Long term dynamics of the subgradient method for Lipschitz path differentiable functions

Jérôme Bolte, Edouard Pauwels, Rodolfo Ríos-Zertuche

Long term dynamics of the subgradient method for Lipschitz path differentiable functions

Jérôme Bolte, Edouard Pauwels, and Rodolfo Ríos-Zertuche

January 24, 2023

Abstract

We consider the long-term dynamics of the vanishing stepsize subgradient method in the case when the objective function is neither smooth nor convex. We assume that this function is locally Lipschitz and path differentiable, i.e., admits a chain rule. Our study departs from other works in the sense that we focus on the behavior of the oscillations, and to do this we use closed measures, a concept that complements the technique of asymptotic pseudotrajectories developed in this setting by Benaïm–Hofbauer–Sorin. We recover known convergence results, establish new ones, and show a local principle of oscillation compensation for the velocities. Roughly speaking, the time average of gradients *around* one limit point vanishes. Various cases are discussed providing new insight into the oscillation and the stabilization phenomena.

Contents

1 Introduction

The predominance of huge scale complex nonsmooth nonconvex problems in the development of certain artificial intelligence methods, has brought back rudimentary, numerically cheap, robust methods, such as subgradient algorithms, to the forefront of contemporary numerics, see e.g., [6, 13, 27, 36, 37]. We investigate here some of the properties of the archetypical algorithm within this class, namely, the vanishing stepsize subgradient method of Shor. Given $f : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz, it reads

$$x_{i+1} \in x_i - \varepsilon_i \partial^c f(x_i), \;\; x_0 \in \mathbb{R}^n,$$

where $\partial^c f$ is the Clarke subgradient, $\varepsilon_i \to 0$, and $\sum_{i=0}^{\infty} \varepsilon_i = \infty$. This dynamics, illustrated in Figure 1, has its roots in Cauchy's gradient method and seems to originate in Shor's thesis [52]. The idea is natural at first sight: one accumulates small subgradient steps to make good progress on average while hoping that oscillations will be tempered by the vanishing steps. For the convex case, the theory was developed by Ermol'ev [29], Poljak [47], Ermol'ev–Shor [30]. It is a quite mature theory, see e.g. [43, 44], which still has a considerable success through the famous mirror descent of Nemirovskii–Yudin [8, 43] and its numerous variants. In the nonconvex case, developments of more sophisticated methods were made, see e.g. [35, 38, 45], yet little was known for the raw method until recently.

The work of Davis et al. [25], see also [12], revolving around the fundamental paper of Benaïm–Hofbauer–Sorin [9], brought the first breakthroughs. It relies on a classical idea of Euler: small-step discrete dynamics resemble their continuous counterparts. As established by Ljung [40], this observation can be made rigorous for large times in the presence of good Lyapunov functions. Benaïm–Hofbauer–Sorin [9] showed further that the transfer of asymptotic properties from continuous differential inclusions to small-step discrete methods is valid under rather weak compactness and dissipativity assumptions. This general result, combined with features specific to the subgradient case, allowed to establish several optimization results such as the convergence to the set of critical points, the convergence in value, convergence in the long run in the presence of noise [13, 16, 25, 51].

Usual properties expected from an algorithm are diverse: convergence of iterates, convergence in values, rates, quality of optimality, complexity, or prevalence of minimizers. Although in our setting some aspects seem hopeless without strong assumptions, most of them remain largely unexplored. Numerical successes suggest however that the apparently erratic process of subgradient dynamics has appealing stability properties beyond the already delicate subsequential convergence to critical points.

In order to address some of these issues, this paper avoids the use of the theory of [9] and focuses on the delicate question of oscillations,[1] which is illustrated on Figures 1 and 2.

In general, as long as the sequence $\{x_i\}_i$ remains bounded and satisfies $x_{i+1} - x_i = \varepsilon_i v_i$ for some vectors v_i and positive scalars ε_i satisfying $\sum_{i=0}^{\infty} \varepsilon_i = +\infty$, we always have

$$\frac{\sum_{i=0}^{N} \varepsilon_i v_i}{\sum_{i=0}^{N} \varepsilon_i} = \frac{x_N - x_0}{\sum_{i=0}^{N} \varepsilon_i} \to 0 \quad \text{as} \quad i \to +\infty. \tag{1}$$

This fact, that could be called "global oscillation compensation," does not prevent the trajectory to oscillate fast around a limit cycle, as illustrated in [24], and is therefore unsatisfying from the stabilization perspective of minimization. The phenomenon (1) remains true even when $\{x_i\}_i$ is not a gradient sequence, as in the case of discrete game theoretical dynamical systems [9].

[1] In Figure 1, we see the sequence descending along a ridge. The jumps $x_{i+1} - x_i$ can be decomposed into two components, one that is parallel to the "ridge" and another one that is perpendicular to it; we will informally refer to these components, respectively, as the *drift* and the *bouncing*, without attempting to define these concepts formally, (see however the discussion before Lemma 8). Similarly we shall often use the term oscillations to evoke notable and persistent variations of the directional term $x_{i+1} - x_i/\|x_{i+1} - x_i\|$ over time.

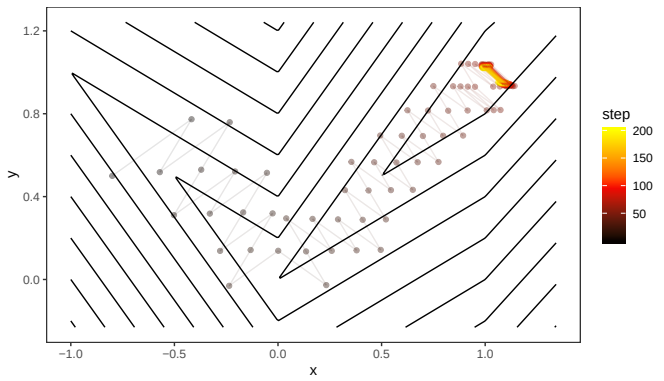

Figure 1: Contour plot of a Lipschitz function with a subgradient sequence. The color reflects the iteration count. The sequence converges to the unique global minimum, but is constantly oscillating.

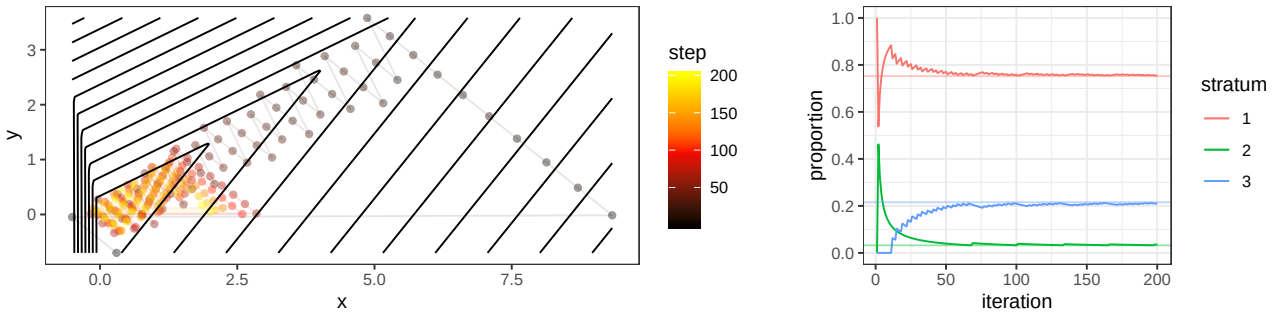

Figure 2: On the left, the contour plot of a convex polyhedral function with three strata, where the gradient is constant. A subgradient sequence starts at $(0.3, -0.7)$ and converges to the origin with an apparent erratic behavior. On the right, we discover that the behavior is not completely erratic. The oscillation compensation phenomenon contributes some structure: the proportions λ_i of time spent in each region where the function has constant gradient g_i, $i = 1, 2, 3$, converge so that we have precisely $\lambda_1 g_1 + \lambda_2 g_2 + \lambda_3 g_3 = 0$.

In this work, we adapt the theory of closed measures, which was originally developed in the calculus of variations (see for example [5, 10]), to the study of discrete dynamics. Using it, we establish several local oscillation compensation results for path differentiable functions. Morally, our results in this direction say that for limit points x we have

$$\text{“} \lim_{\substack{\delta \searrow 0 \\ N \to +\infty}} \frac{\displaystyle\sum_{\substack{0 \leqslant i \leqslant N \\ \|x - x_i\| \leqslant \delta}} \varepsilon_i v_i}{\displaystyle\sum_{\substack{0 \leqslant i \leqslant N \\ \|x - x_i\| \leqslant \delta}} \varepsilon_i} = 0 \text{”} \tag{2}$$

See Theorems 4 and 5 for precise statements, and a discussion in Section 3.3.

While this does not imply the convergence of $\{x_i\}_i$, it does mean that the drift emanating from the average velocity of the sequence vanishes as time elapses. This is made more explicit in the parts of those theorems that show that, given two distinct limit points x and y of the sequence $\{x_i\}_i$, the time it takes for the sequence to flow from a small ball around x to a small ball around y must eventually grow infinitely long, so that the overall average speed of the sequence as it traverses the accumulation

set becomes extremely slow.

With these types of results, we evidence new phenomena:

— while the sequence may not converge, it will spend most of the time oscillating near the critical set of the objective function, and it appears that there are persistent accumulation points whose importance is predominant;

— under weak Sard assumptions, we recover the convergence results of [25] and improve them by oscillation compensations results,

— oscillation structures itself orthogonally to the limit set, so that the incremental drift along this set is negligible with respect to the time increment ε_i.

These results are made possible by the use of closed measures. These measures capture the accumulation behavior of the sequence $\{x_i\}_i$ along with the "velocities" $\{v_i\}_i$. The simple idea of not throwing away the information of the vectors v_i allows one to recover a lot of structure in the limit, that can be interpreted as a portrait of the long-term behavior of the sequence. The theory that we develop in Section 4.1 should apply to the analysis of the more general case of small-step algorithms. Along the way, for example, we are able to establish a new connection between the discrete and continuous gradient flows (Proposition 18) that complements the point of view of [9].

Notations and organization of the paper. Let n be a positive integer, and $\mathbb{R}^n$ denote n-dimensional Euclidean space. The space $\mathbb{R}^n \times \mathbb{R}^n$ of couples (x, v) is seen as the phase space consisting of positions $x \in \mathbb{R}^n$ and velocities $v \in \mathbb{R}^n$. For two vectors $u = (u_1, \dots, u_n)$ and $v = (v_1, \dots, v_n)$, we let $u \cdot v = \sum_{i=1}^n u_i v_i$. The norm $\|v\| = \sqrt{v \cdot v}$ induces the distance $\operatorname{dist}(x, y) = \|x - y\|$, and similarly on $\mathbb{R}^n \times \mathbb{R}^n$. An open ball of center x and radius r is denoted $B(x, r)$. The Euclidean gradient of f is denoted by $\nabla f(x)$. The set $\mathbb{N}$ contains all the nonnegative integers.

In Section 2 we give the definitions necessary to state our results, which we do in Section 3. The proofs of our results will be given in Section 5. Before we broach those arguments, we need to develop some preliminaries regarding our main tool, the so-called closed measures; we do this in Section 4.

2 Algorithm and framework

2.1 The vanishing step subgradient method

Consider a locally Lipschitz functions $f : \mathbb{R}^n \to \mathbb{R}$, denote by $\operatorname{Reg} f$ the set of its differentiability points which is dense by Rademacher's theorem (see for example [31, Theorem 3.2]). The *Clarke subdifferential* of f is defined by

$$\partial^c f(x) = \operatorname{conv}\left\{v \in \mathbb{R}^n : \text{ there is a sequence } \{y_k\}_k \subset \operatorname{Reg} f \text{ with } y_k \to x \text{ and } \nabla f(y_k) \to v\right\}$$

where $\operatorname{conv} S$ denotes the closed convex envelope of a set $S \subset \mathbb{R}^n$; see [22]. A point x such that $0 \in \partial^c f(x)$, is called *critical*. The *critical set* is

$$\operatorname{crit} f = \{x \in \mathbb{R}^n : 0 \in \partial^c f(x)\}.$$

It contains local minima and maxima. The algorithm of interest in this work is:

Definition 1 (Small step subgradient method)**.** Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and $\{\varepsilon_i\}_{i \in \mathbb{N}}$ be a sequence of positive step sizes such that

$$\sum_{i=0}^{\infty} \varepsilon_i = +\infty \quad \text{and} \quad \varepsilon_i \searrow 0. \tag{3}$$

Given $x_0 \in \mathbb{R}^n$, consider the recursion, for $i \geqslant 0$,

$$x_{i+1} = x_i - \varepsilon_i v_i, \qquad v_i \in \partial^c f(x_i).$$

Here, v_i is chosen freely among $\partial^c f(x_i)$. The sequence $\{x_i\}_{i\in\mathbb{N}}$ is called a *subgradient sequence*.

In what follows the sequence ε_i is interpreted as a sequence of time increments, and it naturally defines a time counter through the formula:

$$t_i = \sum_{j=0}^{i} \varepsilon_j$$

so that $t_i \to \infty$ as $i \to \infty$. Given a sequence $\{x_i\}_i$ and a subset $U \subseteq \mathbb{R}^n$, we set

$$t_i(U) = \sum_{x_j \in U,\, j \leqslant i} \varepsilon_j,$$

which corresponds to the time spent by the sequence in U between times 0 and t_i.

When f has a locally Lipschitz continuous gradient, bounded sequences are eventually descent sequences, i.e. $f(x_i)$ is nonincreasing and they approach the critical set, see e.g. [1] and references therein. When f is nonsmooth, the descent property does not hold anymore and oscillations appear both in values $f(x_i)$ and in space x_i. The objective of this article is precisely to study these oscillations.

Recall that the *accumulation set* $\operatorname{acc}\{x_i\}_i$ of the sequence $\{x_i\}_i$ is the set of points $x \in \mathbb{R}^n$ such that, for every neighborhood U of x, the intersection $U \cap \{x_i\}_i$ is an infinite set. Its elements are known as *limit points*.

If the sequence $\{x_i\}_i$ is bounded and comes from the subgradient method as in Definition 1, then $\|x_i - x_{i+1}\| \to 0$ because $\varepsilon_i \to 0$ and $\partial^c f$ is locally bounded by local Lipschitz continuity of f, so $\operatorname{acc}\{x_i\}_i$ is compact and connected, see e.g., [17].

Accumulation points are the manifestation of recurrent behaviors of the sequence but the frequency of the recurrence is ignored. In the presence of a time counter, here $\{t_i\}_i$, this persistence phenomenon may be measured through presence duration in the neighborhood of a recurrent point. This idea is formalized in the following definition:

Definition 2 (Essential accumulation set)**.** Given a step size sequence $\{\varepsilon_i\}_i \subset \mathbb{R}_{\geqslant 0}$ and a subgradient sequence $\{x_i\}_i \subset \mathbb{R}^n$ as in Definition 1, the *essential accumulation set* $\operatorname{ess\,acc}\{x_i\}_i$ is the set of points $x \in \mathbb{R}^n$ such that, for every neighborhood $U \subseteq \mathbb{R}^n$ of x,

$$\limsup_{N\to+\infty} \frac{\displaystyle\sum_{\substack{1\leqslant i\leqslant N\\ x_i\in U}} \varepsilon_i}{\displaystyle\sum_{1\leqslant i\leqslant N} \varepsilon_i} > 0, \quad \text{that is,} \quad \limsup_{N\to+\infty} \frac{t_N(U)}{t_N} > 0.$$

Analogously, considering the increments $\{v_i\}_i \subset \mathbb{R}^n$, we say that the point (x, w) is in the *essential accumulation set* $\operatorname{ess\,acc}\{(x_i, v_i)\}_i$ if for every neighborhood $U \subset \mathbb{R}^n \times \mathbb{R}^n$ of (x, w) satisfies

$$\limsup_{N\to+\infty} \frac{\displaystyle\sum_{\substack{1\leqslant i\leqslant N\\ (x_i,v_i)\in U}} \varepsilon_i}{\displaystyle\sum_{1\leqslant i\leqslant N} \varepsilon_i} > 0.$$

As explained previously, the set $\operatorname{ess\,acc}\{x_i\}_i$ encodes significantly recurrent behavior; it ignores sporadic escapades of the sequence $\{x_i\}_i$. Essential accumulation points are accumulation points but the converse is not true. If the sequence $\{x_i\}_i$ is bounded, $\operatorname{ess\,acc}\{x_i\}_i$ is nonempty and compact, but not necessarily connected.

2.2 Regularity assumptions on the objective function

Lipchitz continuity and pathologies. Recall that, given a locally Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}$, a *subgradient curve* is an absolutely continuous curve satisfying,

$$\gamma'(t) \in -\partial^c f(\gamma(t)), \text{ a.e. on } (0, +\infty) \text{ and } \gamma(0) = x_0.$$

By general results these curves exist, see e.g., [9] and references therein. In our context they embody the ideal behavior we could hope from subgradient sequences.

First let us recall that pathological Lipschitz functions are generic in the Baire sense, as established in [19, 56]. In particular, generic 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$ satisfy $\partial^c f \equiv [-1, 1]$ everywhere on $\mathbb{R}$. This means that any absolutely continuous curve $\gamma : \mathbb{R} \to \mathbb{R}$ with $\|\gamma'\| \leqslant 1$ is a subgradient curve of these functions, regardless of their specifics. Note that this implies that a curve may constantly remain away from the critical set.

The examples by Danillidis–Drusvyatskiy [24] make this erratic behaviour even more concrete. For instance, they provide a Lipschitz function $f\colon \mathbb{R}^2 \to \mathbb{R}$ and a bounded subgradient curve γ having the "absurd" roller coaster property

$$(f \circ \gamma)(t) = \sin t, \quad t \in \mathbb{R}.$$

Although not directly matching our framework, these examples show that we cannot hope for satisfying convergence results under the spineless general assumption of Lipschitz continuity.

Path differentiability. We are thus led to consider functions avoiding pathologies. We choose to pertain to the *fonctions saines*[2] of Valadier [55] (1989), rediscovered in several works, see e.g. [16, 18, 25]. We use the terminology of [16]; note however that the equivalent definition proposed in [16] (see [16, Corollary 2]) is not limited to chain rules involving the Clarke subgradient.

Definition 3 (Path differentiable functions)**.** A locally Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}$ is *path differentiable* if, for each Lipschitz curve $\gamma\colon \mathbb{R} \to \mathbb{R}^n$, for almost every $t \in \mathbb{R}$, the composition $f \circ \gamma$ is differentiable at t and the derivative is given by

$$(f \circ \gamma)'(t) = v \cdot \gamma'(t) \text{ for all } v \in \partial^c f(\gamma(t)).$$

In other words, all vectors in $\partial^c f(\gamma(t))$ share the same projection onto the subspace generated by $\gamma'(t)$.

Note that the definition of path differentiable functions proposed in [16] is slightly different but turns out to be equivalent. There, the condition in the definition is required to be satisfied by all absolutely continuous curves γ; here, instead, we restrict to Lipschitz curves γ. The equivalence follows from the fact that absolutely continuous curves can be parameterized by arc-length —hence becoming Lipschitz curves— without affecting their role in the definition.

The class of path differentiable functions is very large and includes many cases of interest functions that are semi-algebraic, tame (definable in an o-minimal structure, see [25] and references therein. Tame functions encompass most models and loss functions used in machine learning, such as, for example, those occurring in neural network training with all the activation functions that have been

[2] Literally from the French, "healthy functions", as opposed to pathological.

considered in the literature, see e.g., [21, 25]. Note that convex, concave, or semi-convex functions (such as lower or upper C^k functions) are path differentiable; adapt the proof of [20, Lemme 3.3] or see [25].

3 Main results: accumulation, convergence, oscillation compensation

We present our main results: first we only assume path differentiability (Section 3.1)), in a second time this assumption is reinforced by a Sard's like property (Section 3.2). Our results complement those of [25] and [9].

The significance of the results is discussed in Section 3.3. The proofs are presented in Section 5.

3.1 The vanishing subgradient method for path differentiable functions

Theorem 4 (Large-time regime for path differentiable functions)**.** *Assume that $f\colon \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz path differentiable, and that $\{x_i\}_i$ is a bounded subgradient sequence. Then we have:*

i. (Slow evolution regime) Let x and y be two distinct points in $\operatorname{acc}\{x_i\}_i$ such that $f(x) \leqslant f(y)$. Let $\{x_{i_k}\}_k$ be a subsequence such that $x_{i_k} \to x$ as $k \to +\infty$, and for each k choose $i'_k > i_k$ such that $x_{i'_k} \to y$. Then

$$\bar{T}_k = \sum_{p=i_k}^{i'_k} \varepsilon_p \to +\infty.$$

ii. (Oscillation compensation) Let $\psi\colon \mathbb{R}^n \to [0,1]$ be a continuous function. Then, for every integer sequence $N_j \to +\infty$,

$$\liminf_{j\to+\infty} \frac{\sum_{i=0}^{N_j} \varepsilon_i \psi(x_i)}{\sum_{i=0}^{N_j} \varepsilon_i} > 0 \Longrightarrow \lim_{j\to+\infty} \frac{\sum_{i=0}^{N_j} \varepsilon_i v_i \psi(x_i)}{\sum_{i=0}^{N_j} \varepsilon_i \psi(x_i)} = 0.$$

iii. (Criticality). Each essential accumulation point is critical, $\operatorname{ess\,acc}\{x_i\}_{i\in\mathbb{N}} \subseteq \operatorname{crit} f$.

3.2 The vanishing subgradient method for path differentiable functions with a weak Sard property

We now assume in addition that f is constant on the connected components of its critical set. A Sard-type property which is automatically valid for some important cases, as for instance, semialgebraic or tame functions [15] or lower or upper-C^k functions [7] (for k sufficiently large).

Theorem 5 (Large-time regime for path differentiable functions: weak Sard case)**.** *In the setting of Theorem 4, and if additionally f is constant on the connected components of its critical set, then:*

i. (Slow evolution regime 2) Let x and y be two distinct points in $\operatorname{acc}\{x_i\}_i$, $x \neq y$, and take $\delta > 0$ small enough that the balls $B_\delta(x)$ and $B_\delta(y)$ are at a positive distance from each other, that is,

$\|x-y\| > 2\delta$. *Consider the successive time duration the sequence needs to go from the ball* $B_\delta(x)$ *to the ball* $B_\delta(y)$*, namely,*

$$T_j = \inf\{\textstyle\sum_{p=i}^{\ell} \varepsilon_p : j \leqslant i < \ell, x_i \in B_\delta(x), x_\ell \in B_\delta(y)\}.$$

Then $T_j \to +\infty$ *as* $j \to +\infty$.

ii. (Long intervals) Let U, V *be open neighborhoods of some accumulation point of* $\{x_i\}_i$*, such that* $\overline{U} \subset V$*. Consider the sequences of indices* $\{i_k\}_k \subset \mathbb{N}$ *and* $\{j_k\}_k \subset \mathbb{N} \cup \{+\infty\}$ *such that (refer to Figure 3):*

- $i_k < j_k < i_{k+1}$,
- $x_i \in V$ *for* $i \in I_k := [i_k, j_k] \cap \mathbb{N}$,
- $x_{i_k - 1}$ *and* x_{j_k+1} *are not in* V*, and*
- *there is some* $j \in I_k$ *such that* $x_j \in U$.

Then either there is some k *for which* $j_k = +\infty$*, i.e.,* I_k *is unbounded, or*

$$\lim_{k\to+\infty} |I_k| = \lim_{k\to+\infty} \sum_{i\in I_k} \varepsilon_i = +\infty.$$

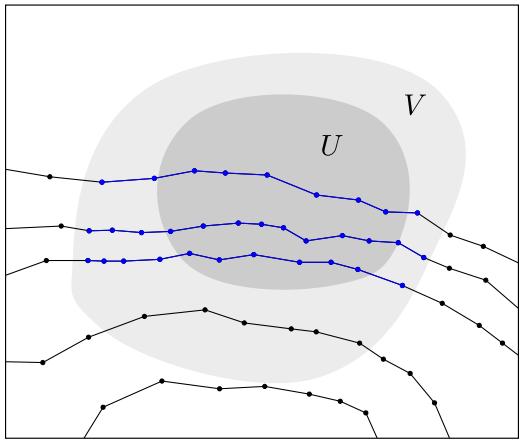

Figure 3: The intervals I_k in the statement of item (ii) correspond to fragments of the sequence contained in V and meeting U displayed here in blue.

iii. (Oscillation compensation version 2) Let $U \subset V$ *be two open sets as in item (ii), and* $A = \bigcup_i I_i$ *be the corresponding union of maximal intervals. Then*

$$\lim_{N\to+\infty} \frac{\displaystyle\sum_{\substack{0\leqslant i\leqslant N \\ i\in A}} \varepsilon_i v_i}{\displaystyle\sum_{\substack{0\leqslant i\leqslant N \\ i\in A}} \varepsilon_i} = 0.$$

iv. (Criticality) Each accumulation point is critical, $\operatorname{acc}\{x_i\}_i \subseteq \operatorname{crit} f$.

v. (Convergence of the values) The values sequence $f(x_i)$ *converge to a Clarke critical value as* $i \to +\infty$.

Remark 6. Items (iv) and (v) of Theorem 5 can also be deduced from [9, Proposition 3.27] using a different approach. The Sard-like assumption on f in [9] is that $f(\operatorname{crit} f)$ has empty interior, which is equivalent, once f is locally Lipschitz, to f being constant on the connected components of $\operatorname{crit} f$; see the proof of Lemma 23.

Up to our knowledge, items (i)–(iii) of Theorem 5 as well as Theorem 4 do not have counterparts in the optimization literature.

Remark 7 (Oscillations and V-shaped valleys). Consider a path differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ that is C^∞ both when restricted to a smooth submanifold $S \subset \mathbb{R}^n$ with $\dim S < n$ and when restricted to its complement $\mathbb{R}^n \setminus S$, and assume the gradient of f is bounded away from zero on $\mathbb{R}^n \setminus S$.

Thus, near S, f forms a V-shaped valley. In this case, we can provide more insight into the oscillation compensation phenomenon: roughly speaking, the "bouncing" (jumps between to strata adjacent to S) of $\{x_i\}_i$ gets more and more orthogonal to S around x, suggesting that the drift of the whole sequence should be parallel to S.

Lemma 8 (Normal bouncing in a V-shaped valley). *Consider a function $f\colon \mathbb{R}^n \to \mathbb{R}$ as in the previous paragraph, and a bounded subgradient sequence $\{x_i\}_i$. Assume that there is a subsequence $x_{i_j} \to x$ with*

$$\limsup_{j\to+\infty} \|v_{i_j}\| > 0 \text{ (actual bouncing).}$$

Then x is a critical point contained in the submanifold S with

$$\operatorname{acc}\{v_{i_j}\}_j \subset N_S(x) \text{ (normal bouncing)}$$

where $N_S(x) = T_S^\perp(x)$ is the normal set to S.

Proof. Let K be a compact set that contains $\{x_i\}_i$ in its interior. By the Morse–Sard theorem applied independently on the submanifold S, $f(\operatorname{crit} f)$ is a compact set of measure zero. Thus, it must be a totally-separated subset of $\mathbb{R}$. It follows that f is constant on each connected component of $\operatorname{crit} f$. In other words, we are in the setting of Theorem 5. From item (iv) of Theorem 5 we know that $x \in \operatorname{acc}\{x_i\}_i \subset \operatorname{crit} f$, and the additional condition we have on x tells us that $\partial^c f(x) \neq \{0\}$, so x cannot be a smooth point of f, whence $x \in S$. Let $(x, v) = \lim_{j\to+\infty}(x_{i_j}, v_{i_j})$ be an accumulation point of the sequence (x_i, v_i), and let $\alpha\colon \mathbb{R} \to \mathbb{R}^n$ be a smooth curve with $\alpha(0) = x$, $\alpha(t) \in S$ for all t, and $\alpha'(0) = w$. The path differentiability of f means that the choice of element of $\partial^c f(\alpha(0))$ is immaterial when we compute $(f \circ \alpha)'(0)$. So we have

$$\lim_{j\to+\infty} w \cdot v_{i_j} = v \cdot w = v \cdot \alpha'(0) = (f \circ \alpha)'(0) = 0. \qquad \square$$

This geometrical setting is reminiscent of the partial smoothness assumptions of Lewis [39] (a smooth manifold lies in between the slopes of a sharp valley). While proximal-like methods end up in a finite time on the smooth locus [34, Theorem 4.1], our result suggests that the explicit subgradient method keeps on bouncing, approaching the smooth part without actually attaining it. This confirms the intuition that finite identification does not occur, although oscillations eventually provide some information on active sets by delineating progressively their normal sets. The observation above can be extended to the semialgebraic or definable setting using Whitney stratifications.

3.3 Further discussion

Theorems 4 and 5 describe the long-term dynamics of the algorithm. While Theorem 4 only describes what happens close to $\operatorname{ess\,acc}\{x_i\}_i$ focusing on persistent behavior, Theorem 5 covers all of $\operatorname{acc}\{x_i\}_i$ that is all recurrent behaviors.

The paper [49] explores the ways in which the results presented above are sharp in the context of the class of locally-Lipschitz, path-differentiable objective functions f. The paper gives examples of functions f with corresponding non-converging subgradient sequences $\{x_i\}_i$, also with non-convergent $\{f(x_i)\}_i$, lack of oscillation compensation outside of $\operatorname{ess\,acc}\{x_i\}_i$, and other interesting properties.

Oscillation compensation. While the high-frequency oscillations will, in many cases, be considerable, they almost cancel out. This is what we refer to as *oscillation compensation*. The intuitive picture the reader should have in mind is a statement that the oscillations cancel out locally, as in (2). Yet, because of small technical minutia, we do not have exactly (2) and obtain instead very good approximations. Let us provide some explanations.

Letting, in item (ii) of Theorem 4, $\psi = \psi_{\delta,\eta} \colon \mathbb{R}^n \to [0,1]$ be a continuous cutoff function equal to 1 on a ball $B_\eta(x)$ of radius $\eta > 0$ around a point $x \in \operatorname{ess\,acc}\{x_i\}_i$ and vanishing outside the ball $B_\delta(x)$ for $\delta > \eta$, then we get, for appropriate subsequences $\{N_j\}_j \subset \mathbb{N}$,

$$\lim_{\delta \searrow 0} \lim_{\eta \nearrow \delta} \lim_{j \to +\infty} \frac{\displaystyle\sum_{i=0}^{N_j} \varepsilon_i v_i \psi_{\delta,\eta}(x_i)}{\displaystyle\sum_{i=0}^{N_j} \varepsilon_i \psi_{\delta,\eta}(x_i)} = 0,$$

which is indeed a very good approximation of (2).

Similarly, setting, in item (iii) of Theorem 5, $U = B_\eta(x)$ and $V = B_\delta(x)$ the balls centered at x with radius $0 < \eta < \delta$, we obtain this local version of the oscillation cancellation phenomenon: in the setting of Theorem 5 if $x \in \operatorname{acc}\{x_i\}_i$ and if $A_{\eta,\delta} \subset \mathbb{N}$ is the union of maximal intervals $I \subset \mathbb{N}$ such that $\{x_i\}_{i \in I} \in B_\delta(x)$ and $\{x_i\}_{i \in I} \cap B_\eta(x) \neq \emptyset$, then

$$\lim_{\delta \searrow 0} \lim_{\eta \nearrow \delta} \lim_{N \to +\infty} \frac{\displaystyle\sum_{\substack{0 \leqslant i \leqslant N \\ i \in A_{\eta,\delta}}} \varepsilon_i v_i}{\displaystyle\sum_{\substack{0 \leqslant i \leqslant N \\ i \in A_{\eta,\delta}}} \varepsilon_i} = 0.$$

Note that as we take the limit $\eta \nearrow \delta$, we cover almost all x_i in the ball $B_\delta(x)$, so we again get a statement very close to (2).

Convergence. While Theorem 5 tells us that $f(x_i)$ converges, we conjecture that this is no longer true in the context of Theorem 4, which is a matter for future research. Similarly, in the setting of path differentiable functions, the question of determining whether all limit points of bounded sequences are critical remains open.

In all cases, including the smooth case, the sequence $\{x_i\}_i$ may not converge. A well-known example of such a situation was provided for the case of smooth f by Palis–de Melo [46].

However, our results show that the drift that causes the divergence of $\{x_i\}_i$ is very slow in comparison with the local oscillations. This slowness can be immediately appreciated in the statement of item (i) of Theorem 4 and items (i) and (ii) of Theorem 5. In substance, these results express that even if the sequence diverges, it takes longer and longer to connect disjoint neighborhoods of different limit points.

4 A closed measure theoretical approach

Given a nonempty open subset of U of $\mathbb{R}^n$, denote by $C^0(U)$ the set of continuous real-valued functions on U while $C^p(U)$ is the set of $p \in [1, \infty]$ continuously differentiable real-valued functions on U. The set $\mathsf{LipCurv}(U)$ denotes the space of Lipschitz curves $\gamma\colon \mathbb{R} \to U$. When U is bounded, $\mathsf{LipCurv}(U)$ is endowed with the supremum norm $\|\gamma\|_\infty = \sup_{t\in\mathbb{R}} \|\gamma(t)\|$.

4.1 A compendium on closed measures

Elementary facts and density. Given a measure ξ on some set $X \neq \emptyset$ and a measurable map $g\colon X \to Y$, where $Y \neq \emptyset$ is another set, the *pushfoward* $g_*\xi$ is defined to be the measure on Y such that, for $A \subset Y$ measurable, $g_*\xi(A) = \xi(g^{-1}(A))$.

Recall that the *support* $\operatorname{supp}\mu$ of a positive Radon measure μ on $\mathbb{R}^m$, $m \geqslant 0$, is the set of points $x \in \mathbb{R}^m$ such that $\mu(U) > 0$ for every neighborhood U of x. It is a closed set.

The origin of the concept of closed measures, sometimes also called *holonomic measures* or *Young measures*, can be traced back to the work of L.C. Young [57, 58] in the context of the calculus of variations. It has developed in parallel to the closely related normal currents [32, 33] and varifolds [2, 3], and has found applications in several areas of mathematics, especially Lagrangian and Hamiltonian dynamics [23, 41, 42, 54], the calculus of variations [4, Section 4.3] and also optimal transport [10, 11].

The definition of closed measures is inspired from the following observations. Given a Lipschitz curve $\gamma\colon [a, b] \to \mathbb{R}^n$, its position-velocity information can be encoded by a measure μ_γ on $\mathbb{R}^n \times \mathbb{R}^n$ that is the pushforward of the Lebesgue measure on the interval $[a, b]$ into $\mathbb{R}^n \times \mathbb{R}^n$ through the mapping $t \mapsto (\gamma(t), \gamma'(t))$, that is,

$$\mu_\gamma = \frac{1}{b-a}(\gamma, \gamma')_*\mathsf{Leb}_{[a,b]}.$$

This notation extends readily to the case of curves defined on an arbitrary measurable set J in the domain of γ:

$$\mu_{\gamma|_J} = \frac{1}{|J|}(\gamma, \gamma')_*\mathsf{Leb}_J.$$

If $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a measurable function, then the integral with respect to μ_γ is given by

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} \phi(x, v)\, d\mu_\gamma(x, v) = \frac{1}{b-a}\int_a^b \phi(\gamma(t), \gamma'(t))\, dt.$$

With this definition of μ_γ, it follows that γ is a closed loop, that is, $\gamma(a) = \gamma(b)$ if, and only if, for all smooth $g\colon \mathbb{R}^n \to \mathbb{R}$, we have

$$\begin{aligned}\int_{\mathbb{R}^n\times\mathbb{R}^n} \nabla g(x)\cdot v\, d\mu_\gamma(x, v) &= \frac{1}{b-a}\int_a^b \nabla g(\gamma(t))\cdot\gamma'(t)\, dt \\ &= \frac{1}{b-a}\int_a^b (g\circ\gamma)'(t)dt = \frac{g\circ\gamma(b) - g\circ\gamma(a)}{b-a} = 0.\end{aligned}$$

In other words, the integral of $\nabla g(x)\cdot v$ with respect to μ_γ is the circulation of the gradient vector field ∇f along the curve γ, and so it vanishes when γ is a closed loop. This generalizes into:

Definition 9 (Closed measure)**.** A compactly-supported, positive, Radon measure μ on $\mathbb{R}^n \times \mathbb{R}^n$ is *closed* if, for all functions $g \in C^\infty(\mathbb{R}^n)$,

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} \nabla g(x)\cdot v\, d\mu(x, v) = 0.$$

Let $\pi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ be the projection $\pi(x,v) = x$. To a measure μ on $\mathbb{R}^n \times \mathbb{R}^n$ we can associate its *projected measure* $\pi_*\mu$. As an immediate consequence we have that $\operatorname{supp} \pi_*\mu = \pi(\operatorname{supp} \mu) \subseteq \mathbb{R}^n$.

The disintegration theorem [26] implies that there are probability measures μ_x, $x \in \mathbb{R}^n$, on $\mathbb{R}^n$ such that, if $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is measurable, we have

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \phi \, d\mu = \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}^n} \phi(x,v) \, d\mu_x(v) \right] d(\pi_*\mu)(x). \tag{4}$$

We shall refer to the couple $(\pi_*\mu, \pi_x)$ as to the *desintegration of* μ. Thus

$$\mu = \int_{\mathbb{R}^n} \mu_x \, d(\pi_*\mu)(x).$$

Definition 10 (Centroid field)**.** Let μ be a positive, compactly-supported, Radon measure on $\mathbb{R}^n \times \mathbb{R}^n$ desintegrated according to (4). The *centroid field* $\bar{v}_x$ *of* μ is

$$\bar{v}_x = \int_{\mathbb{R}^n} v \, d\mu_x(v), \;\; x \in \mathbb{R}^n.$$

The centroid field gives an average velocity at each point; it plays a significant role in our work. As a consequence of the disintegration theorem [26], $x \mapsto \bar{v}_x$ is measurable, and for every measurable $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ linear in the second variable, we have

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \phi(x,v) \, d\mu(x,v) = \int_{\mathbb{R}^n} \phi(x, \bar{v}_x) \, d(\pi_*\mu)(x). \tag{5}$$

Given the measure μ with centroid field $\bar{v}_x$, we may define its *centroidal measure* $\hat{\mu}$ on $\mathbb{R} \times \mathbb{R}$ given by

$$\int_{\mathbb{R} \times \mathbb{R}} \phi \, d\hat{\mu} = \int_{\mathbb{R}} \phi(x, \bar{v}_x) \, d\pi_*\mu(x),$$

for measurable functions $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. With this definition, μ is closed if, and only if $\hat{\mu}$ is closed. Thus the closedness property only depends on the centroid field rather than on the whole constellation of velocities in the support of μ. Observe that, if a positive Radon measure μ has a centroid field $\bar{v}_x$ that vanishes $\pi_*\mu$-almost everywhere, then μ is closed because $\hat{\mu}$ is obviously closed.

Young superposition principle. The following result, known as the *Young superposition principle* [10, 58] or as the *Smirnov solenoidal representation* [5, 53], see also [50, Example 6]. What this result tells us is basically that, not only can closed measures be approximated by measures induced by curves, but actually the centroidal measure $\hat{\mu}$ can be decomposed into a combination of measures induced by Lipschitz curves. This decomposition is very useful theoretically, as there are no limits involved.

Let U be a nonempty bounded open subset of $\mathbb{R}^n$ and set $\mathsf{LipCurv}(U) = \mathsf{LipCurv}$. For $t \in \mathbb{R}$, let $\tau_t\colon \mathsf{LipCurv} \to \mathsf{LipCurv}$ be the time-translation $\tau_t(\gamma)(s) = \gamma(s+t)$.

Theorem 11 (Young superposition principle or Smirnov solenoidal representation)**.** *For every closed probability measure μ supported in $U \times \mathbb{R}^n$ with centroid field $\bar{v}_x$, there is a Borel probability measure ν on the space* $\mathsf{LipCurv}$ *that is invariant under τ_t for all $t \in \mathbb{R}$ and such that*

$$\int_{\mathbb{R}^n} \phi(x, \bar{v}_x) \, d(\pi_*\mu)(x) = \int_{\mathsf{LipCurv}} \phi(\gamma(0), \gamma'(0)) \, d\nu(\gamma) \tag{6}$$

for any measurable $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

For details on how to obtain Theorem 11 from [10], please see [14].

Curves lying in $\operatorname{supp}\nu$ have an appealing property:

Corollary 12 (Centroidal representation of $\operatorname{supp}\nu$)**.** *With the notation of the previous theorem, we have for ν almost all γ in* $\mathsf{LipCurv}$*:*

$$\gamma'(t) = \bar{v}_{\gamma(t)}$$

for almost all $t \in \mathbb{R}$.

Proof. Take indeed $\phi \geqslant 0$ vanishing only on the measurable set consisting of points of the form $(x, \bar{v}_x)$, $x \in \mathbb{R}^n$. Then both sides of (6) must vanish, which means that for ν-almost all γ, the point $(\gamma(0), \gamma'(0))$ must be of the form $(x, \bar{v}_x)$. The conclusion follows from the τ_t-invariance of the measure ν. □

As an example, take the case in which μ is the closed measure

$$\mu = \frac{1}{2\pi}(\beta, \beta')_* \mathsf{Leb}_{[0,2\pi)}$$

on $\mathbb{R}^2 \times \mathbb{R}^2$ for

$$\beta(t) = (\cos t, \sin t).$$

In this simple example, the centroid coincides with the derivative, $\bar{v}_{\beta(t)} = \beta'(t)$. Each time-translate $\tau_t(\beta)$ is still a parameterization of the circle, and the probability measure ν we obtain in Theorem 11 is

$$\nu = \frac{1}{2\pi}\int_0^{2\pi} \delta_{\tau_t(\beta)} dt,$$

where δ_γ is the Dirac delta function whose mass is concentrated at the curve γ in the space $\mathsf{LipCurv}$.

As a general fact, the measure ν in Theorem 11 can be understood as a decomposition of the closed measure μ into a convex superposition of measures induced by Lipschitz curves. Although at first sight each γ on the right-hand side of (6) only participates at $t = 0$, the τ_t-invariance of ν means that in fact the entire curve γ is involved in the integral through its time translates $\tau_t\gamma$. Observe that another consequence of the τ_t-invariance is that the integral in the right-hand side of (6) actually writes, for all $t \in \mathbb{R}$,

$$\begin{aligned}\int_{\mathsf{LipCurv}} \phi(\gamma(0), \gamma'(0))\, d\nu(\gamma) &= \int_{\mathsf{LipCurv}} \phi(\gamma(t), \gamma'(t))\, d\nu(\gamma) \\ &= \frac{1}{|I|}\int_I \int_{\mathsf{LipCurv}} \phi(\gamma(t), \gamma'(t))\, d\nu(\gamma)\, dt \qquad (7) \\ &= \frac{1}{|I|}\int_{\mathsf{LipCurv}} \int_I \phi(\gamma(t), \gamma'(t))\, dt\, d\nu(\gamma).\end{aligned}$$

where I is any nontrivial, bounded interval . Thus (6) has the more explicit lamination or superposition form :

$$\begin{aligned}\int_{\mathbb{R}^n} \phi(x, \bar{v}_x)\, d(\pi_*\mu)(x) &= \frac{1}{|I|}\int_{\mathsf{LipCurv}} \int_I \phi(\gamma(t), \gamma'(t))\, dt\, d\nu(\gamma) \qquad (8) \\ &= \int_{\mathsf{LipCurv}} \int_{\mathbb{R}^n \times \mathbb{R}^n} \phi(x, v)\, d\mu_{\gamma|_I}(x, v)\, d\nu(\gamma) \qquad (9)\end{aligned}$$

for any bounded interval I with nonempty interior.

Although the left-hand side of (6) does not involve the full measure μ, it will turn out to be similar enough: if the integrand $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is linear in the second variable v, we still have (5) and this will be enough for the applications we have in mind.

We remark that the measure ν in Theorem 11 is not unique in general. For example, if γ is a closed curve intersecting itself once so as to form the figure 8, then, just as the figure 8 can be drawn in several ways —on a single stroke without lifting the pencil from the paper, or by drawing two circles separately—, so also the possibilities of different measures ν decomposing $\mu = \mu_\gamma$ reflect this diversity; ν can be taken to be supported on all the τ_t-translates of γ itself, or it could be taken to be supported on the curves traversing each of the loops of the 8.

4.2 Preliminaries on set-valued vector fields and circulation for a subdifferential field

In the following, we consider set-valued maps $Z\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ with the following standing assumption:

Assumption 13. For every $x \in \mathbb{R}^n$, the set $Z(x) \subseteq \mathbb{R}^n$ is nonempty, convex and locally bounded (meaning that for every compact subset $K \subseteq \mathbb{R}^n$ there is a constant $N > 0$ such that $\|y\| \leqslant N$ for all $y \in Z(x)$ and all $x \in K$), and such that the *graph* of Z, defined by

$$\operatorname{graph} Z = \{(x, p) \in \mathbb{R}^n \times \mathbb{R}^n : p \in Z(x)\},$$

is a closed subset of $\mathbb{R}^n \times \mathbb{R}^n$.

Note that if μ is a closed measure with desintegration $(\pi_*\mu, \mu_x)$ and centroid field $\bar{v}_x$, we obviously have

$$a \in \mathbb{R}, x \in \mathbb{R}^n \text{ and } a \operatorname{supp} \mu_x \subset Z(x) \Rightarrow a\bar{v}_x \in Z(x). \tag{10}$$

A major example of a set-valued mapping satisfying the above assumption is given by the Clarke subgradient $\partial^c f$ of a locally Lipschitz continuous mapping as defined in Section 2.1. Path-differentiability of a function ensures that the circulation of its subdifferential along any loop vanishes. This generalizes into:

Proposition 14 (Circulation of subdifferential for path differentiable functions)**.** *If $g\colon \mathbb{R}^n \to \mathbb{R}$ is a path differentiable function and μ is a closed probability measure, then for each open set $U \subset \mathbb{R}^n$ and each measurable function $\sigma\colon U \to \mathbb{R}^n$ with $\sigma(x) \in \partial^c g(x)$ for $x \in U$, the integral*

$$\int_{U\times\mathbb{R}^n} \sigma(x) \cdot v \, d\mu(x, v)$$

is well defined, and its value is independent of the choice of σ. We define the symbol

$$\int_{U\times\mathbb{R}^n} \partial^c g(x) \cdot v \, d\mu(x, v)$$

to be equal to this value. If $\pi(\operatorname{supp} \mu) \subset U$,

$$\int_{U\times\mathbb{R}^n} \partial^c g(x) \cdot v \, d\mu(x, v) = 0.$$

Proof. Denote by $\chi_U\colon \mathbb{R}^n \to \{0, 1\}$ the indicator function for the open set $U \subset \mathbb{R}^n$ that is equal to $\chi_U(x) = 1$ for $x \in U$ and vanishes elsewhere. Let $\sigma_1, \sigma_2\colon \mathbb{R}^n \to \mathbb{R}^n$ be two measurable functions such

that $\sigma_i(x) \in \partial^c g(x)$ for each $x \in U$, $i = 1, 2$. From Theorem 11 we get a τ_t-invariant, Borel probability measure ν on the space $\mathsf{LipCurv}$ of Lipschitz curves. Then, using (8) for any interval $I \subset \mathbb{R}$,

$$\begin{aligned}\int_{U\times\mathbb{R}^n} \sigma_1(x)\cdot v\, d\mu(x,v) - \int_{U\times\mathbb{R}^n} \sigma_2(x)\cdot v\, d\mu(x,v)\\ &= \int_{\mathbb{R}^n\times\mathbb{R}^n} \chi_U(x)(\sigma_1(x)-\sigma_2(x))\cdot v\, d\mu(x,v)\\ &= \int_{\mathsf{LipCurv}} \frac{1}{|I|}\int_I \chi_U(\gamma(t))(\sigma_1(\gamma(t))-\sigma_2(\gamma(t)))\cdot\gamma'(t)\,dt\,d\nu(\gamma).\end{aligned}$$

Since g is path differentiable, for each $\gamma \in \mathsf{LipCurv}$ and for almost every $t \in \mathbb{R}$ with $\gamma(t) \in U$,

$$\sigma_1(\gamma(t))\cdot\gamma'(t) = \sigma_2(\gamma(t))\cdot\gamma'(t).$$

From the τ_t-invariance of ν it follows then that the integrand above vanishes ν-almost everywhere.

Let us now analyze the case when $\pi(\operatorname{supp}\mu) \subset U$. Let $\psi\colon \mathbb{R}^n \to \mathbb{R}$ be a mollifier, that is, a compactly-supported, nonnegative, rotationally-invariant, C^∞ function such that $\int_{\mathbb{R}^n}\psi = 1$, and let $\psi_r(x) = r^{-n}\psi(x/r)$ for $r > 0$, so that ψ_r tends to the Dirac delta at 0 as $r \to 0$. Denote by $\psi_r * g$ the convolution of ψ_r and g. Observe that if $\beta \in \mathsf{LipCurv}$ and $a < b$, then

$$\begin{aligned}\int_a^b (g\circ\beta)'(t)\,dt &= g\circ\beta(b) - g\circ\beta(a)\\ &= \lim_{r\searrow 0}\left[(\psi_r * g)\circ\beta(b) - (\psi_r * g)\circ\beta(a)\right] = \lim_{r\searrow 0}\int_a^b ((\psi_r * g)\circ\beta)'(t)\,dt.\end{aligned}$$

This observation, together with (8), justifies the following calculation: for any bounded interval $I \subset \mathbb{R}$,

$$\begin{aligned}\int_{\mathbb{R}^n\times\mathbb{R}^n} \partial^c g(x).v\, d\mu(x,v) &= \int_{\mathsf{LipCurv}} \frac{1}{|I|}\int_I (g\circ\beta)'(t)\,dt\,d\nu(\beta) =\\ &= \lim_{r\searrow 0}\int_{\mathsf{LipCurv}} \frac{1}{|I|}\int_I ((\psi_r * g)\circ\beta)'(t)\,dt\,d\nu(\beta)\\ &= \lim_{r\searrow 0}\int_{\mathsf{LipCurv}} \frac{1}{|I|}\int_I \nabla(\psi_r * g)(\beta(t))\cdot\beta'(t)\,dt\,d\nu(\beta)\\ &= \lim_{r\searrow 0}\int_{\mathbb{R}^n\times\mathbb{R}^n} \nabla(\psi_r * g)(x)\cdot v\, d\mu(x,v),\end{aligned}$$

which vanishes because μ is closed and $\psi_r * g$ is C^∞. □

4.3 Interpolant curves and their limit measures

Given a set-valued map Z satisfying Assumption 13, we shall consider sequences $\{x_i\}_i$ satisfying

$$x_{i+1} - x_i \in \varepsilon_i Z(x_i). \tag{11}$$

Thus, for example, if $Z = \partial^c f$, the sequence $\{x_i\}_i$ is a subgradient sequence (Definition 1).

In order to analyze the asymptotics of sequences generated by dynamical systems of the form (11), we shall use the following definition.

Definition 15 (Interpolant curves)**.** Given a sequence $\{x_i\}_{i\in\mathbb{N}} \subseteq \mathbb{R}^n$ satisfying $x_{i+1} = x_i - \varepsilon_i v_i$ for some uniformly bounded vectors $v_i \in \mathbb{R}^n$ and some scalars $\varepsilon_i > 0$ that satisfy $\sum_i \varepsilon_i = +\infty$, its *interpolating curve* is the continuous piecewise affine curve $\gamma\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}^n$ with $\gamma(t_i) = x_i$ for $t_i = \sum_{j=0}^i \varepsilon_i$ and $\gamma'(t) = v_i$ for $t_i < t < t_{i+1}$.

Interpolant curves correspond to continuous-time piecewise-affine interpolation of sequences $\{x_i\}_i$, as the ones produced by the dynamical system (11). They are extremely useful to study the asymptotic behavior of these sequences.

For a bounded measurable set $B \subset \mathbb{R}_{\geqslant 0}$, we define a measure on $\mathbb{R}^n \times \mathbb{R}^n$ by

$$\mu_{\gamma|_B} = \frac{1}{|B|}(\gamma, \gamma')_* \mathsf{Leb}_B,$$

where $|B| = \int_B 1\, dt$ is the length of B, and Leb_B is the Lebesgue measure on B. If $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is measurable, then

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \phi \, d\mu_{\gamma|_B} = \frac{1}{|B|} \int_B \phi(\gamma(t), \gamma'(t))\, dt.$$

Lemma 16 (Limiting closed measures associated to bounded sequences)**.** *Let γ be the interpolating curve of a bounded sequence $\{x_i\}_i$ as in Definition 15. Let $A = \{I_i\}_{i\in\mathbb{N}}$ be a collection of intervals $I_i \subset \mathbb{R}$, with disjoint interior, such that $|I_i| \to +\infty$ as $i \to +\infty$. Set $B_N = \cup_{i=0}^N I_i$. Then the set of weak* limit points of the sequence $\{\mu_{\gamma|_{B_N}}\}_N$ is nonempty, and its elements are closed probability measures.*

Proof. Let $\phi \in C^0(\mathbb{R}^n \times \mathbb{R}^n)$. For $i \in \mathbb{N}$, write $I_i = [t_1^i, t_2^i]$ and $d_i = \|\gamma(t_1^i) - \gamma(t_2^i)\|$, and let $\alpha_i\colon [0, d_i] \to \mathbb{R}^n$ be the segment joining $\gamma(t_2^i)$ to $\gamma(t_1^i)$ with unit speed. Also, let

$$\nu_i = (\alpha_i, \alpha_i')_* \mathsf{Leb}_{[0,d_i]}$$

be the measure on $\mathbb{R}^n \times \mathbb{R}^n$ encoding α_i. Let $K \subset \mathbb{R}^n \times \mathbb{R}^n$ be a convex, compact set that contains the image of (γ, γ') and (α_i, α_i') for all i, so that $d_i \leqslant \operatorname{diam} K$. Estimate

$$\begin{aligned}\left|\frac{\sum_{i=0}^N \int_{\mathbb{R}^n\times\mathbb{R}^n} \phi\, d\nu_i}{|B_N|}\right| &= \left|\frac{\sum_{i=0}^N \int_0^{d_i} \phi(\alpha_i(t), \alpha_i'(t))\, dt}{\sum_{i=0}^N |I_i|}\right| \\ &\leqslant \frac{N(\operatorname{diam} K) \sup_{(x,v)\in K} |\phi(x,v)|}{\sum_{i=0}^N |I_i|} \to 0\end{aligned}$$

since $|I_i| \to +\infty$. Thus the measures in the accumulation sets of the sequences $\{\mu_{\gamma|_{B_N}}\}_N$ and

$$\left\{\mu_{\gamma|_{B_N}} + \frac{\sum_{i=0}^N \nu_i}{|B_N|}\right\}_N \tag{12}$$

coincide. The measures in the latter sequence are all closed since, for all $\varphi \in C^\infty(\mathbb{R}^n)$, we have, by the fundamental theorem of calculus,

$$\begin{aligned}&\int_{t_1^i}^{t_2^i} \nabla\varphi(\gamma(t))\cdot\gamma'(t)\, dt + \int_0^{d_i} \nabla\varphi(\alpha(t))\cdot\alpha'(t)\, dt \\ &\qquad = \int_{t_1^i}^{t_2^i} (\varphi\circ\gamma)'(t)\, dt + \int_0^{d_i} (\varphi\circ\alpha)'(t)\, dt \\ &\qquad = [\varphi(\gamma(t_2^i)) - \varphi(\gamma(t_1^i))] + [\varphi(\alpha(d_i)) - \varphi(\alpha_i(0))] \\ &\qquad = [\varphi(\gamma(t_2^i)) - \varphi(\gamma(t_1^i))] + [\varphi(\gamma(t_1^i)) - \varphi(\gamma(t_2^i))] = 0,\end{aligned}$$

and the measures in the sequence (12) are sums of multiples of these.

By Prokhorov's theorem [48], the set of probability measures on K is compact, so the set of limit points is nonempty. The set of closed measures is itself closed, as it is defined by a weak* closed condition.

Thus the limit points must also be closed measures. □

Note that two bounded sequences $\{x_i\}_i$ and $\{y_i\}_i$ having similar asymptotic behavior may give rise to the same set of limiting closed measures through their interpolanting curves; this is the case for example when we start the subgradient method (Definition 1) for the function $f(x) = |x|$, $x \in \mathbb{R}$, at two different initial positions $x_0 \neq y_0$, as in this case the only possible limiting measures are $\delta_{(0,0)}$ and $\frac{1}{2}\delta_{(0,-1)} + \frac{1}{2}\delta_{(0,1)}$. The coincidence of limiting measures is a manifestation of the concentration phenomenon discussed in [4, Section 4.3].

Lemma 17 (Limit points and limiting measure supports)**.** *Let γ be the interpolating curve as in Definition 15. Consider the set $\operatorname{acc}\{\mu_{\gamma|_{[0,N]}}\}_N$ of limit points of the sequence $\{\mu_{\gamma|_{[0,N]}}\}_N$ in the weak* topology. We have*

$$\overline{\bigcup \left\{ \pi(\operatorname{supp}\mu) : \mu \in \operatorname{acc}\{\mu_{\gamma|_{[0,N]}}\}\}_N \right\}} = \operatorname{ess\,acc}\{x_i\}_i.$$

Proof. Assume $x \in \operatorname{ess\,acc}\{x_i\}_i$ and let $B \subset \mathbb{R}^n$ be a closed ball whose interior contains x. Let $C > 0$ be a uniform bound of $\|v_i\| = \|x_{i+1} - x_i\|/\varepsilon_i \leqslant C$, $i \in \mathbb{N}$, which exists by Definition 15. Let $\psi\colon \mathbb{R}^n \to \mathbb{R}_{\geqslant 0}$ be a continuous function with $\operatorname{supp}\psi \subseteq B$. Since ψ is uniformly continuous on B, given $\varepsilon > 0$, there is $n_0 > 0$ such that $i > n_0$, $x, y \in B$, and $\|x - y\| \leqslant \varepsilon_i C$ imply $|\psi(x) - \psi(y)| \leqslant \varepsilon$. We hence have $|\psi(x_i) - \psi(\gamma(t))| \leqslant \varepsilon$ for $t_i \leqslant t \leqslant t_{i+1}$ and $i > n_0$. Thus, for $S > n_0$,

$$\left| \sum_{i=n_0}^{S} \varepsilon_i \psi(x_i) - \int_{t_{n_0}}^{t_S} \psi(\gamma(t))\, dt \right| \leqslant \varepsilon(t_S - t_{n_0}).$$

Since $x \in \operatorname{ess\,acc}\{x_i\}_i \subset B$, there is $\delta > 0$, such that for all $n_1 > n_0$ there is $S > n_1$ with

$$\frac{\sum_{1 \leqslant i \leqslant S} \varepsilon_i \psi(x_i)}{\sum_{i=0}^{S} \varepsilon_i} > \delta.$$

Then, taking $\varepsilon = \delta/2$,

$$\begin{aligned} \int_{\mathbb{R}^n \times \mathbb{R}^n} \psi(x)\, d\mu_{\gamma|_{[0,t_S]}}(x,v) &\geqslant \frac{1}{t_S} \int_{t_{n_0}}^{t_S} \psi(\gamma(t))\, dt \\ &\geqslant \frac{\sum_{i=n_0}^{S} \varepsilon_i \psi(x_i)}{\sum_{i=0}^{S} \varepsilon_i} - \varepsilon \frac{t_S - t_{n_0}}{t_S} \\ &> \delta - \varepsilon = \delta/2 > 0. \end{aligned}$$

It follows that there is an infinite number of integers S satisfying the previous inequality so that there is some $\mu \in \operatorname{acc}\{\mu_{\gamma|_{[0,N]}}\}_N$ with $\pi(\operatorname{supp}\mu) \cap \operatorname{supp}\psi \neq \emptyset$.

Observe that we can take the support of ψ to be contained inside any neighborhood of x, so the argument above proves that there are measures in $\operatorname{acc}\{\mu_{\gamma|_{[0,N]}}\}_N$ whose supports are arbitrarily close to x. This proves the first inclusion.

Conversely, assume that $x \in \overline{\bigcup_{\mu \in \operatorname{acc}\{\mu_{\gamma|_{[0,N]}}\}_N} \pi(\operatorname{supp}\mu)}$. For a nonnegative continuous function ψ with $x \in \operatorname{supp}\psi$, there is $\mu \in \operatorname{acc}\{\mu_{\gamma|_{[0,N]}}\}_N$ with $\int \psi\, d\mu > 0$. There is a subsequence of $\{\mu_{\gamma|_{[0,N]}}\}_N$ converging to μ, hence such that $\sum_{1 \leqslant i \leqslant S} \varepsilon_i \psi(x_i) / \sum_{i=0}^{S} \varepsilon_i$ converges to a positive quantity, so that $x \in \operatorname{ess\,acc}\{x_i\}_i$, and we obtain the opposite inclusion. □

The following proposition gives some connection between the discrete and the continuous dynamics of the differential inclusion associated to the map $-Z$.

Proposition 18 (Limiting dynamics)**.** *Take Z a field satisfying Assumption 13 together with a sequence $\{x_i\}_i$ as in* (11)*. Let $\{I_i\}_i$ be a sequence of disjoint, bounded intervals in $\mathbb{R}$ with $\lim_{i\to+\infty}|I_i| = +\infty$ and write $G_k = I_1 \cup I_2 \cup \cdots \cup I_k$.*
Suppose that for some sequence $\{k_i\}_i \subset \mathbb{N}$, the limit

$$\lim_{i\to+\infty} \mu_{\gamma|_{G_{k_i}}}$$

exists, so that, by Lemma 16, it is a closed probability measure μ. Let then ν be a Borel probability measure on the space LipCurv *of Lipschitz curves that is invariant under the time-translation τ_t and satisfies* (6)*.*
Then ν-almost every curve β satisfies

$$-\beta'(t) \in Z(\beta(t))$$

for almost every $t \in \mathbb{R}$. Moreover $\mu(\mathrm{graph}[-Z]) = 1$ and the centroid field $\bar{v}_x$ satisfies $-\bar{v}_x \in Z(x)$ for $\pi_*\mu$-almost every $x \in \mathbb{R}^n$, that is $\hat{\mu}(\mathrm{graph}[-Z]) = 1$.

Proof. The existence of ν follows from Theorem 11. By Corollary 12, we know that ν-almost every curve $\beta \in$ LipCurv satisfies, $\beta'(t) = \bar{v}_{\beta(t)}$ for almost every t. So we just need to prove that $-\bar{v}_x \in Z(x)$ for $\pi_*\mu$-almost every $x \in \mathbb{R}^n$.

Recall that $\mathrm{graph}[-Z] = \{(x,v) \in \mathbb{R}^n \times \mathbb{R}^n : -v \in Z(x)\}$. Let $K \subset \mathbb{R}^n$ be a closed ball that contains the sequence $\{x_i\}_i$ as well as $\pi(\mathrm{supp}\,\mu)$, and let $N > 0$ be such that, for all $x \in K$ and $v \in Z(x)$, $\|v\| \leqslant N$. Let $t_i \leqslant t < t_{i+1}$, using the triangle inequality,

the fact that $-\gamma'(t)$ is constant equal to v_i in the interval $t \in [t_i, t_{i+1}]$ and belongs to $Z(\gamma(t_i))$, we have

$$\begin{aligned}
&\mathrm{dist}((\gamma(t),\gamma'(t)),\mathrm{graph}[-Z]) \\
&\quad\leqslant \|(\gamma(t),\gamma'(t)) - (\gamma(t_i),-v_i)\| + \mathrm{dist}((\gamma(t_i),-v_i),\mathrm{graph}[-Z]) \\
&\quad= \|(\gamma(t),-v_i) - (\gamma(t_i),-v_i)\| + 0 \\
&\quad= \|\gamma(t) - \gamma(t_i)\| \\
&\quad\leqslant \mathrm{Lip}(\gamma)\varepsilon_i \\
&\quad\leqslant N\varepsilon_i.
\end{aligned}$$

Now

$$\begin{aligned}
&\int_{\mathbb{R}^n\times\mathbb{R}^n} \mathrm{dist}((x,v),\mathrm{graph}[-Z])\, d\mu_{\gamma|_{G_{k_i}}}(x,v) \\
&\qquad= \frac{1}{\sum_{j=1}^{k_i}|I_j|}\sum_{j=1}^{k_i}\int_{I_j} \mathrm{dist}\left((\gamma(t),\gamma'(t)),\mathrm{graph}[-Z]\right) dt \\
&\qquad\leqslant \frac{N\sum_{j=1}^{k_i}|I_j|\max_{t_\ell\in I_j}\varepsilon_\ell}{\sum_{j=1}^{k_i}|I_j|}.
\end{aligned}$$

This implies that

$$\lim_{i\to+\infty}\int_{\mathbb{R}^n\times\mathbb{R}^n} \mathrm{dist}((x,v),\mathrm{graph}[-Z])\, d\mu_{\gamma|_{G_{k_i}}}(x,v) = 0$$

by the Stolz-Cesàro theorem using the fact that, for k large enough, $\sum_{j=1}^{k}|I_j| \geqslant ck$ for a positive constant c, and the fact that ε_i converges to 0 as $i \to +\infty$. This, in turn, implies that

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} \mathrm{dist}((x,v),\mathrm{graph}[-Z])\, d\mu(x,v) = 0$$

because the convergence of measures occurs in the weak* topology and the integrand is continuous. Since graph$[-Z]$ is a closed set, the support of μ must be contained in it. From 10 with $a = -1$, we know that $-\bar{v}_x \in Z(x)$, which is what we wanted to prove. □

Theorem 19 (Subgradient-like closed measures are trivial)**.** *Assume that $g\colon \mathbb{R}^n \to \mathbb{R}$ is a path differentiable function. Let μ be a closed probability measure on $\mathbb{R}^n \times \mathbb{R}^n$ such that $\mu(\mathrm{graph}[-\partial^c g]) = 1$. Then the centroid field $\bar{v}_x$ of μ vanishes for $\pi_*\mu$-almost every x.*

Proof. The condition on μ implies, by Remark 10 with $a = -1$, that $-\bar{v}_x \in \partial^c g(x)$ for $\pi_*\mu$-almost every x. By Proposition 14 we may choose $\sigma(x) = -\bar{v}_x$ to compute

$$\begin{aligned}\int_{\mathbb{R}^n\times\mathbb{R}^n} \partial^c g(x).v \, d\mu(x,v) = \int_{\mathbb{R}^n\times\mathbb{R}^n} \sigma(x)\cdot v \, d\mu(x,v) &= \int_{\mathbb{R}^n} \sigma(x)\cdot\left[\int_{\mathbb{R}^n} v\, d\mu_x\right] d(\pi_*\mu)(x)\\ &= \int_{\mathbb{R}^n} \sigma(x)\cdot\bar{v}_x \, d(\pi_*\mu)(x) = -\int_{\mathbb{R}^n} \bar{v}_x\cdot\bar{v}_x \, d(\pi_*\mu)(x).\end{aligned}$$

Proposition 14 also implies that the left-hand side vanishes because μ is closed. □

In our proofs below, Theorem 19 will be applied in conjunction with Proposition 18 with $Z = \partial^c f$. Observe that Theorem 19 could as well have been presented just after Proposition 14, as not much more is needed for its proof.

5 Proofs of main results

Our proofs use two basic techniques: sometimes we use Theorem 19 to deal with long subsequences of $\{x_i\}_i$, and sometimes we use shorter subsequences and instead use Lemmas 21 and 22, which exploit the Arzelà-Ascoli theorem to obtain curves that describe the asymptotic flow.

The structure of the proofs is described in Figure 4. In contrast, in the paper [9], item (v) of Theorem 5 is proven first, and item (iv) is deduced from it.

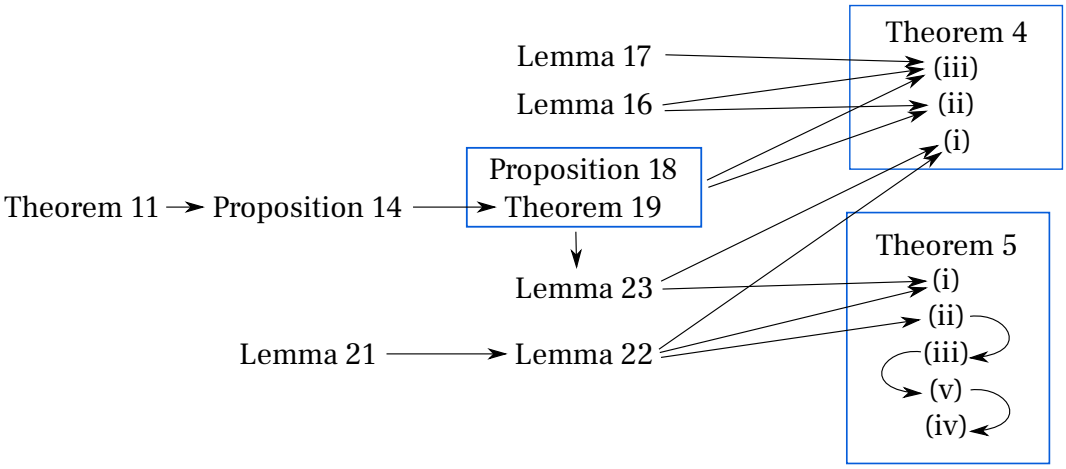

Figure 4: Arrows indicate results that are used in the proofs of the statements they point to.

Remark 20. An alternative route to the proof of items (iv) and (v) of Theorem 5, closer to the one already given in [9], is to use Lemma 23 to prove (v) and deduce (iv). Item (v) also follows from [9, Proposition 3.27].

5.1 Lemmas on the convergence of curve segments for general multivalued dynamics

In this section, we consider a set-valued map $Z : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ satisfying Assumption 13 and x_i a corresponding sequence as in (11).

Lemma 21 (Approximate solutions of differential inclusions)**.** *For each $i \in \mathbb{N}$, let $T_i > 0$ and assume that $T_i \to T$ for some $T > 0$. Let, for each $i \in \mathbb{N}$, $\gamma_i \colon [0, T_i] \to \mathbb{R}^n$ be a Lipschitz curve. Assume that the sequence $\{\gamma_i\}_i$ converges to some bounded, Lipschitz curve $\gamma \colon [0, T] \to \mathbb{R}$, $\gamma_i \to \gamma$, in the sense that $\sup_{t\in[0,\min(T_i,T)]} \|\gamma(t) - \gamma_i(t)\| \to 0$ and*

$$\lim_{i\to+\infty} \int_0^{T_i} \operatorname{dist}((\gamma_i(t), \gamma_i'(t)), \operatorname{graph}[-Z])\, dt = 0. \tag{13}$$

Then $-\gamma'(t) \in Z(\gamma(t))$ for almost all $t \in [0, T]$.

Proof. We follow classical arguments; see for example [9, Theorem 4.2]. Let $0 < T' < T$. For i large enough, $T_i > T'$ because $T_i \to T$. In particular, we eventually have uniform convergence of γ_i on $[0, T']$ to the restriction of γ to $[0, T']$. For each i, the derivative γ_i' is an element of $L^\infty = L^\infty([0, T']; \mathbb{R}^d)$, and being uniformly bounded with compact domain, belong to $L^2 = L^2([0, T']; \mathbb{R}^d)$ as well. Recall that, since L^2 is reflexive, the weak and weak* topologies coincide in L^2. So by the Banach–Alaoglu compactness theorem, by passing to a subsequence we may assume that γ_j' converge weakly in L^2 and weak* in L^∞ to some $u \in L^2 \cap L^\infty$.

Since γ_j converges to γ uniformly, $\gamma_j \to \gamma$ also in L^2. Hence γ_j' tends to γ' in the sense of distributions on $[0, T']$; indeed, for all C^∞ functions $g \colon [0, T'] \to \mathbb{R}$ with compact support in $(0, T')$, we have

$$\int_0^{T'} \gamma_j'(t) g(t)\, dt = -\int_0^{T'} \gamma_j(t) g'(t)\, dt \to -\int_0^{T'} \gamma(t) g'(t)\, dt = \int_0^{T'} \gamma'(t) g(t)\, dt$$

since we have convergence in L^2. By uniqueness of the limit, $u = \gamma'$ almost everywhere on $[0, T']$.

It follows from Mazur's lemma [28, p. 6] that there is a function $N \colon \mathbb{N} \to \mathbb{N}$ and, for each $p \leqslant k \leqslant N(p)$, a number $a(p, k) \geqslant 0$ such that $\sum_{k=p}^{N(p)} a(p, k) = 1$, and such that the convex combinations

$$\sum_{k=p}^{N(p)} a(p, k) \gamma_k' \to \gamma' \tag{14}$$

strongly in L^2 as $p \to +\infty$ (and also in the weak* sense in L^∞).

Since the set $Z(x)$ is convex at each x, the function

$$g(x, v) = \operatorname{dist}(-v, Z(x))$$

is convex in its second argument for fixed $x \in \mathbb{R}^n$. Using the fact that the convergence (14) happens pointwise almost everywhere, we have, by continuity of g and by the fact that countable union of zero measure sets has zero measure, for almost all $t \in [0, T']$

$$\begin{aligned} g(\gamma(t), \gamma'(t)) &= g(\gamma(t), \lim_{p\to+\infty} \textstyle\sum_{k=p}^{N(p)} a(p, k) \gamma_k'(t)) \\ &= \lim_{p\to+\infty} g(\gamma(t), \textstyle\sum_{k=p}^{N(p)} a(p, k) \gamma_k'(t)) \\ &\leqslant \liminf_{p\to+\infty} \sum_{k=p}^{N(p)} a(p, k) g(\gamma(t), \gamma_k'(t)), \end{aligned}$$

where the last step follows from Jensen's inequality and convexity of g in its second argument. Since g is non negative, integrating on $[0, T']$, we have using Fatou's Lemma,

$$\begin{aligned}
0 &\leqslant \int_0^{T'} g(\gamma(t), \gamma'(t))\, dt \\
&\leqslant \liminf_{p\to+\infty} \int_0^{T'} \sum_{k=p}^{N(p)} a(p,k) g(\gamma(t), \gamma_k'(t))\, dt \\
&\leqslant \liminf_{p\to+\infty} \int_0^{T'} \sum_{k=p}^{N(p)} a(p,k) [\operatorname{dist}((\gamma(t), \gamma_k'(t)), (\gamma_k(t), \gamma_k'(t))) \\
&\qquad\qquad + g(\gamma_k(t), \gamma_k'(t))]\, dt \\
&= \liminf_{p\to+\infty} \int_0^{T'} \sum_{k=p}^{N(p)} a(p,k) [\operatorname{dist}(\gamma(t), \gamma_k(t)) + g(\gamma_k(t), \gamma_k'(t))]\, dt
\end{aligned}$$

where we have used the triangle inequality. Now, using a uniform bound on the integral, we have

$$\begin{aligned}
0 &\leqslant \int_0^{T'} g(\gamma(t), \gamma'(t))\, dt \\
&\leqslant \liminf_{p\to+\infty} \sum_{k=p}^{N(p)} a(p,k) \left(T' \sup_{t\in[0,T']} [\operatorname{dist}(\gamma(t), \gamma_k(t))] + \int_0^{T'} g(\gamma_k(t), \gamma_k'(t)) \right) \\
&\leqslant \liminf_{p\to+\infty} \sup_{p\leqslant k\leqslant N(p)} \left(T' \sup_{t\in[0,T']} [\operatorname{dist}(\gamma(t), \gamma_k(t))] + \int_0^{T'} g(\gamma_k(t), \gamma_k'(t)) \right) \\
&\leqslant \limsup_{k\to+\infty} \left(T' \sup_{t\in[0,T']} [\operatorname{dist}(\gamma(t), \gamma_k(t))] + \int_0^{T'} g(\gamma_k(t), \gamma_k'(t)) \right) = 0,
\end{aligned}$$

where we used the fact that $\sum_{k=p}^{N(p)} a(p,k) = 1$, the fact that $\gamma_k \to \gamma$ uniformly and the hypothesis in (13). Hence we have $-\gamma'(t) \in Z(\gamma(t))$ for almost all $t \in [0, T']$, and this proves the lemma since T' was taken arbitrary in $(0, T)$. □

Lemma 22 (Limiting dynamics for discrete sequences)**.** *Let γ be the interpolant curve of the bounded sequence $\{x_i\}_i$, and $\{I_j\}_j$ a collection of pairwise-disjoint intervals of $\mathbb{R}_{\geqslant 0}$ of length $1/C \leqslant |I_j| \leqslant C$ for some $C > 1$. Then, there is a subsequence $\{j_k\}_k \subset \mathbb{N}$ such that the restrictions $\gamma|_{I_{j_k}}$ converge uniformly to a Lipschitz curve $\bar\gamma\colon [a, b] \to \mathbb{R}$ that satisfies $-\bar\gamma'(t) \in Z(\gamma(t))$ for almost every $t \in [a, b]$.*

If we additionally assume that the sequence $\{x_i\}_i$ is generated by the subgradient method, so that $-v_i \in \partial^c f(x_i)$ for some locally Lipschitz, path differentiable function f, then the curve $\bar\gamma$ satisfies

$$-\int_a^b \|\bar\gamma'(t)\|^2 dt = f \circ \bar\gamma(b) - f \circ \bar\gamma(a).$$

Proof. By passing to a subsequence, we may assume that the lengths $|I_j|$ converge to a positive number. By the Lipschitz version of the Arzelà–Ascoli theorem, we may pass to a subsequence such that $\gamma|_{I_{j_k}}$ converges uniformly to a curve $\bar\gamma$ on an interval $[a, b]$ of length $\lim_{j\to+\infty} |I_j| > 0$. Condition (13) holds if we let γ_i be the appropriate translate of $\gamma|_{I_i}$, so by Lemma 21, $-\bar\gamma'(t) \in Z(\gamma(t))$ for almost every $t \in [a, b]$.

Let us now prove the second statement, so that $Z = \partial^c f$. By the path differentiability of f, we have

$$-\int_a^b \|\bar\gamma'(t)\|^2 dt = \int_a^b \partial^c f(\bar\gamma(t))\cdot\bar\gamma'(t)dt = \int_a^b (f\circ\bar\gamma)'(t)\,dt = f\circ\bar\gamma(b) - f\circ\bar\gamma(a). \qquad \square$$

5.2 Proof of Theorem 4

5.2.1 Item (i)

Let γ be the interpolant curve of the sequence $\{x_i\}_i$, and consider the intervals $I_k = [t_{i_k}, t_{i'_k}]$, so that the endpoints of the restriction $\gamma|_{I_k}$ are precisely $\gamma(t_{i_k}) = x_{i_k}$ and $\gamma(t_{i'_k}) = x_{i'_k}$. Aiming for a contradiction, assume that the numbers $T_k = t_{i'_k} - t_{i_k}$ remain bounded. Apply Lemma 22 to obtain a curve $\bar\gamma\colon [a,b]\to\mathbb{R}^n$ joining $\bar\gamma(a) = \lim_k x_{i_k} = x$ and $\bar\gamma(b) = \lim_k x_{i'_k} = y$. We have that the arc length of $\bar\gamma$ must be positive because $x\neq y$, while $\bar\gamma$ also satisfies, as part of the conclusion of Lemma 22,

$$0 > -\int_a^b \|\bar\gamma'(t)\|^2 dt = f\circ\bar\gamma(b) - f\circ\bar\gamma(a) = f(y) - f(x) \geqslant 0.$$

Whence we get the contradiction we were aiming at.

5.2.2 Item (ii)

Let $B\subset\mathbb{R}^n$ be a closed ball containing the sequence $\{x_i\}_i$. By convexity, B contains also the image of the interpolating curve γ.

Fix $\varepsilon > 0$. By uniform continuity of ψ over B, there exists $n_0 > 0$ such that $i > n_0$, $x, y\in B$ and $|x-y|\leqslant \varepsilon_i \operatorname{Lip}(f)$ imply $|\psi(x)-\psi(y)|\leqslant\varepsilon$ and $\varepsilon_i\leqslant 1$. We hence have $|\psi(x_i) - \psi(\gamma(t))|\leqslant\varepsilon$ for $t_i\leqslant t\leqslant t_{i+1}$ and $i > n_0$. Thus

$$\left|\sum_{i=n_0}^{N_j}\varepsilon_i v_i\psi(x_i) - \int_{t_{n_0}}^{t_{N_j}}\gamma'(t)\psi(\gamma(t))dt\right| \leqslant \varepsilon\operatorname{Lip}(f)(t_{N_j} - t_{n_0})$$

and

$$\begin{aligned}\frac{1}{\sum_{i=0}^{N_j}\varepsilon_i}&\left|\sum_{i=0}^{N_j}\varepsilon_i v_i\psi(x_i) - \int_0^{t_{N_j}}\gamma'(t)\psi(\gamma(t))dt\right| \\ &\leqslant \frac{1}{\sum_{i=0}^{N_j}\varepsilon_i}\left[\left|\sum_{i=0}^{n_0-1}\varepsilon_i v_i\psi(x_i) - \int_0^{t_{n_0}}\gamma'(t)\psi(\gamma(t))dt\right| + \varepsilon\operatorname{Lip}(f)(t_{N_j} - t_{n_0})\right].\end{aligned}$$

Since $\sum_{i=0}^{\infty}\varepsilon_i = +\infty$ and $\varepsilon > 0$ was arbitrary, it follows that the latter becomes arbitrarily small as N_j grows.

Whence the quotient in the limit in the statement of item (ii) is very close, for large j, to

$$\frac{\sum_{i=0}^{N_j}\varepsilon_i}{\sum_{i=0}^{N_j}\varepsilon_i\psi(x_i)}\int_{\mathbb{R}^n\times\mathbb{R}^n} v\psi(x)d\mu_{\gamma|_{[0,t_{N_j+1}]}}(x,v).$$

We now prove that the above quantity converges to 0 as $j\to+\infty$. Taking a subsequence so that $\mu_{\gamma|_{[0,t_{N_j+1}]}}$ converges to some probability measure μ, the quotient on the left converges to

$$1\Big/\int\psi(x)\,d\pi_*\mu(x),$$

and our hypothesis on the subsequence $\{N_j\}_j$ thus guarantees that $\int \psi(x)\, d\pi_*\mu(x) > 0$.

Thus, it suffices to show that, for every limit point μ of the sequence $\{\mu_{\gamma|_{[0,t_{N+1}]}}\}_N$ satisfying

$$\int \psi(x)\, d\pi_*\mu(x) > 0,$$

we have

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} v\,\psi(x)\, d\mu(x,v) = \int_{\mathbb{R}^n} \bar{v}_x \psi(x)\, d(\pi_*\mu)(x) = 0, \tag{15}$$

where $\bar{v}_x$ is the centroid field of μ. By Lemma 16 we know that μ is closed so that Proposition 18 and Theorem 19 apply and give $\bar{v}_x = 0$ for $\pi_*\mu$-almost every x. This immediately implies (15).

5.2.3 Item (iii)

To prove item (iii), consider the interpolation curve constructed in Section 4.3. Consider a limit point μ of the sequence $\{\mu_{\gamma|_{[0,N]}}\}_N$. By Lemma 16, μ is closed. By Proposition 18 and Theorem 19, the centroid field $\bar{v}_x$ of μ vanishes for $\pi_*\mu$-almost every x, so from (10) we know that $0 = -\bar{v}_x \in \partial^c f(x)$, and hence $x \in \operatorname{crit} f$ for a dense subset of $\pi(\operatorname{supp}\mu)$. Since this is true for all limit points μ, by Lemma 17 we know that it is true throughout $\operatorname{ess\,acc}\{x_i\}_i$.

5.3 Proof of Theorem 5

5.3.1 The function is constant on the accumulation set

Lemma 23 (f is constant on its limit set)**.** *Assume that the path differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is constant on each connected component of its critical set, and let $\{x_i\}_i$ be a bounded sequence produced by the subgradient method. Then f is constant on the set $\operatorname{acc}\{x_i\}_i$ of limit points of $\{x_i\}_i$.*

Proof. Assume instead that f takes two values $J_1 < J_2$ within $\operatorname{acc}\{x_i\}_i$.

Let K be a compact set that contains the closure $\overline{\{x_i\}_i}$ in its interior. Since f is constant on the connected components of $\operatorname{crit} f$ and since f is Lipschitz, the set $f(K \cap \operatorname{crit} f)$ has measure zero because, given $\varepsilon > 0$, the connected components C_i of $K \cap \operatorname{crit} f$ of positive measure $|C_i| > 0$ —of which there are only countably many— can be covered with open sets

$$f^{-1}((f(C_i) - \varepsilon/2^{i+1}, f(C_i) + \varepsilon/2^{i+1}))$$

with image under f of length $\varepsilon/2^i$; the rest of $K \cap \operatorname{crit} J$ has measure zero, so it is mapped to another set of measure zero. The set $f(K \cap \operatorname{crit} f)$ is also compact, so we conclude that it is not dense on any open interval of $\mathbb{R}$.

We may thus assume, without loss of generality, that the values J_1 and J_2 are such that there are no critical values of $f|_K$ between them.

Pick $c_1, c_2 \in \mathbb{R}$ such that

$$J_1 < c_1 < c_2 < J_2.$$

Let $W_1 = f^{-1}(-\infty, c_1)$ and $W_2 = f^{-1}(c_2, +\infty)$. Clearly $W_j \cap \operatorname{acc}\{x_i\}_i \neq \emptyset$ because the value J_j is attained in $\operatorname{acc}\{x_i\}_i$, $j = 1, 2$.

Consider the curve $\gamma\colon \mathbb{R}_{\geqslant 0} \to K \subset \mathbb{R}^n$ interpolating the sequence $\{x_i\}_i$. Let A be the set of intervals

$$A = \{[t_1, t_2] \subset \mathbb{R} : t_1 < t_2, \gamma(t_1) \in \partial W_1, \gamma(t_2) \in \partial W_2, \gamma(t) \notin \overline{W_1 \cup W_2} \text{ for } t \in (t_1, t_2)\}$$

Write $A = \{I_j\}_{j\in\mathbb{N}}$ for maximal, disjoint intervals I_j. Observe that if $I_j = [t_1^j, t_2^j]$, then we have, by the path differentiability of f, that

$$\int_{t_1^j}^{t_2^j} \partial^c f(\gamma(t)) \cdot \gamma'(t)\, dt = \int_{t_1^j}^{t_2^j} (f\circ\gamma)'(t)\, dt = f\circ\gamma(t_2^j) - f\circ\gamma(t_1^j) = c_2 - c_1. \tag{16}$$

Let μ be a probability measure that is a limit point of the sequence $\{\mu_{\gamma|_{\cup_{i=0}^N I_i}}\}_N$.

Now, since f is Lipschitz and $\overline{W}_1$ and $\overline{W}_2$ are compact, $|I_i|$ is bounded from below, let us say

$$|I_i| > \alpha.$$

It is also bounded from above, because if not then there is a subset $\{I_{i_j}\}_j$ of A consisting of intervals with length $|I_{i_j}| \to +\infty$, and we can apply Lemma 16, Proposition 18, and Theorem 19 to get closed measures $\tilde{\mu}$ with $\operatorname{supp}\pi_*\tilde{\mu} \subset \operatorname{crit} f$. Since the support of each such $\pi_*\tilde{\mu}$ is contained in $K\setminus(W_1\cup W_2)$, this would mean the existence of a critical value between c_1 and c_2, which contradicts our choice of J_1 and J_2. We conclude that the size of the intervals in A is also bounded from above, say,

$$|I_i| < \beta.$$

Apply the first part of Lemma 22 to obtain a subsequence $\{j_k\}_k$ such that $\{\gamma|_{I_{j_k}}\}_k$ converges uniformly to a Lipschitz curve $\bar{\gamma}\colon [a,b]\to\mathbb{R}^n$ joining $\bar{\gamma}(a)\in\overline{W}_1$ with $\bar{\gamma}(b)\in\overline{W}_2$. By (16), we have

$$\int_a^b \partial^c f(\bar{\gamma}(t))\cdot\bar{\gamma}'(t)\, dt = \lim_{k\to+\infty}\int_{t_1^{j_k}}^{t_2^{j_k}} \partial^c f(\gamma(t))\cdot\gamma'(t)\, dt \geqslant c_2 - c_1 > 0.$$

We also know that γ interpolates a gradient sequence, so we may use the path differentiability of f and the second part of Lemma 22 to see that this integral must be nonpositive, a contradiction that proves that statement of the lemma. □

5.3.2 Proof of item (i)

For $j\in\mathbb{N}$, let $I_j = [t_{i_j}, t_{i_j+1}] \subset \mathbb{R}$ be the interval closest to 0 with $t_j \leqslant t_{i_j} < t_{i_j+1}$, $\gamma(t_{i_j})\in B_\delta(x)$, and $\gamma(t_{i_j+1})\in B_\delta(y)$, so that $T_j = t_{i_j+1} - t_{i_j}$. Let $\gamma|_{I_j}$ be the restriction of the interpolant curve γ. Since the two balls $B_\delta(x)$ and $B_\delta(y)$ are at positive distance from each other, and since the velocity is bounded uniformly $\|\gamma'\| \leqslant \operatorname{Lip}(f)$, we know that the numbers $T_j = |I_j|$ are uniformly bounded from below by a positive number.

Assume, looking for a contradiction, that there is a subsequence of $\{T_j\}_j$ that remains bounded from above. Apply Lemma 22 to obtain a curve $\bar{\gamma}\colon[a,b]\to\mathbb{R}$ such that $\bar{\gamma}(a)\in B_\delta(x)\cap\operatorname{acc}\{x_i\}_i$ and $\bar{\gamma}(b)\in B_\delta(y)\cap\operatorname{acc}\{x_i\}_i$, while also satisfying

$$-\int_a^b \|\bar{\gamma}'(t)\|^2 dt = f\circ\bar{\gamma}(b) - f\circ\bar{\gamma}(a) = 0 \tag{17}$$

by Lemma 23. This contradicts the fact that the distance between the balls $B_\delta(x)$ and $B_\delta(y)$ —and hence also the arc length of $\bar{\gamma}$— is positive.

5.3.3 Proof of item (ii)

Aiming at a contradiction, we assume instead that there is some $x\in\overline{U}\cap\operatorname{acc}\{x_i\}_i$ and some subsequence $\{i_j\}_j$ such that $\operatorname{dist}(x,\gamma(I_{i_j}))\to 0$ and $|I_{i_j}| \leqslant C$ for some $C>0$ and all $j\in\mathbb{N}$.

We may thus apply Lemma 22 to get a curve $\bar{\gamma}:[a,b]\to\mathbb{R}^n$ whose endpoints $\bar{\gamma}(a)$ and $\bar{\gamma}(b)$ are contained in $\operatorname{acc}\{x_i\}_i\setminus V$, and $\bar{\gamma}$ passes through $x\in\overline{U}$, so it has positive arc length. However, it is also a conclusion of Lemma 22, together with Lemma 23, that $\bar{\gamma}$ satisfies (17), which makes it impossible for its arc length to be positive, so we have arrived at the contradiction we were looking for.

5.3.4 Proof of item (iii)

Let U, V, and A be as in the statement of item (iii). Let $B = \bigcup_{i\in A}[t_i, t_{i+1})$. The statement of item (iii) is equivalent to the statement that

$$\lim_{N\to+\infty} \int v \, d\mu_{\gamma|_{B\cap[0,N]}} = 0. \tag{18}$$

It follows from item (ii) that the maximal intervals $I_i \subset \mathbb{N}$ comprising $A = \bigcup_i I_i$ satisfy $|I_i| \to +\infty$ as do the lengths $\sum_{j\in I_i} \varepsilon_j$ of the intervals $\bigcup_{j\in I_i}[t_j, t_{j+1}) \subset B$. Hence, from Lemma 16 we know that any limit point μ of the sequence $\{\mu_{\gamma|_{B\cap[0,N]}}\}_N$ is closed. This implies (18) because each coordinate v_k of the integrand $v = (v_1, \dots, v_n)$ is a gradient: $v_k = \nabla p_k \cdot v$ for $p_k(x) = x_k$, so each entry of the integral (18) vanishes.

5.3.5 Proof of item (iv)

Let $x \in \operatorname{acc}\{x_i\}_i$. For any neighborhood U of x, we can take a slightly larger neighborhood V and repeat the construction described in the proof of item (iii) (Section 5.3.4) of a closed measure μ whose support intersects U, and whose centroid field vanishes $\pi_*\mu$-almost everywhere. By Remark 10 we know that the centroid field satisfies $-\bar{v}_x \in \partial^c f(x)$. In sum, we have that in every neighborhood U of x, there is a point $y \in U$ with $0 \in \partial^c f(y)$, which implies that $0 \in \partial^c f(x)$ because the graph of $\partial^c f$ is closed in $\mathbb{R}^n \times \mathbb{R}^n$.

5.3.6 Proof of item (v)

Recall that $\operatorname{acc}\{x_i\}_i$ is connected. We know from item (iv) that $\operatorname{acc}\{x_i\}_i \subseteq \operatorname{crit} f$. So it is contained in a single connected component of $\operatorname{crit} f$. Hence f must be constant on $\operatorname{acc}\{x_i\}_i$, and $\{f(x_i)\}_i$ converges.

Acknowledgements. The authors acknowledge the support of ANR-3IA Artificial and Natural Intelligence Toulouse Institute. JB and EP also thank Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA9550-19-1-7026, FA9550-18-1-0226, and ANR MasDol. JB acknowledges the support of ANR Chess, grant ANR-17-EURE-0010, TSE-P and ANR OMS.

References

[1] Pierre-Antoine Absil, Robert Mahony, and Benjamin Andrews. "Convergence of the iterates of descent methods for analytic cost functions". In: *SIAM Journal on Optimization* 16.2 (2005), pp. 531–547.

[2] William K. Allard. "On the first variation of a varifold". In: *Ann. of Math. (2)* 95 (1972), pp. 417–491.

[3] Frederick J. Almgren Jr. *Plateau's problem: An invitation to varifold geometry*. W. A. Benjamin, Inc., New York-Amsterdam, 1966, pp. xii+74.

[4] Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. Second edition. MOS-SIAM Series on Optimization. SIAM, 2014.

[5] Victor Bangert. "Minimal measures and minimizing closed normal one-currents". In: *Geometric And Functional Analysis* 9.3 (1999), pp. 413–427.

[6] Anas Barakat and Pascal Bianchi. "Convergence analysis of a momentum algorithm with adaptive step size for non convex optimization". Preprint. arXiv:1911.07596. 2019.

[7] Luc Barbet, Marc Dambrine, Aris Daniilidis, and Ludovic Rifford. "Sard theorems for Lipschitz functions and applications in optimization". In: *Israël Journal of Mathematics* 212.2 (2016), pp. 757–790.

[8] Amir Beck and Marc Teboulle. "Mirror descent and nonlinear projected subgradient methods for convex optimization". In: *Operations Research Letters* 31.3 (2003), pp. 167–175.

[9] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. "Stochastic approximations and differential inclusions". In: *SIAM Journal on Control and Optimization* 44.1 (2005), pp. 328–348.

[10] Patrick Bernard. "Young measures, superposition and transport". In: *Indiana Univ. Math. J.* 57.1 (2008), pp. 247–275.

[11] Patrick Bernard and Boris Buffoni. "Optimal mass transportation and Mather theory". In: *Journal of the European Mathematical Society* 9.1 (2007), pp. 85–121.

[12] Pascal Bianchi, Walid Hachem, and Adil Salim. "Constant step stochastic approximations involving differential inclusions: Stability, long-run convergence and applications". In: *Stochastics* 91.2 (2019), pp. 288–320.

[13] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. "Convergence of constant step stochastic gradient descent for non-smooth non-convex functions". Preprint. arXiv:2005.08513. 2020.

[14] Pascal Bianchi and Rodolfo Rios-Zertuche. "A closed-measure approach to stochastic approximation". Preprint. arXiv:2112.05482.

[15] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. "Clarke Subgradients of Stratifiable Functions". In: *SIAM Journal on Optimization* 18.2 (Jan. 2007), pp. 556–572.

[16] Jérôme Bolte and Edouard Pauwels. "Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning". In: *Mathematical Programming* (2020).

[17] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. "Proximal alternating linearized minimization for nonconvex and nonsmooth problems". In: *Mathematical Programming* 146.1-2 (2014), pp. 459–494.

[18] Jonathan M Borwein and Warren B Moors. "A chain rule for essentially smooth Lipschitz functions". In: *SIAM Journal on Optimization* 8.2 (1998), pp. 300–308.

[19] Jonathan Borwein, Warren Moors, and Xianfu Wang. "Generalized subdifferentials: a Baire categorical approach". In: *Transactions of the American Mathematical Society* 353.10 (2001), pp. 3875–3893.

[20] Haïm Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Mathematics Studies, Notas de Matemática no. 5. North-Holland Publishing Company, 1973.

[21] Camille Castera, Jérôme Bolte, Cédric Févotte, and Edouard Pauwels. "An inertial newton algorithm for deep learning". In: *arXiv preprint arXiv:1905.12278* (2019).

[22] Frank H. Clarke. *Optimization and nonsmooth analysis*. Vol. 5. Classics in Applied Mathematics. SIAM/Wiley, 1990.

[23] Gonzalo Contreras and Renato Iturriaga. *Global minimizers of autonomous Lagrangians*. 22º Colóquio Brasileiro de Matemática. [22nd Brazilian Mathematics Colloquium]. Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 1999, p. 148.

[24] Aris Daniilidis and Dmitriy Drusvyatskiy. "Pathological subgradient dynamics". In: *SIAM Journal on Optimization* 30.2 (2020), pp. 1327–1338.

[25] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. "Stochastic Subgradient Method Converges on Tame Functions". In: *Foundations of Computational Mathematics* (Jan. 2019).

[26] Claude Dellacherie and Paul-André Meyer. *Probabilities and potential, vol. 29 of North-Holland Mathematics Studies*. 1978.

[27] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization". In: *Journal of machine learning research* 12.Jul (2011), pp. 2121–2159.

[28] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. Vol. 28. Siam, 1999.

[29] Yu.M. Ermol'ev. "Methods for solving nonlinearextremal problems". In: *Kibernetika (Kiev)* 1.4 (1966), pp. 1–17.

[30] Yu.M. Ermol'ev and N.Z. Shor. "On the minimization of nondifferentiable functions". In: *Cybernetics* 3.1 (1967), pp. 72–72.

[31] Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015.

[32] Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969, pp. xiv+676.

[33] Herbert Federer. "Real flat chains, cochains and variational problems". In: *Indiana Univ. Math. J.* 24 (1974), pp. 351–407.

[34] Warren L. Hare and Adrian S. Lewis. "Identifying active constraints via partial smoothness and prox-regularity". In: *Journal of Convex Analysis* 11.2 (2004), pp. 251–266.

[35] Warren Hare and Claudia Sagastizábal. "A redistributed proximal bundle method for nonconvex optimization". In: *SIAM Journal on Optimization* 20.5 (2010), pp. 2442–2473.

[36] Catherine F Higham and Desmond J Higham. "Deep learning: An introduction for applied mathematicians". In: *SIAM Review* 61.4 (2019), pp. 860–891.

[37] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[38] Krzysztof C Kiwiel. "Convergence and efficiency of subgradient methods for quasiconvex minimization". In: *Mathematical programming* 90.1 (2001), pp. 1–25.

[39] Adrian S. Lewis. "Active sets, nonsmoothness, and sensitivity". In: *SIAM Journal on Optimization* 13.3 (2002), pp. 702–725.

[40] Lennart Ljung. "Analysis of recursive stochastic algorithms". In: *IEEE transactions on automatic control* 22.4 (1977), pp. 551–575.

[41] Ricardo Mañé. *Ergodic theory and differentiable dynamics*. Vol. 8. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Translated from the Portuguese by Silvio Levy. Berlin: Springer-Verlag, 1987, pp. xii+317.

[42] John N. Mather. "Action minimizing invariant measures for positive definite Lagrangian systems". In: *Math. Z.* 207.2 (1991), pp. 169–207.

[43] AS Nemirovskii and DB Yudin. *Complexity of problems and efficiency of optimization methods*. Nauka, Moscow, 1979.

[44] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.

[45] Dominikus Noll. "Bundle method for non-convex minimization with inexact subgradients and function values". In: *Computational and Analytical Mathematics*. Springer, 2013, pp. 555–592.

[46] Jacob Palis Junior and Welington de Melo. *Geometric theory of dynamical systems*. Springer-Verlag, 1982.

[47] Boris T. Poljak. "A general method of solving extremum problems". In: *SMD* 8 (1967), pp. 14–29.

[48] Yuri Vasilyevich Prokhorov. "Convergence of random processes and limit theorems in probability theory". In: *SIAM Theory of Probability and its Applications* 1.2 (1956), pp. 157–214.

[49] Rodolfo Rios-Zertuche. "Examples of pathological dynamics of the subgradient method for Lipschitz path-differentiable functions". In: *Mathematics of Operations Research* (2022).

[50] Rodolfo Ríos-Zertuche. "Characterization of minimizable Lagrangian action functionals and a dual Mather theorem". In: *Discrete & Continuous Dynamical Systems – A* 40.5 (2020), pp. 2615–2639.

[51] Adil Salim. "Random monotone operators and application to stochastic optimization". PhD thesis. Université Paris-Saclay (ComUE), 2018.

[52] Naum Zuselevich Shor. "On the structure of algorithms for numerical solution of problems of optimal planning and design". PhD thesis. V.M. Glushkova Cybernetics Institute, 1964.

[53] Stanislav Konstantinovich Smirnov. "Decomposition of solenoidal vector charges into elementary solenoids, and the structure of normal one-dimensional flows". In: *Algebra i Analiz* 5.4 (1993). Translated in: St. Petersburg Math. J. 5 (1994), 841-867, pp. 206–238.

[54] Alfonso Sorrentino. *Action-minimizing Methods in Hamiltonian Dynamics (MN-50): An Introduction to Aubry-Mather Theory*. Princeton University Press, 2015.

[55] Michel Valadier. "Entrainement unilateral, lignes de descente, fonctions lipschitziennes non pathologiques". In: *CRAS Paris* 308 (1989), pp. 241–244.

[56] Shawn Xianfu Wang. "Fine and Pathological Properties of Subdifferentials". PhD thesis. Simon Fraser University, 1999.

[57] Laurence Chisholm Young. "Generalized curves and the existence of an attained absolute minimum in the calculus of variations". In: *Comptes Rendus de la Societe des Sci. et des Lettres de Varsovie* 30 (1937), pp. 212–234.

[58] Laurence Chisholm Young. *Lectures on the calculus of variations and optimal control theory*. Foreword by Wendell H. Fleming. Philadelphia: W. B. Saunders Co., 1969, pp. xi+331.