Optimization for Machine Learning

Lecture 1: Basics of optimization

Clément W. Royer

M2 IASD - 2025/2026

September 18, 2025

Outline

Outline

Formulation of an (unconstrained) optimization problem

$$\operatorname*{minimize}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$$

Formulation of an (unconstrained) optimization problem

$$\operatorname*{minimize}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$$

- $\boldsymbol{w}$ represents the optimization variable(s);
- d is the dimension of the problem (we will assume $d \geq 1$);
- $f(\cdot)$ is the objective/cost/loss function.

Formulation of an (unconstrained) optimization problem

$$\operatorname*{minimize}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$$

- $\boldsymbol{w}$ represents the optimization variable(s);
- d is the dimension of the problem (we will assume $d \geq 1$);
- $f(\cdot)$ is the objective/cost/loss function.

Maximizing f is equivalent to minimizing $-f$!

Local and global solutions

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w})$$

- $\text{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$: Set of solutions (can be empty).
- $\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$: Optimal value (can be infinite).

Local and global solutions

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w})$$

- $\text{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$: Set of solutions (can be empty).
- $\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$: Optimal value (can be infinite).

Global and local minima

- $\boldsymbol{w}^*$ is a solution or a **global minimum** of f if $f(\boldsymbol{w}^*) \le f(\boldsymbol{w})\ \forall \boldsymbol{w} \in \mathbb{R}^d$.
- $\boldsymbol{w}^*$ is a **local minimum** of f if $f(\boldsymbol{w}^*) \le f(\boldsymbol{w})\ \forall \boldsymbol{w}, \|\boldsymbol{w} - \boldsymbol{w}^*\|_2 \le \epsilon$ for some $\epsilon > 0$.

Local and global solutions

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w})$$

- $\text{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$: Set of solutions (can be empty).
- $\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$: Optimal value (can be infinite).

Global and local minima

- $\boldsymbol{w}^*$ is a solution or a **global minimum** of f if $f(\boldsymbol{w}^*) \le f(\boldsymbol{w}) \ \forall \boldsymbol{w} \in \mathbb{R}^d$.
- $\boldsymbol{w}^*$ is a **local minimum** of f if $f(\boldsymbol{w}^*) \le f(\boldsymbol{w}) \ \forall \boldsymbol{w}, \|\boldsymbol{w} - \boldsymbol{w}^*\|_2 \le \epsilon$ for some $\epsilon > 0$.

- Finding global/local minima is hard in general!
- Regularity of f is needed.

First notion of regularity: Smoothness

Class of $\mathcal{C}^1$ functions

$f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable/$\mathcal{C}^1$ if

- For any $\boldsymbol{w} \in \mathbb{R}^d$, the **gradient** $\nabla f(\boldsymbol{w})$ exists.
- $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is continuous.

First notion of regularity: Smoothness

Class of $\mathcal{C}^1$ functions

$f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable/$\mathcal{C}^1$ if

- For any $\boldsymbol{w} \in \mathbb{R}^d$, the **gradient** $\nabla f(\boldsymbol{w})$ exists.
- $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is continuous.

Class of $\mathcal{C}^{1,1}_L$ functions $(L > 0)$

f is $\mathcal{C}^{1,1}_L$ if it is $\mathcal{C}^1$ and ∇f is L-Lipschitz continuous, i.e.

$$\forall (\boldsymbol{v}, \boldsymbol{w}) \in (\mathbb{R}^d)^2, \qquad \|\nabla f(\boldsymbol{v}) - \nabla f(\boldsymbol{w})\| \le L \|\boldsymbol{v} - \boldsymbol{w}\|.$$

Ex) Linear regression, logistic regression, etc.

Important for today

Function $f(\boldsymbol{w}) \in \mathbb{R}$	Gradient $\nabla f(\boldsymbol{w}) \in \mathbb{R}^d$
$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{w} + \boldsymbol{b}$	$\boldsymbol{a}$
$\frac{1}{2}\|\boldsymbol{w} + \boldsymbol{b}\|_2^2$	$\boldsymbol{w} + \boldsymbol{b}$

Important for today

Function $f(\boldsymbol{w}) \in \mathbb{R}$	Gradient $\nabla f(\boldsymbol{w}) \in \mathbb{R}^d$
$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{w} + \boldsymbol{b}$	$\boldsymbol{a}$
$\frac{1}{2}\|\boldsymbol{w} + \boldsymbol{b}\|_2^2$	$\boldsymbol{w} + \boldsymbol{b}$

Next week How to compute derivatives in ML (bring laptops!).

Problem: minimize$_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}), \;\; f\mathcal{C}^1$.

First-order necessary condition

If $\boldsymbol{w}^*$ is a local minimum of the problem, **then**

$$\|\nabla f(\boldsymbol{w}^*)\|_2 \;=\; 0.$$

- This condition is only necessary;
- A point such that $\|\nabla f(\boldsymbol{w}^*)\|_2 = 0$ can also be a local maximum or a saddle point.

Smoothness and optimality conditions

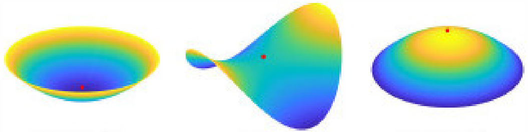
Problem: minimize$_{\boldsymbol{w}\in\mathbb{R}^d}$ $f(\boldsymbol{w})$, $f\mathcal{C}^1$.

First-order necessary condition

If $\boldsymbol{w}^*$ is a local minimum of the problem, **then**

$$\|\nabla f(\boldsymbol{w}^*)\|_2 \;=\; 0.$$

- This condition is only necessary;
- A point such that $\|\nabla f(\boldsymbol{w}^*)\|_2 = 0$ can also be a local maximum or a saddle point.

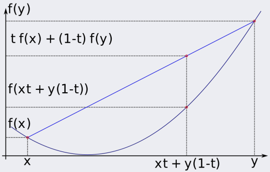


Picture from (Wright and Ma '22).

Generic definition (+Wikicommons picture)

A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is **convex** if

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2,\ \forall t \in [0,1],$$
$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \leq t\, f(\boldsymbol{u}) + (1-t)\, f(\boldsymbol{v}).$$

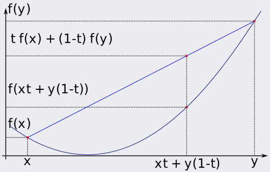


Another notion of regularity: Convexity

Generic definition (+Wikicommons picture)

A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2, \ \forall t \in [0,1],$$
$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \le t\, f(\boldsymbol{u}) + (1-t)\, f(\boldsymbol{v}).$$



Examples in ML

- Linear function $\boldsymbol{w} \mapsto \boldsymbol{a}^{\mathrm{T}}\boldsymbol{w} + b$.
- ℓ_2 loss $\|\boldsymbol{w}\|_2^2 = \sum_{j=1}^{d} w_j^2$.
- Logistic loss.

Showing convexity (from Hardt and Recht '21)

Showing convexity with more than two variables is hard.

Showing convexity (from Hardt and Recht '21)

Showing convexity with more than two variables is hard.

Basic blocks

- All norms (and $\|\boldsymbol{w}\|_2^2$) are convex.
- All linear functions $\boldsymbol{w} \mapsto \boldsymbol{A}\boldsymbol{w} + \boldsymbol{b}$ are convex.
- f convex $\Rightarrow \alpha\, f$ convex $\forall\ \alpha \ge 0$.
- f, g convex $\Rightarrow f + g$ convex.
- f, g convex $\Rightarrow \max(f, g)$ convex.
- f convex $\Rightarrow \boldsymbol{w} \mapsto f(\boldsymbol{A}\boldsymbol{w} + \boldsymbol{b})$ convex.

Convexity and gradient

A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{u}).$$

Smooth convex functions

Convexity and gradient

A continuously differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if and only if

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{u}).$$

A key inequality in optimization.

Convex optimization problem

$$\operatorname*{minimize}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}), \; f \text{ convex.}$$

Convex optimization problem

$$\operatorname*{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}),\ f \text{ convex.}$$

Theorem

Every local minimum of f is a global minimum.

Convex optimization problem

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w}),\ f \text{ convex.}$$

Theorem

Every local minimum of f is a global minimum.

Corollary

If f is $\mathcal{C}^1$,

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{argmin}}\, f(\boldsymbol{w}) = \left\{\ \bar{\boldsymbol{w}}\ \middle|\ \|\nabla f(\bar{\boldsymbol{w}})\|_2 = 0\ \right\}.$$

Any point with a zero gradient is a global minimum!

Strong convexity

Definition

A function $f \ : \mathbb{R}^d \rightarrow \mathbb{R}$ in $\mathcal{C}^1$ is μ-strongly convex (or *strongly convex of modulus* $\mu > 0$) if for all $(\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2$ and $t \in [0, 1]$,

$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \ \leq \ t\, f(\boldsymbol{u}) + (1-t) f(\boldsymbol{v}) - \frac{\mu}{2} t(1-t) \|\boldsymbol{v} - \boldsymbol{u}\|_2^2.$$

Strong convexity

Definition

A function $f \ : \mathbb{R}^d \rightarrow \mathbb{R}$ in $\mathcal{C}^1$ is μ-strongly convex (or *strongly convex of modulus* $\mu > 0$) if for all $(\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2$ and $t \in [0,1]$,

$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \ \leq \ t\, f(\boldsymbol{u}) + (1-t) f(\boldsymbol{v}) - \frac{\mu}{2} t(1-t) \|\boldsymbol{v} - \boldsymbol{u}\|_2^2.$$

Theorem

Any strongly convex function in $\mathcal{C}^1$ has a unique global minimizer.

Gradient and strong convexity

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $f \in \mathcal{C}^1$. Then,

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{u}) + \frac{\mu}{2} \|\boldsymbol{v} - \boldsymbol{u}\|_2^2.$$

Key rules

- For any $\mu > 0$ and $\boldsymbol{w}_0 \in \mathbb{R}^d$, $\boldsymbol{w} \mapsto \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|_2^2$ is μ-strongly convex.

Examples of strongly convex functions

Key rules

- For any $\mu > 0$ and $\boldsymbol{w}_0 \in \mathbb{R}^d$, $\boldsymbol{w} \mapsto \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|_2^2$ is μ-strongly convex.
- If f is μ-strongly convex and g is convex, $f + g$ is μ-strongly convex.

References

- M. Hardt and B. Recht, *Patterns, Predictions and Actions*, Princeton University Press, 2021.
- J. Wright and Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models*, Cambridge University Press, 2022.
- S. J. Wright and B. Recht, *Optimization for Data Analysis*, Cambridge University Press, 2022.

Outline

Exercise 1.a - Convexity

Show that the SVM objective

$$\boldsymbol{w} \in \mathbb{R}^d \;\longmapsto\; \frac{1}{n}\sum_{i=1}^{n} \max\left\{1 - y_i\, \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w}, 0\right\} + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

is a convex function for any $\lambda \geq 0$.

Exercise 1.b - Strong convexity

Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^1$ and μ-strongly convex, and denote by $\boldsymbol{w}^*$ the minimum of f.

1. For any $\boldsymbol{w} \in \mathbb{R}^d$, show that the function

$$\varphi_{\boldsymbol{w}} : \boldsymbol{z} \longmapsto f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{z} - \boldsymbol{w}) + \frac{\mu}{2}\|\boldsymbol{z} - \boldsymbol{w}\|^2$$

   is strongly convex.

2. Compute $\min_{\boldsymbol{z}} \varphi_{\boldsymbol{w}}(\boldsymbol{z})$ and $\operatorname{argmin}_{\boldsymbol{z}} \varphi_{\boldsymbol{w}}(\boldsymbol{z})$.
3. Show that

$$\|\nabla f(\boldsymbol{w})\|_2^2 \quad \geq \quad 2\mu \left(f(\boldsymbol{w}) - f(\boldsymbol{w}^*)\right).$$

Let $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\|_2 \neq 0$ and $\boldsymbol{y} \in \mathbb{R}^d$.

1. Consider the problem
$$\text{minimize}_{w \in \mathbb{R}} \frac{1}{2}\|w\boldsymbol{x} - \boldsymbol{y}\|^2.$$
Is it convex? Is the minimum value 0?
2. Consider now the problem
$$\boldsymbol{W} \in \mathbb{R}^{d \times d} \longmapsto \frac{1}{2}\|\boldsymbol{W}\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$
Is this a convex problem?
3. Justify that
$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times d}} \frac{1}{2}\|\boldsymbol{W}\boldsymbol{x} - \boldsymbol{y}\|_2^2 = 0,$$
and find a global minimum. Is the minimum unique?

Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^{1,1}_L$ and convex. Suppose that $\boldsymbol{w}^* \in \operatorname{argmin}_{\boldsymbol{w}} f(\boldsymbol{w})$ and let $f^* = f(\boldsymbol{w}^*)$.

1. Let $\boldsymbol{w} \in \mathbb{R}^d$. Show that $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \geq \frac{1}{2L}\|\nabla f(\boldsymbol{w})\|_2^2$.
2. Let $(\boldsymbol{w}, \boldsymbol{v}) \in (\mathbb{R}^d)^2$. Show that

$$(\nabla f(\boldsymbol{v}) - \nabla f(\boldsymbol{w}))^{\mathrm{T}} (\boldsymbol{v} - \boldsymbol{w}) \geq \frac{1}{L} \|\nabla f(\boldsymbol{v}) - \nabla f(\boldsymbol{w})\|_2^2 .$$

Consider $\boldsymbol{z} \mapsto f(\boldsymbol{z}) - \nabla f(\boldsymbol{v})^{\mathrm{T}} \boldsymbol{z}$ and $\boldsymbol{z} \mapsto f(\boldsymbol{z}) - \nabla f(\boldsymbol{w})^{\mathrm{T}} \boldsymbol{z}$.

Outline

1. Optimization theory
2. Exercises
3. Bonus

Nonconvex problems

- Convex problems: All local minima are global!
- Nonconvex problems: May have local, non-global (aka spurious minima).

Landscape analysis

Identify classes of **nonconvex problems** for which there are no spurious minima (and possibly more).

Illustration: Linear neural networks

$$\underset{\boldsymbol{W}_1,\dots,\boldsymbol{W}_L}{\text{minimize}} \frac{1}{2} \|\boldsymbol{W}_L \boldsymbol{W}_{L-1} \cdots \boldsymbol{W}_2 \boldsymbol{W}_1 \boldsymbol{X} - \boldsymbol{Y}\|_F^2$$

- $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1} \times d_i}$.
- $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_0}$, $\boldsymbol{Y} \in \mathbb{R}^{d_{L+1} \times d_0}$.
- $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{ij}^2$.

Illustration: Linear neural networks

$$\underset{\boldsymbol{W}_1,\dots,\boldsymbol{W}_L}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}_L\boldsymbol{W}_{L-1}\cdots\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}-\boldsymbol{Y}\|_F^2$$

- $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1}\times d_i}$.
- $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_0}$, $\boldsymbol{Y} \in \mathbb{R}^{d_{L+1}\times d_0}$.
- $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{ij}^2$.

- Also called deep matrix factorization.
- Initially used to better understand neural networks.
- Numerous landscape results, especially between 2016-2022.

Landscape of deep linear networks

Case $L = 1$ (One-layer)

$$\underset{\boldsymbol{W}_1}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y}\|_F^2$$

- Convex problem!
- Explicit form of a solution (often costly to compute).

Landscape of deep linear networks

Case $L = 1$ (One-layer)

$$\underset{\boldsymbol{W}_1}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y}\|_F^2$$

- Convex problem!
- Explicit form of a solution (often costly to compute).

Case $L = 2$ (two-layer network)

$$\underset{\substack{\boldsymbol{W}_1 \in \mathbb{R}^{d_2 \times d_1} \\ \boldsymbol{W}_2 \in \mathbb{R}^{d_3 \times d_2}}}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{Y}\|_F^2$$

- If $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ full rank, there are no spurious minima.
- If $d_2 \geq \max\{d_1, d_3\}$, the optimal value is 0!

Beyond two layers

Bad example for $L = 3$

$$\underset{\boldsymbol{W}_1 \in \mathbb{R}^{1\times 2}, \boldsymbol{W}_2 \in \mathbb{R}, \boldsymbol{W}_3 \in \mathbb{R}^{2\times 1}}{\text{minimize}} \frac{1}{2} \left\| \boldsymbol{W}_3 \boldsymbol{W}_2 \boldsymbol{W}_1 - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\|_F^2$$

Beyond two layers

Bad example for $L = 3$

$$\underset{\boldsymbol{W}_1 \in \mathbb{R}^{1\times 2}, \boldsymbol{W}_2 \in \mathbb{R}, \boldsymbol{W}_3 \in \mathbb{R}^{2\times 1}}{\text{minimize}} \frac{1}{2} \left\| \boldsymbol{W}_3 \boldsymbol{W}_2 \boldsymbol{W}_1 - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\|_F^2$$

→ The point $\left(\begin{bmatrix} 1 & 0 \end{bmatrix}, 0, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right)$ is a local, non-global minimum!

→ Due to intermediate dimensions.

Beyond two layers

Bad example for $L = 3$

$$\operatorname*{minimize}_{\boldsymbol{W}_1 \in \mathbb{R}^{1\times 2}, \boldsymbol{W}_2 \in \mathbb{R}, \boldsymbol{W}_3 \in \mathbb{R}^{2\times 1}} \frac{1}{2} \left\| \boldsymbol{W}_3 \boldsymbol{W}_2 \boldsymbol{W}_1 - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\|_F^2$$

→ The point $\left(\begin{bmatrix} 1 & 0 \end{bmatrix}, 0, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right)$ is a local, non-global minimum!

→ Due to intermediate dimensions.

A positive result (informal)

$$\operatorname*{minimize}_{\boldsymbol{W}_1 \in \mathbb{R}^{d\times 2}, \boldsymbol{W}_2 \in \mathbb{R}^{d\times d}, \boldsymbol{W}_3 \in \mathbb{R}^{2\times d}} \frac{1}{2} \left\| \boldsymbol{W}_3 \boldsymbol{W}_2 \boldsymbol{W}_1 - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\|_F^2$$

If $d \geq 2$ (overparameterized regime), no spurious minima!

More general study

$$\underset{\boldsymbol{W}_1,\dots,\boldsymbol{W}_L}{\text{minimize}} \frac{1}{2} \|\boldsymbol{W}_L \boldsymbol{W}_{L-1} \cdots \boldsymbol{W}_2 \boldsymbol{W}_1 \boldsymbol{X} - \boldsymbol{Y}\|_F^2$$

- $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1} \times d_i}$.
- $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_0}$, $\boldsymbol{Y} \in \mathbb{R}^{d_{L+1} \times d_0}$.
- $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{ij}^2$.

More general study

$$\underset{\boldsymbol{W}_1,\dots,\boldsymbol{W}_L}{\text{minimize}} \frac{1}{2} \|\boldsymbol{W}_L \boldsymbol{W}_{L-1} \cdots \boldsymbol{W}_2 \boldsymbol{W}_1 \boldsymbol{X} - \boldsymbol{Y}\|_F^2$$

- $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1} \times d_i}$.
- $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_0}$, $\boldsymbol{Y} \in \mathbb{R}^{d_{L+1} \times d_0}$.
- $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{ij}^2$.

- Full characterization of the landscape possible (Achour et al '22).
- IF all dimensions are equal and $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ full rank, no spurious local minima!

Summary

Basic block in optimization

- Derivatives (more on that next week).
- Convexity and strong convexity.

Both help characterize solutions of a problem!

Towards the nonconvex case

- Challenge: Presence of spurious minima.
- Overparameterization helps (often the case in ML)!
- Still a lot to be understood (optional course, internships?).

Summary

Basic block in optimization

- Derivatives (more on that next week).
- Convexity and strong convexity.

Both help characterize solutions of a problem!

Towards the nonconvex case

- Challenge: Presence of spurious minima.
- Overparameterization helps (often the case in ML)!
- Still a lot to be understood (optional course, internships?).

Summary

Basic block in optimization

- Derivatives (more on that next week).
- Convexity and strong convexity.

Both help characterize solutions of a problem!

Towards the nonconvex case

- Challenge: Presence of spurious minima.
- Overparameterization helps (often the case in ML)!
- Still a lot to be understood (optional course, internships?).

For now

- Material available online by tomorrow (with corrections if needed).
- Questions are always welcome.