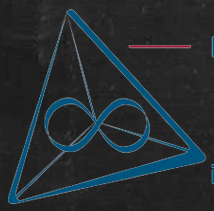
Some aspects in the mathematics of machine learning

Benjamin Gess
Universität Bielefeld & MPI MIS Leipzig

September 2021 .



Max-Planck-Institut für
Mathematik
in den Naturwissenschaften

Universität Bielefeld

Some aspects in math of machine learning:

- Stochastic optimization algorithms
- Information geometry
- Scaling limits and universality

Stochastic optimization algorithms

Supervised learning:

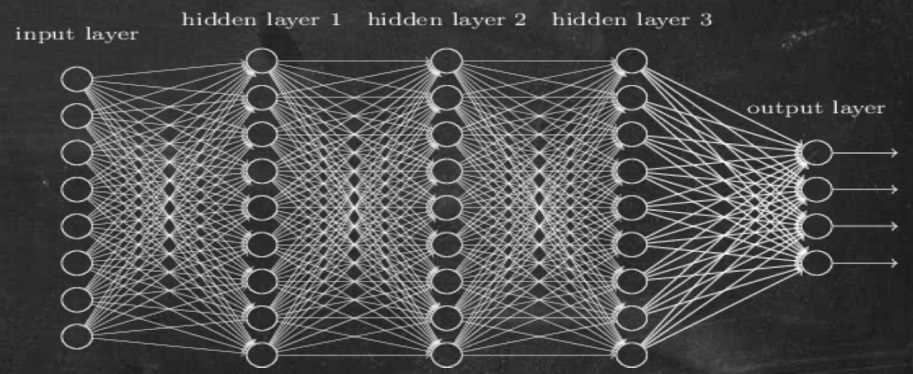
- X is the configuration space (e.g. images)
- $\varphi : X \to \mathbb{R}$ is the function to learn (e.g. classification)
- Training set $S \subseteq X$ with a probability measure μ (e.g. empirical measure)

Parametrizing Ansatz space:

Feed-forward neural network

$$\{\mathrm{input}\} = \{x \in \mathbb{R}^N\} \mapsto R_{\mathbf{1}} = \begin{pmatrix} R(\theta_{\mathbf{11}} \cdot x + b_{\mathbf{11}}) \\ \vdots \\ R(\theta_{n_{\mathbf{1}}\mathbf{1}} \cdot x + b_{n_{\mathbf{1}}\mathbf{1}}) \end{pmatrix} \mapsto \dots$$

$$\mapsto R_m = \begin{pmatrix} R(\theta_{\mathbf{1}m} \cdot R_{m-\mathbf{1}} + b_{\mathbf{1}m}) \\ \vdots \\ R(\theta_{n_m m} \cdot R_{m-\mathbf{1}} + b_{n_m m}) \end{pmatrix} \mapsto \{\mathrm{output}\} \in \mathbb{R}.$$



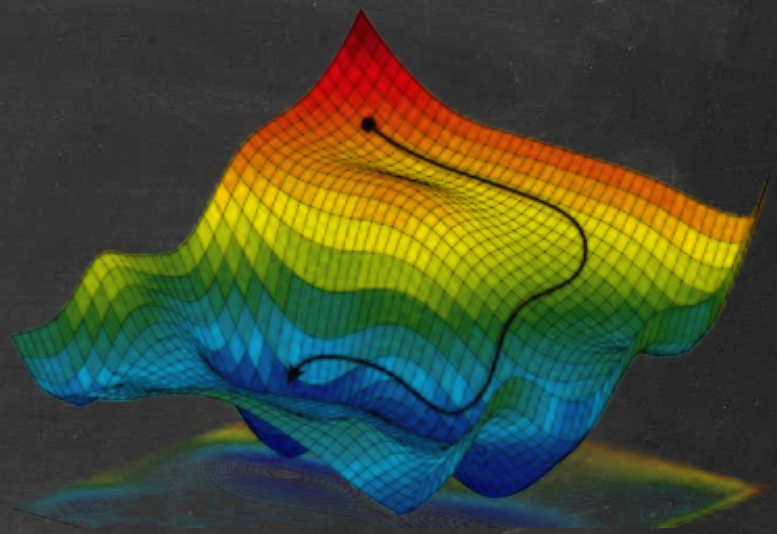
$u_\theta : x \mapsto u_\theta(x)$ an artificial neural network

Empirical loss

$$f(\theta) := \int_S |u_\theta(x) - \varphi(x)|^2 \, \mu(\mathrm{d}x).$$

To minimize, over $\theta \in \mathbb{R}^d$,

$$f(\theta) = \int_S |u_\theta(x) - \varphi(x)|^2 \, \mu(\mathrm{d}x),$$



The difficulties are numerous:

- f is neither convex nor globally attracting
  - "...while SGD has been rigorously analyzed only for convex loss functions...in deep learning the loss is a non-convex function of the network parameters..." Vidal, et al. 2017..
- f and ∇f are computationally expensive
  - Stochastic optimization / stochastic gradient descent

$$\theta_n = \theta_{n-1} - \eta \nabla F(\theta_{n-1}, X_n)$$
$$\mathbb{E}\left[\nabla_\theta F(\theta, X_1)\right] = \nabla f(\theta)$$

- symmetries in the network create degenerate critical points
  - minima are not isolated and may form a manifold

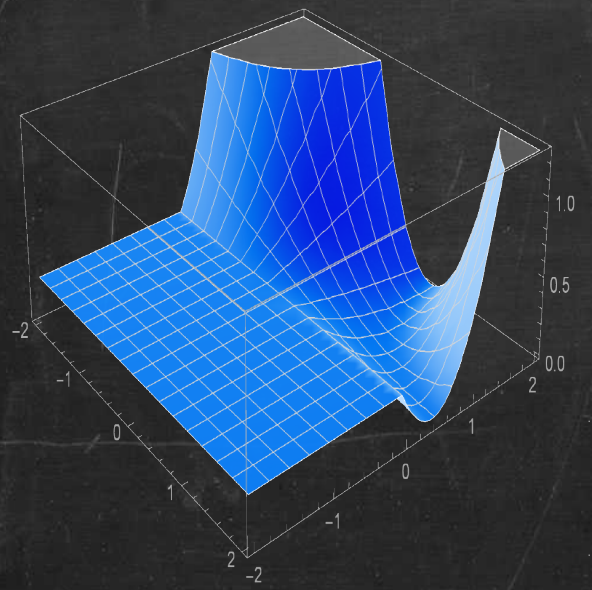
Example

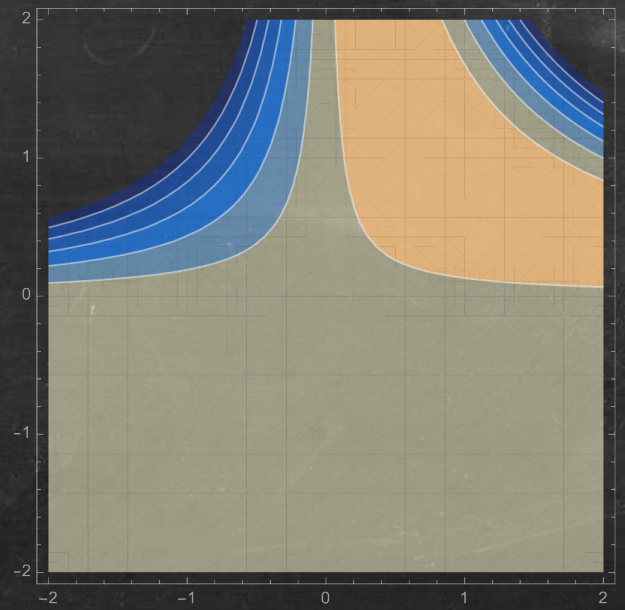
- The rectifier function: $R(x) = \max(0, x)$.
- The neural network, for $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$,

$$u_\theta = \theta_2 R(\theta_1 x).$$

- The objective function: $f(\theta) = \int_{\mathbb{R}} |u_\theta(x) - \sin(x)|^2 \, \mathrm{d}x$.
- The manifold of minima

$$\mathcal{M} = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : \theta_1 \theta_2 = 3\,(\sin(1) - \cos(1))\,\}.$$

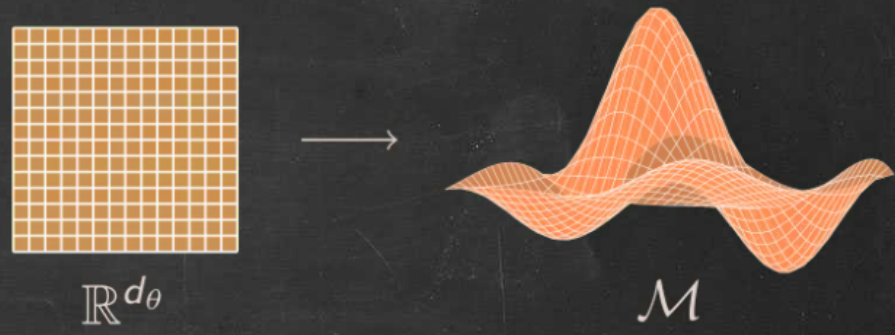




Our take on this: Introduce methods from (stochastic) dynamical systems, e.g. center manifold theorem. Prove quantitative convergence estimates. E.g. [Fehrman, G., Jentzen; JMLR 2020].

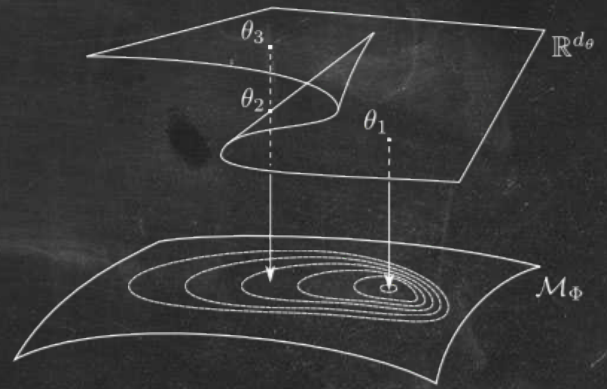
Information geometry

$$\mathbb{R}^{d_\theta} \to C(X;\mathbb{R})$$
$$\theta \mapsto u_\theta(\cdot)$$



is a parametrization of a "manifold"

$$\mathcal{M} = \{u_\theta(\cdot) : \theta \in \mathbb{R}^{d_\theta}\} \subseteq C(X;\mathbb{R})$$



This can introduce "spurious" minima

Deterministic gradient descent

$$\theta_n = \theta_{n-1} - \eta \nabla f(\theta_{n-1})$$

The gradient ∇ is not intrinsic to the problem/Ansatz.

Instead: Use an intrinsic geometry with associated gradient on $\mathcal{M}$: E.g. natural gradient descent

Scaling limits and universality

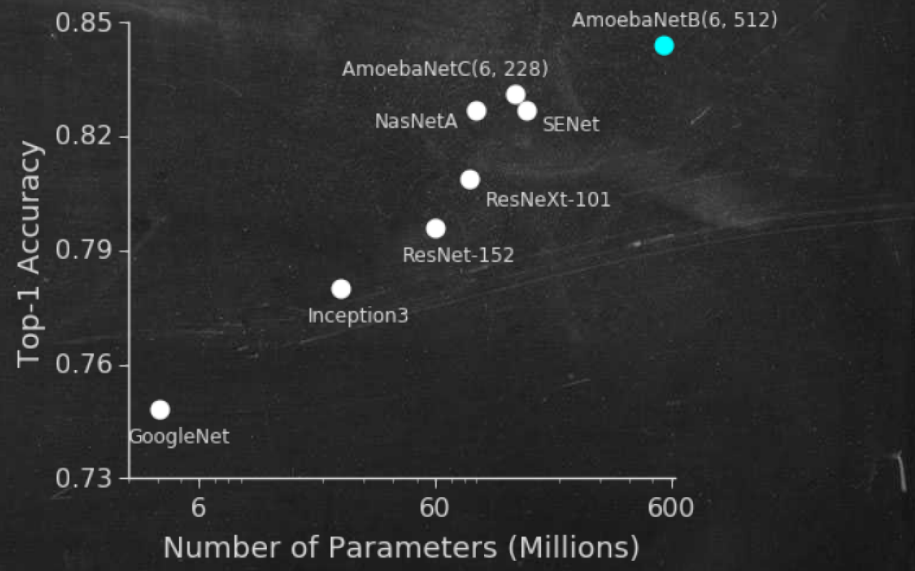
Supervised learning comes with several parameters:

- Number of parameters in the network (depth, width): N
- Size of training data: M
- Learning rate: $\eta = \Delta t$
- Batch size: P

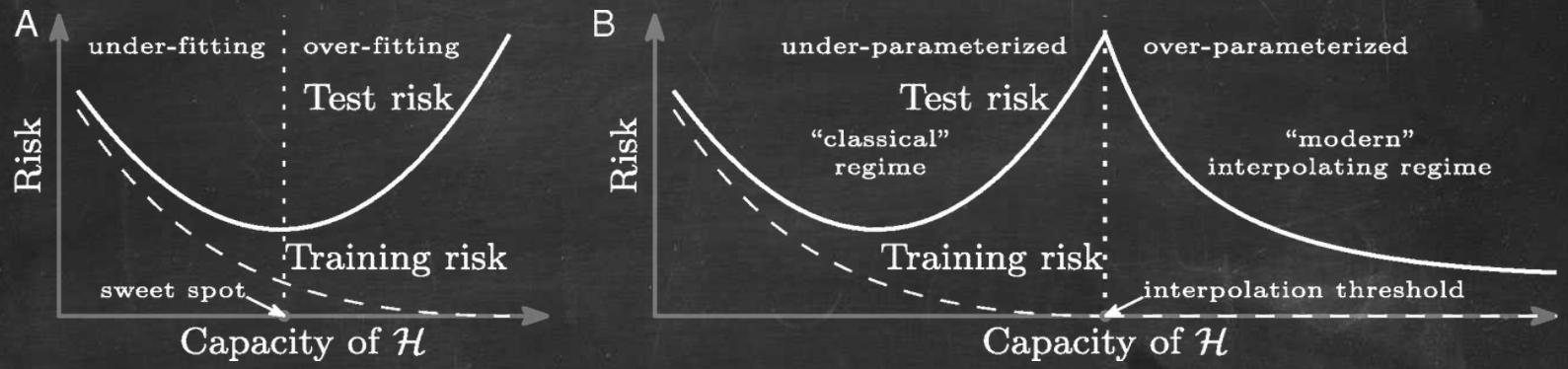
Scaling limits / universal behavior when parameters become large / small?

E.g. Overfitting in machine learning ($N \to \infty$, M, η, P fixed): Dynamics converge to

- Nonlinear PDE (mean-field dynamics, optimal transport) [e.g. Chizat, Bach; 2018]
- Linear dynamics / Neural tangent kernel (lazy training) [e.g. Jacot, Gabriel, Hongler; 2018]



Double descent



Can be explained in certain joint scaling limits ($N, M \to \infty$). [see Mei, Montanari; 2020]

Our take on this: Role of fluctuations in the scaling limits?