Optimization on Measures with Over-parameterized Gradient Descent

Lénaïc Chizat*, joint work with Francis Bach+

Nov. 22nd 2019 - Optimization on Measures Conference - Toulouse

*CNRS and Université Paris-Sud +INRIA and ENS Paris

Introduction

Optimization on Measures

Setting

- Θ compact d-Riemannian manifold without boundaries
- $\mathcal{M}_+(\Theta)$ nonnegative finite Borel measures
- $\phi : \Theta \to \mathcal{F}$ smooth, $\mathcal{F}$ separable Hilbert space
- $R : \mathcal{F} \to \mathbb{R}_+$ convex and smooth, $\lambda \geq 0$

$$\min_{\nu \in \mathcal{M}_+(\Theta)} J(\nu) := R\left(\int_\Theta \phi(\theta)\,\mathrm{d}\nu(\theta)\right) + \lambda\nu(\Theta)$$

In this talk

Simple non-convex gradient descent algorithms reaching ϵ-accuracy in $O(\log(1/\epsilon))$ complexity under non-degeneracy assumptions.

Dealing with the signed case

Consider a function $\tilde{\phi} : \tilde{\Theta} \to \mathcal{F}$ and $\tilde{J} : \mathcal{M}(\tilde{\Theta}) \to \mathbb{R}$ defined as

$$\tilde{J}(\nu) = R\left(\int \tilde{\phi}\, \mathrm{d}\mu\right) + \lambda|\nu|(\tilde{\Theta})$$

- define Θ the disjoint union of two copies $\tilde{\Theta}_+$ and $\tilde{\Theta}_-$ of $\tilde{\Theta}$
- define $\phi : \Theta \to \mathcal{F}$ as

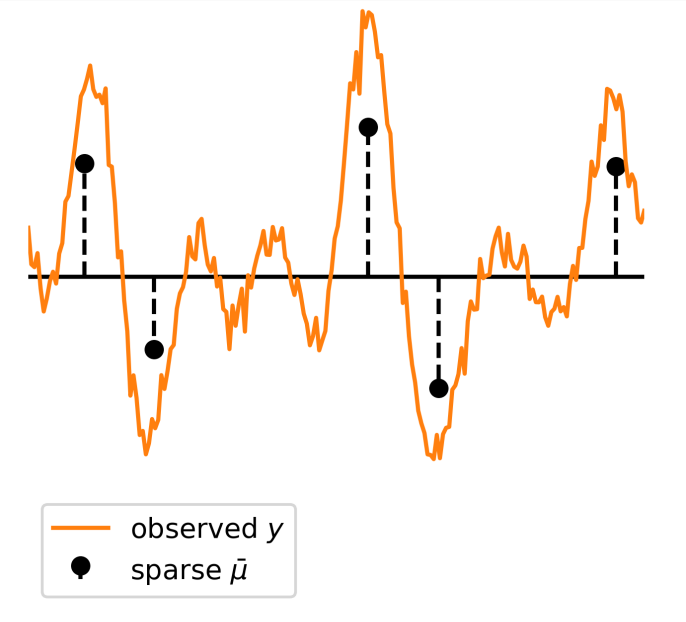
$$\phi(\theta) = \begin{cases} +\tilde{\phi}(\theta) & \text{if } \theta \in \tilde{\Theta}_+ \\ -\tilde{\phi}(\theta) & \text{if } \theta \in \tilde{\Theta}_- \end{cases}$$

Proposition

The signed and the nonnegative formulations are equivalent:

- *minima are the same*
- *given a minimizer for one problem, we can build a minimizer for the other*

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 $y = \varphi * \overline{\mu} + \text{noise}$

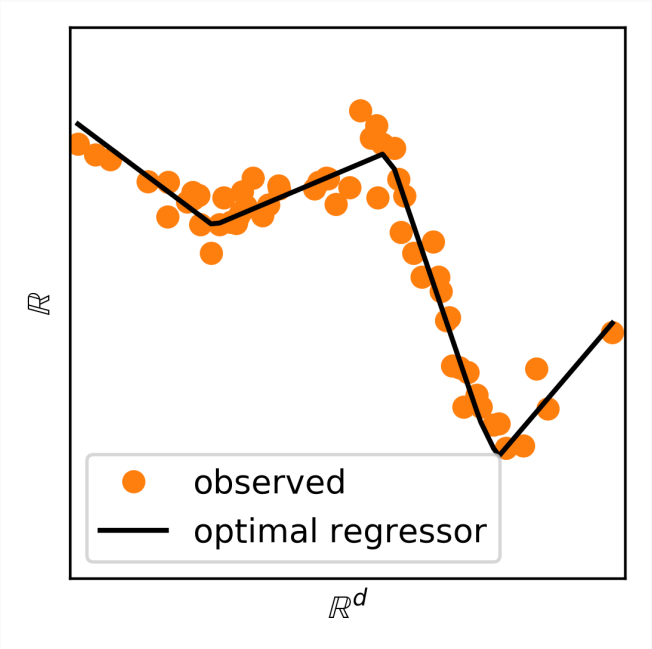
Variational approach (B-LASSO)

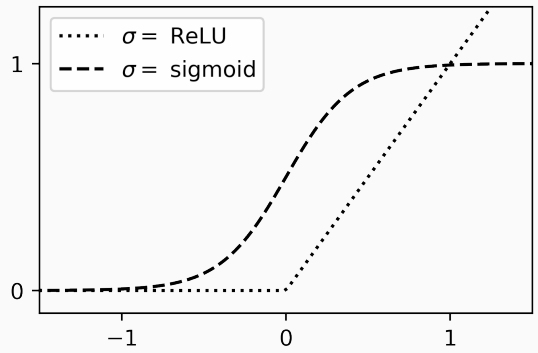
$$\min_{\mu \in \mathcal{M}(\Theta)} \underbrace{\frac{1}{2}\|y - \varphi * \mu\|_{L^2}^2}_{\text{loss}} + \underbrace{\lambda \|\mu\|_{TV}}_{\text{regularization}}$$

[Refs]

Candès, Fernandez-Granda (2014). *Towards a mathematical theory of super-resolution.*
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





Supervised 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

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

[Refs:]
Bengio, Roux, Vincent, Delalleau, Marcotte (2006). *Convex neural networks.*
Bach (2017). *Breaking the curse of dimensionality with convex neural networks.*

Demotivating Example

Continuous optimization

Let $\phi : \Theta \to \mathbb{R}$ be an arbitrary smooth function with minimum $\phi^\star < 0$. Solve

$$\min_{\theta \in \Theta} \phi(\theta)$$

Convex formulation

$$\min_{\nu \in \mathcal{M}_+(\Theta)} \frac{1}{2} \left(2 + \int_\Theta \phi(\theta) \, \mathrm{d}\nu(\theta) \right)^2 + \lambda \nu(\Theta)$$

Proposition (Equivalence)

It if $\lambda < -2\phi^\star$ *then* $\operatorname{spt} \nu^\star \subset \arg\min \phi$ *and* $\nu(\Theta) > 0$.

$\rightsquigarrow$ exponential complexity in the dimension d is unavoidable

Algorithm

Particle Gradient Descent

Algorithm (general case)

- initialize with discrete measure $\nu = \frac{1}{m}\sum_{i=1}^{m} r_i^p \delta_{\theta_i}$, with $p \geq 1$
- run gradient descent (or variant) on $(r_i, \theta_i)^m \in (\mathbb{R}_+ \times \Theta)^m$

Questions for theory

1. What choice for p? for the metric on $\mathbb{R}_+ \times \Theta$?
2. Is it a consistent method? for which initialization?
3. Are there computational complexity guarantees?

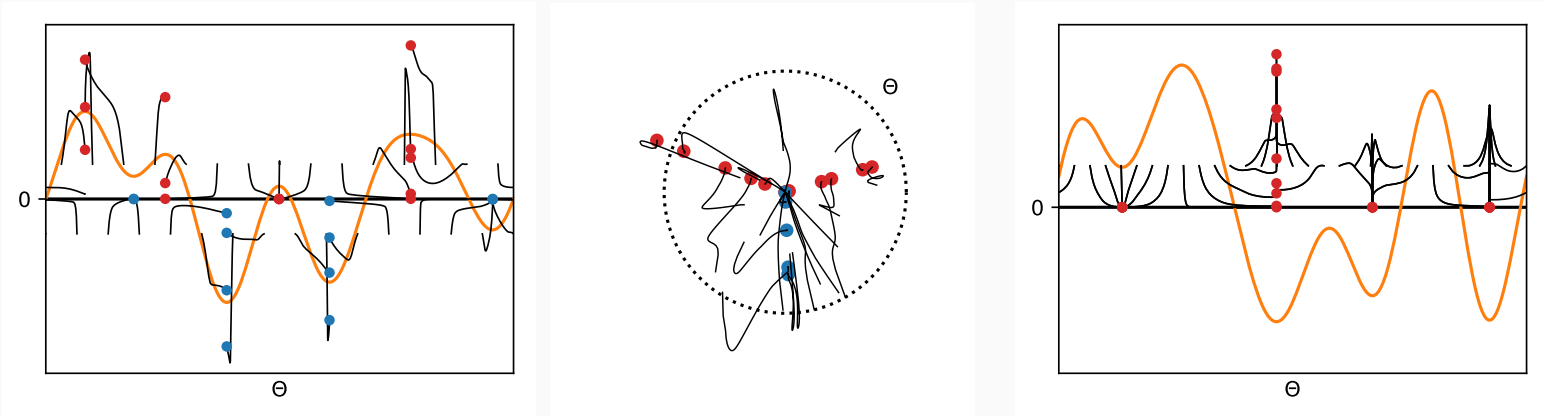
Conic Particle Gradient Descent

Algorithm (conic particle gradient descent)

Take $p = 2$ and discretize (with retractions) the gradient flow

$$\begin{cases} r_i'(t) = -2\alpha r_i(t) J'_{\nu_t}(\theta_t(t)) \\ \theta_i'(t) = -\beta \nabla J'_{\nu_t}(\theta_i(t)). \end{cases}$$

where $J'_\nu(\theta) = \langle \phi(\theta), \nabla R(\int \phi \, \mathrm{d}\nu) \rangle + \lambda$ "is" the Fréchet derivative of J at ν and $\nu_t = \frac{1}{m} \sum_{i=1}^m r_i(t)^2 \delta_{\theta_i(t)}$.



(a) Sparse deconv. **(b)** 2-layer neural net. **(c)** Generic optim.

Illustration

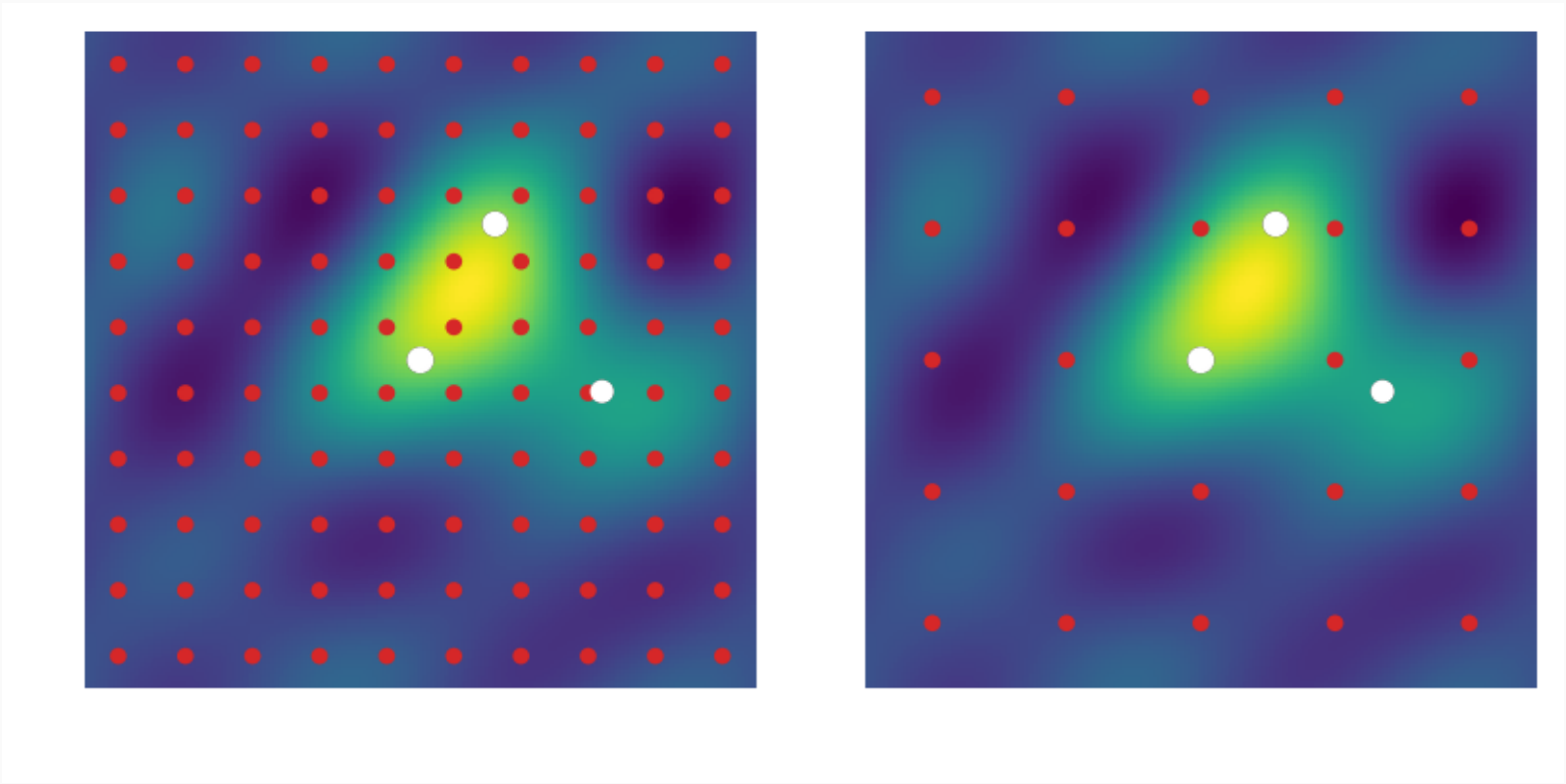


Figure 2: Sparse deconvolution on $\mathbb{T}^2$ with Dirichlet kernel (white) sources (red) particles.

Interpretations

The conic particle gradient flow can be seen as...

- gradient flow in $(\mathbb{R}_+ \times \Theta)^m$ with the (product) cone metric
- when Θ is the sphere, gradient flow on $(\mathbb{R}^{d+1})^m$
- Wasserstein gradient flow in $\mathcal{P}_2(\mathbb{R}_+ \times \Theta)$ with the cone metric
- Wasserstein-Fisher-Rao gradient flow on $\mathcal{M}_+(\Theta)$

Structure on $\mathbb{R}_+ \times \Theta$

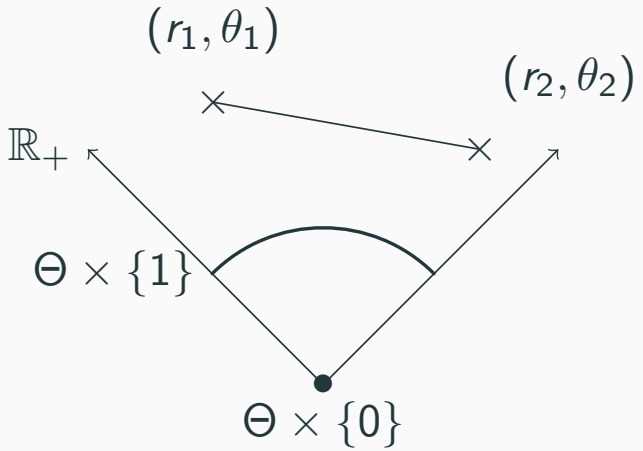
Definition (Cone metric)

$$\langle (\delta r_1, \delta\theta_1), (\delta r_2, \delta\theta_2) \rangle_{(r,\theta)} := \frac{1}{\alpha} \delta r_1 \cdot \delta r_2 + \frac{r^2}{\beta} \langle \delta\theta_1, \delta\theta_2 \rangle_\theta$$

- posing $\cos_\pi(z) = \cos(\min\{\pi, z\})$, the distance is given by

$$\mathrm{dist}((r_1, \theta_1), (r_1, \theta_2))^2 = r_1^2 + r_2^2 - 2 r_1 r_2 \cos_\pi(\mathrm{dist}(\theta_1, \theta_2))$$

- when $\Theta = \mathbb{S}^d$, the map $(r, \theta) \to r\theta \in \mathbb{R}^{d+1}$ is an isometry
- automatic structure when ϕ is 2-homogeneous on $\mathbb{R}^{d+1}$



Gradient flow in Wasserstein space

Consider, on the space $\mathcal{P}_2(\mathbb{R}_+ \times \Theta)$:

- the trajectory $\mu_t = \frac{1}{m} \sum_{i=1}^m \delta_{(r_i(t), \theta_i(t))}$
- the objective $F(\mu) = J(\mathrm{h}\mu)$ where $\mathrm{h}\mu(B) = \int r^2 \mu(\mathrm{d}r, B)$

Optimal transport interpretation

$(\mu_t)_t$ is the Wasserstein gradient flow of F, where $\mathbb{R}_+ \times \Theta$ is endowed with the cone metric.

- Wasserstein gradient flows are weak solutions of

$$\partial_t \nu_t = -\mathrm{div}\left(-\nabla F'_{\mu_t} \mu_t\right)$$

  where $F'_\mu \in \mathcal{C}^1(\Theta)$ "is" the Fréchet derivative of F at μ
- div/∇ defined in the cone metric on $\mathbb{R}_+ \times \Theta$

$\rightsquigarrow$ gives existence and uniqueness for initialization in $\mathcal{P}_2(\mathbb{R}_+ \times \Theta)$

[Refs]

Ambrosio, Gigli, Savaré (2008). *Gradient flows in metric spaces and in the space of probability measures.*

Gradient flow in Wasserstein Fisher-Rao space

Consider, on the space $\mathcal{M}_+(\Theta)$:

- the trajectory $\nu_t = \frac{1}{m}\sum_{i=1}^m r_i(t)^2 \delta_{\theta_i(t)}$
- the objective $J(\nu)$

Unbalanced optimal transport interpretation

$(\nu_t)_t$ is the Wasserstein-Fisher-Rao gradient flow of J.

- Wasserstein-Fisher-Rao gradient flows are weak solutions of

$$\partial_t \nu_t = -\mathrm{div}\left(-\nabla J'_{\nu_t} \nu_t\right) - 4 J'_{\nu_t} \nu_t$$

- this metric can be defined as (W_2 in the cone metric)

$$WFR(\nu_1, \nu_2) = \min\left\{W_2(\mu_1, \mu_2)\ ;\ (\mathrm{h}\mu_1, \mathrm{h}\mu_2) = (\nu_1, \nu_2)\right\}$$

- all statements could be made alternatively on μ_t or ν_t.

[Refs]
Liero, Mielke, Savare (2015). *Optimal Entropy-Transport Problems and a new Hellinger-Kantorovich metric [...].*
Kondratiev, Monsaingeon, Vorotnikov (2015), *A new optimal transport distance on the space of [...] measures.*
Chizat, Peyré, Schmitzer, Vialard (2015). *An interpolating distance between optimal transport and Fisher-Rao.*
Gallouët, Monsaingeon (2017). *A JKO Splitting Scheme for Kantorovich–Fisher–Rao Gradient Flows.*

Other Methods

	pros	cons
conditional gradient	known rate, sparse	1 iter. is hard
moment methods	asymptotically exact	heavy, not generic
particle gradient flow	easy, cheap iteration	guarantees ?

[Refs]:
Lasserre (2010). *Moments, positive polynomials and their applications.*
Bredies, Pikkarainen (2013).*Inverse problems in spaces of measures.*
Frank, Wolfe (1956). *An algorithm for quadratic programming.*
Bach (2017). *Breaking the curse of dimensionality with convex neural networks.*
Catala, Duval, Peyré (2017). *A low-rank approach to off-the-grid sparse deconvolution.*
...

Global convergence

Many-particle / mean-field limit

Theorem

Assume that $\nu_{0,m} \to \nu_0$ weakly. Then $\nu_{t,m} \to \nu_t$ weakly, uniformly on $[0, T]$, where $(\nu_t)_{t \geq 0}$ is the Wasserstein-Fisher-Rao gradient flow of J starting from ν_0.

[Refs]:

Nitanda, Suzuki (2017). *Stochastic particle gradient descent for infinite ensembles.*

Mei, Montanari, Nguyen (2018). *A Mean Field View of the Landscape of Two-Layers Neural Networks.*

Rotskoff, Vanden-Eijndem (2018). *Parameters as Interacting Particles [...].*

Sirignano, Spiliopoulos (2018). *Mean Field Analysis of Neural Networks.*

Chizat, Bach (2018). *On the Global Convergence of Gradient Descent for Over-parameterized Models [...]*

Asymptotic Global Convergence

Assumptions: $\lambda \geq 0$, Sard-type property (e.g. $\phi \in \mathcal{C}^d(\Theta)$)

Theorem (Chizat and Bach, 2018)

If ν_0 has full support on Θ and $(\nu_t)_{t\geq 0}$ converges as $t \to \infty$, then the limit is a global minimizer of J.

Moreover, if $\nu_{m,0} \to \nu_0$ weakly as $m \to \infty$, then

$$\lim_{m,t\to\infty} J(\nu_{m,t}) = \min_{\nu\in\mathcal{M}_+(\Theta)} J(\nu).$$

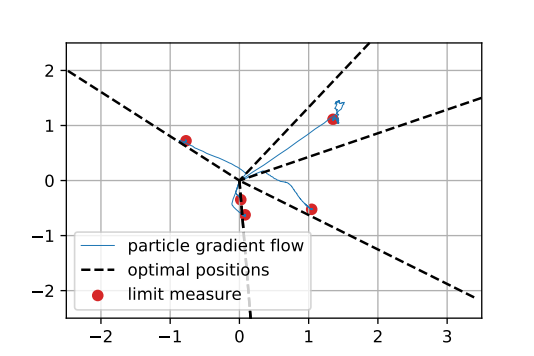
Remarks

- bad stationnary point *exist*, but are avoided thanks to the init.
- such results hold for more general particle gradient flows
- can we say more for the *conic* and *sparse* case?

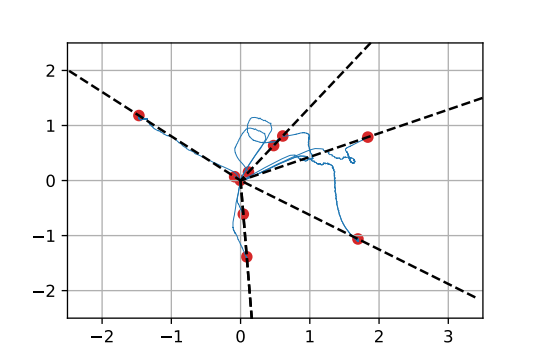
[Refs]
Chizat, Bach (2018). *On the Global Convergence of Gradient Descent for Over-parameterized models [...].*

Numerical Illustrations

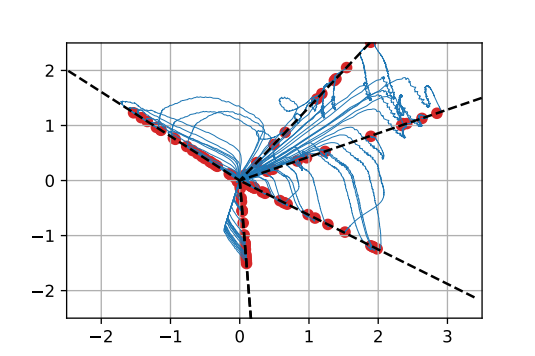
ReLU network, $d = 2$, optimal predictor has 5 neurons (pop. risk)



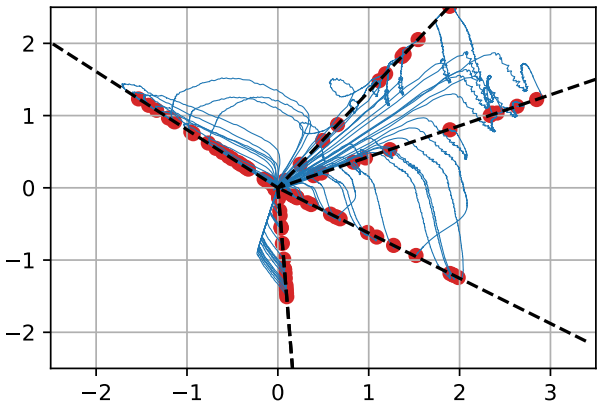
5 neurons



10 neurons



100 neurons



Local convergence

Assumptions for the local analysis

Sparse minimizer

Assume that $\nu^\star = \sum_{i=1}^{m^\star} r_i^2 \delta_{\theta_i}$ is the unique minimizer of J

Non-degeneracy

Typical in the analysis of such sparse problems

Assume (everything in normal coordinates around θ_i):

- (coercivity) $\nabla^2 R \succeq \sigma \mathrm{Id}$ at the optimum
- (local curvature) $H_i = \nabla^2 J'_{\nu^\star}(\theta_i) \succ 0$
- (strict slackness) $J'_{\nu^\star}$ does not vanish except at $\theta_1, \dots, \theta_{m^\star}$
- (global interaction) one has $K \succ 0$, where

$$K_{(i,j),(i',j')} = \left\langle r_i \bar{\nabla}_j \Phi(\theta_i), \nabla^2 R\left(\int \Phi \, \mathrm{d}\mu^\star\right)(r_{i'} \bar{\nabla}_{j'} \Phi(\theta_{i'}))\right\rangle$$

where $\bar{\nabla}\Phi = (2\alpha\Phi, \beta\nabla\Phi)$.

Sharpness/Polyak-Łojasiewicz Inequality

Theorem (C., 2019)

Under these assumptions, there exists $J_0, \kappa_0 > 0$ *such that for any* $\nu \in \mathcal{M}_+(\Theta)$ *satisfying* $J(\nu) \leq J_0$*, it holds*

$$\underbrace{\int \left(4\alpha|J'_\nu|^2 + \beta\|\nabla J'_\nu\|_\theta^2\right) \mathrm{d}\nu}_{\text{Squared-norm of gradient}} \geq \kappa_0 \min\{\alpha, \beta\} \underbrace{\left(J(\nu) - J^\star\right)}_{\text{Optimality gap}}.$$

- J_0 and κ_0 polynomial in the problem characteristics
- crucially, J_0 is independent of the over-parameterization

Corollary

If $J(\nu_0) \leq J_0$*, gradient flow and gradient descent (with* $\max\{\alpha, \beta\}$ *small enough) converge exponentially fast to the global minimizer, in value and in distance (e.g. Bounded-Lipschitz, WFR).*

Proof idea (I)

Fix a small radius τ, partition Θ as

$$\begin{cases} \Theta_i := \{\theta \in \Theta \,;\, \mathrm{dist}(\theta, \theta_i) < \tau\} \\ \Theta_0 := \Theta \setminus \cup_{i=1}^m \Theta_i \end{cases}$$

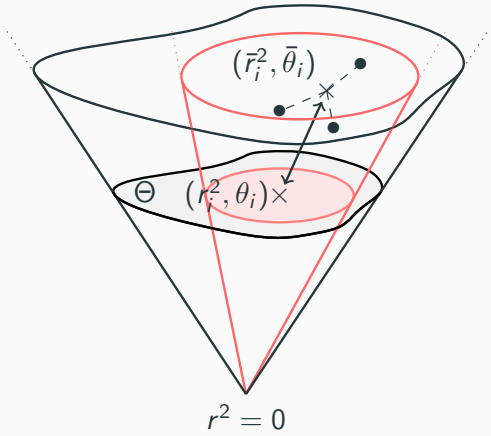
For $\nu \in \mathcal{M}_+(\Theta)$, let (in normal coordinates):

$$\bar{r}_i^2 := \int_{\Theta_i} \mathrm{d}\nu(\theta) \qquad \text{(mean mass)}$$

$$\bar{\theta}_i := \frac{1}{\bar{r}_i^2} \int_{\Theta_i} \theta \, \mathrm{d}\nu(\theta) \qquad \text{(mean position)}$$

$$\Sigma_i := \frac{1}{\bar{r}_i^2} \int_{\Theta_i} (\theta - \bar{\theta}_i)(\theta - \bar{\theta}_i)^\intercal \, \mathrm{d}\nu(\theta)$$

(centered covariance of position)



Proof idea (II)

Then, Taylor expansions lead to controls on $J(\nu)$ and the gradient norm that only depend on $r_i, \bar{r}_i, \theta_i, \bar{\theta}_i, \Sigma_i$.

Remarks

- Łojasiewicz inequalities in Wasserstein space recently studied (e.g. logarithmic Sobolev inequality & relative entropy), but in the geodesically convex case
- analysis enabled by the specific structure related to $p = 2$ and the cone metric

[Refs]
Wisibono (2018). *Sampling as optimization in the space of measures*
Blanchet, Bolte (2018). *A Family of Functional Inequalities: Lojasiewicz inequalities and displacement convex functions.*

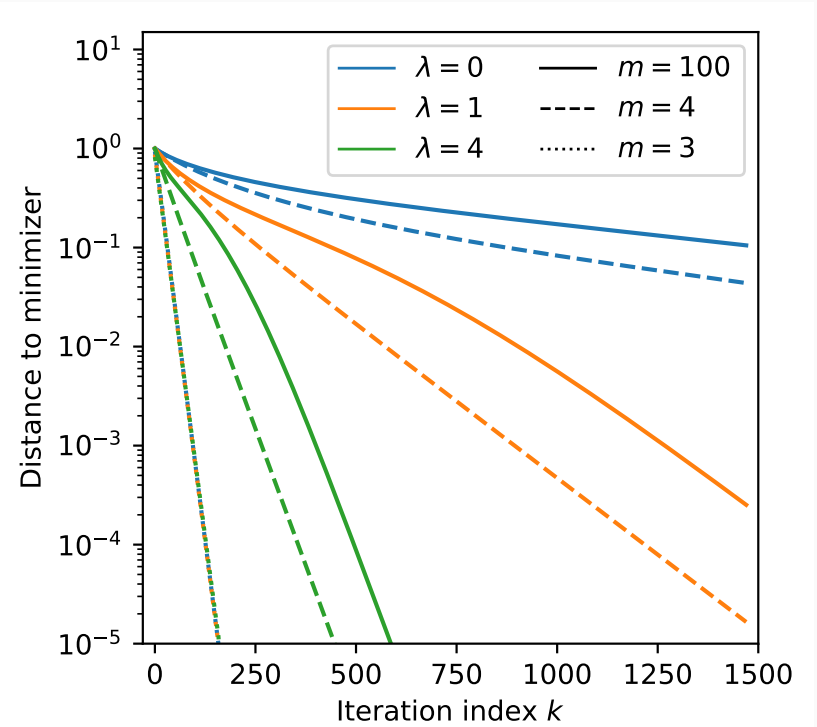
Insights on local rates

We get a local expansion of J (b vector of biases, Σ_i local covariances):

$$J(\nu) - J^\star \approx \frac{1}{2} b^\intercal (K + \lambda H) b + \frac{\lambda}{2} \sum_{i=1}^{m} r_i^2 \operatorname{tr}(\Sigma_i H_i) + \int_{\Theta_0} J'_{\nu^\star} d\nu$$

Local rate vs regularization and over-parameterization

Local convergence for 2D sparse deconvolution



Quantitative Global Convergence

Quantitative Global Convergence

Fine tuning

Particule gradient descent can be used after any optimization algorithm : discrete convex optimization, conditional gradient...

$\rightsquigarrow$ Can we use a single algorithm?

Quantitative Global Convergence

Fine tuning

Particule gradient descent can be used after any optimization algorithm : discrete convex optimization, conditional gradient...

$\rightsquigarrow$ Can we use a single algorithm?

Theorem (C., 2019)

For $M, \epsilon > 0$ fixed, there exists $C_1, C_2 > 0$ such that if for $\eta > 0$,

$$W_\infty(\nu_0, M\mathrm{vol}) < C_1\eta \qquad \text{and} \qquad \frac{\beta}{\alpha} < C_2\eta^{2(1+\epsilon)}$$

then for $\alpha t \geq C_3/\eta^{1+\epsilon}$ it holds $J(\nu_t) - J^\star \leq \eta$. It particular, if this holds for $\eta = J_0 - J^\star$, then $(\nu_t)_{t\geq 0}$ converges to a global minimizer.

Via samples or a grid: $W_\infty(\nu_0, \mathrm{vol}) \asymp m^{-1/d}$ with m particles.

Proof Idea: Perturbed Mirror Descent

Lemma (Mirror descent rate)

The dynamic with $\beta = 0$ satisfies, for some $C > 0$,

$$J(\nu_t) - J(\nu^\star) \lesssim \inf_{\nu \in \mathcal{M}_+(\Theta)} \left\{ \|\nu^\star - \nu\|_{\mathrm{BL}} + \frac{1}{Ct} \mathcal{H}(\nu, \nu_0) \right\}$$
$$\lesssim \frac{\log t}{t} \qquad \text{if } \nu_0 \propto \mathrm{d}\,\mathrm{vol}$$

where $\mathcal{H}$ is the relative entropy (Kullback-Leibler divergence).

Proof idea of the theorem.

Adapt this lemma to deal with $\beta > 0$. □

- discrete time dynamics & choice of retraction: see paper

Importance of the geometry

Comparison of $p = 1$ and $p = 2$ for finding one spike

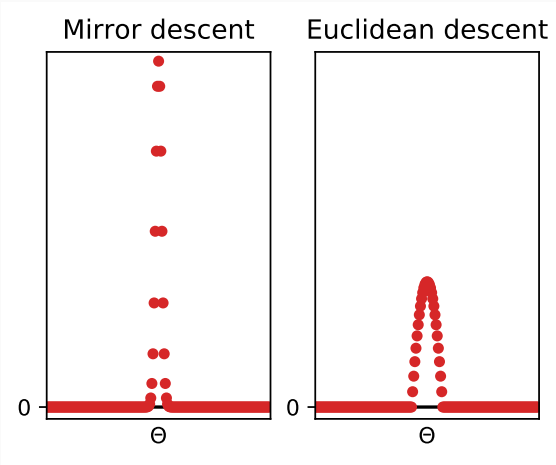
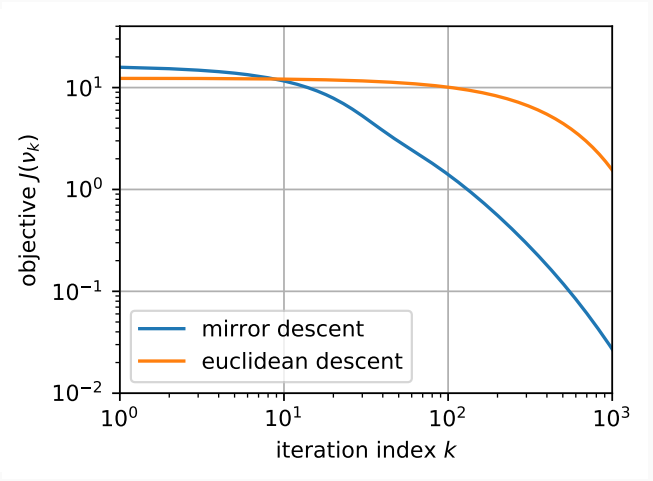


illustration
$(d = 1, \beta = 0)$

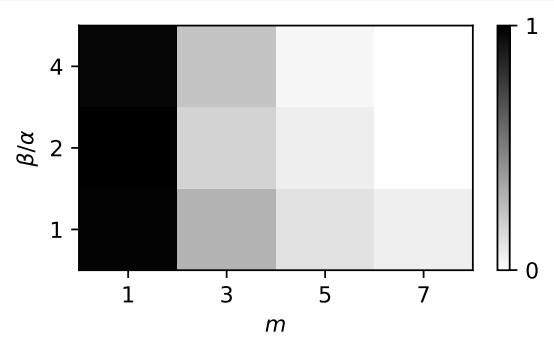


rate in objective
$(d = 2, \beta = 0)$

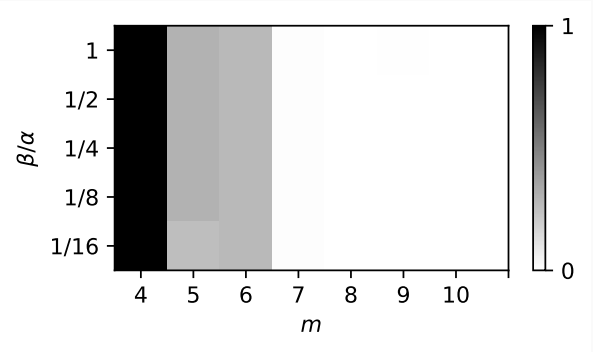
- multiplicative updates are crucial for efficiency
- moreover, conic geometry is crucial for the local analysis
- still, global convergence also true for $p = 1$

Rates for the fully non-convex case ?

Sub-optimality gap after convergence



ReLU network

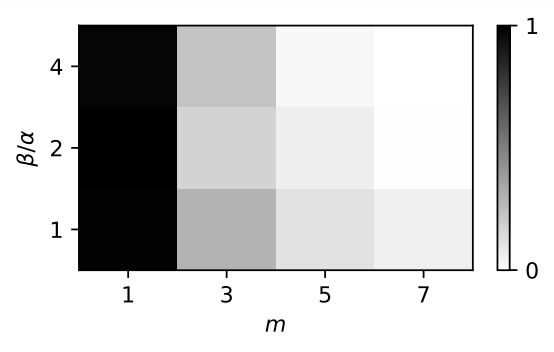


Sparse deconvolution

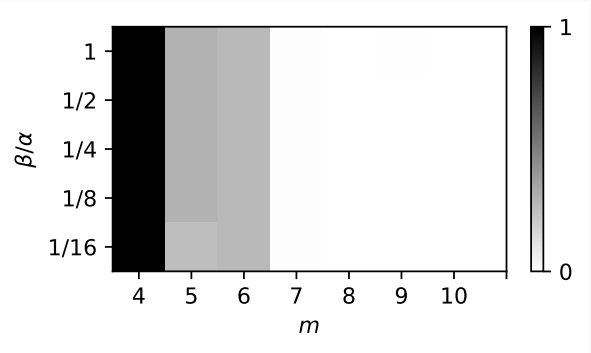
$\rightsquigarrow$ the condition $\beta/\alpha \ll 1$ does not seem important: removing it is an open problem

Rates for the fully non-convex case ?

Sub-optimality gap after convergence



ReLU network



Sparse deconvolution

$\rightsquigarrow$ the condition $\beta/\alpha \ll 1$ does not seem important: removing it is an open problem

Proposition (No condition on β/α)

There exists $C > 0$ such that for all $t, \eta > 0$, as long as $\nu_s \geq \eta$ vol for $0 \leq s \leq t$, then

$$J(\nu_t) - J^\star \leq \frac{C}{\sqrt{\eta t}}.$$

High-level remarks on the algorithm

Comparison with conditional gradient

Both algorithms typically involve two types of computational costs:

- computing $\nabla R(\int \phi \, \mathrm{d}\mu_t)$
- given this, updating the position and/or mass of m particles

$\rightsquigarrow$ other trade-offs are possible

Potential variants:

- inexact gradients (stochastic, delayed)
- resampling
- adaptive methods

[Refs]

Wei, Lee, Liu, Ma (2019). *Regularization Matters: Generalization and Optimization of Neural Nets v.s. their Induced Kernel*.

Conclusion

Conic particle gradient descent with theoretical garanties

Perspectives

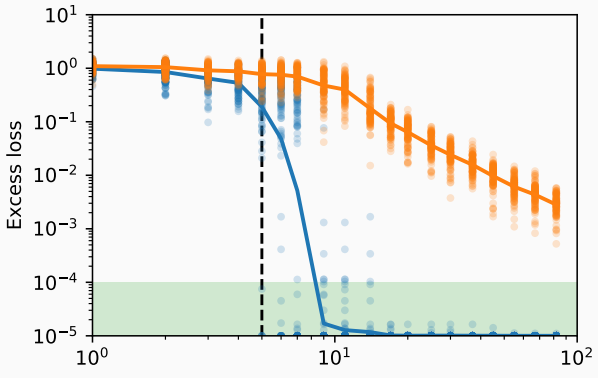
- **extension to the constrained case**
- **high dimensional quantitative results**
- **implicit bias for the unregularized case**

[Papers :]
- Chizat and Bach (2018). On the Global Convergence of Over-parameterized Models using Optimal Transport.
- Chizat (2019). Solving Sparse optimization on Measures with Over-Parameterized Gradient Descent.

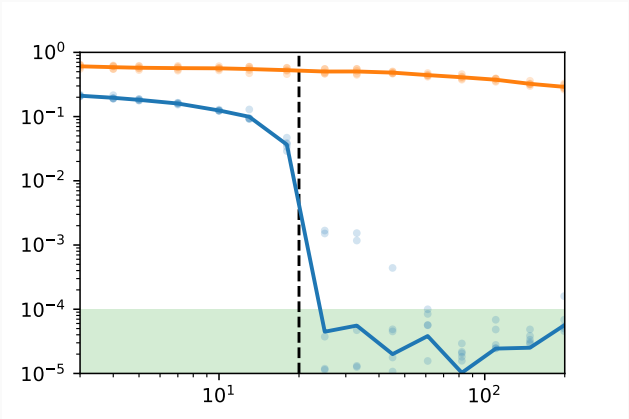
Convex vs non-convex

Comparison

Excess loss at convergence vs number of particles m



(a) Sparse deconvolution ($d = 1$)



(b) Neural net (sigmoid, $d = 100$)

Vertical dashed line shows the nb of particles of simplest minimizer