WORKING PAPERS

N° 1110

June 2020

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

Jérôme Bolte, Edouard Pauwels, and Rodolfo Rios-Zertuche

Toulouse School of Economics

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

Jérôme Bolte, Edouard Pauwels, and Rodolfo Ríos-Zertuche

June 5, 2020

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. 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. This allows us to further analyze the structure of oscillations, and establish their perpendicularity to the general drift.

Contents

1 Introduction **3**

2 Algorithm and framework **5**
2.1 The vanishing step subgradient method 5
2.2 Regularity assumptions on the objective function 6

3 Main results: accumulation, convergence, oscillation compensation **8**
3.1 Asymptotic dynamics for path differentiable functions 8
3.2 Asymptotic dynamics for path differentiable functions with a weak Sard property . . . 9
3.3 Oscillation structure and asymptotics for Whitney stratifiable functions 10
3.4 Further discussion 10

4 A closed measure theoretical approach **11**
4.1 A compendium on closed measures 12
4.2 Interpolant curves of subgradient sequences and their limit measures 16

5 Proofs of main results **20**
5.1 Lemmas on the convergence of curve segments 20
5.2 Proof of Theorem 6 22
5.2.1 Item (i) 22
5.2.2 Item (ii) 22
5.2.3 Item (iii) 23
5.3 Proof of Theorem 7 23
5.3.1 The function is constant on the accumulation set 23
5.3.2 Proof of item (i) 25
5.3.3 Proof of item (ii) 25
5.3.4 Proof of item (iii) 25
5.3.5 Proof of item (iv) 25
5.3.6 Proof of item (v) 25
5.4 Proof of Corollary 9 26
5.5 Proof of Corollary 10 26

A Proof of Theorem 15 **26**

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., [5, 12, 23, 33, 34]. 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 [48]. 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 [26], Poljak [43], Ermol'ev–Shor [25]. It is a quite mature theory, see e.g. [39, 40], which still has a considerable success through the famous mirror descent of Nemirovskii–Yudin [7, 39] and its endless variants. In the nonconvex case, developments of more sophisticated methods were made, see e.g. [32, 35, 41], yet little was known for the raw method until recently.

The work of Davis et al. [21], see also [11], revolving around the fundamental paper of Benaïm–Hofbauer–Sorin [8], brought the first breakthroughs. It relies on a classical idea of Euler: small-step discrete dynamics resemble their continuous counterparts. As established by Ljung [36], this observation can be made rigorous for large times in the presence of good Lyapunov functions. Benaïm–Hofbauer–Sorin [8] 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 [12, 13, 21, 46].

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 [8] and focuses on the delicate question of oscillations, which is illustrated on Figures 1 and 2.

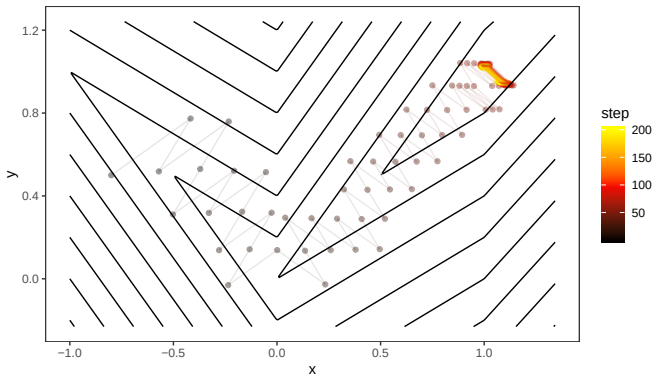

Figure 1: Contour plot of a Lipschitz function with a subgradient sequence. The color reflects the iteration count. The sequence converges to the unique global minimum, but is constantly oscillating.

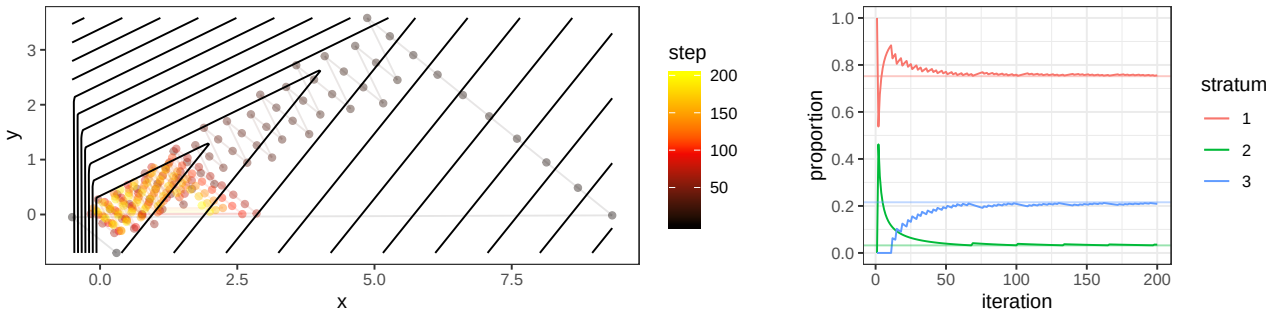

Figure 2: On the left, the contour plot of a convex polyhedral function with three strata, where the gradient is constant. A subgradient sequence starts at $(0.3, -0.7)$ and converges to the origin with an apparent erratic behavior. On the right, we discover that the behavior is not completely erratic. 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 general, as long as the sequence $\{x_i\}_i$ remains bounded, we always have

$$\frac{x_N - x_0}{\sum_{i=0}^N \varepsilon_i} = \frac{\sum_{i=0}^N \varepsilon_i v_i}{\sum_{i=0}^N \varepsilon_i} \to 0, \text{ where } v_i \in \partial^c f(x_i). \tag{1}$$

This fact, that could be called "global oscillation compensation," does not prevent the trajectory to oscillate fast around a limit cycle, as illustrated in [20], 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 [8].

In this work, we adapt the theory of closed measures, which was originally developed in the calculus of variations (see for example [4, 9]), 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 6 and 7 for precise statements, and a discussion in Section 3.4.

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 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 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 [21] 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 (Corollary 22) that complements the point of view of [8].

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 $\mathbb{R}^n \times \mathbb{R}^n$. 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 [27, 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 [18].

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.

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., [14].

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,

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

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., [8] 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 [17, 52]. In particular, generic 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$ satisfy $\partial f \equiv [-1, 1]$ everywhere on $\mathbb{R}$. This means that any absolutely 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 [20] 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*[1] of Valadier [51] (1989), rediscovered in several works, see e.g. [13, 16, 21]. We use the terminology of [13].

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

$$(f \circ \gamma)'(t) = v \cdot \gamma'(t)$$

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 proposed in [13] is not limited to chain rules involving the Clarke subgradient, but it turns out to be equivalent to the a definition very much like the one we give here, with Lipschitz curves replaced by absolutely-continuous curves, the equivalence being furnished by [13, Corollary 2]. The current definition is slightly more general than the original one [13], that is, our class of functions contains the one discussed in [13], because we require a condition only for Lipschitz curves, which are all absolutely continuous.

The class of path differentiable functions is very large and includes many cases of interest, such as functions that are semi-algebraic, tame (definable in an o-minimal structure), or Whitney stratifiable [21] (in particular, 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 considered in the literature), as well as functions that are convex, concave, see e.g., [13, 47].

Whitney stratifiable functions. Due to their ubiquity we detail here the properties of Whitney stratifiability and illustrate their utility. They were first used in [15] in the variational analysis context in order to establish Sard's theorem and Kurdyka-Łojasiewicz inequality for definable functions, two properties which appears to be essential in the study of many subgradient related problems, see e.g., [3, 14].

[1] Literally, "healthy functions" (as opposed to pathological) in French.

Definition 4 (Whitney stratification)**.** Let X be a nonempty subset of $\mathbb{R}^m$ and $p > 0$. A C^p *stratification* $\mathcal{X} = \{X_i\}_{i\in I}$ of X is a locally finite partition of $X = \bigsqcup_i X_i$ into connected submanifolds X_i of $\mathbb{R}^m$ of class C^p such that for each $i \neq j$

$$\overline{X_i} \cap X_j \neq \emptyset \Longrightarrow X_j \subset \overline{X_i} \setminus X_i.$$

A C^p stratification $\mathcal{X}$ of X *satisfies Whitney's condition (a)* if, for each $x \in \overline{X_i} \cap X_j$, $i \neq j$, and for each sequence $\{x_k\}_k \subset X_i$ with $x_k \to x$ as $k \to +\infty$, and such that the sequence of tangent spaces $\{T_{x_k}X_i\}_k$ converges (in the usual metric topology of the Grassmanian) to a subspace $V \subset T_x\mathbb{R}^m$, we have that $T_xX_j \subset V$. A C^p stratification is *Whitney* if it satisfies Whitney's condition (a).

Definition 5 (Whitney stratifiable function)**.** With the same notations as above, a function $f\colon \mathbb{R}^n \to \mathbb{R}^k$ is *Whitney C^p-stratifiable* if there exists a Whitney C^p stratification of its graph as a subset of $\mathbb{R}^{n+k}$.

Examples of Whitney stratifiable functions are semialgebraic or tame functions, but much less structured functions are covered. This class covers most known finite dimensional optimization problems as for instance those met in the training of neural networks. Let us mention here that the subclass of tame functions have led to many results through the nonsmooth Kurdyka–Łojasiewicz inequality, see e.g., [3], while mere Whitney stratifiability combined with the Ljung-like theory developed in [8] has also provided several interesting openings [12, 13, 21].

3 Main results: accumulation, convergence, oscillation compensation

We now present our main, results which rely on three types of increasingly demanding assumptions:

— path differentiability (Section 3.1),

— path differentiable functions with a weak Sard property (Section 3.2),

— Whitney stratifiable functions (Section 3.3).

Section 3.3 also contains a general result pertaining the structure of the oscillations.

The significance of the results is discussed in Section 3.4. The proofs are presented in Section 5.

3.1 Asymptotic dynamics for path differentiable functions

Theorem 6 (Asymptotic dynamics 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 sequence generated by the subgradient method (Definition 1) that remains bounded. Then we have:*

i. (Lengthy separations) Let x and y be two distinct points in $\operatorname{ess\,acc}\{x_i\}_i$ 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 $x_{i'_k} \to y$. Consider

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

Then $T_k \to +\infty$.

ii. (Oscillation compensation) Let $\psi\colon \mathbb{R}^n \to [0,1]$ be a continuous function. Then for every subsequence $\{N_i\}_i \subset \mathbb{N}$ such that

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

we have

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

iii. (Criticality) For all $x \in \operatorname{ess\,acc}\{x_i\}_i$, $0 \in \partial^c f(x)$. In other words, $\operatorname{ess\,acc}\{x_i\}_{i\in\mathbb{N}} \subseteq \operatorname{crit} f$.

3.2 Asymptotic dynamics for path differentiable functions with a weak Sard property

With slightly more stringent hypotheses, which are automatically valid for some important cases of lower or upper-C^k functions [6] (for k sufficiently large), semialgebraic or tame functions [15], we have:

Theorem 7 (Asymptotic dynamics for path differentiable functions: weak Sard case)**.** *In the setting of Theorem 6, and if additionally f is constant on the connected components of its critical set, then we also have:*

i. (Lengthy separations version 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 form each other, that is, $\|x-y\| > 2\delta$. Consider the successive amounts of time it takes for the sequence 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 neighborhoods of $\bar{x} \in \operatorname{acc}\{x_i\}_i$ such that $\overline{U} \subset V$. Let $A \subset \mathbb{N}$ be the union $A = \bigcup_i I_i$ of the maximal intervals $I_i \subset \mathbb{N}$ of the form $I_i = [a_i, b_i] \cap \mathbb{N}$ for some $a_i < b_i$, such that $\{x_i\}_{i\in I_j} \subset U$ and $\{x_i\}_{i\in I_j} \cap V \neq \emptyset$. Then either there is some I_j that is unbounded or

$$\lim_{j\to+\infty} |I_j| = \lim_{j\to+\infty} \sum_{i\in I_j} \varepsilon_i = +\infty.$$

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

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

iv. (Criticality) For all x in the (traditional) accumulation set $\operatorname{acc}\{x_i\}_i$, $0 \in \partial^c f(x)$. *That is to say,* $\operatorname{acc}\{x_i\}_i \subseteq \operatorname{crit} f$.

v. (Convergence of the values) The values $f(x_i)$ converge to a real number as $i \to +\infty$.

Remark 8. Items (iv) and (v) of Theorem 7 can also be deduced from [8, Proposition 3.27] using a different approach. Up to our knowledge, items (i)–(iii) of Theorem 7 as well as Theorem 6 do not have counterparts in the optimization literature.

3.3 Oscillation structure and asymptotics for Whitney stratifiable functions

The two next corollaries express that oscillations happen perpendicularly to the singular set of f, whenever it makes sense. In particular, they are perpendicular to $\operatorname{ess\,acc}\{x_i\}_i$ and $\operatorname{acc}\{x_i\}_i$, respectively, wherever this is well defined.

Corollary 9 (Perpendicularity of the oscillations)**.** *In the setting of Theorem 7 (resp. Theorem 6), let $(x, v) \in \mathbb{R}^n \times \mathbb{R}^n$ be in the accumulation set (resp. essential accumulation set) of $\{(x_i, v_i)\}_i$, and $\alpha\colon (-1, 1) \to \mathbb{R}^n$ be a Lipschitz curve with the property that $\alpha(0) = x, \alpha'(0) = w$, and $t \to (f \circ \alpha)(t)$ is differentiable at $t = 0$ and $(f \circ \alpha)' = v \cdot \alpha'(0)$ for all $v \in \partial^c f(\alpha(0))$. Then*

$$w \cdot v = 0,$$

for all $v \in \partial^c f(x)$. In other words $w \in [\partial^c f(x)]^{\perp}$.

Stratifiable functions (cf. Definition 5) allow to provide much more insight into the oscillation compensation phenomenon: we have seen that substantial oscillations, i.e., those generated by non vanishing subgradients, must be structured orthogonally to the limit point locus. Whitney rigidity then forces the following intuitive phenomenon: substantial bouncing drives the sequence to have limit points lying in the bed of V-shaped valleys formed by the graph of f.

Corollary 10 (Oscillations and V-shaped valleys)**.** *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ is a Whitney C^n stratifiable function, and let x be a point in the accumulation set of a sequence $\{x_i\}_i$ generated by the subgradient method as in Definition 1. Assume that there is a subsequence $x_{i_j} \to x$ with*

$$\limsup_{j \to +\infty} \|v_{i_j}\| > 0,$$

then x is contained in a stratum S of dimension less than n, and if w is tangent to S at x then

$$\lim_{j \to +\infty} w \cdot v_{i_j} = 0.$$

This geometrical setting is reminiscent of the partial smoothness assumptions of Lewis: a smooth path lies in between the slopes of a sharp valley. While proximal-like methods end up in a finite time on the smooth locus [31, 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 their "orthogonality features."

3.4 Further discussion

Theorems 6 and 7 describe the long-term dynamics of the algorithm. While Theorem 6 only talks about what happens close to $\operatorname{ess\,acc}\{x_i\}_i$ and explains only what the most frequent persistent behavior is, Theorem 7 covers all of $\operatorname{acc}\{x_i\}_i$ and hence all recurrent behaviors.

Oscillation compensation. While the high-frequency oscillations (i.e., bouncing) 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 6, $\psi = \psi_{\delta,\eta}\colon \mathbb{R}^n \to [0,1]$ 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 7, $U = B_\eta(x)$ and $V = B_\delta(x)$ the balls centered at x with radius $0 < \eta < \delta$, we obtain this local version of the oscillation cancelation phenomenon: in the setting of Theorem 7 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

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

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 7 tells us that $f(x_i)$ converges, we conjecture that this is no longer true in the context of Theorem 6, 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 Whitney stratifiable 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 [42].

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 6 and items (i) and (ii) of Theorem 7. 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 an open subset of $\mathbb{R}^n$, denote by $C^0(U)$ the set of continuous functions while $C^p(U)$ is the set of $p \in [1,\infty]$ continuously differentiable functions. The set $\mathsf{Lip}(U)$ denotes the space of Lipschitz curves $\gamma\colon \mathbb{R} \to U$. When U is bounded it is endowed with the supremum norm $\|\gamma\|_\infty = \sup_{t\in\mathbb{R}} \|\gamma(t)\|$.

4.1 A compendium on closed measures

General results. 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 [53, 54] in the context of the calculus of variations. It has developed in parallel to the closely related normal currents [28, 29] and varifolds [1, 2], and has found applications in several areas of mathematics, especially Lagrangian and Hamiltonian dynamics [19, 37, 38, 50], and also optimal transport [9, 10].

The definition of closed measures is inspired from the following observations. Given a 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,

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

In other words, 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

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

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

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \nabla f(x) \cdot v\, d\mu_\gamma(x,v) = \int_a^b \nabla f(\gamma(t)) \cdot \gamma'(t)\, dt = \int_a^b (f \circ \gamma)'(t) dt = f \circ \gamma(b) - f \circ \gamma(a) = 0.$$

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

Definition 11 (Closed measure)**.** A compactly-supported, positive, Radon measure μ on $\mathbb{R}^n \times \mathbb{R}^n$ is *closed* if, for all functions $f \in C^\infty(\mathbb{R}^n)$,

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

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 in $\mathbb{R}^n \times \mathbb{R}^n$ 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 [22] implies that there are probability measures μ_x, $x \in \mathbb{R}^n$, on $\mathbb{R}^n$ such that

$$\mu = \int_{\mathbb{R}^n} \mu_x\, d(\pi_*\mu)(x). \tag{4}$$

We shall refer to the couple $(\pi_*\mu, \pi_x)$ as to the *desintegration* of μ. Thus if $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is measurable, we have

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

Definition 12 (Centroid field)**.** Let μ be a positive, compactly-supported, Radon measure on $\mathbb{R}^n \times \mathbb{R}^n$. The *centroid field* $\bar{v}_x$ of μ is, for $x \in \mathbb{R}^n$ and with the decomposition (4),

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

The centroid field gives the average velocity, that is, the average of the velocities encoded by the measure at each point. As a consequence of the disintegration theorem [22], $x \mapsto \bar{v}_x$ 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

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

It plays a significant role in our work. For later use, we record the following facts that follow from the definition of the centroid field, the convexity of $\partial^c f(x)$, and the fact that μ_x is a probability:

Lemma 13 (Quasi-stationary bundle measures)**.** *If a positive Radon measure μ has a centroid field $\bar{v}_x$ that vanishes $\pi_*\mu$-almost everywhere, then μ is closed.*

Proof. Indeed, if $\bar{v}_x = 0$ for $\pi_*\mu$-almost every x, and if $f \in C^\infty(\mathbb{R}^n)$, we have

$$\begin{aligned}
\int_{\mathbb{R}^n \times \mathbb{R}^n} \nabla f(x) \cdot v \, d\mu(x, v) &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \nabla f(x) \cdot v \, d\mu_x(v) \, d(\pi_*\mu)(x) \\
&= \int_{\mathbb{R}^n} \nabla f(x) \cdot \int_{\mathbb{R}^n} v \, d\mu_x(v) \, d(\pi_*\mu)(x) \\
&= \int_{\mathbb{R}^n} \nabla f(x) \cdot \bar{v}_x \, d(\pi_*\mu)(x) = 0,
\end{aligned}$$

so μ is closed. □

Recall that the weak* topology in the space of Radon measures on an open set U is the one induced by the family of seminorms

$$|\mu|_f = \left| \int_U f \, d\mu \right|, \quad f \in C^0(U).$$

Thus a sequence $\{\mu_i\}_i$ of measure converges in this topology to a measure μ if, and only if, for all $f \in C^0(U)$,

$$\int_U f \, d\mu_i \to \int_U f \, d\mu.$$

The following result can be regarded as a consequence of the forthcoming Theorem 15. It can also be seen as a special case of the results of [29] that are very well described in [30, Theorem 1.3.4.6]. Specifically it is shown in [30, Theorem 1.3.4.6] that it is possible to approximate, in a weak* sense, objects (namely, currents) intimately related to closed measures, by simpler objects (namely, closed polyhedral chains), which in our case correspond to combinations of finitely-many piecewise-smooth, closed curves.

Proposition 14 (Weak* density of closed curves)**.** *Consider the set of measures of the form $\beta\mu_\gamma$ for some $\beta > 0$ and a measure $\mu_\gamma = (\gamma, \gamma')_* \mathsf{Leb}_{[a,b]}$ induced by some closed, smooth curve $\gamma\colon [a, b] \to \mathbb{R}^n$, $\gamma(a) = \gamma(b)$, defined on an interval $[a, b] \subset \mathbb{R}$ (which is not fixed). In the weak* topology, this set is dense in the set of closed measures.*

Since the space of measures is sequential, this proposition means that for any closed measure μ, we can find a sequence of closed curves $\gamma_1, \gamma_2, \dots$ that approximate μ in the sense that $\mu_{\gamma_i} \to \mu$ in the weak* topology.

The following result, known as the *Young superposition principle* [9, 53] or as the *Smirnov solenoidal representation* [4, 49], is a strong refinement of the assertion of Proposition 14; see also [45, 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

$$\int \delta_{(x,\bar{v}_x)}\, d(\pi_*\mu)(x),$$

which captures much of the properties of μ, can be decomposed into a combination of measures induced by Lipschitz curves. This decomposition is very useful theoretically, as there are no limits involved. For completeness, the following is proved in Section A.

Theorem 15 (Young superposition principle/Smirnov solenoidal representation)**.** *Let U be a nonempty bounded open subset of $\mathbb{R}^n$ and set $\mathsf{Lip}(U) = \mathsf{Lip}$. For $t \in \mathbb{R}$, let $\tau_t \colon \mathsf{Lip} \to \mathsf{Lip}$ be the time-translation $\tau_t(\gamma)(s) = \gamma(s+t)$. For every closed probability measure μ supported in U with centroid field $\bar{v}_x$, there is a Borel probability measure ν on the space Lip that is invariant under τ_t for all $t \in \mathbb{R}$ and such that*

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

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

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

Corollary 16 (Centroid representation)**.** *With the notation of the previous theorem, we have for ν almost all γ in Lip:*

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

for almost all t.

Proof. Take indeed $\phi \geqslant 0$ vanishing only on the measureable 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 15 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 Lip.

The measure ν in Theorem 15 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) satisfies, for all $t \in \mathbb{R}$,

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

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

$$\int_{\mathbb{R}^n} \phi(x, \bar{v}_x)\, d(\pi_*\mu)(x) = \frac{1}{|I|} \int_{\mathsf{Lip}} \int_I \phi(\gamma(t), \gamma'(t))\, dt\, d\nu(\gamma) \tag{8}$$

for any 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 15 is not unique in general. For example, if γ is a closed curve intersecting itself once so as to form the figure 8, then the measure ν decomposing $\mu = \mu_\gamma$ could 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.

Circulation for a subdifferential field. We provide here some results related to subdifferentials, and that will be useful to the study of the vanishing step subgradient method.

Lemma 17. *Let f be a locally Lipschitz continuous function and μ a closed measure with desintegration $(\pi_*\mu, \mu_x)$ and centroid field $\bar{v}_x$. If for some $a \in \mathbb{R}$ and some $x \in \mathbb{R}^n$ we have $a \operatorname{supp} \mu_x \subset \partial^c f(x)$, then $a\bar{v}_x \in \partial^c f(x)$.*

Proof. Assume $a \operatorname{supp} \mu_x \subset \partial^c f(x)$. Let $g(v) = \operatorname{dist}(av, \partial^c f(x))$, so that $g(v) = 0$ for all $v \in \operatorname{supp} \mu_x$. Since $\partial^c f(x)$ is a convex set, g is a convex function. Then by Jensen's inequality we have

$$g(\bar{v}_x) = g\left(\int_{\mathbb{R}^n} v\, d\mu_x(v)\right) \leqslant \int_{\mathbb{R}^n} g(v)\, d\mu_x(v) = 0. \qquad \square$$

Proposition 18 (Circulation of subdifferential for path differentiable functions)**.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ 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 $\sigma(x) \in \partial^c f(x)$ 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

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

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

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

Proof. Let $\sigma_1, \sigma_2 \colon U \times \mathbb{R}^n \to \mathbb{R}$ be two measurable functions such that $\sigma_i(x) \in \partial^c f(x)$ for each $x \in U$. From Theorem 15 we get a τ_t-invariant, Borel probability measure ν on the space Lip of Lipschitz curves. Then

$$\begin{aligned}\int_{U\times\mathbb{R}^n} \sigma_1(x)\cdot v\, d\mu(x,v) - \int_{U\times\mathbb{R}^n} \sigma_2(x)\cdot v\, d\mu(x,v) \\ = \int_{\mathbb{R}^n\times\mathbb{R}^n} \chi_U(x)(\sigma_1(x)-\sigma_2(x))\cdot v\, d\mu(x,v) \\ = \int_{\mathsf{Lip}} \chi_U(\gamma(0))(\sigma_1(\gamma(0))-\sigma_2(\gamma(0)))\cdot\gamma'(0)\, d\nu(\gamma).\end{aligned}$$

Since f is path differentiable, for each $\gamma \in \mathsf{Lip}$ 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 in which $\pi(\operatorname{supp}\mu) \subset U$. 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 $\psi_r * f$ the convolution of ψ_r and f. Observe that if $\beta \in \mathsf{Lip}$ and $a < b$, then

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

This justifies the following calculation:

$$\begin{aligned}\int_{T\mathbb{R}^n} \partial^c f\, d\mu = \int_{\mathsf{Lip}} (f\circ\beta)'(0)\, d\nu(\beta) = \\ = \lim_{r\searrow 0}\int_{\mathsf{Lip}} ((\psi_r * f)\circ\beta)'(0)\, d\nu(\beta) \\ = \lim_{r\searrow 0}\int_{\mathsf{Lip}} \nabla(\psi_r * f)(\beta(0))\cdot\beta'(0)\, d\nu(\beta) \\ = \lim_{r\searrow 0}\int_{T\mathbb{R}^n} \nabla(\psi_r * f)(x)\cdot v\, d\mu(x,v),\end{aligned}$$

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

4.2 Interpolant curves of subgradient sequences and their limit measures

In this section we fix $f \colon \mathbb{R}^n \to \mathbb{R}$ to be a locally Lipschitz function, and let $\{x_i\}_i$ be a bounded sequence generated by the subgradient method.

Definition 19 (Subgradient sequence interpolants)**.** Given a sequence $\{x_i\}_{i\in\mathbb{N}}$ generated by the subgradient method, with the same notations as in Definition 1, its *interpolating curve* is the 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 $t_i < t < t_{i+1}$. This curve corresponds to a continuous-time piecewise-affine interpolation of the sequence.

For a bounded 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 20 (Limiting closed measures associated to subgradient sequences)**.** *Let γ be the interpolating curve (as in Definition 19) and $A = \{I_i\}_{i\in\mathbb{N}}$ 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

$$\left\{\mu_{\gamma|_{B_N}} + \frac{\sum_{i=0}^N \nu_i}{|B_N|}\right\}_N \tag{9}$$

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,

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

and the measures in the sequence (9) are sums of multiples of these.

By Prokhorov's theorem [44], 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. □

Lemma 21 (Limit points and limiting measure supports)**.** *Let γ be the interpolating curve as in Definition 19. 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*

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

Proof. Let $B\subset\mathbb{R}^n$ be a closed ball containing a neighborhood of the sequence $\{x_i\}_i$. 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 $\|x-y\|\leqslant\varepsilon_i\operatorname{Lip}(f)$ 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 $S>n_0$,

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

Assume $x\in\operatorname{ess\,acc}\{x_i\}_i\subset B$. Take a nonnegative continuous function ψ, as above, so that for all $R>0$ there is $S>R$ such that

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

Then, for $\varepsilon=\delta/2$, n_0 as above, and $S>R>n_0$,

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

It follows that there is 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 positive, continuous function ψ with $x\in\pi(\operatorname{supp}\psi)$, 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 corollary gives some connection between the discrete and the continuous subgradient systems.

Corollary 22 (Limiting dynamics)**.** *Let $\{I_i\}_i$ be a sequence of disjoint, bounded intervals in $\mathbb{R}$ with*

$$\lim_{i\to+\infty}|I_i|=+\infty.$$

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

$$-\beta'(t) \in \partial^c f(\beta(t))$$

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

Proof. The existence of ν follows from Theorem 15. By Corollary 16, we know that ν-almost every curve $\beta \in \mathsf{Lip}$ satisfies, $\beta'(t) = \bar{v}_{\beta(t)}$ for almost every t. So we just need to prove that $\bar{v}_x \in -\partial^c f(x)$ for $\pi_*\mu$-almost every $x \in \mathbb{R}^n$.

Recall that $\operatorname{graph} -\partial^c f = \{(x, v) \in \mathbb{R}^n \times \mathbb{R}^n : -v \in \partial^c f(x)\}$. 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 $\partial^c f(\gamma(t_i))$, we have

$$\begin{aligned}
&\operatorname{dist}((\gamma(t), \gamma'(t)), \operatorname{graph} -\partial^c f) \\
&\quad \leqslant \|(\gamma(t), \gamma'(t)) - (\gamma(t_i), -v_i)\| + \operatorname{dist}((\gamma(t_i), -v_i), \operatorname{graph} -\partial^c f) \\
&\quad = \|(\gamma(t), -v_i) - (\gamma(t_i), -v_i)\| + 0 \\
&\quad = \|\gamma(t) - \gamma(t_i)\| \\
&\quad \leqslant \operatorname{Lip}(\gamma)\varepsilon_i \\
&\quad \leqslant \operatorname{Lip}(f)\varepsilon_i.
\end{aligned}$$

Now

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

This implies that

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

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

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \operatorname{dist}((x, v), \operatorname{graph} -\partial^c f)\, d\mu(x, v) = 0$$

because the convergence of measures occurs in the weak* topology and the integrand is continuous. Since $\operatorname{graph} -\partial^c f$ is a closed set, the support of μ must be contained in it. From Lemma 17 with $a = -1$, we know that $-\bar{v}_x \in \partial^c f(x)$, which is what we wanted to prove. □

Theorem 23 (Subgradient-like closed measures are trivial)**.** *Assume that $f\colon \mathbb{R}^n \to \mathbb{R}$ is a path differentiable function. Let μ be a closed measure on $\mathbb{R}^n \times \mathbb{R}^n$, and assume that every $(x, v) \in \operatorname{supp} \mu$ satisfies $-v \in \partial^c f(x)$. Then the centroid field $\bar{v}_x$ of μ vanishes for $\pi_*\mu$-almost every x.*

Proof. The condition on μ implies, by Lemma 17 with $a = -1$, that $-\bar{v}_x \in \partial^c f(x)$. By Proposition 18 we may choose $\sigma(x) = -\bar{v}_x$ to compute

$$\int \partial^c f \, d\mu = \int \sigma(x) \cdot v \, d\mu(x,v) = \int \sigma(x) \cdot \left[\int v \, d\mu_x \right] d(\pi_* \mu)(x)$$
$$= \int \sigma(x) \cdot \bar{v}_x \, d(\pi_* \mu)(x) = - \int \bar{v}_x \cdot \bar{v}_x \, d(\pi_* \mu)(x).$$

Proposition 18 also implies that the left-hand side vanishes because μ is closed. □

5 Proofs of main results

5.1 Lemmas on the convergence of curve segments

Lemma 24. *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 satisfies*

$$\lim_{i \to +\infty} \int_0^{T_i} \operatorname{dist}((\gamma_i(t), \gamma_i'(t)), \operatorname{graph} -\partial^c f) \, dt = 0. \tag{10}$$

Then $-\gamma'(t) \in \partial^c f(\gamma(t))$ for almost all $t \in [0,T]$.

Proof. We follow classical arguments; see for example [8, 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 [24, 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

$$\sum_{k=p}^{N(p)} a(p,k) \gamma_k' \to \gamma' \tag{11}$$

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

Since the Clarke subdifferential $\partial^c f(x)$ is convex at each x, the function

$$g(x,v) = \operatorname{dist}(-v, \partial^c f(x))$$

is convex in its second argument for fixed $x \in \mathbb{R}^n$. Using the fact that the convergence (11) 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,

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

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

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

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 (10). Hence we have $-\gamma'(t) \in \partial^c f(\gamma(t))$ for almost all $t \in [0, T']$, and this proves the lemma since T' was taken arbitrary in $(0, T)$. □

Lemma 25. *Let γ be the interpolant curve of the bounded gradient sequence $\{x_i\}_i$, and let $\{I_j\}_j$ be 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*

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

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 10 holds if we let γ_i be the appropriate translate of $\gamma|_{I_i}$, so by Lemma 24, $-\bar\gamma'(t) \in \partial^c f(\bar\gamma(t))$ for almost every $t \in [a,b]$. By the path differentiability of f, we have

$$\begin{aligned}-\int_a^b \|\bar\gamma'(t)\|^2 dt &= \int_a^b \partial^c f(\bar\gamma(t))\cdot\bar\gamma'(t)dt \\ &= \int_a^b (f\circ\bar\gamma)'(t)\,dt = f\circ\bar\gamma(b) - f\circ\bar\gamma(a).\end{aligned}$$

□

5.2 Proof of Theorem 6

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 $\bar T_k = t_{i'_k} - t_{i_k}$ remain bounded. Apply Lemma 25 to obtain a curve $\bar\gamma\colon [a,b] \to \mathbb{R}^n$ joining $\bar\gamma(a) = \lim_k x_{i_k} = x$ and $\bar\gamma(b) = \lim_k x_{i'_k} = y$. So 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 25,

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

Whence we get the contradiction we were aiming for.

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 $|\psi(x) - \psi(y)| \leqslant \varepsilon$. 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

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

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

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} v\psi(x)\, d\mu(x,v) = \int_{\mathbb{R}^n} \bar{v}_x \psi(x)\, d(\pi_*\mu)(x) = 0, \tag{12}$$

where $\bar{v}_x$ is the centroid field of μ. By Lemma 20 we know that μ is closed so that Theorem 23 applies which gives $\bar{v}_x = 0$ for $\pi_*\mu$-almost every x. This immediately implies (12).

5.2.3 Item (iii)

To prove item (iii), consider the interpolation curve constructed in Section 4.2. Consider a limit point μ of the sequence $\{\mu_{\gamma|_{[0,N]}}\}_N$. By Lemma 20, μ is closed. By Theorem 23, the centroid field $\bar{v}_x$ of μ vanishes for $\pi_*\mu$-almost every x, so from Lemma 17 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 21 we know that it is true throughout $\operatorname{ess\,acc}\{x_i\}_i$.

5.3 Proof of Theorem 7

5.3.1 The function is constant on the accumulation set

Lemma 26. *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

$$W_1 = f^{-1}(-\infty, c_1) \quad \text{and} \quad W_2 = f^{-1}(c_2, +\infty).$$

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

$$\int_{t_1^j}^{t_2^j} \partial^c f(\gamma(t)) \cdot \gamma'(t)\, dt = \int_{t_1^j}^{t_2^j} (f \circ \gamma)'(t)\, dt = f \circ \gamma(t_2^j) - f \circ \gamma(t_1^j) = c_2 - c_1. \tag{13}$$

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 we 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 20 and Theorem 23 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,

$$|I_i| < \beta.$$

By (13) we have

$$\int \partial^c f\, d\mu_{\gamma|_{\cup_{i=0}^N I_i}} = \frac{N(c_2 - c_1)}{\sum_{i=0}^N |I_i|}$$

and

$$0 < \frac{c_2 - c_1}{\beta} \leqslant \frac{N(c_2 - c_1)}{\sum_{i=0}^N |I_i|} \leqslant \frac{c_2 - c_1}{\alpha}.$$

Whence we also have at the limit

$$\int \partial^c f\, d\mu \geqslant \frac{c_2 - c_1}{\beta} > 0. \tag{14}$$

Let $\bar{v}_x$ denote the centroid velocity vector field for μ. By the construction of $\{x_i\}_i$, the vector $-\bar{v}_x$ is contained in the Clarke subdifferential of each point of $\operatorname{supp} \pi_*\mu$, and since the path-differentiability of f allows us to choose any representative of this differential, we have, as in the proof of Theorem 23,

$$\int \partial^c f\, d\mu = -\int \bar{v}_x \cdot \bar{v}_x\, d\pi_*\mu \leqslant 0,$$

which contradicts (14). □

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 $\|\gamma'\| \leqslant \mathrm{Lip}(f)$, 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 25 to obtain a curve $\bar{\gamma}\colon [a,b] \to \mathbb{R}$ such that $\bar{\gamma}(a) \in B_\delta(x) \cap \mathrm{acc}\{x_i\}_i$ and $\bar{\gamma}(b) \in B_\delta(y) \cap \mathrm{acc}\{x_i\}_i$, while also satisfying

$$-\int_a^b \|\bar{\gamma}'(t)\|^2 dt = f \circ \bar{\gamma}(b) - f \circ \bar{\gamma}(a) = 0 \tag{15}$$

by Lemma 26. 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 $x \in \overline{U} \cap \mathrm{acc}\{x_i\}_i$ and some subsequence $\{i_j\}_j$ such that $\mathrm{dist}(x, \gamma(I_{i_j})) \to 0$ and $|I_{i_j}| \leqslant C$ for some $C > 0$ and all $j \in \mathbb{N}$.

We may thus apply Lemma 25 to get a curve $\bar{\gamma} : [a,b] \to \mathbb{R}^n$ whose endpoints $\bar{\gamma}(a)$ and $\bar{\gamma}(b)$ are contained in $\mathrm{acc}\{x_i\}_i \setminus V$, and $\bar{\gamma}$ passes through $x \in \overline{U}$, so it has positive arc length. However, it is also a conclusion of Lemma 25, together with Lemma 26, that $\bar{\gamma}$ satisfies (15), which makes it impossible for its arc length to postive, 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 $B = \cup_{i \in A}[t_i, t_{i+1})$. The statement of item (iii) is equivalent to the statement that

$$\lim_{N \to +\infty} \int v \, d\mu_{\gamma|_{B \cap [0,N]}} = 0. \tag{16}$$

It follows from item (ii) that the maximal intervals $I_i \subset \mathbb{R}$ comprising $B = \bigcup_i I_i$ satisfy $|I_i| \to +\infty$. Hence, from Lemma 20 we know that any limit point μ of the sequence $\{\mu_{\gamma|_{B\cap[0,N]}}\}_N$ is closed, and from Theorem 23 we know that the centroid field of μ vanishes $\pi_*\mu$-almost everywhere, which implies (16).

5.3.5 Proof of item (iv)

Let $x \in \mathrm{acc}\{x_i\}_i$. 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 Lemma 17 we know that the centroid field is contained in the Clarke subdifferential. 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 $\mathrm{acc}\{x_i\}_i$ is connected. We know from item (iv) that $\mathrm{acc}\{x_i\}_i \subseteq \mathrm{crit}\, f$. So it is contained in a single connected component of $\mathrm{crit}\, f$. Hence f must be constant on $\mathrm{acc}\{x_i\}_i$, and $\{f(x_i)\}_i$ converges.

5.4 Proof of Corollary 9

Let α be as in the statement of the corollary. By item (iii) of Theorem 6 or of Theorem 7, we know that $0 \in \partial^c f(\alpha(0))$. Since $-v_i \in \partial^c f(x_i)$ and because the graph of $\partial^c f$ is closed in $\mathbb{R}^n \times \mathbb{R}^n$, we have that $-v \in \partial^c f(\alpha(0))$. 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

$$w \cdot \alpha'(0) = (f \circ \alpha)'(0) = 0 \cdot \alpha'(0) = 0.$$

5.5 Proof of Corollary 10

Let K be a compact set that contains $\{x_i\}_i$ in its interior. By the Morse–Sard theorem applied independently on each stratum of the stratification of f, $f(\operatorname{crit} f)$ 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 7. From item (iv) of Theorem 7 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 must be contained in a stratum of smaller dimension. The last statement of the corollary follows from Corollary 9.

A Proof of Theorem 15

We follow the exposition of [4].

We remark at the outset that if ν is a probability measure on Lip that is τ_t-invariant for all $t \in \mathbb{R}$, then in view of (7), for $\varphi \in C^\infty(\mathbb{R}^n)$,

$$\begin{aligned}\int_{\mathsf{Lip}} \nabla\varphi(\gamma(0)) \cdot \gamma'(0)\, d\nu(\gamma) &= \int_0^1 \int_{\mathsf{Lip}} \nabla\varphi(\gamma(t)) \cdot \gamma'(t)\, d\nu(\gamma)\, dt \\ &= \int_{\mathsf{Lip}} \int_0^1 (\varphi \circ \gamma)'(t)\, dt\, d\nu(\gamma) = \int_{\mathsf{Lip}} (\varphi \circ \gamma(1) - \varphi \circ \gamma(0))\, d\nu(\gamma) = 0.\end{aligned}$$

Thus the probability measure induced by ν on $\mathbb{R}^n \times \mathbb{R}^n$ by pushing forward through the map $\gamma \mapsto (\gamma(0), \gamma'(0))$ (as in the right-hand side of (6)) is automatically closed.

Smooth case. Assume first that there is a C^∞ compactly-supported vector field $X\colon U \to \mathbb{R}^n$ such that μ is given by $\delta_{(x,X(x))} \otimes \rho(x)\mathsf{Leb}_U(x)$, with some smooth probability density $\rho\colon \mathbb{R}^n \to \mathbb{R}_{\geqslant 0}$,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \phi\, d\mu = \int_{\mathbb{R}^n} \phi(x, X(x))\, \rho(x)\, dx,$$

for measurable $\phi\colon \mathbb{R}^n \to \mathbb{R}$. For μ, the centroid field is $\bar{v}_x = X(x)$. Without loss of generality we may assume that X vanishes in a neighborhood of the boundary ∂U.

Denote by $\Phi\colon U \times \mathbb{R} \to U$ the flow of X, so that, for all $t \in \mathbb{R}$ and writing $\Phi_t(x) = \Phi(x, t)$,

$$\Phi_0(x) = x \quad \text{and} \quad \frac{d}{dt}\Phi_t(x) = X(\Phi_t(x)).$$

Since $\operatorname{supp} X$ is compact, by the Picard–Lindelöf theorem we know that $\Phi_t(x)$ is defined for all $t \in \mathbb{R}$ for all $x \in U$. The measure μ is Φ_t-invariant because, integrating by parts, we get that for all $\varphi \in C^\infty(\mathbb{R}^n)$,

$$0 = \int_{\mathbb{R}^n \times \mathbb{R}^n} \nabla\varphi(x) \cdot v\, d\mu(x, v) = \int_{\mathbb{R}^n} \nabla\varphi(x) \cdot X(x)\, \rho(x)\, dx = -\int \varphi(x) \operatorname{div}(\rho X)(x)\, dx,$$

so ρX is a divergence-free vector field; thus, the flow Φ_t of X preserves ρ.

For $L > 0$, let $\mathsf{Lip}_L \subset \mathsf{Lip}$ be the set of Lipschitz curves γ with Lipschitz constant at most L. Then

$$\Gamma = \{\gamma\colon \mathbb{R} \to U |\ \gamma(t) = \Phi_t(x) \text{ for some } x \in U \text{ and all } t \in \mathbb{R}\} \subset \mathsf{Lip}_{\|X\|_\infty}$$

and Γ is a Borel subset of Lip because it can be expressed as the countable intersection of unions of the closed balls around a dense subset of Γ.

Let $\mathrm{ev}\colon \mathsf{Lip} \to \mathbb{R}^n$ be the evaluation at 0, namely, $\mathrm{ev}(\gamma) = \gamma(0)$. Denote by $\mathrm{ev}|_\Gamma^{-1}\colon U \to \Gamma$ the inverse of the one-to-one map that results from restricting ev to Γ. For $x \in U$, $\mathrm{ev}\,|_\Gamma^{-1}(x)$ is exactly the curve $\beta \in \Gamma$ given by $\beta(t) = \Phi_t(x)$ that satisfies, in particular, $\mathrm{ev}(\beta) = \beta(0) = x$ and $\beta'(0) = X(x)$.

Let ν be the pushforward

$$\nu = (\mathrm{ev}|_\Gamma^{-1})_*(\rho(x)\mathsf{Leb}_{\mathbb{R}^n}(x))$$

so that, for measurable $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$\int_{\mathsf{Lip}} \phi(\gamma(0), \gamma'(0))\, d\nu = \int_{\mathbb{R}^n \times \mathbb{R}^n} \phi(x, X(x))\rho(x)\, dx.$$

The measure ν is supported in Γ and, being the pushforward of a probability, it is a probability as well.

General case. Let μ be an arbitrary closed probability measure on $\mathbb{R}^n$. Let $L > 0$ be such that if $(x, v) \in \operatorname{supp}\mu$ then $\|v\| \leqslant L$. Let U be a bounded, open subset of $\mathbb{R}^n$ that contains $\operatorname{supp}\mu$ and satisfies $\operatorname{dist}(\operatorname{supp}\mu, \partial U) \geqslant 1$. Let $\psi\colon \mathbb{R}^{2n} \to \mathbb{R}$ be a mollifier, that is, a C^∞, compactly supported, radially symmetric $\psi(x) = \psi(\|x\|)$, nonnegative function with $\int_{\mathbb{R}^n} \psi = 1$ and $\operatorname{supp}\psi \subseteq B_1(0) \subset \mathbb{R}^n$, and let $\psi_r(x) = r^{-2n}\psi(x/r)$ for $0 < r < 1$. The probability measure $\psi_r * \mu$ is smooth and compactly supported; in fact,

$$\operatorname{supp}\psi_r * \mu \subset U \times \{v \in \mathbb{R}^n : \|v\| \leqslant L + r\}.$$

Denote $\bar{v}^\eta$ the centroid field of the measure η. Then the centroid field and the projected densities of the convolution, $\bar{v}^{\psi_r * \mu}$ and $\pi_*(\psi_r * \mu)$, are smooth and converge to $\bar{v}^\mu$ and $\pi_*\mu$, respectively, as $r \searrow 0$.

Analogously to the definition of Γ in the smooth case, let Γ_r be the subset of Lip_{L+r} that consists of all flow lines of $\bar{v}^{\psi_r * \mu}$ that are defined on all of $\mathbb{R}$, and let

$$\nu_r = (\mathrm{ev}\,|_{\Gamma_r}^{-1})_*(\pi_*(\psi_r * \mu)(x)\mathsf{Leb}_{\mathbb{R}^n}(x))$$

so that, for measurable $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$\int_{\mathrm{Lip}} \phi(\gamma(0), \gamma'(0))\, d\nu_r(\gamma) = \int_{\mathbb{R}^n \times \mathbb{R}^n} \phi(x, \bar{v}_x^{\psi_r * \mu})\, d(\pi_*(\psi_r * \mu))(x).$$

The probability measure ν_r is supported in the set Lip_{L+r}.

The set Lip_{L+1}, which contains Lip_{L+r} for $0 \leqslant r < 1$, is sequentially compact. Indeed, if we have a family $\{\gamma_i\}_{i \in I} \subset \mathsf{Lip}_{L+1}$, then it is equibounded (as the image of each curve is contained in the bounded set U) and equicontinuous (because all its members have Lipschitz constant at most $L + 1$), so by the Arzelà–Ascoli theorem we can extract a subsequence $\{\gamma_i^1\}_{i \in \mathbb{N}}$ that converges in the interval $[-1, 1]$. We then produce, by induction, a sequence of subsequences: assuming we already extracted a subsequence $\{\gamma_i^j\}_i$ that converges in $[-j, j]$, the Arzelà–Ascoli theorem tells us that there is a further subsequence $\{\gamma_i^{j+1}\}_i \subseteq \{\gamma_i^j\}_i$ of curves that converge in the interval $[-j-1, j+1]$. We then pick the diagonal sequence $\{\gamma_i^i\}_i$, which converges throughout $\mathbb{R}$ to a curve in Lip_{L+1}.

Since it is also metrizable with $\operatorname{dist}(\gamma_1, \gamma_2) = \|\gamma_1 - \gamma_2\|_\infty$, Lip_{L+1} is also compact. Prokhorov's theorem [44] implies that there is a weakly convergent sequence $\{\nu_{r_i}\}_i \subset \mathsf{Lip}_{L+1}$ with $r_i \searrow 0$. It is then a routine procedure to check that the limit probability measure $\nu = \lim_i \nu_{r_i}$ satisfies (6).

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 and ANR OMS.

References

[1] William K. Allard. "On the first variation of a varifold". In: *Ann. of Math. (2)* 95 (1972), pp. 417–491. ISSN: 0003-486X.

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

[3] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods". In: *Mathematical Programming* 137.1-2 (2013), pp. 91–129.

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

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

[6] Luc Barbet et al. "Sard theorems for Lipschitz functions and applications in optimization". In: *Israël Journal of Mathematics* 212.2 (2016), pp. 757–790.

[7] 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.

[8] 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.

[9] Patrick Bernard. "Young measures, superposition and transport". In: *Indiana Univ. Math. J.* 57.1 (2008), pp. 247–275. ISSN: 0022-2518. DOI: 10.1512/iumj.2008.57.3163. URL: http://dx.doi.org/10.1512/iumj.2008.57.3163.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] Jérôme Bolte et al. "Clarke Subgradients of Stratifiable Functions". In: *SIAM Journal on Optimization* 18.2 (Jan. 2007), pp. 556–572. DOI: 10.1137/060670080. URL: https://doi.org/10.1137/060670080.

[16] 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.

[17] 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.

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

[19] 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. ISBN: 85-244-0151-6.

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

[21] Damek Davis et al. "Stochastic Subgradient Method Converges on Tame Functions". In: *Foundations of Computational Mathematics* (Jan. 2019). ISSN: 1615-3383. DOI: 10.1007/s10208-018-09409-5. URL: https://doi.org/10.1007/s10208-018-09409-5.

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

[23] 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.

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

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

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

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

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

[29] Herbert Federer. "Real flat chains, cochains and variational problems". In: *Indiana Univ. Math. J.* 24 (1974), pp. 351–407. ISSN: 0022-2518.

[30] Mariano Giaquinta, Giuseppe Modica, and Jiří Souček. *Cartesian currents in the calculus of variations. II*. Vol. 38. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]. Variational integrals. Springer-Verlag, Berlin, 1998, pp. xxiv+697. ISBN: 3-540-64010-X. DOI: 10.1007/978-3-662-06218-0. URL: http://dx.doi.org/10.1007/978-3-662-06218-0.

[31] 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.

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

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

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

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

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

[37] 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. ISBN: 3-540-15278-4.

[38] John N. Mather. "Action minimizing invariant measures for positive definite Lagrangian systems". In: *Math. Z.* 207.2 (1991), pp. 169–207. ISSN: 0025-5874. DOI: 10.1007/BF02571383. URL: http://dx.doi.org/10.1007/BF02571383.

[39] AS Nemirovskii and DB Yudin. *Complexity of problems and efficiency of optimization methods*. 1979.

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

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

[42] Jacob Palis Junior and Welington de Melo. *Geometric theory of dynamical systems*. Springer-Verlag, 1982. ISBN: 978-1-4612-5705-9.

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

[44] Yu.V. 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. DOI: 10.1137/1101016. URL: https://doi.org/10.1137/1101016.

[45] 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. DOI: 10.3934/dcds.2020143. URL: http://aimsciences.org//article/id/a9da8fc5-0029-4a47-93dc-56f46759edba.

[46] Adil Salim. "Random monotone operators and application to stochastic optimization". PhD thesis. Paris Institute of Technology, 2018.

[47] Stefan Scholtes. *Introduction to piecewise differentiable equations*. Springer Science & Business Media, 2012.

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

[49] 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.

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

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

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

[53] L. C. 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.

[54] 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.