

WORKING PAPERS

N° 1735

April 2026

“Gradient Extremals, Talwegs, Valleys, and Directional
Alignment for Generic Gradient Descent”

Pascal Bégout, Jérôme Bolte, Thomas Mariotti,
Francisco J. Silva

Gradient extremals, talwegs, valleys, and directional alignment for generic gradient descent

Pascal Bégout* Jérôme Bolte† Thomas Mariotti‡ Francisco J. Silva§

April 11, 2026

Abstract

Gradient extremals are loci along which the gradient is an eigenvector of the Hessian. These objects provide a natural geometric framework connecting several notions, notably valleys and talwegs, which we analyze from a variational viewpoint in the generic case. We then show that trajectories of the gradient flow and of its discrete counterpart exhibit directional alignment with the tangent spaces to gradient extremals, and generically to the talweg. Under non-resonance assumptions, and in contrast with the quadratic case, alignment rates are governed either by the first spectral gap or by the smallest eigenvalue of the Hessian at the limit point. Nonlinearities and the step length may both distort these rates in a complex manner. We further prove a volume concentration phenomenon emphasizing the structuring role of gradient extremals: for large times, the images of sets of initial conditions concentrate inside valleys and asymptotically around talwegs.

1 Introduction

Cauchy's gradient descent method is central in many branches of mathematics and has gained renewed prominence because of its fundamental role in applications such as modern machine learning, see e.g. [2], or optimal transport and PDEs, see e.g. [23].

In this paper we investigate the directional properties of the gradient descent dynamics and of its discrete counterpart. This is not a new theme: it was, for instance, part of the program surrounding what was known as Thom's gradient conjecture, eventually established in [17]. Our objective here is rather different as we aim to clarify several geometric structures that govern the behavior of trajectories near critical points in the generic case, that is, for functions having a nondegenerate Hessian with simple spectrum at their critical points.

To this end we formalize a number of natural objects that have appeared in various contexts [11, 8], sometimes in other fields (for instance in chemistry, see e.g. [24]), but whose presentation in the literature remains scattered. We argue that these objects may

*Toulouse School of Economics, Université Toulouse Capitole, Institut de Mathématiques de Toulouse, Toulouse, France, Pascal.Begout@math.cnrs.fr, <https://orcid.org/0000-0002-9172-3057>.

†Toulouse School of Economics, Université Toulouse Capitole, École Normale Supérieure, France, jbolte@ut-capitole.fr.

‡Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France, CEPR, and CESifo. thomas.mariotti@tse-fr.eu.

§Université de Limoges, XLIM, UMR CNRS 7252, France, francisco.silva@unilim.fr.

be unified through the notion of gradient extremals. Gradient extremals are paths along which the gradient is an eigenvector of the Hessian, and they are not, in general, gradient curves. Equivalently, they may be interpreted as loci where the slope of the function is critical relative to its level sets, under suitable nondegeneracy assumptions. This framework naturally leads to the notions of valleys and crest extremals, and calls for a clarification of their connection with the concept of a talweg. We study these notions in the generic case and show in particular that the variational definition of a talweg—minimal slope locus on each level set—corresponds to the gradient extremal associated with the minimal eigenvalue Theorem 2.4. We also investigate the infinitesimal structure of valleys, see Proposition 2.9.

Talwegs have a long history in topography and mathematics, starting with the pioneering works of de Saint-Venant [9], Cayley [6], Maxwell [21], and Boussinesq [5] on contour and slope lines, and the definition of hills and dales. More recently, talwegs have played a fundamental role in the theory of convergence, in particular through Lojasiewicz-type inequalities [19]. Their definition is not uniform across the literature; here we view them as paths along which the slope is locally minimal along the corresponding level set. The works of D’Acunto [7] and Kurdyka [16] are pioneering in the use of this concept and in establishing its connection with the convergence of gradient curves, although it also appears in Lojasiewicz work [20]. In [4] a transparent link between the two notions is provided as it is shown that the existence of a talweg of finite length is both necessary and sufficient for the Kurdyka-Lojasiewicz inequality to hold.

Our analysis departs from these general situations and focuses on the generic case, where gradient trajectories are known to converge generically to local minimizers [26, 22, 18]. In this stability framework, gradient extremals, talwegs, and crest extremals form locally one-dimensional manifolds that intersect orthogonally at their common critical point. They provide a geometric skeleton for the gradient flow, much as eigenspaces provide natural coordinates for gradient descent quadratic functions, i.e., for the exponential function of a symmetric matrix. In a second part of the paper we show that talwegs act as secondary attractors to gradient dynamics: generically, directions and secants of gradient trajectories, both in continuous and discrete time, asymptotically align with the tangent directions of the talwegs, see Section 3 and Section 4. Alignment occurs exponentially in the first spectral gap and the first eigenvalue. Actually the phenomenon is not only generic as it happens for all initial conditions, and at many scales with alignment to other gradient extremals with convergence rates governed by spectral gaps, see Theorem 3.3 and Theorem 4.6. We thus, partially, recover the behavior observed in the quadratic case with a convergence structured by eigenspaces. Let us observe, however, that the nonlinear case differs from the linearized one, as nonlinearities may disrupt the dominant effect of the first spectral gap when the conditioning along the talweg is too poor, i.e., when the smallest eigenvalue at a local minimizer is too small compared to the others; see Remark 3.4 and Remark 4.7. Finally, we establish volume concentration phenomena within the valley: for sets of initial conditions of positive measure, discrete and continuous flow concentrates along talwegs as time grows, see Theorems 3.6 and 4.9.

The tools used in this work rely on optimization techniques together with classical results from the theory of dynamical systems around hyperbolic points, in particular in the case where the real eigenvalues have the same sign, see [14, 25]. To facilitate the reading easy or routine proofs are postponed to the Appendix.

2 Gradient extremals, talweg, and valleys

2.1 Preliminaries

We assume $d \geq 2$, \mathbb{R}^d is endowed with the Euclidean scalar product $\langle \cdot, \cdot \rangle$ and associated norm $|\cdot|$. For a nonempty subset $F \subset \mathbb{R}^d$, we write $\text{dist}(x, F) = \inf\{|x - y| : y \in F\}$.

Given $f: \mathbb{R}^d \rightarrow \mathbb{R}$, for every $r \in \mathbb{R}$, we denote by $[f = r]$, $[f \leq r]$, $[f < r]$ level and sublevel sets. When some point a belongs to $[f < r]$, we denote by $[f < r]_a$ the connected component of $[f < r]$ containing a . If f is C^2 and the Hessian at x^* is positive definite, the family $[f < r]_{x^*}$ forms a family of connected open neighborhoods of x^* for $r > r^*$ with $r^* := f(x^*)$. For this, one can, for instance, observe that

$$f(x) \geq f(x^*) + c_f |x - x^*|^2, \text{ with } c_f > 0, \quad (2.1)$$

on a nonempty open ball centered at x^* . For $r > r^*$, one defines similarly $[f \leq r]_{x^*}$ and set $[f = r]_{x^*} := \{x \in [f \leq r]_{x^*} : f(x) = r\}$.

Let us consider the following generic situation [12]:

Assumption 2.1 (Blanket assumption). $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is of class C^3 and $x^* \in \mathbb{R}^d$ is a nondegenerate critical point with simple spectrum, i.e.,

$$\nabla f(x^*) = 0 \text{ and } \nabla^2 f(x^*) \text{ has pairwise distinct nonzero eigenvalues,} \quad (2.2)$$

where ∇f and $\nabla^2 f$ denote the gradient and the Hessian of f , respectively.

When Assumption 2.1 holds, we adopt some notations and convention (on $\eta, \lambda_i, v_i, i = 1, \dots, d$) that we now proceed to describe. Locally the spectrum of $\nabla^2 f$ varies smoothly near x^* (see Corollary A.5 in the Appendix): there indeed exist $\eta > 0$ and C^1 maps $\lambda_i: B(x^*, \eta) \rightarrow \mathbb{R}$ and $v_i: B(x^*, \eta) \rightarrow \mathbb{R}^d$ ($i = 1, \dots, d$) such that, for every $x \in B(x^*, \eta)$, $\lambda_i(x)$ is a simple eigenvalue of $\nabla^2 f(x)$ with associated eigenvector $v_i(x)$ and the family $\{v_i\}_{i=1}^d$ is orthonormal. We denote by $P(x)$ the orthogonal matrix whose i^{th} column is $v_i(x)$, and by $\mathcal{D}(x)$ the diagonal matrix whose i^{th} diagonal entry is $\lambda_i(x)$. After relabeling, if necessary, we may assume

$$\lambda_1(x) < \lambda_2(x) < \dots < \lambda_d(x), \quad \text{for all } x \text{ in } B(x^*, \eta), \quad (2.3)$$

and thus, by the inverse function theorem, we may further assume that

$$\nabla f(x) \neq 0, \quad \text{for all } x \in B(x^*, \eta) \setminus \{x^*\}. \quad (2.4)$$

We shall also consider the assumption

Assumption 2.2 (Strong local minimizer). For x^* as in Assumption 2.1, we have $\lambda_1 > 0$ throughout $B(x^*, \eta)$.

Under Assumption 2.2, x^* is a strict local minimum satisfying (2.1) and there exists $\bar{r} \in (r^*, \infty)$ such that $[f \leq \bar{r}]_{x^*} \subset B(x^*, \eta)$.

2.2 Gradient extremals

For a C^1 function, *gradient extremals* are rectifiable curves $\gamma(r) \in E(r)$ selected in the multivalued mapping:

$$r \rightrightarrows E(r) := \{x \in [f = r] : \nabla f(x) \neq 0 \text{ and } x \text{ is a critical point of } |\nabla f| \text{ over } [f = r]\}$$

where r ranges over the values of f . But under Assumption 2.1, which is a local assumption, and using Lagrange multipliers conditions, it is more convenient to adopt here the following:

Definition 2.3 (Gradient extremals). Under Assumption 2.1, for $i = 1, \dots, d$, the i^{th} gradient extremal is the set

$$\text{Ext}_i = \{x \in B(x^*, \eta) : \nabla^2 f(x) \nabla f(x) = \lambda_i(x) \nabla f(x)\}.$$

Theorem 2.4 (Gradient extremals are one dimensional smooth manifolds). Under Assumption 2.1, for all $i = 1, \dots, d$, the gradient extremal Ext_i is a one-dimensional C^1 embedded submanifold passing through x^* , its tangent space at x^* is given by $T_{x^*} \text{Ext}_i := \mathbb{R}v_i(x^*)$.

Proof. Defining $B(x^*, \eta) \ni x \mapsto z(x) = P^\top(x) \nabla f(x) \in \mathbb{R}^d$, Definition 2.3 implies that, for every $i = 1, \dots, d$ and $x \in B(x^*, \eta)$,

$$x \in \text{Ext}_i \Leftrightarrow \mathcal{D}(x)z(x) = \lambda_i(x)z(x) \Leftrightarrow z_j(x) = 0 \text{ for all } j \neq i. \quad (2.5)$$

Notice that z is of class C^1 with

$$Dz(x^*) = P^\top(x^*) \nabla^2 f(x^*) + DP^\top(x^*) \nabla f(x^*) = P^\top(x^*) \nabla^2 f(x^*) \quad (2.6)$$

invertible. Therefore, z is a local diffeomorphism. Let us show the result for $i = 1$: let π be the canonical projection on the last $d - 1$ coordinates. Define $S: B(x^*, \eta) \ni x \mapsto \pi(z(x)) \in \mathbb{R}^{d-1}$. By (2.5), we have that $\text{Ext}_1 = S^{-1}(\{0\}) \ni x^*$. Moreover, S is a submersion. Indeed, $DS(x) = \pi(Dz(x))$ and z is a local diffeomorphism. As a conclusion, Ext_1 is a one-dimensional C^1 embedded submanifold with $T_{x^*} \text{Ext}_1 = \ker DS(x^*)$. Since, by (2.6),

$$DS(x^*) = \pi \left(P^\top(x^*) \nabla^2 f(x^*) \right) = \pi \left(\mathcal{D}(x^*) P^\top(x^*) \right),$$

the result follows from $\ker DS(x^*) = \text{span}\{v_2(x^*), \dots, v_d(x^*)\}^\perp = \mathbb{R}v_1(x^*)$. \square

2.3 Talweg and crest extremal

Let us recall the notion of a *talweg* around x^* ,¹ see, e.g., [7, 8, 4] and references therein.

Definition 2.5 (Talweg: “path in the valley”²). Under Assumptions 2.1 and 2.2, the *talweg* of f around x^* is the set-valued map $\text{Tal}: [r^*, \bar{r}] \rightrightarrows [f < \bar{r}]_{x^*}$ defined through:

$$\theta(r) \in \text{Tal}(r) \iff \theta(r) \text{ is a local minimizer of } |\nabla f| \text{ over } [f = r]_{x^*}. \quad (2.7)$$

Since $[f = r]_{x^*}$ is compact, we have $\text{Tal}(r) \neq \emptyset$.

Theorem 2.6 (The talweg is the gradient extremal of minimal eigenvalue). Under Assumptions 2.1 and 2.2,

$$\text{Tal}([r^*, \bar{r}]) = \text{Ext}_1 \cap [f < \bar{r}]_{x^*}, \quad (2.8)$$

for \bar{r} close enough to r^* . In particular, $\text{Tal}([r^*, \bar{r}])$ is a one-dimensional C^1 embedded submanifold containing x^* and

$$T_{x^*} [\text{Tal}([r^*, \bar{r}])] = \mathbb{R}v_1(x^*). \quad (2.9)$$

¹We restrict to the case when x^* is a local minimizer.

²English translation of the German word borrowed from topography: Talweg or Thalweg.

Proof. Let $x \in \text{Tal}([r^*, \bar{r}])$. If $f(x) = r^*$, then $x = x^*$, whence $x \in \text{Ext}_1 \cap [f < r_\eta]$. Thus, let us assume that $r := f(x) \in (r^*, \bar{r})$. The Lagrangian $L_r : [f < \bar{r}]_{x^*} \times \mathbb{R} \rightarrow \mathbb{R}$ associated to the optimization problem in (2.7) is given by

$$L_r(y, \lambda) = \frac{1}{2} |\nabla f(y)|^2 + \lambda(r - f(y))$$

for all $y \in [f < \bar{r}]_{x^*}$, $\lambda \in \mathbb{R}$. Since $\nabla f(x) \neq 0$, there exists a unique $\bar{\lambda}(x) \in \mathbb{R}$ such that $\nabla_x L_r(x, \bar{\lambda}(x)) = 0$, or, equivalently,

$$\nabla^2 f(x) \nabla f(x) = \bar{\lambda}(x) \nabla f(x). \quad (2.10)$$

Moreover, the following second-order necessary condition holds (see, e.g., [10, Chapter 2, Theorem 2]):

$$\langle \nabla_{xx}^2 L_r(x, \bar{\lambda}(x)) h, h \rangle \geq 0, \quad \text{for all } h \in \mathbb{R}^d \text{ such that } \langle \nabla f(x), h \rangle = 0, \quad (2.11)$$

which can be rewritten as

$$\langle (D^3 f(x) \nabla f(x) + \nabla^2 f(x)^2 - \bar{\lambda}(x) \nabla^2 f(x)) h, h \rangle \geq 0, \quad (2.12)$$

for all $h \in \mathbb{R}^d$ such that $\langle \nabla f(x), h \rangle = 0$. If $\bar{\lambda}(x) \neq \lambda_1(x)$, we can take $h = v_1(x)$ in (2.12). Indeed v_1 has to be orthogonal to $\nabla f(x)$ which is also an eigenvector. This yields

$$\langle D^3 f(x) \nabla f(x) v_1(x), v_1(x) \rangle + \lambda_1(x) (\lambda_1(x) - \bar{\lambda}(x)) \geq 0. \quad (2.13)$$

Since the first term tends to zero when x approaches x^* and $\bar{\lambda}(x) \in \{\lambda_2(x), \dots, \lambda_d(x)\}$, the continuity of the spectrum of $\nabla^2 f$ at x^* shows that (2.13) can not hold for $|x - x^*|$ small enough. Thus, reducing $\bar{r} - r^*$, if necessary, we deduce that $\bar{\lambda}(x) = \lambda_1(x)$ and, hence, $x \in \text{Ext}_1 \cap [f < \bar{r}]_{x^*}$.

Conversely, let $x \in \text{Ext}_1 \cap [f < \bar{r}]_{x^*}$ and $h \in \mathbb{R}^d$ be such that $\langle \nabla f(x), h \rangle = 0$. Since $\nabla f(x)$ is in $\mathbb{R}v_1(x)$, there exist $\mu_2, \dots, \mu_d \in \mathbb{R}$ such that $h = \sum_{i=2}^d \mu_i v_i(x)$. Thus,

$$\begin{aligned} & \langle (\nabla^2 f(x)^2 - \lambda_1(x) \nabla^2 f(x)) h, h \rangle \\ &= \sum_{i=2}^d (\lambda_i^2(x) - \lambda_1(x) \lambda_i(x)) \mu_i^2 \\ &\geq \lambda_2(x) (\lambda_2(x) - \lambda_1(x)) \sum_{i=2}^d \mu_i^2 \\ &= \lambda_2(x) (\lambda_2(x) - \lambda_1(x)) |h|^2. \end{aligned} \quad (2.14)$$

On the other hand, since $D^3 f$, ∇f , and $\lambda_2 - \lambda_1$ are continuous and $\nabla f(x^*) = 0$, reducing $\bar{r} - r^*$, if necessary, there exists $c > 0$ such that

$$\lambda_2(x) (\lambda_2(x) - \lambda_1(x)) - \|D^3 f(x)\| |\nabla f(x)| > c.$$

Then we deduce from (2.14) that

$$\left\langle \left(D^3 f(x) \nabla f(x) + (\nabla^2 f(x))^2 - \lambda_1(x) \nabla^2 f(x) \right) h, h \right\rangle \geq c |h|^2. \quad (2.15)$$

Thus, setting $r = f(x)$, we get

$$\nabla_x L_r(x, \lambda_1(x)) = 0 \quad \text{and} \quad \langle \nabla_{xx}^2 L_r(x, \lambda_1(x)) h, h \rangle > 0, \quad (2.16)$$

for all $h \in \mathbb{R}^d \setminus \{0\}$ such that $\langle \nabla f(x), h \rangle = 0$, which by [10, Theorem 4] implies that $x \in \text{Tal}([r^*, \bar{r}])$. Altogether, we have established (2.8). Finally, since $[f < \bar{r}]_{x^*}$ is an open subset of \mathbb{R}^d , we deduce (2.9) from (2.8) and Theorem 2.4. \square

Remark 2.7. (a) (Talweg structure) As shown in Proposition A.1 (see Appendix), if $r > r^*$ is close enough to r^* , $\text{Tal}(r)$ has exactly two points, while $\text{Tal}(r^*) = \{x^*\}$.
(b) (Crest extremal) One could also define a notion of a *crest extremal* $\text{Cre}: [r^*, \bar{r}] \rightrightarrows \mathbb{R}^d$ through

$$\gamma(r) \in \text{Cre}(r) \iff \gamma(r) \text{ is a local maximizer of } |\nabla f| \text{ over } [f = r]_{x^*}. \quad (2.17)$$

Using similar arguments one can prove that the crest extremal is associated to the gradient extremal of maximal eigenvalue Ext_d .

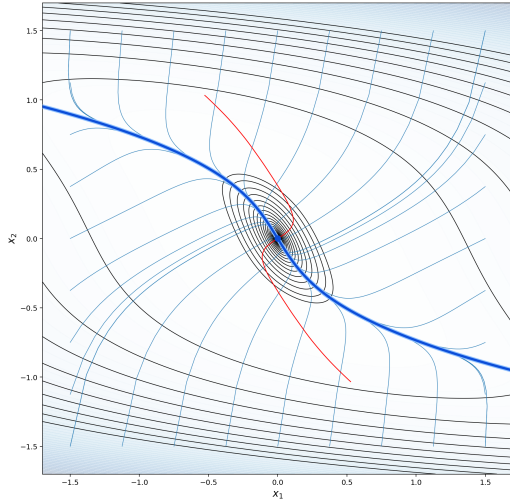


Figure 1: Black level lines, blue talweg, and red crest extremal for $f(x_1, x_2) = x_1^2 + (x_1 + x_2 + x_2^3)^2$. Gradient curves, in light blue, ultimately align with the talweg, see Theorem 3.2 and Theorem 3.3.

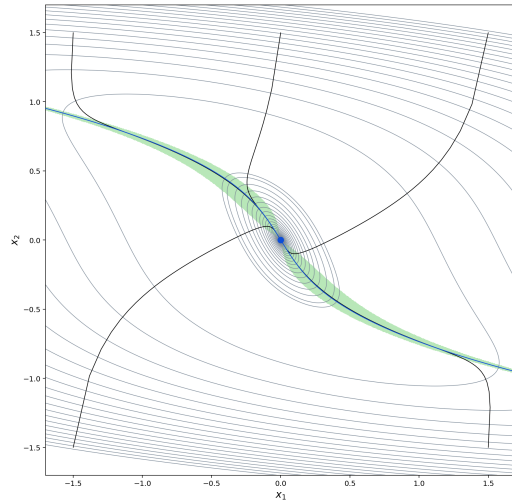


Figure 2: The valley has a simple infinitesimal structure near a local minimizer as it has the structure of a second-order cone, see Proposition 2.8. Farther away, it may become more complex.

2.4 Valleys and their variational structure

Under Assumptions 2.1 and 2.2, and given $w \geq 0$, define the *valley of width w* around x^* by

$$\text{Val}_{f,w}(x^*) = \{x \in [f < \bar{r}]_{x^*} : |\nabla^2 f(x)\nabla f(x) - \lambda_1(x)\nabla f(x)| \leq w|\nabla f(x)|\}. \quad (2.18)$$

The “width measure” w evaluates the relative error of being an eigenvector for the Hessian and quantifies a form of enlargement of the talweg. Other measures could be proposed as in [4], but as we will see, this one has interesting properties.

By Theorem 2.6, we have $\text{Tal}([r^*, \bar{r}]) \subseteq \text{Val}_{f,w}(x^*)$ for all $w \geq 0$, with equality when $w = 0$. Note also that when the width is such that $w \geq \lambda_d(x) - \lambda_1(x)$ for all x in $[f < \bar{r}]_{x^*}$, a simple computation shows that the valley is so wide that it coincides with the whole neighborhood $[f < \bar{r}]_{x^*}$.

Consider the Taylor expansion at order two of $f - f(x^*)$ at x^* , i.e. $q(y) = f(x^*) + \frac{1}{2}\langle \nabla^2 f(x^*)y, y \rangle$, with y in \mathbb{R}^d , and introduce its valley (at 0)

$$\text{Val}_{q,w} = \{x \in \mathbb{R}^d : |(\nabla^2 f(x^*)^2 - \lambda_1(x^*)\nabla^2 f(x^*))x| \leq w|\nabla^2 f(x^*)x|\}.$$

One easily sees that $\text{Val}_{q,w}$ is a non-empty closed cone whose generator is the eigenspace $\mathbb{R}v_1(x^*)$. For every $0 < w_1 < w_2$, one has $\text{Val}_{q,w_1} \subset \text{Val}_{q,w_2}$.

Proposition 2.8 (Valley and conic approximation). *Under Assumptions 2.1, 2.2, and given $w > w' > 0$, the following hold:*

- (i) *Diminishing $\bar{r} > r^*$ if necessary, we have $(x^* + \text{Val}_{q,w'}) \cap [f < \bar{r}]_{x^*} \subset \text{Val}_{f,w}(x^*)$.*
- (ii) *Diminishing $\bar{r} > r^*$ if necessary, we have $\text{Val}_{f,w'} \subset (x^* + \text{Val}_{q,w}) \cap [f < \bar{r}]_{x^*}$.*

Proof. The proof of (ii) being analogous to the proof of (i), we only show the first assertion. Since f is C^3 , $\nabla f(x^*) = 0$, and $\nabla^2 f(x^*)$ are invertible, by Taylor expansions we have the existence of c_1, c_2 in $(0, \infty)$ such that, if $\delta > 0$ is small enough and x in $B(x^*, \delta)$, $|x - x^*| \leq c_1 |\nabla f(x)|$ and $|\nabla f(x) - \nabla^2 f(x^*)(x - x^*)| \leq c_2 |x - x^*|^2$. Thus, setting $c = c_1 c_2 > 0$, we have

$$|\nabla f(x) - \nabla^2 f(x^*)(x - x^*)| \leq c|x - x^*| |\nabla f(x)|, \quad \text{for all } x \in B(x^*, \delta).$$

The continuity of $\nabla^2 f$ and λ_1 implies then the existence of an increasing and continuous function $m: \mathbb{R}_+ \rightarrow \mathbb{R}_+$, with $m(0) = 0$, such that for any $x \in B(x^*, \delta)$,

$$\begin{aligned} & \left| |\nabla^2 f(x) \nabla f(x) - \lambda_1(x) \nabla f(x)| - |(\nabla^2 f(x^*) - \lambda_1(x^*) I_d) \nabla^2 f(x^*)(x - x^*)| \right| \\ & \leq m(|x - x^*|) |\nabla f(x)|. \end{aligned} \quad (2.19)$$

Let $0 < w' < w$ and $0 < \delta < \eta$ be given and let $x \in x^* + \text{Val}_{q,w'} \cap B(0, \delta)$, i.e., $|x - x^*| < \delta$ and

$$|(\nabla^2 f(x^*) - \lambda_1(x^*) I_d) \nabla^2 f(x^*)(x - x^*)| \leq w' |\nabla^2 f(x^*)(x - x^*)|. \quad (2.20)$$

By reducing δ , if necessary, we can assume that $m(\delta) \leq w - w'$. Thus, (2.19) yields

$$|\nabla^2 f(x) \nabla f(x) - \lambda_1(x) \nabla f(x)| \leq (w' + m(|x - x^*|)) |\nabla f(x)| \leq w |\nabla f(x)|,$$

i.e., $x \in \text{Val}_{f,w}$. Diminishing \bar{r} so that $[f < \bar{r}]_{x^*} \subset B(x^*, \delta)$ gives (i). \square

Recall that given a non-empty closed set $A \subseteq \mathbb{R}^d$, the (Bouligand) tangent cone $\mathcal{T}_A(x)$ to A at $x \in A$ is defined by

$$\begin{aligned} \mathcal{T}_A(x) = \{h \in \mathbb{R}^d : \exists (h_n)_{n \in \mathbb{N}} \subset \mathbb{R}^d, (\tau_n)_{n \in \mathbb{N}} \subset (0, \infty), \\ \tau_n \rightarrow 0, h_n \rightarrow h, \text{ and } x + \tau_n h_n \in A, \text{ for } n \in \mathbb{N}\}. \end{aligned}$$

Notice that, for every $w > 0$, $\mathcal{T}_{\text{Val}_q^w}(0) = \text{Val}_{q,w}$.

Proposition 2.9 (Tangent cone to the valley). *Under Assumptions 2.1, 2.2, and given $w > 0$, one has*

$$\mathcal{T}_{\text{Val}_{f,w}(x^*)}(x^*) = \text{Val}_{q,w}.$$

Proof. Fix $w' > w$ and let h in $\mathcal{T}_{\text{Val}_{f,w}(x^*)}(x^*)$, $(h_n)_{n \in \mathbb{N}} \subset \mathbb{R}^d$ and $(\tau_n)_{n \in \mathbb{N}}$ a positive sequence such that $x^* + \tau_n h_n \in \text{Val}_{f,w}(x^*)$, $h_n \rightarrow h$ and $\tau_n \rightarrow 0$. Since $\tau_n h_n \rightarrow 0$, Proposition 2.8(ii) implies the existence of $\delta \in (0, \eta)$ such that, for $n \in \mathbb{N}$ large enough, $\tau_n h_n \in \text{Val}_{q,w'}$ and, hence, $h_n \in \text{Val}_{q,w'}$. Since $\text{Val}_{q,w'}$ is closed, we get that $h \in \text{Val}_{q,w'}$ and, since $w' > w$ is arbitrary, we obtain that $h \in \text{Val}_{q,w}$. Now, let $h \in \text{Val}_{q,w}$ and take two sequences $(w_n)_{n \in \mathbb{N}} \subset [0, w)$ and $(h_n)_{n \in \mathbb{N}} \subset \mathbb{R}^d$ such that, as $n \rightarrow \infty$, $w_n \uparrow w$, $h_n \rightarrow h$ and $h_n \in \text{Val}_{q,w_n}$, for all $n \in \mathbb{N}$. Proposition 2.8(i) then implies the existence of $(\delta_n)_{n \in \mathbb{N}} \subset (0, \eta)$ such that $x^* + \text{Val}_{q,w_n} \cap B(0, \delta_n) \subset \text{Val}_{f,w}(x^*)$. Let $(t_n)_{n \in \mathbb{N}} \subset (0, \infty)$ be such that $t_n \downarrow 0$ and $t_n h_n \in \text{Val}_{q,w_n} \cap B(0, \delta_n)$, for all $n \in \mathbb{N}$, then $x^* + t_n h_n \in \text{Val}_{f,w}(x^*)$ and hence $h \in \mathcal{T}_{\text{Val}_{f,w}(x^*)}(x^*)$. \square

3 Directional convergence of gradient trajectories

3.1 Preliminaries

Radial projection We make repeated use of *the radial projection*

$$\rho : \mathbb{R}^d \setminus \{0\} \longrightarrow S(0,1), \quad \rho(x) = \frac{x}{|x|}.$$

where $S(0,1)$ is the unit sphere. For later use, let us record the following elementary properties of ρ . First, we have $\rho(\ell x) = \rho(x)$ for all $\ell > 0$ and $x \neq 0$. Moreover:

$$|\rho(a+b) - \rho(a)| \leq 2 \frac{|b|}{|a|}, \quad \text{for all } a, b \in \mathbb{R}^d, a \neq 0, a+b \neq 0. \quad (3.1)$$

Observe indeed $\frac{a+b}{|a+b|} - \frac{a}{|a|} = \frac{a+b}{|a+b|} - \frac{a+b}{|a|} + \frac{b}{|a|} = (a+b) \left(\frac{1}{|a+b|} - \frac{1}{|a|} \right) + \frac{b}{|a|}$, and thus,

$$\left| \frac{a+b}{|a+b|} - \frac{a}{|a|} \right| \leq |a+b| \frac{||a| - |a+b||}{|a+b||a|} + \frac{|b|}{|a|} \leq 2 \frac{|b|}{|a|}.$$

Gradient trajectories In view of our regularity assumptions one considers the *gradient flow* $\Phi_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $t \geq 0$, of f , which is defined through

$$\begin{aligned} \partial_t \Phi_t(x) &= -\nabla f(\Phi_t(x)) \quad \text{for } t \in \mathbb{R}, \\ \Phi_0(x) &= x, \quad \text{for } x \in \mathbb{R}^d. \end{aligned}$$

If f is coercive, and generic in the sense of C^2 topology [12, Theorem 6.2], i.e. f is a Morse function, then any critical point x^* of f satisfies Assumption 2.1. In that case gradient curves generically converge with respect to initial conditions towards local minimizers, see Theorem 3.3 (i). The latter seems to be due, in essence, to [26]; it was rediscovered several times, e.g., by [22, 13, 18] under different forms and settings.

3.2 Directional convergence: the generic case

In this section we focus our attention to local minimizers.

Based on some fundamental results by Hartman (see [14] and Chapter 9 in [15]), we establish in Theorem 3.2 below that, generically in terms of the initial condition x near x^* , the asymptotic direction of the trajectory $(\Phi_t(x))_{t \geq 0}$ belongs to the eigenspace associated with the smallest eigenvalue of $\nabla^2 f(x^*)$. Theorem 3.3 shows that this phenomenon is actually generic over \mathbb{R}^d under coercivity assumptions, and that for sufficiently smooth functions generic convergence rates may be provided.

As in Section 2, consider a local parametrization $\{(\lambda_i(x), v_i(x)) : i = 1, \dots, d\}$ of eigenvalues and eigenvectors of $\nabla^2 f(x)$, for x near x^* , satisfying Assumption 2.2. Let us introduce a useful partition of \mathbb{R}^d . Let $\mathcal{S}_0 = \{0\}$ and, for every $1 \leq i \leq d$, set

$$\mathcal{S}_i = \left\{ \sum_{j=d-i+1}^d \alpha_j v_j(x^*) : \alpha_j \in \mathbb{R} \text{ for all } j = d-i+1, \dots, d \text{ with } \alpha_{d-i+1} \neq 0 \right\}. \quad (3.2)$$

Note that each \mathcal{S}_i is a manifold of dimension i and $\{\mathcal{S}_i\}_{i=0}^d$ forms a partition of \mathbb{R}^d . Moreover, \mathcal{S}_d is dense and the radial projection ρ maps \mathcal{S}_d onto a dense open subset of the unit sphere.

For (t, y) in $[0, \infty) \times \mathbb{R}^d$ set

$$\Psi_t(y) = \exp(-\nabla^2 f(x^*)t) \quad (3.3)$$

the flow associated with the linearized system $y \mapsto -\nabla^2 f(x^*)y$ on \mathbb{R}^d . Note that for every $i = 1, \dots, d$ and $y = \sum_{j=d-i+1}^d \alpha_j v_j(x^*) \in \mathcal{S}_i$ we have

$$\forall t \in [0, \infty), \quad \Psi_t(y) = \sum_{j=d-i+1}^d \alpha_j \exp(-\lambda_j(x^*)t) v_j(x^*), \quad (3.4)$$

and hence, by (2.3),

$$\forall t \in [0, \infty), \quad |\Psi_t(y)| \leq |y| \exp(-\lambda_{d-i+1}(x^*)t). \quad (3.5)$$

The following result describes the asymptotic directional behavior of the flow Ψ .

Lemma 3.1 (Directional convergence quadratic case). *Let $i = 2, \dots, d$, $y = \sum_{j=d-i+1}^d \alpha_j v_j(x^*) \in \mathcal{S}_i$. Then, for every $t \in (0, \infty)$, we have*

$$|\rho(\Psi_t(y)) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \leq \frac{2|y|}{|\alpha_{d-i+1}|} \exp(-(\lambda_{d-i+2}(x^*) - \lambda_{d-i+1}(x^*))t), \quad (3.6)$$

$$\begin{aligned} & |\rho(\partial_t \Psi_t(y)) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \\ & \leq \frac{2|y|}{|\alpha_{d-i+1}|} \left(\frac{\lambda_d(x^*)}{\lambda_{d-i+1}(x^*)} \right) \exp(-(\lambda_{d-i+2}(x^*) - \lambda_{d-i+1}(x^*))t). \end{aligned} \quad (3.7)$$

Proof. We only prove (3.6) as the proof of (3.7) is analogous. Since $\exp(\lambda_{d-i+1}(x^*)t) > 0$, we have $\rho(\Psi_t(y)) = \rho\left(\frac{1}{|\alpha_{d-i+1}|} \exp(\lambda_{d-i+1}(x^*)t) \Psi_t(y)\right)$. It follows from (3.1) and (3.4) that

$$\begin{aligned} & |\rho(\Psi_t(y)) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \\ & = \left| \rho\left(\frac{1}{|\alpha_{d-i+1}|} \exp(\lambda_{d-i+1}(x^*)t) \Psi_t(y)\right) - \rho(\text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)) \right| \\ & \leq \frac{2}{|\alpha_{d-i+1}|} \left| \sum_{j=d-i+2}^d \alpha_j \exp(-(\lambda_j(x^*) - \lambda_{d-i+1}(x^*))t) v_j(x^*) \right| \\ & \leq \frac{2|y|}{|\alpha_{d-i+1}|} \exp(-(\lambda_{d-i+2}(x^*) - \lambda_{d-i+1}(x^*))t). \end{aligned}$$

Hence the result. \square

The next result shows, in particular, that, as $t \rightarrow \infty$, the directions of the trajectories $(\Phi_t(x))_{t \geq 0}$ generically align, in terms of the initial condition $x \in B(x^*, \eta)$, with the tangent space to the talweg at x^* .

Theorem 3.2 (Directional convergence of the gradient flow). *Under Assumptions 2.1 and 2.2, the following results hold:*

(i) (C^1 conjugacy) *For $\eta > 0$ small enough there exists a local C^1 -diffeomorphism $H: B(x^*, \eta) \rightarrow \mathbb{R}^d$ satisfying $H(x^*) = 0$ and $DH(x^*) = I_d$, such that*

$$H(\Phi_t(x)) = \Psi_t(H(x)), \quad \text{for all } t \in [0, \infty) \text{ and } x \in B(x^*, \eta). \quad (3.8)$$

(ii) (Stable stratification) Setting $\mathcal{S}_i^H(x^*) := H^{-1}(\mathcal{S}_i)$ for all $i = 0, \dots, d$, $\mathcal{S}_i^H(x^*)$ is a manifold of dimension i , $\{\mathcal{S}_i^H(x^*)\}_{i=0}^d$ forms a partition of $B(x^*, \eta)$ such that $\Phi_t(\mathcal{S}_i^H(x^*)) \subset \mathcal{S}_i^H(x^*)$ for all $t \geq 0$ ³.

(iii) (Directional convergence) For all $t \in [0, \infty)$, and, for every $i = 1, \dots, d$ and $x \in \mathcal{S}_i^H$, the following hold:

$$|\Phi_t(x) - x^*| = O(\exp(-\lambda_{d-i+1}(x^*)t)), \quad (3.9)$$

$$\lim_{t \rightarrow \infty} \rho(\Phi_t(x) - x^*) = - \lim_{t \rightarrow \infty} \rho(\partial_t \Phi_t(x)) \in \{-v_{d-i+1}(x^*), v_{d-i+1}(x^*)\}. \quad (3.10)$$

Proof. Let $\eta > 0$ be as in Lemma A.3. It follows from [14, Theorem (I)] that, reducing η if necessary, there exists $h: B(x^*, \eta) \rightarrow \mathbb{R}^d$ of class C^1 such that $h(x^*) = 0$, $Dh(x^*) = I_d$, and $h(\Phi_1(x)) = \Psi_1(h(x))$ for all $x \in B(x^*, \eta)$. Set $A = -\nabla^2 f(x^*)$ and define $H: B(x^*, \eta) \rightarrow \mathbb{R}^d$ by

$$H(x) = \int_0^1 e^{-As} h(\Phi_s(x)) ds, \quad \text{for all } x \in B(x^*, \eta).$$

From classical results H is C^1 . A change of variable as in the proof of [25, Lemma 4] (see also [27, Section 9.3.3]), shows that H satisfies (3.8). Moreover, since $\Phi_s(x^*) = x^*$ standard computations show that $D_x \Phi_s(x^*) = e^{As}$ for all $s \in [0, \infty)$. On the other hand, $H(x^*) = 0$ and, since $Dh(x^*) = I_d$, we have

$$DH(x^*) = \int_0^1 e^{-As} Dh(\Phi_s(x^*)) D_x \Phi_s(x^*) ds = I_d.$$

In particular, by the inverse function theorem, H is a local C^1 -diffeomorphism. This proves (i).

Let us prove (ii) and (iii). Note that (3.4) implies that, for every $i = 1, \dots, d$ and $t \in [0, \infty)$, $\Psi_t(\mathcal{S}_i) \subset \mathcal{S}_i$, which, by (3.8), implies that $\Phi_t(\mathcal{S}_i^H) \subset \mathcal{S}_i^H$ for all $t \in [0, \infty)$. Moreover, for every $x \in B(x^*, \eta)$ and $t \in [0, \infty)$, we have $\Phi_t(x) - x^* = H^{-1}(\Psi_t(H(x))) - H^{-1}(0)$ and, hence, (3.9) follows from the boundedness of DH^{-1} and (3.5).

Let $x \in \mathcal{S}_i^H$ with $i \geq 1$. Note that since $i \neq 0$, we have $\Phi_t(x) - x^* \neq 0$ and $\partial_t \Phi_t(x) \neq 0$ for all $t \geq 0$ so that both directions can be evaluated. For every $j = 1, \dots, d$, set $\alpha_j = \langle H(x), v_j(x^*) \rangle$. For every $t \in (0, \infty)$ we have

$$\begin{aligned} & |\rho(\Phi_t(x) - x^*) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \\ & \leq |\rho(\Phi_t(x) - x^*) - \rho(\Psi_t(H(x)))| + |\rho(\Psi_t(H(x))) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)|. \end{aligned} \quad (3.11)$$

By (3.4) and the orthogonality of the family $\{v_i(x^*)\}_{i=1}^d$ we have

$$|\exp(\lambda_{d-i+1}(x^*)t)\Psi_t(H(x))| \geq |\alpha_{d-i+1}| > 0.$$

Thus, (3.1) yields

$$\begin{aligned} & |\rho(\Phi_t(x) - x^*) - \rho(\Psi_t(H(x)))| \\ & = |\rho(\exp(\lambda_{d-i+1}(x^*)t)(\Phi_t(x) - x^*)) - \rho(\exp(\lambda_{d-i+1}(x^*)t)\Psi_t(H(x)))| \\ & \leq \frac{2 \exp(\lambda_{d-i+1}(x^*)t)}{|\alpha_{d-i+1}|} |\Phi_t(x) - x^* - \Psi_t(H(x))|. \end{aligned} \quad (3.12)$$

On the other hand, it follows from (3.8) and $DH^{-1}(0) = I_d$ that

$$\Phi_t(x) - x^* - \Psi_t(H(x)) = \left(\int_0^1 (DH^{-1}(\tau\Psi_t(H(x))) - DH^{-1}(0)) d\tau \right) \Psi_t(H(x)), \quad (3.13)$$

³When no confusion can occur, we shall write $\mathcal{S}_i^H = \mathcal{S}_i^H(x^*)$

which, together with (3.12), (3.5), and the continuity of H^{-1} , implies that

$$\begin{aligned} & |\rho(\Phi_t(x) - x^*) - \rho(\Psi_t(H(x)))| \\ & \leq \frac{2|H(x)|}{|\alpha_{d-i+1}|} \int_0^1 \|DH^{-1}(\tau\Psi_t(H(x))) - DH^{-1}(0)\| d\tau \xrightarrow{t \rightarrow \infty} 0. \end{aligned} \quad (3.14)$$

Therefore, we deduce from (3.11) and Lemma 3.1 that

$$|\rho(\Phi_t(x) - x^*) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \xrightarrow{t \rightarrow \infty} 0. \quad (3.15)$$

Similarly, we have

$$\begin{aligned} & |\rho(\partial_t\Phi_t(x)) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \\ & \leq |\rho(\partial_t\Phi_t(x)) - \rho(\partial_t\Psi_t(H(x)))| + |\rho(\partial_t\Psi_t(H(x))) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)|. \end{aligned} \quad (3.16)$$

By (3.4), we have

$$\forall t \in [0, \infty), \quad \partial_t\Psi_t(H(x)) = - \sum_{j=d-i+1}^d \alpha_j \lambda_j(x^*) \exp(-\lambda_j(x^*)t) v_j(x^*).$$

The orthogonality of the family $\{v_i(x^*)\}_{i=1}^d$ implies that

$$|\lambda_{d-i+1}(x^*)^{-1} \exp(\lambda_{d-i+1}(x^*)t) \Psi_t(H(x))| \geq |\alpha_{d-i+1}| > 0.$$

Thus, (3.1) yields

$$|\rho(\partial_t\Phi_t(x)) - \rho(\partial_t\Psi_t(H(x)))| \leq \frac{2 \exp(\lambda_{d-i+1}(x^*)t)}{\lambda_{d-i+1}(x^*)|\alpha_{d-i+1}|} |\partial_t\Phi_t(x) - \partial_t\Psi_t(H(x))|. \quad (3.17)$$

Since (3.8) implies $DH(\Phi_t(x))\partial_t\Phi_t(x) = \partial_t\Psi_t(H(x))$, we deduce that

$$|\partial_t\Phi_t(x) - \partial_t\Psi_t(H(x))| \leq |\partial_t\Psi_t(H(x))| \|DH^{-1}(\Psi_t(H(x))) - DH^{-1}(0)\|, \quad (3.18)$$

which, together with (3.16), (3.17), (3.5), the continuity of H^{-1} , and Lemma 3.1, implies that

$$|\rho(\partial_t\Phi_t(x)) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \xrightarrow{t \rightarrow \infty} 0. \quad (3.19)$$

Therefore, (3.10) follows from (3.15) and (3.19). \square

Theorem 3.3. *If Assumption 2.1 holds at each critical point then:*

(i) (Generic alignment) *If f is coercive, the set of initial conditions whose gradient flow trajectories converge to local minimizers of f with the velocities and the secant aligning to the talweg has full Lebesgue measure.*

(ii) (Alignment rate) *Assume that f is C^∞ ⁽⁴⁾, let x be an initial condition so that, $\Phi_t(x) \rightarrow x^*$ as $t \rightarrow \infty$ with $\Phi_t(x) \in \mathcal{S}_d^H(x^*)$ for some $t \geq 0$ ⁽⁵⁾, and suppose, in addition, the non resonance conditions at x^* : for all integers m_1, \dots, m_d such that $\sum_{j=2}^d m_j \geq 2$, we have⁶*

$$\forall i = 1, \dots, d, \quad \lambda_i(x^*) \neq \sum_{j=1}^d m_j \lambda_j(x^*).$$

⁴Actually we may assume that f is of class C^k with k sufficiently large, see [25] for more details.

⁵This corresponds to the generic case in (i).

⁶Observe this is a generic condition as the m_i are integers.

Then the trajectory aligns with the tangent to the talweg at a rate governed by:

$$\begin{aligned} \text{dist}\left(\rho(\partial_t \Phi_t(x)), T_{x^*} \text{Tal}([r^*, \bar{r}])\right) &= O(\exp(-(\lambda_2 - \lambda_1)t)) + O(\exp(-\lambda_1 t)), \\ \text{dist}\left(\rho(\Phi_t(x) - x^*), T_{x^*} \text{Tal}([r^*, \bar{r}])\right) &= O(\exp(-(\lambda_2 - \lambda_1)t)) + O(\exp(-\lambda_1 t)), \end{aligned}$$

where $\bar{r} > r^*$ is sufficiently close to r^* .

Proof. Item (i) extends [26] and also [22, 18].

In view of Assumption 2.1 all the critical points of f are isolated; hence the set of critical points is at most countable. Local uniqueness of critical points entail the well known fact:

Claim 1. By coercivity, f is inf-bounded so the flow Φ_t is defined for all $t \geq 0$, and moreover one can prove that for all x in \mathbb{R}^d , $\Phi_t(x)$ converges to some x^* belonging to the critical set $\text{crit} f = (\nabla f)^{-1}(\{0\})$.

Let us partition $\text{crit} f$ into the set of local minimizers $\text{argmin-loc} f$ and the set of unstable critical points $\text{crit}^u f$, i.e., those whose Hessian has a negative eigenvalue. Whence $\text{crit} f = \text{argmin-loc} f \cup \text{crit}^u f$.

Let us recall two facts:

- (a) For each $x^* \in \text{argmin-loc} f$, with the notation $\eta := \eta_{x^*}$ of Theorem 3.2, we have

$$B(x^*, \eta) = \bigcup_{i=1}^d \mathcal{S}_i^H(x^*).$$

- (b) For each $x^* \in \text{crit}^u f$, using Assumption 2.1, we may apply the Hadamard–Perron decomposition into stable and unstable manifolds, see [27]. Let $W_{x^*}^s$ denote the local stable manifold of x^* , which has dimension strictly less than d as x^* is not a local minimizer, and which is defined locally as

$$W_{x^*}^s = \{x \in B(x^*, \varepsilon_{x^*}) : \Phi_t(x) \rightarrow x^* \text{ as } t \rightarrow \infty\} \text{ for } \varepsilon_{x^*} > 0 \text{ small enough.}$$

We then define the following two subsets of \mathbb{R}^d

$$\mathcal{N} = \bigcup_{x^* \in \text{argmin-loc} f} \bigcup_{i=1}^{d-1} \bigcup_{t \in \mathbb{N}} \Phi_t^{-1}(\mathcal{S}_i^H(x^*)) \cup \bigcup_{x^* \in \text{crit}^u f} \bigcup_{t \in \mathbb{N}} \Phi_t^{-1}(W_{x^*}^s), \quad (3.20)$$

$$\mathcal{F} = \bigcup_{x^* \in \text{argmin-loc} f} \bigcup_{t \in \mathbb{N}} \Phi_t^{-1}(\mathcal{S}_d^H(x^*)). \quad (3.21)$$

Let us observe (and check) that we have the partition

$$\mathbb{R}^d = \mathcal{N} \cup \mathcal{F}. \quad (3.22)$$

By the fact that $\{\mathcal{S}_i^H\}_{i=1, \dots, d}$ is a partition, we have $\mathcal{N} \cap \mathcal{F} = \emptyset$. Let x in \mathbb{R}^d . We know by Claim 1 that $t \mapsto \Phi_t(x)$ converges to some critical point $x^* \in \text{crit} f$. Two cases may occur:

Case 1: $x^* \in \text{crit}^u f$. Since $\Phi_t(x) \rightarrow x^*$, by definition of the stable manifold of x^* , there exists $T > 0$ such that $\Phi_t(x)$ in $W_{x^*}^s$ for all $t \geq T$. Thus, for any integer $q \geq T$, $\Phi_q(x) \in W_{x^*}^s$, and therefore

$$x \in \Phi_q^{-1}(W_{x^*}^s) \subset \mathcal{N}.$$

Case 2: $x^* \in \operatorname{argmin}\text{-loc } f$. There exists $T > 0$ such that, for any $t \geq T$,

$$\Phi_t(x) \in B(x^*, \eta_{x^*}) = \bigcup_{i=1}^d \mathcal{S}_i^H(x^*).$$

Take an integer $q \geq T$. Since $\{\mathcal{S}_i^H(x^*)\}_{i=1}^d$ forms a partition, we have either $\Phi_q(x) \in \mathcal{S}_d^H(x^*)$ or $\Phi_q(x) \in \bigcup_{i=1}^{d-1} \mathcal{S}_i^H(x^*)$. Hence either $x \in \mathcal{F}$ or $x \in \mathcal{N}$. This proves (3.22).

Let us observe that \mathcal{N} has Lebesgue measure zero. This follows from the fact that each Φ_t is a diffeomorphism, and that the sets $\mathcal{S}_i^H(x^*)$ for $i = 1, \dots, d-1$ and $W_{x^*}^s$ are manifolds of dimension strictly less than d . Hence \mathcal{N} is a countable union of sets of Lebesgue measure zero, and therefore has Lebesgue measure zero.

To conclude we just need to observe that $x \in \mathcal{F}$ implies that x converges to some local minimizer x^* with in addition $\Phi_t(x) \in \mathcal{S}_d^H$, so that Theorem 3.2 applies and we have directional alignment of the secant and the velocity with the talweg.

Let us prove (ii). Recall that since x^* is a local minimizer the tangent space to the talweg is $\mathbb{R}v_1(x^*)$. Shifting time if necessary, we may assume $x \in \mathcal{S}_d^H$. Thus with the notation of Theorem 3.2, we have $\alpha_1 \neq 0$ in the expression $H(x) = \sum_1^d \alpha_i v_i(x^*)$.

By Sternberg's theorem [25, Theorem 2], under the non-resonance assumptions the conjugating diffeomorphism H may be assumed to be of class C^2 . In particular, shrinking neighborhoods if necessary, there exists $C > 0$ such that DH and DH^{-1} are locally C -Lipschitz continuous.

We estimate the alignment rate for the direction of the secant. From (3.11), we have

$$\begin{aligned} & \left| \rho(\Phi_t(x) - x^*) - \operatorname{sign}(\alpha_1)v_1(x^*) \right| \\ & \leq \left| \rho(\Phi_t(x) - x^*) - \rho(\Psi_t(H(x))) \right| + \left| \rho(\Psi_t(H(x))) - \operatorname{sign}(\alpha_1)v_1(x^*) \right|. \end{aligned} \quad (3.23)$$

Using (3.14) and (3.6), we obtain

$$\begin{aligned} & \left| \rho(\Phi_t(x) - x^*) - \operatorname{sign}(\alpha_1)v_1(x^*) \right| \\ & \leq \frac{2|H(x)|}{|\alpha_1|} \int_0^1 \|DH^{-1}(\tau\Psi_t(H(x))) - DH^{-1}(0)\| \, d\tau + \frac{2|H(x)|}{|\alpha_1|} \exp(-(\lambda_2 - \lambda_1)t). \end{aligned} \quad (3.24)$$

Since DH^{-1} is locally Lipschitz,

$$\int_0^1 \|DH^{-1}(\tau\Psi_t(H(x))) - DH^{-1}(0)\| \, d\tau \leq C|\Psi_t(H(x))|.$$

Thus, using $|\Psi_t(H(x))| = O(\exp(-\lambda_1 t))$, yields

$$\left| \rho(\Phi_t(x) - x^*) - \operatorname{sign}(\alpha_1)v_1(x^*) \right| \leq O(\exp(-\lambda_1 t)) + \frac{2|H(x)|}{|\alpha_1|} \exp(-(\lambda_2 - \lambda_1)t). \quad (3.25)$$

Since $\operatorname{sign}(\alpha_1)v_1(x^*) \in T_{x^*}\operatorname{Tal}([r^*, \bar{r}])$, we infer

$$\operatorname{dist}(\rho(\Phi_t(x) - x^*), T_{x^*}\operatorname{Tal}([r^*, \bar{r}])) = O(\exp(-(\lambda_2 - \lambda_1)t)) + O(\exp(-\lambda_1 t)). \quad (3.26)$$

The proof for the velocity alignment is similar as it also relies on the local Lipschitz continuity of DH^{-1} , on the inequalities in the proof of Theorem 3.2 and on the estimate (3.7). \square

Remark 3.4 (On the sharpness of the alignment rate). In the quadratic case the directional convergence is governed by the spectral gap $\lambda_2 - \lambda_1$. One might therefore wonder whether the alignment rate is always $\exp(-(\lambda_2 - \lambda_1)t)$. In the nonlinear world the slower coordinate, corresponding to the small eigenvalue, can act back on the fastest one through the nonlinearity and eventually mitigate the influence of λ_2 . This is what happens with $f: \mathbb{R}^2 \rightarrow \mathbb{R}$, defined by

$$f(x_1, x_2) = \frac{1}{2}\lambda_1 x_1^2 + \frac{1}{2}\lambda_2 x_2^2 + a x_1^2 x_2, \quad a \neq 0, \text{ where } 0 < 2\lambda_1 < \lambda_2.$$

The origin is a nondegenerate critical point and $\nabla^2 f(0) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. Thus, the tangent space to the talweg is $T_0 \text{Ext}_1 = \mathbb{R}(1, 0)$. Let $x_0 = (\alpha_1, \alpha_2) \in \mathbb{R}^d$, with $\alpha_1 \neq 0$, and for every $t \in [0, \infty)$, set $x(t) = \Phi_t(x_0)$. We prove in Section A.2 in the Appendix that, provided that $|x_0|$ is small enough,

$$\text{dist}(\rho(x(t)), T_0 \text{Ext}_1) = C \exp(-\lambda_1 t) + o(\exp(-\lambda_1 t)), \text{ for some } C > 0. \quad (3.27)$$

Since here $\lambda_2 - \lambda_1 > \lambda_1$, the spectral gap would predict a faster decay of order $\exp(-(\lambda_2 - \lambda_1)t)$. This example shows that such a rate cannot be expected in general.

Corollary 3.5 (Valley is generically reached in finite time). *Under Assumptions 2.1 and 2.2, let $\eta > 0$ and $H: B(x^*, \eta) \rightarrow \mathbb{R}^d$ be as in Theorem 3.2. Then, for every $w > 0$ and $x \in \mathcal{S}_d^H$, there exists $t_0 \in [0, \infty)$ such that*

$$\Phi_t(x) \in \text{Val}_{f,w}(x^*), \quad \text{for all } t \geq t_0.$$

Proof. Let $0 < w_1 < w$. By Proposition 2.8 (i), there exists $\delta \in (0, \eta)$ such that

$$x^* + (\text{Val}_{q,w_1} \cap B(0, \delta)) \subset \text{Val}_{f,w}(x^*). \quad (3.28)$$

Assume for instance that $\rho(\Phi_t(x) - x^*) \rightarrow v_1$ as $t \rightarrow +\infty$. Since Val_{q,w_1} is a cone generated by v_1 , this implies that for t large enough, $\Phi_t(x) - x^* \in \text{Val}_{q,w_1}$. On the other hand, we know that $\Phi_t(x) \rightarrow x^*$, as $t \rightarrow +\infty$, therefore, for t sufficiently large, $\Phi_t(x) - x^* \in \text{Val}_{q,w_1} \cap B(0, \delta)$. Using (3.28), we conclude that $\Phi_t(x) \in \text{Val}_{f,w}(x^*)$ for all t large enough. \square

Given a Lebesgue measurable set S , we denote by $\text{vol}(S) = \int_S 1 dx$ its volume.

Theorem 3.6 (Volume concentration in valleys). *Suppose that Assumptions 2.1 and 2.2 hold and let $\eta > 0$ be small enough. Then for every measurable set $S \subset B(x^*, \eta)$ of initial conditions, with $\text{vol}(S) > 0$, and $w > 0$, we have*

$$\lim_{t \rightarrow \infty} \frac{\text{vol}(\Phi_t(S) \cap \text{Val}_{f,w}(x^*))}{\text{vol}(\Phi_t(S))} = 1. \quad (3.29)$$

Proof. For every $t \in [0, \infty)$, Φ_t is a local diffeomorphism and, for every $x \in \mathbb{R}^d$, setting $Z(t, x) = \det \partial_x \Phi_t(x)$, Jacobi's formula implies

$$\begin{aligned} \partial_t Z(t, x) &= -\text{tr}(\nabla^2 f(\Phi_t(x))) Z(t, x) = -\left(\sum_{i=1}^d \lambda_i(\Phi_t(x))\right) Z(t, x), \\ Z(0, x) &= 1. \end{aligned}$$

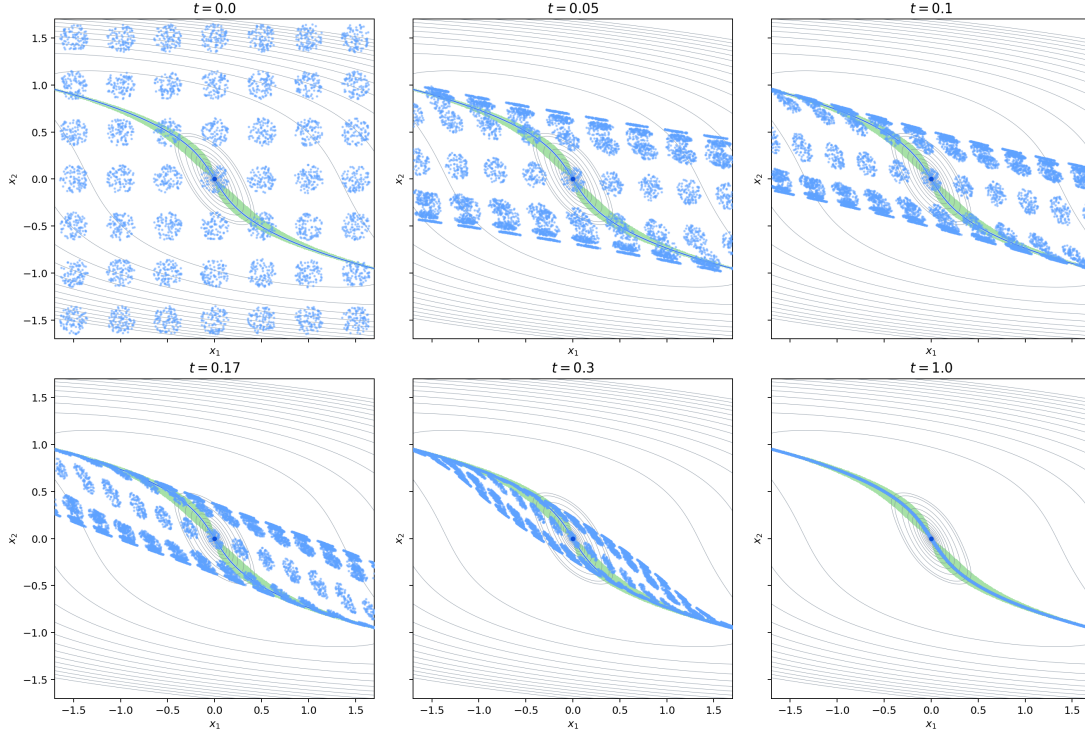


Figure 3: Uniformly distributed balls form the set S on $[-5, 5]^2$ of Theorem 3.6. They are randomly sampled and deformed by the gradient flow. This illustrates the concentration phenomenon of formula (3.29) and provides some mathematical insights into the “formation of rivers”.

In turn,

$$\det \partial_x \Phi_t(x) = \exp \left(- \int_0^t \sum_{i=1}^d \lambda_i(\Phi_s(x)) ds \right), \quad \text{for all } (t, x) \in [0, \infty) \times \mathbb{R}^d.$$

On the other hand, it follows from Theorem 3.2, Lemma A.3, and Corollary A.5 in the Appendix, one can shrink η to obtain a local diffeomorphism $H: B(x^*, \eta) \rightarrow \mathbb{R}^d$ such that $H(x^*) = 0$, $DH(x^*) = I_d$, $\Phi_t(B(x^*, \eta)) \subset B(x^*, \eta)$, for all $t \in [0, \infty)$, the conjugacy relation (3.8) holds, and

$$\sup_{x \in B(x^*, \eta)} |H(x)| + \max_{i=1, \dots, d} \sup_{x \in B(x^*, \eta)} |\nabla \lambda_i(x)| + \sup_{y \in N} \|DH^{-1}(y)\| < \infty, \quad (3.30)$$

where N is a closed convex neighborhood containing $H(B(x^*, \eta))$ and on which DH^{-1} is well defined. In particular, by the mean value theorem there exists $C_1 > 0$ such that, for every $x \in B(x^*, \eta)$ and $t \in [0, \infty)$,

$$\begin{aligned} \exp \left(-t \sum_{i=1}^d \lambda_i(x^*) - C_1 \int_0^t |\Phi_s(x) - x^*| ds \right) &\leq \det \partial_x \Phi_t(x) \\ &\leq \exp \left(-t \sum_{i=1}^d \lambda_i(x^*) + C_1 \int_0^t |\Phi_s(x) - x^*| ds \right). \end{aligned}$$

Using (3.9) and (3.30), we get $\sup_{x \in B(x^*, \eta)} \int_0^\infty |\Phi_s(x) - x^*| ds < \infty$ and, hence, there exists $\zeta > 0$ such that, for every $x \in B(x^*, \eta)$ and $t \in (0, \infty)$,

$$\exp \left(-t \sum_{i=1}^d \lambda_i(x^*) - \zeta \right) \leq \det \partial_x \Phi_t(x) \leq \exp \left(-t \sum_{i=1}^d \lambda_i(x^*) + \zeta \right). \quad (3.31)$$

Finally, let $S \subset B(x^*, \eta)$ be a measurable set, with $\text{vol}(S) > 0$, and $w > 0$. Since, for every $t \in [0, \infty)$, Φ_t is a local diffeomorphism, we have $\text{vol}(\Phi_t(S)) > 0$ and

$$\frac{\text{vol}(\Phi_t(S) \cap \text{Val}_{f,w}(x^*))}{\text{vol}(\Phi_t(S))} = 1 - R(t), \quad \text{for all } t \in [0, \infty), \quad (3.32)$$

where, by the change of variable formula,

$$R(t) := \frac{\text{vol}(\Phi_t(S) \setminus \text{Val}_{f,w}(x^*))}{\text{vol}(\Phi_t(S))} = \frac{\int_{S \setminus \Phi_t^{-1}(\text{Val}_{f,w}(x^*))} \det(\partial_x \Phi_t(x)) \, dx}{\int_S \det(\partial_x \Phi_t(x)) \, dx}.$$

Estimate (3.31) yields

$$R(t) \leq \exp(2\zeta) \frac{\text{vol}(S \setminus \Phi_t^{-1}(\text{Val}_{f,w}(x^*)))}{\text{vol}(S)}, \quad \text{for all } t \in (0, \infty).$$

In turn, it follows from Corollary 3.5, $\text{vol}(B(x^*, \eta) \setminus \mathcal{S}_d^H) = 0$, and dominated convergence that $R(t) \rightarrow 0$, as $t \rightarrow \infty$, and, hence, (3.29) follows from (3.32). \square

3.3 Directional convergence for saddle points

The previous section was centered on the behavior of the flow for the convergence toward local minimizers; the same analysis may be applied, in a Riemannian fashion, to any generic saddle point x^* , i.e., for which $\lambda_1(x^*)\lambda_d(x^*) < 0$, after restricting f to its local stable manifold $W_{x^*}^s$; see [27] for theoretical elements.

Indeed, if we endow $W_{x^*}^s$ with the induced Euclidean metric of \mathbb{R}^d , then $W_{x^*}^s$ is a smooth Riemannian manifold and the restricted function $\hat{f} := f|_{W_{x^*}^s}$ has x^* as a strong local minimizer. Moreover, the Riemannian gradient flow of \hat{f} is exactly the restriction of the ambient gradient flow to $W_{x^*}^s$. Hence the notions introduced in Section 2, gradient extremals, talweg, and valleys, are Riemannian objects that are well defined on $W_{x^*}^s$, $\langle \cdot, \cdot \rangle$, and the alignment and concentration results of this section may be developed in this non-Euclidean setting.

Recent results also show that maxima, minima, and alignment phenomena share interesting connections. For this, we also refer the reader to [3], where, using Hartman's theorem as in the proof of Theorem 3.2, it is shown that if x^* is a nondegenerate local maximum, then, generically, one can associate to x^* two local minimizers, $m_-(x^*)$ and $m_+(x^*)$, such that gradient trajectories starting near x^* converge, with high probability, to one of these two minima.

4 Directional convergence for the gradient method

As convergence of the gradient method generically occurs to a local minimizer [18], we work throughout this section under Assumption 2.2 with x^* being a local minimizer. We shall systematically use the constants $L_1, L_2 > 0$ defined through

$$L_2 := \sup_{x \in B(0, \eta)} \|\nabla^2 f(x)\| < \infty, \quad (4.1)$$

and

$$\langle \nabla^2 f(x)h, h \rangle \geq L_1|h|^2, \quad (4.2)$$

for any $x \in B(x^*, \eta)$ and $h \in \mathbb{R}^d$.

Given $\gamma > 0$, we consider the constant step gradient method

$$x \in B(x^*, \eta), \quad x_0 := x, \quad x_{k+1} = x_k - \gamma \nabla f(x_k), \quad (4.3)$$

for any $k \geq 0$. Given $k \in \mathbb{N}$ and $x \in B(x^*, \eta)$, we denote by $\Phi_k(x)$ the k^{th} iterate in (4.3).

We will need the following standard result on the discrete flow Φ_k . For the sake of completeness, we provide its short proof.

Lemma 4.1 (Gradient descent stability for strong minimizers). *Suppose that $\gamma \in (0, 2/L_2)$ and define $s(\gamma) \in (0, 1)$ by*

$$s(\gamma) = \max\{|1 - \gamma L_1|, |1 - \gamma L_2|\} = \begin{cases} 1 - L_1\gamma, & \text{if } \gamma \in \left(0, \frac{2}{L_1+L_2}\right], \\ L_2\gamma - 1, & \text{if } \gamma \in \left(\frac{2}{L_1+L_2}, \frac{2}{L_2}\right). \end{cases}$$

Then Φ_1 is $s(\gamma)$ -Lipschitz on $B(x^*, \eta)$ and

$$(\forall k \geq 0) \quad |\Phi_{k+1}(x) - x^*| \leq s(\gamma) |\Phi_k(x) - x^*|. \quad (4.4)$$

In particular, $\Phi_k(B(x^*, \eta)) \subset B(x^*, \eta)$ for all $k \in \mathbb{N}$.

Proof. It follows from (4.1) and (4.2) that, for every $x \in B(x^*, \eta)$, the spectrum of the symmetric matrix $D\Phi_1(x) = I_d - \gamma \nabla^2 f(x)$ is contained in $[-s(\gamma), s(\gamma)]$, which implies that Φ_1 is $s(\gamma)$ -Lipschitz on $B(x^*, \eta)$. In particular, for every $x \in B(x^*, \eta)$,

$$|\Phi_{k+1}(x) - x^*| = |\Phi_1(\Phi_k(x)) - \Phi_1(x^*)| \leq s(\gamma) |\Phi_k(x) - x^*|,$$

from which (4.4) follows. \square

For every $k \in \mathbb{N}$ and $y \in \mathbb{R}^d$, denote by $\Psi_k(y)$ the k^{th} iterate of the sequence

$$y_0 = y, \quad y_{k+1} = y_k - \gamma \nabla^2 f(x^*) y_k, \quad (4.5)$$

i.e. $y_k = (I_d - \gamma \nabla^2 f(x^*))^k y_0$ for all $k \in \mathbb{N}$. Recall that the manifolds \mathcal{S}_i ($i = 0, \dots, d$) are defined in (3.2). Since $\nabla^2 f(x^*) = P(x^*) \mathcal{D}(x^*) P^\top(x^*)$, for every $i = 1, \dots, d$ and $y = \sum_{j=d-i+1}^d \alpha_j v_j(x^*) \in \mathcal{S}_i$, we have for all $k \geq 0$,

$$\Psi_k(y) = \sum_{j=d-i+1}^d (1 - \gamma \lambda_j(x^*))^k \alpha_j v_j(x^*) \quad (4.6)$$

and hence, if $\gamma \in (0, 2/\lambda_d(x^*))$, one has

$$|\Psi_k(y)| \leq \kappa_i(\gamma)^k |y|, \quad \text{for all } k \geq 0, \quad (4.7)$$

where $\kappa_i(\gamma) \in (0, 1)$ is given by

$$\kappa_i(\gamma) = \begin{cases} 1 - \gamma \lambda_{d-i+1}(x^*), & \text{if } \gamma \in \left(0, \frac{2}{\lambda_{d-i+1}(x^*) + \lambda_d(x^*)}\right], \\ \gamma \lambda_d(x^*) - 1, & \text{if } \gamma \in \left(\frac{2}{\lambda_{d-i+1}(x^*) + \lambda_d(x^*)}, \frac{2}{\lambda_d(x^*)}\right). \end{cases} \quad (4.8)$$

The following result provides the asymptotic directional behavior of the discrete flow Ψ .

Lemma 4.2. Let $i = 2, \dots, d$, let $y = \sum_{j=d-i+1}^d \alpha_j v_j(x^*) \in \mathcal{S}_i$, and take $\gamma \in \left(0, \frac{2}{\lambda_{d-i+1}(x^*) + \lambda_d(x^*)}\right)$. Then, for every $k \geq 0$, we have

$$|\rho(\Psi_k(y)) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \leq \frac{2|y|}{|\alpha_{d-i+1}|} \kappa_{\text{dir},i}(\gamma)^k, \quad (4.9)$$

$$\left| \rho\left(\frac{\Psi_{k+1}(y) - \Psi_k(y)}{\gamma}\right) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*) \right| \leq \frac{2|y|}{|\alpha_{d-i+1}|} \left(\frac{\lambda_d(x^*)}{\lambda_{d-i+1}(x^*)}\right) \kappa_{\text{dir},i}(\gamma)^k, \quad (4.10)$$

where $\kappa_{\text{dir},i}(\gamma) \in (0, 1)$ is given by

$$\kappa_{\text{dir},i}(\gamma) = \begin{cases} \frac{1 - \gamma\lambda_{d-i+2}(x^*)}{1 - \gamma\lambda_{d-i+1}(x^*)}, & \text{if } \gamma \in \left(0, \frac{2}{\lambda_{d-i+2}(x^*) + \lambda_d(x^*)}\right], \\ \frac{\gamma\lambda_d(x^*) - 1}{1 - \gamma\lambda_{d-i+1}(x^*)}, & \text{if } \gamma \in \left(\frac{2}{\lambda_{d-i+2}(x^*) + \lambda_d(x^*)}, \frac{2}{\lambda_{d-i+1}(x^*) + \lambda_d(x^*)}\right). \end{cases} \quad (4.11)$$

Proof. We only show (4.10) as (4.9) follows from similar computations. By (4.6), we have

$$\begin{aligned} \frac{\Psi_{k+1}(y) - \Psi_k(y)}{\gamma} &= -\alpha_{d-i+1}\lambda_{d-i+1}(x^*)(1 - \gamma\lambda_{d-i+1}(x^*))^k v_{d-i+1}(x^*) \\ &\quad - \sum_{j=d-i+2}^d \alpha_j \lambda_j(x^*)(1 - \gamma\lambda_j(x^*))^k v_j(x^*). \end{aligned} \quad (4.12)$$

On the other hand, since $\lambda_{d-i+1}(x^*) > 0$ and $1 - \gamma\lambda_{d-i+1}(x^*) > 0$, we have

$$\rho\left(\frac{\Psi_{k+1}(y) - \Psi_k(y)}{\gamma}\right) = \rho\left(\lambda_{d-i+1}(x^*)^{-1}(1 - \gamma\lambda_{d-i+1}(x^*))^{-k} \frac{\Psi_{k+1}(y) - \Psi_k(y)}{\gamma}\right)$$

and, hence, (3.1) yields

$$\begin{aligned} &\left| \rho\left(\frac{\Psi_{k+1}(y) - \Psi_k(y)}{\gamma}\right) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*) \right| \\ &\leq \left| \rho\left(\frac{\Psi_{k+1}(y) - \Psi_k(y)}{\gamma}\right) - \rho(-\alpha_{d-i+1}v_{d-i+1}(x^*)) \right| \\ &\leq \frac{2}{|\alpha_{d-i+1}|} \left| \lambda_{d-i+1}(x^*)^{-1}(1 - \gamma\lambda_{d-i+1}(x^*))^{-k} \frac{\Psi_{k+1}(y) - \Psi_k(y)}{\gamma} + \alpha_{d-i+1}v_{d-i+1}(x^*) \right|. \end{aligned} \quad (4.13)$$

Combining with (4.12) and noting that

$$\kappa_{\text{dir},i}(\gamma) = \max \left\{ \left| \frac{1 - \gamma\lambda_j(x^*)}{1 - \gamma\lambda_{d-i+1}(x^*)} \right| : j = d - i + 2, \dots, d \right\},$$

we get

$$\begin{aligned} &\left| \rho\left(\frac{\Psi_{k+1}(y) - \Psi_k(y)}{\gamma}\right) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*) \right| \\ &\leq \frac{2}{|\alpha_{d-i+1}|} \left| \sum_{j=d-i+2}^d \alpha_j \frac{\lambda_j(x^*)}{\lambda_{d-i+1}(x^*)} \left(\frac{1 - \gamma\lambda_j(x^*)}{1 - \gamma\lambda_{d-i+1}(x^*)}\right)^k v_j(x^*) \right| \\ &\leq \frac{2|y|}{|\alpha_{d-i+1}|} \left(\frac{\lambda_d(x^*)}{\lambda_{d-i+1}(x^*)}\right) \kappa_{\text{dir},i}(\gamma)^k. \end{aligned} \quad (4.14)$$

This ends the proof. \square

Remark 4.3 (Link with the power method). One can easily recover estimate (4.9) by interpreting the sequence of directions $(\rho(\Psi_k(y)))_{k \in \mathbb{N}}$ as the iterates of the power method to approximate $v_{d-i+1}(x^*)$ (see, e.g., [1, Theorem 10.3.1]).

The following result deals with the asymptotic directional behavior of the discrete flow Φ .

Theorem 4.4 (Directional convergence of the gradient method). *Under Assumptions 2.1 and 2.2:*

(i) (C^1 conjugacy) *For $\eta > 0$ small enough there exists a local C^1 -diffeomorphism $H: B(x^*, \eta) \rightarrow \mathbb{R}^d$ satisfying $H(x^*) = 0$ and $DH(x^*) = I_d$, such that, if $\gamma \in (0, 2/\lambda_d(x^*))$,*

$$H(\Phi_k(x)) = \Psi_k(H(x)) \text{ for all } x \text{ in } B(x^*, \eta) \text{ and } k \geq 0. \quad (4.15)$$

(ii) (Partition into stable manifolds) *Setting $\mathcal{S}_i^H = H^{-1}(\mathcal{S}_i)$ for $i = 0, \dots, d$, the family $\{\mathcal{S}_i^H\}_{i=0}^d$ forms a partition of $B(x^*, \eta)$ into manifolds such that, for all $i = 0, \dots, d$, $\dim \mathcal{S}_i^H = i$, and for all $k \geq 0$,*

$$\Phi_k(\mathcal{S}_i^H) \subset \mathcal{S}_i^H \text{ whenever } \gamma \neq \frac{1}{\lambda_{d-i+1}(x^*)}.$$

(iii) (Convergence and directional alignment) *There exists $c > 0$ such that, for all $x \in \mathcal{S}_i^H$ and $k \geq 0$,*

$$|\Phi_k(x) - x^*| \leq c |H(x)| \kappa_i(\gamma)^k, \quad (4.16)$$

where κ_i is defined in (4.8). *In addition, if $\gamma \in (0, 2/(\lambda_{d-i+1}(x^*) + \lambda_d(x^*)))$, then, for every $x \in \mathcal{S}_i^H$,*

$$\lim_{k \rightarrow \infty} \rho(\Phi_k(x) - x^*) = - \lim_{k \rightarrow \infty} \rho(\Phi_{k+1}(x) - \Phi_k(x)) \in \{-v_{d-i+1}(x^*), v_{d-i+1}(x^*)\}. \quad (4.17)$$

Proof. Ideas are similar to Theorem 3.2. By [14, Theorem (I)] there exist $\eta > 0$ and a local diffeomorphism $H: B(x^*, \eta) \rightarrow \mathbb{R}^d$ of class C^1 such that $H(x^*) = 0$, $DH(x^*) = I_d$, and $H(\Phi_1(x)) = \Psi_1(H(x))$ for all $x \in B(x^*, \eta)$. Since $\gamma < 2/\lambda_d(x^*)$ and λ_d is continuous in $B(x^*, \eta)$, by reducing η , if necessary, we can assume that $\gamma < 2/L_2$, where L_2 satisfies (4.1). Thus, Lemma 4.1 implies that, for every $x \in B(x^*, \eta)$, we have $\Phi_k(x) \in B(x^*, \eta)$ for all $k \geq 0$. Using that $\Phi_{k_1+k_2} = (\Phi_{k_1} \circ \Phi_{k_2})(x)$ for all $x \in B(x^*, \eta)$ and $k_1, k_2 \geq 0$, we deduce that the conjugacy (4.15) of (i) holds.

Since (4.6) implies that for every $i = 1, \dots, d$, $k \geq 0$, and $\gamma \in (0, 2/\lambda_d(x^*)) \setminus \{1/\lambda_{d-i+1}(x^*)\}$, $\Psi_k(\mathcal{S}_i) \subset \mathcal{S}_i$, it follows from (4.15) that $\Phi_k(\mathcal{S}_i^H) \subset \mathcal{S}_i^H$. Moreover, for every $k \geq 0$ and $x \in B(x^*, \eta)$, we have $\Phi_k(x) - x^* = H^{-1}(\Psi_k(H(x))) - H^{-1}(0)$ and, hence, (4.16) follows from the boundedness of DH^{-1} and (4.7).

Let us establish (ii). In order to show (4.17), let $x \in \mathcal{S}_i^H$ and, for every $j = 1, \dots, d$, set $\alpha_j := \langle H(x), v_j(x^*) \rangle$. Since $\gamma \in (0, 2/(\lambda_{d-i+1}(x^*) + \lambda_d(x^*)))$, one has, by definition, $\kappa_i(\gamma) = 1 - \gamma \lambda_{d-i+1}(x^*)$. For every $k \geq 0$, we have

$$\begin{aligned} & |\rho(\Phi_k(x) - x^*) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \\ & \leq |\rho(\Phi_k(x) - x^*) - \rho(\Psi_k(H(x)))| + |\rho(\Psi_k(H(x))) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)|. \end{aligned} \quad (4.18)$$

Note that (4.6) and the orthogonality of the family $\{v_i(x^*)\}_{i=1}^d$ imply that

$\left| \kappa_i^{-k}(\gamma) \Psi_k(H(x)) \right| \geq |\alpha_{d-i+1}| > 0$. In turn, (3.1) yields

$$\begin{aligned} & \left| \rho(\Phi_k(x) - x^*) - \rho(\Psi_k(H(x))) \right| \\ &= \left| \rho\left(\kappa_i^{-k}(\gamma)(\Phi_k(x) - x^*)\right) - \rho\left(\kappa_i^{-k}(\gamma)\Psi_k(H(x))\right) \right| \\ &\leq \frac{2}{|\alpha_{d-i+1}|} \left| \kappa_i^{-k}(\gamma) [\Phi_k(x) - x^* - \Psi_k(H(x))] \right|. \end{aligned} \quad (4.19)$$

Since $\Phi_k(x) - x^* = H^{-1}(\Psi_k(H(x))) - H^{-1}(0)$ and $DH^{-1}(0) = I_d$, we also have

$$\Phi_k(x) - x^* - \Psi_k(H(x)) = \left(\int_0^1 (DH^{-1}(\tau\Psi_k(H(x))) - DH^{-1}(0)) \, d\tau \right) \Psi_k(H(x)), \quad (4.20)$$

which, together with (4.19), (4.7), and the continuity of DH^{-1} , yields

$$\begin{aligned} & \left| \rho(\Phi_k(x) - x^*) - \rho(\Psi_k(H(x))) \right| \\ &\leq \frac{2|H(x)|}{|\alpha_{d-i+1}|} \int_0^1 \|DH^{-1}(\tau\Psi_k(H(x))) - DH^{-1}(0)\| \, d\tau \xrightarrow[k \rightarrow \infty]{} 0. \end{aligned} \quad (4.21)$$

It then follows from (4.18) and Lemma 4.2 that

$$\left| \rho(\Phi_k(x) - x^*) - \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*) \right| \xrightarrow[k \rightarrow \infty]{} 0. \quad (4.22)$$

Similarly, one has

$$\begin{aligned} & \left| \rho\left(\frac{\Phi_{k+1}(x) - \Phi_k(x)}{\gamma}\right) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*) \right| \\ &\leq \left| \rho\left(\frac{\Phi_{k+1}(x) - \Phi_k(x)}{\gamma}\right) - \rho\left(\frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma}\right) \right| \\ &+ \left| \rho\left(\frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma}\right) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*) \right|. \end{aligned} \quad (4.23)$$

From (4.12) and the orthogonality of the family $\{v_i(x^*)\}_{i=1}^d$, we have

$$\left| \lambda_{d-i+1}(x^*)^{-1} \kappa_i^{-k}(\gamma) \left(\frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma} \right) \right| \geq |\alpha_{d-i+1}| > 0,$$

which, by (3.1), implies after proper rescaling

$$\begin{aligned} & \left| \rho\left(\frac{\Phi_{k+1}(x) - \Phi_k(x)}{\gamma}\right) - \rho\left(\frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma}\right) \right| \\ &\leq \frac{2\kappa_i^{-k}(\gamma)}{\lambda_{d-i+1}(x^*)|\alpha_{d-i+1}|} \left| \frac{\Phi_{k+1}(x) - \Phi_k(x)}{\gamma} - \frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma} \right|. \end{aligned} \quad (4.24)$$

Note that, by (4.3) and (4.5), we have

$$\left| \frac{\Phi_{k+1}(x) - \Phi_k(x)}{\gamma} - \frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma} \right| = \left| \nabla f(\Phi_k(x)) - \nabla^2 f(x^*) \Psi_k(H(x)) \right|.$$

Since f is of class C^3 and $\nabla f(x^*) = 0$, there exists $c > 0$ such that

$$\left| \nabla f(\Phi_k(x)) - \nabla^2 f(x^*)(\Phi_k(x) - x^*) \right| \leq c|\Phi_k(x) - x^*|^2$$

and, hence, by the triangular inequality, (4.1), and (4.20),

$$\begin{aligned}
& \left| \frac{\Phi_{k+1}(x) - \Phi_k(x)}{\gamma} - \frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma} \right| \\
& \leq c|\Phi_k(x) - x^*|^2 + L_2 |\Phi_k(x) - x^* - \Psi_k(H(x))| \\
& \leq c|\Phi_k(x) - x^*|^2 + L_2 \int_0^1 \|DH^{-1}(\tau\Psi_k(H(x))) - DH^{-1}(0)\| d\tau |\Psi_k(H(x))|.
\end{aligned} \tag{4.25}$$

It then follows from (4.23), (4.24), (4.16), (4.7), the continuity of DH^{-1} , and Lemma 4.2 that

$$|\rho(\Phi_{k+1}(x) - \Phi_k(x)) + \text{sign}(\alpha_{d-i+1})v_{d-i+1}(x^*)| \xrightarrow[k \rightarrow \infty]{} 0. \tag{4.26}$$

Therefore, (4.17) follows from (4.22) and (4.26). \square

Remark 4.5 (The strict stepsize condition is sharp). Indeed, in the quadratic case, if $\gamma \in [2/(\lambda_{d-i+1}(x^*) + \lambda_d(x^*)), 2/\lambda_d)$ and $x \in \mathcal{S}_i^H$, the directions $\rho(\Psi_k(H(x)))$ and $\rho(\Psi_{k+1}(H(x)) - \Psi_k(H(x)))$ do not converge in general.

For instance, in the case of $(\rho(\Psi_k(H(x))))_{k \in \mathbb{N}}$, if $\gamma > 2/(\lambda_{d-i+1}(x^*) + \lambda_d(x^*))$ and $\alpha_d = \langle H(x), v_d(x^*) \rangle \neq 0$, arguing as in the proof of Lemma 4.11, one gets that

$$\begin{aligned}
\rho(\Psi_{2k}(H(x))) & \xrightarrow[k \rightarrow \infty]{} \text{sign}(\alpha_d)v_d(x^*), \\
\rho(\Psi_{2k+1}(H(x))) & \xrightarrow[k \rightarrow \infty]{} -\text{sign}(\alpha_d)v_d(x^*),
\end{aligned}$$

i.e., $\rho(\Psi_k(x))$ aligns, in an oscillatory manner, with $\mathbb{R}v_d(x^*)$, the eigenspace associated with the largest eigenvalue $\lambda_d(x^*)$ of $\nabla^2 f(x^*)$. If $\gamma = 2/(\lambda_{d-i+1}(x^*) + \lambda_d(x^*))$, which is the stepsize for which one has the optimal rate of convergence in (4.7), then

$$\begin{aligned}
\rho(\Psi_{2k}(H(x))) & \xrightarrow[k \rightarrow \infty]{} \frac{\alpha_{d-i+1}v_{d-i+1}(x^*) + \alpha_d v_d(x^*)}{\sqrt{\alpha_{d-i+1}^2 + \alpha_d^2}}, \\
\rho(\Psi_{2k+1}(H(x))) & \xrightarrow[k \rightarrow \infty]{} \frac{\alpha_{d-i+1}v_{d-i+1}(x^*) - \alpha_d v_d(x^*)}{\sqrt{\alpha_{d-i+1}^2 + \alpha_d^2}},
\end{aligned}$$

i.e. $\rho(\Psi_k(x))$ oscillates between two elements of $\text{span}\{v_{d-i+1}(x^*), v_d(x^*)\}$. The same behavior occurs for $\rho(\Psi_{k+1}(H(x)) - \Psi_k(H(x)))$. The sequences $\rho(\Phi_k(x))$ and $\rho(\Phi_{k+1}(x) - \Phi_k(x))$ can be studied as in the proof of Theorem 4.4.

Theorem 4.6. (i) (Generic alignment) *Suppose that Assumption 2.1 holds at each critical point x^* and that ∇f is L -Lipschitz continuous with $L > 0$. If $\gamma \in (0, 1/L)$ and f is coercive, the set of initial conditions for which the gradient method converges to local minimizers of f , with both velocities and secants aligning with the associated talweg, has full Lebesgue measure.*

(ii) (Alignment rate) *Suppose that f is C^∞ , let x be such that $\Phi_k(x) \rightarrow x^*$ as $k \rightarrow \infty$, where x^* is a minimizer as in Assumption 2.2 and $\Phi_k(x) \in \mathcal{S}_d^H(x^*)$ for some k . Assume in addition the non-resonance conditions at x^* : for all integers m_1, \dots, m_d such that $\sum_{j=1}^d m_j \geq 2$,*

$$\lambda_i(x^*) \neq \sum_{j=1}^d m_j \lambda_j(x^*) \text{ for all } i = 1, \dots, d.$$

If $\gamma \in \left(0, \frac{2}{\lambda_1(x^*) + \lambda_2(x^*)}\right)$, then alignment with the tangent to the talweg goes at a rate:

$$\begin{aligned} \text{dist}\left(\rho(\Phi_k(x) - x^*), T_{x^*}\text{Tal}([r^*, \bar{r}])\right) &= O\left(\left(\frac{1 - \gamma\lambda_2(x^*)}{1 - \gamma\lambda_1(x^*)}\right)^k\right) + O((1 - \gamma\lambda_1(x^*))^k), \\ \text{dist}\left(\rho(\Phi_{k+1}(x) - \Phi_k(x)), T_{x^*}\text{Tal}([r^*, \bar{r}])\right) &= O\left(\left(\frac{1 - \gamma\lambda_2(x^*)}{1 - \gamma\lambda_1(x^*)}\right)^k\right) + O((1 - \gamma\lambda_1(x^*))^k). \end{aligned}$$

Proof. Item (i) is a mere adaptation of the proof of Theorem 3.3 (i), using Theorem 4.4, and the fact that, for every critical point x^* ,

$$\frac{2}{\lambda_1(x^*) + \lambda_d(x^*)} \geq 1/L.$$

We prove (ii). Observe that, by assumption $\Phi_k(x) \neq \Phi_{k+1}(x)$ for $k \geq 0$, both being different from x^* . Under the non-resonance condition and the smoothness of f , Sternberg's theorem [25, Theorem 2] ensures that the conjugacy H may be chosen C^k with $k \geq 2$, so that there exists $C_H > 0$ such that

$$\|DH^{-1}(y) - DH^{-1}(0)\| \leq C_H|y| \text{ for } y \text{ in a neighborhood of } 0.$$

Write $H(x) = \sum_{j=1}^d \alpha_j v_j(x^*)$, with $\alpha_1 \neq 0$.

Let us estimate for $\rho(\Phi_k(x) - x^*)$. Using the same triangular decomposition as in the proof of Theorem 4.4,

$$\begin{aligned} &|\rho(\Phi_k(x) - x^*) - \text{sign}(\alpha_1)v_1(x^*)| \\ &\leq |\rho(\Phi_k(x) - x^*) - \rho(\Psi_k(H(x)))| + |\rho(\Psi_k(H(x))) - \text{sign}(\alpha_1)v_1(x^*)|. \end{aligned}$$

The second term is controlled by Lemma 4.2: $|\rho(\Psi_k(H(x))) - \text{sign}(\alpha_1)v_1(x^*)| = O(\kappa_{\text{dir},1}(\gamma)^k)$. For the first term, using (4.21) and (4.7) together with the Lipschitz continuity of DH^{-1} yields

$$|\rho(\Phi_k(x) - x^*) - \rho(\Psi_k(H(x)))| = O(\kappa_1(\gamma)^k).$$

Recalling that $\text{sign}(\alpha_1)v_1(x^*) \in T_{x^*}\text{Tal}([r^*, \bar{r}])$ proves the secant estimate. Let us estimate now the direction of the ‘‘discrete velocity’’. Arguing as in (4.23),

$$\begin{aligned} &\left| \rho\left(\frac{\Phi_{k+1}(x) - \Phi_k(x)}{\gamma}\right) + \text{sign}(\alpha_1)v_1(x^*) \right| \\ &\leq \left| \rho\left(\frac{\Phi_{k+1}(x) - \Phi_k(x)}{\gamma}\right) - \rho\left(\frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma}\right) \right| \\ &\quad + \left| \rho\left(\frac{\Psi_{k+1}(H(x)) - \Psi_k(H(x))}{\gamma}\right) + \text{sign}(\alpha_1)v_1(x^*) \right|. \end{aligned}$$

The second term is again controlled by the quadratic case Lemma 4.2. The first term is estimated using (4.24), (4.25), (4.16), (4.7), and the Lipschitz property of DH^{-1} , yielding a term of order $O(\kappa_1(\gamma)^k)$. Combining the above estimates concludes the proof. \square

Remark 4.7 (Sharpness of the alignment rate in the discrete case). Unlike in the quadratic case, considered in Lemma 4.2, one cannot, in general, expect a pure spectral-gap rate. Indeed, as in Remark 3.4, consider $f(x_1, x_2) = \frac{1}{2}\lambda_1 x_1^2 + \frac{1}{2}\lambda_2 x_2^2 + ax_1^2 x_2$ with $0 < \lambda_1 < \lambda_2$ and $a \neq 0$. Assume that $\lambda_2 > 2\lambda_1$ and let $\gamma \in (0, 1/\lambda_2)$. Given $x_0 = (\alpha_1, \alpha_2)$, with $\alpha_1 \neq 0$, we consider the gradient descent iterates $x_k = (x_{k,1}, x_{k,2})$ defined through

$$\begin{aligned} x_{1,k+1} &= (1 - \gamma\lambda_1)x_{1,k} - 2a\gamma x_{1,k}x_{2,k}, \\ x_{2,k+1} &= (1 - \gamma\lambda_2)x_{2,k} - a\gamma x_{1,k}^2, \end{aligned} \quad (4.27)$$

for all $k \in \mathbb{N}$. We show in Section A.2 in the Appendix that, provided that $|x_0|$ is small enough,

$$\text{dist}(\rho(x_k), T_0\text{Ext}_1) = C(1 - \gamma\lambda_1)^k + o\left((1 - \gamma\lambda_1)^k\right), \text{ for some } C > 0. \quad (4.28)$$

Corollary 4.8 (Valley is generically reached in finite time (discrete case)). *Suppose that Assumptions 2.1 and 2.2 hold, let $\eta > 0$, $H: B(x^*, \eta) \rightarrow \mathbb{R}^d$ be as in Theorem 4.4(i), and let $\gamma \in (0, 2/(\lambda_1(x^*) + \lambda_d(x^*)))$. Then, for every $w > 0$ and $x \in \mathcal{S}_d^H(x^*)$, there exists $k_0 \in \mathbb{N}$ such that*

$$\Phi_k(x) \in \text{Val}_{f,w}(x^*) \quad \text{for all } k \geq k_0.$$

Proof. Using directional convergences to the talweg and Proposition 2.8, the proof is similar to that of Corollary 3.5. \square

Let us conclude by a result on gradient iterates concentration within valleys:

Theorem 4.9 (Volume concentration in valleys for the gradient method). *Suppose that Assumptions 2.1 and 2.2 hold. Then, if $\eta > 0$ is small enough and $\gamma \in (0, 1/\lambda_d(x^*))$, for every measurable set $S \subset B(x^*, \eta)$, with $\text{vol}(S) > 0$, and $w > 0$, we have*

$$\lim_{k \rightarrow \infty} \frac{\text{vol}(\Phi_k(S) \cap \text{Val}_{f,w}(x^*))}{\text{vol}(\Phi_k(S))} = 1. \quad (4.29)$$

Proof. Since $\gamma < 2/L_2$ for η small enough, Lemma 4.1 implies that, for every $k \geq 0$, $\Phi_k(B(x^*, \eta)) \subset B(x^*, \eta)$. Using that $\gamma \in (0, 1/\lambda_d(x^*))$, one has that $\Phi_1: B(x^*, \eta) \rightarrow \mathbb{R}^d$ is a local diffeomorphism, and hence $\Phi_k = \Phi_1 \circ \dots \circ \Phi_1$ (k times) is also a local diffeomorphism. Moreover, for every $x \in B(x^*, \eta)$ and $k \geq 1$, the chain rule yields

$$\begin{aligned} \det(\partial_x \Phi_k(x)) &= \det\left((I_d - \gamma \nabla^2 f(\Phi_{k-1}(x))) \partial_x \Phi_{k-1}(x)\right) \\ &= \prod_{\ell=0}^{k-1} \det(I_d - \gamma \nabla^2 f(\Phi_\ell(x))). \end{aligned} \quad (4.30)$$

Set $\bar{p} := \det(I_d - \gamma \nabla^2 f(x^*)) = \prod_{i=1}^d (1 - \gamma \lambda_i(x^*)) > 0$ and notice that, since $\mathbb{R}^d \ni x' \mapsto \det(I_d - \gamma \nabla^2 f(x')) \in \mathbb{R}$ is of class C^1 and $\Phi_\ell(B(x^*, \eta)) \subset B(x^*, \eta)$, for all $\ell \geq 0$, there exists a constant $C_1 > 0$ such that

$$\bar{p} - C_1 |\Phi_\ell(x) - x^*| \leq \det(I_d - \gamma \nabla^2 f(\Phi_\ell(x))) \leq \bar{p} + C_1 |\Phi_\ell(x) - x^*| \quad \text{for all } x \in B(x^*, \eta). \quad (4.31)$$

Shrink η so that $0 < \eta < \bar{p}/C_1$, set $a = \bar{p} - C_1\eta > 0$ and $b = \bar{p} + C_1\eta$. For every $x \in B(x^*, \eta)$ and $\ell \geq 0$, setting $p_\ell(x) = C_1 |\Phi_\ell(x) - x^*|$, Lemma 4.1 implies that $\bar{p} \pm p_\ell(x) \in [a, b] \subset (0, \infty)$ for all $\ell \geq 0$. It follows from (4.30), (4.31), and the inequality

$|\log(\alpha) - \log(\beta)| \leq |\alpha - \beta|/a$ for all $\alpha, \beta \in [a, b]$, that, for every $x \in B(x^*, \eta)$ and $k \geq 0$, we have

$$\det(\partial_x \Phi_k(x)) \geq \exp\left(\sum_{\ell=0}^{k-1} \log(\bar{p} - p_\ell(x))\right) \geq \exp\left(k \log(\bar{p}) - \frac{C_1}{a} \sum_{\ell=0}^{k-1} |\Phi_\ell(x) - x^*|\right) \quad (4.32)$$

and

$$\det(\partial_x \Phi_k(x)) \leq \exp\left(\sum_{\ell=0}^{k-1} \log(\bar{p} + p_\ell(x))\right) \leq \exp\left(k \log(\bar{p}) + \frac{C_1}{a} \sum_{\ell=0}^{k-1} |\Phi_\ell(x) - x^*|\right). \quad (4.33)$$

Let H be given by Theorem 4.4. From (4.16), the fact that H can be taken to be bounded, and (4.8), we deduce that

$$\zeta := \frac{C_1}{a} \sum_{\ell=0}^{\infty} \sup_{x \in B(x^*, \eta)} |\Phi_\ell(x) - x^*| < +\infty.$$

In turn, (4.32) and (4.33) yield

$$\exp(k \log(\bar{p}) - \zeta) \leq \det(\partial_x \Phi_k(x)) \leq \exp(k \log(\bar{p}) + \zeta), \quad (4.34)$$

for all $x \in B(x^*, \eta)$ and $k \geq 0$. Since Φ_k is a local diffeomorphism, if $S \subset B(x^*, \eta)$ is measurable, with $\text{vol}(S) > 0$, one also has $\text{vol}(\Phi_k(S)) > 0$. Moreover, for every $w > 0$, we have

$$\frac{\text{vol}(\Phi_k(S) \cap \text{Val}_{f,w}(x^*))}{\text{vol}(\Phi_k(S))} = 1 - R_k \quad \text{for all } k \in \mathbb{N},$$

where

$$R_k := \frac{\text{vol}(\Phi_k(S) \setminus \text{Val}_{f,w}(x^*))}{\text{vol}(\Phi_k(S))} = \frac{\int_{\Phi_k(S) \setminus \text{Val}_{f,w}(x^*)} dx}{\int_{\Phi_k(S)} dx} = \frac{\int_{S \setminus \Phi_k^{-1}(\text{Val}_{f,w}(x^*))} \det(\partial_x \Phi_k(x)) dx}{\int_S \det(\partial_x \Phi_k(x)) dx},$$

which, together with (4.34), implies that

$$R_k \leq \exp(2\zeta) \frac{\text{vol}(S \setminus \Phi_k^{-1}(\text{Val}_{f,w}(x^*)))}{\text{vol}(S)}.$$

In turn, Corollary 4.8, $\text{vol}(B(x^*, \eta) \setminus \mathcal{S}_d^H(x^*)) = 0$, and Lebesgue dominated convergence imply that $R_k \rightarrow 0$, as $k \rightarrow \infty$, from which (4.29) follows. \square

Acknowledgements: PB, JB, TM thank TSE-P and acknowledge financial support of the ANR (Programmes d'Investissements d'Avenir CHESS ANR-17-EURE-0010).

FJS was partially supported by l'Agence Nationale de la Recherche (ANR), project ANR-22-CE40-0010, by KAUST through the subaward agreement ORA-2021-CRG10-4674.6, and by Ministère de l'Europe et des Affaires étrangères (MEAE), project MATH AmSud 23-MATH-17.

JB thanks AI Interdisciplinary Institute ANITI funding, through the ANR under the France 2030 program (grant ANR-23-IACL-0002), Chair TRIAL, ANR MAD, ANR Regulia, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA8655-22-1-7012.

A Appendix

A.1 Additional facts

Proposition A.1 (More properties of the talweg). *Make Assumptions 2.1 and 2.2. If \bar{r} is sufficiently close to r^* , then for every $r \in (r^*, \bar{r})$ the set $\text{Tal}(r)$ has exactly two elements:*

$$\text{Tal}(r) = \{\theta^-(r), \theta^+(r)\} \text{ and they satisfy } \langle \theta^\pm(r) - x^*, \pm v_1(x^*) \rangle > 0. \quad (\text{A.1})$$

Moreover, the curves $(r^*, \bar{r}) \ni r \rightarrow \theta^\pm(r) \in \mathbb{R}^d$ are of class C^1 and, denoting by $\dot{\theta}^\pm$ their derivatives, we have $\lim_{r \rightarrow f(x^*), r > f(x^*)} \theta^\pm(r) = x^*$,

$$\lim_{r \rightarrow f(x^*), r > f(x^*)} \rho \left(\dot{\theta}^\pm(r) \right) = \pm v_1(x^*), \quad \text{and} \quad \lim_{r \rightarrow f(x^*), r > f(x^*)} |\dot{\theta}^\pm(r)| |\nabla f(\theta^\pm(r))| = 1. \quad (\text{A.2})$$

Proof. By Theorem 2.6, for $\eta > 0$ small enough, there exist $\bar{t} > 0$ and a C^1 and a curve $\gamma: (-\bar{t}, \bar{t}) \rightarrow E_1 \cap [f < \bar{r}]_{x^*}$ satisfying $\gamma(0) = x^*$ and $\dot{\gamma}(0) = v_1(x^*)$. Hence, $\gamma(t) = x^* + tv_1(x^*) + o(t)$ for $t \in (-\bar{t}, \bar{t})$. Let us define $\phi(t) := f(\gamma(t))$ on $(-\bar{t}, \bar{t})$. Since $\nabla f(\gamma(t)) = \nabla^2 f(x^*)(\gamma(t) - x^*) + o(t)$ and $\dot{\phi}(t) = \langle \nabla f(\gamma(t)), \dot{\gamma}(t) \rangle = \langle \nabla f(\gamma(t)), v_1(x^*) + o(1) \rangle$, we deduce that

$$\dot{\phi}(t) = \langle \nabla^2 f(x^*)(\gamma(t) - x^*) + o(t), v_1(x^*) + o(1) \rangle = \lambda_1(x^*)t + o(t), \quad \text{for all } t \in (-\bar{t}, \bar{t}).$$

Reducing \bar{t} , if necessary, we have $\dot{\phi} < 0$ in $(-\bar{t}, 0)$ and $\dot{\phi} > 0$ in $(0, \bar{t})$. In particular, for every $r \in (f(x^*), \min\{\lim_{t \rightarrow -\bar{t}} \phi(t), \lim_{t \rightarrow \bar{t}} \phi(t)\})$ one has the existence of $t^- \in (-\bar{t}, 0)$ and $t^+ \in (0, \bar{t})$ such that $\theta^-(r) := \gamma(t^-)$ and $\theta^+(r) := \gamma(t^+)$ satisfy $\text{Tal}(r) = \{\theta^-(r), \theta^+(r)\}$. Notice that $\langle \theta^+(r) - x^*, v_1(x^*) \rangle = t^+ + o(t^+)$ and $\langle \theta^-(r) - x^*, v_1(x^*) \rangle = t^- + o(t^-)$, and hence, reducing \bar{t} if necessary, we get $\langle \theta^+(r) - x^*, v_1(x^*) \rangle > 0$ and $\langle \theta^-(r) - x^*, v_1(x^*) \rangle < 0$. In conclusion, if $\bar{r} - r^* > 0$ is small enough, (A.1) holds.

Define $F: B(x^*, \eta) \times \mathbb{R} \times (r^*, \bar{r}) \rightarrow \mathbb{R}^d \times \mathbb{R}$ as

$$F(x, \lambda, r) = (\nabla^2 f(x) \nabla f(x) - \lambda \nabla f(x), f(x) - r).$$

Setting $(\theta_r^\pm, \lambda_{1,r}^\pm) := (\theta^\pm(r), \lambda_1(\theta^\pm(r)))$, for all $r \in (r^*, \bar{r})$, one has

$$F(\theta_r^\pm, \lambda_{1,r}^\pm, r) = 0.$$

In view of applying the implicit function theorem, we consider the linearized system of equations, with unknowns (δ_1, δ_2) in $\mathbb{R}^d \times \mathbb{R}$, (θ', λ') in $\mathbb{R}^d \times \mathbb{R}$,

$$\begin{aligned} & (\delta_1, \delta_2) \\ &= D_{(x,\lambda)} F(\theta_r^\pm, \lambda_{1,r}^\pm, r) \cdot (\theta', \lambda') \\ &= \left(D^3 f(\theta_r^\pm) \nabla f(\theta_r^\pm) \theta' + (\nabla^2 f(\theta_r^\pm))^2 \theta' - \lambda_{1,r}^\pm \nabla^2 f(\theta_r^\pm) \theta' - \lambda' \nabla f(\theta_r^\pm), \nabla f(\theta_r^\pm) \cdot \theta' \right). \end{aligned} \quad (\text{A.3})$$

Back to the proof of Theorem 2.6 (see (2.16)), if $\bar{r} - r^* > 0$ is small enough, we have

$$\left\langle \nabla_{xx}^2 L_r(\theta_r^\pm, \lambda_{1,r}^\pm) h, h \right\rangle > 0, \quad \text{for all } h \in \mathbb{R}^d \setminus \{0\} \text{ such that } \langle \nabla f(\theta_r^\pm), h \rangle = 0,$$

which implies that

$$\begin{aligned} \min_{\theta' \in \mathbb{R}^d} \quad & \frac{1}{2} \left\langle \nabla_{xx}^2 L_r(\theta_r^\pm, \lambda_{1,r}^\pm) \theta', \theta' \right\rangle - \delta_1 \cdot \theta' \\ \text{s.t.} \quad & \langle \nabla f(\theta_r^\pm), \theta' \rangle = \delta_2. \end{aligned} \quad (\text{A.4})$$

admits a unique solution $\hat{\theta}$ and, thanks to (2.4), a unique Lagrange multiplier $\hat{\lambda}$. Therefore, by the implicit function theorem, the curves $r \rightarrow \theta_r^\pm$ and $r \rightarrow \lambda_{1,r}^\pm$ are C^1 on (r^*, \bar{r}) , with derivatives $\dot{\theta}_r^\pm$ and $\dot{\lambda}_r^\pm$ satisfying

$$\begin{aligned} \left(D^3 f(\theta_r^\pm) \nabla f(\theta_r^\pm) + (\nabla^2 f(\theta_r^\pm))^2 - \lambda_{1,r}^\pm \nabla^2 f(\theta_r^\pm) \right) \dot{\theta}_r^\pm &= \dot{\lambda}_r^\pm \nabla f(\theta_r^\pm), \\ \langle \nabla f(\theta_r^\pm), \dot{\theta}_r^\pm \rangle &= 1. \end{aligned} \quad (\text{A.5})$$

Notice that, by (2.1) and the continuity of λ_1 , $\theta_r^\pm \rightarrow x^*$. As a consequence, since $\rho(\dot{\theta}_r^\pm) \in T_{\theta_r^\pm} E_1$, one deduces that, as $r \rightarrow r^*$, with $r > r^*$, any cluster point of $\rho(\dot{\theta}_r^\pm)$ belongs to $T_{x^*} E_1 = \mathbb{R}v_1(x^*)$ and hence must belong to $\{-v_1(x^*), v_1(x^*)\}$. On the other hand by Theorem 2.6, the gradient is an eigenvector along the talweg thus $\rho(\nabla f(\theta_r^\pm)) = \mathcal{S}_r^\pm v_1(\theta_r^\pm)$, with $\mathcal{S}_r^\pm \in \{-1, 1\}$. It then follows from the second equation in (A.5) that $\langle v_1(\theta_r^\pm), \dot{\theta}_r^\pm \rangle = \pm 1$ and hence $\rho(\dot{\theta}_r^\pm)$ has only one cluster point given by $\pm v_1(x^*)$, from which the first equality in (A.2) follows. Finally, from the second equation in (A.5), we also have

$$1 = \langle v_1(x^*), v_1(x^*) \rangle = \lim_{r \rightarrow r^*, r > r^*} \left\langle \frac{\nabla f(\theta_r^\pm)}{|\nabla f(\theta_r^\pm)|}, \frac{\dot{\theta}_r^\pm}{|\dot{\theta}_r^\pm|} \right\rangle = \lim_{r \rightarrow r^*, r > r^*} \frac{1}{|\nabla f(\theta_r^\pm)| |\dot{\theta}_r^\pm|},$$

from which the second equality in (A.2) follows. \square

Remark A.2. As a consequence of the second relation in (A.2), one can provide an estimate for the blow-up of $|\dot{\theta}^\pm(r)|$ as $r \rightarrow f(x^*)$, with $r > f(x^*)$. Indeed, since $\nabla f(\theta^\pm(r)) = \nabla^2 f(x^*)(\theta^\pm(r) - x^*) + o(|\theta^\pm(r) - x^*|)$, if η is small enough, (A.2) and (2.1) imply that

$$|\dot{\theta}^\pm(r)| \geq \frac{1}{2\lambda_d(x^*)|\theta^\pm(r) - x^*|} \geq \frac{1}{2\lambda_d(x^*)\sqrt{r - r^*}}, \quad \text{for all } r \in (r^*, \bar{r}).$$

Observe that this estimate can be combined with the results of [4], to show the existence of $c > 0$ such that

$$|\nabla \sqrt{f - r^*}|(x) \geq c \quad \text{over} \quad [f < \bar{r}]_{x^*} \setminus \{x^*\}.$$

Lemma A.3 (Balls are absorbing). *Under Assumption 2.2, if $\eta > 0$ is small enough, then $\Phi_t(B(x^*, \eta)) \subset B(x^*, \eta)$, for all $t \in [0, \infty)$.*

Proof. Just observe that if $\eta > 0$ is small enough $f: B(x^*, \eta) \rightarrow \mathbb{R}$ is convex, as $\lambda_1 > 0$ on $B(x^*, \eta)$. It suffices then observe that $|\Phi_t(x) - x^*|$ is decreasing as $\frac{d}{dt} |\Phi_t(x) - x^*|^2 = 2\langle \nabla f(x^*) - \nabla f(\Phi_t(x)), \Phi_t(x) - x^* \rangle \leq 0$. \square

Denote by $\mathcal{D}_d(\mathbb{R})$ (respectively, $\text{O}_d(\mathbb{R})$, $\mathcal{S}_d(\mathbb{R})$) the set of diagonal (respectively, orthogonal, symmetric) matrices. We have:

Lemma A.4 (Parametric diagonalization). *Let $F: \text{O}_d(\mathbb{R}) \times \mathcal{D}_d(\mathbb{R}) \rightarrow \mathcal{S}_d(\mathbb{R})$ be defined by $F(P, \mathcal{D}) = P\mathcal{D}P^\top$ and let $(P_0, \mathcal{D}_0) \in \text{O}_d(\mathbb{R}) \times \mathcal{D}_d(\mathbb{R})$ be such that for any $i \neq j$, $(\mathcal{D}_0)_{ii} \neq (\mathcal{D}_0)_{jj}$. Then there exist two open sets $U \subset \text{O}_d(\mathbb{R}) \times \mathcal{D}_d(\mathbb{R})$ containing (P_0, \mathcal{D}_0) and $V \subset \mathcal{S}_d(\mathbb{R})$ such that $F|_U: U \rightarrow V$ is a C^∞ -diffeomorphism.*

Proof. Let $G : \mathcal{M}_d(\mathbb{R}) \times \mathcal{M}_d(\mathbb{R}) \rightarrow \mathcal{M}_d(\mathbb{R})$ be defined by $G(P, \mathcal{D}) = P\mathcal{D}P^\top$. The function G is of class C^∞ and, for every $(M, \Delta) \in \mathcal{M}_d(\mathbb{R}) \times \mathcal{M}_d(\mathbb{R})$,

$$DG(P_0, \mathcal{D}_0)(M, \Delta) = M\mathcal{D}_0P_0^\top + P_0\mathcal{D}_0M^\top + P_0\Delta P_0^\top.$$

It follows that $F = G|_{\mathcal{O}_d(\mathbb{R}) \times \mathcal{D}_d(\mathbb{R})}$ is differentiable at (P_0, \mathcal{D}_0) . Take an element H of the tangent space $T_{P_0}\mathcal{O}_d(\mathbb{R})$ at P_0 . Then $H = P_0A$ for some antisymmetric matrix A . Let $\Delta \in \mathcal{D}_d(\mathbb{R})$. From above, we get

$$DF(P_0, \mathcal{D}_0)(P_0A, \Delta) = P_0(A\mathcal{D}_0 - \mathcal{D}_0A + \Delta)P_0^\top.$$

If $DF(P_0, \mathcal{D}_0)(P_0A, \Delta) = 0$ then $A\mathcal{D}_0 - \mathcal{D}_0A + \Delta = 0$. Since, for any $i \neq j$,

$$\begin{aligned} (A\mathcal{D}_0 - \mathcal{D}_0A + \Delta)_{ij} &= (A\mathcal{D}_0 - \mathcal{D}_0A)_{ij} \\ &= \sum_{k=1}^d (A_{ik}(\mathcal{D}_0)_{kj} - (\mathcal{D}_0)_{ik}A_{kj}) \\ &= A_{ij}(\mathcal{D}_0)_{jj} - (\mathcal{D}_0)_{ii}A_{ij} \\ &= A_{ij}((\mathcal{D}_0)_{jj} - (\mathcal{D}_0)_{ii}) \end{aligned}$$

and $(\mathcal{D}_0)_{jj} \neq (\mathcal{D}_0)_{ii}$, it follows that $A_{ij} = 0$. But A is antisymmetric so that $A_{ii} = 0$. Hence $A = 0$. This yields, $0 = A\mathcal{D}_0 - \mathcal{D}_0A + \Delta = \Delta$. Hence $DF(P_0, \mathcal{D}_0)$ is injective and since $\dim(\mathcal{O}_d(\mathbb{R}) \times \mathcal{D}_d(\mathbb{R})) = \frac{d(d-1)}{2} + d = \frac{d(d+1)}{2} = \dim \mathcal{S}_d(\mathbb{R})$, $DF(P_0, \mathcal{D}_0)$ is an isomorphism. We conclude with the inverse function theorem. \square

Corollary A.5 (Parametric diagonalization of Hessian). *Under Assumption 2.1, let $(P^*, \mathcal{D}^*) \in \mathcal{O}_d(\mathbb{R}) \times \mathcal{D}_d(\mathbb{R})$ be such that $\nabla^2 f(x^*) = P^* \mathcal{D}^* (P^*)^\top$. Then there exists $\eta > 0$, and a unique couple $(P, \mathcal{D}) \in C^1(B(x^*, \eta); \mathcal{O}_d(\mathbb{R})) \times C^1(B(x^*, \eta); \mathcal{D}_d(\mathbb{R}))$ such that $(P(x^*), \mathcal{D}(x^*)) = (P^*, \mathcal{D}^*)$ and*

$$\nabla^2 f(x) = P(x)\mathcal{D}(x)P(x)^\top, \quad \text{for all } x \in B(x^*, \eta). \quad (\text{A.6})$$

Proof. Let F be defined as in Lemma A.4. Let $U \ni (P^*, \mathcal{D}^*)$ and V be given by Lemma A.4. Since $F(P^*, \mathcal{D}^*) = P^* \mathcal{D}^* (P^*)^\top = \nabla^2 f(x^*) \in V$ and $f \in C^3(\mathbb{R}^d; \mathbb{R})$, we may choose $\eta > 0$ small enough to have $\nabla^2 f(x) \in V$, for any $x \in B(x^*, \eta)$. By Lemma A.4, F is a C^∞ -diffeomorphism from U onto V . We infer that for every $x \in B(x^*, \eta)$, there exist unique $P(x) \in \mathcal{O}_d$ and $\mathcal{D}(x) \in \mathcal{D}_d(\mathbb{R})$ with $(P(x), \mathcal{D}(x)) \in U$ such that

$$\nabla^2 f(x) = F(P(x), \mathcal{D}(x)) = P(x)\mathcal{D}(x)P(x)^\top.$$

In particular, $(P(x^*), \mathcal{D}(x^*)) = (P^*, \mathcal{D}^*)$ and for any $x \in B(x^*, \eta)$, $(P(x), \mathcal{D}(x)) = F^{-1}(\nabla^2 f(x))$. Since $f \in C^3(\mathbb{R}^d; \mathbb{R})$ and F^{-1} is C^∞ , it follows that P and \mathcal{D} are C^1 on $B(x^*, \eta)$. \square

A.2 Some proofs

Proof of Remark 3.4. For $x(t) = (x_1(t), x_2(t))$ the gradient flow reads

$$\dot{x}_1 = (-\lambda_1 - 2ax_2)x_1, \quad \dot{x}_2 = -\lambda_2 x_2 - ax_1^2. \quad (\text{A.7})$$

From the first equation and $x(0) = (\alpha_1, \alpha_2)$, we have

$$x_1(t) = \alpha_1 \exp\left(-\lambda_1 t - 2a \int_0^t x_2(s) ds\right).$$

By Theorem 3.2, if $|x(0)| = |(\alpha_1, \alpha_2)|$ is small enough, then the trajectory converges exponentially to 0, and hence the function x_2 is integrable on $[0, \infty)$. Therefore, $L := \int_0^\infty x_2(s) ds$ is well defined. Setting $C := x_1(0) \exp(-2aL) > 0$, we may rewrite the above identity as

$$x_1(t) = C \exp(-\lambda_1 t) \exp\left(2a \int_t^\infty x_2(s) ds\right).$$

Substituting this expression into the second equation in (A.7) yields

$$\dot{x}_2 + \lambda_2 x_2 = -aC^2 \exp(-2\lambda_1 t) \exp\left(4a \int_t^\infty x_2(s) ds\right).$$

By variation of constants,

$$x_2(t) = \exp(-\lambda_2 t) \alpha_2 - aC^2 \int_0^t \exp(-\lambda_2(t-s)) \exp(-2\lambda_1 s) \exp\left(4a \int_s^\infty x_2(u) du\right) ds.$$

We split this as

$$x_2(t) = \exp(-\lambda_2 t) \alpha_2 - aC^2 \int_0^t \exp(-\lambda_2(t-s)) \exp(-2\lambda_1 s) ds + R(t),$$

where

$$R(t) := -aC^2 \int_0^t \exp(-\lambda_2(t-s)) \exp(-2\lambda_1 s) \left[\exp\left(4a \int_s^\infty x_2(u) du\right) - 1 \right] ds. \quad (\text{A.8})$$

Since $\lambda_2 > 2\lambda_1$, we have

$$\int_0^t \exp(-\lambda_2(t-s)) \exp(-2\lambda_1 s) ds = \frac{\exp(-2\lambda_1 t) - \exp(-\lambda_2 t)}{\lambda_2 - 2\lambda_1}.$$

Hence $x_2(t) = -\frac{aC^2}{\lambda_2 - 2\lambda_1} \exp(-2\lambda_1 t) + \tilde{C} \exp(-\lambda_2 t) + R(t)$ for some constant $\tilde{C} \in \mathbb{R}$.

It remains to estimate $R(t)$. Since $\int_0^\infty |x_2(u)| du < \infty$, by the mean value theorem,

$$\left| \exp\left(4a \int_s^\infty x_2(u) du\right) - 1 \right| = O\left(\left| \int_s^\infty x_2(u) du \right|\right) = O(\exp(-\lambda_1 s))$$

and hence, there exists $\hat{C} > 0$ such that

$$|R(t)| \leq \hat{C} \exp(-\lambda_2 t) \int_0^t \exp((\lambda_2 - 3\lambda_1)s) ds = o(\exp(-2\lambda_1 t)).$$

Thus, we have

$$x_2(t) = -\frac{aC^2}{\lambda_2 - 2\lambda_1} \exp(-2\lambda_1 t) + o(\exp(-2\lambda_1 t))$$

and, consequently,

$$\frac{x_2(t)}{x_1(t)} = -\frac{aC}{\lambda_2 - 2\lambda_1} \exp(-\lambda_1 t) + o(\exp(-\lambda_1 t)). \quad (\text{A.9})$$

Finally, since $x_2(t)/x_1(t) \rightarrow 0$ as $t \rightarrow \infty$, (A.9) implies

$$\begin{aligned} \text{dist}(\rho(x(t)), T_0 \text{Ext}_1) &= \frac{|x_2(t)|}{|x(t)|} = \frac{|x_2(t)|}{x_1(t)} + o\left(\frac{|x_2(t)|}{x_1(t)}\right) \\ &= \frac{aC}{\lambda_2 - 2\lambda_1} \exp(-\lambda_1 t) + o(\exp(-\lambda_1 t)), \end{aligned}$$

which shows (3.27). \square

Proof of Remark 4.7. Set $\mu_i = 1 - \gamma\lambda_i \in (0, 1)$ for $i = 1, 2$ and notice that, since $\lambda_2 > 2\lambda_1$, one has $\mu_1^2 > \mu_2$. System (4.27) takes the form

$$\begin{aligned} x_{1,k+1} &= (\mu_1 - 2a\gamma x_{2,k})x_{1,k}, \\ x_{2,k+1} &= \mu_2 x_{2,k} - a\gamma x_{1,k}^2. \end{aligned} \quad (\text{A.10})$$

On the one hand, let us show the existence of $\bar{\alpha}_1 \neq 0$ such that, as $k \rightarrow \infty$,

$$x_{1,k} = \bar{\alpha}_1 \mu_1^k + o\left(\mu_1^k\right). \quad (\text{A.11})$$

Indeed, by the first equation in (4.27), for every $k \geq 1$, one has

$$x_{1,k} = \alpha_1 \mu_1^k \prod_{i=0}^{k-1} (1 - 2a\gamma x_{2,i}/\mu_1).$$

From (4.16), if $|x_0|$ is small enough, there exists $c > 0$ such that $|x_{2,i}| \leq c\mu_1^i$ for all $i \geq 0$, and hence $\sum_{i=0}^{\infty} |x_{2,i}| < \infty$, which implies that

$$0 < P = \prod_{i=0}^{\infty} (1 - 2a\gamma x_{2,i}/\mu_1) < \infty.$$

In turn, (A.11) holds with $\bar{\alpha} := \alpha_1 P > 0$. On the other hand, let us show that, as $k \rightarrow \infty$,

$$x_{2,k} = -\frac{a\gamma\bar{\alpha}_1^2}{\mu_1^2 - \mu_2} \mu_1^{2k} + o(\mu_1^{2k}). \quad (\text{A.12})$$

Setting $y_k = x_{2,k}/\mu_1^{2k}$, (A.10), and (A.11) yield

$$(\forall k \geq 0) \quad y_{k+1} = \frac{\mu_2}{\mu_1^2} y_k - \frac{a\gamma x_{1,k}^2}{\mu_1^{2(k+1)}} = qy_k + b + \varepsilon_k,$$

where $q = \frac{\mu_2}{\mu_1^2} \in (0, 1)$, $b = -a\gamma\bar{\alpha}_1^2/\mu_1^2$, and $\varepsilon_k \rightarrow 0$. Letting $\bar{y} = b/(1 - q) = -a\gamma\bar{\alpha}_1^2/(\mu_1^2 - \mu_2)$ and $\eta_k = y_k - \bar{y}$, we have

$$(\forall k \geq 0) \quad \eta_{k+1} = q\eta_k + \varepsilon_k = q^{k+1}\eta_0 + \sum_{i=0}^k q^{k-i}\varepsilon_i. \quad (\text{A.13})$$

Let $\delta > 0$ and pick $N \in \mathbb{N}$ such that $|\varepsilon_i| \leq \delta$ for all $i \geq N$. Then, for $k > N$,

$$\begin{aligned} \left| \sum_{i=0}^k q^{k-i}\varepsilon_i \right| &= \left| \sum_{i=0}^{N-1} q^{k-i}\varepsilon_i + \sum_{i=N}^k q^{k-i}\varepsilon_i \right| \\ &\leq q^k \left| \sum_{i=0}^{N-1} q^{-i}\varepsilon_i \right| + \delta \sum_{i=N}^k q^{k-i} \leq q^k \left| \sum_{i=0}^{N-1} q^{-i}\varepsilon_i \right| + \delta \sum_{i=0}^{\infty} q^i. \end{aligned}$$

Using that $q \in (0, 1)$, we get that $\limsup_{k \rightarrow \infty} \left| \sum_{i=0}^k q^{k-i}\varepsilon_i \right| \leq \delta \sum_{i=0}^{\infty} q^i$, and, since $\sum_{i=0}^{\infty} q^i < \infty$ and δ is arbitrary, we deduce that $\sum_{i=0}^k q^{k-i}\varepsilon_i \rightarrow 0$ and, hence, by (A.13), $\eta_k \rightarrow 0$, which implies that

$$x_{2,k} = \bar{y}\mu_1^{2k} + o(\mu_1^{2k}),$$

from which (A.12) follows.

Finally, it follows from (A.11) and (A.12) that

$$x_k = \left(\bar{\alpha}_1 \mu_1^k + o(\mu_1^k) \right) e_1 + \left(-\frac{a\gamma\bar{\alpha}_1^2}{\mu_1^2 - \mu_2} \mu_1^{2k} + o(\mu_1^{2k}) \right) e_2,$$

which yields $|x_k| = |\bar{\alpha}_1| \mu_1^k + o(\mu_1^k)$, and hence

$$\text{dist}(\rho(x_k), T_0 \text{Ext}_1) = \frac{|x_{2,k}|}{|x_k|} = \frac{|a|\gamma\alpha_1^2}{|\bar{\alpha}_1|(\mu_1^2 - \mu_2)} \mu_1^k + o(\mu_1^k),$$

which shows (4.28). \square

References

- [1] G. Allaire and S. M. Kaber. *Numerical Linear Algebra*, volume 55 of *Texts Appl. Math.* Springer, New York, 2008. Translated from the 2002 French original by K. Trabelsi.
- [2] F. Bach. *Learning Theory from First Principles*. MIT Press, 2024.
- [3] M-A. Belabbas. On gradient flows initialized near maxima. *SIAM J. Control Optim.*, 61(5):2826–2848, 2023.
- [4] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362(6):3319–3363, 2010.
- [5] M.J. Boussinesq. Sur une propriété remarquable des points où les lignes de plus grande pente d’une surface ont leurs plans osculateurs verticaux, et sur la différence qui existe généralement, à la surface de la terre, entre les lignes de faite ou de thalweg et celles le long desquelles la pente du sol est un minimum. *C. R. H. Acad. Sci.*, 73:1368—1371, 1871.
- [6] A. Cayley. On contour and slope lines. *London Edinburgh Dublin Philos. Mag. J. Sci. Ser. 4*, 18:264–268, 1859.
- [7] D. D’Acunto. *Sur les courbes intégrales du champ de gradient*. PhD thesis, Université de Savoie, 2001.
- [8] D. D’Acunto and K. Kurdyka. Explicit bounds for the Lojasiewicz exponent in the gradient inequality for polynomials. *Ann. Polon. Math.*, 87(1):51–61, 2005.
- [9] M. de Saint-Venant. Surfaces à plus grande pente constituées sur des lignes courbes. *Extraits des Procès-Verbaux des Séances de la Société Philomathique de Paris, Ser. 5*, 17:24–30, 1852.
- [10] A. V. Fiacco and G. P. McCormick. *Nonlinear Programming*, volume 4 of *Classics Appl. Math.* SIAM, Philadelphia, PA, 2 edition, 1990. Sequential unconstrained minimization techniques.
- [11] I. M. Gelfand and M. L. Tsetlin. On the principle of nonlocal search in automatic optimization systems. *Dokl. Akad. Nauk SSSR*, 137:295–298, 1961. In Russian.
- [12] M. Golubitsky and V. Guillemin. *Stable mappings and their singularities*, volume Vol. 14 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Heidelberg, 1973.

- [13] X. Goudou and J. Munier. The gradient and heavy ball with friction dynamical systems: the quasiconvex case. *Math. Program.*, 116(1–2):173–191, 2009.
- [14] P. Hartman. On local homeomorphisms of Euclidean spaces. *Bol. Soc. Mat. Mexicana*, 5:220–241, 1960.
- [15] P. Hartman. *Ordinary Differential Equations*, volume 38 of *Classics Appl. Math.* SIAM, Philadelphia, PA, 2002. Corrected reprint of the second (1982) edition.
- [16] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier*, 48(3):769–783, 1998.
- [17] K. Kurdyka, T. Mostowski, and A. Parusiński. Proof of the gradient conjecture of R. Thom. *Ann. of Math.*, 152(3):763–792, 2000.
- [18] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *Proc. 29th Annual Conf. Learn. Theory*, volume 49 of *Proc. Mach. Learn. Res.*, pages 1246–1257. PMLR, 2016.
- [19] S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.
- [20] S. Lojasiewicz. Sur les trajectoires du gradient d’une fonction analytique. *Semin. Geom.*, 1983(1984):115–117, 1982.
- [21] J.C. Maxwell. On hills and dales. *London Edinburgh Dublin Philos. Mag. J. Sci. Ser. 4*, 40:421–427, 1870.
- [22] R. Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *Ann. Probab.*, 18(2):698–712, 1990.
- [23] F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.
- [24] H. B. Schlegel. Following gradient extremal paths. *Theoret. Chim. Acta*, 83(1):15–20, 1992.
- [25] S. Sternberg. Local contractions and a theorem of Poincaré. *Amer. J. Math.*, 79:809–824, 1957.
- [26] R. Thom. Sur une partition en cellules associée à une fonction sur une variété. *C. R. Acad. Sci. Paris*, 228(12):973–975, 1949.
- [27] M. Viana and J. M. Espinar. *Differential Equations—A Dynamical Systems Approach to Theory and Practice*, volume 212 of *Grad. Stud. Math.* Amer. Math. Soc., Providence, RI, 2021.