THE MULTIPLE TIMESCALES OF GRADIENT DESCENT ON THE EDGE OF STABILITY: A PERTURBATIVE DERIVATION OF THE CENTRAL FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

The central flow of Cohen et al. (2025) is an empirically accurate model of gradient descent on the edge of stability in deep learning. However, its derivation is heuristic. We propose a perturbative regime in which the central flow is the limit of gradient descent: we assume that the loss decomposes as $f = g + \varepsilon h$; in the limit $\varepsilon \to 0$, the dynamics of gradient descent with learning rate η converge to the gradient flow of h constrained to the minimizers of g of sharpness at most $2/\eta$. Our approach is formal rather than rigorous; it treats gradient descent as a singularly perturbed dynamical system in ε. Three timescales emerge: a fast timescale of oscillations along the sharpest direction, an intermediate timescale of the self-stabilization at the edge of stability, and a slow timescale of the dynamics along the minimizers of g—the central flow. Using the method of multiple scales, a classical formal method from singular perturbation theory, we derive the expansion of the dynamics in ε: the central flow emerges as the leading-order term in the expansion, while the self-stabilization mechanism appears in the next-order term. We study this mechanism beyond previous analyses: with a single eigenvalue at the edge of stability, we compute the slow drift of the energy of the fluctuations; with several eigenvalues at the edge of stability, we derive the self-stabilization system and explain why fluctuations persist.

1 INTRODUCTION

Context. Recent works on gradient descent in deep learning have pointed out a surprising phenomenology. A large part of the training occurs at the so-called *edge of stability* (Jastrzębski et al., 2019; 2020; Cohen et al., 2021): the *sharpness* of the loss—the maximal eigenvalue of the Hessian—fluctuates around the value $2/\eta$, where η is the learning rate. Above this threshold, gradient descent on the local quadratic approximation of the loss develops diverging oscillations; however, on the loss itself, the oscillations induce an effect in the local cubic approximation that pushes back the dynamics in the direction of sharpness reduction. This effect was called the *self-stabilization* at the edge of stability (Damian et al., 2023).

The gradient flow approximation of gradient descent fails to incorporate the effect of these oscillations. Instead, Cohen et al. (2025) proposed the *central flow* approximation: the gradient flow constrained to the stable region where the sharpness is at most $2/\eta$. Numerical experiments across a wide range of datasets and architectures have shown that the central flow accurately approximates the trajectory of gradient descent when averaged over oscillations and fluctuations.

Problem. However, the derivation of the central flow by Cohen et al. (2025) is heuristic. Several approximation steps are left informal, for instance using a local cubic approximation of the loss, or the meaning of the "local time-averaging" operator $\mathbb{E}$. This raises the following question: is there a regime in which the central flow is the limit of gradient descent?

The regime $\eta \to 0$ of vanishing learning rate, used to derive the gradient flow, is not appropriate to answer this question. In this regime, the sharpness is always below $2/\eta$, and thus edge of stability never occurs.

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.

Contributions. In this work, we propose a perturbative regime—the *sharp valley assumption*—from which we derive the central flow and the self-stabilization mechanism. Mathematically, we assume that we run gradient descent with learning rate η on a loss function $f(z) = g(z) + \varepsilon h(z)$. The function g defines a valley: it has several minimizers and we denote $V = \{z : g(z) = \min g\}$ the valley of its minimizers. Further, $\varepsilon \ll 1$ is a small parameter and h is a perturbation that induces the dynamics along the valley. Equivalently, gradient descent with learning rate $\varepsilon\eta$ on the loss function $\varepsilon^{-1}g + h$ gives the same sequence of iterates. The parameter ε can thus be interpreted as the inverse sharpness of the valley.

The sharp valley picture has been present in several works, see e.g. (Xing et al., 2018; Jastrzębski et al., 2019; Wen et al., 2025; Liu et al., 2025b;c; Song et al., 2025), as it is consistent with several empirical observations, such as the anti-alignment of the gradient at consecutive iterates or the approximate low-rank structure of the Hessian (see App. A for a more complete discussion). A contribution of this work is to show that it is also consistent with the edge of stability, self-stabilization and the central flow.

Indeed, in the regime $\varepsilon \to 0$, the central flow is the formal limit of gradient descent. More precisely, in this regime, gradient descent is a singularly perturbed dynamical system in which three timescales emerge: a fast timescale $t_1 = k$, the iteration number k, which is the typical timescale of the oscillations along the sharpest direction; an intermediate timescale $t_2 = \varepsilon^{1/2}k$, which is the typical timescale of the self-stabilization mechanism; and a slow timescale $t_3 = \varepsilon k$, which is the typical timescale of the central flow—the evolution along the valley.

We then use the *method of multiple scales* (Holmes, 2012) to derive an approximation of gradient descent that decouples the three timescales t_1, t_2, and t_3. The leading term in the approximation is the central flow, which evolves on the slow timescale t_3. It consists in the gradient flow of h in the valley, constrained to the stable region where the sharpness of g is below $2/\eta$. The next-order term contains the oscillations along the sharpest direction as a function of t_1 and the self-stabilization fluctuations as a function of t_2. The method of multiple scales is a *formal* method in the sense of asymptotic analysis: it produces the expansion by matching orders, without proving convergence or bounding the remainder. Our results should thus be interpreted as formal derivations, i.e., conjectures with strong supporting evidence.

Further consequences. Beyond the justification of previous works, the framework of this paper enables some new discussions.

We study the self-stabilization mechanism in more depth: with a single eigenvalue on the edge of stability, we compute the slow drift of the energy of the fluctuations; with multiple eigenvalues on the edge of stability, we derive the self-stabilization system (left open by Damian et al. (2023)) and explain the persistence of fluctuations.

Finally, we compare the central flow with competing continuous-time approximations of gradient descent (Regis & Chewi, 2026; Marion, 2026; Chodron de Courcel, 2026). We obtain that the central flow is a more coarse-grained approximation than the competing approximations, as it depends only on t_3, while the other approximations depend on t_2 and t_3.

Outline. In Sec. 2, we introduce our mathematical setting, including the sharp valley assumption. The strongest result of this paper is provided in Sec. 5; however, for the sake of clarity and to enable additional discussions, previous sections provide simpler versions of the result. In Sec. 3, we analyze a minimal example. The simplicity in this example makes it possible to discuss potential failures of the central flow approximation and the slow drift of the self-stabilization energy. Further, in Sec. 4, we discuss the more general case of valleys of codimension one. In Sec. 5, we discuss the general case of valleys of arbitrary codimension. Finally, in Sec. 6, we discuss the comparison with competing approximations. Additional related work is provided in App. A.

2 SETTING

Let $f(z)$, $z \in Z$, be a loss function, where Z is a finite-dimensional vector space. We study the dynamics of gradient descent on f with learning rate $\eta > 0$: $z(k+1) = z(k) - \eta\nabla f(z(k))$. We assume that the loss function f is decomposed as $f(z) = g(z) + \varepsilon h(z)$, where g defines a valley and $\varepsilon \ll 1$ is a small parameter. We denote $V = \{z \in Z : g(z) = \min g\}$ the valley of minimizers of g.

To simplify the analyses, we assume that the valley defined by g is a linear space. More precisely, we assume there exists an orthogonal decomposition $Z = X \oplus^{\perp} Y$, in which we decompose $z = (x, y)$, such that the valley defined by g is $V = \{(0, y), y \in Y\}$. In a more complex setting where V would be a manifold, the decomposition $z = (x, y)$ would represent a local set of coordinates to parametrize the manifold V of minimizers of g.

Further, for all z, we denote $S(z) = \nabla_x^2 g(z)$ the Hessian of g in the sharp directions—those orthogonal to the valley. We assume that for all $(0, y) \in V$, $S(0, y) \succ 0$: there is some curvature of g in the sharp directions along the valley. Finally, we denote $V_{\mathrm{S}} = \{z \in V : \eta S(z) \preccurlyeq 2 \operatorname{Id}_X\}$ the *stable* valley of g, in which the sharpness is below $2/\eta$.

Notations. When $y \in \mathbb{R}^d$ is a vector, we denote $y[1], \dots, y[d]$ its components. Let A and B be Euclidean spaces. When $L : A \to B$ is a linear operator, we denote $L^* : B \to A$ its adjoint. We denote $\operatorname{Sym} A$ the set of symmetric linear operators on A, and $\operatorname{Sym}_+ A$ the set of symmetric positive semidefinite linear operators on A. We denote Id_A the identity operator on A. When C is a linear subspace of A, we denote $C^{\perp}$ its orthogonal complement, $P_C : A \to C$ the orthogonal projection onto C, and $\iota_C : C \to A$ the canonical inclusion of C into A. Note that $\iota_C = P_C^*$.

3 Minimal example

3.1 Introduction and experiments

Consider the case $Z = \mathbb{R}^2$, $X = Y = \mathbb{R}$, thus $z = (x, y) \in \mathbb{R}^2$. Further, we consider the following choices determining the loss function $f = g + \varepsilon h$:

$$g(x, y) = \frac{1}{2}(y + \zeta x)x^2\,, \qquad h(x, y) = -x - y\,. \tag{3.1}$$

Here, $\zeta \in \mathbb{R}$ is a parameter whose role is discussed in Sec. 3.3. This function g does not rigorously satisfy the valley structure required in Sec. 2. However, this is benign: below, we consider the function g only in the region where $y + \zeta x > 0$ where $g(x, y)$ is minimal if and only if $x = 0$. In the valley $V = \{(0, y), y > 0\}$, h decreases as y increases, thus the gradient descent dynamics move along the valley in the direction of increasing y. However, the sharpness of g in the valley is $S(0, y) = \partial_x^2 g(0, y) = y$, thus there is progressive sharpening, until the dynamics hit the edge of stability at $y = 2/\eta$. There are then self-stabilization fluctuations around the value $y = 2/\eta$.

As the gradient of $f = g + \varepsilon h$ in the valley is of order ε, a typical timescale of the dynamics of the iterates $z(k) = (x(k), y(k))$ along the valley is $t_3 = \varepsilon k$. In Figure 1, we plot the iterates of gradient descent as a function of t_3, for various values of ε. We observe that the iterates converge to the central flow (x_0, y_0) as $\varepsilon \to 0$, that is, to the gradient flow of h constrained to the stable valley $V_{\mathrm{S}} = \{(0, y), y \leqslant 2/\eta\}$.

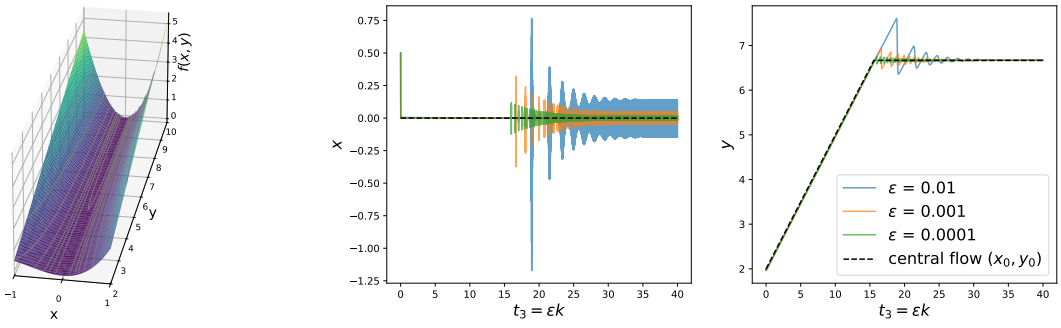


Figure 1: Left: plot of the loss function $f = g + \varepsilon h$ for $\varepsilon = 0.01$. Right: iterates of gradient descent on the minimal example of Eqs. (3.1) with learning rate $\eta = 0.3$, $\zeta = 1$.

At the edge of stability, the oscillations of x and the self-stabilization fluctuations appear in the next-order term of the approximation. As confirmed by the method of multiple scales below, this next order term is of order $\varepsilon^{1/2}$. In Figure 2, we plot the rescaled iterates $\varepsilon^{-1/2}(x - x_0, y - y_0)$ as functions of the slow timescale $t_3 = \varepsilon k$ and the intermediate timescale $t_2 = \varepsilon^{1/2} k$. In the limit

$\varepsilon \to 0$, we observe that the self-stabilization loop and its slow damping decouple on the timescales t_2 and t_3. While the typical timescale of a self-stabilization fluctuation is t_2, the typical timescale of the damping is t_3.

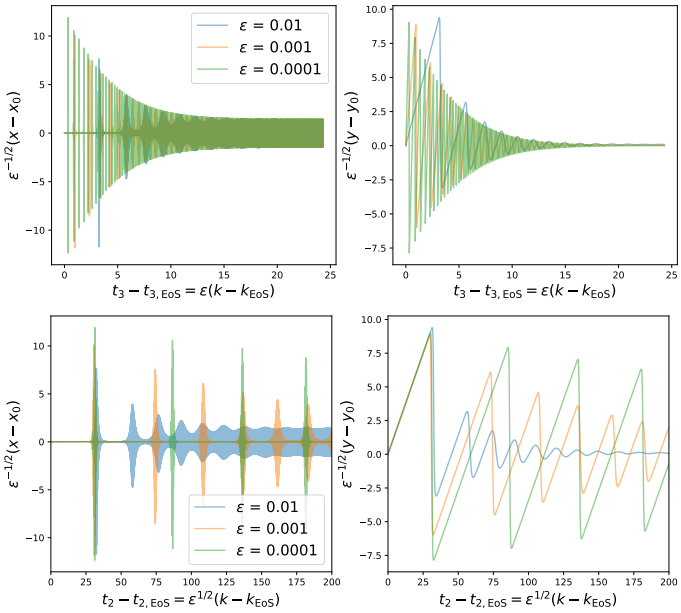


Figure 2: Rescaled iterates of gradient descent on the minimal example of Eqs. (3.1) with $\eta = 0.3$, $\zeta = 1$. We denote k_{EoS} the first iteration number for which $y(k) > 2/\eta$: it encodes the beginning of the edge of stability. We denote $t_{2,\text{EoS}} = \varepsilon^{1/2} k_{\text{EoS}}$ and $t_{3,\text{EoS}} = \varepsilon k_{\text{EoS}}$ the corresponding rescaled times, and we start the plots at these times. Top: as a function of t_3; bottom: as a function of t_2.

3.2 Result by the method of multiple scales

The goal of this section is to confirm and refine the empirical observations above using the method of multiple scales. It is a perturbation method: in our case, the perturbation parameter is $\varepsilon \ll 1$, and the method seeks an expansion of the form

$$x = x_0 + \varepsilon^{1/2} x_1 + \varepsilon x_2 + \dots , \qquad y = y_0 + \varepsilon^{1/2} y_1 + \varepsilon y_2 + \dots . \tag{3.2}$$

The goal is to obtain properties for the leading terms of this expansion. The challenge is that the terms depend simultaneously on the different timescales $t_1 = k$, $t_2 = \varepsilon^{1/2} k$, and $t_3 = \varepsilon k$, that depend themselves on ε. This is a typical situation where the method of multiple scales is useful.

Let us first give an overview of the method. The method postulates that the timescales t_1, t_2, and t_3 can be decoupled by treating them as independent parameters. In our case, t_1 is a discrete parameter while t_2 and t_3 are continuous parameters. It then treats all functions of k as functions of the three independent timescales: $z = z(t_1, t_2, t_3)$, $x = x(t_1, t_2, t_3)$, $y = y(t_1, t_2, t_3)$, $x_0 = x_0(t_1, t_2, t_3)$, etc. The difference equation $\Delta_k z = z(k+1) - z(k) = -\eta \nabla f(z)$ in $z = z(k)$ is transformed into a partial difference-differential equation in $z = z(t_1, t_2, t_3)$. By identifying the different terms in the expansion of this equation, we infer properties for the terms $x_0 = x_0(t_1, t_2, t_3)$, $y_0 = y_0(t_1, t_2, t_3)$, $x_1 = x_1(t_1, t_2, t_3)$, etc. The final approximation is provided by using the expansion of Eq. (3.2) and substituting the actual timescales $t_1 = k$, $t_2 = \varepsilon^{1/2} k$, and $t_3 = \varepsilon k$.

We obtain the following result for our minimal example.

Result 3.1. *Assume that (x_0, y_0) remains in the stable valley, i.e., $x_0 = 0$ and $\eta y_0 \leqslant 2$. Then $y_0 = y_0(t_3)$, i.e., y_0 does not depend on t_1 or t_2; and $y_1 = y_1(t_2, t_3)$, i.e., y_1 does not depend on t_1. Further, we have the following:*

(i) If $\eta y_0 < 2$ (relative interior of the stable valley), then $\partial_{t_3} y_0 = -\eta \partial_y h(x_0, y_0) = \eta$.

(ii) If $\eta y_0 = 2$ (edge of stability), then $x_1 = (-1)^{t_1} a_1$ for some $a_1 = a_1(t_2, t_3)$. Further, we have the self-stabilization system

$$\partial_{t_2} a_1 = \eta a_1 y_1 , \qquad \partial_{t_2} y_1 = \eta \left(1 - \frac{a_1^2}{2}\right) . \tag{3.3}$$

Again, the method of multiple scales is formal, thus the above result should not be interpreted as a mathematical theorem but rather as a conjecture with strong supporting evidence. All the results of this paper share this status.

This result confirms that (x_0, y_0) follows the central flow prediction. It moves in the stable valley $V_{\mathrm{S}} = \{(0, y) : \eta y \leqslant 2\}$, following the constrained gradient flow of h, until it hits the constraint $\eta y_0 \leqslant 2$. It then stays on the edge of stability threshold. This is consistent with Fig. 1.

This result also gives information about the next-order term (x_1, y_1) at the edge of stability. While y_1 does not depend on t_1, x_1 oscillates in t_1, with amplitude $a_1(t_2, t_3)$. Again, this is consistent with the experiments of Figs. 1 and 2. Moreover the joint evolution of a_1 and y_1 in t_2 is provided by the self-stabilization system (3.3), already derived by Damian et al. (2023). As noted by them, these self-stabilization dynamics preserve the energy $E = \frac{y_1^2}{2} + \frac{a_1^2}{4} - \frac{1}{2}\log a_1^2$ resulting in (a_1, y_1) being periodic in t_2. As a consequence, $E = E(t_3)$, i.e., E evolves only on the slow timescale t_3. This is consistent with Fig. 2. The evolution of E in t_3 is discussed in Sec. 3.3.

We now discuss the assumptions of Result 3.1. The statement $x_0 = 0$ is not strictly needed and could instead be obtained as a consequence of the method of multiple scales. However, treating x_0 in the method derivations results in more complicated computations, thus we prefer making the reasonable assumption that the limiting dynamics are in the valley V. Moreover, note that we could initialize x at a non-zero value, which would result in a relaxation of x_0 to 0 on the timescale t_1. (This can be seen in Fig. 1.) This early part of the dynamics is not interesting for our study of the edge of stability, thus we do not include it here.

On the contrary, we need the assumption $\eta S(0, y_0) \leqslant 2$ to derive the result. In fact, in App. C, we build an artificial example in which this condition breaks and the central flow prediction fails. Thus this assumption cannot be removed (but might be replaceable by some weaker condition).

We now turn to the derivation of Result 3.1. The situation in which we apply the method of multiple scales is rather sophisticated, due to the presence of three timescales (two is more common) and the presence of a discrete timescale t_1. See Holmes (2012) for more elementary uses of the method.

Derivation of Result 3.1. The expansion of the gradient descent iteration $\Delta_k z = -\eta \nabla f(z)$ is computed in App. B. From Eqs. (B.1), (B.3) and keeping terms up to order ε, we obtain

$$\begin{aligned}
&\varepsilon^{1/2}\Delta_{t_1} x_1 + \varepsilon\left[\Delta_{t_1} x_2 + \partial_{t_2} x_1 + \partial_{t_2}\Delta_{t_1} x_1\right] \\
&\qquad = -\varepsilon^{1/2}\eta y_0 x_1 - \varepsilon\eta\left[y_0 x_2 + \frac{3}{2}\zeta x_1^2 + x_1 y_1 - 1\right] + o(\varepsilon)\,, \qquad (3.4)\\
&\Delta_{t_1} y_0 + \varepsilon^{1/2}\left[\Delta_{t_1} y_1 + \partial_{t_2} y_0 + \partial_{t_2}\Delta_{t_1} y_0\right] \\
&\qquad + \varepsilon\left[\Delta_{t_1} y_2 + \partial_{t_2} y_1 + \partial_{t_2}\Delta_{t_1} y_1 + \partial_{t_3} y_0 + \frac{1}{2}\partial_{t_2}^2 y_0 + \partial_{t_3}\Delta_{t_1} y_0 + \frac{1}{2}\partial_{t_2}^2\Delta_{t_1} y_0\right] \qquad (3.5)\\
&\qquad = -\varepsilon\eta\left[\frac{x_1^2}{2} - 1\right] + o(\varepsilon)\,.
\end{aligned}$$

We identify the different terms in the expansion.

Order 1**.** The y equation (3.5) gives $\Delta_{t_1} y_0 = 0$, thus $y_0 = y_0(t_2, t_3)$ does not depend on t_1.

Order $\varepsilon^{1/2}$**.** The x equation (3.4) gives $\Delta_{t_1} x_1 = -\eta y_0 x_1$. Here, y_0 should be interpreted as the sharpness of g at (x_0, y_0); this difference equation describes the linearized dynamics of x around (x_0, y_0). Its behavior depends on whether the edge of stability is reached. Fix t_2 and t_3. If $\eta y_0(t_2, t_3) < 2$, then $x_1 = x_1(t_1, t_2, t_3) \to 0$ as $t_1 \to \infty$. If $\eta y_0(t_2, t_3) = 2$, then $x_1 = (-1)^{t_1} a_1(t_2, t_3)$ for some a_1 that does not depend on t_1: at the edge of stability, x_1 oscillates with an amplitude which is a function of t_2 and t_3 only.

We now turn to the y equation (3.5). Recall that y_0 does not depend on t_1. The equation gives $\Delta_{t_1} y_1 + \partial_{t_2} y_0 = 0$. Thus y_1 is an affine function of t_1, with slope $-\partial_{t_2} y_0$ (independent of t_1). However, having y_1 unbounded is not acceptable as it would imply that $\varepsilon^{1/2} y_1$ eventually become of order 1, thus interfering with the leading order term y_0 in the expansion (3.2). In the method of multiple scales, such a term is called *secular*, and the method prescribes to build an approximation that forbids secular terms, to keep the reasoning consistent (Holmes, 2012). This prescription is

the core mechanism of the method; we will use it repeatedly. Here, we can thus conclude that $\Delta_{t_1} y_1 = -\partial_{t_2} y_0 = 0$. As a consequence, $y_1 = y_1(t_2, t_3)$ and $y_0 = y_0(t_3)$.

Order ε. The x equation (3.4) gives

$$\Delta_{t_1} x_2 + \partial_{t_2} x_1 + \partial_{t_2} \Delta_{t_1} x_1 = -\eta \left[y_0 x_2 + \frac{3}{2} \zeta x_1^2 + x_1 y_1 - 1 \right] .$$

Here, we focus on understanding what happens at the edge of stability, i.e., when $\eta y_0 = 2$. We then know that $x_1 = (-1)^{t_1} a_1(t_2, t_3)$, and thus $\Delta_{t_1} x_1 = -2(-1)^{t_1} a_1(t_2, t_3)$. We obtain

$$\Delta_{t_1} x_2 + 2x_2 = (-1)^{t_1} \partial_{t_2} a_1 - \eta \left[\frac{3}{2} \zeta a_1^2 + (-1)^{t_1} a_1 y_1 - 1 \right] . \tag{3.6}$$

Consider the above equation as a difference equation for x_2 in t_1, with fixed t_2 and t_3. The right-hand side depends on t_1 only through the $(-1)^{t_1}$ terms, as the terms a_1 and y_1 were shown to be independent of t_1. This suggests considering the following lemma.

Lemma 3.2 (Corollary of Lemma G.1)**.** *Let $\alpha, \beta \in \mathbb{R}$. The solutions $\varphi = \varphi(t_1)$ of the difference equation $\Delta_{t_1} \varphi + 2\varphi = \alpha + \beta(-1)^{t_1}$ are $\varphi = \frac{\alpha}{2} + a(-1)^{t_1} - \beta t_1 (-1)^{t_1}$, $a \in \mathbb{R}$. In particular, there exists a bounded solution if and only if $\beta = 0$.*

By this lemma applied to Eq. (3.6), we must have $\partial_{t_2} a_1 = \eta y_1 a_1$ to avoid a secular term.

We now turn to the y equation (3.5). It gives $\Delta_{t_1} y_2 = -\partial_{t_2} y_1 - \partial_{t_3} y_0 - \eta \left[\frac{x_1^2}{2} - 1 \right]$. We distinguish whether we are at the edge of stability.

If $\eta y_0 < 2$, then $x_1 \to 0$ as $t_1 \to \infty$, and all other quantities of the right-hand side are independent of t_1. To avoid a secular term, we must have $-\partial_{t_2} y_1 - \partial_{t_3} y_0 + \eta = 0$. We consider this as an equation for y_1 in t_2. As y_0 is independent of t_2, we have a secular term unless $\partial_{t_3} y_0 = \eta$. This finishes our derivation of item (i) of the result.

If $\eta y_0 = 2$, then we have $x_1 = (-1)^{t_1} a_1(t_2, t_3)$ and $\partial_{t_3} y_0 = 0$. We obtain $\Delta_{t_1} y_2 = -\partial_{t_2} y_1 - \eta \left[\frac{a_1^2}{2} - 1 \right]$. The right-hand side is independent of t_1. To avoid a secular term, we must have $\partial_{t_2} y_1 = \eta \left(1 - \frac{a_1^2}{2} \right)$ (and $y_2 = y_2(t_2, t_3)$). This completes the derivation of item (ii) of the result. □

3.3 Energy variations in self-stabilization

The self-stabilization dynamics preserve the energy

$$E = E(t_3) = \frac{y_1^2}{2} + \frac{a_1^2}{4} - \frac{1}{2} \log a_1^2 ,$$

as a function of t_2. However, the energy still varies as a function t_3, as observed in Figs. 1 and 2. The method of multiple scales makes these variations tractable.

Result 3.3. *Under the assumptions of Result 3.1, at the edge of stability, we have*

$$\partial_{t_3} E = \eta^2 (2 - 9\zeta^2) \left\langle y_1^2 \right\rangle_{t_2} ,$$

where $\left\langle y_1^2 \right\rangle_{t_2} = \left\langle y_1^2(t_2, t_3) \right\rangle_{t_2}$ denotes the average of y_1^2 over one period in t_2.

Note that $\left\langle y_1^2 \right\rangle_{t_2}$ is a function of E only, thus the above result is an autonomous ordinary differential equation for E in t_3. If $\zeta^2 > \frac{2}{9}$, the energy decreases (as in Figs. 1 and 2) while if $\zeta^2 < \frac{2}{9}$, the energy increases. In App. D, we validate numerically this result for moderate energy levels and discuss its failure at large energy.

The analysis at moderate energy levels is most relevant for practice since neural networks seem to systematically dissipate energy when only one eigenvalue is on the edge of stability, see Fig. 3 of this paper or Figs. 5, 6, 10, 14, 15 of Cohen et al. (2025). We leave it as an open problem to understand the structure of the loss of neural networks that causes this mechanism.

Analytically, studying the dependence of E in t_3 requires pushing the method of multiple scales to the order $\varepsilon^{3/2}$. Indeed, as the reader might have noticed, the order 1 gives the dependence of z_0 in t_1;

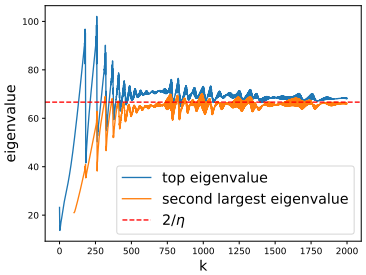


Figure 3: Evolution of the top eigenvalues of the Hessian along the training path of a neural network. Here, we train a multilayer perceptron with two hidden layers of width 256 and GELU activation on a subset of CIFAR-10 with mean-squared error loss. The learning rate is $\eta = 0.03$. See reproducibility statement for more details.

the order $\varepsilon^{1/2}$ gives the dependencies of z_1 in t_1 and of z_0 in t_2; the order ε gives the dependencies of z_2 in t_1, of z_1 in t_2, and of z_0 in t_3. Continuing this pattern, the dependence of z_1 (and thus of E) in t_3 is obtained at order $\varepsilon^{3/2}$. This leads to tedious computations provided in App. E. We have done these derivations only in the minimal example of this section and leave the more general case of Sec. 4 for future work.

4 Valleys of Codimension One

We return to the general setting of Sec. 2. We are interested in the fact that Y can be multidimensional, thus y_0 can move at the edge of stability. However, in this section, we assume that $X = \mathbb{R}$, so that $\nabla^2 g(0, y)$ has exactly one non-zero eigenvalue $S(0, y) = \partial_x^2 g(0, y) \geqslant 0$. In Fig. 4, we provide an example of such dynamics where the valley is of dimension 2. We observe the edge of stability phenomenon, the self-stabilization fluctuations, and the convergence to the central flow as $\varepsilon \to 0$.

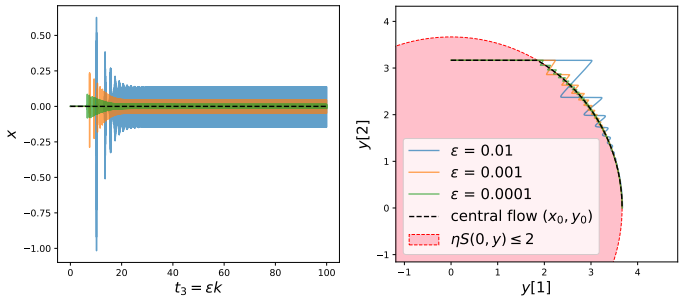


Figure 4: Iterates of gradient descent with learning rate $\eta = 0.3$ on $f = g + \varepsilon h$. We take $X = \mathbb{R}$ and $Y = \mathbb{R}^2$, thus the valley is of dimension 2 and codimension 1. Further, we take $g(x, y) = \frac{1}{2}(\|y\| + 3e^{x/2})x^2$ and $h(x, y) = -x - y[1]$. Note that $S(0, y) = \|y\| + 3$ and thus V_S is a closed disk: $V_S = \{(0, y), \|y\| \leqslant 2/\eta - 3\}$.

We now state a general result for such valleys of codimension 1.

Result 4.1. *Assume that X is of dimension 1, and that (x_0, y_0) remains in the stable valley, i.e., $x_0 = 0$ and $\eta S(0, y_0) \leqslant 2$. Then $y_0 = y_0(t_3)$, and $y_1 = y_1(t_2, t_3)$. Further, we have the following:*

(i) z_0 follows the gradient flow of h constrained to the stable valley V_S, i.e., there exists $\sigma^2 = \sigma^2(t_3) \geqslant 0$ such that

$$\partial_{t_3} y_0 = -\eta \left[\nabla_y h(0, y_0) + \frac{\sigma^2}{2} \nabla_y S(0, y_0) \right], \qquad \text{where } \sigma^2 = 0 \text{ or } \eta S(0, y_0) = 2.$$

(ii) If $\eta S(0, y_0) = 2$ (edge of stability), then $x_1 = (-1)^{t_1} a_1$ for some $a_1 = a_1(t_2, t_3)$. Further, denote $u_1 = \langle y_1, \nabla_y S(0, y_0) \rangle$. Then we have the self-stabilization system

$$\partial_{t_2} a_1 = \eta a_1 u_1, \quad \partial_{t_2} u_1 = -\eta \left[\langle \nabla_y h(0, y_0), \nabla_y S(0, y_0) \rangle + \|\nabla_y S(0, y_0)\|^2 \frac{a_1^2}{2} \right]. \quad (4.1)$$

Moreover, in that case, we have $\sigma^2 = \langle a_1^2 \rangle_{t_2}$, i.e. $\sigma^2(t_3)$ is the average of $a_1^2(t_2, t_3)$ over a period in t_2.

The derivation of this result is provided in App. F. Point (i) shows that the leading term (x_0, y_0) follows the central flow prediction: it follows the gradient flow of h constrained to the stable valley.

The function $\sigma^2 = \sigma^2(t_3)$ should be interpreted as the Lagrange multiplier associated with the constraint $\eta S(0, y_0) \leqslant 2$: it satisfies dual feasibility ($\sigma^2 \geqslant 0$) and complementarity slackness with the constraint (for all t_3, $\sigma^2(t_3) = 0$ or $\eta S(0, y_0(t_3)) = 2$).

Note that in the self-stabilization system (4.1), $\langle \nabla_y h(0, y_0), \nabla_y S(0, y_0) \rangle$ and $\|\nabla_y S(0, y_0)\|^2$ are independent of t_2. Moreover, in a situation where the dynamics stay at the edge of stability, we must have $\langle \nabla_y h(0, y_0), \nabla_y S(0, y_0) \rangle < 0$: the negative gradient of h pushes towards the stability boundary. The self-stabilization system (4.1) is thus the same as the one of (3.3), up to a rescaling. A self-stabilization loop appears, preserving an energy $E = E(t_3)$. For fixed t_3, (a_1, u_1) is periodic in t_2, thus the average $\langle a_1^2 \rangle_{t_2}$ over t_2 is well-defined.

The derivation of the central flow by Cohen et al. (2025) already interpreted $\sigma^2 = \mathbb{E}x^2$, where $\mathbb{E}$ denotes "local time-averages". Here, thanks to the timescale decoupling, we can be more precise: $\mathbb{E}$ corresponds to averaging over t_2, for fixed t_3.

5 Valleys of arbitrary codimension

This section tackles the case where the dimension of X is arbitrary. Recall that we denote $S(z) = \nabla_x^2 g(z) \in \operatorname{Sym} X$. Further, we denote $D_y S(x, y) : \delta y \in Y \mapsto D_y S(x, y)[\delta y] \in \operatorname{Sym} X$ the differential of S in the y variable, and consequently, $D_y S(x, y)^* : \operatorname{Sym} X \to Y$ its adjoint. Finally, we denote $\langle ., . \rangle$ the canonical inner product on $\operatorname{Sym} X$.

Result 5.1. *Assume that (x_0, y_0) remains in the stable valley, i.e., $x_0 = 0$ and $\eta S(0, y_0) \preccurlyeq 2 \operatorname{Id}_X$. Then $y_0 = y_0(t_3)$, and $y_1 = y_1(t_2, t_3)$. Further, we have the following:*

(i) z_0 follows the gradient flow of h constrained to the stable valley V_S, i.e., there exists $\Sigma = \Sigma(t_3) \in \operatorname{Sym}_+ X$ such that

$$\partial_{t_3} y_0 = -\eta \left[\nabla_y h(0, y_0) + \frac{1}{2} D_y S(0, y_0)^*[\Sigma] \right] , \quad \text{where } \langle \Sigma, 2 \operatorname{Id}_X - \eta S(0, y_0) \rangle = 0.$$

(ii) Let $C = C(t_3) = \ker(2 \operatorname{Id}_X - \eta S(0, y_0)) \subset X$ denote the critical subspace. Recall that $\iota_C : C \to X$ denotes the canonical inclusion, and $P_C : X \to C$ the orthogonal projection onto C. We have $P_C x_1 = (-1)^{t_1} a_1$ for some $a_1 = a_1(t_2, t_3) \in C$ and $P_{C^\perp} x_1 \xrightarrow[t_1 \to \infty]{} 0$. Further, denote $\Psi = \Psi(t_3) : \delta y \in Y \to \Psi[\delta y] = P_C D_y S(0, y_0)[\delta y] \iota_C \in \operatorname{Sym} C$ and $U_1 = \Psi[y_1] \in \operatorname{Sym} C$. Then we have the self-stabilization system

$$\partial_{t_2} a_1 = \eta U_1 a_1 , \qquad \partial_{t_2} U_1 = -\eta \left[\Psi[\nabla_y h(0, y_0)] + \frac{1}{2} \Psi \Psi^*[a_1 a_1^\top] \right] . \tag{5.1}$$

Moreover, provided that the average $\langle a_1 a_1^\top \rangle_{t_2} = \lim_{T_2 \to \infty} \frac{1}{T_2} \int_0^{T_2} a_1(t_2, t_3) a_1(t_2, t_3)^\top \mathrm{d}t_2$ is well-defined, we have $\Sigma = \iota_C \langle a_1 a_1^\top \rangle_{t_2} P_C$.

The derivation is provided in App. G. The function $\Sigma = \Sigma(t_3)$ should be interpreted as the Lagrange multiplier associated with the constraint $\eta S(0, y_0) \preccurlyeq 2 \operatorname{Id}_X$: it satisfies dual feasibility ($\Sigma \in \operatorname{Sym}_+ X$) and complementarity slackness with the constraint (for all t_3, $\langle \Sigma(t_3), 2 \operatorname{Id}_X - \eta S(0, y_0(t_3)) \rangle = 0$).

The derivation of the central flow by Cohen et al. (2025) already interpreted $\Sigma = \mathbb{E} x x^\top$, where $\mathbb{E}$ denotes "local time-averages". Here, thanks to the timescale decoupling, we can be more precise: $\mathbb{E}$ corresponds to averaging over t_2, for fixed t_3 (and taking the limit $t_1 \to \infty$).

Result 5.1 is a generalization of Result 4.1. When only one eigenvalue of $S(0, y_0)$ has reached the edge of stability, we recover the self-stabilization system (4.1). Indeed, then C is of dimension 1, thus $a_1 \in C$ and $U_1 \in \operatorname{Sym} C$ are scalars, and $\Psi \Psi^* : \operatorname{Sym} C \to \operatorname{Sym} C$ is a scalar multiplication. Thus in that case, the self-stabilization system (5.1) reduces to the self-stabilization system (4.1). We have a periodic behavior in t_2. When more eigenvalues enter the edge of stability, the behavior of the self-stabilization system (5.1) is much richer, see Fig. 8 in App. H.

The self-stabilization system (4.1) admits the fixed point $u_1 = 0, a_1^2 = -\frac{2\langle \nabla_y h, \nabla_y S \rangle}{\|\nabla_y S\|^2}$. Meanwhile, finding a fixed point for the self-stabilization system (5.1) would require solving the equation $\Psi \Psi^*[a_1 a_1^\top] = -2\Psi[\nabla_y h(0, y_0)]$. As the right hand side is in $\operatorname{im} \Psi = \operatorname{im} \Psi \Psi^*$, there exists

$\widetilde{\Sigma} \in \operatorname{Sym} C$ such that $\Psi\Psi^*[\widetilde{\Sigma}] = -2\Psi[\nabla_y h(0, y_0)]$. However, it is non-generic that a rank-one solution $\widetilde{\Sigma}$ exists. Thus, generically, there is no fixed point, and to avoid a secular term, $a_1 a_1^\top$ must evolve persistently so that its average value is $\widetilde{\Sigma}$.

This explains the observed behavior in Fig. 8 of App. H. When only one eigenvalue is on the edge of stability, the self-stabilization system is observed to lose energy (on the t_3 timescale), converging to its fixed point. However, when a second eigenvalue enters the edge of stability, there is a significant increase in the edge of stability fluctuations, and these fluctuations are persistent. This phenomenology is observed systematically in deep learning experiments, see Fig. 3 of this paper or Figs. 5, 6, 10, 14, 15 of Cohen et al. (2025).

The implications of Result 5.1 for the loss $f(x, y)$ along training are discussed in App. I.

6 Comparison with other approximations

Several works (Regis & Chewi, 2026; Marion, 2026; Chodron de Courcel, 2026) subsequent to the central flow (Cohen et al., 2025) have proposed other approximations of gradient descent, see Fig. 5 for a comparison. In addition to the slow dynamics in t_3, these approximations attempt at capturing the self-stabilization dynamics in t_2. In fact, the derivations of Regis & Chewi (2026) and Marion (2026) explicitly average only two subsequent iterates, thus over the timescale t_1 only. All of these approximations report similar or slightly better accuracy than the central flow in approximating gradient descent. While qualitatively, these approximations correctly generate self-stabilization loops, these loops are often off in magnitude or in phase, leading to only minor quantitative improvement.

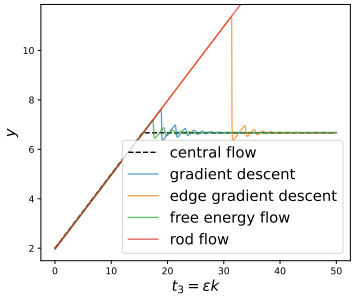


Figure 5: Comparison of the central flow of Cohen et al. (2025), the rod flow of Regis & Chewi (2026), the edge gradient descent of Marion (2026) and the flow of Chodron de Courcel (2026) (labeled "free energy flow" here) in the setting of Fig. 1 with $\varepsilon = 0.01$. The rod flow has not been designed to handle the transition to the edge of stability; the fact that it may fail to detect the start of the edge of stability phase had already been pointed out by Marion (2026, Fig. 6).

Overall, in the presence of the several timescales t_1, t_2, t_3, effective models can average at different levels, leading to different tradeoffs between accuracy and conceptual simplicity. The central flow, with its constrained gradient flow formulation, is the simplest and coarsest version, averaging over t_1 and t_2. The other approximations, averaging over t_1 only, contain more information on the self-stabilization dynamics but have a more complex formulation. This paper advocates for seeking *expansions* of the dynamics, where each individual term is simple and interpretable, while their combination can have high accuracy.

7 Conclusion

The contributions of this paper are both conceptual and analytical. Conceptually, we clarify how the central flow and the self-stabilization mechanism are related to a sharp valley structure of the loss landscape. However, understanding why such a loss landscape structure appears when training deep learning models is still largely an open problem (see Wen et al. (2025) for a first approach).

Analytically, we provide a framework for studying the dynamics of gradient descent on the edge of stability, based on the method of multiple scales. This framework could be improved by formalizing the derivations through more rigorous techniques of singular perturbation theory. Computing the effect of a curved valley V is also an interesting direction for future work. MacDonald et al. (2025); MacDonald & Vidal (2026) provide interesting elements in these directions.

The strength of the framework stems from its systematic nature. As a consequence, we plan to use it to analyze inertial and adaptive algorithms. We hope that understanding these complex dynamics across many optimizers will help to design new methods.

AI USE STATEMENT

In this work, AI-based tools were used for line completion in the text editor, for proofreading and for formatting the paper. They were not used to generate the scientific content of the paper, or to draft its prose beyond phrase-level completion. The code generating the figures (except Fig. 3) has been written by the authors, but AI has been used to clean the code before sharing it. We have verified all AI-assisted output. In particular, we have checked that the cleaned code produces the same figures as the original code. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The code that generated all figures of this paper, except Fig. 3, is available in the supplementary material, see `code/README.md`. Fig. 3 was obtained using the publicly available code of Cohen et al. (2025), using the command:

```
python main.py opt:gd data:cifar10 arch:mlp \
  --data.classes=2 --data.n=1000 --data.criterion=mse \
  --opt.lr=0.03 --runs discrete --steps=2000 --eig.frequency=1
```

Results 3.3, 4.1 and 5.1 are not derived in the paper but in the appendices. A technical expansion of the gradient descent update is also deferred to App. B.

REFERENCES

Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In *International Conference on Machine Learning*, 2023.

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International Conference on Machine Learning*, 2022.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, 2022.

Gerard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. High-dimensional SGD aligns with emerging outlier eigenspaces. In *International Conference on Learning Representations*, 2024.

Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. *Advances in Neural Information Processing Systems*, 2024.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *International Conference on Machine Learning*, 2023.

Antonin Chodron de Courcel. Gradient descent at the edge of stability: free energy model and kinetic description of the two-layer network. *arXiv preprint arXiv:2606.05326*, 2026.

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

Jeremy Cohen, Alex Damian, Ameet Talwalkar, Zico Kolter, and Jason Lee. Understanding optimization in deep learning with central flows. In *International Conference on Learning Representations*, 2025.

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

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *International Conference on Machine Learning*, 2019.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep linear networks beyond the edge of stability. *International Conference on Learning Representations*, 2025.

Guy Gur-Ari, Daniel Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Mark Holmes. *Introduction to perturbation methods*. Springer Science & Business Media, 2012.

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

Stanislaw Jastrzębski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020.

Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos. In *International Conference on Learning Representations*, 2025.

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. In *International Conference on Machine Learning*, 2023.

Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of SGD for deep nets: Dynamics and generalization. In *SIAM International Conference on Data Mining*, 2020.

Elon Litman. The origin of edge of stability. *arXiv preprint arXiv:2604.20446*, 2026.

Liming Liu, Zixuan Zhang, Simon Du, and Tuo Zhao. A minimalist example of edge-of-stability and progressive sharpening. *Advances in Neural Information Processing Systems*, 2025a.

Yizhou Liu, Ziming Liu, and Jeff Gore. Focus: First order concentrated updating scheme. *arXiv preprint arXiv:2501.12243*, 2025b.

Ziming Liu, Yizhou Liu, Jeff Gore, and Max Tegmark. Neural thermodynamic laws for large language model training. *arXiv preprint arXiv:2505.10559*, 2025c.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *arXiv preprint arXiv:2204.11326*, 2022.

Lachlan MacDonald and René Vidal. Dynamics of gradient descent with large step size near a manifold of flat minima. *arXiv preprint arXiv:2607.08380*, 2026.

Lachlan MacDonald, Hancheng Min, Leandro Palma, Salma Tarmoun, Ziqing Xu, and René Vidal. Convergence rates for gradient descent on the edge of stability in overparametrised least squares. *Advances in Neural Information Processing Systems*, 2025.

Pierre Marion. Edge flow: A tractable and predictive continuous-time model for gradient descent at the edge of stability. *arXiv preprint arXiv:2606.18080*, 2026.

Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet Hessians. In *International Conference on Machine Learning*, 2019.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.

Eric Regis and Sinho Chewi. Rod flow: A continuous-time model for gradient descent at the edge of stability. *arXiv preprint arXiv:2602.01480*, 2026.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the Hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Jun-Ho So and Dongwook Shin. Mini-batch noise lowers sharpness via dominant-subspace fluctuations. In *High-dimensional Learning Dynamics*, 2026.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. *Advances in Neural Information Processing Systems*, 2023.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations*, 2025.

Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. *Journal of Machine Learning Research*, 26(273):1–68, 2025.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems*, 2022.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. In *International Conference on Learning Representations*, 2025.

Jingfeng Wu, Vladimir Braverman, and Jason Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 2023.

Jingfeng Wu, Peter Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *Conference on Learning Theory*, 2024.

Jingfeng Wu, Pierre Marion, and Peter Bartlett. Large stepsizes accelerate gradient descent for regularized logistic regression. *Advances in Neural Information Processing Systems*, 2025.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with SGD. *arXiv preprint arXiv:1802.08770*, 2018.

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. *arXiv preprint arXiv:2506.06940*, 2025.

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

A Additional related work

Sharp valley structure of the loss. Arguments for the sharp valley structure of the loss landscape are widespread in the literature. To give a few examples, this picture is consistent with plots of the loss along the sharpest direction of the Hessian around an iterate (Jastrzębski et al., 2019), plots of the loss along the segment between two consecutive iterates (Xing et al., 2018), and the anti-alignment of the gradients of consecutive iterates (Xing et al., 2018). Along the training path, the loss Hessian exhibits an approximate low-rank structure (Sagun et al., 2016; 2017; Ghorbani et al., 2019; Papyan, 2019; 2020; Ben Arous et al., 2024). The gradients are mostly aligned with the sharpest directions of the Hessian (Gur-Ari et al., 2018; Li et al., 2020), however the component of the gradients in these dominant directions does not contribute to any loss decrease (Jastrzębski et al., 2019; Song et al., 2025), but to a sharpness reduction effect (So & Shin, 2026). In a toy problem, the valley structure has been mathematically analyzed and the sharpest directions are associated with uncertain features of the problem (Wen et al., 2025). Finally, given this accumulation of empirical evidence, the sharp valley structure has already been used as a working hypothesis in some works (Liu et al., 2025b;c).

Analyses of the edge of stability phenomenon. After the initial empirical observations of the edge of stability phenomenon (Jastrzębski et al., 2019; 2020; Cohen et al., 2021), a large body of work has been devoted to explaining it (Ahn et al., 2022; Arora et al., 2022; Wang et al., 2022; Damian et al., 2023; Zhu et al., 2023; Ma et al., 2022; Agarwala et al., 2023; Chen & Bruna, 2023; Kreisler et al., 2023; Song & Yun, 2023; Wu et al., 2023; 2024; Cai et al., 2024; Wang et al., 2025; Kalra et al., 2025; Ghosh et al., 2025; Yoo et al., 2025; Liu et al., 2025a; Wu et al., 2025; Litman, 2026). In contrast, our results *assume* that the leading-term of dynamics remains in the stable region, and study the consequences of this assumption. To the best of our knowledge, no previous work had provided additional justification for the central flow approximation of Cohen et al. (2025).

From a modeling perspective, our work bears similarities with MacDonald et al. (2025); MacDonald & Vidal (2026) that study the dynamics of gradient descent around a valley of minimizers. They show that gradient descent implicitly performs a Riemannian gradient descent on the sharpness in the valley of minimizers, and describe the bifurcating dynamical system in the orthogonal direction. In their "subcritical regime", their analysis corresponds to the case $h = 0$ in our work. However, in the absence of ∇h pushing towards the stability boundary, the edge of stability phenomenon is transient.

B Expansions

B.1 Expansion of the discrete difference operator

To expand the discrete difference $\Delta_k z$, we first need to express the discrete difference operator Δ_k applied to a function $z = z(t_1, t_2, t_3)$. As $t_1 = k$, $t_2 = \varepsilon^{1/2}k$, and $t_3 = \varepsilon k$ depend on k, the expression of Δ_k depends on the discrete difference operator Δ_{t_1} and the partial derivatives ∂_{t_2} and ∂_{t_3}.

$$\begin{aligned}\Delta_k z &= z(t_1+1, t_2+\varepsilon^{1/2}, t_3+\varepsilon) - z(t_1, t_2, t_3)\\ &= z(t_1+1, t_2+\varepsilon^{1/2}, t_3+\varepsilon) - z(t_1, t_2+\varepsilon^{1/2}, t_3+\varepsilon)\\ &\quad + z(t_1, t_2+\varepsilon^{1/2}, t_3+\varepsilon) - z(t_1, t_2, t_3)\\ &= (\Delta_{t_1} z)(t_1, t_2+\varepsilon^{1/2}, t_3+\varepsilon) + z(t_1, t_2+\varepsilon^{1/2}, t_3+\varepsilon) - z(t_1, t_2, t_3)\end{aligned}$$

We first decompose

$$\begin{aligned}&z(t_1, t_2+\varepsilon^{1/2}, t_3+\varepsilon) - z(t_1, t_2, t_3)\\ &\quad = \varepsilon^{1/2}\partial_{t_2} z + \varepsilon \partial_{t_3} z + \frac{\varepsilon}{2}\partial_{t_2}^2 z + \varepsilon^{3/2}\partial_{t_2}\partial_{t_3} z + \frac{\varepsilon^{3/2}}{6}\partial_{t_2}^3 z + o(\varepsilon^{3/2})\,.\end{aligned}$$

Note that we do not write the evaluations of the partial derivatives at (t_1, t_2, t_3) for readability. Using a similar expansion for $(\Delta_{t_1} z)(t_1, t_2+\varepsilon^{1/2}, t_3+\varepsilon)$, we obtain

$$\begin{aligned}\Delta_k z &= \Delta_{t_1} z + \varepsilon^{1/2}\partial_{t_2} z + \varepsilon\partial_{t_3} z + \frac{\varepsilon}{2}\partial_{t_2}^2 z + \varepsilon^{3/2}\partial_{t_2}\partial_{t_3} z + \frac{\varepsilon^{3/2}}{6}\partial_{t_2}^3 z\\ &\quad + \varepsilon^{1/2}\partial_{t_2}\Delta_{t_1} z + \varepsilon\partial_{t_3}\Delta_{t_1} z + \frac{\varepsilon}{2}\partial_{t_2}^2\Delta_{t_1} z + \varepsilon^{3/2}\partial_{t_2}\partial_{t_3}\Delta_{t_1} z + \frac{\varepsilon^{3/2}}{6}\partial_{t_2}^3\Delta_{t_1} z + o(\varepsilon^{3/2})\\ &= \Delta_{t_1} z + \varepsilon^{1/2}\left[\partial_{t_2} z + \partial_{t_2}\Delta_{t_1} z\right] + \varepsilon\left[\partial_{t_3} z + \frac{1}{2}\partial_{t_2}^2 z + \partial_{t_3}\Delta_{t_1} z + \frac{1}{2}\partial_{t_2}^2\Delta_{t_1} z\right]\\ &\quad + \varepsilon^{3/2}\left[\partial_{t_2}\partial_{t_3} z + \frac{1}{6}\partial_{t_2}^3 z + \partial_{t_2}\partial_{t_3}\Delta_{t_1} z + \frac{1}{6}\partial_{t_2}^3\Delta_{t_1} z\right] + o(\varepsilon^{3/2})\,.\end{aligned}$$

We combine the above decomposition with the expansion of z

$$z = z_0 + \varepsilon^{1/2} z_1 + \varepsilon z_2 + \varepsilon^{3/2} z_3 + o(\varepsilon^{3/2})\,.$$

We obtain

$$\begin{aligned}\Delta_k z = \Delta_{t_1} z_0 + \varepsilon^{1/2}\left[\Delta_{t_1} z_1 + \partial_{t_2} z_0 + \partial_{t_2}\Delta_{t_1} z_0\right] \\ + \varepsilon\left[\Delta_{t_1} z_2 + \partial_{t_2} z_1 + \partial_{t_2}\Delta_{t_1} z_1 + \partial_{t_3} z_0 + \frac{1}{2}\partial_{t_2}^2 z_0 + \partial_{t_3}\Delta_{t_1} z_0 + \frac{1}{2}\partial_{t_2}^2\Delta_{t_1} z_0\right] \\ + \varepsilon^{3/2}\Big[\Delta_{t_1} z_3 + \partial_{t_2} z_2 + \partial_{t_2}\Delta_{t_1} z_2 + \partial_{t_3} z_1 + \frac{1}{2}\partial_{t_2}^2 z_1 + \partial_{t_3}\Delta_{t_1} z_1 + \frac{1}{2}\partial_{t_2}^2\Delta_{t_1} z_1 \\ + \partial_{t_2}\partial_{t_3} z_0 + \frac{1}{6}\partial_{t_2}^3 z_0 + \partial_{t_2}\partial_{t_3}\Delta_{t_1} z_0 + \frac{1}{6}\partial_{t_2}^3\Delta_{t_1} z_0\Big] + o(\varepsilon^{3/2})\,.\end{aligned} \tag{B.1}$$

B.2 Expansion of the gradient step

In this section, we compute the expansion of $-\eta\nabla f(z)$.

$$\begin{aligned}-\eta\nabla f(z) &= -\eta\nabla g(z) - \eta\varepsilon\nabla h(z) \\ &= -\eta\left[\nabla g\left(z_0 + \varepsilon^{1/2} z_1 + \varepsilon z_2 + \varepsilon^{3/2} z_3 + o(\varepsilon^{3/2})\right) + \varepsilon\nabla h\left(z_0 + \varepsilon^{1/2} z_1 + o(\varepsilon^{1/2})\right)\right] \\ &= -\eta\Big[\nabla g(z_0) + \nabla^2 g(z_0)\left(\varepsilon^{1/2} z_1 + \varepsilon z_2 + \varepsilon^{3/2} z_3\right) \\ &\qquad + \frac{1}{2}\nabla^3 g(z_0)\left[\varepsilon^{1/2} z_1 + \varepsilon z_2, \varepsilon^{1/2} z_1 + \varepsilon z_2\right] \\ &\qquad + \frac{1}{6}\nabla^4 g(z_0)\left[\varepsilon^{1/2} z_1, \varepsilon^{1/2} z_1, \varepsilon^{1/2} z_1\right] \\ &\qquad + \varepsilon\nabla h(z_0) + \varepsilon\nabla^2 h(z_0)\left(\varepsilon^{1/2} z_1\right) + o(\varepsilon^{3/2})\Big] \\ &= -\eta\nabla g(z_0) - \varepsilon^{1/2}\eta\nabla^2 g(z_0) z_1 - \varepsilon\eta\left[\nabla^2 g(z_0) z_2 + \frac{1}{2}\nabla^3 g(z_0)[z_1, z_1] + \nabla h(z_0)\right] \\ &\qquad - \varepsilon^{3/2}\eta\left[\nabla^2 g(z_0) z_3 + \nabla^3 g(z_0)[z_1, z_2] + \frac{1}{6}\nabla^4 g(z_0)[z_1, z_1, z_1] + \nabla^2 h(z_0) z_1\right] \\ &\qquad + o(\varepsilon^{3/2})\,.\end{aligned} \tag{B.2}$$

In the minimal example of Sec. 3, and using the assumption $x_0 = 0$, this gives

$$\begin{aligned}-\eta\partial_x f(x,y) &= -\varepsilon^{1/2}\eta y_0 x_1 - \varepsilon\eta\left[y_0 x_2 + \frac{3}{2}\zeta x_1^2 + x_1 y_1 - 1\right] \\ &\qquad - \varepsilon^{3/2}\eta\left[y_0 x_3 + 3\zeta x_1 x_2 + x_1 y_2 + x_2 y_1\right] + o(\varepsilon^{3/2})\,, \\ -\eta\partial_y f(x,y) &= -\varepsilon\eta\left[\frac{x_1^2}{2} - 1\right] - \varepsilon^{3/2}\eta x_1 x_2 + o(\varepsilon^{3/2})\,.\end{aligned} \tag{B.3}$$

C Metastability beyond the edge of stability

In the derivation of Result 3.1, without making the assumption $\eta y_0 \leqslant 2$, we would have considered a third case $\eta y_0 > 2$ in the derivations of order $\varepsilon^{1/2}$. In this case, x_1 develops a secular term (exponentially growing), unless $x_1 = 0$. Thus, to avoid a secular term, we must have the dichotomy: $\eta y_0 \leqslant 2$ or $x_1 = 0$. It is not excluded that we leave the stable valley, however this can be done only in a way where the perturbation x_1 of $x_0 = 0$ is exactly 0. This situation seems non-generic; in the minimal example of Eqs. (3.1), the term $-x$ of h maintains a small drift from $x = 0$ that prevents this situation. In Fig. 6, we observe that removing the term $-x$ from h results in the iterates leaving the stable valley for a significant slow time t_3, thus breaking the central flow prediction. More generally, we conjecture that a perturbation in x (transversal gradient of h, noise, numerical errors, etc.) is necessary for the central flow prediction to hold.

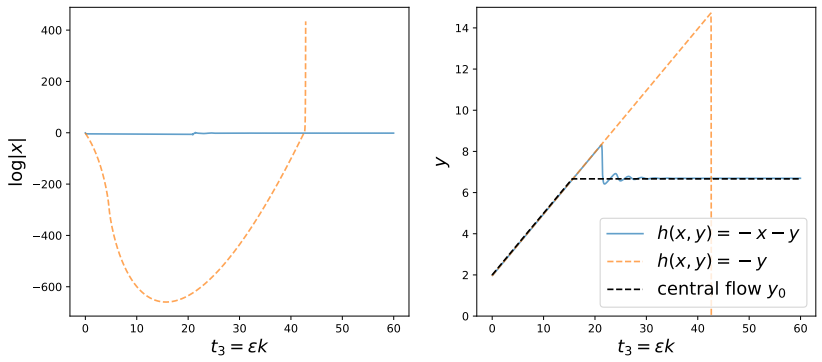
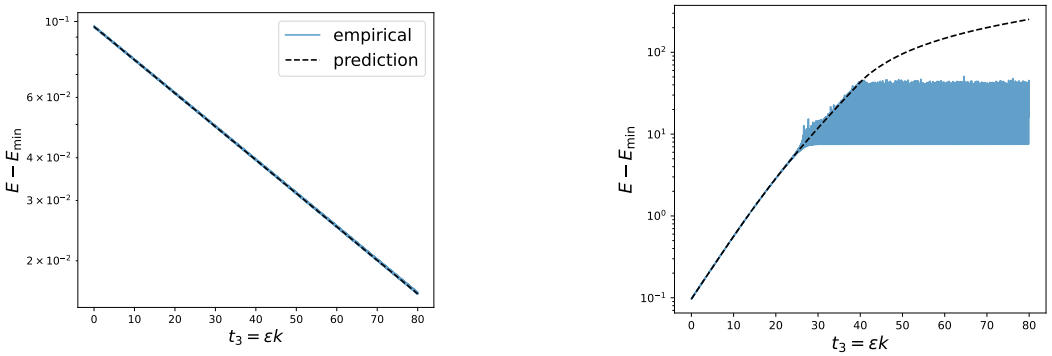


Figure 6: Iterates of gradient descent where g is chosen as in Eqs. (3.1) and $h(x, y) = -x - y$ or $h(x, y) = -y$. Here, $\eta = 0.3$, $\zeta = 1$ and $\varepsilon = 0.03$. In the absence of the term $-x$ in h, x becomes exponentially small in the first part of the dynamics, and thus it is possible to leave the stable valley, with the self-stabilization mechanism kicking in only after a significant slow time t_3. The self-stabilization mechanism is then so strong that the trajectory diverges.

D EMPIRICAL VALIDATION OF RESULT 3.3

In Fig. 7, we present an empirical validation of Result 3.3. The result is excellent at predicting the rate of decay or growth of the energy, up to a certain point: in the latter case, the energy does not diverge but saturates at a large value. In fact, E becomes so large that the scales of $\varepsilon^{1/2}a_1, \varepsilon^{1/2}y_1$ are no longer of order $\varepsilon^{1/2}$, breaking the asymptotic expansion ansatz used to derive Result 3.3. Our derivations are then no longer predictive of the actual dynamics and more refined computations would be required to study the saturation of the energy.

To summarize, the self-stabilization system (3.3) and Result 3.3 should be understood as an analysis at energy levels of order 1.



(a) $\zeta = 0.5$

(b) $\zeta = 0$

Figure 7: Comparison between the empirical variation of the energy and the theoretical prediction of Result 3.3. This experiment is initialized at the edge of stability. Here, $\eta = 0.3$ and $\varepsilon = 10^{-5}$. More precisely, we define $y_1^{\text{emp}} = \varepsilon^{-1/2}(y - y_0)$ and $a_1^{\text{emp}} = \varepsilon^{-1/2}(-1)^{t_1}(x - x_0)$. This defines an empirical energy $E^{\text{emp}} = (y_1^{\text{emp}})^2/2 + (a_1^{\text{emp}})^2/4 - \log((a_1^{\text{emp}})^2)/2$, that we compare against the theoretical prediction $E^{\text{pred}}(t_3) = E^{\text{emp}}(0) + \eta^2(2 - 9\zeta^2)\int_0^{t_3}(y_1^{\text{emp}})^2 \mathrm{d}s_3$. We denote $E_{\min} = \frac{1}{2}(1 - \log 2)$ the minimal possible energy.

E Derivation of Result 3.3

We identify the terms of order $\varepsilon^{3/2}$ in the expansion of the iteration $\Delta_k z = -\eta \nabla f(z)$. Using the expressions of Eqs. (B.1) and (B.3), we obtain for the x-component

$$\begin{aligned}\Delta_{t_1} x_3 + \partial_{t_2} x_2 + \partial_{t_2}\Delta_{t_1} x_2 + \partial_{t_3} x_1 + \frac{1}{2}\partial_{t_2}^2 x_1 + \partial_{t_3}\Delta_{t_1} x_1 + \frac{1}{2}\partial_{t_2}^2\Delta_{t_1} x_1 \\ = -\eta\left[y_0 x_3 + 3\zeta x_1 x_2 + x_1 y_2 + x_2 y_1\right] .\end{aligned}$$

Here, we are interested in the edge of stability phase where $\eta y_0 = 2$, $x_1 = (-1)^{t_1} a_1(t_2, t_3)$. Moreover, applying Lemma 3.2 to Eq. (3.6), we have $x_2 = -\frac{\eta}{2}\left[\frac{3}{2}\zeta a_1^2 - 1\right] + (-1)^{t_1} a_2$ for some $a_2 = a_2(t_2, t_3)$. This gives

$$\begin{aligned}\Delta_{t_1} x_3 + 2x_3 = \frac{3\eta}{4}\zeta\partial_{t_2}\left(a_1^2\right) - (-1)^{t_1}\partial_{t_2} a_2 + 2(-1)^{t_1}\partial_{t_2} a_2 - (-1)^{t_1}\partial_{t_3} a_1 \\ - \frac{1}{2}(-1)^{t_1}\partial_{t_2}^2 a_1 + 2(-1)^{t_1}\partial_{t_3} a_1 + (-1)^{t_1}\partial_{t_2}^2 a_1 \\ - \eta\Big[3\zeta(-1)^{t_1} a_1\left(-\frac{\eta}{2}\left[\frac{3}{2}\zeta a_1^2 - 1\right] + (-1)^{t_1} a_2\right) + (-1)^{t_1} a_1 y_2 \\ + \left(-\frac{\eta}{2}\left[\frac{3}{2}\zeta a_1^2 - 1\right] + (-1)^{t_1} a_2\right) y_1\Big] \\ = \frac{3\eta}{4}\zeta\partial_{t_2}\left(a_1^2\right) + (-1)^{t_1}\partial_{t_2} a_2 + (-1)^{t_1}\partial_{t_3} a_1 + \frac{1}{2}(-1)^{t_1}\partial_{t_2}^2 a_1 \\ - \eta\Big[3\zeta(-1)^{t_1} a_1\left(-\frac{\eta}{2}\left[\frac{3}{2}\zeta a_1^2 - 1\right] + (-1)^{t_1} a_2\right) + (-1)^{t_1} a_1 y_2 \\ + \left(-\frac{\eta}{2}\left[\frac{3}{2}\zeta a_1^2 - 1\right] + (-1)^{t_1} a_2\right) y_1\Big] .\end{aligned}$$

The right hand side depends on t_1 only through the $(-1)^{t_1}$ terms. By Lemma 3.2, to avoid a secular term, we must have

$$\begin{aligned}\partial_{t_2} a_2 + \partial_{t_3} a_1 + \frac{1}{2}\partial_{t_2}^2 a_1 = \eta\left(-3\zeta a_1\frac{\eta}{2}\left[\frac{3}{2}\zeta a_1^2 - 1\right] + a_1 y_2 + a_2 y_1\right) \\ = -\eta^2\frac{3}{2}\zeta a_1\left[\frac{3}{2}\zeta a_1^2 - 1\right] + \eta\left(a_1 y_2 + a_2 y_1\right) .\end{aligned} \tag{E.1}$$

We now identify terms of order $\varepsilon^{3/2}$ in the y equation (3.5) in Eqs. (B.1) and (B.3). Recall that $y_0 = 2/\eta$, $y_1 = y_1(t_2, t_3)$, $y_2 = y_2(t_2, t_3)$, leading to many simplifications. We obtain for the y-component

$$\Delta_{t_1} y_3 + \partial_{t_2} y_2 + \partial_{t_3} y_1 + \frac{1}{2}\partial_{t_2}^2 y_1 = -\eta x_1 x_2 = -\eta(-1)^{t_1} a_1\left(-\frac{\eta}{2}\left[\frac{3}{2}\zeta a_1^2 - 1\right] + (-1)^{t_1} a_2\right) .$$

We consider this as a difference equation for y_3 in t_1.

$$\Delta_{t_1} y_3 = -\partial_{t_2} y_2 - \partial_{t_3} y_1 - \frac{1}{2}\partial_{t_2}^2 y_1 - \eta(-1)^{t_1} a_1\left(-\frac{\eta}{2}\left[\frac{3}{2}\zeta a_1^2 - 1\right] + (-1)^{t_1} a_2\right) .$$

The right hand side depends on t_1 only through the $(-1)^{t_1}$ terms. To avoid a secular term, we must have

$$\partial_{t_2} y_2 + \partial_{t_3} y_1 + \frac{1}{2}\partial_{t_2}^2 y_1 = -\eta a_1 a_2 . \tag{E.2}$$

We now have derived our two working Eqs. (E.1) and (E.2). Recall that we seek to compute

$$\partial_{t_3} E = y_1\partial_{t_3} y_1 + \left(\frac{a_1}{2} - \frac{1}{a_1}\right)\partial_{t_3} a_1 .$$

This suggests considering the quantity $F = y_1 y_2 + \left(\frac{a_1}{2} - \frac{1}{a_1}\right) a_2$. Indeed, using Eqs. (E.1) and (E.2), we have

$$\begin{aligned}
\partial_{t_2} F &= (\partial_{t_2} y_1) y_2 + y_1 (\partial_{t_2} y_2) + \left(\frac{1}{2} + \frac{1}{a_1^2}\right)(\partial_{t_2} a_1) a_2 + \left(\frac{a_1}{2} - \frac{1}{a_1}\right)(\partial_{t_2} a_2) \\
&= \eta \left(1 - \frac{a_1^2}{2}\right) y_2 - y_1 \left(\partial_{t_3} y_1 + \frac{1}{2}\partial_{t_2}^2 y_1 + \eta a_1 a_2\right) + \left(\frac{1}{2} + \frac{1}{a_1^2}\right)(\eta a_1 y_1) a_2 \\
&\quad + \left(\frac{a_1}{2} - \frac{1}{a_1}\right)\left(-\partial_{t_3} a_1 - \frac{1}{2}\partial_{t_2}^2 a_1 - \eta^2 \frac{3}{2}\zeta a_1 \left[\frac{3}{2}\zeta a_1^2 - 1\right] + \eta(a_1 y_2 + a_2 y_1)\right) \\
&= -\partial_{t_3} E - \frac{y_1}{2}\partial_{t_2}^2 y_1 - \left(\frac{a_1}{2} - \frac{1}{a_1}\right)\frac{1}{2}\partial_{t_2}^2 a_1 - \eta^2 \frac{3}{2}\zeta \left(\frac{a_1^2}{2} - 1\right)\left(\frac{3}{2}\zeta a_1^2 - 1\right) .
\end{aligned}$$

The right-hand side depends on t_2 only through the functions a_1 and y_1, which are periodic in t_2. We then use the following lemma.

Lemma E.1. *Let $t \mapsto \varphi(t)$ be a periodic real function. Then $\varphi(t)$ has a bounded primitive if and only if the average of φ is 0.*

To avoid a secular term in F (which would induce a secular term in y_2 or a_2), we must have

$$\partial_{t_3} E = \left\langle -\frac{y_1}{2}\partial_{t_2}^2 y_1 + \left(\frac{1}{2a_1} - \frac{a_1}{4}\right)\partial_{t_2}^2 a_1 - \eta^2 \frac{3}{2}\zeta \left(\frac{a_1^2}{2} - 1\right)\left(\frac{3}{2}\zeta a_1^2 - 1\right)\right\rangle_{t_2} .$$

We now simplify this expression. We compute

$$\partial_{t_2}^2 y_1 = \eta \partial_{t_2}\left(1 - \frac{a_1^2}{2}\right) = -\eta a_1 \partial_{t_2} a_1 = -\eta^2 a_1^2 y_1 ,$$

$$\partial_{t_2}^2 a_1 = \eta \partial_{t_2}(a_1 y_1) = \eta\left[(\partial_{t_2} a_1) y_1 + a_1 (\partial_{t_2} y_1)\right] = \eta^2 \left[a_1 y_1^2 + a_1\left(1 - \frac{a_1^2}{2}\right)\right] .$$

This gives

$$\begin{aligned}
\partial_{t_3} E &= \eta^2 \left\langle \frac{a_1^2}{2} y_1^2 + \left(\frac{1}{2a_1} - \frac{a_1}{4}\right)\left[a_1 y_1^2 + a_1 - \frac{a_1^3}{2}\right] - \frac{3}{2}\zeta\left(\frac{a_1^2}{2} - 1\right)\left(\frac{3}{2}\zeta a_1^2 - 1\right)\right\rangle_{t_2} \\
&= \eta^2 \left\langle \frac{a_1^2}{2} y_1^2 + \frac{y_1^2}{2} + \frac{1}{2} - \frac{a_1^2}{4} - \frac{a_1^2}{4} y_1^2 - \frac{a_1^2}{4} + \frac{a_1^4}{8} - \frac{9}{8}\zeta^2 a_1^4 + \frac{3}{4}\zeta a_1^2 + \frac{9}{4}\zeta^2 a_1^2 - \frac{3}{2}\zeta\right\rangle_{t_2} \\
&= \eta^2 \left(\frac{\langle a_1^2 y_1^2\rangle_{t_2}}{4} + \frac{\langle y_1^2\rangle_{t_2}}{2} + \frac{1}{2} - \frac{3}{2}\zeta + \left(-\frac{1}{2} + \frac{3\zeta}{4} + \frac{9\zeta^2}{4}\right)\langle a_1^2\rangle_{t_2} + \frac{1}{8}\left(1 - 9\zeta^2\right)\langle a_1^4\rangle_{t_2}\right) .
\end{aligned}$$

We now compute the average values that appear above. Note that, as y_1 is periodic in t_2,

$$0 = \langle \partial_{t_2} y_1\rangle_{t_2} = \eta\left(1 - \frac{\langle a_1^2\rangle_{t_2}}{2}\right) ,$$

thus $\langle a_1^2\rangle_{t_2} = 2$. Similarly,

$$0 = \left\langle \partial_{t_2}(y_1^3)\right\rangle_{t_2} = 3\left\langle y_1^2 \partial_{t_2} y_1\right\rangle_{t_2} = 3\eta\left\langle y_1^2\left(1 - \frac{a_1^2}{2}\right)\right\rangle_{t_2} = 3\eta\left(\langle y_1^2\rangle_{t_2} - \frac{\langle y_1^2 a_1^2\rangle_{t_2}}{2}\right) ,$$

thus $\langle y_1^2 a_1^2\rangle_{t_2} = 2\langle y_1^2\rangle_{t_2}$. Finally,

$$\begin{aligned}
0 &= \left\langle \partial_{t_2}(a_1^2 y_1)\right\rangle_{t_2} = \eta\left\langle 2a_1(\partial_{t_2} a_1) y_1 + a_1^2 \partial_{t_2} y_1\right\rangle_{t_2} = \eta\left\langle 2a_1^2 y_1^2 + a_1^2\left(1 - \frac{a_1^2}{2}\right)\right\rangle_{t_2} \\
&= \eta\left[2\left\langle a_1^2 y_1^2\right\rangle_{t_2} + \left\langle a_1^2\right\rangle_{t_2} - \frac{\left\langle a_1^4\right\rangle_{t_2}}{2}\right] = \eta\left[4\left\langle y_1^2\right\rangle_{t_2} + 2 - \frac{\left\langle a_1^4\right\rangle_{t_2}}{2}\right] ,
\end{aligned}$$

thus $\langle a_1^4\rangle_{t_2} = 4 + 8\langle y_1^2\rangle_{t_2}$. Overall, we obtain

$$\begin{aligned}
\partial_{t_3} E &= \eta^2\left[\frac{\langle y_1^2\rangle_{t_2}}{2} + \frac{\langle y_1^2\rangle_{t_2}}{2} + \frac{1}{2} - \frac{3}{2}\zeta - 1 + \frac{3\zeta}{2} + \frac{9\zeta^2}{2} + (1 - 9\zeta^2)\left(\frac{1}{2} + \langle y_1^2\rangle_{t_2}\right)\right] \\
&= \eta^2 (2 - 9\zeta^2)\langle y_1^2\rangle_{t_2} .
\end{aligned}$$

Remark E.2. In the above derivations, ensuring that $F = y_1 y_2 + \left(\frac{a_1}{2} - \frac{1}{a_1}\right) a_2$ is non-secular suffices to compute the variations of the energy E. However, ensuring that y_2 and a_2 are both non-secular would require a condition on an independent combination of y_2 and a_2. This additional condition would compute the phase shift of the self-stabilization fluctuations in the slow timescale t_3. This phase shift can be observed in Fig. 2, bottom plots.

F Derivation of Result 4.1

As before, we identify terms in the expansion of

$$\Delta_k x = -\eta \partial_x f(x, y)\,, \tag{F.1}$$
$$\Delta_k y = -\eta \nabla_y f(x, y)\,, \tag{F.2}$$

using the expansions of Eqs. (B.1) and (B.2). Eq. (B.1) gives that

$$\begin{aligned}
\Delta_k x &= \varepsilon^{1/2} \Delta_{t_1} x_1 + \varepsilon \left[\Delta_{t_1} x_2 + \partial_{t_2} x_1 + \partial_{t_2} \Delta_{t_1} x_1\right] + o(\varepsilon)\,, \\
\Delta_k y &= \Delta_{t_1} y_0 + \varepsilon^{1/2} \left[\Delta_{t_1} y_1 + \partial_{t_2} y_0 + \partial_{t_2} \Delta_{t_1} y_0\right] \\
&\quad + \varepsilon \left[\Delta_{t_1} y_2 + \partial_{t_2} y_1 + \partial_{t_2} \Delta_{t_1} y_1 + \partial_{t_3} y_0 + \frac{1}{2} \partial_{t_2}^2 y_0 + \partial_{t_3} \Delta_{t_1} y_0 + \frac{1}{2} \partial_{t_2}^2 \Delta_{t_1} y_0\right] + o(\varepsilon)\,.
\end{aligned}$$

Recall that for all $y \in Y$, we have $g(0, y) = \min g$. In particular, this implies that for all $p \geqslant 1$, $\nabla_y^p g(0, y) = 0$. Moreover, it also implies that for all $y \in Y$, $\partial_x g(0, y) = 0$. As a consequence, for all $p \geqslant 0$, $\nabla_y^p \partial_x g(0, y) = 0$. This simplifies many terms in Eq. (B.2). Combining with the assumption $x_0 = 0$ and the notation $S(0, y) = \partial_x^2 g(0, y)$, we obtain

$$\begin{aligned}
-\eta \partial_x f(x, y) &= -\varepsilon^{1/2} \eta S(0, y_0) x_1 \\
&\quad - \varepsilon \eta \left[S(0, y_0) x_2 + \frac{1}{2} \partial_x S(0, y_0) x_1^2 + x_1 \left\langle \nabla_y S(0, y_0), y_1 \right\rangle + \partial_x h(0, y_0)\right] \\
&\quad + o(\varepsilon)\,, \\
-\eta \nabla_y f(x, y) &= -\varepsilon \eta \left[\frac{x_1^2}{2} \nabla_y S(0, y_0) + \nabla_y h(0, y_0)\right] + o(\varepsilon)\,.
\end{aligned} \tag{F.3}$$

In the derivations that follow, we do not write explicitly the evaluations at $(0, y_0)$ anymore. We can now proceed to the identification of terms in the expansion of Eqs. (F.1)–(F.2).

Order 1**.** As in the derivation of Result 3.1, this order gives $y_0 = y_0(t_2, t_3)$.

Order $\varepsilon^{1/2}$**.** The x equation (F.1) gives

$$\Delta_{t_1} x_1 = -\eta S x_1\,.$$

We distinguish whether we are at the edge of stability.

- If $\eta S < 2$, recall that, by assumption from Sec. 2, we also have $\eta S > 0$. Thus $x_1 \to 0$ as $t_1 \to \infty$.
- If $\eta S = 2$, then $x_1 = (-1)^{t_1} a_1(t_2, t_3)$ for some a_1.

As before, the y equation (F.2) gives $y_1 = y_1(t_2, t_3)$ and $y_0 = y_0(t_3)$.

Order ε**.** The x equation (F.1) gives

$$\Delta_{t_1} x_2 + \partial_{t_2} x_1 + \partial_{t_2} \Delta_{t_1} x_1 = -\eta \left[S x_2 + \frac{1}{2} (\partial_x S) x_1^2 + x_1 \left\langle \nabla_y S, y_1 \right\rangle + \partial_x h\right]\,.$$

Again, we focus on understanding what happens at the edge of stability, i.e., when $\eta S = 2$. We obtain

$$\Delta_{t_1} x_2 + 2 x_2 = (-1)^{t_1} \partial_{t_2} a_1 - \eta \left[\frac{1}{2} (\partial_x S) a_1^2 + (-1)^{t_1} a_1 \left\langle \nabla_y S, y_1 \right\rangle + \partial_x h\right]\,.$$

By Lemma 3.2, we must have $\partial_{t_2} a_1 = \eta a_1 \left\langle \nabla_y S, y_1 \right\rangle$ to avoid a secular term.

We now turn to the y equation (F.2). It gives

$$\Delta_{t_1} y_2 + \partial_{t_2} y_1 + \partial_{t_3} y_0 = -\eta \left[\frac{x_1^2}{2} \nabla_y S + \nabla_y h \right] .$$

We distinguish whether we are at the edge of stability.

- If $\eta S < 2$, then $x_1 \to 0$ as $t_1 \to \infty$. To prevent y_2 from developing a secular term on the timescale t_1, we must have $\partial_{t_2} y_1 + \partial_{t_3} y_0 = -\eta \nabla_y h$. Here, all quantities but y_1 are independent of t_2. To prevent a secular term, we must have

  $$\partial_{t_3} y_0 = -\eta \nabla_y h \, .$$

  In the relative interior of the stable region, the dynamics follow the gradient flow of h constrained to the valley.

- If $\eta S = 2$, then we have $x_1 = (-1)^{t_1} a_1(t_2, t_3)$. We obtain

  $$\Delta_{t_1} y_2 = -\partial_{t_2} y_1 - \partial_{t_3} y_0 - \eta \left[\frac{a_1^2}{2} \nabla_y S + \nabla_y h \right] .$$

  The right-hand side is independent of t_1. To avoid a secular term, we must have

  $$\partial_{t_2} y_1 = -\partial_{t_3} y_0 - \eta \left[\frac{a_1^2}{2} \nabla_y S + \nabla_y h \right] . \tag{F.4}$$

  Differentiating the edge of stability condition $\eta S(0, y_0) = 2$ with respect to t_3, we obtain $\langle \partial_{t_3} y_0, \nabla_y S \rangle = 0$. This suggests taking the inner product of Eq. (F.4) with $\nabla_y S$. Recall that we denote $u_1 = \langle y_1, \nabla_y S \rangle$. We obtain

  $$\partial_{t_2} u_1 = -\eta \left[\|\nabla_y S\|^2 \, \frac{a_1^2}{2} + \langle \nabla_y h, \nabla_y S \rangle \right] .$$

  This finishes the derivation of Eqs. (4.1).

  As discussed below the result, the self-stabilization system (4.1) implies that a_1 is periodic in t_2 (for fixed t_3). Thus in the right hand side of Eq. (F.4), all quantities are independent of t_2 except for a_1 which is periodic in t_2.

  By Lemma E.1, to avoid y_1 developing a secular term in t_2, we must have

  $$\partial_{t_3} y_0 = -\eta \left[\nabla_y h + \frac{\langle a_1^2 \rangle_{t_2}}{2} \nabla_y S \right] .$$

This completes the derivation of item (i) by taking $\sigma^2 = \langle a_1^2 \rangle_{t_2}$ if $\eta S = 2$ and $\sigma^2 = 0$ if $\eta S < 2$, and thus completes the derivation of the result.

G Derivation of Result 5.1

As before, we identify terms in the expansion of

$$\Delta_k x = -\eta \nabla_x f(x, y) \, , \tag{G.1}$$
$$\Delta_k y = -\eta \nabla_y f(x, y) \, , \tag{G.2}$$

using the expansions of Eqs. (B.1) and (B.2). The expansion of the right hand side is the same as in Eq. (F.3), except for a few modifications due to the fact that X is multidimensional:

- ∂_x is replaced by ∇_x,
- $S(0, y) = \partial_x^2 g(0, y) \in \mathbb{R}$ is replaced by $S(0, y) = \nabla_x^2 g(0, y) \in \mathrm{Sym}_+ X$,
- $\partial_x S(0, y) \in \mathbb{R}$ is replaced by $D_x S(0, y) : X \to \mathrm{Sym}\, X$,
- $\nabla_y S(0, y) \in Y$ is replaced by $D_y S(0, y) : Y \to \mathrm{Sym}\, X$.

We obtain:

$$\begin{aligned}
-\eta\nabla_x f(x,y) &= -\varepsilon^{1/2}\eta S(0,y_0)x_1 \\
&\quad - \varepsilon\eta\left[S(0,y_0)x_2 + \frac{1}{2}D_xS(0,y_0)[x_1]x_1 + D_yS(0,y_0)[y_1]x_1 + \nabla_x h(0,y_0)\right] \\
&\quad + o(\varepsilon)\,, \\
-\eta\nabla_y f(x,y) &= -\varepsilon\eta\left[\frac{1}{2}D_yS(0,y_0)^*[x_1x_1^\top] + \nabla_y h(0,y_0)\right] + o(\varepsilon)\,.
\end{aligned}$$

In the derivations that follow, we do not write explicitly the evaluations at $(0, y_0)$ anymore. We can now proceed to the identification of terms in the expansion of Eqs. (G.1)–(G.2).

Order 1**.** As before, this order gives $y_0 = y_0(t_2, t_3)$.

Order $\varepsilon^{1/2}$**.** The x equation (G.1) gives

$$\Delta_{t_1} x_1 = -\eta S x_1\,.$$

We distinguish directions depending on whether they are critical. Recall that $C = C(t_3) = \ker(2\,\mathrm{Id}_X - \eta S)$ denotes the critical subspace. We use a spectral decomposition of ηS. By assumption from Sec. 2, all eigenvalues are positive. The eigenvectors associated the eigenvalue 2 form a basis of C, while the other eigenvectors are associated with eigenvalues in $(0, 2)$ and form a basis of $C^\perp$.

Denoting $x_1^\parallel = P_C x_1$ and $x_1^\perp = P_{C^\perp} x_1$, this gives $x_1^\perp \xrightarrow[t_1\to\infty]{} 0$, with $\sum_{t_1=0}^{+\infty} \|x_1^\perp(t_1,t_2,t_3)\| < +\infty$ and $x_1^\parallel = (-1)^{t_1} a_1$ for some $a_1 = a_1(t_2, t_3) \in C$.

As before, the y equation (G.2) gives $y_1 = y_1(t_2, t_3)$ and $y_0 = y_0(t_3)$.

Order ε**.** The x equation (G.1) gives

$$\Delta_{t_1} x_2 + \partial_{t_2} x_1 + \partial_{t_2}\Delta_{t_1} x_1 = -\eta\left[Sx_2 + \frac{1}{2}D_xS[x_1]x_1 + D_yS[y_1]x_1 + \nabla_x h\right]\,.$$

We focus on the critical directions, i.e., we apply P_C to both sides of the equation. Denoting $x_2^\parallel = P_C x_2$, we obtain

$$\Delta_{t_1} x_2^\parallel + \partial_{t_2} x_1^\parallel + \partial_{t_2}\Delta_{t_1} x_1^\parallel = -\eta\left[P_C S x_2 + \frac{1}{2}P_C D_xS[x_1]x_1 + P_C D_yS[y_1]x_1 + P_C\nabla_x h\right]\,.$$

Recall that C is the eigenspace of ηS associated with the eigenvalue 2, thus $\eta P_C S = 2P_C$. Moreover, we decompose $x_1 = \iota_C x_1^\parallel + \iota_{C^\perp} x_1^\perp = (-1)^{t_1}\iota_C a_1 + \iota_{C^\perp} x_1^\perp$. This gives

$$\begin{aligned}
\Delta_{t_1} x_2^\parallel + 2x_2^\parallel = (-1)^{t_1}\partial_{t_2} a_1 - \eta\Big[&\frac{1}{2}P_C D_xS[\iota_C a_1]\iota_C a_1 + \frac{1}{2}(-1)^{t_1}P_C D_xS[\iota_{C^\perp}x_1^\perp]\iota_C a_1 \\
&+ \frac{1}{2}(-1)^{t_1}P_C D_xS[\iota_C a_1]\iota_{C^\perp}x_1^\perp + \frac{1}{2}P_C D_xS[\iota_{C^\perp}x_1^\perp]\iota_{C^\perp}x_1^\perp \\
&+ (-1)^{t_1}P_C D_yS[y_1]\iota_C a_1 + P_C D_yS[y_1]\iota_{C^\perp}x_1^\perp + P_C\nabla_x h\Big]\,.
\end{aligned}$$

The right hand side depends on t_1 only through the $(-1)^{t_1}$ terms and $x_1^\perp$ terms, but the latter are vanishing as $t_1 \to \infty$, with $\sum_{t_1=0}^{+\infty}\|x_1^\perp\| < +\infty$. This suggests studying the following generalization of Lemma 3.2.

Lemma G.1. *Let $\alpha, \beta \in \mathbb{R}$ and $\gamma = \gamma(t_1) \in \mathbb{R}$ such that $\sum_{t_1=0}^{+\infty}|\gamma(t_1)| < +\infty$. The solutions $\varphi = \varphi(t_1)$ of the difference equation*

$$\Delta_{t_1}\varphi + 2\varphi = \alpha + \beta(-1)^{t_1} + \gamma(t_1)$$

are

$$\varphi = \frac{\alpha}{2} + a(-1)^{t_1} - \beta t_1(-1)^{t_1} - \sum_{s=0}^{t_1-1}(-1)^{t_1+s}\gamma(s)\,, \qquad a \in \mathbb{R}\,.$$

In particular, there exists a bounded solution if and only if $\beta = 0$.

Proof. As $(-1)^{t_1}$ is solution to the homogeneous equation, we seek solutions of the form $\varphi = A(-1)^{t_1}$ for some $A = A(t_1)$. We compute

$$\begin{aligned}\Delta_{t_1}\varphi + 2\varphi &= \varphi(t_1+1) - \varphi(t_1) + 2\varphi(t_1) = \varphi(t_1+1) + \varphi(t_1) \\ &= A(t_1+1)(-1)^{t_1+1} + A(t_1)(-1)^{t_1} = -(-1)^{t_1}\Delta_{t_1} A\,.\end{aligned}$$

The function $\varphi = A(-1)^{t_1}$ solves the difference equation if and only if A solves

$$\Delta_{t_1} A = -\alpha(-1)^{t_1} - \beta - (-1)^{t_1}\gamma(t_1)\,.$$

The discrete primitives are

$$A = a + \frac{\alpha}{2}(-1)^{t_1} - \beta t_1 - \sum_{s=0}^{t_1-1}(-1)^s\gamma(s)\,, \qquad a \in \mathbb{R}\,.$$

This gives the result. □

Applying Lemma G.1 to the equation above, we obtain that to avoid a secular term, we must have

$$\partial_{t_2} a_1 = \eta P_C D_y S[y_1]\iota_C a_1 = \eta\Psi[y_1]a_1 = \eta U_1 a_1\,.$$

We now turn to the y equation (G.2). It gives

$$\Delta_{t_1} y_2 + \partial_{t_2} y_1 + \partial_{t_3} y_0 = -\eta\left[\frac{1}{2}(D_y S)^*[x_1 x_1^\top] + \nabla_y h\right]\,.$$

Again, we decompose $x_1 = (-1)^{t_1}\iota_C a_1 + \iota_{C^\perp} x_1^\perp$.

$$\begin{aligned}&\Delta_{t_1} y_2 + \partial_{t_2} y_1 + \partial_{t_3} y_0 \\ &\quad= -\eta\Big[\frac{1}{2}(D_y S)^*[\iota_C a_1 a_1^\top P_C] + \frac{(-1)^{t_1}}{2}(D_y S)^*\left[\iota_C a_1 x_1^{\perp\top} P_{C^\perp} + \iota_{C^\perp} x_1^\perp a_1^\top P_C\right] \\ &\qquad + \frac{1}{2}(D_y S)^*[\iota_{C^\perp} x_1^\perp x_1^{\perp\top} P_{C^\perp}] + \nabla_y h\Big]\,.\end{aligned}$$

Here, the right hand side depends on t_1 only through the $(-1)^{t_1}$ terms and $x_1^\perp$ terms, but the latter are vanishing as $t_1 \to \infty$. To avoid a secular term, we must have

$$\partial_{t_2} y_1 + \partial_{t_3} y_0 = -\eta\left[\frac{1}{2}(D_y S)^*[\iota_C a_1 a_1^\top P_C] + \nabla_y h\right]\,.$$

Note that for $M \in \operatorname{Sym} C$, $\Psi^*[M] = (D_y S)^*[\iota_C M P_C]$. Thus we have

$$\partial_{t_2} y_1 + \partial_{t_3} y_0 = -\eta\left[\frac{1}{2}\Psi^*[a_1 a_1^\top] + \nabla_y h\right]\,. \tag{G.3}$$

Here, we want to use that the evolution of $S(0, y_0)$ is constrained in the critical subspace C. We use the following lemma.

Lemma G.2. *Let $t \mapsto M(t) \succcurlyeq 0$ be a smooth curve of positive semidefinite matrices, $P = P(t)$ denote the orthogonal projection onto $\ker M(t)$, and $\iota = \iota(t) = P^*$ the corresponding inclusion operator. Then, for all t, we have $P\partial_t M\iota = 0$.*

Proof. Let $a \in \ker M(t)$. Then $a^\top M(t) a = 0$. Moreover, for all s, we have $a^\top M(s) a \geqslant 0$. We thus have, for all $a \in \ker M(t)$, $\partial_t(a^\top M(t) a) = a^\top \partial_t M(t) a = 0$.

Now, consider $a, b \in \ker M(t)$. Then, by polarization ($\partial_t M(t)$ is symmetric),

$$2a^\top \partial_t M(t) b = (a+b)^\top \partial_t M(t)(a+b) - a^\top \partial_t M(t) a - b^\top \partial_t M(t) b = 0\,.$$

Thus $P\partial_t M\iota = 0$. □

Here, we apply the lemma to $M(t_3) = 2\,\mathrm{Id}_X - \eta S(0, y_0(t_3))$. We obtain

$$0 = P_C \partial_{t_3} S \iota_C = P_C D_y S[\partial_{t_3} y_0] \iota_C = \Psi[\partial_{t_3} y_0]\,.$$

This suggests evaluating Ψ at Eq. (G.3). We obtain

$$\partial_{t_2} U_1 = -\eta \left[\frac{1}{2} \Psi\Psi^*[a_1 a_1^\top] + \Psi[\nabla_y h] \right] .$$

This completes the derivation of the self-stabilization system (5.1).

We do not know whether the self-stabilization system induces periodic solutions in t_2 when more than one eigenvalue is on the edge of stability. To give a meaning to the average $\langle . \rangle_{t_2}$, we take the limit $T_2 \to \infty$ of the average over $[0, T_2]$ (if the limit exists). For instance, to avoid a secular term, y_1 is uniformly bounded in t_2, thus we have

$$\langle \partial_{t_2} y_1 \rangle_{t_2} = \lim_{T_2 \to \infty} \frac{1}{T_2} \int_0^{T_2} \partial_{t_2} y_1 \mathrm{d}t_2 = \lim_{T_2 \to \infty} \frac{y_1(T_2, t_3) - y_1(0, t_3)}{T_2} = 0\,.$$

Taking the average of Eq. (G.3) in t_2, we obtain

$$\partial_{t_3} y_0 = -\eta \left[\nabla_y h + \frac{1}{2} \langle \Psi^*[a_1 a_1^\top] \rangle_{t_2} \right] .$$

Provided that the average $\langle a_1 a_1^\top \rangle_{t_2}$ is well-defined, we have

$$\langle \Psi^*[a_1 a_1^\top] \rangle_{t_2} = \Psi^*[\langle a_1 a_1^\top \rangle_{t_2}] = (D_y S)^*[\iota_C \langle a_1 a_1^\top \rangle_{t_2} P_C]\,,$$

thus the result holds with $\Sigma = \iota_C \langle a_1 a_1^\top \rangle_{t_2} P_C$, which is complementary to $2\,\mathrm{Id}_X - \eta S$ (by definition of $C = \ker(2\,\mathrm{Id}_X - \eta S)$). If we do not make the assumption that the average $\langle a_1 a_1^\top \rangle_{t_2}$ is well-defined, we use that a_1 is bounded in t_2 (it is non-secular) and thus the averages $\frac{1}{T_2} \int_0^{T_2} a_1 a_1^\top \mathrm{d}t_2$ have an accumulation point $\widetilde{\Sigma} \in \mathrm{Sym}_+ C$ as $T_2 \to \infty$. We then have $\langle \Psi^*[a_1 a_1^\top] \rangle_{t_2} = \Psi^*[\widetilde{\Sigma}]$, and we can define $\Sigma = \iota_C \widetilde{\Sigma} P_C \in \mathrm{Sym}_+ X$. This gives the derivation of item (i) and completes the derivation of Result 5.1.

H Illustration of Result 5.1

In Fig. 8, we plot the iterates of gradient descent with learning rate $\eta = 0.3$ on $f = g + \varepsilon h$, with $\varepsilon = 0.01$. In this example, we take $Z = \mathbb{R}^5$, $X = \mathbb{R}^2$ and $Y = \mathbb{R}^3$, thus the valley is of dimension 3 and codimension 2. Further, we take $g(x, y) = \frac{1}{2} x^\top H[y] x + \frac{1}{2}(x[1]^3 + x[2]^3)$ with $H[y] = \begin{pmatrix} y[1] & \frac{y[3]}{\sqrt{2}} \\ \frac{y[3]}{\sqrt{2}} & y[2] \end{pmatrix}$ and $h(x, y) = -x[1] - 2y[1] - y[2] - y[3]$. Note that the assumptions of Sec. 2 are satisfied in the region where $H[y] + \mathrm{diag}(x[1], x[2]) \succ 0$ (where the iterates remain). Moreover, $S(0, y) = H[y]$ and $V_{\mathrm{S}} = \{(0, y), \eta H[y] \preccurlyeq 2\,\mathrm{Id}_2\}$.

I Implications for the loss

We discuss here the implications of Result 5.1 for the loss $f(x, y)$, as it is the most commonly observed quantity in practice.

Corollary I.1. *Under the assumptions of Result 5.1, we have*

$$f(x, y) = \min g + \varepsilon f_2 + o(\varepsilon)$$

where

$$f_2 \xrightarrow[t_1 \to \infty]{} h(0, y_0) + \frac{\|a_1\|^2}{\eta}\,.$$

The dominant non-constant term of the loss is of order ε. In the relative interior of the stable valley, this term depends only on t_3 (up to a potential initial transient term in t_1), through the component h of the loss. At the edge of stability, an additional increase $\|a_1\|^2/\eta$ appears, depending on both t_2

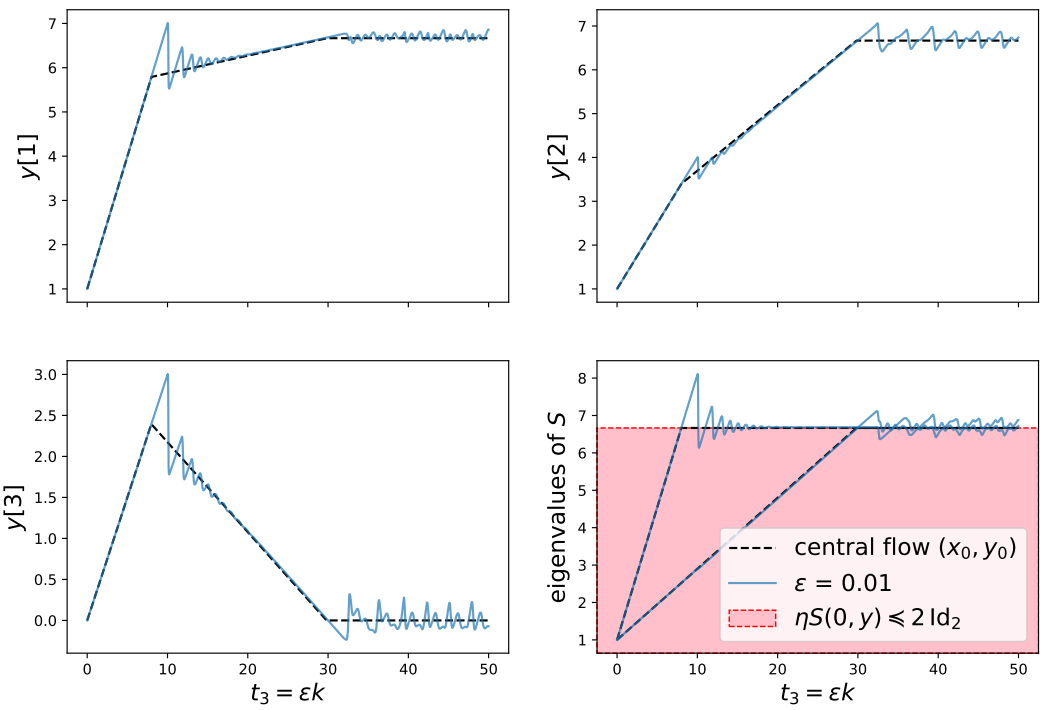


Figure 8: Iterates of gradient descent on a problem with a valley of dimension 3 and codimension 2. See App. H for details.

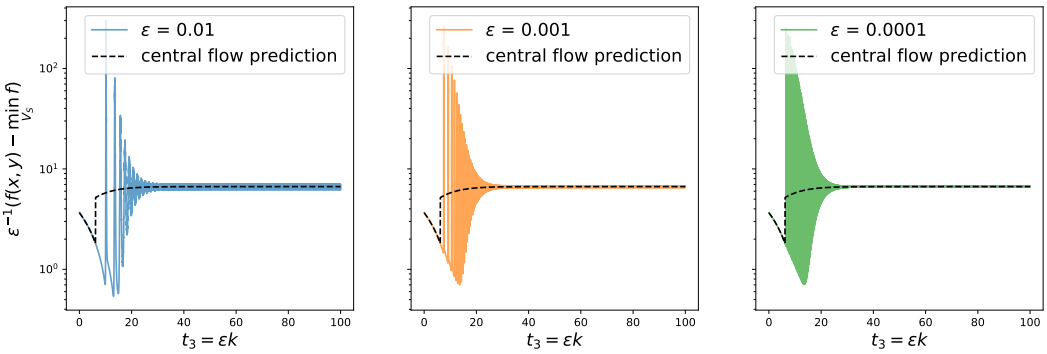


Figure 9: Rescaled loss $\varepsilon^{-1}\left(f(x(k), y(k)) - \min_{(x,y)\in V_S} f(x, y)\right)$ along the iterates of gradient descent and its central flow prediction $h(0, y_0) - \min_{(x,y)\in V_S} h(x, y) + \sigma^2/\eta$, in the same experiment as in Fig. 4. The loss progress made on h in the valley and the loss increase due to oscillations are both of the same order ε.

and t_3. In t_2, the self-stabilization fluctuations induce spikes in the loss. Averaging over these, and assuming that the average $\langle a_1 a_1^\top \rangle_{t_2}$ is well-defined, we obtain

$$\langle f_2 \rangle_{t_2} \xrightarrow[t_1 \to \infty]{} h(0, y_0) + \frac{\operatorname{Tr} \Sigma}{\eta},$$

consistently with the result of Cohen et al. (2025, Sec. 3.2.3). These results are illustrated in Fig. 9.

Derivation of Corollary I.1. We compute

$$\begin{aligned} f(x,y) &= g(x,y) + \varepsilon h(x,y) \\ &= g(\varepsilon^{1/2}x_1 + \varepsilon x_2 + o(\varepsilon), y_0 + \varepsilon^{1/2}y_1 + \varepsilon y_2 + o(\varepsilon)) + \varepsilon h(o(1), y_0 + o(1)) \\ &= g(0,y_0) + \left\langle \nabla_x g(0,y_0), \varepsilon^{1/2}x_1 + \varepsilon x_2 \right\rangle + \left\langle \nabla_y g(0,y_0), \varepsilon^{1/2}y_1 + \varepsilon y_2 \right\rangle \\ &\quad + \frac{1}{2}\left\langle \varepsilon^{1/2}x_1, \nabla_x^2 g(0,y_0)\varepsilon^{1/2}x_1 \right\rangle + \left\langle \varepsilon^{1/2}x_1, \nabla_x\nabla_y g(0,y_0)\varepsilon^{1/2}y_1 \right\rangle \\ &\quad + \frac{1}{2}\left\langle \varepsilon^{1/2}y_1, \nabla_y^2 g(0,y_0)\varepsilon^{1/2}y_1 \right\rangle + \varepsilon h(0,y_0) + o(\varepsilon)\,. \end{aligned}$$

As in the derivations above, we have $\nabla_x g(0,y_0) = 0$, $\nabla_y g(0,y_0) = 0$, $\nabla_x^2 g(0,y_0) = S(0,y_0)$, $\nabla_x\nabla_y g(0,y_0) = 0$, and $\nabla_y^2 g(0,y_0) = 0$. Moreover, we have $g(0,y_0) = \min g$. We obtain

$$f(x,y) = \min g + \varepsilon f_2 + o(\varepsilon)\,, \qquad f_2 = h(0,y_0) + \frac{1}{2}\langle x_1, S(0,y_0)x_1\rangle\,.$$

Here, we decompose $x_1 = (-1)^{t_1}\iota_C a_1 + \iota_{C^\perp}x_1^\perp$. As $x_1^\perp \to 0$ as $t_1 \to \infty$, and C is the eigenspace of $S(0,y_0)$ associated to the eigenvalue $2/\eta$, we have

$$f_2 \xrightarrow[t_1\to\infty]{} h(0,y_0) + \frac{1}{2}\langle \iota_C a_1, S(0,y_0)\iota_C a_1\rangle = h(0,y_0) + \frac{\|a_1\|^2}{\eta}\,.$$

□