# On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport


**Lénaïc Chizat and Francis Bach**

**INRIA, École Normale Supérieure, PSL, Paris**


## A classical non-convex problem

**Euclidean formulation.** Minimize a loss $R$ on a Hilbert space $\mathcal{F}$ over all possible combinations of features $\phi(\theta,\cdot) \in \mathcal{F}$ with $\theta \in \mathbb{R}^d$:

$$\inf_{\substack{m\in\mathbb{N}\\ \theta_1,\dots,\theta_m\in\mathbb{R}^d}} \underbrace{R\left(\frac{1}{m}\sum_{i=1}^m \phi(\theta_i,\cdot)\right)}_{\text{Loss}} + \underbrace{\frac{1}{m}\sum_{i=1}^m V(\theta_i)}_{\text{Optional regularizer}}$$

- feature function $\theta \mapsto \phi(\theta,\cdot)$ is **differentiable** (e.g. neuron or filter)
- **convex** smooth loss $R : \mathcal{F} \to \mathbb{R}$ (e.g. quadratic or logistic)
- regularizer $V : \mathbb{R}^d \to \mathbb{R}$ possibly **non-smooth** ($\ell_1$, $\ell_2^2$ penalties)
- minimization **also on the number** $m$ **of features**/particles

**Measure formulation.** Rewrites as a **convex problem in the space of probability measures** by setting $\mu = \frac{1}{m}\sum \delta_{\theta_i} \in \mathcal{P}(\mathbb{R}^d)$:

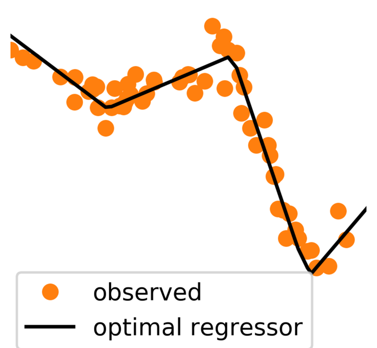
$$\min_{\mu\in\mathcal{P}(\mathbb{R}^d)} F(\mu) := R\left(\int_{\mathbb{R}^d} \phi(\theta,\cdot)d\mu(\theta)\right) + \int_{\mathbb{R}^d} V(\theta)d\mu(\theta) \qquad (1)$$

### Example 1: Neural networks with 2 layers

Input/output random data $(X,Y)$, loss $\ell$ and activation $\sigma$:

$$\min_{m,(a_i,b_i,w_i)_i} \mathbb{E}_{(X,Y)}\left[\ell\left(\frac{1}{m}\sum_{i=1}^m a_i\sigma(w_i\cdot X + b_i), Y\right)\right]$$

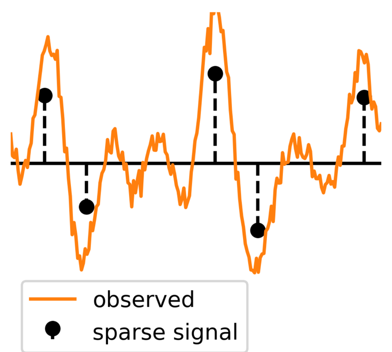
- features $\phi(\theta,x) = a\cdot\sigma(w\cdot x + b)$ with parameters $\theta = (a,b,w)$
- $R$ is the population loss accessed through stochastic gradients
- global minimizer here means best possible generalization among all hidden layer sizes




### Example 2: Sparse inverse problems

Recovering a sparse signal from filtered and noisy observations with BLASSO:

- $\phi(\theta,\cdot)$ are weighted filter impulse responses
- $R$ is the mean square error
- $V$ is a non-smooth sparsity inducing penalty
- our viewpoint corresponds in practice to forward-backward algorithm on the positions and weights of $m$ spikes




## Contributions in a nutshell

**New insight.** For these **non-convex** problems, we prove **a consistency result for gradient based optimization** methods: under assumptions, they converge to **global minimizers** in the **over-parameterization** limit.

**Key assumptions.** Mainly relies on $2$ structural assumptions:

- **homogeneity** of $\phi$ (full or partial)
  - $\to$ leads to selection of the correct magnitude for each feature
- **diversity in the initialization** of parameters
  - $\to$ turns out sufficient to explore all combinations of features

**Approach.** We make a **qualitative analysis** of the optimization path using tools from **optimal transport theory** and **topology**.

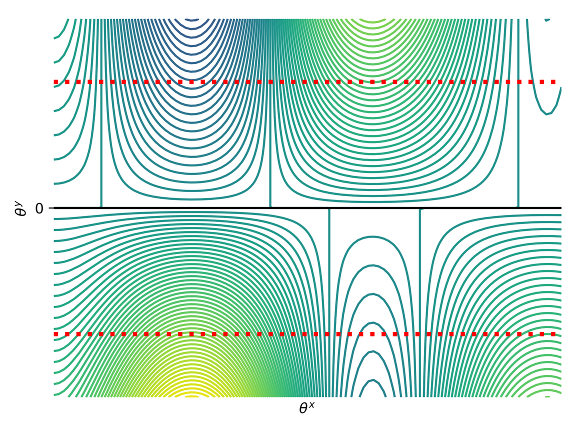
## Global convergence result

**Main theorem (simplified)**

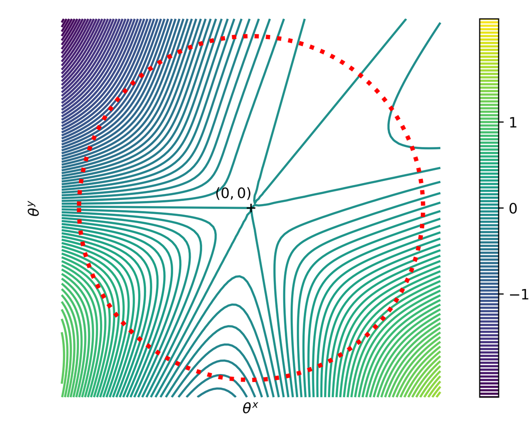
Assume that the initializations $\theta_1(0), \theta_2(0), \dots$ are drawn randomly according to a measure $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ that satisfies a support condition (see below). Then the gradient flow $(\theta_1(t), \theta_2(t), \dots)$ of the objective function satisfies

$$\lim_{m,t\to\infty} F\left(\frac{1}{m}\sum_{i=1}^m \delta_{\theta_i(t)}\right) = \min_{\mu\in\mathcal{P}(\mathbb{R}^d)} F(\mu).$$

- see paper for precise statements with technical assumptions and statements for ReLU/sigmoid neural networks and sparse deconvolution
- diversity at initialization is crucial: this is captured by a support condition on $\mu_0$, that also reflects the homogeneity properties of $\phi$



Partially homogeneous case: $\phi((\theta^x, \lambda\theta^y),\cdot) = \lambda^p\phi((\theta^x,\theta^y),\cdot)$



Fully homogeneous case: $\phi((\lambda\theta^x, \lambda\theta^y),\cdot) = \lambda^p\phi((\theta^x,\theta^y),\cdot)$

**Figure:** Dotted lines show admissible supports on 2d exemples. Also plotted: level lines of the Fréchet derivative $F'$ of $F$ at $\mu_0$ ($\lambda, p > 0$).

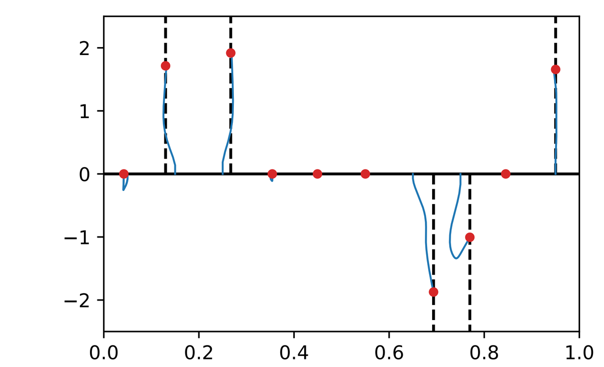
## Many-particle limit of gradient flows

**Gradient flow.** Gradient-based methods use estimates $\tilde g$ of the gradient $g$ of the objective function and a step-size $\eta$. We consider the gradient flow, their **idealized continuous-time** counter-part

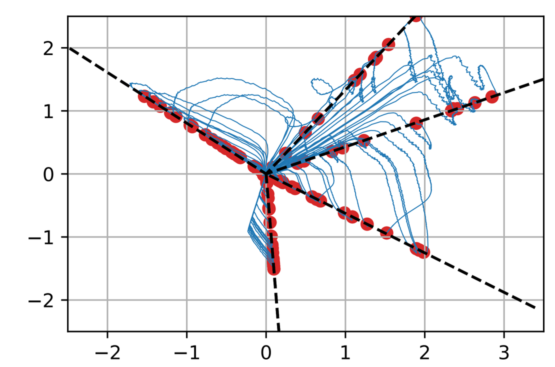
$$\theta_{t+1} = \theta_t - \eta\cdot\tilde g(\theta_t) \quad \underset{\eta\to 0}{\Rightarrow} \quad \theta'(t) \propto -g(\theta(t)).$$

**Many-particle limit.** When $m \to \infty$, the gradient flow is described by a time-dependent density $\mu_t \in \mathcal{P}(\mathbb{R}^d)$ obeying a partial differential equation: the **optimal transport/Wasserstein gradient flow** of the objective function $F(\mu)$ in Equation (**??**):

$$\partial_t\mu_t = -\nabla\cdot(\mu_t\nabla F'(\mu_t)).$$



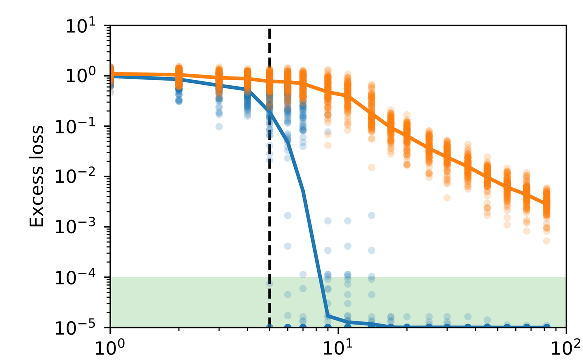
Trajectory of forward-backward algorithm for sparse deconvolution with $m = 10$ (attains optimum)



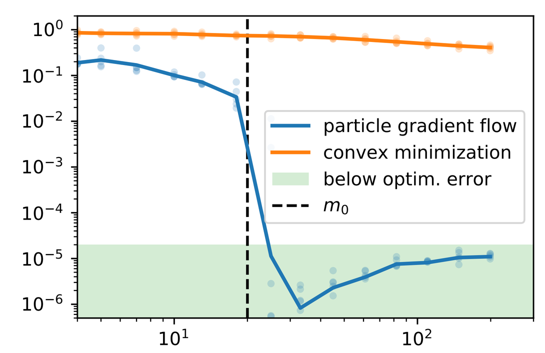
Trajectory of stochastic gradient method on a ReLU neural network with $m = 100$ (attains optimum)

## Experimental results

In practice, a **slight** over-parameterization is **sufficient** for optimality.




Sparse deconvolution ($d = 2$)




ReLU neural network ($d = 100$)

**Figure:** Excess loss at convergence versus number of particles $m$ for the non-convex gradient flows (in blue) and convex minimization on the magnitude only, initialized with random features (in orange). Synthetic problems where simplest minimizer has $m_0$ components.

## Main references


- Ambrosio et al., *Gradient flows: in metric spaces and in the space of probability measures.* 2008.
- Bredies and Pikkarainen. *Inverse problems in spaces of measures.* 2013.
- Bach. *Breaking the curse of dimensionality with convex neural networks.* 2017.
- Nitanda and Suzuki, *Stochastic Particle Gradient Descent for Infinite Ensembles*, 2017.