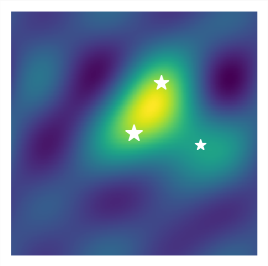
Sparse Optimization on Measures with Over-parameterized Gradient Descent

Lénaïc Chizat*

SIAM Activity Group on Imaging Science Best Paper Lecture 2022

*EPFL (work carried while at CNRS)

A Motivating Problem : Spikes Deconvolution



Blurred and noisy observation of stars on a domain $\mathcal{X}$
(here Dirichlet blurring kernel on the 2-torus)

Questions

- **Statistics.** Is recovery of positions, weights and number of particles possible? With which estimator?
- **Optimization.** Can we compute this estimator accurately and efficiently ? $\rightsquigarrow$This talk.

Estimator

Setting (simplified for this talk)

- ambiant space $\mathcal{X}$ (compact Riemannian d-manifold)
- observed signal $f \in L^2(\mathcal{X})$
- known impulse response $\phi(\cdot,\cdot) \in \mathcal{C}^3(\mathcal{X} \times \mathcal{X})$

Optimization problem

- Take $m \in \mathbb{N}$ particles with weight/position $(a, x) \in \mathbb{R}_+ \times \mathcal{X}$
- Parameterize with $\theta = \big((a_1, x_1), \dots, (a_m, x_m)\big) \in (\mathbb{R}_+ \times \mathcal{X})^m$
- Find the minimizer (in θ and m) of

$$F_m(\theta) := \underbrace{\int_{\mathcal{X}} \Big(\frac{1}{m}\sum_{i=1}^m a_i \phi(x, x_i) - f(x)\Big)^2 \mathrm{d}x}_{\text{Data fitting}} + \underbrace{\frac{\lambda}{m}\sum_{i=1}^m a_i}_{\text{Regularization}}$$

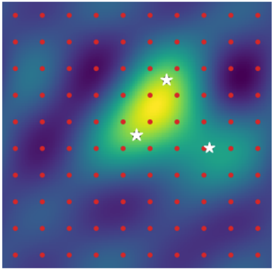
NB: F_m is not convex and admits spurious local minima

Conic Particle Gradient Descent

Algorithm (continuous time version)

- Initialize $(x_i)_i$ uniformly in $\mathcal{X}$ (at random/on a grid), $a_i = 1$
- Compute $(\theta(t))_{t\geq 0}$ by following

$$\begin{cases} \frac{d}{dt} a_i(t) = -4m\, a_i(t) \nabla_{a_i} F_m(\theta(t)) \\ \frac{d}{dt} x_i(t) = -\alpha m \nabla_{x_i} F_m(\theta(t)) \end{cases}$$



Why multiplicative updates for weights?

Initializing with $\theta(0) = (a_0, x_0)$

$\Leftrightarrow$

Initializing with

$\theta(0) = ((a_0/2, x_0), (a_0/2, x_0))$

Summary of results

Let $F^* := \inf_{m \geq 1, \theta} F_m(\theta)$ the optimal value

Theorem (Local convergence)

If the problem is *non-degenerate*, there exists $C_0, C_1 > 0$ such that

$$F_m(\theta(0)) \leq F^* + C_0 \quad \Rightarrow \quad F_m(\theta(t)) - F^* \leq C_0 e^{-C_1 t}.$$

Theorem (Global convergence)

If the problem is *non-degenerate*, there exists $C_0', C_1' > 0$ such that

$$\begin{cases} \alpha \leq C_0' \\ \sup_{x \in \mathcal{X}} \inf_{i=1,\dots,m} \mathrm{dist}(x, x_i(0)) \leq C_1' \end{cases} \quad \Rightarrow \quad \lim_{t \to \infty} F_m(\theta(t)) = F^*.$$

Applications and related algorithms

General problem: Find a sparse decomposition of an observed signal using a smoothly parameterized dictionary

Sampled applications

- **Imaging.** Astronomy (2D) [Puschmann 2017], Neuro-imaging with EEG (3D) [Gramfort 2013], Fluorescence microscopy (3D) [Betzig 2006]
- **Machine Learning.** 2-layer Relu neural networks, where CPGD $\Leftrightarrow$ backpropagation, Mixture models fitting [Keriven 2017] [Boyd et al 2015]

Other approaches for optimization on measures

- Moment methods: parameterize with moments [Lasserre]
- Stochastic algorithms: generalized Langevin dynamics
- Frank-Wolfe: add one particle per iteration [Bredies, 2013]

Outline

Statics: Sparse optimization over measures

Dynamics: Local convergence

Dynamics: Global convergence

Statics: Sparse optimization over measures

Formulation over measures

Symmetries lead to a natural reformulation:

$$\theta = (a_i, x_i)_{i=1}^m \in (\mathbb{R}_+ \times \mathcal{X})^m \quad \Rightarrow \quad \mu_m := \frac{1}{m}\sum_{i=1}^m a_i \delta_{x_i} \in \mathcal{M}_+(\mathcal{X})$$

Objective over the space of nonnegative measures $\mathcal{M}_+(\mathcal{X})$

$$F(\mu) = \underbrace{\frac{1}{2}\int_{\mathcal{X}} \Big(\int_{\mathcal{X}} \phi(x,y)\,\mathrm{d}\mu(y) - f(x) \Big)^2 \mathrm{d}x}_{\text{Data fitting}} + \underbrace{\lambda \mu(\mathcal{X})}_{\text{Regularization}}$$

Basic properties of F

- $F(\mu_m) = F_m(\theta)$
- convex
- admits a minimizer μ^*

Signed case ($a_i \in \mathbb{R}$)

Set $\begin{cases} \tilde{\phi} = (+\phi, -\phi) \\ \tilde{\mu} = (\mu_+, \mu_-) \end{cases}$

$\rightsquigarrow$ regularization by $\lambda \|\tilde{\mu}\|_{\mathrm{TV}}$ [De Castro & Gamboa, 2012]

Sparsity and optimality

Assumption 1 (Uniqueness)

There exists a unique minimizer which is sparse: $\mu^* = \sum_{i=1}^{m^*} a_i^* \delta_{x_i^*}$.

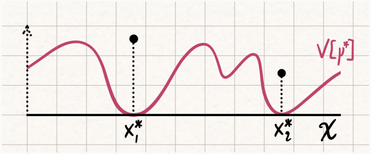
Let $V[\mu] \in \mathcal{C}^3(\mathcal{X})$ be the first variation of F at μ, characterized by

$$F(\mu+\epsilon\nu) = F(\mu) + \epsilon \int_{\mathcal{X}} V[\mu](x)\,\mathrm{d}\nu(x) + o(\epsilon), \quad \forall \nu \in \mathcal{M}(\mathcal{X}) \text{ adm.}$$

Proposition (Optimality conditions)

The first variation of F at μ^* satisfies

$$V[\mu^*] \geq 0 \quad \text{and} \quad \operatorname{spt}(\mu^*) = \{x_1^*, \dots, x_{m^*}^*\} \subset \{V[\mu^*] = 0\}.$$



Non-degeneracy

Definition (Interaction kernels)

Global interaction kernel $K \in \mathbb{R}^{(m^*(d+1))^2}$ (convention $\nabla_0 \phi = 2\phi$):

$$K_{(i,j),(i',j')} = \langle \sqrt{a_i^*} \nabla_j \phi(x_i^*, \cdot), \sqrt{a_{i'}^*} \nabla_{j'} \phi(x_{i'}^*, \cdot) \rangle_{L^2}$$

Local interaction kernel $H = \operatorname{diag}(H_i)_{i=1}^{m^*} \in \mathbb{R}^{(m^*(d+1))^2}$ with

$$H_i := \nabla^2 V[\mu^*](x_i^*)$$

Definition (Non-degeneracy)

We say that F is **non-degenerate** iff:

- $K \succ 0$
- $\arg\min V[\mu^*] = \{x_1^*, \dots, x_{m^*}^*\}$
- $H_i \succ 0$, $i \in \{1, \dots, m^*\}$

Can be guaranteed a priori under spikes separation & noise level conditions [Duval & Peyré, 2015] [Poon et al, 2019] [Akiyama & Suzuki, 2021]

Non-degeneracy vs. stability

Unbalanced L_2-Wasserstein metric (e.g. [Liero et al. 2020])

Define, for $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$:

$$\widehat{W}_2^2(\mu,\nu) := \min_{\gamma} \mathrm{KL}(\gamma_1|\mu) + \mathrm{KL}(\gamma_2|\nu) + \int c(x,y)\,\mathrm{d}\gamma(x,y)$$

where $\gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{X})$ has marginals γ_1, γ_2 and $c(x,y) \approx \mathrm{dist}(x,y)^2/\alpha^2$

Theorem (stability)

F is non-degenerate

$$\Rightarrow$$

$$\exists C_0, C_1 > 0 \text{ s.t. } F(\mu) - F^* \leq C_0 \Rightarrow \widehat{W}_2^2(\mu,\mu^*) \leq C_1\big(F(\mu) - F^*\big)$$

The opposite inequality $\widehat{W}_2^2(\mu,\mu^*) \geq C'\big(F(\mu) - F^*\big)$ holds, hence:

$F(\mu) - F^*$ small $\Leftrightarrow$ μ close to μ^*

Back to dynamics

Using the first-variation V, conic particle gradient descent solves:

$$\begin{cases} \frac{d}{dt}a_i(t) = -4m\, a_i(t) V[\mu_t](x_i(t)) \\ \frac{d}{dt}x_i(t) = -\alpha m \nabla V[\mu_t](x_i(t)) \end{cases}$$

where $\mu_t := \frac{1}{m}\sum_{i=1}^m a_i(t)\delta_{x_i(t)} \in \mathcal{M}_+(\mathcal{X})$.

Proposition (Dynamics in the space of measures)

The curve $(\mu_t)_t$ solves (distributionally) the PDE:

$$\partial_t \mu_t = \underbrace{\alpha \nabla \cdot \big(\mu_t \nabla V[\mu_t]\big)}_{\text{Drift}} - \underbrace{4\mu_t V[\mu_t]}_{\text{Reaction}}$$

This is the gradient flow of F under the metric $\widehat{W_2}$.

Dynamics: Local convergence

Energy dissipation

Let $f : \mathbb{R}^d \to \mathbb{R}$ a smooth function and $x : \mathbb{R}_+ \to \mathbb{R}^d$ a gradient flow of f, i.e.

$$\frac{d}{dt}x(t) = -\nabla f(x(t)), \quad \forall t \geq 0$$

Energy dissipation formula: Euclidean case

$$\frac{d}{dt}f(x(t)) = \nabla f(x(t))^\top x'(t) = -\|\nabla f(x(t))\|^2$$

In our context, let

$$\|\nabla_{\widehat{W_2}} F(\mu)\|^2 := \int_{\mathcal{X}} \left(\alpha \|\nabla V[\mu](x)\|^2 + 4|V[\mu](x)|^2\right) \mathrm{d}\mu(x)$$

Proposition (Energy dissipation for $(\mu_t)_t$)

$$\frac{d}{dt}F(\mu_t) = -\|\nabla_{\widehat{W_2}} F(\mu_t)\|^2$$

Main local convergence result

Theorem (A Łojasiewicz gradient inequality)

$$F \text{ is non-degenerate}$$
$$\Rightarrow$$
$$\exists C_0, C_1 > 0 \text{ s.t. } F(\mu) - F^* < C_0 \Rightarrow \|\nabla_{\widehat{W_2}} F[\mu]\|^2 \geq C_1(F(\mu) - F^*)$$

Corollary

If F is non-degenerate then there exists $C_0, C_1 > 0$ such that

$$F(\mu_0) - F^* \leq C_0 \quad \Rightarrow \quad F(\mu_t) - F^* \leq C_0 e^{-C_1 t}.$$

Proof.

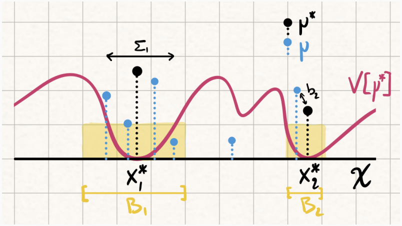
$$\frac{\mathrm{d}}{\mathrm{d}t}\big(F(\mu_t) - F^*\big) = -\|\nabla_{\widehat{W_2}} F[\mu_t]\|^2 \leq -C_1\big(F(\mu_t) - F^*\big)$$

and we conclude by integrating in time. □

Proof idea and local expansion

Decompose μ into local moments in small balls B_i around each x_i^*:

- local biases $b_i \in \mathbb{R}^{d+1}$
- local covariances $\Sigma_i \in \mathbb{R}^{d\times d}$



Local Taylor expansion of F around μ^*

$$F(\mu) - F^* \approx \underbrace{\frac{1}{2} b^\intercal (K+H) b}_{\text{Bias term (local+global)}} + \underbrace{\sum_{i=1}^{m^*} a_i \operatorname{tr}(\Sigma_i H_i)}_{\text{Variance term (local)}} + \underbrace{\int_{\mathcal{X}\setminus(\cup B_i)} V[\mu^*]\,\mathrm{d}\mu}_{\text{Mass sent to 0}}$$

Dynamics: Global convergence

Convergence with fixed grid ($\alpha = 0$)

Consider an infinitely dense grid. What are the convergence rates?

Proposition (Convergence rate, multiplicative updates)

Let $\mu_0 \propto \mathrm{vol}$ *and* $\partial_t \mu_t = -4\mu_t V[\mu_t]$. *It holds* $F(\mu_t) - F^* \lesssim \frac{\log(t)}{t}$.

- proof via mirror descent + approximation argument
- in practice discretization error quickly takes over
- compare with the L^2 gradient flow:

Proposition (Convergence rate, additive updates)

Let $\mu_0 \propto \mathrm{vol}$ *and* $\partial_t \mu_t = -V[\mu_t]\mathrm{vol}$. *If* F *is non-degenerate, then*

$$F(\mu_t) - F^* \asymp t^{-2/(d+2)}.$$

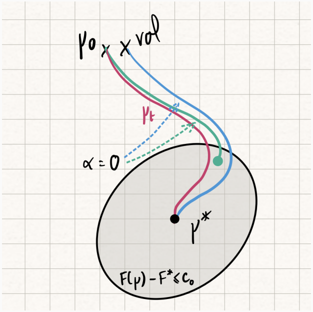
See [Chizat, 2021] for a complete analysis of convergence rates.

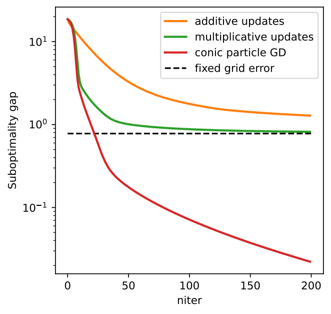
Global convergence

Theorem (Global convergence)

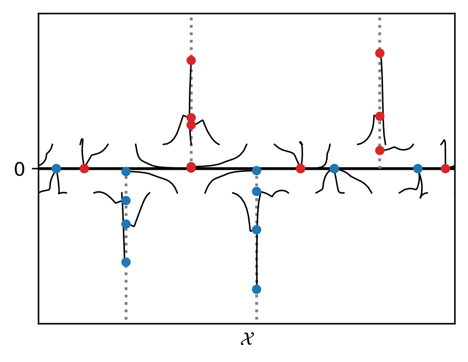
If the problem is *non-degenerate*, there exists $C_0', C_1' > 0$ such that

$$\begin{cases} \alpha \leq C_0' \\ \sup_{x\in\mathcal{X}} \inf_{i=1,\dots,m} \mathrm{dist}(x, x_i(0)) \leq C_1' \end{cases} \Rightarrow \quad \lim_{t\to\infty} F_m(\theta(t)) = F^*.$$





An illustration



Signed 1D spikes deconvolution: trajectory of μ_t

Concluding remarks

- **Extensions**
  We focused on GD but one could explore more advanced algorithms (pre-conditioning, acceleration, SGD)
- **Curse of dimensionality**
  The guarantees require $\exp(d)$ particles, which is unavoidable under our assumptions.
- **Can we change assumptions?**
  - dealing with the degenerate case (see [Zhou, Ge, Jin, 2021])
  - dealing with non-sparse minimizers (open)