A non-autonomous center-stable set theorem for saddle avoidance in optimization

Andreea-Alexandra Musat

Joint work with Nicolas Boumal.

ELLIIT
Excellence Center at Linköping – Lund in Information Technology

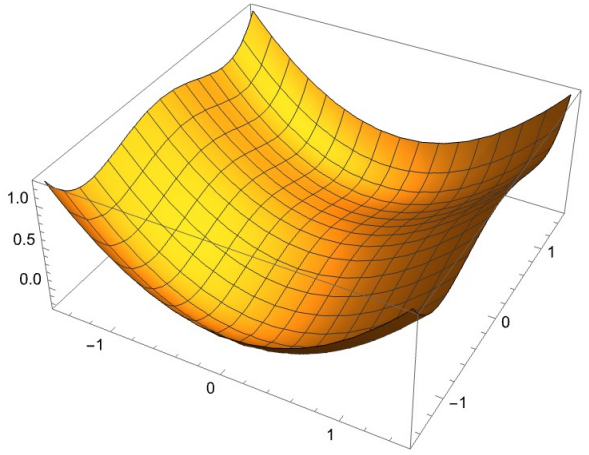
EPFL

GD in the non-convex world

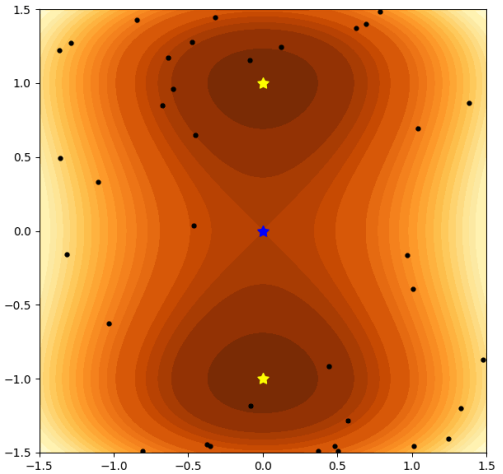
- Usual guarantee: $\|\nabla f(x_k)\| \to 0$.
- In practice: not strict saddle.

GD in the non-convex world

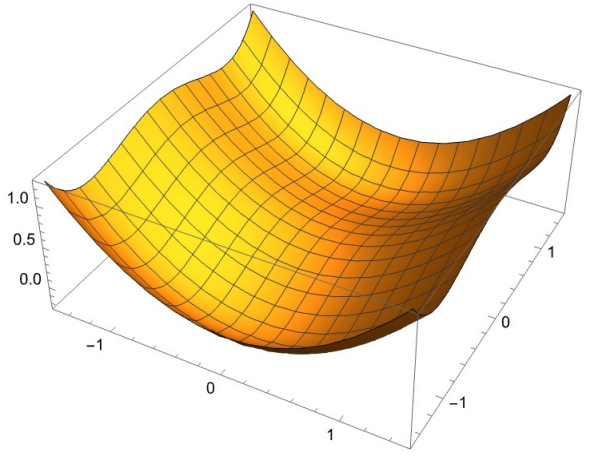
- Usual guarantee: $\|\nabla f(x_k)\| \to 0$
- In practice: not strict saddle.



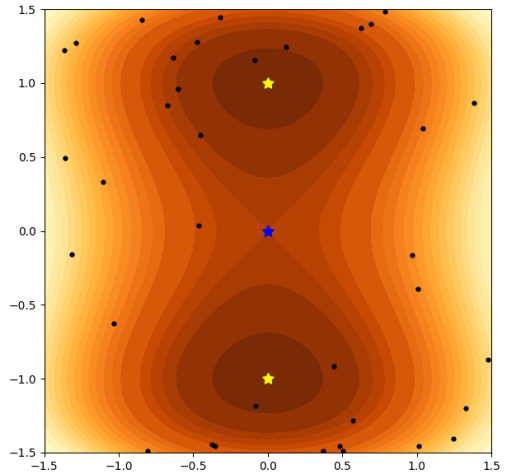
$$f(x) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$$



GD in the non-convex world



$f(x) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$



- Usual guarantee: $\|\nabla f(x_k)\| \to 0$
- In practice: not strict saddle.
- *Strict* saddle avoidance is well-understood ([1], [2], [3]). Strategy:
  - Write the algorithm as an iteration map g.
  - Obtain a local result using the center-stable manifold theorem (CSMT) on g,
  - Then globalize the argument.
- There are many extensions ([4]), but all of them require writing the algorithm as a time-independent dynamical system.

[1] X. Goudou and J. Munier - The gradient and heavy ball with friction dynamical systems: the quasiconvex case (2009).

[2] J. Lee, M. Simchowitz, M. Jordan, and B. Recht - Gradient descent only converges to minimizers (2016)

[3] I. Panageas and G. Piliouras - Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions (2017)

[4] J. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. Jordan, and B. Recht - First-order methods almost always avoid strict saddle points (2019)

The question

- What happens if we allow the maps to change ? i.e. $x_{k+1} = g_k(x_k)$

The question

- What happens if we allow the maps to change ? i.e. $x_{k+1} = g_k(x_k)$

- For example:
  - GD on a cost $f \in C^2$ where ∇f is L-Lipschitz and $\alpha_k \in (\delta, 2/L)$
  - GD on a cost $f \in C^2$ (no Lipschitz gradient) and step sizes $\alpha_k \to 0$ ([1]).
  - Others (Riemannian versions, proximal point methods with variable step sizes)
- Need a new CSMT to cover these!

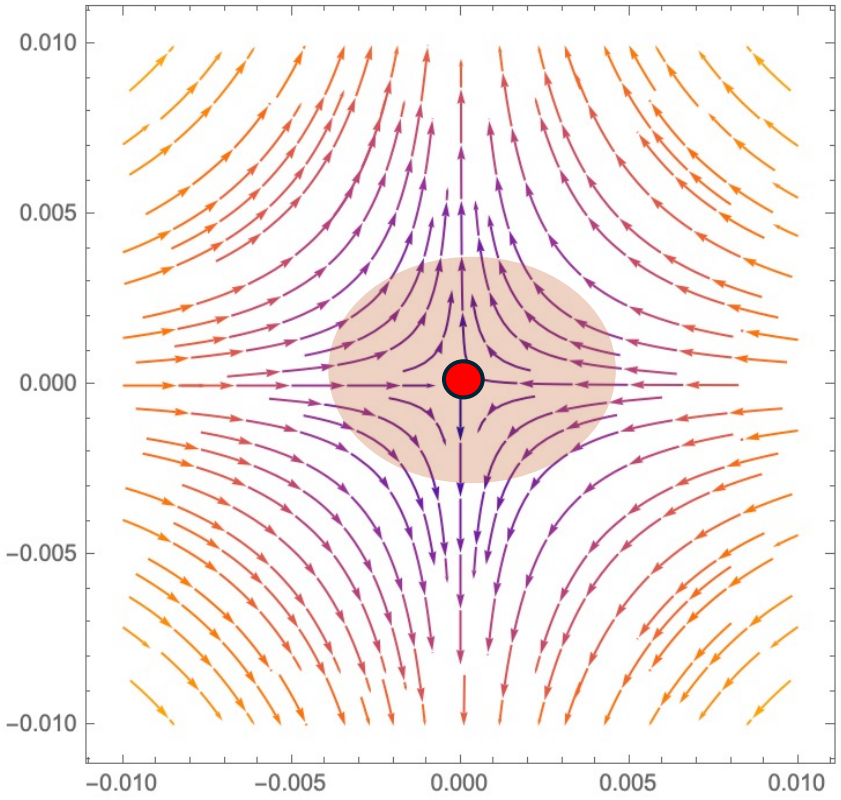
[1] I. Panageas, G. Piliouras, X. Wang - First-order methods almost always avoid saddle points: The case of vanishing step-sizes (2019)

The standard CSMT: statement and intuition

Theorem: Let x^* be a fixed point of the C^1 map $g\colon \mathbb{R}^d \to \mathbb{R}^d$. Assume $Dg(x^*)$ has at least one eigenvalue λ with magnitude $|\lambda| > 1$. Then, there exist

- an open neighborhood B of x^*, and
- a measure zero set W_s^{loc} in $\mathbb{R}^d$

such that if $\{x_t\}_{t\geq 0}$ is a sequence generated by $x_{t+1} = g(x_t)$ and x_t is in B for all $t \geq 0$, then x_t is also in W_s^{loc} for all $t \geq 0$.



[1] M. Shub - Global Stability of Dynamical Systems (1987)

[2] M. W. Hirsch, C. C. Pugh, and M. Shub - Invariant Manifolds (1977)

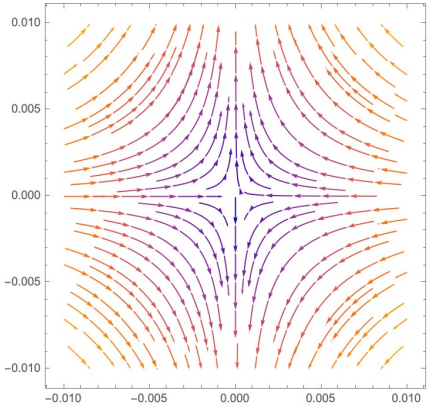
Towards the CSMT: Pseudo-hyperbolicity

- Consider a map $g\colon \mathbb{R}^d \to \mathbb{R}^d$ and a linear map $T\colon \mathbb{R}^d \to \mathbb{R}^d$
- Let $1 \leq \lambda < \mu$ and $0 < \varepsilon < (\mu - \lambda)/4$.
- We say that the pair (g, T) is $(\mu, \lambda, \varepsilon)$ **pseudo-hyperbolic** on $\mathbb{R}^d$ with respect to the splitting $\mathbb{R}^d = E_{cs} \oplus E_u$ if

Towards the CSMT: Pseudo-hyperbolicity

- Consider a map $g\colon \mathbb{R}^d \to \mathbb{R}^d$ and a linear map $T\colon \mathbb{R}^d \to \mathbb{R}^d$.
- Let $1 \leq \lambda < \mu$ and $0 < \varepsilon < (\mu - \lambda)/4$.
- We say that the pair (g, T) is $(\mu, \lambda, \varepsilon)$ **pseudo-hyperbolic** on $\mathbb{R}^d$ with respect to the splitting $\mathbb{R}^d = E_{cs} \oplus E_u$ if

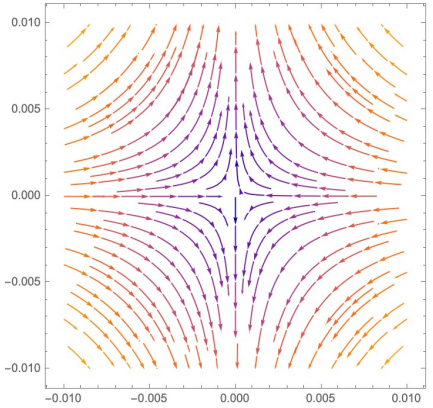
    1. $g(0) = 0$,
    2. The subspaces E_{cs}, E_u are invariant under T,
    3. $\|Ty\| \leq \lambda \|y\|$ and $\|Tz\| \geq \mu \|z\|$ for all $y \in E_{cs}, z \in E_u$,
    4. $\mathrm{Lip}(g - T) \leq \varepsilon$.



Towards the CSMT: Pseudo-hyperbolicity

- Consider a map $g\colon \mathbb{R}^d \to \mathbb{R}^d$ and a linear map $T\colon \mathbb{R}^d \to \mathbb{R}^d$
- Let $1 \le \lambda < \mu$ and $0 < \varepsilon < (\mu - \lambda)/4$.
- We say that the pair (g, T) is $(\mu, \lambda, \varepsilon)$ **pseudo-hyperbolic** on $\mathbb{R}^d$ with respect to the splitting $\mathbb{R}^d = E_{cs} \oplus E_u$ if

  1. $g(0) = 0$,
  2. The subspaces E_{cs}, E_u are invariant under T,
  3. $\|Ty\| \le \lambda \|y\|$ and $\|Tz\| \ge \mu \|z\|$ for all $y \in E_{cs}, z \in E_u$,
  4. $\mathrm{Lip}(g - T) \le \varepsilon$.



- Example to keep in mind: GD on C^2 cost f and step size α near a strict saddle x^*

$$g(x) = x - \alpha \nabla f(x)$$
$$T := Dg(x^*) = I - \alpha \nabla^2 f(x^*)$$

The standard CSMT: inside the proof

- Show that there exists $\varphi\colon E_{cs} \to E_u$ s.t. if $x \in \mathrm{graph}(\varphi)$, then $g(x) \in \mathrm{graph}(\varphi)$.

The standard CSMT: inside the proof

- Show that there exists $\varphi\colon E_{cs} \to E_u$ s.t. if $x \in \text{graph}(\varphi)$, then $g(x) \in \text{graph}(\varphi)$.
- Define the potential $V(x) = \|p_u(x) - \varphi(p_{cs}(x))\|$ and show that

$$V(g(x)) \geq (\mu - 2\varepsilon)V(x)$$

The standard CSMT: inside the proof

- Show that there exists $\varphi\colon E_{cs} \to E_u$ s.t. if $x \in \text{graph}(\varphi)$, then $g(x) \in \text{graph}(\varphi)$.
- Define the potential $V(x) = \|p_u(x) - \varphi(p_{cs}(x))\|$ and show that

$$V(g(x)) \geq (\mu - 2\varepsilon)V(x)$$

- Show that if $x_{k+1} = g(x_k)$ is a sequence with $V(x_{\bar{k}}) > 0$ for some $\bar{k}$, then $V(x_k) \to \infty$.
- Therefore, if $V(x_k)$ remains bounded, then $V(x_k) = 0$ for all k.

Towards a new *CSST*: non-uniform pseudo-hyperbolicity

- A non-autonomous dynamical systems is a sequence of maps $(g_k : \mathbb{R}^d \to \mathbb{R}^d)_{k \geq 0}$.
- A fixed point x^* of a non-autonomous system is a point fixed by all maps.

Towards a new *CSST*: non-uniform pseudo-hyperbolicity

- A non-autonomous dynamical systems is a sequence of maps $(g_k : \mathbb{R}^d \to \mathbb{R}^d)_{k\geq 0}$.
- A fixed point x^* of a non-autonomous system is a point fixed by all maps.

- An **NPH unstable fixed point** x^* of $(g_k)_{k\geq 0}$ is a point for which
  - we can find corresponding linear maps T_k s.t. the pairs (g_k, T_k) are jointly pseudo-hyperbolic with the same splitting $\mathbb{R}^d = E_{cs} \oplus E_u$; $\dim(E_u) \geq 1$
  - with "contraction"/expansion rates λ_k, μ_k that are allowed to vary
  - and need to satisfy a mild non-summability assumption

Towards a new *CSST*: non-uniform pseudo-hyperbolicity

- A non-autonomous dynamical systems is a sequence of maps $(g_k : \mathbb{R}^d \to \mathbb{R}^d)_{k\geq 0}$.
- A fixed point x^* of a non-autonomous system is a point fixed by all maps.

- An **NPH unstable fixed point** x^* of $(g_k)_{k\geq 0}$ is a point for which
  - we can find corresponding linear maps T_k s.t. the pairs (g_k, T_k) are jointly pseudo-hyperbolic with the same splitting $\mathbb{R}^d = E_{cs} \oplus E_u$; $\dim(E_u) \geq 1$
  - with "contraction"/expansion rates λ_k, μ_k that are allowed to vary
  - and need to satisfy a mild non-summability assumption
- Example to keep in mind: GD on C^2 cost f and step sizes $\alpha_k \to 0, \sum_k \alpha_k = \infty$ near a strict saddle x^*

$$g_k(x) = x - \alpha_k \nabla f(x)$$
$$T_k := Dg_k(x^*) = I - \alpha_k \nabla^2 f(x^*)$$

Constants μ_k, λ_k go to 1!

A new CSST

Theorem: Let $x^* \in \mathbb{R}^d$ be an NPH unstable fixed point of a non-autonomous dynamical system $\left(g_k \colon \mathbb{R}^d \to \mathbb{R}^d\right)_{k\geq 0}$. Then, there exist

- an open neighborhood B of x^*,
- a measure zero set W_{cs}^{loc} in $\mathbb{R}^d$,

and an integer $K \geq 0$ with the following property: if $(x_k)_{k\geq 0}$ is a sequence such that, for some $\bar{k} \geq K$, we have $x_{k+1} = g_k(x_k)$ and $x_k \in B$ for all $k \geq \bar{k}$, then $x_k \in W_{cs}^{loc}$ for all $k \geq \bar{k}$.

- No single graph, but a sequence of graphs!

Obtaining a global result

- The center-stable set theorem only gives that the measure of the "bad" set is 0 *locally*.

Obtaining a global result

- The center-stable set theorem only gives that the measure of the "bad" set is 0 *locally*.
- To preserve this globally, a sufficient property on each map g_k is:

$$\mu(S) = 0 \quad \Longrightarrow \quad \mu(g^{-1}(S)) = 0 \qquad \text{(Luzin } N^{-1}\text{)}$$

Obtaining a global result

- The center-stable set theorem only gives that the measure of the "bad" set is 0 *locally*.
- To preserve this globally, a sufficient property on each map g_k is:

$$\mu(S) = 0 \quad \Longrightarrow \quad \mu(g^{-1}(S)) = 0 \qquad \text{(Luzin } N^{-1}\text{)}$$

- If $g \in C^1$, then this is equivalent to $Dg(x)$ invertible a.e.

Obtaining a global result

- The center-stable set theorem only gives that the measure of the "bad" set is 0 *locally*.
- To preserve this globally, a sufficient property on each map g_k is:

$$\mu(S) = 0 \quad \Longrightarrow \quad \mu(g^{-1}(S)) = 0 \qquad \text{(Luzin } N^{-1}\text{)}$$

- If $g \in C^1$, then this is equivalent to $Dg(x)$ invertible a.e.

Theorem: If each map g_k satisfies the Luzin N^{-1} property, then the system avoids its set of NPH unstable fixed points, i.e.

$$\mu\Big(x_0 \in \mathbb{R}^d \mid \lim_{k\to\infty} (g_k \circ g_{k-1} \circ \cdots \circ g_0)(x_0) \text{ is NPH unstable}\Big) = 0$$

Simplest example

- GD on a C^2 function: $g_k(x) = x - \alpha_k \nabla f(x)$.
- If ∇f is L-Lipschitz and $\alpha_k \in (\delta, 2/L)$ for all k, then strict saddles of f are NPH unstable fixed points of $(g_k)_{k \geq 0}$.
- If $\alpha_k \to 0$ and $\sum_k \alpha_k = \infty$, then every strict saddle of f is an NPH unstable fixed point of $(g_k)_{k \geq 0}$.

[1] A-A. Musat, N. Boumal - Gradient descent avoids strict saddles with a simple line-search method too (2025)

Simplest example

- GD on a C^2 function: $g_k(x) = x - \alpha_k \nabla f(x)$.
- If ∇f is L-Lipschitz and $\alpha_k \in (\delta, 2/L)$ for all k, then strict saddles of f are NPH unstable fixed points of $(g_k)_{k\geq 0}$.
- If $\alpha_k \to 0$ and $\sum_k \alpha_k = \infty$, then every strict saddle of f is an NPH unstable fixed point of $(g_k)_{k\geq 0}$.

- For almost all step size α, the GD map satisfies the Luzin N^{-1} property [1].

[1] A-A. Musat, N. Boumal - Gradient descent avoids strict saddles with a simple line-search method too (2025)

End of Tokens!

The standard CSMT: inside the proof (2)

- Main ingredient: finding $\varphi\colon E_{cs} \to E_u$ s.t. if $x \in \text{graph}(x)$, then $g(x) \in \text{graph}(\varphi)$.

- Given some function 1-Lipschitz function φ, construct another 1-Lipschitz function $\tilde{\varphi}$ s.t.

$$x \in \text{graph}(\tilde{\varphi}) \implies g(x) \in \text{graph}(\varphi)$$

- Write this out as an equation: $(p_u \circ g)(y, \tilde{\varphi}(y)) = (\varphi \circ p_{cs} \circ g)(y, \tilde{\varphi}(y))$.
- Rearrange and show $z := \tilde{\varphi}(y)$ is a solution iff it is a fixed point of a certain function.
- Show that the respective function is a contraction, so z is well-defined.

- Define $\tilde{\varphi}$ as the function $y \mapsto z(y)$
- Define Γ to be the operator that given φ, return $\tilde{\varphi}$ with the invariance property.
- Show that Γ is a contraction.