On the Global Convergence of Gradient Descent for Over-parameterized Models

using Wasserstein gradient flows

Lénaïc Chizat*, joint work with Francis Bach*

Sept. 2018 - Laboratoire de Mathématiques d'Orsay

*INRIA and ENS Paris

What is this talk about?

Minimize a **convex** function over **measures** (domain $\Theta \subset \mathbb{R}^d$):

$$\min_{\mu \in \mathcal{M}(\Theta)} J(\mu)$$

Challenges

- infinite dimensional $\Rightarrow$ parameterization

What is this talk about?

Minimize a **convex** function over **measures** (domain $\Theta \subset \mathbb{R}^d$):

$$\min_{\mu \in \mathcal{M}(\Theta)} J(\mu)$$

Challenges

- infinite dimensional $\Rightarrow$ parameterization
- covers all of continuous optimization $\Rightarrow$ more structure

What is this talk about?

Minimize a **convex** function over **measures** (domain $\Theta \subset \mathbb{R}^d$):

$$\min_{\mu \in \mathcal{M}(\Theta)} J(\mu)$$

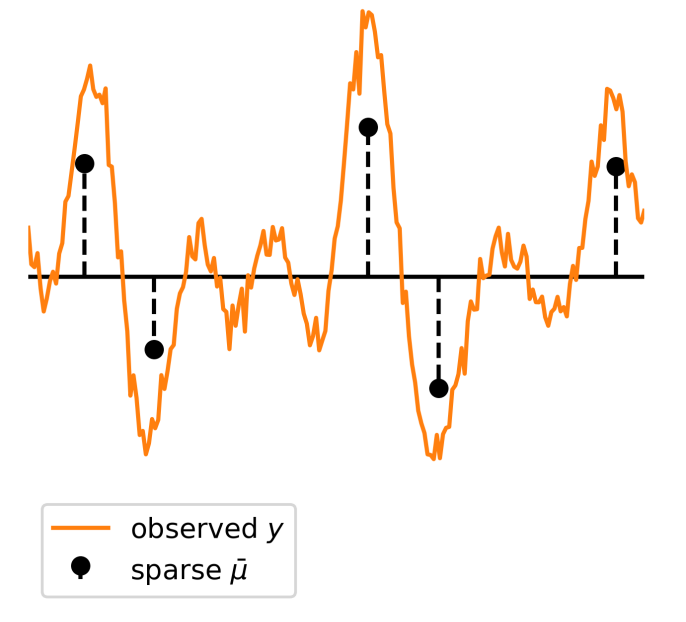
Challenges

- infinite dimensional $\Rightarrow$ parameterization
- covers all of continuous optimization $\Rightarrow$ more structure

Content

Qualitative study of a *practical* method that provably reaches global optimality asymptotically.

Motivating example I: signal processing

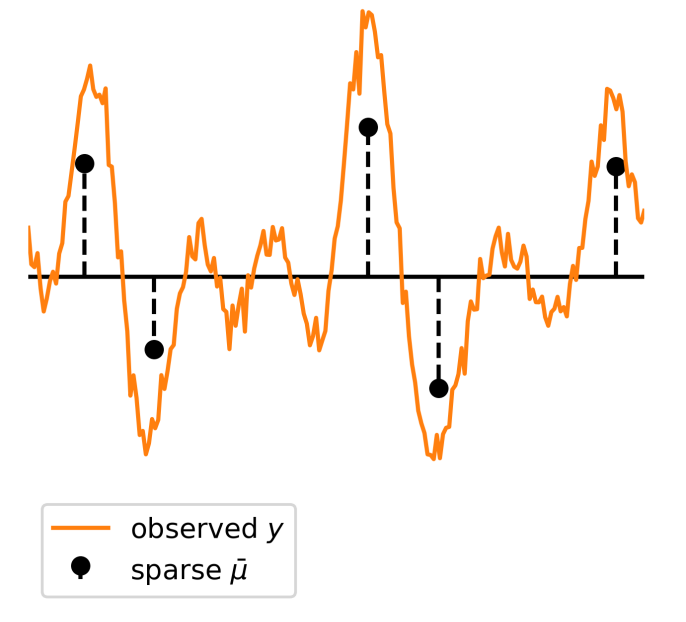


Sparse deconvolution

Recover a sparse signal $\overline{\mu} = \sum_{i=1}^{m} w_i \delta_{\theta_i}$

from a filtered version $y = \varphi * \overline{\mu} + \text{noise}$

Motivating example I: signal processing



Sparse deconvolution

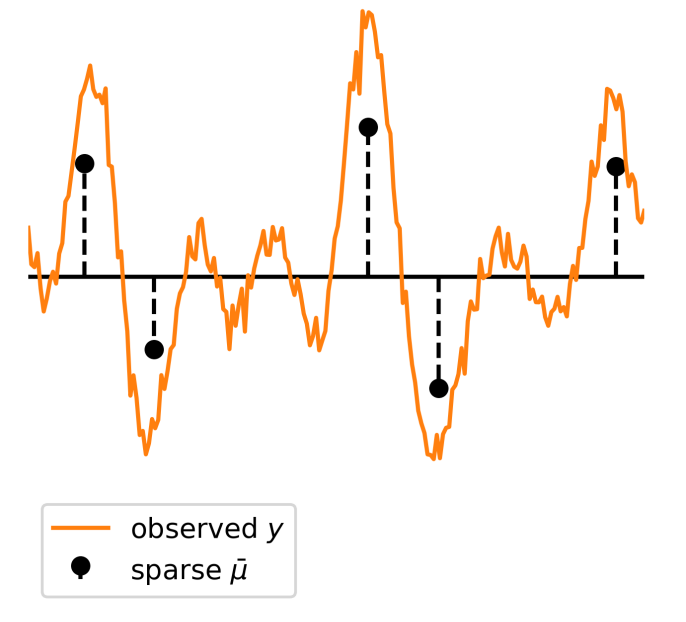
Recover a sparse signal $\overline{\mu} = \sum_{i=1}^{m} w_i \delta_{\theta_i}$

from a filtered version $y = \varphi * \overline{\mu} + \text{noise}$

Variational approach (B-LASSO)

$$\min_{\mu \in \mathcal{M}(\Theta)} \underbrace{\frac{1}{2}\|y - \varphi * \mu\|_{L^2}^2}_{\text{loss}} + \underbrace{\lambda \|\mu\|_{TV}}_{\text{regularization}}$$

Motivating example I: signal processing



Sparse deconvolution

Recover a sparse signal $\overline{\mu} = \sum_{i=1}^{m} w_i \delta_{\theta_i}$
from a filtered version $y = \varphi * \overline{\mu} + \text{noise}$

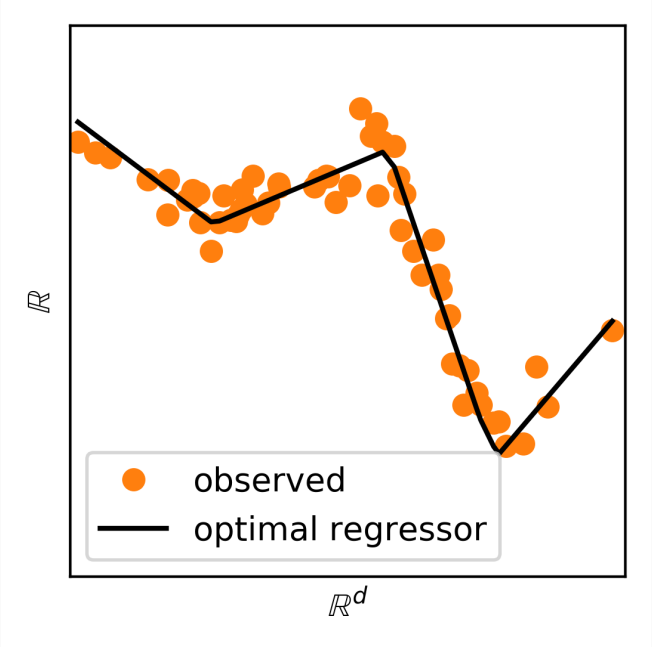
Variational approach (B-LASSO)

$$\min_{\mu \in \mathcal{M}(\Theta)} \underbrace{\frac{1}{2}\|y - \varphi * \mu\|_{L^2}^2}_{\text{loss}} + \underbrace{\lambda \|\mu\|_{TV}}_{\text{regularization}}$$

$\rightarrow$ illustrative example in this talk : signal with 5 spikes on the 1-torus and φ Dirichlet (low pass) filter of order 7.

[Refs]
Azaïs, De Castro & Gamboa (2015). *Spike detection from inaccurate samplings.*
Duval & Peyré (2015). *Exact support recovery for sparse spikes deconvolution.*

Motivating example II: machine learning

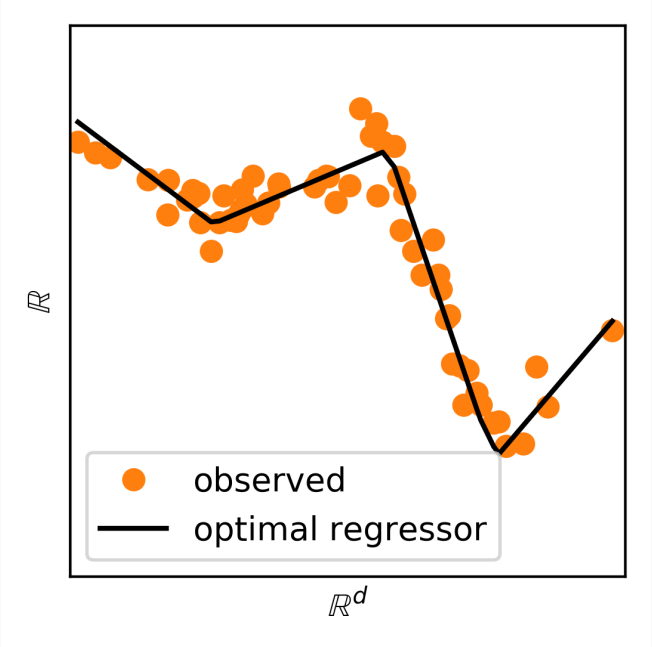


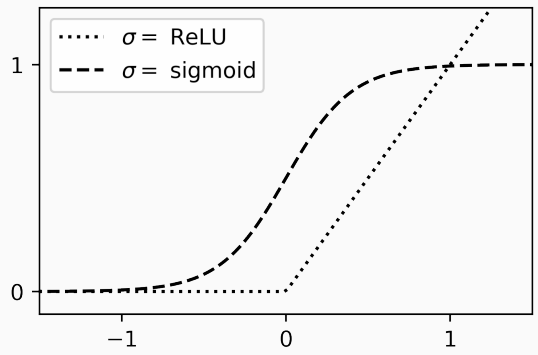
Statistical learning

Let (X, Y) a couple of r.v. on $\mathbb{R}^d \times \mathbb{R}$ and a smooth convex loss $\ell : \mathbb{R}^2 \to \mathbb{R}_+$.
Given n samples $(x_i, y_i)_{i=1}^n$, "solve"

$$\min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}\, \ell(f(X), Y)$$

Motivating example II: machine learning





Statistical learning

Let (X, Y) a couple of r.v. on $\mathbb{R}^d \times \mathbb{R}$ and a smooth convex loss $\ell : \mathbb{R}^2 \to \mathbb{R}_+$.
Given n samples $(x_i, y_i)_{i=1}^n$, "solve"

$$\min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}\, \ell(f(X), Y)$$

Neural network with 1 hidden layer

Write $f_\mu(x) = \int \sigma(\theta \cdot x) \mathrm{d}\mu(\theta)$ and solve

$$\min_{\mu \in \mathcal{M}(\Theta)} \frac{1}{n} \sum_{i=1}^n \underbrace{\ell(f_\mu(x_i), y_i)}_{\text{loss}} + \underbrace{\lambda \|\mu\|_{TV}}_{\text{regularization}}$$

[Refs] Bach (2017). *Breaking the curse of dimensionality with convex neural networks.*

Common structure

- Differentiable dictionary $(\phi(\theta))_{\theta \in \Theta} \subset \mathcal{F}$ in Hilbert space $\mathcal{F}$
- Smooth convex loss $R : \mathcal{F} \to \mathbb{R}_+$

$$J^* = \min_{\mu \in \mathcal{M}(\Theta)} J(\mu), \qquad J(\mu) := \underbrace{R\left(\int \phi \, \mathrm{d}\mu\right)}_{\text{loss}} + \underbrace{\lambda \|\mu\|_{TV}}_{\text{regularization}}$$

Common structure

- Differentiable dictionary $(\phi(\theta))_{\theta\in\Theta} \subset \mathcal{F}$ in Hilbert space $\mathcal{F}$
- Smooth convex loss $R : \mathcal{F} \to \mathbb{R}_+$

$$J^* = \min_{\mu\in\mathcal{M}(\Theta)} J(\mu), \qquad J(\mu) := \underbrace{R\left(\int \phi \,\mathrm{d}\mu\right)}_{\text{loss}} + \underbrace{\lambda\|\mu\|_{TV}}_{\text{regularization}}$$

Examples

- sparse deconvolution: $\phi(\theta) : x \mapsto \varphi(x-\theta)$
- neural networks with 1 hidden layer: $\phi(\theta) : x \mapsto \sigma(\theta \cdot x)$
- inversion of sketches, low rank tensor decomposition, linear system identification, matrix completion...

[Refs] Boyd, Schiebinger, Recht (2017). *The alternating descent conditional gradient method for sparse inverse problems.*

Particle gradient flow

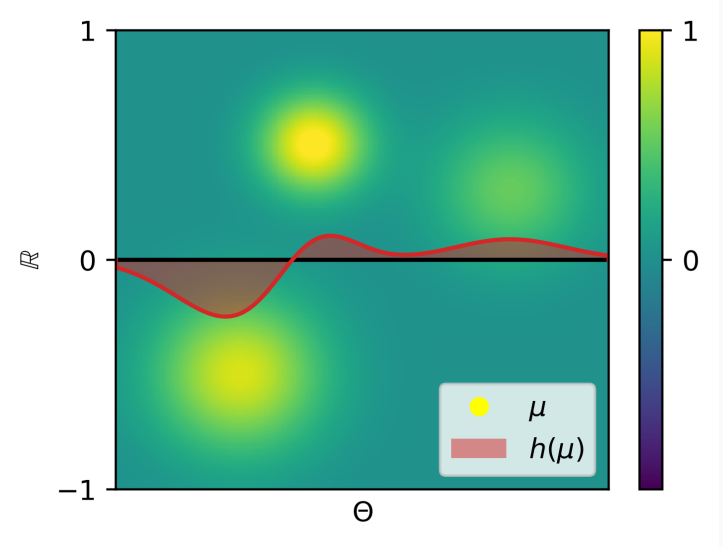
Construction

Change of unknown (lifting)

Let, for $\mu \in \mathcal{P}(\mathbb{R} \times \Theta)$,

$$h(\mu)(\mathrm{d}\theta) = \int_{\mathbb{R}} w \, \mu(\mathrm{d}w, \mathrm{d}\theta)$$

and pose $F(\mu) \approx J(h(\mu))$.



Remark: $\|\nu\|_{TV} = \min_{h(\mu)=\nu} \int |w| \mathrm{d}\mu(w, \theta)$

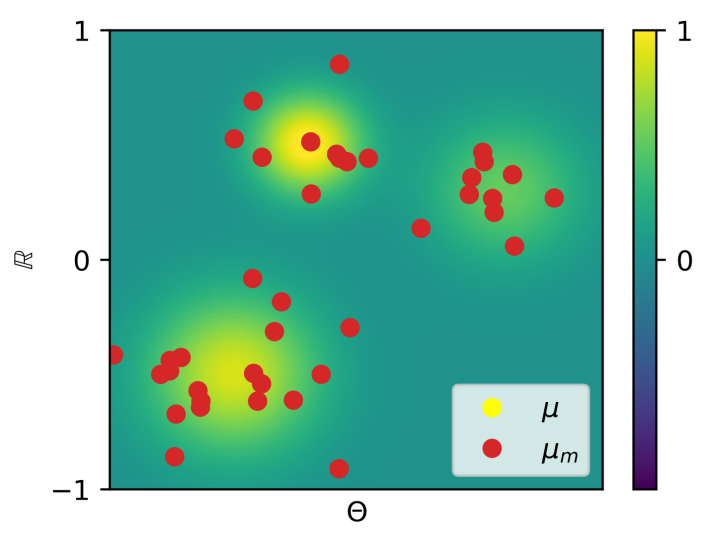
Construction

Change of unknown (lifting)

Let, for $\mu \in \mathcal{P}(\mathbb{R} \times \Theta)$,

$$h(\mu)(\mathrm{d}\theta) = \int_{\mathbb{R}} w \, \mu(\mathrm{d}w, \mathrm{d}\theta)$$

and pose $F(\mu) \approx J(h(\mu))$.



Remark: $\|\nu\|_{TV} = \min_{h(\mu)=\nu} \int |w| \mathrm{d}\mu(w, \theta)$

Definition (Particle gradient flow)

Gradient flow $\mathbf{u}'(t) = -m \nabla F_m(\mathbf{u}(t))$ in $(\mathbb{R} \times \Theta)^m$, where

$$F_m(\mathbf{u}(t)) := F(\mu_{m,t}), \qquad \mu_{m,t} := \frac{1}{m} \sum_{i=1}^{m} \delta_{u_i(t)}$$

Remark: in practice, gradient descent or its variants (stochastic, proximal, fast, etc).

Comparison of methods

	pros	cons
conditional gradient	known rate, sparse	sometimes NP hard
moment methods	asymptotically exact	heavy, not generic
particle gradient flow	practical, cheap	global optimality ?

Contribution

Proof of asymptotic global convergence in the $m, t \to \infty$ limit.

[Refs]

Bach (2017). *Breaking the curse of dimensionality with convex neural networks.*

Lasserre (2010). *Moments, positive polynomials and their applications.*

Catala, Duval, Peyré (2017). *A low-rank approach to off-the-grid sparse deconvolution.*

Main result

Objective: ϕ differentiable and bounded, R smooth and convex:

$$F(\mu) = R(\int w\phi(\theta)\mathrm{d}\mu(w,\theta)) + \lambda \int |w|\mathrm{d}\mu$$

Main result

Objective: ϕ differentiable and bounded, R smooth and convex:

$$F(\mu) = R(\int w\phi(\theta)\mathrm{d}\mu(w,\theta)) + \lambda \int |w|\mathrm{d}\mu$$

Theorem (Global convergence, informal[1])

If the initialization is such that $\mu_{m,0}$ converges to a measure which support separates $\{-\infty\} \times \Theta$ from $\{+\infty\} \times \Theta$, then

$$\lim_{m,t\to\infty} F(\mu_{m,t}) = \min_{\mu\in\mathcal{M}_+(\mathbb{R}\times\Theta)} F(\mu)$$

$$\lim_{m,t\to\infty} J(h(\mu_{m,t})) = \min_{\mu\in\mathcal{M}(\Theta)} J(\mu).$$

[1] see paper for precise statement. We also assume boundary conditions, existence of $\lim_{t\to\infty} h(\mu_{\infty,t})$, and a technical "Sard-type" non-degeneracy.

Main result

Objective: ϕ differentiable and bounded, R smooth and convex:

$$F(\mu) = R(\int w\phi(\theta)\mathrm{d}\mu(w,\theta)) + \lambda \int |w|\mathrm{d}\mu$$

Theorem (Global convergence, informal[1])

If the initialization is such that $\mu_{m,0}$ converges to a measure which support separates $\{-\infty\} \times \Theta$ from $\{+\infty\} \times \Theta$, then

$$\lim_{m,t\to\infty} F(\mu_{m,t}) = \min_{\mu \in \mathcal{M}_+(\mathbb{R}\times\Theta)} F(\mu)$$

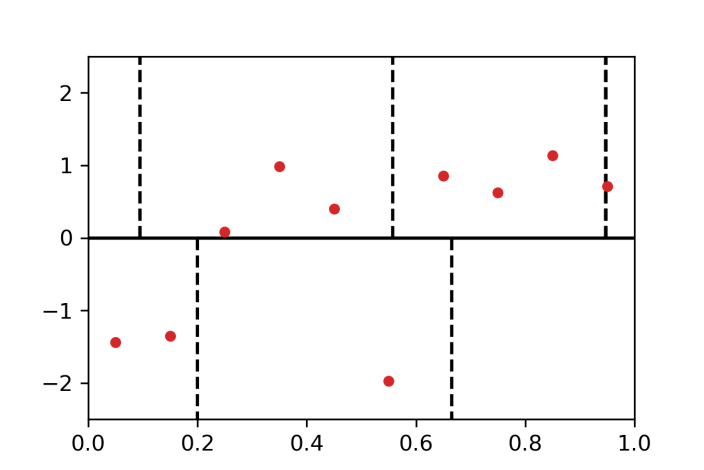
$$\lim_{m,t\to\infty} J(h(\mu_{m,t})) = \min_{\mu \in \mathcal{M}(\Theta)} J(\mu).$$

Remarks

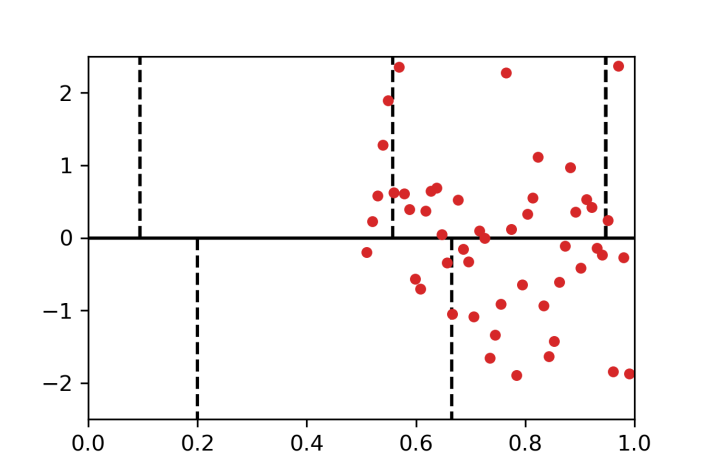
- bad local minima *exist*, but are avoided if good initialization
- also a statement if ϕ homogeneous and unbounded

[1] see paper for precise statement. We also assume boundary conditions, existence of $\lim_{t\to\infty} h(\mu_{\infty,t})$, and a technical "Sard-type" non-degeneracy.

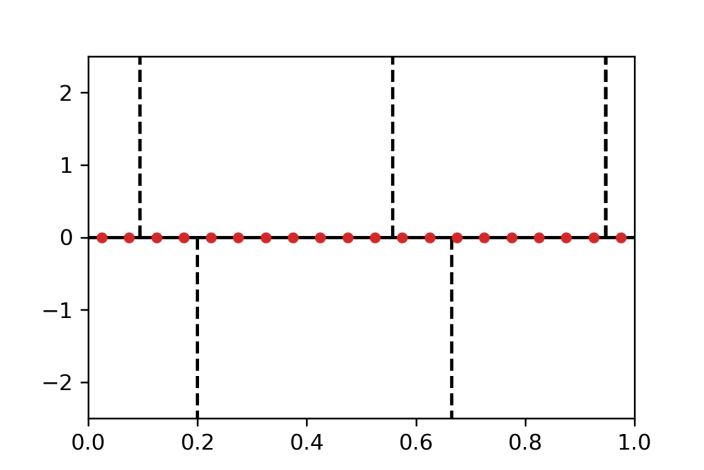
Illustration



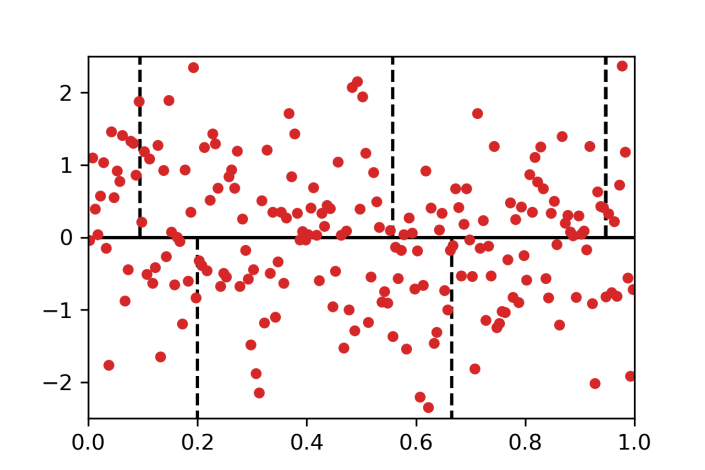
(a) Not enough particles



(b) Breaks separation assumption



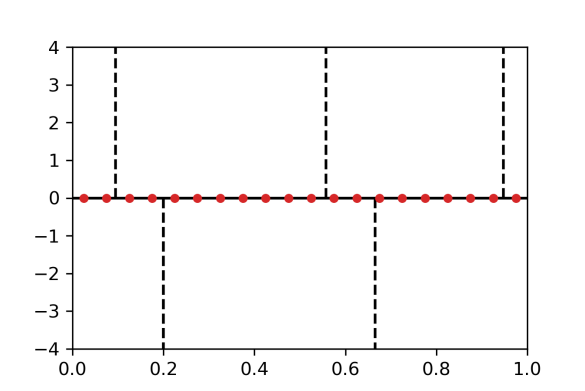
(c) Success



(d) Success

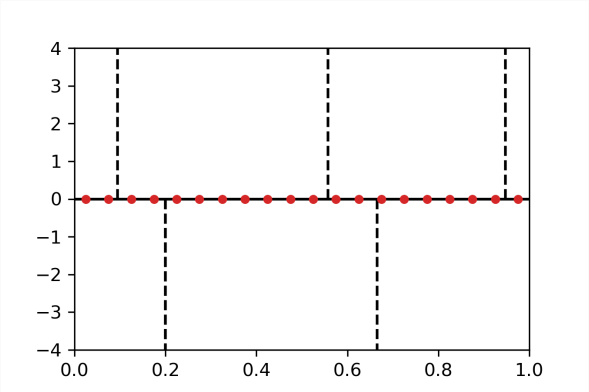
Particle-complexity

What about convex weights minimization ?

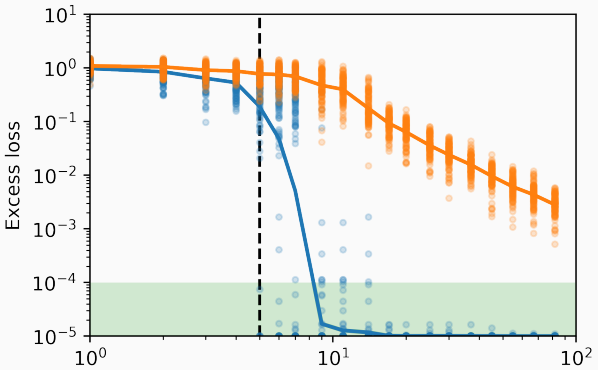


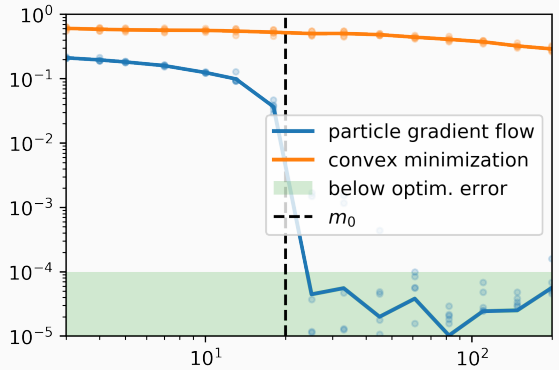
Particle-complexity

What about convex weights minimization ?



Particle-complexity





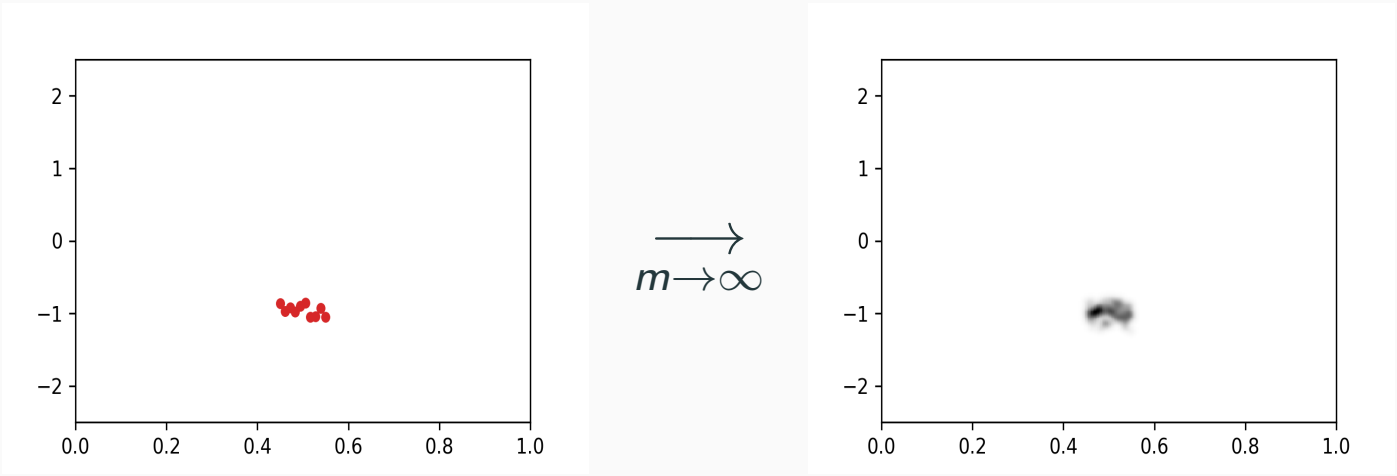
(a) Sparse deconvolution ($d = 1$) **(b)** Neural net (sigmoid, $d = 100$)

Excess loss at convergence vs nb of particles m.

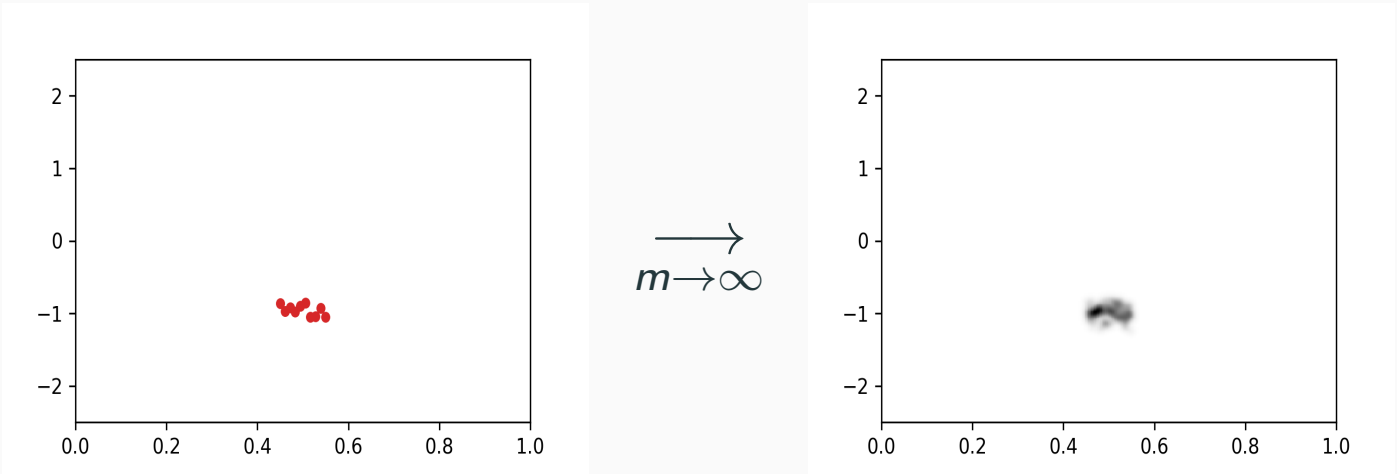
Simplest minimizer has m_0 particles.

Proof arguments

Many-particle limit



Many-particle limit



Proposition (Wasserstein gradient flow)

If $\mu_{m,0}$ converges to μ_0 (in Wasserstein) then $(\mu_{m,t})_t$ converges to the unique Wasserstein gradient flow $(\mu_t)_t$ of F starting from μ_0, characterized by

$$\partial_t \mu_t + \operatorname{div}(v_t \mu_t) = 0 \quad \textit{with} \quad v_t = -\nabla F'_{\mu_t}.$$

[See also]

Mei, Montanari, Nguyen (2018). *A Mean Field View of the Landscape of Two-Layers Neural Networks.*
Rotskoff, Vanden-Eijnden (2018). *Neural Networks as Interacting Particle Systems.*
Sirignano, Spiliopoulos (2018). *Mean Field Analysis of Neural Networks.*

Remarks

Proof technique

- existence and uniqueness through "local" geodesic *semi*-convexity and general theory of Ambrosio et al.
- the term of global interaction $\mu \mapsto R(\int \phi \mathrm{d}\mu)$ is new

Remarks

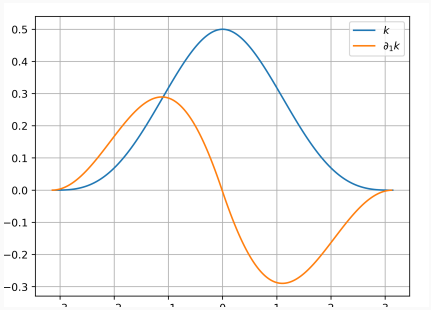
Proof technique

- existence and uniqueness through "local" geodesic *semi*-convexity and general theory of Ambrosio et al.
- the term of global interaction $\mu \mapsto R(\int \phi \mathrm{d}\mu)$ is new

Quadratic loss

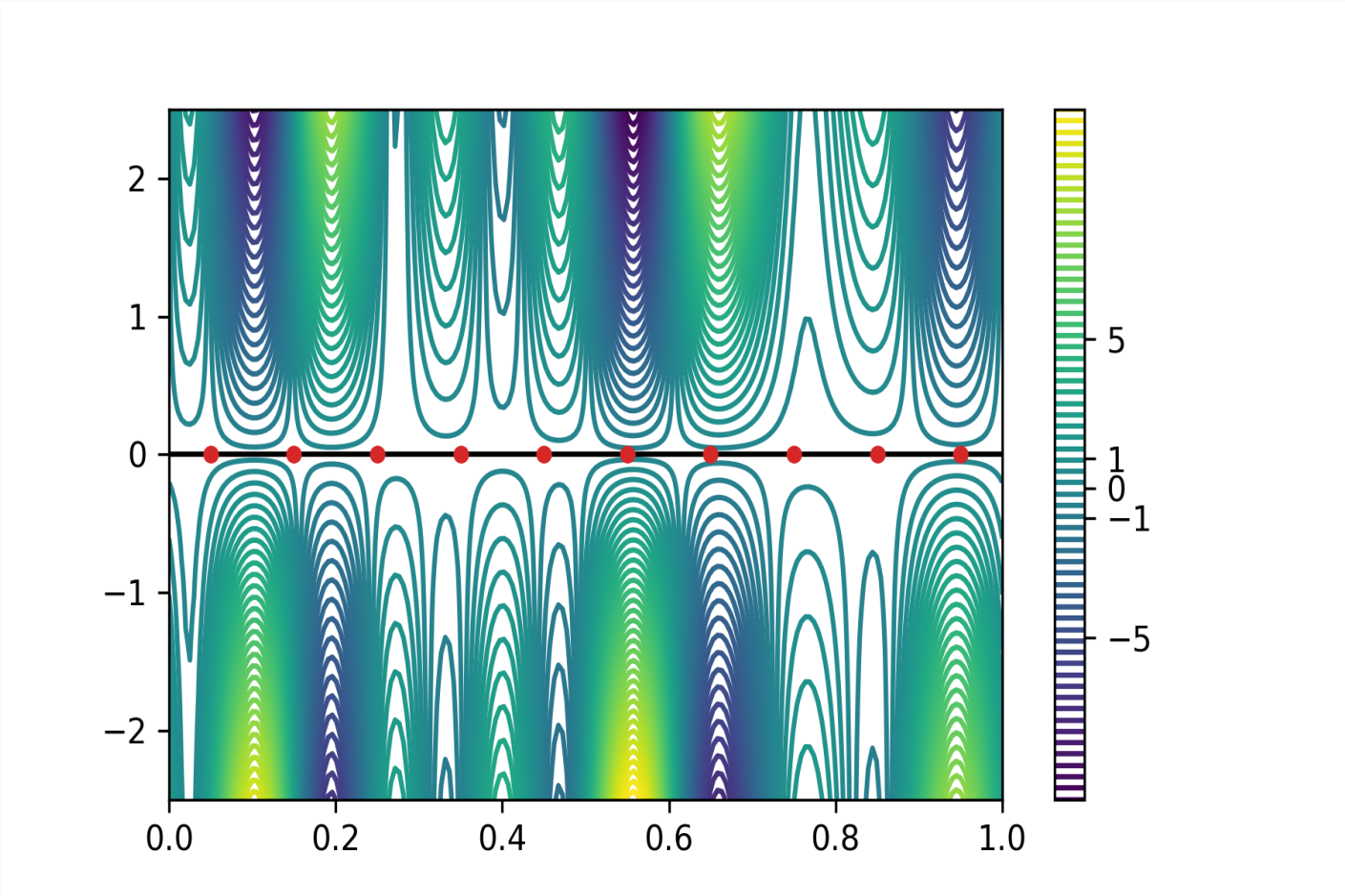
When $R(f) = \frac{1}{2}\|f - f^*\|^2$ for some $f^* \in \mathcal{F}$, interpretation as a system of charged particles with *varying* charge and interaction kernel

$$k((w_1, \theta_1), (w_2, \theta_2)) = w_1 w_2 \langle \phi(\theta_1), \phi(\theta_2) \rangle_{\mathcal{F}}.$$



Differential and Wasserstein (sub)-differential

The differential of $F : \mathcal{M}(\mathbb{R} \times \Theta) \to \mathbb{R}$ at μ is a function F'_μ.
The velocity of a particle located at u at time t is $-\nabla F'_{\mu_t}(u)$.



Level sets of F'_{μ_t} (a.k.a *mean field potential*) and particles of μ_t.

Optimality conditions

Global minimizers of F on $\mathcal{M}_+(\mathbb{R} \times \Theta)$ are characterized by

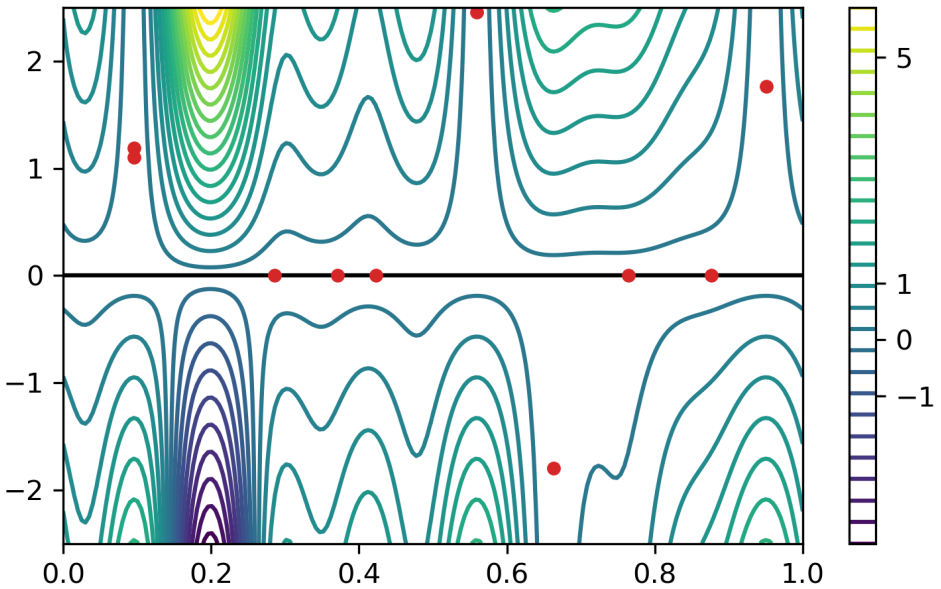
$$\begin{cases} F'_\mu \geq 0 \text{ everywhere on } \mathbb{R} \times \Theta \\ F'_\mu = 0 \ \mu\text{-a.e.} \end{cases}$$

Wasserstein gradient flow stationary points

Stationary points of (μ_t) are characterized by $\nabla F'_\mu = 0$, μ-a.e.

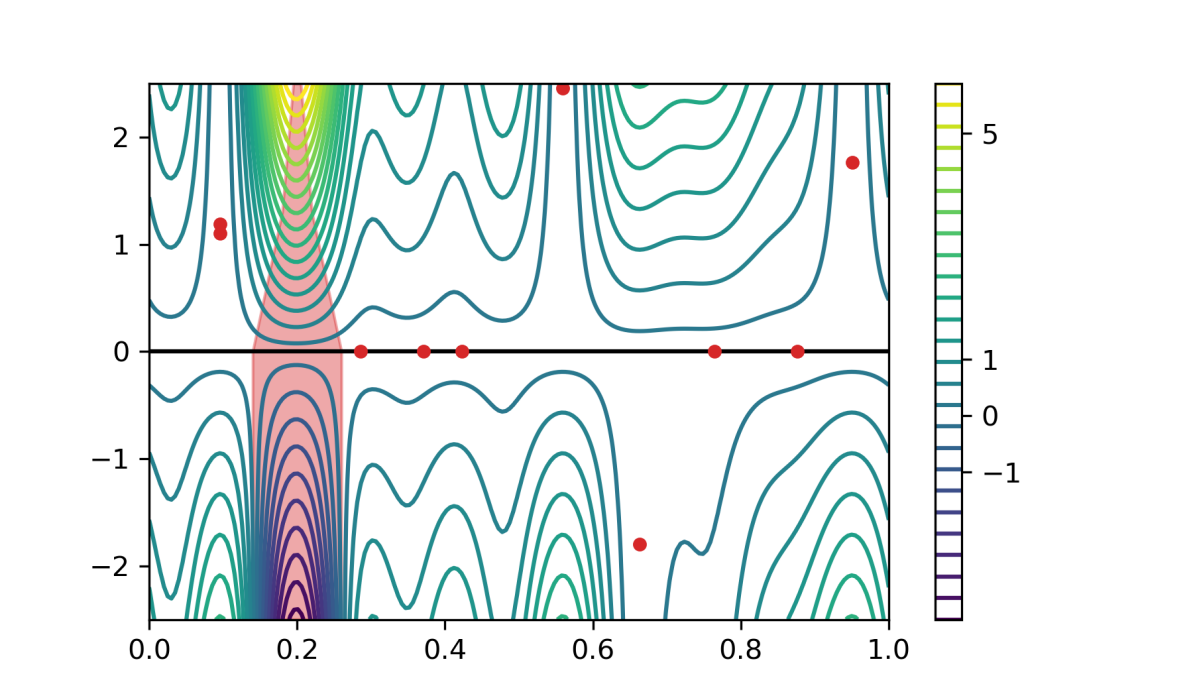
$$\text{Homogeneity} \quad \Rightarrow \quad F'_\mu = 0, \ \mu\text{-a.e.}$$

Avoiding bad local minima



Non-optimal stationary point $\tilde{\mu}$ and $F'_{\tilde{\mu}}$.

Avoiding bad local minima



Avoided if (arbitrary small) mass of $\mu_{t_0} \in \mathcal{N}(\tilde{\mu})$ lies in the red set.

Proposition (Escape criterion)

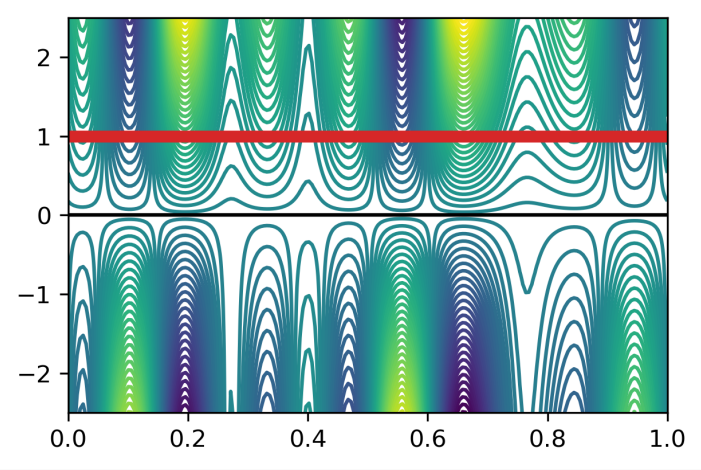
Let $\tilde{\mu}$ such that $F'_{\tilde{\mu}} \not\geq 0$. There is a neighborhood $\mathcal{N}(\tilde{\mu})$ and a "red set" (as above) such that:

If $\mu_{t_0} \in \mathcal{N}(\tilde{\mu})$ and μ_{t_0} gives mass to the set, then μ_t subsequently escapes from $\mathcal{N}(\tilde{\mu})$.

Separation property

Definition (Separation property)

Any continuous path joining $\{-\infty\} \times \Theta$ to $\{+\infty\} \times \Theta$ intersects support(μ).

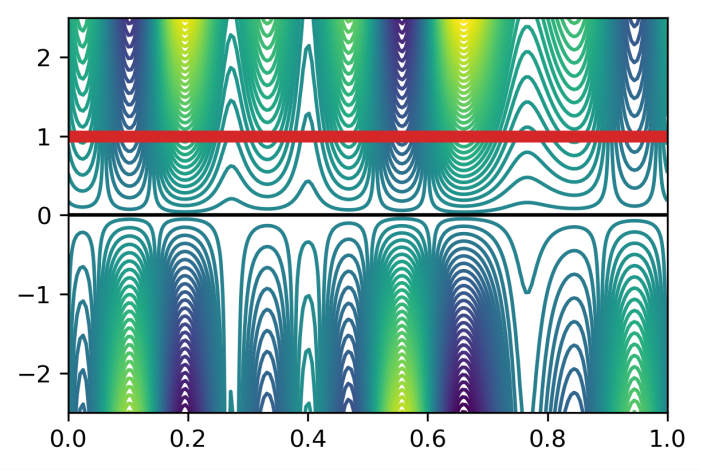


Support(μ_t) in red

Separation property

Definition (Separation property)

Any continuous path joining $\{-\infty\} \times \Theta$ to $\{+\infty\} \times \Theta$ intersects support(μ).



Support(μ_t) in red

Proposition (Stability)

If μ_0 satisfies the separation property, so does μ_t for $t > 0$.

Proof.

- easy if v_t is continuous : the flow of characteristics is a diffeo
- otherwise, topological degree theory □

Proof conclusion

"Proof" of the main theorem.

If μ_0 satisfies the separation property, then (μ_t) avoids all bad local minima. □

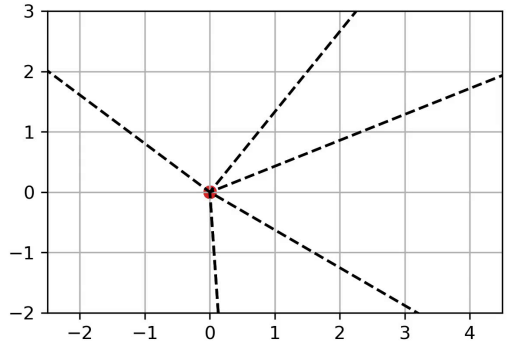
Proof conclusion

"Proof" of the main theorem.

If μ_0 satisfies the separation property, then (μ_t) avoids all bad local minima. □

Remarks

- quite insensitive to the choice of metric
- the 2-homogeneous case involves spherical geometry



Training ReLU neural net with SGD to optimality

Conclusion

- practical method, global convergence
- non-convex: initialization matters

Perspectives

- quantitative statements
- other models: more than one layer

[Paper] Chizat, Bach (2018). On the Global Convergence of Over-parameterized Models using Optimal Transport.