MIRROR DESCENT AT THE EDGE OF STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In many deep learning settings, gradient descent has been shown to operate at the *edge of stability* – a regime where the largest eigenvalue of the loss Hessian hovers around $2/\eta$. We study this phenomenon in *mirror descent*, a generalization of gradient descent. As mirror descent is non-Euclidean, existing explanations of the edge of stability phase do not extend to this setting. We generalize the existing notion of sharpness to introduce *mirror sharpness*, which measures the curvature of the loss with respect to the mirror potential. Empirically, we observe that the mirror sharpness hovers slightly above the $2/\eta$ threshold. Furthermore, we derive a dual central flow ODE for mirror descent that models iterates in continuous time at the onset of edge of stability. Our ODE provides insights into the difficulties of extending existing arguments for why sharpness stabilizes at $2/\eta$ from gradient descent to mirror descent in full generality.

1 INTRODUCTION

First-order methods, such as gradient descent, are the de facto optimization methods for much of deep learning. Despite their empirical successes, their behaviors on realistic *non-convex* objectives, such as neural network losses, are poorly understood and significantly depart from the classical picture of convex optimization.

For instance, the famed *descent lemma* (Nesterov, 2004) guarantees that if the maximum Hessian eigenvalue is uniformly bounded above by M, then gradient descent monotonically decreases the loss when the step size satisfies $\eta < 2/M$. However, in realistic settings, gradient descent has been shown to operate close to the boundary of this "stable regime." In particular, Cohen et al. (2020) observed across a variety of architectures that, for a fixed step size η, the *sharpness*, $S(x) := \lambda_{\max}(\nabla^2 L(x))$, progressively increases until it reaches approximately $2/\eta$, after which it oscillates near this threshold while optimization continues. The first phase is known as *progressive sharpening*, and the second is termed the *edge of stability* (EoS).

While full-batch gradient descent is typically the testbed for the analysis of the EoS regime, recent works have shown that EoS also manifests in several fundamental optimizers beyond this setting. A key thread through some of these works is that the relevant notion of sharpness is connected to the geometry induced by the optimizer. For instance, Cohen et al. (2023) formulates a *preconditioned sharpness* and observes an EoS phase in adaptive preconditioning methods like AdaGrad and Adam. Most recently, Islamov et al. (2026) formulates a *generalized sharpness* and observe EoS behavior in *steepest descent* methods in arbitrary norm.

In this paper, we consider EoS behavior in *mirror descent*, a non-Euclidean extension of gradient descent that fundamentally differs from the steepest descent studied in Islamov et al. (2026). Whereas steepest descent "fixes" a geometry, mirror descent introduces a Bregman geometry that can change at every point. Furthermore, mirror descent differs from general adaptive preconditioning algorithms (such as AdaGrad and Adam, as studied in Cohen et al. (2023)) because the Bregman geometry is *explicitly* given by the iterate instead of an additional state that depends on the history of the trajectory. We hope that understanding mirror descent at EoS may help clarify why self-stabilization appears beyond just the setting of gradient descent. Mirror descent provides a simple and *explicit* adaptive geometry and, at the same time, appears to display EoS behavior that "adapts" to the local geometry.

In (stochastic) gradient descent, why does sharpness tend to hover around $2/\eta$? Several existing explanations of this phenomenon for gradient descent (Arora et al., 2022; Damian et al., 2022;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Cohen et al., 2024; Liao et al., 2026) rely on the observation that at EoS, the gradient descent iterates oscillate along "unstable" eigendirections u_k with eigenvalue $2/\eta$, and these oscillations induce a "self-stabilizing" effect. It is unclear the extent to which this is true for mirror descent, and we explore this problem via the framework of central flows (Cohen et al., 2024). We find that such an explanation is *not* sufficient to justify the self-stabilization phenomenon in mirror descent in full generality, and we believe that further exploration may help establish the extent to which self-stabilization is a universal phenomenon across optimizers.

Our contributions are as follows:

- In Section 3, we define *mirror sharpness*, which generalizes existing notions of sharpness to the point-wise geometry induced by mirror descent. We connect this definition to an approximate descent lemma (Proposition 3) as well as optimization over a quadratic (Proposition 5).
- In Section 4, across a series of mirror maps and architectures, we empirically observe that mirror sharpness hovers around $2/\eta$.
- In Section 5, we derive a central flow ODE that empirically describes the time-averaged iterates at the onset of EoS via the method of (Cohen et al., 2024), improving over Riemannian gradient flow.

1.1 Related Works

Edge of stability Even in the simplest case of full-batch gradient descent, the EoS phenomenon remains poorly understood. Many works attempt to study EoS dynamics in the case of specific simplified loss functions (Zhu et al., 2022; Ahn et al., 2023; Kreisler et al., 2023; Ghosh et al., 2024; Liu et al., 2025; Yoo et al., 2025), even those beyond neural network losses. Another line of work (Arora et al., 2022; Damian et al., 2022; Cohen et al., 2024; MacDonald et al., 2025; Liao et al., 2026) analyzes EoS behavior for general losses. To our knowledge, all works of this sort exploit the *Euclidean* fact that the gradient of the sharpness can be expressed with the third derivative of the loss, which Damian et al. (2022, Lemma 2) dubs the "self-stabilization property." Unfortunately, for mirror sharpness, such a simple identity does not hold, and we find an additional third-order geometry correction (see Lemma 14). This creates difficulties in analysis, which we make explicit in Section 5.3.

Central flows Central flows, introduced by Cohen et al. (2024), are ODEs that model optimization algorithms in continuous time. They differ from "small step size" approximations (e.g., Riemannian gradient flow) because they use the EoS phase as an ansatz to construct a flow. Some recent work propose modifications to the central flow (Regis & Chewi, 2026a; Marion, 2026) and Regis & Chewi (2026b) extends a model to algorithms with momentum. We extend the central flow model of Cohen et al. (2024) to mirror descent in the regime where one eigenvalue is at $2/\eta$ by modeling oscillations in the *dual* space, which we empirically show to track the mirror descent iterates better.

Implicit bias of mirror descent Mirror descent, introduced in Nemirovski & Yudin (1983), has been a central first-order optimization paradigm for decades, as it is a surprisingly broad framework. Recent works within the past decade have focused on the *implicit bias* of mirror descent. For instance, in various regression settings, Gunasekar et al. (2018); Azizan & Hassibi (2018); Azizan et al. (2022) argue that mirror descent finds a solution close to initialization in Bregman divergence. Sun et al. (2022; 2023) show that in classification settings, mirror descent (with an ℓ_p potential) finds a solution that maximizes a margin with respect to an ℓ_p norm. Most recently, Vansover-Hager et al. (2026) studies how sensitive mirror descent is to initializations, and find that *non-quadratic* mirror maps exponentially amplify perturbations at initialization. Our paper instead studies the implicit bias of mirror descent through the complementary lens of sharpness.

2 Background

We use standard notation, detailed in Appendix A. We will work with multi-linear operators though they are not required to understand the high-level contents of the paper. Throughout this paper, unless otherwise indicated, we will use $H(x) = \nabla^2 L(x)$ and $G(x) = \nabla^2 \Phi(x)$, where L is the loss

and Φ is the mirror potential (as defined below). **In all formal statements and definitions, we will assume L, Φ to have appropriate regularity where needed.** Assuming that these are C^3 is usually sufficient (unless otherwise noted).

Generalized eigenvalues We say λ_i is a *generalized eigenvalue* of (H, G) if $Hv_i = \lambda_i G v_i$ for some nonzero eigenvector v_i. When G is the identity, this reduces to an eigenvalue of H.

Mirror descent We say a function $\Phi : E \to \mathbb{R}$ is a *mirror potential* if it is differentiable, strictly convex, and its gradient $\nabla\Phi : E \to E^*$ is a bijection. We call $\nabla\Phi : E \to E^*$ the *mirror map*. Its inverse is given by $(\nabla\Phi)^{-1} = \nabla\Phi^*$, where Φ^* denotes the convex conjugate of Φ. In this paper, we set $E = \mathbb{R}^n$.[1]

The natural notion of "distance" induced by Φ on E is the *Bregman divergence* $D_\Phi(x, y) := \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y\rangle$. Note that the Bregman divergence is not symmetric (i.e., in general, $D_\Phi(x, y) \neq D_\Phi(y, x)$). When Φ is strongly convex, G is a positive-definite metric (i.e., $G(x) \succ 0$ for all $x \in E$).

The mirror descent (MD) algorithm is given by the update

$$x_{t+1} = \arg\min_{y \in E} \{\eta \langle \nabla L(x_t), y - x_t\rangle + D_\Phi(y, x_t)\},$$

or equivalently

$$x_{t+1} = \nabla\Phi^*(\nabla\Phi(x_t) - \eta \nabla L(x_t)). \tag{MD}$$

Conceptually, we may also think of the above as updates in dual coordinates $z_t = \nabla\Phi(x_t)$ given by

$$z_{t+1} = z_t - \eta(\nabla L \circ \nabla\Phi^*)(z_t).$$

We may refer to both z_t and x_t as iterates because the mirror map $\nabla\Phi$ is bijective. As an example, when $\Phi(x) = \frac{1}{2}\|x\|_2^2$, we retrieve ordinary gradient descent, with $x_t = z_t$.

3 Defining Sharpness

In this section, we generalize the notion of sharpness to mirror descent. Recall that for ordinary gradient descent, the sharpness is given by

$$S^{\ell_2}(x) := \lambda_{\max}(\nabla^2 L(x)) = \max_{\|u\|_2 = 1} \langle u, \nabla^2 L(x) u\rangle.$$

A loose geometric interpretation of this quantity is the maximum curvature of L along any direction u. We will refer to this notion of sharpness as the ℓ_2 sharpness.

Our definition is inspired by the work of Gunasekar et al. (2021), which interprets mirror descent *in the primal space* as a partial discretization of a Riemannian gradient flow with geometry $G(x) := \nabla^2\Phi(x)$. Thus, it is natural to measure the curvature of L under this geometry:

Definition 1 (Mirror sharpness) *Let $\nabla^2\Phi(x) \succ 0$. For $x \in E$, the* mirror sharpness[2] *$S^\Phi : E \to \mathbb{R}$ is given by*

$$S^\Phi(x) := \max_{\langle u, \nabla^2\Phi(x) u\rangle = 1} \langle u, \nabla^2 L(x) u\rangle.$$

Equivalently, $S^\Phi(x)$ is the top generalized eigenvalue of $(\nabla^2 L(x), \nabla^2\Phi(x))$.

Similar to the definition for ℓ_2 sharpness, S^Φ is the maximum curvature of L along any direction u *with respect to the local metric* $G(x) = \nabla^2\Phi(x)$. Note that when $\Phi(x) = \frac{1}{2}\|x\|_2^2$, we retrieve ℓ_2 sharpness.

Mirror sharpness also retrieves previous notions of curvature studied for preconditioned gradient descent (Cohen et al., 2023; 2024; Islamov et al., 2026), which we show in Section 4.1. Furthermore,

[1] This is not the case for certain choices of mirror descent (most notably, the multiplicative weights update induced by the negative entropy $\Phi_{\text{ent}}(x) = \sum_i x_i \log x_i$).

[2] More generally, one may define the sharpness with respect to an arbitrary metric G by replacing $\nabla^2\Phi(x)$ with $G(x)$.

while our definition resembles the *generalized sharpness* presented in Islamov et al. (2026), their work studies *steepest descent* algorithms in arbitrary norm, which induce a different *fixed* geometry on the space than mirror descent. In this sense, mirror sharpness can be viewed as generalized sharpness with respect to a point-wise norm.

One benefit of mirror sharpness is that it is relatively simple to compute, as it can be expressed as a generalized eigenvalue. If we assume that $\nabla^2\Phi(x)$ is positive-definite, then the mirror sharpness can be written as a symmetric eigenvalue problem via Rayleigh–Ritz:

Lemma 2 *Let $x \in E$ and $G = \nabla^2\Phi(x) \succ 0$. Then, $S^\Phi(x) = \lambda_{\max}(G^{-1/2}(x)\nabla^2 L(x)G^{-1/2}(x))$.*

3.1 Second-order pictures of mirror sharpness

We show how, at the second order, we can retrieve familiar results in terms of the mirror sharpness.

The ℓ^2 sharpness in ordinary gradient descent is significant because it can characterize when L decreases between two iterates x_t and x_{t+1}. In the case of mirror descent, this relationship is murkier because the geometry induced by Φ changes between x_t and x_{t+1}. However, this asymmetry only arises at the *third* order locally:

$$D_\Phi(x_t, x_{t+1}) = D_\Phi(x_{t+1}, x_t) + O(\|x_{t+1} - x_t\|^3).$$

Consequently, at second order, using standard arguments, we recover a descent lemma (Bolte et al., 2018) with the familiar $2/\eta$ threshold.

Proposition 3 (Second-order (approximate) descent lemma) *Let $\nabla^2\Phi(x_t) \succ 0$. Let x_{t+1} be a mirror descent update with step size η. Then,*

$$L(x_{t+1}) - L(x_t) \le \left(S^\Phi(x_t) - \frac{2}{\eta}\right) D_\Phi(x_{t+1}, x_t) + O\left(\|x_{t+1} - x_t\|^3\right).$$

Next, we restrict the loss to a quadratic objective, and we show that for any mirror map Φ, mirror descent can be reinterpreted as gradient descent. To our knowledge, this appears not to be a readily known result in the literature. Using this equivalence, we show that mirror sharpness naturally emerges as the sharpness of the reparameterized loss function.

Theorem 4 (Quadratic mirror descent as gradient descent) *Let $L(x) := \frac{1}{2}\langle x - x^\star, A(x - x^\star)\rangle$ with $A \succ 0$. Let $\nabla^2\Phi(x) \succ 0$ for all $x \in E$. Define $z^\star := \nabla\Phi(x^\star)$ and $y := A^{-1/2}(z - z^\star)$, and let*

$$\widehat{L}(y) := \Phi^*(z^\star + A^{1/2}y) - \Phi^*(z^\star) - \langle x^\star, A^{1/2}y\rangle = D_{\Phi^*}(z^\star + A^{1/2}y, z^\star).$$

Then, the dual mirror descent updates z_k on L correspond to the gradient descent updates y_k on $\widehat{L}$ (i.e., $z_k = z^\star + A^{1/2}y_k$).

This equivalence shows that the quadratic picture of mirror descent in the dual is essentially the same as gradient descent. Thus, any "new" behavior that is different from gradient descent must arise through third-order interactions of the loss, as we later argue in Section 5.3. Lastly, we identify mirror sharpness with ℓ^2 sharpness in the reparameterized coordinates:

Proposition 5 *Let $\widehat{L}$ be defined as in Theorem 4 and $y \in E$. Set $x := \nabla\Phi^*(z^\star + A^{1/2}y)$ (i.e., x is the corresponding coordinate of y) and take $\nabla^2\Phi(x) \succ 0$. Then, $\lambda_{\max}(\nabla^2\widehat{L}(y)) = S^\Phi(x)$.*

4 Mirror sharpness stabilizes at $2/\eta$

In this section, we empirically demonstrate that just as sharpness hovers around $2/\eta$ in gradient descent, our notion of mirror sharpness also hovers around $2/\eta$.

4.1 Preconditioned gradient descent (PGD)

As a sanity check, we apply our definition of mirror sharpness to preconditioned gradient descent (PGD), which can be rewritten as mirror descent with a quadratic potential $\Phi_{\text{PGD}}(x) = \frac{1}{2}\langle x, Px\rangle$ for a matrix $P \succ 0$. In this setting, because $\nabla\Phi_{\text{PGD}}$ is linear, Equation (MD) reduces to the update $x_{t+1} = x_t - \eta P^{-1}\nabla L(x_t)$. Mirror sharpness recovers previous notions of sharpness for PGD as previously studied in (Cohen et al., 2024; Islamov et al., 2026):

$$S^{\Phi_{\text{PGD}}}(x) = \max_{\langle u, Pu\rangle = 1} \langle u, \nabla^2 L(x)u\rangle = \max_{\|v\|_2=1} \langle P^{-1/2}v, \nabla^2 L(x) P^{-1/2}v\rangle,$$

where the second equality follows from a change of variables $u \mapsto P^{-1/2}v$.

4.2 Non-Euclidean mirror potentials

We demonstrate EoS behavior with various architectures trained on a subset of CIFAR-10. All experiments in this section are run with full-batch mirror descent, ReLU activations, and MSE loss. Specific architectural and implementation details can be found in Appendix E. In general, we have found these behaviors rather consistent across several architectures, with further experiments shown in Appendix F.

Hypentropy We consider the class of *hypentropy* mirror potentials

$$\Phi_{\text{hyp},\gamma}(x) = \sum_i \left(x_i \operatorname{arcsinh}\left(\frac{x_i}{\gamma}\right) - \sqrt{x_i^2 + \gamma^2} \right), \quad \gamma > 0$$

proposed by Ghai et al. (2019). The mirror descent update induced by this class interpolates between a Euclidean update and the multiplicative weights update. As γ approaches 0, hypentropy acts similar to entropy, and as γ grows large, hypentropy acts as a scaled Euclidean distance.

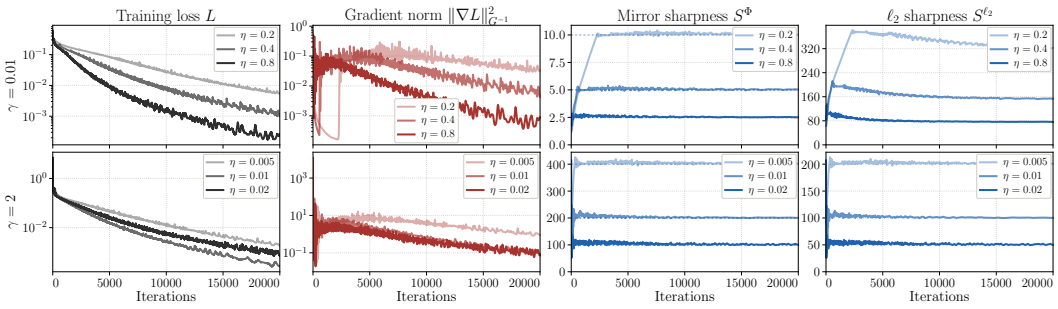


Figure 1: CNN with a hypentropy potential. Columns show training loss, norm of the loss gradient with respect to G, mirror sharpness, and ℓ_2 sharpness. The $2/\eta$ thresholds are indicated by the horizontal dashed lines. Top: $\gamma = 0.01$, bottom: $\gamma = 2$.

In Figure 1, we see that for a variety of step sizes η, the mirror sharpness S^Φ stabilizes just over $2/\eta$. This is consistent with previous findings for full-batch gradient descent in Cohen et al. (2020) and non-Euclidean steepest descent in Islamov et al. (2026). In contrast, the ℓ_2 sharpness does not stabilize at $2/\eta$; furthermore, in the very non-Euclidean case of $\gamma = 0.01$, the ℓ_2 sharpness appears to drift substantially even after EoS.

ℓ_p norm We consider the class of mirror potentials given by $\Phi_{\ell_p}(x) = \frac{1}{p}\|x\|_p^p$ for $p \in (1, \infty)$, whose implicit regularization properties are studied in (Azizan et al., 2022; Sun et al., 2022; Julistiono et al., 2026). This class of maps has received attention because it is implementable coordinate-wise and also has been empirically observed to generalize better than ordinary gradient descent ($p = 2$) in some settings (Sun et al., 2022).

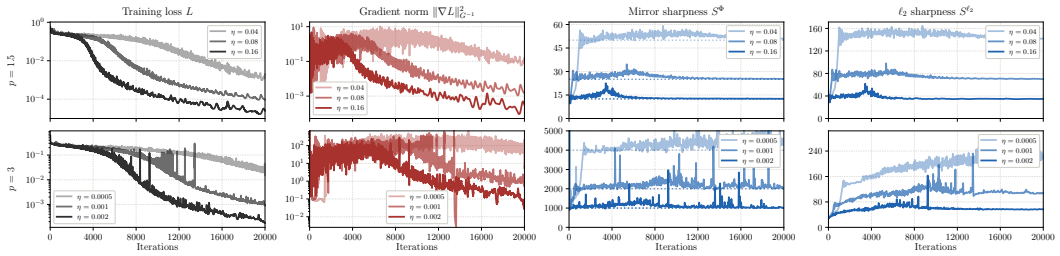


Figure 2: Vision transformer (ViT) with an ℓ_p potential. Columns show training loss, norm of the loss gradient with respect to G, mirror sharpness, and ℓ_2 sharpness. The $2/\eta$ thresholds are indicated by the horizontal dashed lines. Top: $p = 1.5$, bottom: $p = 3$.

One peculiarity of this class of maps for $p \neq 2$ is that the Hessian $\nabla^2\Phi(x)$ becomes problematic whenever $x_i = 0$, so mirror sharpness cannot be defined on this set. In practice, this is typically not a significant issue unless parameters are zero at initialization because this region is measure zero. More specifically, for $p \in (1, 2)$, Φ_{ℓ_p} fails to be C^2 here, so $\nabla^2\Phi$ is not defined. For $p > 2$, $\nabla^2\Phi(x)$ has a zero eigenvalue whenever some $x_i = 0$. Outside of this region, one may compute

$$[\nabla^2\Phi_{\ell_p}(x)]^{-1}_{ij} = \begin{cases} \frac{1}{p-1}|x_i|^{2-p} & i = j \\ 0 & i \neq j. \end{cases}$$

In general, we observe that for $p \in (1, 2)$, as $x_i \to 0$, we have $G^{-1}_{ii} \to 0$. In practice, this suppresses certain coordinates in H when computing the mirror sharpness via Lemma 2. By contrast, in the $p > 2$ regime, as $x_i \to 0$, we have $G^{-1}_{ii} \to \infty$. This *amplifies* certain coordinates when computing mirror sharpness, and thus the mirror sharpness sharply spikes when some x_i is close to zero. In our experiments, these effects are most significant near initialization.

The ill-posed nature of $\nabla^2\Phi$ does not appear to significantly empirically alter EoS dynamics, and our theoretical results remain valid away from the problematic set. We find these empirical results very interesting, and we believe that they may hint that some more "universal" mechanism is at play that manifests in the EoS phenomenon.

4.3 Mirror sharpness and ordinary sharpness

We discuss the behavior of the well-studied ℓ_2 sharpness, whose connection to generalization has been extensively studied, such as in (Kaur et al., 2023; Andriushchenko et al., 2023). Furthermore, Cohen et al. (2023) empirically showed that while Adam maintains a *preconditioned* sharpness[3], Adam has the capability to arbitrarily increases the sharpness. This fact appears connected to how the Frobenius norm of the preconditioner matrix P_t is observed to *increase* throughout training time[4], allowing $\lambda_{\max}(H)$ to increase arbitrarily while fixing $\lambda_{\max}(P_t^{-1}H)$ around Adam's corresponding stability threshold.

In the case for mirror sharpness, can we predict S^{ℓ_2} at EoS? In the most general sense where we can take any Φ, the answer is no: taking $\Phi = \frac{1}{2c}\|x\|_2^2$ yields $S^\Phi = cS^{\ell_2}$. Fixing a mirror map Φ, however, we can make the following naive estimate of the ratio S^{ℓ_2}/S^Φ:

Proposition 6 (Sharpness ratio) *Assume $S^\Phi(x) > 0$. Then,*

$$\lambda_{\min}(\nabla^2\Phi(x)) \leq S^{\ell_2}(x)/S^\Phi(x) \leq \lambda_{\max}(\nabla^2\Phi(x)).$$

For some mirror maps, we can bound $\lambda_{\max}(\nabla^2\Phi)$. For instance, in the case of the hypentropy potential $\Phi_{\text{hyp},\gamma}$, we can compute $\lambda_{\max}(\nabla^2\Phi(x)) \leq \frac{1}{\gamma}$; we empirically observe this in all of our

[3]This notion of sharpness is the same as discussed in Section 4.1, only that P may vary with time.

[4]This is, however, not universally true for any arbitrary loss, as there exists convex optimization problems such that this preconditioner norm does not monotonically increase (Reddi et al., 2018).

experiments. In these cases, if we assume that S^Φ is approximately $2/\eta$, then this implies that S^{ℓ_2} *cannot* arbitrarily increase– in contrast to Adam. In fact, as mentioned previously, in certain runs (e.g, Figure 1), we found that S^{ℓ_2} may *decrease* after the onset of EoS. In light of Proposition 6, we find it surprising that, in our experiments, the ratio S^{ℓ_2}/S^Φ remains relatively stable and occupies a very narrow range within what is predicted by Proposition 6.

5 A DUAL CENTRAL FLOW FOR MIRROR DESCENT

In the remainder of this paper, we attempt to study mirror descent through the *central flow* framework (Cohen et al., 2024), which is an ODE that describes a *time-averaged* (i.e., locally smooth) trajectory $\bar{x}(t)$ of iterates at EoS. The Riemannian gradient flow fails to be a good approximation at EoS because it does not fix the sharpness at around $2/\eta$; thus, the Riemannian gradient flow significantly diverges from the mirror descent iterates after the onset of EoS. We show this empirically in Section 5.2.

While this ODE may be of independent interest as a more tractable object to study than mirror descent, our main goal is to understand the extent to which the self-stabilization theory holds in the non-Euclidean setting of mirror descent, so we use this picture as an ansatz for our central flow. We find that such a framework is not necessary well posed without adding extra constraints on the *third-order* interactions between the loss and Φ; such a result is interesting because it raises a challenge in extending the self-stabilization argument to mirror descent (Arora et al., 2022; Damian et al., 2022; Cohen et al., 2024; Liao et al., 2026).

5.1 DERIVING THE CENTRAL FLOW

We will assume that L, Φ are C^4. We will restrict our central flow in the case when only one eigenvalue of H is at (approximately) $2/\eta$. Let $u = u(\bar{x})$ be the top eigenvector of (H, G) that solves the maximization problem defined in Definition 1. Even in this case, there are complexities caused by the non-Euclidean nature of mirror descent.

The main idea behind central flows is modeling the iterates x_t as an oscillation about a *time-averaged*[5] $\bar{x}(t)$ along the unstable eigenvector $u(t)$. More concretely, Cohen et al. (2024) models $x_t = \bar{x}(t) + \xi(t)u(t)$.[6] It is unclear to what extent such a phenomenon holds for the case of mirror descent *especially* because neither the primal nor the dual space is implicitly endowed with an ℓ_2 geometry in general. To make progress towards answering this question, let us naively assume that such a model is valid.

Specifically, we model oscillations in the *dual* space via $z_t = \bar{z}(t) + \xi(t)r(t)$, where $r = Gu$ is the pushforward of u under the mirror map $\nabla\Phi$ (i.e., the natural way tangent vectors in the primal space are mapped to tangent vectors in the dual space).[7] Furthermore, we impose $\mathbb{E}[\xi(t)] = 0$ and $\mathbb{E}[\xi(t)r(t)] \approx \mathbb{E}[\xi(t)]r(t) = 0$. We believe that the dual is a more natural choice to model than the primal because the primal update introduces additional non-linearities via the mirror map. Recall that the dual update is given by:

$$z_{t+1} - z_t = -\eta(\nabla L \circ \nabla\Phi^*)(z_t).$$

We abbreviate $F := \nabla L \circ \nabla\Phi^*$. With these assumptions and regularity where appropriate, approximating $d\bar{z}/dt \approx \bar{z}_{t+1} - \bar{z}_t$ and expanding all relevant terms to $O(\xi^3)$ yields the *dual central flow* ODE:

$$\frac{d\bar{z}}{dt} = -\eta\left(F(\bar{z}) + \frac{\sigma^2}{2}D^2F(\bar{z})[r, r]\right), \tag{DCF}$$

[5]This expectation represents a smoothening/averaging of $\xi(t)r(t)$, though this is left deliberately vague because it is not easy to tell what time scale to average the terms with. We lift the same assumptions about the expectation operator $\mathbb{E}[\cdot]$ from Cohen et al. (2024, Appendix A.1.5), where one may find a more detailed discussion.

[6]Even in the case of gradient descent, such a model is a simplification, as $x_t - \mathbb{E}[x_t]$ does not necessarily have to lie in the span of $u(t)$.

[7]More formally, we view u as an element of the tangent space T_xE. We can rewrite the differential $D(\nabla\Phi)_x = \nabla^2\Phi(x) = G$, so $r = Gu$ is the push-forward of u and lives in $T_{\nabla\Phi(x)}E^*$.

where $\sigma^2 := \mathbb{E}[\xi(t)^2]$. By further constraining $\widetilde{S}^\Phi(\bar{z}) := (S^\Phi \circ \nabla\Phi^*)(\bar{z}) = 2/\eta$ to model iterates at EoS, we can solve for a unique value of σ^2 to pin down the trajectory:

$$\sigma^2 = \frac{-2\langle \nabla \widetilde{S}^\Phi(\bar{z}), F(\bar{z})\rangle}{\langle \nabla \widetilde{S}^\Phi(\bar{z}), D^2F(\bar{z})[r,r]]\rangle}.$$

As much of the details of this derivation is similar to Cohen et al. (2024), we defer a complete derivation to Section B.1. We observe that in the case where $\Phi = \frac{1}{2}\|x\|_2^2$, we retrieve the central flow for ordinary gradient descent in Cohen et al. (2024).

Finally, we retrieve a flow for the primal quantity $\bar{x}(t) := \nabla\Phi^*(\bar{z}(t))$. Again, as $d\bar{x}/dt$ is a tangent vector on the primal space, we use the pullback G^{-1} to relate it with $d\bar{z}/dt$:

$$\frac{d\bar{x}}{dt} = G^{-1}\frac{d\bar{z}}{dt} = -\eta G^{-1}\left(\nabla L(\bar{x}) + \frac{\sigma^2}{2}\left(\nabla^3 L(\bar{x})[u,u] - HG^{-1}\nabla^3\Phi(\bar{x})[u,u]\right)\right) \tag{PCF}$$

We call this the *primal central flow*.[8] In the second equality, we re-express the dual quantities in Equation (DCF) in primal quantities.

5.2 Empirical verification

We verify empirically that Equation (DCF) describes iterates at the edge of stability more accurately than the Riemannian gradient flow in the dual space, which is given by

$$\frac{d\bar{z}}{dt} = -\eta\nabla L(\bar{x}). \tag{RGF}$$

For comparison, we also plot the trajectory of gradient flow (i.e., where we let G be the identity) to show that it diverges sharply from the mirror descent trajectory.

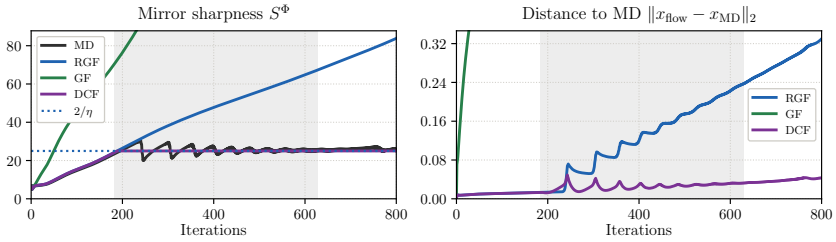


Figure 3: CNN with an $\ell_{1.5}$ potential, GeLU activations, MSE loss, and $\eta = 0.08$ on a subset of CIFAR-10. Along with the mirror descent iterates (MD), we track the Riemannian gradient flow (RGF), gradient flow (GF), and our dual central flow (DCF). We indicate, in gray, the timesteps where the DCF iterates only have one unstable eigenvector with eigenvalue greater than $1.95/\eta$. Left: Mirror sharpness. Right: Distance to mirror descent iterates.

We find that the dual central flow tracks the mirror descent iterates fairly well at EoS, improving over the Riemannian gradient flow, even when there is more than one unstable eigenvalue. Furthermore, our dual central flow straddles the $2/\eta$ threshold for mirror sharpness very closely. We also remark that discretizing Equation (DCF) is non-trivial since we require that the mirror sharpness is constrained *exactly* at $2/\eta$. We comment more on this in Section B.2. Further experiments can be found in Appendix G.

5.3 The central flow encodes self-stabilization… sometimes

Let us make more explicit how the central flow Equation (DCF) "encodes" the past result of Damian et al. (2022) explaining why ℓ_2 sharpness stabilizes at $2/\eta$ in the case of gradient descent. Assuming a well-separated gap between the first and second eigenvalues, one can solve explicitly for $\frac{d}{dt}S^\Phi$:

[8] Note however, that $\bar{x} \neq \mathbb{E}[x_t]$ in general due to the non-linearities of $\nabla\Phi^*$.

Proposition 7 *Let $\bar{z}$ evolve according to Equation (DCF). Let L, Φ be C^4 and $\nabla^2\Phi \succ 0$ for all x along $\bar{x}$. Then,*

$$\frac{d}{dt}\widetilde{S}^{\Phi}(\bar{z}) = \eta\langle\nabla\widetilde{S}(\bar{z}), -F(\bar{z})\rangle - \frac{\eta\sigma^2}{2}\langle\nabla\widetilde{S}(\bar{z}), D^2F(\bar{z})[r,r]\rangle,$$

where $r = Gu$ and u is the top eigenvector of (H, G).

The first term admits a very simple interpretation as the *progressive sharpening coefficient*. In dual space, the gradient update is given by $-F(z)$, so $\widetilde{S}$ increases if and only if $\langle\nabla\widetilde{S}(z), -F(z)\rangle > 0$.

Definition 8 *The* progressive sharpening coefficient *is given by*

$$\alpha^{\Phi}(z) := -\langle\nabla\widetilde{S}(z), F(z)\rangle.$$

In the case of gradient descent, α^{Φ} is assumed to be positive by Damian et al. (2022) and verified empirically.

To counteract the positive contribution to S^{Φ}, Damian et al. (2022) explains that the oscillations along the top eigenvector "self-stabilizes" the sharpness. In the context of gradient descent, the strength of this restoring force is quantified by $\beta^{\ell_2}(x) = \|\nabla S^{\ell_2}(x)\|_2^2 \geq 0$. As we may expect, the second term in Proposition 7 reveals the mirror descent analogue of β_{ord}, as $D^2F(\bar{z})[r,r]$ reduces to $\nabla S(x)$ in the case of gradient descent.

Definition 9 *The* mirror stabilization coefficient $\beta^{\Phi}(z)$ *is given by:*

$$\beta^{\Phi}(z) := \langle\nabla\widetilde{S}(\bar{z}), D^2F(\bar{z})[r,r]\rangle,$$

where $r = Gu$ and u is the top eigenvector of (H, G). Equivalently, by Proposition 16,

$$\beta^{\Phi}(\nabla\Phi(x)) = \|\nabla S\|_{G^{-1}}^2 + \langle G^{-1}\nabla S, (SI - HG^{-1})\nabla^3\Phi[u,u]\rangle.$$

Using the primal expansion of β^{Φ}, we observe that for quadratic potentials, $\beta^{\Phi} \geq 0$. However, this need not be true in general; we show the "geometry correction" term can be made arbitrarily large:

Lemma 10 (β^{Φ} can be negative) *Fix $x^\star \in E$ for $n > 1$, and some mirror map Φ with $\nabla^3\Phi(x^\star) \neq 0$. There exists some L with bounded $\|\nabla^3 L[\cdot,\cdot,\cdot]\|$ such that $\beta^{\Phi}(x^\star) < 0$.*

Why is this problematic? If we assume that progressive sharpening happens with $\alpha^{\Phi} > 0$, having $\beta^{\Phi} < 0$ *forces* $\widetilde{S}^{\Phi}(\bar{z})$ to increase via Proposition 7. Thus, in this case, when we solve for the value of $\sigma^2 = 2\alpha^{\Phi}/\beta^{\Phi}$, we would find that σ^2, which represents the variance of the oscillations, *is negative* and thus the central flow fails to be physically relevant. Thus, the universal self-stabilization mechanism via oscillations is not currently sufficient to explain EoS in mirror descent.

In light of this negative result, we believe it is quite difficult, given the current theory, to argue for a self-stabilization result without enforcing substantial conditions that are not easy to interpret. While it may be possible to sidestep the obstacle mentioned in Lemma 10 by arguing that the central flow is repelled from regions where $\beta^{\Phi} < 0$, we find this difficult because even arguing why $\alpha^{\Phi} > 0$ holds is a major open question for ordinary gradient descent. In our experiments in Appendix G, we plot values of $\alpha^{\Phi}, \beta^{\Phi}$ along the dual central flow trajectory, and we find that $\beta^{\Phi} > 0$ always.

6 Conclusion and Future Work

In this work, we studied edge of stability behavior in the non-Euclidean setting of mirror descent, by introducing a mirror sharpness that recovers the familiar $2/\eta$ threshold for the edge of stability. Empirically, we show that the EoS phase manifests through various step sizes, architectures, and mirror maps. However, through the framework of central flows, we identify that a "self-stabilization" picture in the dual space does not sufficiently explain why mirror descent should stabilize due to third-order interactions between the mirror potential and the loss.

It remains unclear if mirror descent "universally" self-stabilizes assuming progressive sharpening and an arbitrary loss. A positive answer would likely require an argument that the scenarios in Lemma 10 are pathological, or substantially better understandings of what kinds of conditions realistic losses satisfy. Towards a negative answer, one may try to exhibit an example of successful optimization where $\beta^{\Phi}(z) < 0$ and the mirror sharpness climbs arbitrarily.

AI USE STATEMENT

In this work, we used generative AI tools for:

- implementing methods. Almost all code was generated by AI tools; however, the first author has carefully checked the code and takes full responsibility. In reviewing the code, we found that one implementation detail was autonomously added by the AI. Specifically, the "floor" for the ℓ_3 mirror potential that we kept is described in Appendix E.1, and it was implemented independently of our instructions. In addition, we used AI tools to prototype our choices of hyperparameters for numerical computation (e.g., ODE step size) to determine if they had a significant effect on the error.
- assistance in the writing of proofs. More specifically, we used AI for some preliminary computations of values to those in (and similar to those in) Appendix D.9. Such prompts were usually of the flavor: "express $D^2F[r, r]$ in terms of primal quantities $H(x), G(x), L(x), u(x)$"; moreover, we independently derived them, not necessarily with the same method as the AI. We have also used AI to proofread our proofs, wherein the AI found some algebra errors and clerical mistakes.

We have not used generative AI tools for:

- helping develop theoretical models or conceptual frameworks;
- formulating mathematical claims;
- providing critical ingredients for proving mathematical claims;
- proposing or refining hypotheses;
- supporting qualitative and thematic data analysis;
- interpreting results.

The following items are not applicable to this work:

- generating synthetic data sets;
- assisting with translation;
- cleaning and reformatting dataset.

Additionally, we used generative AI tools for:

- plotting;
- suggesting experiment parameters (specifically the choice of η);
- searching for related literature. Citations were not made using generative AI.
- proofreading the text.

We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All assumptions and definitions necessary for the proofs are included in the main paper and appendix, or otherwise properly referenced with a citation. We have detailed our implementations in Appendix E, and for each mirror map, we run experiments using seeds 1 through 6 to ensure reproducibility. Wherever possible, we have leveraged existing code, rather than rewriting. We plan to make code available and open-source at a later date.

REFERENCES

Kwangjun Ahn, Sebastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A Modern Look at the Relationship between Sharpness and Generalization. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 840–902. PMLR, July 2023.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding Gradient Descent on the Edge of Stability in Deep Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 948–1024. PMLR, June 2022.

Navid Azizan and Babak Hassibi. Stochastic Gradient/Mirror Descent: Minimax Optimality and Implicit Regularization. In *International Conference on Learning Representations*, September 2018.

Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic Mirror Descent on Overparameterized Nonlinear Models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7717–7727, December 2022. ISSN 2162-2388. doi: 10.1109/TNNLS.2021.3087480.

Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28:2131–2151, July 2018. doi: 10.1137/17M1138558.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *International Conference on Learning Representations*, October 2020.

Jeremy Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive Gradient Methods at the Edge of Stability. In *NeurIPS 2023 Workshop Heavy Tails in Machine Learning*, November 2023.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding Optimization in Deep Learning with Central Flows. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability. In *The Eleventh International Conference on Learning Representations*, September 2022.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated Gradient Meets Gradient Descent. In *ALT*. arXiv, February 2019. doi: 10.48550/arXiv.1902.01903.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning Dynamics of Deep Matrix Factorization Beyond the Edge of Stability. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing Implicit Bias in Terms of Optimization Geometry. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1832–1841. PMLR, July 2018.

Suriya Gunasekar, Blake Woodworth, and Nathan Srebro. Mirrorless Mirror Descent: A Natural Derivation of Mirror Descent. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 2305–2313. PMLR, March 2021.

Rustem Islamov, Michael Crawshaw, Jeremy Cohen, and Robert M. Gower. Non-Euclidean Gradient Descent Operates at the Edge of Stability. In *Forty-Third International Conference on Machine Learning*, June 2026.

Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing Attention with Mirror Descent: Generalized Max-Margin Token Selection. *Journal of Machine Learning Research*, 27(36):1–61, 2026. ISSN 1533-7928.

Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. On the Maximum Hessian Eigenvalue and Generalization, May 2023.

Andrew V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001. doi: 10.1137/S1064827500366124.

Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient Descent Monotonically Decreases the Sharpness of Gradient Flow Solutions in Scalar Networks and Beyond. June 2023.

Serge Lang. *Real and Functional Analysis*, volume 142 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 3 edition, 1993. ISBN 0-387-94001-4. doi: 10.1007/978-1-4612-0897-6.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, 2 edition, 2013. ISBN 978-1-4419-9981-8.

Fangshuo Liao, Afroditi Kolomvaki, and Anastasios Kyrillidis. SGD at the Edge of Stability: The Stochastic Sharpness Gap, April 2026.

Liming Liu, Zixuan Zhang, Simon Shaolei Du, and Tuo Zhao. A Minimalist Example of Edge-of-Stability and Progressive Sharpening. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

LucidRains. Vision transformer – PyTorch, 2024.

Lachlan Ewen MacDonald, Hancheng Min, Leandro Palma, Salma Tarmoun, Ziqing Xu, and Rene Vidal. Convergence Rates for Gradient Descent on the Edge of Stability for Overparametrised Least Squares. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

Pierre Marion. Edge flow: A tractable and predictive continuous-time model for gradient descent at the edge of stability, 2026.

Arkadi S. Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the Convergence of Adam and Beyond. In *International Conference on Learning Representations*, February 2018.

Eric Regis and Sinho Chewi. Rod Flow: A Continuous-Time Model for Gradient Descent at the Edge of Stability, February 2026a.

Eric Regis and Sinho Chewi. A Rod Flow Model for Adam at the Edge of Stability, May 2026b.

Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror Descent Maximizes Generalized Margin and Can Be Implemented Efficiently. In *Advances in Neural Information Processing Systems*, October 2022.

Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24(1), January 2023. ISSN 1532-4435.

Shira Vansover-Hager, Matan Schliserman, Ofir Schlisselberg, and Tomer Koren. Mirror Descent Beyond Euclidean Stability: An Exponential Separation in Initialization Sensitivity, June 2026.

Geonhui Yoo, Minhak Song, and Chulhee Yun. Understanding Sharpness Dynamics in NN Training with a Minimalist Example: The Effects of Dataset Difficulty, Depth, Stochasticity, and More. In *Forty-Second International Conference on Machine Learning*, June 2025.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding Edge-of-Stability Training Dynamics with a Minimalist Example. In *The Eleventh International Conference on Learning Representations*, September 2022.

A NOTATION

Notation To make the distinction between primal coordinates and dual coordinates more clear, we will work over the parameter space $E = \mathbb{R}^n$ with dual E^*. Let $L : E \to \mathbb{R}$ be a C^3 loss function throughout this paper. We will use $H(x)$ to denote the Hessian matrix of L. We use $A \succ 0$ to denote that A is positive definite. For a matrix A, let $\lambda_{\max}(A)$ to denote the largest eigenvalue of A, and let $\lambda_{\min}(A)$ denote the smallest. We use $\|x\|_2^2 = \sum x_i^2$ to denote the squared ℓ^2 norm.

Derivatives In deriving the central flows for mirror descent, we will often use higher-order derivatives that are multilinear. We try to keep such notation at a minimum, opting to use the more familiar gradient ∇^k where possible. We refer the reader to (Lang, 1993, Chapter XIII, §6) for explicit constructions of these objects.

Let $F : E \to E$ be a C^k function (such as the loss L or mirror potential Φ). We use $D^k(x) : E^k \to E$ to denote the derivative viewed as a k-linear map, which is symmetric since we assume F is C^k. We will frequently invoke Taylor's theorem; a proof can be found in (Lang, 1993, Chapter XIII, §6).

Theorem 11 (Taylor's theorem) *Let $f : E \to E$ be C^p. Denote by $y^{(k)}$ the k-tuple $[y, \dots, y]$. Then*

$$f(x+y) = f(x) + \sum_{k=1}^{p-1} \frac{D^k f(x)[y^{(k)}]}{k!} + O(\|y\|^p).$$

In the case when $f : E \to \mathbb{R}$ (as in the case of the loss L or mirror potential Φ) and f is C^k, we opt to use the more familiar gradient notation ∇^k. We will only invoke the $k \le 3$ cases. As the $k = 1$ case corresponds to the gradient and the $k = 2$ case corresponds to the Hessian, we will describe the $k = 3$ case. Let $\nabla^3 f(x) : E^2 \to E^*$ be the symmetric, bilinear operator that satisfies

$$\langle \nabla^3 f(x)[u, v], w \rangle = D^3 f(x)[u, v, w]$$

for all $u, v, w \in E$.

B MORE ON THE CENTRAL FLOW

B.1 CENTRAL FLOW DERIVATION

We will consider the appropriate regularity $L, \Phi \in C^4$. We will work in the dual coordinates $z = \nabla\Phi(\bar{x})$. Recalling $F = \nabla\mathcal{L} \circ \nabla\Phi^*$, we may write the dual mirror descent update:

$$z_{k+1} = z_k - \eta F(z_k),$$

We model iterates z_k as oscillating along the direction r, which is the pushforward of u through $\nabla\Phi$, given by $r = Gu$. Then, $r \in E^*$ satisfies $\langle r, \nabla^2\Phi^*(z)r \rangle = 1$, and we model $z(t) := \bar{z}(t) + \xi(t)r(t)$ with $\mathbb{E}[\xi(t)] = 0$. Let $\sigma^2(t) := \mathbb{E}[\xi(t)^2]$. One may Taylor expand F at $\bar{z}$ along the r direction:

$$F(\bar{z} + \xi r) = F(\bar{z}) + \xi DF(\bar{z})[r] \; + \frac{\xi^2}{2} D^2 F(\bar{z})[r, r] + O(\xi^3).$$

After averaging, we have:

$$\begin{aligned}\mathbb{E}\big[F(\bar{z} + \xi r)\big] &= \mathbb{E}\big[F(\bar{z}) + \xi DF(\bar{z})[r] \; + \frac{\xi^2}{2} D^2 F(\bar{z})[r, r]\big] + O(\mathbb{E}\big[|\xi|^3\big]) \\ &\approx F(\bar{z}) + \frac{\sigma^2}{2} D^2 F(\bar{z})[r, r] + O(\mathbb{E}\big[|\xi|^3\big]).\end{aligned}$$

Here, we take some care in handling the second term

$$\mathbb{E}[\xi DF(\bar{z})[r]] = DF(\bar{z})[\mathbb{E}[\xi r]] \approx DF(\bar{z})[\mathbb{E}[\xi] r] = 0,$$

as $DF(\bar{z})$ acts linearly on r. At the time t that corresponds to an iterate z_k, we may approximate

$$\frac{d\bar{z}}{dt} \approx \bar{z}_{k+1} - \bar{z}_k = -\eta \mathbb{E}[F(\bar{z}_k + \xi r)].$$

Combining the two equations yields the dual central flow ODE:

$$\frac{d\bar{z}}{dt} = -\eta \left(F(\bar{z}) + \frac{\sigma^2}{2} D^2 F(\bar{z})[r, r] \right).$$

Define $\widetilde{S} = S \circ \nabla\Phi^*$.

In the continuous-time limit, to solve for $\sigma^2(t)$, we may simply assume that the iterates $\bar{z}$ are already at the edge of stability:

$$\widetilde{S}(\bar{z}(t)) = \frac{2}{\eta},$$

and therefore to ensure that $\bar{z}$ stays here, we enforce

$$\frac{d}{dt}(S \circ \nabla\Phi^*(\bar{z})) = 0$$

Then, by the chain rule, the LHS can be re-expressed as $\langle \nabla\widetilde{S}, \frac{d\bar{z}}{dt} \rangle$, which implies that

$$\left\langle \nabla\widetilde{S}(\bar{z}), -\eta \left(F(\bar{z}) + \frac{\sigma^2}{2} D^2 F(\bar{z})[r, r] \right) \right\rangle = 0.$$

Solving for σ^2 yields:

$$\sigma^2 = \frac{-2\langle \nabla\widetilde{S}(\bar{z}), F(\bar{z}) \rangle}{\langle \nabla\widetilde{S}(\bar{z}), D^2 F(\bar{z})[r, r]] \rangle}.$$

Observe that the numerator is exactly $2\alpha^\Phi$ as $\nabla\widetilde{S}(\bar{z}) = G^{-1}\nabla S(\bar{x})$ by Lemma 13 and the denominator is equal to β^Φ by Proposition 16. To retrieve Equation (PCF), we directly apply Theorem 15.

B.2 Discretizing the central flow

In practice, however, we cannot assume that $\widetilde{S}(\bar{z}(t))$ will rise continuously until it hits $2/\eta$ due to discretization effects. Let us suppose that the top eigenvalue of (H, G) is close to, but not over, $2/\eta$.

Define the sharpness residue $A(z) := 2/\eta - \widetilde{S}(z)$. Let us write down the following *differential complementarity problem* (DCP) in ζ:

$$\frac{d\bar{z}}{dt} = -\eta \left(F(\bar{z}) + \frac{\zeta}{2} D^2 F(\bar{z})[r, r] \right), \qquad \text{where} \quad 0 \preceq \zeta \perp A(\bar{z}) \succeq 0.$$

We use $\zeta \perp A(\bar{z})$ (where $\perp$ denotes complementarity) to enforce the condition that $\zeta A(\bar{z}) = 0$.

Discretizing this DCP is discussed in good detail in Cohen et al. (2024, Appendix A.6.1), and we adapt their method. While this is not practical in general in the case of multiple unstable eigenvectors as Cohen et al. (2024) describes, for our purposes of modeling one unstable eigenvector, it is satisfactory. Define the linearization of A:

$$\begin{aligned} A^{\text{lin}}(z) &:= A(\bar{z}) + \langle \nabla A(\bar{z}), z - \bar{z} \rangle \\ &= 2/\eta - \widetilde{S}(\bar{z}) - \langle \nabla\widetilde{S}(\bar{z}), z - \bar{z} \rangle \end{aligned}$$

Let t' denote the current time step, and $t' + \epsilon$ denote the next timestep. We would like to enforce the DCP at $\bar{z}^{(t'+\epsilon)} = \bar{z}^{(t')} + \epsilon \frac{d\bar{z}^{(t')}}{dt}$. Thus,

$$\begin{aligned} A^{\text{lin}}(\bar{z}^{(t'+\epsilon)}) &= 2/\eta - \widetilde{S}(\bar{z}^{(t')}) - \left\langle \nabla\widetilde{S}(\bar{z}^{(t')}), \epsilon \frac{d\bar{z}^{(t')}}{dt} \right\rangle. \\ &= 2/\eta - \widetilde{S}(\bar{z}^{(t')}) - \eta\epsilon\alpha^\Phi(\bar{z}^{(t')}) + \frac{\eta\epsilon\zeta}{2}\beta^\Phi(\bar{z}^{(t')}), \end{aligned}$$

where the last equality follows from Proposition 7. Thus, enforcing the complementarity condition at $\bar{z}^{(t'+\epsilon)}$ yields

$$\zeta = 0 \text{ or } \zeta = \frac{2}{\beta^\Phi(z^{(t')})} \left(\alpha^\Phi(z^{(t')}) - \frac{2/\eta - \widetilde{S}(z^{(t')})}{\eta\epsilon} \right).$$

Enforcing the complementarity and $\zeta \geq 0$ and $A^{\text{lin}}(\bar{z}^{(t'+\epsilon)}) \geq 0$ yields:

$$\zeta = \max\left(0, \frac{2}{\beta}\left(\alpha^{\Phi}(z^{(t')}) - \frac{2/\eta - \widetilde{S}(z^{(t')})}{\eta\epsilon}\right)\right).$$

This becomes our choice of ζ, and we implement the Euler timestep

$$\bar{z}^{(t'+\epsilon)} \leftarrow \bar{z}^{(t')} - \epsilon\left(F(\bar{z}^{(t')}) + \frac{\zeta}{2}D^2F(\bar{z}^{(t')})[r^{(t')}, r^{(t')}]\right).$$

B.3 Comparison against Cohen et al. (2024)

We compare our model against Cohen et al. (2024)'s central flow model for gradient descent. Surprisingly, even though we only account for the top eigenvector, our model provides very similar results. We hypothesize that this might be because we run 4 timesteps per mirror descent iteration, so our solver can "take turns" and push down different eigenvalues on each timestep.

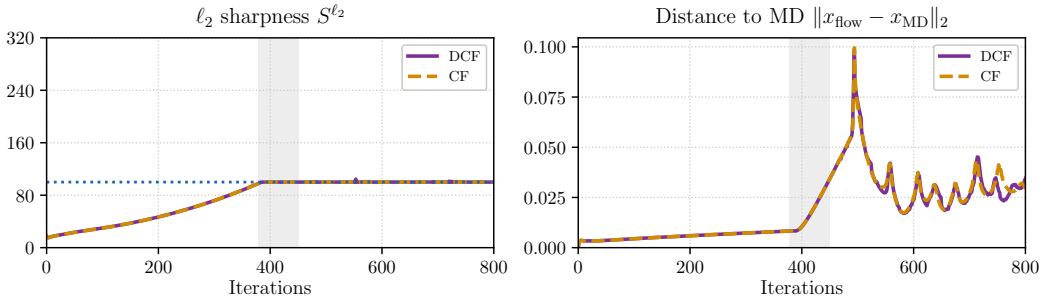


Figure 4: MLP with a MSE loss on a subset of CIFAR-10; gradient descent; $\eta = 0.02$. We compare our dual central flow (DCF) against the central flow (CF) of Cohen et al. (2024). We indicate, in gray, the timesteps where the DCF iterates only have one unstable eigenvector with eigenvalue greater than $1.95/\eta$.

C A spectral interpretation of the stabilization coefficient

Proposition 12 (Expansion in the generalized eigenbasis) *Let u_i be an orthonormal basis of generalized eigenvectors satisfying*

$$Hu_i = \lambda_i G u_i, \qquad \langle u_i, Gu_j\rangle = \delta_{ij},$$

with eigenvalues ordered so that $\lambda_1 = S$. Then,

$$\beta^{\Phi}(x) = \|\nabla S^{\Phi}(x)\|^2_{G^{-1}} + \sum_i (S^{\Phi} - \lambda_i)\langle \nabla S^{\Phi}, u_i\rangle\langle \nabla^3\Phi[u,u], u_i\rangle.$$

While we can guarantee $\|\nabla S^{\Phi}(x)\|^2_{G^{-1}}$ is never negative, it appears difficult to control the geometry correction caused by the alignment of ∇S and $\nabla^3\Phi[u,u]$. Interestingly, the eigenvalues at the edge of stability do not contribute to the geometry correction, and *smaller* eigenvalues are weighted heavier in the geometry correction. This suggests that understanding EoS dynamics in mirror descent may require tracking the evolution of the entire spectrum during training, rather than only the largest eigenvalue.

D Deferred proofs and some identities

D.1 Proof of Lemma 2

Via a homogeneity argument,

$$\begin{aligned} S(x) &= \max_{\langle u, Gu\rangle = 1} \langle u, Hu\rangle \\ &= \max_{u \neq 0} \frac{\langle u, Hu\rangle}{\langle u, Gu\rangle}. \end{aligned}$$

Since $G \succ 0$, let $v := G^{1/2}u$. Then, one can rewrite $\langle u, Gu\rangle = \langle v, v\rangle$ and $\langle u, Hu\rangle = \left\langle v, G^{-1/2}HG^{-1/2}v\right\rangle$. Therefore,

$$S(x) = \max_{v \neq 0} \frac{\left\langle v, G^{-1/2}HG^{-1/2}v\right\rangle}{\langle v, v\rangle}.$$

Since $G^{-1/2}HG^{-1/2}$ is symmetric, Rayleigh–Ritz gives

$$S^{\Phi}(x) = \lambda_{\max}(G^{-1/2}HG^{-1/2}).$$

This shows the first claim. Observe that the top eigenvector v_1 solves

$$S^{\Phi}v_1 = G^{-1/2}HG^{-1/2}v_1.$$

Let $u_1 = G^{-1/2}v_1$, and rewrite

$$S^{\Phi}G^{1/2}u_1 = G^{-1/2}Hu_1.$$

As $G \succ 0$, we find that $SGu_1 = Hu_1$, showing that mirror sharpness is the top generalized eigenvalue of (H, G).

D.2 Proof of Proposition 3

Let $\Delta_t := x_{t+1} - x_t$. Taylor expanding L gives

$$L(x_{t+1}) - L(x_t) = \langle \nabla L(x_t), \Delta_t\rangle + \frac{1}{2}\langle \Delta_t, H(x_t)\Delta_t\rangle + O(\|\Delta_t\|^3),$$

and we will bound the terms on the RHS. Writing the eq. (MD) update gives:

$$\nabla\Phi(x_{t+1}) - \nabla\Phi(x_t) = -\eta\nabla L(x_t).$$

Taking the inner product with Δ_t, we obtain

$$\langle \nabla L(x_t), \Delta_t\rangle = -\frac{1}{\eta}\left(D_{\Phi}(x_{t+1}, x_t) + D_{\Phi}(x_t, x_{t+1})\right).$$

Note that $D_{\Phi}(x_{t+1}, w_t)$ and $D_{\Phi}(x_t, x_{t+1})$ have the same Taylor expansion up to order two:

$$\begin{aligned} D_{\Phi}(x_{t+1}, x_t) &= \frac{1}{2}\langle \Delta_t, G(x_t)\Delta_t\rangle + O(\|\Delta_t\|^3), \\ D_{\Phi}(x_t, x_{t+1}) &= \frac{1}{2}\langle \Delta_t, G(x_t)\Delta_t\rangle + O(\|\Delta_t\|^3). \end{aligned}$$

Thus, one can rewrite:

$$\langle \nabla L(x_t), \Delta_t\rangle = -\frac{2}{\eta}D_{\Phi}(x_{t+1}, x_t) + O(\|\Delta_t\|^3).$$

By the definition of mirror sharpness,

$$\langle \Delta_t, H(x_t)\Delta_t\rangle \leq S(x_t)\langle \Delta_t, G(x_t)\Delta_t\rangle.$$

To compute $\langle \Delta_t, G(x_t)\Delta_t\rangle$, we can Taylor expand $\Phi(x_{t+1})$ about x_t:

$$\Phi(x_{t+1}) = \Phi(x_t) + \langle \nabla\Phi(x_t), \Delta_t\rangle + \frac{1}{2}\langle \Delta_t, G(x_t)\Delta_t\rangle + O(\|\Delta_t\|^3).$$

Rearranging and substituting the Bregman divergence $D_{\Phi}(x_{t+1}, x_t)$ yields:

$$\frac{1}{2}\langle \Delta_t, G(x_t)\Delta_t\rangle = D_{\Phi}(x_{t+1}, x_t) + O(\|\Delta_t\|^3).$$

Combining these estimates yields the result.

D.3 Proof of Theorem 4

Write $z_t := \nabla\Phi(x_t)$. Since $\nabla L(x) = A(x - x^\star)$, the mirror descent update in dual coordinates is

$$z_{t+1} = z_t - \eta A(x_t - x^\star) = z_t - \eta A\big(\nabla\Phi^*(z_t) - \nabla\Phi^*(z^\star)\big).$$

The key observation is that, because the gradient of the quadratic objective is linear, this vector field becomes a Euclidean gradient after a linear change of coordinates. Define

$$y_t := A^{-1/2}(z_t - z^\star), \qquad \text{such that} \qquad z_t = z^\star + A^{1/2}y_t.$$

Multiplying the dual update by $A^{-1/2}$ gives

$$y_{t+1} = y_t - \eta A^{1/2}\big(\nabla\Phi^*(z^\star + A^{1/2}y_t) - x^\star\big).$$

On the other hand, differentiating $\widehat{L}$ yields

$$\nabla\widehat{L}(y) = A^{1/2}\big(\nabla\Phi^*(z^\star + A^{1/2}y) - x^\star\big).$$

Thus,

$$y_{t+1} = y_t - \eta\nabla\widehat{L}(y_t).$$

D.4 Proof of Proposition 5

Differentiating $\widehat{L}$ twice yields

$$\nabla^2\widehat{L}(y) = A^{1/2}\nabla^2\Phi^*(z^\star + A^{1/2}y)A^{1/2}.$$

Note that $\nabla^2\Phi^*(z) = (\nabla^2\Phi(x))^{-1} = G^{-1}$, which follows from differentiating both sides of $\nabla\Phi^*(\nabla\Phi(x)) = x$. Thus, we may rewrite this as

$$\nabla^2\widehat{L}(y) = A^{1/2}G^{-1}A^{1/2}$$

We claim that the eigenvectors $\tilde{v}_i$ of $\nabla^2\widehat{L}(y)$ are in bijection with the generalized eigenvectors v_i of (H, G) via $\tilde{v}_i \mapsto A^{1/2}v_i$; furthermore, their eigenvalues are preserved. Indeed,

$$A^{1/2}G^{-1}A^{1/2}\tilde{v}_i = \lambda_i\tilde{v}_i \quad\iff\quad G^{-1}Av_i = \lambda_i v_i \quad\iff\quad Av_i = \lambda_i Gv_i.$$

This proves the claim by taking the top eigenvalue.

D.5 Proof of Proposition 6

The proof follows from several applications of Rayleigh–Ritz. Throughout these proofs, let u be the top generalized eigenvalue of (H, G), and let v be the top eigenvalue of H. For the first inequality:

$$\begin{aligned} S^{\ell_2} &= \max_{w\neq 0}\frac{\langle w, Hw\rangle}{\langle w, w\rangle} \\ &\geq \frac{\langle u, Hu\rangle}{\langle u, u\rangle} \\ &= S^{\Phi}\frac{\langle u, Gu\rangle}{\langle u, u\rangle} \\ &\geq S^{\Phi}\lambda_{\min}(G). \end{aligned}$$

For the second inequality:

$$\begin{aligned} S^{\Phi} &= \max_{w\neq 0}\frac{\langle w, Hw\rangle}{\langle w, Gw\rangle} \\ &\geq \frac{\langle v, Hv\rangle}{\langle v, Gv\rangle} \\ &= S^{\ell_2}\cdot\frac{\langle v, v\rangle}{\langle v, Gv\rangle} \\ &\geq S^{\ell_2}\cdot\frac{1}{\lambda_{\max}(G)}. \end{aligned}$$

D.6 Proof of Proposition 7

By chain rule,

$$\begin{aligned}\frac{d}{dt}S^\Phi &= \left\langle \frac{d\bar{z}}{dt}, \nabla\widetilde{S}^\Phi(\bar{z})\right\rangle \\ &= \left\langle -\eta\left(F(\bar{z}) + \frac{\sigma^2}{2}D^2F(\bar{z})[r,r]\right), \nabla\widetilde{S}^\Phi(\bar{z})\right\rangle && \text{(Equation (DCF))} \\ &= \eta\langle\nabla\widetilde{S}(\bar{z}), -F(\bar{z})\rangle - \frac{\eta\sigma^2}{2}\langle\nabla\widetilde{S}(\bar{z}), D^2F(\bar{z})[r,r]\rangle\end{aligned}$$

D.7 Proof of Lemma 10

The overall idea is that we can arbitrarily pick H and $\nabla^3 L[u,u]$ to be adversarial, where u is the maximum eigenvector that satisfies $SGu = Hu$ and $\langle u, Gu\rangle = 1$, which is equivalent to the problem defined in Definition 1. Let

$$\mathcal{L} = \{\text{losses } L : \|\nabla^3 L[\cdot,\cdot,\cdot]\| \text{ bounded}\}.$$

We present a *affine* reparameterization that allows us to take $G(x^\star) = I$ without loss of generality. Consider the coordinate transformation $x = x^\star + G(x^\star)^{-1/2}y$, and the reparameterized loss and mirror map

$$\widehat{L}(y) := L\left(x^\star + G(x^\star)^{-1/2}y\right).$$
$$\widehat{\Phi}(y) := \Phi\left(x^\star + G(x^\star)^{-1/2}y\right).$$

One may observe that when $x = x^\star$, we have $y = 0$ with

$$\nabla^2\widehat{\Phi}(0) = I,$$

or we find that the mirror metric in the reparameterized coordinates is precisely Euclidean. We can compute that β remains unchanged under these coordinates, and $\widehat{L}$ remains in $\mathcal{L}$.

Without loss of generality, take $G(x^\star) = I$ and $x^\star = 0$. Let $T := \nabla^3\Phi(x^\star)$. Since $T \neq 0$ and $d \geq 2$, we claim that there exists a unit vector u such that

$$T[u,u] \notin \text{span}\{u\}.$$

To see this, suppose not, and write $f(x) = T[x,x,x]$. Then, one may compute $\nabla f(x) = 3T[x,x]$. We observe that $\nabla f(x)$ is parallel to x everywhere, so f must be constant on a unit sphere. However, because f is odd, the only case must be $f = 0$ everywhere on the sphere. We require $d \geq 2$ for this argument because the unit sphere is connected here. This directly implies $T = 0$. Let us decompose $T[u,u] = au + b$ for some $b \perp u$ and $b \neq 0$.

We construct H by considering its projections on $\text{span}\{u\}$ and $u^\perp$, the orthogonal complement. Fix some $\lambda < S$, and let L be such that the Hessian H satisfies

$$H = Suu^T + \lambda(I - uu^T).$$

Now, we may compute

$$\begin{aligned}\beta^\Phi(x) &= \|\nabla S(x)\|^2_{G^{-1}} + \langle G^{-1}\nabla S, (SI - HG^{-1})\nabla^3\Phi[u,u]\rangle \\ &= \|\nabla S(x)\|_2^2 + \langle\nabla S, (SI - H)(au + b)\rangle\end{aligned}$$

Recall that u satisfies $Hu = Su$ and observe that b satisfies $Hb = \lambda b$. Thus,

$$\beta^\Phi(x) = \|\nabla S(x)\|_2^2 + \langle\nabla S, (S - \lambda)b\rangle. \tag{1}$$

While keeping the Hessian H fixed at $x^\star$, we may prescribe $\nabla^3 L[u,u]$ arbitrarily; for instance, this can be achieved by adding an appropriate cubic polynomial, whose third derivative is constant. By extension, we can control ∇S arbitrarily by Lemma 14, and we opt to minimize eq. (1). We choose $L \in \mathcal{L}$ so that

$$\nabla S = -\frac{S - \lambda}{2}b.$$

Substituting this into the stabilization coefficient gives

$$\begin{aligned}\beta^{\Phi} &= \frac{(S-\lambda)^2}{4}\|b\|_2^2 - \frac{(S-\lambda)^2}{2}\|b\|_2^2 \\ &= -\frac{(S-\lambda)^2}{4}\|b\|_2^2 < 0.\end{aligned}$$

D.8 Proof of Proposition 12

Observe that $q_i = G^{1/2}u_i$ forms an orthonormal basis in the L^2 inner product. Thus, we may decompose:

$$G^{-1/2}\nabla^3\Phi[u,u] = \sum_i \left\langle G^{-1/2}\nabla^3\Phi[u,u], q_i \right\rangle q_i.$$

Furthermore, observe that the q_i's are eigenvectors of $G^{-1/2}HG^{-1/2}$ with $G^{-1/2}HG^{-1/2}u_i = \lambda_i u_i$. Expanding $\beta^{\Phi}(x)$ from Proposition 16, we find:

$$\begin{aligned}\beta^{\Phi}(x) &= \|\nabla S\|_{G^{-1}}^2 + \left\langle G^{-1/2}\nabla S, G^{-1/2}(SI - HG^{-1})G^{1/2}G^{-1/2}\nabla^3\Phi[u,u] \right\rangle \\ &= \|\nabla S\|_{G^{-1}}^2 + \left\langle G^{-1/2}\nabla S, (SI - G^{-1/2}HG^{-1/2})G^{-1/2}\nabla^3\Phi[u,u] \right\rangle \\ &= \|\nabla S\|_{G^{-1}}^2 + \left\langle G^{-1/2}\nabla S, (SI - G^{-1/2}HG^{-1/2}) \sum_i \left\langle G^{-1/2}\nabla^3\Phi[u,u], q_i \right\rangle q_i \right\rangle \\ &= \|\nabla S\|_{G^{-1}}^2 + \sum_i \left\langle G^{-1/2}\nabla S, (S-\lambda_i)q_i \right\rangle \left\langle G^{-1/2}\nabla^3\Phi[u,u], q_i \right\rangle \\ &= \|\nabla S\|_{G^{-1}}^2 + \sum_i (S-\lambda_i)\langle G^{-1/2}\nabla S, q_i\rangle \left\langle G^{-1/2}\nabla^3\Phi[u,u], q_i \right\rangle \\ &= \|\nabla S\|_{G^{-1}}^2 + \sum_i (S-\lambda_i)\langle \nabla S, u_i\rangle \left\langle \nabla^3\Phi[u,u], u_i \right\rangle.\end{aligned}$$

D.9 Auxiliary computations

Throughout this section, we will let $u(x)$ be the top unit eigenvector, and further assume that it is unique. More specifically, u is the maximum eigenvector that satisfies $SGu = Hu$ and $\langle u, Gu\rangle = 1$, as defined in Definition 1. Define $r = Gu$, which is the pushforward of u under $\nabla\Phi$.

Lemma 13 (Gradient in dual coordinates) *Let $f : E \to \mathbb{R}$ and define*

$$\widetilde{f} := f \circ \nabla\Phi^*.$$

For $z = \nabla\Phi(x)$,

$$\nabla\widetilde{f}(z) = \nabla^2\Phi^*(z)\nabla f(x) = G(x)^{-1}\nabla f(x).$$

By the chain rule,

$$\nabla\widetilde{f}(z) = \nabla^2\Phi^*(z)\nabla f(x).$$

Moreover, differentiating $\nabla\Phi^*(\nabla\Phi(x)) = x$ gives

$$\nabla^2\Phi^*(z)\nabla^2\Phi(x) = I,$$

and hence $\nabla^2\Phi^*(z) = G(x)^{-1}$.

Lemma 14 (Computation of sharpness gradient)

$$\nabla S^{\Phi}(x) = \nabla^3 L(x)[u,u] - S(x)\nabla^3\Phi(x)[u,u].$$

Define $H := \nabla^2 L(x)$. Let u satisfy the eigenvalue problem

$$SGu = Hu, \qquad \langle u, Gu\rangle = 1$$

Take gradients on both sides,

$$(\nabla H)u + H(\nabla u) = (\nabla S)Gu + S(\nabla G)u + SG\nabla u.$$

Taking an inner product with u,

$$\langle u, (\nabla H)u\rangle + \langle u, H(\nabla u)\rangle = \nabla S\langle u, Gu\rangle + S\langle u, (\nabla G)u\rangle + S\langle u, G(\nabla u)\rangle$$

Notice that

$$\langle u, H(\nabla u)\rangle = \langle Hu, \nabla u\rangle = S\langle Gu, \nabla u\rangle.$$

Rearranging and writing the identity in tensor notation yields the result.

Proposition 15 *Let $x \in E, v \in E$, $z = \nabla\Phi(x)$.*

$$\nabla^3\Phi^*(z)[Gv, Gv] = -G^{-1}\nabla^3\Phi(x)[v, v].$$

It is well-known that the left-hand side can be expressed as the acceleration of a curve Lee (2013). In particular, let $\alpha(t) = \nabla\Phi^*(z + tGv)$. Then, $\alpha''(0) = \nabla^3\Phi^*(z)[Gv, Gv]$. By applying $\nabla\Phi$, one can write

$$\nabla\Phi \circ \alpha(t) = z + tGv$$

and taking derivatives

$$\nabla^2\Phi(\alpha(t))\alpha'(t) = Gv$$
$$\nabla^3\Phi(\alpha(t))[\alpha'(t), \alpha'(t)] + \nabla^2\Phi(\alpha(t))\alpha''(t) = 0.$$

Plugging in $t = 0$ yields $\alpha'(0) = v$ and $\alpha''(0) = -G^{-1}\nabla^3\Phi[v, v]$.

Proposition 16 *Let $x \in E$ with $z = \nabla\Phi(x)$. Let $\widetilde{S}(z) = S \circ \nabla\Phi^*$. Then,*

$$\langle \nabla\widetilde{S}(z), D^2F[r, r]\rangle = \|\nabla S(x)\|^2_{G^{-1}} + \langle G^{-1}\nabla S, (SI - HG^{-1})\nabla^3\Phi[u, u]\rangle$$

The above quantity is the same as $\beta^\Phi(x)$ as defined in Definition 9.

First, we compute $\nabla S(x)$. By Lemma 14,

$$\nabla S = \nabla^3 L[u, u] - S\nabla^3\Phi[u, u].$$

Next, differentiating $F = \nabla L \circ \nabla\Phi^*$ twice yields

$$\begin{aligned} D^2F[r, r] &= \nabla^3 L[u, u] - HG^{-1}\nabla^3\Phi[u, u] \\ &= \nabla S + S\nabla^3\Phi[u, u] - HG^{-1}\nabla^3\Phi[u, u] \end{aligned}$$

By Lemma 13, $\nabla\widetilde{S}(z) = G^{-1}\nabla S(x)$. Putting these together yields:

$$\begin{aligned} \langle \nabla\widetilde{S}(z), D^2F[r, r]\rangle &= \langle G^{-1}\nabla S(x), \nabla S + S\nabla^3\Phi[u, u] - HG^{-1}\nabla^3\Phi[u, u]\rangle \\ &= \|\nabla S(x)\|^2_{G^{-1}} + \langle G^{-1}\nabla S, (SI - HG^{-1})\nabla^3\Phi[u, u]\rangle. \end{aligned}$$

E Experimental Details

Experiments were run on a shared cluster of V100s, A100s, and H200s.

E.1 Computational details

Eigenvalues and sharpness All eigenvalues are calculated through a LOBPCG algorithm (Knyazev, 2001) and use the implementation of Cohen et al. (2024) (keeping their condition for termination). This method is particularly nice for computing sharpness because it uses the previous eigenvector to "warm start" the solver for the next computation. All computations are done in `Float64`, as we observe that `Float32` can incur substantial numerical errors. In our experiments in Appendix F, we plot the sharpness every 30 steps.

For the ℓ_3 mirror potential, if any parameter $|x_i| < 10^{-30}$, we set $|x_i| = 10^{-30}$ for the purposes of computing G^{-1} and $G^{-1/2}$, which are singular when $x_i = 0$ as described in Section 4.2. In practice, we observe that this route is never used *except* when layers are initialized at zero. This happens only in our vision transformer, where some parameters are set to zero at initialization.

Central flows We try to follow the choices of Cohen et al. (2024) in general. For the flows that represent mirror descent, we do all updates in the dual, which we observe to incur less error. The dual central flow is computed using an Euler method, with step size 0.25 (so that four steps correspond to a mirror descent update, which takes unit time). We discuss our discretization in Section B.2 as it is non-trivial; this discretization is used when $S^{\Phi} > 1.95/\eta$. The other stable flows are implemented with the conservative choice of adaptive step size $1/\max(4, \lceil 2S^{\Phi} \rceil)$. We terminate the gradient flow when the step size becomes smaller than $1/1000$.

Some of our choices also differ from Cohen et al. (2024) (other than only dealing with the regime of one unstable eigenvalue). We do not find it necessary to run mirror descent for a few steps before computing the flows, as the mirror sharpness at initialization falls below $2/\eta$. We also compare with the iterates *directly* instead of a "second-order" average of the iterates, as it is unclear how to "average" the primal iterates in non-Euclidean geometries.

E.2 Architectural/dataset details

We use ReLU activation functions for the experiments in Appendix F, and we use GeLU activation functions for the central flow experiments to ensure differentiability of the loss.

MLP We use the MLP architecture of Cohen et al. (2020). We use PyTorch's standard initializations. This architecture has 655604 parameters.

CNN We use the CNN architecture of Cohen et al. (2020). We use PyTorch's standard initializations. This architecture has 30634 parameters.

ViT Our ViT follows the same architecture as that of Cohen et al. (2024), whose implementation is provided in (LucidRains, 2024). Following Cohen et al. (2024), in our experiments, we multiply the initialization weights and biases of the final layer by 0.5 to reduce the sharpness at initialization. This architecture has 496996 parameters.

CIFAR-10 details We used a training set consisting of 1000 images, which are obtained by taking the first 250 of the first four classes of CIFAR-10. We use the standard pre-processing of centering and normalizing the data by subtracting the mean and then dividing by the standard deviation for each channel across the *full* CIFAR-10K dataset.

F Varying architectures

In this section, we observe the $2/\eta$ mirror sharpness threshold in a variety of architectures beyond this settings described in the main text. For each setting, we graph the training loss, the gradient norm $\|\nabla L\|^2_{G^{-1}}$, the mirror sharpness S^{Φ}, and the ratio between ℓ_2 sharpness and mirror sharpness S^{ℓ_2}/S^{Φ}.

F.1 MLP

F.1.1 MSE

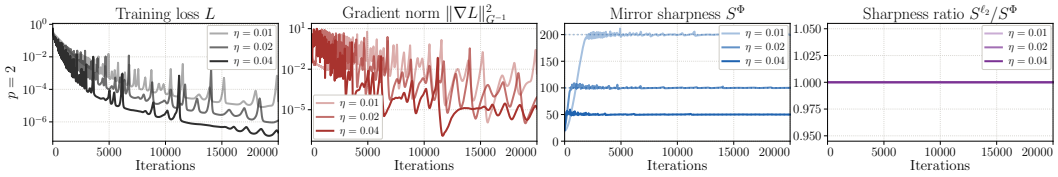


Figure 5: MLP with a MSE loss on a subset of CIFAR-10; ordinary gradient descent.

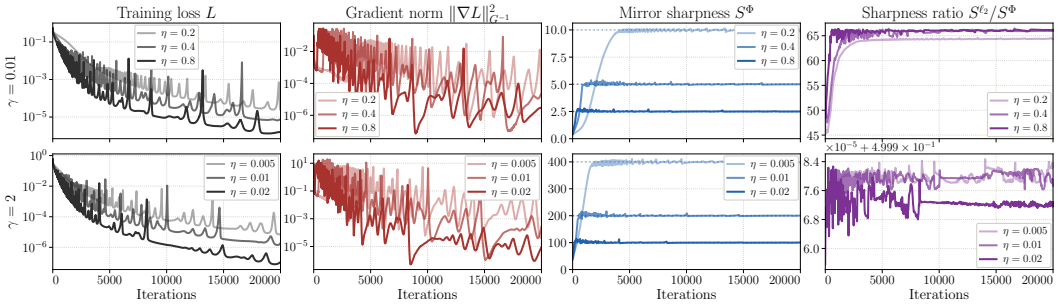


Figure 6: MLP with a MSE loss on a subset of CIFAR-10; hypentropy potential. Top: $\gamma = 0.01$, bottom: $\gamma = 2$.

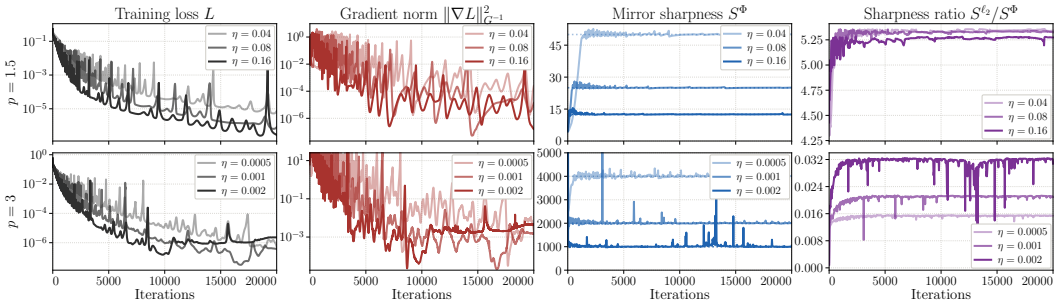


Figure 7: MLP with a MSE loss on a subset of CIFAR-10; ℓ_p potential. Top: $\ell_{1.5}$, bottom: ℓ_3.

F.1.2 Cross-entropy

For cross-entropy losses, we observe that the mirror sharpness hovers slightly above the $2/\eta$ threshold before returning to zero. This is well-explained in the case for gradient descent in Cohen et al. (2020, Appendix C), and we anticipate that a similar explanation may be at play here.

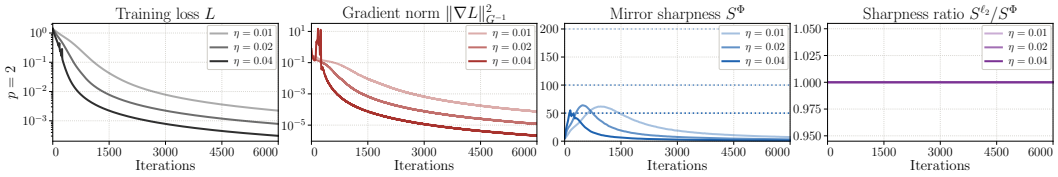


Figure 8: MLP with a cross-entropy loss on a subset of CIFAR-10; ordinary gradient descent.

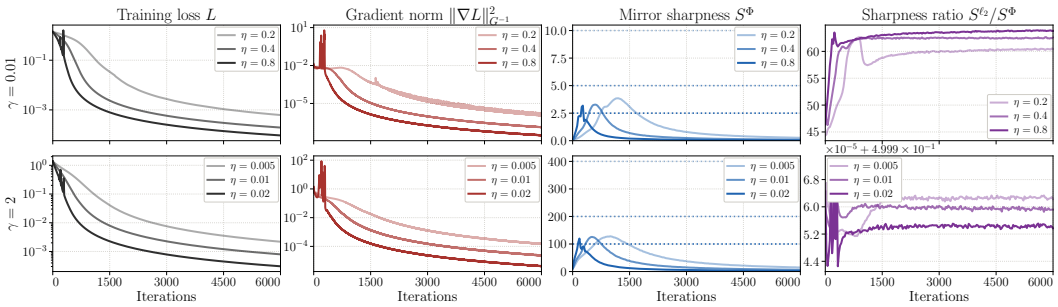


Figure 9: MLP with a cross-entropy loss on a subset of CIFAR-10; hypentropy potential. Top: $\gamma = 0.01$, bottom: $\gamma = 2$.

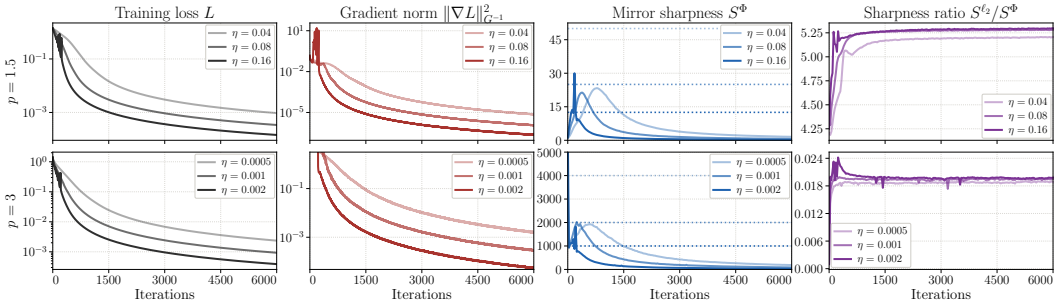


Figure 10: MLP with a cross-entropy loss on a subset of CIFAR-10; ℓ_p potential. Top: $\ell_{1.5}$, bottom: ℓ_3.

F.2 CNN

F.2.1 MSE

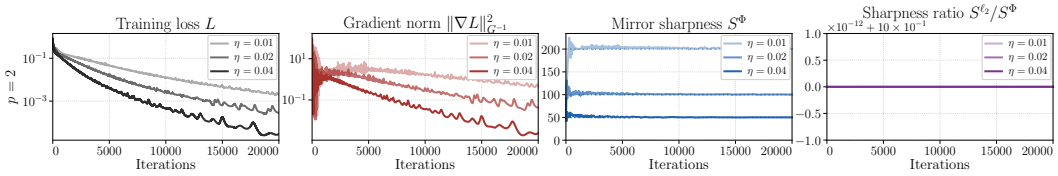


Figure 11: CNN with a MSE loss on a subset of CIFAR-10; gradient descent.

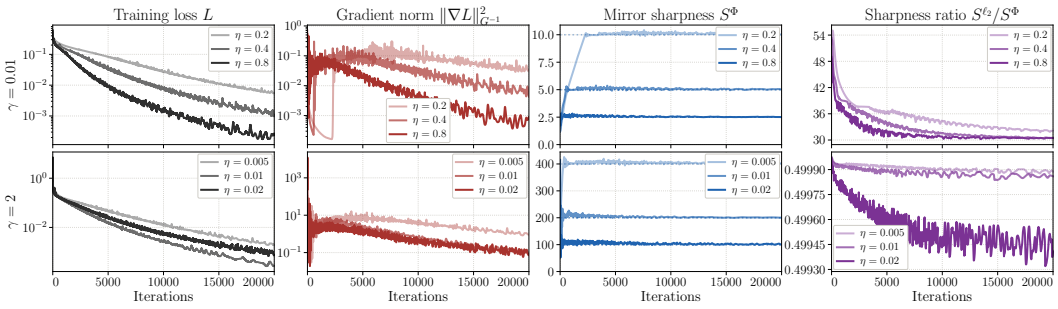


Figure 12: CNN with a MSE loss on a subset of CIFAR-10; hypentropy potential. Top: $\gamma = 0.01$, bottom: $\gamma = 2$. These are the same experiments as Figure 1.

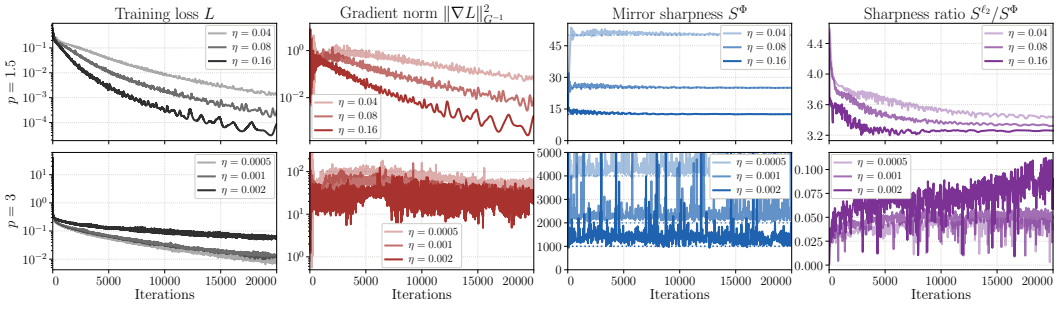


Figure 13: CNN with a MSE loss on a subset of CIFAR-10; ℓ_p potential. Top: $\ell_{1.5}$, bottom: ℓ_3.

F.2.2 Cross-entropy

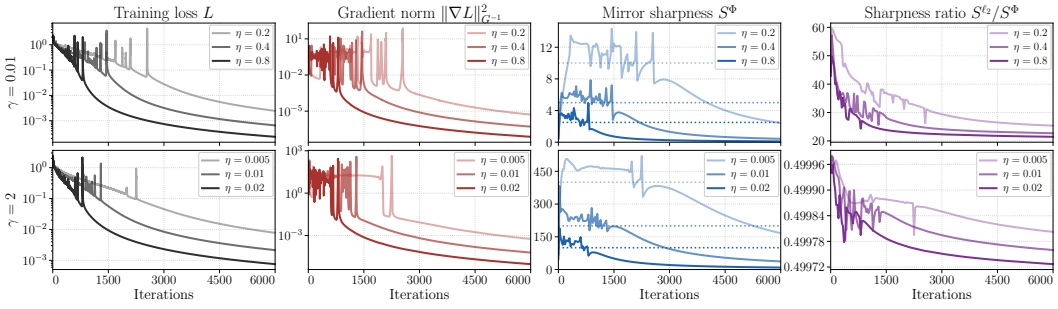


Figure 14: CNN with cross-entropy loss on a subset of CIFAR-10; hypentropy potential. Top: $\gamma = 0.01$, bottom: $\gamma = 2$.

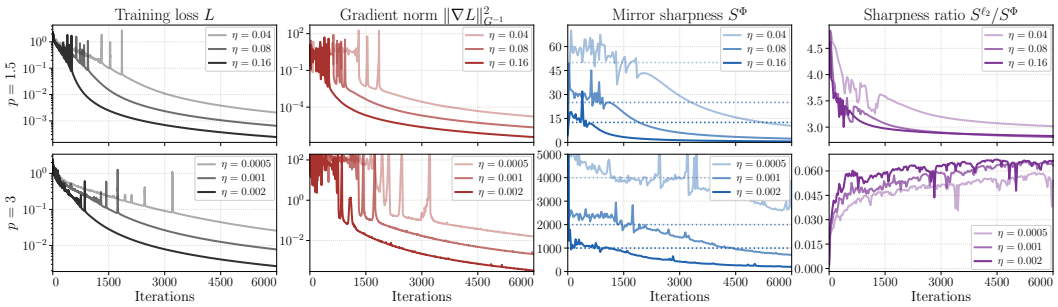


Figure 15: CNN with cross-entropy loss on a subset of CIFAR-10; ℓ_p potential. Top: $\ell_{1.5}$, bottom: ℓ_3.

F.3 ViT

F.3.1 MSE

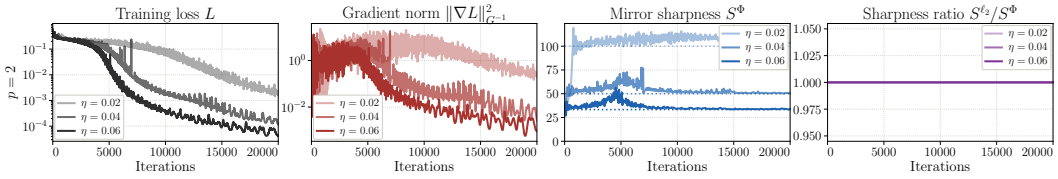


Figure 16: ViT with a MSE loss on a subset of CIFAR-10; ordinary gradient descent.

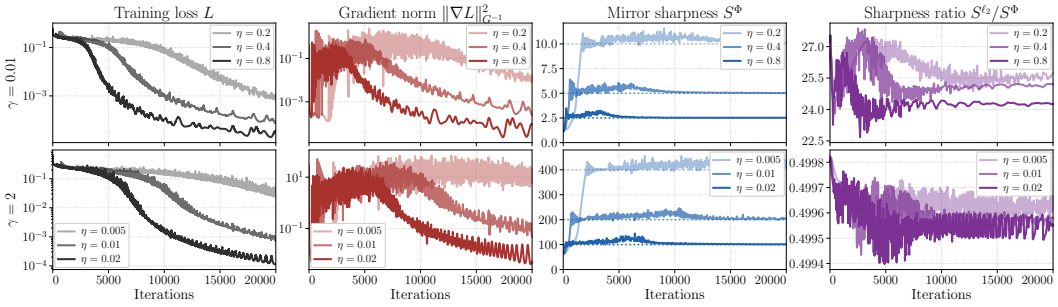


Figure 17: ViT with a MSE loss on a subset of CIFAR-10; hypentropy potential. Top: $\gamma = 0.01$, bottom: $\gamma = 2$.

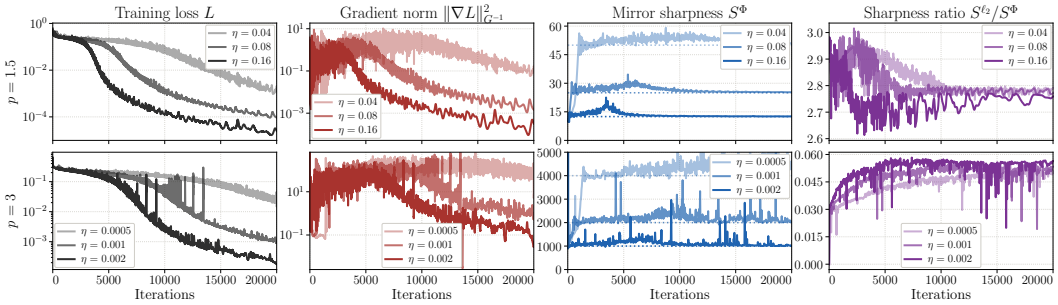


Figure 18: ViT with a MSE loss on a subset of CIFAR-10; ℓ_p potential. Top: $\ell_{1.5}$, bottom: ℓ_3. These are the same experiments as Figure 2.

F.3.2 CROSS-ENTROPY

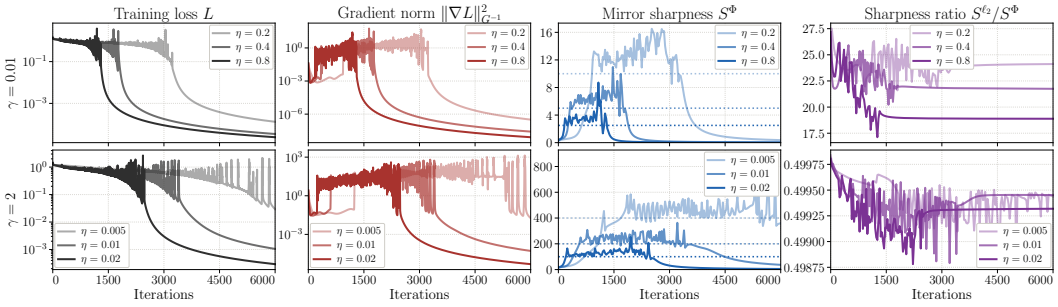


Figure 19: ViT with cross-entropy loss on a subset of CIFAR-10; hypentropy potential. Top: $\gamma = 0.01$, bottom: $\gamma = 2$.

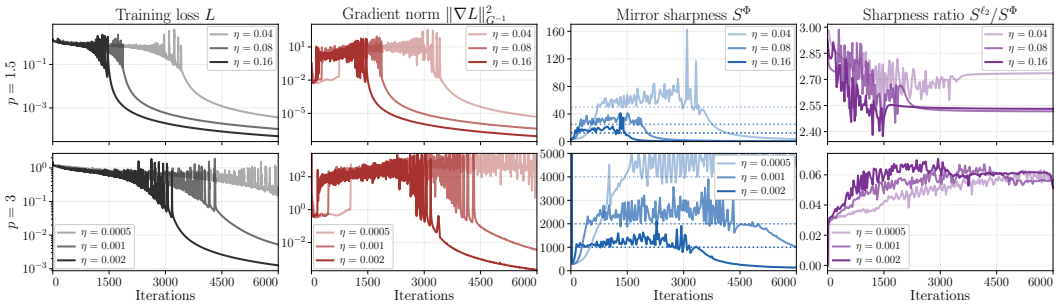


Figure 20: ViT with cross-entropy loss on a subset of CIFAR-10; ℓ_p potential. Top: $\ell_{1.5}$, bottom: ℓ_3.

G CENTRAL FLOW EXPERIMENTS

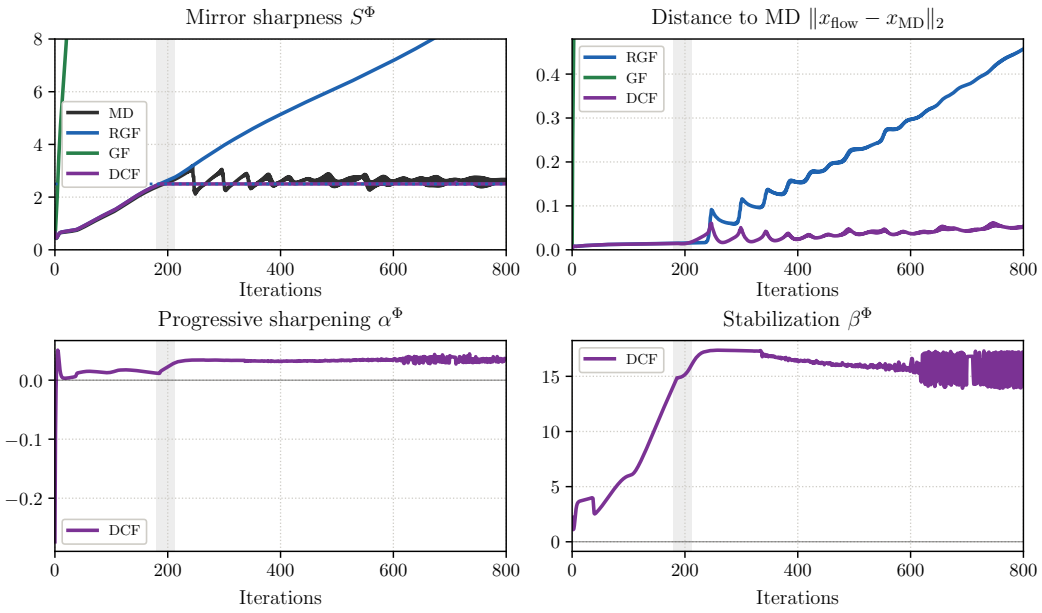


Figure 21: CNN with a MSE loss on a subset of CIFAR-10; hypentropy potential with $\gamma = 0.01$; $\eta = 0.8$. We indicate, in gray, the timesteps where the DCF iterates only have one unstable eigenvector with eigenvalue greater than $1.95/\eta$.

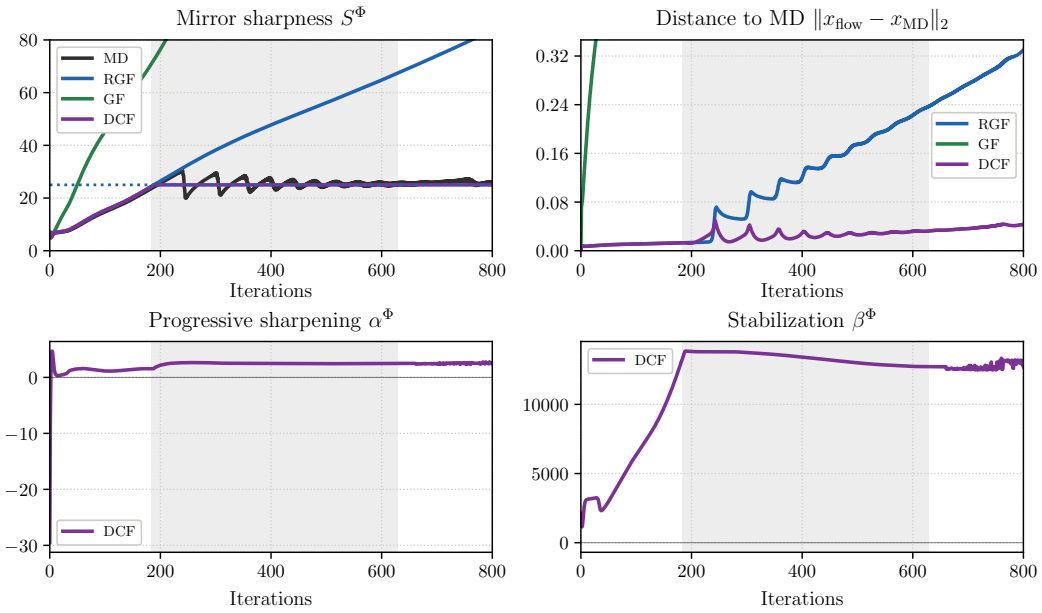


Figure 22: CNN with a MSE loss on a subset of CIFAR-10; $\ell_{1.5}$ potential; $\eta = 0.08$. This is the same experiment as Section 5.2. We indicate, in gray, the timesteps where the DCF iterates only have one unstable eigenvector with eigenvalue greater than $1.95/\eta$.

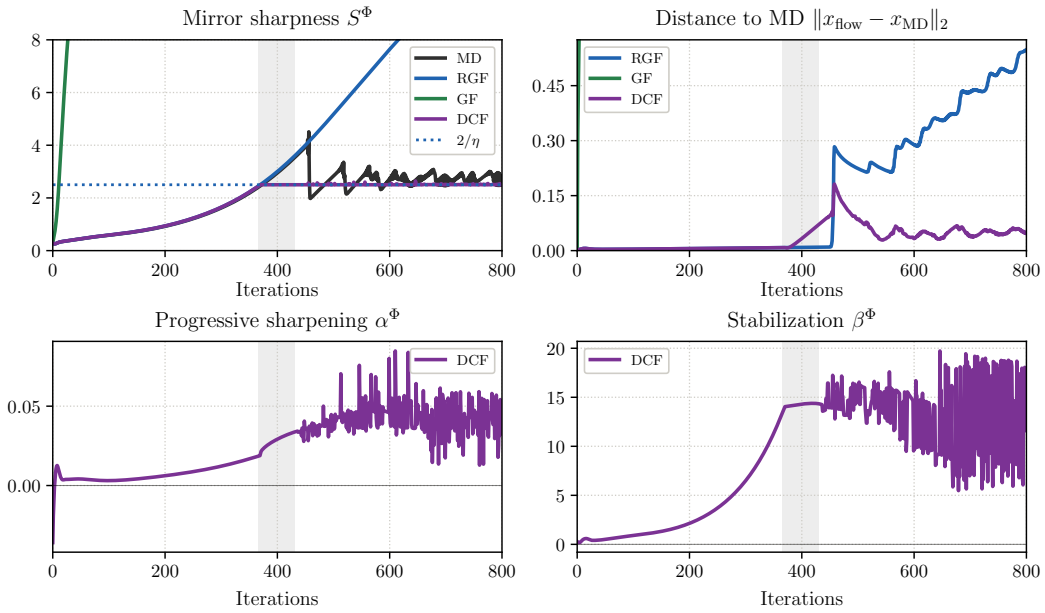


Figure 23: MLP with a MSE loss on a subset of CIFAR-10; hypentropy potential with $\gamma = 0.01$; $\eta = 0.8$. We indicate, in gray, the timesteps where the DCF iterates only have one unstable eigenvector with eigenvalue greater than $1.95/\eta$.

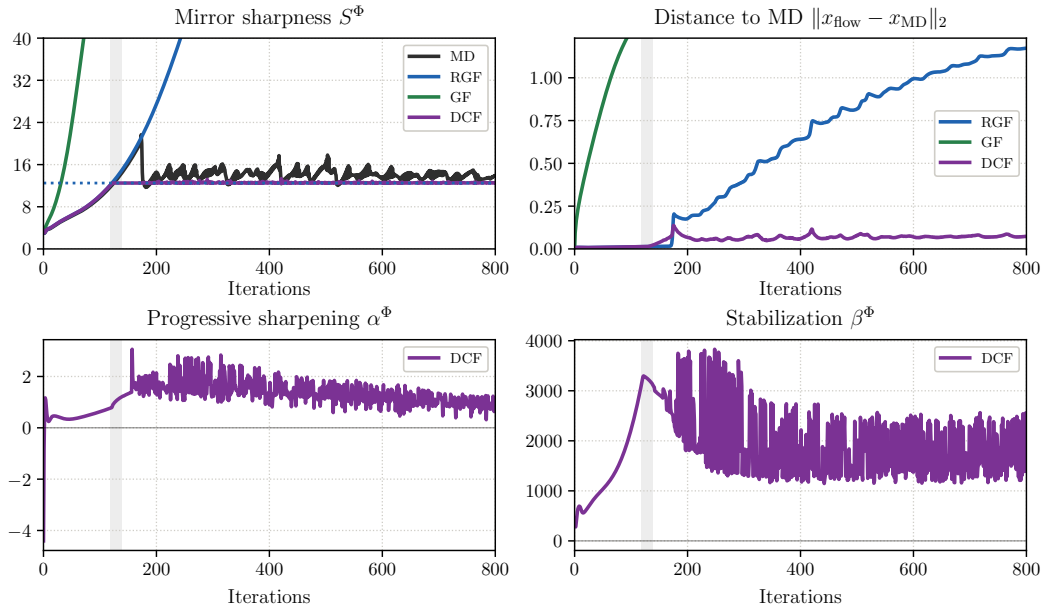


Figure 24: MLP with a MSE loss on a subset of CIFAR-10; $\ell_{1.5}$ potential; $\eta = 0.08$. We indicate, in gray, the timesteps where the DCF iterates only have one unstable eigenvector with eigenvalue greater than $1.95/\eta$.