Non-convex Subgradient Descent

Mert Pilanci

EE364b, Stanford University

Outline

- semialgebraic functions, Clarke subgradients, stationary points
- bounded subgradient descent with $\alpha_k \asymp 1/k$ converges
- simple low-dimensional examples and convergence basins
- outside the semialgebraic world, convergence can fail

Definitions

We study

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}, \qquad g^{(k)} \in \partial_C f(x^{(k)}).$$

The set $\partial_C f(x)$ is the **Clarke subdifferential**:

$$\partial_C f(x) = \operatorname{conv}\left\{ \lim \nabla f(x_i) : x_i \to x,\ f \text{ differentiable at } x_i \right\}.$$

A point x is **stationary** if

$$0 \in \partial_C f(x).$$

- smooth case: $\partial_C f(x) = \{\nabla f(x)\}$
- convex case: $\partial_C f(x)$ is the usual convex subdifferential

Algebraic sets

A set $S \subseteq \mathbf{R}^n$ is **algebraic** if it is defined only by polynomial equalities:

$$S = \{x \in \mathbf{R}^n : \ p_1(x) = 0, \ldots, p_m(x) = 0\},$$

where each p_i is a polynomial.

Examples:

$$\{(x, y) : x^2 + y^2 - 1 = 0\}, \qquad \{(x, y) : y - x^2 = 0\}.$$

These are the unit circle and the parabola.

Algebraic sets are cut out only by **equations**. They are rigid objects such as curves, surfaces, and intersections of such objects.

Semialgebraic sets

A set $S \subseteq \mathbf{R}^n$ is **semialgebraic** if it can be described using finitely many polynomial equalities and inequalities, together with finite **and/or** operations.

Examples:

$$\{(x, y) : x^2 + y^2 \leq 1\}, \qquad \{(x, y) : x^2 + y^2 \geq 1\}.$$

The second set is **nonconvex**, but it is still semialgebraic.

Semialgebraic functions

A set $S \subseteq \mathbf{R}^n$ is **semialgebraic** if it can be described using finitely many polynomial equalities and inequalities, together with finite **and/or** operations.
A function $f : \mathbf{R}^n \to \mathbf{R}$ is semialgebraic if its graph

$$\{(x, t) : t = f(x)\}$$

is semialgebraic.
Examples:

$x^4 - y^2$, $|x|$, $\max\{x^2, |y|\}$, finite piecewise polynomial functions.

Semialgebraic functions have **finite geometric complexity**. They can have kinks, but they cannot oscillate in a wild way at infinitely many scales.

Example: From ReLU to neural network losses

The ReLU activation

$$\sigma(z) = \max\{z, 0\}$$

is semialgebraic because its graph is

$$\{(z,t) : t = z,\ z \geq 0\} \cup \{(z,t) : t = 0,\ z \leq 0\}.$$

Now consider a two-layer ReLU network

$$f_\theta(x) = \sum_{j=1}^{m} a_j \, \sigma(w_j^T x + b_j) + c,$$

with parameters

$$\theta = (a_1, \ldots, a_m, w_1, \ldots, w_m, b_1, \ldots, b_m, c).$$

For each hidden unit, the quantity

$$t_j = \sigma(w_j^T x + b_j)$$

is semialgebraic in (x, θ, t_j), since it is defined by polynomial equalities and inequalities.

Summing and multiplying by parameters preserves semialgebraicity, so the graph

$$\{(x, \theta, t) : t = f_\theta(x)\}$$

is semialgebraic.

Therefore a two-layer ReLU network is a semialgebraic function of both the input and the parameters.

Loss of a neural network

Given data (x_i, y_i) for $i = 1, \ldots, N$, define the squared loss

$$L(\theta) = \sum_{i=1}^{N} \big(f_\theta(x_i) - y_i\big)^2.$$

Since each x_i, y_i is fixed, each term

$$\big(f_\theta(x_i) - y_i\big)^2$$

is obtained from a semialgebraic function by polynomial operations. Hence the graph

$$\{(\theta, t) : t = L(\theta)\}$$

is semialgebraic.

So the training objective of a finite ReLU networks with polynomial-type losses such as squared loss is semialgebraic.

This is why semialgebraic geometry naturally appears in the analysis of neural network optimization.

Main theorem

Theorem. Let $f : \mathbf{R}^n \to \mathbf{R}$ be locally Lipschitz and semialgebraic. Suppose

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}, \qquad g^{(k)} \in \partial_C f(x^{(k)}),$$

the iterates stay in a bounded set, and there exist constants $c_1, c_2 > 0$ such that

$$\frac{c_1}{k+1} \leq \alpha_k \leq \frac{c_2}{k+1} \qquad \text{for all } k.$$

Then

$$x^{(k)} \to \bar{x}, \qquad 0 \in \partial_C f(\bar{x}).$$

So bounded subgradient descent with a $1/k$-type stepsize converges to a stationary point.

Why we assume boundedness

Consider

$$f(x) = -|x|.$$

It is semialgebraic and locally Lipschitz, with

$$\partial_C f(x) = \begin{cases} \{-1\}, & x > 0, \\ [-1, 1], & x = 0, \\ \{1\}, & x < 0. \end{cases}$$

If $x^{(1)} > 0$, then $g^{(k)} = -1$ for every k, so

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} = x^{(k)} + \alpha_k, \qquad x^{(k)} = x^{(1)} + \sum_{j=1}^{k-1} \alpha_j.$$

If $\sum_k \alpha_k = \infty$ (in particular, if $\alpha_k \asymp 1/k$), then

$$x^{(k)} \to +\infty.$$

Similarly, if $x^{(1)} < 0$, then $x^{(k)} \to -\infty$.

So semialgebraicity and a $1/k$-type stepsize do **not** prevent escape to infinity: the boundedness assumption in the theorem is essential.

How can boundedness be guaranteed?

Common sufficient mechanism:

Compact feasible set. For the projected method

$$x^{(k+1)} = P_Q\big(x^{(k)} - \alpha_k g^{(k)}\big), \qquad Q \text{ compact},$$

every iterate lies in Q.

Leaving the semialgebraic world

Outside the semialgebraic world, there exist very rough 1-Lipschitz functions $a : \mathbf{R} \to \mathbf{R}$ such that

$$\partial_C a(t) = [-1, 1] \qquad \text{for every } t \in \mathbf{R}.$$

This means: at every point, every slope between -1 and 1 is legal.

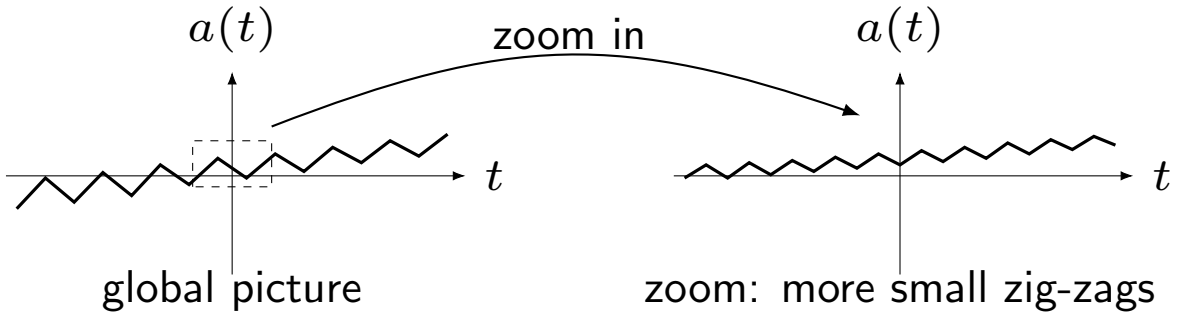
Such a function is impossible in the semialgebraic world. A one-dimensional semialgebraic function is piecewise smooth with only finitely many kinks.

We use this rough function to build a counterexample with

- a global minimizer, and
- a bounded subgradient descent sequence that never converges.

Counterexample picture: the bad function a

The exact function a is built from a splitting set, so there is no simple closed formula. The picture below is a **schematic cartoon**: a 1-Lipschitz graph with more and more tiny zig-zags.



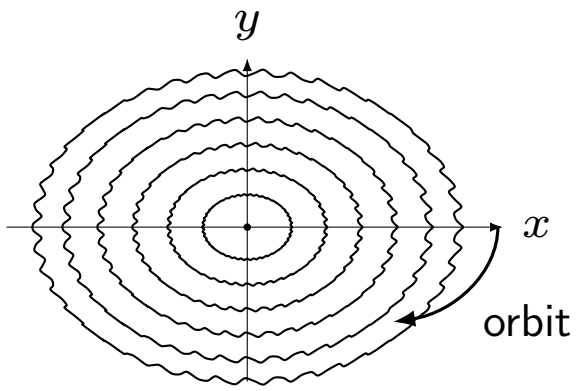
- near every point, slopes $+1$ and -1 both show up arbitrarily close by
- this is the geometric reason why $\partial_C a(t) = [-1, 1]$ for every t

Counterexample picture: the objective F

Define

$$F(x,y) = \frac{1}{2}(x^2+y^2) + 4a(x) + 4a(y).$$

The figure is a **schematic contour picture**: a quadratic bowl with rough wrinkles.



Think of F as a rough bowl: **the bowl keeps the orbit bounded, and the rough part lets us choose a tangential subgradient.** Hence, subgradient descent iterates keeps rotating forever: the sequence is bounded but it does not converge.

Takeaways

- semialgebraic geometry rules out wild infinite-scale oscillation
- bounded subgradient descent with $\alpha_k \asymp 1/k$ converges to a stationary point
- without semialgebraicity, bounded nonconvergent subgradient orbits can exist even when a global minimum is attained

The theorem and the counterexample match each other exactly: finite geometry gives convergence; pathological roughness destroys it.

Examples: a small zoo

The theorem says **what happens** under bounded $1/k$-type steps. The examples below show **how the dynamics can look**.

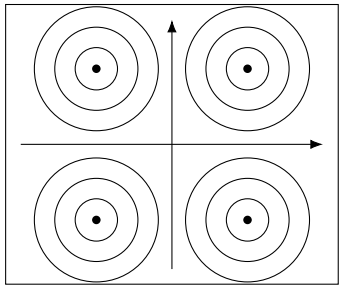
We will see three patterns:

- finitely many wells with clean basin boundaries
- a continuum of minimizers
- nonsmooth nonconvex objectives whose unit-step update is a familiar algorithm

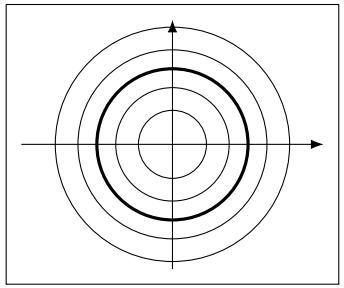
In the last pattern we use unit steps to reveal the algorithm. That is **not** the theorem regime; it is an intuition and modeling point.

Four wells and a ring

These are **level-set pictures of the objectives**.



four wells

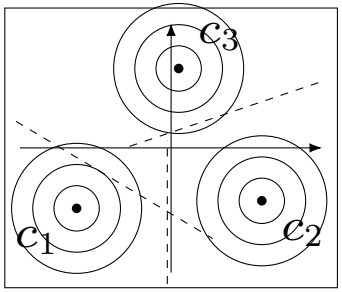


ring of minimizers

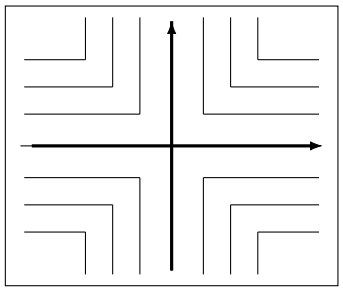
- left: $f_\square(x, y) = \frac{1}{2}(|x| - 1)^2 + \frac{1}{2}(|y| - 1)^2$: four groups of contours, one around each minimizer
- right: $f_\circ(x) = \frac{1}{2}(\|x\|_2 - 1)^2$: the dark circle is the whole minimizer set

Nearest codeword and hard thresholding

These are objective level sets.



nearest codeword


union of axes

- left: $f_C(x) = \frac{1}{2} \min_j \|x - c_j\|_2^2$: three bowls glued along Voronoi boundaries
- right: $\frac{1}{2} d_C(x, y)^2 = \frac{1}{2} \min\{x^2, y^2\}$: cross-shaped contours around the union of axes

Example 1: four wells

Consider

$$f_{\square}(x, y) = \frac{1}{2}(|x| - 1)^2 + \frac{1}{2}(|y| - 1)^2.$$

- semialgebraic, locally Lipschitz, nonconvex, nondifferentiable on the axes
- global minimizers are the four corners

$$(\pm 1, \pm 1)$$

- away from the axes, a Clarke subgradient is

$$g(x, y) = \big(x - \mathbf{sign}(x),\ y - \mathbf{sign}(y)\big)$$

So if $x^{(k)} \neq 0$ and $y^{(k)} \neq 0$,

$$x^{(k+1)} - \mathbf{sign}(x^{(k)}) = (1 - \alpha_k)\big(x^{(k)} - \mathbf{sign}(x^{(k)})\big),$$

and the same recursion holds for y.

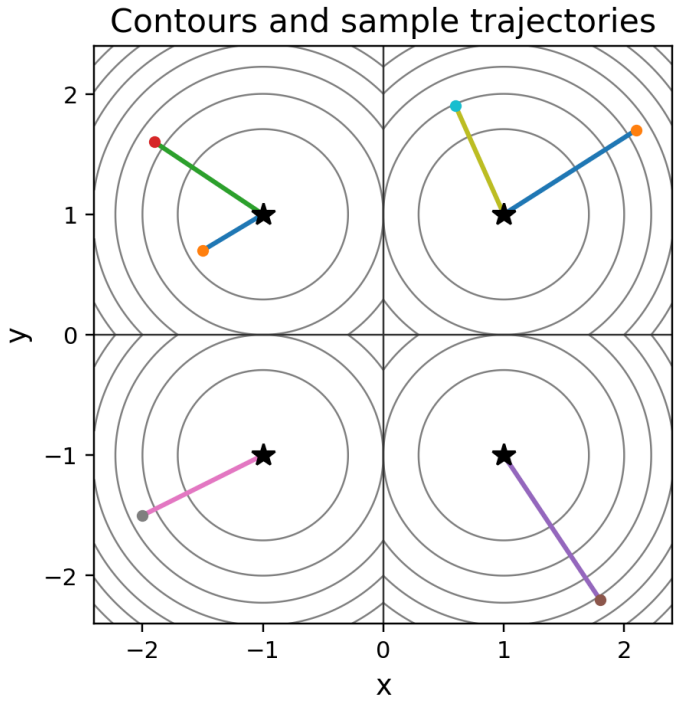
Example 1: basin picture

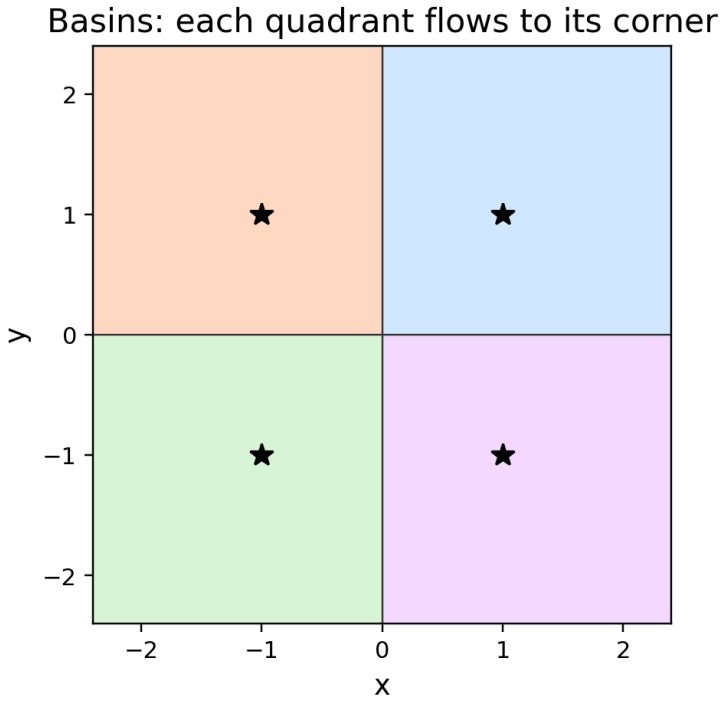
If $0 < \alpha_k \leq 1$ and $\sum_k \alpha_k = \infty$, then each nonzero coordinate keeps its sign and moves to ± 1.

So every open quadrant is a basin:

$$\text{sign pattern of } (x^{(1)}, y^{(1)}) \quad \Longrightarrow \quad \text{limit corner } (\pm 1, \pm 1).$$

The axes are the boundary cases. There the Clarke subdifferential is set-valued, so the first subgradient choice can decide the basin.


Contours and sample trajectories
x
y
−2
−1
0
1
2


Basins: each quadrant flows to its corner
x
y
−2
−1
0
1
2

Example 2: a ring of minimizers

Consider

$$f_\circ(x) = \frac{1}{2}(\|x\|_2 - 1)^2 = \frac{1}{2}d_{S^1}(x)^2, \qquad x \in \mathbf{R}^2.$$

- semialgebraic, locally Lipschitz, nonconvex
- every point on the unit circle S^1 is a global minimizer
- for $x \neq 0$,

$$g(x) = \left(1 - \frac{1}{\|x\|_2}\right) x \in \partial_C f_\circ(x)$$

Writing $r_k = \|x^{(k)}\|_2$, we get

$$x^{(k+1)} = \left(1 - \alpha_k + \frac{\alpha_k}{r_k}\right) x^{(k)}, \qquad r_{k+1} - 1 = (1 - \alpha_k)(r_k - 1).$$

Example 2: Convergence Analysis

- Defining the *residual* $\delta_k := r_{k+1} - 1$ and noting $f_\circ(x^{(k)}) = \frac{1}{2}\delta_k^2$

$$\delta_{k+1} = (1 - \alpha_k)\delta_k$$

- after K steps we have

$$\delta_K = \delta_0 \prod_{k=1}^{K} (1 - \alpha_k)$$

  - constant step size $\alpha_k = \alpha$

$$\delta_K = \delta_0 \alpha^K$$

– step size $\alpha_k = \frac{1}{k+1}$

$$\delta_K = \delta_0 \prod_{k=1}^{K} (1 - \frac{1}{k+1}) = \delta_0 \prod_{k=1}^{K} \frac{k}{k+1} = \frac{\delta_0}{K+1}$$

– diminishing non-summable $\alpha_k \in [0, 1]$, $\sum_i \alpha_i \to \infty$

$$\delta_K = \delta_0 \prod_{k=1}^{K} (1 - \alpha_k) \leq \delta_0 \prod_{k=1}^{K} (1 - e^{-\alpha_k}) = \delta_0 e^{-\sum_{k=1}^{K} \alpha_k}$$

Example 2: every ray is a basin

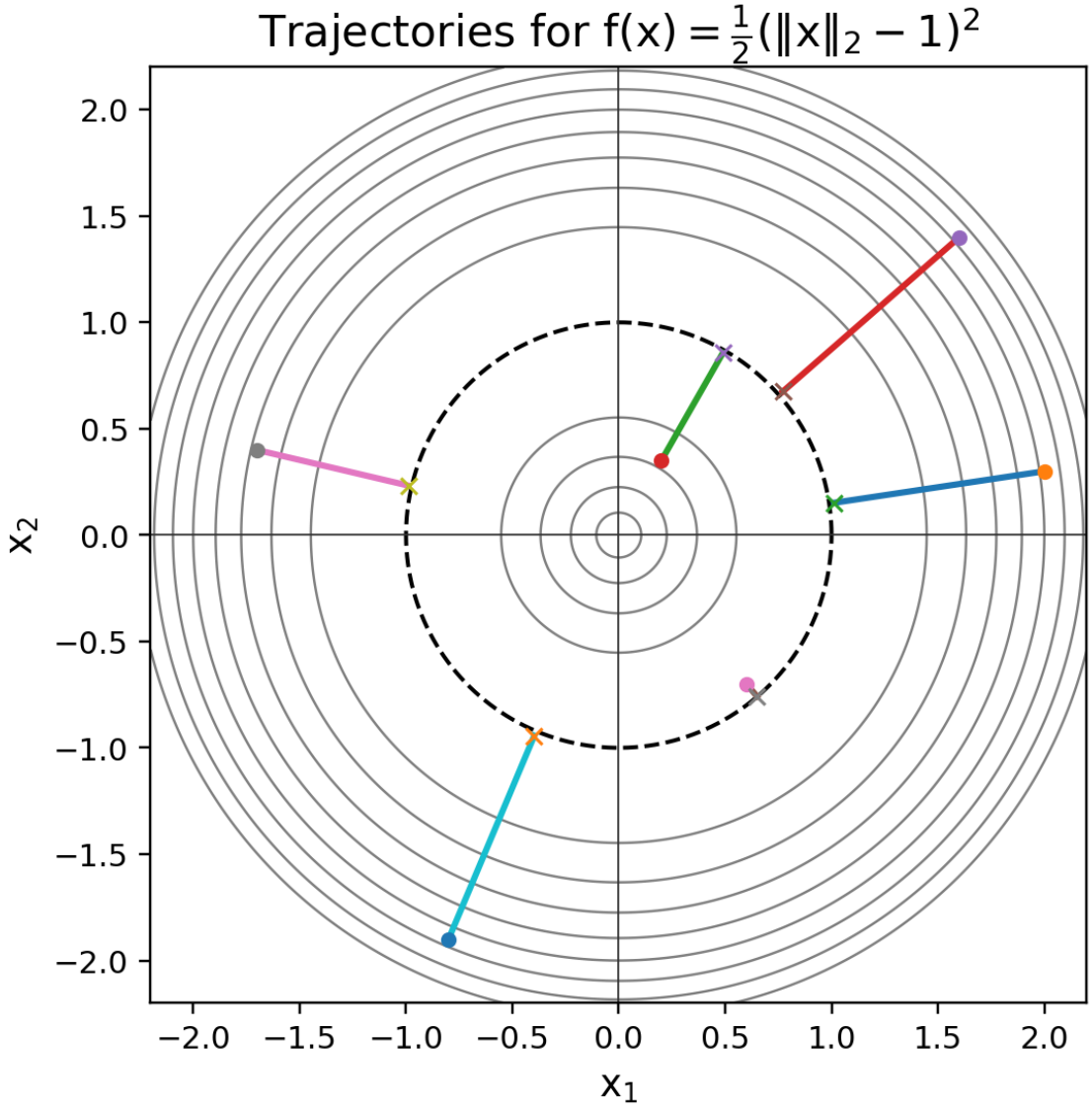
Because the multiplier in the update is a scalar, the **direction never changes**.

So every nonzero trajectory stays on its initial ray and moves only in radius. If $0 < \alpha_k \leq 1$ and $\sum_k \alpha_k = \infty$, then

$$r_k \to 1.$$

Hence every nonzero starting point converges to the unique point on S^1 with the same angle.

So this problem has a **continuum of minimizers** and a **continuum of basins**.


Trajectories for f(x) = $\frac{1}{2}(\|x\|_2 - 1)^2$
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
x_2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
x_1

Example 3: a finite set of target points

Let

$$C = \{c_1, \ldots, c_m\} \subset \mathbf{R}^2, \qquad f_C(x) = \frac{1}{2} \min_{j=1,\ldots,m} \|x - c_j\|_2^2.$$

- f_C is semialgebraic, nonconvex, and nondifferentiable on Voronoi boundaries
- away from a boundary, there is one nearest point $c_{j^\star(x)}$
- there

$$g(x) = x - c_{j^\star(x)} \in \partial_C f_C(x)$$

With unit stepsize $\alpha = 1$,

$$x^+ = x - g(x) = c_{j^\star(x)}.$$

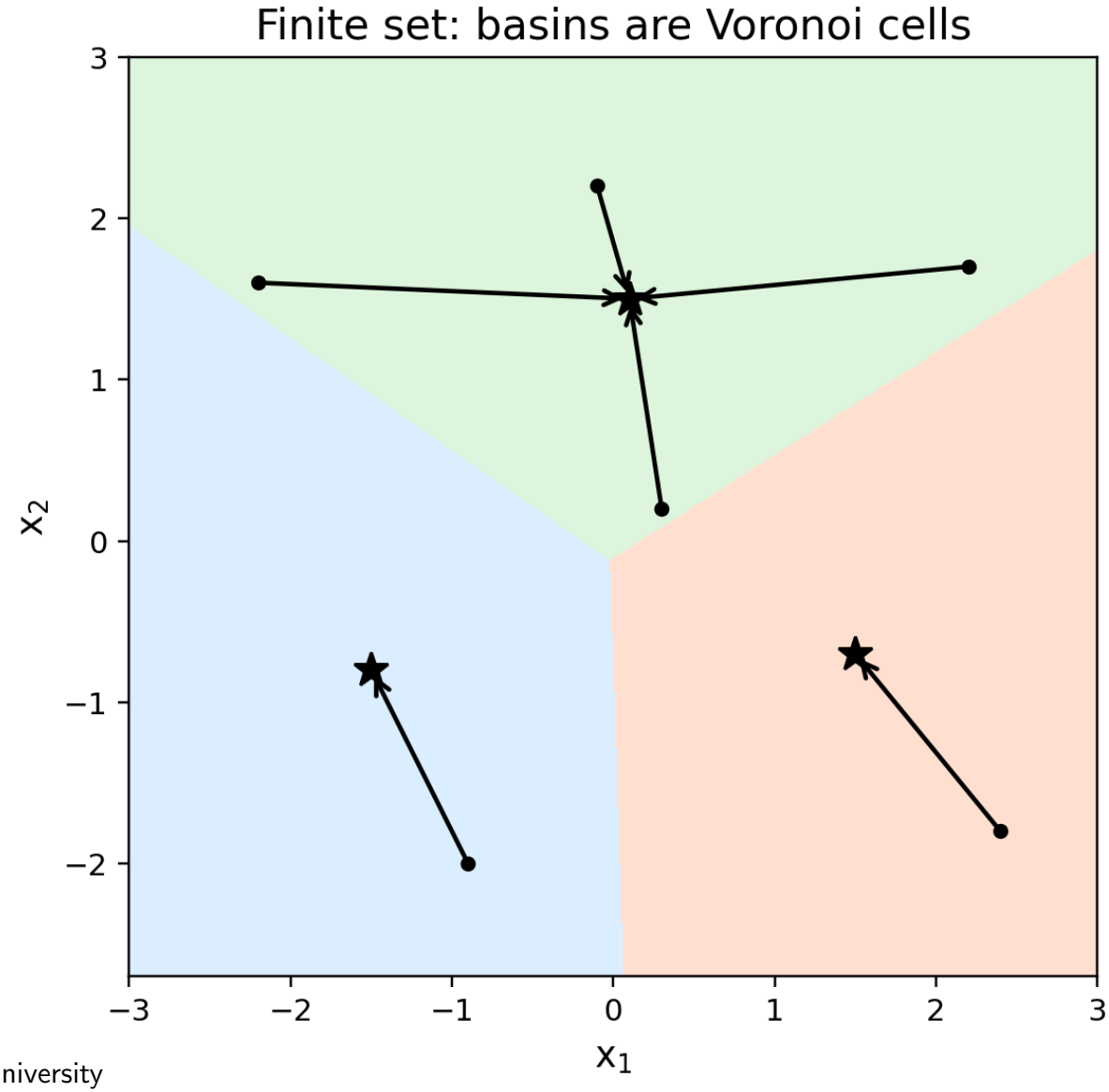
So one step of subgradient descent is simply:

jump to the nearest codeword.

Example 3: Voronoi basins

For a finite set C, the convergence basins are exactly the Voronoi cells.

- every point in one cell jumps to the same minimizer in one step
- on the cell boundaries there are several legal nearest points
- so the basin boundary is exactly the nondifferentiable set


Finite set: basins are Voronoi cells
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
x_2
x_1

Example 4: union of axes = hard thresholding in 2D

Let

$$C = \{(u, 0) : u \in \mathbf{R}\} \ \cup \ \{(0, v) : v \in \mathbf{R}\}.$$

Then

$$f(x, y) = \frac{1}{2} d_C(x, y)^2 = \frac{1}{2} \min\{x^2, y^2\}.$$

Away from the diagonals $|x| = |y|$,

$$g(x, y) = \begin{cases} (0, y), & |x| > |y|, \\ (x, 0), & |y| > |x|. \end{cases}$$

With unit stepsize,

$$(x^+, y^+) = \begin{cases} (x, 0), & |x| > |y|, \\ (0, y), & |y| > |x|. \end{cases}$$

Example 4: basin picture and meaning

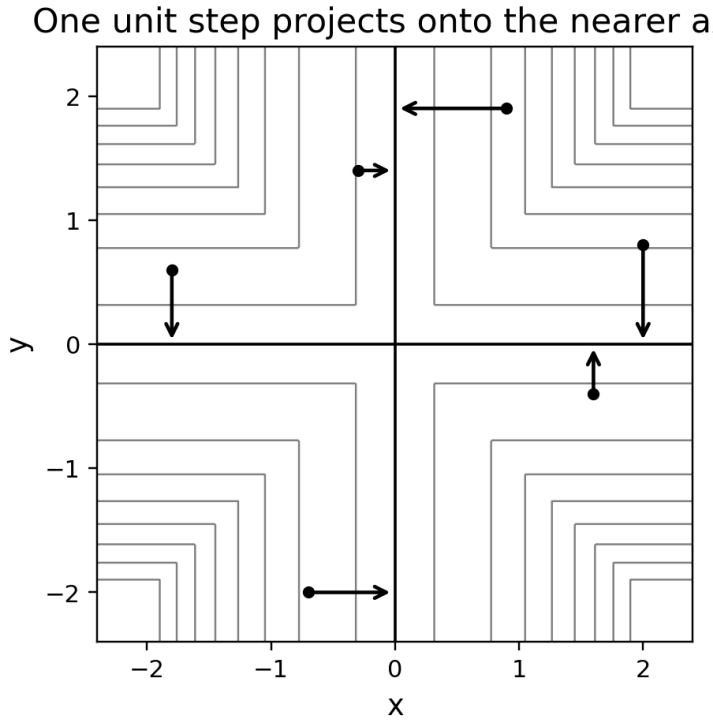
So unit-step subgradient descent keeps the larger coordinate in magnitude and kills the smaller one.

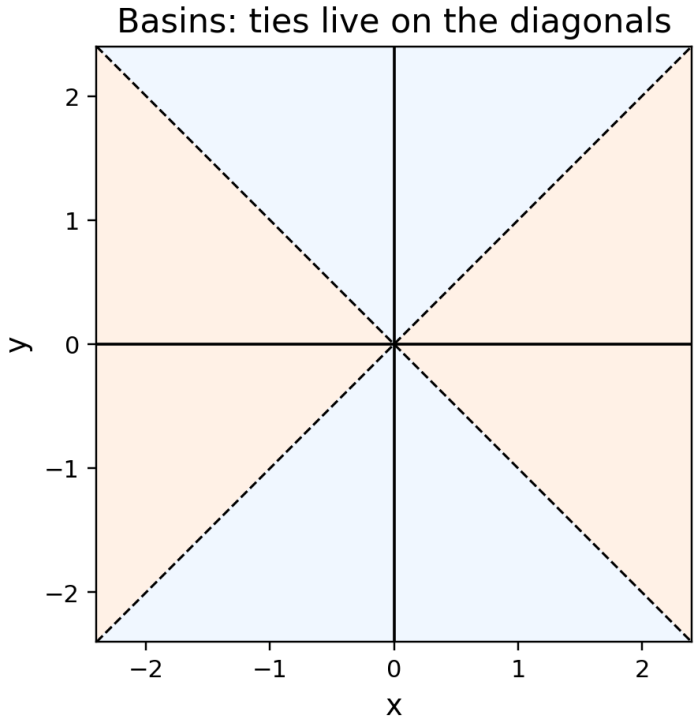
That is exactly **hard thresholding** in two dimensions.

The two basins are the wedge regions

$$|x| > |y| \qquad \text{and} \qquad |y| > |x|,$$

with ties on the diagonals.


One unit step projects onto the nearer axis
2
1
0
−1
−2
y
−2
−1
0
1
2
x


Basins: ties live on the diagonals
2
1
0
−1
−2
y
−2
−1
0
1
2
x

Example 5: sparse vectors

The same idea gives familiar algorithms.
sparse vectors. Let

$$C_s = \{x \in \mathbf{R}^n : \|x\|_0 \leq s\}.$$

Away from magnitude ties, if $S(x)$ keeps the s largest entries of x in magnitude, then

$$\frac{1}{2} d_{C_s}(x)^2 = \frac{1}{2} \sum_{i \notin S(x)} x_i^2,$$

and a unit-step subgradient update is

$$x^+ = H_s(x),$$

the hard-thresholding operator.
The 2D union-of-axes example is exactly the case $s = 1$.

Example 5: low-rank matrices

Now let

$$\mathcal{M}_r = \{X : \mathrm{rank}(X) \leq r\}.$$

Assume the rth and $(r+1)$st singular values of X are different, so the best rank-r approximation is unique. Then the unit-step subgradient update for

$$\frac{1}{2} d_{\mathcal{M}_r}(X)^2$$

is:

$$X^+ = \text{truncated SVD of } X$$

(keep the top r singular values, zero out the rest).

So a classical low-rank projection step can be viewed as **nonconvex nonsmooth subgradient descent**.

What these examples show us

- different starting regions can lead to different limits
- the basin boundary is can be exactly where the function is nondifferentiable
- a whole set of minimizers is not a problem; the method can still choose one limit
- some nonconvex nonsmooth objectives hide a very simple projection rule

This is why low-dimensional examples are worth studying: they show the geometry behind the abstract convergence theorem.

References

1. Davis, D., Drusvyatskiy, D., Kakade, S. et al. Stochastic Subgradient Method Converges on Tame Functions. Found Comput Math 20, 119–154 (2020).

2. Lai, L., Song, M. (2025). On the diameter of subgradient sequences in o-minimal structures. arXiv preprint arXiv:2511.06868.

Appendix: building the bad function a

A measurable set $A \subseteq \mathbf{R}$ is a **splitting set** if every nonempty interval I satisfies

$$0 < m(A \cap I) < m(I),$$

where m is Lebesgue measure.
Given such a set, define

$$a(t) = \int_0^t \left(\chi_A(\tau) - \chi_{A^c}(\tau)\right) d\tau.$$

Then a is 1-Lipschitz.

Appendix: why $\partial_C a(t) = [-1, 1]$

Because A splits every interval, slopes $+1$ and -1 both appear arbitrarily near every point. Hence the nearby gradients of a approach both $+1$ and -1. Taking convex hulls gives

$$\partial_C a(t) = [-1, 1] \qquad \text{for every } t.$$

This is the source of the pathological freedom in the counterexample.