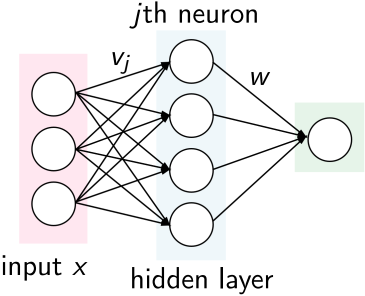
On Convex Neural Networks

Mathematical Foundations of Data Science

Eloïse BERTHIER
(MVA)

January 7, 2019

Convex neural networks



output $y = f_{w,v}(x) = w \cdot \sum_j (v_j^\top x)_+$

$$\min_{w,v} \frac{1}{n} \sum_{i=1}^{n} \ell(f_{w,v}(x_i), y_i) \qquad \min_{f \in \mathcal{F}_1} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

non-convex problem $\longrightarrow$ convex... but infinite-dimensional

Regularized convex problem

Variation norm of f: $\gamma_1(f) = ||w||_1$ if finite number of neurons.
extends to any number of neurons using Radon measures.

$$\min_{\gamma_1(f)<\delta} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Generalization results:

- f decomposed with at most $O(\gamma_1(f)^2/\varepsilon^2)$ neurons with error ε
- $\mathcal{F}_1$ contains non-smooth functions
- convex neural networks are adaptive to structure

  $\rightarrow$ *breaking the curse of dimensionality*

Optimizing with Frank-Wolfe

How to solve a constrained problem in high-dimension?

$$\min_{\gamma_1(f)<\delta} \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(x_i), y_i\big)$$

Frank-Wolfe algorithm: incrementally add neurons.

- convergence rate $O(1/t)$ after t steps
- but each step requires to add a new neuron
- and this is NP-hard!

Optimizing with Gradient Descent

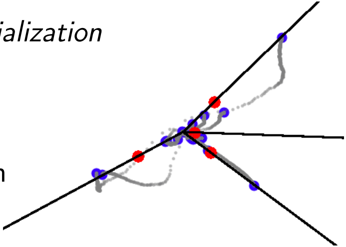
Neural network as a mixture of particles v with weight w

In the space of measures: $\mu = \sum_j w_j \delta_{v_j}$

Result from optimal transport using Wasserstein gradient flow:

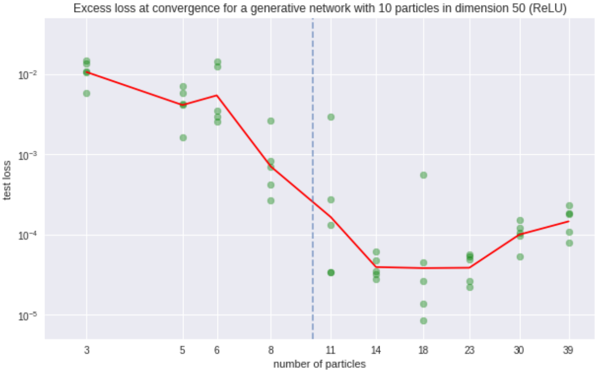
- assumptions on *homogeneity* and *initialization*
- infinite particle limit
- continuous time limit

$\rightarrow$ Convergence to a *global* minimum



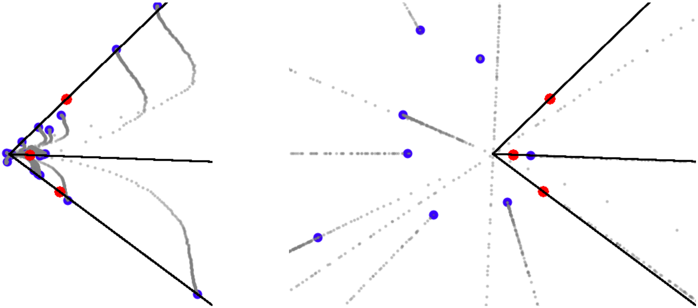
Experiments

How much overparametrization is needed for global convergence?



Experiments

Influence of the initialization?



$r_0 = 0.05$ $r_0 = 10$

Initial norm of each particle r_0

Conclusion & Perspectives

A novel way to understand the convergence of neural networks; linking the original non-convex problem to the space of *measures*, hence to optimal transport.

Studying gradient descent through *gradient flows* is promising.

$\rightarrow$ How far from practical applications is this *ideal* dynamics?

- link between overparametrization and dimension
- iterative algorithm as a continuous time process
- practical rules to initialize neural networks

References

- Bengio, Y., Roux, N. L., Vincent, P., Delalleau, O., and Marcotte, P. (2006). Convex neural networks. In *Advances in neural information processing systems*
- Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. In *Journal of Machine Learning Research*
- Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*