FastPart: Over-Parametrized Stochastic Gradient Descent for Sparse optimization on Measures

Yohann De Castro[1,4], Sébastien Gadat[2,4] and Clément Marteau[3]

[1] *École Centrale de Lyon, CNRS UMR 5208, Institut Camille Jordan, Écully, France.*
[2] *Université Toulouse 1 Capitole, Toulouse School of Economics, France.*
[3] *Univ. Claude Bernard, CNRS UMR 5208, Institut Camille Jordan, Villeurbane, France.*
[4] *Institut Universitaire de France (IUF)*

version as of September 4, 2025

Corresponding Author: Sébastien Gadat
Affiliation: Toulouse School of Economics
E-mail Address: sebastien.gadat@tse-fr.eu

Abstract

This paper presents a novel algorithm that leverages Stochastic Gradient Descent strategies in conjunction with Random Features to augment the scalability of Conic Particle Gradient Descent (CPGD) specifically tailored for solving sparse optimization problems on measures. By formulating the CPGD steps within a variational framework, we provide rigorous mathematical proofs demonstrating the following key findings: (i) The total variation norms of the solution measures along the descent trajectory remain bounded, ensuring stability and preventing undesirable divergence; (ii) We establish a global convergence guarantee with a convergence rate of $\mathcal{O}(\log(K)/\sqrt{K})$ over K iterations, showcasing the efficiency and effectiveness of our algorithm, (iii) Additionally, we analyse and establish local control over the first-order condition discrepancy, contributing to a deeper understanding of the algorithm's behaviour and reliability in practical applications.

1 Introduction

1.1 Convex programming for sparse optimization on measures

Convex optimization on the space of measures has gained attention during the past decade, *e.g.,* Bach and Chizat [2021], Chizat and Bach [2018], Chizat [2022], De Castro et al. [2021], Poon et al. [2021], De Castro and Gamboa [2012], Candès and Fernandez-Granda [2014] and references therein. It is also a popular field of investigation to derive global optimization methods (a.k.a. simulated annealing) through the embedding of $\mathbb{R}^d$ into the space of measures [Bolte et al., 2023, Miclo, 2023].

At his core, it can be viewed as a fruitful way of expressing many non-convex signal processing and machine learning tasks into a convex one, where one searches for an element of a Hilbert space $\mathbb{H}$ that can be described as a linear combination of a few, say $\bar{s}$, elements:

$$\bar{\boldsymbol{y}} = \sum_{j=1}^{\bar{s}} \bar{\omega}_j \varphi_{\bar{\boldsymbol{t}}_j} \,, \tag{1.1}$$

from a given parameterized set $\left\{\varphi_{\boldsymbol{t}} \,:\, \boldsymbol{t} \in \mathcal{X}\right\}$ where $\bar{\omega}_j \in \mathbb{R} \setminus \{0\}$ and $\mathcal{X}$ is a closed compact and convex set of $\mathbb{R}^d$. From now on, and throughout the remainder of this paper, we shall assume without loss of generality that $\mathcal{X} = \bar{B}(0, R_{\mathcal{X}})$, the closed centered ball in $\mathbb{R}^d$ with some radius $R_{\mathcal{X}} > 0$.

Given an empirical observation $\boldsymbol{y}$, we would like to find a sparse representation, akin to (1.1), that explains $\boldsymbol{y}$ and for which the learnt parameters $(\omega_j, \boldsymbol{t}_j)_{j=1}^s$ encode an output solution for which generalization properties can be proven. A common practice is to minimize:

$$(\omega_j, \boldsymbol{t}_j)_{j=1}^s \mapsto \Big\| \boldsymbol{y} - \sum_{j=1}^{s} \omega_j \varphi_{\boldsymbol{t}_j} \Big\|_{\mathbb{H}}^2 , \tag{1.2}$$

which is a non-convex program. In the expression (1.2) above, $s \geq 1$ is a tuning parameter quantifying the so called *sparsity* of the solution.

A substantial body of literature pertains to the minimization of Mean Squared Error (MSE) and aligns with the framework outlined herein. In this paper, we will expound upon our methodology in a generalized format that can effectively encompass a majority of fields. Notably, certain specific instances will be elaborated upon within this paper, including but not limited to sparse de-convolution [De Castro and Gamboa, 2012, Candès and Fernandez-Granda, 2014], infinitely wide neural networks [Bach and Chizat, 2021, Chizat and Bach, 2018], or Mixture Models [De Castro et al., 2021]. Our approach is deployed according to the following steps.

Lifting on the space of signed measures First, we lift Program (1.2) onto the space $(\mathcal{M}(\mathcal{X}), \|\cdot\|_{\mathrm{TV}})$ of Radon measures with finite total variation norm on $\mathcal{X}$, defined a the dual space of the continuous functions $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$. Consider the positive definite kernel $\mathbb{K}$ defined as the dot product $\mathbb{K}(\boldsymbol{t}, \boldsymbol{t}') := \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}'} \rangle_{\mathbb{H}}$ for all $t, t' \in \mathcal{X}$, and assume that it is continuous:

Assumption ($\mathbf{A}_{\mathrm{C}}$). *The feature map function*

$$\boldsymbol{t} \in \mathcal{X} \mapsto \varphi_{\boldsymbol{t}} \in \mathbb{H} \tag{$\mathbf{A}_{\mathrm{C}}$}$$

is continuous.

Consider the *kernel measure embedding* (we refer to Appendix A.2 for further details),

$$\Phi : \ \nu \in \mathcal{M}(\mathcal{X}) \mapsto \int_{\mathcal{X}} \varphi_{\boldsymbol{t}} \mathrm{d}\nu(\boldsymbol{t}) \in \mathbb{H} . \tag{1.3}$$

It is proven in the appendix (see Lemma A.2) that Φ is a bounded linear map under ($\mathbf{A}_{\mathrm{C}}$). We deduce that:

$$\nu \in \mathcal{M}(\mathcal{X}) \mapsto \big\| \boldsymbol{y} - \Phi(\nu) \big\|_{\mathbb{H}}^2 , \tag{1.4}$$

is a convex function on $\mathcal{M}(\mathcal{X})$. Taking the set of discrete measures given by $\nu = \sum_{j=1}^s \omega_j \delta_{\boldsymbol{t}_j}$, where $\delta_{\boldsymbol{t}}$ is the Dirac mass at point $\boldsymbol{t}$, we uncover the parametrization (1.2). Hence, we have lifted a non-convex program on $(\omega_j, \boldsymbol{t}_j)_{j=1}^s$ onto a convex program over a much larger space, the space of signed measures.

Total variation norm regularization The second step is regularization. One key parameter is s, the number of learnt parameters, that should be considered to estimate (an approximation of) the true function (1.1). In practice, it can be cumbersome to tune this parameter and it might be better to resort to regularization. One benefit of the lifting on the space of measures is that this can be simply done by the TV-norm. Inspired by L^1-regularization in (high-dimensional) inverse problems, we study the so-called *Beurling LASSO* [De Castro and Gamboa, 2012, Candès and Fernandez-Granda, 2014] referred to as BLASSO below, whose convex objective function is given by:

$$J(\nu) := \frac{1}{2} \big\| \boldsymbol{y} - \Phi(\nu) \big\|_{\mathbb{H}}^2 + \lambda \|\nu\|_{\mathrm{TV}} , \tag{1.5}$$

where $\lambda > 0$ is a tuning parameter. We denote by $\mu^\star \in \mathcal{M}(\mathcal{X})$ a solution to BLASSO

$$J(\mu^\star) = \min_{\mu \in \mathcal{M}(\mathcal{X})} J(\mu) . \tag{$\mathcal{B}$}$$

The existence of a solution to the problem at hand is not manifestly evident. Seminal contributions in the field, as articulated in [Bredies and Pikkarainen, 2013, Proposition 3.1] and [Hofmann et al., 2007, Theorem 3.1], have shown the existence of solutions upon continuity prerequisites imposed on the operator Φ or its pre-dual counterpart. Nevertheless, the arduousness associated with ascertaining the continuity and well-defined attributes of the operator Φ as expounded in Equation (1.3), within a given framework, underscores the complexity involved. In contrast, Condition ($\mathbf{A}_\mathrm{C}$) affords a more tractable means of validation. We present a result displayed in Theorem 1.1 below (established in Appendix A.3) that demonstrates the existence of solutions for any convex optimization problem formulated in the manner of Equation (1.6), subject to the condition of continuity stipulated in Equation ($\mathbf{A}_\mathrm{C}$). This result is aligned with a nice recent work [Bredies et al., 2024, Proposition 2.3] for general regularization terms which assume sequentially weak*-to-strong continuity of Φ. As shown in the proof (see Appendix A.3), we establish weak*-to-weak continuity for Φ when using the TV-norm regularization.

Theorem 1.1. *Let $\mathbb{H}$ be separable Hilbert space and let $\mathcal{X}$ be compact metric space. Consider the problem*

$$\inf_{\mu\in\mathcal{M}(\mathcal{X})}\Big\{L(\Phi\mu)+\lambda\|\mu\|_{\mathrm{TV}}\Big\} \tag{1.6}$$

where $L:\boldsymbol{h}\in\mathbb{H}\to L(\boldsymbol{h})\in[0,\infty]$ is convex and lower semi-continuous. If ($\mathbf{A}_\mathrm{C}$) holds then there exists a measure $\mu^\star\in\mathcal{M}(\mathcal{X})$ solution to (1.6). Furthermore, if L is strictly convex, then the vector $\Phi(\mu^\star)\in\mathbb{H}$ is unique (it does not depend on the choice of the solution $\mu^\star$).

Remark 1.1. *By choosing $L(\boldsymbol{h})=(1/2)\|\boldsymbol{y}-\boldsymbol{h}\|^2_{\mathbb{H}}$ in Theorem 1.1, it is established that there exists a signed measure $\mu^\star\in\mathcal{M}(\mathcal{X})$ solution to BLASSO ($\mathcal{B}$).*

Over the last decade, several investigations on the performances of the estimator associated to the solution of (1.6) have been proposed in several specific situations. This solution can be proven to be close, for some partial Wasserstein 2 distance [Poon et al., 2021], to the target measure $\bar\mu=\sum_{j=1}^{\bar s}\bar\omega_j\delta_{\bar t_j}$ involved in Equation (1.1) in some cases of interest (*e.g.,* Mixture Models [De Castro et al., 2021] or sparse deconvolution [Poon et al., 2021, De Castro and Gamboa, 2012]) as soon as the support points $\bar t_j$ of the target $\bar\mu$ are sufficiently separated. Moreover, if the bounded linear map Φ has finite rank $m\geq 1$, then there exists a solution to ($\mathcal{B}$) with at most m atoms, as proven by [Boyer et al., 2019, Section 4].

1.2 Conic Particle Gradient Descent (CPGD)

Solving ($\mathcal{B}$) from a practical point of view is not an immediate task, due to the infinite dimensional nature of the target. In this context, the Sliding Franck-Wolfe algorithm (see, e.g., Denoyelle et al. [2019]) provides an answer to this question. In this paper, we focus instead on the convergence of a Stochastic and Random Feature version of the Conic Particle Gradient Descent (CPGD) [Chizat, 2022] towards a minimum of Program ($\mathcal{B}$).

Writing the weights $\mathbb{W}:=(\omega_1,\dots,\omega_p)$ and the positions $\mathbb{T}:=(\boldsymbol{t}_1,\dots,\boldsymbol{t}_p)\in\mathcal{X}^p$, we consider a generic measure with p weighted particles by:

$$\nu(\mathbb{W},\mathbb{T}):=\sum_{j=1}^{p}\varepsilon_j\omega_j\delta_{\boldsymbol{t}_j}, \tag{1.7}$$

where $\omega_j>0$ (resp. $\varepsilon_j=\pm1$) refers to the weight (resp. the sign) of the particle j. The signs are fixed along the descent while the positions $\mathbb{T}$ and weights $\mathbb{W}$ are updated at each gradient step. By a symmetrization argument, see for instance [Chizat, 2022, Appendix A], we consider, without loss of generality, that $\varepsilon=1$. It holds that minimizing ($\mathcal{B}$) or minimizing J defined by:

$$J(\mu^\star)=\min_{\mu\in\mathcal{M}(\mathcal{X})_+}J(\mu)\,. \tag{$\mathcal{B}_+$}$$

are equivalent, in a sense made precise by [Chizat, 2022, Proposition A.1] for instance; where $\mathcal{M}(\mathcal{X})_+$ is the set of non-negative measures with finite TV-norm. The attentive reader can uncover the next results on $\mathcal{M}(\mathcal{X})$ by replacing ω_j by $\varepsilon_j\omega_j$. The gradient descent dynamics are the same and our results also holds in this latter case.

Our algorithm makes use of particles measures as a proxy for solving the problem ($\mathcal{B}$). To this end, we adapt the notation of objective functions and related quantities accordingly. Denoting $\boldsymbol{\lambda} := (\lambda, \ldots, \lambda)$, the definition of $\nu(\mathbb{W}, \mathbb{T})$ in (1.7) then yields

$$J(\nu(\mathbb{W}, \mathbb{T})) = \frac{1}{2}\Big\| \boldsymbol{y} - \sum_{j=1}^{p} \omega_j \varphi_{\boldsymbol{t}_j} \Big\|_{\mathbb{H}}^2 + \lambda \sum_{j=1}^{p} \omega_j := F(\mathbb{W}, \mathbb{T}) + \frac{1}{2}\|\boldsymbol{y}\|_{\mathbb{H}}^2 ,$$

where $\boldsymbol{k}_{\mathbb{T}} := (\langle \boldsymbol{y}, \varphi_{\boldsymbol{t}_1}\rangle_{\mathbb{H}}, \ldots, \langle \boldsymbol{y}, \varphi_{\boldsymbol{t}_p}\rangle_{\mathbb{H}}) \in \mathbb{R}^p$, $\mathbb{K}_{\mathbb{T}}$ is a $(p \times p)$ matrix with entries $\mathbb{K}(\boldsymbol{t}_i, \boldsymbol{t}_j)$ defined by:

$$\mathbb{K}(\boldsymbol{t}_i, \boldsymbol{t}_j) = \langle \varphi_{\boldsymbol{t}_i}, \varphi_{\boldsymbol{t}_j}\rangle_{\mathbb{H}}, \tag{1.8}$$

and

$$F(\mathbb{W}, \mathbb{T}) := \langle \boldsymbol{\lambda} - \boldsymbol{k}_{\mathbb{T}}, \mathbb{W}\rangle + \frac{1}{2}\mathbb{W}^T \mathbb{K}_{\mathbb{T}} \mathbb{W} , \tag{1.9}$$

is equal to $J(\nu(\mathbb{W}, \mathbb{T}))$ up to the additive constant term $(1/2)\|\boldsymbol{y}\|_{\mathbb{H}}^2$ (that only depends on the observations and not on the parameters of the measure we are optimizing on).

In contrast to the original problem presented in ($\mathcal{B}$), the optimization process now operates within a different domain. Instead of working within the space of measures, it focuses on particle measures with a set of fixed size p. This shift in perspective involves optimizing over both the positions $\mathbb{T}$ and weights $\mathbb{W}$. Although this adjustment serves to simplify the model's complexity to some extent, it introduces certain computational challenges. Firstly, for each pair of parameters $(\mathbb{W}, \mathbb{T})$, the computation of $F(\mathbb{W}, \mathbb{T})$ necessitates the evaluation of $\boldsymbol{k}_{\mathbb{T}}$ and $\mathbb{K}_{\mathbb{T}}$. Depending on the structure of the Hilbert space $\mathbb{H}$ and the associated scalar products, this computation can be time-consuming. The need to calculate these quantities at each iteration of a gradient descent algorithm can become problematic, especially when dealing with high-dimensional spaces (d being significant). Furthermore, considering a large number of particles during the optimization process, which is often essential, can substantially escalate the computational burden. This computational overhead needs to be carefully managed and optimized to ensure the efficiency of the optimization procedure. In this paper, we address these issues by presenting a novel algorithm that incorporates Stochastic Gradient Descent (SGD) iterations. To the best of our knowledge, this approach has not been previously explored in the context of sparse optimization within the domain of measures. We rigorously examine the properties of this algorithm, conducting a comprehensive investigation from both theoretical and practical perspectives.

The paper is organized as follows. Sections 1.3 and 1.4 present our stochastic algorithm and our main convergence results. Section 2 then describes the rationale behind the construction of the algorithm. Section 3 is illustrated by the example of mixture models, an important topic in unsupervised learning. A numerical illustration on some toy examples are discussed in Section 4. Proofs and related technical results are gathered in Section 5 and in Appendices A, B and C.

1.3 Stochastic & Random Feature CPGD (`FastPart`)

Iterative algorithms solving ($\mathcal{B}_+$) have already been at the core of theoretical investigations. We refer for instance to Chizat [2022] among others. The latter investigates an algorithm that requires some frequent calls to the gradient ∇F of the objective F defined in (1.9). This gradient can be related to the Fréchet differential function $\boldsymbol{t} \mapsto J'_\nu(\boldsymbol{t})$ of $J(.)$ at point $\nu \in \mathcal{M}(\mathcal{X})_+$ and its gradient $\nabla_{\boldsymbol{t}} J'_\nu$. The Fréchet differential J'_ν is defined through the following first order Taylor expansion:

$$\forall \nu \in \mathcal{M}(\mathcal{X})_+, \quad \nu + \sigma \in \mathcal{M}(\mathcal{X})_+ \qquad J(\nu + \sigma) - J(\nu) = \langle J'_\nu, \sigma\rangle_{\mathcal{M}(\mathcal{X})^\star, \mathcal{M}(\mathcal{X})} + q(\sigma) , \tag{1.10}$$

where J'_ν is the Fréchet gradient and q is a second order term. We refer to Proposition B.1 for details. According to Proposition B.3, for any $\boldsymbol{t} \in \mathcal{X}$, we have by Equation (B.3) that, given $\nu = \sum_{j=1}^{p} \omega_j \delta_{\boldsymbol{t}_j}$,

$$J'_\nu(\boldsymbol{t}) = \sum_{j=1}^{p} \omega_j \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}_j}\rangle_{\mathbb{H}} - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda . \tag{1.11}$$

The computation of these functions is time-consuming for the three reasons listed below. For each of these challenges, we outline our strategy to address them, and we introduce three notation—namely, the three random variables U, V, and T—which will be elaborated upon in the subsequent section. In Section 3.1, we will provide a concrete example illustrating this phenomenon.

Kernel evaluation: random variable U for Random feature strategies Firstly, it is important to note that these functions may necessitate integral approximations due to their lack of closed-form expressions. For instance, the computation of $J'_\nu(\boldsymbol{t})$ within each iteration of a gradient descent algorithm involves multiple evaluations of the kernel $\mathbb{K}(\boldsymbol{t},\boldsymbol{t}')$, defined in Equation (1.8). In many cases, these evaluations rely on non-explicit integrals, posing computational challenges. To circumvent this issue, we will present two strategies: random Fourier feature and convolution by a non-negative function. These strategies aim to approximate the kernel $\mathbb{K}$ by using a low-rank random kernel, achieved by evaluating the integral defining the kernel through independent Monte Carlo sampling. We need the following assumption, which is standard in the theory of Reproducing Kernel Hilbert Spaces (RKHS).

Assumption ($\mathbf{A}_{\text{T.I.}}$)**.** *We assume that the kernel $\mathbb{K}$ is translation invariant:*

$$\forall \boldsymbol{t},\boldsymbol{t}' \in \mathcal{X}\,,\quad \mathbb{K}(\boldsymbol{t},\boldsymbol{t}') = k(\boldsymbol{t}-\boldsymbol{t}')\,, \tag{$\mathbf{A}_{\text{T.I.}}$}$$

for some function $k(\cdot)$.

- **Random Fourier feature strategy** By Bochner's theorem to decompose it into its spectral form and use a Monte Carlo approximation as follows:

$$\mathbb{K}(\boldsymbol{t},\boldsymbol{t}') = k(\boldsymbol{t}-\boldsymbol{t}') = k(0)\int e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}-\boldsymbol{t}'\rangle}\mathrm{d}\sigma(\boldsymbol{u}) \simeq \frac{k(0)}{m}\sum_{j=1}^{m} \mathcal{R}\mathrm{e}\big(e^{-\mathrm{i}\langle U_j,\boldsymbol{t}\rangle}\overline{e^{-\mathrm{i}\langle U_j,\boldsymbol{t}'\rangle}}\big) = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{g}_{\boldsymbol{t},\boldsymbol{t}'}(U_j)\,,$$

where $\mathcal{R}\mathrm{e}$ denotes the real part of a complex, σ is the spectral probability measure associated to the kernel, U_i are i.i.d. with law σ and

$$\boldsymbol{g}_{\boldsymbol{t},\boldsymbol{t}'}(\boldsymbol{u}) := \mathcal{R}\mathrm{e}\big(e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}\rangle}\overline{e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}'\rangle}}\big)\,,$$

the bar denoting complex conjugation. Hence, the $(p\times p)$ matrix $\mathbb{K}_{\mathbb{T}}$ can be approximated by the following explicit rank m approximation

$$\mathbb{K}_{\mathbb{T}} \simeq \mathbb{U}_{\mathbb{T}}\mathbb{U}_{\mathbb{T}}^{\star} \quad \text{where} \quad \mathbb{U}_{\mathbb{T}} := \sqrt{\frac{k(0)}{m}}\big(e^{-\mathrm{i}\langle U_j,\boldsymbol{t}_i\rangle}\big)_{\substack{1\leq i\leq p\\ 1\leq j\leq m}}\,.$$

- **Convolution by non-negative function strategy** A second way is to consider that $k(\cdot)$ can be written as a convolution $k=\tilde{k}\star\sigma$ for some non-negative integrable function σ, say a probability density function without loss of generality. Invoke a Monte Carlo approximation as follows:

$$\mathbb{K}(\boldsymbol{t},\boldsymbol{t}') = (\tilde{k}\star\sigma)(\boldsymbol{t}-\boldsymbol{t}') = \int_{\mathbb{R}^d} \tilde{k}(\boldsymbol{t}-\boldsymbol{t}'-\boldsymbol{u})\sigma(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \simeq \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{g}_{\boldsymbol{t},\boldsymbol{t}'}(U_j)\,,$$

where U_i are i.i.d. with law σ and

$$\boldsymbol{g}_{\boldsymbol{t},\boldsymbol{t}'}(\boldsymbol{u}) := \tilde{k}(\boldsymbol{t}-\boldsymbol{t}'-\boldsymbol{u})\,, \tag{1.12}$$

to get an approximation of the kernel. This strategy can be used in Mixture Models (MM) as shown in (3.2). As a side note, it is not hard to see that if $\tilde{k}(\cdot)$ is a positive type function (*i.e.*, it defines a RKHS kernel) so is $k(\cdot)$.

Large samples: random variable V for picking a data sample at random A second strategy is to employ stochastic gradient computation with batch sub-sampling, which entails selecting a single data point at random–a fundamental component of various stochastic algorithms. Within our framework, this can be realized by utilizing the observation vector $\boldsymbol{y}$. It is worth emphasizing that in specific scenarios, the observed signal $\boldsymbol{y}$ can itself be a random variable. For instance, consider the case where

$$\boldsymbol{y} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i \quad \text{and} \quad \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} = \boldsymbol{h}_{\boldsymbol{t}}(V)\,, \tag{1.13}$$

consisting of N i.i.d. samples $\boldsymbol{y}_i$, V uniformly distributed over $\{1, \dots, N\}$ and

$$\boldsymbol{h}_{\boldsymbol{t}}(\boldsymbol{v}) := \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}_{\boldsymbol{v}}\rangle_{\mathbb{H}}\,.$$

In such cases, it becomes feasible to generate an unbiased stochastic version of $\boldsymbol{y}$ by randomly selecting a single or a mini-batch data sample. Moreover, even if $\boldsymbol{y}$ is deterministic or does not match the aforementioned identity (1.13), it is always possible to employ a similar strategy as described above (Random Fourier feature or convolution by a probability density function) to approximate $\langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}}$ itself. Certainly, note that the unbiased stochastic version of $\boldsymbol{y}$ and random kernel approximation of $\langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}}$ should be jointly considered whenever possible.

Many particles: random variable T for picking a particle at random The necessity for a large number of particles to attain convergence guarantees increases exponentially with the dimension d of $\mathcal{X}$. All these particles are simultaneously involved in the evaluation of gradient terms. As a final ingredient, we opt to use only a particle or a mini-batch of particles, selected randomly, to evaluate these terms at each step of the algorithm. The time complexity of gradient computations can be diminished by a factor p (number of particles) as one can see using (1.11) where the sum over p particles is replaced by one evaluation at a random particle's location:

$$J'_\nu(\boldsymbol{t}) = \sum_{j=1}^{p} \omega_j \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}_j}\rangle_{\mathbb{H}} - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda = \|\nu\|_{\mathrm{TV}}\mathbb{E}\big[\langle \varphi_{\boldsymbol{t}}, \varphi_T\rangle_{\mathbb{H}}\big] - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda\,,$$

with $T \sim \nu/\|\nu\|_{\mathrm{TV}}$. Instead of choosing one particle at random, one can also pick a mini-batch of particles at random, say m_T of them using, for instance the law $T = (T_1, \dots, T_{m_T}) \sim \big(\nu/\|\nu\|_{\mathrm{TV}}\big)^{\otimes m_T}$ with the gradient rule based on

$$J'_\nu(\boldsymbol{t}) = \sum_{j=1}^{p} \omega_j \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}_j}\rangle_{\mathbb{H}} - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda = \frac{\|\nu\|_{\mathrm{TV}}}{m_T}\mathbb{E}\sum_{i=1}^{m_T}\big[\langle \varphi_{\boldsymbol{t}}, \varphi_{T_i}\rangle_{\mathbb{H}}\big] - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda\,.$$

1.3.1 Stochastic & Random Feature approximations of the gradient

We provide in this paragraph a general framework allowing the construction of the stochastic conic gradient particle algorithm we are studying. Specific examples are discussed in Section 3 below.

The formulation of a stochastic approximation for the gradient ∇F necessitates certain general assumptions, which are outlined below. Of particular significance is the introduction of a random variable $Z = (T, U, V)$, which serves as a way to alleviate the computational burden.

Assumption ($\mathbf{A}_1$). *There exists a pair of random variables (U, V) (not necessarily independent) such that, for any $\boldsymbol{t}, \boldsymbol{t}' \in \mathcal{X}$,*

$$\langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}'}\rangle_{\mathbb{H}} = \mathbb{E}_U \boldsymbol{g}_{\boldsymbol{t},\boldsymbol{t}'}(U) \quad \text{and} \quad \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} = \mathbb{E}_V \boldsymbol{h}_{\boldsymbol{t}}(V)\,, \tag{$\mathbf{A}_1$}$$

for some explicit **bounded** *functions* $\boldsymbol{g}$ *and* $\boldsymbol{h}$.

The latter assumption allows for a stochastic approximation of the functional J'_ν. It exactly corresponds to what happens in Equation (3.8) below for the mixture model. Indeed, if we consider from now on the random variable T with distribution $\nu/\nu(\mathcal{X})$, sampled independently from (U, V), and introduce:

$$\boldsymbol{J}'_\nu(\boldsymbol{t}, Z) := \|\nu\|_{\mathrm{TV}}\, \boldsymbol{g}_{\boldsymbol{t},T}(U) - \boldsymbol{h}_{\boldsymbol{t}}(V) + \lambda \quad \text{where} \quad Z := (T, U, V)\,. \tag{1.14}$$

According to ($\mathbf{A}_1$), we can write that

$$\boldsymbol{J}'_\nu(\boldsymbol{t}, Z) := J'_\nu(\boldsymbol{t}) + \xi_\nu(\boldsymbol{t}, Z) \quad \text{with} \quad \mathbb{E}_Z \xi_\nu(\boldsymbol{t}, Z) = 0 \quad \forall \boldsymbol{t} \in \mathcal{X}\,. \tag{1.15}$$

We can construct a similar stochastic approximation for the gradient (w.r.t. the position parameter) of the functional J'_ν. First remark that

$$\nabla_{\boldsymbol{t}} J'_\nu(\boldsymbol{t}) = \sum_{j=1}^{p} \omega_j \nabla_{\boldsymbol{t}} \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}_j} \rangle_{\mathbb{H}} - \nabla_{\boldsymbol{t}} \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y} \rangle_{\mathbb{H}}\,.$$

The following assumption allows the commutativity between derivation and expectation and is satisfied in many situations, including batch or mini-batch strategies and smooth integral computations. The associated term in the mixture model is (3.9) and its stochastic counterpart is (3.10).

Assumption ($\mathbf{A}_2$). *The couple* (U, V) *and the functions* $\boldsymbol{g}, \boldsymbol{h}$ *introduced in Assumption* ($\mathbf{A}_1$) *satisfy*

$$\nabla_{\boldsymbol{t}} \mathbb{E}_U \boldsymbol{g}_{\boldsymbol{t},\boldsymbol{t}'}(U) = \mathbb{E}_U \nabla_{\boldsymbol{t}} \boldsymbol{g}_{\boldsymbol{t},\boldsymbol{t}'}(U) \quad \text{and} \quad \nabla_{\boldsymbol{t}} \mathbb{E}_V \boldsymbol{h}_{\boldsymbol{t}}(V) = \mathbb{E}_V \nabla_{\boldsymbol{t}} \boldsymbol{h}_{\boldsymbol{t}}(V)\,, \tag{$\mathbf{A}_2$}$$

for any $\boldsymbol{t}, \boldsymbol{t}' \in \mathcal{X}$ and $\boldsymbol{g}$ and $\boldsymbol{h}$ have **bounded** derivatives.

Then, introduce

$$\boldsymbol{D}_\nu(\boldsymbol{t}, Z) := \|\nu\|_{\mathrm{TV}}\, \nabla_{\boldsymbol{t}} \boldsymbol{g}_{\boldsymbol{t},T}(U) - \nabla_{\boldsymbol{t}} \boldsymbol{h}_{\boldsymbol{t}}(V)\,, \tag{1.16}$$

where $Z = (T, U, V)$ still denotes the same variable as above. We can then observe that

$$\boldsymbol{D}_\nu(\boldsymbol{t}, Z) := \nabla_{\boldsymbol{t}} J'_\nu(\boldsymbol{t}) + \zeta_\nu(\boldsymbol{t}, Z) \quad \text{with} \quad \mathbb{E}_Z \zeta_\nu(\boldsymbol{t}, Z) = 0 \quad \forall \boldsymbol{t} \in \mathcal{X}\,, \tag{1.17}$$

where we have used Assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_2$).

Remark 1.2. *Lemma B.1 in Appendix provides a explicit and uniform upper bound on* J'_ν *for any measure* $\nu \in \mathcal{M}(\mathcal{X})_+$. *From this remark, the boundedness of the functions* $\boldsymbol{g}$ *and* $\boldsymbol{f}$ *is indeed a reasonable assumption within this context. Furthermore, in addition to the conclusions presented in Equations* (1.15) *and* (1.17), *that establish unbiased estimations of* J'_ν *and* $\nabla_{\boldsymbol{t}} J'_\nu$, *it is imperative to highlight that our assumptions* ($\mathbf{A}_1$) *and* ($\mathbf{A}_2$) *also result in almost sure upper bounds on* $|\boldsymbol{J}'_\nu(., Z)|$ *and* $\|\mathbf{D}_\nu(., Z)\|$ *(for more comprehensive details, we refer to Lemma B.1 and Proposition C.1).*

Remark 1.3. *Assumption* ($\mathbf{A}_2$) *implicitly assumes that the left hand side gradient exists. This is the case when the kernel is continuously differentiable, see [Steinwart and Christmann, 2008, Definition 4.35] for a proper definition. This later property is equivalent to the continuous differentiability of the feature map, see [Steinwart and Christmann, 2008, Lemma 4.34].*

1.3.2 Mini-batch and variance reduction

Introducing randomness in the CPGD algorithm creates some issues in terms of convergence. It is indeed necessary to control the fluctuation of the random terms $\boldsymbol{J}'_\nu(., Z)$ and $\boldsymbol{D}_\nu(., Z)$ w.r.t. their deterministic counterparts at each iteration. Assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_2$) already provides some tools in this direction: it allow to obtain almost sure bounds on these quantities. Nevertheless, our analysis will require slightly stronger results and in particular some control on the variance of theses stochastic approximations. To this end, we introduce in this section a mini-batch step.

The principle is to draw a m-sample of i.i.d. random variables $Z_1, \dots, Z_m$ having the same law than Z introduced in (1.14). The term $m \in \mathbb{N}^*$ denotes the mini-batch sample size and might depends on the iteration step. Then, we introduce, for any $\nu \in \mathcal{M}_+(\mathcal{X})$ and $\boldsymbol{t} \in \mathcal{X}$

$$\widehat{\boldsymbol{J}'_\nu}(\boldsymbol{t}) := \frac{1}{m} \sum_{l=1}^{m} \boldsymbol{J}'_\nu(\boldsymbol{t}, Z_l) \quad \text{and} \quad \widehat{\boldsymbol{D}_\nu}(\boldsymbol{t}) := \frac{1}{m} \sum_{l=1}^{m} \boldsymbol{D}_\nu(\boldsymbol{t}, Z_l). \tag{1.18}$$

The terms $\widehat{\boldsymbol{J}'_\nu}(.)$ and $\widehat{\boldsymbol{D}_\nu}(.)$ simply correspond to averaged version of the stochastic approximations introduced in the previous section. Their second order moments are controlled by the mini-batch size m. To alleviate notation, we do not highlight the dependency of the average terms introduced in (1.18) with respect to the value of m. This dependency will be clear following the context.

1.3.3 Stochastic gradient updates

Both weights and position updates are based on the CPGD principle which will be discussed in details in Section 2 below. The weights optimization is performed employing an exponential weights update on a per-particle basis:

$$\forall j \in \{1, \dots, p\}, \qquad \omega_j^{k+1} = \omega_j^k e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}. \tag{1.19}$$

In a similar way, the positions updates satisfies, for any $j \in \{1, \dots, p\}$

$$\begin{aligned} \boldsymbol{t}_j^{k+1} &= \arg\min_{t \in \mathcal{X}} \left\{ \langle \boldsymbol{t}, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k))_j \rangle + \frac{1}{2\eta} \|\boldsymbol{t} - \boldsymbol{t}_k^j\|^2 \right\} \\ &= \arg\min_{t \in \mathcal{X}} \left\{ \left\| \boldsymbol{t} - (\boldsymbol{t}_j^k - \eta \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k))_j) \right\|^2 \right\} \end{aligned} \tag{1.20}$$

which leads to projected gradient descent, after simplifying the similar weighting terms appearing on the one hand in the gradient, and on the other hand the conic metric. This update is then given as follows:

$$\forall j \in \{1, \dots, p\}, \qquad \boldsymbol{t}_j^{k+1} = \pi_{\mathcal{X}} \left(\boldsymbol{t}_j^k - \eta \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) \right), \tag{1.21}$$

where $\pi_{\mathcal{X}}$ denotes the projection operator over $\mathcal{X}$. Recall that $\mathcal{X} = \bar{B}(0, R_{\mathcal{X}})$ for some radius $R_{\mathcal{X}} > 0$ and, in this setting, the projection $\pi_{\mathcal{X}}$ is uniquely defined and may be easily computed as:

$$\forall u \in \mathbb{R}^d \qquad \pi_{\mathcal{X}}(u) = \frac{R_{\mathcal{X}}}{\|u\|} u \mathbf{1}_{\|u\| > R_{\mathcal{X}}} + u \mathbf{1}_{\|u\| \leq R_{\mathcal{X}}}.$$

We refer again to Section 2 for a complete presentation of the rationale behind this strategy.

Algorithm 1 below summarizes our stochastic optimization strategy that uses unbiased stochastic realizations, as outlined in the previous assumptions. At each iteration k, the elements Z_l^{k+1} of the mini-batch sample are made of i.i.d. random vectors $(T_\ell^{k+1}, U_\ell^{k+1}, V_\ell^{k+1})$ where the $(U_\ell^{k+1}, V_\ell^{k+1})$ have the same distribution as (U, V) introduced in Assumption $(\mathbf{A}_1)$, and $T_l^{k+1} \sim \nu_k / \nu_k(\mathcal{X})$.

Remark 1.4 (Importance of the projection step)**.** *The projection operator $\pi_{\mathcal{X}}$ appearing in the position updates (1.21) constraints the particles to stay on the set $\mathcal{X}$ along the iterations. In the deterministic version of the algorithm (namely by using directly $\boldsymbol{J}'_{\nu_k}(\boldsymbol{t}_j^k)$ and $\boldsymbol{D}_{\nu_k}(\boldsymbol{t}_j^k)$ in Step 5 of Algorithm 1), this projection step appears to be necessary only for the global convergence results (analog of Theorem 1.2) while control of the TV-norm (Proposition 1.1) along with local convergence (Theorem 1.3) still hold. Some issues arise when introducing randomness during the gradient descent. In particular, the introduction of the variable T in Section 1.3.1 does not allow to manage a control of $\|\nu_k\|_{\mathrm{TV}}$ along the iteration. In particular, the terms $\langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{s}} \rangle$ is no more lower bounded when $\boldsymbol{t}$ and $\boldsymbol{s}$ do not necessarily belong to $\mathcal{X}$. We stress that these issues may correspond to an artefact effect related to our proof techniques. We can indeed check that for a particle whose location $\boldsymbol{t}_j^k$ is far away from $\mathcal{X}$ (in a sense to made precise) at step k, then $\omega_j^{k+1} \leq \omega_j^k e^{-\lambda/2}$. From an heuristic point of view, such a particle should not disturb the convergence process after few steps.*

Algorithm 1 Stochastic & Random Feature Conic Particle Gradient Descent (`FastPart`)

```
Require: Learning rates $(\alpha, \eta)$, Initialization $(\mathbb{W}^0, \mathbb{T}^0)$, Projection radius $R_{\mathcal{X}}$, Mini-batch size $(m_k)_{k\geq 1}$;
Weights: $\mathbb{W}^k$ and Positions: $\mathbb{T}^k$;
for $k = 1, \dots, K$ do                                                          ▷ $K$ gradient steps
    Set $\nu_k \longleftarrow \nu(\mathbb{W}^k, \mathbb{T}^k)$;                    ▷ Particles
    Sample
        $(Z_\ell^{k+1})_{1\leq \ell \leq m_k} \longleftarrow ((T_\ell^{k+1}, U_\ell^{k+1}, V_\ell^{k+1}))_{1\leq \ell\leq m_k}$ where $T_l^{k+1} \sim \nu_k/\nu_k(\mathcal{X})$;
                                                                 ▷ Stochastic mini-batch variables;
    Use Equations (1.14) and (1.16) and compute
        $\widehat{J'_{\nu_k}}(\boldsymbol{t}_j^k) := \frac{1}{m_k}\sum_{\ell=1}^{m_k} J'_{\nu_k}(\boldsymbol{t}_j^k, Z_\ell^{k+1})$ and $\widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) := \frac{1}{m_k}\sum_{\ell=1}^{m_k} \boldsymbol{D}_{\nu_k}(\boldsymbol{t}_j^k, Z_\ell^{k+1})$;
    Update the weights and the positions with Equations (1.19) and (1.21):
        $\forall j \in \{1, \dots, p\}, \quad \omega_j^{k+1} = \omega_j^k e^{-\alpha \widehat{J'_{\nu_k}}(\boldsymbol{t}_j^k)}$ and $\boldsymbol{t}_j^{k+1} = \pi_{\mathcal{X}}\left(\boldsymbol{t}_j^k - \eta \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)\right)$,   (1.22)
                                                                 ▷ Projected Mirror descent
end for
```

1.4 Main results

We denote by ν_k the measure of particles produced at step k by Algorithm 1. Our main contributions are threefold:

- A control on the total-variation norm of the measures ν_k along the different iterations;
- A global minimization result;
- A local investigation on the evolution of J'_{ν_k} and its gradient.

We emphasize that these results are stated in a finite horizon setting and are therefore non-asymptotic in terms of K.

Here and below, we will require additional notation. We set

$$\|\underline{\varphi}\|_{\mathbb{H}} := \inf_{s,t\in\mathcal{X}} \langle \varphi_t, \varphi_s\rangle_{\mathbb{H}} \qquad \|\mathbf{g}\|_{\mathrm{Inf}} := \operatorname*{ess\,inf}_{s,t\in\mathcal{X}} \boldsymbol{g}_{t,s}$$

$$\|\boldsymbol{g}\|_\infty := \operatorname*{ess\,sup}_{s,t\in\mathcal{X}} |\boldsymbol{g}_{s,t}| \qquad \|\boldsymbol{h}\|_\infty := \operatorname*{ess\,sup}_{t\in\mathcal{X},v} |\boldsymbol{h}_t|$$

where the functions $\boldsymbol{g}$ and $\boldsymbol{h}$ have been introduced in ($\mathbf{A}_1$) and the essential infimum and supremum are taken with respect to the probability space defining the random variable $Z = (T, U, V)$. Furthermore, Assumption ($\mathbf{A}_{\mathrm{T.I.}}$) implies the following bound

$$\forall t \in \mathcal{X}, \quad \|\varphi_{\boldsymbol{t}}\|_{\mathbb{H}}^2 = \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}}\rangle_{\mathbb{H}} = k(0).$$

and we will use below the notation $\|\varphi\|_{\infty,\mathbb{H}} := \sqrt{k(0)}$ to refer to the norm of any element $\varphi_{\boldsymbol{t}}$ in $\mathbb{H}$. We also define $\|\varphi\|_{\mathrm{Lip}}$ as the Lipschitz constant associated to φ, namely

$$\|\varphi\|_{\mathrm{Lip}} = \sup_{s\neq t\in\mathcal{X}} \frac{\|\varphi_t - \varphi_s\|_{\mathbb{H}}}{\|s - t\|}.$$

1.4.1 Boundedness of the sequence

The next result establishes a preliminary upper bound of the total variation norm of $(\nu_k)_{k\geq 1}$ that holds *uniformly* over the iterations. It will be the key to obtain the convergence towards minimizers.

Proposition 1.1. *Assume* ($\mathbf{A}_1$) *and let* $\mathbf{g}$ *and* $\mathbf{h}$ *be the functions appearing in this assumption. Assume that* $\alpha \leq 1$, *that* $\|\boldsymbol{g}\|_{\mathrm{Inf}} > 0$ *and define* r_0 *as:*

$$r_0 := \frac{\max(0, \|\mathbf{h}\|_\infty - \lambda)}{\|\mathbf{g}\|_{\mathrm{Inf}}} e^{\max(0, \|\mathrm{h}\|_\infty - \lambda)} \tag{1.23}$$

Then, for any $k \in \mathbb{N}$, *we have*

$$\|\nu_k\|_{\mathrm{TV}} \leq R_0 := \|\nu_0\|_{\mathrm{TV}} \vee (p\, r_0). \tag{1.24}$$

Provided the mass of the measure ν_0 at the initialization step is not too large, $\|\nu_k\|_{\mathrm{TV}}$ remains bounded along the iterations of the algorithm. We stress that the control is deterministic although we consider a stochastic algorithm. The main ingredient of the proof is to take advantage of the relationship between ω_j^k and J'_{ν_k} (together with its stochastic counterpart). The update displayed in Algorithm 1 then allows to conclude. The complete proof is postponed to Section 5.1.

Remark 1.5. *Note that* $|\langle \boldsymbol{y}, \varphi_{\boldsymbol{t}} \rangle| \leq \|\boldsymbol{h}\|_\infty$ *and that the KKT condition reads*

$$\forall \boldsymbol{t} \in \mathcal{X}\,, \quad J'_{\mu^\star}(\boldsymbol{t}) = \sum_{j=1}^{p^\star} \omega_j^\star \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}_j^\star} \rangle_{\mathbb{H}} - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y} \rangle_{\mathbb{H}} + \lambda \geq 0$$

We deduce that if $\lambda \geq \|\boldsymbol{h}\|_\infty$ *then the KKT conditions are satisfied by the null solution* ($\mu^\star = 0$), *which is the unique solution to* ($\mathcal{B}_+$). *Hence, the condition* $r_0 = 0$ *implies that the null measure is solution. Proposition 1.1 is informative when* $r_0 > 0$ *which occurs as soon as there exists a non-zero solution.*

Remark 1.6. *The assumption* $\|\boldsymbol{g}\|_{\mathrm{Inf}} > 0$ *appears to be quite reasonable as soon as* $\|\underline{\varphi}\|_{\mathbb{H}}>0$*. For any* $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{X}$, *the variable* $\boldsymbol{g}_{\boldsymbol{s},\boldsymbol{t}}(U)$ *is indeed a stochastic approximation of* $\langle \varphi_s, \varphi_t \rangle_{\mathbb{H}}$. *We get that* $\|\boldsymbol{g}\|_{\mathrm{Inf}} > 0$ *with common assumptions on the construction of this approximation.*

In the case of an approximation by convolution, displayed in (1.12), *note that if* $\inf_{\boldsymbol{t},\boldsymbol{t}',\boldsymbol{u}} \tilde{k}(\boldsymbol{t} - \boldsymbol{t}' - \boldsymbol{u}) > 0$ *then the assumption* $\|\boldsymbol{g}\|_{\mathrm{Inf}} > 0$ *is satisfied. For instance, if* $\tilde{k}(\cdot)$ *is a Gaussian density and* $\boldsymbol{u} \mapsto \sigma(\boldsymbol{u})$ *has compact support then* $\inf_{\boldsymbol{t},\boldsymbol{t}',\boldsymbol{u}} \tilde{k}(\boldsymbol{t} - \boldsymbol{t}' - \boldsymbol{u}) > 0$ *as* $\boldsymbol{t}, \boldsymbol{t}' \in \mathcal{X}$ *are also bounded.*

1.4.2 Global minimization with projected swarm stochastic optimization

Consider $\mu^\star$ a measure that *globally* minimizes J, obtained from Theorem 1.1. The aim of this section is to show that our algorithm can produce a solution close to $\mu^\star$ (in a sense made precise in Theorem 1.2 below), under some specific conditions. Given a fixed number K of iterations of Algorithm 1, we define the Cesàro average of our sequence $(\nu_k)_{k \geq 0}$ by:

$$\bar{\nu}_K = \frac{1}{K+1} \sum_{k=0}^{K} \nu_k\,. \tag{1.25}$$

We then obtain the following global minimization result whose proof is displayed in Section 5.2.

Theorem 1.2. *Assume that* $\mu^\star$ *has a support contained in* $\mathcal{X}$. *Consider an integer* K *and the sequence of* $(\nu_k)_{1 \leq k \leq K}$ *defined in Algorithm 1. We set the learning rates as:*

$$\alpha = \sqrt{\frac{d\|\mu^\star\|_{\mathrm{TV}}}{R_0^3 K}} \quad \text{and} \quad \eta = \sqrt{\frac{d R_0}{K^3 \|\mu^\star\|_{\mathrm{TV}}}}\,,$$

where R_0 *is introduced in* (1.24). *Assume that the measure* ν_0 *is uniformly distributed over a uniform grid of step-size* $\delta = 2\sqrt{\frac{d}{\|\mu^\star\|_{\mathrm{TV}} KM}}$, *then:*

$$\mathbb{E}\left[J(\bar{\nu}_K) - J(\mu^\star)\right] \leq \mathfrak{C}\sqrt{\frac{d\|\mu^\star\|_{\mathrm{TV}} R_0^3}{K}} \left[\log(d\|\mu^\star\|_{\mathrm{TV}} R_0^3 K) + \frac{\log(|\mathcal{X}|)}{d}\right]\,,$$

for some positive constant $\mathfrak{C}$ *depending only on* $\|\varphi\|_{\infty,\mathbb{H}}$, $\|\boldsymbol{y}\|_{\mathbb{H}}$, $\|\mathbf{g}'\|_{\mathbb{H}}$, $\|\boldsymbol{h}'\|_{\mathbb{H}}$.

A careful inspection of the previous upper bound shows that to obtain an ϵ approximation with our Cesàro averaged measure $\bar{\nu}_K$, (while removing the effect of the log term) we need to choose K as:

$$K_\varepsilon = dR_0^3 \|\mu^\star\|_{\mathrm{TV}} \varepsilon^{-2}.$$

Then, the grid step-size is then of the order δ_ε given by:

$$\delta_\varepsilon = \frac{\varepsilon}{\|\mu^\star\|_{\mathrm{TV}}}.$$

We finally observe that the number of particles p needed to obtain an ε approximation is then of the order:

$$p_\varepsilon = |\mathcal{X}| \|\mu^\star\|_{\mathrm{TV}}^d \varepsilon^{-d}.$$

Hence, if the number of iteration varies polynomially in terms of ε^{-2}, we observe the degradation of the number of particles needed to well approximate any distribution over $\mathcal{X}$ in terms of the dimension d.

Remark 1.7. *Results displayed in Theorem 1.2 (together with Proposition 1.1) hold for every possible value for the mini-batch sample size $(m_k)_{k\in\mathbb{N}^*}$. In particular, one can set $m_k = 1$ for any $k \in \{1, \ldots, K\}$, which corresponds to the case where a single variable Z^{k+1} is drawn at each step. Indeed, the control on the TV-norm and the global convergence results only require an almost sure bounded stochastic approximation of our gradients. In the next section (local convergence), we need stronger constraint, and in particular control on the variance of these approximations.*

Remark 1.8. *The bound (5.17) displayed at the end of the proof of Theorem 1.2 can be considered as the analogue of Lemma F1 in Chizat [2022] which is obtained with the deterministic instance of our algorithm. Taking advantage of this lemma, Theorem 4.2 of Chizat [2022] then provide a global convergence result in some kind of asymptotic context. This result (with similar calibration for α and β but associated exponential convergence rates) rely in particular on a Polyak–Łojasiewicz inequality that appears quite difficult to manage in stochastic context. Theorem 1.2 should hence be understood as a non-asymptotic version of this control which appears to be of order $\mathcal{O}(\log(K)/\sqrt{K})$.*

1.4.3 Local minimization with swarm stochastic optimization

To conclude this contribution, we state a complementary result that quantifies the behaviour of Algorithm 1 when the number of particles used is not "as large" as the one indicated in Theorem 1.2.

Theorem 1.3. *Assume that the kernel is two times continuously differentiable. Set $\alpha = \eta = 1/\sqrt{K}$, $m_k = \sqrt{K}$ for all $k \in \{1, \ldots, K\}$ and assume that $\alpha C_1(R_0 + 1) < 1$ where $\mathcal{C}_1$ is introduced in Lemma B.1. For any initial measure satisfying the requirement of Proposition 1.1, if τ_K refers to a random variable uniformly distributed over $\{1, \ldots, K\}$, independent from $(\nu_k)_{k\geq 1}$, then:*

$$\mathbb{E}\left[\|J'_{\nu_{\tau_K}}\|^2_{\nu_{\tau_K}} + \|\nabla J'_{\nu_{\tau_K}}\|^2_{\nu_{\tau_K}}\right] \leq \frac{J(\nu_0) + \mathfrak{C}(1 + R_0^4)}{\sqrt{K}},$$

for some constant $\mathfrak{C}$ that depends on $\|\varphi\|_{\mathrm{Lip}}$.

Even though weaker than Theorem 1.2, the previous "local" result deserves several comments as it raises some meaningful and challenging questions.

Remark 1.9. *The previous result is a strong indicator of the convergence of our swarm stochastic particle algorithms towards a minimizer of J. Indeed, Proposition B.2 stated in Appendix B states that any minimizer $\mu^\star$ of J necessarily satisfies $J'_{\mu^\star} = 0$ on the support of $\mu^\star$ and $J'_{\mu^\star} \geq 0$ everywhere, which implicitly means that $\nabla J'_{\mu^\star} = 0$ on the support of $\mu^\star$ otherwise the previous positivity condition would not hold. In Theorem 1.3, we obtain that $\|J'_{\nu_k}\|^2_{\nu_{\tau_K}}$ and $\|\nabla J'_{\nu_k}\|^2_{\nu_{\tau_K}}$ become arbitrarily small when k becomes larger and larger, which is perfectly in accordance with the previous conditions. Nevertheless, the nature of our algorithm* *does not permit to push further our conclusions, especially about the positivity of J'_{ν_k} everywhere (and not only on the support of ν_k). In particular, to extend our conclusion towards a global minimization result, we need to be able to address a kind of "density of the support", which is absolutely unattainable in our present analysis without multiplying the number of particles all over the state space, which is in some sense what is done in Theorem 1.2. Our convergence statements are expressed in terms of objective values $J(\nu_k) - J(\mu^\star)$ (Theorem 1.2) or derivative-based quantities such as $|J'(\nu_k)|^2 + |\nabla J'(\nu_k)|^2$ (Theorem 1.3). Earlier works (Chizat [2022], De Castro et al. [2021]) reported convergence in terms of proximity between ν_k and $\mu^\star$, which requires stronger local assumptions ("strong source conditions") that yield a form of local strong convexity for J. Establishing analogous results in the stochastic setting (convergence in a measure divergence) is an interesting direction for future research.*

Remark 1.10. *The bound* (5.22) *displayed in the proof of Theorem 1.3 can be related to the descent property provided in deterministic case (see Chizat [2022], Lemma 2.5) and holds for a wide range a possible values for α and β. These parameters are then specifically calibrated to obtain the local result displayed in Theorem 1.3.*

Remark 1.11. *The stochastic setting yields the rate $K^{-1/2}$ in Theorem 1.3, which is minimax optimal for stochastic convex problems. Deterministic CPGD can achieve faster rates (even linear convergence under strong assumptions), as discussed in Chizat (2022).*

2 Rationale of a Stochastic CPGD algorithm.

The primary objective of this contribution is to introduce a stochastic algorithm designed for tackling the optimization Program ($\mathcal{B}$), followed by an exploration of its underlying theoretical properties. Our approach initially stems from a deterministic algorithm, which is subsequently adapted into a stochastic variant to enhance computational efficiency. Considering Program ($\mathcal{B}$) as an optimization challenge, it can be effectively addressed through the application of gradient descent (GD) techniques within the domain of measures. A straightforward algorithm would entail performing a discretized gradient descent on the space of non-negative measures $\mathcal{M}(\mathcal{X})_+$. However, in the absence of a significant conceptual breakthrough pertaining to an efficient parametrization of the preceding iteration within $\mathcal{M}(\mathcal{X})_+$ (or its dual spaces and Hilbert basis), we resort to emulating this gradient descent process through a collection of measures encoded with particles. This method of approximating optimization over measures through particles finds its conceptual foundation in swarm optimization approaches, as exemplified in recent works such as those of Bolte et al. [2023] and Miclo [2023].

2.1 A mirror principled conic gradient descent

2.1.1 The Mirror descent principle

We first introduce a basic ingredient related to optimization problems on geometric spaces, that permits to adapt the evolution of an algorithm to some constrained sets where the problem is embedded. Mirror Descent (MD below) originates from the pioneering work of Nemirovskij and Yudin [1983] and permits to naturally handle optimization problems especially when the mirror/proximal mapping is explicit, which is indeed the case for a convex problem constrained on measures as $\mathcal{M}(\mathcal{X})_+$ (see *e.g.*, Lan et al. [2012], Bubeck et al. [2015]).

Consider a strongly convex function h on the convex set $\mathbb{R}_+^p \times \mathcal{X}^p$, we define the Bregman divergence associated to h as follows: for two pairs $(\mathbb{W}_1, \mathbb{T}_1)$ and $(\mathbb{W}_2, \mathbb{T}_2)$ in $\mathbb{R}_+^p \times \mathcal{X}^p$, we denote:

$$D_h((\mathbb{W}_1, \mathbb{T}_1), (\mathbb{W}_2, \mathbb{T}_2)) = h(\mathbb{W}_1, \mathbb{T}_1) - h(\mathbb{W}_2, \mathbb{T}_2) - \langle \nabla h(\mathbb{W}_2, \mathbb{T}_2), (\mathbb{W}_1, \mathbb{T}_1) - (\mathbb{W}_2, \mathbb{T}_2) \rangle \,. \qquad (2.1)$$

The Bregman divergence D_h is then used to define the MD with the following variational characterisation

$$(\mathbb{W}^{k+1}, \mathbb{T}^{k+1}) = \arg\min_{(\mathbb{W},\mathbb{T})} \left\{ \langle \nabla F(\mathbb{W}^k, \mathbb{T}^k), (\mathbb{W}, \mathbb{T}) - (\mathbb{W}^k, \mathbb{T}^k) \rangle + \frac{1}{\kappa} D_h\left((\mathbb{W}, \mathbb{T}), (\mathbb{W}^{k+1}, \mathbb{T}^{k+1})\right) \right\}, \quad (2.2)$$

where $\kappa > 0$ is a gradient step size.

2.1.2 A conic descent

In this contribution, we will consider the entropy function Ent on $\{\mathbb{R}_+\}^p$ as:

$$\mathrm{Ent}(\mathbb{W}) := \sum_{j=1}^{p} \omega_j \log(\omega_j) \quad (2.3)$$

It induces the Bregman divergence defined on the set of positive weights as

$$D_{\mathrm{Ent}}(\mathbb{W}^1, \mathbb{W}^2) = \sum_{j=1}^{p} \omega_j^1 \left(\frac{\omega_j^2}{\omega_j^1} - 1 - \log \frac{\omega_j^2}{\omega_j^1} \right). \quad (2.4)$$

We then define the global divergence on the set $\mathbb{R}_+^p \times \mathcal{X}^p$, as

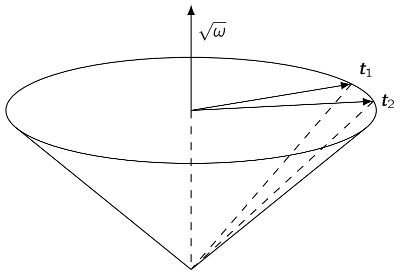
$$\Delta_{\alpha,\eta}((\mathbb{W}^1, \mathbb{T}^1), (\mathbb{W}^2, \mathbb{T}^2)) := \frac{1}{\alpha} D_{\mathrm{Ent}}(\mathbb{W}^1, \mathbb{W}^2) + \frac{1}{\eta} D_{\mathrm{Conic}}((\mathbb{W}^1, \mathbb{T}^1), (\mathbb{W}^2, \mathbb{T}^2)), \quad (2.5)$$

where

$$D_{\mathrm{Conic}}((\mathbb{W}^1, \mathbb{T}^1), (\mathbb{W}^2, \mathbb{T}^2)) := \frac{1}{2} \sum_{j=1}^{p} \omega_j^2 \|\boldsymbol{t}_j^1 - \boldsymbol{t}_j^2\|^2. \quad (2.6)$$

In Equation (2.5) above, the parameter $\alpha > 0$ is the gradient step size for the weights update and $\eta > 0$ for the positions updates. Without loss of generality, we will consider from now on that $\kappa = 1$ as the divergence $\Delta_{\alpha,\eta}$ already incorporates gradients step sizes α and η. This global divergence $\Delta_{\alpha,\eta}$ is used to compute the gradient descent updates thanks to the variational formulation (2.2). We incorporate the term D_{Conic} with the specific intention of aligning it with the gradient updates associated with Conic Particle Gradient Descent (CPGD), as presented for instance by [Chizat, 2022, Section 2.2].

Remark 2.1 (The conic metric...). *We recall that in the CPGD framework [Chizat, 2022, Section 2.2], the set of particles $(\omega, \boldsymbol{t}) \in \mathbb{R}_+ \times \mathcal{X}$ is equipped with the Riemannian metric defined by $(1/2)\nabla_\omega^2 + (\omega/2)\|\nabla_{\boldsymbol{t}}\|^2$ where $(\nabla_\omega, \nabla_{\boldsymbol{t}}) \in \mathbb{R} \times T_{(\omega,\boldsymbol{t})}(\mathcal{X})$ is a tangent vector at point $(\omega, \boldsymbol{t})$. As displayed below, we recognize a "conic" metric where two given fixed points $\boldsymbol{t}_1, \boldsymbol{t}_2$ of $\mathcal{X}$ gets linearly closer as $\sqrt{\omega}$ goes to zero.*



Then a mirror retraction is applied to ω (see Chizat [2022][Definition 2.3]), corresponding to the Bregman divergence term D_{Ent} in our variational formulation. We notice that the term D_{Conic} comes from the latter conic metric.

Remark 2.2 (...is not a Bregman divergence). *It is noteworthy that D_{Conic} does not conform to the definition of a Bregman divergence, as outlined in Definition (2.1). Specifically, there exists no global function h such that D_{Conic} can be expressed in the form of D_h. This is evident, for instance, by considering the boundedness* *of Bregman balls, a property that is conspicuously absent in the case of D_{Conic}. Furthermore, a direct proof of this assertion can be established through a contradiction argument. If such a function h were to exist for $p = 1$, it would imply the following relationship*

$$h(\omega_1, \boldsymbol{t}_1) - h(\omega_2, \boldsymbol{t}_2) - \langle \nabla h(\omega_2, t_2), (\omega_1 - \omega_2, \boldsymbol{t}_1 - \boldsymbol{t}_2) \rangle = \omega_1 \|\boldsymbol{t}_1 - \boldsymbol{t}_2\|^2 \, .$$

Then, consider $\boldsymbol{t}_2 = 0$ and observe that necessarily $h(\omega_1, \boldsymbol{t}_1) = h(0,0) + \omega_1 \|t_1\|^2$, by removing the linear part that is necessarily vanishing. Subsequent straightforward calculations would lead to a contradiction, thereby confirming the incompatibility of D_{Conic} with the Bregman divergence framework.

According to the above remark, our descent algorithm should rather be understood as a Riemannian (stochastic) gradient descent instead of a purely mirror descent. We will keep the MD abuse of terms in what follows as it refers essentially to the evolution of the weights and since it is the commonly used term in machine learning with exponentially parametrized weights. Nevertheless, the difference between $\Delta_{\alpha,\eta}$ in (2.5) and a Bregman divergence prevents the use of standard arguments of convergence for mirror descent algorithm.

2.2 Projected Stochastic conic particle gradient descent (`FastPart`)

With all these essential components in place, we are now prepared to construct our algorithm. The fundamental concept behind this approach is to substitute the deterministic gradient ∇F of F in the mirror descent (2.2) with its stochastic counterpart, as derived from the stochastic gradients on weights (1.14) and positions (1.16) under Assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_2$).

Mirror descent updates An important remark for the tractability of the stochastic CPGD is that Equation (2.2) may be made explicit. In particular, the optimization with respect to $\mathbb{W}$ can be carried out independently of $\mathbb{T}$. This leads to the updates

$$\mathbb{W}^{k+1} = \arg\min_{\mathbb{W}} \left\{ \left\langle \mathbb{W} - \mathbb{W}^k, (\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k))_{j=1\dots p} \right\rangle + \frac{1}{\alpha} D_{Ent}(\mathbb{W}, \mathbb{W}^k) \right\} \tag{2.7}$$

and

$$\mathbb{T}^{k+1} = \arg\min_{\mathbb{T}} \left\{ \left\langle \mathbb{T} - \mathbb{T}^k, (\omega_j^k \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k))_{j=1\dots p} \right\rangle + \frac{1}{\eta} D_{Conic}((\mathbb{W}, \mathbb{T}), (\mathbb{W}^k, \mathbb{T}^k)) \right\} . \tag{2.8}$$

which gives the expressions (1.19) and (1.21) of the algorithm.

Proximal methods guarantees Analysis of Algorithm 1 will be addressed by using some standard tools of proximal methods, and in particular the *generalized projected gradient*. We briefly sketch these tools below. For any $\boldsymbol{t} \in \mathcal{X}$, any "direction" d and any step-size η, we introduce:

$$\boldsymbol{t}^+ := \arg\min_{u \in \mathcal{X}} \left\{ \langle u, d \rangle + \frac{1}{2\eta} \|u - \boldsymbol{t}\|^2 \right\} .$$

The generalized projected gradient associated to a direction d is then defined as

$$P_{\mathcal{X}}(\boldsymbol{t}, d, \eta) := \frac{\boldsymbol{t} - \boldsymbol{t}^+}{\eta} \, ,$$

which depends implicitly on d through t^+ as we also have that $d - P_{\mathcal{X}}(\boldsymbol{t}, d, \eta)$ is a vector of the normal cone of $\mathcal{X}$ at point t^+. Hence, leading to the key identity

$$\boldsymbol{t}^+ = \boldsymbol{t} - \eta P_{\mathcal{X}}(\boldsymbol{t}, d, \eta). \tag{2.9}$$

Some useful properties of this generalized projected gradient can be found for instance in Ghadimi et al. [2016]. In particular Lemma 1 of Ghadimi et al. [2016] may be stated as follows:

Lemma 2.1 (Correlation of the projected gradient and gradient lower bound, Ghadimi et al. [2016])**.** *For any $\boldsymbol{t}_j \in \mathcal{X}$, $d \in \mathbb{R}^d$ and $\eta > 0$:*

$$\langle d, P_{\mathcal{X}}(\boldsymbol{t}_j, d, \eta)\rangle \geq \|P_{\mathcal{X}}(\boldsymbol{t}_j, d, \eta)\|^2.$$

A second key property is the generalization of the 1-Lipschitz inequality for projection, which is obvious in our framework here:

Lemma 2.2 (Lipschitz continuity of generalized projected gradients Ghadimi et al. [2016])**.** *For any $\boldsymbol{t}_j \in \mathcal{X}$, $(d_1, d_2) \in \mathbb{R}^d$ and $\eta > 0$:*

$$\|P_{\mathcal{X}}(\boldsymbol{t}, d_1, \eta) - P_{\mathcal{X}}(\boldsymbol{t}, d_2, \eta)\| \leq \|d_1 - d_2\|.$$

Using (1.20) and (1.21), we have, for any $k \in \mathbb{N}$ and $j \in \{1, \ldots, p\}$

$$\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k = -\eta P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta). \tag{2.10}$$

A direct application of Lemma 2.2 (setting $\boldsymbol{t} = \boldsymbol{t}_j^k$, $d_1 = \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)$ and $d_2 = 0$) then leads to

$$\|\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k\| \leq \eta \|\widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)\|. \tag{2.11}$$

Despite the projection step over $\mathcal{X}$, inequality (2.11) allows to connect the distance between $\boldsymbol{t}_j^{k+1}$ and $\boldsymbol{t}_j^k$ to the norm of the gradient $\widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)$. This inequality will be a cornerstone in our analysis.

3 Some examples from Unsupervised learning and Signal processing

3.1 Mixture Models (GMM)

3.1.1 Introduction

For the sake of clarity, we discuss briefly in this section the specific case of statistical mixture models. They are a class of statistical models that can be used for various purposes such as inference, testing, and modelling, have garnered significant attention in recent years due to their versatility and simplicity. However, the estimation of mixture models remains a complex task, with many aspects of the process not yet fully understood. The Expectation-Maximization (E.M.) algorithm, introduced by Dempster et al. [1977], and its subsequent generalization to stochastic variants in Delyon et al. [1999], have played a crucial role in the development of M.M.. Notably, the E.M. algorithm has been reinterpreted on exponential families as a descent algorithm with a surrogate in Kunstner et al. [2021], which has led to a renewed interest in M.M. within the machine learning and optimization communities. Our work is also related to preliminary experiments conducted in De Castro et al. [2021], which employed a deterministic version of particle gradient descent. These experiments demonstrated the potential of using such methods in the context of M.M., and have inspired further research in this area. While M.M. may appear straightforward at first glance, their estimation poses significant challenges. However, recent advances in the field, including the reinterpretation of the EM algorithm and the use of descent algorithms with surrogates, have reignited interest in these models and their potential applications.

In this setting, the data $\boldsymbol{X} = (x_1, \ldots, x_N)$ are i.i.d. random variables having a density ρ in $\mathbb{R}^d$ (w.r.t. the Lebesgue measure) verifying:

$$\rho = \sigma \star \bar{\mu} = \sum_{j=1}^{\bar{s}} \bar{\omega}_j \sigma_{\bar{\boldsymbol{t}}_j} \text{ with } \bar{\mu} := \sum_{j=1}^{\bar{s}} \bar{\omega}_j \delta_{\bar{\boldsymbol{t}}_j},$$

where $\star$ denotes the convolution product, $\bar{\mu}$ is an *unknown* mixing distribution, σ is a *known* bounded even probability density function on $\mathbb{R}^d$ (*e.g.*, a Gaussian) and we denote by $\sigma_{\boldsymbol{t}}(\cdot) := \sigma(\boldsymbol{t} - \cdot)$. The goal is to recover the target $\bar{\mu}$ and/or the corresponding weights $\overline{\mathbb{W}} := (\bar{\omega}_1, \ldots, \bar{\omega}_{\bar{s}})$ and positions $\overline{\mathbb{T}} := (\bar{\boldsymbol{t}}_1, \ldots, \bar{\boldsymbol{t}}_{\bar{s}})$.

3.1.2 Model specification

We consider the Hilbert space $\mathbb{H}$ defined as the RKHS associated to the Gaussian kernel, denoted by γ_m, with variance parameter $m^2\mathrm{Id}$ and leading to

$$\mathbb{H} := \Big\{ f : \mathbb{R}^d \to \mathbb{R} \ : \ \|f\|_{\mathbb{H}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\mathcal{F}[f](\boldsymbol{t})|^2}{\mathcal{F}[\gamma_m](\boldsymbol{t})} \mathrm{d}\boldsymbol{t} < +\infty \Big\} \quad \text{with} \quad \mathcal{F}[\gamma_m] = \exp\big(-\frac{m^2}{2}\|\cdot\|_2^2\big)\,,$$

where $\mathcal{F}$ is the Fourier transform defined as

$$\forall \boldsymbol{u} \in \mathbb{R}^d\,, \quad \forall f \in \mathbb{H}\,, \qquad \mathcal{F}[f](\boldsymbol{u}) := \int_{\mathbb{R}^d} f(\boldsymbol{t}) e^{-\mathrm{i}\langle \boldsymbol{u}, \boldsymbol{t}\rangle} \mathrm{d}\boldsymbol{t},$$

and i refers to the complex number. The inner product associated to the space $\mathbb{H}$ hence verifies

$$\forall f, g \in \mathbb{H}\,, \qquad \langle f, g\rangle_{\mathbb{H}} = \frac{1}{(2\pi)^d} \mathcal{R}\mathrm{e}\Big(\int_{\mathbb{R}^d} \frac{\mathcal{F}[f](\boldsymbol{t}) \times \overline{\mathcal{F}[g](\boldsymbol{t})}}{\mathcal{F}[\gamma_m](\boldsymbol{t})} \mathrm{d}\boldsymbol{t} \Big)\,, \tag{3.1}$$

where $\mathcal{R}\mathrm{e}(z)$ denotes the real part of a complex number z. We embed the sample $\boldsymbol{X}$ and the density ρ in the same Hilbert space $\mathbb{H}$ taking a convolution with γ_m. It yields

$$\boldsymbol{y} = \frac{1}{N} \sum_{i=1}^{N} \gamma_m(x_i - \cdot) \quad \text{and} \quad \bar{\boldsymbol{y}} := \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{y}] = (\gamma_m \star \sigma) \star \bar{\mu} = \sum_{j=1}^{\bar{s}} \bar{\omega}_j (\gamma_m \star \sigma)(\bar{\boldsymbol{t}}_j - \cdot)\,.$$

We uncover that the feature map (1.1) and the measure embedding (1.3) are given by

$$\varphi_{\boldsymbol{t}}(\cdot) := (\gamma_m \star \sigma)(\boldsymbol{t} - \cdot) \quad \text{and} \quad \Phi(\mu) := (\gamma_m \star \sigma) \star \mu\,.$$

Note that the observation $\boldsymbol{y} \in \mathbb{H}$ corresponds to the non-parametric kernel estimator of $\bar{\boldsymbol{y}} = \Phi(\bar{\mu})$ based of the sample $\boldsymbol{X}$. We also have the following kernel expression matching the expression (1.12)

$$\mathbb{K}(\boldsymbol{t}, \boldsymbol{t}') = \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}'}\rangle_{\mathbb{H}} = (\tilde{k} \star \sigma)(\boldsymbol{t} - \boldsymbol{t}') \quad \text{where} \quad \tilde{k}(\boldsymbol{t}) := (\gamma_m \star \sigma)(\boldsymbol{t})\,, \tag{3.2}$$

using the Fourier inversion formula. Last but not least, we can derive the following expression

$$\langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} = \frac{1}{N} \sum_{i=1}^{N} \tilde{k}(x_i - \boldsymbol{t}) = \boldsymbol{y}(\boldsymbol{t})\,,$$

by the same means. For the sake of simplicity, we will omit the dependency with the standard deviation m as we are only interested in the optimization problem in the sequel.

3.1.3 Gradients

For any $\mu \in \mathcal{M}_+(\mathbb{R}^d)$, the observation $\boldsymbol{y}$ is compared to $\Phi(\mu) = (\gamma \star \sigma) \star \mu$ inside the criterion $J(\mu)$. By (1.11), one has

$$J'_\nu(\boldsymbol{t}) = \langle \varphi_{\boldsymbol{t}}, \Phi(\nu)\rangle_{\mathbb{H}} - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda = \sum_{j=1}^{p} \omega_j \langle \varphi_{\boldsymbol{t}}, \varphi_{\boldsymbol{t}_j}\rangle_{\mathbb{H}} - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda\,, \tag{3.3}$$

which can be re-written as

$$J'_\nu(\boldsymbol{t}) = \sum_{j=1}^{p} \omega_j \int_{\mathbb{R}^d} \tilde{k}(\boldsymbol{t} - \boldsymbol{t}_j - \boldsymbol{u}) \sigma(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} - \frac{1}{N} \sum_{i=1}^{N} \tilde{k}(x_i - \boldsymbol{t}) + \lambda\,, \tag{3.4}$$

following (1.12).

3.1.4 Assumptions

We can identify the $\mathbf{g}$ and $\mathbf{h}$ functions appearing in Assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_2$). It holds

$$\begin{aligned}\mathbf{g}_{\boldsymbol{t},\boldsymbol{t}'}(\boldsymbol{u}) &:= \tilde{k}(\boldsymbol{t}-\boldsymbol{t}'-\boldsymbol{u})\\ \mathbf{h}_{\boldsymbol{t}}(\boldsymbol{v}) &:= \tilde{k}(\boldsymbol{t}-\boldsymbol{v})\,,\end{aligned}$$

and the laws of U, V are independently distributed according to $U \sim \sigma$ and V picking a data sample at random as we will see below. Now, observe that $\tilde{k}$ is bounded as γ is, hence both functions are bounded and Assumption ($\mathbf{A}_1$) is satisfied. Also

$$\begin{aligned}\nabla_{\boldsymbol{t}}\mathbf{g}_{\boldsymbol{t},\boldsymbol{t}'}(\boldsymbol{u}) &= \nabla\tilde{k}(\boldsymbol{t}-\boldsymbol{t}'-\boldsymbol{u})\\ \nabla_{\boldsymbol{t}}\mathbf{h}_{\boldsymbol{t}}(\boldsymbol{v}) &= \nabla\tilde{k}(\boldsymbol{t}-\boldsymbol{v})\,,\end{aligned}$$

and since γ has bounded gradient and that the convolution is a contraction for the sup norm, it holds that Assumption ($\mathbf{A}_2$) is satisfied. Finally, it is not hard to prove that Assumption ($\mathbf{A}_\mathrm{C}$) is satisfied since the kernel is a convolution by a smooth function.

3.1.5 Stochastic counterpart

Remark from (3.4) that J'_ν is obtained through some integral computations. Given $\nu = \nu(\mathbb{W}, \mathbb{T})$, we introduce a random variable $Z = (T, U, V)$ built as follows. Consider three independent random variables given by

$$T \sim \frac{\nu}{\|\nu\|_{\mathrm{TV}}}\,,\quad U \sim \sigma\,,\quad \text{and}\quad V \sim \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}\,, \tag{3.5}$$

where ν is assumed to be a discrete non-negative measure as displayed in Algorithm 1, hence $\nu/\|\nu\|_{\mathrm{TV}}$ is a discrete probability measure. We have

$$\begin{aligned}\boldsymbol{J}'_{\nu,\boldsymbol{t}}(Z) &= \|\nu\|_{\mathrm{TV}}\tilde{k}(\boldsymbol{t}-T-U) - \tilde{k}(\boldsymbol{t}-V) + \lambda\,, &(3.6)\\ &= \|\nu\|_{\mathrm{TV}}\mathbf{g}_{\boldsymbol{t},T}(U) - \mathbf{h}_{\boldsymbol{t}}(V) + \lambda\,. &(3.7)\end{aligned}$$

and we uncover (1.14). According to (3.3) and (3.5), we can verify that once ν is fixed, $\boldsymbol{J}'_{\nu,\boldsymbol{t}}(Z)$ is an unbiased estimation of J'_ν and we can observe that that there exists a random variable $\xi_\nu(\boldsymbol{t}, Z)$ such that

$$\boldsymbol{J}'_{\nu,\boldsymbol{t}}(Z) = J'_\nu(\boldsymbol{t}) + \xi_\nu(\boldsymbol{t}, Z) \quad \text{with} \quad \mathbb{E}_Z[\xi_\nu(\boldsymbol{t}, Z)] = 0\,. \tag{3.8}$$

A same remark occurs for the gradient of J'_ν since for any $\boldsymbol{t} \in \mathbb{R}^d$,

$$\nabla_{\boldsymbol{t}}J'_\nu(\boldsymbol{t}) = \sum_{j=1}^{p}\omega_j\int_{\mathbb{R}^d}\nabla\tilde{k}(\boldsymbol{t}-\boldsymbol{t}_j-\boldsymbol{u})\sigma(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - \frac{1}{N}\sum_{i=1}^{N}\nabla\tilde{k}(x_i-\boldsymbol{t})\,. \tag{3.9}$$

We shall then define

$$\begin{aligned}\boldsymbol{D}_{\nu,\boldsymbol{t}}(Z) &= \|\nu\|_{\mathrm{TV}}\nabla\tilde{k}(\boldsymbol{t}-T-U) - \nabla\tilde{k}(\boldsymbol{t}-V)\,,\\ &= \|\nu\|_{\mathrm{TV}}\nabla_{\boldsymbol{t}}\mathbf{g}_{\boldsymbol{t},T}(U) - \nabla_{\boldsymbol{t}}\mathbf{h}_{\boldsymbol{t}}(V)\,.\end{aligned}$$

and we uncover (1.16). We get that

$$\boldsymbol{D}_{\nu,\boldsymbol{t}}(Z) = \nabla_{\boldsymbol{t}}J'_\nu(\boldsymbol{t}) + \zeta_\nu(\boldsymbol{t}, Z) \quad \text{with} \quad \mathbb{E}_Z[\zeta_\nu(\boldsymbol{t}, Z)] = 0, \tag{3.10}$$

for some centred random vector $\zeta_\nu(t, Z)$.

Therefore, in this example of mixture deconvolution, our situation perfectly fits the stochastic gradient setting where we can easily access some unbiased random realization of the true gradient of F for any position of a current algorithm $\nu = \nu(\mathbb{W}, \mathbb{T})$.

3.2 Interlude, a comparison with sketching

In the field of machine learning, replacing the computation of the complex integrals present in Equation (3.4) is commonly referred to as sketching [Keriven et al., 2018]. The key distinction between our approach and sketching-type methods lies in how the frequencies (U_k) are sampled. In the sketching approach, the frequencies are initially sampled only once at the beginning of the algorithm. Consequently, the gradients $J'_\nu(\boldsymbol{t})$ and $\nabla_{\boldsymbol{t}} J'_\nu(\boldsymbol{t})$ are approximated using a Monte-Carlo version of (3.4) that employs the same Monte-Carlo sample (U_k) along the descent and the size of U_k *needs to be large* to guarantee a good Monte-Carlo approximation of the several integrals involved in the iterations of the method. Therefore, the sketching strategy of [Keriven et al., 2018] leads to costly iterations as the size of (U_k) is not negligible. In contrast, our method involves sampling a unique Monte-Carlo sample U_k at each step k, and we replace the integrals of (3.4) with an evaluation at sample U_k (or we could also adopt a mini-batch strategy). It is important to note that in our case, the number of frequencies (U_k) is equal to the number of steps K (or times the size of mini-batch), whereas in sketching approaches, the number of frequencies is directly proportional to the number of parameters to be estimated, up to logarithmic factors. For more information on this topic, refer to Keriven et al. [2018].

3.3 Sparse deconvolution with positive definite kernel

3.3.1 Introduction

The statistical analysis of the ℓ_1-regularization in the space of measures was initiated by Donoho [Donoho, 1992] and then investigated by Gamboa and Gassiat [1996]. Recently, this problem has attracted a lot of attention in the "*Super-Resolution*" community and its companion formulation in "*Line spectral estimation*". In the Super-Resolution frame, one aims at recovering fine scale details of an image from few low frequency measurements, ideally the observation is given by a low-pass filter. The novelty of this body of work lies in new theoretical guarantees of the ℓ_1-minimization over the space of discrete measures in a gridless manner, referred to as "off-the-grid" methods. Some recent work on this topic can be found in Bredies and Pikkarainen [2013], Tang et al. [2013], Candès and Fernandez-Granda [2014, 2013], Fernandez-Granda [2013], Duval and Peyré [2015], De Castro and Gamboa [2012], Azais et al. [2015]. More precisely, pioneering work can be found in Bredies and Pikkarainen [2013], which treats of inverse problems on the space of Radon measures and Candès and Fernandez-Granda [2014], which investigates the Super-Resolution problem via Semi-Definite Programming and the ground breaking construction of a "*dual certificate*". Exact Reconstruction property (in the noiseless case), minimax prediction and localization (in the noisy case) have been performed using the "*Beurling Lasso*" estimator ($\mathcal{B}$) introduced in Azais et al. [2015] and also studied in Tang et al. [2013], Fernandez-Granda [2013], Tang et al. [2015] which minimizes the total variation norm over complex Borel measures. Noise robustness (as the noise level tends to zero) has been investigated in the captivating paper Duval and Peyré [2015]. A sketching formulation and a construction of dual certificates with respect to the Fisher metric has been pioneering studied in Poon et al. [2021].

3.3.2 Model specification

In sparse deconvolution, the convolution $\bar{\boldsymbol{y}}$ of some measure $\bar{\mu}$ with a $\mathcal{C}^1$-continuous **positive definite** function φ is given by

$$\forall \boldsymbol{t} \in \mathcal{X}\,, \quad \bar{\boldsymbol{y}}(\boldsymbol{t}) = \sum_{j=1}^{\bar{s}} \bar{\omega}_j \varphi(\boldsymbol{t} - \bar{\boldsymbol{t}}_j)\,,$$

and we observe $\boldsymbol{y}$, a noisy version of $\bar{\boldsymbol{y}}$, given by $\boldsymbol{y} = \bar{\boldsymbol{y}} + \boldsymbol{e}$, where $\boldsymbol{e} : \mathcal{X} \to \mathbb{R}$ is some function. The feature map (1.1) is the convolution kernel and the measure embedding (1.3) are given by

$$\varphi_{\boldsymbol{t}}(\cdot) := \varphi(\boldsymbol{t} - \cdot) \text{ and } \Phi(\mu) := \varphi \star \mu\,.$$

Recall that φ is a positive definite function. We assume that φ is defined on the d-Torus $\mathbb{T}^d$ or on $\mathbb{R}^d$, and we further assume that $\varphi(0) = 1$, without log of generality on this latter point. By Bochner's theorem, there exists a probability measure σ, referred to as the spectral measure of φ, such that

$$\forall \boldsymbol{t}, \boldsymbol{s} \in \mathcal{X}\,, \quad \mathcal{F}[\varphi_{\boldsymbol{t}}](\boldsymbol{s}) = \sigma(\boldsymbol{s}) e^{-\mathrm{i}\langle \boldsymbol{s}, \boldsymbol{t}\rangle}\,. \tag{3.11}$$

We consider the following assumption.

Assumption $(\mathbf{A}_{\text{conv}})$**.** *The spectral measure σ of φ has compact support.*

Super Resolution In this case, $\mathcal{X} = \mathbb{T}^d$ and $\sigma = \sum_{\boldsymbol{u}\in\mathbb{Z}^d} \sigma_{\boldsymbol{u}}\delta_{\boldsymbol{u}}$ is a probability measure on $\mathbb{Z}^d$. When σ is the uniform measure on $[-f_c, f_c]^d$, with $f_c \geq 1$, we uncover the Super-Resolution framework and φ is the Dirichlet kernel. For latter use, we denote

$$\int e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}-\boldsymbol{s}\rangle}\mathrm{d}\sigma(\boldsymbol{u}) := \sum_{\boldsymbol{u}\in\mathbb{Z}^d} \sigma_{\boldsymbol{u}} e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}-\boldsymbol{s}\rangle}\,.$$

Continuous sampling Fourier transform In this case, $\mathcal{X}$ is a compact set of $\mathbb{R}^d$ and σ is a probability measure on $\mathbb{R}^d$. For latter use, we denote

$$\int e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}-\boldsymbol{s}\rangle}\mathrm{d}\sigma(\boldsymbol{u}) := \int_{\mathbb{R}^d} e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}-\boldsymbol{s}\rangle}\mathrm{d}\sigma(\boldsymbol{u})\,.$$

By Lemma A.1, we can assume that $\mathbb{H}$ is the RKHS associated with the kernel defined by φ. In particular,

$$\forall \boldsymbol{t} \in \mathcal{X}\,,\quad \langle \varphi_{\boldsymbol{t}}, \varphi_{t_j}\rangle_{\mathbb{H}} = \varphi_{t_j}(\boldsymbol{t}) \text{ and } \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} = \boldsymbol{y}(\boldsymbol{t}). \tag{3.12}$$

About the noise term and the regularity of the observation Using the projection Π and the isometry $\sqsupset$ of Lemma A.1, we can assume without loss of generality that the noise term $\boldsymbol{e}$ belongs to $\mathbb{H}$, the RKHS associated with φ. By Assumption $(\mathbf{A}_{\text{conv}})$, it can shown that φ is smooth, hence all the elements of $\mathbb{H}$ are smooth and so is $\boldsymbol{e}$. Since $\mathcal{X}$ is compact, we deduce that $\boldsymbol{y}$ and its gradient are bounded.

3.3.3 Gradients

For any $\mu \in \mathcal{M}(\mathbb{R}^d)$, the observation $\boldsymbol{y}$ is compared to $\Phi(\mu) = \varphi \star \mu$ inside the criterion $J(\mu)$. A direct application of Proposition B.1 leads to $J'_\nu = \Phi^\star(\Phi(\nu) - \boldsymbol{y}) + \lambda$, and one can check that, for any $\boldsymbol{t} \in \mathbb{R}^d$,

$$J'_\nu(\boldsymbol{t}) = \langle \varphi_{\boldsymbol{t}}, \Phi(\nu) - \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda\,.$$

The latter formula can be re-written as

$$J'_\nu(\boldsymbol{t}) = \langle \varphi_{\boldsymbol{t}}, \Phi(\nu)\rangle_{\mathbb{H}} - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda = \sum_{j=1}^{p} \omega_j \langle \varphi_{\boldsymbol{t}}, \varphi_{t_j}\rangle_{\mathbb{H}} - \langle \varphi_{\boldsymbol{t}}, \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda\,. \tag{3.13}$$

By (3.11) and (3.12), it yields that

$$J'_\nu(\boldsymbol{t}) = \sum_{j=1}^{p} \omega_j \int e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}-t_j\rangle}\mathrm{d}\sigma(\boldsymbol{u}) - \boldsymbol{y}(\boldsymbol{t}) + \lambda\,. \tag{3.14}$$

3.3.4 Assumptions

We can identify the $\mathbf{g}$ and $\mathbf{h}$ functions appearing in Assumptions $(\mathbf{A}_1)$ and $(\mathbf{A}_2)$. It holds

$$\begin{aligned}\mathbf{g}_{\boldsymbol{t},\boldsymbol{t}'}(\boldsymbol{u}) &:= e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}-\boldsymbol{t}'\rangle}\\ \mathbf{h}_{\boldsymbol{t}} &:= \boldsymbol{y}(\boldsymbol{t})\,,\end{aligned}$$

which are bounded functions and Assumption $(\mathbf{A}_1)$ is satisfied. Using the dominated convergence theorem for the bounded gradients

$$\begin{aligned}\nabla_{\boldsymbol{t}}\mathbf{g}_{\boldsymbol{t},\boldsymbol{t}'}(\boldsymbol{u}) &:= -\mathrm{i}\boldsymbol{u}e^{-\mathrm{i}\langle \boldsymbol{u},\boldsymbol{t}-\boldsymbol{t}'\rangle}\mathbf{1}_{\{\boldsymbol{u}\in\mathrm{Supp}(\sigma)\}}\\ \nabla_{\boldsymbol{t}}\mathbf{h}_{\boldsymbol{t}} &:= \nabla\boldsymbol{y}(\boldsymbol{t})\,,\end{aligned}$$

where $\mathrm{Supp}(\sigma)$ denotes the support of the spectral measure σ. One can check that Assumption $(\mathbf{A}_2)$ is satisfied under Assumption $(\mathbf{A}_{\text{conv}})$.

3.3.5 Stochastic counterpart

Given $\nu = \nu(\mathbb{W}, \mathbb{T})$, we introduce a random variable $Z = (T, U)$ built as follows (there is no V random variable in this case). Consider two independent random variables given by

$$T \sim \frac{\nu}{\|\nu\|_{\mathrm{TV}}} \quad \text{and} \quad U \sim \sigma\,, \tag{3.15}$$

where ν assumed non-negative without loss of generality as discussed in (1.7), hence $\nu/\|\nu\|_{\mathrm{TV}}$ is a discrete probability measure. We then define

$$\begin{aligned} \boldsymbol{J}'_{\nu,\boldsymbol{t}}(Z) &= \|\nu\|_{\mathrm{TV}} e^{-\mathrm{i}\langle U, \boldsymbol{t}-T\rangle} - \boldsymbol{y}(\boldsymbol{t}) + \lambda\,, \\ &= \|\nu\|_{\mathrm{TV}} \mathbf{g}_{\boldsymbol{t},T}(U) - \mathbf{h}_{\boldsymbol{t}}(V) + \lambda\,, \end{aligned}$$

writing, with a slight abuse of notation, $\boldsymbol{h}_t(V) = \boldsymbol{y}(t)$. We hence uncover (1.14). According to (3.13) and (3.15), we can verify that once ν is fixed, $\boldsymbol{J}'_{\nu,\boldsymbol{t}}(Z)$ is an unbiased estimation of J'_ν and we can observe that that there exists a random variable $\xi_\nu(\boldsymbol{t}, Z)$ such that

$$\boldsymbol{J}'_{\nu,\boldsymbol{t}}(Z) = J'_\nu(\boldsymbol{t}) + \xi_\nu(\boldsymbol{t}, Z) \quad \text{with} \quad \mathbb{E}_Z[\xi_\nu(\boldsymbol{t}, Z)] = 0\,. \tag{3.16}$$

A same remark occurs for the gradient of J'_ν since for any $\boldsymbol{t} \in \mathbb{R}^d$,

$$\nabla_{\boldsymbol{t}} J'_\nu(\boldsymbol{t}) = -\mathrm{i} \sum_{j=1}^{p} \omega_j \int \boldsymbol{u} e^{-\mathrm{i}\langle \boldsymbol{u}, \boldsymbol{t}-\boldsymbol{t}_j\rangle} \mathrm{d}\sigma(\boldsymbol{u}) - \nabla \boldsymbol{y}(\boldsymbol{t})\,. \tag{3.17}$$

We shall then define

$$\begin{aligned} \boldsymbol{D}_{\nu,\boldsymbol{t}}(Z) &= -\mathrm{i}\|\nu\|_{\mathrm{TV}} U e^{-\mathrm{i}\langle \boldsymbol{u}, \boldsymbol{t}-T\rangle} - \nabla \boldsymbol{y}(\boldsymbol{t})\,, \\ &= \|\nu\|_{\mathrm{TV}} \nabla_{\boldsymbol{t}} \mathbf{g}_{\boldsymbol{t},T}(U) - \nabla_{\boldsymbol{t}} \mathbf{h}_{\boldsymbol{t}}(V)\,. \end{aligned}$$

and we uncover (1.16). We get that

$$\boldsymbol{D}_{\nu,\boldsymbol{t}}(Z) = \nabla_{\boldsymbol{t}} J'_\nu(\boldsymbol{t}) + \zeta_\nu(\boldsymbol{t}, Z) \quad \text{with} \quad \mathbb{E}_Z[\zeta_\nu(\boldsymbol{t}, Z)] = 0, \tag{3.18}$$

for some centered random vector $\zeta_\nu(t, Z)$.

4 Numerical experiments on FastPart

4.1 Experimental setup

In this short section, we develop a brief numerical study, that may be seen as a proof of concept, to assess the efficiency of our stochastic gradient descent, when using some sketched randomized evaluations of J'_ν, with the help of sampling involved in Equations (3.4) and (3.5).[1]

For this purpose, we consider the Supermix problem introduced in De Castro et al. [2021], which is described in Section 3.1, when considering a mixture of Gaussian densities. We consider three toy situations in 1D. Figure 1 represents the mixture densities considered in this study, that contains for two of them 3 components, and for the last one 5 components.

In Section **??**, we present a different setting on a real dataset in dimension $d = 8$ using two layers neural network and S-CPGD.

[1] Our simulations greatly benefit from the previous work of Nicolas Jouvin https://nicolasjouvin.github.io/, while the original numerical Python code is made available here https://forgemia.inra.fr/njouvin/particle_blasso.

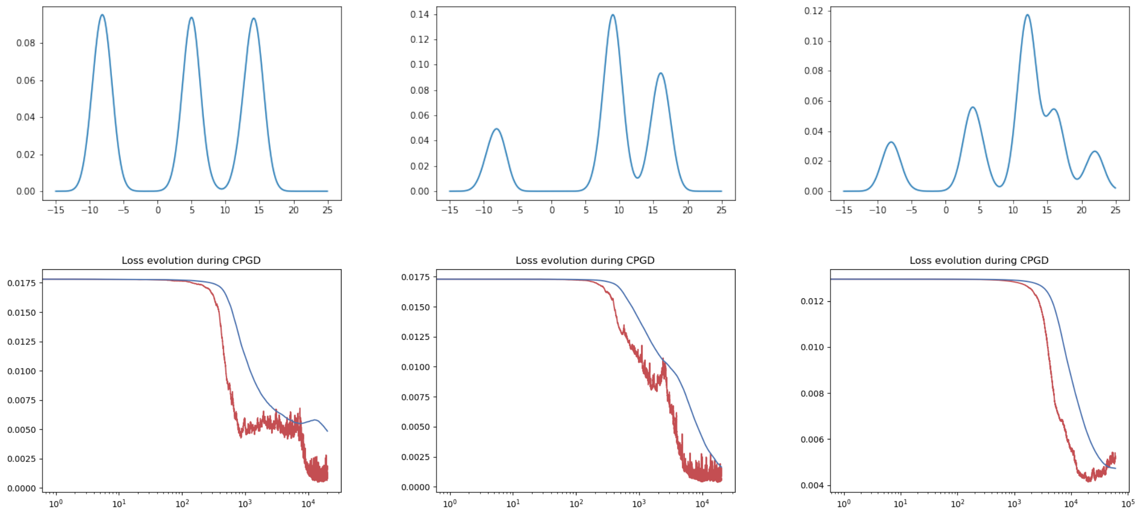


Figure 1: *Top: Three 1-D Gaussian mixture distributions to be learnt by Supermix and Stochastic Conic Particle Gradient Descent. Bottom: Evolution of the loss (log scale) with our S-CPGD algorithm (red) and its averaged counterpart (in blue) when using our methods on the mixture problems.*

4.2 Benchmark

Our experimental setup is essentially built with the help of three versions of the Conic Particle Gradient Descent.

- The first method we will use is the deterministic CPGD introduced in Chizat [2022], implemented by N. Jouvin `https://forgemia.inra.fr/njouvin/particle_blasso`. This method depends on the number of particles we use, the learning rate that encodes the gain of the algorithm at each iteration, and the number of iterations.

- The second method is the Stochastic-CPGD introduced in this work. Our method depends on the same previous set of parameters (number of particles, learning rate, number of iterations) and of the batch size of the data we sample per iteration and the number of randomly sketched frequencies.

- The last method is simply the Cesàro averaged counterpart of our Stochastic-CPGD, but this method raises some technical computational difficulties since averaging a sequence of measures seriously complicates the final estimates $\bar{\nu}_K = \frac{1}{K+1}\sum_{k=0}^{K} \nu_k$. To overcome this difficulty, we have chosen to use instead the measures supported by the averaged means all along the trajectory of the S-CPGD, and weighted by the averaged weights of the S-CPGD. For this purpose, we introduce

$$\forall j \in \{1,\dots,p\} \quad \forall K \geq 0 \qquad \bar{\mathbf{t}}_j^K = \frac{1}{K+1}\sum_{k=0}^{K} \mathbf{t}_j^k \qquad \text{and} \qquad \bar{\omega}_j^K = \frac{1}{K+1}\sum_{k=0}^{K} \omega_j^k$$

  and we approximate $\bar{\nu}_K$ with the help of the sequence $\hat{\bar{\nu}}_K$, defined by:

$$\hat{\bar{\nu}}_K = \sum_{j=1}^{p} \bar{\omega}_j^K \delta_{\bar{\mathbf{t}}_j^K}. \tag{4.1}$$

  Again, this sequence $\hat{\bar{\nu}}_K$ depends on several parameters, the learning rate, the number of particles and iterations, and the size of batches and sketches as well.

4.3 Results

Loss function: Averaging vs no averaging We show in Figure 1 the evolution of the loss function J_ν over the iterations of the algorithm. We emphasize that the complexity of the S-CPGD and of the approximated

Cesàro average are almost the same, since the sequence $\hat{\bar{\nu}}_K$ introduced in (4.1) is a cheap approximation of the true Cesàro averaged sequence $\bar{\nu}_K$.

First, as indicated in Figure 1, we shall observe that the sequence $\hat{\bar{\nu}}_K$ always produces the desired smoothing effect all over the iterations of the algorithm, while slowing a bit the decrease of the loss function J_ν over the iterations. As a consequence, it seems more appropriate to use several long-range parallelized S-CPGD instead of a unique average thread of S-CPGD. In the same time, it may be remarked that our sequence $(\hat{\bar{\nu}}_K)_{K\geq 1}$ is a rough approximation of the true Cesàro averaging that is studied in our paper and the numerical approximation introduced in (4.1) may not be as good as the true Cesàro sequence $(\bar{\nu}_K)_{K\geq 1}$.

Second, as a classical phenomenon in machine learning when using stochastic approximation algorithm, or over-parametrized neural networks, our S-CPGD commonly generates some double-descent phenomena (see the 3 sub-figures of Figure 1) that translates some local minimizer escape of the swarm of particles.

Loss function: Averaging vs no averaging vs Deterministic Figure 2 represents the evolution of the cost function with respect to the numerical cost which is a far better indicator than the number of iterations of the algorithm in our case since the S-CPGD algorithm is designed to be much more cheaper than the deterministic CPGD. Figure 2 clearly illustrates the efficiency of our method with regards to the deterministic one as the red curve shows that the non-averaged S-CPGD produces comparable results as those obtained by CPGD with a significantly lower needs of computational cost: the red curve is clearly shifted on the left when compared to the blue one. It is furthermore possible to quantitatively assess the numerical gain produced by the S-CPGD when compared to the deterministic one: on our toy example, the deterministic CPGD requires approximately 4 more computations to attain the same decrease of the J_ν, this effect being even amplified when the number of particles is increasing.

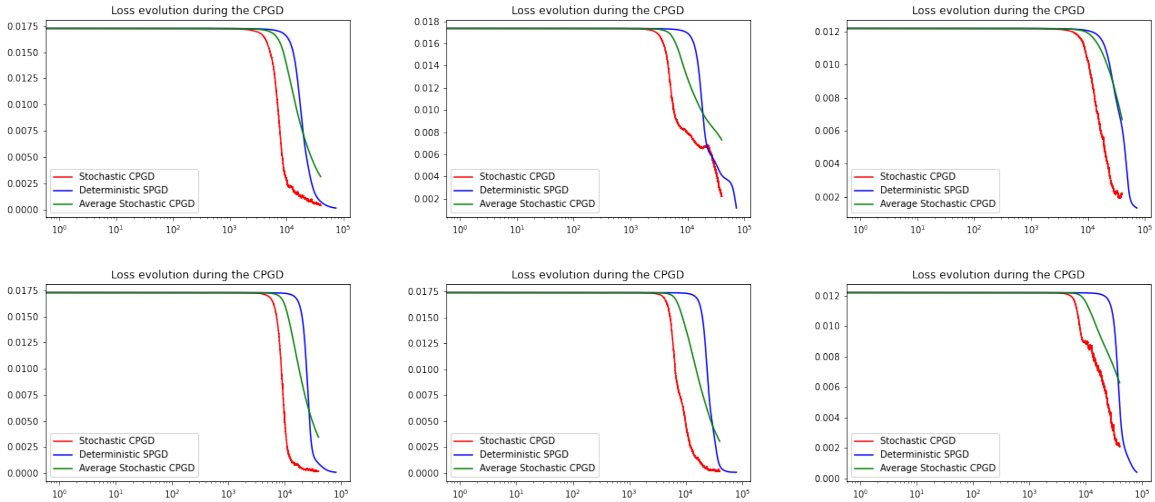


Figure 2: *Evolution of the loss (log scale of the* **computational time***) with our S-CPGD algorithm (red), its averaged counterpart (in green) and the deterministic CPGD (in blue) on the mixture problem of Figure 1 with 20 particles (top) and with 50 particles (bottom). As indicated in the text, and observed with the shift to the right of the blue curve when compared to the red one, the cost of our S-CPGD is much cheaper than the deterministic CPGD.*

Loss function: Effect of the number of particles In the meantime, we observe that the loss function benefits from a large number of particles (see the comparison between top and bottom lines of Figure 2) but this should be tempered by the increasing number of simulations, which varies linearly with the number of particles. We should finally observe that using a large number of particles seems to be important especially in difficult situations (as illustrated in the right column of Figure 2 where using 50 particles instead of 20 significantly improves the loss function, which is not the case on the right column of Figure 2.

The effect of the number of particles can also be illustrated while looking at the trajectories themselves of the particles as shown in Figure 3. We observe that the number of particles is a clear key parameter that strongly influences the success of the method. In our example of 5 components GMM, (last example in Figure 1), we see that a too small number of particles completely miss some components of the mixture while using a strongly over-parametrized set of particles permit to fully recover the support of the mixing distribution, even in the situation where some components of the mixture overlap.

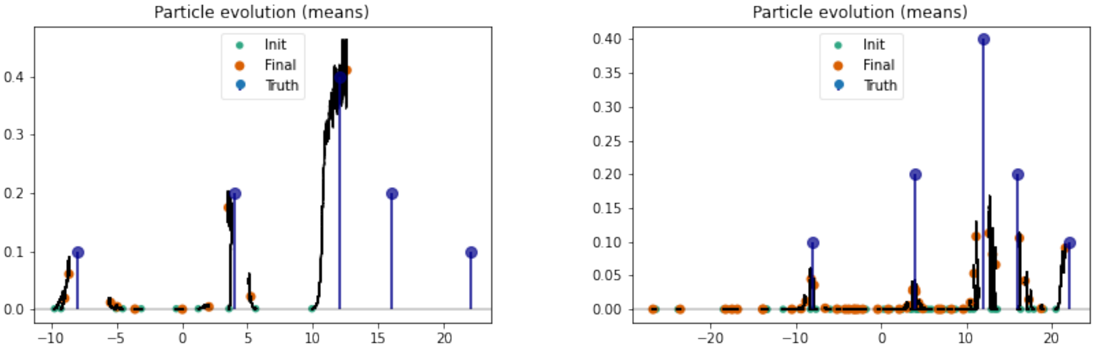


Figure 3: *Trajectories of the particles of our S-CPGD algorithm in the third example of the Gaussian mixture problem of Figure 1 with 5 modes. Left: trajectories using 10 particles. Right: Same with 50 particles. The l.h.s. shows the behaviour of S-CPGD when a too small number of particles is used. Particles concentrate around good positions but may miss some of the important locations of the Gaussian mixture. The r.h.s. demonstrates that a sufficiently large number of particles is necessary to guarantee an exhaustive reconstruction of the mixing distribution.*

From our brief numerical study, we can conclude that both sketching and batch subsampling with a stochastic gradient strategy appears to strongly improve the numerical cost of the Conic Particle Gradient Descent, which permits to increase the number of particles used in the mean field approximation. We also have shown that in some difficult inverse problem examples, a large number of particles seems necessary to perfectly recover the solution of the optimization problem. It appears that the problem seriously benefits from a strong over-parametrisation, that may be handled with our cheap stochastic computing approach, which is not the case in a reasonable time with the deterministic CPGD.

4.4 An example on a real dataset using two-layers neural network

We illustrate S-CPGD on the California Housing dataset, a standard dataset in Machine Learning see `https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_california_housing.html`. It has $d = 8$ features (dimension) and $20,640$ data points. We use a training set of size $N = 18,576$, the remaining data points being the test set. Stochastic gradient descent was performed over batches of size $\mathrm{BS} = 256$.

We use a simple neural network given by a hidden layer of size $p = 500$ (number of particles) with ReLU activation function given by

$$\sum_{i=1}^{500} \varepsilon_i \omega_i \operatorname{ReLu}(\langle \boldsymbol{t}_i, \cdot \rangle)$$

where $\boldsymbol{t}_i \in \mathbb{R}^8$ are the location of the particles and $\varepsilon_i \omega_i \in \mathbb{R}$ their weights. Since the ReLu is one homogeneous, we project the locations onto the unit ball (radius $R_{\mathcal{X}} = 1$) at each iteration of the algorithm, the weights ω_i being scaled by $\|\boldsymbol{t}_i\|_2$ in light of the identity:

$$\sum_{i=1}^{500} \varepsilon_i \omega_i \operatorname{ReLu}(\langle \boldsymbol{t}_i, \cdot \rangle) = \sum_{i=1}^{500} \varepsilon_i \omega_i \|\boldsymbol{t}_i\|_2 \operatorname{ReLu}(\langle \boldsymbol{t}_i / \|\boldsymbol{t}_i\|_2, \cdot \rangle) \,.$$

This setup matches the one of Bach and Chizat [2021] and their experiments.

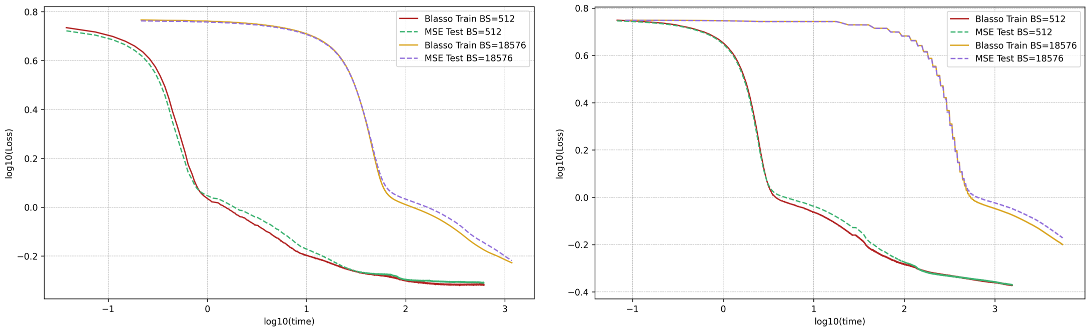


Figure 4: *Stochastic/deterministic CPGD on California Housing dataset* ($d = 8$) *using two-layers neural network with ReLu activation function. The stochastic version has a batch size* BS $= 512$ *while the deterministic CPGD computes gradients over the whole training set of size* BS $= 18,576$. We use $p = 10$ particles on the right and $p = 500$ on the left. The dotted line are the MSE on the test set of size $2,064$ over time. The plain line represents the BLASSO loss over time.

In Figure 4, we observe a gain of two orders of magnitude when using the stochastic version of CPGD. The experiments were done on a single CPU of a computer (iMac with M3 chip) and shows that it takes around a few minutes to optimize $p = 500$ particles in dimension $d = 8$.

5 Proof of the main results

In this section, $\mathfrak{C}$ will refer to a constant independent on K and p, whose value may change from line to line.

5.1 Proof of Proposition 1.1

Proof. The principle of this proofs is as follows. We identify two radii r and r_0 for which, for any $k \in \mathbb{N}^*$ and $j \in \{1, \ldots, p\}$, the two followings properties hold:

- If $\omega_j^k \leq r < r_0$, then $\omega_j^{k+1} \leq r_0$.
- If $r \leq \omega_j^k \leq r_0$, then $\omega_j^{k+1} \leq \omega_j^k$.

Let $k \in \mathbb{N}$ and $j \in \{1, \ldots, p\}$ be fixed. First recall from (1.19) that

$$\forall j \in \{1, \ldots, p\}, \qquad \omega_j^{k+1} = \omega_j^k e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)},$$

where

$$\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k) = \frac{1}{m_k}\sum_{l=1}^{m_k} \boldsymbol{J}'_{\nu}(\boldsymbol{t}_j^k, Z_l^{k+1}) = \frac{1}{m_k}\sum_{l=1}^{m_k}\Big(\|\nu_k\|_{\mathrm{TV}}\, \boldsymbol{g}_{\boldsymbol{t}_j^k, T_l^{k+1}}(U_l^{k+1}) - \boldsymbol{h}_{\boldsymbol{t}_j^k}(V_l^{k+1}) + \lambda\Big).$$

In particular, we have

$$\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k) \geq \omega_j^k \|\mathbf{g}\|_{\mathrm{Inf}} - \|\mathbf{h}\|_\infty + \lambda. \tag{5.1}$$

<u>1^{st} case:</u> If we have $\lambda - \|\mathbf{h}\|_\infty \geq 0$, then $\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k) \geq 0$ and $(\omega_j^k)_{k\geq 0}$ is a decreasing sequence, which yields

$$\forall k \geq 1 \qquad \|\nu_k\|_{TV} \leq \|\nu_0\|_{TV}.$$

<u>2^{nd} case:</u> If we have $\lambda - \|\mathbf{h}\|_\infty \leq 0$, then (5.1) together with (1.19) entails that

$$\omega_j^k \geq \frac{\|\mathbf{h}\|_\infty - \lambda}{\|\mathbf{g}\|_{\mathrm{Inf}}} \quad \Longleftrightarrow \quad \omega_j^k \|\mathbf{g}\|_{\mathrm{Inf}} - \|\mathbf{h}\|_\infty + \lambda \geq 0 \Longrightarrow \omega_j^{k+1} \leq \omega_j^k,$$

whereas

$$\omega_j^k \leq \frac{\|\mathbf{h}\|_\infty - \lambda}{\|\mathbf{g}\|_{\text{Inf}}} \Longrightarrow \omega_j^{k+1} \leq \frac{\|\mathbf{h}\|_\infty - \lambda}{\|\mathbf{g}\|_{\text{Inf}}} e^{\|\mathbf{h}\|_\infty - \lambda}.$$

Finally, we end the proof using an induction argument. □

5.2 Proof of Theorem 1.2

5.2.1 The shadow sequence

Consider an integer $k \geq 0$ and the map $\mathcal{T}_{k+1}$ defined as:

$$\forall t \in \mathcal{X} \qquad \mathcal{T}_{k+1}(t) = \pi_{\mathcal{X}}(t - \eta \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)). \tag{5.2}$$

The sequence of maps $(\mathcal{T}_{k+1})_{k\geq 0}$ only acts on the positions of $\mathcal{X}$ and is built with the random sequence $(\nu_k)_{k\geq 1}$.

Using $(\mathcal{T}_{k+1})_{k\geq 0}$, we then define the shadow sequence $(\nu_k^\varepsilon)_{k\geq 1}$ obtained through an iterative push-forward from a given initialisation measure $\nu_0^\varepsilon \in \mathcal{M}(\mathcal{X})_+$ with the sequence of maps $(\mathcal{T}_k)_{k\geq 1}$. More formally, we set-up this definition in an iterative way:

$$\nu_{k+1}^\varepsilon = \mathcal{T}_{k+1}^{\#}(\nu_k^\varepsilon) \qquad \forall k \in \mathbb{N}^\star, \tag{5.3}$$

where, for any continuous function ψ,

$$\int_{\mathcal{X}} \psi d\mathcal{T}_{k+1}^{\#}(\nu) = \int \psi(\mathcal{T}_{k+1}(.)) d\nu \quad \forall \nu \in \mathcal{M}(\mathcal{X}).$$

The measure ν_0^ε will be defined carefully at the very end of our study. Roughly speaking, the shadow sequence $(\nu_k^\varepsilon)_{k\geq 1}$ moves exactly like $(\nu_k)_{k\geq 1}$ and will share the same support, but the weights on the particles for the sequence $(\nu_k^\varepsilon)_{k\geq 1}$ will be optimised to allow for a good approximation of $\mu^\star$. In particular, if we decompose the initial measure ν_0 as

$$\nu_0 = \sum_{j=1}^p \omega_j \delta_{\boldsymbol{t}_j^0},$$

then we can write $\nu_0^\varepsilon = \nu(\mathbb{W}^\varepsilon, \boldsymbol{t}^0)$, so that:

$$\nu_0^\varepsilon = \sum_{j=1}^p \omega_j^\varepsilon \delta_{\boldsymbol{t}_j^0},$$

for some weights $(\omega_j^\varepsilon)_{j=1..p}$ that will be chosen in an appropriate way.

5.2.2 Excess risk decomposition

The starting point is Proposition B.1 that is used with $\nu = \nu_k$ and $\sigma = \mu^\star - \nu_k$. We write:

$$\begin{aligned} J(\nu_k) - J(\mu^\star) &= \int_{\mathcal{X}} J'_{\nu_k} \mathrm{d}[\nu_k - \mu^\star] - \frac{1}{2}\|\Phi(\mu^\star - \nu_k)\|_{\mathbb{H}}^2 \\ &= \underbrace{\int_{\mathcal{X}} J'_{\nu_k} \mathrm{d}[\nu_k - \nu_k^\varepsilon]}_{:=①} + \underbrace{\int_{\mathcal{X}} J'_{\nu_k} \mathrm{d}[\nu_k^\varepsilon - \mu^\star]}_{:=②} - \frac{1}{2}\|\Phi(\mu^\star - \nu_k)\|_{\mathbb{H}}^2 \end{aligned} \tag{5.4}$$

where $(\nu_k^\varepsilon)_{k\geq 1}$ is the auxiliary shadow sequence of measures introduced in (5.3). First, we establish that the mirror descent adapts the weights of $(\nu_k)_{k\geq 1}$ to those of the shadow sequence $(\nu_k^\varepsilon)_{k\geq 1}$. For any $\mu_1, \mu_2 \in \mathcal{M}(\mathcal{X})_+$, we introduce the following entropy:

$$\mathcal{H}(\mu_1, \mu_2) = -\int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right) \mathrm{d}\mu_2 - \|\mu_2\|_{\mathrm{TV}} + \|\mu_1\|_{\mathrm{TV}}. \tag{5.5}$$

The next proposition focuses on the first term of Equation (5.4).

Proposition 5.1. *Term* $\textcircled{1}$ *of Equation* (5.4) *may be decomposed as:*

$$\textcircled{1} = \int_{\mathcal{X}} J'_{\nu_k} d[\nu_k - \nu_k^{\varepsilon}] = \frac{1}{\alpha}\left[\mathcal{H}(\nu_k, \nu_k^{\varepsilon}) - \mathcal{H}(\nu_{k+1}, \nu_{k+1}^{\varepsilon})\right] \\ + \frac{1}{\alpha}\sum_{j=1}^{p} \omega_j^k \left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k) + e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(t_j^k)} - 1\right] + \sum_{j=1}^{p} \omega_j^{\varepsilon} \widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k)),$$

where for all $k \in \{1, \dots, K\}$ *and for all* $j \in \{1, \dots, p\}$

$$\widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k) := \frac{1}{m_k}\sum_{l=1}^{m_k} \xi_{\nu_k}(\boldsymbol{t}_j^k, Z_l^{k+1}).$$

Proof. Since both measures ν_k and ν_k^{ε} share the same particle locations $\boldsymbol{t}^k$, we can remark that

$$\textcircled{1} = \int_{\mathcal{X}} J'_{\nu_k} \mathrm{d}[\nu_k - \nu_k^{\varepsilon}] = \sum_{j=1}^{p} [\omega_j^k - \omega_j^{\varepsilon}] J'_{\nu_k}(\boldsymbol{t}_j^k) \tag{5.6}$$

We then observe from Equation (1.19) that:

$$\omega_j^{k+1} = \omega_j^k e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k} \quad \Rightarrow \quad \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k) = -\frac{1}{\alpha}\log\left(\frac{\omega_j^{k+1}}{\omega_j^k}\right).$$

Using now Equation (1.15), we observe that:

$$J'_{\nu_k}(\boldsymbol{t}_j^k) = -\frac{1}{\alpha}\log\left(\frac{\omega_j^{k+1}}{\omega_j^k}\right) - \frac{1}{m_k}\sum_{l=1}^{m_k} \xi_{\nu_k}(\boldsymbol{t}_j^k, Z_l^{k+1}).$$

We then use the previous equality in (5.6) and obtain that:

$$\begin{aligned} \textcircled{1} &= \sum_{j=1}^{p}\left(\omega_j^k J'_{\nu_k}(\boldsymbol{t}_j^k) - \omega_j^{\varepsilon}\left[-\frac{1}{\alpha}\log\left(\frac{\omega_j^{k+1}}{\omega_j^k}\right) - \widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k)\right]\right), \\ &= \sum_{j=1}^{p}\left(\omega_j^k J'_{\nu_k}(\boldsymbol{t}_j^k) + \omega_j^{\varepsilon}\frac{1}{\alpha}\log\left(\frac{\omega_j^{k+1}}{\omega_j^k}\right)\right) + \sum_{j=1}^{p}\omega_j^{\varepsilon}\widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k) \\ &= \frac{1}{\alpha}\sum_{j=1}^{p}\left(\alpha\omega_j^k J'_{\nu_k}(\boldsymbol{t}_j^k) + \omega_j^{\varepsilon}\left[\log\left(\frac{\omega_j^{k+1}}{\omega_j^{\varepsilon}}\right) - \log\left(\frac{\omega_j^k}{\omega_j^{\varepsilon}}\right)\right]\right) + \sum_{j=1}^{p}\omega_j^{\varepsilon}\widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k). \end{aligned}$$

We use the entropy $\mathcal{H}$ introduced in Equation (5.5) and deduce that:

$$\begin{aligned} \textcircled{1} &= \frac{1}{\alpha}\sum_{j=1}^{p}\left[\alpha\omega_j^k J'_{\nu_k}(\boldsymbol{t}_j^k) + \omega_j^{k+1} - \omega_j^k\right] + \frac{\mathcal{H}(\nu_k, \nu_k^{\varepsilon}) - \mathcal{H}(\nu_{k+1}, \nu_{k+1}^{\varepsilon})}{\alpha} + \sum_{j=1}^{p}\omega_j^{\varepsilon}\widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k) \\ &= \frac{1}{\alpha}\sum_{j=1}^{p}\omega_j^k\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k) + e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(t_j^k)} - 1\right] + \frac{\mathcal{H}(\nu_k, \nu_k^{\varepsilon}) - \mathcal{H}(\nu_{k+1}, \nu_{k+1}^{\varepsilon})}{\alpha} + \sum_{j=1}^{p}\omega_j^{\varepsilon}\widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k). \end{aligned}$$

We then obtain the conclusion of the proof. □

Now, we study the second term of Equation (5.4), which is an "approximation" term. We essentially follow the same methodology proposed in Chizat [2022] but we use the specificity of our model to properly analyze this term. For this purpose, we use the BL norm (over functions) and dual norm (over measures) introduced in Chizat [2022], defined as:

$$\forall f : \mathcal{X} \to \mathbb{R} \qquad \|f\|_{BL} = \|f\|_{\infty} + \|f\|_{\mathrm{Lip}}, \tag{5.7}$$

where $\|.\|_\infty$ refers to the supremum norm over $\mathcal{X}$, $\|.\|_{\mathrm{Lip}}$ to the Lipschitz constant for f, and

$$\forall \nu \in \mathcal{M}(\mathcal{X})_+ \qquad \|\nu\|^*_{BL} = \sup_{\|f\|_{BL}\leq 1} \int f \mathrm{d}\nu. \tag{5.8}$$

Using these notation, we can propose a bound on the second term of Equation (5.4) as displayed in the following proposition.

Proposition 5.2. *The approximation term* $\textcircled{2}$ *satisfies:*

$$\forall k \geq 1 \qquad \textcircled{2} = \int_{\mathcal{X}} J'_{\nu_k} d[\nu_k^\varepsilon - \mu^\star] \leq \mathfrak{A}_k \|\nu_k^\varepsilon - \mu^\star\|^*_{BL} \quad a.s.$$

where $\mathfrak{A}_k$ *is given by:*

$$\mathfrak{A}_k = \mathcal{C}_0(\|\nu_k\|_{\mathrm{TV}} + 1) + Lip(\varphi)\left[\|\nu_k\|_{\mathrm{TV}}\|\varphi\|_{\infty,\mathbb{H}} + \|y\|_{\mathbb{H}}\right], \tag{5.9}$$

where

$$\mathcal{C}_0 = \max(\lambda + \|\varphi\|_{\infty,\mathbb{H}}\|y\|_{\mathbb{H}}; \|\varphi\|^2_{\infty,\mathbb{H}}).$$

Proof. We can immediately remark that

$$\textcircled{2} = \int_{\mathcal{X}} J'_{\nu_k} \mathrm{d}[\nu_k^\varepsilon - \mu^\star] \leq \|J'_{\nu_k}\|_{BL}\|\nu_k^\varepsilon - \mu^\star\|^*_{BL}.$$

Then, according to Lemma B.1,

$$\|J'_{\nu_k}\|_{BL} = \|J'_{\nu_k}\|_\infty + \|J'_{\nu_k}\|_{\mathrm{Lip}} \leq \mathcal{C}_0(\|\nu_k\|_{\mathrm{TV}} + 1) + \|J'_{\nu_k}\|_{\mathrm{Lip}}.$$

To conclude the proof, we have to propose an upper bound on $\|J'_{\nu_k}\|_{\mathrm{Lip}}$. For any $s, t \in \mathcal{X}$ we have

$$\begin{aligned}
|J'_{\nu_k}(s) - J'_{\nu_k}(t)| &= \left|\sum_{j=1}^p \omega_j^k \langle \varphi_t - \varphi_s, \varphi_{t_j^k}\rangle_{\mathbb{H}} - \langle \varphi_t - \varphi_s, \mathbf{y}\rangle_{\mathbb{H}}\right|, \\
&\leq \sum_{j=1}^p \omega_j^k \|\varphi_t - \varphi_s\|_{\mathbb{H}}\|\varphi_{t_j^k}\|_{\mathbb{H}} + \|\varphi_t - \varphi_s\|_{\mathbb{H}}\|\mathbf{y}\|_{\mathbb{H}}, \\
&\leq Lip(\varphi)\left[\|\nu_k\|_{\mathrm{TV}}\|\varphi\|_{\infty,\mathbb{H}} + \|y\|_{\mathbb{H}}\right] \times \|t - s\|_{\mathcal{X}}.
\end{aligned}$$

The results is obtained by gathering the previous bounds. □

We finally introduce a key term that quantifies the way where $\mu^\star$ can be approximated by a discrete measure. This term, denoted by $\mathcal{Q}$, is defined as

$$\mathcal{Q}_{\mu^\star,\nu_0}(\tau) := \inf_{\mu\in\mathcal{M}(\mathcal{X})_+}\left[\|\mu^\star - \mu\|^*_{BL} + \frac{1}{\tau}\mathcal{H}(\mu,\nu_0)\right] \quad \forall \tau > 0. \tag{5.10}$$

5.2.3 Proof of Theorem 1.2

Proof. The proof is decomposed into three steps.

Step 1: Decomposition of the excess risk with the convexity of J.
We denote by $(\mathfrak{F}_k)_{k\geq 0}$ the natural canonical filtration associated to the sequence of random variables $(Z^k)_{k\geq 0}$. The next upper bound is a consequence of the relationship (5.4) and Propositions 5.1, 5.2. We have

$$\begin{aligned}
J(\nu_k) - J(\mu^\star) \leq{}& \frac{\mathcal{H}(\nu_k,\nu_k^\varepsilon) - \mathcal{H}(\nu_{k+1},\nu_{k+1}^\varepsilon)}{\alpha} + \mathfrak{A}_k\|\nu_k^\varepsilon - \mu^\star\|^*_{BL} \\
&+ \frac{1}{\alpha}\sum_{j=1}^p \omega_j^k\left[\alpha J'_{\nu_k}(\mathbf{t}_j^k) + e^{-\alpha\widehat{J'_{\nu_k}}(t_j^k)} - 1\right] + \sum_{j=1}^p \omega_j^\varepsilon\widehat{\xi_{\nu_k}}(\mathbf{t}_j^k).
\end{aligned}$$

We then use a telescopic sum argument and obtain that:

$$\begin{aligned}\sum_{k=0}^{K}(J(\nu_k)-J(\mu^\star)) \leq \frac{\mathcal{H}(\nu_0,\nu_0^\varepsilon)}{\alpha} &+ \sum_{k=0}^{K}\mathfrak{A}_k\|\nu_k^\varepsilon-\mu^\star\|_{BL}^* \\ &+ \frac{1}{\alpha}\sum_{k=0}^{K}\sum_{j=1}^{p}\omega_j^k\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)+e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}-1\right]+\sum_{k=0}^{K}\sum_{j=1}^{p}\omega_j^\varepsilon\widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k)).\end{aligned}$$

Finally, using the convexity of J, the Cesàro average defined by:

$$\bar{\nu}_K=\frac{1}{K+1}\sum_{k=0}^{K}\nu_k, \tag{5.11}$$

satisfies:

$$\begin{aligned}J(\bar{\nu}_K)-J(\mu^\star) &\leq \frac{\mathcal{H}(\nu_0,\nu_0^\varepsilon)}{\alpha(K+1)}+\frac{\sum_{k=0}^{K}\mathfrak{A}_k\|\nu_k^\varepsilon-\mu^\star\|_{BL}^*}{K+1} \\ &\quad+\frac{\sum_{k=0}^{K}\sum_{j=1}^{p}\omega_j^k\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)+e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}-1\right]}{\alpha(K+1)}+\frac{\sum_{k=0}^{K}\sum_{j=1}^{p}\omega_j^\varepsilon\widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k)}{K+1} \\ &\leq \frac{\mathcal{H}(\nu_0,\nu_0^\varepsilon)}{\alpha K}+\frac{\sum_{k=0}^{K}\mathfrak{A}_k\left[\|\nu_0^\varepsilon-\mu^\star\|_{BL}^*+\sum_{\ell=0}^{k-1}\|\nu_{\ell+1}^\varepsilon-\nu_\ell^\varepsilon\|_{BL}^*\right]}{K} \\ &\quad+\frac{\sum_{k=0}^{K}\sum_{j=1}^{p}\omega_j^k\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)+e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}-1\right]}{\alpha K}+\frac{\sum_{k=0}^{K}\sum_{j=1}^{p}\omega_j^\varepsilon\widehat{\xi_{\nu_k}}(\boldsymbol{t}_j^k)}{K}\end{aligned}$$

where we used the triangle inequality on the telescopic decomposition

$$\nu_k^\varepsilon-\mu^\star=(\nu_k^\varepsilon-\nu_{k-1}^\varepsilon)+(\nu_{k-1}^\varepsilon-\nu_{k-2}^\varepsilon)+\ldots+(\nu_0^\varepsilon-\mu^\star).$$

We then take the expectation and use in particular a standard conditional expectation argument. Since $\mathbb{E}[\xi_{\nu_k}(\boldsymbol{t}_j^k,Z_l^{k+1})|\mathfrak{F}_k]=0$ for any $l\in\{1,\ldots,m_k\}$, we deduce that:

$$\begin{aligned}\mathbb{E}\left[J(\bar{\nu}_K)-J(\mu^\star)\right] \leq \frac{\mathcal{H}(\nu_0,\nu_0^\varepsilon)}{\alpha(K+1)} &+ \overbrace{\|\nu_0^\varepsilon-\mu^\star\|_{BL}^*\frac{\sum_{k=0}^{K}\mathbb{E}[\mathfrak{A}_k]}{K+1}}^{:=A_1}+\overbrace{\frac{\sum_{k=0}^{K}\mathbb{E}\left[\mathfrak{A}_k\sum_{\ell=0}^{k-1}\|\nu_{\ell+1}^\varepsilon-\nu_\ell^\varepsilon\|_{BL}^*\right]}{K+1}}^{:=A_2} \\ &+\underbrace{\frac{\sum_{k=0}^{K}\sum_{j=1}^{p}\mathbb{E}\left[\omega_j^k\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)+e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}-1\right]\right]}{\alpha(K+1)}}_{:=A_3}.\end{aligned} \tag{5.12}$$

Step 2a: Study of A_1. We use the definition of $\mathfrak{A}_k$ in Equation (5.9) and observe that $\mathfrak{A}_k\leq\mathfrak{C}(1+\|\nu_k\|_{\mathrm{TV}})$. We then use Proposition 1.1 and conclude that:

$$\mathbb{E}[A_1]=\|\nu_0^\varepsilon-\mu^\star\|_{BL}^*\frac{\sum_{k=0}^{K}\mathbb{E}[\mathfrak{A}_k]}{K+1}\leq\mathfrak{C}R_0\|\nu_0^\varepsilon-\mu^\star\|_{BL}^*. \tag{5.13}$$

Step 2b: Study of A_2. We focus on the shadow sequence that involves $\nu_{\ell+1}^\varepsilon-\nu_\ell^\varepsilon$. Using (2.11), observe that:

$$
\begin{aligned}
\|\nu^{\varepsilon}_{\ell+1}-\nu^{\varepsilon}_{\ell}\|^{*}_{BL} &= \sup_{\|\psi\|_{BL}\leq 1}\int_{\mathcal{X}}\psi(t)\mathrm{d}[\nu^{\varepsilon}_{\ell+1}-\nu^{\varepsilon}_{\ell}](t)\\
&= \sup_{\|\psi\|_{BL}\leq 1}\sum_{j=1}^{p}\omega^{\varepsilon}_{j}[\psi(\boldsymbol{t}_j^{\ell+1})-\psi(\boldsymbol{t}_j^{\ell})]\\
&\leq \sum_{j=1}^{p}\omega^{\varepsilon}_{j}\|\boldsymbol{t}_j^{\ell+1}-\boldsymbol{t}_j^{\ell}\|_{\mathcal{X}}\\
&= \eta\sum_{j=1}^{p}\omega^{\varepsilon}_{j}\|\widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^{k}\|_{\mathcal{X}}\\
&\leq \mathfrak{C}\eta\sum_{j=1}^{p}\omega^{\varepsilon}_{j}(1+\|\nu_\ell\|_{\mathrm{TV}})\\
&= \mathfrak{C}\eta\|\nu^{\varepsilon}_0\|_{\mathrm{TV}}(1+\|\nu_\ell\|_{\mathrm{TV}}), \qquad (5.14)
\end{aligned}
$$

where we used the almost sure upper bound in Proposition C.1 and the fact that $\pi_{\mathcal{X}}$ is 1-Lipschitz. A simple sum yields:

$$
\begin{aligned}
\frac{\sum_{k=0}^{K}\mathfrak{A}_k\sum_{\ell=0}^{k-1}\|\nu^{\varepsilon}_{\ell+1}-\nu^{\varepsilon}_{\ell}\|^{*}_{BL}}{K+1} &\leq \mathfrak{C}\eta\|\nu^{\varepsilon}_0\|_{\mathrm{TV}}\frac{\sum_{k=0}^{K}\mathfrak{A}_k\sum_{\ell=1}^{k}(1+\|\nu_\ell\|_{\mathrm{TV}})}{K+1}\\
&\leq \mathfrak{C}\eta\|\nu^{\varepsilon}_0\|_{\mathrm{TV}}\frac{\sum_{k=0}^{K}(1+\|\nu_k\|_{\mathrm{TV}})\sum_{\ell=1}^{k}(1+\|\nu_\ell\|_{\mathrm{TV}})}{K+1},
\end{aligned}
$$

for $\mathfrak{C}$ large enough. We apply Proposition 1.1 and compute the expectation of the previous term. We then observe that:

$$
\mathbb{E}[A_2]=\frac{\sum_{k=0}^{K}\mathbb{E}\left[\mathfrak{A}_k\sum_{\ell=0}^{k-1}\|\nu^{\varepsilon}_{\ell+1}-\nu^{\varepsilon}_{\ell}\|^{*}_{BL}\right]}{K+1}\leq \mathfrak{C}R_0^2\eta\|\nu^{\varepsilon}_0\|_{\mathrm{TV}}K. \qquad (5.15)
$$

<u>Step 2c: Study of A_3.</u> This last term deserves a specific study. We use a conditional expectation argument and observe that, for any $k\in\{0,\dots,K\}$, $j\in\{1,\dots,p\}$,

$$
\mathbb{E}\left[\omega_j^k\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)+e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}-1\right]\middle|\mathfrak{F}_k\right]=\omega_j^k\mathbb{E}\left[\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)+e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}-1\right]\middle|\mathfrak{F}_k\right].
$$

We then apply Proposition C.2 ang get:

$$
\mathbb{E}\left[\omega_j^k\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)+e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}-1\right]\middle|\mathfrak{F}_k\right]\leq \mathfrak{C}\omega_j^k\alpha^2(1+\|\nu_k\|^2_{\mathrm{TV}}).
$$

We then sum the previous upper bounds from 0 to K with a global expectation and Proposition 1.1. We obtain that:

$$
\mathbb{E}[A_3]=\frac{\sum_{k=0}^{K}\sum_{j=1}^{p}\mathbb{E}\left[\omega_j^k\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)+e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)}-1\right]\right]}{\alpha(K+1)}\leq \mathfrak{C}\alpha R_0^3. \qquad (5.16)
$$

<u>Step 3: End of the proof.</u>
We gather Equations (5.13), (5.15), (5.16) and obtain that:

$$
\mathbb{E}\left[J(\bar{\nu}_K)-J(\mu^\star)\right]\leq\frac{\mathcal{H}(\nu_0,\nu_0^{\varepsilon})}{\alpha K}+\mathfrak{C}R_0\|\nu_0^{\varepsilon}-\mu^\star\|^{*}_{BL}+\mathfrak{C}R_0^2\left[\eta\|\nu_0^{\varepsilon}\|_{\mathrm{TV}}K+\alpha R_0\right].
$$

We then use the definition of $\mathcal{Q}$ given in Equation (5.10) and observe that if ν_0^{ε} is chosen in an optimal way, as given in Proposition C.3, then:

$$
\begin{aligned}
\mathbb{E}\left[J(\bar{\nu}_K)-J(\mu^\star)\right] &\leq R_0\mathcal{Q}_{\mu^\star,\nu_0}(\alpha K R_0)+\mathfrak{C}R_0^2\left[\eta\|\mu^\star\|_{\mathrm{TV}}K+\alpha R_0\right], \qquad (5.17)\\
&\leq \mathfrak{C}\|\mu^\star\|_{\mathrm{TV}}\left[\frac{d\left(1+\log\frac{\alpha K R_0}{2d}+\frac{\log|\mathcal{X}|}{d}\right)}{\alpha K}+\frac{\alpha R_0^3}{\|\mu^\star\|_{\mathrm{TV}}}+R_0^2\eta K\right],
\end{aligned}
$$

where we have used the relationship $\|\nu_0^\varepsilon\|_{\mathrm{TV}} = \|\mu^\star\|_{\mathrm{TV}}$. The choices

$$\alpha = \sqrt{\frac{d\|\mu^\star\|_{\mathrm{TV}}}{R_0^3 K}} \quad \text{and} \quad \eta = \sqrt{\frac{dR_0}{K^3\|\mu^\star\|_{\mathrm{TV}}}}$$

then leads to

$$\mathbb{E}\left[J(\bar{\nu}_K) - J(\mu^\star)\right] \leq \mathfrak{C}\sqrt{\frac{d\|\mu^\star\|_{\mathrm{TV}} R_0^3}{K}}\left[\log(d\|\mu^\star\|_{\mathrm{TV}} R_0^3 K) + \frac{\log(|\mathcal{X}|)}{d}\right].$$

□

5.3 Proof of Theorem 1.3

Proof. The proof is slitted into three parts, and relies on a contraction argument with conditional expectation. In what follows, we will choose α such that $\alpha\mathcal{C}_1(R_0+1) < 1$ where $\mathcal{C}_1$ is involved in Lemma B.1 and R_0 has been introduced in Proposition 1.1. We shall often use the inequality $|e^h - 1| \leq 2|h|$ which is valid when $|h| \leq 1$. We will frequently apply this inequality with $h = -\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)$.

Step 1: One-step evolution and second order term. Let $k \in \mathbb{N}^\star$ be fixed. According to Proposition B.1:

$$J(\nu_{k+1}) - J(\nu_k) = \int_{\mathcal{X}} J'_{\nu_k} d(\nu_{k+1} - \nu_k) + \frac{1}{2}\|\Phi(\nu_{k+1} - \nu_k)\|_{\mathbb{H}}^2. \tag{5.18}$$

Introducing the measure $\tilde{\nu}_{k+1} = \sum_{j=1}^p \omega_j^{k+1}\delta_{\boldsymbol{t}_j^k}$, we deduce that:

$$\begin{aligned} \|\Phi(\nu_{k+1} - \nu_k)\|_{\mathbb{H}}^2 &= \|\Phi(\nu_{k+1} - \tilde{\nu}_{k+1} + \tilde{\nu}_{k+1} - \nu_k)\|_{\mathbb{H}}^2, \\ &\leq 2\|\Phi(\nu_{k+1} - \tilde{\nu}_{k+1})\|_{\mathbb{H}}^2 + 2\|\Phi(\tilde{\nu}_{k+1} - \nu_k)\|_{\mathbb{H}}^2 \end{aligned}$$

First remark that,

$$\begin{aligned} \|\Phi(\nu_{k+1} - \tilde{\nu}_{k+1})\|_{\mathbb{H}}^2 &= \left\|\sum_{j=1}^p \omega_j^{k+1}(\varphi_{\boldsymbol{t}_j^{k+1}} - \varphi_{\boldsymbol{t}_j^k})\right\|_{\mathbb{H}}^2, \\ &\leq \sum_{j=1}^p \omega_j^{k+1} \times \sum_{j=1}^p \omega_j^{k+1}\|\varphi_{\boldsymbol{t}_j^{k+1}} - \varphi_{\boldsymbol{t}_j^k}\|_{\mathbb{H}}^2, \\ &\leq Lip(\varphi)\|\nu_{k+1}\|_{\mathrm{TV}}\sum_{j=1}^p \omega_j^{k+1}\|\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k\|^2. \end{aligned}$$

According to (1.21) and (2.10), we obtain that

$$\|\Phi(\nu_{k+1} - \tilde{\nu}_{k+1})\|_{\mathbb{H}}^2 \leq Lip(\varphi)\|\nu_{k+1}\|_{\mathrm{TV}}\ \eta^2\sum_{j=1}^p \omega_j^{k+1}\|\widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)\|^2. \tag{5.19}$$

In the same time, using (1.19),

$$\begin{aligned} \|\Phi(\tilde{\nu}_{k+1} - \nu_k)\|_{\mathbb{H}}^2 &= \left\|\sum_{j=1}^p (\omega_j^{k+1} - \omega_j^k)\varphi_{\boldsymbol{t}_j^k}\right\|_{\mathbb{H}}^2, \\ &= \left\|\sum_{j=1}^p \omega_j^k(e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)} - 1)\varphi_{\boldsymbol{t}_j^k}\right\|_{\mathbb{H}}^2, \\ &\leq \sum_{j=1}^p \omega_j^k \times \sum_{j=1}^p \omega_j^k(e^{-\alpha\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)} - 1)^2\|\varphi_{\boldsymbol{t}_j^k}\|_{\mathbb{H}}^2, \end{aligned}$$

where the last line comes from the Jensen inequality. We then observe from Lemma B.1 that the terms $J'_{\nu_k}(\boldsymbol{t}_j^k, Z_l^{k+1})$ are bounded. Hence, provided $\alpha \mathcal{C}_1(R_0+1) < 1$, a constant C_φ large enough exists such that:

$$\|\Phi(\tilde{\nu}_{k+1} - \nu_k)\|_{\mathbb{H}}^2 \leq C_\varphi \|\nu_k\|_{\mathrm{TV}} \alpha^2 \sum_{j=1}^p \omega_j^k |\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)|^2. \tag{5.20}$$

Gathering Equations (5.19) and (5.20), we then deduce that

$$\|\Phi(\nu_{k+1} - \nu_k)\|_{\mathbb{H}}^2 \leq C_\varphi \left(\eta^2 \|\nu_{k+1}\|_{\mathrm{TV}} \sum_{j=1}^p \omega_j^{k+1} \|\widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)\|^2 + \alpha^2 \|\nu_k\|_{\mathrm{TV}} \sum_{j=1}^p \omega_j^k |\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)|^2 \right). \tag{5.21}$$

Step 2: Study of the drift first order term. We expand the first order term and observe that:

$$\begin{aligned}
\int_{\mathcal{X}} J'_{\nu_k} d(\nu_{k+1} - \nu_k) = &\sum_{j=1}^p \left[(\omega_j^{k+1} - \omega_j^k) J'_{\nu_k}(\boldsymbol{t}_j^k) + \omega_j^k (J'_{\nu_k}(\boldsymbol{t}_j^{k+1}) - J'_{\nu_k}(\boldsymbol{t}_j^k)) \right] \\
&+ \sum_{j=1}^p (\omega_j^{k+1} - \omega_j^k)(J'_{\nu_k}(\boldsymbol{t}_j^{k+1}) - J'_{\nu_k}(\boldsymbol{t}_j^k)), \\
= &\sum_{j=1}^p \left[(\omega_j^{k+1} - \omega_j^k) J'_{\nu_k}(\boldsymbol{t}_j^k) + \omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k) \rangle \right] \\
&+ \sum_{j=1}^p \left[\omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla^2 J'_{\nu_k}(\upsilon_j^k)(\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k) \rangle + (\omega_j^{k+1} - \omega_j^k) \langle \nabla J'_{\nu_k}(\tilde{\upsilon}_j^k), \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k \rangle \right],
\end{aligned}$$

where υ_j^k and $\tilde{\upsilon}_j^k$ are some auxiliary points that belong to $(\boldsymbol{t}_j^k, \boldsymbol{t}_j^{k+1})$ obtained with the help of first and second order Taylor expansions. Using Proposition C.1, we deduce that:

$$\begin{aligned}
\int_{\mathcal{X}} J'_{\nu_k} d(\nu_{k+1} - \nu_k) \leq &\sum_{j=1}^p \left[(\omega_j^{k+1} - \omega_j^k) J'_{\nu_k}(\boldsymbol{t}_j^k) + \omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k) \rangle \right] \\
&+ \sum_{j=1}^p \omega_j^k \|\nabla^2 J'_{\nu_k}\|_{\infty,op} \|\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k\|^2 \\
&+ \sum_{j=1}^p |\omega_j^{k+1} - \omega_j^k| \times \|\nabla J'_{\nu_k}\| \|\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k\| \,.
\end{aligned}$$

Using the total variation upper bound stated in Proposition 1.1 by R_0, we then define for the sake of readability the constant A as:

$$A = (\|\nu\|_{\mathrm{TV}} \|\varphi\|_{\infty,\mathbb{H}} + \|y\|_{\mathbb{H}}) \|\nabla^2 \varphi\|_{\infty,op} \vee (\|\nu_k\|_{\mathrm{TV}} \|\varphi\|_{\infty,\mathbb{H}} + \|y\|_{\mathbb{H}}) \|\varphi'\|_{\mathbb{H}}.$$

We then derive:

$$\begin{aligned}
\int_{\mathcal{X}} J'_{\nu_k} d(\nu_{k+1} - \nu_k) \leq &\sum_{j=1}^p (\omega_j^{k+1} - \omega_j^k) J'_{\nu_k}(\boldsymbol{t}_j^k) + \omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k) \rangle \\
&+ A \sum_{j=1}^p \left[\omega_j^k \|\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k\|^2 + (\omega_j^{k+1} - \omega_j^k) \|\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k\| \right].
\end{aligned}$$

We now use the surrogate update (1.21) on the previous inequality and obtain that:

$$\begin{aligned}
\int_{\mathcal{X}} J'_{\nu_k} d(\nu_{k+1} - \nu_k) \leq &\sum_{j=1}^p \omega_j^k (e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)} - 1) J'_{\nu_k}(\boldsymbol{t}_j^k) + \omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k) \rangle \\
&+ A \sum_{j=1}^p \omega_j^k \|\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k\|^2 + \omega_j^k (e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)} - 1) \|\boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k\|.
\end{aligned}$$

We pay a specific attention to the second term of the right hand side. Using the generalized projected gradient introduced in Section 2.2 and in particular (2.10), we get for any $j \in \{1, \dots, p\}$,

$$\begin{aligned}
&\omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\rangle \\
&\quad = -\eta\omega_j^k \left\langle P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta), \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)\right\rangle + \eta\omega_j^k \left\langle P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta), \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\rangle \\
&\quad \le -\eta\omega_j^k \left\| P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta)\right\|^2 + \eta\omega_j^k \left\langle P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta), \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\rangle \\
&\quad = -\eta\omega_j^k \left\| P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta)\right\|^2 + \eta\omega_j^k \left\langle P_{\mathcal{X}}(\boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k), \eta), \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\rangle \\
&\qquad + \eta\omega_j^k \left\langle P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta) - P_{\mathcal{X}}(\boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k), \eta), \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\rangle
\end{aligned}$$

where for the first inequality, we have used Lemma 2.1 and simple algebraic manipulations hereafter. Using that $P_{\mathcal{X}}$ is 1-Lip (see Lemma 2.2) and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\rangle &\le -\eta\omega_j^k \left\| P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta)\right\|^2 \\
&\quad + \eta\omega_j^k \left\langle P_{\mathcal{X}}(\boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k), \eta), \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\rangle + \eta\omega_j^k \left\| \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\|^2.
\end{aligned}$$

Finally, for any $j \in \{1, \dots, p\}$, we have

$$\begin{aligned}
\mathbb{E}\left[\omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\rangle | \mathfrak{F}_k\right] &\le -\eta\omega_j^k \mathbb{E}\left[\left\| P_{\mathcal{X}}(\boldsymbol{t}_j^k, \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k), \eta)\right\|^2 | \mathfrak{F}_k\right] + \eta\omega_j^k \mathbb{E}\left[\left\| \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\|^2 | \mathfrak{F}_k\right], \\
&\le -\eta\omega_j^k \mathbb{E}\left[\left\| \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k)\right\|^2 | \mathfrak{F}_k\right] + \eta\omega_j^k \mathbb{E}\left[\left\| \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\|^2 | \mathfrak{F}_k\right], \\
&\le -\eta\omega_j^k \mathbb{E}\left[\left\| \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\|^2 | \mathfrak{F}_k\right] + 2\eta\omega_j^k \mathbb{E}\left[\left\| \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\|^2 | \mathfrak{F}_k\right].
\end{aligned}$$

At this step, we can take advantage of the mini-batch step. Indeed, according to (1.18), we have

$$\begin{aligned}
\mathbb{E}\left[\left\| \widehat{\boldsymbol{D}_{\nu_k}}(\boldsymbol{t}_j^k) - \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\right\|^2 | \mathfrak{F}_k\right] &= \frac{1}{m_k^2}\sum_{l=1}^{m_k} \mathbb{E}\left[\left\| D_{\nu_k}(\boldsymbol{t}_j^k, Z_l^k)\right\|^2 | \mathfrak{F}_k\right], \\
&\le \frac{\|\nu_k\|_{\mathrm{TV}}\|\mathbf{g}'\|_{\infty,\mathbb{H}} + \|\mathbf{h}'\|_{\mathbb{H}}}{m_k}.
\end{aligned}$$

Hence

$$\mathbb{E}\left[\omega_j^k \langle \boldsymbol{t}_j^{k+1} - \boldsymbol{t}_j^k, \nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\rangle | \mathfrak{F}_k\right] \le 2\eta\omega_j^k \frac{\|\nu_k\|_{\mathrm{TV}}\|\mathbf{g}'\|_{\infty,\mathbb{H}} + \|\mathbf{h}'\|_{\mathbb{H}}}{m_k}.$$

We then consider the conditional expectation at time k and apply Proposition C.1 (to upper bound some rest terms) and Proposition C.2 (to control the drift at iteration k). We deduce that a large enough $\mathfrak{C}$ such that:

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathcal{X}} J'_{\nu_k} d(\nu_{k+1} - \nu_k) | \mathfrak{F}_k\right] \\
&\le -\alpha \sum_{j=1}^p \omega_j^k J'_{\nu_k}(\boldsymbol{t}_j^k)^2 + \mathfrak{C}\alpha^2 \sum_{j=1}^p \omega_j^k (1 + \|\nu_k\|_{\mathrm{TV}})^3 - \eta \sum_{j=1}^p \omega_j^k \|\nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\|^2 \\
&\qquad + \mathfrak{C}\eta^2 \sum_{j=1}^p \omega_j^k (1 + \|\nu_k\|_{\mathrm{TV}})^2 + \mathfrak{C}\eta\alpha \sum_{j=1}^p \omega_j^k (1 + \|\nu_k\|_{\mathrm{TV}})^3 + \mathfrak{C}\frac{\eta}{m_k}\sum_{j=1}^p \omega_j^k, \\
&\le -\alpha \sum_{j=1}^p \omega_j^k J'_{\nu_k}(\boldsymbol{t}_j^k)^2 - \eta \sum_{j=1}^p \omega_j^k \|\nabla J'_{\nu_k}(\boldsymbol{t}_j^k)\|^2 \\
&\qquad + \mathfrak{C}\|\nu_k\|_{\mathrm{TV}}(1 + \|\nu_k\|_{\mathrm{TV}})^3 \left(\alpha^2 + \eta^2 + \frac{\eta}{m_k}\right),
\end{aligned}$$

where in the last line we used the Young inequality $2\eta\alpha \leq \alpha^2 + \eta^2$ and some rough upper bounds on the rest terms. We now associate this last inequality with Equations (5.21) and obtain the descent property:

$$\mathbb{E}\left[J(\nu_{k+1})\middle|\mathfrak{F}_k\right] \leq J(\nu_k) - \alpha\|J'_{\nu_k}\|^2_{\nu_k} - \eta\|\nabla J'_{\nu_k}\|^2_{\nu_k} + \mathfrak{C}(1+R_0^4)\left(\alpha^2+\eta^2+\frac{\eta}{m_k}\right) \tag{5.22}$$

Step 3: Conclusion of the proof. The rest of the proof proceeds with a standard argument. We use a telescopic sum + conditional expectation strategy and observe that:

$$\alpha\sum_{k=1}^{K}\mathbb{E}[\|J'_{\nu_k}\|^2_{\nu_k}] + \eta\sum_{k=1}^{K}\mathbb{E}[\|\nabla J'_{\nu_k}\|^2_{\nu_k}] \leq J(\nu_0) + \mathfrak{C}(1+R_0^4)\left(K\alpha^2 + K\eta^2 + \eta\sum_{k=1}^{K}\frac{1}{m_k}\right).$$

Choosing $\alpha = \eta$, we deduce that:

$$\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}[\|J'_{\nu_k}\|^2_{\nu_k}] + \mathbb{E}[\|\nabla J'_{\nu_k}\|^2_{\nu_k}]\right) \leq \frac{J(\nu_0)}{\alpha K} + \mathfrak{C}(1+R_0^4)\alpha + \frac{1}{K}\sum_{k=1}^{K}\frac{\mathfrak{C}}{m_k}.$$

Finally, if τ_K refers to a random variable uniformly distributed over $\{1,\dots,K\}$, independent from the sequence $(\nu_k)_{k\geq 1}$, the tuning $\alpha = \eta = 1/\sqrt{K}$ and $m_k = \sqrt{K}$ for all $k \in \{1,\dots,K\}$ yields:

$$\mathbb{E}\left[\|J'_{\nu_{\tau_K}}\|^2_{\nu_{\tau_K}} + \|\nabla J'_{\nu_{\tau_K}}\|^2_{\nu_{\tau_K}}\right] \leq \frac{J(\nu_0) + \mathfrak{C}(1+R_0^4)}{\sqrt{K}}.$$

□

Acknowledgements The authors would like to thank N. Jouvin for its remarks and time on the numerical aspects of this project. They are also in debt with L. Chizat for valuable discussions on CPGD during a seminar at Institut Henri Poincaré.

References

Jean-Marc Azais, Yohann De Castro, and Fabrice Gamboa. Spike detection from inaccurate samplings. Applied and Computational Harmonic Analysis, 38(2):177–195, 2015.

Francis Bach and Lénaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. arXiv preprint arXiv:2110.08084, 2021.

Jérôme Bolte, Laurent Miclo, and Stéphane Villeneuve. Swarm gradient dynamics for global optimization: the mean-field limit case. Mathématical Programming, to appear, 2023.

Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. On representer theorems and convex regularization. SIAM Journal on Optimization, 29(2):1260–1281, 2019.

Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. ESAIM: Control, Optimisation and Calculus of Variations, 19(01):190–218, 2013.

Kristian Bredies, Marcello Carioni, Silvio Fanzon, and Daniel Walter. Asymptotic linear convergence of fully-corrective generalized conditional gradient methods. Mathematical programming, 205(1):135–202, 2024.

Haim Brézis. Functional analysis, Sobolev spaces and partial differential equations, volume 2. Springer, 2011.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. Foundations and Trends® in Machine Learning, 8(3-4):231–357, 2015.

Emmanuel J. Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. Journal of Fourier Analysis and Applications, 19(6):1229–1254, 2013.

Émmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. Communications on pure and applied Mathematics, 67(6):906–956, 2014.

Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. Mathematical Programming, 194(1-2):487–532, 2022.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. Advances in neural information processing systems, 31, 2018.

Yohann De Castro and Fabrice Gamboa. Exact reconstruction using beurling minimal extrapolation. Journal of Mathematical Analysis and applications, 395(1):336–354, 2012.

Yohann De Castro, Sébastien Gadat, Clément Marteau, and C Maugis-Rabusseau. Supermix: sparse regularization for mixtures. The Annals of Statistics, 49(3):1779–1809, 2021.

Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. The Annals of Statistics, 27(1):94–128, 1999.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding frank-wolfe algorithm and its application to super-resolution microscopy. Inverse Problems, 2019.

David L. Donoho. Superresolution via sparsity constraints. SIAM Journal on Mathematical Analysis, 23(5): 1309–1331, 1992.

Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. Foundations of Computational Mathematics, 15(5):1315–1355, 2015.

Carlos Fernandez-Granda. Support detection in super-resolution. In The 10th International Conference on Sampling Theory and Applications (SampTA 2013), pages 145–148, 2013.

Fabrice Gamboa and E Gassiat. Sets of superresolution and the maximum entropy method on the mean. SIAM journal on mathematical analysis, 27(4):1129–1152, 1996.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math. Program., 155(1–2):267–305, January 2016.

Evarist Giné and Richard Nickl. Mathematical foundations of infinite-dimensional statistical models. Cambridge university press, 2021.

Bernd Hofmann, Barbara Kaltenbacher, Christiane Poeschl, and Otmar Scherzer. A convergence rates result for tikhonov regularization in banach spaces with non-smooth operators. Inverse Problems, 23(3): 987, 2007.

Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. Sketching for large-scale learning of mixture models. Information and Inference: A Journal of the IMA, 7(3):447–508, 2018.

Frederik Kunstner, Raunak Kumar, and Mark Schmidt. Homeomorphic-invariance of em: Non-asymptotic convergence in kl divergence for exponential families via mirror descent. In Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pages 3295–3303. PMLR, 13–15 Apr 2021.

Guanghui Lan, Arkadij Semenovič Nemirovskij, and Alexander Shapiro. Validation analysis of mirror descent stochastic approximation method. Mathematical programming, 134(2):425–458, 2012.

Laurent Miclo. On the convergence of global-optimization fraudulent stochastic algorithms. Preprint, 2023.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning, 10 (1-2):1–141, 2017.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. Wiley-Interscience, 1983.

Clarice Poon, Nicolas Keriven, and Gabriel Peyré. The geometry of off-the-grid compressed sensing. Foundations of Computational Mathematics, pages 1–87, 2021.

Walter Rudin. Real and complex analysis. Mcgraw hill International Book Company, 1974.

Ingo Steinwart and Andreas Christmann. Support vector machines. Springer Science & Business Media, 2008.

Gongguo Tang, Badri N. Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. Information Theory, IEEE Transactions on, 59(11):7465–7490, 2013.

Gongguo Tang, Badri N. Bhaskar, and Benjamin Recht. Near minimax line spectral estimation. Information Theory, IEEE Transactions on, 61(1):499–512, 2015.

Appendix of “FastPart: Over-Parameterized Stochastic Gradient Descent for Sparse optimization on Measures"

A Reformulations, proofs and technical lemmas

A.1 The non-separable case and the interesting reformulation of the objective

There is a little subtlety on the properties that one should require on $\mathbb{H}$. At first glance, we need $\mathbb{H}$ separable to prove Bochner integrability in Lemma A.2. But K is continuous on a compact space $\mathcal{X}$, hence its RKHS is separable [Steinwart and Christmann, 2008, Lemma 4.3.3]. This RKHS is isometric to a separable subspace of $\mathbb{H}$ as proven by the next lemma.

Lemma A.1. *Let $\mathbb{H}$ be Hilbert space and let $\mathcal{X}$ be compact space. Under ($\mathbf{A}_{\mathrm{C}}$), there exists a separable cloded vector subspace $(\mathbb{H}_{\mathcal{F}}, \|\cdot\|_{\mathbb{H}})$ of $\mathbb{H}$ which is isometric to $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$, the RKHS defined by K. Denote by Π the orthogonal projection onto $\mathbb{H}_{\mathcal{F}}$ then for any $\nu \in \mathcal{M}(\mathcal{X})$*

$$J(\nu) = \frac{1}{2}\|\boldsymbol{y} - \Pi(\boldsymbol{y})\|_{\mathbb{H}}^2 + \frac{1}{2}\|\Pi(\boldsymbol{y}) - \Phi(\nu)\|_{\mathbb{H}_{\mathcal{F}}}^2 + \lambda\|\nu\|_{\mathrm{TV}}\,. \tag{A.1}$$

Proof. We denote by $\mathcal{F}$ the RKHS defined by K. By [Steinwart and Christmann, 2008, Theorem 4.21], one has that

$$\mathcal{F} := \Big\{ f : \mathcal{X} \to \mathbb{H} \;:\; \exists h \in \mathbb{H}\,,\ \forall t \in \mathcal{X}\,,\ f(x) = \langle h, \Phi(t)\rangle_{\mathbb{H}} \Big\}\,.$$

is the only RKHS defined by K and

$$||f||_{\mathcal{F}} = \inf \Big\{ ||h||_{\mathbb{H}} \;:\; h \in \mathbb{H} \;\text{ s.t. }\; f(\cdot) = \langle h, \Phi(\cdot)\rangle_{\mathbb{H}} \Big\}\,,$$

Consider the functions

$$f_h \,:\, x \mapsto \langle h,\, \Phi(x)\rangle_{\mathbb{H}}\,, \quad h = \sum_{j=1}^{r} \omega_j \Phi(t_j)$$

defining the pre-dual of $\mathcal{F}$. Observe that

$$\langle f_{h_1},\, f_{h_2}\rangle_{\mathcal{F}} = \langle h_1,\, h_2\rangle_{\mathbb{H}}\,,$$

where we denote by $\langle\cdot,\cdot\rangle_{\mathcal{F}}$, the dot product of $\mathcal{F}$. Define $\mathbb{H}_{\mathcal{F}}$ the vector subspace of $\mathbb{H}$ defined as the closure (in $\mathbb{H}$) of the span of $\Phi(\mathcal{X})$. The aforementioned equality shows that $h \mapsto f_h$ is an isometry from $(\mathbb{H}_{\mathcal{F}}, \|\cdot\|_{\mathbb{H}})$ onto $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$. Since $\mathcal{F}$ is separable [Steinwart and Christmann, 2008, Lemma 4.3.3], we deduce that $\mathbb{H}_{\mathcal{F}}$ is separable. The last statement is a consequence of the Pythagorean theorem. □

Note that in (A.1), the term $\frac{1}{2}\|\boldsymbol{y} - \Pi(\boldsymbol{y})\|_{\mathbb{H}}^2$ is constant. Hence, up to a constant term, and without loss of generality, one can assume that $\mathbb{H}$ is separable.

Remark A.1. *The proof of Lemma A.1 is a consequence of [Steinwart and Christmann, 2008, Theorem 4.21] and we choose to maintain it in this paper for sake of completeness. Moreover, it sheds light on an interesting reformulation of the quadratic term in (1.5), the objective J. Indeed, it holds*

$$J(\nu) = \frac{1}{2}\Big\|(\beth \circ \Pi)(\boldsymbol{y}) - (\beth \circ \Phi)(\nu)\Big\|_{\mathcal{F}}^2 + \lambda\|\nu\|_{\mathrm{TV}}\,, \tag{A.2}$$

up to a constant term and where $\beth$ denotes the isometry between $(\mathbb{H}_{\mathcal{F}}, \|\cdot\|_{\mathbb{H}})$ and $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$.

A.2 Existence of the kernel measure embedding

Kernel mean embedding is a standard notion in Machine Learning, see for instance Muandet et al. [2017]. Extending this notion of measure with finite total variation norm is straightforward. We referred to this notion as *Kernel Measure Embedding* as the two notions coincides on probability measures.

Lemma A.2. *Let $\mathbb{H}$ be separable Hilbert space and let $\mathcal{X}$ be compact metric space. Under ($\mathbf{A}_{\mathrm{C}}$), the operator Φ defined by (1.3) is well defined and bounded linear as a function from $\mathcal{M}(\mathcal{X})$ to $\mathbb{H}$. Furthermore, the dual of Φ is given by*

$$\Phi^\star \,:\, h \in \mathbb{H} \mapsto \big(t \mapsto \langle h, \varphi_t\rangle_{\mathbb{H}}\big) \in \big(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty\big)\,, \tag{A.3}$$

and for any $(h, \nu) \in \mathbb{H} \times \mathcal{M}(\mathcal{X})$,

$$\langle h, \Phi(\nu)\rangle_{\mathbb{H}} = \int_{\mathcal{X}} \langle h, \varphi_t\rangle_{\mathbb{H}} \mathrm{d}\nu(t) = \langle \Phi^\star(h), \nu\rangle_{\mathcal{C}(\mathcal{X}),\mathcal{M}(\mathcal{X})} \leq \sup_t \sqrt{\mathbb{K}(t,t)}\|h\|_{\mathbb{H}}\|\nu\|_{\mathrm{TV}}\,. \tag{A.4}$$

Remark A.2. *A key result of Lemma A.2 is that $\mathrm{Im}(\Phi^\star) \subseteq \big(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty\big)$, this latter being a subset of $\mathcal{M}(\mathcal{X})^\star$, the topological dual of $\mathcal{M}(\mathcal{X})$. Strictly speaking, the dual of Φ maps to the dual of the space of measures and the right manner to expose this results is as it is done by Bredies and Pikkarainen [2013], using the predual operator $\Phi_\star$ which satisfies $(\Phi_\star)^\star = \Phi$ and $\Phi^\star = \imath\Phi_\star$ where $\imath$ denotes the canonical embedding of continuous functions $\big(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty\big)$ into the dual of measures.*

Proof. Let $\nu \in \mathcal{M}(\mathcal{X})$. We say that $t \in \mathcal{X} \mapsto f(x) \in \mathbb{H}$ is *simple* if it is finitely valued, namely

$$f(t) = \sum_{i=1}^n h_i \mathbf{1}_{\{t \in B_i\}}\,,$$

for some $n \geq 1$, $h_i \in \mathbb{H}$, and B_i Borel set of $\mathcal{X}$. In this case, one has

$$\int_{\mathcal{X}} f \mathrm{d}\nu = \sum_{i=1}^n h_i \nu(B_i)\,.$$

Note that $\|\varphi_t\|_{\mathbb{H}} = \sqrt{\mathbb{K}(t,t)}$ and, it holds that

$$\int_{\mathcal{X}} \|\varphi_t\|_{\mathbb{H}} \mathrm{d}\nu(t) \leq \sup_t \sqrt{\mathbb{K}(t,t)}\|\nu\|_{\mathrm{TV}} < \infty\,, \tag{A.5}$$

using the fact that $t \in \mathcal{X} \mapsto \sqrt{\mathbb{K}(t,t)}$ is a bounded continuous function by ($\mathbf{A}_{\mathrm{C}}$). We emphasize that this function not need to be vanishing at infinity.

From (A.5), we deduce that the map $m \,:\, A \mapsto m(A) := \int_{\mathcal{X}} \|\varphi_t\|_{\mathbb{H}} \mathbf{1}_{\{\varphi_t \in A\}} \mathrm{d}\nu(t)$ is a finite measure on the Borel sets of $\mathbb{H}$ and hence, by Oxtoby-Ulam theorem (see [Giné and Nickl, 2021, Proposition 2.1.4] for instance), a tight Borel measure. Given $0 < \varepsilon_n \to 0$, let K_n be a compact set such that $m(K_n^c) < \varepsilon_n/2$, let $A_{n,1}, \ldots, A_{n,k_n}$ be a finite partition of K_n consisting of sets of diameter at most $\varepsilon_n/2$, pick up a point $h_{n,k} \in A_{n,k}$ for each k and define the simple function

$$f_n(t) = \sum_{k=1}^{k_n} h_{n,k} \mathbf{1}_{\{\varphi_t \in A_{n,k}\}}\,.$$

Then

$$\int_{\mathcal{X}} \|\varphi_t - f_n(t)\|_{\mathbb{H}} \mathrm{d}\nu(t) \leq \varepsilon_n/2 + m(K_n^c) < \varepsilon_n \to 0\,,$$

showing that $t \in \mathcal{X} \mapsto \varphi_t \in \mathbb{H}$ is Bochner integrable, hence Petti's integrable, and both integrals coincide (see for instance [Giné and Nickl, 2021, Section 2.6.1]). We deduce that Φ is well defined, using Bochner integration. Furthermore, one can deduce that

$$\Big\| \int_{\mathcal{X}} \varphi_t \mathrm{d}\nu(t) \Big\|_{\mathbb{H}} \leq \int_{\mathcal{X}} \|\varphi_t\|_{\mathbb{H}} \mathrm{d}\nu(t) \leq \sup_t \sqrt{\mathbb{K}(t,t)}\|\nu\|_{\mathrm{TV}}\,,$$

showing that Φ is bounded linear.

Also, if $h \in \mathbb{H}$ then

$$\Big| \int_{\mathcal{X}} \langle h, f_n(t)\rangle_{\mathbb{H}} - \langle h, \varphi_t\rangle_{\mathbb{H}} \mathrm{d}\nu(t) \Big| \leq \|h\|_{\mathbb{H}} \int_{\mathcal{X}} \|\varphi_t - f_n(t)\|_{\mathbb{H}} \mathrm{d}\nu(t) \to 0\,.$$

Hence, $\displaystyle\int_{\mathcal{X}} \langle h, \varphi_t\rangle_{\mathbb{H}} \mathrm{d}\nu(t) = \lim_n \int_{\mathcal{X}} \langle h, f_n(t)\rangle_{\mathbb{H}} \mathrm{d}\nu(t)$ exists and is finite. We deduce that

$$\int_{\mathcal{X}} \langle h, \varphi_t\rangle_{\mathbb{H}} \mathrm{d}\nu(t) = \langle h, \Phi(\nu)\rangle_{\mathbb{H}}\,, \tag{A.6}$$

using that $\int_{\mathcal{X}} \langle h, f_n(t)\rangle_{\mathbb{H}} \mathrm{d}\nu(t) = \langle h, \int_{\mathcal{X}} f_n \mathrm{d}\nu\rangle_{\mathbb{H}}$.

Using (A.6) and Cauchy-Schwarz inequality, one gets that

$$\langle h, \Phi(\nu)\rangle_{\mathbb{H}} = \int_{\mathcal{X}} \langle h, \varphi_t\rangle_{\mathbb{H}} \mathrm{d}\nu(t) \leq \sup_t \sqrt{\mathbb{K}(t,t)} \|h\|_{\mathbb{H}} \|\nu\|_{\mathrm{TV}}\,, \tag{A.7}$$

and hence, we can write

$$\langle h, \Phi(\nu)\rangle_{\mathbb{H}} = \langle \langle h, \varphi_t\rangle_{\mathbb{H}}, \nu\rangle_{\mathcal{M}(\mathcal{X})^\star, \mathcal{M}(\mathcal{X})}$$

where $\mathcal{M}(\mathcal{X})^\star$ is the topological dual of $\mathcal{M}(\mathcal{X})$. It shows that the dual $\Phi^\star$ is given by $\Phi^\star(h)(t) = \langle h, \varphi_t\rangle_{\mathbb{H}}$. As a function of t, it is clear that it is continuous by ($\mathbf{A}_{\mathrm{C}}$) and that $\|\Phi^\star(h)\|_\infty \leq \sup_t \sqrt{\mathbb{K}(t,t)} \|h\|_{\mathbb{H}} < \infty$, showing that it belongs to the space of bounded continuous functions. □

A.3 Proof of Theorem 1.1

Let (ν_n) be a minimizing sequence of measures of Program (1.6). Up to an extraction we can consider that $L(\Phi(\nu_n)) + \lambda\|\nu_n\|_{\mathrm{TV}} \leq 1 + \inf_\nu\{L(\Phi(\nu)) + \lambda\|\nu\|_{\mathrm{TV}}\}$. In particular, it holds that

$$\lambda\|\nu_n\|_{\mathrm{TV}} \leq 1 + \inf_\nu\{L(\Phi(\nu)) + \lambda\|\nu\|_{\mathrm{TV}}\}\,.$$

Up to an extraction, by Banach-Alaoglu theorem, we can consider that the sequence (ν_n) converges for the weak-$\star$ topology. We denote by $\mu^\star \in \mathcal{M}(\mathcal{X})$ its limit. Using [Brézis, 2011, Proposition 3.13(iii)], the TV-norm is l.s.c. for the weak-$\star$ topology, and we get that

$$\liminf_n \|\nu_n\|_{\mathrm{TV}} \geq \|\mu^\star\|_{\mathrm{TV}}\,.$$

Using Lemma A.2, it holds that for any $h \in \mathbb{H}$ and for any convergent sequence $\nu_n \to \mu^\star$ for the weak-$\star$ topology,

$$\langle h, \Phi(\nu_n)\rangle_{\mathbb{H}} = \langle \Phi^\star(h), \nu_n\rangle_{\mathcal{C}(\mathcal{X}),\mathcal{M}(\mathcal{X})} \to \langle \Phi^\star(h), \mu^\star\rangle_{\mathcal{C}(\mathcal{X}),\mathcal{M}(\mathcal{X})}\,,$$

proving that Φ is continuous from $\mathcal{M}(\mathcal{X})$ weak-$\star$ to $\mathbb{H}$ weak (see Remark A.2). Since L is l.s.c for the weak topology of $\mathbb{H}$ [Brézis, 2011, Corollary 3.9], we get that

$$\liminf_n L(\Phi(\nu_n)) \geq L(\Phi(\mu^\star))\,.$$

Combining the aforementioned limits, we deduce that

$$(1.6) = \liminf_n \big\{L(\Phi(\nu_n)) + \lambda\|\nu_n\|_{\mathrm{TV}}\big\} \geq L(\Phi(\mu^\star)) + \lambda\|\mu^\star\|_{\mathrm{TV}} \geq (1.6)\,,$$

hence equality. The uniqueness of $\Phi(\mu^\star)$ follows by strict convexity.

Remark A.3. *In this paper, we assume that $\mathcal{X}$ is compact. Some of our bounds depend on the size of $\mathcal{X}$ and do not hold for non-compact spaces. But, the existence of $\mu^\star$ can be proven in the non-compact case.*

The subtlety is in (A.3). To get the proof of Theorem 1.1 work when $\mathcal{X}$ is a Polish space (not necessarily compact), one needs that $\mathrm{Im}(\Phi^\star) \in \big(\mathcal{C}_0(\mathcal{X}), \|\cdot\|_\infty\big)$, the space of continuous functions vanishing at infinity. We already know that $\mathrm{Im}(\Phi^\star) \in \big(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty\big)$ by Condition ($\mathbf{A}_{\mathrm{C}}$). We have the following result.

Theorem A.3. *Let $\mathbb{H}$ be Hilbert space and let $\mathcal{X}$ be Polish space. Assume that*

- *Assumption* ($\mathbf{A}_{\mathrm{C}}$) *holds;*
- *the RKHS $\mathcal{F}$ (defined by $\mathbb{K}$) is contained in $\mathcal{C}_0(\mathcal{X})$;*
- *and $\sup_t \sqrt{\mathbb{K}(t,t)} < \infty$;*

then the there exists a measure $\mu^\star \in \mathcal{M}(\mathcal{X})$ such that

$$J(\mu^\star) = \min_{\mu \in \mathcal{M}(\mathcal{X})} J(\mu)\,.$$

Furthermore, the vector $\Phi(\mu^\star) \in \mathbb{H}$ is unique.

Remark A.4. *The same argument can be used to prove that Program (1.6) restricted to $\mathcal{M}(\mathcal{X})_+$ admits solutions. Indeed, take (ν_n) a sequence of nonnegative measures such that the objective converges towards the infimum. We can use the above proof to show the existence of $\mu^\star$. The only point left to prove is that the measure $\mu^\star$ is nonnegative, which is straight forward using weak-star convergence and Riesz representation theorem [Rudin, 1974, Chapter 2] of nonnegative linear functional defined by nonnegative continuous functions with compact support (which are included in $\mathcal{C}_0(\mathcal{X})$).*

Remark A.5. *A similar result can be found in [Chizat, 2022, Proposition 3.1] using Prokorov's theorem.*

B Gradients of the objective

B.1 In the space of (nonnegative) measures

We first consider the variation of J in $\mathcal{M}(\mathcal{X})_+$ in terms of its Fréchet differential.

Proposition B.1. *If $\nu + \sigma \in \mathcal{M}(\mathcal{X})_+$ and $\nu \in \mathcal{M}(\mathcal{X})_+$ then*

$$J(\nu+\sigma) - J(\nu) = \int_{\mathcal{X}} J'_\nu \mathrm{d}\sigma + \frac{1}{2}\|\Phi(\sigma)\|_{\mathbb{H}}^2\,, \tag{B.1}$$

where $J'_\nu := \Phi^\star(\Phi(\nu) - \mathbf{y}) + \lambda$ and $\Phi^\star : (\mathbb{H}, \|\cdot\|_{\mathbb{H}}) \to (\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$ is the dual of Φ.

Proof of B.1. The proof follows from the expansion of $J(\nu+\sigma)$:

$$\begin{aligned}
J(\nu+\sigma) &= \frac{1}{2}\big\|\mathbf{y} - \Phi(\nu+\sigma)\big\|_{\mathbb{H}}^2 + \lambda\|\mu+\sigma\|_{\mathrm{TV}} \\
&= \frac{1}{2}\big\|\mathbf{y} - \Phi(\nu) - \Phi(\sigma)\big\|_{\mathbb{H}}^2 + \lambda\|\nu+\sigma\|_{\mathrm{TV}} \\
&= \frac{1}{2}\big\|\mathbf{y} - \Phi(\nu)\big\|_{\mathbb{H}}^2 - \langle \mathbf{y} - \Phi(\nu), \Phi(\sigma)\rangle_{\mathbb{H}} + \frac{1}{2}\big\|\Phi(\sigma)\big\|_{\mathbb{H}}^2 + \lambda\|\nu+\sigma\|_{\mathrm{TV}} \\
&= J(\nu) - \langle \Phi^\star(\mathbf{y} - \Phi(\nu)), \sigma\rangle_{\mathbb{H}} + \frac{1}{2}\big\|\Phi(\sigma)\big\|_{\mathbb{H}}^2 + \lambda\left[\|\nu+\sigma\|_{\mathrm{TV}} - \|\nu\|_{\mathrm{TV}}\right].
\end{aligned}$$

Using $\mathrm{Sign}(\nu)$ as a subgradient of the TV-norm at point ν (ν-almost everywhere equal to the sign of ν and with infinity norm less than one), we then observe that:

$$\|\nu+\sigma\|_{\mathrm{TV}} - \|\nu\|_{\mathrm{TV}} = \langle \mathrm{Sign}(\nu), \sigma\rangle_{\mathcal{M}(\mathcal{X})^\star, \mathcal{M}(\mathcal{X})} + \mathcal{D}_\nu(\sigma),$$

where $\mathcal{D}_\nu(\sigma)$ is the second order Bregman divergence of the TV-norm between ν and $\nu+\sigma$ using the subgradient $\mathrm{Sign}(\nu)$, given by:

$$\mathcal{D}_\nu(\sigma) := \|\nu+\sigma\|_{\mathrm{TV}} - \|\nu\|_{\mathrm{TV}} - \langle \mathrm{Sign}(\nu), \sigma\rangle_{\mathcal{M}(\mathcal{X})^\star, \mathcal{M}(\mathcal{X})}\,,$$

with $\mathcal{M}(\mathcal{X})^\star \subseteq (L^\infty(\mathcal{X}), \|\cdot\|_\infty)$ the topological dual of $\mathcal{M}(\mathcal{X})$. Gathering all the pieces, we obtain that:

$$J(\nu+\sigma) - J(\nu) = \langle J'_\nu, \sigma\rangle_{\mathcal{M}(\mathcal{X})^\star, \mathcal{M}(\mathcal{X})} + q(\sigma)\,, \tag{B.2}$$

where J'_ν is given in the statement of Proposition B.1 and q is a second order term given by:

$$q(\sigma) := \frac{1}{2}\|\Phi(\sigma)\|_{\mathbb{H}}^2 + \lambda \mathcal{D}_\nu(\sigma).$$

Finally, we remark that when ν is nonnegative, one possible choice for the TV subgradient is $\mathrm{Sign}(\nu) = 1$. In this case, the previous decomposition may be simplified as:

$$J'_\nu = \Phi^\star(\Phi(\nu) - \boldsymbol{y}) + \lambda$$

and $\mathcal{D}_\nu(\sigma) = 0$ when $\nu + \sigma \in \mathcal{M}(\mathcal{X})_+$ and $\nu \in \mathcal{M}(\mathcal{X})_+$. □

Remark B.1. *The above proof shows that for any $\mu, \nu \in \mathcal{M}(\mathcal{X})$,*

$$J(\nu + \sigma) - J(\nu) = \int_{\mathcal{X}} J'_\nu \mathrm{d}\sigma + \frac{1}{2}\|\Phi(\sigma)\|_{\mathbb{H}}^2 + \lambda \mathcal{D}_\nu(\sigma)\,,$$

where $\mathcal{D}_\nu(\sigma) := \|\nu+\sigma\|_{\mathrm{TV}} - \|\nu\|_{\mathrm{TV}} - \langle \mathrm{Sign}(\nu), \sigma\rangle_{\mathcal{M}(\mathcal{X})^\star, \mathcal{M}(\mathcal{X})}$ and $\mathcal{M}(\mathcal{X})^\star \subseteq (L^\infty(\mathcal{X}), \|\cdot\|_\infty)$ the topological dual of $\mathcal{M}(\mathcal{X})$.

Besides the expression of the Fréchet differential of J on the space $\mathcal{M}(\mathcal{X})_+$, it is possible to explicit the value of J'_ν at any point $t \in \mathcal{X}$. Using Lemma A.2, it holds,

$$J'_\nu(t) = \langle \varphi_t, \Phi(\nu) - \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda, \quad \forall t \in \mathcal{X}. \tag{B.3}$$

Note that J'_ν depends on ν through $\Phi(\nu)$. Recall that $\Phi(\mu^\star)$ is constant across all solutions $\mu^\star$ of Program ($\mathcal{B}_+$), see Theorem 1.1. Hence, the function

$$J'_\star(x) := \langle \varphi_x, \Phi(\mu^\star) - \boldsymbol{y}\rangle_{\mathbb{H}} + \lambda\,, \quad x \in \mathcal{X}\,,$$

is well defined and does not depend on the choice of the solution $\mu^\star$ (it is the same function across all possible choice of $\mu^\star$ solution to ($\mathcal{B}_+$)). The next proposition gives the first order condition of Program ($\mathcal{B}_+$).

Proposition B.2. *It holds that $J'_\star \geq 0$ and, for any solution $\mu^\star$ to Program ($\mathcal{B}_+$),*

$$\mathrm{Supp}(\mu^\star) \subseteq \left\{x \in \mathcal{X} \,:\, J'_\star(x) = 0\right\}. \tag{B.4}$$

Conversely, if a measure $\nu \in \mathcal{M}(\mathcal{X})_+$ is such that $J'_\nu \geq 0$ and it satisfies the condition

$$\mathrm{Supp}(\nu) \subseteq \left\{x \in \mathcal{X} \,:\, J'_\nu(x) = 0\right\},$$

then ν is a solution to Program ($\mathcal{B}_+$) and $J'_\star = J'_\nu$.

Proof. Let $x \in \mathcal{X}$ and let $\varepsilon > 0$ be defined later. By Proposition B.1, one has

$$J'_\star(x) = \frac{J(\mu^\star + \varepsilon\delta_x) - J(\mu^\star)}{\varepsilon} - \frac{\varepsilon}{2}\|\varphi_x\|_{\mathbb{H}}^2\,,$$

hence

$$J'_\star(x) \geq \liminf_{\varepsilon \downarrow 0} \left\{ \frac{J(\mu^\star + \varepsilon\delta_x) - J(\mu^\star)}{\varepsilon} \right\} \geq 0\,,$$

since $J(\mu^\star + \varepsilon\delta_x) - J(\mu^\star) \geq 0$.

Assume now that there exists a point $x \in \mathcal{X}$ such that $x \in \mathrm{Supp}(\mu^\star)$ and $J'_\star(x) > 0$. Since $J'_\star$ is continuous and $J'_\star(x) > 0$ there exists $\varepsilon > 0$ and a open neighborhood U_x of x such that

$$\forall t \in U_x\,, \quad J'_x(t) > \sqrt{\varepsilon}$$

By Jordan decomposition theorem, there exists two nonnegative measures $\mu^\star_+$ and $\mu^\star_-$ with disjoints supports such that $\mu^\star = \mu^\star_+ - \mu^\star_-$ and $\mathrm{Supp}(\mu^\star) = \mathrm{Supp}(\mu^\star_+) \sqcup \mathrm{Supp}(\mu^\star_-)$. Without loss of generality, we assume that $x \in \mathrm{Supp}(\mu^\star_+)$. Taking $\varepsilon > 0$ and U_x sufficiently smalls, one has $U_x \cap \mathrm{Supp}(\mu^\star_-) = \emptyset$ and

$$\mu^\star(U_x) > \sqrt{\varepsilon}\,.$$

Let B be a Borelian of $\mathcal{X}$ and define $\sigma \in \mathcal{M}(\mathcal{X})$ by $\sigma(B) := -\mu_+^\star(B \cap U_x)$. Remark that $\mathcal{D}_{\mu^\star}(\sigma) = 0$, this latter being straightforward when $\mu^\star \in \mathcal{M}(\mathcal{X})_+$ (in this case $\mu^\star + \sigma \in \mathcal{M}(\mathcal{X})_+$). By Proposition B.1, one has

$$0 \leq J(\mu^\star + \sigma) - J(\mu^\star) = \int_{\mathcal{X}} J'_\star \mathrm{d}\sigma = -\int_{\mathcal{X} \cap U_x} J'_\star \mathrm{d}\mu_+^\star \,,$$

and also

$$\int_{\mathcal{X} \cap U_x} J'_\star \mathrm{d}\mu_+^\star \geq \varepsilon \,,$$

which is a contradiction. The converse result is a consequence of (B.1) as

$$J(\mu^\star) - J(\nu) = \int_{\mathcal{X}} J'_\nu \mathrm{d}(\mu^\star - \nu) + \frac{1}{2}\|\Phi(\mu^\star - \nu)\|_{\mathbb{H}}^2 = \int_{\mathcal{X}} J'_\nu \mathrm{d}\mu^\star + \frac{1}{2}\|\Phi(\mu^\star - \nu)\|_{\mathbb{H}}^2 \geq 0$$

and hence ν is a minimizer. Finally, $J'_\star$ does not depend on the choice of the solution $\mu^\star$ hence $J'_\star = J'_\nu$. □

The lemma displayed below provides some bounds on the Frechet differential of the objective function and on its stochastic estimate.

Lemma B.1. *There exists a positive constant $\mathcal{C}_0 = \mathcal{C}_0(y, \varphi, \lambda)$ such that*

$$\|J'_\nu\|_\infty := \sup_{t \in \mathcal{X}} |J'_\nu(t)| \leq \mathcal{C}_0(\|\nu\|_{\mathrm{TV}} + 1) \quad \forall \nu \in \mathcal{M}(\mathcal{X})_+.$$

*Moreover, provided Assumption (**A**$_1$) is satisfied, we have almost surely for any $\nu \in \mathcal{M}(\mathcal{X})_+$*

$$\sup_{t \in \mathcal{X}} |J'_\nu(t, Z)| \leq \mathcal{C}_1(\|\nu\|_{\mathrm{TV}} + 1) \quad \text{and} \quad \sup_{t \in \mathcal{X}} |\xi_\nu(t, Z)| \leq \mathcal{C}_2(\|\nu\|_{\mathrm{TV}} + 1),$$

for some constants $\mathcal{C}_1$ and $\mathcal{C}_2$ depending only on y, φ and λ.

Proof. Let $\nu \in \mathcal{M}(\mathcal{X})_+$ be fixed. We denote by $\tilde{\nu}$ the normalized measure $\tilde{\nu} = \nu / \|\nu\|_{\mathrm{TV}} \in \mathcal{M}(\mathcal{X})_+$. According to (B.3), we have

$$\begin{aligned}
\sup_{t \in \mathcal{X}} |J'_\nu(t)| &\leq \lambda + \sup_{t \in \mathcal{X}} |\langle \varphi_t, y \rangle_{\mathbb{H}}| + \|\nu\|_{\mathrm{TV}} \sup_{t \in \mathcal{X}} |\langle \varphi_t, \Phi(\tilde{\nu}) \rangle_{\mathbb{H}}|, \\
&\leq \lambda + \sup_{t \in \mathcal{X}} |\langle \varphi_t, y \rangle_{\mathbb{H}}| + \|\nu\|_{\mathrm{TV}} \sup_{t,s \in \mathcal{X}} |\langle \varphi_t, \varphi_s \rangle_{\mathbb{H}}|, \\
&\leq \lambda + \|\varphi\|_{\infty,\mathbb{H}} \|y\|_{\mathbb{H}} + \|\nu\|_{\mathrm{TV}} \|\varphi\|_{\infty,\mathbb{H}}^2, \\
&\leq \mathcal{C}_0(\|\nu\|_{\mathrm{TV}} + 1),
\end{aligned}$$

with

$$\mathcal{C}_0 = \max\left(\lambda + \|\varphi\|_{\infty,\mathbb{H}} \|y\|_{\mathbb{H}}; \|\varphi\|_{\infty,\mathbb{H}}^2\right). \tag{B.5}$$

Concerning the second part of the lemma, we first remark that, provided Assumption (**A**$_1$) is satisfied, we have for any $\nu \in \mathcal{M}(\mathcal{X})_+$

$$\sup_{t \in \mathcal{X}} |J'_\nu(t, Z)| \leq \|\nu\|_{\mathrm{TV}} \sup_{t \in \mathcal{X}} |\boldsymbol{g}_{t,T}(U)| + \sup_{t \in \mathcal{X}} |\boldsymbol{h}_t(V)| + \lambda \leq \mathcal{C}_1(\|\nu\|_{\mathrm{TV}} + 1),$$

with

$$\mathcal{C}_1 := \max\left(\|\boldsymbol{g}\|_\infty; \|\boldsymbol{h}\|_\infty + \lambda\right). \tag{B.6}$$

The last results is obtained thanks to a basic triangle inequality

$$\sup_{t \in \mathcal{X}} |\xi_\nu(t, Z)| \leq \sup_{t \in \mathcal{X}} |J'_\nu(t)| + \sup_{t \in \mathcal{X}} |J'_\nu(t, Z)| \leq \mathcal{C}_2(\|\nu\|_{\mathrm{TV}} + 1),$$

with

$$\mathcal{C}_2 = \mathcal{C}_0 + \mathcal{C}_1 = \max\left(\lambda + \|\varphi\|_{\infty,\mathbb{H}} \|y\|_{\mathbb{H}}; \|\varphi\|_{\infty,\mathbb{H}}^2\right) + \max\left(\|\boldsymbol{g}\|_\infty; \|\boldsymbol{h}\|_\infty + \lambda\right). \tag{B.7}$$

□

B.2 In the space of particles

We consider any set of positions $\mathbb{T}$ and their associate weights $\mathbb{W}$. In order to compute the derivatives of F w.r.t. $\mathbb{W}$ and $\mathbb{T}$, our starting point is Equation (1.9) and we observe that the gradient with respect to $\mathbb{W}$ is easily computed:

$$\nabla_{\boldsymbol{\omega}} F(\boldsymbol{\omega}, \mathbb{T}) = \boldsymbol{\lambda} - \boldsymbol{k}_{\mathbb{T}} + \mathbb{K}_{\mathbb{T}} \mathbb{W}.$$

Nevertheless, the interpretation in terms of Fréchet derivative and of J allows to obtain the next result.

Proposition B.3. *For any $\mathbb{W}$ and $\mathbb{T}$, denote $\nu = \nu(\mathbb{W}, \mathbb{T})$, one has:*

- (i) *Gradient w.r.t. weights: for any $j \in \{1, \dots, p\}$, one has $\nabla_{\omega_j} F(\boldsymbol{\omega}, \mathbb{T}) = J'_\nu(\boldsymbol{t}_j)$*
- (ii) *Gradient w.r.t. positions: for any $j \in \{1, \dots, p\}$, one has $\nabla_{\boldsymbol{t}_j} F(\boldsymbol{\omega}, \mathbb{T}) = \omega_j \nabla_{\boldsymbol{t}_j} J'_\nu(\boldsymbol{t}_j)$*

Proof. The starting point is the Fréchet derivative that, if we consider $\sigma \in \mathcal{M}(\mathcal{X})_+$ and $\varepsilon > 0$ small enough:

$$J(\nu + \varepsilon\sigma) = J(\nu) + \varepsilon \langle J'_\nu, \sigma \rangle_{\mathbb{H}} + o(\varepsilon).$$

Proof of (i): Considering any particle $j \in \{1, \dots, p\}$ and $\sigma = \delta_{t_j}$, we then obtain that:

$$\lim_{\varepsilon \to 0} \frac{J(\nu + \varepsilon\sigma) - J(\nu)}{\varepsilon} = \langle J'_\nu, \delta_{\boldsymbol{t}_j} \rangle_{\mathbb{H}} = J'_\nu(\boldsymbol{t}_j),$$

where the last equality comes from the reproducing kernel property. In the meantime, we observe that

$$\lim_{\varepsilon \to 0} \frac{J(\nu + \varepsilon\sigma) - J(\nu)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{F(\mathbb{W} + \varepsilon \mathbf{1}_j, \mathbb{T}) - F(\mathbb{W}, \mathbb{T})}{\varepsilon} = \frac{\partial F(\mathbb{W}, \mathbb{T})}{\partial \omega_j}.$$

We then conclude using the Fréchet derivative of J that:

$$J'_\nu(\boldsymbol{t}_j) = \nabla_{\omega_j} F(\boldsymbol{\omega}, \mathbb{T}).$$

Proof of (ii): Using the same consideration on the positions of the particles, we then consider any pertubed set of positions $\tilde{\mathbb{T}}_{\varepsilon,j} = \mathbb{T} + \varepsilon \mathbf{1}_j$ where only the coordinate j of $\mathbb{T}$ is modified. We then write the partial derivative of F:

$$\lim_{\varepsilon \to 0} \frac{F(\mathbb{W}, \tilde{\mathbb{T}}_{\varepsilon,j}) - F(\mathbb{W}, \mathbb{T})}{\varepsilon} = \frac{\partial F(\mathbb{W}, \mathbb{T})}{\partial t_j}.$$

In the meantime, we observe that with the Fréchet derivative of J that:

$$\begin{aligned}
\lim_{\varepsilon \to 0} \frac{F(\mathbb{W}, \tilde{\mathbb{T}}_{\varepsilon,j}) - F(\mathbb{W}, \mathbb{T})}{\varepsilon} &= \lim_{\varepsilon \to 0} \frac{J(\nu(\mathbb{W}, \tilde{\mathbb{T}}_{\varepsilon,j})) - J(\nu(\mathbb{W}, \mathbb{T}))}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{J(\nu(\mathbb{W}, \mathbb{T}) + \omega_j[\delta_{\boldsymbol{t}_j+\varepsilon} - \delta_{\boldsymbol{t}_j}]) - J(\nu(\mathbb{W}, \mathbb{T}))}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{\langle J'_{\nu(\mathbb{W},\mathbb{T})}, \omega_j \delta_{\boldsymbol{t}_j+\varepsilon} \rangle_{\mathbb{H}} - \langle J'_{\nu(\mathbb{W},\mathbb{T})}, \omega_j \delta_{\boldsymbol{t}_j} \rangle_{\mathbb{H}}}{\varepsilon} \\
&= \omega_j \nabla_{t_j} J'_{\nu(\mathbb{W},\mathbb{T})}(\boldsymbol{t}_j).
\end{aligned}$$

□

Finally, it is possible to quantify the way F is modified when we change $(\mathbb{W}, \mathbb{T})$ to $(\mathbb{W}', \mathbb{T}')$ thanks to the next proposition. where $\nu = \sum_{k=1}^p \omega_k \delta_{t_k}$.

Proposition B.4. *Consider two pairs $(\mathbb{W}, \mathbb{T})$ and $(\mathbb{W}', \mathbb{T}')$ of weights/positions and denote $\nu = \nu(\mathbb{W}, \mathbb{T})$ defined in Equation (1.7), then:*

$$F(\boldsymbol{\omega}', \mathbb{T}') - F(\boldsymbol{\omega}, \mathbb{T}) = \sum_{j=1}^p (\omega'_j J'_\nu(\boldsymbol{t}'_j) - \omega_j J'_\nu(\boldsymbol{t}_j)) + \frac{1}{2} (\boldsymbol{\omega}', -\boldsymbol{\omega})^\top \mathbb{K}_{(\mathbb{T}',\mathbb{T})} (\boldsymbol{\omega}', -\boldsymbol{\omega}), \tag{B.8}$$

where $\mathbb{K}_{(\mathbb{T}',\mathbb{T})}$ is a $(2p \times 2p)$ symmetric matrix with $(p \times p)$ diagonal blocks $\langle \varphi_{t'_i}, \varphi_{t'_j} \rangle_{\mathbb{H}}$ and $\langle \varphi_{t_i}, \varphi_{t_j} \rangle_{\mathbb{H}}$, and $(p \times p)$ off-diagonal block $\langle \varphi_{t'_i}, \varphi_{t_j} \rangle_{\mathbb{H}}$.

Proof. We denote $\nu' = \nu(\mathbb{W}', \mathbb{T}')$ and apply Equation (B.1) with $\nu' = \nu + \sigma$ with $\sigma = \nu' - \nu$, we obtain that:

$$F(\mathbb{W}', \mathbb{T}') - F(\mathbb{W}, \mathbb{T}) = J(\nu + \sigma) - J(\nu) = \int_{\mathcal{X}} J'_\nu \mathrm{d}\sigma + \frac{1}{2}\|\Phi(\sigma)\|^2_{\mathbb{H}} = \int_{\mathcal{X}} J'_\nu \mathrm{d}\sigma + \frac{1}{2}\|\Phi(\nu') - \Phi(\nu)\|^2_{\mathbb{H}}$$

Note that the last term rewrites

$$(\boldsymbol{\omega}', -\boldsymbol{\omega})^\top \mathbb{K}_{(\mathbb{T}', \mathbb{T})}(\boldsymbol{\omega}', -\boldsymbol{\omega}) = \|\Phi(\nu') - \Phi(\nu)\|^2_{\mathbb{H}} \,,$$

which is the squared Maximum Mean Discrepancy (MMD) between ν and ν' for the kernel K. □

C Technical results for Theorem 1.2

Proposition C.1. *Consider $(\boldsymbol{t}, \tilde{\boldsymbol{t}}) \in \mathcal{X}^2$, the following technical inequalities hold:*

i) $|\nabla J'_\nu(\boldsymbol{t}) - \nabla J'_\nu(\tilde{\boldsymbol{t}})| \leq [Lip(\nabla\varphi)\|\nu\|_{\mathrm{TV}} + Lip(\boldsymbol{\nabla} y)]\, |\boldsymbol{t} - \tilde{\boldsymbol{t}}|$,

ii) $\|\nabla J'_\nu(t)\|_{\mathcal{X}} \leq (\|\nu\|_{\mathrm{TV}}\|\varphi\|_{\infty,\mathbb{H}} + \|y\|_{\mathbb{H}})\|\varphi'\|_\infty$ *with* $\|\varphi'\|_\infty := \sup_t \sup_{\psi:\|\psi\|_{\mathbb{H}}\leq 1} \|\nabla_{\boldsymbol{t}}\langle\varphi_{\boldsymbol{t}}, \psi\rangle\|_{\mathcal{X}}$,

iii) $\|D_\nu(t, Z)\|_{\mathcal{X}} \leq \|\nu\|_{\mathrm{TV}}\|\boldsymbol{g}'\|_{\infty,\mathbb{H}} + \|\boldsymbol{h}'\|_{\mathbb{H}}$,

iv) $\|\nabla^2 J'_\nu(\boldsymbol{t})\|_{op} \leq (\|\nu\|_{\mathrm{TV}}\|\varphi\|_{\infty,\mathbb{H}}+\|y\|_{\mathbb{H}})\|\nabla^2\varphi\|_{\infty,op}$ *with* $\|\nabla^2\varphi\|_{\infty,op} := \sup_t \sup_{\psi:\|\psi\|_{\mathbb{H}}\leq 1} \|\nabla^2_{\boldsymbol{t}}\langle\varphi_{\boldsymbol{t}}, \psi\rangle\|_{op}$,

where Assumption ($\mathbf{A}_2$) *is required for inequality iii).*

Proof of Proposition C.1. We consider any finite measure ν.
Proof of *i)* We provide an upper bound on the Lipschitz constant of $\nabla J'_\nu$: consider $(\boldsymbol{t}, \tilde{\boldsymbol{t}}) \in \mathcal{X}^2$, we repeat the same arguments as above and observe that:

$$|\nabla J'_\nu(\boldsymbol{t}) - \nabla J'_\nu(\tilde{\boldsymbol{t}})| \leq [Lip(\nabla\varphi)\|\nu\|_{\mathrm{TV}} + Lip(\boldsymbol{\nabla} y)]\, |\boldsymbol{t} - \tilde{\boldsymbol{t}}|.$$

Proof of *ii)* Using a rough bound, we get

$$\begin{aligned}
\|\nabla_{\boldsymbol{t}} J'_\nu(t)\| &= \left\| \sum_{j=1}^p \omega_j \nabla_{\boldsymbol{t}}\langle\varphi_{\boldsymbol{t}}, \varphi_{t_j}\rangle_{\mathbb{H}} - \nabla_{\boldsymbol{t}}\langle\varphi_{\boldsymbol{t}}, y\rangle_{\mathbb{H}} \right\|, \\
&\leq \sum_{j=1}^p \omega_j \|\nabla_{\boldsymbol{t}}\langle\varphi_{\boldsymbol{t}}, \varphi_{t_j}\rangle_{\mathbb{H}}\|_{\mathcal{X}} + \|\nabla_{\boldsymbol{t}}\langle\varphi_{\boldsymbol{t}}, y\rangle_{\mathbb{H}}\|_{\mathcal{X}}, \\
&\leq (\|\nu\|_{\mathrm{TV}}\|\varphi\|_{\infty,\mathbb{H}} + \|y\|_{\mathbb{H}}) \times \sup_{\psi:\|\psi\|_{\mathbb{H}}\leq 1} \|\nabla_{\boldsymbol{t}}\langle\varphi_{\boldsymbol{t}}, \psi\rangle\|_{\mathcal{X}},
\end{aligned}$$

which provides the desired result.
Proof of *iii)* Using Assumption ($\mathbf{A}_2$), and in particular the boundedness of the derivative of $\boldsymbol{g}$ and $\boldsymbol{h}$, we get

$$\begin{aligned}
\|D_\nu(\boldsymbol{t}, Z)\|_{\mathcal{X}} &= \|\|\nu\|_{\mathrm{TV}}\nabla_{\boldsymbol{t}}\boldsymbol{g}_{\boldsymbol{t},T}(U) - \nabla_t \boldsymbol{h}_{\boldsymbol{t}}(V)\|_{\mathcal{X}}\,, \\
&\leq \|\nu\|_{\mathrm{TV}} \sup_{t,s,u} \|\nabla_t \boldsymbol{g}_{t,s}(u)\|_{\mathcal{X}} + \sup_{t,v} \|\nabla_t \boldsymbol{h}_t(v)\|_{\mathcal{X}}.
\end{aligned}$$

Proof of *iv)* Similarly to item *iii)*, we have

$$\begin{aligned}
\|\nabla^2 J'_\nu(\boldsymbol{t})\|_{op} &= \left\| \sum_{j=1}^p \omega_j \nabla^2\langle\varphi_{\boldsymbol{t}}, \varphi_{t_j}\rangle_{\mathbb{H}} - \nabla^2\langle\varphi_{\boldsymbol{t}}, y\rangle_{\mathbb{H}} \right\|, \\
&\leq \sum_{j=1}^p \omega_j \|\nabla^2\langle\varphi_{\boldsymbol{t}}, \varphi_{t_j}\rangle_{\mathbb{H}}\|_{op} + \|\nabla^2\langle\varphi_{\boldsymbol{t}}, y\rangle_{\mathbb{H}}\|_{op}, \\
&\leq (\|\nu\|_{\mathrm{TV}}\|\varphi\|_{\infty,\mathbb{H}} + \|y\|_{\mathbb{H}}) \times \sup_{\psi:\|\psi\|_{\mathbb{H}}\leq 1} \|\nabla^2\langle\varphi_{\boldsymbol{t}}, \psi\rangle\|_{op},
\end{aligned}$$

□

Proposition C.2. *A large enough constant $\mathfrak{C}$ exists such that for any iteration $k \in \mathbb{N}$ and any particle $j \in \{1, \ldots, p\}$, :*

$$\left|\mathbb{E}\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k) + e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)} - 1 \middle| \mathfrak{F}_k\right]\right| \leq \mathfrak{C}\alpha^2(1 + \|\nu_k\|_{\mathrm{TV}})^2.$$

Proof. This key technical argument relies on the Hoeffding inequality. We shall write:

$$\begin{aligned}
\mathbb{E}\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k) + e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)} - 1 \middle| \mathfrak{F}_k\right] &= \alpha J'_{\nu_k}(\boldsymbol{t}_j^k) - 1 + \mathbb{E}\left[e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)} \middle| \mathfrak{F}_k\right] \\
&= \left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k) - 1 + e^{-\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)} \mathbb{E}\left[e^{-\alpha[\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k) - J'_{\nu_k}(\boldsymbol{t}_j^k)]} \middle| \mathfrak{F}_k\right]\right] \\
&= \left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k) - 1 + e^{-\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)}\right] \\
&\quad + e^{-\alpha J'_{\nu_k}(\boldsymbol{t}_j^k)} \mathbb{E}\left[e^{-\alpha[\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k) - J'_{\nu_k}(\boldsymbol{t}_j^k)]} - 1 \middle| \mathfrak{F}_k\right]
\end{aligned}$$

To derive an upper bound, we apply the Hoeffding Lemma to the random variable $J'_{\nu_k}(\boldsymbol{t}_j^k, Z^{k+1}) - J'_{\nu_k}(\boldsymbol{t}_j^k)$ that is centered and bounded by $\boldsymbol{T} = \mathfrak{C}(1 + \|\nu_k\|_{\mathrm{TV}})$ from according to Lemma B.1. We obtain that:

$$\left|\mathbb{E}\left[e^{-\alpha[\widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)) - J'_{\nu_k}(\boldsymbol{t}_j^k)]} - 1 \middle| \mathfrak{F}_k\right]\right| \leq e^{\frac{T^2\alpha^2}{8}} - 1 \leq \mathfrak{C}\alpha^2(1 + \|\nu_k\|_{\mathrm{TV}}^2).$$

Using that $|e^x - 1 - x| \leq c|x|^2$ for bounded $x = \alpha J'_{\nu_k}(\boldsymbol{t}_j^k)$ and c large enough, we finally obtain that:

$$\left|\mathbb{E}\left[\alpha J'_{\nu_k}(\boldsymbol{t}_j^k) + e^{-\alpha \widehat{\boldsymbol{J}'_{\nu_k}}(\boldsymbol{t}_j^k)} - 1 \middle| \mathfrak{F}_k\right]\right| \leq \mathfrak{C}\alpha^2(1 + \|\nu_k\|_{\mathrm{TV}}^2).$$

□

We recall here the result essentially due to Chizat [2022], which is stated in a simplest way for our purpose.

Proposition C.3. *Assume that $\mu^\star$ is discrete and that ν_0 is a uniform distribution over a grid of size $\delta = \frac{2d}{\tau}$ where d is the dimension of $\mathcal{X}$, then:*

$$Q_\tau(\mu^\star, \nu_0) \leq \frac{\|\mu^\star\|_{\mathrm{TV}} d}{\tau}\left(1 + \log\frac{\tau}{2d} + \frac{\log|\mathcal{X}|}{d}\right)$$

Moreover, the measure ν_0^δ that meets this upper bound satisfies $\|\nu_0^\delta\|_{\mathrm{TV}} = \|\mu^\star\|_{\mathrm{TV}}$.

Proof. We define m as the size of the support of ν_0 and

$$\nu_0 = m^{-1}\sum_{i=1}^{m} \delta_{x_i},$$

where $(x_i)_{1 \leq i \leq m}$ refers to the uniform grid of size δ on $\mathcal{X}$. Since $\mu^\star$ is discrete, it may be written as:

$$\mu^\star = \sum_{j=1}^{m^\star} \mu_j^\star \delta_{z_j^\star}.$$

For any support point $z_j^\star$ of $\mu^\star$, we then consider $i_j \in \{1, ots, m\}$ such that $\|x_{i_j} - z_j\| \leq \delta/2$ and we define ν_0^δ as:

$$\nu_0^\delta := \sum_{j=1}^{m^\star} \mu_j^\star \delta_{x_{i_j}}$$

We observe that by construction, $\|\nu_0^\delta\|_{\mathrm{TV}} = \|\mu^\star\|_{\mathrm{TV}}$ and

$$\begin{aligned}
\mathcal{H}(\nu_0, \nu_0^\delta) &= \sum_{j=1}^{m^\star} \nu_0^\delta(x_{i_j}) \log\left(\frac{\nu_0^\delta(x_{i_j})}{\nu_0(x_{i_j})}\right) \\
&= \sum_{j=1}^{m^\star} \mu_j^\star \log\left(\frac{\mu_j^\star}{\nu_0(x_{i_j})}\right) \\
&\leq -\mathrm{Ent}(\mu^\star) + \|\mu^\star\|_{\mathrm{TV}}\left[d\log\left(\frac{1}{\delta}\right) + \log|\mathcal{X}|\right], \qquad \text{(C.1)}
\end{aligned}$$

where we used the entropy of a discrete measure defined as

$$\mathrm{Ent}(\mu) = -\sum_{x \in Supp(\mu)} \mu(x)\log(\mu(x))$$

and a lower bound of $\nu_0(x_{i_j})$, which is of the order $\delta^d|\mathcal{X}|^{-1}$ where $|\mathcal{X}|$ refers to the Lebesgue measure of $\mathcal{X}$.

In the meantime, we also observe that the BL dual norm between ν_0^δ and $\mu^\star$ can be easily upper bounded. Indeed

$$\begin{aligned} \|\nu_0^\delta - \mu^\star\|_{BL}^* &= \sup_{\|f\|_{BL}\leq 1} \int_{\mathcal{X}} f \mathrm{d}[\nu_0^\delta - \mu^\star] \\ &= \sup_{\|f\|_{BL}\leq 1} \sum_{j=1}^{m^\star} \mu_j^\star [f(x_{i_j}) - f(z_j)] \\ &\leq \frac{\|\mu^\star\|_{\mathrm{TV}}\delta}{2} \end{aligned} \tag{C.2}$$

We then add the two upper bounds (C.1) and (C.2) and minimize

$$\delta \longmapsto \frac{\|\mu^\star\|_{\mathrm{TV}}\delta}{2} + \frac{1}{\tau}\left(-\mathrm{Ent}(\mu^\star) + \|\mu^\star\|_{\mathrm{TV}}\left[d\log\left(\frac{1}{\delta}\right) + \log|\mathcal{X}|\right]\right).$$

We are led to choose $\delta = 2d\tau^{-1}$ and we obtain the following upper bound:

$$\frac{1}{\tau}\mathcal{H}(\nu_0, \nu_0^\delta) + \|\nu_0^\delta - \mu^\star\|_{BL}^* \leq \frac{\|\mu^\star\|_{\mathrm{TV}} d}{\tau}\left(1 + \log\frac{\tau}{2d} + \frac{\log|\mathcal{X}|}{d}\right),$$

which ends the proof of the proposition. □