Behind fast convergence rates in non convex optimization

Oleksii Kachaiev, **Hippolyte Labarrière**, Cesare Molinari, Silvia Villa

Journées SMAI-MODE
Nice, France

March 19, 2026

UniGe | DIBRIS | MaLGa MACHINE LEARNING GENOA CENTER

Context

My old life: For some **convex** f (and potentially more than convex):

$$\min_{x \in \mathcal{H}} \; f(x)$$

What can we get? Methods that provide

- convergence to a minimum (which is **global**),
- explicit **convergence rates**,
- fancy techniques for **acceleration** (inertia, preconditioning, Newton).

Context

Optimization nowadays: Training overparameterized models:

$$\min_{w \in \mathcal{W}} \; L(w)$$

Context

Optimization nowadays: Training overparameterized models:

$$\min_{w \in \mathcal{W}} \; L(w)$$

$\rightarrow L$ is **non convex**!

What can be ensured in general?

- convergence to . . . a **critical point**?
- convergence rates?

Context

Optimization nowadays: Training overparameterized models:

$$\min_{w \in \mathcal{W}} L(w)$$

$\rightarrow$ **However...**

[Submitted on 30 Mar 2022 (v1), last revised 20 Feb 2026 (this version, v5)]

Convergence of gradient descent for deep neural networks

Sourav Chatterjee

[1] See also [Oymak '19, Liu et al. '22, Buskulic et al. '24]

Context

Optimization nowadays: Training overparameterized models:

$$\min_{w \in \mathcal{W}} L(w)$$

$\rightarrow$ **However...**

[Submitted on 30 Mar 2022 (v1), last revised 20 Feb 2026 (this version, v5)]

Convergence of gradient descent for deep neural networks

Sourav Chatterjee

- Simple methods converge to a (potentially global) **minimizer**,
- **Linear convergence**(!)

$\rightarrow$ **What is hidden?**

[1] See also [Oymak '19, Liu et al. '22, Buskulic et al. '24]

Preliminaries

Gradient Flow (GF): For some initialization point $x_0 \in \mathbb{R}^d$:

$$\forall t \geq 0, \quad \dot{x}(t) + \nabla f(x(t)) = 0, \quad x(0) = x_0$$

Preliminaries

Gradient Flow (GF): For some initialization point $x_0 \in \mathbb{R}^d$:

$$\forall t \geq 0, \quad \dot{x}(t) + \nabla f(x(t)) = 0, \quad x(0) = x_0$$

- Discretization give **Gradient Descent**:

$$\forall k \in \mathbb{N}, \quad x_{k+1} = x_k - s\nabla f(x_k), \quad s > 0$$

- Always brings you **down**:

$$\frac{d}{dt} f(x(t)) = \langle \dot{x}(t), \nabla f(x(t)) \rangle = -\|\nabla f(x(t))\|^2 \leq 0$$

Geometric conditions

(Polyak)-Lojasiewicz (PL) inequality [Lojasiewicz '63, Polyak '63]:

$$\exists \mu > 0,\ \forall x \in \mathbb{R}^d, \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Geometric conditions

(Polyak)-Lojasiewicz (PL) inequality [Lojasiewicz '63, Polyak '63]:

$$\exists \mu > 0, \ \forall x \in \mathbb{R}^d, \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Getting linear rates without convexity:

$$\forall t \geq 0, \quad \dot{x}(t) + \nabla f(x(t)) = 0, \quad x(0) = x_0$$

Geometric conditions

(Polyak)-Lojasiewicz (PL) inequality [Lojasiewicz '63, Polyak '63]:

$$\exists \mu > 0,\ \forall x \in \mathbb{R}^d, \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Getting linear rates without convexity:

$$\forall t \geq 0, \quad \dot{x}(t) + \nabla f(x(t)) = 0, \quad x(0) = x_0$$

Supposing **PL** holds:

$$\frac{d}{dt}\left(f(x(t)) - f^*\right) = -\|\nabla f(x(t))\|^2 \leq -2\mu \left(f(x(t)) - f^*\right)$$

Geometric conditions

(Polyak)-Lojasiewicz (PL) inequality [Lojasiewicz '63, Polyak '63]:

$$\exists \mu > 0, \ \forall x \in \mathbb{R}^d, \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Getting linear rates without convexity:

$$\forall t \geq 0, \quad \dot{x}(t) + \nabla f(x(t)) = 0, \quad x(0) = x_0$$

Supposing **PL** holds:

$$\frac{d}{dt}\left(f(x(t)) - f^*\right) = -\|\nabla f(x(t))\|^2 \leq -2\mu\left(f(x(t)) - f^*\right)$$

Therefore,

$$f(x(t)) - f^* \leq \exp(-2\mu t)(f(x_0) - f^*)$$

Geometric conditions

(Polyak)-Lojasiewicz (PL) inequality [Lojasiewicz '63, Polyak '63]:

$$\exists \mu > 0,\ \forall x \in \mathbb{R}^d, \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Getting linear rates without convexity:

$$f(x(t)) - f^* \leq \exp(-2\mu t)(f(x_0) - f^*)$$

Are we already done?

Geometric conditions

(Polyak)-Lojasiewicz (PL) inequality [Lojasiewicz ’63, Polyak ’63]:

$$\exists \mu > 0,\ \forall x \in \mathbb{R}^d, \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Getting linear rates without convexity:

$$f(x(t)) - f^* \leq \exp(-2\mu t)(f(x_0) - f^*)$$

Are we already done?

No. **PL** is a restrictive assumption to hold globally:

$\rightarrow$ No critical points: ∇f only cancels at the global minimizer!

Geometric conditions

(Polyak)-Lojasiewicz (PL) inequality [Lojasiewicz '63, Polyak '63]:

$$\exists \mu > 0,\ \forall x \in \mathbb{R}^d, \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Where does $x(t)$ go? Under **PL**,

$$\forall t \geq 0, \quad \|x(t) - x_0\| \leq \int_0^t \|\dot{x}(s)\| ds \leq \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}$$

Geometric conditions

(Polyak)-Lojasiewicz (PL) inequality [Lojasiewicz '63, Polyak '63]:

$$\exists \mu > 0,\ \forall x \in \mathbb{R}^d, \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Where does $x(t)$ go? Under **PL**,

$$\forall t \geq 0, \quad \|x(t) - x_0\| \leq \int_0^t \|\dot{x}(s)\| ds \leq \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}$$

The trajectory is trapped in a ball!

Semilocal convergence

It is sufficient to have PL on the corresponding ball $\to$ **Semilocal PL**

Theorem [Oymak et al. '19, Kachaiev et al. '26]: Suppose that for some $\mu > 0$:

$$\forall x \in \overline{\mathcal{B}}\left(x_0, \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}\right), \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Semilocal convergence

It is sufficient to have PL on the corresponding ball $\rightarrow$ **Semilocal PL**

Theorem [Oymak et al. '19, Kachaiev et al. '26]: Suppose that for some $\mu > 0$:

$$\forall x \in \overline{\mathcal{B}}\left(x_0, \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}\right), \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Then, the solution $x(t)$ of GF starting from x_0 is such that

- it stays in the ball,

Semilocal convergence

It is sufficient to have PL on the corresponding ball $\rightarrow$ **Semilocal PL**

Theorem [Oymak et al. '19, Kachaiev et al. '26]: Suppose that for some $\mu > 0$:

$$\forall x \in \overline{\mathcal{B}}\left(x_0, \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}\right), \quad 2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2$$

Then, the solution $x(t)$ of GF starting from x_0 is such that

- it stays in the ball,
- it converges to a global minimizer x^* (!) and:

$$f(x(t)) - f^* \leq \exp(-2\mu t)(f(x_0) - f^*)$$

Semilocal convergence

Why *Semilocal*? PL is required to hold on the ball

$$\overline{\mathcal{B}}\left(x_0, \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}\right)$$

Semilocal convergence

Why *Semilocal*? PL is required to hold on the ball

$$\overline{\mathcal{B}}\left(x_0, \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}\right)$$

Classical conditions in optimization:

- **Global** $\rightarrow$ holds everywhere
- **Local** $\rightarrow$ holds around minimizers

Semilocal convergence

Why *Semilocal*? PL is required to hold on the ball

$$\overline{\mathcal{B}}\left(x_0, \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}\right)$$

Classical conditions in optimization:

- **Global** $\rightarrow$ holds everywhere
- **Local** $\rightarrow$ holds around minimizers

Here, **PL** holds around the initialization!

Semilocal convergence

How to enforce this assumption? Take

$$f : x \mapsto \frac{1}{2}\|H(x) - y^*\|^2,\ H : \mathbb{R}^D \to \mathbb{R}^d,\ y^* \in \mathbb{R}^d$$

Semilocal convergence

How to enforce this assumption? Take

$$f : x \mapsto \frac{1}{2}\|H(x) - y^*\|^2,\ H : \mathbb{R}^D \to \mathbb{R}^d,\ y^* \in \mathbb{R}^d$$

Theorem [Kachaiev et al. ’26, Chatterjee ’22]: If for some $r > 0$, there exists $\sigma_r > 0$ such that

$$\forall x \in \overline{\mathcal{B}}(x_0, r), \quad \sigma_{\min}(J_H(x)) \geq \sigma_r \quad \text{and} \quad y^* \in \overline{\mathcal{B}}(H(x_0), r\sigma_r).$$

Semilocal convergence

How to enforce this assumption? Take

$$f : x \mapsto \frac{1}{2}\|H(x) - y^*\|^2,\ H : \mathbb{R}^D \to \mathbb{R}^d,\ y^* \in \mathbb{R}^d$$

Theorem [Kachaiev et al. '26, Chatterjee '22]: If for some $r > 0$, there exists $\sigma_r > 0$ such that

$$\forall x \in \overline{\mathcal{B}}(x_0, r), \quad \sigma_{\min}(J_H(x)) \geq \sigma_r \quad \text{and} \quad y^* \in \overline{\mathcal{B}}(H(x_0), r\sigma_r).$$

Then,

- f satisfies semilocal **PL** for $\mu = \sigma_r^2$.
- GF converges to x^* s.t. $H(x^*) = y^*$.

A simple example

A simple neural network: Let $x = (a, \mathbf{w})$ with $a \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$.

$$f : x \mapsto \frac{1}{2}\|H(x) - y^*\|^2, \; H : x \mapsto a\mathbf{w}, \; y^* \in \mathbb{R}^d$$

A simple example

A simple neural network: Let $x = (a, \mathbf{w})$ with $a \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$.

$$f : x \mapsto \frac{1}{2}\|H(x) - y^*\|^2, \; H : x \mapsto a\mathbf{w}, \; y^* \in \mathbb{R}^d$$

When does GF converge?

A simple example

A simple neural network: Let $x = (a, \mathbf{w})$ with $a \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$.

$$f : x \mapsto \frac{1}{2}\|H(x) - y^*\|^2, \; H : x \mapsto a\mathbf{w}, \; y^* \in \mathbb{R}^d$$

When does GF converge?

Let $\mathbf{w}_0 = 0_d$. Find $a_0 \in \mathbb{R}$, $r > 0$ and $\sigma_r > 0$ such that:

$$\forall x \in \overline{\mathcal{B}}(x_0, r), \; \sigma_{\min}(J_H(x)) \geq \sigma_r$$

$$y^* \in \overline{\mathcal{B}}(H(x_0), r\sigma_r)$$

A simple example

A simple neural network: Let $x = (a, \mathbf{w})$ with $a \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$.

$$f : x \mapsto \frac{1}{2}\|H(x) - y^*\|^2, \; H : x \mapsto a\mathbf{w}, \; y^* \in \mathbb{R}^d$$

When does GF converge?

Let $\mathbf{w}_0 = 0_d$. Find $a_0 \in \mathbb{R}$, $r > 0$ and $\sigma_r > 0$ such that:

$$\forall (a, \mathbf{w}) \in \overline{\mathcal{B}}((a_0, \mathbf{w}_0), r), \; \sqrt{a^2 + \|\mathbf{w}\|^2} \geq \sigma_r$$

$$y^* \in \overline{\mathcal{B}}(\underbrace{H(a_0, \mathbf{w}_0)}_{=0_d}, r\sigma_r)$$

A simple example

A simple neural network: Let $x = (a, \mathbf{w})$ with $a \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$.

$$f : x \mapsto \frac{1}{2}\|H(x) - y^*\|^2, \; H : x \mapsto a\mathbf{w}, \; y^* \in \mathbb{R}^d$$

When does GF converge?

Let $\mathbf{w}_0 = 0_d$. Fix $r = \frac{|a_0|}{2}$ and $\sigma_r = |a_0| - r = \frac{|a_0|}{2}$. For $|a_0|$ sufficiently large:

$$\forall (a, \mathbf{w}) \in \overline{\mathcal{B}}((a_0, \mathbf{w}_0), r), \; \sqrt{a^2 + \|\mathbf{w}\|^2} \geq |a| \geq |a_0| - r = \sigma_r > 0$$

$$r\sigma_r = \frac{a_0^2}{4} \geq \|y^*\|$$

Linear convergence!

Discussion

Is non convex optimization solved?

Discussion

Is non convex optimization solved?

No! This approach is **restrictive** in various ways:

- **Optimization:** artificially increasing μ (or σ_r) $\implies$ increasing the Lipschitz constant of ∇f!

$\rightarrow$ Critical issue for **discrete algorithms**!

Discussion

Is non convex optimization solved?

No! This approach is **restrictive** in various ways:

- **Optimization:** artificially increasing μ (or σ_r) $\implies$ increasing the Lipschitz constant of ∇f!

  $\rightarrow$ Critical issue for **discrete algorithms**!

- **Learning:** rescaling f through initialization $\implies$ entering **lazy training regime** [Chizat et al. '19]

  $\rightarrow$ Behaves as a **linearized model**!

Conclusion

Takeaways:

- Non convex optimization on overparameterized models can be fine!
- Everything happens at initialization

Limitations:

- Impractical for discrete algorithms
- Everything happens at initialization!

Open questions:

$\rightarrow$ Are there practical cases where Semilocal **PL** holds?

$\rightarrow$ Could inertial techniques help in this context? [Buskulic et al. '25]

Thank you for your attention!

Questions?

Related paper:

Kachaiev, O., Labarrière, H., Molinari, C., Villa, S., *On the Semilocal Convergence of Overparameterized Models*, in preparation.

My Website:

https://hippolytelbrrr.github.io/pages/index_eng.html

References

- Oymak, Soltanolkotabi, *Overparameterized nonlinear learning: Gradient descent takes the shortest path?*, ICML, 2019.
- Chizat, Oyallon, Bach, *On lazy training in differentiable programming*, NEURIPS, 2019.
- Liu, Zhu, Belkin, *Loss landscapes and optimization in over-parameterized non-linear systems and neural networks*, Applied and Computational Harmonic Analysis, 2022.
- Chatterjee, *Convergence of gradient descent for deep neural networks*, arXiv preprint arXiv:2203.16462, 2022.
- Buskulic, Fadili, Quéau, *Convergence and recovery guarantees of unsupervised neural networks for inverse problems*, Journal of Mathematical Imaging and Vision, 2024.
- $x$ojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les équations aux dérivées partielles, 1963.
- Polyak, *Gradient methods for the minimisation of functionals*, USSR Computational Mathematics and Mathematical Physics, 1963.
- Buskulic, Fadili, Quéau, *Implicit regularization of the deep inverse prior trained with inertia*, arXiv preprint arXiv:2506.02986, 2025.

Appendix

Where does $x(t)$ go?

$$\|x(t) - x_0\| \leq \int_0^t \|\dot{x}(s)\| ds = \int_0^t \|\nabla f(x(s))\| ds \leq \int_0^t \frac{\|\nabla f(x(s))\|^2}{\sqrt{2\mu(f(x(s)) - f^*)}} ds$$

Appendix

Where does $x(t)$ go?

$$\|x(t) - x_0\| \leq \int_0^t \|\dot{x}(s)\| ds = \int_0^t \|\nabla f(x(s))\| ds \leq \int_0^t \frac{\|\nabla f(x(s))\|^2}{\sqrt{2\mu(f(x(s)) - f^*)}} ds$$

Since $\frac{d}{dt}\sqrt{f(x(t)) - f^*} = \frac{1}{2\sqrt{f(x(t))-f^*}} \underbrace{\frac{d}{dt}(f(x(t)) - f^*)}_{=-\|\nabla f(x(t))\|^2}$,

$$\|x(t) - x_0\| \leq \sqrt{\frac{2}{\mu}} \left[\sqrt{f(x_0) - f^*} - \sqrt{f(x(t)) - f^*}\right] \leq \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}$$

Appendix

Where does $x(t)$ go?

$$\|x(t) - x_0\| \leq \int_0^t \|\dot{x}(s)\| ds = \int_0^t \|\nabla f(x(s))\| ds \leq \int_0^t \frac{\|\nabla f(x(s))\|^2}{\sqrt{2\mu(f(x(s)) - f^*)}} ds$$

Since $\frac{d}{dt}\sqrt{f(x(t)) - f^*} = \frac{1}{2\sqrt{f(x(t))-f^*}} \underbrace{\frac{d}{dt}(f(x(t)) - f^*)}_{=-\|\nabla f(x(t))\|^2}$,

$$\|x(t) - x_0\| \leq \sqrt{\frac{2}{\mu}} \left[\sqrt{f(x_0) - f^*} - \sqrt{f(x(t)) - f^*}\right] \leq \sqrt{\frac{2\left(f(x_0) - f^*\right)}{\mu}}$$

The trajectory always stays in a ball around x_0!

Appendix

Lazy regime [Chizat, Oyallon, Bach, '19]: Let $f(x) := \ell(H(x))$ and $\alpha > 0$,

$$f_\alpha(x) := \frac{1}{\alpha^2}\,\ell(\alpha H(x)), \quad \dot{x}_\alpha(t) = -\nabla f_\alpha(x_\alpha(t))$$

In lazy training regime:

$$\sup_{t\geq 0}\|x_\alpha(t) - x_0\| = O(1/\alpha) \quad \text{or, equivalently,} \quad x_\alpha(t) \in \overline{B\Big(x_0; O(1/\alpha)\Big)},$$

and

$$\sup_{t\geq 0}\|x_\alpha(t) - \bar{x}_\alpha(t)\| = O\left(\frac{\log\alpha}{\alpha^2}\right)$$

where $\bar{x}_\alpha$ follows GF on the linearization of f at x_0.