2}x\right) \text{ and } G^{-1}(y) = \Theta\left((log\ x)^{2-\alpha/L}x\right),$$

*where $x = G^{-1}(y)$.*

*Proof.* Denote $\beta := \frac{\alpha}{L} - 2$ By $u = e^z$ substitution and iterative integration by parts after we get:

$$\begin{aligned}
\int_{1/\mathcal{L}(t_0)}^{x} (\log(u))^{\beta}\, du &= \int^{\log(x)} z^{\beta} e^z dz \\
&= \log(x)^{\beta} x - \beta \int^{\log x} z^{\beta-1} e^z dz \\
&= \log(x)^{\beta} x \left(1 - \frac{\beta}{\log(x)} + ...\right) \\
&= \Theta\left((\log\ x)^{\alpha/L-2}x\right)
\end{aligned}$$

Now for $G^{-1}(y)$, let $x = G^{-1}(y)$ for $y \geq 0$. $G(x)$ is finite for $x$ finite, thus if $x \to \infty$ we have $y \to \infty$. From the first part we get: $y = \Theta\left((\log\ x)^{\alpha/L-2}x\right)$. Taking logarithms on each side gives: $\log\ y = \Theta\left(\log\ x\right)$. Therefore, we have $x = \Theta\left((\log\ x)^{2-\alpha/L}x\right)$. □

Now we can prove Theorem 4.6.

*Proof of Theorem 4.6.* **Upper bounding $\mathcal{L}$.** It follows from Lemma D.2 that we have $\frac{1}{\mathcal{L}} \geq G^{-1}(\Omega(t))$. Using Lemma F.1 we have $\frac{1}{\mathcal{L}} = \Omega\left((\log\ t)^{2-\alpha/L}t\right)$.

**Bounding $(\alpha Q)^{1/\alpha}$ in terms of $\mathcal{L}$.** We have $\tilde{\gamma}_Q(t) \geq \tilde{\gamma}_Q(t_0)$, so $(\alpha Q)^{L/\alpha} \leq \frac{1}{\tilde{\gamma}_Q(t_0)} \log\frac{1}{\mathcal{L}}$. On the other hand, $\log\frac{1}{\mathcal{L}} \leq q_{\min} \leq B_0(\alpha Q)^{L/\alpha}$, so $(\alpha Q)^{L/\alpha} = \Omega(\log\frac{1}{\mathcal{L}})$. This implies that we have:

$$(\alpha Q)^{L/\alpha} = \Theta(\log\frac{1}{\mathcal{L}}). \tag{8}$$

**Lower bounding $\mathcal{L}$.** Let $h_1, \ldots, h_K$ be a set of vectors such that $h_i \in \frac{\partial q_i}{\partial \theta}$ and

$$\nabla^2 R(\theta)\frac{d\theta}{dt} = \sum_{i=1}^{K} e^{-q_i} h_i.$$

We have that $\|h_i\|_2 \leq B_1 Q^{\frac{L-1}{\alpha}} = O((\log\frac{1}{\mathcal{L}})^{1-1/L})$. Note that we have $|\partial_i^2 R^{-1}(\theta)| = O(\max\{1, (\log\frac{1}{\mathcal{L}})^{(2-\alpha)/L}\})$ by asymptotic $\alpha$-homogeneity and separability of the potential $R$. Thus:

$$\begin{aligned}
-\frac{d\mathcal{L}}{dt} &= \|\frac{d\theta}{dt}\|^2_{\nabla^2 R} = \langle \nabla^2 R(\theta)\frac{d\theta}{dt}, \nabla^2 R^{-1}(\theta)\nabla^2 R(\theta)\frac{d\theta}{dt}\rangle_2 \leq \max_{i\in[K]} |\nabla^2 R^{-1}(\theta)|\ \|\nabla^2 R(\theta)\frac{d\theta}{dt}\|_2^2 \\
&\leq O(\max\{1, (\log\frac{1}{\mathcal{L}})^{(2-\alpha)/L}\})\|\sum_{i=1}^{K} e^{-q_i} h_i\|_2^2 \\
&\leq O(\max\{1, (\log\frac{1}{\mathcal{L}})^{(2-\alpha)/L}\}) \left(\sum_{i=1}^{K} e^{-q_i} \max_{i\in[K]} \|h_i\|_2\right)^2 \\
&\leq O(\max\{1, (\log\frac{1}{\mathcal{L}})^{(2-\alpha)/L}\})\mathcal{L}^2 O((\log\frac{1}{\mathcal{L}})^{2-2/L})) = \mathcal{L}^2 O((\log\frac{1}{\mathcal{L}})^{2-\min\{2,\alpha\}/L})).
\end{aligned}$$

Now if $\alpha \leq 2$, rearranging and with the definition of $G$, there exist a constant $c$ such that $\frac{d}{dt}G(\frac{1}{\mathcal{L}}) \leq c$ for any $\mathcal{L}$ that is small enough. The proof follows directly from applying Lemma F.1.

**Bounding $(\alpha Q)^{1/\alpha}$ in terms of $t$.** By Eq. (8) and the tight bounds for $\mathcal{L}$ for $\alpha \leq 2$ we have that $(\alpha Q)^{L/\alpha} = \Theta(\log \frac{1}{\mathcal{L}}) = \Theta(\log t)$. For $\alpha > 2$ we have that the lowerbound on $\mathcal{L}$ fails, however, we still have an upperbound for $\mathcal{L}$ giving us $(\alpha Q)^{L/\alpha} = \Theta(\log \frac{1}{\mathcal{L}}) = \Omega(\log t)$. □

# G Proof of Theorem 4.7

Here we show the KKT conditions for the Homogeneous potentials with $\alpha = p \geq 2$. We can be more general and replace it by the following condition:

**Assumption G.1.** For $R$ and corresponding $\phi_\alpha$ there exists a fixed constant $C > 0$ such that we have when $\eta \sim \|\theta\| \to \infty$:

$$\|\partial^\circ \phi_\alpha^2(\theta)/\eta - C(\alpha Q(\nabla R(\theta)))^{2/\alpha - 1} \nabla^2 R(\theta)\theta/\eta\|_2^2 \to 0.$$

Assumption G.1 is satisfied by the homogeneous potentials but not by the hyperbolic entropy as detailed in Appendix G.4.

Now we want to construct similarly as in [46, 33] a limiting sequence of KKT points. For this we first recall the definitions of KKT points, approximate KKT points and constraint qualification. Next, we bound the approximate KKT points by the "angle" of the iterates with respect to the evolution and denoted by $\beta_t$ in Lemma G.6. We can then control the angle decay in terms of the iterate growth in Corollary G.8. Finally, we can construct the sequence of KKT points for which the angle decays.

In this section, we will show the following result, which covers and extends Theorem 4.7.

**Theorem G.2.** *Assume that $R$ satisfies Assumption G.1 and $\alpha \geq 2$, under the same assumptions as in Theorem 4.5, the limit points $\bar{\theta}$ of $\{\frac{\theta_t}{\|\theta_t\|_2}, t > 0\}$ converge in direction to a solution to the following optimization problem:*

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\phi_\alpha^2(\theta) \qquad \textit{such that} \qquad y_i f(\theta, x_i) \geq 1, \textit{ for all } i \in [K], \tag{9}$$

*where $\phi_\alpha := \lim_{\eta \to 0} \eta(\alpha Q(\nabla R(\theta/\eta)))^{1/\alpha}$ is the horizon function.*

### G.1 KKT conditions

To show the result we first need to state what the corresponding optimality conditions i.e. Karush Kuhn Tucker (KKT). The main idea is to show is to create a sequence of appproximate KKT points which under the Mangasarian-Fromovitz Constraint Qualification (MFCQ) implies the that limit is a KKT point.

**Definition G.3.** (KKT point) A feasible point $\theta$ of Eq.(9) is a KKT point if there exists $\lambda_1, \ldots, \lambda_K \geq 0$ such that

- $\partial^\circ \frac{1}{2}\phi_\alpha^2(\theta) - \sum_{i=1}^K \lambda_i h_i$ for some $h_1, \ldots, h_K$ satisfying $h_i \in \partial^\circ q_i(\theta)$.
- For all $i \in [K]$: $\lambda_i(q_i(\theta) - 1) = 0$.

**Definition G.4.** (Approximate KKT point) A feasible point $\theta$ of Eq. (9) is an $(\epsilon, \delta)$-KKT point of Eq. (9) if there exists $\lambda_1, \ldots, \lambda_K \geq 0$ such that

- $\|\partial^\circ \frac{1}{2}\phi_\alpha^2(\theta) - \sum_{i=1}^K \lambda_i h_i\|_2 \leq \epsilon$ for some $h_1, \ldots, h_K$ satisfying $h_i \in \partial^\circ$.
- For all $i \in [K]$: $\lambda_i(q_i(\theta) - 1) \leq \delta$.

**Lemma G.5.** *Eq. (9) satisfies the MFCQ at every feasible point $\theta$.*

*Proof.* See Lemma C.7 in [33]. □

### G.2 Key lemmas

In order to show that the sequence of KKT points converges we need to show that the angle quantified by the constant $\beta_t$ goes to 1:

$$\beta_t := \frac{1}{\|\frac{d\theta_t}{dt}\|_{\nabla^2 R}} \langle \frac{\theta_t}{\|\theta_t\|_{\nabla^2 R}}, \nabla^2 R(\theta)\frac{d\theta_t}{dt} \rangle \leq 1.$$

**Lemma G.6.** *Let $C_1$ and $C_2$ be two constants defined as:*

$$C_1 := \sqrt{\frac{2}{\tilde{\gamma}_Q^{2/L}}\hat{\mu}}, \qquad C_2 := \frac{\tilde{\gamma}_Q^{-2/L}}{L} K/e.$$

*where $\hat{\mu}$ is a growth constant depending on R. Then $\tilde{\theta} := \theta/q_{min}(\theta)^{1/L}$ is an $(\epsilon, \delta)-$KKT point of Eq. (9) where $\epsilon := C_1\sqrt{1-\beta}$ and $\delta := C_2/\log(1/\mathcal{L})$.*

*Proof.* For the first condition we approximate the horizon function, then we show that the approximation can be bounded by the angle $\beta_t$. Notice that by construction the horizon function is 1-homogeneous this gives us:

$$\partial^\circ \frac{1}{2}\phi_\alpha^2(\theta/q_{\min}(\theta)^{1/L}) = \partial^\circ \frac{1}{2}\phi_\alpha^2(\theta)/q_{\min}(\theta)^{1/L}.$$

This for $\alpha \geq 2$ then is well approximated by $(\alpha Q(\nabla R(\theta)))^{2/\alpha-1}\nabla^2 R(\theta)\theta/q_{\min}(\theta)^{1/L}$ as $\|\theta\| \to \infty$ by Assumption G.1 and the fact that $q_{\min} \sim \|\theta\|$ as $\|\theta\| \to \infty$.

Now, denote $h_t := \frac{d\theta_t}{dt}$ for a.e. $t > 0$. Then, by the chain rule, there exists $h_1, \ldots, h_N$ such that $h_i \in \partial^\circ q_i$ and $h = \nabla^2 R^{-1}(\theta_t)\sum_{i=1}^N e^{-q_i}h_i$. Denote $\tilde{h}_i := h_i/q_{\min}^{1-1/L}$ and $\zeta := (\alpha Q(\nabla R(\theta)))^{2/\alpha-1}\nabla^2 R(\theta)\theta/q_{\min}^{1/L}$. Construct $\lambda_i = (\alpha Q(\nabla R(\theta)))^{2/\alpha-1} q_{\min}^{1-2/L}\|\theta\|_{\nabla^2 R} e^{-q_i}/\|h\|_{\nabla^2 R}$. Then

$$\begin{aligned}
\|\zeta - \sum_i \lambda_i \tilde{h}_i\|_2^2 &= (\alpha Q(\nabla R(\theta)))^{4/\alpha-2}/q_{\min}^{2/L}\left\|\nabla^2 R(\theta)\theta - \frac{\|\theta\|_{\nabla^2 R}}{\|h\|_{\nabla^2 R}}\nabla^2 R(\theta)h\right\|_2^2 \\
&\leq (\alpha Q(\nabla R(\theta)))^{4/\alpha-1}/q_{\min}^{2/L}\|\nabla^2 R(\theta)\theta/\|\theta\|_{\nabla^2 R} - \frac{1}{\|h\|_{\nabla^2 R}}\nabla^2 R(\theta)h\|_2^2 \\
&\leq \frac{1}{\tilde{\gamma}_Q^{2/L}}/(\alpha Q(\nabla R(\theta)))^{1-2/\alpha}\left\|\nabla^2 R(\theta)\theta/\|\theta\|_{\nabla^2 R} - \frac{1}{\|h\|_{\nabla^2 R}}\nabla^2 R(\theta)h\right\|_{\nabla^2 R(\theta)}^2 \\
&\leq \frac{1}{\tilde{\gamma}_Q^{2/L}}\mu/(\alpha Q(\nabla R(\theta)))^{1-2/\alpha}\left\|\theta/\|\theta\|_{\nabla^2 R} - \frac{1}{\|h\|_{\nabla^2 R}}h\right\|_{\nabla^2 R(\theta)}^2 \\
&\leq \frac{2}{\tilde{\gamma}_Q^{2/L}}\hat{\mu}(1-\beta),
\end{aligned}$$

where we used the bound $\mu := (\max_i(1, |\nabla^2 R(\theta_i)|))$ and $\hat{\mu} = \mu Q^{\frac{2-\alpha}{\alpha}} < \infty$, when $\alpha \geq 2$.

Similarly we can bound the other quantity with our construction and homogeneity we have (note that $q_i = y_i f(\theta, x_i)$):

$$\sum_{i=1}^K \lambda_i(q_i(\tilde{\theta}) - 1) = \frac{q_{\min}^{-2/L}(\alpha Q)^{2/\alpha-1}\|\theta\|_{\nabla^2 R}}{\|h\|_{\nabla^2 R}}\sum_{i=1}^K e^{-q_i}(q_i - q_{\min}).$$

Now we can use that $\|h\|_{\nabla^2 R} \geq \langle h, \frac{\theta}{\|\theta\|_{\nabla^2 R}}\rangle_{\nabla^2 R} \geq L\nu/\|\theta\|_{\nabla^2 R}$ and note that

$$\nu = \mathcal{L}\,\log(1/\mathcal{L}) \geq e^{-q_{\min}}\log(1/\mathcal{L}) = e^{-q_{\min}}\tilde{\gamma}_Q(\alpha Q)^{L/\alpha}.$$

Combining this gives:

$$\begin{aligned}
\sum_{i=1}^K \lambda_i(q_i(\tilde{\theta}) - 1) &\leq \frac{q_{\min}^{-2/L}(\alpha Q)^{2/\alpha}}{L\tilde{\gamma}_Q(\alpha Q)^{(L)/\alpha}}\sum_{i=1}^K e^{-(q_i - q_{\min})}(q_i - q_{\min}) \\
&\leq \frac{\tilde{\gamma}_Q^{-2/L}}{L\tilde{\gamma}_Q(\alpha Q)^{(L)/\alpha}}\sum_{i=1}^K e^{-(q_i - q_{\min})}(q_i - q_{\min}) \\
&= \frac{\tilde{\gamma}_Q^{-2/L}}{L\tilde{\gamma}_Q(\alpha Q)^{(L)/\alpha}}K/e \\
&= \frac{\tilde{\gamma}_Q^{-2/L}}{L}K/e \cdot \frac{1}{\log 1/\mathcal{L}},
\end{aligned}$$

where we used that $e^{-x}x$ has its maximum $e^{-1}$ at $x = 1$. This concludes the proof. □

**Lemma G.7.** *For all $t_2 > t_1 \geq t_0$,*

$$\int_{t_1}^{t_2} \left(\beta_\tau^{-2} - 1\right) \frac{d}{d\tau} \log Q(\nabla R(\theta_\tau)) d\tau \leq \frac{\alpha}{L} \log \frac{\tilde{\gamma}_Q(t_2)}{\tilde{\gamma}_Q(t_1)}.$$

*Proof.* From Lemma D.1 we have that for all $t \in (t_1, t_2)$:

$$\frac{d}{dt} \log \tilde{\gamma} \geq L \left( \frac{d}{dt} \log Q(\nabla R(\theta_t)) \right) Q^2 / \dot{Q}^2 E_{tan}$$

We can now lowerbound $Q^2/\dot{Q}^2 E_{tan}$ by using the definitions of $E_{\text{tan}}$ and $\dot{Q} =$

$$\begin{aligned} Q^2/\dot{Q}^2 E_{tan} &= \frac{Q}{\langle \theta, \nabla^2 R(\theta_t) \frac{d}{dt}\theta_t \rangle^2} \left( \| \frac{d}{dt}\theta_t \|_{\nabla^2 R}^2 - \frac{\langle \theta, \nabla^2 R(\theta_t) \frac{d}{dt}\theta_t \rangle^2}{\|\theta\|_{\nabla^2 R}^2} \right) \\ &\geq \frac{1}{\alpha} \frac{\|\theta\|_{\nabla^2 R}^2}{\langle \theta, \nabla^2 R(\theta_t) \frac{d}{dt}\theta_t \rangle^2} \left( \| \frac{d}{dt}\theta_t \|_{\nabla^2 R}^2 - \frac{\langle \theta, \nabla^2 R(\theta_t) \frac{d}{dt}\theta_t \rangle^2}{\|\theta\|_{\nabla^2 R}^2} \right) \\ &= \frac{1}{\alpha} \left( \beta^{-2} - 1 \right). \end{aligned}$$

Together with Lemma D.1 we have:

$$\frac{d}{dt} \log \tilde{\gamma}_Q \geq \frac{L}{\alpha} \left( \beta^{-2} - 1 \right) \frac{d}{dt} \log Q.$$

Integrating both sides concludes the result. □

**Corollary G.8.** *For all $t_2 > t_1 \geq t_0$, then there exists $t_* \in (t_1, t_2)$ such that:*

$$\beta_{t_*}^{-2} - 1 \leq \frac{\alpha}{L} \cdot \frac{\log \tilde{\gamma}_Q(t_2) - \log \tilde{\gamma}_Q(t_1)}{\log Q(\nabla R(\theta_{t_2})) - \log Q(\nabla R(\theta_{t_1}))}$$

*Proof.* We follow the exact same steps as [33, Corollary C.10]. Denote the RHS as $C$. Assume the opposite is true i.e. $\beta_\tau^{-2} - 1 > C$ for a.e. $\tau \in (t_1, t_2)$. From Lemma D.1 we know $\log(Q) > 0$ for a.e. $\tau \in (t_1, t_2)$. Then, by Lemma G.7:

$$\frac{\alpha}{L} \log \frac{\tilde{\gamma}_Q(t_2)}{\tilde{\gamma}_Q(t_1)} > \int_{t_1}^{t_2} C \frac{d}{d\tau} \log Q(\nabla R(\theta_\tau)) d\tau = C \, \log Q(\nabla R(\theta_{t_2})) - \log Q(\nabla R(\theta_{t_1})) = \frac{\alpha}{L} \log \frac{\tilde{\gamma}_Q(t_2)}{\tilde{\gamma}_Q(t_1)},$$

which is a contradiction. □

Now we are ready to construct the limiting sequence.

**Lemma G.9.** *For every limit point $(\bar{\theta}, \bar{g})$ of $\{(\frac{\theta_t}{\|\theta_t\|_2}, \frac{\nabla R(\theta_t)}{\|\nabla R(\theta_t)\|_2}) : t \geq 0\}$, there exists a sequence of $\{t_m : m \in \mathbb{N}\}$ such that $t_m \uparrow \infty$, $(\frac{\theta_t}{\|\theta_t\|_2}, \frac{\nabla R(\theta_t)}{\|\nabla R(\theta_t)\|_2}) \to (\bar{\theta}, \bar{g})$ and $\beta_{t_m} \to 1$, where $\bar{g} \in c \, \partial^\circ \phi_\alpha(\bar{\theta})$ for $c > 0$.*

*Proof.* Let $\{\epsilon_m : m \in \mathbb{N}\}$ be an arbitrary sequence with $\epsilon_m \to 0$. Now we construct $\{t_m\}$ by induction. Suppose $t_1 < t_2 < \ldots < t_{m-1}$ have already been constructed. Since $\bar{\theta}$ is a limit point and $\tilde{\gamma}_{Q,t} \uparrow \tilde{\gamma}_{Q,\infty}$, there exists a $s_m > t_{m-1}$ such that:

$$\| \frac{\theta_{s_m}}{\|\theta_{s_m}\|_2} - \bar{\theta} \|_2 \leq \epsilon_m, \quad \| \frac{\nabla(\theta_{s_m})}{\|\nabla R(\theta_{s_m})\|_2} - \bar{g} \|_2 \leq \epsilon_m, \quad \text{and,} \quad \frac{\alpha}{L} \log \frac{\tilde{\gamma}_{Q,\infty}}{\tilde{\gamma}_{Q,s_m}} \leq \epsilon_m^3.$$

The relationship between $\bar{g}$ and $\bar{\theta}$ through $\bar{g} \in c \, \partial^\circ \phi_\alpha(\bar{\theta})$ for $c > 0$ follows from Corollary 2 in [40]. Now let $s'_m > s_m$ be a time such that $\log Q(s'_m) = \log Q(s_m) + \epsilon_m$. According to Theorem 4.5, $\log Q \to \infty$, so $s'_m$ must exist. We construct $t_m \in (s_m, s'_m)$ to be the time that $\beta_{t_m}^{-2} - 1 \leq \epsilon_m^2$ where

the existence follows from Corollary G.8. It follows that $\beta_{t_m} \geq 1/\sqrt{1+\epsilon_m^2} \to 1$. Moreover, by Lemmas E.2 and E.1 we have:

$$\begin{aligned}\|\frac{\nabla R(\theta_{t_m})}{\|\nabla R(\theta_{t_m})\|_2} - \bar{g}\|_2 &\leq \|\frac{\nabla R(\theta_{t_m})}{\|\nabla R(\theta_{t_m})\|_2} - \frac{\nabla R(\theta_{s_m})}{\|\nabla R(\theta_{s_m})\|_2}\|_2 + \|\frac{\nabla R(\theta_{s_m})}{\|\nabla R(\theta_{s_m})\|_2} - \bar{g}\|_2 \\ &\leq \frac{2B_1}{LC_2\tilde{\gamma}_{Q,t_0}}\epsilon_m + \epsilon_m \to 0.\end{aligned}$$

Since $\nabla Q$ is a diffeomorphism that has the same growth in all directions due to being separable and identical for each coordinate we also have $\theta_{t_m}/\|\theta_{t_m}\|_2 \to \bar{\theta}$. To see this, we can express the limit of $\frac{\theta_t}{\|\theta_t\|_2}$ as follows:

$$\begin{aligned}\lim_{t\to\infty}\frac{\theta_t}{\|\theta_t\|_2} &= \lim_{t\to\infty}\frac{\nabla Q(\nabla R(\theta_t))}{\|\nabla Q(\nabla R(\theta_t))\|_2} \\ &= \lim_{t\to\infty}\frac{\frac{1}{\|\nabla R(\theta_t)\|_2}\nabla Q(\nabla R(\theta_t))}{\|\frac{1}{\|\nabla R(\theta_t)\|_2}\nabla Q(\nabla R(\theta_t))\|_2} \\ &= \lim_{t\to\infty}\frac{\frac{1}{\|\nabla R(\theta_t)\|_2}\nabla Q(\|\nabla R(\theta_t)\|_2\frac{1}{\|\nabla R(\theta_t)\|_2}\nabla R(\theta_t))}{\|\frac{1}{\|\nabla R(\theta_t)\|_2}\nabla Q(\|\nabla R(\theta_t)\|_2\frac{1}{\|\nabla R(\theta_t)\|_2}\nabla R(\theta_t)\|_2)} \\ &= \lim_{t\to\infty}\frac{\frac{1}{\|\nabla R(\theta_t)\|_2}\nabla Q(\|\nabla R(\theta_t)\|_2\bar{g})}{\|\frac{1}{\|\nabla R(\theta_t)\|_2}\nabla Q(\|\nabla R(\theta_t)\|_2\bar{g})\|_2} \\ &= \lim_{t\to\infty}\frac{\frac{Q^{\frac{1}{\alpha}-1}}{\|\nabla R(\theta_t)\|_2}\nabla Q(\|\nabla R(\theta_t))\|_2\bar{g})}{\|\frac{Q^{\frac{1}{\alpha}-1}}{\|\nabla R(\theta_t))\|_2}\nabla Q(\|\nabla R(\theta_t)\|_2\bar{g})\|_2} \\ &= \lim_{t\to\infty}\frac{\bar{\theta}}{\|\bar{\theta}\|_2} \\ &= \bar{\theta}.\end{aligned}$$

Here we used the definition of the horizon function and the fact that $\|\nabla R(\theta_t)\|_2 \to \infty$ as $\|\theta_t\| \to \infty$ by the coercivity property of the mirror map. □

*Proof of Theorem 4.7.* Combining Lemmas G.6 and G.9 implies Theorem G.2 and with that concluding the result as a direct consequence. □

### G.3 Hyperbolic entropy and homogeneous reparameterization

The current proof strategy of the KKT conditions relies on homogeneity $\alpha \geq 2$. This we believe is a technicality of the analysis and not a real obstruction. To highlight this we consider the hyperbolic entropy. We use the observation that the hyperbolic entropy $R_\lambda(\theta)$ corresponds to training with the two-homogeneous parameterization $u \odot v$ under gradient flow or differential inclusion as shown in [15, 29] for differentiable objectives. However the result only relies on the parameteric invariance hence the mirror flow-reparameterization correspondence also holds in our setting as also substantiated by [36, 35] and moreover it follows as a corollary from Lemma 3.1. This allows us to apply the result by [33] to characterize the max-margin. Together with the fact that for all $t \geq 0$ $u_t^2 - v_t^2 = u_{in}^2 - v_{in}^2 = \sqrt{\lambda}$ we can characterize the max margin in terms of $\theta = u \odot v$ giving the $L_1$-max margin. This follows directly from noticing that the normalized iterates $\bar{u}$ and $\bar{v}$ become balanced under the vector level constraint on $u$ and $v$:

$$\bar{u}_t^2 - \bar{v}_t^2 = (u_t^2 - v_t^2)/(\|u_t, v_t\|_2^2) = (u_{in}^2 - v_{in}^2)/(\|u_t, v_t\|_2^2) \to 0.$$

Therefore the max-margin solution has additional constraint $\bar{u}^2 = \bar{v}^2$. Changing the objective from $\frac{1}{2}\|\bar{u}, \bar{v}\|_2^2 = \|\bar{\theta}\|_1$ where $\bar{\theta} := \bar{u} \odot \bar{v}$. By the same argument we have that there exist a positive constant $b > 0$ such that $\|u_t, v_t\|_2^2/\|\theta_t\|_1 \to b$, thus $\bar{\theta} \simeq \frac{\theta}{\|\theta\|_1}$, which concludes the result. This implies that the normalized iterates $\theta_t/\|\theta_t\|_1$ (and therefore also $\theta_t/\|\theta_t\|_2$ ) corresponding to hyperbolic entropy mirror flow converge to the direction of the following optimization problem:

$$\min_{\theta\in\mathbb{R}^n} \|\theta\|_1 \qquad \text{such that} \qquad y_i f(\theta, x_i) \geq 0.$$

This shows that other strategies may exist to show the KKT conditions for $1 \leq \alpha \leq 2$.

**Relation between the reparameterization and hyperbolic entropy.** We provide some more details on the relation between the mirror flow and the reparameterization. We want to study the trajectory of $\theta = u \odot v$, this according to the chain rule can be written as:

$$d\theta_t = -(u_t^2 + v_t^2)\nabla_\theta \mathcal{L}(\theta_t)dt.$$

We now can connect $u^2 + v^2$ to the metric tensor of the hyperbolic entropy, by using the invariance and the fact that $\theta = u \odot v$ we can solve a system of equations:

$$\begin{cases} u^2 - v^2 = \sqrt{\lambda} \\ u \odot v = \theta. \end{cases}$$

This can be solved by applying the quadratic formula giving $u^2 + v^2 = \sqrt{4\theta^2 + \lambda}$. Note that the constant $4$ in front of the parameter can be dealt with by rescaling.

### G.4 KKT assumption

To show the KKT conditions for the max margin problem [46, 33] rely on the 1-homogeneity of the derivative of the norm squared. We would have $\partial^\circ \frac{1}{2}\|\theta\|^2 = \|\theta\|\partial^\circ\|\theta\|$. This allows them to extract the normalization term $1/q_{\min}^{1/L}$ and having the object $\|\theta\|\partial^\circ\|\theta\|/q_{\min}^{1/L}$. We can apply the same principle for the corresponding mirror descent objective in terms of the horizon function $\partial^\circ \frac{1}{2}\phi_\alpha^2(\theta)$. By construction we have that the horizon function is equivalent to an $L_p$ norm. This was formalized in Assumption G.1.

**Lemma G.10.** *For all $\lambda > 0$ and $\alpha = p \geq 2$ there exists a constant $C > 0$ such that for $\eta(\theta) \sim \|\theta\| \to \infty$*

$$\|C(\alpha Q(\nabla R(\theta)))^{2/\alpha - 1}\nabla^2 R(\theta)\theta/\eta(\theta) - \phi_\alpha(\theta)\partial^\circ\phi_\alpha(\theta)/\eta(\theta)\|_2 \to 0.$$

*Proof.* We have:

$$(\alpha Q(\nabla R(\theta)))^{2/\alpha - 1}\nabla^2 R(\theta)\theta = \frac{1}{((p-1)\|\theta\|_p^p + \frac{p}{2}\lambda\|\theta\|_2^2)^{1-2/p}}((p-1)|\theta|^{p-2} + \lambda)\theta$$

and

$$\phi_\alpha(\theta) = (p-1)^{1/p}\|\theta\|_p,$$

giving

$$\phi_\alpha(\theta)\partial^\circ\phi_\alpha(\theta) = (p-1)^{2/p}\|\theta\|_p^{2-p}|\theta|^{p-2}\theta.$$

Note that both terms constant normalization decay at rate $\|\theta\|_p^{2-p}$ as the term depending on $\lambda$ grows less fast. Moreover the terms $|\theta|^{p-2}\theta$ match upto a multiplicative constant. Knowing this we can bound the difference:

$$\begin{aligned} &\leq \|(p-1)^{2/p}|\theta|^{p-2}\theta\|_2^2 \left| \frac{C}{((p-1)\|\theta\|_p^p + \frac{p}{2}\lambda\|\theta\|_2^2)^{1-2/p}} - \|\theta\|_p^{2-p} \right|^2 / \eta^2(\theta) \\ &\sim \|\theta\|^{2p-2}\|\theta\|^{8-4p}/\eta^2(\theta) \to 0 \end{aligned}$$

iff $p > 2$, where we choose $C = (p-1)^{1-2/p}$ such that we can cancel the leading term in the expansion of:

$$\left| \frac{C}{((p-1)\|\theta\|_p^p + \frac{p}{2}\lambda\|\theta\|_2^2)^{1-2/p}} - \|\theta\|_p^{2-p} \right|^2 \sim \|\theta\|^{8-4p}.$$

Note that if we would set $\lambda = 0$ we get exactly zero for this particular choice of $C$. Note that the case $p = 2$ can be directly matched with $\phi_\alpha(\theta)\partial^\circ\phi_\alpha(\theta) = \theta$ and $\alpha Q(\nabla R(\theta)))^{2/\alpha-1}\nabla^2 R(\theta)\theta = (1+\lambda)\theta$, so $C = (1 + \lambda)$ yields the result. □

**Hyperbolic entropy.** We now show that such an approximation can not hold or the hyperbolic entropy. We first have the expression:

$$(\alpha Q(\nabla R(\theta)))^{2/\alpha-1}\nabla^2 R(\theta)\theta/q_{\min}^{1/L} = \left(\sum_i \sqrt{\theta_i^2+\lambda}\right)\frac{\theta}{\sqrt{\theta^2+\lambda}}/q_{\min}^{1/L}.$$

Comparing this to the subgradient of $\frac{1}{2}\|\theta\|_1^2$ we get:

$$\|\theta\|_1 \text{sign}(\theta)/q_{\min}^{1/L}.$$

These do not approach each other in the Euclidean distance as we can not bound

$$\|\frac{\theta}{\sqrt{\theta^2+\lambda}} - \text{sign}(\theta)\|,$$

which needs to go to zero for a good approximation. This is not possible in any metric.

### G.5 Two-layer optimization problem reformulation

Here we provide the details of the two-layer neural network $f(a, w, x) := \sum_{j=1}^N a_j\, \sigma(w_j^T x)$ setting with ReLU activation $\sigma(\cdot) := \max\{\cdot, 0\}$.

*Proof of Theorem 4.9.* The result follows from the balance equation for the unnormalized iterates and the rescale invariance of the two layer neural network, i.e., for $c > 0$ we have $a\sigma(w^T x) = a/c\, \sigma(cw^T x)$ by homogeneity of the ReLU activation.

The balance equation for the hyperbolic entropy gives us for each neuron $j \in [N]$ and $t \geq 0$:

$$\sqrt{a_{j,t}^2+\lambda} - \sum_{i=1}^d \sqrt{w_{j,i,t}^2+\lambda} = \text{constant}$$

Now divide both sides by the norm $\|a_t, w_t\|_2$ which gives for $t \to \infty$:

$$|\bar{a}_j| - \|\bar{w}_j\|_1 = 0.$$

Using this additional constrained, the objective $\phi_1 \sim \|\theta\|_1$ where $\theta = (a, w)$ can be rewritten as:

$$\|\theta\|_1 = 2\|a\|_1 = 2\sum_{j=1}^N \sqrt{|a_j|\|w_j\|_1}.$$

We can use a change of variable $\tilde{a}_j = a_j\|w_j\|_1$ and $\tilde{w_j}/\|w_j\|_1$ and the rescale invariance to get the objective and constrained in Eq. (5). The proof for the homogenous potentials is analogous but with dividing the balance equation with $\|a_t, w_t\|_p^p$. □

## H Margin alignment

Here we show how fast reaching the margin depends on the hyperparameter $\lambda$. We first provide a general result under Assumption H.1 and then apply it to our specific cases.

**Assumption H.1.** Assume there exists independent constants $a > 0$ and $c \geq 1$ such that:

$$\phi_\alpha \leq (\alpha Q)^{1/\alpha} \leq c\, \phi_\alpha + a.$$

**Lemma H.2.** *If $Q$ and $\phi_\alpha$ satisfy Assumption H.1, then after $\Omega(\exp(a^L))$ time the relative difference is $\mathcal{O}(1)$.*

*Proof.* From Theorem 4.6 we know for a general $\alpha$ that $(\alpha Q)^{1/\alpha} = \Omega(\log(t)^{1/L})$. Under Assumption H.1 we can bound the relative difference:

$$0 \leq \frac{(\alpha Q)^{1/\alpha} - \phi_\alpha}{(\alpha Q)^{1/\alpha}} \leq (c-1) + aO(1/\log(t)^{1/L})$$

Hence $t \geq \exp(a^L)$ the difference is $\mathcal{O}(1)$. □

**Corollary H.3.** *(Hyperbolic entropy) For the hyperbolic entropy mirror flow, after $\Omega(\exp((\sqrt{\lambda}n)^L))$ time the relative difference between $Q_\lambda(\nabla R_\lambda(\theta))$ and $\phi_1(\theta)$ is $\mathcal{O}(1)$.*

*Proof.* We can verify Assumption H.1 using the definition of $Q_\lambda(\nabla R_\lambda(\theta)) = \sum_{i=1}^n \sqrt{\theta_i^2 + \lambda}$ and $\phi_1(\theta) = \|\theta\|_1$:

**Lower bound**: We know that $\sqrt{\theta^2 + \lambda} \geq |\theta|$ for all $\theta \in \mathbb{R}$ and $\lambda \geq 0$ summing over all entries implies the lower bound holds.

**Upper bound**: For $\lambda \geq 0$ we can bound $\sqrt{\theta^2 + \lambda} \leq |\theta| + \sqrt{\lambda}$ summing now gives the upper bound with $a = n\sqrt{\lambda}$ and $c = 1$.

Applying Lemma H.2 concludes the result. □

**Corollary H.4.** *(Smoothened Homogeneous potentials) For the smoothened homogeneous mirror flow with $p \geq 2$, after $\Omega(\exp(\left(\frac{p\lambda n}{2}\right)^{L/p}))$ time the relative difference between $Q_\lambda(\nabla R_\lambda(\theta))$ and $\phi_p(\theta)$ is $\mathcal{O}(1)$.*

*Proof.* We can verify Assumption H.1 using the definition of $Q_\lambda(\nabla R_\lambda(\theta)) = 1/q\|\theta\|_p^p + \frac{\lambda}{2}\|\theta\|_2^2$ and $\phi_p(\theta) = (p-1)^{1/p}\|\theta\|_p$. Recall that $q \geq 1$ is such that $\frac{1}{q} + \frac{1}{p} = 1$:

**Lower bound**: We can use that $pQ_\lambda(\nabla R_\lambda(\theta)) \geq pQ_0(\nabla R_0(\theta)) = (p-1)||\theta||_p^p$ taking the power $1/p$ on both sides gives the lower bound.

**Upper bound**: We can use that for all $z \in \mathbb{R}$ we have that $z^2 \leq |z|^p + 1$. This gives the upper bound:

$$\begin{aligned} pQ_\lambda(\nabla R_\lambda(\theta)) &= (p-1)\|\theta\|_p^p + \frac{p\lambda}{2}\|\theta\|_2^2 \\ &\leq (p - 1 + \frac{p\lambda}{2})\|\theta\|_p^p + \frac{p\lambda n}{2} \end{aligned}$$

Now taking both sides to the power $1/p$ gives:

$$\begin{aligned} (pQ_\lambda(\nabla R_\lambda(\theta)))^{1/p} &\leq \left((p - 1 + \frac{p\lambda}{2})\|\theta\|_p^p + \frac{p\lambda n}{2}\right)^{1/p} \\ &\leq \left((p - 1 + \frac{p\lambda}{2})/(p-1)\right)^{1/p}(1-p)^{1/p}\|\theta\|_p + \left(\frac{p\lambda n}{2}\right)^{1/p} \\ &= c\phi_p(\theta) + a \end{aligned}$$

where $c = \left((p - 1 + \frac{p\lambda}{2})/(p-1)\right)^{1/p}$ and $a = \left(\frac{p\lambda n}{2}\right)^{1/p}$, concluding the upper bound.

Applying Lemma H.2 with the given $a$ and $c$ concludes the result. □

*Proof of Lemma 4.10* . The result now follows from combining Corollaries H.3 and H.4. □

# I NTK and modularity implications

Here we provide two additional implications of the main theorems. We recover the support vector machine with a kernel characterization as in [33]. Moreover, we can extend the max-margin result to the case where we use different mirror maps for each layer (i.e., modularity [3]).

**Neural tangent kernel.** The Neural Tangent Kernel (NTK) is a helpful tool to characterize the training dynamics of deep learning models.

**Corollary I.1** (Corollary of Theorem 4.7)**.** *Assume $f \in C^2$-smooth on $\mathbb{R}^n \setminus \{0\}$. Then for mirror flow under the same assumptions as Theorem 4.7 or 4.8, any limit point $\bar{\theta}$ of $\{\frac{\theta_t}{\|\theta_t\|_2}, t \geq 0\}$ points in the max-margin direction for the hard-margin SVM with kernel $K_{\bar{\theta}}(x, x') = \langle \nabla f_x(\bar{\theta}), \nabla f_{x'}(\bar{\theta}) \rangle_2$, where $f_x(\theta) := f(\theta, x)$. That is, there exists some $\beta > 0$ such that $\beta\bar{\theta}$ is the optimal solution for the following constrained optimization problem for $\alpha \geq 2$ or $\alpha = 1$:*

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\phi_\alpha^2(\theta) \qquad \textit{such that} \qquad y_i \langle \theta, \nabla f_{x_i}(\bar{\theta}) \rangle_2 \geq 1, \textit{ for all } i \in [K].$$

*If we do not assume $f \in C^2$, for mirror flow, then there exists a mapping $h(x) \in \partial^\circ f_x(\bar{\theta})$ such that the same conclusion holds for $K_{\bar{\theta}}(x, x') = \langle h(x), h(x') \rangle_2$.*

*Remark* I.2. Corollary I.1 indicates that, in general, the max-margin characterization cannot be captured by a Reproducing Kernel Hilbert Space (RKHS), as the underlying objective does not respect an inner product structure. This highlights a fundamental distinction between mirror flow and gradient flow. A natural framework for capturing the behavior of mirror flow is provided by Reproducing Kernel Banach Spaces (RKBS) [14].

By the homogeneity of $q_i$, we can characterize KKT points using kernel SVM.

**Lemma I.3.** *If $\theta^*$ is KKT point of the optimization problem in either Theorem 4.7 or 4.8, then there exists $h_i \in \partial^\circ f_{x_i}(\theta^*)$ for $i \in [K]$ such that $\frac{1}{L}\theta^*$ is an optimal solution for the following constrained optimization problem:*

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\phi_\alpha^2(\theta) \qquad \textit{such that} \qquad y_i \langle \theta, h_i \rangle_2 \geq 1, \textit{ for all } i \in [K].$$

*Proof.* For $\theta = \frac{2}{L}\theta^*$, from homogeneity, we can see that $y_i\langle \theta, h_i \rangle_2 = 2q_i(\theta^*) \geq 2 > 1$, which implies Slater's condition so there is a feasible solution. Thus, we only need to show that $\frac{1}{L}\theta^*$ satisfies KKT conditions for of the optimization problem. By the KKT conditions for the optimization problem in either Theorem 4.7 or 4.8, we can construct $h_i \in \partial^\circ q_i(\theta^*)$ for $i \in [K]$ such that $\partial^\circ \frac{1}{2}\phi_\alpha^2(\theta^*) - \sum_{i=1}^K \lambda_i y_i h_i = 0$ for some $\lambda_i \geq 0$ for $i \in [K]$ and $\lambda_i(q_i(\theta^*) - 1) = 0$. Thus $\frac{1}{L}\theta^*$ satisfies by homogeneity of $\phi_\alpha$ and Eulers identity:

$$\begin{cases} \frac{1}{L}\partial^\circ \frac{1}{2}\phi_\alpha^2(\theta) - \sum_{i=1}^K \frac{1}{L}\lambda_i y_i h_i = 0 \\ \frac{1}{L}\lambda_i((\langle \frac{1}{L}\theta^*, h_i \rangle) - 1) = \frac{1}{L}\lambda_i(q_i(\theta) - 1) \geq 0 \end{cases}$$

So $\frac{1}{L}\theta^*$ satisfies the KKT conditions for the optimization problem. □

Now we prove Corollary I.1.

*Proof of Corollary I.1.* The proof is analogous to [33, Corollary 4.5]. By Theorem 4.7 or 4.8, every limit point $\bar{\theta}$ is along the direction of a KKT point of the optimization problem. Combining this with Lemma I.3, we know that every limit point $\bar{\theta}$ is also along the max-margin direction of the new optimization problem.

For smooth models, $h_i$ is exactly the gradient $\nabla f_{x_i}(\bar{\theta})$ So, it becomes the optimization problem for SVM with kernel $K_{\bar{\theta}}(x, x') = \langle \nabla f_x(\bar{\theta}), \nabla f_{x'}(\bar{\theta}) \rangle_2$. For non-smooth models, we can construct an arbitrary function $h(x) \in \partial^\circ f_x(\bar{\theta})$ that ensures $h(x_i) = h_i$. Then, the optimization problem for SVM with kernel $K_{\bar{\theta}}(x, x') = \langle h(x), h(x') \rangle_2$. □

**Modularity result.** Motivated by the modular perspective on modern optimization [3, 26], we consider applying different mirror maps to different layers and aim to characterize the resulting max-margin solution. The proof follows the arguments of [33, Lemma H.2]. We specify the notion of margin and the associated constrained optimization problem. Let the mirror map potential be layer-wise separable, i.e., $R(\theta) := \sum_{j=1}^{L} R_j(W_j)$ with corresponding homogeneity degrees $\alpha_i \geq 1$, and assume $f$ is $(1, 1, \ldots, 1)$-homogeneous, i.e., $f(c_1 W_1, c_2 W_2, \ldots, c_L W_L) = (\Pi_{j=1}^{L} c_j) f(W_1, \ldots, W_L)$.

**Definition I.4** (Multi-$Q$-margin)**.** Let $f$ be $(1, \ldots, 1)$-homogeneous and let $R$ be an asymptotically $\alpha$-positively homogeneous mirror map that is layer-wise separable, i.e., $R(\theta) := \sum_{j=1}^{L} R_j(W_j)$, with associated functions $Q_j$ and degrees $\alpha_j \geq 1$. The multi-$Q$-margin corresponding to $R$ is defined as

$$\begin{aligned}\gamma_Q &:= \min_i \, y_i \, f\left( \frac{W_1}{\left(\alpha_1 Q_1(\nabla R_1(W_1))\right)^{1/\alpha_1}}, \ldots, \frac{W_L}{\left(\alpha_L Q_L(\nabla R_L(W_L))\right)^{1/\alpha_L}}, x_i \right) \\ &= \min_i \frac{y_i f(\theta, x_i)}{\prod_{j=1}^{L} \left(\alpha_j Q_j(\nabla R_j(W_j))\right)^{1/\alpha_j}}.\end{aligned}$$

This then leads to the constrained optimization problem:

$$\min_{\theta \in \mathbb{R}^n} \; \frac{1}{2} \sum_{j=1}^{L} \phi_{\alpha_j}^2(W_j) \qquad \text{such that} \qquad y_i f(W_1, \ldots, W_L, x_i) \geq 1, \text{ for all } i \in [K].$$

with directional vector $(W_1/\|W_1\|_2, \ldots, W_L/\|W_L\|_2)$.

# J Experimental details and ablations

Here we provide the details of all the experiments in the main text and additional ablations. For the toy example the precision used is float64, to prevent underflow.

## J.1 Student-teacher two-layer neural network

Here we provide the details and ablations on the student teacher setting. We randomly initialize a teacher with 3 neurons and no biases. Then we train a 100 neuron two layer student neural network also without biases. We consider 3 mirror maps hyperbolic, gradient descent, and smoothened homogeneous with $p = 3$. The data is sampled from the unit circle, $N = 200$ points. For teacher network $f_t(\theta)$, the neurons are randomly generated such that $|a_j|||w_j||_2 = 1$ for each $j \in [3]$. The labels are then generated by the indicator $\mathbb{I}_{f_t>0} - \mathbb{I}_{f_t\leq 0}$. The students weights are initialized in the mean field regime. The experiments are all executed on a NVIDIA GeForce RTX 4090 Laptop GPU.

**Experiment main text.** For the first set of experiments for the main text we use time rescaling i.e. when the loss is small enough ($\mathcal{L} < 0.1$) and all training data points have been classified we set the learning rate $0.1\eta/\mathcal{L}$. We use starting learning rates $\eta \in \{0.001, 0.1, 1\}$, for hyperbolic, gradient descent and $L_3$ in that order. We train until the training loss is below $1e-50$. We repeat the experiment over 6 seeds. Here $\lambda = 0.1$ for the hyperbolic entropy and $\lambda = 1$ for the smoothened homogeneous potential. The max margin values are reported in Table 2. Every mirror map maximizes their margin value. In the main text this lead to learning different representations which all exhibit feature learning (Figure 2).

Table 2: Reported max margin values for the student teacher setup with time rescaling over 6 seeds.

| | $L_1$ | $L_2$ | $L_3$ |
|---|---|---|---|
| Hyp. | $\mathbf{3.31 \times 10^{-5} \pm 2.00 \times 10^{-6}}$ | $2.76 \times 10^{-4} \pm 8.80 \times 10^{-8}$ | $4.55 \times 10^{-4} \pm 3.50 \times 10^{-6}$ |
| GD | $1.83 \times 10^{-6} \pm 1.02 \times 10^{-7}$ | $\mathbf{2.94 \times 10^{-4} \pm 4.42 \times 10^{-8}}$ | $1.34 \times 10^{-3} \pm 4.59 \times 10^{-5}$ |
| $p = 3$ | $1.19 \times 10^{-6} \pm 1.27 \times 10^{-8}$ | $2.91 \times 10^{-4} \pm 7.77 \times 10^{-7}$ | $\mathbf{1.76 \times 10^{-3} \pm 5.18 \times 10^{-6}}$ |

**The influence of $\lambda$.** We now do not rescale time but consider the same fixed learning rate. We now want to investigate the role of $\lambda$ in both the hyperbolic and smoothened homogeneous potentials. For the hyperbolic entropy we consider $\lambda \in \{1, 0.1, 0.01\}$ and for the smoothened homogeneous we consider $\lambda \in \{10, 1, 0.1\}$. We train for $T = 10000$ epochs and repeat for 6 seeds. We both report the final margin values and the representation reached.

We illustrate this in Figures 5 and 6 for seed 42. The main observation is that feature learning still occurs but the representation found has more smeared out neurons as in the case for gradient descent. In Tables 3 and 4 we report the max margin values reached. Indeed the corresponding margin to each mirror map becomes smaller when using a larger $\lambda$.

**Reaching the margin.** Next we empirically verify that for different values of $\lambda > 0$, the time required to reach the corresponding max-margin solution can vary substantially. To this end, we train for a long but fixed duration without time rescaling. The results in Figure 6 show that, for large $\lambda$, the returned representation remains similar to that of gradient descent, with neurons more widely spread. This behavior persists even when the training time is increased tenfold, as shown in Figure 7. These findings indicate that, to obtain representations that differ from those of gradient descent, one must either train for extremely long times or appropriately tune the hyperparameter $\lambda$.

Table 3: Reported margin values for different $\lambda$ values for the hyperbolic entropy.

| $\lambda$ | $L_1$ | $L_2$ |
|---|---|---|
| 1 | $9.37 \times 10^{-6} \pm 1.14 \times 10^{-6}$ | $\mathbf{2.48 \times 10^{-4} \pm 1.54 \times 10^{-5}}$ |
| 0.1 | $1.21 \times 10^{-5} \pm 7.51 \times 10^{-7}$ | $2.29 \times 10^{-4} \pm 2.09 \times 10^{-5}$ |
| 0.01 | $\mathbf{1.55 \times 10^{-5} \pm 1.23 \times 10^{-6}}$ | $2.20 \times 10^{-4} \pm 2.58 \times 10^{-5}$ |

**Long training.** Here we investigate the exponential separation for $\lambda = 10$ for $T \in \{1000, 10000, 100000\}$. We observe that margin grows but the representation does not change

Table 4: Reported margin values for different $\lambda$ values for the smoothened homogeneous potential $p = 3$.

| $\lambda$ | $L_2$ | $L_3$ |
|---|---|---|
| 10 | $2.24 \times 10^{-4} \pm 3.54 \times 10^{-6}$ | $1.21 \times 10^{-3} \pm 1.83 \times 10^{-5}$ |
| 1 | $2.32 \times 10^{-4} \pm 2.53 \times 10^{-6}$ | $1.43 \times 10^{-3} \pm 1.35 \times 10^{-5}$ |
| 0.1 | $\mathbf{2.34 \times 10^{-4} \pm 2.43 \times 10^{-6}}$ | $\mathbf{1.43 \times 10^{-3} \pm 1.46 \times 10^{-5}}$ |

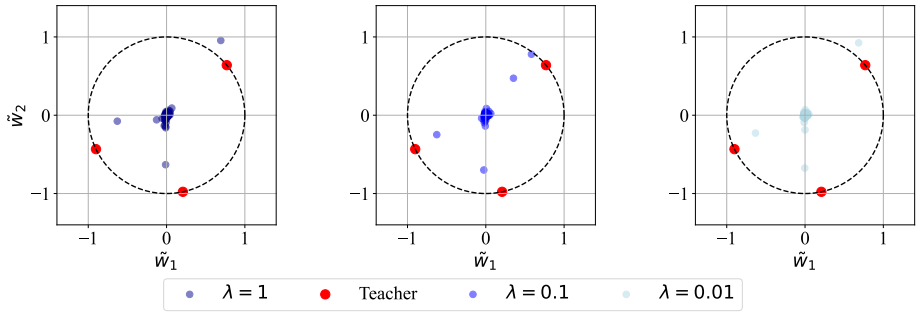


Figure 5: Learned representation by mirror descent with hyperbolic entropy for various $\lambda$. Larger $\lambda$ leads to a different representation that looks more like gradient descent solution with more neurons spread out.

indicating it is really hard to reach the time rescaled solution or small $\lambda$ solution when starting with large $\lambda$.

Table 5: Reported $L_3$ margin values for different epoch times $T$ for the smoothened homogeneous mirror.

| $T$ | $L_3$ |
|---|---|
| 100000 | $\mathbf{1.39 \times 10^{-3} \pm 1.84 \times 10^{-5}}$ |
| 10000 | $1.21 \times 10^{-3} \pm 1.83 \times 10^{-5}$ |
| 1000 | $8.03 \times 10^{-4} \pm 2.35 \times 10^{-5}$ |

### J.2 Three-layer neural network student-teacher setup

We conduct an additional experiment on a 3 layer student network with the same 2 layer teacher as in the previous section. The 3 layer neural network has hidden dimension 100. We again train with the hyperbolic mirror map, gradient descent, and smoothened homogeneous ($p = 3$). The students weights are initialized in the mean field regime.

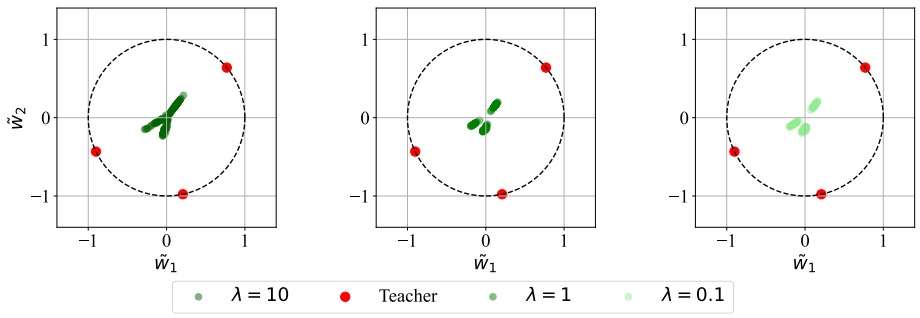


Figure 6: Learned representation by mirror descent with smoothened homogeneous ($p = 3$) for various $\lambda$. Larger $\lambda$ leads to a different representation that looks more like gradient descent solution with more neurons spread out.

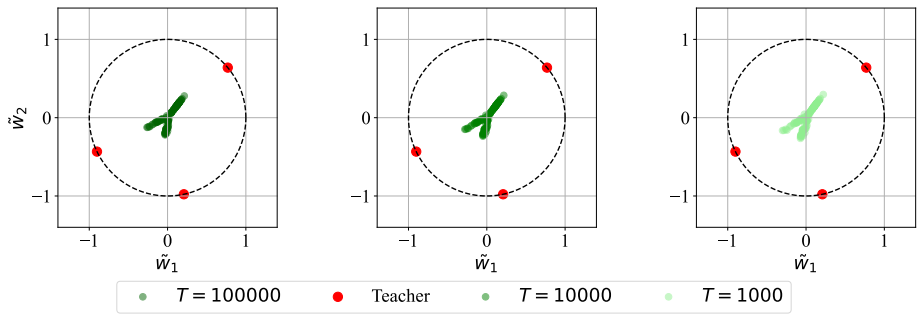


Figure 7: Training ten times longer does not change the representation much when $\lambda = 10$ is large.

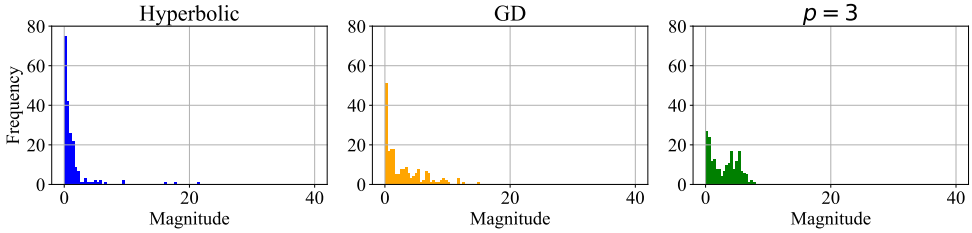


Figure 8: Weight magnitude distribution of the first layer after time rescaling for all mirror maps.

**Time rescaling.** To reach closer to the margin we again employ time rescaling. In order to ensure stable time rescaling we start rescaling when the loss is below $0.1$ and then we set the learning rate to $\eta\, 0.1/\mathcal{L}$ instead. Moreover, we set the learning rates to $\eta \in \{0.001, 0.01, 0.1\}$ for the hyperbolic, GD, smoothened homogeneous mirror maps in that order. For the hyperbolic mirror map we use $\lambda = 0.1$ and for the smoothened homogeneous we use $\lambda = 1$. We report the weight distributions of all 3 layers in Figures 8, 9, and 10. We observe in all layers that the weight magnitude distribution has shifted accordingly to the corresponding KKT problem. To illustrate the effect in function space we also plot the input-output map highlighting the decision boundary in Figure 11. Moreover, we report the final margin value found in Table 6. We observe that the margin values in case of $L_1$ and $L_3$ have different orders of magnitude, which is in line with the change in representation.

Table 6: Max margins reached by the 3 mirror descent algorithms after time rescaling for the 3 layer student and 2 layer teacher setup.

| | $L_1$ | $L_2$ | $L_3$ |
|---|---|---|---|
| Hyp. | $\mathbf{(4.01 \pm 0.34) \times 10^{-8}}$ | $(6.21 \pm 0.72) \times 10^{-5}$ | $(1.03 \pm 0.26) \times 10^{-4}$ |
| GD | $(1.13 \pm 0.07) \times 10^{-9}$ | $\mathbf{(1.10 \pm 0.00) \times 10^{-4}}$ | $(1.73 \pm 0.07) \times 10^{-3}$ |
| Hom. | $(2.39 \pm 0.08) \times 10^{-10}$ | $(9.36 \pm 0.04) \times 10^{-5}$ | $\mathbf{(3.42 \pm 0.06) \times 10^{-3}}$ |

**Additional experiments.** We conduct the same 2 experiments as in case of the 2 layer student: influence of $\lambda$ in both the case of hyperbolic entropy and smoothened homogeneous. In the first experiment, we use $\eta = 0.01$ and $T = 10000$ for the hyperbolic entropy. We report the margins reached in Tables 7 and 8. Moreover, we report the changed input representation in Figures 12 and 13.

Table 7: Max margins for different $\lambda$ values with the hyperbolic entropy mirror in the 3 layer student setting.

| | $L_1$ | $L_2$ |
|---|---|---|
| $\lambda = 1$ | $(1.67 \pm 0.24) \times 10^{-9}$ | $\mathbf{(6.43 \pm 0.51) \times 10^{-5}}$ |
| $\lambda = 0.1$ | $(3.62 \pm 0.30) \times 10^{-9}$ | $(4.76 \pm 0.60) \times 10^{-5}$ |
| $\lambda = 0.01$ | $\mathbf{(4.87 \pm 0.23) \times 10^{-9}}$ | $(4.44 \pm 0.76) \times 10^{-5}$ |

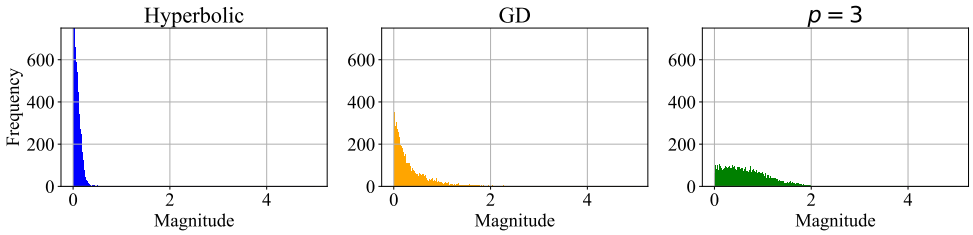


Figure 9: Weight magnitude distribution of the second layer after time rescaling for all mirror maps.

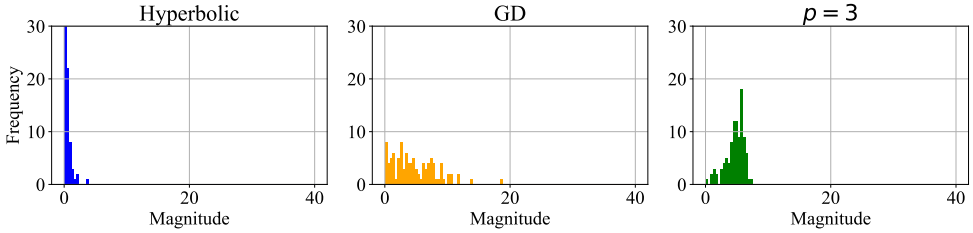


Figure 10: Weight magnitude distribution of the third layer after time rescaling for all mirror maps.

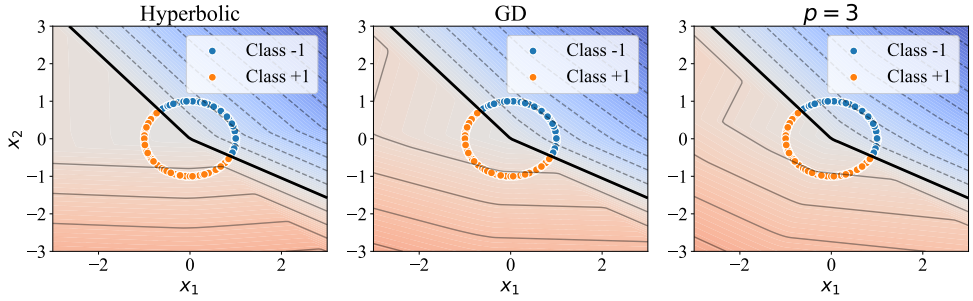


Figure 11: Illustration of the decision boundary and function activation value reached for the 3 layer neural network experiment.

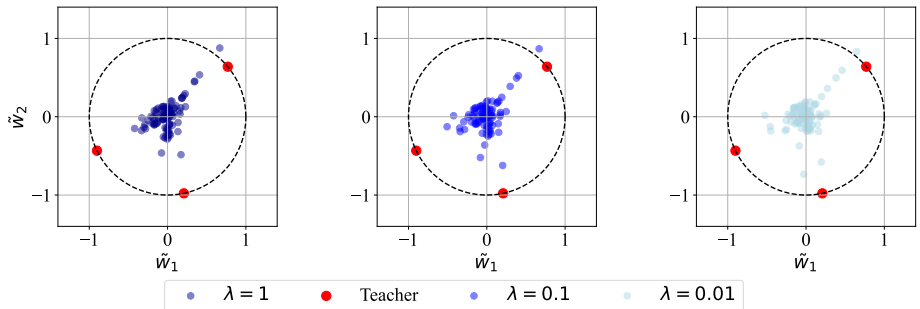


Figure 12: Learned representation in the input layer $w$ by mirror descent with hyperbolic entropy for various $\lambda$ in the 3 layer student setting. Larger $\lambda$ leads to a different representation that looks more like gradient descent solution with more neurons spread out.

Table 8: Max margins for different $\lambda$ values with the smoothened homogeneous mirror ($p = 3$).

| | $L_2$ | $L_3$ |
|---|---|---|
| $\lambda = 10$ | $\mathbf{(8.76 \pm 0.09) \times 10^{-5}}$ | $(1.75 \pm 0.04) \times 10^{-3}$ |
| $\lambda = 1$ | $(8.69 \pm 0.06) \times 10^{-5}$ | $(2.62 \pm 0.03) \times 10^{-3}$ |
| $\lambda = 0.1$ | $(7.47 \pm 0.04) \times 10^{-5}$ | $\mathbf{(3.44 \pm 0.03) \times 10^{-3}}$ |

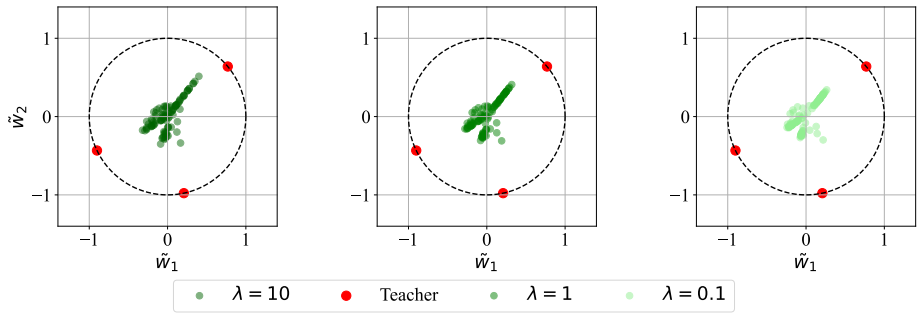


Figure 13: Learned representation in the input layer $w$ by mirror descent with smoothened homogeneous ($p = 3$) for various $\lambda$ in the 3 layer student setting. Larger $\lambda$ leads to a different representation that looks more like gradient descent solution with more neurons spread out.

### J.3 Classification on vision tasks

We run mirror descent with stochastic gradient estimates on single A100 GPU. We do not use weight decay or momentum. The inverse metric tensor is created in the following way to ensure comparability and layerwise stability:

$$\nabla^2 R^{-1}_{\text{hyp}}(\theta) = \sqrt{\theta^2 * \text{width} + \lambda}/(\sqrt{1+\lambda}) \text{ and } \nabla^2 R^{-1}_{\text{hom}}(\theta) = \frac{\lambda}{|\theta|^{p-2} * \text{width}^{(2-p)/2} + \lambda}.$$

This ensures the average update for each element in the same layer gets a similar update. Note we do not do learning rescaling with the loss or multiplicative coefficient as for the toy example. For tracking the margin we use the layerwise product of the norms and the logits similar as in [33]. We use He initialization for the weights and zero initialization for the biases of the first layer. In other words we are using the standard parameterization (SP). Note that this setting is not in the feature learning regime, therefore, we can not expect all layers to move equally as much. For the experiment we sweep the hyper parameters $\lambda \in \{1e-05, 0.01, 0.1, 1\}$ and learning rate $\eta \in \{0.1, 0.2\}$. We use batch-size 100, we do not use weight decay or momentum. We train a VGG16 [43] without biases (except the first layer) for 1000 epochs. The first layer having biases does not violate homogeneity as it can be seen as augmented data input. The numerical values are over 3 seeds, the illustrations are for one seed: 42. Furthermore, we note that training either GD, or Smoothened homogeneous becomes unstable and diverges when trained with learning rate 0.2.

We report the best accuracy values for each method in Table 9. We observe that the hyperbolic entropy leads to an improvement over the other methods. Note that the standard baseline accuracy is higher due to the use of momentum, batch norm and weight decay. Besides the first layer magnitude plot in the main text we plot a histogram of the last layer in Figure 14. We observe that the weight distributions are very similar. This is due to the standard parameterization (SP) [53]. This highlights the importance of knowing in which regime we are training, irrespective of validation accuracy performance, which representation we end up with. Note that in [45, 16] the mirror maps lead to very different representations and with that different magnitude distributions per layer in modern architectures using residual connections and batch/layer norm. Moreover, it also leads to improved performance in these more general settings notably for using a mirror map associated with the hyperbolic entropy [16] and the homogeneous mirror map with ($p = 3$) [45].

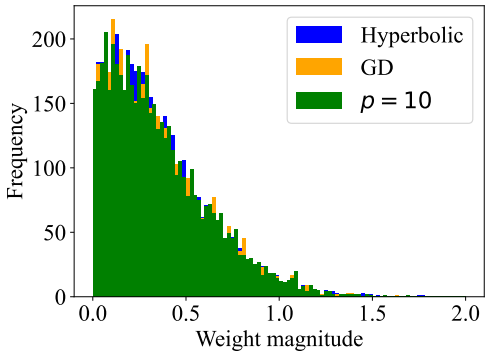


Figure 14: The weight magnitude distribution of a VGG16's last layer. We observe that all algorithms get a similar distribution due to the standard parameterization.

Table 9: Validation accuracy for a VGG16 on CIFAR 10 without using batch normalization, weight decay or momentum.

| Mirror | Val. Acc. |
|---|---|
| GD | $81.26 \pm 0.15$ |
| $p = 10$ | $81.26 \pm 0.12$ |
| Hyperbolic | $81.96 \pm 0.41$ |

### J.4 First step analysis

We furthermore conduct a first step analysis. In order for the parameter norm to grow and reach the max-margin direction we need the parameters to move from initialization. For gradient descent this has been studied under the NTK and mean-field regimes. Under the mean-field regime we have that $\nabla_i \mathcal{L} = O(1)$ for all $i \in [L]$. This means that the weights of all layers are moving. Now for mirror flow we have an additional contribution trough the metric tensor. To account for this, we can study the random variable transform $\nabla_i^2 R^{-1}(\theta_0)$ with the standard initialization $\theta_0 \sim N(0, I_n \sigma^2)$. If we treat this pointwise and together with the learning rate we would need to have

$$\mathbb{E}[\eta \nabla_i^2 R^{-1}(\theta_0)] = O(1)$$

to keep the maximal update like in gradient descent. We will refer to this as mirror-$\mu$P, which is needed to preserve the update magnitude. We now know what mirror descent has to satisfy in order to give a maximal parameter update at initialization and reach the margin.

Table 10: Training recipes for mirror descent with learning rate $\eta$ and hyperparameter $\lambda$, denoted as $(\eta, \lambda)$.

| Mirror potential | Mirror-$\mu$P | Margin reachable |
|---|---|---|
| GD | $(1, -)$ | $(1, -)$ |
| Hyp | $(1, 1), (\sigma^{-1}, 0), (\sigma^{-1}, \sigma^2)$ | $(\sigma^{-1}, 0), (\sigma^{-1}, \sigma^4)$ |
| $L^p_{p,\lambda}, p > 2$ | $(1, 1), (\sigma^{p-2}, \sigma^{p-2})$ | $(\sigma^2, \sigma^2)$ |

This allows us to summarize the hyperparameter for designing training recipes in Table 10. Observe that how Mirror-$\mu$P is decoupled from reaching a max-margin solution, implying that we can have large parameter changes at the start of training without eventually reaching the corresponding margin solution.

**Example J.1.** Consider the hyperbolic entropy, which has 2 mirror-$\mu$P solutions. The expectation is given by $\mathbb{E}\left[\eta\sqrt{\theta_0^2 + \lambda}\right] = O(\eta(\sigma + \sqrt{\lambda}))$, so $\lambda = 1, \eta = 1$ is a solution for decaying $\sigma$. Moreover, $\eta = \sigma^{-1}, \lambda = \sigma^2$ is a solution for all $n \in \mathbb{N}^+$.

*Remark* J.2. The random variables corresponding to the metric tensor $\nabla^2 R^{-1}(\theta_0)$ and gradient $\nabla f(\theta_0)$ are coupled. However we can justify using free probability theory to treat them as decoupled.