# Long term dynamics of small step algorithms

Rodolfo Ríos-Zertuche

Artificial and Natural Intelligence
Toulouse Institute

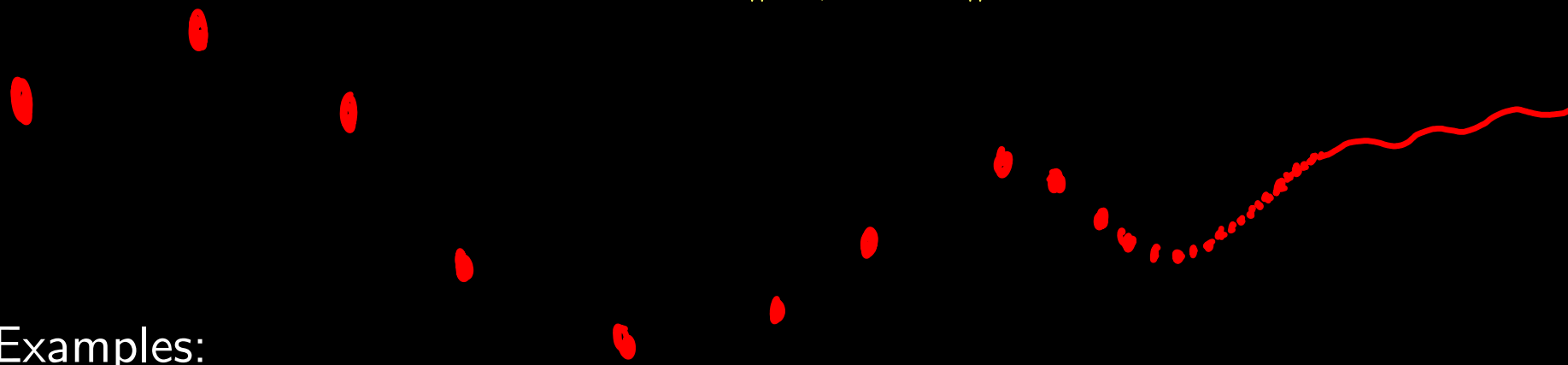
*Joint work with Jérôme Bolte and Edouard Pauwels*

BrainPOP March 8, 2021

# Discrete processes with vanishing step size

Discrete process with vanishing step size

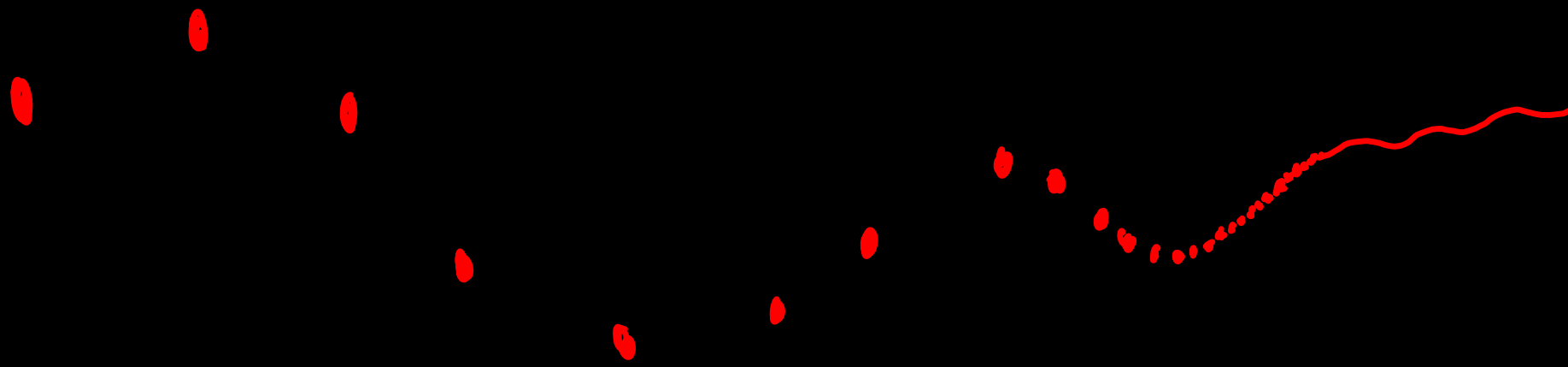
$x_1, x_2, \ldots$ $\qquad\qquad$ $\|x_{i+1} - x_i\| \to 0$



Examples:

- subgradient method of Shor
- stochastic subgradient method
- adaptive gradient, Adam, AdaDelta, AdaGrad
- inertial subgradient / heavyball
- subdifferential inclusions
- Lyapunov function analysis
- games

# How to study this?



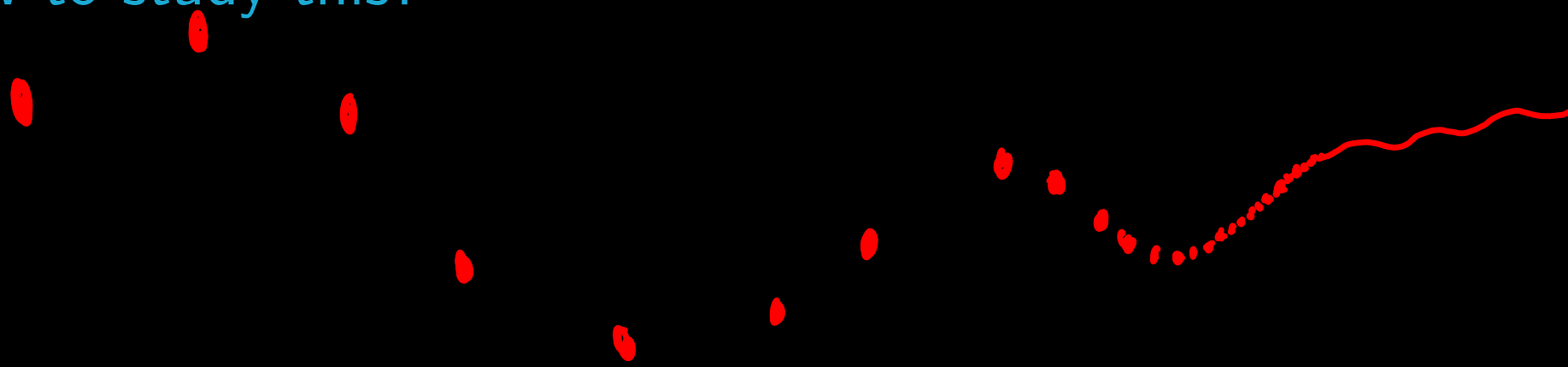
**Idea 1**

*Look at the accumulation behavior of* $\{x_i\}_i$.

This is hard!

- Not the usual dynamical system (the rule changes at each step), so tools like ergodic theory are not available.
- The points $\{x_i\}_i$ don't carry much information about their relationship to each other.

# How to study this?



**Idea 2**

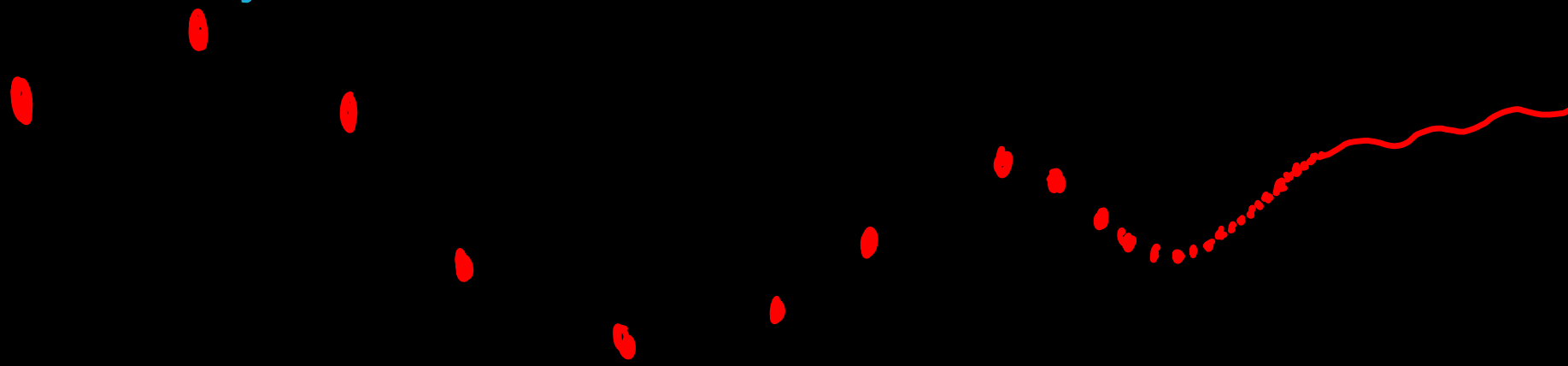
*Study the accumulation measures*

$$\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}.$$

Still hard!

- Still not enough information, no tools from other parts of the theory.
- Get only statistical results about the mean.

# How to study this?



**Idea 3 (Ljung)**

*Study the continuous flow*

$$\gamma'(t) \in \mathcal{F}(\gamma(t))$$

*for a multivalued vector field* $\mathcal{F} \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$.

- Works well when $\mathcal{F}$ is regular.
- Breaks down when $\mathcal{F}$ causes bouncing.

# Study the continuous flow

We get a lot of insight from the continuous flow though:

- *Palis–de Melo example*
  Even the continuous gradient flow may not converge.

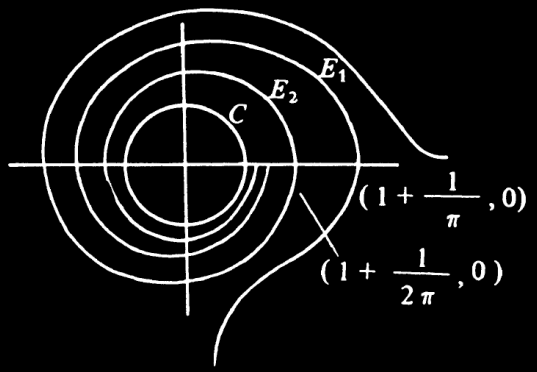


Figure 7

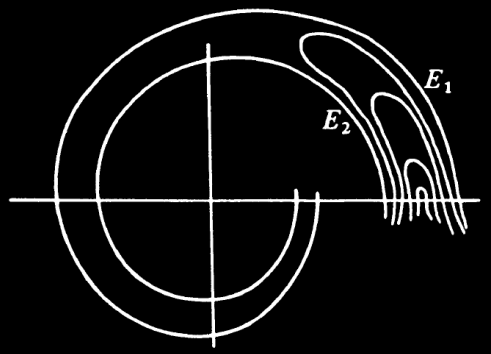
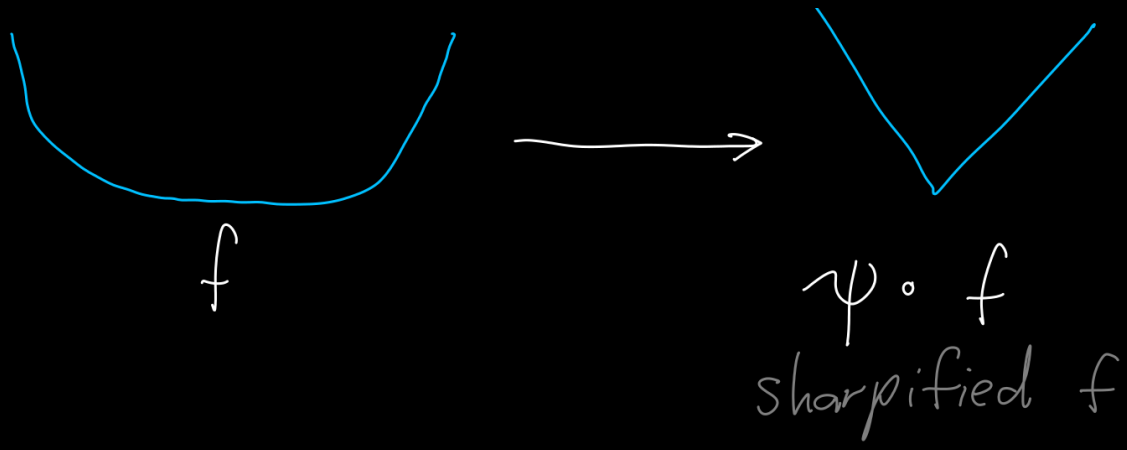


Figure 8

- *Kurdyka-Łojasiewicz theory* $\|\nabla(\psi \circ f)\| \geq 1$
  If we can "sharpify," the gradient flow converges.

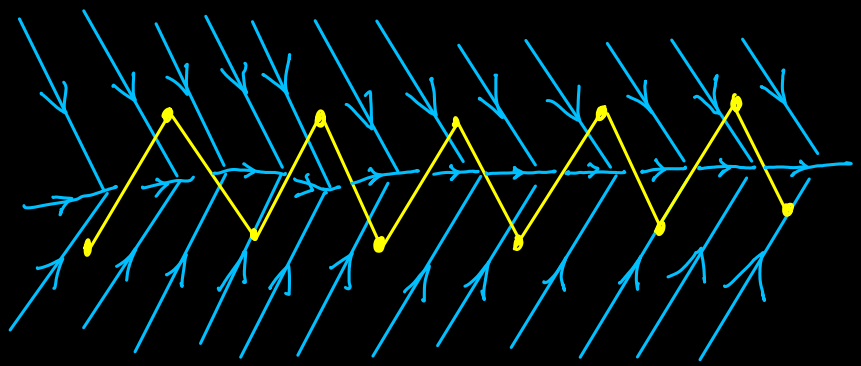


# How to study this?

**Idea 4 (Benaïm-Hofbauer-Sorin)**

*Asymptotic Pseudotrajectories*

- Powerful theory that formalizes the intuition that *"The continuous flow asymptotically approximates the discrete flow — but we may need to constantly update the continuous orbit we approximate with."*



- Flexible.
- It's a bit complicated.

# Asymptotic Pseudotrajectories

Metric in the space of Lipschitz curves

$$D(\gamma,\eta)=\sum_{k=1}^{\infty}\frac{1}{2^k}\min\left(\sup_{t\in[-k,k]}\|\gamma(t)-\eta(t)\|,1\right)$$

Set of orbits passing through one point

$$S_x=\{\text{solutions } \gamma\colon\mathbb{R}\to\mathbb{R}^n \text{ to } \gamma'(t)\in\mathcal{F}(\gamma(t)) \text{ with } \gamma(0)=x\}$$

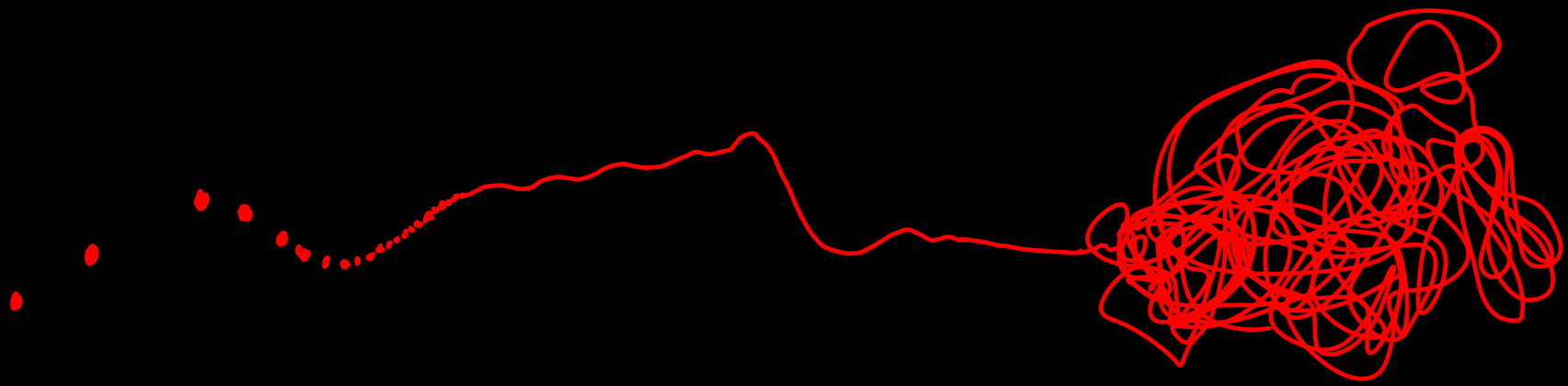
**Definition 1**

$\gamma$ is an *asymptotic pseudotrajectory* if

$$\lim_{t\to+\infty} D(\tau_t\gamma, S_{\gamma(t)})=0,$$

where $\tau_t\gamma(s)=\gamma(s+t)$.

# Wait up – take another look!



# Main principle

**"Theorem"**

*The most significant part of the asymptotic dynamics of a bounded small step process can be decomposed into quasiperiodic Lipschitz cycles.*

# How does that work?

Interpolant curve: $\gamma(\sum_{j=1}^{n-1} \varepsilon_j) = x_n$, interpolate linearly.

Probability measure that encodes the position and speed of the process

$$\mu_T = \frac{1}{T}(\gamma, \gamma')_* \operatorname{Leb}_{[0,T]}$$

that is

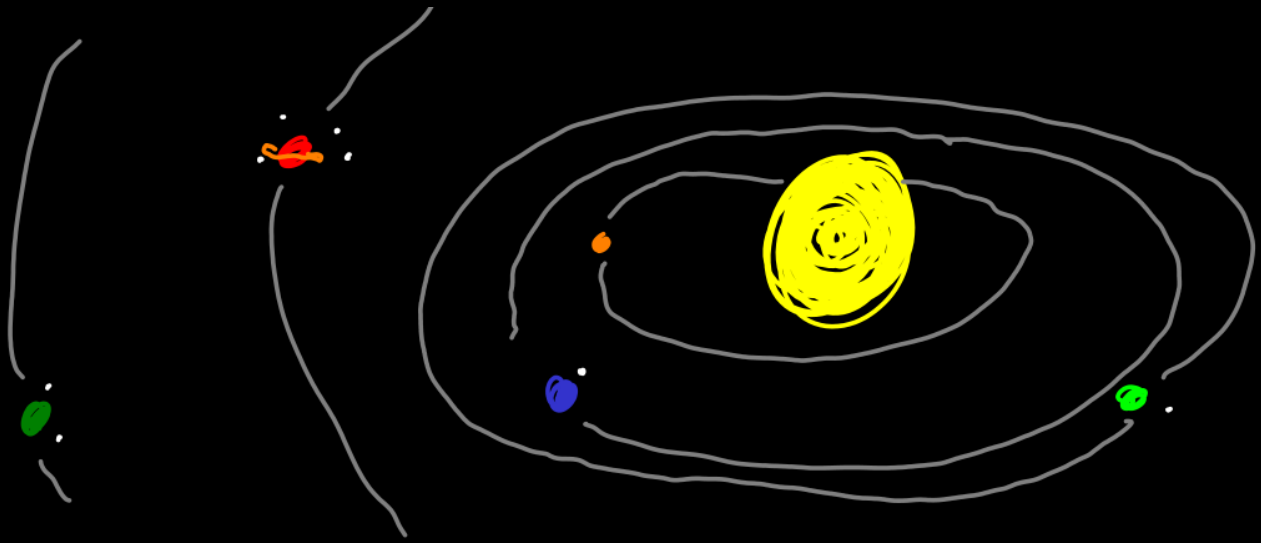
$$\int_{\mathbb{R}^n \times \mathbb{R}^n} g \, d\mu_T = \frac{1}{T} \int_0^T g(\gamma(t), \gamma'(t)) \, dt, \quad g \in C^0(\mathbb{R}^n \times \mathbb{R}^n).$$

# How does that work?

$$\mu_T = \frac{1}{T}(\gamma, \gamma')_* \operatorname{Leb}_{[0,T]}$$

- The sequence $\{\mu_T\}_T$ always has a nonempty accumulation set by Prokhorov's theorem.
- The cluster points are *closed measures*.

# Closed measures



Closed measures

- Used to study Aubry–Mather theory in Hamiltonian dynamics (e.g. celestial mechanics).
- Naturally encode quasiperiodic behavior.
- Very robust to noisy/irregular phenomena.
- Useful also for variational analysis and optimal control.

# Closed measures

**The idea behind the definition.**
Take the case of $\mu_\gamma$ defined by $\mu_\gamma = \frac{1}{T}(\gamma, \gamma')_* \operatorname{Leb}_{[0,T]}$, i.e.,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} f \, d\mu_\gamma = \frac{1}{T} \int_0^T f(\gamma(t), \gamma'(t)) \, dt$$

in the case $f(x, v) = \langle \nabla g(x), v \rangle$ (circulation of $\nabla g$ along $\gamma$):

$$\begin{aligned} \int \langle \nabla g(x), v \rangle \, d\mu_\gamma(x, v) &= \frac{1}{T} \int_0^T \langle \nabla g(\gamma(t)), \gamma'(t) \rangle \, dt \\ &= \frac{1}{T} \int_0^T (g \circ \gamma)'(t) \, dt \\ &= \frac{1}{T} (g \circ \gamma(T) - g \circ \gamma(0)) \end{aligned}$$

This vanishes for all $g \in C^\infty(\mathbb{R}^n)$ iff $\gamma$ is a closed loop, $\gamma(0) = \gamma(T)$.

# Closed measures

## Definition 2

A probability measure $\mu$ on $\mathbb{R}^n \times \mathbb{R}^n$ is *closed* if either of the following equivalent conditions are verified:

- for all $g \in C^\infty(\mathbb{R}^n)$

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

- for some sequence of closed loops $\gamma_i : [0, T_i] \to \mathbb{R}^n$,

$$\mu = \lim_{i \to +\infty} \mu_{\gamma_i} = \lim_{i \to +\infty} \frac{1}{T_i} (\gamma_i, \gamma_i')_* \, \mathrm{Leb}_{[0, T_i]} \, .$$

# The superposition principle

Let $\mathcal{L}$ be the space of Lipschitz curves $\gamma\colon \mathbb{R} \to \mathbb{R}^n$.
Let $\tau_s$ be the time translation in $\mathcal{L}$, defined by

$$\tau_s\gamma(t) = \gamma(t+s).$$

**Theorem 3 (Superposition principle, due to Young, Smirnov)**

*Let $\mu$ be a closed measure. For every closed probability measure $\mu$ there is a Borel probability measure $\nu$ on the space $\mathcal{L}$ that is invariant under the action of $\tau_s$ and such that*

$$\int_{\mathbb{R}^n} \phi(x,v)\, d\mu(x,v) = \int_{\mathcal{L}} \phi(\gamma(0),\gamma'(0))\, d\nu(\gamma)$$

*for any measurable $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ linear in the second variable.*

# Recap

Small step process

$\Longrightarrow$ closed limit measures

$\Longrightarrow$ superposition of cycles

# Application: large-scale optimization

We strive to find the minimum

$$\min_{x \in \mathbb{R}^n} f(x)$$

with $f : \mathbb{R}^n \to \mathbb{R}$

- not convex
- not smooth
- not structured
- with $n$ very large

# Main motivation: Deep Learning

Minimize the loss function

$$L(X_1, \ldots, X_k) = \sum_{i=1}^{N} (p_i - \hat{p}_i)^2$$

where $\hat{p}_i = \phi_k X_k \ldots \phi_2 X_2 \phi_1 X_1(q_i)$

- $X_i$ = linear + constant
- $\phi_i$ activation function, such as $\phi_i = \mathrm{ReLU}(x) = \max(0, x)$.
- $N$ can be huge.
- Using back-propagation we can compute $\approx \nabla L$ and try to do gradient method.
- In general, $L$ non smooth, non convex.

# Small step gradient methods

The lack of structure forces us to use rather rudimentary methods. As model we take

- Subgradient method

Other examples:

- AdaGrad (adaptive gradient)
- AdaDelta
- Adam (adaptive moment estimation)
- Stochastic subgradient
- Heavyball
- ...

# Subgradient method

## Definition 4 (Clarke subdifferential)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function and $x \in \mathbb{R}^n$. The *Clarke subdifferential* $\partial^c f(x)$ is the closed convex hull of the vectors $v$ such that

- there is a sequence $p_i \to x$
- $f$ is differentiable at $p_i$
- $\nabla f(p_i) \to v$

## Definition 5 (Subgradient method)

Goal: minimize the function $f : \mathbb{R}^n \to \mathbb{R}$

- $x_{i+1} = x_i - \varepsilon_i v_i$, $v_i \in \partial^c f(x_i)$.
- $\varepsilon_i \searrow 0$ with $\sum_i \varepsilon_i = +\infty$
  (often take $\sum_i \varepsilon_i^2 < +\infty$)

# Classical cases

$f$ **smooth.** We know:

- the limit points $x$ are critical, $0 \in \partial^c f(x)$
- $f(x_i)$ converges
- if $f$ satisfies the Kurdyka-Łojasiewicz inequality, then $x_i$ converges

$f$ **convex, nonsmooth.** We know:

- $x_i$ converges
- $f(x_i)$ converges

Otherwise not much is known.

# How does this behave?

- $f$ **nonsmooth, nonconvex.** Chaotic bouncing!

# Very counterinuitive behavior

**Example (Daniliidis–Drusvyatskiy 2019, adapted)**

*There is a Lipschitz function* $f : \mathbb{R}^n \to \mathbb{R}$ *and a curve* $\gamma : \mathbb{R} \to \mathbb{R}^n$ *such that* $\|\gamma(t)\|$ *remains bounded,*

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

*and*

$$f(\gamma(t)) = \sin(t).$$

- By a result of Borwein–Moors–Wang, similar behavior is displayed for generic Lipschitz functions $f$.
- We need to identify a more realistic, more rigid class of functions to work with.

# Path-differentiable functions

**Definition 6 (Valadier 1989)**

A locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ is *path-differentiable* if for all Lipschitz curves $\gamma : \mathbb{R} \to \mathbb{R}^n$, for almost all $t \in \mathbb{R}$, $f \circ \gamma$ is differentiable at $t$ and the derivative is given by

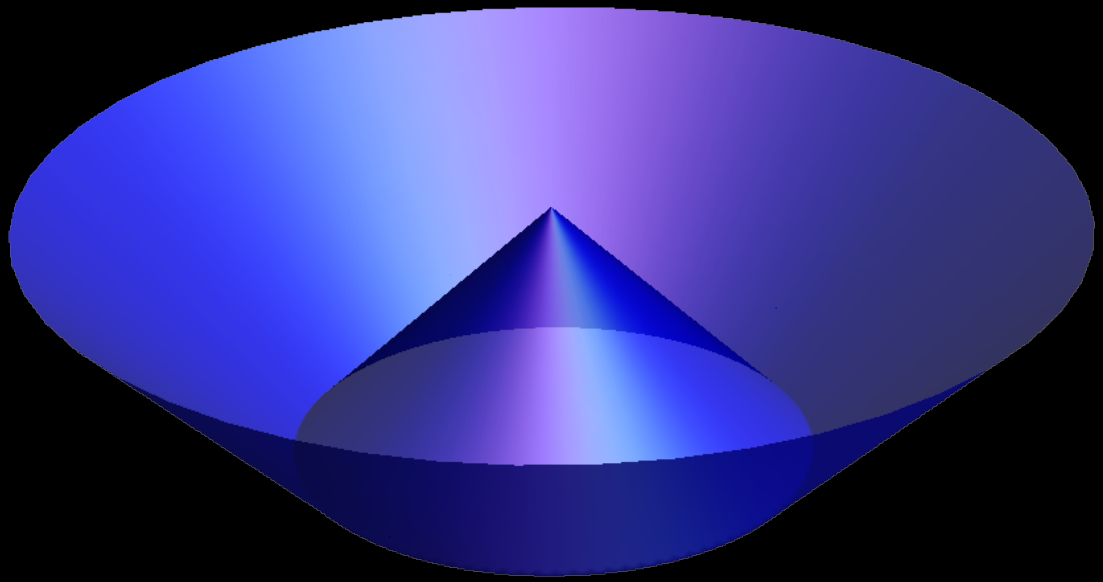
$$(f \circ \gamma)'(t) = \langle v, \gamma'(t) \rangle$$

for all $v \in \partial^c f(\gamma(t))$.

- All vectors $v$ in $\partial^c f$ share the same projection onto the subspace generated by $\gamma'(t)$.
- This class contains all functions of interest for applications, including Deep Learning.

# Path-differentiable functions

Typical example: $f(x) = |1 - \|x\||$.



# Results

**Theorem 7 (Bolte–Pauwels–RZ)**

*$f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz path-differentiable,*
*$\{x_i\}_i$ bounded gradient sequence.*
*. . .*

With this level of generality we can show that it is *impossible* to prove:

- that $x_i$ converges
- that $f(x_i)$ converges
- that the accumulation points are contained in the critical set of $f$
- that $x_i$ stops bouncing

(See paper *Examples of pathological dynamics. . .* )
So what can we prove?

# Essential accumulation set

## Definition 8

The *accumulation set* of the sequence $\{x_n\}_n$ is the set of points $x \in \mathbb{R}^n$ such that, for every neighborhood $U$ of $x$, the intersection $U \cap \{x_n\}_n$ is infinite.

## Definition 9

The *essential accumulation set* of the sequence $\{x_n\}_n$ is the set of points $x \in \mathbb{R}^n$ such that, for every neighborhood $U$ of $x$,

$$\limsup_{N \to +\infty} \frac{\displaystyle\sum_{\substack{1 \le i \le N \\ x_i \in U}} \varepsilon_i}{\displaystyle\sum_{1 \le i \le N} \varepsilon_i} = \frac{\text{time spent in } U}{\text{total time}} > 0.$$

# Essential accumulation set

**Definition 10**

The *essential accumulation set* of the sequence $\{x_n\}_n$ is the set of points $x \in \mathbb{R}^n$ such that, for every neighborhood $U$ of $x$,

$$\limsup_{N\to+\infty} \frac{\sum_{\substack{1\le i\le N \\ x_i\in U}} \varepsilon_i}{\sum_{1\le i\le N} \varepsilon_i} = \frac{\text{time spent in } U}{\text{total time}} > 0.$$

- Encodes persistently recurrent behavior.
- Ignores sporadic, recurrent escapades.
- Compact but not necessarily connected.

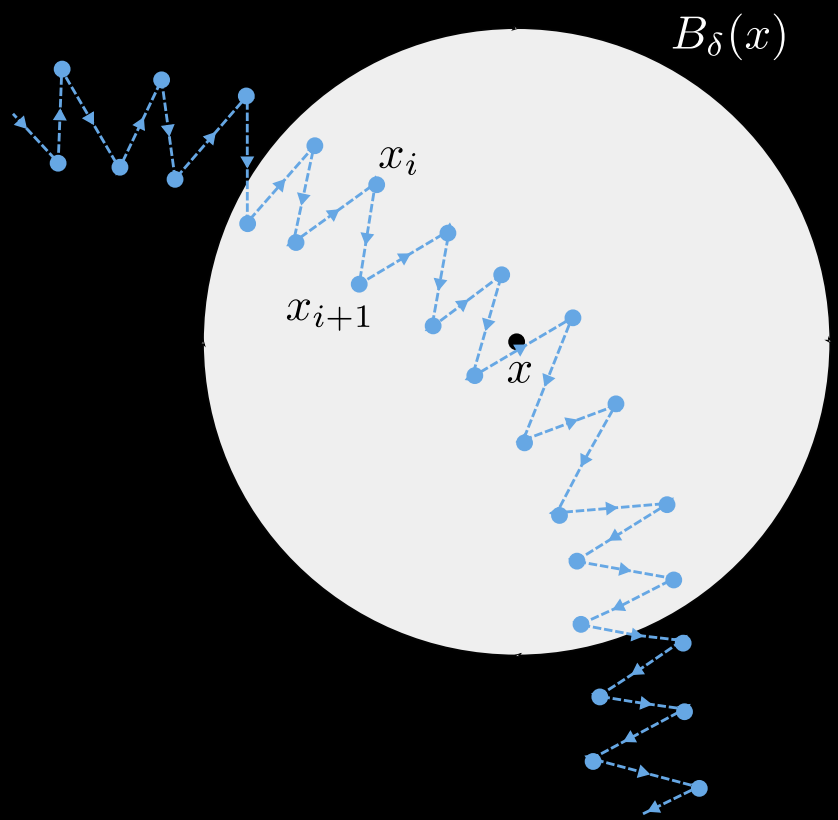
# Criticality

**Theorem 11 (Bolte–Pauwels–RZ)**

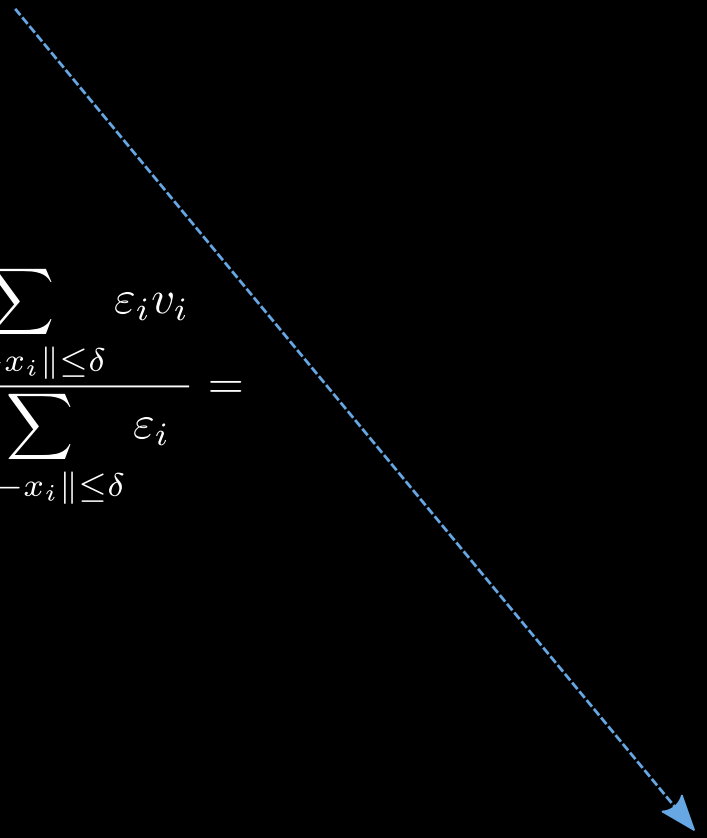
*$f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz path-differentiable,*
*$\{x_i\}_i$ bounded gradient sequence.*
*The essential accumulation set of $\{x_i\}_i$ is contained in the critical set of $f$.*
*If $f$ is constant on each connected component of its critical set, then the accumulation set of $\{x_i\}_i$ is contained in the critical set of $f$ and $\{f(x_i)\}_i$ converges.*

# Oscillation compensation

Pick a point $x$ in the accumulation set of $\{x_i\}_i$ and $\delta > 0$.




$$\frac{\sum_{\|x-x_i\|\leq\delta} \varepsilon_i v_i}{\sum_{\|x-x_i\|\leq\delta} \varepsilon_i} =$$



# Oscillation compensation

## Theorem 12 (Bolte–Pauwels–RZ)

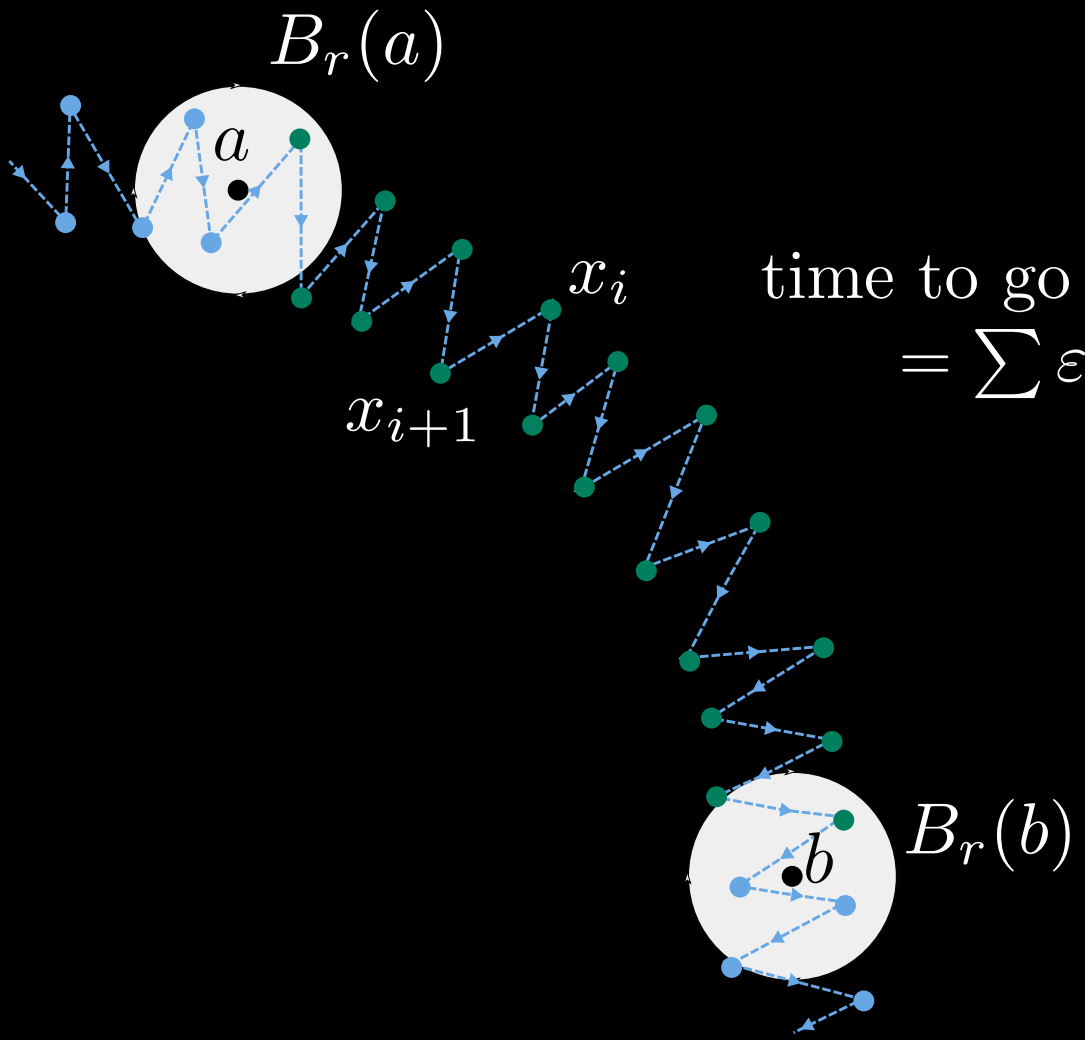
*$f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz path-differentiable,*
*$\{x_i\}_i$ bounded gradient sequence.*
*If $x$ is in the essential accumulation set of $\{x_i\}_i$ (or if $f$ is constant in the connected components of its critical set and $x$ is a point in the accumulation set $\{x_i\}_i$) then for all $\delta > 0$*

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

*(The actual statement is a bit more complicated.)*

# Slow down

Pick points $a, b$ in the accumulation set of $\{x_i\}_i$ and $r > 0$.




time to go from $B_r(a)$ to $B_r(b)$
$= \sum \varepsilon_i \to +\infty$

# Slow down

**Theorem 13 (Bolte–Pauwels–RZ)**

*$f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz path-differentiable,*
*$\{x_i\}_i$ bounded gradient sequence.*
*Let $a$ and $b$ be two distinct points in the accumulation set of $\{x_i\}_i$ such that $f(a) \leq f(b)$. Pick a subsequence $x_{i_k} \to a$ and, for each $k$, $i'_k > i_k$ such that $x_{i'_k} \to b$. Then*

$$\sum_{p=i_k}^{i'_k-1} \varepsilon_p \to +\infty.$$

# Perpendicularity of the oscillations

**Remark**

*The oscillation happens, asymptotically, in the directions perpendicular to the critical set of $f$ (because of the structure of path-differentiable functions).*

# Technique of proof

- Record the position-velocity data with interpolant curve $\gamma\colon [0,+\infty) \to \mathbb{R}^n$ joining $x_i$ with velocity $v_i \in \partial^c f(x_i)$.
- Encode with probability measures $\mu_T$ on $\mathbb{R}^n \times \mathbb{R}^n$ given by

$$\mu_T = \frac{1}{T}(\gamma, \gamma')_* \operatorname{Leb}_{[0,T]}.$$

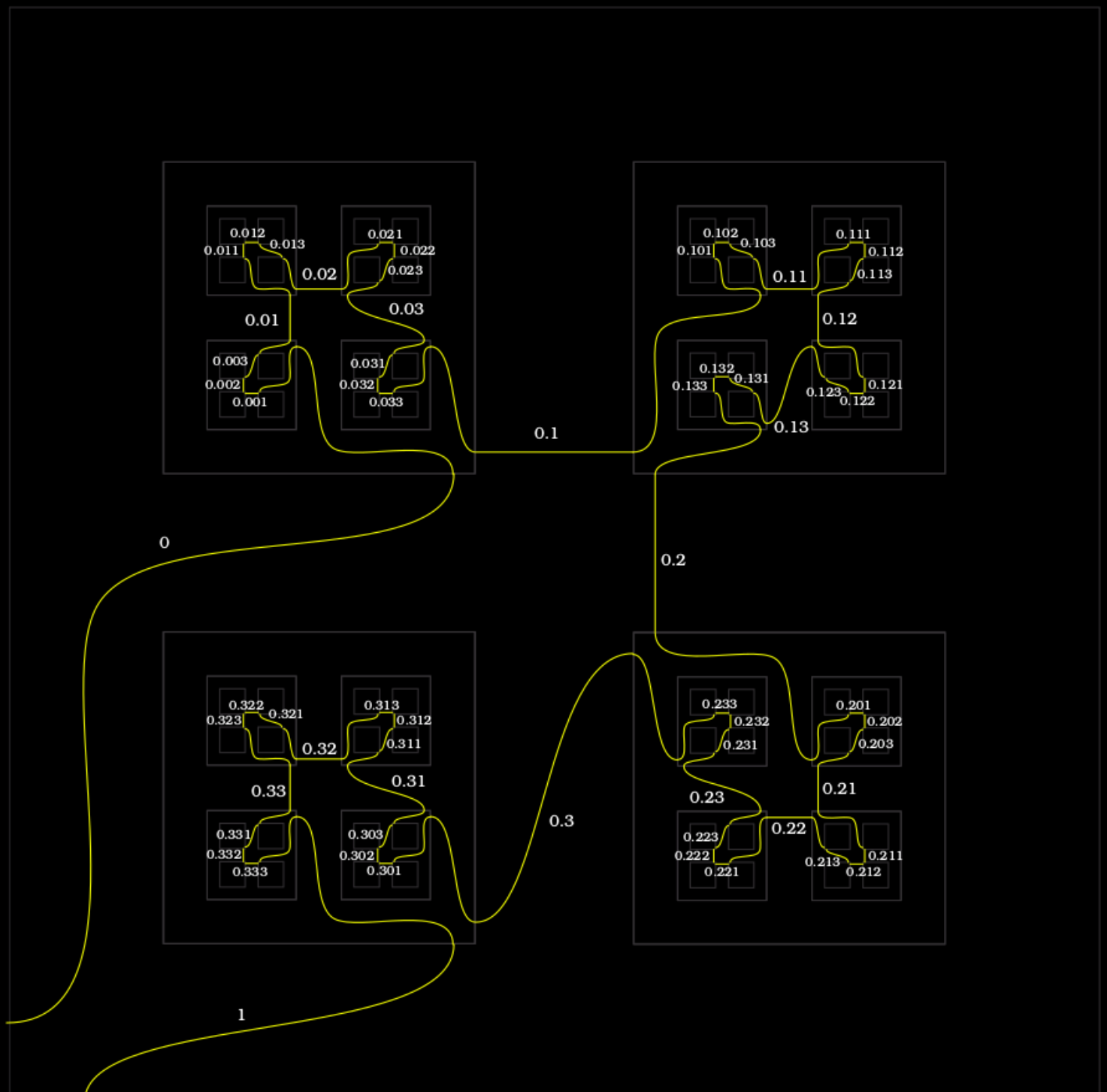
- The accumulation points $\mu$ of the sequence $\{\mu_T\}_T$ have vanishing circulation on gradient vector fields, that is,

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

  for all $h \in C^\infty(\mathbb{R}^n)$ (i.e., these $\mu$ are *closed*).
- The support of $\mu$ is contained in the graph of $-\partial^c f$ ($\mu$ is *subgradient*).
- Can prove "$\mu$ closed and subgradient $\implies$ $\mu$ stationary."
- "$\mu$ stationary $\implies$ our statements."

# Recap


This means we have orderly oscillations around the critical set, almost perpendicular to it, and that almost stay in place.


0.012
0.011
0.013
0.02
0.021
0.022
0.023
0.01
0.03
0.003
0.002
0.001
0.031
0.032
0.033
0.1
0.102
0.101
0.103
0.11
0.111
0.112
0.113
0.12
0.132
0.133
0.131
0.13
0.123
0.122
0.121
0
0.2
0.322
0.323
0.321
0.32
0.313
0.312
0.311
0.33
0.31
0.331
0.332
0.333
0.303
0.302
0.301
0.3
0.233
0.232
0.231
0.23
0.201
0.202
0.203
0.21
0.223
0.222
0.221
0.22
0.213
0.212
0.211
1

# Thank you!

Thank you!

References:


- J. Bolte, E. Pauwels, and R. Ríos-Zertuche. *Long term dynamics of the subgradient method for Lipschitz path differentiable functions*. arXiv:2006.00098 [math.OC]
- R. Ríos-Zertuche. *Examples of pathological dynamics of the subgradient method for Lipschitz path-differentiable functions*. arXiv:2007.11699 [math.OC]