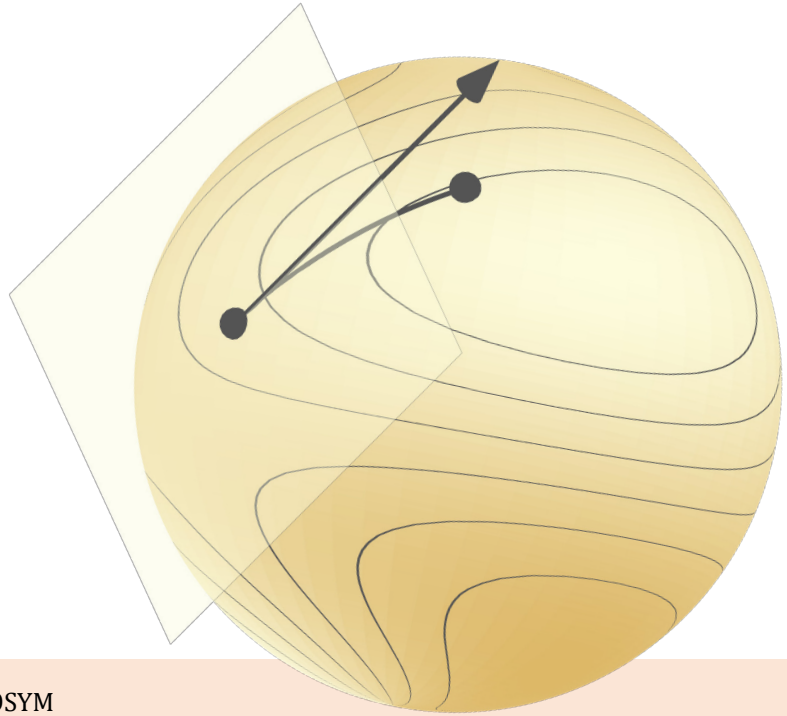
Non-convex optimization when the solution is not unique: A kaleidoscope of favorable conditions

February 2023

Nicolas Boumal, with Quentin Rebjock

OPTIM, Institute of Mathematics, EPFL



The paper on arXiv treats optimization on manifolds. This talk specialized to $\mathbf{R}^n$.

Funds: SEFRI/ERC GEOSYM

This talk is about fast local convergence

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

We use algorithms to compute $x \in \mathbf{R}^n$ such that $f(x) \in \mathbf{R}$ is small.

Ideally, we want a *global* minimum, but that's hard. *Local* will do.

x is a local minimum if $f(x) \leq f(y)$ for all y in a neighborhood of x.

Say f is C^2 (continuous Hessian). Focus on local convergence rates.

A simple look at the one-dimensional case

Say $f(x) = x^4$. Minimizer is $x^* = 0$.

Then $\nabla f(x) = 4x^3$ and $\nabla^2 f(x) = 12x^2$.

Gradient descent: $x_{k+1} = x_k - \alpha \nabla f(x_k) = \left(1 - 4\alpha x_k^2\right) x_k$.

Only sublinear convergence (if $0 < \alpha < 1/2x_0^2$).

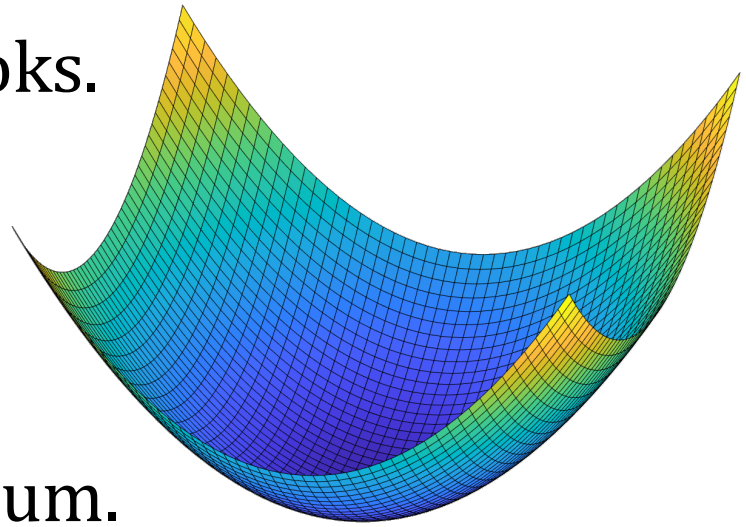
Newton's method: $x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} [\nabla f(x_k)] = \frac{2}{3} x_k$.

Only linear convergence.

The culprit: $\nabla^2 f(x^*) \not\succ 0$ can kill local rates

Indeed, if we assume a positive definite Hessian at x^*, then typical algorithms enjoy their "normal" fast local convergence rates.

This is what we find in classic optimization textbooks.



The issue is: for $f \in C^2$, at a critical point x^*,

$$\nabla^2 f(x^*) \succ 0 \quad \Rightarrow \quad x^* \text{ is an isolated local minimum.}$$

But quite often, minima are *not* isolated

And therefore, there is no way $\nabla^2 f(x^*) \succ 0$ for such applications.

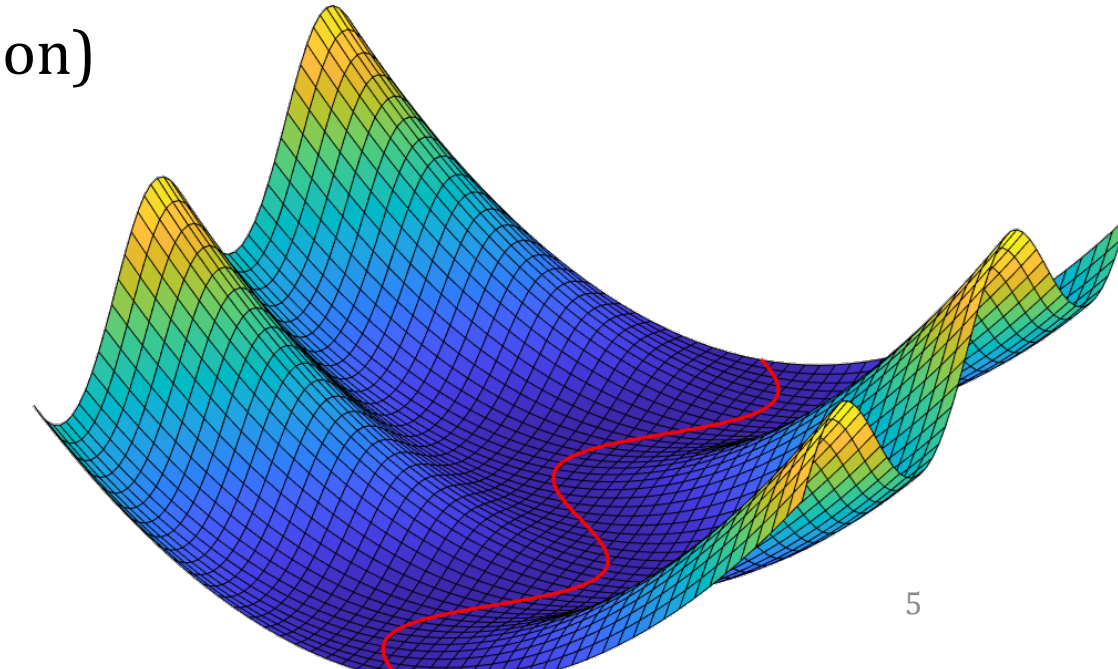
Overparameterized regression / neural network (e.g., $\min_x \|F(x) - b\|^2$)

Redundant parameterization (e.g., $(L, R) \mapsto LR^\top$)

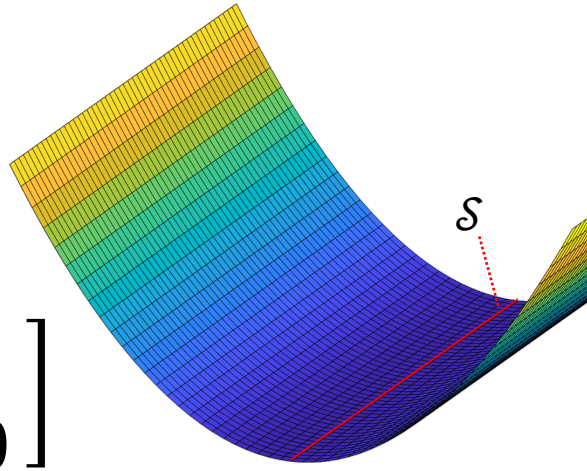
Symmetry (e.g., $f(x)$ invariant to rotation)

Yet, we often still see fast convergence.

Why?



Insights from a simple 2-D example



$$f(x,y) = \frac{\mu}{2}x^2 \qquad \nabla f(x,y) = \begin{bmatrix} \mu x \\ 0 \end{bmatrix} \qquad \nabla^2 f(x,y) = \begin{bmatrix} \mu & \\ & 0 \end{bmatrix}$$

The set of minimizers is a line $\mathcal{S} = \{(0,y) : y \in \mathbf{R}\}$, where $f^* = 0$.

The gradient and Hessian "ignore" the tangent direction.

As a result, typical algorithms only "see" the direction that matters.

$\ker \nabla^2 f(0,y) = \{(0,u) : u \in \mathbf{R}\}$
All other eigenvalues $= \mu$

$$f(x,y) - f^* = \frac{1}{2\mu}\|\nabla f(x,y)\|^2$$

$$f(x,y) - f^* = \frac{\mu}{2}\operatorname{dist}\big((x,y),\mathcal{S}\big)^2$$

$$\|\nabla f(x,y)\| = \mu \operatorname{dist}\big((x,y),\mathcal{S}\big)$$

$\ker \nabla^2 f(0,y) = \{(0,u) : u \in \mathbf{R}\}$
All other eigenvalues $= \mu$

$f(x,y) - f^* = \frac{1}{2\mu}\|\nabla f(x,y)\|^2$

$f(x,y) - f^* = \frac{\mu}{2}\mathrm{dist}\big((x,y),\mathcal{S}\big)^2$

$\|\nabla f(x,y)\| = \mu\ \mathrm{dist}\big((x,y),\mathcal{S}\big)$

Take 1: Morse–Bott

If the minima are not singletons, maybe they are still "nice" sets.

Say the set of minima $\mathcal{S}$ is an embedded submanifold of $\mathbf{R}^n$.

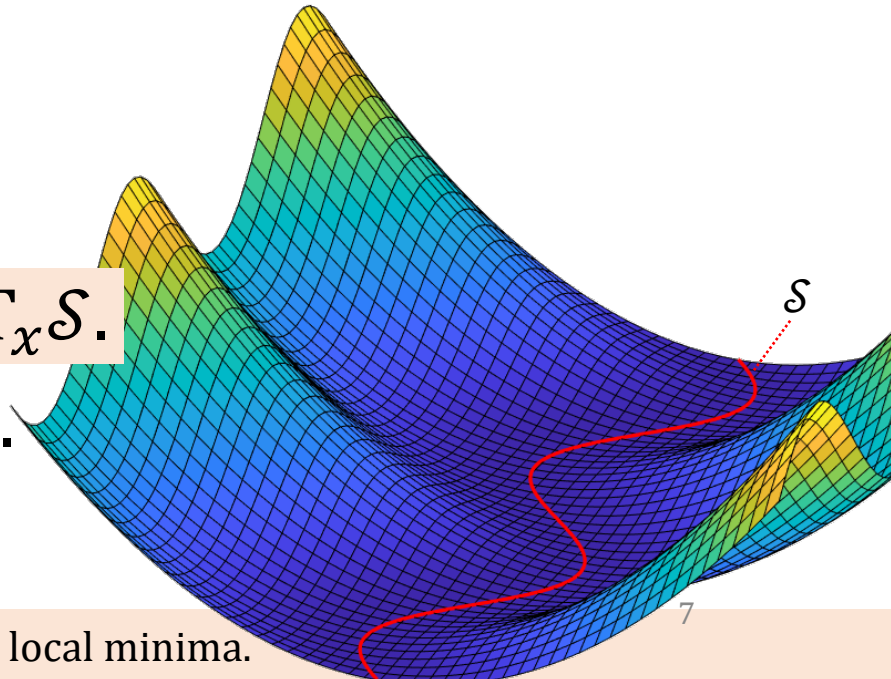
Surely, $\nabla f(x) = 0$ and $\nabla^2 f(x) \succcurlyeq 0$ for $x \in \mathcal{S}$.

Also, $\nabla^2 f(x)[v] = 0$ if $v \in \mathrm{T}_x\mathcal{S}$.

Def.: f is MB if $\mathcal{S}$ is smooth and $\ker \nabla^2 f(x) = \mathrm{T}_x\mathcal{S}$.

Can then show good rates for various methods.

We did not find many refs; see e.g. Fehrman, Gess & Jentzen 2020.



To avoid topological curiosities, here $\mathcal{S}$ is the set of global minimizers of f. The paper treats local minima.

$\ker \nabla^2 f(0,y) = \{(0,u) : u \in \mathbf{R}\}$
All other eigenvalues $= \mu$

$f(x,y) - f^* = \frac{1}{2\mu}\|\nabla f(x,y)\|^2$

$f(x,y) - f^* = \frac{\mu}{2}\operatorname{dist}\big((x,y),\mathcal{S}\big)^2$

$\|\nabla f(x,y)\| = \mu \operatorname{dist}\big((x,y),\mathcal{S}\big)$

Take 2: Polyak–Łojasiewicz

In 1963, Polyak studied f where gradient norm^2 $\sim$ optimality gap.

This is also called gradient dominance and is a particular case of Kurdyka–Łojasiewicz.

Def.: f is PŁ at x^* if $f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2$ for x around x^*.

This includes strongly convex functions, and much more.

Linear cvgce for GD, and superlinear cvgce for cubic regularization.

Polyak 1963; Nesterov & Polyak 2006. Many, many, many recent analyses of algorithms under PŁ.

$\ker \nabla^2 f(0,y) = \{(0,u) : u \in \mathbf{R}\}$ All other eigenvalues $= \mu$	$f(x,y) - f^* = \frac{1}{2\mu}\|\nabla f(x,y)\|^2$
$f(x,y) - f^* = \frac{\mu}{2}\operatorname{dist}\big((x,y),\mathcal{S}\big)^2$	$\|\nabla f(x,y)\| = \mu \operatorname{dist}\big((x,y),\mathcal{S}\big)$

Take 3: Quadratic Growth

We could also assume that f grows fast as we move away from $\mathcal{S}$.

Already in Bonnans & Ioffe 1995, but likely much older.

Def.: f has QG at x^* if $f(x) - f(x^*) \geq \frac{\mu}{2}\operatorname{dist}(x,\mathcal{S})^2$ for x around x^*.

Interestingly, this one is well defined even if f is nonsmooth.

This has been used to show fast convergence in that setting.

Drusvyatskiy & Lewis 2016; Davis & Jiang 2022; Lewis & Tian 2022.

$\ker \nabla^2 f(0,y) = \{(0,u) : u \in \mathbf{R}\}$ All other eigenvalues $= \mu$	$f(x,y) - f^* = \frac{1}{2\mu}\|\nabla f(x,y)\|^2$
$f(x,y) - f^* = \frac{\mu}{2}\operatorname{dist}\big((x,y),\mathcal{S}\big)^2$	$\|\nabla f(x,y)\| = \mu \operatorname{dist}\big((x,y),\mathcal{S}\big)$

Take 4: Error Bound

We could assume the gradient grows fast as we move away from $\mathcal{S}$.

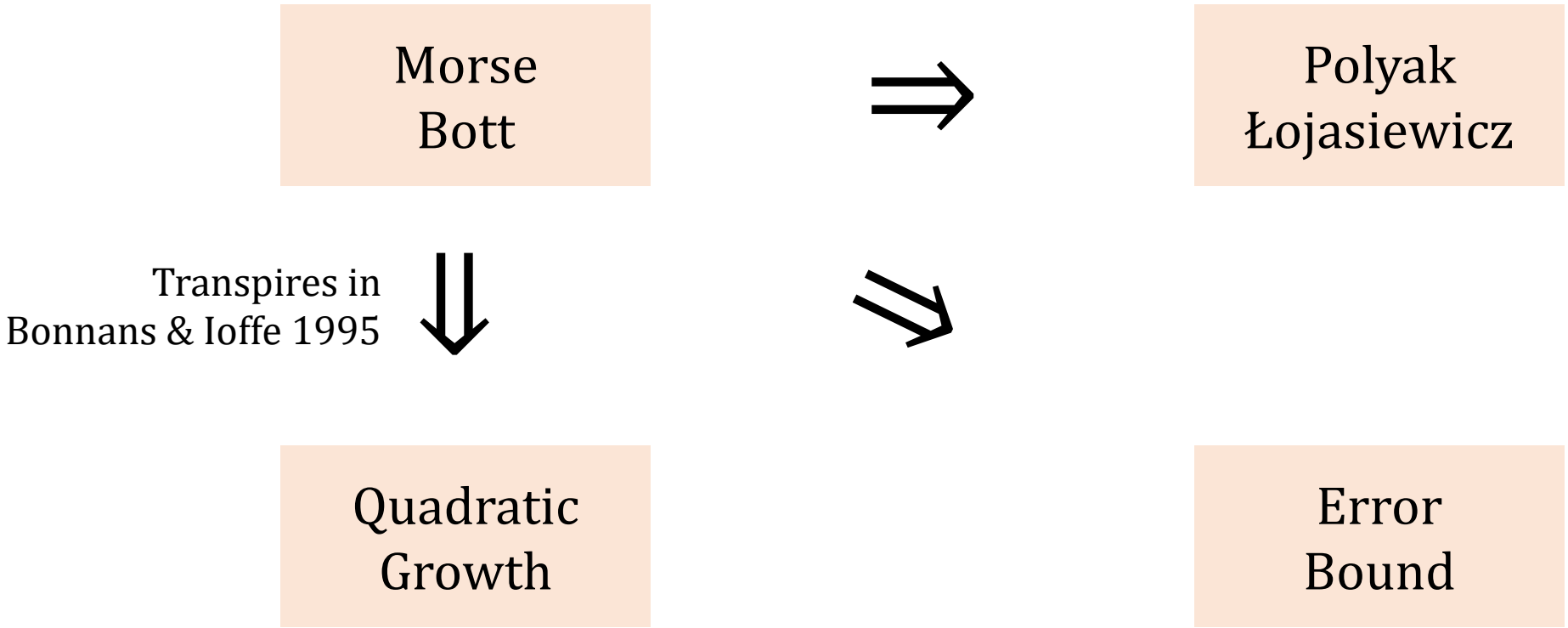
This seems to have originated with Luo & Tseng 1993.

Def.: f has EB at x^* if $\|\nabla f(x)\| \geq \mu \operatorname{dist}(x, \mathcal{S})$ for x around x^*.

Luo & Tseng showed linear convergence for several methods with EB.

Yue, Zhou & Man-Cho So 2019 showed quadratic cvgce for cubic regularization.

Some conditions imply others for $f \in C^2$



Morse–Bott is explicitly strong, partly because it requires a smooth solution set $\mathcal{S}$.

With a few simple Taylor expansion arguments, we can see MB ⇒ QG, EB, PŁ.

Some conditions imply others for $f \in C^2$

Feehan 2020 yields part of it for f analytic.
We show it for C^2. In particular,
PŁ implies smooth $\mathcal{S}$!

Morse
Bott

$\Leftarrow$

Polyak
Łojasiewicz

Transpires in
Bonnans & Ioffe 1995

$\Downarrow$

$\Leftarrow$

$\Updownarrow$

Classic for $f \in C^1$. E.g.,
Karimi, Nutini & Schmidt 2016

* Gradient flow argument ($f \in C^1$),
Ekeland's variational principle

Quadratic
Growth

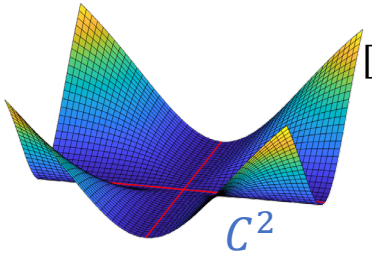
$\Rightarrow$

Error
Bound

We didn't see it stated.
Controls ∇f from f if C^2.

* PŁ ⇒ QG has a rich history: see Ioffe 2000, Bolte et al. 2017 (arXiv 2015); Drusvyatskiy et al. 2015, Karimi et al. 2016, Zhang 2017.

Actually, all conditions coincide for $f \in C^2$!



$[f \in C^1 \text{ and PŁ}] \not\Rightarrow \text{smooth } S$

$$f(x,y) = \frac{x^2y^2}{x^2+y^2}$$

C^2

Morse
Bott

$\Leftarrow$

Polyak
Łojasiewicz

C^2 $\Downarrow$

$\Leftarrow$ C^1

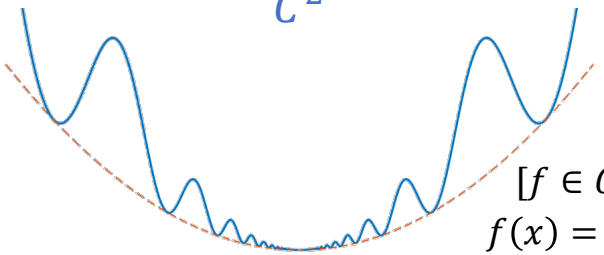
$\Updownarrow$ C^1 with some constant,
C^2 with better constant.

Quadratic
Growth

$\Rightarrow$

C^2

Error
Bound



$[f \in C^1 \text{ and QG}] \not\Rightarrow \text{EB, PŁ}$

$$f(x) = 2x^2 + x^2 \sin\left(1/\sqrt{|x|}\right)$$

Technical details of note

$\ker \nabla^2 f(x) = \mathrm{T}_x \mathcal{S}$ All other eigenvalues $\geq \mu$	$f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$
$f(x) - f^* \geq \frac{\mu}{2} \mathrm{dist}(x, \mathcal{S})^2$	$\|\nabla f(x)\| \geq \mu \, \mathrm{dist}(x, \mathcal{S})$

PŁ, QG and EB hold *in a neighborhood*. It may shrink along "⇒".

If one holds with μ, all hold for all $\mu' < \mu$ (trade-off with ngbhd).

For $f \in C^k$ with $k \geq 2$, we have PŁ ⇒ MB with $\mathcal{S}$ of class C^{k-1}.

PŁ ⇒ MB: Elements of proof

The most interesting bit is to show: PŁ ⇒ $\mathcal{S}$ smooth

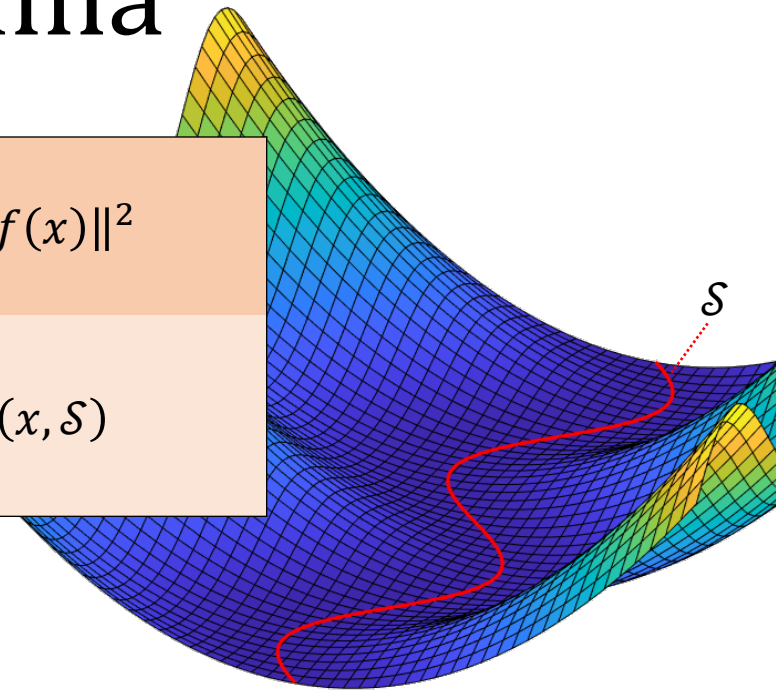
Pick $\bar{x} \in \mathcal{S}$. Let $P(\bar{x})$ be the projector to the image of $\nabla^2 f(\bar{x})$, and:

$$\mathcal{Z} = \{x \text{ close to } \bar{x} : P(\bar{x})\nabla f(x) = 0\}$$

Clearly, $\mathcal{Z}$ contains $\mathcal{S}$ (locally).
Also, $\mathcal{Z}$ is a smooth submanifold: study rank of $P(\bar{x})\nabla^2 f(x)$.
And we can show that PŁ implies $\mathcal{Z} = S$ (locally).

Take away for non-isolated minima

$\ker \nabla^2 f(x) = \mathrm{T}_x \mathcal{S}$ All other eigenvalues $\geq \mu$	$f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$
$f(x) - f^* \geq \frac{\mu}{2} \mathrm{dist}(x, \mathcal{S})^2$	$\|\nabla f(x)\| \geq \mu \, \mathrm{dist}(x, \mathcal{S})$



If f is C^2, those four conditions are equivalent.

Thus, assuming one of MB, PŁ, QG or EB, we can use all.

This helps analysis. In our paper: cubic regularization, trust regions.

E.g.: Nesterov & Polyak ’06 compared to Yue et al. ‘19 for cubic regularization.