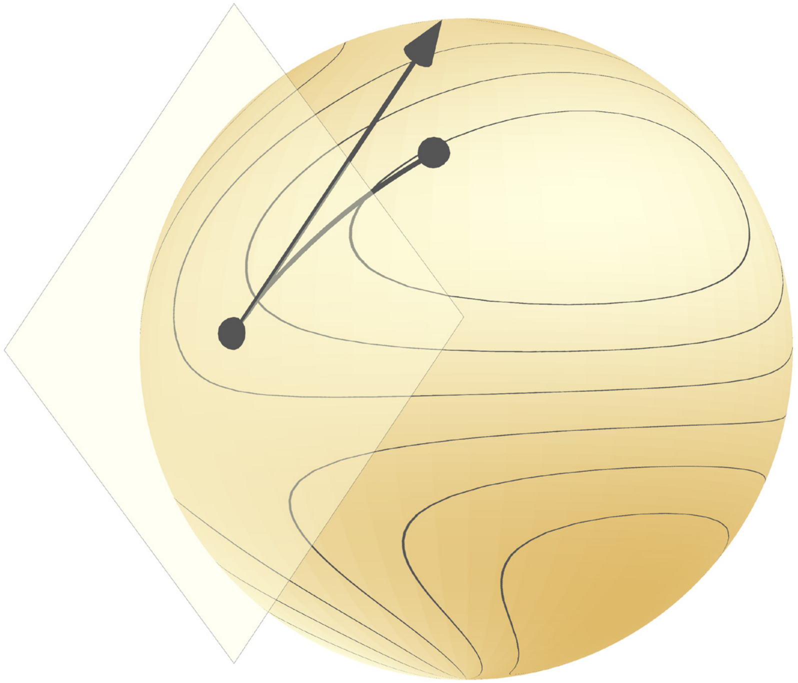
Blog: racetothebottom.xyz

nonconvex just means not convex

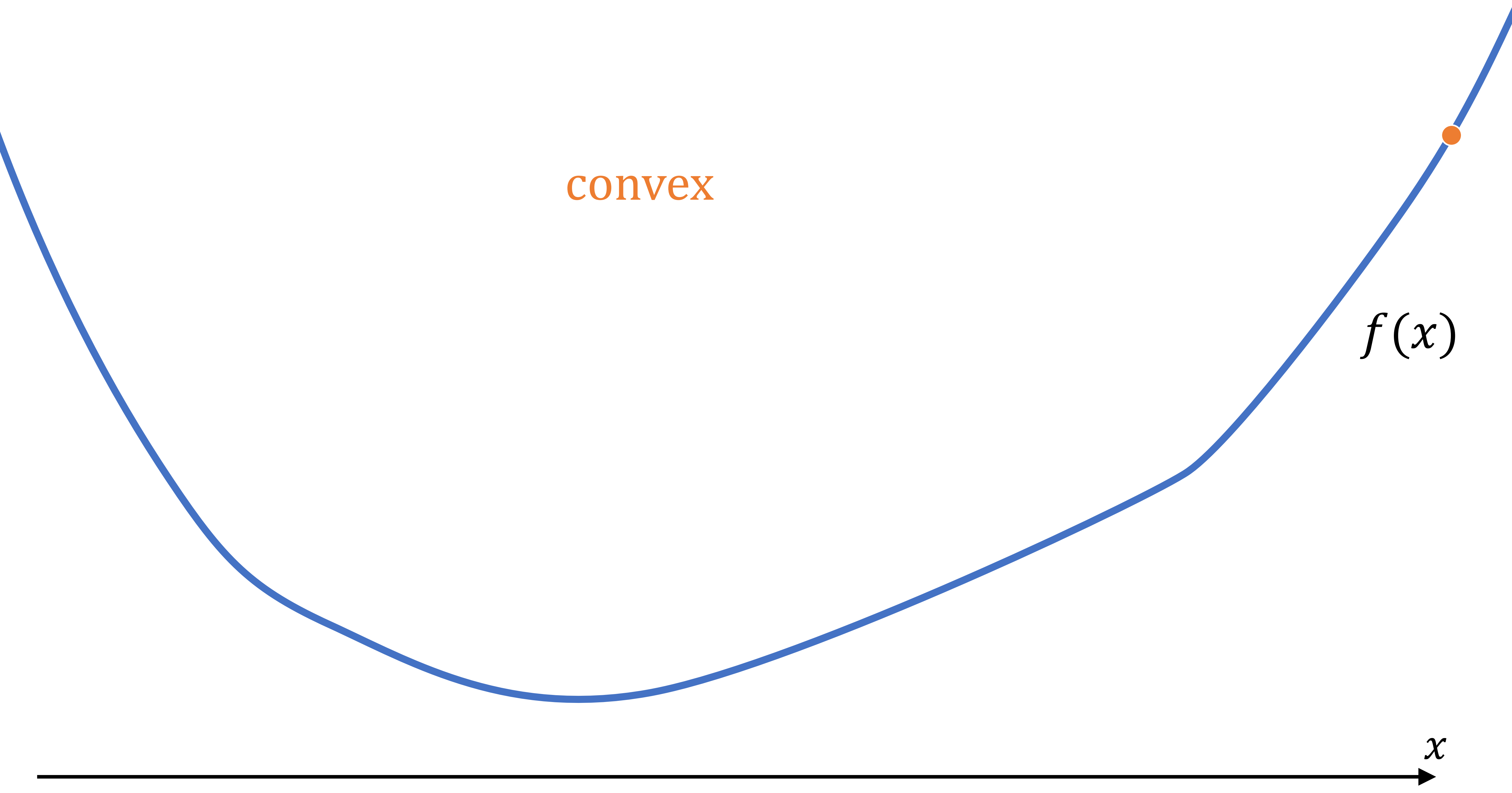
Journée SIGMA–MODE, January 30, 2024

Nicolas Boumal – chair of continuous optimization

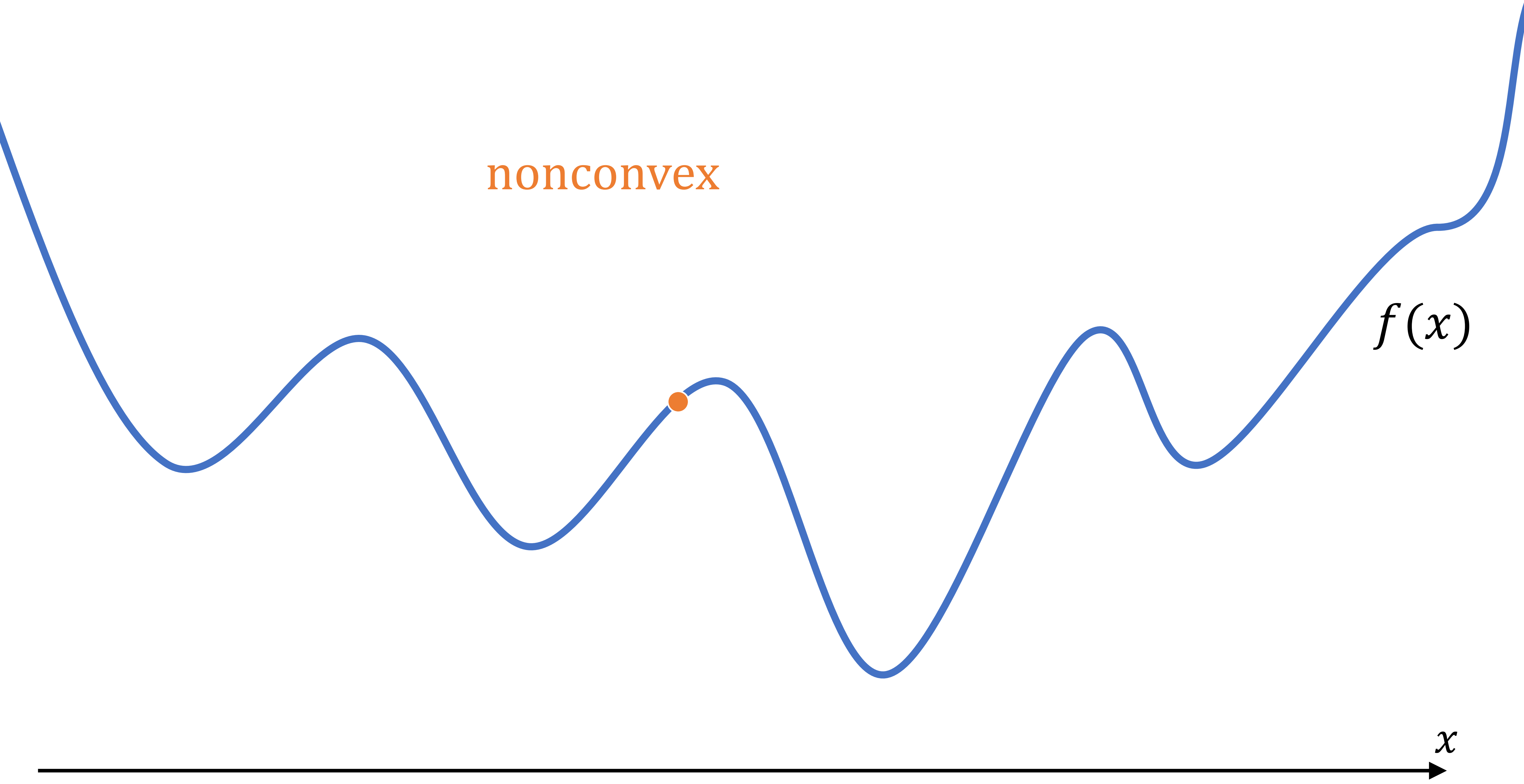
Institute of Mathematics, EPFL



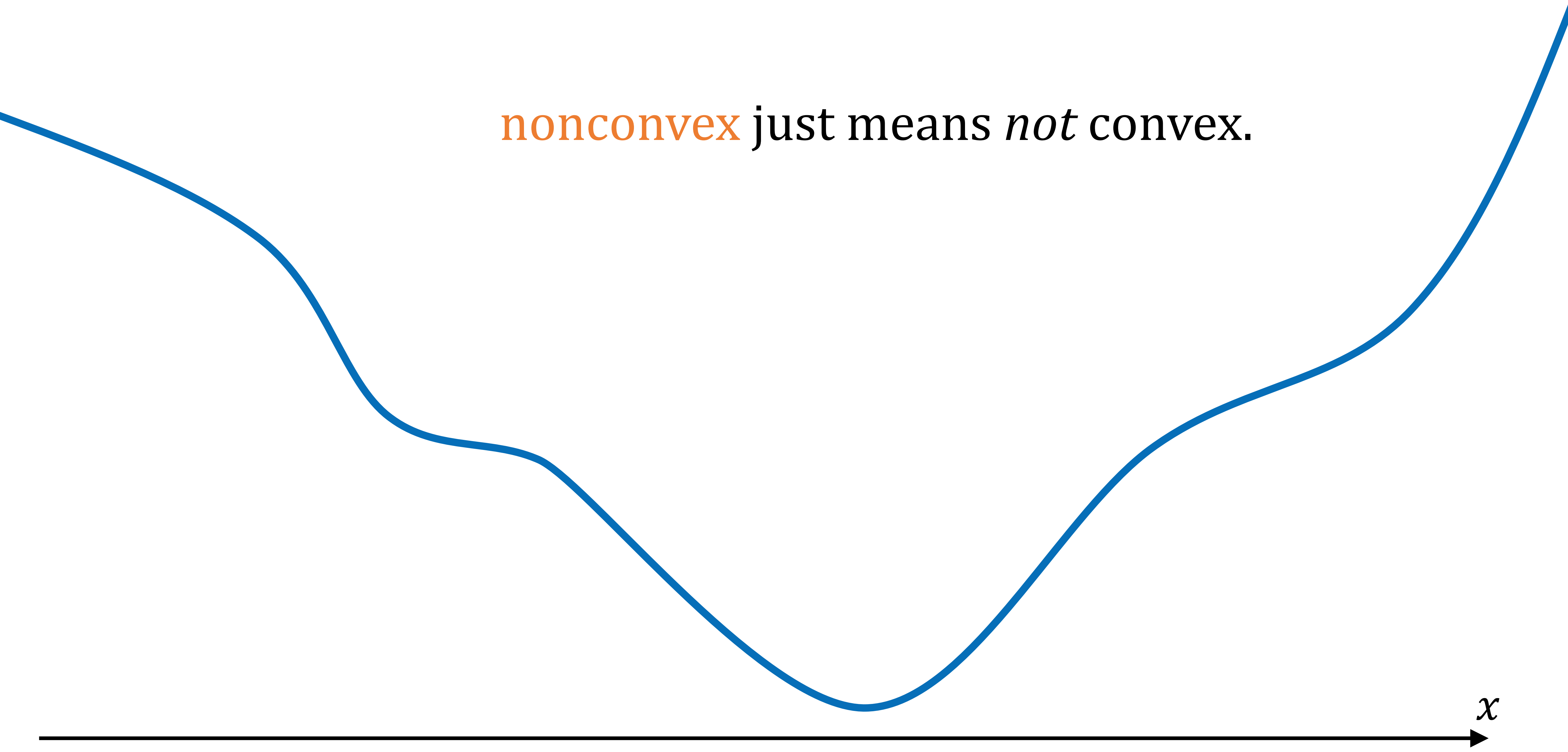

$$\min_{x} f(x)$$



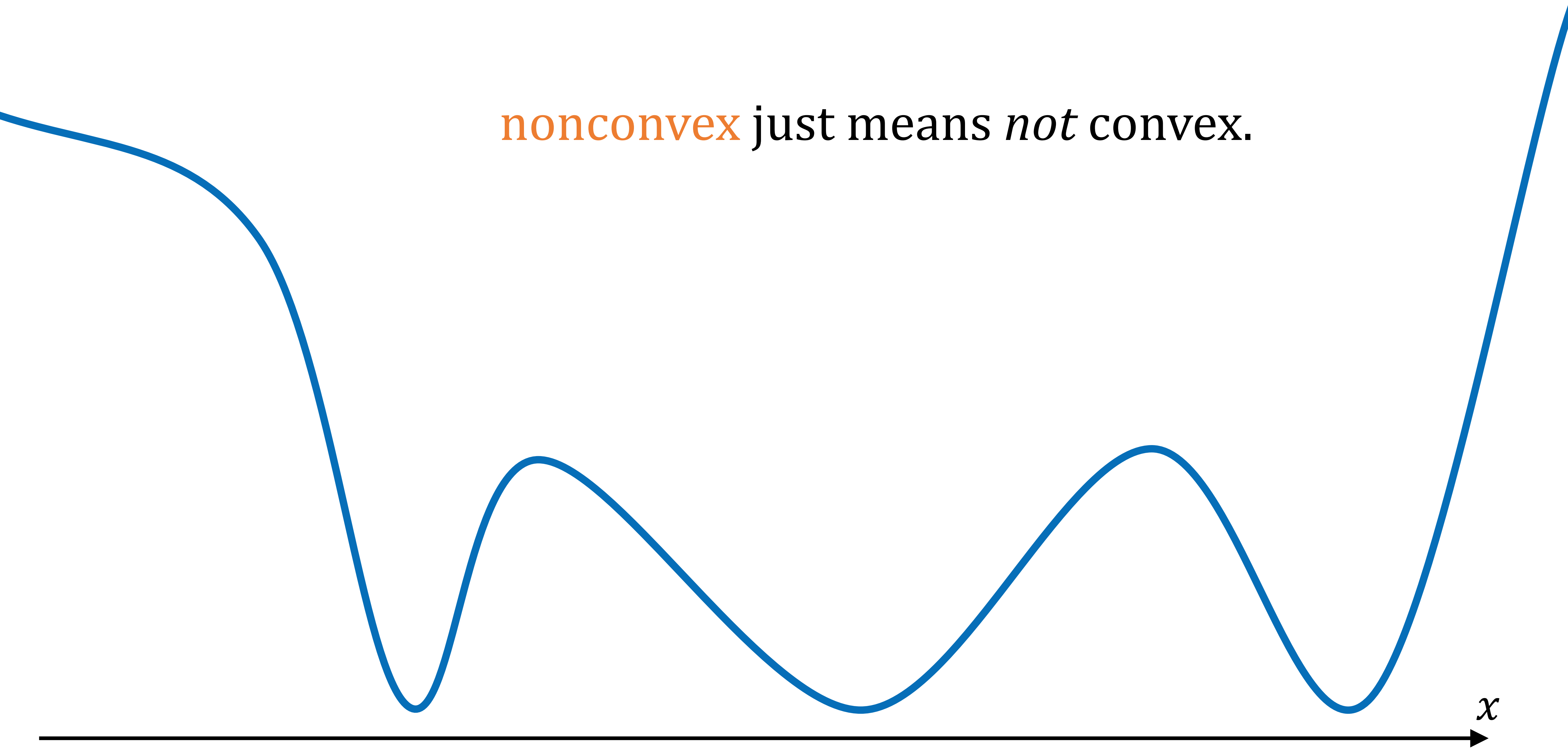
convex
$f(x)$
x


nonconvex
f(x)
x

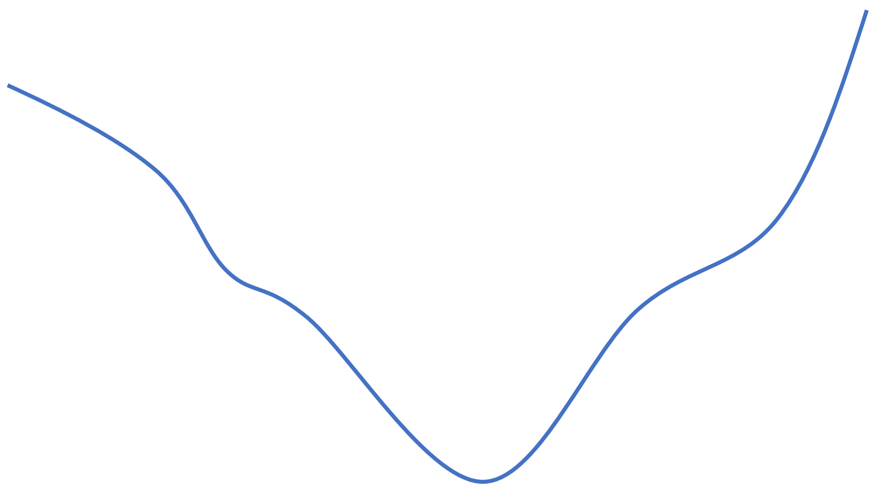
nonconvex just means *not* convex.

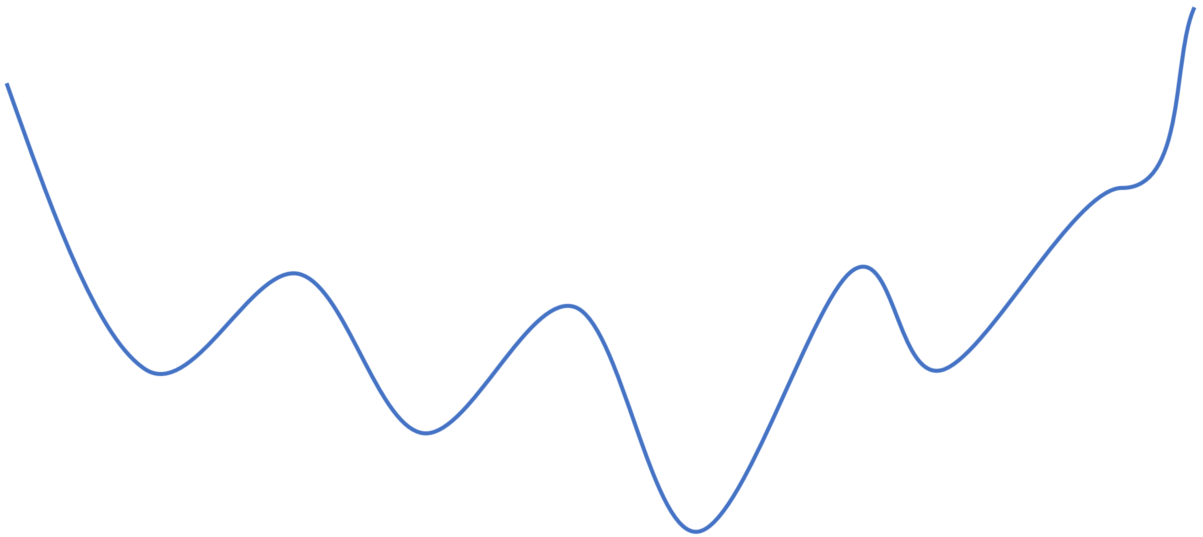


nonconvex just means *not* convex.



> *"Using a term like nonlinear science is like referring to the bulk of zoology as the study of non-elephant animals."*
>
> —Stanisław Ulam





Pockets of benign non-convexity: Ju Sun's list

sunju.org/research/nonconvex, ~900 papers in March 2021; categories:

Matrix Completion/Sensing

Tensor Recovery/Decomposition &
Hidden Variable Models

Phase Retrieval

Dictionary Learning

Deep Learning

Sparse Vectors in Linear Subspaces

Nonnegative/Sparse
Principal Component Analysis

Mixed Linear Regression

Blind Deconvolution/Calibration

Super Resolution

Synchronization Problems
Community Detection

Joint Alignment

Numerical Linear Algebra

Bayesian Inference

Empirical Risk Minimization &
Shallow Networks

System Identification

Burer-Monteiro Style Decomposition Algorithms

Generic Structured Problems

Nonconvex Feasibility Problems

Separable Nonnegative Factorization (NMF)

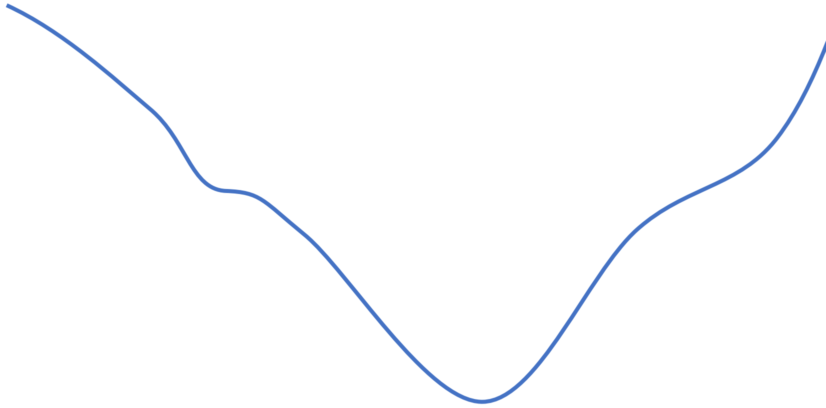
Good things happen but it's hard to tell

In an intro course to optimization, we learn how to spot convexity.

In contrast, for nonconvex problems, analyses are case-by-case.

E.g., some landscapes have strict saddles:

$$\mathrm{grad} f(x) = 0, \mathrm{Hess} f(x) \succcurlyeq 0 \Rightarrow x \text{ optimal}$$

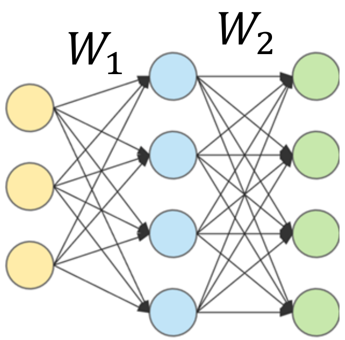


Proofs are often a whole paper…

It would be nice to have more tools to make proofs easier to build.

Tools to study nonconvex landscapes?

Example 1: Shallow linear networks



$$\min_{W_1, W_2} \|W_2 W_1 A - B\|_{\mathrm{F}}^2$$

Example 2: Rayleigh quotient

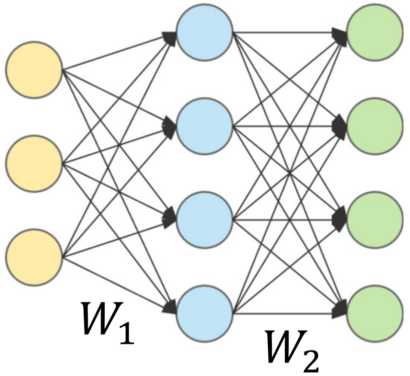
$$\min_{y} y^\top A y \quad \text{subject to} \quad \|y\| = 1$$

These problems are known to be benign (strict saddles).

Could we rediscover that by combining **reusable** facts?

Joint work with
Eitan Levin (Caltech) +
Joe Kileel (UT Austin)

Example 1: Shallow linear networks



$$\min_{W_1,W_2} \|W_2 W_1 A - B\|_F^2 \leftarrow g$$

Nonconvex due to product $W_2 W_1$.

Factor g through $\varphi(W_1, W_2) = W_2 W_1$:

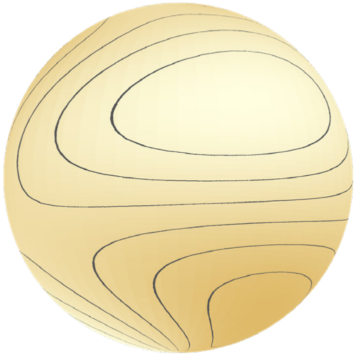
$$\min_X \|XA - B\|_F^2 \leftarrow f$$

$$\begin{tikzcd}
\{(W_1, W_2)\} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\
\{X\} \arrow[r, "f"'] & \mathbf{R}
\end{tikzcd}$$

Key facts:
f is convex, so: critical $\Rightarrow$ optimal
φ maps 2nd order critical points to critical points

Baldi & Hornik 1989, *Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima*
Lu and Kawaguchi 2017, *Depth Creates No Bad Local Minima*
Ha, Liu & Barber 2020, *An Equivalence between Critical Points for Rank Constraints Versus Low-Rank Factorizations*

Example 2: Rayleigh quotient



$$\min_{y \in \mathbf{R}^n} \overbrace{y^\top A y}^{g} \quad \text{s.t.} \quad \|y\|^2 = 1$$

We only know D, V *exist*!

We know $A = VDV^\top$ with
$D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and V orthogonal. So:

$$g(y) = y^\top A y = \left(V^\top y\right)^\top D\left(V^\top y\right) = \textstyle\sum_i \lambda_i \left(V^\top y\right)_i^2$$

Thus, $g(y) = f\big(\varphi(y)\big)$ where

$$f(x) = \textstyle\sum_i \lambda_i x_i \quad \text{and} \quad \varphi(y) = \left(V^\top y\right)^{\odot 2}$$

Notice: $\varphi = (\text{entrywise squaring}) \circ (\text{rotation})$

$$\begin{tikzcd}
\underset{\text{Sphere}}{y_1^2 + \cdots + y_n^2 = 1} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\
\underset{\substack{x_1 + \cdots + x_n = 1 \\ x_1 \geq 0, \dots, x_n \geq 0}}{\text{Simplex}} \arrow[r, "f"'] & \mathbf{R}
\end{tikzcd}$$

Key facts:
f and simplex are convex, so critical $\Rightarrow$ optimal
φ maps 2nd order critical points to critical points

Leake & Vishnoi 2021, *Optimization and Sampling Under Continuous Symmetry: Examples and Lie Theory*
Li, McKenzie & Yin 2021, *From the simplex to the sphere: Faster constrained optimization using the Hadamard parametrization*

General view: problems paired via a lift φ

$$\min_{y \in \mathcal{M}} g(y) \qquad \min_{x \in \mathcal{X}} f(x)$$

$$\begin{tikzcd}
\mathcal{M} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\
\mathcal{X} \arrow[r, "f"'] & \mathbf{R}
\end{tikzcd}$$

How do their landscapes compare?

E.g., if y is a local minimum for $g|_{\mathcal{M}}$,

is $\varphi(y)$ a local minimum for $f|_{\mathcal{X}}$?

Answer: yes for *all* f if and only if φ is open at y.

Example: $Y \mapsto YY^\top$ is open everywhere, but $(L, R) \mapsto LR^\top$ is not.

General view: problems paired via a lift φ

$$\min_{y \in \mathcal{M}} g(y) \qquad \min_{x \in \mathcal{X}} f(x)$$

$$\begin{tikzcd}
\mathcal{M} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\
\mathcal{X} \arrow[r, "f"'] & \mathbf{R}
\end{tikzcd}$$

How do their landscapes compare?

E.g., if y is first-order critical for $g|_{\mathcal{M}}$,

is $\varphi(y)$ first-order critical for $f|_{\mathcal{X}}$?

Answer: yes *for all f* iff $\operatorname{image}\big(\mathrm{D}\varphi(y)\big)$ = tangent cone $\mathrm{T}_{\varphi(y)}\mathcal{X}$.

Rarely true! In particular, requires tangent cones to be linear.

General view: problems paired via a lift φ

$$\min_{y \in \mathcal{M}} g(y) \qquad\qquad \min_{x \in \mathcal{X}} f(x)$$

$$\begin{tikzcd}
\mathcal{M} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\
\mathcal{X} \arrow[r, "f"'] & \mathbf{R}
\end{tikzcd}$$

How do their landscapes compare?

E.g., if y is **second**-order critical for $g|_{\mathcal{M}}$,
is $\varphi(y)$ **first**-order critical for $f|_{\mathcal{X}}$?

Answer: yes for *all* f iff [see paper for characterization].

Frequent: $Y \mapsto YY^\top$, $(L, R) \mapsto LR^\top$, other low-rank lifts, $y \mapsto y^{\odot 2}$, ...

General view: problems paired via a lift φ

$$\min_{y \in \mathcal{M}} g(y) \qquad \min_{x \in \mathcal{X}} f(x)$$

How do their landscapes compare?

E.g., if y is so-and-so for $g|_{\mathcal{M}}$,

is $\varphi(y)$ a this-or-that for $f|_{\mathcal{X}}$?

$$\begin{tikzcd}
\mathcal{M} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\
\mathcal{X} \arrow[r, "f"'] & \mathbf{R}
\end{tikzcd}$$

Key insight:
The relations are largely dictated by φ, independently of cost functions.
Thus, facts about lifts are reusable.

$\mathcal{M}$	$\mathcal{X} = \varphi(\mathcal{M})$	φ	local $\Rightarrow$ local	$1 \Rightarrow 1$	$2 \Rightarrow 1$
Manifold	Submanifold	Submersion	✓	✓	✓
Sphere	Simplex	$x \mapsto x^{\odot 2}$ (Hadamard)	✓	φ^{-1}(interior)	✓
(Sphere in $\mathbf{R}^n$)^n	Stochastic matrices	Hadamard on each col or row	✓	φ^{-1}(interior)	✓
Sphere in $\mathbf{R}^{n+1}$	Ball in $\mathbf{R}^n$	Coordinate projection	✓	φ^{-1}(interior)	✓
Torus in $\mathbf{R}^{n+1}$	Annulus in $\mathbf{R}^n$	See paper	✓	φ^{-1}(interior)	✓
$\mathcal{A}(YY^\top) = b$, smooth	$X \succcurlyeq 0, \mathcal{A}(X) = b$	$Y \mapsto YY^\top$ (Burer-Monteiro)	✓	Y full rank	✓
(L, R) in $\mathbf{R}^{m\times r} \times \mathbf{R}^{n \times r}$	$\mathrm{rank}(X) \leq r$	$(L, R) \mapsto LR^\top$	balanced*	L, R full rank	✓
$X \in \mathbf{R}^{m\times n}, \mathcal{S} \subseteq \ker X$, $\dim \mathcal{S} = n - r$	$\mathrm{rank}(X) \leq r$	$(X, \mathcal{S}) \mapsto X$ (desingularization)	$\mathrm{rank}(X) = r$	$\mathrm{rank}(X) = r$	✓
Linear space of factors	Low-rank tensors	CP, TT, Tucker, …	✗	✗	✗

* balanced means
$\mathrm{rank}(L) = \mathrm{rank}(R) = \mathrm{rank}(LR^\top)$

$$\begin{tikzcd}
\mathcal{M} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\
\mathcal{X} \arrow[r, "f"'] & \mathbf{R}
\end{tikzcd}$$

More in our paper

Blog: racetothebottom.xyz

The effect of smooth parametrizations on nonconvex optimization landscapes

with Eitan Levin and Joe Kileel

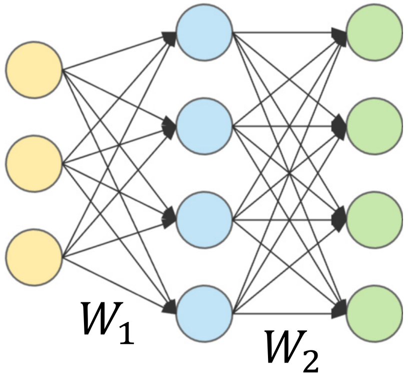
arxiv.org/abs/2207.03512

$$\begin{tikzcd} \mathcal{M} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\ \mathcal{X} \arrow[r, "f"'] & \mathbf{R} \end{tikzcd}$$

Some future directions:

- Explore new lifts
- Study compositionality
- Apply to new landscapes
- Explore other properties
  E.g., local $\Rightarrow$ 1
- Prove no good lift exists for $\mathcal{X}$

Details in December 2023
blog posts: racetothebottom.xyz



Example 1’: Narrow linear networks

$$\min_{W_1 \in \mathbf{R}^{r\times n}, W_2 \in \mathbf{R}^{m \times r}} \underbrace{\|W_2 W_1 A - B\|_F^2}_{g}$$

Nonconvex due to product $W_2 W_1$.

Factor g through $\varphi(W_1, W_2) = W_2 W_1$:

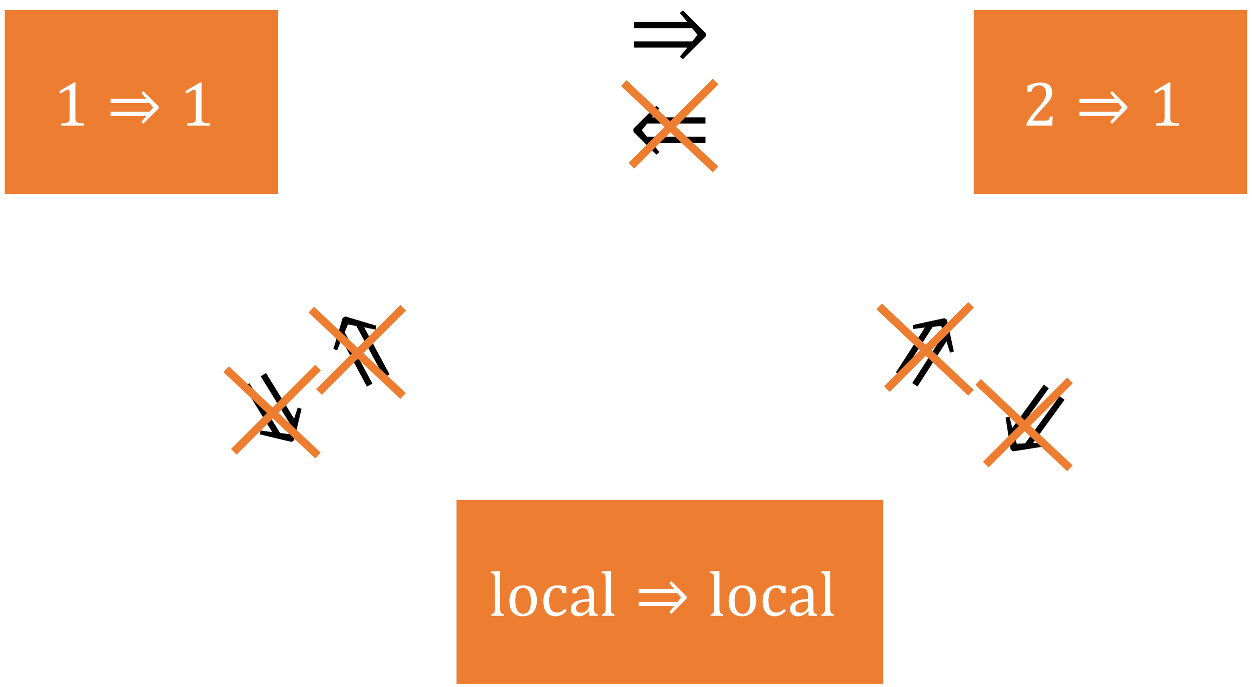
$$\min_{X \in \mathbf{R}^{m\times n}, \operatorname{rank}(X) \le r} \underbrace{\|XA - B\|_F^2}_{f}$$

$$\begin{tikzcd} \mathbf{R}^{r\times n} \times \mathbf{R}^{m\times r} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\ \{X : \operatorname{rank}(X) \le r\} \arrow[r, "f"'] & \mathbf{R} \end{tikzcd}$$

Key facts (see blog; A full row rank):
$\operatorname{rank}(X) < r$ and 1-critical $\Rightarrow$ optimal
$\operatorname{rank}(X) = r$ and 2-critical $\Rightarrow$ optimal
φ maps 2-critical points to 1-critical points
φ maps 2-critical points of rank r to 2-critical points

Baldi & Hornik 1989, *Neural Networks and Principal Component Analysis*
Lu and Kawaguchi 2017, *Depth Creates No Bad Local Minima*
Ha, Liu & Barber 2020, *An Equivalence between Critical Points for Rank Constraints Versus Low-Rank Factorizations*

Lift properties are fairly independent



Remark 2.13 (Relations between lift properties). *Aside from Proposition 2.12, the only relation between the three properties in Definition 2.2 is that "1 $\Rightarrow$ 1" at y implies "2 $\Rightarrow$ 1" at y (since 2-critical points are 1-critical). None of the other possible implications hold in general: The desingularization lift* (Desing) *shows that "2 $\Rightarrow$ 1" at y implies neither "1 $\Rightarrow$ 1" nor "local $\Rightarrow$ local" at y in general. The example $\varphi(x) = x^3$ viewed as a lift from $\mathcal{M} = \mathbb{R}$ to $\mathcal{X} = \mathbb{R}$ satisfies "local $\Rightarrow$ local" at the origin but neither "2 $\Rightarrow$ 1" nor "1 $\Rightarrow$ 1", hence "local $\Rightarrow$ local" does not imply the other two properties. Finally, the standard parametrization of the cochleoid curve [59] satisfies "1 $\Rightarrow$ 1" but not "local $\Rightarrow$ local" at all preimages of the origin, hence "1 $\Rightarrow$ 1" does not imply "local $\Rightarrow$ local".*

$$\begin{tikzcd} y \in \mathcal{M} \arrow[d, "\varphi"'] \arrow[dr, "g = f \circ \varphi"] & \\ x \in \mathcal{X} \arrow[r, "f"'] & \mathbf{R} \end{tikzcd}$$

Upstairs—e.g., $\min_{W_1, W_2} g(W_1, W_2) = f(W_2 W_1)$

Downstairs—e.g., $\min_{\operatorname{rank}(X) \le r} f(X)$

	global min $\Rightarrow$	**local min** $\Rightarrow$	**2-critical** $\Rightarrow$	**1-critical**
global min $\Downarrow$				
local min $\Downarrow$				
2-critical $\Downarrow$				
1-critical				

$\Uparrow$ If x is a [see rows], then $y \in \varphi^{-1}(x)$ is a [see cols].

$\Leftarrow$ If $y \in \mathcal{M}$ is a [see cols], then $x = \varphi(y)$ is a [see rows].