A regularization tour of optimization (for ML)

Lorenzo Rosasco
MaLGa, Università degli Studi di Genova, MIT, IIT

Joint with Silvia Villa, Cesare Molinari (...)

Robustness and Resilience in Stochastic Optimization and Statistical Learning: Mathematical Foundations Erice, May 20th 2022

Outline

Optimization in ML

Learning from data

A least squares interlude

Where we are at

Optimization for machine learning

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} F_i(\theta)$$

Optimization for machine learning

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} F_i(\theta)$$

For example

$$F_i(\theta) = \ell(f(x_i, \theta), y_i) + \frac{1}{n} R(\theta)$$

Typical questions

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} F_i(\theta)$$

F's: smooth, (strongly) convex, composite (e.g. $F = E + R$)?

First order methods: accelerated, stochastic, coordinate-wise, distributed $(\dots)$?

But where do the F_i's come from?

Outline

Optimization in ML

Learning from data

A least squares interlude

Where we are at

Learning from data

Given $\quad (x_i, y_i)_{i=1}^n \quad$ find $\quad f : X \to Y \quad \text{s.t.} \quad f(x) \sim y$

Function models

$$f(x) = f(x, \theta)$$

Function models

$$f(x) = f(x, \theta)$$

Linear parameterization

$$f(x, \theta) = \langle \theta, \Phi(x) \rangle$$

Function models

$$f(x) = f(x, \theta)$$

Linear parameterization

$$f(x, \theta) = \langle \theta, \Phi(x) \rangle$$

Non linear parameterization

$$f(x, \theta) = \langle w, \sigma(Wx) \rangle, \qquad \theta = (w, W)$$

Data model

$$y_i = f(x_i, \theta_*) + \delta_i$$

Data model

$$y_i = f(x_i, \theta_*) + \delta_i$$

- $\delta = \sqrt{\sum_i \delta_i^2}$ is the noise level

Data model

$$y_i = f(x_i, \theta_*) + \delta_i$$

- $\delta = \sqrt{\sum_i \delta_i^2}$ is the noise level
- x_i are deterministic distinct but arbitrarily close

Data model

$$y_i = f(x_i, \theta_*) + \delta_i$$

- $\delta = \sqrt{\sum_i \delta_i^2}$ is the noise level
- x_i are deterministic distinct but arbitrarily close
- $\exists R : \mathbb{R}^d \to \mathbb{R}$ s.t.

$$R(\theta_*) \leqslant r_*$$

Uh-Oh

$$R(\theta_*) \leqslant r_*$$

Neither θ_* nor r_* are known!

Enter regularization

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), y_i) + \lambda R(\theta), \qquad \lambda > 0$$

Rationale

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), y_i) + \lambda R(\theta), \qquad \lambda > 0$$

Rationale

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), y_i) + \lambda R(\theta), \qquad \lambda > 0$$

(stability)

$$\approx$$

$$\theta_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), f(x_i, \theta_*)) + \lambda R(\theta), \qquad \lambda > 0$$

Rationale

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), y_i) + \lambda R(\theta), \qquad \lambda > 0$$

(stability)

$$\approx$$

$$\theta_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), f(x_i, \theta_*)) + \lambda R(\theta), \qquad \lambda > 0$$

(approximation)

$$\downarrow \quad \lambda \to 0$$

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \qquad \text{s.t.} \qquad f(x_i, \theta) = f(x_i, \theta_*)$$

Stability and approximation

Back to optimization

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} F_i(\theta)$$

For example

$$F_i(\theta) = \ell(f(x_i, \theta), y_i) + \frac{\lambda}{n} R(\theta)$$

Recap

The problem

Given $(x_i, y_i)_{i=1}^n$ estimate $\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \quad$ s.t. $\quad f(x_i, \theta) = f(x_i, \theta_*)$

UniGe | MaLGa

Recap

The problem

$$\text{Given} \quad (x_i, y_i)_{i=1}^{n} \qquad \text{estimate} \qquad \theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \quad \text{s.t.} \quad f(x_i, \theta) = f(x_i, \theta_*)$$

Classic approach

UniGe | MaLGa

Recap

The problem

Given $(x_i, y_i)_{i=1}^n$ estimate $\theta_*^\dagger = \operatorname{argmin}_{\theta \in \mathbb{R}^d} R(\theta),$ s.t. $f(x_i, \theta) = f(x_i, \theta_*)$

Classic approach

1. (somebody) designs and studies the F_i's

Recap

The problem

$$\text{Given} \quad (x_i, y_i)_{i=1}^{n} \qquad \text{estimate} \qquad \theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \quad \text{s.t.} \quad f(x_i, \theta) = f(x_i, \theta_*)$$

Classic approach

1. (somebody) designs and studies the F_i's

2. (somebody else) computes a solution

$$\widehat{\theta}_{t+1} = \widehat{\theta}_{t+1} - \gamma_t \nabla F_{i_t}(\widehat{\theta}_t)$$

There are (at least) two *caveats*....

(1) Tuning

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), y_i) + \lambda R(\theta), \qquad \lambda > 0$$

(1) Tuning

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), y_i) + \lambda R(\theta), \qquad \lambda > 0$$

- The *regularization* parameter λ is not known and needs be fixed ...

(1) Tuning

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell(f(x_i, \theta), y_i) + \lambda R(\theta), \qquad \lambda > 0$$

- The *regularization* parameter λ is not known and needs be fixed ...
- ...this requires solving multiple optimization problems!

UniGe | MaLGa

(2) Stability with no regularization

Empirically solutions are often *stable* also when $\lambda = 0$!

Outline

Optimization in ML

Learning from data

A least squares interlude

Where we are at

Interlude

$$f(x,\theta) = \langle \theta, x \rangle \qquad \ell(a,y) = (a-y)^2 \qquad R(\theta) = \|\theta\|^2$$

UniGe | MaLGa

Explicit regularization

$$\widehat{\theta}_\lambda = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (\langle x_i, \theta \rangle - y_i)^2 + \lambda \|\theta\|^2, \qquad \lambda > 0$$

UniGe | MaLGa

Explicit regularization

$$\widehat{\theta}_\lambda = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (\langle x_i, \theta \rangle - y_i)^2 + \lambda \|\theta\|^2, \qquad \lambda > 0$$

$$\approx \frac{\delta}{\lambda}$$

$$\theta_\lambda = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (\langle x_i, \theta \rangle - \langle x_i, \theta_* \rangle)^2 + \lambda \|\theta\|^2, \qquad \lambda > 0$$

Explicit regularization

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (\langle x_i, \theta \rangle - y_i)^2 + \lambda \|\theta\|^2, \qquad \lambda > 0$$

$$\approx \frac{\delta}{\lambda}$$

$$\theta_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (\langle x_i, \theta \rangle - \langle x_i, \theta_* \rangle)^2 + \lambda \|\theta\|^2, \qquad \lambda > 0$$

$$\approx \lambda \|\theta_*\|$$

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \qquad \text{s.t.} \qquad \langle x_i, \theta \rangle = \langle x_i, \theta_* \rangle$$

Implicit (!?) regularization

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \gamma \nabla \sum_{i=1}^{n} \left(\left\langle x_i, \widehat{\theta}_t \right\rangle - y_i \right)^2$$

Implicit (!?) regularization

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \gamma \nabla \sum_{i=1}^{n} \left(\left\langle x_i, \widehat{\theta}_t \right\rangle - y_i \right)^2$$

$$\approx t\delta$$

$$\theta_{t+1} = \theta_t - \gamma \nabla \sum_{i=1}^{n} \left(\langle x_i, \theta_t \rangle - \langle x_i, \theta_* \rangle \right)^2$$

Implicit (!?) regularization

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \gamma \nabla \sum_{i=1}^{n} \left(\left\langle x_i, \widehat{\theta}_t \right\rangle - y_i \right)^2$$

$$\approx t\delta$$

$$\theta_{t+1} = \theta_t - \gamma \nabla \sum_{i=1}^{n} (\langle x_i, \theta_t \rangle - \langle x_i, \theta_* \rangle)^2$$

$$\approx \frac{\|\theta_*\|}{t}$$

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \qquad \text{s.t.} \qquad \langle x_i, \theta_t \rangle = \langle x_i, \theta_* \rangle$$

Convergence & stability

The family of solutions $(\widehat{\theta}_t)_t$ behaves much like $(\widehat{\theta}_\lambda)_\lambda$ with $t \sim 1/\lambda$

Back to the caveats

$$\widehat{\theta}_t \quad \underset{t\delta}{\approx} \quad \theta_t \quad \underset{\frac{\|\theta_*^\dagger\|}{t}}{\longrightarrow} \theta_*^\dagger$$

- # *iterations* plays the role of the regularization parameter
- the iterates are *biased* towards small norms

UniGe | MaLGa

Some context

- Iterative regularization: classic in inverse problems since the 50's
- Implicit regularization: recent trend in machine learning
- Inexact optimization: but the perturbations are non vanishing (!)

Outline

Optimization in ML

Learning from data

A least squares interlude

Where we are at

Some perspectives

1. Beyond GD (acceleration, stochastic gradients...)
2. Beyond Euclidean regularization
3. Beyond least squares
4. Beyond deterministic da models
5. Beyond linear models

(1) Beyond GD: acceleration

$$\widehat{\theta}_t \quad \underset{t^2\delta}{\approx} \quad \theta_t \quad \underset{\frac{\|\theta_*^\dagger\|}{t^2}}{\rightarrow} \quad \theta_*^\dagger$$

- trade-off between convergence and stability
- same accuracy in less iterates

(1) Beyond GD: acceleration

$$\widehat{\theta}_t \quad \underset{t^2\delta}{\approx} \quad \theta_t \quad \underset{\frac{\|\theta_*^\dagger\|}{t^2}}{\rightarrow} \quad \theta_*^\dagger$$

- trade-off between convergence and stability
- same accuracy in less iterates

Compare with Mert's talk + check out results from the 80' s.

[Nemirovski, Polyak '86, Nemirovski '86]

Acceleration illustrated

(2) Beyond Euclidean norms

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \qquad \text{s.t.} \qquad \widehat{X}\theta = \underbrace{\widehat{X}\theta_*}_{\sim \widehat{y}}$$

(2) Beyond Euclidean norms

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \qquad \text{s.t.} \qquad \widehat{X}\theta = \underbrace{\widehat{X}\theta_*}_{\sim \widehat{y}}$$

$R = J + \frac{\alpha}{2}\|\cdot\|^2$ str. convex

(2) Beyond Euclidean norms

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \qquad \text{s.t.} \qquad \widehat{X}\theta = \underbrace{\widehat{X}\theta_*}_{\sim \widehat{y}}$$

$R = J + \frac{\alpha}{2}\|\cdot\|^2$ str. convex

$$\widehat{\theta}_t = \mathrm{Prox}_{\alpha^{-1}J}\left(-\alpha^{-1}\widehat{X}^T\widehat{\Theta}^t\right)$$

$$\widehat{\Theta}_{t+1} = \widehat{\Theta}_t + \gamma\left(\widehat{X}\widehat{\theta}_t - \widehat{y}\right)$$

Dual GD aka MD

[Matet, R., Villa, Vu ’18]

UniGe | MaLGa

(2) Beyond Euclidean norms

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \qquad \text{s.t.} \qquad \widehat{X}\theta = \underbrace{\widehat{X}\theta_*}_{\sim \widehat{y}}$$

$R = J + \frac{\alpha}{2}\|\cdot\|^2$ str. convex

R convex e.g. $R = \|\cdot\|_1$

$$\widehat{\theta}_t = \mathrm{Prox}_{\alpha^{-1}J}\left(-\alpha^{-1}\widehat{X}^T\widehat{\Theta}^t\right)$$
$$\widehat{\Theta}_{t+1} = \widehat{\Theta}_t + \gamma\left(\widehat{X}\widehat{\theta}_t - \widehat{y}\right)$$

Dual GD aka MD

[Matet, R., Villa, Vu ’18]

(2) Beyond Euclidean norms

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \qquad \text{s.t.} \qquad \widehat{X}\theta = \underbrace{\widehat{X}\theta_*}_{\sim \widehat{y}}$$

$R = J + \frac{\alpha}{2}\|\cdot\|^2$ str. convex

$$\widehat{\theta}_t = \mathrm{Prox}_{\alpha^{-1}J}\left(-\alpha^{-1}\widehat{X}^T\widehat{\Theta}^t\right)$$

$$\widehat{\Theta}_{t+1} = \widehat{\Theta}_t + \gamma\left(\widehat{X}\widehat{\theta}_t - \widehat{y}\right)$$

Dual GD aka MD

[Matet, R., Villa, Vu '18]

R convex e.g. $R = \|\cdot\|_1$

$$\widehat{\theta}_{t+1} = \mathrm{Prox}_{\tau R}\left(\widehat{\theta}_t - \tau\widehat{X}^T\left(2\widehat{\Theta}_t - \widehat{\Theta}_{t-1}\right)\right)$$

$$\widehat{\Theta}_{t+1} = \widehat{\Theta}_t + \sigma\left(\widehat{X}\widehat{\theta}_{t+1} - \widehat{y}\right)$$

Prima-Dual Hybrid Gradient

[Massias, Molinari, R., Villa '21]

(2) Beyond Euclidean norms

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R(\theta), \qquad \text{s.t.} \qquad \widehat{X}\theta = \underbrace{\widehat{X}\theta_*}_{\sim \widehat{y}}$$

$\underline{R = J + \frac{\alpha}{2}\|\cdot\|^2 \text{ str. convex}}$

$$\widehat{\theta}_t = \operatorname{Prox}_{\alpha^{-1}J}\left(-\alpha^{-1}\widehat{X}^T\widehat{\Theta}^t\right)$$

$$\widehat{\Theta}_{t+1} = \widehat{\Theta}_t + \gamma\left(\widehat{X}\widehat{\theta}_t - \widehat{y}\right)$$

Dual GD aka MD

[Matet, R., Villa, Vu '18]

$\underline{R \text{ convex}}$ e.g. $R = \|\cdot\|_1$

$$\widehat{\theta}_{t+1} = \operatorname{Prox}_{\tau R}\left(\widehat{\theta}_t - \tau\widehat{X}^T\left(2\widehat{\Theta}_t - \widehat{\Theta}_{t-1}\right)\right)$$

$$\widehat{\Theta}_{t+1} = \widehat{\Theta}_t + \sigma\left(\widehat{X}\widehat{\theta}_{t+1} - \widehat{y}\right)$$

Prima-Dual Hybrid Gradient

[Massias, Molinari, R., Villa '21]

See also [Osher, Burger et al. '05, Guneskar et al. '18, Rebeschini et. '19]

(3) Beyond least squares: classification

$$\widehat{\theta}^{\dagger} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \qquad \text{s.t.} \qquad \langle x_i, \theta \rangle \, y_i \geqslant 1$$

(3) Beyond least squares: classification

$$\widehat{\theta}^{\dagger} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \qquad \text{s.t.} \qquad \langle x_i, \theta \rangle y_i \geqslant 1$$

$$\Leftrightarrow$$

$$\widehat{\theta}^{+} = \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \min_{i=1,\dots,n} \langle x_i, \theta \rangle y_i, \qquad \text{s.t.} \qquad \|\theta\| = 1$$

Min norm $\Leftrightarrow$ max margin

(3) Beyond least squares: classification (cont.)

$$\widehat{\theta}^{\dagger} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \qquad \text{s.t.} \qquad \langle x_i, \theta \rangle \, y_i \geqslant 1$$

(3) Beyond least squares: classification (cont.)

$$\widehat{\theta}^{\dagger} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \qquad \text{s.t.} \qquad \langle x_i, \theta \rangle \, y_i \geqslant 1$$

- for $\ell(a, y) = \log(1 + e^{-y\,a})$ GD converges sub-linearly in direction

$$\frac{\widehat{\theta}_t}{\|\widehat{\theta}_t\|} \to \frac{\widehat{\theta}^{\dagger}}{\|\widehat{\theta}^{\dagger}\|} = \widehat{\theta}^{+}$$

[Soudry et al. ’18]

(3) Beyond least squares: classification (cont.)

$$\widehat{\theta}^{\dagger} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \qquad \text{s.t.} \qquad \langle x_i, \theta \rangle y_i \geqslant 1$$

- for $\ell(a, y) = \log(1 + e^{-ya})$ GD converges sub-linearly in direction

$$\frac{\widehat{\theta}_t}{\|\widehat{\theta}_t\|} \to \frac{\widehat{\theta}^{\dagger}}{\|\widehat{\theta}^{\dagger}\|} = \widehat{\theta}^{+}$$

[Soudry et al. '18]

- for $\ell(a, y) = \max\{1 - ya, 0\}$ a dual diagonal iteration converges linearly

$$\widehat{\theta}_t \to \widehat{\theta}^{\dagger}$$

[Apidopolous, R. Villa ' 22], see also [Molitor, Needell, Ward ' 21]

(4) Beyond deterministic data models

$$(x_i, y)_{i=1}^n \sim P^n$$

UniGe | MaLGa

(4) Beyond deterministic data models

$$(x_i, y)_{i=1}^{n} \sim P^n$$

- $\delta^2 = \mathbb{E}\left[(y - \mathbb{E}[y \mid x])^2\right]$ is the noise level

UniGe | MaLGa

(4) Beyond deterministic data models

$$(x_i, y)_{i=1}^n \sim P^n$$

- $\delta^2 = \mathbb{E}\left[(y - \mathbb{E}[y \mid x])^2\right]$ is the noise level
- $(x_i)_{i=1}^n \sim P_x$ are sample according to the marginal P_x

(4) Beyond deterministic data models

$$(x_i, y)_{i=1}^n \sim P^n$$

- $\delta^2 = \mathbb{E}\left[(y - \mathbb{E}[y \mid x])^2\right]$ is the noise level
- $(x_i)_{i=1}^n \sim P_x$ are sample according to the marginal P_x
- $\langle x, \theta_* \rangle = \mathbb{E}[y \mid x]$ and $\exists\, R : \mathbb{R}^d \to \mathbb{R}$ s.t.

$$R(\theta_*) \leqslant r_*$$

Implicit regularization: learning edition

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \gamma \nabla \frac{1}{n} \sum_{i=1}^{n} \left(\left\langle x_i, \widehat{\theta}_t \right\rangle - y_i \right)^2$$

Implicit regularization: learning edition

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \gamma \nabla \frac{1}{n} \sum_{i=1}^{n} \left(\left\langle x_i, \widehat{\theta}_t \right\rangle - y_i \right)^2$$

$$\approx \frac{t\delta}{n}$$

$$\theta_{t+1} = \theta_t - \gamma \nabla \mathbb{E} \left[(\langle x, \theta_t \rangle - y)^2 \right]$$

Implicit regularization: learning edition

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \gamma \nabla \frac{1}{n} \sum_{i=1}^{n} \left(\left\langle x_i, \widehat{\theta}_t \right\rangle - y_i \right)^2$$

$$\approx \frac{t\delta}{n}$$

$$\theta_{t+1} = \theta_t - \gamma \nabla \mathbb{E} \left[(\langle x, \theta_t \rangle - y)^2 \right]$$

$$\approx \frac{\|\theta_*\|}{t}$$

$$\theta_*^\dagger = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|\theta\|, \qquad \text{s.t.} \qquad \langle x, \theta_t \rangle = \langle x, \theta_* \rangle, \qquad \text{a.s.}$$

UniGe | MaLGa

(5) Beyond linear models

- For classification and $f(x, \theta)$ one-homogenous in θ

  If GD converges, then it converges in direction to the max margin solution

[Nacson et al. '19]

UniGe | MaLGa

(5) Beyond linear models

- For classification and $f(x, \theta)$ one-homogenous in θ

  If GD converges, then it converges in direction to the max margin solution

  [Nacson et al. '19]

- Results can be extended to $f(x, \theta) = \langle \sigma(\cdot, x), \theta \rangle = \int \sigma(\omega, x) d\theta(\omega)$

  [Chizat, Bach '20]

(5) Beyond linear models

- For classification and $f(x, \theta)$ one-homogenous in θ

  If GD converges, then it converges in direction to the max margin solution

  [Nacson et al. '19]

- Results can be extended to $f(x, \theta) = \langle \sigma(\cdot, x), \theta \rangle = \int \sigma(\omega, x) d\theta(\omega)$

  [Chizat, Bach '20]

- $f(x, \theta) = \langle x, \theta \rangle$ with $\theta = \beta^{\odot L} = \underbrace{\beta \odot \cdots \odot \beta}_{\text{L times}}$

$$\textit{GD on } \beta \quad \Leftrightarrow \quad \textit{MD on } \theta$$

[Amid, Warmuth '21, Chou, Maly, Rauhut '22]

Wrapping up

- Iterative regularization: merging modeling and optimization
- Inexact optimization meets regularization
- A new playground (...)

What's next?

- Non linear models
- Beyond ERM
- Zeroth-order optimization

Multiple post-docs/PhD positions @MaLGa!

@lrntzrsc

malga.unige.it

UniGe | MaLGa