

Certifying the absence of spurious local minima at infinity

Xiaopeng Li

Joint work with Cédric Jozs

Columbia University

September 18, 2022

Statistical estimation problems

The following problems have no spurious local minima:

- 1 principal component analysis (Baldi and Hornik 1989)

$$\inf_{X,Y} \|XY^T - M\|_F^2$$

- 2 low-rank matrix recovery (Li et al. 2020)

$$\inf_{X,Y} \sum_{i=1}^m (\langle A_i, XY^T \rangle - y_i)^2$$

- 3 linear neural networks (Kawaguchi 2016; Laurent and Brecht 2018; Zhang 2019)

$$\inf_{W_1, \dots, W_L} \|W_L \cdots W_1 X - Y\|_F^2$$

→ yet the objective functions are **not coercive**

Do they have spurious local minima at infinity?

Matrix completion - landscape analysis

$$M = \begin{pmatrix} 1 & ? \\ 1 & 1 \end{pmatrix}$$

$$\inf_{x_1, x_2, y_1, y_2 \in \mathbb{R}} (x_1 y_1 - 1)^2 + (x_2 y_1 - 1)^2 + (x_2 y_2 - 1)^2$$

1 Compute the gradient and Hessian

$$\nabla f(x_1, x_2, y_1, y_2) = \begin{pmatrix} 2(x_1 y_1 - 1)y_1 \\ 2(x_2 y_1 - 1)y_1 + 2(x_2 y_2 - 1)y_2 \\ 2(x_1 y_1 - 1)x_1 + 2(x_2 y_1 - 1)x_2 \\ 2(x_2 y_2 - 1)x_2 \end{pmatrix}$$
$$\nabla^2 f(x_1, x_2, y_1, y_2) = \begin{pmatrix} 2y_1^2 & 0 & 4x_1 y_1 - 2 & 0 \\ 0 & 2(y_1^2 + y_2^2) & 4x_2 y_1 - 2 & 4x_2 y_2 - 2 \\ 4x_1 y_1 - 2 & 4x_2 y_1 - 2 & 2(x_1^2 + x_2^2) & 0 \\ 0 & 4x_2 y_2 - 2 & 0 & 2x_2^2 \end{pmatrix}$$

2 Conclude that there are **no spurious local minima** and **all saddle points are strict** (Lee *et al.* 2016 Gradient Descent Only Converges to Minimizers).

Matrix completion - numerical experiment

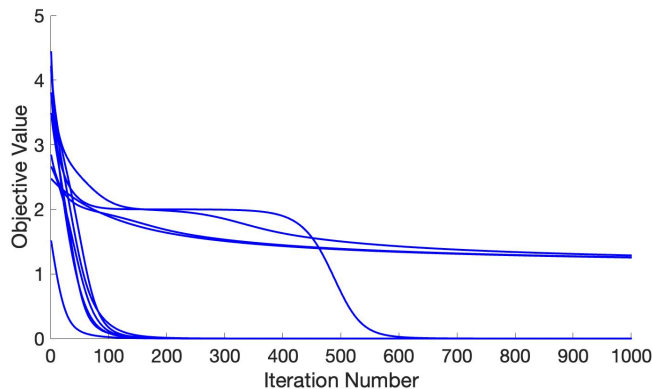


Figure: Gradient method initialized uniformly at random in $[-1, 1]^4$ with constant step size 0.01 sometimes gets stuck at a spurious local minimum at infinity.

Literature review on local minima at infinity

- 1 Blackmore *et al.* 1996: [Local minima and attractors at infinity for gradient descent learning algorithms](#)
- 2 Liang *et al.*: 2018 [Adding One Neuron Can Eliminate All Bad Local Minima](#)
- 3 Sohl-Dickstein and Kawaguchi 2019: [Eliminating All Bad Local Minima from Loss Landscapes Without Even Adding an Extra Unit](#)
- 4 Liang *et al.* 2019: [Revisiting Landscape Analysis in Deep Neural Networks: Eliminating Decreasing Paths to Infinity](#)

Main result

$$\inf_{x \in \mathbb{R}^n} f(x), \quad f : \mathbb{R}^n \rightarrow \mathbb{R}$$

Theorem 1

*Suppose a locally Lipschitz function is bounded below, admits a chain rule, has finitely many critical values, and has bounded subgradient trajectories. Then it has no spurious local minima if and only if it has **no spurious setwise local minima**.*

Assumptions:

- 1 bounded below: $\inf_{\mathbb{R}^n} f > -\infty$
- 2 locally Lipschitz continuous: $\forall x, y \in B(a, r), \quad |f(x) - f(y)| \leq L_a \|x - y\|$
- 3 chain rule: $(f \circ x)'(t) = \langle v, x'(t) \rangle, \forall v \in \partial f(x(t)), \text{ for a.e. } t \geq 0$
- 4 finitely many critical values: $\{f(x) \mid 0 \in \partial f(x)\} = \{v_1, \dots, v_m\}$
- 5 bounded subgradient trajectories: $\forall x_0 \in \mathbb{R}^n, \exists c > 0 : \|x(t)\| \leq c$
- 6 no spurious local minima: $\forall x \in B(x^*, \epsilon), f(x) \geq f(x^*) \Rightarrow \forall x \in \mathbb{R}^n, f(x) \geq f(x^*)$

Conclusion:

- 1 no spurious local minima at infinity

Applications

Corollary 1

The following problems have **no spurious local minima at infinity**:

- 1 *deep linear neural network*

$$\inf_{W_1, \dots, W_L} \|W_L \cdots W_1 X - Y\|_F^2;$$

- 2 *one dimensional deep neural network with sigmoid activation function σ*

$$\inf_{w_1, \dots, w_L} (w_L \sigma(w_{L-1} \cdots \sigma(w_1 x)) - y)^2;$$

- 3 *matrix recovery with restricted isometry property (RIP)*

$$\inf_{X, Y} \sum_{i=1}^m (\langle A_i, XY^T \rangle_F - b_i)^2;$$

- 4 *nonsmooth matrix factorization where $\text{rank}(M) = 1$ and $M_{ij} \neq 0$*

$$\inf_{x, y} \sum_{i=1}^m \sum_{j=1}^n |x_i y_j - M_{ij}|.$$

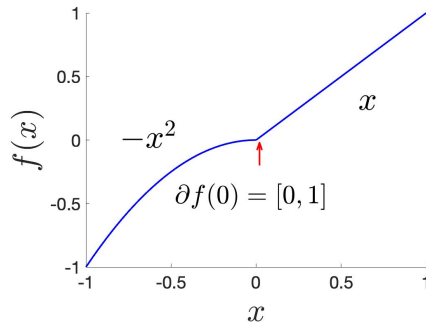
Background - Clarke subdifferential

Definition 1 (Clarke 1990, p. 27)

The Clarke subdifferential of a locally Lipschitz function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined for all $x \in \mathbb{R}^n$ by

$$\partial f(x) := \{s \in \mathbb{R}^n \mid f^\circ(x, d) \geq \langle s, d \rangle, \forall d \in \mathbb{R}^n\}$$

$$f^\circ(x, d) := \limsup_{y \rightarrow x, t \searrow 0} \frac{f(y + td) - f(y)}{t}$$



Background - subgradient trajectories

Definition 2

A subgradient trajectory is an absolutely continuous function $x : [0, \infty) \rightarrow \mathbb{R}^n$ such that

$$x'(t) \in -\partial f(x(t)), \quad \text{for almost every } t \geq 0, \quad x(0) = x_0.$$

- Subgradient trajectory may not exist: consider $f(x) = -\frac{1}{3}x^3$ and $x_0 = 1$.

Proposition 1

If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is locally Lipschitz and bounded below, then there exists a subgradient trajectory $x(\cdot)$ of f starting at arbitrary $x_0 \in \mathbb{R}^n$.

- The trajectory $x(t)$ could go to infinity as $t \rightarrow \infty$.

Definition 3 (Bounded subgradient trajectories)

$\forall x_0 \in \mathbb{R}^n, \exists$ subgradient trajectory $x(\cdot)$ with $x(0) = x_0$: $\exists c > 0$: $\forall t \geq 0, \quad \|x(t)\| \leq c$

Background - setwise local minima

Definition 4 (Setwise local minimum)

A nonempty closed subset $S \subset \mathbb{R}^n$ is a *setwise local minimum* of a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if there exists an open set $U \subset \mathbb{R}^n$ such that $S \subset U$ and $f(x) \leq f(y)$ for all $x \in S$, $y \in U \setminus S$.

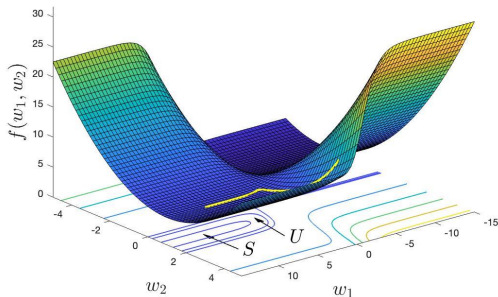


Figure: $\frac{1}{2}[(w_2\sigma(w_1) - 1)^2 + (w_2\sigma(-w_1) + 3)^2]$

Definition 5

A setwise local minimum S is a *spurious local minimum at infinity* if it is unbounded and $\inf_S f > \inf_{\mathbb{R}^n} f$.

Background - setwise local minima

Definition 4 (Setwise local minimum)

A nonempty closed subset $S \subset \mathbb{R}^n$ is a *setwise local minimum* of a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if there exists an open set $U \subset \mathbb{R}^n$ such that $S \subset U$ and $f(x) \leq f(y)$ for all $x \in S$, $y \in U \setminus S$.

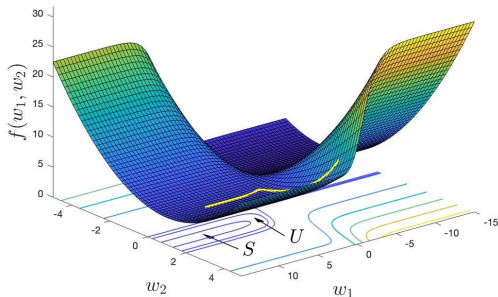


Figure: $\frac{1}{2}[(w_2\sigma(w_1) - 1)^2 + (w_2\sigma(-w_1) + 3)^2]$

Definition 5

A setwise local minimum S is a *spurious local minimum at infinity* if it is unbounded and $\inf_S f > \inf_{\mathbb{R}^n} f$.

Cannot be detected by ∇f and $\nabla^2 f$!

Background - setwise local minima

Related notion (Venturi, Bandeira, and Bruna 2019):

- a *valley* is a path-connected component of a sublevel set of f
- a valley is *spurious* if it does not contain a global minimum

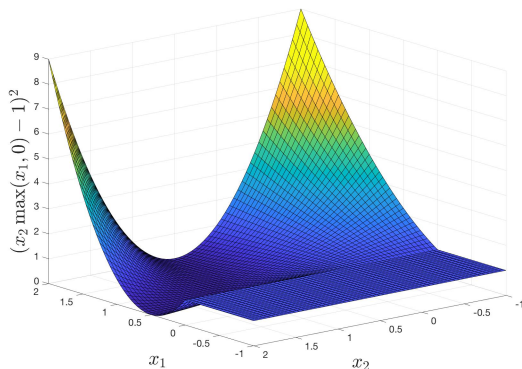


Figure: Function devoid of spurious valleys containing a spurious local minimum at infinity.

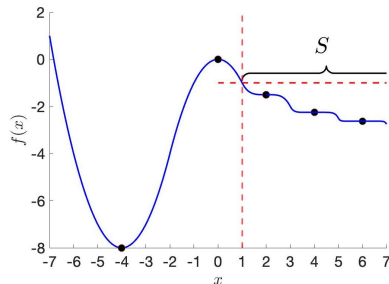
Main result - a closer look

The assumption on finitely many critical values is necessary. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be

$$f(x) := \begin{cases} (x+4)^2 - 8 & \text{if } x \leq -2; \\ -x^2 & \text{if } x \in [-2, 0]; \\ -2^{-k}(x-2k)^{2^{k+1}} - 3(1-2^{-k}) & \text{if } x \in [2k, 2k+1], k \in \mathbb{N}; \\ 2^{-k}(x-2k)^{2^{k+1}} - 3(1-2^{-k}) & \text{if } x \in [2k-1, 2k], k \in \mathbb{N}^+. \end{cases}$$

where \mathbb{N}^+ is the set of all positive integers.

- 1 bounded below ✓
- 2 locally Lipschitz continuous ✓
- 3 chain rule ✓
- 4 finitely many critical values ✗
- 5 bounded subgradient trajectories ✓
- 6 no spurious local minima ✓

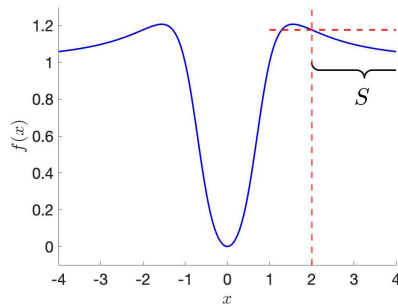


Main result - a closer look

The assumption on bounded subgradient trajectories is necessary. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be

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

- 1 bounded below ✓
- 2 locally Lipschitz continuous ✓
- 3 chain rule ✓
- 4 finitely many critical values ✓
- 5 bounded subgradient trajectories ✗
- 6 no spurious local minima ✓



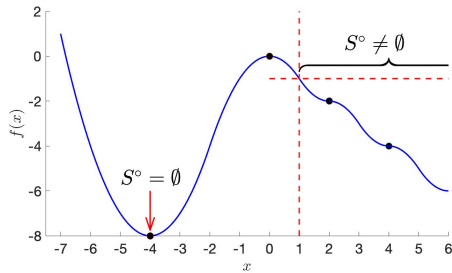
Main result - proof idea

Theorem 1

Suppose a locally Lipschitz function is bounded below, admits a chain rule, has finitely many critical values, and has bounded subgradient trajectories. Then it has no spurious local minima if and only if it has **no spurious setwise local minima**.

Proof sketch: Show that any setwise local minimum S contains a local minimum.

- If $S^\circ = \emptyset$, then every point in S is a local minimum.
- If $S^\circ \neq \emptyset$, then solutions to $\inf_{x \in C_f(S)} f(x)$ are local minima of f .



Applications - existing results

The following problems have no spurious local minima:

- 1 linear neural network ([Kawaguchi 2016](#); [Laurent and Brecht 2018](#); [Zhang 2019](#))

$$f(W_1, \dots, W_L) = \|W_L \cdots W_1 X - Y\|_F^2$$

- 2 one dimensional deep neural network with sigmoid activation function σ

$$f(w_1, \dots, w_L) = (w_L \sigma(w_{L-1} \cdots \sigma(w_1 x)) - y)^2$$

- 3 matrix recovery with restricted isometry property ([Li et al. 2020](#))

$$f(X, Y) = \sum_{i=1}^m (\langle A_i, XY^T \rangle_F - b_i)^2$$

- 4 nonsmooth rank one matrix factorization ([Josz and Lai 2021](#))

$$f(x, y) = \sum_{i=1}^m \sum_{j=1}^n |x_i y_j - M_{ij}|$$

Applications - checking assumptions

Checklist:

- bounded below (nonnegative)
- locally Lipschitz (continuously differentiable, composition of such functions)
- chain rule & finitely many critical values (definable in an o-minimal expansion of the real field \rightarrow definable Morse-Sard theorem)
- bounded subgradient trajectories

For example, consider

1 deep linear neural network

$$f(W_L, \dots, W_1) = \|W_L \cdots W_1 X - Y\|_F^2$$

$$(\text{ODE}) \quad \dot{W}_i = -2(W_L \cdots W_{i+1})^T (W_L \cdots W_1 X - Y) (W_{i-1} \cdots W_1 X)^T$$

Bounded subgradient trajectories - proof illustration

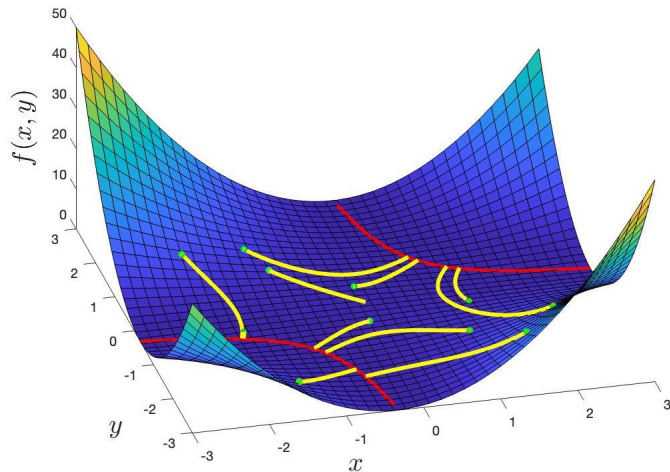


Figure: Red, global minima;
yellow, subgradient
trajectories; green, initial
points; $f(x, y) = (xy - 1)^2$

Bounded subgradient trajectories - pictorial illustration

The proof consists of two steps:

- 1 reduce the problem for general data matrix X to the problem for an invertible data matrix Λ ;
- 2 combine the fact that W_i 's are balanced and their product $W_L \cdots W_1$ is bounded (Du, Hu, and Lee 2018; Bah et al. 2019).

A prototypical example to illustrate the second step: $f(x, y) = (xy - 1)^2$.

$$\begin{cases} \dot{x} = -2(xy - 1)y \\ \dot{y} = -2(xy - 1)x \end{cases} \implies \dot{x}x = \dot{y}y \implies x^2(t) - y^2(t) = c_1$$

Since $f(x(t), y(t))$ is decreasing and nonnegative, $|x(t)y(t)| \leq c_2$. Done!

Bounded subgradient trajectories - proof illustration

For the first step:

- 1 Apply SVD $X = U\Sigma V^T$ with orthogonal U, V and

$$\Sigma = \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_r) \succ 0.$$

$$\longrightarrow \dot{\bar{W}}_i = -2(\bar{W}_L \cdots \bar{W}_{i+1})^T (\bar{W}_L \cdots \bar{W}_1 \Sigma - Z) (\bar{W}_{i-1} \cdots \bar{W}_1 \Sigma)^T.$$

- 2 Partition \bar{W}_1 into dynamic and static components,

$$\bar{W}_1 = [\bar{W}_{11} \quad \bar{W}_{12}], \quad Z = [Z_1 \quad Z_2].$$

$$\longrightarrow \begin{cases} \dot{\bar{W}}_{11} = -2(\bar{W}_L \cdots \bar{W}_2)^T (\bar{W}_L \cdots \bar{W}_2 \bar{W}_{11} \Lambda - Z_1) \Lambda^T, \text{ and } \dot{\bar{W}}_{12} = 0 \\ \dot{\bar{W}}_i = -2(\bar{W}_L \cdots \bar{W}_{i+1})^T (\bar{W}_L \cdots \bar{W}_2 \bar{W}_{11} \Lambda - Z_1) (\bar{W}_{i-1} \cdots \bar{W}_2 \bar{W}_{11} \Lambda)^T \end{cases}$$

- 3 By change of variables,

$$\dot{\widetilde{W}}_i = -2(\widetilde{W}_L \cdots \widetilde{W}_{i+1})^T (\widetilde{W}_L \cdots \widetilde{W}_1 \Lambda - Z_1) (\widetilde{W}_{i-1} \cdots \widetilde{W}_1 \Lambda)^T$$

describes the gradient trajectory of $g(\widetilde{W}_1, \dots, \widetilde{W}_L) := \|\widetilde{W}_L \cdots \widetilde{W}_1 \Lambda - Z_1\|_F^2$.

Thank you for your attention!







Feel free to contact us at

xl3040@columbia.edu


cj2638@columbia.edu

for questions or suggestions.

References I

-  Bah, Bubacarr et al. (2019). “Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers”. In: *arXiv preprint arXiv:1910.05505*.
-  Baldi, Pierre and Kurt Hornik (1989). “Neural networks and principal component analysis: Learning from examples without local minima”. In: *Neural networks* 2.1, pp. 53–58.
-  Clarke, F. H. (1990). *Optimization and Nonsmooth Analysis*. Philadelphia: SIAM Classics in Applied Mathematics.
-  Du, Simon S, Wei Hu, and Jason D Lee (2018). “Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced”. In: *arXiv preprint arXiv:1806.00900*.
-  Josz, Cédric and Lexiao Lai (2021). “Nonsmooth rank-one matrix factorization landscape”. In: *Optimization Letters*, pp. 1–21.
-  Kawaguchi, Kenji (2016). “Deep Learning without Poor Local Minima”. In: *Advances in Neural Information Processing Systems*. Vol. 29. PMLR.

References II

-  Laurent, Thomas and James Brecht (2018). “Deep linear networks with arbitrary loss: All local minima are global”. In: *International conference on machine learning*. PMLR, pp. 2902–2907.
-  Li, Shuang et al. (2020). “The global geometry of centralized and distributed low-rank matrix recovery without regularization”. In: *IEEE Signal Processing Letters* 27, pp. 1400–1404.
-  Venturi, Luca, Afonso S Bandeira, and Joan Bruna (2019). “Spurious valleys in one-hidden-layer neural network optimization landscapes”. In: *Journal of Machine Learning Research* 20, p. 133.
-  Zhang, Li (2019). “Depth creates no more spurious local minima”. In: *arXiv preprint arXiv:1901.09827*.