

# Fast Spawn&Prune (FS&P): Global convergence of stochastic conic particle gradient descent via birth/death process

**Yohann De Castro**

*École Centrale Lyon, CNRS UMR 5208, Institut Camille Jordan, Écully, France.*

YOHANN.DE-CASTRO@EC-LYON.FR

**Sébastien Gadat**

*Toulouse School of Economics, CNRS UMR 5314, TSE-R Toulouse, France.*

SEBASTIEN.GADAT@TSE-FR.EU

**Clément Marteau**

*Université Lyon 1, CNRS UMR 5208, Institut Camille Jordan, Villeurbanne, France.*

CLEMENT.MARTEAU@MATH.UNIV-LYON1.FR

## Abstract

We investigate the global optimization of the objective function arising in continuous sparse regression, specifically the Beurling LASSO (BLASSO), over the space of measures. While Conic Particle Gradient Descent (CPGD) methods are computationally efficient, they may become trapped in local minima due to the non-convexity of the parameterization. To overcome this limitation, we introduce **Fast Spawn&Prune (FS&P)**, a stochastic algorithm that extends **FastPart** introduced in De Castro et al. (2025a) and combines CPGD with a birth–death process. The birth mechanism ensures asymptotic global exploration by introducing particles in regions where first-order optimality conditions are violated, while the death process preserves computational efficiency by pruning non-informative particles. We provide the first theoretical guarantee of global convergence for this class of discrete-time stochastic algorithms, without requiring exponentially large initializations. Furthermore, we derive explicit convergence rates for the excess risk, which scale as  $\mathcal{O}\left(\frac{\log K/K}{2^{(2+d)}}\right)$ , where  $K$  denotes the number of iterations and  $d$  the dimension of the domain, thereby quantifying the trade-off between global exploration and local refinement. Moreover, the sample complexity is  $\mathcal{O}\left(N^{-\frac{1}{4(2+d)}}\right)$  (up to logarithmic factors). We also propose a horizon-free variant that does not require prior knowledge of the iteration budget.

**Keywords:** continuous sparse regression, conic particle gradient descent, birth and death process, global convergence, stochastic optimization

## 1 Introduction

Continuous sparse regression has been at the core of numerous studies in statistics and signal processing. In particular, it encompasses a wide range of models and problems, including statistical *mixture models*, *deconvolution* problems, and *neural networks*. We refer to Candès and Fernandez-Granda (2014); Azais et al. (2015); De Castro et al. (2021); Duval and Peyré (2015); Giard et al. (2025), among others.

In this paper, we do not focus on the statistical properties of the estimator  $\mu^*$ , but rather on the underlying *optimization problem*. Indeed, attaining the exact global minimum is not strictly necessary in practice; approximate solutions are often sufficient to achieve the

desired statistical guarantees. While a thorough analysis of these properties lies beyond the scope of this paper, we refer the interested reader to Appendix F for a brief overview.

**From Over-Parametrization to Global Convergence** The optimization of a non-convex objective  $J$  via particle discretization has undergone significant theoretical advances. The seminal works of Chizat and Bach (2018); Chizat (2022) established that, in the *overparameterized* regime—where the number of particles is very large—gradient descent dynamics can benefit from a *convex optimization landscape*. Specifically, in the mean-field limit, the gradient flow converges to the global optimum, provided that the initialization assigns strictly positive mass to every measurable subset of the domain  $\mathcal{X}$ . This regime effectively convexifies the problem by allowing mass to flow freely toward the optimal support.

However, the computational cost of deterministic particle gradient descent scales poorly with both the number of particles and the dataset size. To address this issue, recent works have investigated stochastic approximations (De Castro et al., 2025a) and sketching techniques (Poon et al., 2023; De Castro et al., 2025b). In previous work, the authors analyzed *Stochastic Conic Particle Gradient Descent (FastPart)* in the overparameterized regime. They showed that replacing exact gradients with unbiased stochastic estimators (via mini-batching and random features) significantly improves time complexity while maintaining strong stability guarantees, in particular the boundedness of the total variation norm along the trajectory and local convergence rates.

Despite these advances, a critical gap remains regarding *global convergence* for discrete-time algorithms with sparse initialization (*i.e.*, with few to a moderate number of particles). Standard gradient descent methods, including their stochastic variants, are primarily local search methods and struggle to transport mass to remote regions of the domain when the current support is far from the optimum. As a result, the discrete algorithm may remain near stationary points where the first-order optimality conditions—namely, the non-negativity of  $J'_\nu$ , the Fréchet derivative of the objective—are violated in regions devoid of particles.

**Fast Spawn&Prune and the Birth/Death Process** In this paper, we introduce **FS&P**, an algorithm that augments stochastic conic particle gradient descent with a *Birth and Death* process. This mechanism is designed to bridge the gap between local descent and global exploration. The **Birth Process** acts as a global corrective mechanism. It detects violations of the first-order optimality conditions—specifically, regions  $\mathcal{N}_\nu$  where the so-called *dual certificate*  $J'_\nu$  satisfies  $J'_\nu < 0$ —and introduces new particles in these areas. This mechanism ensures asymptotic global exploration and prevents the dynamics from becoming trapped in local minima. Unlike greedy methods such as Frank–Wolfe algorithms, which require solving a global minimization problem for  $J'_\nu$  to add a particle, our approach simply samples random points within the negative regions  $\mathcal{N}_\nu$  of the certificate  $J'_\nu$ , making it significantly more computationally tractable. The **Death Process** maintains computational efficiency by pruning non-informative particles. Exploiting the convexity of the objective function over the space of signed measures, we show that removing a particle whose weight is small and for which the dual certificate  $J'_\nu$  at its location is sufficiently large (*i.e.*, in the regions  $\mathcal{P}_\nu$ ) strictly decreases the objective value. This provides a rigorous criterion for reducing the number of particles—and consequently the computational cost—without compromising convergence guarantees.

Our main theoretical contribution is to prove the **global convergence** of this scheme. The *Stochastic Conic Particle Gradient Descent with Birth and Death* is a discrete-time algorithm that constructs a sequence of measures  $(\nu_k)_{k \geq 0}$  from the discretized version of the objective function  $J(\cdot)$  by encoding the measures as a finite sum of *particles* (Dirac masses). Let  $(\varepsilon_k)_k$  be a sequence of *exploration parameters* controlling the intensity of the birth process at iteration  $k$ . At each iteration, the algorithm performs the following steps:

- **Weight and Push-Forward Update:** Each particle’s weight is updated via an exponential weighting scheme based on the local value of the dual certificate  $J'_{\nu_k}$  (the so-called *conic descent*). Simultaneously, the positions of the particles are adjusted using a so-called *generalized gradient descent step* to ensure they remain within the domain  $\mathcal{X}$ . A descent lemma (Proposition 2) quantifies the decrease in the objective function due to these updates, giving the intermediate update  $\nu_k \mapsto \nu_{k+}$ .
- **Birth Process:** New particles are introduced by sampling from regions  $\mathcal{N}_{\nu_{k+}}$  where the (updated) dual certificate  $J'_{\nu_{k+}}$  is negative (which can be done in practice using *rejection sampling*, for instance). The number of new particles added is proportional to the *exploration schedule parameter*  $\varepsilon_k$ , which decays over time to balance exploration and exploitation.
- **Death Process:** Particles located in regions  $\mathcal{P}_{\nu_{k+}}$  where  $J'_{\nu_{k+}}$  is sufficiently large are safely removed (with theoretical guarantees) from the measure. This pruning step helps control the total number of particles, ensuring computational efficiency.
- **Stochastic Gradient Estimation:** To further enhance scalability, the algorithm employs stochastic approximations of the gradient using *mini-batches* of data and *random feature* mappings (sketching). This reduces the computational burden associated with evaluating the full gradient at each iteration.

We start by presenting the continuous sparse regression framework and the key mathematical objects involved in the optimization problem. We then detail the **Fast Spawn&Prune** algorithm, including the weight and push-forward updates, as well as the Birth and Death processes. Next, we derive explicit convergence rates that depend on the dimension of the domain  $\mathcal{X}$ , reflecting the cost of global exploration. We show that, under suitable assumptions on the exploration schedule  $(\varepsilon_k)_k$ , the sequence of iterates converges to the global optimum.

### 1.1 Continuous sparse regression

Let  $\mathcal{X} \subset \mathbb{R}^d$  be a *compact convex* set (equal to the closure of its interior) and consider  $(\mathcal{M}(\mathcal{X}), \|\cdot\|_{TV})$  the space of *signed* measures, defined as the topological dual space of the space  $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$ , the *continuous* functions endowed with the infinity norm. Let  $\mathbb{H}$  be a *separable* Hilbert space and let  $\Phi : \mathcal{M}(\mathcal{X}) \rightarrow \mathbb{H}$  be a linear map, referred to as the (forward) measurement operator. We define the BLASSO problem (Candès and Fernandez-Granda, 2014; Azais et al., 2015) as

$$\mu^* \in \arg \min_{\mu \in \mathcal{M}(\mathcal{X})} J(\mu) \quad \text{where} \quad J(\mu) := \frac{1}{2} \|y - \Phi\mu\|_{\mathbb{H}}^2 + \kappa \|\mu\|_{TV}, \quad (\mathcal{P})$$

where  $\kappa > 0$  is a regularization parameter and  $y \in \mathbb{H}$  is some observation. We assume that  $\Phi$  is a *bounded linear* and *weak- $*$  continuous* operator and one can prove (see Lemma A.1 in the appendix) that

$$\Phi : \nu \in \mathcal{M}(\mathcal{X}) \mapsto \int_{\mathcal{X}} \varphi_t d\nu(t) \in \mathbb{H}, \quad (1a)$$

where  $t \in \mathcal{X} \mapsto \varphi_t \in \mathbb{H}$  denotes the *feature map*. One can define the *model kernel*  $K(\cdot, \cdot)$  as

$$\forall s, t \in \mathcal{X}, \quad K(s, t) := \langle \varphi_s, \varphi_t \rangle_{\mathbb{H}}. \quad (1b)$$

We make the following assumption on the program ( $\mathcal{P}$ ) throughout this paper, which is satisfied for standard kernels on compact sets  $\mathcal{X}$ , such as the Gaussian kernel for instance.

**Assumption ( $\mathcal{H}_{\mathcal{P}}$ ).** There exist constants  $\mathfrak{c}_{\mathcal{P}} > 0$  and  $\mathfrak{C}_{\mathcal{P}} > 0$  such that the observation  $y \in \mathbb{H}$  is **bounded** in  $\mathbb{H}$ , namely:

$$\|y\|_{\mathbb{H}} \leq \mathfrak{C}_{\mathcal{P}},$$

and the kernel  $K(\cdot, \cdot)$  introduced in (1b) is

- **Smooth:** Twice continuously differentiable

$$\max \left\{ \|K(\cdot, \cdot)\|_{\infty}, \|\nabla_s K(\cdot, \cdot)\|_{\infty}, \|\nabla_s^2 K(\cdot, \cdot)\|_{\infty} \right\} \leq \mathfrak{C}_{\mathcal{P}}, \quad (2a-(\mathcal{H}_{\mathcal{P}}))$$

where  $\|\cdot\|_{\infty}$  is the infinity norm (each case over all  $s, t \in \mathcal{X}$ ) of the absolute kernel value, the Euclidean norm of the kernel gradient with respect to the first variable and the operator norm of the kernel Hessian with respect to the first variable, respectively.

- **Normalized and positive:**

$$\forall s, t \in \mathcal{X}, \quad K(s, t) \geq \mathfrak{c}_{\mathcal{P}} > 0 \quad \text{and} \quad K(t, t) = 1. \quad (2b-(\mathcal{H}_{\mathcal{P}}))$$

**Remark 1.1** Note that Equation (2b-( $\mathcal{H}_{\mathcal{P}}$ )) yields  $K(s, t) \in [\mathfrak{c}_{\mathcal{P}}, 1]$ , for any  $s, t \in \mathcal{X}$ . Note also that (see Lemma A.2) the so-called kernel metric  $d_K(\cdot, \cdot)$  (see (Steinwart and Christmann, 2008, Eq. (4.80))) satisfies

$$\forall s, t \in \mathcal{X}, \quad d_K(s, t) := \|\varphi_t - \varphi_s\|_{\mathbb{H}} \leq \sqrt{\mathfrak{C}_{\mathcal{P}}} \|t - s\|, \quad (3)$$

where  $\|\cdot\|$  is the Euclidean norm on  $\mathbb{R}^d$ . We will use the Lipschitz property throughout this paper.

## 1.2 First order optimality condition

**Symmetrization trick** Following Chizat (2022), we address the optimization problem over the space of signed measures  $\mathcal{M}(\mathcal{X})$  by lifting it to the space of *non-negative* measures, denoted by  $\mathcal{M}_+(\mathcal{X})$ . This standard argument is presented in Section A.4 (in the appendix) and, from now on, we assume that the feasible set of ( $\mathcal{P}$ ) is  $\mathcal{M}_+(\mathcal{X})$ .

**Fréchet differentiation: the dual certificate** For any  $\nu \in \mathcal{M}_+(\mathcal{X})$ , the Fréchet derivative of  $J(\cdot)$  is denoted by  $J'_\nu \in \mathcal{C}(\mathcal{X})$  (referred to as the *dual certificate*). By Lemmas A.3 and A.4 (in the appendix), the dual certificate enjoys the following equality, for all  $\nu \in \mathcal{M}_+(\mathcal{X})$  and  $\sigma \in \mathcal{M}(\mathcal{X})$  such that  $\nu + \sigma \in \mathcal{M}_+(\mathcal{X})$ ,

$$J(\nu + \sigma) - J(\nu) = \langle J'_\nu, \sigma \rangle + \frac{1}{2} \|\Phi(\sigma)\|_{\mathbb{H}}^2, \quad (4a)$$

and

$$\forall t \in \mathcal{X}, \quad J'_\nu(t) = \langle \varphi_t, \Phi(\nu) - y \rangle_{\mathbb{H}} + \kappa. \quad (4b)$$

**Remark 1.2 (Lipschitz continuity of the dual certificate)** *As established in Lemma A.5 (see the appendix), the dual certificate  $J'_\nu$  is Lipschitz continuous with respect to the spatial variable  $t \in \mathcal{X}$ . Specifically, its Lipschitz constant  $\mathfrak{L}(\nu)$  satisfies*

$$\mathfrak{L}(\nu) \leq \sqrt{\mathfrak{C}_{\mathcal{P}}}(\mathfrak{C}_{\mathcal{P}} + \|\nu\|_{\text{TV}}),$$

where  $\mathfrak{C}_{\mathcal{P}}$  is the Lipschitz constant of the feature map  $t \mapsto \varphi_t$  (derived from the kernel smoothness in Assumption  $(\mathcal{H}_{\mathcal{P}})$ ). This smoothness is crucial for the descent lemma (Proposition 2), as it ensures the gradient does not vary too much between close particles. Note that this bound depends on the mass of the measure  $\|\nu\|_{\text{TV}}$ , which we prove remains bounded by a constant  $\mathfrak{C}_{\text{TV}} > 0$  throughout the algorithm. We can define the uniform Lipschitz constant  $\mathfrak{L}$  of the dual certificates  $J'_\nu$  as

$$\mathfrak{L} := \sqrt{\mathfrak{C}_{\mathcal{P}}}(\mathfrak{C}_{\mathcal{P}} + \mathfrak{C}_{\text{TV}}). \quad (4c)$$

**Remark 1.3 (Bounds on the dual certificate)** *Note that the dual certificate (4b) satisfies, for any  $\nu \in \mathcal{M}_+(\mathcal{X})$  and any  $t \in \mathcal{X}$ ,*

$$\|\nu\|_{\text{TV}} + \mathfrak{C}_{\mathcal{P}} + \kappa \geq J'_\nu(t) = \int_{\mathcal{X}} K(s, t) d\nu(s) - \langle \varphi_t, y \rangle_{\mathbb{H}} + \kappa \geq \mathfrak{c}_{\mathcal{P}} \|\nu\|_{\text{TV}} - \mathfrak{C}_{\mathcal{P}} + \kappa \quad (5)$$

under (2). The bound (5) will be used to prove that the measure updates  $(\nu_k)_k$  have bounded total variation.

**Optimality condition** The dual certificate  $J'_\nu$  plays a central role in solving the problem. In particular, it enables the characterization of solutions to our optimization problem, thereby emphasizing its importance. We have the following proposition; see, for instance, (Chizat, 2022, Proposition 3.1).

**Proposition 1 (Karush–Kuhn–Tucker (KKT) conditions)** *A measure  $\nu^*$  is a minimizer of  $\nu \mapsto J(\nu)$  if and only if  $J'_{\nu^*}(t) \geq 0$  for all  $t \in \mathcal{X}$  and  $J'_{\nu^*}(t) = 0$  when  $t$  belongs to the support of  $\nu^*$ .*

The KKT conditions in Proposition 1 provide a natural rationale for the particle birth and death dynamics proposed in this algorithm. First, the requirement that  $J'_{\nu^*} \geq 0$  implies that any region where the current dual certificate is negative ( $J'_\nu < 0$ ) corresponds to a local violation of optimality. This motivates a **Birth process** to inject mass into these under-represented areas. Second, the complementary slackness condition implies that the support of the optimal measure is contained within the zero level set of the dual certificate. Consequently, particles located in regions where  $J'_\nu$  is strictly positive should be pruned or down-weighted to reduce the objective value, thereby justifying a **Death process**.

### 1.3 Weight & Push-Forward update

To adopt an optimization perspective on the functional  $J(\cdot)$ , we can formulate an ideal algorithm that recursively generates a sequence of measures  $(\nu_k)_{k \geq 0}$  via a so-called conic gradient descent on  $J(\cdot)$ .

**Generalized descent on  $\mathcal{X}$**  While the dynamics of the weights of the measures are made explicit through an exponential-weight descent, we also need to update the positions of the support points of the measure. Since the measure is supported on  $\mathcal{X}$ , we must therefore constrain the particles to remain in  $\mathcal{X}$ . This is ensured by the following proximal approach, which we briefly describe below. We refer to Ghadimi et al. (2016) for further details. Given a step size  $\beta > 0$ , we define

$$\forall t \in \mathcal{X}, \quad \forall v \in \mathbb{R}^d, \quad t_{t,v,\beta}^+ := \arg \min_{u \in \mathcal{X}} \left\{ \langle u, v \rangle + \frac{1}{2\beta} \|u - t\|^2 \right\}. \quad (6a)$$

The *generalized gradient descent step* associated with a descent vector  $v$  is then defined as

$$\pi_{\mathcal{X}}(t, v, \beta) := \frac{t - t_{t,v,\beta}^+}{\beta} \quad \text{so that} \quad t_{t,v,\beta}^+ = t - \beta \pi_{\mathcal{X}}(t, v, \beta). \quad (6b)$$

**Descent Property** By considering step sizes  $\alpha, \beta > 0$ , whose values will be specified and discussed throughout the paper, we introduce for any measure  $\nu \in \mathcal{M}_+(\mathcal{X})$  the mappings  $\mathbb{W}_{\nu,\alpha} : \mathcal{X} \rightarrow \mathbb{R}$  and  $\mathbb{T}_{\nu,\beta} : \mathcal{X} \rightarrow \mathbb{R}^d$  defined as:

$$\forall t \in \mathcal{X}, \quad \mathbb{W}_{\nu,\alpha}(t) = e^{-\alpha J'_\nu(t)} \quad \text{and} \quad \mathbb{T}_{\nu,\beta}(t) = t - \beta \pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta).$$

**Definition 1.1 (Weight & Push-Forward update)** For any  $\nu \in \mathcal{M}_+(\mathcal{X})$ , define the update as  $\nu^+ := \mathbb{T}_{\nu,\beta}^\# \mathbb{W}_{\nu,\alpha} \nu$  where the weight update  $\mathbb{W}_{\nu,\alpha} \nu \in \mathcal{M}_+(\mathcal{X})$  is given by, for any Borel set  $\mathcal{B} \subseteq \mathcal{X}$ ,

$$(\mathbb{W}_{\nu,\alpha} \nu)(\mathcal{B}) = \int_{\mathcal{B}} \mathbb{W}_{\nu,\alpha}(t) d\nu(t) = \int_{\mathcal{B}} e^{-\alpha J'_\nu(t)} d\nu(t),$$

and the push-forward measure  $\mathbb{T}_{\nu,\beta}^\# \mu \in \mathcal{M}_+(\mathcal{X})$ , for any  $\mu \in \mathcal{M}_+(\mathcal{X})$ , is defined by:

$$\forall \psi \in \mathcal{C}(\mathcal{X}), \quad \int_{\mathcal{X}} \psi(t) d\mathbb{T}_{\nu,\beta}^\# \mu(t) = \int_{\mathcal{X}} \psi(\mathbb{T}_{\nu,\beta}(t)) d\mu(t).$$

The next result provides a quantitative characterization of the effect of the  $W_{\nu,\alpha}$  and  $T_{\nu,\beta}$  updates on the value of the objective  $J(\cdot)$ .

**Proposition 2 (Descent property)** *Assume that (2) holds. Then, for any  $\nu \in \mathcal{M}_+(\mathcal{X})$  and  $\mathfrak{C}_{TV} > 0$  such that  $\|\nu\|_{TV} \leq \mathfrak{C}_{TV}$  and for any  $\alpha \geq 0$  and  $\beta \geq 0$  such that:*

$$\alpha < \frac{1}{10(1 + \mathfrak{C}_{TV} + \mathfrak{C}_{\mathcal{P}} + \kappa)(1 \vee \mathfrak{C}_{TV})} \quad \text{and} \quad \beta \leq \frac{1}{2\mathfrak{C}_{\mathcal{P}}(\mathfrak{C}_{\mathcal{P}} + 3\mathfrak{C}_{TV})e^{1/5}}, \quad (7)$$

it holds that

$$J(\nu^+) - J(\nu) \leq -\frac{3}{4} \left( \alpha \int_{\mathcal{X}} |J'_{\nu}|^2 d\nu + \beta \int_{\mathcal{X}} \|\pi_{\mathcal{X}}(t, \nabla J'_{\nu}(t), \beta)\|^2 d\nu \right).$$

where  $\nu^+ := T_{\nu,\beta}^{\#} W_{\nu,\alpha} \nu$ .

The proof is deferred to Appendix B. This result highlights both the strengths and limitations of the dynamics  $\nu \mapsto \nu^+$ , thereby motivating the algorithmic modifications introduced below. First, the Weight & Push-Forward update (Definition 1.1) enables the definition of an iterative sequence of measures  $(\nu_k)_{k \geq 1}$ . Proposition 2 quantifies the associated “descent” property: it establishes a lower bound on the decrease of the objective, guaranteeing a minimal energy reduction for the transition  $\nu \mapsto \nu^+$ . Secondly, while Chizat (2022); De Castro et al. (2021) prove that the sequence  $(\nu_k)_{k \geq 1}$  converges to a sparse measure  $\nu_{\infty}$  satisfying  $J'_{\nu_{\infty}} = 0$  on its support, they do not guarantee  $J'_{\nu_{\infty}} \geq 0$  elsewhere. Thus, standard conic particle gradient descent fails to satisfy the full optimality conditions of Proposition 1 due to insufficient exploration of  $\mathcal{X}$ , and global convergence cannot be guaranteed in this setting.

#### 1.4 Birth and Death Stochastic conic particle gradient descent

**The particle swarms and their non-convex program** We consider a generic non-negative measure composed of  $p$  Dirac masses, referred to as a *particle swarm*. Let  $\mathbf{W} := (\omega_1, \dots, \omega_p) \in \mathbb{R}^p$ ,  $\mathbf{T} := (t_1, \dots, t_p) \in \mathbb{R}^{p \times d}$ , and  $\boldsymbol{\kappa} := (\kappa_1, \dots, \kappa_p) \in \mathbb{R}^p$ . We introduce

$$\nu(\mathbf{W}, \mathbf{T}) := \sum_{j=1}^p \omega_j \delta_{t_j}, \quad (8a)$$

where, for any  $j$ ,  $\omega_j > 0$  denotes the weight of particle  $j$ . Then, the parametrization of  $\nu$  in (8a) yields:

$$J(\nu(\mathbf{W}, \mathbf{T})) = F(\mathbf{W}, \mathbf{T}) + \frac{1}{2} \|y\|_{\mathbb{H}}^2 \quad \text{with} \quad F(\mathbf{W}, \mathbf{T}) := \langle \boldsymbol{\kappa} - k_{\mathbf{T}}, \mathbf{W} \rangle + \frac{1}{2} \mathbf{W}^{\top} K_{\mathbf{T}} \mathbf{W}, \quad (8b)$$

where  $\varphi_t \in \mathbb{H}$  is the feature map,  $k_{\mathbf{T}} := (\langle y, \varphi_{t_1} \rangle_{\mathbb{H}}, \dots, \langle y, \varphi_{t_p} \rangle_{\mathbb{H}}) \in \mathbb{R}^p$ , and  $K_{\mathbf{T}}$  is a  $(p \times p)$  kernel matrix with entries  $K(s, t) := \langle \varphi_s, \varphi_t \rangle_{\mathbb{H}}$ .

While  $(\mathcal{P})$  is convex over the space of measures, the parametric formulation (8b) is non-convex due to the joint optimization of weights and positions. Nevertheless, if a solution to  $(\mathcal{P})$  can be written as a particle swarm (8a) then the minimizer  $(\mathbf{W}^*, \mathbf{T}^*)$  of  $F(\cdot, \cdot)$  is a global solution to  $(\mathcal{P})$ . The existence of such sparse solutions is well-established under some conditions, see for instance Duval and Peyré (2015); Boyer et al. (2019).

**The stochastic version of the dual-certificate and its gradients** In the following, we assume access to *unbiased* stochastic estimators of both the dual certificate  $J'_\nu$  and its spatial gradient  $\nabla J'_\nu$ . We will denote by  $\widehat{J}'_\nu(\cdot, Z)$  (resp.  $\widehat{D}_\nu(\cdot, Z)$ ) the estimator of  $J'_\nu$  (resp.  $\nabla J'_\nu$ ) for some random variable  $Z$  that captures the randomness of these approximations. We refer to (De Castro et al., 2025a, Section 3) for examples and explicit constructions. In the following, we will require some properties on these estimators, as described in the following assumption.

**Assumption ( $\widehat{\mathcal{H}}_{\text{sto}}$ ): Stochastic unbiased gradients.** A random variable  $Z$  exists such that:

- $\forall \nu \in \mathcal{M}_+(\mathcal{X})$  and  $\forall t \in \mathcal{X}$ ,
 
$$\begin{cases} \widehat{J}'_\nu(t, Z) := J'_\nu(t) + \xi_\nu(t, Z) \\ \widehat{D}_\nu(t, Z) := \nabla J'_\nu(t) + \zeta_\nu(t, Z) \end{cases} \quad \text{with } \mathbb{E}_Z[\xi_\nu(t, Z)] = 0 \text{ and } \mathbb{E}_Z[\zeta_\nu(t, Z)] = 0_{\mathbb{R}^d}. \quad (9a)$$

- There exist positive constants  $\mathbf{H}, \mathbf{G}, \mathbf{E}_\infty > 0$  such that, almost surely,
 
$$\forall t \in \mathcal{X}, \quad |\xi_\nu(t, Z)| \vee \|\zeta_\nu(t, Z)\| \leq \mathbf{E}_\infty \quad \text{and} \quad \widehat{J}'_\nu(t, Z) \geq \mathbf{G} \|\nu\|_{\text{TV}} - \mathbf{H} + \kappa. \quad (9b)$$
- Almost surely, the *stochastic dual-certificate*  $t \mapsto \widehat{J}'_\nu(t, Z)$  is uniformly  $\mathfrak{L}$ -Lipschitz, regardless of the value of  $Z$  (the constant  $\mathfrak{L}$  may be taken to be larger than the one appearing in (4c) if necessary).

**Remark 1.4** In (9b), we assume a Hoeffding-type condition on the centered random variables, along with a lower bound on the stochastic dual certificate. The constant  $\mathbf{E}_\infty$  stands for the maximal size of the admissible noise level that perturbs the computation of  $\widehat{J}'_\nu(t, Z)$  and  $\widehat{D}_\nu(t, Z)$  at each iteration. Such a condition could be replaced by a sub-Gaussian assumption on the distribution of the noise. Regarding the lower bound (affine in  $\|\nu\|_{\text{TV}}$ ), note that it holds for the deterministic dual certificate under  $(\mathcal{H}_\mathcal{P})$ ; see Equation (5). This assumption will be used to show that the stochastic measure updates  $(\hat{\nu}_k)_{k \geq 0}$  remain bounded in total variation norm.

Interestingly, we can use these stochastic counterparts with a mini-batch strategy to reduce the variance of the approximations. More precisely, at each step  $k \geq 1$ , given  $m_k \geq 1$ , a mini-batch sample size, we draw a  $m_k$ -sample of i.i.d. random variables  $\mathbf{Z}_{k+1} := (Z_{1,k+1}, \dots, Z_{m_k,k+1})$  satisfying (9a) and set

$$\widehat{J}'_{\hat{\nu}_k}(t) := \frac{1}{m_k} \sum_{l=1}^{m_k} \widehat{J}'_{\hat{\nu}_k}(t, Z_{l,k+1}) \quad \text{and} \quad \widehat{D}_k(t) := \frac{1}{m_k} \sum_{l=1}^{m_k} \widehat{D}_{\hat{\nu}_k}(t, Z_{l,k+1}). \quad (10a)$$

Given a measure  $\hat{\nu}_k = \sum_{j=1}^{p_k} \omega_j^k \delta_{t_j^k}$  (as in (8a)) composed of  $p_k$  particles, the Push-Forward update (Definition 1.1) leads to the following stochastic update  $\hat{T}_\#^{\alpha, \beta}$  at step  $k \geq 1$ .

**Definition 1.2 (Stochastic push-forward update  $\hat{T}_\#^{\alpha, \beta}$ )** Define  $\hat{\nu}_{k+} = \hat{T}_\#^{\alpha, \beta}(\hat{\nu}_k)$  as  $\hat{\nu}_{k+} = \sum_{j=1}^{p_k} \omega_j^{k+} \delta_{t_j^{k+}}$  with

$$\omega_j^{k+} = \omega_j^k e^{-\alpha \widehat{J}'_{\hat{\nu}_k}(t_j^k)} \quad \text{and} \quad t_j^{k+} = t_j^k - \beta \pi_{\mathcal{X}}(t_j^k, \widehat{D}_k(t_j^k), \beta). \quad (10b)$$

where  $\pi_{\mathcal{X}}(\cdot, \cdot, \cdot)$  denotes the projection operator over  $\mathcal{X}$  introduced in Section 1.3.

**The mass tweaking (Birth and Death)** Given the stochastic push-forward update  $\hat{\nu}_{k+} = \sum_{j=1}^{p_k} \omega_j^{k+} \delta_{t_j^{k+}}$ , we now describe our stochastic update  $\hat{\nu}_{k+} \mapsto \hat{\nu}_{k+1}$ , whose evolution involves both deletion  $\hat{\nu}_{k+} \mapsto \hat{\nu}_k^{++}$  and creation  $\hat{\nu}_k^{++} \mapsto \hat{\nu}_{k+1}$  of weighted particles. To this end, we define the *pushed dual certificate*  $\widehat{J'_{\hat{\nu}_{k+}}}$  as in (10a) with a new independent  $m_k$  mini-batch  $\mathbf{Z}_{k+1}^+$  and the particle swarm  $\hat{\nu}_{k+}$  (see Step 6 of Algorithm 1). We also need a *decision rule*  $\text{DR} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  that takes as input the pushed dual certificate values and the push-forward weights  $(\widehat{J'_{\hat{\nu}_{k+}}}(t_j^{k+}), \omega_j^{k+})$  and outputs a deletion intensity. Finally, we are given some positivity (resp. negativity) schedules  $(\widehat{C}_k)_k$  (resp.  $(\widehat{c}_k)_k$ ).

Again, mass deletion only concerns regions where  $\widehat{J'_{\hat{\nu}_{k+}}} \geq 0$  (given by (10a) with a new independent  $m_k$  mini-batch  $\mathbf{Z}_{k+1}^+$  and the particle swarm  $\hat{\nu}_{k+}$ ) whereas our algorithm adds some mass in regions where  $\widehat{J'_{\hat{\nu}_{k+}}} \leq 0$ . For this purpose, we use the super-level (resp. sub-level) set of positivity (resp. negativity) of  $\widehat{J'_{\hat{\nu}_{k+}}}$ , *i.e.*, we define  $\widehat{\mathcal{P}}_{\nu_{k+}} \subseteq \text{Supp}(\hat{\nu}_{k+})$  and  $\widehat{\mathcal{N}}_{\nu_{k+}} \subseteq \mathcal{X}$  as:

$$\widehat{\mathcal{P}}_{\nu_{k+}} := \left\{ t \in \text{Supp}(\hat{\nu}_{k+}) : \text{DR}(\widehat{J'_{\hat{\nu}_{k+}}}(t), \hat{\nu}_{k+}(\{t\})) \geq \widehat{C}_k \right\} \quad (11a)$$

$$\widehat{\mathcal{N}}_{\nu_{k+}} := \left\{ t \in \mathcal{X} : \widehat{J'_{\hat{\nu}_{k+}}}(t) \leq \widehat{c}_k \right\}, \quad (11b)$$

where  $\text{Supp}(\cdot)$  denotes the support of a measure and  $(\widehat{c}_k)_k$  is a sequence of (small) positive numbers. We construct  $\omega_j^{k+} \rightarrow \omega_j^{k++}$  by removing mass on  $\widehat{\mathcal{P}}_{\nu_{k+}}$  and  $\hat{\nu}_k^{++} \rightarrow \hat{\nu}_{k+1}$  adding mass on  $\widehat{\mathcal{N}}_{\nu_{k+}}$ . Define

$$\omega_j^{k++} := (1 - \mathbf{1}_{\widehat{\mathcal{P}}_{\nu_{k+}}}(t_j^{k+})) \omega_j^{k+}, \quad (12a)$$

$$\hat{\nu}_{k+1} := \sum_{j=1}^{p_k} \omega_j^{k++} \delta_{t_j^{k+}} + \varepsilon_k \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}}(U_{k+1}) \delta_{U_{k+1}}, \quad (12b)$$

where the random variables  $(U_l)_{l \in \mathbb{N}}$  are assumed independent from the other random variables sampled at step  $k$  and uniformly sampled over  $\mathcal{X}$ , and, for any Borel set  $\mathcal{B} \subseteq \mathcal{X}$ ,  $\mathbf{1}_{\mathcal{B}}(\cdot)$  denotes the indicator function of  $\mathcal{B}$ . The scheme (12b) is implementable as it only requires the generation of a uniform random variable over the space  $\mathcal{X}$  and a single evaluation of  $\widehat{J'_{\hat{\nu}_{k+}}}$ . Also, we emphasize that the Positivity  $(\widehat{C}_k)_k$  and Negativity  $(\widehat{c}_k)_k$  schedules can be chosen adaptively depending on the stochastic dual certificate  $\widehat{J'_{\hat{\nu}_{k+}}}$  and the weights  $\omega_j^{k+}$ . For instance, a valid strategy for the Death process is to target the particle with the largest ratio certificate value over weight, *i.e.*, by setting a threshold related to  $\max_j \{ \widehat{J'_{\hat{\nu}_{k+}}}(t_j^{k+}) / \omega_j^{k+} \}$ . Examples of explicit decision rules and specific tuning for constants are provided in Section 3.2.

**The Fast Spawn&Prune Algorithm** We have now all the ingredients to design an implementable procedure. The previous steps are gathered in Algorithm 1.

---

**Algorithm 1** Birth and Death Stochastic Conic Particle Gradient Descent (FS&P)

---

**Require:** Learning rates  $\alpha, \beta > 0$ ; Mini-batch size schedule  $(m_k)_{k \geq 1}$ ; Exploration schedule  $(\varepsilon_k)_{k \geq 1}$ ; Positivity (resp. Negativity) schedules  $(\widehat{C}_k)_k$  (resp.  $(\widehat{c}_k)_k$ ); Decision Rule DR :  $\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ ;

- 1: Weights:  $\mathbf{W}^k$  and Positions:  $\mathbf{T}^k$ ; ▷ No specific initialization required
- 2: **for**  $k = 1, \dots, K$  **do** ▷  $K$  gradient steps
- 3:   Set  $\hat{\nu}_k \leftarrow \nu(\mathbf{W}^k, \mathbf{T}^k)$ ; ▷ Particle swarm
- 4:   Sample  $\mathbf{Z}_{k+1} \leftarrow (Z_{1,k+1}, \dots, Z_{m_k,k+1})$  and compute stochastic values and gradients

$$\widehat{J}'_{\hat{\nu}_k}(t_j^k) := \frac{1}{m_k} \sum_{\ell=1}^{m_k} \widehat{J}'_{\hat{\nu}_k}(t_j^k, Z_{\ell,k+1}) \quad \text{and} \quad \widehat{D}_k(t_j^k) := \frac{1}{m_k} \sum_{\ell=1}^{m_k} \widehat{D}_{\nu_k}(t_j^k, Z_{\ell,k+1});$$

▷ Stochastic mini-batch variables (10a);

- 5:   Update weights and positions  $\hat{\nu}_{k+} \leftarrow \sum_{j=1}^{p_k} \omega_j^{k+} \delta_{t_j^{k+}}$  with

$$\omega_j^{k+} = \omega_j^k e^{-\alpha \widehat{J}'_{\hat{\nu}_k}(t_j^k)} \quad \text{and} \quad t_j^{k+} = t_j^k - \beta \pi_{\mathcal{X}}(t_j^k, \widehat{D}_k(t_j^k), \beta);$$

▷ Stochastic push-forward update (10b);

- 6:   Sample  $\mathbf{Z}_{k+1}^+ \leftarrow (Z_{1,k+1}^+, \dots, Z_{m_k,k+1}^+)$  and compute stochastic pushed dual certificate

$$\widehat{J}'_{\hat{\nu}_{k+}}(\cdot) \leftarrow \frac{1}{m_k} \sum_{\ell=1}^{m_k} \widehat{J}'_{\hat{\nu}_{k+}}(\cdot, Z_{\ell,k+1}^+);$$

▷ Stochastic mini-batch variables (10a);

- 7:   Sample  $U_{k+1} \sim \mathcal{U}_{\mathcal{X}}$  independent from the rest (Uniform measure on  $\mathcal{X}$ ) and compute

$$\begin{aligned} \mathbf{1}_{\widehat{\mathcal{P}}_{\nu_{k+}}}(t_j^{k+}) &\leftarrow 1 \text{ if } \left\{ \text{DR}(\widehat{J}'_{\hat{\nu}_{k+}}(t_j^{k+}), \omega_j^{k+}) \geq \widehat{C}_k \right\} \text{ and } 0 \text{ otherwise;} \\ \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}}(U_{k+1}) &\leftarrow 1 \text{ if } \left\{ \widehat{J}'_{\hat{\nu}_{k+}}(U_{k+1}) \leq \widehat{c}_k \right\} \text{ and } 0 \text{ otherwise;} \\ \hat{\nu}_{k+1} &\leftarrow \sum_{j=1}^{p_k} (1 - \mathbf{1}_{\widehat{\mathcal{P}}_{\nu_{k+}}}(t_j^{k+})) \omega_j^{k+} \delta_{t_j^{k+}} + \varepsilon_k \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}}(U_{k+1}) \delta_{U_{k+1}}; \end{aligned}$$

▷ Mass tweaking (12);

8: **end for**

---

## 1.5 Global convergence results

Our theoretical analysis proceeds in two steps. First, we analyze a deterministic version of the algorithm with continuous updates (Section 2). Theorem 2.1 establishes that this method converges to the global optimum  $\mu^*$ , escaping local minima thanks to the Birth process. We derive an explicit convergence rate of order  $O(K^{-\frac{1}{2(2+d)}})$  for the minimum gap

$\min_{k \leq K} \{J(\nu_k) - J(\mu^*)\}$ , where  $d$  is the dimension of the domain  $\mathcal{X}$ . This dependence on  $d$  reflects the computational cost of global exploration in a non-convex landscape.

Second, we extend these results to the fully stochastic FS&P algorithm (Section 3). Theorem 3.1 proves that, under suitable choices of learning rates, mini-batch sizes, and exploration schedules, the expected excess risk converges to zero, yielding a global minimization result. Specifically, we obtain a global convergence rate of order  $O((\log K/K)^{\frac{1}{2(2+d)}})$ , confirming that FS&P achieves global optimization with computationally efficient stochastic updates. The mini-batch size scales as  $m = K$ , resulting in a total sample complexity of  $\mathcal{O}(N^{-1/(4(2+d))})$  up to logarithmic factors (Corollary 3.1). Finally, Theorem 3.2 provides a horizon-free variant with iteration-dependent schedules ( $m_k = k$ ,  $\varepsilon_k = 1/\sqrt{k}$ ,  $\beta_k = 1/k$ ) that achieves the same sample complexity rate without requiring prior knowledge of the total number of iterations.

## 1.6 Related works

**Convex programming for sparse optimization on measures.** Continuous sparse regression, often framed as the BLASSO problem, has been extensively studied through the lens of convex optimization. Early foundational works by Candès and Fernandez-Granda (2014), Azais et al. (2015), and Duval and Peyré (2015) established exact recovery guarantees using semidefinite programming or grid-free methods, focusing on the statistical properties of the minimizer. More recent contributions, such as Poon et al. (2023), Giard et al. (2025), and De Castro et al. (2025b), have further refined these statistical error bounds and extended the analysis to various geometries and metrics. However, these works generally analyze the static optimization problem rather than the algorithmic dynamics required to solve it efficiently in high dimensions.

**Over-parameterized Gradient Descent and Global Convergence.** The dynamic approach, which involves optimizing particle positions and weights via gradient descent, relies heavily on over-parametrization. The seminal works of Chizat and Bach (2018) and Chizat (2022) analyzed the mean-field limit of these particle systems. They established that in the regime where the number of particles tends to infinity, the gradient flow converges to the global optimum, provided that the initialization covers the entire domain. However, for a finite number of particles, Chizat (2022) only guarantees local convergence to stationary points, or global convergence under restrictive assumptions, such as an exponential number of particles at initialization. The gap between the global convergence of the continuous flow and the local convergence of the discrete algorithm remains a significant theoretical hurdle.

**Stochastic Algorithms and FastPart.** To address the computational complexity of deterministic gradient descent, which scales quadratically with the number of particles, stochastic approximations were introduced in De Castro et al. (2025a). This method, referred to as FastPart, utilizes mini-batching and random features to achieve a time complexity of  $O(1)$  per iteration with respect to the number of particles. While De Castro et al. (2025a) proved the stability of the algorithm (boundedness of the total variation norm) and established convergence rates to stationary points of order  $O(\log K/\sqrt{K})$ , it did not guarantee global convergence from arbitrary sparse initializations. This work builds upon that foundation by integrating a mechanism to escape local minima.

**Birth and Death Processes in Optimization.** The idea of adding mass to ensure global optimality has antecedents in the Frank-Wolfe (conditional gradient) algorithm Bredies and Pikkarainen (2013), where particles are added iteratively to the support. However, Frank-Wolfe methods require locating the new atom at the global minimum of the Fréchet derivative (the dual certificate, see (4b)), which amounts to solving a non-convex optimization problem exactly. Our Birth process is considerably more flexible: it suffices to draw a random point from a sub-level set of the Fréchet derivative (defined by the threshold  $\hat{c}_k$  in Algorithm 1). This flexibility comes at a price—a sub-linear convergence rate—but one that is dimension-free up to the exponent  $1/(2+d)$ , thereby quantifying the cost of global exploration via the Birth process. The Death process is essentially harmless: it does not prevent the loss from decreasing, yet it reduces the per-iteration complexity of the algorithm. Our Death procedure is mathematically grounded in the Fréchet derivative and provides a principled criterion for safely removing particles from the support. To the best of our knowledge, this point of view is new.

## 1.7 Notation

Throughout the paper,  $\mathfrak{C} > 0$  denotes a generic constant used for upper bounds, while  $\mathfrak{c} > 0$  denotes a generic constant used for lower bounds; both  $\mathfrak{C}$  and  $\mathfrak{c}$  may change from line to line. Both constants are independent of  $k$  and  $d$ . The Euclidean norm of a vector  $x \in \mathbb{R}^d$  is denoted by  $\|x\|$ . For any set  $A$ , we denote by  $\mathbf{1}_A$  its indicator function. The support of a measure  $\mu$  is denoted by  $\text{Supp}(\mu)$ . A list of notation is provided in Table 4 in Appendix G.

## 2 Birth process for the deterministic CPGD

Our objective in this section is to develop ideas that enable an effective exploration of the space  $\mathcal{X}$ . We begin by focusing on a simplified setting in which the measures  $\nu_k$  remain continuous throughout the iterative process, and restrict to weight-only updates ( $\beta = 0$  in Definition 1.1). This framework is convenient for developing and understanding the theoretical tools that lead to global convergence of the optimization procedure, which will be extended to the stochastic setting in Section 3. The extension to position updates ( $\beta > 0$ ) is carried out in Appendix E for completeness.

### 2.1 Update evolution

The limitations of Proposition 2 lie in the fact that it provides no information about the sign of  $J'_{\nu_k}$  on the entire domain  $\mathcal{X}$ . The core idea we pursue is to explicitly add some mass, at each iteration  $k$ , on subsets of  $\mathcal{X}$  where  $J'_{\nu_k}$  is negative. To this end, we design an iterative algorithm that generates, at each step  $k$ , a triple of positive measures  $(\nu_k, \nu_{k+}, \nu_{k+1})$ : the intermediate measure  $\nu_{k+}$  is computed by performing a gradient descent step on  $J$  starting from  $\nu_k$ , while  $\nu_{k+1}$  is obtained by modifying the mass of  $\nu_{k+}$  in relevant regions. Specifically, the transition  $\nu_k \mapsto \nu_{k+}$  corresponds to a weight update only—i.e., setting  $\beta = 0$  in Definition 1.1. In contrast, the transition  $\nu_{k+} \mapsto \nu_{k+1}$  involves modifying the mass in regions where  $J'_{\nu_{k+}} \leq 0$ , and possibly removing mass from regions where  $J'_{\nu_{k+}} \geq 0$ . Such a construction ensures that the support of  $\nu_k$  remains included in  $\mathcal{X}$  throughout the iterations, thereby avoiding the need for any projection or correction steps. Formally, the iterative

scheme alternates between two steps. Let  $\nu_0$  be any positive measure supported on  $\mathcal{X}$  with  $\|\nu_0\|_{\text{TV}} < \infty$ .

**Weight update** ( $k \mapsto k^+$ ). The intermediate measure  $\nu_{k^+}$  is obtained through

$$\nu_{k^+} = \mathbb{W}_{\nu_k, \alpha} \nu_k. \quad (13)$$

**Birth-death step** ( $k^+ \mapsto k+1$ ). The measure  $\nu_{k+1}$  is obtained by modifying the mass of  $\nu_{k^+}$ . We require this step to satisfy Assumptions (14) and (15–( $\mathcal{H}_{\text{TV}}^\infty$ )) defined below.

**Assumption** ( $\mathcal{H}_\varepsilon$ ). Let  $(\varepsilon_k)_{k \geq 0}$  be a decreasing sequence with  $\varepsilon_0 = 1$ , and let  $\lambda$  denote the Lebesgue measure. The transition  $\nu_{k^+} \mapsto \nu_{k+1}$  satisfies ( $\mathcal{H}_\varepsilon$ ) if the following three conditions hold for every  $k \geq 1$ :

$$J(\nu_{k+1}) - J(\nu_{k^+}) \leq \mathfrak{C} \varepsilon_k^2, \quad (14\text{a}-(\mathcal{H}_\varepsilon^{\text{smooth},1}))$$

$$\|J'_{\nu_{k+1}} - J'_{\nu_{k^+}}\|_\infty \leq \mathfrak{C} \varepsilon_k, \quad (14\text{b}-(\mathcal{H}_\varepsilon^{\text{smooth},2}))$$

$$\nu_{k+1}(\mathcal{B}) \geq \varepsilon_k \lambda(\mathcal{B} \cap \{J'_{\nu_{k^+}} \leq 0\}), \quad (14\text{c}-(\mathcal{H}_\varepsilon^+))$$

for any Borel set  $\mathcal{B} \subseteq \mathcal{X}$ .

**Assumption** ( $\mathcal{H}_{\text{TV}}^\infty$ ). A constant  $\mathfrak{C}_{\text{TV}} > 0$  exists such that

$$\|\nu_k\|_{\text{TV}} \leq \mathfrak{C}_{\text{TV}} \quad \forall k \in \mathbb{N}. \quad (15-(\mathcal{H}_{\text{TV}}^\infty))$$

Note that Assumption (15–( $\mathcal{H}_{\text{TV}}^\infty$ )) implies that the sequence of values  $J(\nu_k)$  remains bounded throughout the iterations. Indeed, a rough computation yields:

$$J(\nu_k) \leq \frac{1}{2} (\|y\|_{\mathbb{H}} + \|\Phi \nu_k\|_{\mathbb{H}})^2 + \kappa \|\nu_k\|_{\text{TV}} \leq \frac{1}{2} (\|y\|_{\mathbb{H}} + \mathfrak{C}_{\text{TV}})^2 + \kappa \mathfrak{C}_{\text{TV}}. \quad (16)$$

We emphasize that, at this stage, no specific update rule has been defined yet for constructing  $\nu_{k+1}$  from  $\nu_{k^+}$ . In the next paragraph (Section 2.2), we will present a strategy that is algorithmically well-motivated and satisfies both assumptions (14) and (15–( $\mathcal{H}_{\text{TV}}^\infty$ )). Such a strategy may not be unique, and other examples of transition rules could be proposed. Our objective is to show that, as long as properties (14) and (15–( $\mathcal{H}_{\text{TV}}^\infty$ )) hold, we obtain global convergence results with explicit convergence rates. In Section 2.4, we leverage Proposition 2 to derive the global convergence of our method, along with explicit rates.

## 2.2 Transition $\nu_{k^+} \mapsto \nu_{k+1}$

In this section, we describe the key ingredients for designing a transition

$$\nu_{k^+} \mapsto \nu_{k^{++}} \mapsto \nu_{k+1}$$

that ensures our assumptions (14) and (15–( $\mathcal{H}_{\text{TV}}^\infty$ )) are satisfied. This transition is divided into two steps. The first step consists of removing mass from regions where  $J'_{\nu_{k^+}} \geq 0$ , which typically leads to a decrease in the energy  $J$ . The second step involves adding mass in regions where  $J'_{\nu_{k^+}} \leq 0$  in order to enhance the effectiveness of the subsequent weight update in the transition  $\nu_{k^{++}} \mapsto \nu_{k+1}$ .

**Mass deletion**  $\nu_{k+} \rightarrow \nu_{k++}$ . The idea is to remove some mass on the set where  $J'_{\nu_{k+}}$  is positive. More precisely, we can decide to cancel some subset of the domain  $\mathcal{X}$  by considering

$$\nu_{k++} = \nu_{k+}(1 - \mathbf{1}_{\mathcal{P}_{\nu_{k+}}}) \quad \text{with} \quad \mathcal{P}_{\nu_{k+}} = \left\{ J'_{\nu_{k+}} > -2\alpha^{-1} \log \varepsilon_k + \mathfrak{C}_w \right\} \cap \left\{ J'_{\nu_{k+}} > 0 \right\}, \quad (17)$$

where  $\mathfrak{C}_w$  is a constant defined below in Remark 2.1.

**Remark 2.1 (Weight-update perturbation bound)** *The deletion set  $\mathcal{P}_{\nu_{k+}}$  is defined in terms of  $J'_{\nu_{k+}}$ , but the density identity  $d\nu_{k+}(t) = e^{-\alpha J'_{\nu_k}(t)} d\nu_k(t)$  involves  $J'_{\nu_k}$ . The constant  $\mathfrak{C}_w$  in (17) is the precise gauge that bridges the two. We claim:*

$$\|J'_{\nu_{k+}} - J'_{\nu_k}\|_\infty \leq \mathfrak{C}_w, \quad \mathfrak{C}_w := \alpha(\mathfrak{C}_{\mathcal{P}} + \mathfrak{C}_{TV} + \kappa) e^{\alpha(\mathfrak{C}_{\mathcal{P}} + \mathfrak{C}_{TV} + \kappa)} \mathfrak{C}_{TV}, \quad (18a)$$

$$\mathcal{P}_{\nu_{k+}} \subset \{J'_{\nu_k} > -2\alpha^{-1} \log \varepsilon_k\}, \quad \text{so that } e^{-\alpha J'_{\nu_k}(t)} < \varepsilon_k^2 \text{ on } \mathcal{P}_{\nu_{k+}}. \quad (18b)$$

Under condition (7),  $\alpha(\mathfrak{C}_{\mathcal{P}} + \mathfrak{C}_{TV} + \kappa) < 1/10$ , hence  $\mathfrak{C}_w < \mathfrak{C}_{TV}/5$ . The proofs of (18a) and (18b) are deferred to Section C.2.

**Mass creation**  $\nu_{k++} \rightarrow \nu_{k+1}$ . We introduce the set  $\mathcal{N}_{\nu_{k+}}$ , associated with the negative part of  $J'_{\nu_{k+}}$ , and defined as:

$$\mathcal{N}_{\nu_{k+}} := \{J'_{\nu_{k+}} \leq 0\}. \quad (19)$$

We emphasize that the set  $\mathcal{N}_{\nu_{k+}}$  is determined by the values of  $J'_{\nu_{k+}}$ , and this point will be carefully addressed below. We add some mass on  $\mathcal{N}_{\nu_{k+}}$  defined in Equation (19) and define:

$$\nu_{k+1} := \nu_{k++} + \varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda \quad (20)$$

Figure 1 provides a schematic representation to ease the understanding of the evolution from  $\nu_{k+}$  to  $\nu_{k+1}$ . We stress that our assumptions are general enough to allow for alternative update strategies. For the sake of clarity, we have only focused on one specific scheme, but alternative approaches could be investigated.

The next proposition ensures that the corresponding updated measure  $\nu_{k+1}$  also satisfies the required assumptions: the boundedness of the total variation norm (Assumption (15–( $\mathcal{H}_{TV}^\infty$ ))) and the  $\varepsilon$ -smoothness (Assumptions (14)). The proof is deferred to Section C.1.

**Proposition 3** *Assume (2). Let  $(\nu_k)_{k \geq 1}$  defined according to (17) and (20), and assume that the sequence  $(\varepsilon_k)_{k \geq 0}$  satisfies  $\forall k \geq 0 : \varepsilon_k \leq \alpha$ . Let us define*

$$\mathfrak{R} := \frac{\|y\|_{\mathbb{H}}}{\mathfrak{c}_{\mathcal{P}}} e + \sqrt{\frac{e^3 \lambda(\mathcal{X})}{\mathfrak{c}_{\mathcal{P}}}} + \lambda(\mathcal{X})$$

(i) *If we choose  $\alpha \leq \frac{1}{1+\mathfrak{R}}$ , then  $(\mathcal{H}_{TV}^\infty)$  holds and:*

$$\forall k \geq 0 \quad \|\nu_k\|_{TV} \leq \|\nu_0\|_{TV} \vee 2\mathfrak{R} := \mathfrak{C}_{TV}.$$

(ii) *If in addition  $(\alpha, \beta)$  satisfies (7) (with the  $\mathfrak{C}_{TV}$  of item i), then Assumption  $(\mathcal{H}_\varepsilon)$  holds true, i.e. Equations (14) ((14a–( $\mathcal{H}_\varepsilon^{\text{smooth},1}$ ))), (14b–( $\mathcal{H}_\varepsilon^{\text{smooth},2}$ ))) and (14c–( $\mathcal{H}_\varepsilon^+$ ))) are satisfied.*

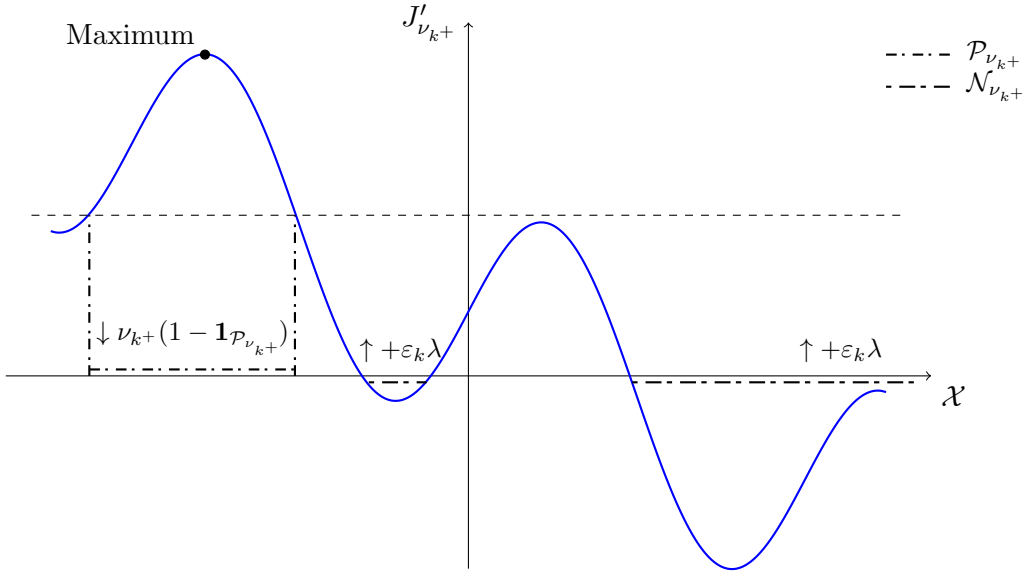


Figure 1: Evolution from  $\nu_{k+}$  to  $\nu_{k+1}$ : decrease on  $\mathcal{P}_{\nu_{k+}}$ , increase on  $\mathcal{N}_{\nu_{k+}}$ .

### 2.3 Discussion: the role of $\alpha$

Among the algorithmic parameters  $(\alpha, \beta, \varepsilon_k, m_k)$ , the weight learning rate  $\alpha$  plays a distinguished role: it is fixed *a priori* from problem data alone, while every other parameter is then calibrated as a function of  $\alpha$  and of intrinsic problem constants. We make this hierarchy explicit, since two upper bounds on  $\alpha$  appear in the analysis and a careless reading suggests a circular dependence.

**An apparent circularity.** The descent property of Proposition 2 requires Inequality (7), which involves the uniform total-variation bound  $\mathfrak{C}_{TV}$ . In turn,  $\mathfrak{C}_{TV}$  is produced by Proposition 3 (i), which itself requires  $\alpha \leq (1 + \mathfrak{R})^{-1}$ . Read in this order, the requirements appear to chain into a loop ( $\alpha$  small with  $\mathfrak{C}_{TV}$ , then  $\mathfrak{C}_{TV}$  from  $\alpha$  small).

**Resolution:  $\alpha$  is fixed first, from intrinsic data.** The loop is only apparent. The constant

$$\mathfrak{R} = \frac{\|y\|_{\mathbb{H}}}{\mathfrak{c}_{\mathcal{P}}} e + \sqrt{\frac{e^3 \lambda(\mathcal{X})}{\mathfrak{c}_{\mathcal{P}}}} + \lambda(\mathcal{X})$$

depends only on the problem data  $(\|y\|_{\mathbb{H}}, \mathfrak{c}_{\mathcal{P}}, \lambda(\mathcal{X}))$  and is independent of  $\alpha$ . As a consequence, so is the candidate TV bound

$$\mathfrak{C}_{TV} := \|\nu_0\|_{TV} \vee 2\mathfrak{R}. \quad (21a)$$

Substituting (21a) into (7) turns the descent constraint into a numerical one. Hence  $\alpha$  may be fixed once and for all according to the explicit, self-contained rule

$$0 < \alpha \leq \min \left\{ \frac{1}{1 + \mathfrak{R}}, \frac{1}{10(1 + \mathfrak{C}_{TV} + \mathfrak{c}_{\mathcal{P}} + \kappa)(1 \vee \mathfrak{C}_{TV})} \right\}, \quad (21b)$$

whose right-hand side depends only on  $(\|y\|_{\mathbb{H}}, \mathfrak{C}_{\mathcal{P}}, \kappa, \lambda(\mathcal{X}), \|\nu_0\|_{\text{TV}}, \mathfrak{C}_{\mathcal{P}})$ . Under (21b), Proposition 3 delivers both  $(\mathcal{H}_{\text{TV}}^{\infty})$  and  $(\mathcal{H}_{\varepsilon})$ , and the descent inequality of Proposition 2 applies along the trajectory  $(\nu_k)_{k \geq 0}$ .

**Cascade of derived parameters.** Once  $\alpha$  is fixed via (21b), the remaining parameters are calibrated downstream:

- the position learning rate must obey  $\beta \leq (2\mathfrak{C}_{\mathcal{P}}(\mathfrak{C}_{\mathcal{P}} + 3\mathfrak{C}_{\text{TV}}) e^{1/5})^{-1}$  (the second part of (7), with  $\mathfrak{C}_{\text{TV}}$  now an explicit numerical constant);
- the birth schedule satisfies  $\varepsilon_k \leq \alpha$  for all  $k$ , with the horizon-dependent or horizon-free calibrations specified in Theorem 2.1; the constraint  $\max_k \varepsilon_k \leq \alpha$  reflects that small learning rates slow the exponential decay induced by  $e^{-\alpha J_{\nu_k}^k}$ , so the injected mass  $\varepsilon_k \lambda$  may accumulate before being absorbed.

The same hierarchy governs the stochastic setting of Section 3: there,  $\mathfrak{R}$  is replaced by  $\widehat{\mathfrak{R}} = \frac{\mathfrak{H}}{\mathfrak{G}} e + \sqrt{e^3/\mathfrak{G}} + 1$ , intrinsic to Assumption (9b), and an additional Hoeffding cap  $\alpha \leq \sqrt{8 \log 8} / \mathbf{E}_{\infty}$  enters via (26a); the downstream calibration of  $\beta$ ,  $\varepsilon_k$  and  $m_k$  proceeds identically.

**Optimal magnitude of  $\alpha$ .** Within the admissible range (21b), the rates of Theorem 2.1 scale as  $\alpha^{-1/(2+d)}$ , and those of Theorem 3.1 as  $\alpha^{-1/2}$ : it is therefore advantageous to take  $\alpha$  as large as the constraints allow. This is the standard situation in deterministic optimization, where the descent inequality enforces a ceiling on the step size and saturating it yields the best worst-case rate.

## 2.4 Global convergence result

We now establish a convergence result for the sequence  $(J(\nu_k) - J(\nu^*))_{k \in \mathbb{N}}$ . In particular, we derive several distinct bounds on the sequence  $(\min_{i \leq k} J(\nu_i) - J(\nu^*))_{k \in \mathbb{N}}$  (assertions i) and ii) of Theorem 2.1) and on the sequence  $(J(\nu_k) - J(\nu^*))_{k \in \mathbb{N}}$  itself (assertion iii) of Theorem 2.1). We also consider the cases where  $\varepsilon$  depends on  $k$  (horizon-dependent convergence) and where it does not (horizon-free convergence). Theorem 2.1 is stated with generic constants. The interested reader can refer to the proofs in Sections C.3 and C.4, where the dependence of these constants with respect to our mathematical framework is made precise.

**Theorem 2.1** *Assume that Assumption  $(\mathcal{H}_{\varepsilon})$  stated in Equation (14) holds, that  $(\alpha, \beta)$  satisfies condition (7) and that  $(\mathcal{H}_{\text{TV}}^{\infty})$  is satisfied. For any final horizon time  $K \geq 2$ , we have:*

i) *If  $(\varepsilon_k)_{k \geq 0}$  is non-adaptive and  $\varepsilon_k = \varepsilon = \sqrt{\frac{\mathfrak{C}}{K}} \leq \alpha, \forall k \in \{1, \dots, K\}$ , then we have:*

$$\forall K \geq 2 \quad \min_{1 \leq k \leq K} \{J(\nu_k) - J(\nu^*)\} \leq \mathfrak{C} \mathfrak{L}^{\frac{2+2d}{2+d}} \alpha^{-\frac{1}{2+d}} K^{-\frac{1}{2(2+d)}}. \quad (22a)$$

ii) *If  $(\varepsilon_k)_{k \geq 0}$  is horizon-free and  $\varepsilon_k = \sqrt{\frac{\mathfrak{C}}{(k+1)}} \leq \alpha$ , then we have:*

$$\min_{1 \leq k \leq K} \{J(\nu_k) - J(\nu^*)\} \leq \mathfrak{C} \mathfrak{L}^{\frac{2+2d}{2+d}} \alpha^{-\frac{1}{2+d}} K^{-\frac{1}{2(2+d)}} \log(K)^{\frac{1}{(2+d)}}. \quad (22b)$$

iii) If  $(\varepsilon_k)_{k \geq 0}$  is  $\varepsilon_k = \varepsilon = \mathfrak{C} \left( \frac{\mathfrak{L}^{2+2d}}{(d+1)\alpha} \right)^{\frac{1}{5+2d}} K^{-\frac{3+d}{5+2d}}$ , then we have

$$J(\nu_k) - J(\nu^*) \leq \mathfrak{C} \left( \frac{\mathfrak{L}^{2+2d}}{(d+1)\alpha} \right)^{\frac{2}{5+2d}} K^{-\frac{1}{5+2d}}.$$

In all three items, the generic constant  $\mathfrak{C}$  may depend polynomially on  $(\|\nu^*\|_{\text{TV}}/\mathfrak{L})$ ; in item iii),  $\mathfrak{C}$  also depends polynomially on the initial excess  $J(\nu_1) - J^*$  (which is finite under  $(\mathcal{H}_{\text{TV}}^\infty)$ ).

Items i) and ii) concern the minimal value of the functional  $J$  along the first  $K$  iterates of the algorithm, differing only in the choice of  $\varepsilon$  at the end. Both results yield similar convergence rates, up to constants and a logarithmic term. In particular, using a horizon-free calibration for  $\varepsilon$  introduces only an additional logarithmic factor in  $k$  in the convergence rate. Item iii) provides a stronger result concerning the value of the last iterate, but the convergence rate obtained is slightly weaker than those in i) and ii). This difference is essentially technical, arising from a proof method based on a compensator/penalty strategy to construct a decreasing sequence—a method that results in a degraded convergence rate.

**Remark 2.2** • *Although the setting considered in this section is specific — focusing on continuous measures and updates — we can still observe that the global convergence rates are slower than those reported in, e.g., Chizat (2022) or De Castro et al. (2025a), which achieve rates of order  $1/\sqrt{k}$ . In our case, the convergence exhibits a dependence on the dimension: the volume of the region where  $J'_{\nu_k} < 0$  at each iteration significantly influences the behavior of the objective  $J$  (see, e.g., Proposition 6 and its proof). This is formalized by a geometric lemma on the volume of the target region, which ensures that  $\lambda(\{J'_{\nu_k} \leq 0\}) \geq C_d \mathfrak{L}^{-d} |\min_x J'_{\nu_k}(x)|^d$ . This bounds the birth process probability from below, preventing it from vanishing arbitrarily fast compared to the loss gap. Nevertheless, unlike the aforementioned works, our method does not impose any local specific constraints on the initialization measure  $\nu_0$ . The algorithm allows for a dynamic evolution of the support of  $\nu_k$  throughout the iterations, which enables global convergence. However, this adjustment of the algorithm to redistribute the mass of the measure in regions where  $J'_{\nu_k}$  is negative affects the convergence speed in a dimension-dependent manner.*

- *In the deterministic optimization literature, it is uncommon to observe convergence rates that depend on the problem's dimension, particularly in convex optimization. The ellipsoid method in convex optimization, introduced by Shor, Yudin, and Nemirovsky (see e.g. Nemirovskij and Yudin (1983)), achieves linear convergence rates, with the rate inversely proportional to the dimension. This degradation also stems from the curse of dimensionality, arising from the geometric optimization strategy—specifically, from the dependence of the ellipsoid's volume on the dimension. Sometimes, particularly in Quasi-Newton and Newton methods, the dependence on the problem dimension is somewhat hidden within the overall computational cost: each iteration requires a number of operations that increases with  $d$ , while the total number of iterations remains essentially independent of  $d$ .*

- When comparing our method with that of Theorem 4.2 of Chizat (2022), we observe that their result is derived under the assumption that the initialization lies within a basin where a Polyak-Lojasiewicz inequality holds — an assumption that is both very strong and restrictive, and which does not hold in full generality. Moreover, Proposition H.1 in Chizat (2022) also exhibits a hidden dependence on the dimension in the setting where the density is uniformly lower bounded for the continuous-time gradient flow — an assumption that cannot be satisfied easily in a discretized counterpart. In particular, the proof of Proposition H.1 in Chizat (2022) is valid only in dimension 1, and the rate also deteriorates as the dimension of the ambient space increases.

**Remark 2.3 (Extension to  $\beta > 0$ )** *The analysis of this section is carried out under the simplification  $\beta = 0$  in order to develop and illustrate the theoretical tools—screening, birth-death dynamics, and descent inequalities—that will be central to the stochastic convergence theory of Section 3. Theorem E.1 in Appendix E extends Theorem 2.1 to  $\beta > 0$  and establishes the same three convergence rates under the learning-rate condition (7). The proof, given there for completeness, follows the same strategy with additional perturbation estimates controlling the effect of the position update  $T_{\nu,\beta}$ .*

### 3 Stochastic algorithm and stochastic convergence properties

The previous section provides the main tools and ideas allowing the conic gradient descent to converge toward a global minimum. The main ingredient consists of adding, at each iteration  $k$ , mass on some specific regions (namely where  $J'_{\nu_k}$  is negative). However, this principle is not feasible in practice since it involves continuous measures. We investigate in this section the implementable Algorithm 1 based on the birth and death process. Similarly to the deterministic analysis, we first provide generic assumptions. Then, we will exhibit some specific updates that will fit our requirements. We finally provide convergence results of the expected risk towards 0, leading to a global optimization result.

#### 3.1 Notation and assumptions on the stochastic update

We consider in this section the generic construction discussed in Section 1.4. Recall that the sequence  $(\hat{\nu}_k)_{k \in \mathbb{N}}$  is built in two steps at each iteration  $k$ : first, a stochastic instance of a CPGD algorithm  $\hat{\nu}_k \rightarrow \hat{\nu}_{k+}$ , associated with a specific descent property; then, an additional update  $\hat{\nu}_{k+} \rightarrow \hat{\nu}_{k+1}$ , which modifies the mass of the current measure at some strategic locations. Both successive updates may involve stochastic computations.

**Assumption ( $\hat{\mathcal{H}}_{\mathfrak{F}}$ ):** There exist two increasing collections of  $\sigma$ -algebras  $(\mathfrak{F}_k)_{k \in \mathbb{N}}$  and  $(\mathfrak{F}_k^+)_{k \in \mathbb{N}}$  such that  $\mathfrak{F}_{k-1} \subset \mathfrak{F}_{k-1}^+ \subset \mathfrak{F}_k \subset \mathfrak{F}_k^+$  for any  $k \in \mathbb{N}$ , and such that  $\hat{\nu}_k$  is  $\mathfrak{F}_k$ -measurable and  $\hat{\nu}_{k+}$  is  $\mathfrak{F}_k^+$ -measurable. Equivalently,  $(\mathfrak{F}_k)_{k \geq 1}$  is adapted to  $(\hat{\nu}_k)_{k \geq 1}$  and  $(\mathfrak{F}_k^+)_{k \geq 1}$  is adapted to  $(\hat{\nu}_{k+})_{k \geq 1}$ .

We impose the following requirements on the updates. Throughout,  $\mathfrak{C} > 0$  (resp.  $\mathfrak{c} > 0$ ) denotes a generic upper-bound (resp. lower-bound) constant that may change from line to line.

- **Assumption** ( $\hat{\mathcal{H}}_{TV}^\infty$ ): There exists a constant  $\mathfrak{C}_{TV}$  such that almost surely:

$$\|\hat{\nu}_k\|_{TV} \leq \mathfrak{C}_{TV} \quad \forall k \in \mathbb{N}.$$

- Iteration  $k \mapsto k^+$ :  $\hat{\nu}_{k^+}$  satisfies the following descent property that is the stochastic counterpart of Proposition 2:

**Assumption** ( $\hat{\mathcal{H}}_D$ ): For any  $k \geq 1$ :

$$\mathbb{E} [J(\hat{\nu}_{k^+}) | \mathfrak{F}_k] - J(\hat{\nu}_k) \leq -\frac{\alpha}{2} \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right),$$

where  $\alpha, \beta$  denote tunable parameters of the algorithm.

- Iteration  $k^+ \mapsto k+1$ : Let  $a > 0$  whose value will be made precise later on. The stochastic update  $\hat{\nu}_{k+1}$  satisfies the following assumptions:

- **Assumption** ( $\hat{\mathcal{H}}_{\varepsilon,a}^+$ ): For any  $k \geq 0$ , for any Borel set  $\mathcal{B} \subseteq \mathcal{X}$ :

$$\mathbb{E} [\hat{\nu}_{k+1}(\mathcal{B}) | \mathfrak{F}_k^+] \geq \mathfrak{c} \varepsilon_k \lambda(\mathcal{B} \cap \{J'_{\hat{\nu}_{k^+}} < 0\}) - \mathfrak{C} \varepsilon_k m_k^{-a}$$

where  $\lambda$  stands for the Lebesgue measure.

- **Assumption** ( $\hat{\mathcal{H}}_\varepsilon^{\text{smooth},1}$ ): For any  $k \geq 1$ :

$$\mathbb{E} [J(\hat{\nu}_{k+1}) - J(\hat{\nu}_{k^+}) | \mathfrak{F}_k^+] \leq \mathfrak{C} \left( \varepsilon_k^2 + \varepsilon_k \sqrt{\frac{\log m_k}{m_k}} \right) \quad (23a-(\hat{\mathcal{H}}_\varepsilon^{\text{smooth},1}))$$

- **Assumption** ( $\hat{\mathcal{H}}_\varepsilon^{\text{smooth},2}$ ): For any  $k \geq 1$ :

$$\|J'_{\hat{\nu}_{k+1}} - J'_{\hat{\nu}_{k^+}}\|_\infty \leq \mathfrak{C} \varepsilon_k \quad \text{a.s.} \quad (23b-(\hat{\mathcal{H}}_\varepsilon^{\text{smooth},2}))$$

The next section provides an example of an implementable stochastic update that satisfies all these assumptions.

## 3.2 Fast Spawn&Prune: a conic particle birth/death process

Although the setting considered in Section 3 is quite general, it encompasses Algorithm 1. The setting of Section 3 may be seen as a minimal sufficient set of conditions for the global convergence of the stochastic dynamic. We provide here specific instances of the mass tweaking step introduced in (12). Then, we prove that the resulting algorithm satisfies all the requirements introduced in Section 3.1.

### 3.2.1 TRANSITION $\hat{\nu}_{k^+} \rightarrow \hat{\nu}_{k+1}$

Inspired by Section 2.2 for the deterministic side, we now describe our sequence of stochastic updates  $\hat{\nu}_k \rightarrow \hat{\nu}_{k^+} \rightarrow \hat{\nu}_{k+1}$ , whose evolution still involves both deletion  $\hat{\nu}_{k^+} \rightarrow \hat{\nu}_k^{++}$  and creation  $\hat{\nu}_k^{++} \rightarrow \hat{\nu}_{k+1}$  of weighted particles. Again, mass deletion should only concern areas where  $J'_{\hat{\nu}_k} \geq 0$  whereas our algorithm adds some mass in areas where  $J'_{\hat{\nu}_k} \leq 0$ . Since

handling these objects might be time-consuming at each iteration, we deal instead with the sets of positivity and negativity of  $J'_{\hat{\nu}_k}$ , *i.e.* we define  $\widehat{\mathcal{P}}_{\nu_{k+}}$  and  $\widehat{\mathcal{N}}_{\nu_{k+}}$  as:

$$\begin{aligned}\widehat{\mathcal{P}}_{\nu_{k+}} &:= \left\{ t_j^{k+} : \widehat{J'_{\hat{\nu}_{k+}}}(t_j^{k+}) \geq 0 \text{ and } \omega_j^{k+} \leq \sqrt{2\varepsilon_k} \right\} \\ \widehat{\mathcal{N}}_{\nu_{k+}} &:= \left\{ t \in \mathcal{X} : \widehat{J'_{\hat{\nu}_{k+}}}(t) \leq c_a \sqrt{\frac{\log m_k}{m_k}} \right\},\end{aligned}\tag{24}$$

where  $c_a$  is a positive constant depending on  $a > 0$  involved in  $(\widehat{\mathcal{H}}_{\varepsilon,a}^+)$ .

**Mass deletion**  $\hat{\nu}_{k+} \longrightarrow \hat{\nu}_k^{++}$  We build  $\hat{\nu}_k^{++}$  from  $\hat{\nu}_{k+}$  while removing some mass on  $\widehat{\mathcal{P}}_{\nu_{k+}}$  and define:

$$\hat{\nu}_k^{++} := \hat{\nu}_{k+} (1 - \mathbf{1}_{\widehat{\mathcal{P}}_{\nu_{k+}}}(V_{k+1}) \delta_{V_{k+1}}) \quad \text{with } V_{k+1} \sim \mathcal{U}_{\text{supp}(\hat{\nu}_{k+})}\tag{25a}$$

**Mass creation**  $\hat{\nu}_k^{++} \longrightarrow \hat{\nu}_{k+1}$  For any  $k \in \mathbb{N}$ , given  $\hat{\nu}_{k+}$ , define

$$\hat{\nu}_{k+1} = \hat{\nu}_k^{++} + \varepsilon_k \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}}(U_{k+1}) \delta_{U_{k+1}} \quad \text{with } U_{k+1} \sim \mathcal{U}_{\mathcal{X}}.\tag{25b}$$

Contrary to the deterministic update described in the previous section, the scheme (25b) is implementable. Indeed, it only requires the generation of a pair of uniform random variables over the spaces  $\text{supp}(\hat{\nu}_{k+})$  and  $\mathcal{X}$  and pointwise evaluations of  $\widehat{J'_{\hat{\nu}_{k+}}}$ .

**Remark 3.1** We stress that the constant  $c_a$  involved in (24) is positive. Hence, the set  $\widehat{\mathcal{N}}_{\nu_{k+}}$  is not exactly defined as the negative part of  $\widehat{J'_{\hat{\nu}_k}}$ . Recall that, according to the KKT conditions (see Proposition 1) our initial target is the negative part of  $J'_{\hat{\nu}_k}$ . Since we use a stochastic approximation of  $J'_{\hat{\nu}_k}$ , we have to ensure a complete exploration of this latter set. In this context, choosing  $c_a > 0$  allows to control some kind of Type II error.

Thanks to the previous definitions, we can now state that the sequence of measures  $(\hat{\nu}_k, \hat{\nu}_{k+})_{k \geq 1}$  satisfies our previous assumptions. The proof of the next proposition is deferred to Section D.

**Proposition 4** Define

$$\widehat{\mathfrak{R}} = \frac{\mathbf{H}}{\mathbf{G}} e + \sqrt{\frac{e^3}{\mathbf{G}}} + 1.$$

Then provided  $\max_k \varepsilon_k \leq \alpha$  and  $\alpha \leq (1 + \widehat{\mathfrak{R}})^{-1}$ , then  $\|\hat{\nu}_k\|_{TV} \leq \widehat{\mathfrak{R}} \wedge \|\nu_0\|_{TV} := \mathfrak{C}_{TV}$  for all  $k \geq 1$ . Assume moreover that  $(\alpha, \beta)$  satisfies (7) and

$$\alpha \leq \frac{1}{\mathbf{E}_\infty} \sqrt{8 \log 8},\tag{26a}$$

and that the mini-batch schedule satisfies  $p_k \leq p_0 + m_k$  (equivalently  $k \leq m_k$ , which holds in particular for  $m_k = K$  and  $m_k = k$ ). Then, for any  $a > 0$ , picking the threshold  $c_a = \mathbf{E}_\infty \sqrt{2a}$  in (24) ensures that the sequence  $(\hat{\nu}_k, \hat{\nu}_{k+})_{k \geq 1}$  of Algorithm 1 enjoys  $(\widehat{\mathcal{H}}_D)$ ,  $(\widehat{\mathcal{H}}_{\varepsilon,a}^+)$ ,  $(\widehat{\mathcal{H}}_\varepsilon^{\text{smooth},1})$ ,  $(\widehat{\mathcal{H}}_\varepsilon^{\text{smooth},2})$  and  $(\widehat{\mathcal{H}}_{TV}^\infty)$ .

### 3.3 Global convergence results

We have now all the ingredients to establish convergence rates for Algorithm 1. The following theorem provides a control on the minimal value of the sub-optimality value sequence  $(J(\hat{\nu}_k) - J(\nu^*))_{k \in \{1, \dots, K\}}$  in a finite horizon scenario.

**Theorem 3.1** *Assume that the stochastic sequences  $(\hat{\nu}_k, \hat{\nu}_{k+})_{k \geq 1}$  satisfy  $\hat{\mathcal{H}}_D$ ,  $(\hat{\mathcal{H}}_{\varepsilon, a}^+)$  with  $a \geq \frac{d}{2(2+d)}$ ,  $\hat{\mathcal{H}}_\varepsilon^{\text{smooth}, 1}$ ,  $\hat{\mathcal{H}}_\varepsilon^{\text{smooth}, 2}$  and  $(\hat{\mathcal{H}}_{\text{TV}}^\infty)$ . For any final horizon iterate  $K$ , we define  $\hat{\rho}_K$  as the lowest value of the excess loss along the  $K$  iterations:*

$$\hat{\rho}_K := \min_{1 \leq k \leq K} \{J(\hat{\nu}_k) - J(\nu^*)\}.$$

If we choose:  $\alpha \leq \frac{1}{\mathbf{E}_\infty} \sqrt{8 \log 8}$ ,  $\beta \leq \frac{1}{\alpha^{d/4} \sqrt{K}}$ ,  $\varepsilon_k = \frac{1}{\sqrt{K}}$  and  $m_k = K$ , then the sequence  $(\hat{\nu}_k, \hat{\nu}_{k+})_{k \geq 1}$  verifies the global convergence rate:

$$\mathbb{E}[\hat{\rho}_K] \leq \mathfrak{c} \alpha^{-1/2} \left( \frac{\log K}{K} \right)^{\frac{1}{2(2+d)}}. \quad (27)$$

The schedule above satisfies  $\varepsilon_k \leq \alpha$  together with (7) (instantiated with  $\hat{\mathfrak{R}}$  as defined in Proposition 4), so the hypotheses of Proposition 4 are met. Note that this requires  $K \geq 1/\alpha^2$ .

**Corollary 3.1 (Sample complexity)** *Under the assumptions of Theorem 3.1, the total number of stochastic oracle calls over  $K$  iterations is  $N = K \times m = K^2$ . Expressed in terms of  $N$ , the convergence rate reads:*

$$\mathbb{E}[\hat{\rho}_K] \leq \mathfrak{c} \alpha^{-1/2} \left( \frac{\log N}{\sqrt{N}} \right)^{\frac{1}{2(2+d)}}.$$

In particular,  $\mathbb{E}[\hat{\rho}_K] = \mathcal{O} \left( N^{-\frac{1}{4(2+d)}} (\log N)^{\frac{1}{2(2+d)}} \right)$ .

**Proof** Since  $m = K$ , the total number of oracle evaluations is  $N = K^2$ , hence  $K = \sqrt{N}$ . Substituting into (27):

$$\mathbb{E}[\hat{\rho}_K] \leq \mathfrak{c} \alpha^{-1/2} \left( \frac{\log \sqrt{N}}{\sqrt{N}} \right)^{\frac{1}{2(2+d)}} \leq \mathfrak{c} \alpha^{-1/2} \left( \frac{\log N}{\sqrt{N}} \right)^{\frac{1}{2(2+d)}},$$

and the result is proven. ■

The horizon-dependent tuning of Theorem 3.1 requires the knowledge of  $K$  in advance to set  $\beta$  and  $m$ . The following result removes this requirement by using iteration-dependent schedules.

**Theorem 3.2 (Horizon-free variant)** *Under the same assumptions as Theorem 3.1, choose the iteration-dependent schedules:*

$$m_k = k \vee 1, \quad \varepsilon_k = \min \left( \alpha, \frac{1}{\sqrt{k \vee 1}} \right), \quad \beta_k = \frac{1}{k \vee 1},$$

with  $\alpha$  a fixed constant satisfying (7) (independent of  $K$ ). Then, for any horizon  $K \geq 4/\alpha^2$ ,

$$\mathbb{E}[\hat{\rho}_K] \leq \mathfrak{C}\alpha^{-1/2} \left( \frac{(\log K)^3}{K} \right)^{\frac{1}{2(2+d)}}. \quad (28)$$

In particular, no prior knowledge of  $K$  is needed: the algorithm is “any-time”. The total number of oracle calls is  $N = \sum_{k=1}^K k = \frac{K(K+1)}{2}$ , yielding the sample complexity

$$\mathbb{E}[\hat{\rho}_K] = \mathcal{O} \left( N^{-\frac{1}{4(2+d)}} (\log N)^{\frac{3}{2(2+d)}} \right).$$

The proofs of Theorems 3.1 and 3.2 are displayed in Section D.4. They follow essentially the same lines as the deterministic scheme discussed in Section 2. However, the stochastic approximation of the Fréchet derivative of  $J$  along the trajectory of the algorithm requires careful control. We stress that the final convergence rate does not significantly differ from (22a). The main difference with the latter is a logarithmic term in  $K$ , which is a consequence of the stochastic approximation. Furthermore, note that the cap  $\varepsilon_k \leq \alpha$  ensures the hypothesis  $\varepsilon_k \leq \alpha$  of Proposition 4 for every  $k \geq 0$ , and the monotonicity  $\beta_k \leq \beta_0$  guarantees that (7) (instantiated with  $\hat{\mathfrak{R}}$ ) holds at every step.

## 4 Numerical experiments

In this section, we illustrate the capabilities of the *FS $\mathcal{E}$ P* algorithm on two distinct tasks: density estimation using Gaussian Mixture Models (GMM) on synthetic data, and a regression task using two-layer neural networks (NN) on the California Housing dataset. In both cases, we compare the standard Full-Batch and Stochastic Conic Particle Gradient Descent (CPGD) against our proposed versions augmented with the Birth and Death (BD) processes. A companion Jupyter notebook reproducing every experiment, figure and table of this section — with explicit cross-references to the paper — is available at <https://github.com/ydecastro/Fast-Spawn-Prune> (see `section4_companion.ipynb`).

### 4.1 Gaussian Mixture Models (GMM) with Fixed Covariance

In this first experiment, we illustrate the ability of our dynamic particle system to *escape local minima thanks to the birth process*. We consider a simple density estimation problem where we observe  $n$  i.i.d. points  $X_1, \dots, X_n \in \mathbb{R}^2$  from a 2D Gaussian mixture ( $n = 24,000$ ). The goal is to recover the means and weights of this mixture by minimizing a regularized  $L^2$  fitting criterion over the space of discrete positive measures as in De Castro et al. (2021). The birth process acts as an *exploration mechanism*. Following our theoretical birth criterion, the algorithm evaluates the Fréchet derivative of the objective (or the dual certificate)  $J'_\nu(t)$  at proposed locations. When the first-order optimality condition is violated—that is, when  $J'_\nu(t)$  is negative in some region of the state space—new particles are spawned in these promising areas. Empirically, the continuous infusion of mass in regions of negative gradient allows the interacting particle system to efficiently escape poor local configurations and quickly recover all the true modes of the synthetic distribution (see Figure 4).

**Loss function and Kernel** The goal of the optimization is to minimize the regularized  $L^2$  distance between the smoothed empirical measure  $\hat{f}_\tau^n = \frac{1}{n} \sum_{i=1}^n \phi_\tau(\cdot - X_i)$  and the model. For a discrete positive measure  $\nu = \sum_j \omega_j \delta_{t_j}$ , the objective is given by:

$$J(\nu) = \frac{1}{2} \|\phi_\tau * \nu - \hat{f}_\tau^n\|_{L^2(\mathbb{R}^2)}^2 + \kappa \|\nu\|_{\text{TV}}, \quad (29)$$

where  $\phi_\tau = \mathcal{N}(0, \tau^2 I_2)$  is a Gaussian smoothing kernel, see De Castro et al. (2021) for further details on smoothing applied to GMM for BLASSO. The associated feature map is  $\varphi_t(x) = \mathcal{N}(x; t, (1 + \tau^2)I_2)$ , which yields the smooth Gram matrix  $K(t_i, t_j) = \mathcal{N}(t_i; t_j, 2(1 + \tau^2)I_2)$ .

**Verification of the assumptions** Because the Gaussian kernel  $K(t_i, t_j)$  is bounded, strictly positive, and twice continuously differentiable, this formulation perfectly satisfies the mathematical assumptions of the paper. Most notably, it respects Assumption ( $\mathcal{H}_P$ ) which requires  $\mathcal{C}^2$  spatial smoothness for the theoretical descent lemma.

### 4.2 Training two-layer neural networks on the California Housing dataset

Next, we turn to a real-world dataset to assess the impact of the death process in typical machine learning applications. We consider the regression task on the California housing dataset using a two-layer neural network with ReLU activations. As the ReLU function is

positively homogeneous, we can equivalently rewrite the network by normalizing the locations (the incoming weights of the neurons) to live on the unit Euclidean sphere  $\mathcal{X} = \mathbb{S}^{d-1}$ . The optimization is performed over this connected space by minimizing the regularized empirical risk with respect to the weights  $\mathbf{W}$  and locations  $\mathbf{T}$ . We use a stochastic gradient descent (SGD) scheme to update the weights and locations over a sequence of mini-batches. In this context, while the birth process can introduce new neurons to increase the capacity of the network, the death process ensures that the size of the network remains controlled. In practice, the death rule removes particles whose weight falls below a given threshold or whose contribution becomes negligible. Our numerical experiments demonstrate that the death process harms neither the convergence nor the quality of the final predictive results. Indeed, *pruning the inactive neurons significantly reduces the computational burden and prevents over-parameterization without deteriorating the generalization error*. This simple experiment perfectly illustrates the theoretical insights of the paper, indicating that the dynamic interactions and systematic particle suppressions do not prevent the system from reaching an optimal configuration.

**Loss function and Kernel** We apply the algorithm to a regression task on the California Housing dataset ( $n = 18,576$  samples,  $d = 8$  features). The objective is to minimize the regularized Mean Squared Error (MSE):

$$J(\nu) = \frac{1}{2} \|y - \Phi(\nu)\|_{\mathbb{H}}^2 + \kappa \|\nu\|_{\text{TV}}, \quad (30)$$

with the prediction function being a two-layer ReLU network  $f_{\nu}(x) = \Phi(\nu)(x)$ . Here, the feature map is given by affine ReLU neurons  $\varphi_t(x) = \text{ReLU}(\langle v, x \rangle + b)$ , where the parameters  $t = (v, b)$  are constrained to the unit ball  $\bar{B}(0, 1) \subset \mathbb{R}^{d+1}$ . Concretely, the incoming weight vector  $v$  is normalized to lie on the unit sphere  $\mathbb{S}^{d-1} \subset \mathbb{R}^d$  (as described above) and the bias  $b \in [-1, 1]$ ; the pair  $(v, b)$  therefore belongs to  $\mathbb{S}^{d-1} \times [-1, 1] \subset \bar{B}(0, 1) \subset \mathbb{R}^{d+1}$ , reconciling the two descriptions. The empirical kernel relies on the inner product over the  $n$  observations:  $K(t_i, t_j) = \frac{1}{n} \sum_{k=1}^n \text{ReLU}(\langle v_i, x_k \rangle + b_i) \text{ReLU}(\langle v_j, x_k \rangle + b_j)$ .

**Verification of the assumptions** Unlike the GMM experiment, the ReLU activation function is continuous but not everywhere differentiable. Strictly speaking, it does not fully satisfy the bounded  $\mathcal{C}^2$  spatial smoothness assumption ( $\mathcal{H}_{\mathcal{P}}$ ) established in the theoretical sections. In practice, the gradients are computed using the ReLU subgradient, and the empirical method remains highly effective and robust despite this theoretical mismatch.

### 4.3 Experimental Results and Discussion

In this section, we present the empirical results of our experiments, highlighting the benefits of integrating the Birth and Death (BD) process into both Full-Batch and Stochastic Conic Particle Gradient Descent (CPGD).

**Birth and Death process** Particles are removed from the support if their contribution becomes negligible or falls into regions where the dual certificate is strongly positive. Specifically, a particle at  $t_i$  with weight  $\omega_i$  is pruned if  $J'_{\nu}(t_i)/\omega_i > \tau_{\text{death}}$ . Particles are added by evaluating  $J'_{\nu}(t)$  on a set of randomly sampled candidate positions to find regions  $\mathcal{N}_{\nu}$ ,

where the first-order optimality condition is violated. For the stochastic setting, the targeted birth level is set to spawn particles where  $J'_\nu(t) < \tau_{\text{birth}} \sqrt{\log(m_k)/m_k}$ , where  $m_k$  is the mini-batch size.

### 4.3.1 DYNAMICS OF THE BIRTH AND DEATH PROCESS

To understand how the network capacity adapts during training, we plot the number of active particles over the optimization iterations in Figure 2.

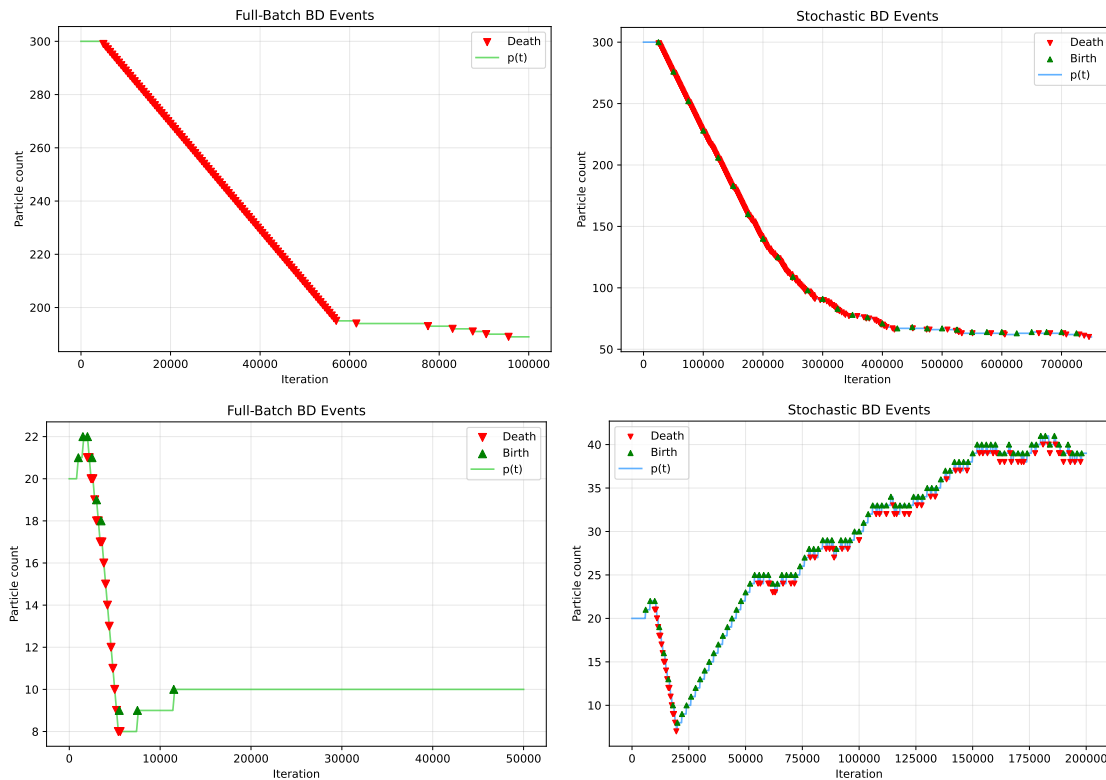


Figure 2: Particle count over iterations for GMM (bottom) and Neural Network (top) regression tasks.

The dynamic capacity of the model is visible through the sharp drops (Death events pruning inactive particles) and spikes (Birth events) in Figure 2. In the Neural Network example (top row) we see that an over-parametrized layer is reduced by 50–70% by the death process. In the bottom row (GMM toy example), the right panel shows a pruning of spurious particles and an exploration phase (as in the left panel) which ends with a stabilization around  $p = 39$  (and  $p = 10$  in the left panel).

### 4.3.2 CONVERGENCE AND GENERALIZATION

Figure 3 illustrates the temporal evolution of the optimization metrics. We compare the standard CPGD methods against our proposed BD-augmented variants in terms of Wall-Clock time.

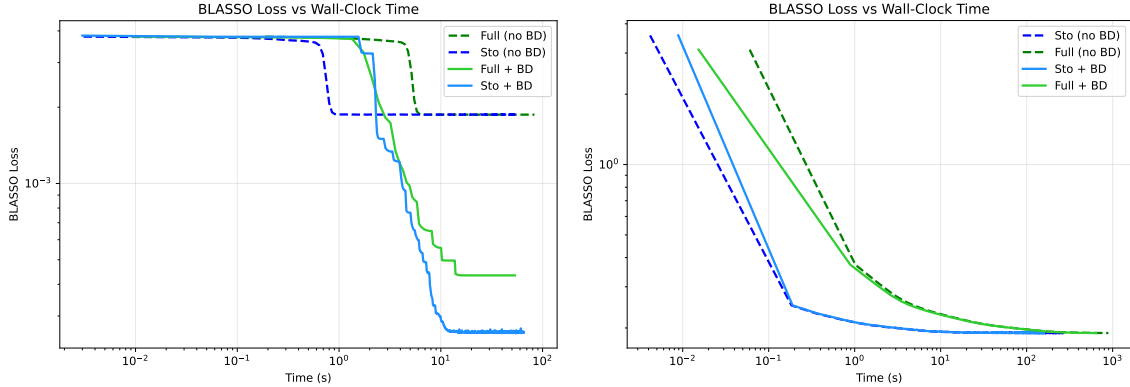


Figure 3: BLASSO loss evaluated on the GMM (left) and Neural Network (right) regression tasks.

As observed in Figure 3 (left panel–GMM), the standard algorithms quickly plateau into local minima. The BD mechanisms, however, periodically inject new particles in regions where the dual certificate is highly negative, allowing the loss to drop further. Also, the convergence is not altered by the death process (right panel–NN): the solution gets sparser and sparser (with fewer and fewer neurons) and achieves the same performance as the large neural network with 300 hidden neurons.

### 4.3.3 SPATIAL DISTRIBUTION OF PARTICLES IN GMM

Finally, we visualize the end-state positions of the particles for the 2D Gaussian Mixture Model experiment in Figure 4. The baseline models (top row) suffer from the presence of spurious particles that fail to align with the true means, keeping the meaningful particle count significantly below  $p = 20$  (the number of initial particles).

In contrast, the Full-Batch model equipped with the BD process (bottom-left) recovers the true means and cleanly prunes all unnecessary components, albeit yielding an under-representation with  $p = 10$  particles. The stochastic counterpart (bottom-right) successfully maps out the target distribution and maintains a matching set of particles (all targets are identified except the smallest one, which has a weight of 0.0005).

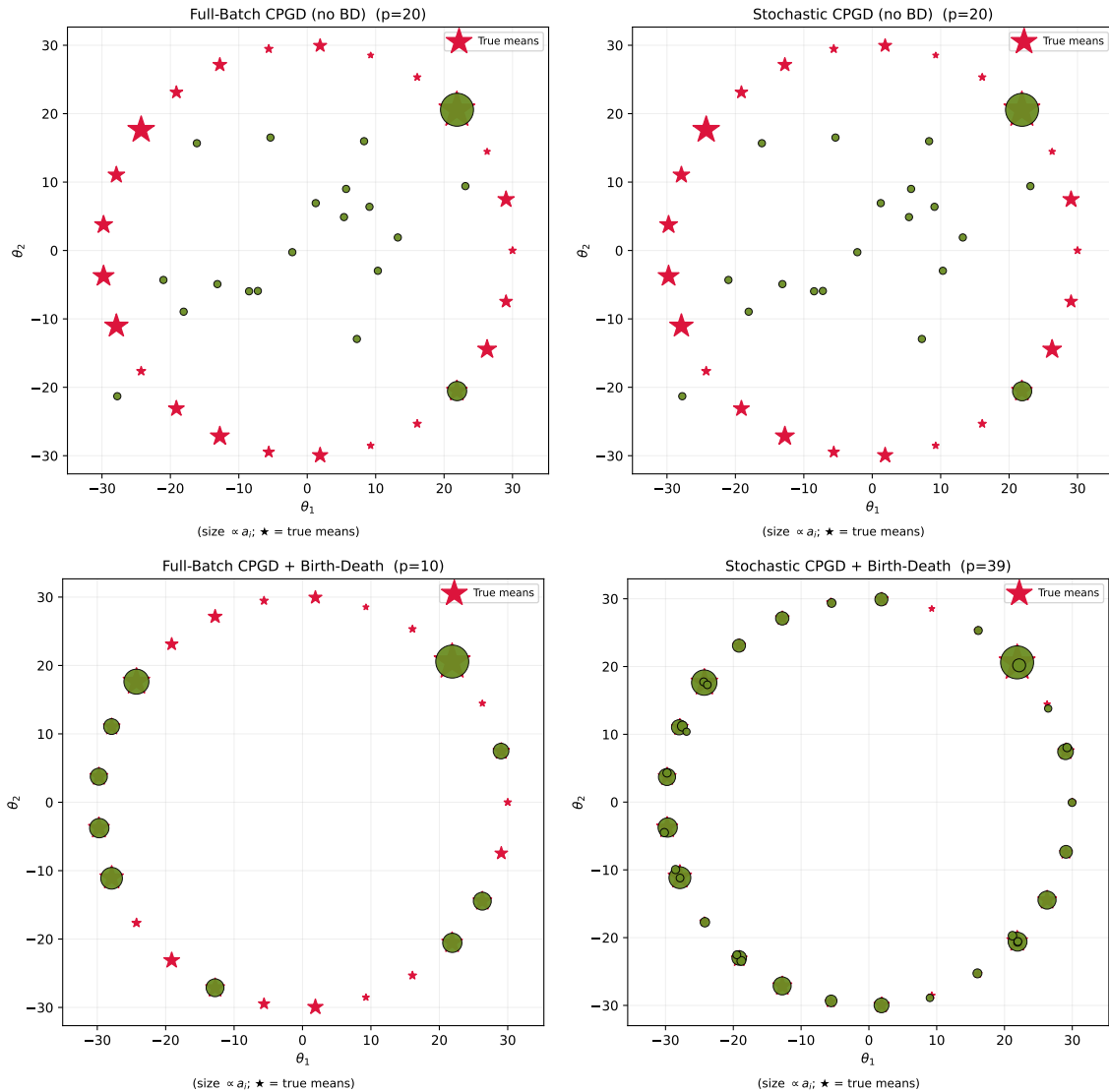


Figure 4: Final positions of the particles  $t = (t_1, t_2)$  relative to the true cluster means (represented by cross marks) in the GMM experiment. Top row: baseline CPGD without BD. Bottom row: CPGD with BD. The size and color of the markers correspond to the weights  $\omega_i$  of the particles. A common random seed has been fixed for shared initialization and batch sampling.

Table 1: Summary of shared hyperparameters across the two main experiments.

Parameter	GMM (Fixed Covariance)	Neural Net. (California)
Train Samples ( $n$ )	24,000	18,576
Input Dimension ( $d$ )	2	8
Initial Particles ( $p$ )	20	300
TV Regularization ( $\kappa$ )	0.0001	0.0005
Full-Batch Size	24,000	18,576
Full-Batch Iterations	50,000	100,000
Stochastic Batch Size	256	256
Stochastic Iterations	200,000	750,000
Death Threshold ( $\tau_{\text{death}}$ )	5.0	5.0

#### 4.3.4 PERFORMANCE ANALYSIS AND EXPERIMENTAL PARAMETERS

To formalize the empirical observations, we report the exact hyper-parameters used for both the GMM and NN regression tasks in Table 1, and the quantitative final results in Tables 2 and 3.

Table 2: Final results for the 2D GMM Experiment. The BD variations significantly lower the final loss and yield a better TV norm (true is 1) while maintaining or reducing computation time.

Method	Loss	TV	$p_{\text{final}}$	Time (s)	Deaths	Births
Full-Batch	0.001869	0.2760	20	83.67	0	0
Stochastic	0.001872	0.2863	20	55.02	0	0
Full-Batch + BD	0.000434	0.7788	10	61.88	19	9
Stochastic + BD	<b>0.000259</b>	<b>0.9786</b>	39	65.23	78	97

In the GMM experiment, the dataset consists of  $n = 24,000$  training samples generated from  $K = 25$  true components. The optimization regularizes the total variation (TV) norm with  $\kappa = 0.0001$ , starting from an initial set of  $p = 20$  particles. As shown in Table 2, integrating the Birth-Death (BD) process yields a drastic improvement in the final BLASSO loss. For example, the Full-Batch variant sees its loss drop from 0.001869 to 0.000434 while trimming the final number of particles down to  $p = 10$ . The Stochastic+BD method provides the lowest overall loss (0.000259) and proxy test error, managing a highly dynamic capacity (78 deaths and 97 births) to settle at  $p = 39$  particles.

Table 3: Final results for the Two-Layer Neural Network regression on the California Housing dataset. The BD process yields highly parsimonious networks and faster training times without sacrificing Test MSE.

Method	MSE	MSE (test)	$p_{\text{final}}$	Time (s)	Deaths	Births
Full-Batch	0.365476	0.393433	300	892.2	0	0
Stochastic	0.366231	0.393154	300	265.1	0	0
Full-Batch + BD	<b>0.365506</b>	0.393495	189	644.0	111	0
Stochastic + BD	0.366369	<b>0.392429</b>	<b>60</b>	<b>158.9</b>	269	29

For the California Housing NN regression, the model starts over-parameterized with  $p = 300$  initial particles (neurons) to learn from  $n = 18,576$  training samples with  $d = 8$  input features. The regularization is set to  $\kappa = 0.0005$ . Table 3 highlights that the Stochastic+BD algorithm prunes the network down to just  $p = 60$  neurons (via 269 death events and 29 births). This significant reduction in parameters translates to a 40% decrease in wall-clock training time (from 265.1s to 158.9s) compared to the standard Stochastic CPGD, all while achieving a slightly better test MSE of 0.392429. The Full-Batch+BD method similarly reduces both the particle count ( $p = 189$ ) and the training time (644.0s compared to 892.2s). In both experimental settings, the death threshold was consistently set to  $\tau_{\text{death}} = 5.0$ .

## References

- J.-M. Azais, Y. De Castro, and F. Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.
- C. Boyer, A. Chambolle, Y. D. Castro, V. Duval, F. De Gournay, and P. Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019.
- K. Bredies and H. K. Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(01):190–218, 2013.
- É. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.
- L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1-2):487–532, 2022.
- L. Chizat and F. R. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3040–3050, 2018.
- Y. De Castro, S. Gadat, C. Marteau, and C. Maugis-Rabusseau. Supermix: sparse regularization for mixtures. *The Annals of Statistics*, 49(3):1779–1809, 2021.
- Y. De Castro, R. Gribonval, and N. Jouvin. Four new results in continuous sparse regularization, with application to sketch mixtures. Technical report, Institut Camille Jordan, 2024.
- Y. De Castro, S. Gadat, and C. Marteau. Fastpart: Over-parameterized stochastic gradient descent for sparse optimisation on measures. *ArXiv preprint*, abs/2312.05993, 2025a.
- Y. De Castro, R. Gribonval, and N. Jouvin. Effective regions and kernels in continuous sparse regularisation, with application to sketched mixtures. *ArXiv preprint*, abs/2507.08444, 2025b.
- V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.
- S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1–2):267–305, 2016.
- R. Giard, Y. De Castro, and C. Marteau. Gaussian mixture model with unknown diagonal covariances via continuous sparse regularization. *arXiv preprint*, 2025.
- T. Hytönen, J. Van Neerven, M. Veraar, and L. Weis. *Analysis in Banach spaces*, volume 12. Springer, 2016.

- A. S. Nemirovskij and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.
- C. Poon, N. Keriven, and G. Peyré. The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*, 23(1):241–327, 2023.
- I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

## Appendix A. Technical Lemmas

### A.1 Existence of the Feature Map

**Lemma A.1** *Let  $\Phi : \mathcal{M}(\mathcal{X}) \rightarrow \mathbb{H}$  be a bounded linear and weak-\* continuous operator, then its dual operator  $\Phi^* : \mathbb{H} \rightarrow \mathcal{C}(\mathcal{X})$  reads*

$$\Phi^* : h \in \mathbb{H} \mapsto (t \in \mathcal{X} \mapsto \langle \varphi_t, h \rangle_{\mathbb{H}}) \in \mathcal{C}(\mathcal{X}),$$

where we identified the pre-dual space  $\mathcal{C}(\mathcal{X})$  as a subspace of the dual  $\mathcal{M}(\mathcal{X})^*$ , and where  $\varphi_t := \Phi \delta_t$  with  $\delta_t$  the Dirac measure at  $t \in \mathcal{X}$ .

Moreover, for every  $\mu \in \mathcal{M}(\mathcal{X})$ , one has the Bochner integral representation in  $\mathbb{H}$ :

$$\Phi \mu = \int_{\mathcal{X}} \varphi_t \, d\mu(t).$$

**Proof** Fix  $h \in \mathbb{H}$  and consider the linear functional  $L_h : \mathcal{M}(\mathcal{X}) \rightarrow \mathbb{R}$  defined by  $L_h(\mu) := \langle h, \Phi \mu \rangle_{\mathbb{H}}$ . Since  $\Phi$  is weak-\* continuous and  $h \mapsto \langle h, \cdot \rangle_{\mathbb{H}}$  is continuous on  $\mathbb{H}$ , the map  $L_h$  is linear and weak-\* continuous on  $\mathcal{M}(\mathcal{X})$ . By definition of the weak-\* topology on a dual space  $E^*$  (here  $E^* = \mathcal{M}(\mathcal{X})$ ), the continuous linear functionals on  $(E^*, \text{weak-}^*)$  are precisely the evaluations by elements of the pre-dual space  $E$  (here  $E = \mathcal{C}(\mathcal{X})$ ). Formally,  $(E^*, \sigma(E^*, E))^* \cong E$ . Hence, there exists a unique  $f_h \in \mathcal{C}(\mathcal{X})$  such that

$$\langle h, \Phi \mu \rangle_{\mathbb{H}} = \langle f_h, \mu \rangle_{\mathcal{C}(\mathcal{X}), \mathcal{M}(\mathcal{X})} \quad \text{for all } \mu \in \mathcal{M}(\mathcal{X}).$$

Define  $\Phi^* h := f_h \in \mathcal{C}(\mathcal{X})$ . Then  $\Phi^* : \mathbb{H} \rightarrow \mathcal{C}(\mathcal{X})$  is linear and bounded, with

$$\|\Phi^* h\|_{\infty} = \sup_{\|\mu\|_{\text{TV}} \leq 1} |\langle \Phi^* h, \mu \rangle| = \sup_{\|\mu\|_{\text{TV}} \leq 1} |\langle h, \Phi \mu \rangle_{\mathbb{H}}| \leq \|h\|_{\mathbb{H}} \|\Phi\|.$$

For  $t \in \mathcal{X}$ , let  $\delta_t$  be the Dirac measure at  $t$  and set  $\varphi_t := \Phi \delta_t \in \mathbb{H}$ . Evaluating the identity at  $\mu = \delta_t$  yields

$$(\Phi^* h)(t) = \langle \Phi^* h, \delta_t \rangle = \langle h, \Phi \delta_t \rangle_{\mathbb{H}} = \langle \varphi_t, h \rangle_{\mathbb{H}}.$$

Hence  $\Phi^* h$  is precisely the continuous function  $t \mapsto \langle \varphi_t, h \rangle_{\mathbb{H}}$ , as claimed.

We now prove the integral representation. First, note that

$$\|\varphi_t\|_{\mathbb{H}} = \|\Phi \delta_t\|_{\mathbb{H}} \leq \|\Phi\| \|\delta_t\|_{\text{TV}} = \|\Phi\|$$

for all  $t \in \mathcal{X}$ , so  $t \mapsto \varphi_t$  is bounded. Next, for each fixed  $h \in \mathbb{H}$ , we already identified  $\Phi^* h \in \mathcal{C}(\mathcal{X})$  and established  $(\Phi^* h)(t) = \langle \varphi_t, h \rangle_{\mathbb{H}}$ . Since  $\Phi^* h \in \mathcal{C}(\mathcal{X})$ , the scalar map  $t \mapsto (\Phi^* h)(t) = \langle \varphi_t, h \rangle_{\mathbb{H}}$  is continuous on  $\mathcal{X}$ . Hence  $t \mapsto \varphi_t$  is weakly continuous (i.e., all scalar evaluations  $\langle \varphi_t, h \rangle_{\mathbb{H}}$  are continuous in  $t$ ). Because  $\mathbb{H}$  is separable, weak measurability/continuity implies strong (Bochner) measurability by Pettis' theorem, see (Hytönen et al., 2016, Theorem 1.1.6). Consequently, for any finite signed measure  $\mu \in \mathcal{M}(\mathcal{X})$ , the Bochner integral  $\int_{\mathcal{X}} \varphi_t \, d\mu(t)$  is well-defined in  $\mathbb{H}$  and satisfies, for all  $h \in \mathbb{H}$ ,

$$\left\langle h, \int_{\mathcal{X}} \varphi_t \, d\mu(t) \right\rangle_{\mathbb{H}} = \int_{\mathcal{X}} \langle h, \varphi_t \rangle_{\mathbb{H}} \, d\mu(t). \quad (31)$$

Using the identity  $\langle h, \Phi\mu \rangle_{\mathbb{H}} = \langle \Phi^*h, \mu \rangle = \int_{\mathcal{X}} (\Phi^*h)(t) d\mu(t) = \int_{\mathcal{X}} \langle h, \varphi_t \rangle_{\mathbb{H}} d\mu(t)$ , we deduce that

$$\langle h, \Phi\mu \rangle_{\mathbb{H}} = \left\langle h, \int_{\mathcal{X}} \varphi_t d\mu(t) \right\rangle_{\mathbb{H}}$$

for all  $h \in \mathbb{H}$ . By uniqueness of the Riesz representation in  $\mathbb{H}$ , this implies

$$\Phi\mu = \int_{\mathcal{X}} \varphi_t d\mu(t),$$

which is the desired representation. ■

## A.2 Lipschitz Continuity of the Feature Map

**Lemma A.2** *Assume that the kernel  $K$  satisfies Assumption  $(\mathcal{H}_{\mathcal{P}})$ , specifically that  $K(t, t) = 1$  for all  $t \in \mathcal{X}$  and that its second derivatives are bounded by  $\mathfrak{C}_{\mathcal{P}}$ . Then, the kernel metric  $d_K$  satisfies the following Lipschitz inequality:*

$$d_K(s, t) := \|\varphi_t - \varphi_s\|_{\mathbb{H}} \leq \sqrt{\mathfrak{C}_{\mathcal{P}}} \|t - s\|, \quad (32)$$

where  $\|\cdot\|_{\mathbb{H}}$  denotes the norm in the Hilbert space  $\mathbb{H}$  and  $\|\cdot\|$  is the Euclidean norm.

**Proof** By the definition of the kernel metric  $d_K$  and the relation  $K(s, t) = \langle \varphi_s, \varphi_t \rangle_{\mathbb{H}}$ , we have:

$$\begin{aligned} d_K(s, t)^2 &= \|\varphi_t - \varphi_s\|_{\mathbb{H}}^2 \\ &= \langle \varphi_t - \varphi_s, \varphi_t - \varphi_s \rangle_{\mathbb{H}} \\ &= \langle \varphi_t, \varphi_t \rangle_{\mathbb{H}} - 2\langle \varphi_t, \varphi_s \rangle_{\mathbb{H}} + \langle \varphi_s, \varphi_s \rangle_{\mathbb{H}} \\ &= K(t, t) - 2K(s, t) + K(s, s). \end{aligned}$$

Using the normalization property  $K(u, u) = 1$  for all  $u \in \mathcal{X}$  from Assumption  $(\mathcal{H}_{\mathcal{P}})$ , this simplifies to:

$$d_K(s, t)^2 = 2(1 - K(s, t)). \quad (33)$$

Consider the function  $g(s) := K(t, s)$ . From the normalization and the Cauchy-Schwarz inequality, we know that  $K(s, t) \leq \sqrt{K(s, s)K(t, t)} = 1$ . Thus,  $g(s)$  achieves its global maximum at  $s = t$ , implying that the gradient vanishes at this point:  $\nabla_s K(t, s)|_{s=t} = 0$ .

Applying a second-order Taylor expansion of  $K(t, s)$  with respect to  $s$  around  $t$ , there exists  $\xi$  on the segment  $[s, t]$  such that:

$$K(t, s) = K(t, t) + \langle \nabla_s K(t, t), s - t \rangle + \frac{1}{2}(s - t)^\top \nabla_s^2 K(t, \xi)(s - t).$$

Substituting  $K(t, t) = 1$  and  $\nabla_s K(t, t) = 0$ :

$$1 - K(t, s) = -\frac{1}{2}(s - t)^\top \nabla_s^2 K(t, \xi)(s - t).$$

Substituting this back into (33):

$$d_K(s, t)^2 = -(s - t)^\top \nabla_s^2 K(t, \xi)(s - t) = |(s - t)^\top \nabla_s^2 K(t, \xi)(s - t)|.$$

By Assumption  $(\mathcal{H}_P)$ , the Hessian is bounded, i.e.,  $\|\nabla_s^2 K(\cdot, \cdot)\|_\infty \leq \mathfrak{C}_P$ . Therefore:

$$d_K(s, t)^2 \leq \mathfrak{C}_P \|s - t\|^2 \implies d_K(s, t) \leq \sqrt{\mathfrak{C}_P} \|s - t\|.$$

■

**Comment on the Kernel Metric  $d_K$ :** The kernel metric  $d_K(s, t) := \|\varphi_t - \varphi_s\|_{\mathbb{H}}$  measures the Hilbertian distance between data points after they have been mapped into the high-dimensional feature space  $\mathbb{H}$ . It defines the “pullback” geometry of the feature space onto the input space  $\mathcal{X}$ . The lemma shows that for smooth kernels (specifically  $C^2$  kernels like the Gaussian kernel), this map is Lipschitz continuous with respect to the Euclidean distance on  $\mathcal{X}$ . This ensures that points close in the input space  $\mathcal{X}$  remain close in the feature space  $\mathbb{H}$ , a critical property for the complexity (with respect to the dimension) of the particle gradient descent algorithms discussed in the paper.

### A.3 Fréchet derivatives

We consider the objective function  $J(\nu)$  defined as the sum of a data-fitting term  $R(\nu)$  and the total variation norm, i.e.,  $J(\nu) = R(\nu) + \kappa \|\nu\|_{\text{TV}}$ . The following lemmas establish the Fréchet derivatives of these components.

**Lemma A.3 (Derivative of the Risk Term)** *Let  $R(\nu) = \frac{1}{2} \|\int_{\mathcal{X}} \varphi_x d\nu(x) - y\|_{\mathbb{H}}^2$  be the risk functional defined on the space of measures  $\mathcal{M}(\mathcal{X})$ , where  $x \mapsto \varphi_x$  is the feature map into a Hilbert space  $\mathbb{H}$  and  $y \in \mathbb{H}$  is the target. Then, the Fréchet derivative of  $R$  at  $\nu$ , denoted by  $R'(\nu)$ , is the function on  $\mathcal{X}$  given by:*

$$\forall t \in \mathcal{X}, \quad R'(\nu)(t) = \left\langle \int_{\mathcal{X}} \varphi_x d\nu(x) - y, \varphi_t \right\rangle_{\mathbb{H}}. \quad (34)$$

**Proof** Let  $\nu \in \mathcal{M}(\mathcal{X})$  and consider a perturbation  $\sigma \in \mathcal{M}(\mathcal{X})$ . We expand the term  $R(\nu + \sigma)$ :

$$\begin{aligned} R(\nu + \sigma) &= \frac{1}{2} \left\| \int_{\mathcal{X}} \varphi_x d(\nu + \sigma)(x) - y \right\|_{\mathbb{H}}^2 \\ &= \frac{1}{2} \left\| \left( \int_{\mathcal{X}} \varphi_x d\nu(x) - y \right) + \int_{\mathcal{X}} \varphi_x d\sigma(x) \right\|_{\mathbb{H}}^2 \\ &= \frac{1}{2} \left\| \int_{\mathcal{X}} \varphi_x d\nu(x) - y \right\|_{\mathbb{H}}^2 + \left\langle \int_{\mathcal{X}} \varphi_x d\nu(x) - y, \int_{\mathcal{X}} \varphi_x d\sigma(x) \right\rangle_{\mathbb{H}} + \frac{1}{2} \left\| \int_{\mathcal{X}} \varphi_x d\sigma(x) \right\|_{\mathbb{H}}^2. \end{aligned}$$

The first term is  $R(\nu)$ . The third term is of order  $O(\|\sigma\|_{\text{TV}}^2)$ . The second term is the linear part in  $\sigma$ . Using (31), we can rewrite the inner product as:

$$\left\langle \int_{\mathcal{X}} \varphi_x d\nu(x) - y, \int_{\mathcal{X}} \varphi_x d\sigma(x) \right\rangle_{\mathbb{H}} = \int_{\mathcal{X}} \left\langle \int_{\mathcal{X}} \varphi_z d\nu(z) - y, \varphi_x \right\rangle_{\mathbb{H}} d\sigma(x).$$

This identifies the Fréchet derivative  $R'(\nu)$  as the function  $t \mapsto \langle \int_{\mathcal{X}} \varphi_x d\nu(x) - y, \varphi_t \rangle_{\mathbb{H}}$ , proving (34). ■

**Lemma A.4 (Derivative of the Regularization Term)** *Consider the regularization term  $H(\nu) = \kappa\nu(\mathcal{X})$  for non-negative measures  $\nu \in \mathcal{M}_+(\mathcal{X})$  (which corresponds to the TV norm for non-negative measures). Its Fréchet derivative is constant:*

$$\forall t \in \mathcal{X}, \quad H'(\nu)(t) = \kappa. \quad (35)$$

**Proof** Let  $\nu \in \mathcal{M}_+(\mathcal{X})$  and let  $\sigma$  be a perturbation such that  $\nu + \sigma \in \mathcal{M}_+(\mathcal{X})$ . The functional  $H$  is linear:

$$H(\nu + \sigma) = \kappa(\nu(\mathcal{X}) + \sigma(\mathcal{X})) = \kappa\nu(\mathcal{X}) + \kappa\sigma(\mathcal{X}).$$

We can write  $\kappa\sigma(\mathcal{X})$  as the integral against the constant function  $\kappa$ :

$$\kappa\sigma(\mathcal{X}) = \int_{\mathcal{X}} \kappa d\sigma(x).$$

Thus, the linear variation is represented by the constant function  $t \mapsto \kappa$ . Therefore, the Fréchet derivative is  $H'(\nu)(t) = \kappa$  for all  $t \in \mathcal{X}$ .  $\blacksquare$

#### A.4 Symmetrization trick

Following Chizat (2022), we address the optimization problem over the space of signed measures  $\mathcal{M}(\mathcal{X})$  by lifting it to the space of non-negative measures on an augmented domain. We introduce the extended space  $\tilde{\mathcal{X}} := \mathcal{X} \times \{-1, +1\}$  and associate to any signed measure  $\mu \in \mathcal{M}(\mathcal{X})$  a non-negative measure  $\nu \in \mathcal{M}_+(\tilde{\mathcal{X}})$ . The correspondence is established through the linear map  $P : \mathcal{M}_+(\tilde{\mathcal{X}}) \rightarrow \mathcal{M}(\mathcal{X})$  defined by:

$$\mu = P(\nu) := \nu(\cdot, +1) - \nu(\cdot, -1).$$

In the context of the linear model where observations are given by  $\int_{\mathcal{X}} \varphi_x d\mu(x)$ , we define the augmented feature map  $\tilde{\varphi} : \tilde{\mathcal{X}} \rightarrow \mathbb{H}$  as  $\tilde{\varphi}(x, s) := s\varphi(x)$  for any  $(x, s) \in \tilde{\mathcal{X}}$ . Consequently, the linear measurements satisfy:

$$\int_{\mathcal{X}} \varphi(x) d\mu(x) = \int_{\tilde{\mathcal{X}}} \tilde{\varphi}(\tilde{x}) d\nu(\tilde{x}).$$

Furthermore, the total variation norm satisfies  $\|\mu\|_{\text{TV}} \leq \nu(\tilde{\mathcal{X}})$ , with equality holding if and only if the positive and negative parts of  $\mu$  have disjoint supports (which is verified for optimal solutions of sparse problems). This symmetrization allows us to solve the signed problem by applying the conic particle gradient descent algorithm to the non-negative measure  $\nu$  on the space  $\tilde{\mathcal{X}}$ .

#### A.5 Lipschitz Continuity of the Fréchet derivative

By smoothness of the kernel (2a-( $\mathcal{H}_{\mathcal{P}}$ )), the Fréchet derivative  $J'_\nu$  is twice continuously differentiable for any  $\nu$ . Define for any twice continuously differentiable  $\psi$ :

$$\|\psi\|_{\mathcal{C}^2(\mathcal{X})} := \max\{\|\psi\|_\infty, \|\nabla\psi\|_\infty, \|\nabla^2\psi\|_\infty\}. \quad (36)$$

The next lemma gives an upper bound on  $\|J'_\nu\|_{\mathcal{C}^2(\mathcal{X})}$ . We already know from (5) that  $\|\nu\|_{\text{TV}} + \mathfrak{C}_{\mathcal{P}} + \kappa \geq \|J'_\nu(t)\|_\infty$ .

**Lemma A.5** *Assume that Assumption  $(\mathcal{H}_{\mathcal{P}})$  holds. For any measure  $\nu \in \mathcal{M}(\mathcal{X})$ , the function  $t \mapsto J'_\nu(t)$  is Lipschitz continuous with constant:*

$$\mathfrak{L}(\nu) = \|\nabla J'_\nu\|_\infty \leq \sqrt{\mathfrak{C}_{\mathcal{P}}} (\mathfrak{C}_{\mathcal{P}} + \|\nu\|_{\text{TV}}).$$

and the dual certificate is gradient-Lipschitz with constant:

$$\|\nabla^2 J'_\nu\|_\infty \leq \mathfrak{C}_{\mathcal{P}} (\|\nu\|_{\text{TV}} + \|y\|_{\mathbb{H}}),$$

so that the full  $\mathcal{C}^2$  norm satisfies:

$$\|J'_\nu\|_{\mathcal{C}^2(\mathcal{X})} \leq \mathfrak{C}_{\mathcal{P}} (\|\nu\|_{\text{TV}} + \mathfrak{C}_{\mathcal{P}}) + \kappa.$$

**Proof** Throughout the proof we use the expression of the Fréchet derivative (4b):

$$J'_\nu(t) = \langle \Phi\nu - y, \varphi_t \rangle_{\mathbb{H}} + \kappa,$$

and the common bound:

$$\|\Phi\nu - y\|_{\mathbb{H}} \leq \|\nu\|_{\text{TV}} + \|y\|_{\mathbb{H}}. \quad (37)$$

• For any  $s, t \in \mathcal{X}$ :

$$|J'_\nu(t) - J'_\nu(s)| = |\langle \Phi\nu - y, \varphi_t - \varphi_s \rangle_{\mathbb{H}}| \leq \|\Phi\nu - y\|_{\mathbb{H}} \|\varphi_t - \varphi_s\|_{\mathbb{H}}.$$

By Lemma A.2,  $\|\varphi_t - \varphi_s\|_{\mathbb{H}} \leq \sqrt{\mathfrak{C}_{\mathcal{P}}} \|t - s\|$ . Combined with (37):

$$|J'_\nu(t) - J'_\nu(s)| \leq \sqrt{\mathfrak{C}_{\mathcal{P}}} (\|\nu\|_{\text{TV}} + \|y\|_{\mathbb{H}}) \|t - s\|,$$

establishing  $\mathfrak{L}(\nu) = \|\nabla J'_\nu\|_\infty \leq \sqrt{\mathfrak{C}_{\mathcal{P}}} (\|\nu\|_{\text{TV}} + \|y\|_{\mathbb{H}})$ .

• Differentiating  $J'_\nu(t) = \langle \Phi\nu - y, \varphi_t \rangle_{\mathbb{H}} + \kappa$  twice under the inner product (justified by the  $\mathcal{C}^2$  smoothness of  $t \mapsto \varphi_t$  under Assumption (2)) gives, for any  $t \in \mathcal{X}$  and unit vectors  $u, v \in \mathbb{R}^d$ :

$$u^\top \nabla_t^2 J'_\nu(t) v = \langle \Phi\nu - y, D_t^2 \varphi_t[u, v] \rangle_{\mathbb{H}}.$$

By Cauchy-Schwarz and (37):

$$|u^\top \nabla_t^2 J'_\nu(t) v| \leq \|\Phi\nu - y\|_{\mathbb{H}} \|D_t^2 \varphi_t[u, v]\|_{\mathbb{H}} \leq (\|\nu\|_{\text{TV}} + \|y\|_{\mathbb{H}}) \|D_t^2 \varphi_t[u, v]\|_{\mathbb{H}}.$$

It remains to bound  $\|D_t^2 \varphi_t[u, v]\|_{\mathbb{H}}$ . Since  $K(s, t) = \langle \varphi_s, \varphi_t \rangle_{\mathbb{H}}$ , differentiating twice with respect to  $t$  gives:

$$u^\top \nabla_t^2 K(s, t) v = \langle \varphi_s, D_t^2 \varphi_t[u, v] \rangle_{\mathbb{H}}.$$

Since  $\varphi_s = \Phi\delta_s$ , the closed linear span  $\overline{\text{span}}\{\varphi_s : s \in \mathcal{X}\} \subset \mathbb{H}$  coincides with the closure of  $\Phi(\mathcal{M}(\mathcal{X}))$ . We assume throughout that this closure equals the whole of  $\mathbb{H}$  (the standard non-degeneracy condition for the RKHS associated with  $K$ ). Then any unit vector  $h \in \mathbb{H}$  is the  $\mathbb{H}$ -limit of finite linear combinations  $\sum_i \alpha_i \varphi_{s_i}$ , and by the Hahn–Banach theorem the operator norm of a continuous linear functional on  $\mathbb{H}$  equals its supremum over  $\overline{\text{span}}\{\varphi_s\}$ . Combined with  $\|\varphi_s\|_{\mathbb{H}} = 1$ , this yields

$$\|D_t^2 \varphi_t[u, v]\|_{\mathbb{H}} = \sup_{\|h\|_{\mathbb{H}}=1} \langle h, D_t^2 \varphi_t[u, v] \rangle_{\mathbb{H}} = \sup_{s \in \mathcal{X}} |\langle \varphi_s, D_t^2 \varphi_t[u, v] \rangle_{\mathbb{H}}| = \sup_{s \in \mathcal{X}} |u^\top \nabla_t^2 K(s, t) v| \leq \|\nabla^2 K\|_\infty \leq \mathfrak{C}_{\mathcal{P}}.$$

Hence  $\|\nabla^2 J'_\nu\|_\infty \leq \mathfrak{C}_P(\|\nu\|_{\text{TV}} + \|y\|_{\mathbb{H}})$ .

- Combining with the  $\|\cdot\|_\infty$  bound from (5) and using  $\|y\|_{\mathbb{H}} \leq \mathfrak{C}_P$ :

$$\|J'_\nu\|_{\mathcal{C}^2(\mathcal{X})} = \max(\|J'_\nu\|_\infty, \|\nabla J'_\nu\|_\infty, \|\nabla^2 J'_\nu\|_\infty) \leq \mathfrak{C}_P(\|\nu\|_{\text{TV}} + \mathfrak{C}_P) + \kappa.$$

■

## A.6 Generalized descent

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 A.6 (Ghadimi et al. (2016))** *The projection operator  $\pi_{\mathcal{X}}(t, v, \beta)$  satisfies the two properties:*

- *Correlation of the projected gradient and gradient lower bound: for any  $t \in \mathcal{X}$ ,  $v \in \mathbb{R}^d$  and  $\beta > 0$ :*

$$\langle v, \pi_{\mathcal{X}}(t, v, \beta) \rangle \geq \|\pi_{\mathcal{X}}(t, v, \beta)\|^2. \quad (38a)$$

- *1-Lipschitz inequality for projection: For any  $t \in \mathcal{X}$ ,  $(v_1, v_2) \in \mathbb{R}^d$  and  $\beta > 0$ :*

$$\|\pi_{\mathcal{X}}(t, v_1, \beta) - \pi_{\mathcal{X}}(t, v_2, \beta)\| \leq \|v_1 - v_2\|. \quad (38b)$$

## Appendix B. Descent properties: technical results

The goal of this Appendix section is to provide some technical proofs of the key descent properties stated in Proposition 2.

### B.1 Proof of Proposition 2

Our argument follows the same lines as in the proof of (Chizat, 2022, Lemma 2.5). We decompose  $\nu^+$  as  $\nu^+ = \nu + (\tilde{\nu} - \nu) + (\nu^+ - \tilde{\nu})$ , where  $\tilde{\nu} = \mathbb{W}_{\nu, \alpha} \nu$  and  $\nu^+ = T_{\nu, \beta}^\# \tilde{\nu}$ . Invoke (4a) with  $\sigma = \nu^+ - \nu$  to get that

$$\begin{aligned} J(\nu^+) - J(\nu) &= \langle J'_\nu, \sigma \rangle + \frac{1}{2} \|\Phi(\sigma)\|_{\mathbb{H}}^2 \\ &= \int_{\mathcal{X}} J'_\nu(d\nu^+ - d\tilde{\nu}) + \int_{\mathcal{X}} J'_\nu(d\tilde{\nu} - d\nu) + \frac{1}{2} \left\| \int_{\mathcal{X}} \varphi_t(d\nu^+(t) - d\nu(t)) \right\|_{\mathbb{H}}^2 \\ &\leq \underbrace{\int_{\mathcal{X}} J'_\nu(d\nu^+ - d\tilde{\nu})}_{:=A_1} + \underbrace{\int_{\mathcal{X}} J'_\nu(d\tilde{\nu} - d\nu)}_{:=A_2} \\ &\quad + \underbrace{\left\| \int_{\mathcal{X}} \varphi_t(d\nu^+(t) - d\tilde{\nu}(t)) \right\|_{\mathbb{H}}^2}_{:=B_1} + \underbrace{\left\| \int_{\mathcal{X}} \varphi_t(d\tilde{\nu}(t) - d\nu(t)) \right\|_{\mathbb{H}}^2}_{:=B_2} \end{aligned}$$

In order to study the previous terms, we define  $\delta = \alpha(\|J'_\nu\|_{\mathcal{C}^2(\mathcal{X})} \vee 1)$ .

Study of  $A_1$ . We use the proximal update defined in Equation (6b):

$$\begin{aligned} A_1 &= \int [J'_\nu(t - \beta\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)) - J'_\nu(t)] d\tilde{\nu}(t) \\ &\leq \int \left( -\beta \langle \pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta), \nabla J'_\nu(t) \rangle + \frac{\beta^2 \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2}{2} \|\nabla^2 J'_\nu\|_\infty \right) d\tilde{\nu}(t). \end{aligned}$$

By Lemma A.5,  $\|\nabla^2 J'_\nu\|_\infty \leq \mathfrak{C}_{\mathcal{P}}(\mathfrak{C}_{TV} + \mathfrak{C}_{\mathcal{P}})$ . By Lemma A.6 *ii*) and Lemma A.5,

$$\|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 \leq \|\nabla J'_\nu\|_\infty^2 \leq \mathfrak{C}_{\mathcal{P}}(\mathfrak{C}_{TV} + \mathfrak{C}_{\mathcal{P}})^2.$$

Using  $\int \|\pi_{\mathcal{X}}\|^2 d\tilde{\nu} \leq e^\delta \int \|\pi_{\mathcal{X}}\|^2 d\nu$  (same ratio argument as (39b)), the Taylor remainder of  $A_1$  satisfies:

$$\frac{\beta^2}{2} \|\nabla^2 J'_\nu\|_\infty \int_{\mathcal{X}} \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 d\tilde{\nu} \leq \frac{\beta}{2} \mathfrak{C}_{\mathcal{P}}(\mathfrak{C}_{TV} + \mathfrak{C}_{\mathcal{P}}) e^\delta \|g_\nu^\beta\|_{L^2(\nu)}^2. \quad (39a)$$

We are led to study the first order term. Starting with (38a), we get

$$\begin{aligned} &\int -\beta \langle \pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta), \nabla J'_\nu(t) \rangle d\tilde{\nu}(t) \\ &\leq -\beta \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 d\tilde{\nu}(t) \\ &\leq -\beta \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 d\nu(t) + \beta \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 \left| e^{-\alpha J'_\nu(t)} - 1 \right| d\nu(t) \\ &\leq -\beta \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 d\nu(t) + \alpha\beta \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 |J'_\nu(t)| e^{\alpha |J'_\nu(t)|} d\nu(t) \\ &\leq -\beta(1 - \alpha(\kappa + \|\nu\|_{TV} + \|y\|_{\mathbb{H}})e^\delta) \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 d\nu(t), \end{aligned} \quad (39b)$$

where we have used (4b) and rough upper bounds for the last inequality.

Study of  $A_2$ . Observe that for any  $\psi \in \mathcal{C}^2(\mathcal{X})$ , one has

$$\int_{\mathcal{X}} \psi(d\tilde{\nu} - d\nu) = \int_{\mathcal{X}} (e^{-\alpha J'_\nu(t)} - 1) \psi(t) d\nu(t) \quad (39c)$$

We will use the standard inequality:

$$\forall u \in \mathbb{R} \quad |e^{-u} - 1 + u| \leq \frac{u^2}{2} e^{|u|} \quad (39d)$$

and a first and second Taylor expansion on  $\psi$ . For any  $(t, t + \mathbf{h}) \in \mathcal{X}$ :

$$|\psi(t + \mathbf{h}) - \psi(t)| \leq \|\mathbf{h}\| \|\nabla \psi\|_\infty \quad \text{and} \quad |\psi(t + \mathbf{h}) - \psi(t) - \langle \mathbf{h}, \nabla \psi(t) \rangle| \leq \frac{\|\mathbf{h}\|^2}{2} \|\nabla^2 \psi\|_\infty. \quad (39e)$$

We then use Equation (39d) and Equation (39e) and observe that the first term of the right hand side of Equation (39c) may be upper-bounded as follows:

$$\begin{aligned} (e^{-\alpha J'_\nu(t)} - 1)\psi(t) &= -\alpha J'_\nu(t)\psi(t) + (e^{-\alpha J'_\nu(t)} - 1 + \alpha J'_\nu(t))\psi(t) \\ &\leq -\alpha J'_\nu(t)\psi(t) + \frac{\alpha^2 J'_\nu(t)^2}{2} e^{\alpha |J'_\nu(t)|} \|\psi\|_\infty. \end{aligned}$$

We then integrate with respect to  $d\nu$  and get (while omitting the variable  $t$  for the sake of readability):

$$\int_{\mathcal{X}} \psi(d\tilde{\nu} - d\nu) \leq - \int_{\mathcal{X}} \alpha J'_\nu \psi d\nu + \underbrace{\frac{\alpha^2 \|\psi\|_\infty}{2} \int_{\mathcal{X}} |J'_\nu|^2 e^{\alpha |J'_\nu|} d\nu}_{\text{Remainder}(\psi)} \quad (39f)$$

We are led to study the remainder term. Replacing  $\psi$  by  $J'_\nu$  in (39f), we get:

$$\text{Remainder}(J'_\nu) \leq \frac{\alpha \|J'_\nu\|_\infty e^{\alpha \|J'_\nu\|_\infty}}{2} \|g_\nu^\alpha\|_{L^2(\nu)}^2.$$

Using  $\delta = \alpha(\|J'_\nu\|_{\mathcal{C}^2(\mathcal{X})} \vee 1)$ , we then deduce the following bound:

$$A_2 = \int_{\mathcal{X}} J'_\nu(d\tilde{\nu} - d\nu) \leq -\|g_\nu^\alpha\|_{L^2(\nu)}^2 \left(1 - \frac{\delta e^\delta}{2}\right). \quad (39g)$$

Study of  $B_1$ . To upper bound the second order terms, we use the  $\sqrt{\mathfrak{C}_{\mathcal{P}}}$ -Lipschitz continuity of  $t \mapsto \varphi_t$  from Lemma A.2, so that  $\|\varphi_t(a) - \varphi_t(b)\|_{\mathbb{H}} \leq \mathfrak{C}_{\mathcal{P}} \|a - b\|^2$ :

$$\begin{aligned} B_1 &= \left\| \int_{\mathcal{X}} \varphi_t(d\nu^+(t) - d\tilde{\nu}(t)) \right\|_{\mathbb{H}}^2 \\ &= \left\| \int_{\mathcal{X}} \varphi_t(t - \beta \pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)) - \varphi_t(t) d\tilde{\nu}(t) \right\|_{\mathbb{H}}^2 \\ &\leq \beta^2 \mathfrak{C}_{\mathcal{P}} \|\tilde{\nu}\|_{TV}^2 \left( \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\| \frac{d\tilde{\nu}(t)}{\|\tilde{\nu}\|_{TV}} \right)^2 \\ &\leq \beta^2 \mathfrak{C}_{\mathcal{P}} \|\tilde{\nu}\|_{TV} \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 d\tilde{\nu}(t) \\ &\leq \beta^2 \mathfrak{C}_{\mathcal{P}} \mathfrak{C}_{TV} e^{2\delta} \int \|\pi_{\mathcal{X}}(t, \nabla J'_\nu(t), \beta)\|^2 d\nu(t) \end{aligned} \quad (39h)$$

where we used Lemma A.6 *ii*) and the Jensen inequality on the normalized measure  $\frac{d\tilde{\nu}(t)}{\|\tilde{\nu}\|_{TV}}$ .

Study of  $B_2$ . Since  $K \leq 1$ , for any signed measure  $\mu = \mu^+ - \mu^-$  one has  $\|\Phi\mu\|_{\mathbb{H}}^2 \leq (\mu^+(\mathcal{X}) + \mu^-(\mathcal{X}))^2 = \|\mu\|_{\text{TV}}^2$ ; hence:

$$\begin{aligned} \left\| \int_{\mathcal{X}} \varphi_t(d\tilde{\nu}(t) - d\nu(t)) \right\|_{\mathbb{H}}^2 &= \int_{\mathcal{X}} \left( \int_{\mathcal{X}} K(s, t) (d\tilde{\nu}(s) - d\nu(s)) \right) (d\tilde{\nu}(t) - d\nu(t)) \\ &\leq \|\tilde{\nu} - \nu\|_{\text{TV}}^2 \\ &= \left( \int_{\mathcal{X}} |e^{-\alpha J'_\nu(t)} - 1| d\nu(t) \right)^2 \\ &\leq \left( \int_{\mathcal{X}} \alpha |J'_\nu(t)| + \frac{\alpha^2 |J'_\nu(t)|^2}{2} e^{\alpha |J'_\nu(t)|} d\nu(t) \right)^2 \\ &\leq 2\|\nu\|_{\text{TV}} \int_{\mathcal{X}} \left( \alpha^2 |J'_\nu(t)|^2 + \frac{\alpha^4 |J'_\nu(t)|^4}{4} e^{2\alpha |J'_\nu(t)|} \right) d\nu(t), \end{aligned}$$

where we used in the last line the Cauchy-Schwarz inequality and  $(u + v)^2 \leq 2u^2 + 2v^2$ . Using again  $\delta$  defined above, we deduce that:

$$B_2 = \left\| \int_{\mathcal{X}} \varphi_t(d\tilde{\nu}(t) - d\nu(t)) \right\|_{\mathbb{H}}^2 \leq 2\alpha \mathfrak{C}_{\text{TV}} \|g_\nu^\alpha\|_{L^2(\nu)}^2 \left( 1 + \frac{\delta^2 e^{2\delta}}{4} \right). \quad (39i)$$

Gathering (39b) with (39a) and (39h) on one side, and (39g) with (39i) on the other side, we deduce the upper bound:

$$\begin{aligned} J(\nu^+) - J(\nu) &\leq -\|g_\nu^\alpha\|_{L^2(\nu)}^2 \left( 1 - \frac{\delta e^\delta}{2} - 2\alpha \mathfrak{C}_{\text{TV}} \left( 1 + \frac{\delta^2 e^{2\delta}}{4} \right) \right) \\ &\quad - \|g_\nu^\beta\|_{L^2(\nu)}^2 \left( 1 - \alpha(\kappa + \|\nu\|_{\text{TV}} + \|y\|_{\mathbb{H}})e^\delta - \beta \mathfrak{C}_{\mathcal{P}} \left( \mathfrak{C}_{\text{TV}} e^{2\delta} + \frac{(\mathfrak{C}_{\text{TV}} + \mathfrak{C}_{\mathcal{P}}) e^\delta}{2} \right) \right). \end{aligned}$$

We are then led to choose

$$\alpha < \frac{1}{10(1 + \|\nu\|_{\text{TV}} + \|y\|_{\mathbb{H}} + \kappa)(1 \vee \mathfrak{C}_{\text{TV}})},$$

Such a constraint on  $\alpha$  ensures that  $\delta e^\delta < \frac{1}{7}$ , which occurs as soon as  $\alpha$  is such that  $\delta < \frac{1}{10}$ . By Lemma A.5, since  $\|\nu\|_{\text{TV}} \leq \mathfrak{C}_{\text{TV}}$  and  $\|y\|_{\mathbb{H}} \leq \mathfrak{C}_{\mathcal{P}}$ :

$$\|J'_\nu\|_{\mathcal{C}^2(\mathcal{X})} \leq \mathfrak{C}_{\mathcal{P}}(\mathfrak{C}_{\text{TV}} + \mathfrak{C}_{\mathcal{P}}) + \kappa,$$

so  $\delta = \alpha(\|J'_\nu\|_{\mathcal{C}^2(\mathcal{X})} \vee 1)$  is small under the constraint on  $\alpha$ . It is then straightforward to verify that:

$$\delta e^\delta > \max \left( \alpha, \alpha \frac{\delta^2 e^{2\delta}}{4} \right).$$

In the meantime, using  $e^\delta \leq e^{1/5}$  and  $e^{2\delta} \leq e^{1/5}$ , the choice  $\beta \leq \frac{1}{2\mathfrak{C}_{\mathcal{P}}(\mathfrak{C}_{\mathcal{P}} + 3\mathfrak{C}_{\text{TV}})e^{1/5}}$  yields

$$\beta \mathfrak{C}_{\mathcal{P}} \left( \mathfrak{C}_{\text{TV}} e^{2\delta} + \frac{(\mathfrak{C}_{\text{TV}} + \mathfrak{C}_{\mathcal{P}}) e^\delta}{2} \right) \leq \frac{\beta \mathfrak{C}_{\mathcal{P}} e^{1/5} (3\mathfrak{C}_{\text{TV}} + \mathfrak{C}_{\mathcal{P}})}{2} \leq \frac{1}{4},$$

which finally entails:

$$J(\nu^+) - J(\nu) \leq -\frac{3}{4} \left( \|g_\nu^\alpha\|_{L^2(\nu)}^2 + \|g_\nu^\beta\|_{L^2(\nu)}^2 \right) \quad (39j)$$

## B.2 Linearization of the mirror descent strategy

This section concerns conditional expectations on the mirror descent strategy. Below, we state a general result on  $\widehat{J}'_\nu(t) = \frac{1}{m} \sum_{l=1}^m \widehat{J}'_\nu(t, Z_l)$  where  $(Z_1, \dots, Z_m)$  stands for a mini-batch sample of size  $m$ .

**Proposition 5** *Assume that  $\alpha \leq \sqrt{8 \log 8} \mathbf{E}_\infty^{-1}$ , then:*

$$\forall t \in \mathcal{X} : \quad \left| \mathbb{E} \left[ e^{-\alpha \widehat{J}'_\nu(t, Z)} - (1 - \alpha J'_\nu(t)) \mid \nu \right] \right| \leq \frac{\alpha^2 \mathbf{E}_\infty^2}{m} e^{-\alpha J'_\nu(t)} + \frac{\alpha^2 J'_\nu(t)^2}{2} e^{\alpha |J'_\nu(t)|}.$$

**Proof** Below, we compute the expectation with respect to the randomness brought by the mini-batch sample. Let  $t \in \mathcal{X}$ . We begin with the following decomposition:

$$\begin{aligned} & \mathbb{E} \left[ \alpha J'_\nu(t) + e^{-\alpha \widehat{J}'_\nu(t)} - 1 \right] \\ &= \alpha J'_\nu(t) - 1 + \mathbb{E} \left[ e^{-\alpha \widehat{J}'_\nu(t)} \right] \\ &= \left[ \alpha J'_\nu(t) - 1 + e^{-\alpha J'_\nu(t)} \mathbb{E} \left[ e^{-\alpha [\widehat{J}'_\nu(t) - J'_\nu(t)]} \right] \right] \\ &= \left[ \alpha J'_\nu(t) - 1 + e^{-\alpha J'_\nu(t)} \right] + e^{-\alpha J'_\nu(t)} \mathbb{E} \left[ e^{-\alpha [\widehat{J}'_\nu(t) - J'_\nu(t)]} - 1 \right] \\ &= \left[ \alpha J'_\nu(t) - 1 + e^{-\alpha J'_\nu(t)} \right] + e^{-\alpha J'_\nu(t)} \mathbb{E} \left[ e^{-\frac{\alpha}{m} \sum_{l=1}^m [\widehat{J}'_\nu(t, Z_l) - J'_\nu(t)]} - 1 \right] \\ &= \left[ \alpha J'_\nu(t) - 1 + e^{-\alpha J'_\nu(t)} \right] + e^{-\alpha J'_\nu(t)} \left( \left( \mathbb{E} \left[ e^{-\frac{\alpha}{m} [\widehat{J}'_\nu(t, Z_l) - J'_\nu(t)]} \right] \right)^m - 1 \right). \end{aligned}$$

To derive an upper bound, we use the inequality  $|e^{-u} - 1 + u| \leq \frac{u^2}{2} e^{|u|}$  which holds for any  $u \in \mathbb{R}$  for the first term and we apply the Hoeffding Lemma to the random variable  $\widehat{J}'_\nu(t, Z) - J'_\nu(t)$ , which is a centered random variable almost surely bounded by  $\mathbf{E}_\infty$ . According to Assumption (9a), we obtain that:

$$\left| \mathbb{E} \left[ e^{-\alpha [\widehat{J}'_\nu(t, Z) - J'_\nu(t)]} - 1 \right] \right| \leq e^{\frac{\mathbf{E}_\infty^2 \alpha^2}{8m}} - 1 \leq \frac{\alpha^2 \mathbf{E}_\infty^2}{8m} e^{\frac{\alpha^2 \mathbf{E}_\infty^2}{8m}} \leq \frac{\alpha^2 \mathbf{E}_\infty^2}{m}$$

where thanks to our assumption on  $\alpha$ , we observe that  $e^{\frac{\alpha^2 \mathbf{E}_\infty^2}{8m}} \leq 8$ . We finally obtain the desired upper bound.  $\blacksquare$

## Appendix C. Proofs of the deterministic results

In this paragraph, we present all the proofs related to the deterministic results introduced at the beginning of our work. In particular, we establish the proof of the key results of Section 2.2 and Section 2.4.

### C.1 Total variation boundedness

**Proof** [Proof of Proposition 3, *i*] We address  $(15 - (\mathcal{H}_{\text{TV}}^\infty))$ . Let  $k \in \mathbb{N}$  be fixed. Our starting point is the relationship

$$\nu_{k+1} = \nu_{k++} + \varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda = \left( 1 - \mathbf{1}_{\mathcal{P}_{\nu_{k+}}} \right) \nu_{k+} + \varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda.$$

Computing the total variation norm, we obtain:

$$\|\nu_{k+1}\|_{\text{TV}} \leq \|\nu_k\|_{\text{TV}} + \varepsilon_k \lambda(\mathcal{X}) \leq \int_{\mathcal{X}} e^{-\alpha J'_{\nu_k}(t)} d\nu_k(t) + \max_{j \geq 0} \varepsilon_j \lambda(\mathcal{X}).$$

According to Assumption (2), observe that for any  $t \in \mathcal{X}$

$$J'_{\nu_k}(t) = \int_{\mathcal{X}} \langle \varphi_t, \varphi_s \rangle d\nu_k(s) - \langle y, \varphi_t \rangle + \kappa \geq \mathfrak{c}_{\mathcal{P}} \|\nu_k\|_{\text{TV}} - \|y\|_{\mathbb{H}} + \kappa \geq -\|y\|_{\mathbb{H}} + \kappa.$$

We then deduce that thanks to the condition  $\varepsilon_k \leq \alpha$ :

$$\|\nu_{k+1}\|_{\text{TV}} \leq e^{-\alpha(\mathfrak{c}_{\mathcal{P}} \|\nu_k\|_{\text{TV}} - \|y\|_{\mathbb{H}})} \|\nu_k\|_{\text{TV}} + \lambda(\mathcal{X})\alpha$$

We define  $\mathfrak{M}$  as  $\mathfrak{M} = \frac{2}{\mathfrak{c}_{\mathcal{P}}} \|y\|_{\mathbb{H}} + \sqrt{\frac{2e\lambda(\mathcal{X})}{\mathfrak{c}_{\mathcal{P}}}}$  and we verify that

$$u \geq \mathfrak{M} \implies \mathfrak{c}_{\mathcal{P}} u - \|y\|_{\mathbb{H}} \geq \frac{\mathfrak{c}_{\mathcal{P}} u}{2}.$$

We now consider the two different cases:

- If  $\|\nu_k\|_{\text{TV}} \geq \mathfrak{M}$ , then

$$\begin{aligned} \|\nu_{k+1}\|_{\text{TV}} &\leq e^{-\alpha \frac{\mathfrak{c}_{\mathcal{P}}}{2} \|\nu_k\|_{\text{TV}}} \|\nu_k\|_{\text{TV}} + \lambda(\mathcal{X})\alpha \\ &= \|\nu_k\|_{\text{TV}} - (1 - e^{-\alpha \frac{\mathfrak{c}_{\mathcal{P}}}{2} \|\nu_k\|_{\text{TV}}}) \|\nu_k\|_{\text{TV}} + \lambda(\mathcal{X})\alpha \\ &= \|\nu_k\|_{\text{TV}} - \frac{2}{\alpha \mathfrak{c}_{\mathcal{P}}} \varphi\left(\alpha \frac{\mathfrak{c}_{\mathcal{P}}}{2} \|\nu_k\|_{\text{TV}}\right) + \lambda(\mathcal{X})\alpha, \end{aligned}$$

where  $\varphi$  is defined by  $\varphi(t) = t(1 - e^{-t})$ . We check that  $\varphi$  is an increasing function so that when  $\|\nu_k\|_{\text{TV}} \geq \mathfrak{M}$ , then

$$\|\nu_{k+1}\|_{\text{TV}} \leq \|\nu_k\|_{\text{TV}} - \frac{2}{\alpha \mathfrak{c}_{\mathcal{P}}} \varphi\left(\alpha \frac{\mathfrak{c}_{\mathcal{P}}}{2} \mathfrak{M}\right) + \lambda(\mathcal{X})\alpha = \|\nu_k\|_{\text{TV}} - \mathfrak{M} \left(1 - e^{-\alpha \frac{\mathfrak{c}_{\mathcal{P}}}{2} \mathfrak{M}}\right) + \lambda(\mathcal{X})\alpha.$$

Thanks to our condition on  $\alpha$ , we know that  $\alpha \frac{\mathfrak{c}_{\mathcal{P}}}{2} \mathfrak{M} \leq 1$ , and the convex inequality when  $t \in [0, 1]$  :  $e^{-t} \leq 1 - t/e$  yields:

$$\|\nu_{k+1}\|_{\text{TV}} \leq \|\nu_k\|_{\text{TV}} - \mathfrak{M} \times \frac{\alpha \frac{\mathfrak{c}_{\mathcal{P}}}{2} \mathfrak{M}}{e} + \lambda(\mathcal{X})\alpha.$$

Thanks to our definition of  $\mathfrak{M}$ , we then observe that in this case  $\|\nu_{k+1}\|_{\text{TV}} \leq \|\nu_k\|_{\text{TV}}$ .

- If  $\|\nu_k\|_{\text{TV}} \leq \mathfrak{M}$ , then we use the straightforward upper bound:

$$\|\nu_{k+1}\|_{\text{TV}} \leq \mathfrak{M} e^{\alpha \|y\|_{\mathbb{H}}} + \lambda(\mathcal{X})\alpha \leq e\mathfrak{M} + \lambda(\mathcal{X})$$

A direct induction argument then shows that

$$\forall k \geq 0 \quad \|\nu_k\|_{\text{TV}} \leq \|\nu_0\|_{\text{TV}} \vee (e\mathfrak{M} + \lambda(\mathcal{X})).$$

Recalling  $\mathfrak{M} = \frac{2\|y\|_{\mathbb{H}}}{\mathfrak{c}_{\mathcal{P}}} + \sqrt{\frac{2e\lambda(\mathcal{X})}{\mathfrak{c}_{\mathcal{P}}}}$  and  $\mathfrak{R} = \frac{e\|y\|_{\mathbb{H}}}{\mathfrak{c}_{\mathcal{P}}} + \sqrt{\frac{e^3\lambda(\mathcal{X})}{\mathfrak{c}_{\mathcal{P}}}} + \lambda(\mathcal{X})$ , we have  $e\mathfrak{M} + \lambda(\mathcal{X}) = \frac{2e\|y\|_{\mathbb{H}}}{\mathfrak{c}_{\mathcal{P}}} + \sqrt{\frac{2e^3\lambda(\mathcal{X})}{\mathfrak{c}_{\mathcal{P}}}} + \lambda(\mathcal{X}) \leq 2\mathfrak{R}$ , since  $\sqrt{2} \leq 2$  and  $\lambda(\mathcal{X}) \leq 2\lambda(\mathcal{X})$ . The proof bound is thus at most the proposition's statement bound  $\|\nu_0\|_{\text{TV}} \vee 2\mathfrak{R} =: \mathfrak{C}_{TV}$ , concluding the proof of the boundedness of the sequence  $(\nu_k)_{k \geq 0}$ .  $\blacksquare$

## C.2 $\varepsilon$ -smoothness evolution

In this paragraph, we now establish that our sequence  $(\nu_k)_{k \geq 0}$  verifies Assumption  $(\mathcal{H}_\varepsilon)$  given by Equation (14), *i.e.* we establish the second part of Proposition 3. We first prove the perturbation bound and the deletion-set inclusion stated without proof in Remark 2.1, on which the proof of Proposition 3 *ii*) relies.

**Proof** [Proof of Remark 2.1]

*Proof of* (18a). The Fréchet identity (4b), with  $\|\varphi_t\|_{\mathbb{H}} = 1$ , gives

$$|J'_{\nu_{k+}}(t) - J'_{\nu_k}(t)| \leq \int_{\mathcal{X}} |e^{-\alpha J'_{\nu_k}(s)} - 1| d\nu_k(s).$$

The uniform bound  $\|J'_{\nu_k}\|_\infty \leq \mathfrak{C}_{\mathcal{P}} + \mathfrak{C}_{TV} + \kappa$  (from (5) and  $(15 - (\mathcal{H}_{TV}^\infty))$ ), the elementary inequality  $|e^{-x} - 1| \leq |x|e^{|x|}$ , and  $\|\nu_k\|_{TV} \leq \mathfrak{C}_{TV}$  now yield (18a).

*Proof of* (18b). For  $t \in \mathcal{P}_{\nu_{k+}}$ , definition (17) gives  $J'_{\nu_{k+}}(t) > -2\alpha^{-1} \log \varepsilon_k + \mathfrak{C}_w$ , and (18a) then yields  $J'_{\nu_k}(t) \geq J'_{\nu_{k+}}(t) - \mathfrak{C}_w > -2\alpha^{-1} \log \varepsilon_k$ .  $\blacksquare$

**Proof** [Proof of Proposition 3, *ii*)] Recall that the update defined in Section 2.1 yields:

$$\nu_{k+1} := \nu_{k++} + \varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda = \nu_{k+}(1 - \mathbf{1}_{\mathcal{P}_{\nu_{k+}}}) + \varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda, \quad (40)$$

where  $\nu_{k++}$  is a positive measure.

**Assumption**  $(\mathcal{H}_\varepsilon^+)$  The definition of  $\mathcal{N}_{\nu_{k+}} = \{J'_{\nu_{k+}} \leq 0\}$  implies Assumption  $\mathcal{H}_\varepsilon^+$ .

**Assumption**  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$  First, using (4a),

$$\begin{aligned} J(\nu_{k+1}) - J(\nu_{k+}) &= \int J'_{\nu_{k+}} d(\nu_{k+1} - \nu_{k+}) + \frac{1}{2} \|\Phi(\nu_{k+1} - \nu_{k+})\|_{\mathbb{H}}^2, \\ &\leq - \int_{\mathcal{P}_{\nu_{k+}}} J'_{\nu_{k+}} d\nu_{k+} + \varepsilon_k \int_{\mathcal{N}_{\nu_{k+}}} J'_{\nu_{k+}} d\lambda \\ &\quad + \|\Phi(\nu_{k+} \mathbf{1}_{\mathcal{P}_{\nu_{k+}}})\|_{\mathbb{H}}^2 + \|\Phi(\varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda)\|_{\mathbb{H}}^2, \\ &\leq \|\Phi(\nu_{k+} \mathbf{1}_{\mathcal{P}_{\nu_{k+}}})\|_{\mathbb{H}}^2 + \|\Phi(\varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda)\|_{\mathbb{H}}^2, \end{aligned}$$

since  $\mathcal{P}_{\nu_{k+}} \subset \{J'_{\nu_{k+}} > 0\}$  and  $\mathcal{N}_{\nu_{k+}} \subset \{J'_{\nu_{k+}} < 0\}$ . Now, by Remark 2.1, it holds the following inclusion  $\mathcal{P}_{\nu_{k+}} \subset \{J'_{\nu_k} > -2\alpha^{-1} \log \varepsilon_k\}$ , so that  $e^{-\alpha J'_{\nu_k}(t)} < \varepsilon_k^2$  on  $\mathcal{P}_{\nu_{k+}}$ . Now, using that  $d\nu_{k+}(t) = e^{-\alpha J'_{\nu_k}(t)} d\nu_k(t)$  and  $\|\varphi_t\|_{\mathbb{H}} = 1$ :

$$\begin{aligned} \|\Phi(\nu_{k+} \mathbf{1}_{\mathcal{P}_{\nu_{k+}}})\|_{\mathbb{H}}^2 &= \left\| \int_{\mathcal{P}_{\nu_{k+}}} \varphi_t d\nu_{k+}(t) \right\|_{\mathbb{H}}^2, \\ &\leq \nu_{k+}(\mathcal{P}_{\nu_{k+}}) \times \int_{\mathcal{P}_{\nu_{k+}}} \|\varphi_t\|_{\mathbb{H}}^2 d\nu_{k+}(t), \\ &= \left[ \int_{\mathcal{P}_{\nu_{k+}}} e^{-\alpha J'_{\nu_k}(t)} d\nu_k(t) \right]^2 \leq \|\nu_k\|_{TV}^2 \varepsilon_k^4 \leq \|\nu_k\|_{TV}^2 \varepsilon_k^2, \end{aligned}$$

where the last inequality uses  $\varepsilon_k \leq 1$ ; the tighter  $\varepsilon_k^4$  bound shows that the prune-set contribution is asymptotically dominated by the birth-set  $\varepsilon_k^2$  term derived below. Similarly, by Cauchy–Schwarz,

$$\begin{aligned} \|\Phi(\varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda)\|_{\mathbb{H}}^2 &= \varepsilon_k^2 \left\| \int_{\mathcal{N}_{\nu_{k+}}} \varphi_s d\lambda(s) \right\|_{\mathbb{H}}^2, \\ &\leq \varepsilon_k^2 \lambda(\mathcal{N}_{\nu_{k+}}) \int_{\mathcal{N}_{\nu_{k+}}} \|\varphi_s\|_{\mathbb{H}}^2 d\lambda(s) = \varepsilon_k^2 \lambda(\mathcal{N}_{\nu_{k+}})^2 \leq \lambda(\mathcal{X})^2 \varepsilon_k^2, \end{aligned}$$

using  $\|\varphi_s\|_{\mathbb{H}} = 1$ . This entails that Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$  is satisfied.

**Assumption**  $(\mathcal{H}_\varepsilon^{\text{smooth},2})$  Concerning  $(\mathcal{H}_\varepsilon^{\text{smooth},2})$ , we use again the inclusion  $\mathcal{P}_{\nu_{k+}} \subset \{J'_{\nu_k} > -2\alpha^{-1} \log \varepsilon_k\}$  (Remark 2.1). For any  $t \in \mathcal{X}$ :

$$\begin{aligned} \left| J'_{\nu_{k+1}}(t) - J'_{\nu_{k+}}(t) \right| &= |\langle \varphi_t, \Phi(\nu_{k+1} - \nu_{k+}) \rangle|, \\ &= \left| \int \langle \varphi_u, \varphi_t \rangle_{\mathbb{H}} d(\nu_{k+1} - \nu_{k+})(u) \right|, \\ &\leq \left| \int_{\mathcal{P}_{\nu_{k+}}} \langle \varphi_u, \varphi_t \rangle_{\mathbb{H}} d\nu_{k+}(u) \right| + \left| \varepsilon_k \int_{\mathcal{N}_{\nu_{k+}}} \langle \varphi_u, \varphi_t \rangle_{\mathbb{H}} d\lambda(u) \right|, \\ &\leq \int_{\mathcal{P}_{\nu_{k+}}} e^{-\alpha J'_{\nu_k}(u)} d\nu_k(u) + \varepsilon_k \lambda(\mathcal{X}), \\ &\leq \|\nu_k\|_{\text{TV}} \varepsilon_k^2 + \varepsilon_k \lambda(\mathcal{X}). \end{aligned}$$

This inequality is uniform in  $t \in \mathcal{X}$ , which proves the desired result.  $\blacksquare$

### C.3 One-step analysis

We introduce:

$$v_{k-1+} = \min(J'_{\nu_{k-1+}}) \wedge 0 \quad \text{and} \quad \Theta_{k-1+} := \left\{ J'_{\nu_{k-1+}} \leq \frac{v_{k-1+}}{2} \right\}. \quad (41)$$

We now present a quantitative result that explicitly connects the weight update  $T_{\nu,\alpha}$  to Assumption (14), by relating the quantity  $v_{k-1+}$  to the evolution of the energy  $J$  under the update  $T_{\nu,\alpha}$ . Proposition 6 below is largely inspired by Proposition H.1 in Chizat (2022). Here, to relate the increments  $J(\nu_{k+}) - J(\nu_k)$  to the minimum of  $J'_{\nu_{k-1+}}$ , we rely on two main ingredients: the descent property (see Proposition 2) and Assumption  $(\mathcal{H}_\varepsilon)$ , which ensures that sufficient mass is present where  $J'_{\nu_k}$  is negative.

**Proposition 6** *Assume  $(\mathcal{H}_\varepsilon)$  and  $(\mathcal{H}_{\text{TV}}^\infty)$  hold so that Equations (14) and (15– $(\mathcal{H}_{\text{TV}}^\infty)$ ) are satisfied, if  $\alpha$  is chosen such that (7) holds and assume that  $v_{k-1+}^2 \geq 24\varepsilon_{k-1}^2 \mathfrak{C}^2$ , then:*

$$J(\nu_{k+}) - J(\nu_k) \leq -\frac{3\alpha}{2} C_d \mathfrak{L}^{-d} |v_{k-1+}|^{2+d} \varepsilon_{k-1}.$$

where  $\mathfrak{L}$  is introduced in (4c) and  $C_d$  in Lemma 7.

**Proof** Consider  $k \in \mathbb{N}^*$ . Our starting point is Proposition 2: we keep only the negative descent contribution produced by  $|J'_{\nu_k}|^2$  and observe that, with  $\alpha$  small enough so that (7) holds and with  $\beta = 0$ , we have:

$$\begin{aligned} J(\nu_{k+}) - J(\nu_k) &\leq -\frac{3}{4} \|g_{\nu_k}^\alpha\|_{L^2(\nu_k)}^2 \\ &= -\frac{3}{4} \alpha \int_{\mathcal{X}} |J'_{\nu_k}|^2 d\nu_k \end{aligned} \quad (42a)$$

$$\leq -\frac{3}{4} \alpha \int_{\Theta_{k-1+}} |J'_{\nu_k}|^2 d\nu_k. \quad (42b)$$

Now, according to Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},2})$ , we know that  $|J'_{\nu_k}(t) - J'_{\nu_{k-1+}}(t)| \leq \mathfrak{C}\varepsilon_{k-1}$  uniformly for all values of  $t$ . In particular, for any  $t \in \mathcal{X}$ , we deduce from the inequality  $(a+b)^2 \geq \frac{a^2}{2} - b^2$  that:

$$|J'_{\nu_k}(t)|^2 \geq \frac{1}{2} |J'_{\nu_{k-1+}}(t)|^2 - \mathfrak{C}^2 \varepsilon_{k-1}^2 \quad (42c)$$

Using (42b) together with (42c) leads to:

$$\begin{aligned} J(\nu_{k+}) - J(\nu_k) &\leq -\frac{3}{8} \alpha \int_{\Theta_{k-1+}} |J'_{\nu_{k-1+}}|^2 d\nu_k + \frac{3}{4} \alpha \mathfrak{C}^2 \varepsilon_{k-1}^2 \int_{\Theta_{k-1+}} d\nu_k, \\ &\leq -\frac{3}{32} \alpha v_{k-1+}^2 \int_{\Theta_{k-1+}} d\nu_k + \frac{3}{4} \alpha \mathfrak{C}^2 \varepsilon_{k-1}^2 \int_{\Theta_{k-1+}} d\nu_k, \\ &= -\frac{3}{8} \alpha \left[ \frac{v_{k-1+}^2}{4} - 2\varepsilon_{k-1}^2 \mathfrak{C}^2 \right] \int_{\Theta_{k-1+}} d\nu_k, \end{aligned}$$

where we used the definition of  $\Theta_{k-1+}$  to lower bound  $|J'_{\nu_{k-1+}}|^2$  in the first line. In particular, in the specific regime considered here, namely when  $24\varepsilon_{k-1}^2 \mathfrak{C}^2 \leq v_{k-1+}^2$ , we deduce that:

$$J(\nu_{k+}) - J(\nu_k) \leq -\frac{3}{2} \alpha v_{k-1+}^2 \int_{\Theta_{k-1+}} d\nu_k.$$

Finally, Assumption  $(\mathcal{H}_\varepsilon^+)$  yields:

$$J(\nu_{k+}) - J(\nu_k) \leq -\frac{3}{2} \alpha \varepsilon_{k-1} v_{k-1+}^2 \lambda(\Theta_{k-1+}), \quad (42d)$$

We are led to lower bound  $\lambda(\Theta_{k-1+})$ . To this end, we introduce the following geometric lemma:

**Lemma 7 (Geometric lower bound on the sub-level set)** *Let  $g : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathbb{R}$  be a  $\mathfrak{L}$ -Lipschitz function achieving a minimum value  $v^* = \min_{x \in \mathcal{X}} g(x) \leq 0$ . There exists a purely dimensional constant  $C_d > 0$  such that the Lebesgue measure of its sub-level sets satisfies:*

$$\lambda(\{x \in \mathcal{X} : g(x) \leq 0\}) \geq \lambda\left(\left\{x \in \mathcal{X} : g(x) \leq \frac{v^*}{2}\right\}\right) \geq C_d \mathfrak{L}^{-d} |v^*|^d.$$

Recall from Lemma A.5 that for any  $\nu \in \mathcal{M}(\mathcal{X})$ ,  $J'_\nu$  is a  $\mathfrak{L}(\nu)$ -Lipschitz function with  $\mathfrak{L}(\nu) \leq \mathfrak{L}$  as soon as  $\|\nu\|_{TV} \leq \mathfrak{C}_{TV}$  (see Remark 1.2). In the meantime, we know from Assumption (15- $(\mathcal{H}_{TV}^\infty)$ ) that  $(\nu_k)_{k \geq 1}$  and  $(\nu_{k+})_{k \geq 1}$  are two bounded sequences in terms of their TV-norm. Therefore,  $\mathfrak{L}(\nu_k) \leq \mathfrak{L}$ . By applying Lemma 7 to  $g = J'_{\nu_{k-1+}}$  and observing that  $v_{k-1+} = \min_x J'_{\nu_{k-1+}}(x) \leq 0$ , we have:

$$\begin{aligned} \forall x \in \mathcal{X} : \quad |x - \arg \min J'_{\nu_{k-1+}}| \leq \frac{|v_{k-1+}|}{2\mathfrak{L}} &\implies |J'_{\nu_{k-1+}}(x) - \min(J'_{\nu_{k-1+}})| \leq \mathfrak{L} \times \frac{|v_{k-1+}|}{2\mathfrak{L}}, \\ &\implies J'_{\nu_{k-1+}}(x) \leq \frac{v_{k-1+}}{2} \leq 0, \\ &\implies x \in \Theta_{k-1+}. \end{aligned}$$

We then deduce that

$$\left\{ x \in \mathcal{X} : |x - \arg \min J'_{\nu_{k-1+}}| \leq \frac{|v_{k-1+}|}{2\mathfrak{L}} \right\} \subset \Theta_{k-1+}.$$

This inclusion entails that the volume is bounded by that of a Euclidean ball of radius  $\frac{|v_{k-1+}|}{2\mathfrak{L}}$ , so for some constant  $C_d > 0$ :

$$\lambda(\Theta_{k-1+}) \geq C_d \mathfrak{L}^{-d} |v_{k-1+}|^d.$$

We finally obtain from the previous lower bound and from (42d) that:

$$J(\nu_{k+}) - J(\nu_k) \leq -\frac{3\alpha}{2} C_d \mathfrak{L}^{-d} |v_{k-1+}|^{2+d} \varepsilon_{k-1}.$$

This concludes the proof. ■

The following proposition describes the evolution of the cost function itself along the iterations. For this purpose, we introduce the non-negative sequence  $(\Delta_{k+})_{k \geq 1}$  and the auxiliary sequence  $(\Delta_k)_{k \geq 1}$  defined by

$$\forall k \geq 1, \quad \Delta_{k+} := J(\nu_k) - J(\nu_{k+}) \quad \text{and} \quad \Delta_k := J(\nu_{k-1}) - J(\nu_k). \quad (43)$$

These two sequences are then associated with our longitudinal evolution as follows.

$$\forall k \geq 1 \quad \underbrace{\nu_{k-1} \xrightarrow{\Delta_{k-1+}} \nu_{k-1+}}_{\Delta_k} \longrightarrow \underbrace{\nu_k \xrightarrow{\Delta_{k+}} \nu_{k+}}.$$

We emphasize that, from Proposition 2, the sequence  $(\Delta_{k+})_{k \geq 1}$  is non-negative. However, we cannot draw the same conclusion for the sequence  $(\Delta_k)_{k \geq 1}$ .

**Proposition 8** *Assuming (14) and (15- $(\mathcal{H}_{TV}^\infty)$ ), and that  $\alpha$  is chosen so that (7) holds, then*

$$\forall k \geq 1 \quad J(\nu_k) - J^* \leq \mathfrak{C} (\|\nu^*\|_{TV} \vee \mathfrak{L}) \max \left( \left[ \frac{\mathfrak{L}^d \Delta_{k+}}{\alpha \varepsilon_{k-1}} \right]^{\frac{1}{2+d}} ; \varepsilon_{k-1} \right).$$

The above result relates the evolution of the sequence  $J(\nu_k)$  to the increments  $J(\nu_k^+) - J(\nu_k)$ , and is primarily based on Proposition 6. This result involves a trade-off in the choice of the parameter  $\varepsilon_k$  (the amount of mass available in regions where  $J'_{\nu_k}$  is negative), which must be selected carefully. On the one hand,  $\varepsilon_k$  needs to be sufficiently large to control the increment of the objective function (see Proposition 6); on the other hand, they contribute an additional term that affects the value of  $J(\nu_k)$ .

**Proof** Using (4a), we have in a first time

$$J^* - J(\nu_k) = \int_{\mathcal{X}} J'_{\nu_k} d(\nu^* - \nu_k) + \frac{1}{2} \|\Phi(\nu_k - \nu^*)\|^2.$$

This implies

$$J(\nu_k) - J^* \leq \int_{\mathcal{X}} J'_{\nu_k} d(\nu_k - \nu^*) = \underbrace{\int_{\mathcal{X}} J'_{\nu_k} d\nu_k}_{:=A} - \underbrace{\int_{\mathcal{X}} J'_{\nu_k} d\nu^*}_{:=B}. \quad (44a)$$

Study of A: We can first observe that (42a) can be written as

$$\int_{\mathcal{X}} |J'_{\nu_k}|^2 d\nu_k \leq 4 \frac{J(\nu_k) - J(\nu_{k+})}{3\alpha} = \frac{4\Delta_{k+}}{3\alpha}. \quad (44b)$$

Then, the Cauchy-Schwarz inequality associated with Equation (44b) yields:

$$|A| = \left| \int_{\mathcal{X}} J'_{\nu_k} d\nu_k \right| \leq \left[ \|\nu_k\|_{\text{TV}} \int_{\mathcal{X}} |J'_{\nu_k}|^2 d\nu_k \right]^{1/2} \leq \left[ 4 \frac{\|\nu_k\|_{\text{TV}}}{3\alpha} \Delta_{k+} \right]^{1/2}. \quad (44c)$$

Study of B: We use the smoothness of  $\nu \mapsto J'_\nu$  induced by  $(\mathcal{H}_\varepsilon^{\text{smooth},2})$  and obtain that:

$$\begin{aligned} B &= \int_{\mathcal{X}} J'_{\nu_k} d\nu^* = \int_{\mathcal{X}} J'_{\nu_{k-1+}} d\nu^* + \int_{\mathcal{X}} (J'_{\nu_k} - J'_{\nu_{k-1+}}) d\nu^* \\ &\geq v_{k-1+} \|\nu^*\|_{\text{TV}} - \|\nu^*\|_{\text{TV}} \|J'_{\nu_k} - J'_{\nu_{k-1+}}\|_{\infty} \\ &\geq v_{k-1+} \|\nu^*\|_{\text{TV}} - \mathfrak{C}\varepsilon_{k-1} \|\nu^*\|_{\text{TV}}. \end{aligned}$$

At this stage, two distinct cases may arise depending on the value of  $v_{k-1+}$ .

- **1<sup>st</sup> case:**  $v_{k-1+} \leq -2\sqrt{6}\mathfrak{C}\varepsilon_{k-1}$ . Then, we get from the previous bound that

$$\begin{aligned} B &= \int_{\mathcal{X}} J'_{\nu_k} d\nu^* \geq v_{k-1+} \|\nu^*\|_{\text{TV}} - \mathfrak{C}\varepsilon_{k-1} \|\nu^*\|_{\text{TV}} \\ &\geq \left(1 + \frac{1}{2\sqrt{6}}\right) v_{k-1+} \|\nu^*\|_{\text{TV}} \\ &\geq - \left(1 + \frac{1}{2\sqrt{6}}\right) \|\nu^*\|_{\text{TV}} \left[ \frac{2\Delta_{k+} \mathfrak{L}^d}{3\alpha\varepsilon_{k-1} C_d} \right]^{\frac{1}{2+d}}, \end{aligned}$$

where the last line is obtained using Proposition 6. We deduce that:

$$\int_{\mathcal{X}} J'_{\nu_k} d\nu^* \geq - \left(1 + \frac{1}{2\sqrt{6}}\right) \|\nu^*\|_{\text{TV}} \left[ \frac{2\mathfrak{L}^d \Delta_{k+}}{3\alpha\varepsilon_{k-1} C_d} \right]^{\frac{1}{2+d}}.$$

- **2<sup>nd</sup> case:**  $v_{k-1+} \geq -2\sqrt{6}\mathfrak{C}\varepsilon_{k-1}$ . In such a situation we immediately have

$$\int_{\mathcal{X}} J'_{\nu_k} d\nu^* \geq v_{k-1+} \|\nu^*\|_{\text{TV}} - \mathfrak{C}\varepsilon_{k-1} \|\nu^*\|_{\text{TV}} \geq -(1 + 2\sqrt{6})\mathfrak{C} \|\nu^*\|_{\text{TV}} \varepsilon_{k-1}.$$

Regardless of the value of  $v_{k-1+}$ , we then get

$$\int_{\mathcal{X}} J'_{\nu_k} d\nu^* \geq -\|\nu^*\|_{\text{TV}} \min \left( (1 + 2\sqrt{6})\mathfrak{C}\varepsilon_{k-1}; \left(1 + \frac{1}{2\sqrt{6}}\right) \left[ \frac{2\mathfrak{L}^d \Delta_{k+}}{3\alpha\varepsilon_{k-1}C_d} \right]^{\frac{1}{2+d}} \right). \quad (44d)$$

Using Equations (44a), (44c) and (44d), we deduce that

$$J(\nu_k) - J^* \leq \mathfrak{C} (\|\nu^*\|_{\text{TV}} \vee \mathfrak{L}) \max \left( \left[ \frac{\mathfrak{L}^d \Delta_{k+}}{\alpha\varepsilon_{k-1}} \right]^{\frac{1}{2+d}}; \left[ \frac{\Delta_{k+}}{\alpha} \right]^{\frac{1}{2}}; \varepsilon_{k-1} \right).$$

We conclude while observing that  $(\Delta_{k+})_{k \geq 1}$  and  $(\varepsilon_k)_{k \geq 1}$  are two bounded sequences, which implies that:

$$J(\nu_k) - J^* \leq \mathfrak{C} (\|\nu^*\|_{\text{TV}} \vee \mathfrak{L}) \max \left( \alpha^{-1/2} \left[ \frac{\mathfrak{L}^d \Delta_{k+}}{\varepsilon_{k-1}} \right]^{\frac{1}{2+d}}; \varepsilon_{k-1} \right).$$

■

#### C.4 Proof of the deterministic global convergence

Below, we finally provide the proof of our global convergence result in the deterministic situation, stated in Theorem 2.1. We will use the key property obtained in Section C.3.

**Proof** [Proof of Theorem 2.1] We introduce the function  $f : \mathbb{N} \rightarrow \mathbb{R}^+$  defined by  $f(k) = J(\nu_k) - J^*$  for all  $k \in \mathbb{N}$ . The triangle inequality and Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$  yields

$$\Delta_{k+1} = J(\nu_k) - J(\nu_{k+1}) = J(\nu_k) - J(\nu_{k+}) + J(\nu_{k+}) - J(\nu_{k+1}) \geq \Delta_{k+} - \mathfrak{C}\varepsilon_k^2. \quad (45a)$$

Using Equation (45a) and the definition of  $f$ , we then obtain:

$$f(k) - f(k+1) = \Delta_{k+1} \geq \Delta_{k+} - \mathfrak{C}\varepsilon_k^2. \quad (45b)$$

Simultaneously, if  $A = \mathfrak{C} (\|\nu^*\|_{\text{TV}} \vee \mathfrak{L})$ , then we can apply Proposition 8 to get

$$[A^{-1}f(k)]^{2+d} \leq \max \left( \frac{\mathfrak{L}^d \Delta_{k+}}{\alpha\varepsilon_{k-1}}, \varepsilon_{k-1}^{2+d} \right) \leq \frac{\mathfrak{L}^d \Delta_{k+}}{\alpha\varepsilon_{k-1}} + \varepsilon_{k-1}^{2+d}.$$

It leads to

$$\Delta_{k+} \geq \alpha \mathfrak{L}^{-d} \varepsilon_{k-1} [A^{-1}f(k)]^{2+d} - \alpha \mathfrak{L}^{-d} \varepsilon_{k-1}^{3+d} \geq \alpha \mathfrak{L}^{-d} \varepsilon_{k-1} [A^{-1}f(k)]^{2+d} - \alpha \mathfrak{L}^{-d} \varepsilon_{k-1}^2, \quad (45c)$$

where the last inequality comes from the bound  $\varepsilon_k \leq 1$ . Hence, using together (45b) and (45c), we get

$$f(k) - f(k+1) \geq \alpha \mathfrak{L}^{-d} \varepsilon_{k-1} [A^{-1} f(k)]^{2+d} - (1 + \alpha \mathfrak{L}^{-d}) \varepsilon_{k-1}^2.$$

We use a telescopic sum argument (summing for  $k = 1, \dots, K-1$ ), leading to the inequality

$$f(1) - f(K) \geq \alpha \mathfrak{L}^{-d} A^{-2-d} \sum_{k=1}^{K-1} \varepsilon_{k-1} f(k)^{2+d} - (1 + \alpha \mathfrak{L}^{-d}) \sum_{k=1}^{K-1} \varepsilon_{k-1}^2,$$

which in turn implies

$$f(K) + \alpha \mathfrak{L}^{-d} A^{-2-d} \sum_{k=1}^{K-1} \varepsilon_{k-1} f(k)^{2+d} \leq f(1) + (1 + \alpha \mathfrak{L}^{-d}) \sum_{k=1}^{K-1} \varepsilon_{k-1}^2. \quad (45d)$$

Proof of *i*) and *ii*): We introduce in the following the quantity  $\rho_K$ , defined as the minimum value of the sequence  $(J(\nu_k) - J(\nu^*))_{1 \leq k \leq K}$  over the first  $K$  iterations:

$$\forall K \in \mathbb{N} \quad \rho_K = \min_{0 \leq k \leq K} J(\nu_k) - J(\nu^*) = \min_{0 \leq k \leq K} f(k).$$

Thanks to the definition of  $\rho_K$ , (45d) implies that

$$\rho_K^{2+d} \alpha \mathfrak{L}^{-d} A^{-2-d} \sum_{k=1}^{K-1} \varepsilon_{k-1} \leq \left( f(1) + (1 + \alpha \mathfrak{L}^{-d}) \sum_{k=1}^{K-1} \varepsilon_{k-1}^2 \right),$$

which can be rewritten as

$$\rho_K \leq \mathfrak{e} \left[ \frac{1 + (1 + \alpha \mathfrak{L}^{-d}) \sum_{k=1}^{K-1} \varepsilon_{k-1}^2}{\alpha \mathfrak{L}^{-d} A^{-2-d} \sum_{k=1}^{K-1} \varepsilon_{k-1}} \right]^{\frac{1}{2+d}}. \quad (45e)$$

Starting from (45e), we now consider two different cases.

• *Horizon dependent step-size sequence i)* Considering the case of a constant step-size sequence  $(\varepsilon_k)_{k \geq 0}$  with  $\varepsilon_k = \varepsilon$  for all  $k \in \mathbb{N}$ , we deduce from (45e) that in this specific case

$$\forall K \geq 0 \quad \rho_K \leq \mathfrak{e} \frac{A \mathfrak{L}^{\frac{d}{2+d}}}{\alpha^{\frac{1}{2+d}}} \left( \frac{1}{(K-1)\varepsilon} + \varepsilon(1 + \alpha \mathfrak{L}^{-d}) \right)^{\frac{1}{2+d}}.$$

It remains to optimize the previous upper bound in terms of  $\varepsilon$ . The trade-off between the two terms appearing in the r.h.s. of the previous bound is attained for

$$\varepsilon = \mathfrak{e} \sqrt{\frac{1}{(K-1)(1 + \alpha \mathfrak{L}^{-d})}},$$

which yields

$$\rho_K \leq \mathfrak{C} \frac{A\mathfrak{L}^{\frac{d}{2+d}}}{\alpha^{\frac{1}{2+d}}} \left( \frac{(1 + \alpha\mathfrak{L}^{-d})}{(K-1)} \right)^{\frac{1}{2(2+d)}}. \quad (45f)$$

• *Horizon-free step-size sequence ii)* It is also possible to derive a convergence rate with a horizon-free step-size sequence that does not depend on the horizon of the simulation. For this purpose, we simply consider the sequence:

$$\forall k \geq 0 \quad \varepsilon_k = \frac{\mathfrak{C}}{\sqrt{(k+1)}}.$$

In this case, we verify that:

$$\sum_{k=1}^{K-1} \varepsilon_k^2 \leq \mathfrak{C} \sum_{k=1}^{K-1} \frac{1}{k} \leq \mathfrak{C}[\log(K) + 1] \quad \text{and} \quad \sum_{k=1}^{K-1} \varepsilon_k = \sum_{k=1}^K \frac{\mathfrak{C}}{\sqrt{k}} \geq 2\mathfrak{C}(\sqrt{K} - 1).$$

Then

$$\rho_K \leq \mathfrak{C} \alpha^{-\frac{1}{(2+d)}} \left( \frac{1}{K} \right)^{\frac{1}{2(2+d)}} \log^{\frac{1}{2+d}}(K).$$

Proof of iii): We first introduce the function  $\bar{f} : \mathbb{N} \rightarrow \mathbb{R}$  defined as

$$\bar{f}(K) = \max \left( J(\nu_K) - J^* - \mathfrak{C} \sum_{k=1}^{K-1} \varepsilon_k^2 ; 0 \right) \quad \forall K \in \mathbb{N}, \quad (45g)$$

where  $\mathfrak{C}$  is here related to Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$  (see (14a- $(\mathcal{H}_\varepsilon^{\text{smooth},1})$ )). This function is non-increasing as, for any  $K \in \mathbb{N}$ ,

$$\begin{aligned} \bar{f}(K+1) &= \max \left( J(\nu_{K+1}) - J^* - \mathfrak{C} \sum_{k=1}^K \varepsilon_k^2 ; 0 \right) \\ &= \max \left( J(\nu_{K+1}) - J(\nu_{K+}) + J(\nu_{K+}) - J(\nu_K) + J(\nu_K) - J^* - \mathfrak{C} \sum_{k=1}^K \varepsilon_k^2 ; 0 \right) \\ &\leq \max \left( \mathfrak{C}\varepsilon_K^2 + J(\nu_K) - J^* - \mathfrak{C} \sum_{k=1}^K \varepsilon_k^2 ; 0 \right) \\ &= \bar{f}(K) \end{aligned}$$

where we applied Assumption  $(\mathcal{H}_\varepsilon)$  in the third line and the fact that  $J(\nu_{k+}) - J(\nu_k) \leq 0$  (see Proposition 2). We first assume that

$$\bar{f}(K) \geq 0. \quad (45h)$$

Remark that (45h) together with the non-increasing property of  $\bar{f}$  entails that  $\bar{f}(k) \geq 0$  for any  $k \in \{1, \dots, K\}$ . We use the discrete integration by part relationship: for  $p = \bar{f}(k+1)^{-1}$  and  $q = \bar{f}(k)^{-1}$ ,

$$p^{1+d} - q^{1+d} = (p - q) \sum_{i=0}^d p^i q^{d-i}. \quad (45i)$$

The relationship  $\bar{f}(k) \geq \bar{f}(k+1)$  then provides, since every term  $p^i q^{d-i}$  in (45i) is bounded below by  $q^d = \bar{f}(k)^{-d}$  (resp. by  $p q^d$  for the  $i \geq 1$  terms),

$$\frac{1}{\bar{f}(k+1)^{1+d}} - \frac{1}{\bar{f}(k)^{1+d}} \geq (\bar{f}(k) - \bar{f}(k+1)) \frac{d+1}{\bar{f}(k+1)\bar{f}(k)^{1+d}}. \quad (45j)$$

Using (45h), we have

$$\begin{aligned} \bar{f}(k) - \bar{f}(k+1) &\geq J(\nu_k) - J(\nu_k^+) + J(\nu_k^+) - J(\nu_{k+1}) + \mathfrak{C}\varepsilon_k^2, \\ &= \Delta_{k^+} + J(\nu_k^+) - J(\nu_{k+1}) + \mathfrak{C}\varepsilon_k^2, \\ &\geq \Delta_{k^+}, \end{aligned} \quad (45k)$$

since

$$J(\nu_k^+) - J(\nu_{k+1}) \geq -\mathfrak{C}\varepsilon_k^2,$$

according to Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$ . Hence, (45j) together with (45k) leads to

$$\frac{1}{\bar{f}(k+1)^{1+d}} - \frac{1}{\bar{f}(k)^{1+d}} \geq \Delta_{k^+} \frac{d+1}{\bar{f}(k+1)\bar{f}(k)^{1+d}} \geq \Delta_{k^+} \frac{d+1}{\bar{f}(k)^{2+d}}, \quad (45l)$$

where the last inequality uses  $\bar{f}(k+1) \leq \bar{f}(k)$ . Since for all  $k \geq 0$ :  $\bar{f}(k) \leq J(\nu_k) - J^* = f(k)$ , (45c) entails that

$$\Delta_{k^+} \geq \alpha \varepsilon_{k-1} \mathfrak{L}^{-d} [A^{-1} \bar{f}(k)]^{2+d} - \alpha \mathfrak{L}^{-d} \varepsilon_{k-1}^2, \quad (45m)$$

This last inequality, together with (45l), leads to

$$\begin{aligned} \frac{1}{\bar{f}(k+1)^{1+d}} - \frac{1}{\bar{f}(k)^{1+d}} &\geq \left( \alpha \mathfrak{L}^{-d} \varepsilon_{k-1} [A^{-1} \bar{f}(k)]^{2+d} - \alpha \mathfrak{L}^{-d} \varepsilon_{k-1}^2 \right) \frac{d+1}{\bar{f}(k)^{2+d}}, \\ &= \left( \alpha \mathfrak{L}^{-d} \varepsilon_{k-1} A^{-(2+d)} - \alpha \mathfrak{L}^{-d} \varepsilon_{k-1}^2 \frac{1}{\bar{f}(k)^{2+d}} \right) (d+1). \end{aligned}$$

We then use a telescopic sum argument to obtain:

$$\begin{aligned} \frac{1}{\bar{f}(K)^{1+d}} - \frac{1}{\bar{f}(1)^{1+d}} &\geq \mathfrak{C}(d+1) \left[ \alpha \mathfrak{L}^{-d} A^{-(2+d)} \sum_{k=1}^{K-1} \varepsilon_{k-1} - \alpha \mathfrak{L}^{-d} \sum_{k=1}^{K-1} \varepsilon_{k-1}^2 \frac{1}{\bar{f}(k)^{2+d}} \right], \\ &\geq \mathfrak{C}(d+1) \mathfrak{L}^{-d} \left[ \alpha A^{-(2+d)} \sum_{k=1}^{K-1} \varepsilon_{k-1} - \alpha \frac{1}{\bar{f}(K)^{2+d}} \sum_{k=1}^{K-1} \varepsilon_{k-1}^2 \right], \end{aligned}$$

since  $\bar{f}$  is non-increasing. This last inequality can be re-written as:

$$\alpha \mathfrak{L}^{-d} (d+1) \frac{1}{\bar{f}(K)^{2+d}} \sum_{k=1}^{K-1} \varepsilon_{k-1}^2 + \frac{1}{\bar{f}(K)^{1+d}} \geq (d+1) \alpha \mathfrak{L}^{-d} A^{-(2+d)} \sum_{k=1}^{K-1} \varepsilon_{k-1} + \frac{1}{\bar{f}(1)^{1+d}}.$$

Using again the monotonicity of  $\bar{f}$ , we deduce that:

$$\begin{aligned} \frac{1}{\bar{f}(K)^{2+d}} \left[ \alpha \mathfrak{L}^{-d} (d+1) \sum_{k=1}^{K-1} \varepsilon_{k-1}^2 + \bar{f}(1) \right] \\ \geq \mathfrak{C}(d+1) \mathfrak{L}^{-d} \alpha A^{-(2+d)} \sum_{k=1}^{K-1} \varepsilon_{k-1} + \frac{1}{\bar{f}(1)^{1+d}}, \end{aligned}$$

where the monotonicity  $\bar{f}(K) \leq \bar{f}(1)$  was used to upper bound  $\bar{f}(K)^{-(1+d)}$  by  $\bar{f}(1) \bar{f}(K)^{-(2+d)}$  (the finiteness of  $\bar{f}(1) = J(\nu_1) - J^*$  itself being ensured by  $(\mathcal{H}_{\text{TV}}^\infty)$ ). The last inequality can be rewritten as

$$\bar{f}(K) \leq \mathfrak{C} \left[ \frac{\alpha \mathfrak{L}^{-d} (d+1) \sum_{k=1}^{K-1} \varepsilon_{k-1}^2 + \bar{f}(1)}{(d+1) \alpha \mathfrak{L}^{-d} A^{-(2+d)} \sum_{k=1}^{K-1} \varepsilon_{k-1} + \bar{f}(1)^{-(1+d)}} \right]^{\frac{1}{2+d}}.$$

Since  $\bar{f}(1) = J(\nu_1) - J^* \geq 0$ , this leads to

$$J(\nu_K) - J^* \leq \mathfrak{C} \left[ \frac{\alpha \mathfrak{L}^{-d} (d+1) \sum_{k=1}^{K-1} \varepsilon_{k-1}^2 + (J(\nu_1) - J^*)}{(d+1) \alpha \mathfrak{L}^{-d} A^{-(2+d)} \sum_{k=1}^{K-1} \varepsilon_{k-1}} \right]^{\frac{1}{2+d}} + \mathfrak{C} \sum_{k=1}^{K-1} \varepsilon_k^2. \quad (45n)$$

In this context, we consider a horizon-dependent strategy, namely we set  $\varepsilon_k = \varepsilon$  for any  $k \in \{1, \dots, K\}$ . The bound (45n) becomes in this case

$$\begin{aligned} J(\nu_K) - J^* &\leq \mathfrak{C} \left[ \frac{\alpha \mathfrak{L}^{-d} (d+1) (K-1) \varepsilon^2 + (J(\nu_1) - J^*)}{(d+1) \alpha \mathfrak{L}^{-d} A^{-(2+d)} (K-1) \varepsilon} \right]^{\frac{1}{2+d}} + \mathfrak{C} K \varepsilon^2, \\ &\leq \mathfrak{C} \left[ \frac{1}{A^{-(2+d)}} \varepsilon + \frac{J(\nu_1) - J^*}{(d+1) \mathfrak{L}^{-d} \alpha A^{-(2+d)}} \times \frac{1}{(K-1) \varepsilon} \right]^{\frac{1}{2+d}} + \mathfrak{C} K \varepsilon^2 \\ &\leq \mathfrak{C} A \left[ \varepsilon + \frac{J(\nu_1) - J^*}{(d+1) \mathfrak{L}^{-d} \alpha (K-1) \varepsilon} \right]^{\frac{1}{2+d}} + \mathfrak{C} K \varepsilon^2. \end{aligned}$$

Then, choosing  $\varepsilon$  such that

$$\left( \frac{1}{(d+1) \alpha \mathfrak{L}^{-d} A^{-(2+d)}} \times \frac{1}{K \varepsilon} \right)^{\frac{1}{2+d}} = K \varepsilon^2 \quad \Leftrightarrow \quad \varepsilon = \left( \frac{A^{2+d}}{(d+1) \alpha \mathfrak{L}^{-d}} \right)^{\frac{1}{5+2d}} K^{-\frac{3+d}{5+2d}},$$

and defining  $\mathfrak{C}$  large enough to absorb  $(J(\nu_1) - J^*)^{\frac{1}{2+d}}$ , we obtain that (the constant  $\mathfrak{C}$  depends polynomially on the initial excess, specifically as  $(J(\nu_1) - J^*)^{1/(2+d)}$ ):

$$J(\nu_K) - J^* \leq \mathfrak{C} \left( \frac{\mathfrak{L}^d A^{2+d}}{(d+1) \alpha} \right)^{\frac{2}{5+2d}} K^{-\frac{1}{5+2d}}.$$

To conclude the proof, we have to investigate the case where (45h) does not hold. Remark that the latter entails that

$$J(\nu_K) - J^* \leq \mathfrak{C} \sum_{k=1}^K \varepsilon_k^2 \leq \mathfrak{C} K \varepsilon^2 \leq \mathfrak{C} \left( \frac{\mathfrak{L}^d A^{2+d}}{(d+1) \alpha} \right)^{\frac{2}{5+2d}} K^{-\frac{1}{5+2d}},$$

keeping the same choice for  $\varepsilon$ . We then use  $A = \mathfrak{C} (\|\nu^*\|_{\text{TV}} \vee \mathfrak{L})$  in the final results.  $\blacksquare$

## Appendix D. Proof of the stochastic results

### D.1 Almost sure total variation bound

We begin with the study of the almost sure TV norm upper bound, that will be then used throughout the rest of the proofs.

**Proposition 9** *Define*

$$\widehat{\mathfrak{K}} = \frac{\mathbf{H}}{\mathbf{G}}e + \sqrt{\frac{e^3}{\mathbf{G}}} + 1$$

and assume that  $\alpha \leq \frac{1}{1+\widehat{\mathfrak{K}}}$  and  $\varepsilon_k \leq \alpha$  for all  $k \geq 1$ , then the sequence  $(\hat{\nu}_k)_{k \geq 0}$  satisfies:

$$\forall k \geq 0 \quad \|\hat{\nu}_k\|_{TV} \leq \widehat{\mathfrak{K}} \quad a.s.$$

**Proof** Our proof follows essentially the same lines as the deterministic case, except that we have to take into account the randomness of our updates. Let  $k \in \mathbb{N}^*$  be fixed. According to (25a) and (25b), we have

$$\hat{\nu}_{k+1} = \hat{\nu}_{k+} - \mathbf{1}_{\widehat{\mathcal{P}}_{\nu_{k+}}} (V_{k+1})\hat{\nu}_{k+}(V_{k+1})\delta_{V_{k+1}} + \varepsilon_k \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}} (U_{k+1})\delta_{U_{k+1}}$$

Computing the total variation norm, we obtain:

$$\|\hat{\nu}_{k+1}\|_{TV} \leq \|\hat{\nu}_{k+}\|_{TV} + \varepsilon_k \leq \int_{\mathcal{X}} e^{-\alpha \widehat{J}'_{\hat{\nu}_k}(t)} d\nu_k(t) + \|\varepsilon\|_{\infty}.$$

Next, observe that for any  $t \in \mathcal{X}$ , according to Assumption (9b), we have

$$\widehat{J}'_{\hat{\nu}_k}(t) = \frac{1}{m_k} \sum_{l=1}^{m_k} \widehat{J}'_{\hat{\nu}_k}(t, Z_{l,k}^+) \geq \mathbf{G} \|\hat{\nu}_k\|_{TV} - \mathbf{H} + \kappa \geq -\mathbf{H} + \kappa \quad a.s.,$$

so that we have the almost sure upper bound:

$$\|\hat{\nu}_{k+1}\|_{TV} \leq e^{-\alpha \mathbf{G} \|\hat{\nu}_k\|_{TV} + \alpha \mathbf{H}} \|\hat{\nu}_k\|_{TV} + \alpha \quad a.s.$$

Then, the rest of the proof proceeds exactly following the same lines as in Proposition 3, i), whose proof is located in Section C.1.  $\blacksquare$

### D.2 Proof of Proposition 4

The proof of Proposition 4 is split into several parts, following all the assumptions we need to verify accordingly. In particular, we establish below Propositions 10, 12, 11 and 13.

**Proposition 10 (Assumption  $\mathcal{H}_{\varepsilon,a}^+$ )** *For any  $a > 0$ , set  $c_a = \mathbf{E}_{\infty} \sqrt{2a}$  in Equation (24). Then for any integer  $k$ , the iterate  $(\hat{\nu}_{k+}, \hat{\nu}_{k+1})$  of Algorithm 1 satisfies for any  $A \subset \mathcal{X}$ :*

$$\mathbb{E} \left[ \hat{\nu}_{k+1}(A \cap \{J'_{\hat{\nu}_{k+}} < 0\}) \mid \mathfrak{F}_k^+ \right] \geq \varepsilon_k \frac{\lambda(A \cap \{J'_{\hat{\nu}_{k+}} < 0\})}{\lambda(\mathcal{X})} - \varepsilon_k m_k^{-a}$$

**Proof** Consider any integer  $k \in \mathbb{N}$  and any measurable set  $A \subset \mathcal{X}$ , we use the definition of  $\hat{\nu}_{k+1}$ :

$$\begin{aligned} & \mathbb{E} \left[ \hat{\nu}_{k+1}(A \cap \{J'_{\hat{\nu}_{k+}} < 0\}) \mid \mathfrak{F}_k^+ \right] \\ &= \hat{\nu}_k^{++}(A \cap \{J'_{\hat{\nu}_{k+}} < 0\}) + \varepsilon_k \mathbb{E} \left[ \mathbf{1}_{\hat{\mathcal{N}}_{\nu_{k+}}} (U_{k+1}) \delta_{U_{k+1}}(A \cap \{J'_{\hat{\nu}_{k+}} < 0\}) \mid \mathfrak{F}_k^+ \right], \\ &\geq \varepsilon_k \int_{A \cap \{J'_{\hat{\nu}_{k+}} < 0\}} \lambda(\mathcal{X})^{-1} dx - \varepsilon_k \mathbb{E} \left[ \int_{A \cap \{J'_{\hat{\nu}_{k+}} < 0\}} \mathbf{1}_{\hat{\mathcal{N}}_{\nu_{k+}}} (x) \lambda(\mathcal{X})^{-1} dx \right], \\ &\geq \varepsilon_k \frac{\lambda(A \cap \{J'_{\hat{\nu}_{k+}} < 0\})}{\lambda(\mathcal{X})} - \varepsilon_k \mathbb{P} \left( J'_{\hat{\nu}_{k+}}(U_{k+1}) < 0 \text{ and } \hat{J}'_{\hat{\nu}_{k+}}(U_{k+1}) > c_a \sqrt{\frac{\log m_k}{m_k}} \right), \end{aligned}$$

where we used the definition of  $\hat{\mathcal{N}}_{\nu_{k+}}$ . Recall in particular that the constant  $c_a$  is positive. The key observation is that the sampled point  $U_{k+1} \sim \text{Uniform}(\mathcal{X})$  is *independent* of the mini-batch  $Z_{k+1}^+$  used to construct the stochastic certificate  $\hat{J}'_{\hat{\nu}_{k+}}$ . Conditioning on  $U_{k+1} = x$  reduces the problem to bounding the deviation of an empirical mean at a single fixed point, where standard Hoeffding's inequality applies without any covering argument. Specifically, for any  $x \in \mathcal{X}$ :

$$\mathbb{P} \left( \hat{J}'_{\hat{\nu}_{k+}}(x) > c_a \sqrt{\frac{\log m_k}{m_k}} \mid U_{k+1} = x \right) \leq \exp \left( -\frac{m_k}{2\mathbf{E}_\infty^2} \left( c_a \sqrt{\frac{\log m_k}{m_k}} \right)^2 \right) = m_k^{-a}. \quad (46)$$

Integrating over  $U_{k+1}$  and setting  $a = \frac{c_a^2}{2\mathbf{E}_\infty^2}$ , we obtain:

$$\mathbb{E} \left[ \hat{\nu}_{k+1}(A \cap \{J'_{\hat{\nu}_{k+}} < 0\}) \mid \mathfrak{F}_k^+ \right] \geq \varepsilon_k \frac{\lambda(A \cap \{J'_{\hat{\nu}_{k+}} < 0\})}{\lambda(\mathcal{X})} - \varepsilon_k m_k^{-a},$$

with  $a = \frac{c_a^2}{2\mathbf{E}_\infty^2}$ . No covering of  $\mathcal{X}$  is needed, so the dimension  $d$  disappears from the birth threshold. This ensures that  $\hat{\mathcal{H}}_{\varepsilon, a}^+$  holds with  $\mathfrak{c} = \lambda(\mathcal{X})^{-1}$  and  $\mathfrak{C} = 1$ .  $\blacksquare$

**Proposition 11 (Assumption  $\hat{\mathcal{H}}_\varepsilon^{\text{smooth}, 2}$ .)** For any integer  $k$ , the iterate  $(\hat{\nu}_k^{++}, \hat{\nu}_{k+1})$  of Algorithm 1 satisfies:

$$\|J'_{\hat{\nu}_{k+1}} - J'_{\hat{\nu}_{k+}}\|_\infty \leq \mathfrak{C}\varepsilon_k.$$

**Proof** Using (4b), we have, for any  $t \in \mathcal{X}$ ,

$$J'_{\hat{\nu}_{k+1}}(t) - J'_{\hat{\nu}_{k+}}(t) = \langle \varphi_t, \Phi(\hat{\nu}_{k+1} - \hat{\nu}_{k+}) \rangle_{\mathbb{H}}.$$

Hence, according to our update scheme,

$$\sup_{t \in \mathcal{X}} |J'_{\hat{\nu}_{k+1}}(t) - J'_{\hat{\nu}_{k+}}(t)| = \sup_{t \in \mathcal{X}} |\langle \varphi_t, \Phi \delta_{V_{k+1}} \rangle_{\mathbb{H}} \hat{\nu}_{k+}(V_{k+1}) + \langle \varphi_t, \Phi \delta_{U_{k+1}} \rangle_{\mathbb{H}} \varepsilon_k| \leq \sqrt{2}\varepsilon_k + \varepsilon_k \leq \mathfrak{C}\varepsilon_k,$$

where we have used the Cauchy-Schwarz Inequality and Assumption  $(\mathcal{H}_P)$ .  $\blacksquare$

**Proposition 12 (Assumption  $\mathcal{H}_\varepsilon^{\text{smooth},1}$ )** For any integer  $k$ , the iterate  $(\hat{\nu}_k^{++}, \hat{\nu}_{k+1})$  of Algorithm 1 satisfies:

$$\mathbb{E} [J(\hat{\nu}_{k+1}) - J(\hat{\nu}_{k+}) | \mathfrak{F}_k^+] \leq \mathfrak{C} \left( \varepsilon_k \sqrt{\frac{\log m_k}{m_k}} + \varepsilon_k^2 \right).$$

**Proof** The proof is inspired by the deterministic case, see Section C.2, but we still need to handle the randomness brought by  $\widehat{\mathcal{P}}_{\nu_{k+}}$  and  $\widehat{\mathcal{N}}_{\nu_{k+}}$ . We decompose the evolution of  $J$  into two terms:

$$J(\hat{\nu}_{k+1}) - J(\hat{\nu}_{k+}) = J(\hat{\nu}_{k+1}) - J(\hat{\nu}_k^{++}) + J(\hat{\nu}_k^{++}) - J(\hat{\nu}_{k+}).$$

**Death part:**  $J(\hat{\nu}_k^{++}) - J(\hat{\nu}_{k+})$ . Using  $\hat{\nu}_k^{++} - \hat{\nu}_{k+} = -\hat{\nu}_{k+}(V_{k+1})\delta_{V_{k+1}}\mathbf{1}_{\widehat{\mathcal{P}}_{\nu_{k+}}}(V_{k+1})$ :

$$\begin{aligned} & J(\hat{\nu}_k^{++}) - J(\hat{\nu}_{k+}) \\ &= \int J'_{\hat{\nu}_k^{++}} d[\hat{\nu}_k^{++} - \hat{\nu}_{k+}] + \frac{1}{2} \|\Phi(\hat{\nu}_k^{++} - \hat{\nu}_{k+})\|_{\mathbb{H}}^2 \\ &= \left[ -\widehat{\mathcal{J}}'_{\hat{\nu}_{k+}}(V_{k+1})\hat{\nu}_{k+}(V_{k+1}) + (\widehat{\mathcal{J}}'_{\hat{\nu}_{k+}} - J'_{\hat{\nu}_{k+}})(V_{k+1})\hat{\nu}_{k+}(V_{k+1}) + \frac{1}{2}\hat{\nu}_{k+}(V_{k+1})^2 \|\Phi\delta_{V_{k+1}}\|_{\mathbb{H}}^2 \right] \\ &\quad \times \mathbf{1}_{\widehat{\mathcal{P}}_{\nu_{k+}}}(V_{k+1}) \\ &\leq \varepsilon_k^2 + \sqrt{2}\varepsilon_k \max_{1 \leq j \leq p_k} \left| \widehat{\mathcal{J}}'_{\hat{\nu}_{k+}}(\hat{t}_j^{k+}) - J'_{\hat{\nu}_{k+}}(\hat{t}_j^{k+}) \right|, \end{aligned}$$

since on  $\widehat{\mathcal{P}}_{\nu_{k+}}$  we have  $\widehat{\mathcal{J}}'_{\hat{\nu}_{k+}}(V_{k+1}) \geq 0$  (so the first term is  $\leq 0$  and dropped),  $\hat{\nu}_{k+}(V_{k+1}) \leq \sqrt{2}\varepsilon_k$ , and  $\|\Phi\delta_{V_{k+1}}\|_{\mathbb{H}}^2 \leq 1$ . The death process only evaluates  $\widehat{\mathcal{J}}'_{\hat{\nu}_{k+}}$  at the  $p_k$  active particle locations  $(\hat{t}_j^{k+})_{1 \leq j \leq p_k}$ , which are  $\mathfrak{F}_k^+$ -measurable. By Hoeffding's inequality applied at each fixed particle position and a sub-Gaussian maximal bound over the  $p_k \leq m_k + p_0$  points:

$$\mathbb{E} \left[ \max_{1 \leq j \leq p_k} \left| \widehat{\mathcal{J}}'_{\hat{\nu}_{k+}}(\hat{t}_j^{k+}) - J'_{\hat{\nu}_{k+}}(\hat{t}_j^{k+}) \right| \middle| \mathfrak{F}_k^+ \right] \leq \mathfrak{C} \sqrt{\frac{\log m_k}{m_k}},$$

since  $\log(p_k) \leq \log(m_k + p_0) = \mathcal{O}(\log m_k)$ . Hence:

$$\mathbb{E} [J(\hat{\nu}_k^{++}) - J(\hat{\nu}_{k+}) | \mathfrak{F}_k^+] \leq \varepsilon_k^2 + \mathfrak{C}\varepsilon_k \sqrt{\frac{\log m_k}{m_k}}.$$

**Birth part:**  $J(\hat{\nu}_{k+1}) - J(\hat{\nu}_k^{++})$ . Since  $\hat{\nu}_{k+1} - \hat{\nu}_k^{++} = \varepsilon_k \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}}(U_{k+1}) \delta_{U_{k+1}}$ :

$$\begin{aligned} & J(\hat{\nu}_{k+1}) - J(\hat{\nu}_k^{++}) \\ &= \int J'_{\hat{\nu}_k^{++}} d[\hat{\nu}_{k+1} - \hat{\nu}_k^{++}] + \frac{1}{2} \|\Phi(\hat{\nu}_{k+1} - \hat{\nu}_k^{++})\|_{\mathbb{H}}^2 \\ &= \varepsilon_k \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}}(U_{k+1}) \left[ \widehat{\mathcal{J}}'_{\hat{\nu}_{k+}}(U_{k+1}) + (J'_{\hat{\nu}_{k+}} - \widehat{\mathcal{J}}'_{\hat{\nu}_{k+}})(U_{k+1}) + (J'_{\hat{\nu}_k^{++}} - J'_{\hat{\nu}_{k+}})(U_{k+1}) \right] \\ &\quad + \frac{\varepsilon_k^2}{2} \|\Phi\delta_{U_{k+1}}\|_{\mathbb{H}}^2 \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}}(U_{k+1}). \end{aligned}$$

We bound each contribution. By the definition of  $\widehat{\mathcal{N}}_{\nu_{k+}}$  in (24), it holds that

$$\widehat{J}'_{\widehat{\nu}_{k+}}(U_{k+1}) \mathbf{1}_{\widehat{\mathcal{N}}_{\nu_{k+}}}(U_{k+1}) \leq c_a \sqrt{(\log m_k)/m_k}.$$

Proposition 11 yields  $\|J'_{\widehat{\nu}_k^{++}} - J'_{\widehat{\nu}_k^+}\|_\infty \leq \mathfrak{C}\varepsilon_k$ , and  $\|\Phi\delta_{U_{k+1}}\|_{\mathbb{H}}^2 \leq 1$ . Since  $U_{k+1} \sim \text{Uniform}(\mathcal{X})$  is independent of the mini-batch, conditioning on  $U_{k+1} = x$  and applying pointwise Hoeffding gives

$$\mathbb{E}[\widehat{J}'_{\widehat{\nu}_{k+}}(U_{k+1}) - J'_{\widehat{\nu}_k^+}(U_{k+1}) \mid \mathfrak{F}_k^+] \leq \frac{\mathfrak{C}}{\sqrt{m_k}}.$$

Taking the conditional expectation and using  $1/\sqrt{m_k} \leq \sqrt{(\log m_k)/m_k}$  for  $m_k \geq 3$ :

$$\mathbb{E}[J(\widehat{\nu}_{k+1}) - J(\widehat{\nu}_k^{++}) \mid \mathfrak{F}_k^+] \leq \mathfrak{C} \left( \varepsilon_k \sqrt{\frac{\log m_k}{m_k}} + \varepsilon_k^2 \right).$$

The finitely many iterates with  $m_k \in \{1, 2\}$  contribute an  $O(1)$  constant absorbed into  $\mathfrak{C}$ . Combining the birth and death parts establishes  $\widehat{\mathcal{H}}_\varepsilon^{\text{smooth}, 1}$ .  $\blacksquare$

**Proposition 13 (Assumption  $(\widehat{\mathcal{H}}_D)$ )** *There exists a large enough constant  $\mathfrak{C}$  such that, for any integer  $k$ , if  $\alpha \leq (4\mathfrak{C})^{-1} \wedge \sqrt{8 \log 8} \mathbf{E}_\infty^{-1}$ , then:*

$$\mathbb{E}[J(\widehat{\nu}_{k+}) \mid \mathfrak{F}_k] \leq J(\widehat{\nu}_k) - \frac{\alpha}{2} \|J'_{\widehat{\nu}_k}\|_{\widehat{\nu}_k}^2 + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right).$$

**Proof** Step 1: One-step evolution and second order term. Let  $k \in \mathbb{N}^*$  be fixed. According to (4a):

$$J(\widehat{\nu}_{k+}) - J(\widehat{\nu}_k) = \int_{\mathcal{X}} J'_{\widehat{\nu}_k} d(\widehat{\nu}_{k+} - \widehat{\nu}_k) + \frac{1}{2} \|\Phi(\widehat{\nu}_{k+} - \widehat{\nu}_k)\|_{\mathbb{H}}^2. \quad (47a)$$

Writing  $\widehat{\nu}_k = \sum_{j=1}^{p_k} \widehat{\omega}_j^k \delta_{\widehat{t}_j^k}$  and  $\widehat{\nu}_{k+} = \sum_{j=1}^{p_k} \widehat{\omega}_j^{k+} \delta_{\widehat{t}_j^{k+}}$ , we introduce the following measure  $\tilde{\nu}_{k+} = \sum_{j=1}^{p_k} \widehat{\omega}_j^{k+} \delta_{\widehat{t}_j^k}$ . Then, we deduce that:

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

First remark that,

$$\begin{aligned} \|\Phi(\widehat{\nu}_{k+} - \tilde{\nu}_{k+})\|_{\mathbb{H}}^2 &= \left\| \sum_{j=1}^{p_k} \widehat{\omega}_j^{k+} (\varphi_{\widehat{t}_j^{k+}} - \varphi_{\widehat{t}_j^k}) \right\|_{\mathbb{H}}^2, \\ &\leq \sum_{j=1}^{p_k} \widehat{\omega}_j^{k+} \times \sum_{j=1}^{p_k} \widehat{\omega}_j^{k+} \|\varphi_{\widehat{t}_j^{k+}} - \varphi_{\widehat{t}_j^k}\|_{\mathbb{H}}^2, \\ &\leq \mathfrak{C}_{\mathcal{P}} \|\widehat{\nu}_{k+}\|_{\text{TV}} \sum_{j=1}^{p_k} \widehat{\omega}_j^{k+} \|\widehat{t}_j^{k+} - \widehat{t}_j^k\|^2, \\ &\leq \mathfrak{C}_{\mathcal{P}} \|\widehat{\nu}_{k+}\|_{\text{TV}} \beta^2 \sum_{j=1}^{p_k} \widehat{\omega}_j^{k+} \|\pi_{\mathcal{X}}(\widehat{t}_j^k, \widehat{D}_k(\widehat{t}_j^k), \beta)\|^2, \end{aligned} \quad (47b)$$

where we have used (3) and (10b). Similarly

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

where the last line comes from the Jensen inequality and from the fact that  $\|\varphi_t\|_{\mathbb{H}}^2 = 1$  for any  $t \in \mathcal{X}$ . Moreover, since the random variable  $\widehat{J}'_k(t)$  is bounded for any  $t \in \mathcal{X}$ ,

$$\begin{aligned}
\|\Phi(\tilde{\nu}_{k+} - \hat{\nu}_k)\|_{\mathbb{H}}^2 &\leq \mathfrak{C} \|\hat{\nu}_k\|_{\text{TV}} \alpha^2 \sum_{j=1}^{p_k} \hat{\omega}_j^k \widehat{J}'_k(\hat{t}_j^k)^2 \\
&\leq \mathfrak{C} \|\hat{\nu}_k\|_{\text{TV}} \alpha^2 \|\widehat{J}'_k\|_{\hat{\nu}_k}^2.
\end{aligned}$$

Taking the conditional expectation, since the particles' weights and positions  $(\hat{\omega}_j^k, \hat{t}_j^k)_{j \in [p_k]}$  are  $\mathfrak{F}_k$ -measurable and independent of the  $k$ -th iteration mini-batch, we can push the expectation inside the sum over the particles:

$$\mathbb{E} \left[ \|\widehat{J}'_k\|_{\hat{\nu}_k}^2 \mid \mathfrak{F}_k \right] = \sum_{j=1}^{p_k} \hat{\omega}_j^k \mathbb{E} \left[ |\widehat{J}'_k(\hat{t}_j^k)|^2 \mid \mathfrak{F}_k \right].$$

Using the pointwise bias-variance decomposition at each fixed particle position  $\hat{t}_j^k \in \mathfrak{F}_k$ :

$$\mathbb{E} \left[ |\widehat{J}'_k(\hat{t}_j^k)|^2 \mid \mathfrak{F}_k \right] = |J'_{\hat{\nu}_k}(\hat{t}_j^k)|^2 + \text{Var} \left( \widehat{J}'_k(\hat{t}_j^k) \mid \mathfrak{F}_k \right) \leq |J'_{\hat{\nu}_k}(\hat{t}_j^k)|^2 + \frac{\mathbf{E}_{\infty}^2}{m_k},$$

where  $\mathbf{E}_{\infty}^2$  provides a uniform bound on the variance of the stochastic gradient evaluations due to assumption (9b). Plugging this into the second-order term and using the almost sure boundedness of  $(\|\hat{\nu}_k\|_{\text{TV}})_{k \geq 0}$ :

$$\mathbb{E} \left[ \|\Phi(\tilde{\nu}_{k+} - \hat{\nu}_k)\|_{\mathbb{H}}^2 \mid \mathfrak{F}_k \right] \leq \mathfrak{C} \alpha^2 \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 + \mathfrak{C} \frac{\alpha^2}{m_k}.$$

Gathering with Equation (47b), we deduce that:

$$\mathbb{E} \left[ \|\Phi(\hat{\nu}_{k+} - \hat{\nu}_k)\|_{\mathbb{H}}^2 \mid \mathfrak{F}_k \right] \leq \mathfrak{C} \alpha^2 \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 + \mathfrak{C} \frac{\alpha^2}{m_k} + \mathfrak{C} \beta^2. \quad (47c)$$

Step 2: Study of the drift. We expand the first order term in (47a) and observe that:

$$\begin{aligned}
\int_{\mathcal{X}} J'_{\hat{\nu}_k} d(\hat{\nu}_{k^+} - \hat{\nu}_k) &= \sum_{j=1}^{p_k} \left[ (\hat{\omega}_j^{k^+} - \hat{\omega}_j^k) J'_{\hat{\nu}_k}(\hat{t}_j^k) + \hat{\omega}_j^k (J'_{\hat{\nu}_k}(\hat{t}_j^{k^+}) - J'_{\hat{\nu}_k}(\hat{t}_j^k)) \right] \\
&\quad + \sum_{j=1}^{p_k} (\hat{\omega}_j^{k^+} - \hat{\omega}_j^k) (J'_{\hat{\nu}_k}(\hat{t}_j^{k^+}) - J'_{\hat{\nu}_k}(\hat{t}_j^k)), \\
&= \sum_{j=1}^{p_k} \left[ (\hat{\omega}_j^{k^+} - \hat{\omega}_j^k) J'_{\hat{\nu}_k}(\hat{t}_j^k) + \hat{\omega}_j^k \langle \hat{t}_j^{k^+} - \hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \rangle \right] \\
&\quad + \sum_{j=1}^{p_k} \left[ \hat{\omega}_j^k \langle \hat{t}_j^{k^+} - \hat{t}_j^k, \nabla^2 J'_{\hat{\nu}_k}(\hat{t}_j^k) (\hat{t}_j^{k^+} - \hat{t}_j^k) \rangle + (\hat{\omega}_j^{k^+} - \hat{\omega}_j^k) \langle \nabla J'_{\hat{\nu}_k}(\tilde{v}_j^k), \hat{t}_j^{k^+} - \hat{t}_j^k \rangle \right],
\end{aligned}$$

where  $\hat{v}_j^k$  and  $\tilde{v}_j^k$  are some auxiliary points that belong to  $(\hat{t}_j^k, \hat{t}_j^{k^+})$  obtained with the help of first and second order Taylor expansions. Using Proposition C.1 in De Castro et al. (2025a), we get

$$\begin{aligned}
\int_{\mathcal{X}} J'_{\hat{\nu}_k} d(\hat{\nu}_{k^+} - \hat{\nu}_k) &\leq \sum_{j=1}^{p_k} \left[ (\hat{\omega}_j^{k^+} - \hat{\omega}_j^k) J'_{\hat{\nu}_k}(\hat{t}_j^k) + \hat{\omega}_j^k \langle \hat{t}_j^{k^+} - \hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \rangle \right] \\
&\quad + \|\nabla^2 J'_{\hat{\nu}_k}\|_{\infty} \sum_{j=1}^{p_k} \hat{\omega}_j^k \|\hat{t}_j^{k^+} - \hat{t}_j^k\|^2 + \sum_{j=1}^{p_k} |\hat{\omega}_j^{k^+} - \hat{\omega}_j^k| \times \|\nabla J'_{\hat{\nu}_k}\| \|\hat{t}_j^{k^+} - \hat{t}_j^k\|, \\
&\leq \sum_{j=1}^{p_k} \left[ (\hat{\omega}_j^{k^+} - \hat{\omega}_j^k) J'_{\hat{\nu}_k}(\hat{t}_j^k) + \hat{\omega}_j^k \langle \hat{t}_j^{k^+} - \hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \rangle \right] \\
&\quad + A \sum_{j=1}^{p_k} \left[ \hat{\omega}_j^k \|\hat{t}_j^{k^+} - \hat{t}_j^k\|^2 + |\hat{\omega}_j^{k^+} - \hat{\omega}_j^k| \|\hat{t}_j^{k^+} - \hat{t}_j^k\| \right],
\end{aligned}$$

where

$$A := (\|\hat{\nu}_k\|_{\text{TV}} + \|y\|_{\mathbb{H}}) \mathfrak{C}_{\mathcal{P}}.$$

Using the weights update (10b), we obtain

$$\begin{aligned}
\int_{\mathcal{X}} J'_{\hat{\nu}_k} d(\hat{\nu}_{k^+} - \hat{\nu}_k) &\leq \sum_{j=1}^{p_k} \hat{\omega}_j^k (e^{-\alpha \widehat{J}'_k(\hat{t}_j^k)} - 1) J'_{\hat{\nu}_k}(\hat{t}_j^k) + \hat{\omega}_j^k \langle \hat{t}_j^{k^+} - \hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \rangle \\
&\quad + A \sum_{j=1}^{p_k} \left[ \hat{\omega}_j^k \|\hat{t}_j^{k^+} - \hat{t}_j^k\|^2 + \hat{\omega}_j^k \left| e^{-\alpha \widehat{J}'_k(\hat{t}_j^k)} - 1 \right| \|\hat{t}_j^{k^+} - \hat{t}_j^k\| \right]. \quad (47d)
\end{aligned}$$

The first term of the right hand side of Equation (47d) is dealt thanks to Proposition 5: a straightforward conditional expectation argument yields:

$$\begin{aligned} \mathbb{E} \left[ \sum_{j=1}^{p_k} \hat{\omega}_j^k (e^{-\alpha \widehat{J}'_k(\hat{t}_j^k)} - 1) J'_{\hat{\nu}_k}(\hat{t}_j^k) \mid \mathfrak{F}_k \right] &\leq -\alpha \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 \\ &+ \frac{\alpha^2 \|J'_{\hat{\nu}_k}\|_{\infty} \|\hat{\nu}_k\|_{TV} \mathbf{E}_{\infty}^2}{m_k} e^{\alpha \|J'_{\hat{\nu}_k}\|_{\infty}} + \frac{\alpha^2 \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2}{2} e^{\alpha \|J'_{\hat{\nu}_k}\|_{\infty}} \end{aligned}$$

We pay a specific attention to the second term of the right hand side of Equation (47d). Using the generalized projected gradient and its related properties (*e.g.* Lemma A.6), we get for any  $j \in \{1, \dots, p\}$ ,

$$\begin{aligned} &\hat{\omega}_j^k \langle \hat{t}_j^{k+} - \hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \rangle \\ &= -\beta \hat{\omega}_j^k \langle \pi_{\mathcal{X}}(\hat{t}_j^k, \widehat{D}_k(\hat{t}_j^k), \beta), \widehat{D}_k(\hat{t}_j^k) \rangle + \beta \hat{\omega}_j^k \langle \pi_{\mathcal{X}}(\hat{t}_j^k, \widehat{D}_k(\hat{t}_j^k), \beta), \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) - \widehat{D}_k(\hat{t}_j^k) \rangle, \\ &\leq -\beta \hat{\omega}_j^k \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \widehat{D}_k(\hat{t}_j^k), \beta) \right\|^2 + \beta \hat{\omega}_j^k \langle \pi_{\mathcal{X}}(\hat{t}_j^k, \widehat{D}_k(\hat{t}_j^k), \beta), \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) - \widehat{D}_k(\hat{t}_j^k) \rangle. \end{aligned}$$

Using the Young inequality, we get

$$\hat{\omega}_j^k \langle \hat{t}_j^{k+} - \hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \rangle \leq -\frac{\beta}{2} \hat{\omega}_j^k \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \widehat{D}_k(\hat{t}_j^k), \beta) \right\|^2 + 2\beta \hat{\omega}_j^k \left\| \widehat{D}_k(\hat{t}_j^k) - \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \right\|^2,$$

where we have used again Lemma A.6. Taking the conditional expectation w.r.t.  $\mathfrak{F}_k$ , we get

$$\begin{aligned} &\mathbb{E} \left[ \hat{\omega}_j^k \langle \hat{t}_j^{k+} - \hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \rangle \mid \mathfrak{F}_k \right] \\ &\leq -\frac{\beta}{2} \hat{\omega}_j^k \mathbb{E} \left[ \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \widehat{D}_k(\hat{t}_j^k), \beta) \right\|^2 \mid \mathfrak{F}_k \right] + 2\beta \hat{\omega}_j^k \mathbb{E} \left[ \left\| \widehat{D}_k(\hat{t}_j^k) - \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \right\|^2 \mid \mathfrak{F}_k \right], \\ &\leq -\frac{\beta}{2} \hat{\omega}_j^k \mathbb{E} \left[ \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta) \right\|^2 \mid \mathfrak{F}_k \right] + 2\beta \hat{\omega}_j^k \mathbb{E} \left[ \left\| \widehat{D}_k(\hat{t}_j^k) - \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \right\|^2 \mid \mathfrak{F}_k \right] \\ &\quad - \beta \hat{\omega}_j^k \mathbb{E} \left[ \langle \pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta), \pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta) - \pi_{\mathcal{X}}(\hat{t}_j^k, \widehat{D}_k(\hat{t}_j^k), \beta) \rangle \mid \mathfrak{F}_k \right] \\ &\leq -\frac{\beta}{4} \hat{\omega}_j^k \mathbb{E} \left[ \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta) \right\|^2 \mid \mathfrak{F}_k \right] + 4\beta \hat{\omega}_j^k \mathbb{E} \left[ \left\| \widehat{D}_k(\hat{t}_j^k) - \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \right\|^2 \mid \mathfrak{F}_k \right]. \end{aligned}$$

At this step, we can take advantage of the mini-batch step to control the second expectation in the previous inequality. Indeed, according to (10a), we have

$$\begin{aligned} \mathbb{E} \left[ \left\| \widehat{D}_k(\hat{t}_j^k) - \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \right\|^2 \mid \mathfrak{F}_k \right] &= \frac{1}{m_k^2} \sum_{l=1}^{m_k} \mathbb{E} \left[ \left\| \zeta_{\hat{\nu}_k}(\hat{t}_j^k, Z_l^k) \right\|^2 \mid \mathfrak{F}_k \right], \\ &\leq \frac{\mathbf{E}_{\infty}^2}{m_k} \leq \frac{\mathbf{c}}{m_k}. \end{aligned}$$

This leads to

$$\mathbb{E} \left[ \hat{\omega}_j^k \langle \hat{t}_j^{k+} - \hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k) \rangle \mid \mathfrak{F}_k \right] \leq -\frac{\beta}{4} \hat{\omega}_j^k \mathbb{E} \left[ \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta) \right\|^2 \mid \mathfrak{F}_k \right] + \frac{\mathbf{c}}{m_k}.$$

Plugging this expression in (47d), using the almost sure TV boundedness and taking the conditional expectation, we get:

$$\begin{aligned}
\mathbb{E} \left[ \int_{\mathcal{X}} J'_{\hat{\nu}_k} d(\hat{\nu}_{k+} - \hat{\nu}_k) | \mathfrak{F}_k \right] &\leq \sum_{j=1}^{p_k} \hat{\omega}_j^k \mathbb{E} \left[ (e^{-\alpha \widehat{J}_k(\hat{t}_j^k)} - 1) | \mathfrak{F}_k \right] J'_{\hat{\nu}_k}(\hat{t}_j^k) + \mathfrak{C} \frac{\beta}{m_k} \\
&\quad - \frac{\beta}{4} \sum_{j=1}^{p_k} \hat{\omega}_j^k \mathbb{E} \left[ \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta) \right\|^2 | \mathfrak{F}_k \right] \\
&\quad + \mathfrak{C} \sum_{j=1}^p \mathbb{E} \left[ \left( \hat{\omega}_j^k \|\hat{t}_j^{k+} - \hat{t}_j^k\|^2 + \hat{\omega}_j^k \left| e^{-\alpha \widehat{J}_k(\hat{t}_j^k)} - 1 \right| \|\hat{t}_j^{k+} - \hat{t}_j^k\| \right) | \mathfrak{F}_k \right], \\
&\leq -\alpha \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 - \frac{\beta}{4} \sum_{j=1}^{p_k} \hat{\omega}_j^k \mathbb{E} \left[ \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta) \right\|^2 | \mathfrak{F}_k \right] \\
&\quad + \mathfrak{C} \left( \alpha^2 \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 + \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right), \tag{47e}
\end{aligned}$$

where we have used the almost sure boundedness of  $(\|\hat{\nu}_k\|_{TV})_{k \geq 0}$  and a large enough  $\mathfrak{C}$ . Considering now Equations (47c) and (47e), using  $\alpha \leq (4\mathfrak{C})^{-1}$ , the quadratic drift term  $\mathfrak{C}\alpha^2 \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2$  is absorbed by half of the linear descent  $-\alpha \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2$ , and we finally obtain that:

$$\begin{aligned}
\mathbb{E} [J(\hat{\nu}_{k+}) - J(\hat{\nu}_k) | \mathfrak{F}_k] &\leq -\frac{\alpha}{2} \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 - \frac{\beta}{4} \sum_{j=1}^{p_k} \hat{\omega}_j^k \mathbb{E} \left[ \left\| \pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta) \right\|^2 | \mathfrak{F}_k \right] \\
&\quad + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right).
\end{aligned}$$

This bound delivers a stronger conclusion than stated in the proposition, since it retains the negative projected-gradient term  $-\frac{\beta}{4} \sum_j \hat{\omega}_j^k \mathbb{E} [\|\pi_{\mathcal{X}}(\hat{t}_j^k, \nabla J'_{\hat{\nu}_k}(\hat{t}_j^k), \beta)\|^2 | \mathfrak{F}_k]$ ; dropping it recovers exactly the statement of Proposition 13.  $\blacksquare$

### D.3 One-step analysis

We introduce below the stochastic counterpart of the set  $\Theta_{k-1+}$  used in the deterministic approach and defined in Equation (41), that is denoted as  $\hat{\Theta}_{k-1+}$  and is given by:

$$\hat{\Theta}_{k-1+} := \left\{ J'_{\hat{\nu}_{k-1+}} \leq \frac{\hat{\nu}_{k-1+}}{2} \right\} \quad \text{with} \quad \hat{\nu}_{k-1+} = \min(J'_{\hat{\nu}_{k-1+}}) \wedge 0.$$

**Proposition 14** *Assume that the sequence  $(\hat{\nu}_k, \hat{\nu}_{k+})_{k \geq 1}$  satisfies  $(\hat{\mathcal{H}}_D)$ ,  $(\hat{\mathcal{H}}_{\varepsilon, a}^+)$ ,  $(\hat{\mathcal{H}}_{\varepsilon}^{\text{smooth}, 1})$ ,  $(\hat{\mathcal{H}}_{\varepsilon}^{\text{smooth}, 2})$  and  $(\hat{\mathcal{H}}_{TV}^{\infty})$ . Then, for any  $k \geq 1$ :*

$$\begin{aligned}
&\mathbb{E} [J(\hat{\nu}_{k+}) | \mathfrak{F}_k] - J(\hat{\nu}_k) \\
&\leq -c\alpha\varepsilon_{k-1} ( (|\hat{\nu}_{k-1+}|^2 - \mathfrak{C}\varepsilon_{k-1}^2) \vee 0 ) \left( |\hat{\nu}_{k-1+}|^d - \mathfrak{C}m_{k-1}^{-a} \right) + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right).
\end{aligned}$$

**Proof** First, we use Assumption  $(\hat{\mathcal{H}}_D)$  and obtain that:

$$\begin{aligned} \mathbb{E} [J(\hat{\nu}_{k+}) | \mathfrak{F}_k] - J(\hat{\nu}_k) &\leq -\frac{\alpha}{2} \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right), \\ &= -\frac{\alpha}{2} \int_{\mathcal{X}} |J'_{\hat{\nu}_k}|^2 d\hat{\nu}_k + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right), \\ &\leq -\frac{\alpha}{2} \int_{\hat{\Theta}_{k-1+}} |J'_{\hat{\nu}_k}|^2 d\hat{\nu}_k + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right) \end{aligned}$$

We apply the Young inequality  $|J'_{\hat{\nu}_k}(t)|^2 \geq \frac{1}{2}|J'_{\hat{\nu}_{k-1+}}(t)|^2 - |J'_{\hat{\nu}_k}(t) - J'_{\hat{\nu}_{k-1+}}(t)|^2$  to get, according to  $(\hat{\mathcal{H}}_{\varepsilon}^{\text{smooth},2})$ ,

$$\forall t \in \hat{\Theta}_{k-1+}, \quad |J'_{\hat{\nu}_k}(t)|^2 \geq \frac{1}{2}|J'_{\hat{\nu}_{k-1+}}(t)|^2 - \mathfrak{C}\varepsilon_{k-1}^2 \geq \frac{|\hat{\nu}_{k-1+}|^2}{8} - \mathfrak{C}\varepsilon_{k-1}^2.$$

Thanks to the positivity of  $\alpha \int_{\hat{\Theta}_{k-1+}} |J'_{\hat{\nu}_k}|^2 d\hat{\nu}_k$ , we then get:

$$\mathbb{E} [J(\hat{\nu}_{k+}) | \mathfrak{F}_k] - J(\hat{\nu}_k) \leq -\frac{\alpha}{2} \left( \left[ \frac{|\hat{\nu}_{k-1+}|^2}{8} - \mathfrak{C}\varepsilon_{k-1}^2 \right] \vee 0 \right) \hat{\nu}_k(\hat{\Theta}_{k-1+}) + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right).$$

The term  $\hat{\nu}_{k-1+}$  is  $\mathfrak{F}_{k-1}^+$  measurable, and since  $\mathfrak{F}_{k-1}^+ \subset \mathfrak{F}_k$ , we get

$$\begin{aligned} &\mathbb{E} [J(\hat{\nu}_{k+}) - J(\hat{\nu}_k) | \mathfrak{F}_{k-1}^+] \\ &= \mathbb{E} [\mathbb{E} [J(\hat{\nu}_{k+}) - J(\hat{\nu}_k) | \mathfrak{F}_k] | \mathfrak{F}_{k-1}^+] \\ &\leq -\frac{\alpha}{2} \left( \left[ \frac{|\hat{\nu}_{k-1+}|^2}{8} - \mathfrak{C}\varepsilon_{k-1}^2 \right] \vee 0 \right) \mathbb{E} [\hat{\nu}_k(\hat{\Theta}_{k-1+}) | \mathfrak{F}_{k-1}^+] + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right), \\ &\leq -\frac{\alpha}{2} \varepsilon_{k-1} \left( \left[ \frac{|\hat{\nu}_{k-1+}|^2}{8} - \mathfrak{C}\varepsilon_{k-1}^2 \right] \vee 0 \right) [\mathfrak{c}\lambda(\hat{\Theta}_{k-1+}) - \mathfrak{C}m_{k-1}^{-a}] \\ &+ \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right), \end{aligned} \tag{48}$$

where we have used Assumption  $(\hat{\mathcal{H}}_{\varepsilon}^+)$  at iteration  $k-1$  for the last inequality. The last part of the proof is very similar to the one displayed in the deterministic case. First, recall that  $J'_{\nu}$  is a  $\mathfrak{L}(\nu)$ -Lipschitz function and thanks to the boundedness of the TV norm stated by  $\hat{\mathcal{H}}_{\text{TV}}$ , we know that for any  $k$ ,  $J'_{\hat{\nu}_{k-1+}}$  is  $\mathfrak{L}$ -Lipschitz. We then get:

$$\begin{aligned} \forall k \geq 1: \quad |x - \arg \min J'_{\hat{\nu}_{k-1+}}| \leq \frac{|\hat{\nu}_{k-1+}|}{2\mathfrak{L}} &\implies |J'_{\hat{\nu}_{k-1+}}(x) - \min(J'_{\hat{\nu}_{k-1+}})| \leq \mathfrak{L} \times \frac{|\hat{\nu}_{k-1+}|}{2\mathfrak{L}} \\ &\implies J'_{\hat{\nu}_{k-1+}}(x) \leq \frac{\hat{\nu}_{k-1+}}{2} \\ &\implies x \in \hat{\Theta}_{k-1+}. \end{aligned}$$

This leads to the  $\mathfrak{F}_{k-1}^+$ -measurable inequality:

$$\lambda(\hat{\Theta}_{k-1+}) \geq \left| \frac{\hat{\nu}_{k-1+}}{2\mathfrak{L}} \right|^d.$$

Using this lower bound in Equation (48)

$$\begin{aligned} & \mathbb{E} [J(\hat{\nu}_{k+}) - J(\hat{\nu}_k) | \mathfrak{F}_{k-1}^+] \\ & \leq -\mathfrak{c}\alpha\varepsilon_{k-1} ( (|\hat{\nu}_{k-1+}|^2 - \mathfrak{c}\varepsilon_{k-1}^2) \vee 0 ) \left( |\hat{\nu}_{k-1+}|^d - \mathfrak{c}m_{k-1}^{-a} \right) + \mathfrak{c} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right). \end{aligned}$$

■

We introduce  $\widehat{\Delta}_{k+}$  that quantifies the amount of decrease on  $J$  through the transportation map:

$$\widehat{\Delta}_{k+} = J(\hat{\nu}_k) - J(\hat{\nu}_{k+}).$$

As the evolution is now randomized, we emphasize that  $\widehat{\Delta}_{k+}$  is not necessarily greater than 0.

**Proposition 15** *Assume that the sequence  $(\hat{\nu}_k, \hat{\nu}_{k+})_{k \geq 1}$  satisfies  $(\widehat{\mathcal{H}}_D)$ ,  $(\widehat{\mathcal{H}}_{\varepsilon, a}^+)$ ,  $(\widehat{\mathcal{H}}_{\varepsilon}^{\text{smooth}, 1})$ ,  $(\widehat{\mathcal{H}}_{\varepsilon}^{\text{smooth}, 2})$  and  $(\widehat{\mathcal{H}}_{\text{TV}}^{\infty})$ . Then, for any  $k \geq 1$ , we have the  $\mathfrak{F}_k$ -measurable inequality:*

$$\begin{aligned} J(\hat{\nu}_k) - J^* & \leq \mathfrak{c} \max \left( \alpha^{-1/2} \left[ \frac{\mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k]}{\varepsilon_{k-1}} \right]^{\frac{1}{2+d}} ; \sqrt{\frac{\alpha}{m_k}} ; \frac{\beta}{\sqrt{\alpha}} ; \sqrt{\frac{\beta}{\alpha m_k}} ; \varepsilon_{k-1} ; m_k^{-a/d} ; \right. \\ & \left. \left( \frac{\alpha}{m_k \varepsilon_{k-1}} \right)^{\frac{1}{2+d}} ; \left( \frac{\beta^2 + \frac{\beta}{m_k}}{\alpha \varepsilon_{k-1}} \right)^{\frac{1}{2+d}} \right). \end{aligned}$$

**Proof** Thanks to the convexity of  $J$ , we have:

$$J(\hat{\nu}_k) - J^* \leq \int_{\mathcal{X}} J'_{\hat{\nu}_k} d(\hat{\nu}_k - \nu^*). \quad (49a)$$

We now bound separately the two terms in the r.h.s. of (49a) and begin with the term  $\int J'_{\hat{\nu}_k} d\hat{\nu}_k$ . The Cauchy-Schwarz inequality and Assumption (15- $(\widehat{\mathcal{H}}_{\text{TV}}^{\infty})$ ) yield:

$$\left| \int_{\mathcal{X}} J'_{\hat{\nu}_k} d\hat{\nu}_k \right| \leq \left[ \|\hat{\nu}_k\|_{\text{TV}} \int_{\mathcal{X}} |J'_{\hat{\nu}_k}|^2 d\hat{\nu}_k \right]^{1/2} \leq \mathfrak{c} \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}.$$

Thanks to Assumption  $(\widehat{\mathcal{H}}_D)$ , we have:

$$\frac{\alpha}{2} \|J'_{\hat{\nu}_k}\|_{\hat{\nu}_k}^2 \leq \mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k] + \mathfrak{c} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right),$$

which in turn implies that

$$\left| \int_{\mathcal{X}} J'_{\hat{\nu}_k} d\hat{\nu}_k \right| \leq \frac{\mathfrak{c}}{\sqrt{\alpha}} \sqrt{\mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k] + \mathfrak{c} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right)}. \quad (49b)$$

The second integral in (49a) is dealt with Assumption  $(\widehat{\mathcal{H}}_{\varepsilon}^{\text{smooth}, 2})$  leading first to the bound

$$\int_{\mathcal{X}} J'_{\hat{\nu}_k} d\nu^* = \int_{\mathcal{X}} J'_{\hat{\nu}_{k-1+}} d\nu^* + \int_{\mathcal{X}} (J'_{\hat{\nu}_k} - J'_{\hat{\nu}_{k-1+}}) d\nu^* \geq (\hat{\nu}_{k-1+} - \mathfrak{c}\varepsilon_{k-1}) \|\nu^*\|_{\text{TV}}.$$

Then, two different situations may occur according to the value of  $\hat{\nu}_{k-1+}$ .

- **1<sup>st</sup> case:**  $\hat{v}_{k-1+} \leq -\left(2\mathfrak{C}\varepsilon_{k-1} \vee (2\mathfrak{C})^{1/d}m_k^{-a/d}\right)$ . Then, we get from the previous bound that

$$\int_{\mathcal{X}} J'_{\hat{v}_k} d\nu^* \geq \frac{3}{2}\hat{v}_{k-1+}\|\nu^*\|_{\text{TV}}.$$

Simultaneously, as soon as  $\hat{v}_{k-1+} \leq -2\mathfrak{C}\varepsilon_{k-1}$  and  $|\hat{v}_{k-1+}|^d \geq 2\mathfrak{C}m_k^{-a}$ , Proposition 14 implies that:

$$\begin{aligned} \mathbb{E}\left[-\widehat{\Delta}_{k+} \mid \mathfrak{F}_k\right] &\leq -\mathfrak{C}\alpha\varepsilon_{k-1} \left(\left[|\hat{v}_{k-1+}|^2 - \mathfrak{C}\varepsilon_{k-1}^2\right] \vee 0\right) \left(|\hat{v}_{k-1+}|^d - \mathfrak{C}m_k^{-a}\right) \\ &\quad + \mathfrak{C} \left(\frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k}\right), \\ &\leq -\mathfrak{C}\alpha\varepsilon_{k-1}|\hat{v}_{k-1+}|^{2+d} + \mathfrak{C} \left(\frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k}\right). \end{aligned} \quad (49c)$$

The last inequality implies that

$$|\hat{v}_{k-1+}|^{2+d} \leq \mathfrak{C} \frac{\left(\mathbb{E}\left[\widehat{\Delta}_{k+} \mid \mathfrak{F}_k\right] + \mathfrak{C} \left(\frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k}\right)\right)}{\alpha\varepsilon_{k-1}}.$$

Hence, we deduce from these computations that:

$$\int_{\mathcal{X}} J'_{\hat{v}_k} d\nu^* \geq -\mathfrak{C} \left[ \frac{\left(\mathbb{E}\left[\widehat{\Delta}_{k+} \mid \mathfrak{F}_k\right] + \mathfrak{C} \left(\frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k}\right)\right)}{\alpha\varepsilon_{k-1}} \right]^{\frac{1}{2+d}}.$$

- **2<sup>nd</sup> case:**  $\hat{v}_{k-1+} \geq -\left(2\mathfrak{C}\varepsilon_{k-1} \vee (2\mathfrak{C})^{1/d}m_k^{-a/d}\right)$ . In such a situation, we immediately have from Assumption  $\widehat{\mathcal{H}}_\varepsilon^{\text{smooth},2}$  that:

$$\int_{\mathcal{X}} J'_{\hat{v}_k} d\nu^* \geq (\hat{v}_{k-1+} - \mathfrak{C}\varepsilon_{k-1})\|\nu^*\|_{\text{TV}} \geq -\mathfrak{C}(\varepsilon_{k-1} \vee m_k^{-a/d}).$$

Regardless of the value of  $\hat{v}_{k-1+}$ , we then get the almost sure inequality:

$$\int_{\mathcal{X}} J'_{\hat{v}_k} d\nu^* \geq -\mathfrak{C} \max \left( \varepsilon_{k-1}; m_k^{-a/d}; \left[ \frac{\left(\mathbb{E}\left[\widehat{\Delta}_{k+} \mid \mathfrak{F}_k\right] + \mathfrak{C} \left(\frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k}\right)\right)}{\alpha\varepsilon_{k-1}} \right]^{\frac{1}{2+d}} \right). \quad (49d)$$

Gathering Equations (49a), (49b) and (49d), we deduce that:

$$\begin{aligned}
& J(\hat{\nu}_k) - J^* \\
& \leq \frac{\mathfrak{C}}{\sqrt{\alpha}} \left( \mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k] + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right) \right)^{1/2} \\
& \quad + \mathfrak{C} \max \left( \varepsilon_{k-1}; m_k^{-a/d}; \left[ \frac{\left( \mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k] + \mathfrak{C} \left( \frac{\alpha^2}{m_k} + \beta^2 + \frac{\beta}{m_k} \right) \right)}{\alpha \varepsilon_{k-1}} \right]^{\frac{1}{2+d}} \right) \\
& \leq \mathfrak{C} \max \left( \left[ \frac{\mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k]}{\alpha \varepsilon_{k-1}} \right]^{\frac{1}{2+d}}; \left[ \frac{\mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k]}{\alpha} \right]^{\frac{1}{2}}; \sqrt{\frac{\alpha}{m_k}}; \frac{\beta}{\sqrt{\alpha}}; \sqrt{\frac{\beta}{\alpha m_k}}; \varepsilon_{k-1}; m_k^{-a/d}; \right. \\
& \quad \left. \left( \frac{\alpha}{m_k \varepsilon_{k-1}} \right)^{\frac{1}{2+d}}; \left( \frac{\beta^2 + \frac{\beta}{m_k}}{\alpha \varepsilon_{k-1}} \right)^{\frac{1}{2+d}} \right).
\end{aligned}$$

Since  $(\widehat{\Delta}_{k+})_{k \geq 1}$  is a bounded sequence, we can verify that the last inequality is translated into:

$$\begin{aligned}
J(\hat{\nu}_k) - J^* & \leq \mathfrak{C} \max \left( \alpha^{-1/2} \left[ \frac{\mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k]}{\varepsilon_{k-1}} \right]^{\frac{1}{2+d}}; \sqrt{\frac{\alpha}{m_k}}; \frac{\beta}{\sqrt{\alpha}}; \sqrt{\frac{\beta}{\alpha m_k}}; \varepsilon_{k-1}; m_k^{-a/d}; \right. \\
& \quad \left. \left( \frac{\alpha}{m_k \varepsilon_{k-1}} \right)^{\frac{1}{2+d}}; \left( \frac{\beta^2 + \frac{\beta}{m_k}}{\alpha \varepsilon_{k-1}} \right)^{\frac{1}{2+d}} \right).
\end{aligned}$$

■

#### D.4 Proof of Theorem 3.1

**Proof** In what follows, we set:

$$F(k) = J(\hat{\nu}_k) - J(\nu^*) \quad \forall k \in \mathbb{N}^*.$$

Let  $k \geq 1$  be fixed. Then, we know from Assumption  $(\widehat{\mathcal{H}}_\varepsilon^{\text{smooth},1})$

$$\begin{aligned}
\mathbb{E} [F(k) - F(k+1) | \mathfrak{F}_k^+] & = \mathbb{E} [J(\hat{\nu}_k) - J(\hat{\nu}_{k+1}) | \mathfrak{F}_k^+] \\
& = J(\hat{\nu}_k) - J(\hat{\nu}_{k+1}) + \mathbb{E} [J(\hat{\nu}_{k+1}) - J(\hat{\nu}_{k+1}) | \mathfrak{F}_k^+] \\
& \geq \widehat{\Delta}_{k+} - \mathfrak{C} \left( \varepsilon_k^2 + \varepsilon_k \sqrt{\frac{\log m_k}{m_k}} \right).
\end{aligned} \tag{50}$$

Applying Proposition 15, we get

$$\begin{aligned}
\mathfrak{C} F(k)^{2+d} & \leq \alpha^{-\frac{2+d}{2}} \left[ \frac{\mathbb{E}[\widehat{\Delta}_{k+} | \mathfrak{F}_k]}{\varepsilon_{k-1}} \right] + \left( \frac{\alpha}{m_k} \right)^{\frac{2+d}{2}} + \left( \frac{\beta}{\sqrt{\alpha}} \right)^{2+d} + \left( \frac{\beta}{\alpha m_k} \right)^{\frac{2+d}{2}} + \varepsilon_{k-1}^{2+d} + m_k^{-\frac{(2+d)a}{d}} + \\
& \quad \frac{\alpha}{m_k \varepsilon_{k-1}} + \frac{\beta^2 + \frac{\beta}{m_k}}{\alpha \varepsilon_{k-1}}.
\end{aligned}$$

Then, using Equation (50), the tower rule and then multiplying each term by  $\varepsilon_{k-1}\alpha^{\frac{2+d}{2}}$ , we obtain:

$$\begin{aligned}\mathbb{E}[F(k) - F(k+1)|\mathfrak{F}_k] &= \mathbb{E}[\mathbb{E}[F(k) - F(k+1)|\mathfrak{F}_k^+]| \mathfrak{F}_k] \\ &\geq \mathbb{E}[\widehat{\Delta}_{k+}|\mathfrak{F}_k] - \mathfrak{C} \left( \varepsilon_k^2 + \varepsilon_k \sqrt{\frac{\log m_k}{m_k}} \right) \\ &\geq \mathfrak{C} \alpha^{\frac{2+d}{2}} \varepsilon_{k-1} F(k)^{2+d} - s_k,\end{aligned}$$

where  $s_k$  is given by:

$$\begin{aligned}s_k &= \frac{\alpha^{2+d}\varepsilon_{k-1}}{m_k^{\frac{2+d}{2}}} + \alpha^{\frac{2+d}{2}}\varepsilon_{k-1} \left( \frac{\beta}{\sqrt{\alpha}} \right)^{2+d} + \varepsilon_{k-1} \left( \frac{\beta}{m_k} \right)^{\frac{2+d}{2}} + \varepsilon_{k-1}^{3+d} \alpha^{\frac{2+d}{2}} + \varepsilon_{k-1} \alpha^{\frac{2+d}{2}} m_k^{\frac{-(2+d)\alpha}{d}} \\ &\quad + \frac{\alpha^{\frac{4+d}{2}}}{m_k} + \alpha^{d/2} \left( \beta^2 + \frac{\beta}{m_k} \right) + \varepsilon_k^2 + \varepsilon_k \sqrt{\frac{\log m_k}{m_k}}.\end{aligned}$$

Taking the global expectation on both side of the inequality, we get:

$$\mathbb{E}[F(k)] - \mathbb{E}[F(k+1)] \geq \mathfrak{C} \alpha^{\frac{2+d}{2}} \varepsilon_{k-1} \mathbb{E}[F(k)^{2+d}] - s_k.$$

We finally fix now a final horizon  $K$  and use a telescopic sum argument to obtain

$$\mathbb{E}[F(1)] - \mathbb{E}[F(K+1)] \geq \sum_{k=1}^K \left( \mathfrak{C} \alpha^{\frac{2+d}{2}} \varepsilon_{k-1} \mathbb{E}[F(k)^{2+d}] - s_k \right),$$

which can be rewritten as

$$\mathbb{E}[F(K+1)] + \mathfrak{C} \alpha^{\frac{2+d}{2}} \sum_{k=1}^K \varepsilon_{k-1} \mathbb{E}[F(k)^{2+d}] \leq \mathbb{E}[F(1)] + \sum_{k=1}^K s_k.$$

Using  $\hat{\rho}_K$  as the minimal value of  $F(k)_{1 \leq k \leq K}$ , this last inequality implies that

$$\mathbb{E}[\hat{\rho}_K^{2+d}] \leq \mathfrak{C} \frac{\mathbb{E}[F(1)] + \sum_{k=1}^K s_k}{\alpha^{\frac{2+d}{2}} \sum_{k=1}^K \varepsilon_{k-1}}.$$

Using the Jensen Inequality, we finally obtain

$$\mathbb{E}[\hat{\rho}_K] \leq \mathfrak{C} \left( \frac{\mathbb{E}[F(1)] + \sum_{k=1}^K s_k}{\alpha^{\frac{2+d}{2}} \sum_{k=1}^K \varepsilon_{k-1}} \right)^{\frac{1}{2+d}}. \quad (51)$$

We now consider the final tuning of our parameters to optimize the upper bound obtained in (51). For the sake of simplicity, we restrict our parametrization to constant step-size sequences that depend on the final horizon  $K$  of simulation: *i.e.*  $\alpha, \beta, \varepsilon_k$  and  $m_k$  are chosen constant with a value that only depends on  $K$  (the final horizon of simulation). A careful

inspection of the terms involved in (51) shows that  $\varepsilon$  needs to balance  $\frac{1}{\alpha^{1+d/2}\varepsilon K}$  and  $\frac{\varepsilon}{\alpha^{1+d/2}}$ . A straightforward argument yields

$$\forall k \in [1, K] \quad \varepsilon_k = \frac{1}{\sqrt{K}}.$$

Then, we observe that the step-size  $\alpha$  of the push-forward has to be chosen small enough (to guarantee the descent property stated in  $\widehat{\mathcal{H}}_D$ ) but has to be lower bounded and cannot be chosen arbitrarily small. Finally,  $\beta$  has to be chosen small enough to make the biggest term  $\frac{\beta^2}{\alpha\varepsilon}$  smaller than  $\frac{1}{\alpha^{1+d/2}\sqrt{K}}$ . We then deduce that:

$$\beta \leq \frac{1}{\sqrt{K}\alpha^{d/4}}.$$

At last, we setup the mini-batch size  $m$ . Thanks to the pointwise bounds established in Propositions 10, 12 and 13, the dominant mini-batch-dependent term in  $s_k$  is now  $\varepsilon\sqrt{\frac{\log m}{m}}$  (from the birth/death process), while the descent lemma contributes only  $\frac{\alpha^{(4+d)/2}}{m}$ . Setting

$$m = K$$

makes both terms of order  $\mathcal{O}(\sqrt{\log K}/K)$  or smaller, so that  $\sum_{k=1}^K s_k = \mathcal{O}(\sqrt{\log K})$ . According to these several choices, we then obtain the *global* convergence rate of our horizon dependent sequence:

$$\mathbb{E}[\hat{\rho}_K] \leq \mathfrak{C} \left( \alpha^{-(1+d/2)} \sqrt{\frac{\log K}{K}} \right)^{\frac{1}{2+d}} \leq \mathfrak{C} \alpha^{-1/2} \left( \frac{\log K}{K} \right)^{\frac{1}{2(2+d)}}.$$

This ends the proof of our final result. ■

**Proof** [Proof of Theorem 3.2] We start from the general bound (51) with the iteration-dependent schedules  $m_k = k \vee 1$ ,  $\varepsilon_k = \min(\alpha, 1/\sqrt{k \vee 1})$  and  $\beta_k = 1/(k \vee 1)$ . Let  $k_\alpha := \lceil 1/\alpha^2 \rceil$ , so that  $\varepsilon_k = \alpha$  for  $k \leq k_\alpha$  and  $\varepsilon_k = 1/\sqrt{k}$  for  $k > k_\alpha$ . The first  $k_\alpha$  iterations contribute only an additive  $\mathcal{O}(1)$  to all sums and are absorbed into  $\mathfrak{C}$ .

**The saturated regime  $k \leq k_\alpha$  is negligible in  $K$ .** We make this  $\mathcal{O}(1)$  statement precise. By Section 2.3,  $\alpha$  is fixed once and for all from intrinsic problem data; in particular,  $\alpha$  does not depend on  $K$ . Hence the threshold  $k_\alpha = \lceil 1/\alpha^2 \rceil$  is itself a constant independent of the horizon  $K$ , and although it may be large when  $\alpha$  is small, it does not grow with  $K$ .

On the saturated range  $\{1, \dots, k_\alpha\}$  each summand entering  $s_k$  in (51) is uniformly bounded by a constant  $\mathfrak{C}(\alpha, d, \mathbf{H}, \mathbf{G}, \mathbf{E}_\infty)$  depending only on intrinsic data: indeed  $\varepsilon_k = \alpha$  is constant,  $\beta_k \leq \beta_1 = 1$ ,  $m_k \geq 1$ , and the assumed bounds in (9b) together with  $(\mathcal{H}_{TV}^\infty)$  provide  $K$ -independent uniform controls on every factor (compare the term-by-term estimates below). Therefore the total contribution of these terms is bounded by

$$\sum_{k=1}^{k_\alpha} s_k \leq k_\alpha \mathfrak{C}(\alpha, d, \mathbf{H}, \mathbf{G}, \mathbf{E}_\infty) = \mathcal{O}(1) \quad \text{in } K,$$

which is additive,  $K$ -independent, and absorbed into the generic constant  $\mathfrak{C}$  of (28). For the lower bound on  $\sum_{k=1}^K \varepsilon_{k-1}$  established below, the saturated terms are nonnegative and may simply be dropped; the dominant  $\sqrt{K}$  growth comes from the regime  $k > k_\alpha$ , which is precisely why the assertion requires  $K \geq 4k_\alpha$ . Equivalently, the theorem assumes  $K \geq 4/\alpha^2$ , ensuring both that  $K$  exceeds  $k_\alpha$  enough to make the lower bound effective and that the saturated regime is dominated by the  $\sqrt{K}$  contribution.

**Lower bound on  $\sum \varepsilon_{k-1}$ .** For  $K \geq k_\alpha + 4$ :

$$\sum_{k=1}^K \varepsilon_{k-1} \geq \sum_{j=k_\alpha}^{K-1} \frac{1}{\sqrt{j}} \geq \int_{k_\alpha}^K x^{-1/2} dx = 2(\sqrt{K} - \sqrt{k_\alpha}) \geq \sqrt{K},$$

provided  $K \geq 4k_\alpha$ .

**Upper bound on  $\sum s_k$ .** We bound each group of terms in  $s_k$  separately. The dominant contributions are:

- *Birth/death term:*  $\sum_{k=1}^K \varepsilon_k \sqrt{\frac{\log m_k}{m_k}} \leq \sum_{k=1}^K \frac{\sqrt{\log k}}{k} \leq \frac{2}{3}(\log K)^{3/2} + \mathfrak{C}$ , using the integral bound  $\int_1^K \frac{\sqrt{\log x}}{x} dx = \frac{2}{3}(\log K)^{3/2}$ .
- *Exploration term:*  $\sum_{k=1}^K \varepsilon_k^2 \leq \sum_{k=1}^K \frac{1}{k} = \mathcal{O}(\log K)$ .
- *Descent lemma term:*  $\sum_{k=1}^K \frac{\alpha^{(4+d)/2}}{m_k} = \alpha^{(4+d)/2} \sum_{k=1}^K \frac{1}{k} = \mathcal{O}(\log K)$ .
- *Regularization term:*  $\sum_{k=1}^K \alpha^{d/2} \left( \beta_k^2 + \frac{\beta_k}{m_k} \right) = \alpha^{d/2} \sum_{k=1}^K \frac{2}{k^2} = \mathcal{O}(1)$ .
- *Hoeffding term:*  $\sum_{k=1}^K \varepsilon_{k-1} \alpha^{(2+d)/2} m_k^{-(2+d)a/d}$ . With  $a \geq \frac{d}{2(2+d)}$ , we have  $m_k^{-(2+d)a/d} \leq 1/\sqrt{k}$ , so the summand is at most  $\alpha^{(2+d)/2}/k$ , yielding  $\mathcal{O}(\log K)$  at the boundary value of  $a$  (and  $\mathcal{O}(1)$  for any  $a > \frac{d}{2(2+d)}$ ).
- All remaining terms in  $s_k$  involve strictly higher negative powers of  $k$  (each  $\leq k^{-(3+d)/2}$ ) and yield convergent sums bounded by a constant.

Therefore,  $\sum_{k=1}^K s_k = \mathcal{O}((\log K)^{3/2})$ .

**Conclusion.** Substituting into (51):

$$\mathbb{E}[\hat{\rho}_K] \leq \mathfrak{C} \left( \frac{(\log K)^{3/2}}{\alpha^{(2+d)/2} \sqrt{K}} \right)^{\frac{1}{2+d}} = \mathfrak{C} \alpha^{-1/2} \left( \frac{(\log K)^3}{K} \right)^{\frac{1}{2(2+d)}}.$$

Since  $N = \sum_{k=1}^K k = K(K+1)/2$ , we have  $K = \Theta(\sqrt{N})$ , and the sample complexity follows. ■

## Appendix E. Extension to $\beta > 0$ : position updates in the deterministic CPGD

In Section 2, the deterministic analysis is carried out under the restriction  $\beta = 0$ , meaning that only the weights of the measure are updated at each iteration. In this appendix, we extend the analysis to the case  $\beta > 0$ , where the Push-Forward update (Definition 1.1) also moves the positions of the particles. Throughout this section, we assume that  $(\alpha, \beta)$  satisfies condition (7). The transition  $\nu_k \mapsto \nu_{k+}$  now reads:

$$\nu_{k+} = \mathbb{T}_{\nu, \beta}^{\sharp} \mathbb{W}_{\nu_k, \alpha} \nu_k, \quad (52)$$

where  $\mathbb{T}_{\nu, \beta}^{\sharp}$  denotes the push-forward by the proximal position update  $\mathbb{T}_{\nu, \beta}$  (Definition 1.1).

### E.1 Total Variation boundedness for $\beta > 0$

The TV-norm boundedness established in Proposition 3 *i*) carries over to  $\beta > 0$  without modification. Indeed, the push-forward map preserves the total variation norm: for any  $\mu \in \mathcal{M}_+(\mathcal{X})$ ,

$$\|\mathbb{T}_{\nu, \beta}^{\sharp} \mu\|_{\text{TV}} = \|\mu\|_{\text{TV}},$$

since the push-forward merely redistributes mass without creating or destroying it. Therefore:

$$\|\nu_{k+}\|_{\text{TV}} = \|\mathbb{T}_{\nu, \beta}^{\sharp} \mathbb{W}_{\nu_k, \alpha} \nu_k\|_{\text{TV}} = \|\mathbb{W}_{\nu_k, \alpha} \nu_k\|_{\text{TV}} = \int_{\mathcal{X}} e^{-\alpha J'_{\nu_k}(t)} d\nu_k(t),$$

which is the same starting point as in the proof of Proposition 3 *i*) (Section C.1). The remainder of that proof—the case analysis on whether  $\|\nu_k\|_{\text{TV}}$  exceeds the threshold  $\mathfrak{M}$  defined there, leading to the bound  $\|\nu_k\|_{\text{TV}} \leq \mathfrak{C}_{\text{TV}}$ —applies verbatim, since it depends on  $\nu_k$  only through its TV-norm.

### E.2 Smoothness assumptions for $\beta > 0$

We verify that Assumptions (14) remain valid when the transition  $\nu_{k+} \mapsto \nu_{k+1}$  is applied after a push-forward update with  $\beta > 0$ .

**Assumption** ( $\mathcal{H}_{\varepsilon}^+$ ). This assumption concerns the birth process  $\nu_{k+} \mapsto \nu_{k+1}$  only, and requires that

$$\nu_{k+1} \mathbf{1}_{\{J'_{\nu_{k+}} \leq 0\}} \geq \lambda \varepsilon_k \mathbf{1}_{\{J'_{\nu_{k+}} \leq 0\}}.$$

The definition of the sets  $\mathcal{N}_{\nu_{k+}} = \{J'_{\nu_{k+}} \leq 0\}$  depends on the post-update measure  $\nu_{k+}$ , which now incorporates the push-forward. This is the most delicate point and is discussed separately in Section E.5 below.

**Assumption** ( $\mathcal{H}_{\varepsilon}^{\text{smooth}, 1}$ ). The proof of Proposition 3 *ii*) (Section C.2) relies on the decomposition

$$J(\nu_{k+1}) - J(\nu_{k+}) \leq \|\Phi(\nu_{k+} \mathbf{1}_{\mathcal{P}_{\nu_{k+}}})\|_{\mathbb{H}}^2 + \|\Phi(\varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda)\|_{\mathbb{H}}^2,$$

where  $\mathcal{P}_{\nu_{k+}} \subset \{J'_{\nu_{k+}} > 0\}$  and  $\mathcal{N}_{\nu_{k+}} \subset \{J'_{\nu_{k+}} < 0\}$ . These set inclusions hold by definition regardless of  $\beta$ . The subsequent bounds rely on the TV-norm of  $\nu_{k+}$  (which is preserved by the push-forward, as shown above) and on the inclusion  $\mathcal{P}_{\nu_{k+}} \subset \{J'_{\nu_k} > -2\alpha^{-1} \log \varepsilon_k\}$

of Remark 2.1; the latter inclusion is sensitive to  $\beta$  through the relationship between  $\nu_k$  and  $\nu_{k+}$ , and its  $\beta > 0$  counterpart is established in Section E.5 below. Modulo this re-verification, Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$  holds with the same constant  $\mathfrak{C}$  up to a bounded multiplicative factor  $e^{\alpha\beta\mathfrak{L}^2}$ .

**Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},2})$ .** The bound

$$\|J'_{\nu_{k+1}} - J'_{\nu_{k+}}\|_\infty \leq \|\nu_k\|_{\text{TV}}\varepsilon_k^2 + \varepsilon_k\lambda(\mathcal{X})$$

is obtained via the Cauchy–Schwarz inequality applied to the feature map  $\Phi$ , and depends on  $\nu_{k+}$  only through its TV-norm. The argument is therefore identical to the  $\beta = 0$  case.

### E.3 Extended one-step descent for $\beta > 0$

**Proposition 16** *Under Assumptions (14) and  $(15-(\mathcal{H}_{\text{TV}}^\infty))$ , if  $(\alpha, \beta)$  satisfies (7) and  $v_{k-1+}^2 \geq 24\varepsilon_{k-1}^2\mathfrak{C}^2$ , then:*

$$J(\nu_{k+}) - J(\nu_k) \leq -\frac{3\alpha}{2}(2\mathfrak{L})^{-d}|v_{k-1+}|^{2+d}\varepsilon_{k-1}.$$

**Proof** The proof follows that of Proposition 6 with a single modification: we apply Proposition 2 with  $\beta > 0$  instead of  $\beta = 0$ . The descent property yields:

$$J(\nu_{k+}) - J(\nu_k) \leq -\frac{3}{4}\left(\alpha \int_{\mathcal{X}} |J'_{\nu_k}|^2 d\nu_k + \beta \int_{\mathcal{X}} \|\pi_{\mathcal{X}}(t, \nabla J'_{\nu_k}(t), \beta)\|^2 d\nu_k\right).$$

Since the second term is non-positive, dropping it only weakens the upper bound by an amount of additional descent, and we obtain:

$$J(\nu_{k+}) - J(\nu_k) \leq -\frac{3}{4}\alpha \int_{\mathcal{X}} |J'_{\nu_k}|^2 d\nu_k = -\frac{3}{4}\|g_{\nu_k}^\alpha\|_{L^2(\nu_k)}^2.$$

This is exactly inequality (42a), and the remainder of the proof of Proposition 6 (Eqs (42b)–(42d)) proceeds without change.  $\blacksquare$

### E.4 Convergence rates for $\beta > 0$

Since Proposition 16 yields the same bound as Proposition 6, the downstream results—namely Proposition 8 and Theorem 2.1—extend to  $\beta > 0$  with identical rates.

The proof of Theorem 2.1 (Section C.4) uses two ingredients: (i) Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$ , which is verified in Section E.2, and (ii) the descent property  $J(\nu_{k+}) - J(\nu_k) \leq 0$ , which follows from Proposition 2 for any  $(\alpha, \beta)$  satisfying (7). The telescoping sum argument and the optimization over  $\varepsilon_k$  are unchanged.

### E.5 The screening issue for $\beta > 0$

We now address the most delicate point: verifying that the birth process construction of Section 2.2 still satisfies Assumption  $(\mathcal{H}_\varepsilon^+)$  and that the smoothness bounds (14) remain valid when  $\nu_{k+}$  is computed with  $\beta > 0$ .

**Assumption** ( $\mathcal{H}_\varepsilon^+$ ) **is satisfied by construction.** Recall that the mass creation step (20) adds  $\varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}}$   $\lambda$  on  $\mathcal{N}_{\nu_{k+}} = \{J'_{\nu_{k+}} \leq 0\}$ . Since  $\nu_{k+1} = \nu_{k++} + \varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}}$   $\lambda$  with  $\nu_{k++} \geq 0$ , we obtain

$$\nu_{k+1} \mathbf{1}_{J'_{\nu_{k+}} \leq 0} \geq \varepsilon_k \lambda \mathbf{1}_{J'_{\nu_{k+}} \leq 0},$$

which is precisely Assumption ( $\mathcal{H}_\varepsilon^+$ ). This holds regardless of  $\beta$ , since the birth process is defined in terms of the actual measure  $\nu_{k+}$ , whatever its construction.

**Perturbation of  $J'_{\nu_{k+}}$  by the push-forward.** The more subtle issue arises in the smoothness proofs (Section C.2), which use the specific relationship between  $\nu_k$ ,  $\nu_{k+}$ , and the set  $\mathcal{P}_{\nu_{k+}}$ . We denote by  $\tilde{\nu}_{k+} = \mathbb{W}_{\nu_k, \alpha} \nu_k$  the intermediate reweighted measure (before position update), so that  $\nu_{k+} = \mathbb{T}_{\nu, \beta}^\# \tilde{\nu}_{k+}$ .

We control the perturbation induced by the push-forward using a Taylor expansion. For any  $t \in \mathcal{X}$ :

$$J'_{\nu_{k+}}(t) - J'_{\tilde{\nu}_{k+}}(t) = \langle \varphi_t, \Phi(\nu_{k+} - \tilde{\nu}_{k+}) \rangle_{\mathbb{H}}. \quad (53)$$

Since  $\nu_{k+} = \mathbb{T}_{\nu, \beta}^\# \tilde{\nu}_{k+}$ , the change-of-variables formula gives:

$$\Phi(\nu_{k+} - \tilde{\nu}_{k+}) = \int_{\mathcal{X}} \left[ \varphi_{\mathbb{T}_{\nu, \beta}(t)} - \varphi_t \right] d\tilde{\nu}_{k+}(t).$$

By the Lipschitz property of the feature map (Lemma A.2),  $\|\varphi_{\mathbb{T}_{\nu, \beta}(t)} - \varphi_t\|_{\mathbb{H}} \leq \sqrt{\mathfrak{C}_{\mathcal{P}}} \|\mathbb{T}_{\nu, \beta}(t) - t\|$ . Recalling the definition  $\mathbb{T}_{\nu, \beta}(t) = t - \beta \pi_{\mathcal{X}}(t, \nabla J'_{\nu_k}(t), \beta)$  and the bound  $\|\pi_{\mathcal{X}}(t, \nabla J'_{\nu_k}(t), \beta)\| \leq \|\nabla J'_{\nu_k}\|_{\infty} \leq \mathfrak{L}$  (Lemma A.5), we obtain:

$$\|\Phi(\nu_{k+} - \tilde{\nu}_{k+})\|_{\mathbb{H}} \leq \beta \sqrt{\mathfrak{C}_{\mathcal{P}}} \mathfrak{L} \|\tilde{\nu}_{k+}\|_{\text{TV}} \leq \beta \sqrt{\mathfrak{C}_{\mathcal{P}}} \mathfrak{L} \mathfrak{C}_{TV}.$$

Combined with (53) and  $\|\varphi_t\|_{\mathbb{H}} = 1$ :

$$\|J'_{\nu_{k+}} - J'_{\tilde{\nu}_{k+}}\|_{\infty} \leq \beta \sqrt{\mathfrak{C}_{\mathcal{P}}} \mathfrak{L} \mathfrak{C}_{TV}. \quad (54)$$

**Extension of the smoothness bounds.** It remains to verify that Assumptions ( $\mathcal{H}_\varepsilon^{\text{smooth}, 1}$ ) and ( $\mathcal{H}_\varepsilon^{\text{smooth}, 2}$ ) still hold with the same order. In both cases, the key observation is the following: with  $\beta > 0$ , integrals against  $\nu_{k+}$  over a set  $\mathcal{A} \subset \mathcal{X}$  are expressed via the push-forward as

$$\int_{\mathcal{A}} d\nu_{k+}(t') = \int_{\mathbb{T}_{\nu, \beta}^{-1}(\mathcal{A})} e^{-\alpha J'_{\nu_k}(t)} d\nu_k(t).$$

On the preimage  $\mathbb{T}_{\nu, \beta}^{-1}(\mathcal{P}_{\nu_{k+}})$ , since  $\mathbb{T}_{\nu, \beta}(t) \in \mathcal{P}_{\nu_{k+}} \subset \{J'_{\nu_k} > -2\alpha^{-1} \log \varepsilon_k\}$  (by Remark 2.1), the Lipschitz continuity of  $J'_{\nu_k}$  (Lemma A.5) yields:

$$J'_{\nu_k}(t) \geq J'_{\nu_k}(\mathbb{T}_{\nu, \beta}(t)) - \mathfrak{L} \|\mathbb{T}_{\nu, \beta}(t) - t\| > -2\alpha^{-1} \log \varepsilon_k - \beta \mathfrak{L}^2, \quad (55)$$

using the bound  $\|\mathbb{T}_{\nu, \beta}(t) - t\| = \beta \|\pi_{\mathcal{X}}(t, \nabla J'_{\nu_k}(t), \beta)\| \leq \beta \mathfrak{L}$  (Lemma A.5). Therefore:

$$\forall t \in \mathbb{T}_{\nu, \beta}^{-1}(\mathcal{P}_{\nu_{k+}}) : \quad e^{-\alpha J'_{\nu_k}(t)} < e^{\alpha \beta \mathfrak{L}^2} \varepsilon_k^2. \quad (56)$$

Under condition (7), the product  $\alpha \beta \mathfrak{L}^2$  is uniformly bounded by a constant depending only on  $\mathfrak{C}_{\mathcal{P}}$ ,  $\mathfrak{C}_{TV}$ , and  $\kappa$ .

Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$ . As in the proof of Proposition 3 *ii*) (Section C.2), the Cauchy–Schwarz inequality gives:

$$\|\Phi(\nu_{k+} \mathbf{1}_{\mathcal{P}_{\nu_{k+}}})\|_{\mathbb{H}}^2 = \left\| \int_{\mathcal{P}_{\nu_{k+}}} \varphi_{t'} d\nu_{k+}(t') \right\|_{\mathbb{H}}^2 \leq \nu_{k+}(\mathcal{P}_{\nu_{k+}}) \times \int_{\mathcal{P}_{\nu_{k+}}} \|\varphi_{t'}\|_{\mathbb{H}}^2 d\nu_{k+}(t').$$

Since  $\|\varphi_{t'}\|_{\mathbb{H}} = 1$ , both factors equal  $\nu_{k+}(\mathcal{P}_{\nu_{k+}})$ . Using the push-forward and the bound (56):

$$\nu_{k+}(\mathcal{P}_{\nu_{k+}}) = \int_{\Gamma_{\nu,\beta}^{-1}(\mathcal{P}_{\nu_{k+}})} e^{-\alpha J'_{\nu_k}(t)} d\nu_k(t) \leq e^{\alpha\beta\Omega^2} \varepsilon_k^2 \|\nu_k\|_{\text{TV}}.$$

Therefore  $\|\Phi(\nu_{k+} \mathbf{1}_{\mathcal{P}_{\nu_{k+}}})\|_{\mathbb{H}}^2 \leq e^{2\alpha\beta\Omega^2} \varepsilon_k^4 \|\nu_k\|_{\text{TV}}^2 \leq \mathfrak{C} \varepsilon_k^2 \|\nu_k\|_{\text{TV}}^2$ , where the last inequality uses  $\varepsilon_k \leq 1$ . The bound on  $\|\Phi(\varepsilon_k \mathbf{1}_{\mathcal{N}_{\nu_{k+}}} \lambda)\|_{\mathbb{H}}^2 \leq \varepsilon_k^2 \lambda(\mathcal{X})^2$  is unchanged (it depends only on  $\varepsilon_k$  and  $\lambda(\mathcal{X})$ ). Therefore, Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},1})$  holds with a modified constant  $\mathfrak{C}$ .

Assumption  $(\mathcal{H}_\varepsilon^{\text{smooth},2})$ . Following the proof in Section C.2:

$$\left| J'_{\nu_{k+1}}(t) - J'_{\nu_{k+}}(t) \right| \leq \int_{\mathcal{P}_{\nu_{k+}}} |\langle \varphi_u, \varphi_t \rangle_{\mathbb{H}}| d\nu_{k+}(u) + \varepsilon_k \int_{\mathcal{N}_{\nu_{k+}}} |\langle \varphi_u, \varphi_t \rangle_{\mathbb{H}}| d\lambda(u).$$

The second term is bounded by  $\varepsilon_k \lambda(\mathcal{X})$  as before. For the first term, using  $|\langle \varphi_u, \varphi_t \rangle| \leq 1$  and the push-forward:

$$\int_{\mathcal{P}_{\nu_{k+}}} d\nu_{k+}(u) = \int_{\Gamma_{\nu,\beta}^{-1}(\mathcal{P}_{\nu_{k+}})} e^{-\alpha J'_{\nu_k}(t)} d\nu_k(t) \leq e^{\alpha\beta\Omega^2} \varepsilon_k^2 \|\nu_k\|_{\text{TV}},$$

where we used (56). This yields:

$$\|J'_{\nu_{k+1}} - J'_{\nu_{k+}}\|_{\infty} \leq e^{\alpha\beta\Omega^2} \|\nu_k\|_{\text{TV}} \varepsilon_k^2 + \varepsilon_k \lambda(\mathcal{X}) \leq \mathfrak{C} \varepsilon_k,$$

since  $e^{\alpha\beta\Omega^2}$  is bounded under (7) and  $\varepsilon_k \leq 1$ .

## E.6 Global convergence with position updates

We now state the main result of this section, which extends Theorem 2.1 to the case  $\beta > 0$ . The transition  $\nu_k \mapsto \nu_{k+}$  is given by the full Weight & Push-Forward update (52), and the birth process  $\nu_{k+} \rightarrow \nu_{k+1}$  follows the construction of Section 2.2.

**Theorem E.1 (Global convergence for  $\beta > 0$ )** *Assume (2) with  $\kappa \geq 0$ . Let the sequence  $(\nu_k)_{k \geq 1}$  be generated by the update  $\nu_{k+} = \Gamma_{\nu,\beta}^{\sharp} \mathbb{W}_{\nu_k, \alpha} \nu_k$  followed by the birth process (17)–(20), with  $(\varepsilon_k)_{k \geq 0}$  satisfying  $\varepsilon_k \leq \alpha$  for every  $k \geq 0$ . If  $(\alpha, \beta)$  satisfies condition (7), then Assumptions (14) and (15– $(\mathcal{H}_{\text{TV}}^{\infty})$ ) hold (with constants  $\mathfrak{C}$  depending on  $\alpha, \beta, \mathfrak{C}_{\mathcal{P}}, \mathfrak{C}_{\text{TV}}$ , and  $\kappa$ , and differing from the  $\beta = 0$  case by bounded multiplicative factors of order  $e^{\alpha\beta\Omega^2}$ ), and for any final horizon  $K \geq 2$ :*

*i) If  $(\varepsilon_k)_{k \geq 0}$  is non-adaptive and  $\varepsilon_k = \varepsilon = \sqrt{\mathfrak{C}/K} \leq \alpha$  for all  $k \in \{1, \dots, K\}$ , then:*

$$\min_{1 \leq k \leq K} \{J(\nu_k) - J(\nu^*)\} \leq \mathfrak{C} \mathfrak{L}^{\frac{2+2d}{2+d}} \alpha^{-\frac{1}{2+d}} K^{-\frac{1}{2(2+d)}}.$$

ii) If  $(\varepsilon_k)_{k \geq 0}$  is horizon-free and  $\varepsilon_k = \sqrt{\mathfrak{C}/(k+1)} \leq \alpha$ , then:

$$\min_{1 \leq k \leq K} \{J(\nu_k) - J(\nu^*)\} \leq \mathfrak{C} \mathfrak{L}^{\frac{2+2d}{2+d}} \alpha^{-\frac{1}{2+d}} K^{-\frac{1}{2(2+d)}} \log(K)^{\frac{1}{(2+d)}}.$$

iii) If  $\varepsilon_k = \varepsilon = \mathfrak{C} \left( \frac{\mathfrak{L}^{2+2d}}{(d+1)\alpha} \right)^{\frac{1}{5+2d}} K^{-\frac{3+d}{5+2d}}$ , then:

$$J(\nu_K) - J(\nu^*) \leq \mathfrak{C} \left( \frac{\mathfrak{L}^{2+2d}}{(d+1)\alpha} \right)^{\frac{2}{5+2d}} K^{-\frac{1}{5+2d}}.$$

In all three items, the generic constant  $\mathfrak{C}$  may depend polynomially on  $(\|\nu^*\|_{\text{TV}}/\mathfrak{L})$ ; in item iii),  $\mathfrak{C}$  also depends polynomially on the initial excess  $J(\nu_1) - J^*$  (which is finite under  $(\mathcal{H}_{\text{TV}}^\infty)$ ).

**Proof** We have established in Sections E.1–E.5 that Assumptions (14) and (15– $(\mathcal{H}_{\text{TV}}^\infty)$ ) hold under the update (52) with  $\beta > 0$ , with constants  $\mathfrak{C}$  that may differ from the  $\beta = 0$  case by bounded multiplicative factors (specifically, factors of  $e^{\alpha\beta\mathfrak{L}^2}$ , which are uniformly bounded under (7)).

With these assumptions verified, Proposition 16 yields the same one-step descent bound as Proposition 6 (up to the modified constants), and the proof of Theorem 2.1 (Section C.4) applies verbatim: the telescoping sum argument (45d) and the optimization over  $\varepsilon_k$  depend only on the structure of Assumptions (14), not on the specific value of  $\beta$ . The three convergence rates therefore coincide with those of Theorem 2.1.  $\blacksquare$

## Appendix F. Statistical guarantees

**Definition F.1 (Sparse target measures)** Let  $\Delta^0$  be positive and let  $s^0$  be greater than 1. We define the class of  $s^0$ -sparse measures with minimal Euclidean separation  $\Delta^0$  as

$$\mathbb{M}_{s^0, \Delta^0} := \left\{ \mu^0 : \mu^0 = \sum_{k=1}^{s^0} a_k^0 \delta_{x_k^0} \text{ and } \min_{k \neq l} \|x_k^0 - x_l^0\|_2 \geq \Delta^0 \right\}, \quad (57)$$

where  $\delta_x$  denotes the Dirac mass at point  $x \in \mathcal{X}$  and at least one  $a_k^0 \in \mathbb{R}$  is non-zero.

In statistical learning theory, one considers a target sparse measure  $\mu^0 \in \mathbb{M}_{s^0, \Delta^0}$  and one defines the noise term and, respectively, the noise level as

$$\Gamma := y - \Phi \mu^0 \quad (58a)$$

$$\text{resp. } \gamma := \|\Gamma\|_{\mathbb{H}}. \quad (58b)$$

The statistical estimation error bounds of  $(\mathcal{P})$  are defined by means of the so-called far and near regions.

**Definition F.2 (Far and Near regions)** Let  $\mu^0 \in \mathbb{M}_{s^0, \Delta^0}$  and let  $r > 0$ . Define the near region of  $x_k^0$  of radius  $r$  as

$$\mathbb{N}_k(r) := \{x \in \mathcal{X}, \quad \|x - x_k^0\|_2 \leq r\}, \quad (59a)$$

and the far region as

$$\mathbb{F}(r) := \mathcal{X} \setminus \mathbb{N}(r), \quad \text{with: } \mathbb{N}(r) := \bigcup_{k=1}^{s^0} \mathbb{N}_k(r). \quad (59b)$$

In the literature (Azais et al., 2015; Candès and Fernandez-Granda, 2014; Poon et al., 2023; De Castro et al., 2025b), the estimation errors (with respect to the Euclidean metric) are proven to be, under some conditions, for some radius  $r > 0$  such that  $r < \Delta^0/2$ ,

- Control of the far region:

$$|\mu^*|(\mathbb{F}(r)) \lesssim_d \gamma \sqrt{s^0}, \quad (60a)$$

- Control of all the near regions:

$$|\mu^*(\mathbb{N}_k(r)) - a_k^0| \lesssim_d \gamma \sqrt{s^0}, \quad (60b)$$

- Detection level: For every Borel set  $A \subset \mathcal{X}$  such that  $|\mu^*|(A) \gtrsim_d \gamma \sqrt{s^0}$ , there exists  $x_k^0$  such that

$$\min_{t \in A} \|t - x_k^0\|_2 \lesssim_d r, \quad (60c)$$

where  $\lesssim_d$  denotes the inequality up to a multiplicative constant that may depend on the dimension  $d$ ,  $\mu^*$  is a solution to  $(\mathcal{P})$  with regularization parameter  $\kappa \sim \gamma/\sqrt{s^0}$  and  $|\mu^*|$  denotes the absolute part of  $\mu^*$ . These statistical estimation error bounds hold for all  $\mu \in \mathcal{M}(\mathcal{X})$  such that

$$0 \leq J(\mu) - J(\mu^*) \leq J(\mu^0) - J(\mu^*) =: \varepsilon^*, \quad (61)$$

see for instance (De Castro et al., 2024, Theorem 1), under some conditions (see (Poon et al., 2023, Assumption 1)). Extensions to Fisher-Rao metrics are given in Poon et al. (2023) and Giard et al. (2025).

We call an  $\varepsilon$ -solution any  $\mu \in \mathcal{M}(\mathcal{X})$  such that  $0 \leq J(\mu) - J(\mu^*) \leq \varepsilon$ . From an optimization point of view, Equation (61) shows that there exists  $\varepsilon^* > 0$  such that any  $\varepsilon$ -solution satisfies the statistical error bounds (60), for  $0 \leq \varepsilon \leq \varepsilon^*$ . While the parameter  $\varepsilon^*$  is not observed in practice, it shows that any gradient descent path converging towards a solution will satisfy the statistical error (60) after a finite number of steps.

## Appendix G. List of notation

<i>General Notation &amp; Measure Spaces</i>	
$d, \mathcal{X}$	Dimension and domain ( $\mathcal{X} \subset \mathbb{R}^d$ is compact convex)
$c, \mathfrak{C}$	Generic positive constants
$\lambda$	Lebesgue measure on $\mathcal{X}$
$\mathcal{C}(\mathcal{X})$	Space of continuous functions on $\mathcal{X}$ equipped with $\ \cdot\ _\infty$
$\mathcal{M}(\mathcal{X}), \mathcal{M}_+(\mathcal{X})$	Spaces of signed and non-negative Radon measures on $\mathcal{X}$ equipped with $\ \cdot\ _{\text{TV}}$
<i>Continuous Sparse Regression Framework</i>	
$\mathbb{H}, y$	Separable Hilbert space and observation vector in $\mathbb{H}$
$\Phi$	Linear measurement operator $\mathcal{M}(\mathcal{X}) \rightarrow \mathbb{H}$
$\varphi_t$	Feature map $t \mapsto \varphi_t \in \mathbb{H}$
$K(\cdot, \cdot)$	Model kernel defined by $K(s, t) = \langle \varphi_s, \varphi_t \rangle_{\mathbb{H}}$
$\kappa$	Regularization parameter of problem ( $\mathcal{P}$ )
$J(\nu)$	Objective function over measures
$J'_\nu$	Dual certificate (Fréchet derivative of $J$ at $\nu$ )
$\mu^*, \nu^*$	Optimal signed and non-negative measures (minimizers of $J$ )
$\rho_K$	Minimum excess loss along the deterministic trajectory, $\min_{0 \leq k \leq K} [J(\nu_k) - J(\nu^*)]$
<i>Regularity &amp; Complexity Constants</i>	
$c_{\mathcal{P}}, \mathfrak{C}_{\mathcal{P}}$	Kernel smoothness bounding constants (Assumption $\mathcal{H}_{\mathcal{P}}$ )
$\mathfrak{C}_{\text{TV}}$	Uniform bound on TV norm along trajectories (Assumption (15- $(\mathcal{H}_{\text{TV}}^\infty)$ ))
$\mathfrak{L}$	Uniform Lipschitz constant for $J'_\nu$ , see (4c)
$\mathbf{E}_\infty$	Uniform $L^\infty$ bound on the per-sample dual-certificate estimator (Assumption (A2))
<i>Conic Particle Gradient Descent Algorithm</i>	
$K$	Total number of iterations (horizon)
$p_k, p_0$	Number of active particles at iteration $k$ and at initialization
$\alpha, \beta$	Learning rates for weight and position updates
$\nu_k, \nu_{k+}, \nu_{k+1}$	Measure estimates at iteration $k$ , post-weight update, and next iteration
$\mathbf{W}_{\nu, \alpha}, \mathbf{T}_{\nu, \beta}$	Weight exponential update and position push-forward update operators
$\pi_{\mathcal{X}}$	Generalized gradient descent step on $\mathcal{X}$
$\Delta_k, \Delta_{k+}$	Potential energy descents along iterations
<i>Exploration: Birth and Death Processes</i>	
$\mathcal{P}_\nu, \mathcal{N}_\nu$	Regions of positivity (particle death) and negativity (particle birth) of $J'_\nu$
$\varepsilon_k$	Exploration schedule / birth rate parameter
$\widehat{C}_k, \widehat{c}_k$	Threshold parameters for particle death and birth processes
$U_{k+1}, V_{k+1}$	Random candidate points proposed for birth and for death at iteration $k+1$
<i>FSEP Stochastic Estimators</i>	
$\mathbf{W}, \mathbf{T}$	Vectors of particle weights ( $p_k$ components) and positions ( $p_k \times d$ )
$\hat{\nu}_k, \hat{\nu}_{k+}, \hat{\nu}_k^{++}, \hat{\nu}_{k+1}$	Stochastic counterparts of $\nu_k, \nu_{k+}, \nu_{k++}, \nu_{k+1}$ (current iterate, post-weight update, post-death, and next iterate)
$m_k, N$	Mini-batch size at iteration $k$ and total oracle count $N = \sum_{k=1}^K m_k$
$\widehat{J}'_k, \widehat{D}_k$	Stochastic estimators of the dual certificate and its gradient
$\widehat{\mathcal{P}}_{\nu_{k+}}, \widehat{\mathcal{N}}_{\nu_{k+}}$	Empirical regions of particle death and birth
$\widehat{\Delta}_k, \widehat{\Delta}_{k+}$	Empirical potential energy descents
$\mathfrak{F}_k, \mathfrak{F}_k^+$	Filtrations generated by the algorithm up to step $k$ and post-update $k^+$
$\hat{\rho}_K$	Minimum stochastic excess loss along $K$ iterations, see (27)