EPFL

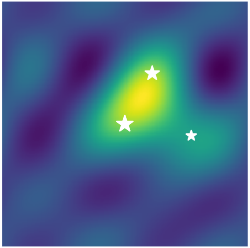
Sparse Optimization on Measures with Over-parameterized Gradient Descent

Lénaïc Chizat*

CANUM 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 $g \in L^2(\mathcal{X})$
- 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

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

NB: F_m is not convex and admits spurious local minima

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 $\mathcal{M}_+(\mathcal{X})$

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

Basic properties of F

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

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

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

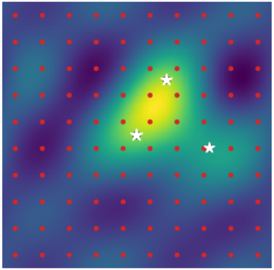
$\rightsquigarrow$ regularization by $\lambda\|\tilde{\mu}\|_{\mathrm{TV}}$ [De Castro & Gamboa, 2012]

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)

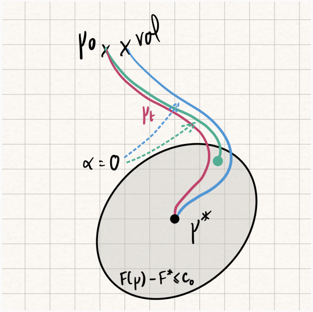
If the problem is *non-degenerate*, there exists $C_0', C_1' > 0$ such that

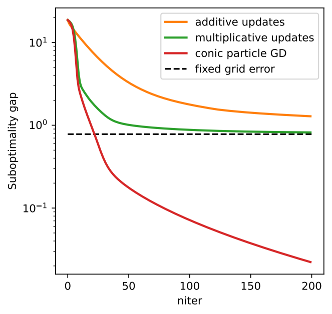
$$\begin{cases} \alpha \leq C_0' \\ \sup_{x\in\mathcal{X}} \inf_{i=1,\dots,m} \operatorname{dist}(x, x_i(0)) \leq C_1' \end{cases} \quad \Rightarrow \quad \lim_{t\to\infty} F_m(\theta(t)) = F^*.$$

$\rightsquigarrow$ These results are uniform in $m > 0$.

Chizat (2019). *Sparse Optimization on the Space of Measures with Over-parameterized Gradient Descent.*

Two-phase analysis





- **global phase**: convex approach, approximates $\alpha = 0$
- **local phase**: non-convex finish, exponential convergence

$\rightsquigarrow$ *this talk*: behavior of 1st order methods on (infinitely) thin grids

Sparsity and optimality

Assumption 1 (Uniqueness)

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

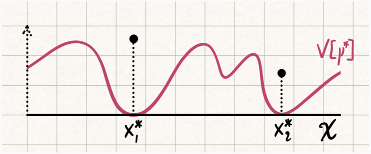
Let $V[\mu] \in \mathcal{C}^3(\mathcal{X})$ be the first variation of F at μ, characterized by

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

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 $H = \mathrm{diag}(H_i)_{i=1}^{m^*} \in \mathbb{R}^{(m^*(d+1))^2}$ 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]

Rates of Convex Optimization on Thin Grids

General framework & algorithms

- Fix ref. measure τ and pose $\mu = f\tau$ with $f \in L^1(\tau)$
- Minimize $F(f) = G(f) + H(f)$, G smooth and H prox-tractable
- Power entropy Bregman divergences $D_{\bar{\eta}}$, $\eta(s) = \begin{cases} s^p, & p \in]1,2] \\ s\log(s), & p = 1 \end{cases}$

Algorithm 1: (Bregman) Proximal Gradient Method (PGM)

Initialization: $f_0 \in \operatorname{dom} H$, *step-size* $s > 0$
for k=0,1,... **do**
  $f_{k+1} = \arg\min_f \left\{ G(f_k) + \int G'[f_k](f - f_k)\mathrm{d}\tau + H(f) + \frac{1}{s} D_{\bar{\eta}}(f, f_k) \right\}$
end
Output: f_{k+1}

Algorithm 2: Accelerated (Bregman) Proximal Gradient Method (APGM)

Initialization: $f_0 = h_0 \in \operatorname{dom} H$, $\gamma_0 = 1$, *step-size* $s > 0$
for k=0,1,... **do**
  $g_k = (1 - \gamma_k) f_k + \gamma_k h_k$
  $h_{k+1} = \arg\min_f \left\{ G(g_k) + \langle \nabla G(g_k), f - g_k \rangle + H(f) + \frac{\gamma_k}{s} D_{\bar{\eta}}(f, h_k) \right\}$
  $f_{k+1} = (1 - \gamma_k) f_k + \gamma_k h_{k+1}$
  $\gamma_{k+1} = \frac{1}{2}\left(\sqrt{\gamma_k^4 + 4\gamma_k^2} - \gamma_k^2 \right)$
end
Output: f_{k+1}

Paul Tseng (2010). *Approximation accuracy, gradient methods [...] for structured convex optimization.*

Known guaranties and How to use them

Theorem [Tseng, 2010, adapted]

For a small enough step-size s, if bounded iterates, it holds

$$F(f_k) - F(f) \leq \underbrace{\frac{4}{s(k+1)^{\beta}}}_{\xi_k} D_{\bar{\eta}}(f, f_0), \quad \forall f \in L^1(\tau), \forall k \geq 0$$

where $\beta = 1$ for PGM and $\beta = 2$ for APGM.

- **Problem**: $D_{\bar{\eta}}(f^*, f_0) = \infty$ (in fact $f^* \notin L^1(\tau)$)
- **Workaround**: use instead

$$F(f_k) - \inf F \leq \inf_{f \in L^1(\tau)} \Big(F(f) - \inf F \Big) + \xi_k D_{\bar{\eta}}(f, f_0)$$

Often used in the literature about (S)GD in Hilbert spaces...
Jacobs, Léger, Li, Osher (2019). *Solving large-scale optimization problems with a convergence rate [...].*

Main results

Convergence rates [Chizat' 2021]

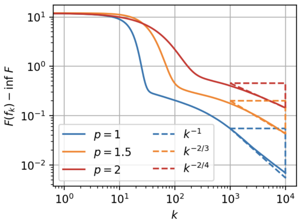
For non-degenerate sparse problems, (A)PGM satisfies

$$F(f_k) - \inf F \lesssim \begin{cases} k^{-\frac{2\beta}{(p-1)d+2}} & \text{if } p > 1 \\ \log(k) k^{-\beta} & \text{if } p = 1 \end{cases}$$

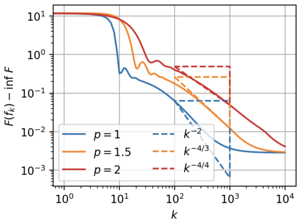
- rates are exact up to log factors (lower bounds)
- beyond non-degenerate cases: the rate depends on the structure at optimality (see paper)
- for signed problems: use hyperbolic entropy ($p = 1$)

Chizat (2021). *Convergence Rates of Gradient Methods for Convex Optimization in the Space of Measures*

Numerics



(a) PGM ($d = 2$, $q = 2$)



(b) APGM ($d = 2$, $q = 2$)

Observed vs. theoretical rates on a non-degenerate sparse $2D$ deconvolution problem

$\rightsquigarrow$ $p = 1$ (APGM with hyperbolic entropy) is one order of magnitude faster than $p = 2$ (FISTA) on a large range of accuracies!

Concluding remarks

- **Extensions**
  We focused on GD but one could explore more advanced algorithms (pre-conditioning, SGD)
- **Curse of dimensionality**
  The guarantees require $\exp(d)$ particles, which is unavoidable under our assumptions.