Riemannian Gradient Descent for Gaussian Mixture Models with unknown diagonal covariances

Romane Giard [1], Yohann De Castro [1,2], Roland Denis [1] and Clément Marteau [1]

[1] *Centrale Lyon, INSA Lyon, Université Lyon 1, Université Jean Monnet, CNRS, ICJ UMR5208, 69130 Ecully, France.*
[2] *Institut Universitaire de France (IUF)*

version as of September 25, 2026

Abstract

This paper investigates the numerical resolution of the Beurling-LASSO (BLASSO), a convex optimization framework that promotes sparsity in the space of measures. We consider its application to the estimation of Gaussian mixture models (GMMs) with an unknown number of components and unknown diagonal covariance matrices. Our approach combines the Conic Particle Gradient Descent (CPGD) principle with Riemannian gradient descent, to account for the underlying Fisher-Rao geometry of Gaussian distributions. Our contributions are twofold. First, we provide theoretical guarantees for the convergence of our algorithm. In particular, we establish exponential local convergence under a non-degeneracy condition on the solution and relate this assumption to a separation condition on the underlying statistical target. Second, we address practical implementation aspects of CPGD and present numerical experiments illustrating its performance. On the test cases considered, these experiments suggest that CPGD is more robust to overspecification of the number of components than the EM algorithm. We also investigate the impact of component separation on recovery accuracy.

Corresponding author: Romane Giard
E-mail address: `romane.giard@ec-lyon.fr`

1 Introduction

Our aim in this paper is to investigate the numerical and theoretical properties of a new algorithm designed to address a specific instance of the BLASSO. This first section recalls the motivation underlying this optimization problem and sheds light on the setting considered.

The Beurling-LASSO Let $\mathbb{L}$ be an Hilbert space with associated norm $\|\cdot\|_{\mathbb{L}}$. We consider an observation $y \in \mathbb{L}$ that we seek to decompose or approximate as a nonnegative linear combination of $p \in \mathbb{N}^*$ parameterized signals:

$$\sum_{j=1}^{p} \omega_j f_{x_j} \,, \tag{1}$$

where for any $j \in \{1, \dots, p\}$, the amplitude ω_j is nonnegative, and the location x_j is a real vector belonging to a compact set $\mathcal{X}$, indexing $f_x \in \mathbb{L}$. Our goal is to recover a sparse representation, in the sense that only a small number of weights ω_j are nonzero. This question can be naturally reformulated as the optimization problem

$$\underset{(\omega_j)_{j=1}^p \subset \mathbb{R}_+, (x_j)_{j=1}^p \subset \mathcal{X}}{\arg\min} \frac{1}{2} \left\| y - \sum_{j=1}^{p} \omega_j f_{x_j} \right\|_{\mathbb{L}}^2 . \tag{$\mathcal{P}'$}$$

However, we stress that ($\mathcal{P}'$) is highly nonconvex due to the joint optimization over amplitudes and locations. Moreover, nothing guarantees that the corresponding solution will be sparse.

To overcome these difficulties, the Beurling–LASSO (BLASSO), introduced by [De Castro and Gamboa, 2012] and further studied in, *e.g.*, [Duval and Peyré, 2015; Poon et al., 2023], lifts the problem to the space of measures via the embedding

$$(\omega_j)_{j=1}^p, (x_j)_{j=1}^p \mapsto \mu = \sum_{j=1}^{p} \omega_j \delta_{x_j} \,.$$

Defining the linear operator $\Psi : \mu \mapsto \int_{\mathcal{X}} f_x \, \mathrm{d}\mu(x)$, the BLASSO formulation then reads as

$$\operatorname*{arg\,min}_{\mu \in \mathcal{M}(\mathcal{X})^+} \frac{1}{2} \|y - \Psi\mu\|_{\mathbb{L}}^2 + \kappa \|\mu\|_{\mathrm{TV}} \ , \tag{$\mathcal{P}'_\kappa$}$$

where $\mathcal{M}(\mathcal{X})^+$ denotes the space of nonnegative Radon measures on $\mathcal{X}$. The problem ($\mathcal{P}'_\kappa$) includes a penalty term expressed in the total variation norm and controlled by the regularization parameter $\kappa > 0$. The total variation norm plays a role analogous to the ℓ^1-norm in finite dimensions and promotes sparsity of the solution (see, *e.g.*, [Candès and Fernandez-Granda, 2014]).

The formulation ($\mathcal{P}'_\kappa$) is convex over the space of measures and does not require specifying the number of components p in advance. As such, it has been the subject of several investigations, focusing either on the statistical performance of the solution $\mu^\star$ (w.r.t. to a given ground truth μ^0) or on algorithmic aspects of the underlying optimization problem. Nevertheless, almost all of these contributions have been developed in the setting where the kernel $(x, x') \mapsto K(x, x') = \langle \Psi\delta_x, \Psi\delta_{x'} \rangle_{\mathbb{L}}$ induced by the feature map Ψ is translation-invariant (see Section 2 for further details). In this work, we extend these investigations to a specific kernel that is not translation-invariant. Our motivation stems from the classical Gaussian mixture estimation problem. This extension raises interesting new challenges, both from analytical and algorithmic perspectives.

Gaussian mixtures and non-translation-invariant kernels The particular case of Gaussian mixture models (GMMs) provides an interesting example in which the kernel is not translation-invariant. This situation arises when the covariance structures are unknown. In the GMM setting, y represents an approximation of an underlying target density built from i.i.d. measurements $X_1, \ldots, X_n \in \mathbb{R}^d$. The terms f_x in (1) then correspond to Gaussian densities. The parameter x encodes both the mean and the diagonal covariance matrix.

Classically, GMM estimation is performed using the Expectation–Maximization (EM) algorithm [McLachlan and Krishnan, 1997], which maximizes the likelihood of the data. However, this optimization problem is inherently nonconvex and may suffer from local minima and sensitivity to initialization. In contrast, the BLASSO provides a convex alternative, offering a principled framework for sparse mixture recovery together with theoretical guarantees. The properties of the corresponding estimator have been investigated in [Giard et al., 2025].

From a methodological perspective, previous implementations in the BLASSO literature (see, *e.g.*, [De Castro et al., 2021; De Castro et al., 2023]) have focused on translation-invariant kernels. In the GMM framework, such a setting corresponds to the case where the covariance matrices associated with all components are equal (and known in advance). Imposing a known common covariance structure is quite restrictive and unrealistic from a practical standpoint. The framework developed in this contribution allows us to relax these constraints within the BLASSO methodology. In particular, we consider a specific operator Ψ that allows us to handle Gaussian components with unknown (diagonal) covariance structures. In this setting, we present an implementable algorithm supported by a theoretical convergence analysis, together with competitive numerical results.

Conic Particle Gradient Descent To address the BLASSO problem ($\mathcal{P}'_\kappa$), we develop and investigate, in Section 2.2, a new instance of the Conic Particle Gradient Descent (CPGD) algorithm introduced by [Chizat, 2022].

The CPGD approach relies on a discretization of the space $\mathcal{M}(\mathcal{X})^+$. Since this space cannot be directly parameterized, the optimization is carried out over discrete measures of the form $\sum_{j=1}^p \omega_j \delta_{x_j}$. The algorithm then updates the weights ω_j and positions x_j through gradient-based iterations. To enforce the nonnegativity constraint on the weights, a conic retraction step is applied. In our setting, particular attention is paid to the update of the parameters x_j. We exploit the underlying geometry induced by the Gaussian structure and consider updates based on the Fisher–Rao metric [Amari, 2016], leading to Riemannian gradient descent and natural gradient descent schemes.

Related works Conic Particle Gradient Descent belongs to a broader lineage of overparameterized particle methods for optimization over the space of measures. [Chizat and Bach, 2018] analyze the Wasserstein gradient flow that arises as the mean-field limit of gradient descent on the weights and positions of an overparameterized model. Under a homogeneity assumption and a separation property of the initialization that can only be attained in the infinite-width limit, any limit point of the flow is a global minimizer. This yields consistency-type guarantees for the continuous-time flow, without providing a convergence rate for a finite step size. The conic version studied by [Chizat, 2022] couples a multiplicative (mirror) update of the particle weights with an additive update of their positions; its mean-field limit is the gradient flow of the objective with respect to the Wasserstein-Fisher-Rao, or Hellinger-Kantorovich, metric. This cone geometry was introduced and developed independently by three groups [Liero et al., 2016; Liero et al., 2018; Chizat et al., 2018a; Chizat et al., 2018b; Kondratyev et al., 2016]. The Hellinger-Kantorovich distance is characterized through a lifting to the metric cone over the base space, and interpolates between pure transport (Wasserstein) and pure correlation

(Hellinger/Fisher-Rao). [Gallouët and Monsaingeon, 2017] realize the corresponding gradient flows through a JKO (Jordan-Kinderlehrer-Otto) splitting scheme that alternates between a Wasserstein transport substep and a Fisher–Rao correlation substep.

The theory available for CPGD is, however, sharply divided between the flow and the discrete scheme. At the continuous-time level, [Chizat, 2022] establishes asymptotic global convergence under a full-support initialization, but without a rate. For the iterates, a sharpness (Polyak-Lojasiewicz) inequality valid within a sublevel set yields only *local* exponential convergence; global convergence of the discrete descent additionally requires an initialization that densely samples the parameter space, which forces the number of particles to grow exponentially with the dimension d, together with a small position-to-weight step-size ratio $\frac{\beta}{\alpha}$ that can make the guaranteed local rate slow. A faster $O\left(\frac{\ln k}{k}\right)$ rate is available only for the pure-mirror update ($\beta = 0$, *i.e.* frozen positions) initialized from a smooth full-support density, whereas the general descent rate is $O\left(\frac{\ln k}{\sqrt{k}}\right)$. Stochastic and birth-death variants address scalability and this initialization burden: [De Castro et al., 2023] combine CPGD with stochastic gradient descent and random features, proving bounded total-variation norms along the trajectory and convergence to stationary points at rate $O\left(\frac{\ln k}{\sqrt{k}}\right)$; [De Castro et al., 2026] add a spawn-and-prune (birth-death) process, in the spirit of the birth-death dynamics of [Rotskoff et al., 2019], a non-local reweighting term that keeps the mean-field flow mass-conserving, and thereby obtain global convergence without an exponentially large initialization.

A widely used grid-based competitor is the Sliding Frank-Wolfe algorithm [Denoyelle et al., 2019], which interleaves conditional-gradient steps that add a spike at the global maximizer of the dual certificate with a non-convex "sliding" step jointly re-optimizing all amplitudes and positions. This algorithm terminates in finitely many iterations under uniqueness of the BLASSO solution together with a non-degeneracy hypothesis on its dual certificate. Both ingredients are costly in our setting: the per-iteration global maximization is typically performed by a grid search. The corresponding cost grows with the dimension of the parameter domain $\mathcal{X} \subset \mathbb{R}^d \times [u_{\min}, u_{\max}]^d$, whose points $x = (t, u)$ encode both means and diagonal scales, while the sliding subproblem is non-convex. Conic particle gradient descent instead moves particles continuously and avoids any per-iteration grid search. Such moving-particle dynamics also deliberately steer clear of the lazy, or tangent-kernel, regime [Chizat et al., 2019], in which a large output scaling keeps the parameters frozen near their initialization so that no features are learned. Relatedly, high-dimensional analyses of the effective dynamics of stochastic gradient descent characterize when fixed-step trajectories converge to suboptimal solutions from a random initialization [Ben Arous et al., 2022].

Against this backdrop, the non-translation-invariant kernel associated with unknown diagonal covariances induces a Fisher-Rao location-scale geometry into the position updates, making the dynamics Riemannian. In this setting, we establish local exponential convergence with explicit constants for the joint weight-and-position dynamics. We further connect the underlying non-degeneracy condition to an interpretable separation condition on the statistical target through a dual-certificate and Fisher-metric analysis in the spirit of off-the-grid super-resolution [Duval and Peyré, 2015; Poon et al., 2023].

Outline of the paper and main results Our contribution begins in Section 2, where we provide details on the framework under consideration. We first introduce a generic algorithm that encompasses several possible strategies.

Section 3 provides theoretical guarantees for the convergence of the CPGD algorithm. Building on the works of [Chizat, 2022; De Castro et al., 2023], we first establish convergence of the weights ω_j, provided that the initialization assigns sufficient mass near the BLASSO solution. We then establish exponential local convergence of CPGD under a non-degeneracy condition on the solution, relying on the convergence of the position parameters x_j. We further connect this non-degeneracy assumption—which depends on the (unknown) solution—to a more interpretable separation condition on the underlying statistical target. This connection relies on the analysis of dual certificates and their behavior, following ideas developed in [Duval and Peyré, 2015; Poon et al., 2023; Giard et al., 2025].

In Section 4, we turn to practical aspects of the method. We make explicit the relationship between GMMs and the BLASSO problem. We then present numerical experiments illustrating the performance of the algorithm in this setting. We implement CPGD using natural gradient descent to account for the underlying geometry of the problem. Our experiments highlight the robustness of CPGD to misspecification of the number of components, in contrast to the EM algorithm. We also investigate the influence of component separation and sample size on recovery accuracy: while larger sample sizes improve performance, insufficient separation between components significantly degrades recovery, and increasing the sample size in this regime brings only limited gains.

The code to reproduce the numerical illustrations of this article can be found online at [Giard and Denis, 2026]. All proofs and technical details are provided in the appendix to this paper.

2 Optimization problem

2.1 Framework

Functional framework We introduce the space $\mathbb{L}$ involved in the data fidelity term in $(\mathcal{P}'_\kappa)$. Let $\tau \in (0, u_{\min}]$ (playing the role of a smoothing parameter, see Section 4). We consider the Hilbert space $\mathbb{L}$ as the RKHS associated with the Gaussian kernel

$$\lambda : z \in \mathbb{R}^d \mapsto \frac{e^{-\frac{\|z\|_2^2}{2\tau^2}}}{(2\pi\tau^2)^{d/2}} .$$

In particular, by the Fourier characterization of translation-invariant RKHS [Christmann and Steinwart, 2008, Chapter 4]:

$$\mathbb{L} \coloneqq \left\{ f : \mathbb{R}^d \to \mathbb{R} \ : \ f \in L^2(\mathbb{R}^d) \text{ s.t. } \|f\|_{\mathbb{L}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\mathcal{F}[f](\xi)|^2}{\mathcal{F}[\lambda](\xi)} \mathrm{d}\xi < +\infty \right\} ,$$

with dot product

$$\forall f, g \in \mathbb{L}, \quad \langle f, g \rangle_{\mathbb{L}} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\overline{\mathcal{F}[f](\xi)} \mathcal{F}[g](\xi)}{\mathcal{F}[\lambda](\xi)} \mathrm{d}\xi , \tag{2}$$

where for the Fourier transform we adopt the convention $\mathcal{F}[f](\xi) = \int f(z) e^{-i\langle z, \xi \rangle} \, \mathrm{d}z$.

We denote by $x = (t_1, \ldots, t_d, u_1, \ldots, u_d)$ any element of $\mathbb{R}^d \times [u_{\min}, +\infty)^d$. We write $t = (t_k)_{k=1}^d$ and $u = (u_k)_{k=1}^d$. In the specific Gaussian mixture setting, $t \in \mathbb{R}^d$ (resp. $u \in [u_{\min}, +\infty)^d$) denotes the mean (resp. the vector of componentwise standard deviations, *i.e.* the square root of the diagonal of the covariance matrix) of a given component. We define the linear operator $\Psi : \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d)^+ \mapsto \mathbb{L}$ by

$$\Psi\mu : z = (z_1, \ldots, z_d) \in \mathbb{R}^d \mapsto \int_{\mathbb{R}^d \times [u_{\min}, +\infty)^d} \prod_{k=1}^d \frac{(2u_k^2 + \tau^2)^{1/4}}{(2\pi)^{1/4}(u_k^2 + \tau^2)^{1/2}} e^{-\frac{(z_k - t_k)^2}{2(u_k^2 + \tau^2)}} \, \mathrm{d}\mu(x) .$$

This operator maps a measure μ to a function in $\mathbb{L}$, obtained as a superposition of Gaussian-shaped functions. Each point $x = (t, u)$ indexes a function $\Psi\delta_x$. Note that these functions are not probability densities, but rather normalized features adapted to the RKHS structure (*i.e.* $\|\Psi\delta_x\|_{\mathbb{L}} = 1$ for any $x \in \mathcal{X}$). We will make explicit the relationship between this setting and the GMM estimation problem in Section 4.1. We refer also to [Giard et al., 2025] for more details and related properties. In particular, the kernel associated with Ψ naturally arises as a scalar product between two Gaussian components, while the lower bound $u_{\min}$ on the covariance matrices appears to be a standard constraint to avoid singular models.

Optimization problem Let $\mathcal{X}$ be a compact of $\mathbb{R}^d \times [u_{\min}, u_{\max}]^d$. We denote $\mathcal{M}(\mathcal{X})^+$ the space of nonnegative Radon measures on $\mathcal{X}$. Let $\kappa > 0$ (regularization parameter). We aim to approach $\mu_\omega^\star$ a solution to the BLASSO problem $(\mathcal{P}'_\kappa)$ that can be rewritten as

$$\operatorname*{arg\,min}_{\mu \in \mathcal{M}(\mathcal{X})^+} J_W(\mu) \quad \text{where} \quad J_W(\mu) \coloneqq \frac{1}{2} \|\Psi\mu - y\|_{\mathbb{L}}^2 + \kappa \|\mu\|_{\mathrm{TV}} , \tag{$\mathcal{P}_\kappa$}$$

with $y \in \mathbb{L}$ and $\kappa > 0$.

This problem can be interpreted as an infinite-dimensional sparse recovery problem over the space of measures. The first summand of J_W enforces fidelity to the data y, while the total variation norm $\|\mu\|_{\mathrm{TV}}$ promotes sparsity, favoring measures that are concentrated on a small number of points.

The operator Ψ plays the role of a forward model: solving $(\mathcal{P}_\kappa)$ amounts to recovering a sparse representation of y in terms of the parameterized family of functions $\Psi\delta_x$. The underlying statistical motivation for this problem is presented in Section 4.1.

This problem has a solution (not necessarily unique), see for instance [Giard et al., 2025, Proposition B.1]. In general, solutions are not necessarily discrete.

Kernel and geometry Problem $(\mathcal{P}_\kappa)$ is associated with the feature map Ψ. We consider the corresponding (normalized) kernel: for $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$,

$$K_{\mathrm{norm}}(x, x') = \langle \Psi\delta_x, \Psi\delta_{x'} \rangle_{\mathbb{L}} = \prod_{k=1}^d \frac{(2u_k^2 + \tau^2)^{1/4} (2{u'_k}^2 + \tau^2)^{1/4}}{(u_k^2 + {u'_k}^2 + \tau^2)^{1/2}} e^{-\frac{(t_k - t'_k)^2}{2(u_k^2 + {u'_k}^2 + \tau^2)}} . \tag{3}$$

Interestingly, this kernel induces a specific geometry over $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$. We will take advantage of this underlying geometry when designing our algorithm (see Section 2.2). Formally, the space $\mathbb{R}^d \times [u_{\min}, +\infty)^d$

can be endowed with the Fisher–Rao metric associated with K_{norm} [Giard et al., 2025]. For $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$, the corresponding metric tensor is defined by

$$\mathfrak{g}_x \coloneqq \nabla_1 \nabla_2 K_{\text{norm}}(x,x) = \operatorname{diag}\left(\frac{1}{2u_1^2+\tau^2}, \dots, \frac{1}{2u_d^2+\tau^2}, \frac{2u_1^2}{(2u_1^2+\tau^2)^2}, \dots, \frac{2u_d^2}{(2u_d^2+\tau^2)^2} \right). \tag{4}$$

This metric naturally induces norm and scalar product, defined by

$$\langle v, v' \rangle_x = v^T \mathfrak{g}_x v' \quad \text{and} \quad \|v\|_x = \sqrt{v^T \mathfrak{g}_x v} \qquad \forall\, v, v' \in \mathbb{R}^{2d},\ x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d .$$

We denote by $\gamma_{x \to x'}$ the Fisher-Rao geodesic parameterized between 0 and 1 connecting $x = \gamma_{x\to x'}(0)$ to $x' = \gamma_{x\to x'}(1)$. In particular, it minimizes $\int_0^1 \|\dot\gamma(y)\|^2_{\gamma(y)}\, \mathrm{d}y$ among smooth functions from $[0,1]$ to $\mathbb{R}^d \times [u_{\min}, +\infty)^d$ such that $\gamma(0) = x, \gamma(1) = x'$. The corresponding distance between any $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$ can then be defined as

$$\mathfrak{d}_{\mathfrak{g}}(x,x') = \|\dot\gamma_{x\to x'}(y)\|_{\gamma_{x\to x'}(y)} \quad \text{for any} \quad y \in [0,1].$$

Finally, we can define the associated Riemannian gradient, which will be of particular interest when designing our algorithm.

Definition 2.1 (Riemannian gradient)**.** Let $\psi \in \mathcal{C}^2(\mathbb{R}^d \times [u_{\min}, +\infty)^d)$. Let $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$. We define

$$\nabla^{\mathfrak{g}} \psi(x) = \mathfrak{g}_x^{-1} \nabla \psi(x) .$$

Additional details on the metric $\mathfrak{g}_x$ can be found in [Giard et al., 2025, Appendix H].

2.2 Optimization procedure

We have now all the ingredients to design our algorithm based on the CPGD principle. In this context, the Fréchet derivative of the target objective function plays an important role. We first recall its main properties and discuss how we can take advantage of this quantity to address our problem.

Fréchet derivative Let $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$ and $\mu \in \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d)^+$. The Fréchet derivative J'_μ of J_W verifies

$$J_W(\mu + \sigma) - J_W(\mu) = \int_{\mathbb{R}^d \times [u_{\min}, +\infty)^d} J'_\mu \, \mathrm{d}\sigma + \frac{1}{2} \|\Psi\sigma\|^2_{\mathbb{L}} . \tag{5}$$

for any $\sigma \in \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d)$ such that $\mu + \sigma \in \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d)^+$. In particular, there exists an explicit expression for this derivative, namely

$$J'_\mu(x) = -\langle \Psi\delta_x, y \rangle_{\mathbb{L}} + \langle \Psi\delta_x, \Psi\mu \rangle_{\mathbb{L}} + \kappa . \tag{6}$$

The Fréchet derivative is strongly connected to our optimization problem. Indeed, as proved in [Chizat, 2022, Proposition 3.1], the KKT conditions associated with the problem ($\mathcal{P}_\kappa$) impose that for $\mu^\star_\omega$ a solution to ($\mathcal{P}_\kappa$), $J'_{\mu^\star_\omega} \geq 0$ on $\mathcal{X}$ and $J'_{\mu^\star_\omega} = 0$ on $\operatorname{supp}(\mu^\star_\omega)$.

Discretization To numerically solve ($\mathcal{P}_\kappa$), we must discretize the space $\mathcal{M}(\mathcal{X})^+$, as it cannot be directly parameterized. We therefore restrict our attention to discrete measures supported on p particles.

The gradients of J_W with respect to the weights and the positions of these particles are linked to J'. This relationship is formalized in [Chizat, 2022, Section 2.1]. In particular, for any discrete measure $\mu = \sum_{j=1}^p \omega_j \delta_{x_j}$, we can check that

$$\partial_{\omega_j} J_W(\mu) = J'_\mu(x_j) \quad \text{and} \quad \nabla_{x_j} J_W(\mu) = \omega_j \nabla_{x_j} J'_\mu(x_j) \quad \forall j \in \{1, \dots, p\} .$$

Starting with an initialization $\mu^1_\omega = \sum_{j=1}^p \omega_j^k \delta_{x_j^k}$ with $p \in \mathbb{N}^*$ particles, the general CPGD principle consists in designing iteratively a sequence of measures (μ^k_ω) where at each step $k > 1$, weights and positions of the current measure $\mu^k_\omega = \sum_{j=1}^p \omega_j^k \delta_{x_j^k}$ are (roughly speaking) updated according to a gradient descent. Since our target is a positive measure, this gradient descent includes a conic retraction to constrain the nonnegativity of the weights. In the same time, the updates on the positions will take advantage of the underlying geometry of the considered problem.

Weight update Following [Chizat, 2022], we use a conic retraction for the weight updates, with step size $\alpha \geq 0$. In particular, given μ_ω^k for $k \geq 1$, we write

$$\omega_j^{k+1} = \omega_j^k e^{-\alpha J'_{\mu_\omega^k}(x_j^k)} \quad \forall j \in \{1, \dots, p\}\,. \tag{7}$$

Contrary to a direct gradient descent, this update ensures that the weights remain positive throughout the iterations. Note that ω_j^{k+1} can be written as the solution of

$$\operatorname*{arg\,min}_{\omega \in \mathbb{R}^+} \left[J'_{\mu_\omega^k}(x_j^k)(\omega - \omega_j^k) + \frac{1}{\alpha}\left(\omega_j^k - \omega + \omega \ln\left(\frac{\omega}{\omega_j^k}\right)\right)\right],$$

where the second term is the Bregman divergence of the entropy, leading to an entropic mirror descent step for the weights. We stress that the update (7) is in accordance with the KKT condition. Particles x_j associated with a negative Fréchet derivative are promoted (*i.e.* $\omega_j^{k+1} > \omega_j^k$) to secure mass guidance in regions where we observe a default in the KKT condition, while particles for which $J'_{\mu_\omega^k}(x_j^k) > 0$ are scaled down.

Position update The underlying Riemannian geometry associated with the statistical problem (see Section 2.1 above) gives different strategies for the position update. Formally, we use a custom operator $\mathcal{T}$ and a step size $\beta \geq 0$: $\mathcal{T}_\beta(x_j^k, J'_{\mu_\omega^k})$ denotes an update on the location-scale parameter of the jth particle, with step size β, depending on $J'_{\mu_\omega^k}$. We can think of three different approaches for this update:

$$\mathcal{T}_\beta(x_j^k, J'_{\mu_\omega^k}) = x_j^k - \beta \nabla_x J'_{\mu_\omega^k}(x_j^k)\,, \tag{EGD}$$

$$\mathcal{T}_\beta(x_j^k, J'_{\mu_\omega^k}) = x_j^k - \beta \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x_j^k)\,, \tag{NGD}$$

$$\mathcal{T}_\beta(x_j^k, J'_{\mu_\omega^k}) = \exp_{x_j^k}(-\beta \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x_j^k))\,. \tag{RGD}$$

These updates (illustrated in Figure 1) correspond respectively to Euclidean gradient descent (EGD), natural gradient descent (NGD) and Riemannian gradient descent (RGD). In the expression (RGD) above, $\exp_\cdot$ denotes the exponential map, *i.e.* $\exp_{x_0}(v) = \gamma(1)$ where γ is a Fisher-Rao geodesic with starting point $\gamma(0) = x_0$ and initial direction $\dot\gamma(0) = v$. In dimension $d = 1$, the Fisher-Rao geodesics are either straight lines parallel to $\{t = 0\}$ or semicircles centered on $\{u = 0\}$ [Giard et al., 2025, Appendix H.2].

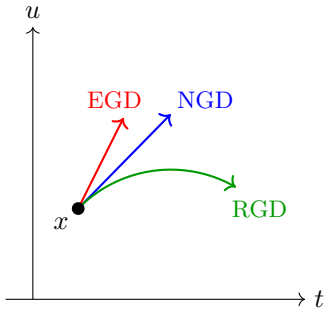

Figure 1: Illustration of the three update strategies in dimension $d = 1$. The parameters t (mean) and u (standard deviation) define a two-dimensional location–scale space (t, u). Each arrow starts at x and points to the corresponding updated position.

The updates (NGD) and (RGD) account for the intrinsic geometry of the space of Gaussian distributions (means and diagonal covariances). We refer to [Nielsen, 2020, Section 4.1] and [Amari, 1998]. Furthermore, the RGD update is invariant under any invertible smooth reparameterization of the parameter space. These two schemes are closely related: natural gradient descent can be viewed as a first-order approximation of Riemannian gradient descent when the step size β is small.

We emphasize that the existence of these different strategies is specific to our setting. In particular, when considering GMM with *known* and *identical* covariance structures, the Fisher-Rao metric reduces to the Euclidean one, and (EGD), (NGD), (RGD) lead in this case to similar descent schemes (up to the choice of the learning rate β). This is a strong difference with previous implementations of the BLASSO problem (see, *e.g.*, [De Castro et al., 2023; De Castro et al., 2026]), all of them being restricted to translation-invariant kernels. Considering a non-translation-invariant setting generates new questions that are addressed in the following.

Remark 2.1 (Ensuring well-defined updates). We stress that these updates are well-defined under certain conditions or modifications. We can for instance think of some projected version of (EGD) or (NGD), to keep

the location-scale parameters in $\mathbb{R}^d \times [u_{\min}, +\infty)^d$, or in $\mathcal{X}$. We could use the Euclidean projection, or the projection w.r.t. to the Fisher-Rao distance. Thinking about a projection step is particularly important for (RGD), because the exponential map is not defined everywhere (*i.e.* $\exp_{x_0}(v)$ is defined for all v such that $\|v\|_{x_0} \leq c_{x_0}$, with $c_{x_0} > 0$ depending on x_0, but not for all $v \in \mathbb{R}^{2d}$). Moreover, alternative instances can be considered using alternative optimization tools or strategies. A solution involving freezing the particles outside a certain set will be presented in Section 3.2, while numerical experiments displayed in Section 4 use AdaGrad strategies for the calibration of the learning rates α and β.

A general algorithm Gathering the weights and position update strategies discussed above, Algorithm 1 proposes a generic instance of the CPGD algorithm adapted to our framework.

Algorithm 1 CPGD

Require: Step sizes $\alpha, \beta \geq 0$, number of iterates K, initialization $\mu_\omega^1 = \sum_{j=1}^p \omega_j^1 \delta_{x_j^1}$

for $k = 1, \dots K-1$ **do**
  for $j = 1, \dots, p$ **do**
    $\omega_j^{k+1} \leftarrow \omega_j^k e^{-\alpha J'_{\mu_\omega^k}(x_j^k)}$ ▷ Weight update
    $x_j^{k+1} \leftarrow \mathcal{T}_\beta(x_j^k, J'_{\mu_\omega^k})$ ▷ Location-scale update
  end for
  $\mu_\omega^{k+1} \leftarrow \sum_{j=1}^p \omega_j^{k+1} \delta_{x_j^{k+1}}$
end for

Geometry and gradient descent The relative updates for the weights and positions are adapted from [Chizat, 2022]. They rely on the cone metric, whose metric tensor at point $(\omega, x) \in \mathbb{R}_+^* \times \mathbb{R}^{2d}$, and for step sizes $\alpha, \beta > 0$ is given by

$$\operatorname{diag}\left(\frac{\alpha^{-1}}{\omega}, \beta^{-1}\omega, \dots, \beta^{-1}\omega\right) \in \mathbb{R}^{(1+2d)\times(1+2d)}.$$

This choice induces multiplicative updates in ω, while the updates in x are independent of ω. In addition, a mirror retraction is applied to ω in order to ensure its positivity. The retraction (EGD) together with the weight update (7) is no more than a natural gradient descent for the cone metric and using a logarithmic reparameterization ($\tilde{\omega} = \ln \omega$).

On the other hand, using (NGD) corresponds to the more intricate metric

$$\operatorname{diag}\left(\frac{\alpha^{-1}}{\omega}, \beta^{-1}\omega\frac{1}{2u^2+\tau^2}, \dots, \beta^{-1}\omega\frac{1}{2u^2+\tau^2}, \beta^{-1}\omega\frac{2u^2}{(2u^2+\tau^2)^2}, \dots, \beta^{-1}\omega\frac{2u^2}{(2u^2+\tau^2)^2}\right),$$

which combines both the cone and Fisher-Rao metrics.

For more details on the cone metric and related properties, we refer to [Liero et al., 2016; Liero et al., 2018; Chizat et al., 2018a; Chizat et al., 2018b; Kondratyev et al., 2016].

3 Theoretical analysis of the algorithm

This section provides a theoretical analysis of the CPGD principle in our setting. We investigate the convergence properties of Algorithm 1 and one of its variant under generic conditions.

3.1 Convergence of the objective values via the weights dynamic

We first establish a sublinear $O\left(\frac{\ln K}{\sqrt{K}}\right)$ rate on the objective values of the Cesàro average, driven by the weight dynamics only. More precisely, if μ_ω^1 is constructed on an grid dense enough in $\mathcal{X}$, then we can expect to recover the target measure. We revisit here [De Castro et al., 2023, Theorem 1.2] and part of the proof of [Chizat, 2022, Theorem 4.1] in our specific setting. At this step, we do not select a specific position update among (EGD), (NGD), (RGD), but work instead with a generic scheme under general conditions (in particular, freezing positions would work). The proof of the following proposition is displayed in Appendix B.

Proposition 3.1. *Choice of initialization: Assume that* $\mu_\omega^1 = \sum_{j=1}^p \omega_j^1 \delta_{x_j^1}$ *with* $\omega_j^1 > 0$ *and* $x_j^1 \in \mathcal{X}$. *Moreover, we suppose that there exists disjoint regions* $\{\mathcal{X}_j\}_{j=1}^p \subset \mathcal{X}$ *such that for all* $j = 1, \dots, p$, $x_j^1 \in \mathcal{X}_j$ *and* $\sup_{\substack{x \in \mathcal{X}_j, \\ j=1,\dots,p}} \mathfrak{d}_{\mathfrak{g}}(x_j^1, x) \leq r_{\mu_\omega^1, \mathcal{X}}$ *for a radius* $r_{\mu_\omega^1, \mathcal{X}} > 0$. *We write* $\mathcal{X}_0 \coloneqq \mathcal{X} \setminus \left(\cup_{j=1}^p \mathcal{X}_j\right)$.
Choice of location-scale update: We choose $\mathcal{T}$ *and* β *such that for* $\mu_\omega^k = \sum_{j=1}^p \omega_j^k \delta_{x_j^k} \in \mathcal{M}(\mathcal{X})^+$, $x_j^{k+1} =$

$\mathcal{T}_\beta(x_j^k, J'_{\mu_\omega^k}) \in \mathcal{X}$ *and* $\mathfrak{d}_{\mathfrak{g}}(x_j^k, x_j^{k+1}) \le C_{\mathcal{T}}\beta$.

Control of the Cesàro average: Let $K \ge 2$. *Let* $\kappa \le 1$. *If* $\alpha = \frac{1}{\sqrt{K}}$ *and* $0 \le \beta \le \frac{1}{K^{3/2} C_{\mathcal{T}}}$, *then*

$$J_W\left(\frac{1}{K}\sum_{k=1}^{K}\mu_\omega^k\right) - J_W(\mu_\omega^\star) \lesssim_{\|\mu_\omega^\star\|_{\mathrm{TV}}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \frac{H_{\mu_\omega^1, \mu_\omega^\star} + 1}{\sqrt{K}} + r_{\mu_\omega^1, \mathcal{X}} + \mu_\omega^\star(\mathcal{X}_0) \tag{8a}$$

where

$$H_{\mu_\omega^1, \mu_\omega^\star} = \sum_{\substack{j \in \{1,\dots,p\}, \\ \mu_\omega^\star(\mathcal{X}_j) > 0}} \left(-\ln\left(\frac{\omega_j^1}{\mu_\omega^\star(\mathcal{X}_j)}\right)\mu_\omega^\star(\mathcal{X}_j) - \mu_\omega^\star(\mathcal{X}_j)\right) + \left\|\mu_\omega^1\right\|_{\mathrm{TV}} . \tag{8b}$$

Proposition 3.1 describes the convergence of the algorithm. In particular, it provides a control on the performances of the Cesaro average $\frac{1}{K}\sum_{k=1}^{K}\mu_\omega^k$ in terms of the quantities displayed in the r.h.s. of (8a). Inside this bound, the term $H_{\mu_\omega^1, \mu_\omega^\star}$ measures the divergence between the target $\mu_\omega^\star$ and the initialization. The smaller this divergence, the faster the convergence. This upper bound is completed by the mass of $\mu_\omega^\star$ outside the search space $\mathcal{X}_0$ and the radius $r_{\mu_\omega^1, \mathcal{X}}$. In particular, the grid should be thin enough to guarantee a good recovery precision.

Note that Proposition 3.1 only establishes convergence of the objective values, in the Cesàro sense. In particular, it does not guarantee that μ_ω^k converges to $\mu_\omega^\star$. When the descent property holds, *i.e.* when $k \mapsto J_W(\mu_\omega^k)$ is decreasing (see Lemma 3.1 for sufficient conditions), the bound in Proposition 3.1 applies directly to $J_W(\mu_\omega^k) - J_W(\mu_\omega^\star)$.

Remark 3.1 (Choice of β and limits of Proposition 3.1). In Proposition 3.1, the step size β can be taken equal to zero without affecting the bound on $J_W\left(\frac{1}{K}\sum_{k=1}^{K}\mu_\omega^k\right) - J_W(\mu_\omega^\star)$. The choice of location–scale updates is likewise largely immaterial, as it only needs to satisfy a very mild condition. In particular, one may simply use the trivial update $\mathcal{T}_\beta(x_j^k, J'_{\mu_\omega^k}) = x_j^k$. This reflects the fact that this convergence analysis depends solely on the behavior of the weights, and not on the evolution of the location–scale parameters. Consequently, the initialization must place sufficient mass near the support of the solution $\mu_\omega^\star$. More precisely, it requires that there exists $\mathcal{X}_0 \subset \mathcal{X}$ such that $\mu_\omega^\star(\mathcal{X}_0)$ is small, and for every $x \in \mathrm{supp}(\mu_\omega^\star) \setminus \mathcal{X}_0$, $\min_{j=1,\dots,p} \mathfrak{d}_{\mathfrak{g}}(x, x_j^1)$ is small.

Remark 3.2 (Construction of the initialization). The initialization μ_ω^1 can be specifically designed to control the terms $H_{\mu_\omega^1, \mu_\omega^\star}$, $r_{\mu_\omega^1, \mathcal{X}}$ and $\mu_\omega^\star(\mathcal{X}_0)$ appearing in Proposition 3.1. Setting $\mu_\omega^1 = \sum_{j=1}^{p}\frac{1}{p}\delta_{x_j^1}$ where the points x_j^1 form a grid and the regions $\mathcal{X}_j$ correspond to the associated cells, then we have $\mu_\omega^\star(\mathcal{X}_0) = 0$ and we can expect $r_{\mu_\omega^1, \mathcal{X}}$ to be of order $p^{-1/(2d)}$ (omitting the dependence on the total volume of $\mathcal{X}$). Moreover, recalling (8b), with this choice of μ_ω^1,

$$\begin{aligned} H_{\mu_\omega^1, \mu_\omega^\star} &\le \max\left\{\|\mu_\omega^\star\|_{\mathrm{TV}} \ln\left(\frac{\|\mu_\omega^\star\|_{\mathrm{TV}}}{\min_j \omega_j^1}\right), 0\right\} - \|\mu_\omega^\star\|_{\mathrm{TV}} + \mu_\omega^\star(\mathcal{X}_0) + \left\|\mu_\omega^1\right\|_{\mathrm{TV}} , \\ &\le \|\mu_\omega^\star\|_{\mathrm{TV}} \left(\max\left\{\ln\left(p\,\|\mu_\omega^\star\|_{\mathrm{TV}}\right), 0\right\} - 1\right) + 1 . \end{aligned}$$

In particular, choosing $p = \lceil K^d \rceil$, we get $H_{\mu_\omega^1, \mu_\omega^\star} \lesssim_{\|\mu_\omega^\star\|_{\mathrm{TV}}} d\ln(K)$ and

$$J_W\left(\frac{1}{K}\sum_{k=1}^{K}\mu_\omega^k\right) - J_W(\mu_\omega^\star) \lesssim_{\|\mu_\omega^\star\|_{\mathrm{TV}}, d, \|y\|_{\mathbb{L}}, \mathcal{X}} \frac{\ln(K)}{\sqrt{K}} .$$

Note that with our choice of initialization, $\left\|\mu_\omega^1\right\|_{\mathrm{TV}} = 1$ does not depend on p.

Remark 3.3 (Need to work on a compact set, positivity of the kernel). Working on a compact set ensures the existence of a solution to the BLASSO. Moreover, in the absence of prior information on $\mu_\omega^\star$, restricting the problem to a compact set is necessary in order to construct the initialization as described in Remark 3.2.

In addition, we used in the proof the explicit form of K_{norm} and in particular the fact that $\inf_{x,x' \in \mathcal{X}} K_{\mathrm{norm}} =: C_{\mathcal{X}} > 0$ since $\mathcal{X}$ is compact. This control allows to establish a bound on $\left\|\mu_\omega^k\right\|_{\mathrm{TV}}$. An alternative approach, which does not rely on the positivity of $C_{\mathcal{X}}$, can still yield a bound on $\left\|\mu_\omega^k\right\|_{\mathrm{TV}}$, but this bound then depends on p (see Remark A.1 in appendix). In particular, without prior information on $\mu_\omega^\star$, one cannot expect convergence as p grows, since the bound on $J_W\left(\frac{1}{K}\sum_{k=1}^{K}\mu_\omega^k\right) - J_W(\mu_\omega^\star)$ then involves the term $pr_{\mu_\omega^1, \mathcal{X}}$, which may scale as $p^{1-\frac{1}{(2d)}}$ when constructing μ_ω^1 on a grid.

3.2 Exponential local convergence

The investigation presented in this section relies on [Chizat, 2022, Corollary 3.5]. We leverage here not only the dynamic of the weights but also that of the location–scale parameters. To this end, we employ the Riemannian

gradient descent updates defined in (RGD) together with the weights update (7). To avoid difficulties related to projection, we adopt a simple strategy: freezing particles outside a region of interest. More precisely, we assume that the compact set $\mathcal{X}$ satisfies $\mathcal{X} \subset \mathbb{R}^d \times [u_{\min}, u_{\max}]^d$, for some $\tau \leq u_{\min} \leq u_{\max}$. We also consider a known compact set $\tilde{\mathcal{X}} \subset \mathring{\mathcal{X}}$. Particles lying outside $\tilde{\mathcal{X}}$ are frozen during the optimization, which ensures that the iterates remain in $\mathcal{X}$ (see Lemma 3.1). This strategy is formalized in Algorithm 2.

Algorithm 2 Conic Particles Riemannian Descent with freezing process

Require: Step sizes $\alpha, \beta \geq 0$, number of iterates K, initialization $\mu_\omega^1 = \sum_{j=1}^p \omega_j^1 \delta_{x_j^1}$

for $k = 1, \dots K-1$ **do**
  for $j = 1, \dots, p$ **do**
    $\omega_j^{k+1} \leftarrow \omega_j^k e^{-\alpha J'_{\mu_\omega^k}(x_j^k)}$ ▷ Weight update
    if $x_j^k \in \mathcal{X} \setminus \tilde{\mathcal{X}}$ **then**
      $x_j^{k+1} \leftarrow x_j^k$
    else
      $x_j^{k+1} \leftarrow \exp_{x_j^k}(-\beta \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x_j^k))$ ▷ Location-scale update
    end if
  end for
  $\mu_\omega^{k+1} \leftarrow \sum_{j=1}^p \omega_j^{k+1} \delta_{x_j^{k+1}}$
end for

The update (RGD) is well-defined, and x_j^{k+1} remains in $\mathcal{X}$ provided that β is sufficiently small. Note that the underlying manifold is not geodesically complete, so $\exp_x(v)$ is not defined for all $v \in \mathbb{R}^d$. In practice, this can be addressed by restricting to vectors v with bounded norm $\|v\|_x$. This requirement, together with the need to ensure that x_j^k remains in $\mathcal{X}$, motivates the assumption below that β is sufficiently small.

Descent property The following lemma (whose proof is postponed to Appendix C) provides a first description of the performances of Algorithm 2. In particular, it establishes that the objective function J_W is non-increasing along the trajectory $(\mu_\omega^k)_{k \geq 1}$.

Lemma 3.1. *Assume that $\kappa \leq 1$. For $0 \leq \alpha, \beta \leq 1$ small enough (depending on $\mathcal{X}, \tilde{\mathcal{X}}, \|y\|_{\mathbb{L}}, \left\|\mu_\omega^1\right\|_{\mathrm{TV}}, \tau$), the position updates of Algorithm 2 are well-defined and for all $k \geq 1$,*

$$J_W(\mu_\omega^{k+1}) - J_W(\mu_\omega^k) \leq -\frac{1}{2} \int_{\tilde{\mathcal{X}}} \left(\alpha |J'_{\mu_\omega^k}(x)|^2 + \beta \left\| \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x) \right\|_x^2 \right) \mathrm{d}\mu_\omega^k(x) - \frac{1}{2} \int_{\mathcal{X} \setminus \tilde{\mathcal{X}}} \alpha |J'_{\mu_\omega^k}(x)|^2 \, \mathrm{d}\mu_\omega^k(x) \, .$$

The inequality displayed in Lemma 3.1 describes the behavior of the increment of the sequence $J_W(\mu_\omega^k)$ in terms of some energy terms related to the Fréchet derivative $J'_{\mu_\omega^k}$ and its gradient. This result is in line with the KKT conditions associated with our problem. In particular, the descent property holds as long as $J'_{\mu_\omega^k} \neq 0$ on a set where μ_ω^k is strictly positive. However, the descent breaks off as soon as $J'_{\mu_\omega^k}$ cancels out in the support of μ_ω^k, which does not guarantee that μ_ω^k is indeed a global minimum.

Non-degeneracy of the solution Establishing strong convergence guarantees requires structural assumptions on the solution $\mu_\omega^\star$. Specifically, the latter must be a discrete measure satisfying a *non-degeneracy* condition (see Assumption 1 below). This condition involves the Riemannian Hessian, which we define next.

Definition 3.1 (Riemannian Hessian)**.** Let $\psi \in \mathcal{C}^2(\mathbb{R}^d \times [u_{\min}, +\infty)^d)$. Let $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$. We define

$$H^{\mathfrak{g}} \psi(x) = \nabla^2 \psi(x) - \sum_{k=1}^d \Gamma^{t_k} \partial_{t_k} \psi(x) - \sum_{k=1}^d \Gamma^{u_k} \partial_{u_k} \psi(x) \, ,$$

with $\Gamma^{t_k}, \Gamma^{u_k}$ defined by (19).

Assumption 1 (Non-degenerate certificate for the solution)**.** ***Location of the solution:*** *The solution to ($\mathcal{P}_\kappa$) is unique and of the form $\mu_\omega^\star = \sum_{j=1}^{p^\star} \omega_j^\star \delta_{x_j^\star}$. Moreover, it satisfies $\omega_j^\star \geq \omega_\star > 0$ and $x_j^\star \in \mathring{\tilde{\mathcal{X}}}$.*
Non-degeneration: *In addition, there exist $\varepsilon_2 > 0$, $C_{\tilde{\Upsilon}} > 0$ and a non-degenerate certificate associated with $\mu_\omega^\star$, namely*

$$\eta^\star : x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d \mapsto -\frac{1}{\kappa} \langle \Psi \delta_x, \Psi \mu_\omega^\star - y \rangle_{\mathbb{L}} = 1 - \frac{J'_{\mu_\omega^\star}(x)}{\kappa}$$

satisfying

1. $\eta^\star \leq 1$ *on* $\mathcal{X}$,

2. $\eta^\star(x) = 1 \iff x \in \{x_j^\star\}_{j=1}^{p^\star}$,

3. $\nabla\eta^\star(x_j^\star) = 0$ *for all* $j = 1, \dots, p^\star$,

4. $-v^T H^{\mathfrak{g}}\eta^\star(x_j^\star)v \geq \varepsilon_2 \|v\|^2_{x_j^\star}$ *for all* $v \in \mathbb{R}^{2d}$, $j \in \{1, \dots, p^\star\}$,

5. $\tilde{\Upsilon} \succeq C_{\tilde{\Upsilon}} I$ *with* $C_{\tilde{\Upsilon}} > 0$, *where*

$$\tilde{\Upsilon} \coloneqq \begin{pmatrix} (K_{\mathrm{norm}}(x_j^\star, x_i^\star))_{j,i=1,\dots,p^\star} & (\nabla_2 K_{\mathrm{norm}}(x_j^\star, x_i^\star)^T \mathfrak{g}_{x_i^\star}^{-1/2})_{j,i=1,\dots,p^\star} \\ (\mathfrak{g}_{x_j^\star}^{-1/2}\nabla_1 K_{\mathrm{norm}}(x_j^\star, x_i^\star))_{j,i=1,\dots,p^\star} & (\mathfrak{g}_{x_j^\star}^{-1/2}\nabla_1\nabla_2 K_{\mathrm{norm}}(x_j^\star, x_i^\star)\mathfrak{g}_{x_i^\star}^{-1/2})_{j,i=1,\dots,p^\star} \end{pmatrix}. \tag{9}$$

The non-degeneracy assumption is in line with [Poon et al., 2023; Duval and Peyré, 2015], where it is introduced to establish the closeness between $\mu_\omega^\star$ and the statistical target measure μ^0. It is referred to as the non-degenerate source condition (NDSC), which postulates the existence of a non-degenerate certificate associated with μ^0.

Here, however, the non-degenerate certificate is associated with $\mu_\omega^\star$ rather than with μ^0. The connection between the non-degeneracy of μ^0 and that of the BLASSO solution is discussed in [Duval and Peyré, 2015, Section 3].

Note that any solution of the BLASSO satisfies $\eta^\star \leq 1$ on $\mathcal{X}$ and $\eta^\star = 1$ on $\mathrm{supp}(\mu_\omega^\star)$ (this follows from the KKT conditions, *cf*. [Chizat, 2022, Proposition 3.1]).

Convergence guarantees Under Assumption 1, combined with a strong condition on the initialization, we establish exponential convergence of the objective values, as stated in the following theorem whose proof is given in Appendix D.

Theorem 3.1 (Exponential local convergence under non-degeneracy of the solution)**.** *Let* $0 < \kappa \leq 1$.
Conditions on the solution: We suppose that Assumption 1 is satisfied.
Choice of initialization, updates: Assume that $\mu_\omega^1 = \sum_{j=1}^p \omega_j^1 \delta_{x_j^1}$ *with* $\omega_j^1 > 0$ *and* $x_j^1 \in \mathcal{X}$. *Assume that it satisfies*

$$J_W(\mu_\omega^1) - J_W(\mu_\omega^\star) \leq J_0 \kappa^{12}, \tag{10a}$$

with $J_0 > 0$ *depending on* $\mathcal{X}$, $\tilde{\mathcal{X}}$, τ, $\mu_\omega^\star$, $\eta^\star$, $\|y\|_{\mathbb{L}}$, $\left\|\mu_\omega^1\right\|_{\mathrm{TV}}$. *We use Algorithm 2 together with the parameter choices* α, β *described in Lemma 3.1.*
Convergence of CPGD: There exists $R_0 > 0$, *depending on* $\mathcal{X}$, $\tilde{\mathcal{X}}$, τ, $\mu_\omega^\star$, $\eta^\star$, $\|y\|_{\mathbb{L}}$ *and* $\left\|\mu_\omega^1\right\|_{\mathrm{TV}}$, *such that* $\frac{1}{2}\min\{\alpha, \beta\} R_0 \kappa^6 < 1$ *and, for all* $k \geq 1$,

$$J_W(\mu_\omega^k) - J_W(\mu_\omega^\star) \leq \left(1 - \frac{1}{2}\min\{\alpha, \beta\} R_0 \kappa^6\right)^{k-1} \left(J_W(\mu_\omega^1) - J_W(\mu_\omega^\star)\right). \tag{10b}$$

In contrast to Lemma 3.1, Theorem 3.1 describes the behavior of the difference between the current state $J_W(\mu_\omega^k)$ and the target $J_W(\mu_\omega^\star)$ in terms of the iteration number k. In particular, it establishes exponential convergence of the objective values, provided that κ is small enough and that the initialization condition (10a) holds; the convergence is hence said to be local.

Three limitations should be kept in mind when reading Theorem 3.1. First, the constants J_0 and R_0 depend on the solution $\mu_\omega^\star$ and on its associated certificate $\eta^\star$, both of which are unknown. Condition (10a) should therefore be understood as a qualitative locality requirement, and not as a criterion that can be evaluated on a given initialization; likewise, the rate in (10b) is not a quantity the user can compute in advance. Second, for the sake of proof readability, the exponents for κ in (10a) and (10b) are not optimized, as discussed in Remark D.1. For κ small, the resulting basin of attraction is correspondingly narrow, and the guarantee is qualitative rather than quantitative in nature. Third, Theorem 3.1 says nothing about how the initialization condition may be met: Proposition 3.1 controls the Cesàro average rather than the last iterate, and we do not claim that combining the two yields a complete convergence theory. We note that the numerical behavior reported in Section 4 is considerably more favorable than these constants would suggest.

Since the problem ($\mathcal{P}_\kappa$) restricted to discrete measures is non-convex, obtaining global convergence guarantees remains challenging. We refer, for instance, to [De Castro et al., 2026] for a similar problem involving a translation-invariant kernel.

Checking the non-degeneracy assumption Assumption 1 allows to obtain an exponential local convergence for Algorithm 2 and plays a crucial role in our analysis. Nevertheless, it is impossible to check in practice. We can however try to give some hints on this condition in a specific setting, by shifting the condition on the solution $\mu^\star_\omega$ to an underlying statistical ground truth. More precisely, we assume that the measurement y can be written as

$$y = \Psi\mu^0_\omega + b\,, \tag{11}$$

where $b \in \mathbb{L}$ and $\mu^0_\omega \coloneqq \sum_{j=1}^s \omega^0_j \delta_{x^0_j}$, with $\omega^0_j > 0$ and $x^0_j \in \mathcal{X}$. We interpret b as noise and refer to μ^0_ω as the statistical target. Section 4.1 discusses the relevance of this model in a statistical context.

We prove below that Assumption 1 holds provided that the noise b and the regularization parameter κ are sufficiently small, and that the particles of μ^0_ω are well separated with respect to a tailored semi-distance. This semi-distance is defined, for $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$, by

$$\mathrm{d}(x,x')^2 = -2\ln(K_{\mathrm{norm}}(x,x')) = \sum_{k=1}^d \left(\frac{(t_k - t'_k)^2}{{u'_k}^2 + u_k^2 + \tau^2} + \ln\left(\frac{{u'_k}^2 + u_k^2 + \tau^2}{\sqrt{2u_k^2+\tau^2}\sqrt{2{u'_k}^2+\tau^2}} \right) \right). \tag{12}$$

These conditions are formalized in the following theorem whose proof is postponed to Appendix E.

Theorem 3.2 (Checking Assumption 1)**.** *Assume that y is of the form (11), that $s \geq 2$ and $\{x^0_j\}_{j=1}^s \subset \mathring{\tilde{\mathcal{X}}}$. Assume that $\min_{i\neq j} \mathrm{d}(x^0_j, x^0_i) \geq \Delta_\tau$, where*

$$\Delta_\tau \coloneqq \max\left\{ \frac{\sqrt{u_{\max}^2 + \frac{1}{4}\frac{0.3025^2}{d}(2u_{\max}^2+\tau^2)}}{u_{\min}} \left(\Delta + \frac{0.3025}{\sqrt{d}}\right), 2\frac{u_{\max}}{u_{\min}}\Delta \right\} + \sqrt{d\ln\left(\frac{u_{\max}^2}{u_{\min}^2}\right)} \tag{13a}$$

and

$$\Delta = 2\sqrt{11.9 + 3\ln(d+6.62) + \ln(s-1)}\,. \tag{13b}$$

There exists $\gamma_0 > 0$ depending on d, and $C_\kappa > 0$ depending on $\mathcal{X}, \tilde{\mathcal{X}}, \tau, \mu^0_\omega$ such that if $0 < \kappa \leq C_\kappa$ and $\|b\|_{\mathbb{L}} \leq \gamma_0\kappa$, then Assumption 1 is satisfied with $C_{\tilde{\Upsilon}} = \frac{1}{4}$.

This result relies on the connection between the Non-Degenerate Source Condition introduced by [Duval and Peyré, 2015], studied in our setting by [Giard et al., 2025], which concerns the non-degeneracy of the target μ^0_ω, and the non-degeneracy of the BLASSO solution.

The numerical constants appearing in (13a) and (13b) are those of the local positive curvature property satisfied by K_{norm}: the radius $\frac{0.3025}{\sqrt{d}}$ and the associated curvature parameters are established in [Giard et al., 2025, Theorem 5.3 and Lemma 5.4], and the separation (13a) is the corresponding condition of [Giard et al., 2025, Theorem 5.1].

It is worth making explicit what is imported from that work and what is established here. The construction of non-degenerate dual certificates for the statistical target μ^0_ω, together with the separation condition under which it succeeds, is the object of [Giard et al., 2025]. Assumption 1, by contrast, bears on the BLASSO solution $\mu^\star_\omega$, and it is this form that the analysis of Section 3.2 requires: the descent argument is driven by the behavior of $\eta^\star$ in the vicinity of $\mathrm{supp}(\mu^\star_\omega)$, a set that is not known in advance and does not coincide with $\mathrm{supp}(\mu^0_\omega)$. Theorem 3.2 accordingly performs two steps that have no counterpart in [Giard et al., 2025]: it transfers non-degeneracy from μ^0_ω to $\mu^\star_\omega$, which calls for quantitative control of both the locations and the weights of the solution in the small-κ, small-noise regime; and it upgrades the resulting condition to a Riemannian one, expressed through the Fisher-Rao Hessian $H^{\mathfrak{g}}$ and the metric-normalized matrix $\tilde{\Upsilon}$ of (9). The normalization by $\mathfrak{g}_{x^\star_j}^{-1/2}$ is what renders the curvature bound uniform in the location-scale parameters, and hence usable to produce the explicit rate of Theorem 3.1.

Note that Theorems 3.1 and 3.2 establish pointwise convergence rates. In Theorem 3.2, the condition on b required for Assumption 1 to hold depends on the statistical target μ^0_ω. Meanwhile, in Theorem 3.1, both the condition on the initialization and the resulting convergence rate depend on the solution $\mu^\star_\omega$.

Positioning relative to [Chizat, 2022] Our contribution relies on [Chizat, 2022, Corollary 3.5]. A difference with the setting of [Chizat, 2022] is that the domain of the particle positions is not a compact manifold without boundary. The presence of a boundary in $\mathcal{X}$ requires a specific treatment, which we handle by freezing particles outside $\tilde{\mathcal{X}}$.

In addition, we provide an interpretable condition under which [Chizat, 2022, Assumption 5], is satisfied. This yields a concrete criterion for verifying the assumption in our setting. To the best of our knowledge, this is the first work showing that, in a statistical framework, the non-degeneracy assumption of [Chizat, 2022] is met under a more explicit condition on the support separation and noise boundedness.

4 Numerical experiments

In this section, we discuss the practical implementation of CPGD for the statistical estimation of a Gaussian mixture model with unknown (diagonal) covariances and present numerical experiments on simulated data. The code is shared in the archived release [Giard and Denis, 2026]. We first provide a proof of concept in a favorable setting, illustrating the good convergence properties of CPGD. We then compare its performance to the EM algorithm. Finally, we investigate the influence of the separation between target components on recovery accuracy.

4.1 Statistical motivation

The optimization problem $(\mathcal{P}_\kappa)$ is particularly relevant in a statistical setting. In this framework, the goal is to recover the target probability measure encoding a Gaussian mixture distribution,

$$\mu^0 = \sum_{j=1}^{s} a_j^0 \delta_{x_j^0} \quad \text{where} \quad a_1^0, \dots a_s^0 > 0\,, \ \sum_{j=1}^{s} a_j^0 = 1\,.$$

We assume that the location-scale parameters $(x_1^0, \dots, x_s^0)$ are distinct, where for any $j \in \{1, \dots, s\}$, $x_j^0 = (t_j^0, u_j^0)$, with $t_j^0 = (t_{j,k}^0)_{k=1}^d \in \mathbb{R}^d$ and $u_j^0 = (u_{j,k}^0)_{k=1}^d \in [u_{\min}, +\infty)^d$. The parameters t_j^0 and u_j^0 represent respectively the mean and square root of the diagonal covariance of the j^{th} Gaussian component with weight a_j^0. The lower bound $u_{\min}$ on the minimal eigenvalue of each covariance matrix is a standard constraint for Gaussian mixture estimation based on likelihood maximization (see, *e.g.*, [Bouveyron et al., 2019]).

We observe a sample $X_1, \dots, X_n \overset{i.i.d.}{\sim} f_0$ drawn according to a Gaussian mixture distribution, whose density f_0 is associated with the measure μ^0. In particular, it can be rewritten as $f_0 \coloneqq \Phi\mu^0$ where $\Phi : \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d) \longrightarrow L^2(\mathbb{R}^d)$ is the linear operator defined by

$$\Phi\mu : z \in \mathbb{R}^d \mapsto \int_{\mathbb{R}^d \times [u_{\min}, +\infty)^d} \prod_{k=1}^{d} \frac{1}{\sqrt{2\pi} u_k} e^{\frac{-(z_k - t_k)^2}{2u_k^2}} \, \mathrm{d}\mu(t,u) \quad \forall\, \mu \in \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d)\,.$$

To compare the observations with a candidate density, we use a Gaussian convolution with smoothing parameter τ: introducing $\lambda : z \in \mathbb{R}^d \mapsto \frac{e^{-\frac{\|z\|_2^2}{2\tau^2}}}{(2\pi\tau^2)^{d/2}}$, we define $y = \frac{1}{n}\sum_{i=1}^n \lambda(\bullet - X_i)$. The term y involved in $(\mathcal{P}'_\kappa)$ corresponds in such a case to a kernel estimator associated with the density f_0.

We then introduce a reparameterized version of the target measure, $\mu_\omega^0 = \sum_{j=1}^s \omega_j^0 \delta_{x_j^0}$ with $\omega_j^0 = W(x_j^0) a_j^0$, where

$$W(x) \coloneqq \prod_{k=1}^{d} (2\pi)^{-1/4} (2u_k^2 + \tau^2)^{-1/4} \quad \forall\, x = ((t_1, \dots, t_d), (u_1, \dots, u_d)) \in \mathbb{R}^d \times [u_{\min}, +\infty)^d\,. \tag{14}$$

With this notation, one has

$$\Psi\mu = \lambda * \Phi\frac{\mu}{W} \quad \forall\, \mu \in \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d)\,,$$

which makes explicit the connection between the BLASSO problem $(\mathcal{P}_\kappa)$ and the recovery of the parameters of a Gaussian mixture model. In this formulation, a solution to $(\mathcal{P}_\kappa)$ approximates μ_ω^0 rather than μ^0. This renormalization is essential for obtaining recovery guarantees [Giard et al., 2025], as it leads to a kernel that is normalized (*i.e.*, $x \mapsto K_{\mathrm{norm}}(x,x)$ is constant).

Note that in this statistical setting, $y = \lambda * \hat{f}_n$ where $\hat{f}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. The updates (EGD), (NGD), (RGD) admit closed-form expressions when y is of the form $\lambda * \nu$ with $\nu \in \mathcal{M}(\mathbb{R}^d)$ a discrete measure. In this setting, the integrals appearing in the norms and inner products on $\mathbb{L}$ can be computed exactly, and no Monte Carlo approximation is required (the formal computation is given in `cpgd_sympy.ipynb`, and the implementation is available in `methods.py` in [Giard and Denis, 2026]).

Remark 4.1 (Extension to general covariance structures). We focus our attention on the case where the covariance matrices associated with each component of the mixture are diagonal. These investigations are grounded in the statistical analysis displayed by [Giard et al., 2025] in the same framework. Extending this theoretical analysis to general covariance structures is still an open problem. From a computational perspective, the extension of the CPGD principle to general covariance matrices is not straightforward. It indeed requires to handle a non-diagonal metric tensor. Guaranteeing the positive definiteness of the estimated matrices is also a key point. Such an extension hence goes outside the scope of this paper.

4.2 AdaGrad with natural gradient descent

In Section 3, theoretical results are presented, when the Riemannian gradient descent with fixed step sizes is used, together with a freezing mechanism (see Algorithm 2). In practice, the step sizes α, β must be chosen sufficiently small to ensure that the updates are well defined (Lemma 3.1). Moreover, fixing a prior region outside of which the location–scale parameters are frozen might be cumbersome. It requires identifying, at each iteration, the particles that lie outside this region, and selecting β sufficiently small to ensure that the iterates remain in $\mathbb{R}^d \times (0, +\infty)^d$. Finally, computing the geodesic trajectories—even though they are available in closed form—requires more computation than a classical Euclidean gradient descent.

For these reasons, we use the natural gradient descent (NGD) for our numerical experiments. We stress that the (NGD) scheme can be considered as an approximation of the Riemannian gradient descent (RGD) when the step sizes are small. This choice is theoretically justified (see Section 2.2), and also leads to better results than Euclidean gradient descent in practice. Moreover, we do not perform any projection or freezing outside a prescribed domain; instead, we retain only the componentwise absolute value of u_j^k, noting that the loss J_W is even in $u_{j,l}$, for $1 \leq l \leq d$. Finally, we use AdaGrad [Duchi et al., 2011] to avoid dealing with step sizes, and to improve convergence. This construction leads to Algorithm 3.

Algorithm 3 CPGD with NGD and AdaGrad

Require: Numerical stability term $\epsilon > 0$, number of iterates K, initialization $\mu_\omega^1 = \sum_{j=1}^p \omega_j^1 \delta_{x_j^1}$

$v_{\omega_j}^0 \leftarrow 0, \quad v_{x_j}^0 \leftarrow 0 \quad$ for $j = 1, \ldots, p$ ▷ Initialize AdaGrad accumulators

for $k = 1, \ldots K-1$ **do**

  for $j = 1, \ldots, p$ **do**

    $v_{\omega_j}^{k+1} \leftarrow v_{\omega_j}^k + |J'_{\mu_\omega^k}(x_j^k)|^2, \quad v_{x_j}^{k+1} \leftarrow v_{x_j}^k + \frac{1}{2d} \left\| \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x_j^k) \right\|_{x_j^k}^2$ ▷ Update AdaGrad accumulators

    $\omega_j^{k+1} \leftarrow \omega_j^k \exp\left(- \frac{J'_{\mu_\omega^k}(x_j^k)}{\sqrt{v_{\omega_j}^{k+1}} + \epsilon} \right)$ ▷ Weight update

    $x_j^{k+1} = (t_j^{k+1}, u_j^{k+1}) \leftarrow x_j^k - \frac{\nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x_j^k)}{\sqrt{v_{x_j}^{k+1}} + \epsilon}$ ▷ Location-scale update

    $u_j^{k+1} \leftarrow |u_j^{k+1}|$ ▷ Componentwise absolute value

  end for

  $\mu_\omega^{k+1} \leftarrow \sum_{j=1}^p \omega_j^{k+1} \delta_{x_j^{k+1}}$

end for

Note that we do not use separate adaptive steps for each coordinate. Instead, we accumulate the squared norms of the gradients *per particle*. The underlying metric is essential: the Riemannian gradients determine how the relative updates of $(t_1, \ldots, t_d, u_1, \ldots, u_d)$ should be scaled. In practice, the adaptation operates at the level of particles, controlling the relative step sizes between weights and positions, as well as across different particles.

We emphasize that Algorithm 3, which produces all the numerical results reported below, differs from Algorithm 2, to which the analysis of Section 3 applies: it replaces the Riemannian update (RGD) by (NGD), the fixed step sizes of Lemma 3.1 by AdaGrad, and the freezing mechanism by the componentwise absolute value. None of the guarantees of Section 3 therefore covers Algorithm 3; the experiments below should be read as an empirical study of a practical variant.

Hyperparameters The estimation procedure is sensitive to both the smoothing parameter τ and the regularization parameter κ. If κ is chosen too large, the weights tend to shrink to zero. Conversely, if κ is too small, additional local minima may appear, and some irrelevant particles may persist in the solution.

Regarding τ, excessively large values may deteriorate the accuracy of the estimation when the components are not well separated. In particular, τ should not be too large relative to $u_{\min}$. On the other hand, if τ is taken too small, the theoretical bounds governing the estimation become less favorable [Giard et al., 2025, Theorem 1.1]. Moreover, in certain settings, the estimator returned by CPGD is not robust with respect to τ, as even small variations in this parameter can result in significant changes in estimation quality.

In practice, the parameters κ and τ can be tuned by cross-validation, for instance using a Monte Carlo estimate of the KL divergence on a test sample.

4.3 Scores

We aim to evaluate the recovery of μ^0 achieved by CPGD, by comparing the last estimate μ^K produced by Algorithm 3 to the ground truth μ^0. Since many discrepancy measures require probability measures, we will in

fact compare μ^0 with the renormalized version of μ^K defined as

$$\mu^K_{\mathrm{norm}} = \frac{\mu^K}{\|\mu^K\|_{\mathrm{TV}}} = \sum_{j=1}^{p} a^K_{j,\mathrm{norm}} \delta_{x^K_j}\,, \quad \text{where} \quad a^K_{j,\mathrm{norm}} = \frac{a^K_j}{\sum_{j=1}^p a^K_j} \ \forall j \in \{1,\dots,p\}.$$

We consider several quantitative criteria to illustrate the discrepancy between the reconstructed measure μ^K_{norm} and the target μ^0. First, we compute a Monte–Carlo approximation of the Kullback–Leibler divergence $D_{\mathrm{KL}}(\Phi\mu^0, \Phi\mu^K_{\mathrm{norm}})$ between $\Phi\mu^0$ and $\Phi\mu^K_{\mathrm{norm}}$. Formally, given a sample $Z_1,\dots,Z_m \overset{i.i.d.}{\sim} \Phi\mu^0$, we use the approximation

$$D_{\mathrm{KL}}(\Phi\mu^0, \Phi\mu^K_{\mathrm{norm}}) \approx \frac{1}{m}\sum_{i=1}^{m} \ln\left(\frac{\Phi\mu^0(Z_i)}{\Phi\mu^K_{\mathrm{norm}}(Z_i)}\right) =: S^{\mathrm{KL}}\,,$$

with $m = 10000$. Second, we compute a Wasserstein-type distance between μ^0 and μ^K_{norm}, defined by

$$S = \min_{\substack{\gamma_{ij}\geq 0,\\ \sum_j \gamma_{ij} = a^0_i,\\ \sum_i \gamma_{ij} = a^K_{j,\mathrm{norm}}}} \sum_{ij} \gamma_{ij} d(x^0_i, x^K_j)\,.$$

Several choices of ground distance d, independent of τ, can be considered:

- The semi-distance defined for $x, x' \in \mathbb{R}^d \times (0,+\infty)^d$ by

$$\mathrm{d}_0(x,x')^2 = \sum_{k=1}^{d} \frac{(t_k - t'_k)^2}{u_k^2 + u_k'^2} + \ln\left(\frac{u_k^2 + u_k'^2}{2u_k u'_k}\right)\,.$$

  We refer to (12) or [Giard et al., 2025] for more details on d_0. Please note that the associated Wasserstein discrepancy S^{d} is only a semi-distance in such a case.

- The Fisher-Rao distance with $\tau = 0$, *i.e.* the Poincaré half-plane distance

$$\mathfrak{d}_{\mathfrak{g}0}(x,x')^2 = 2\sum_{k=1}^{d} \ln\left(\frac{\sqrt{(t_k - t'_k)^2 + (u_k - u'_k)^2} + \sqrt{(t_k - t'_k)^2 + (u_k + u'_k)^2}}{2\sqrt{u_k u'_k}}\right)^2\,.$$

  In such a case, $S^{\mathfrak{d}_{\mathfrak{g}}}$ is a distance.

- The Euclidean distance, for which S^e reduces to the classical Wasserstein–1 distance.

In the following, we report results using the scores S^{d}, $S^{\mathfrak{d}_{\mathfrak{g}}}$, S^e and S^{KL}.

4.4 Proof of concept

We first illustrate the behavior of the algorithm in a favorable setting where successful recovery is expected. More precisely, we consider the one-dimensional case with well-separated components. The target measure consists of two particles with equal weights and identical scale parameters, but distinct means, namely

$$\mu^0 = \frac{1}{2}\delta_{(-1,1)} + \frac{1}{2}\delta_{(1,1)}\,. \tag{15}$$

We consider a sample of size $n = 5000$ drawn from the Gaussian mixture distribution associated with μ^0 (Figure 2). The hyperparameters are set to $\tau = 0.2$ and $\kappa = 10^{-5}$. Algorithm 3 is initialized with three particles and run for 5000 iterations.

The performance of our approach is illustrated in Figure 3. The first curve, on the left, illustrates the descent behavior. In particular, the loss decreases sharply over the iterations (on a logarithmic scale). Although the overall behavior is not monotone, the loss decreases for the most part before reaching a plateau after approximately 10^2 iterations. Starting from three particles at initialization, we observe that two remain significant after a few iterations, while the third is associated with a small final weight (around 0.15). The plot on the right illustrates the trajectories of the individual particles: one converges to one of the target atoms (located in $(-1,1)$), while the other two particles, including the one associated with a small weight, converge to the same target atom. Although this issue is beyond the scope of this paper, one could consider introducing a post-processing procedure to merge particles with similar positions, based on an appropriate similarity criterion.

To conclude this discussion, we can notice that our algorithm produces the scores

$$S^{\mathrm{d}} = 0.064\,, \quad S^{\mathfrak{d}_{\mathfrak{g}}} = 0.059\,, \quad S^e = 0.091\,, \quad \text{and} \quad S^{\mathrm{KL}} = 0.00088\,,$$

for this experiment.

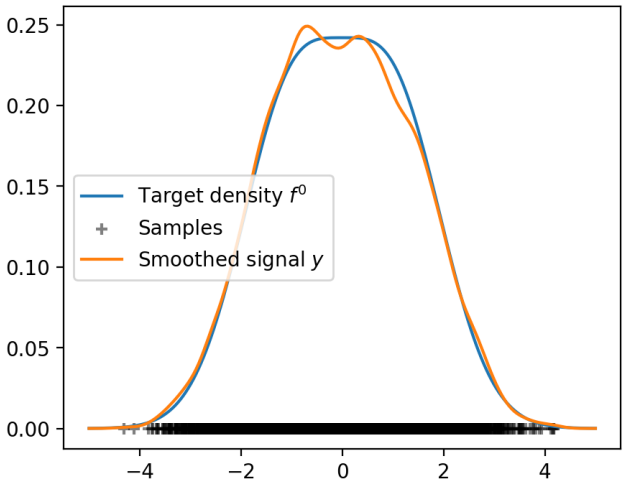

Figure 2: Target mixture and observation y where μ^0 is defined in (15).

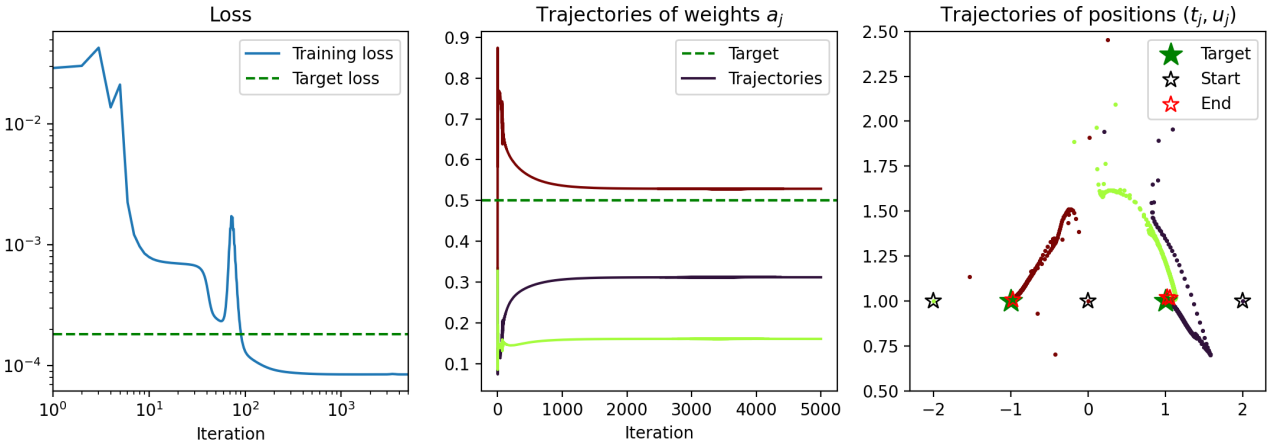

Figure 3: Illustration of the convergence of Algorithm 3. From left to right: evolution of the loss in terms of the iteration number, trajectories of the weights according to the iteration number and trajectories of the positions (mean t on the x-axis, scale parameter u on the y-axis). Target: μ^0 of (15), $n = 5000$, $p = 3$ particles, $\tau = 0.2$, $\kappa = 10^{-5}$, $K = 5000$ iterations.

4.5 Comparison with EM

The EM algorithm is the standard approach for estimating Gaussian mixture models and can be considered as a benchmark for numerical comparisons. Specifically, we use here the vanilla diagonal EM algorithm (which estimates only diagonal covariance matrices) as implemented in scikit-learn. No regularization is applied to the covariance estimates.

The comparison displayed in this section is not intended to be exhaustive. It rather illustrates performance in a representative setting. We consider below a two-dimensional problem with three target components sharing the same mean and equal weights, but with different covariance structures. More precisely, our target measure is

$$\mu^0 = \frac{1}{3}\delta_{(0,0,\sigma_1,\sigma_2)} + \frac{1}{3}\delta_{(0,0,\sigma_2,\sigma_1)} + \frac{1}{3}\delta_{(0,0,\sigma_3,\sigma_3)} \quad \text{with } \sigma_1 = 0.5,\ \sigma_2 = 2,\ \sigma_3 = \sqrt{\sigma_1^2 + \sigma_2^2}\,. \tag{16}$$

The first two components have anisotropic covariances forming a cross-shaped pattern in the observation space, while the third one is isotropic. We consider a sample of size $n = 1000$ drawn from $\Phi\mu^0$.

In this specific setting, we set $\tau = 0.7$, $\kappa = 0.01$ in Algorithm 3. Both EM and Algorithm 3 are initialized identically with p particles of equal weights, whose means are uniformly distributed on a circle of radius 1 centered at $(0, 0)$, and with identical initial scale parameters $u_j^1 = (1, 1)$. Both algorithms are then run for 3000 iterations.

Initialization with 3 particles The case $p = 3$ is particularly favorable to the EM algorithm, since it is specifically designed for situations in which the number of mixture components is known in advance.

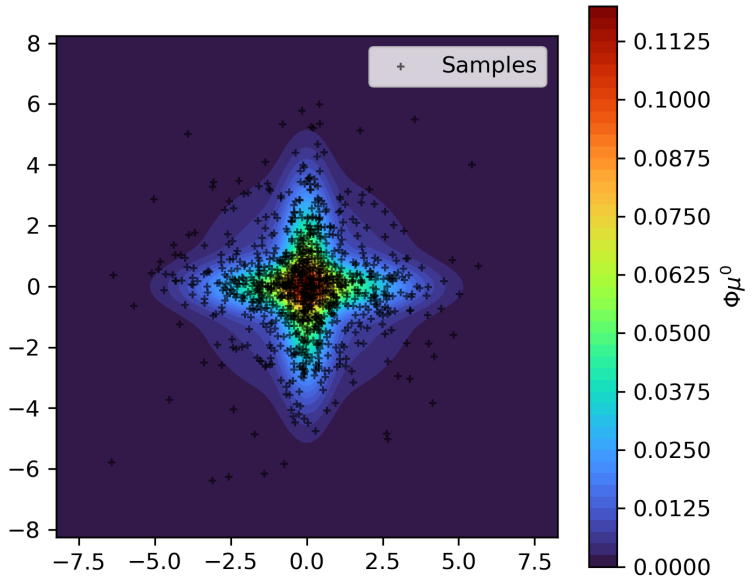

Figure 4: Target mixture and samples when μ^0 is defined according to (16).

In this setting, both algorithms yield comparable results. For CPGD (Figure 6), we obtain

$$S^{\mathcal{d}} = 0.11\,, \quad S^{\mathfrak{d}_{\mathfrak{g}}} = 0.11\,, \quad S^{e} = 0.28\,, \quad S^{\mathrm{KL}} = 0.023\,,$$

while for EM (Figure 5), the corresponding scores are

$$S^{\mathcal{d}} = 0.14\,, \quad S^{\mathfrak{d}_{\mathfrak{g}}} = 0.15\,, \quad S^{e} = 0.25\,, \quad S^{\mathrm{KL}} = 0.0087\,.$$

The Wasserstein-type scores are of similar magnitude for both methods, whereas the KL divergence is smaller for the EM algorithm.

Concerning the trajectories of the parameter estimates, we also notice globally similar behaviors. Note that the estimated weights for Algorithm 3 are (as expected) smaller than those of the target (Figure 6). This is a consequence of the ℓ^1 regularization of the weights induced by the total variation norm in the loss.

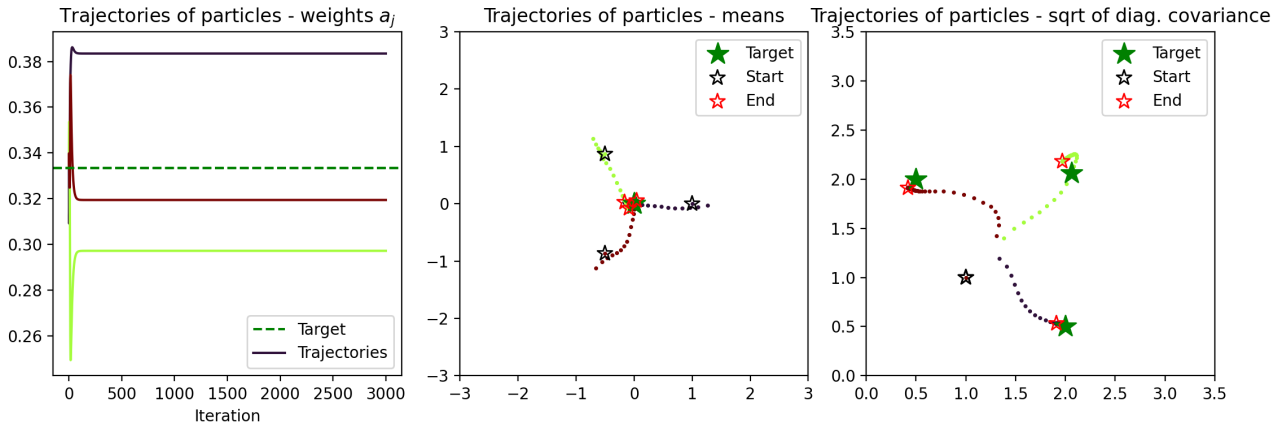

Figure 5: Convergence of the diagonal EM when the target measure is defined in (16). From left to right: evolution of the weights in terms of the iteration number, evolution of the mean along the iteration, evolution of the variance along the iteration.

Influence of overparameterization To extend these investigations, we considered a similar analysis with several values for the particle number p. For the sake of readability, we focus our attention on the evolution of the metrics introduced in Section 4.3 in terms of p. Corresponding results are displayed in Figure 7 (Wasserstein) and Figure 8 (KL). As p increases, Algorithm 3 exhibits stable and consistent behavior, whereas the performance of EM deteriorates.

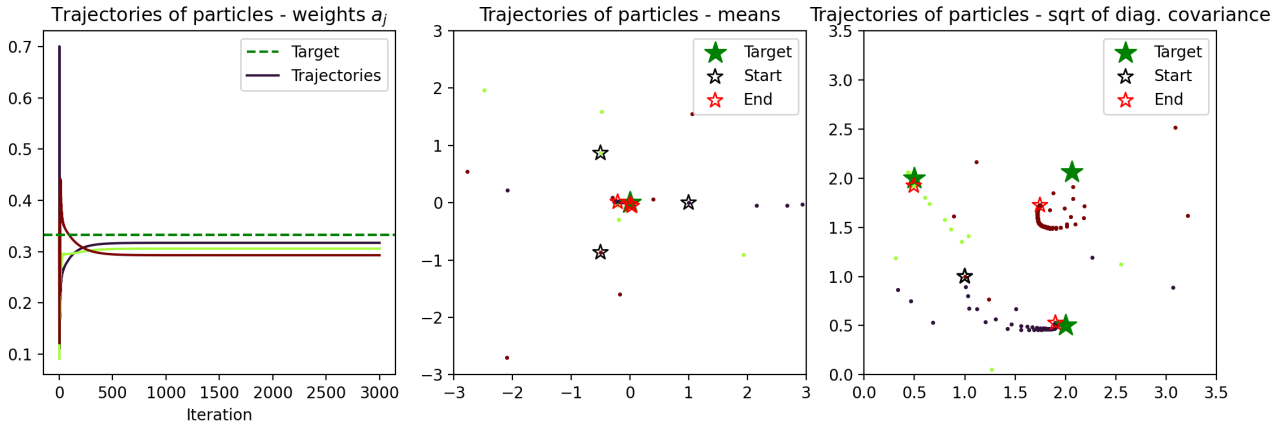

Figure 6: Convergence of Algorithm 3 when the target measure is defined in (16). From left to right: evolution of the weights in terms of the iteration number, evolution of the mean along the iteration, evolution of the variance along the iteration.

The EM algorithm aims to maximize the log-likelihood associated with the observed sample for a mixture model with a prescribed number of p components. It is therefore particularly sensitive to misspecification of the number of components. This is a well-known issue that can be addressed by introducing an additional model-selection step, in which the log-likelihood is penalized by a function of p designed to promote sparsity. We refer, among others, to [Maugis and Michel, 2011] for a description and analysis of such a strategy.

In contrast, the CPGD principle offers greater flexibility in the choice of the initial number of components p. It is based on the optimization problem $(\mathcal{P}_\kappa)$, which includes a regularization term (the total variation norm) that promotes sparsity of the solution. In the experiments reported here, the accuracy of Algorithm 3 is therefore largely insensitive to the number of particles, provided that p exceeds the number of target components. We stress that this statement concerns the accuracy of the reconstructed measure: the algorithm does not by itself return exactly s particles, and the surviving particles may remain split across a single target atom, as observed in Section 4.

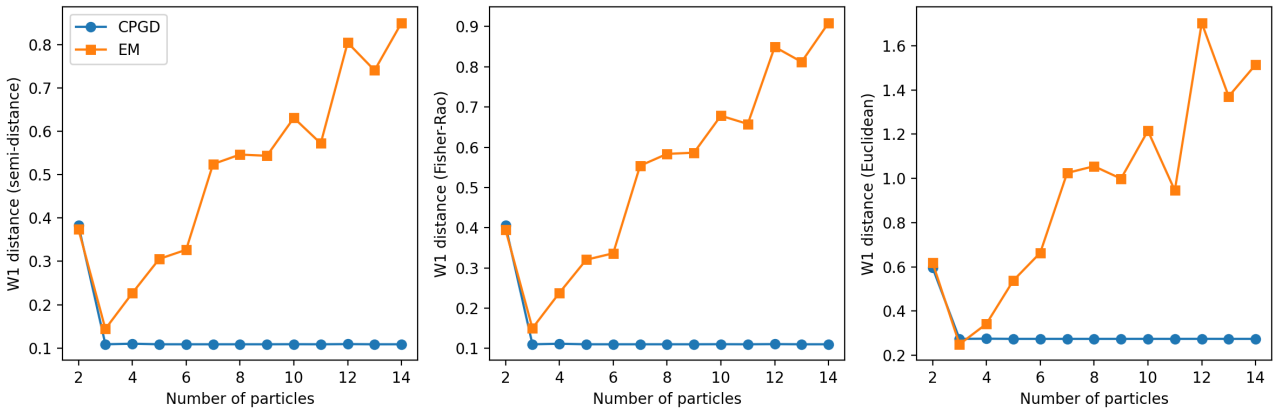

Figure 7: Wasserstein-type scores for EM and Algorithm 3 expressed in terms of p.

4.6 Influence of the separation and sample size

To conclude these numerical investigations, we illustrate the impact of the separation between target components and the sample size on the performance of Algorithm 3.

Experimental design We work in dimension $d = 1$. The target is

$$\mu^0 = \frac{1}{2}\delta_{(-1,u^0)} + \frac{1}{2}\delta_{(1,u^0)}$$

with $u^0 > 0$. We consider different samples of size n drawn from $\Phi\mu^0$.

We study the convergence of CPGD for various values of u^0 and n. As u^0 gets larger, the separation between components gets smaller (measured by $\mathfrak{d}_{\mathfrak{g}}$ or $\mathcal{d}$).

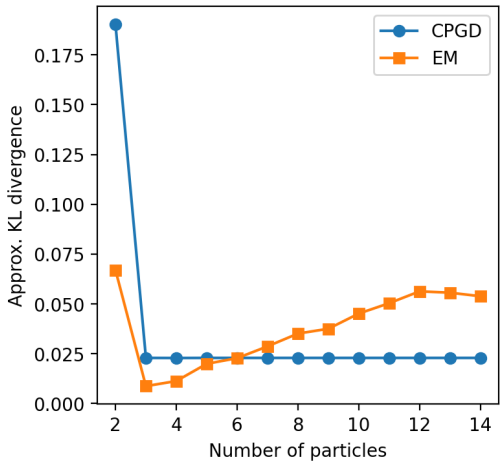

Figure 8: KL divergence for EM and Algorithm 3 expressed in terms of p.

For each pair (u^0, n), Algorithm 3 is run with 10 independently generated samples, and three different initializations. For each sample, we retain the solution achieving the smallest final loss among the initializations. The results are then aggregated over the samples, using the median performance. We get 3 aggregated Wasserstein-type scores depending on the ground distance used in S. For each configuration (u^0, n, d) (with d the ground distance), we then select the hyperparameters τ, κ that yield the best (aggregated) score (this choice depend on the score used, via the ground distance). The parameters τ and κ are searched over a grid: τ ranges from 10^{-5} to 100, and κ from 10^{-14} to 0.1. Since this selection uses the score itself, which depends on the ground truth μ^0, the results reported below are an *oracle* envelope. They describe the best accuracy attainable by the method over the hyperparameter grid, not the accuracy a practitioner would obtain from a data-driven choice of (τ, κ) (for instance selecting hyperparameters using an approximation of the KL divergence on a test sample).

Results The selected hyperparameters vary significantly with n and u^0, and this dependence does not follow a clear or predictable pattern. Still, larger values of u^0 tend to correspond to larger τ.

Moreover, the recovery scores themselves exhibit substantial variability across different initializations and sampled observations, indicating that CPGD lacks strong stability in this setting.

Despite this variability, a meaningful trend emerges from the experiment. We report the best median recovery scores, selected over the different initializations and choices of τ and κ (Figure 9).

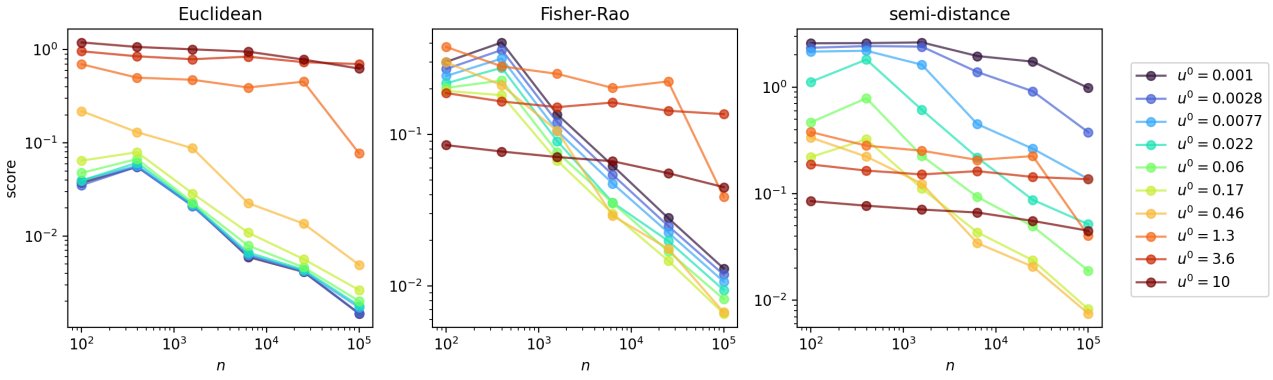

Figure 9: Recovery scores for Algorithm 3 in dimension $d = 1$, for the two-component target $\mu^0 = \frac{1}{2}\delta_{(-1,u^0)} + \frac{1}{2}\delta_{(1,u^0)}$, as a function of the sample size n (horizontal axis, logarithmic) and of the scale parameter u^0 (legend). From left to right: the Wasserstein-type scores S^e (Euclidean ground distance), $S^{\mathfrak{d}_\mathfrak{g}}$ (Fisher-Rao distance) and S^d (semi-distance) introduced in Section 4.3; lower is better. Each point is the median, over 10 independently generated samples, of the best score over three initializations, minimized over the grid of hyperparameters (τ, κ). As discussed above, this selection of (τ, κ) uses the score itself, so that the reported values form an oracle envelope. The vertical scales differ across the three panels.

A phase transition phenomenon becomes apparent in the recovery behavior, and it is visible in all three panels of Figure 9. The quantity to follow is the *slope* of the curves rather than their level. For the smaller

values of u^0, that is when the components are well separated relative to their own width, the score decreases steadily with n, as expected from a consistent procedure: larger samples improve recovery accuracy.

In contrast, beyond a critical scale located around $u^0 \approx 1$—precisely the regime in which the standard deviation of each component becomes comparable to the half-distance between the two means—the curves flatten: CPGD no longer recovers the target measure accurately, and increasing the sample size by three orders of magnitude yields only marginal improvement. The location of this transition is consistent across the Euclidean, Fisher-Rao and semi-distance scores, which suggests that it reflects a genuine difficulty of the recovery problem in this regime rather than an artifact of a particular discrepancy measure.

When interpreting these results, it is important to keep in mind that both the semi-distance and the Fisher–Rao distance tend to yield larger values in regimes characterized by small variances. This effect should be taken into consideration when comparing scores across different configurations. It accounts for the reversal of curve rankings between the Euclidean distance plots and the Fisher–Rao (as well as semi-distance) plots.

5 Discussion and open problems

The CPGD algorithm has been studied in a general setting by [Chizat, 2022], and with stochastic gradient methods in [De Castro et al., 2023]. Its specific application to the Gaussian mixture problem with unknown covariance matrices has been overlooked. Three aspects of this contribution deserve emphasis.

From a methodological standpoint, unknown diagonal covariances endow the parameter space with a location-scale Fisher-Rao geometry, which we exploit directly in the position update, beyond its role in the analysis. The resulting Riemannian scheme (RGD) is invariant under any invertible smooth reparameterization of the parameter space, a property that the Euclidean update (EGD) does not share, and one that becomes meaningful precisely because the metric is no longer constant over $\mathcal{X}$.

On the theoretical side, we establish exponential local convergence with explicit constants for the joint weight-and-position dynamics. The non-degeneracy condition underlying this rate bears on the BLASSO solution, which is not known in advance; Theorem 3.2 connects it to an interpretable separation condition on the statistical target, allowing the assumption to be read in terms of the model rather than the optimizer, with an explicit dependence on the regularization parameter κ.

Numerically, on the test cases considered here, our approach performs on a par with the well-established EM algorithm, with an advantage in robustness to an over-specified number of components and a disadvantage on the Kullback-Leibler score when the number of components is known. The accuracy of the method proves largely insensitive to overparameterization, and the experiments of Section 4 exhibit a phase transition governed by the separation between components, consistently across the three ground distances considered.

We now turn to the limitations of the approach and to the questions they leave open. We successively discuss the convergence guarantees and their local nature, the estimation of the scale parameters, the gap between the algorithm covered by our analysis and the one used in practice, and the evaluation of the optimality gap along the iterations.

Convergence issues The discretization of the space of measures destroys convexity, which prevents us from establishing global convergence guarantees. At present, we can only prove the exponential convergence of CPGD when the initialization is close to the solution, provided strong assumptions on the solution (Theorem 3.1). This convexity issue could be bypassed by introducing a spawn-and-prune step to our algorithm (see, *e.g.*, [De Castro et al., 2026]).

Scale parameter estimation Estimating the scale parameters of the components requires handling projection strategies. When using natural gradient descent, one can exploit the symmetry of J_W with respect to u_j and apply the simple transformation $u_j \mapsto |u_j|$. Alternatively, to enforce the nonnegativity of u_j, one may take advantage of the invariance of the Fisher-Rao metric under reparameterization. For instance, we could consider the parameterization $\ln u$, or $\ln(u - u_{\min})$, and use the associated natural gradient descent. The related covariance updates are then

$$u_j^{k+1} = (u_j^k - u_{\min}) \exp\left(-\beta \frac{(2(u_j^k)^2 + \tau^2)^2}{(u_j^k - u_{\min}) 2(u_j^k)^2} \partial_{u_j} J'(x_j^k) \right) + u_{\min} .$$

This issue becomes more challenging when using Riemannian gradient descent, as the exponential map is not globally defined. Freezing particles (see Lemma 3.1) provides a theoretical workaround, but it is not easily implementable in practice.

Optimality gap A natural direction for further investigation is to exploit the KKT conditions associated with the first variation J' (Section 2.2). Formally, let μ_ω^k denote an approximation of the minimizer $\mu_\omega^\star$. The function $J'_{\mu_\omega^k}$ can then be used to derive an upper bound on the optimality gap $J_W(\mu_\omega^k) - J_W(\mu_\omega^\star)$, which is generally sharper than the trivial bound $J_W(\mu_\omega^k)$ obtained from the nonnegativity of J_W. More precisely, assume that $\alpha > 0$ verifies

$$\alpha J'_{\mu_\omega^k}(x) \geq (\alpha - 1)\kappa \quad \forall\, x \in \mathcal{X}\,. \tag{17}$$

Setting $h = \alpha(\Psi\mu_\omega^k - y)$, we get

$$J_W(\mu_\omega^k) - J_W(\mu_\omega^\star) \leq J_W(\mu_\omega^k) + \frac{1}{2}\left\|h\right\|_{\mathbb{L}}^2 + \langle h, y\rangle_{\mathbb{L}}\,. \tag{18}$$

We refer to Appendix F for the proof. Observe that the case $\alpha = 1$ corresponds exactly to the KKT feasibility condition $J'_{\mu_\omega^k} \geq 0$.

In practice, to determine α satisfying the condition displayed in (17), one may compute either an approximation $\underline{J'}$ of $\inf_{x\in\mathcal{X}} J'_{\mu_\omega^k}(x)$, or a rigorous lower bound, for instance $-\left\|\Psi\mu_\omega^k - y\right\|_{\mathbb{L}} + \kappa$. A valid choice is then $\alpha = \frac{\kappa}{-\min\{\underline{J'}\,,\,0\}+\kappa}$, which is the *largest* value of α ensuring the above condition. Indeed, (17) bounds α from above, while (18) degenerates to the trivial bound $J_W(\mu_\omega^k)$ when $\alpha = 0$: the bound is therefore all the more informative as α is large. When no estimate $\underline{J'}$ is available, the rigorous fallback $\alpha = \frac{\kappa}{\max\{\|\Psi\mu_\omega^k - y\|_{\mathbb{L}}\,,\,\kappa\}}$ of Appendix F is always admissible.

This construction is an instance of a general duality-gap mechanism, which clarifies both its scope and its limitations. Writing the objective as $F(\Psi\mu) + \kappa R(\mu)$ with $F = \frac{1}{2}\left\|\bullet - y\right\|_{\mathbb{L}}^2$ and $R = \|\bullet\|_{\mathrm{TV}}$, [Aubin-Frankowski and De Castro, 2026] establishes the exact identity

$$\Delta(\mu, h) = L_F(\Psi\mu \parallel h) + \kappa\, L_R(\mu \parallel \eta)\,, \qquad \eta = -\Psi^\star h/\kappa\,,$$

valid for *any* primal μ and *any* dual h, in which L_F and L_R are the Fenchel-Young losses of F and R. Evaluated at a dual-feasible h, the left-hand side is computable without any knowledge of $\mu_\omega^\star$ and upper bounds the suboptimality of μ: the bound (18) is exactly $\Delta(\mu_\omega^k, h)$ for the choice $h = \alpha(\Psi\mu_\omega^k - y)$, and condition (17) is precisely the requirement that this h be dual-feasible.

This reading explains the role played by α. For the quadratic fidelity F, the data-fidelity loss $L_F(\Psi\mu \parallel h)$ vanishes exactly at $h = \nabla F(\Psi\mu) = \Psi\mu - y$, that is, at $\alpha = 1$; the whole gap then reduces to the regularizer term. Shrinking α below 1 is what restores dual feasibility when $J'_{\mu_\omega^k}$ takes negative values, but it moves h away from that alignment and inflates L_F. The bound is therefore informative only when α can be taken close to 1, which requires $\inf_{\mathcal{X}} J'_{\mu_\omega^k}$ to be close to nonnegative. Computing an informative lower bound $\underline{J'}$ is, in this respect, the practical bottleneck.

Preliminary experiments are consistent with this analysis. On the setting of Section 4, the bound (18) evaluated at the output of Algorithm 3 remains close to the trivial bound $J_W(\mu_\omega^K)$, and becomes informative only when the components are well separated and κ is taken large enough—a regime that in turn requires more particles at initialization. Larger values of κ also relax (17) and thus allow α closer to 1, so this improvement is not merely an artifact of the term $\kappa\,\|\mu\|_{\mathrm{TV}}$ in the objective. We do not report these experiments here.

A comprehensive study of this duality gap, including the construction of certified dual-feasible points and the resulting early-stopping rules, is carried out in [Aubin-Frankowski and De Castro, 2026]. In particular, a constructive form of the Brøndsted-Rockafellar theorem turns an approximately feasible pair into a dual-feasible one at controlled distance, which removes the need to compute $\underline{J'}$ exactly and makes the stopping criterion certified. The Beurling-LASSO with quadratic fidelity considered here is a special case of the Generalized Beurling-LASSO studied therein, so the questions raised in this section fall squarely within the scope of that work.

References

Amari S.-i. (1998). "Natural Gradient Works Efficiently in Learning". *Neural Computation* 10.2, pp. 251–276. DOI: 10.1162/089976698300017746.

— (2016). *Information geometry and its applications*. Springer. DOI: 10.1007/978-4-431-55978-8.

Aubin-Frankowski P.-C. and De Castro Y. (2026). *Fenchel-Young Duality Gaps: Certified Early Stopping for Regularized Inverse Problems*. arXiv: 2609.17629 [math.OC].

Ben Arous G., Gheissari R., and Jagannath A. (2022). "High-dimensional limit theorems for SGD: Effective dynamics and critical scaling". *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., pp. 25349–25362.

Bouveyron C., Celeux G., Murphy T. B., and Raftery A. E. (2019). *Model-based clustering and classification for data science: with applications in R*. Cambridge series in statistical and probabilistic mathematics. DOI: 10.1017/9781108644181.

Candès E. J. and Fernandez-Granda C. (2014). "Towards a Mathematical Theory of Super-resolution". *Communications on Pure and Applied Mathematics* 67.6, pp. 906–956. DOI: 10.1002/cpa.21455.

Chizat L. (2022). "Sparse optimization on measures with over-parameterized gradient descent". *Mathematical Programming* 194.1, pp. 487–532. DOI: 10.1007/s10107-021-01636-z.

Chizat L. and Bach F. (2018). "On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport". *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.

Chizat L., Oyallon E., and Bach F. (2019). "On Lazy Training in Differentiable Programming". *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Chizat L., Peyré G., Schmitzer B., and Vialard F.-X. (2018a). "An Interpolating Distance Between Optimal Transport and Fisher—Rao Metrics". 18.1, 1–44. DOI: 10.1007/s10208-016-9331-y.

Chizat L., Peyré G., Schmitzer B., and Vialard F.-X. (2018b). "Unbalanced optimal transport: Dynamic and Kantorovich formulations". *Journal of Functional Analysis* 274.11, pp. 3090–3123. DOI: 10.1016/j.jfa.2018.03.008.

Christmann A. and Steinwart I. (2008). *Support vector machines*. DOI: 10.1007/978-0-387-77242-4.

De Castro Y., Gadat S., Marteau C., and Maugis C. (2021). "SuperMix: Sparse Regularization for Mixtures". *Annals of Statistics* 49.3, pp. 1779 –1809. DOI: 10.1214/20-AOS2022.

De Castro Y., Gadat S., and Marteau C. (2023). *FastPart: Over-Parameterized Stochastic Gradient Descent for Sparse optimisation on Measures*. arXiv: 2312.05993 [math.OC].

De Castro Y., Gadat S., and Marteau C. (2026). *Fast Spawn&Prune (FS&P): Global convergence of stochastic conic particle gradient descent via birth/death process*. arXiv: 2605.19784 [math.OC].

De Castro Y. and Gamboa F. (2012). "Exact reconstruction using Beurling minimal extrapolation". *Journal of Mathematical Analysis and Applications* 395.1, pp. 336–354. DOI: 10.1016/j.jmaa.2012.05.011.

Denoyelle Q., Duval V., Peyré G., and Soubies E. (2019). "The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy". *Inverse Problems* 36.1, p. 014001. DOI: 10.1088/1361-6420/ab2a29.

Duchi J., Hazan E., and Singer Y. (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". *Journal of Machine Learning Research* 12.61, pp. 2121–2159.

Duval V. and Peyré G. (2015). "Exact Support Recovery for Sparse Spikes Deconvolution". *Foundations of Computational Mathematics* 15.5, pp. 1315–1355. DOI: 10.1007/s10208-014-9228-6.

Gallouët T. O. and Monsaingeon L. (2017). "A JKO Splitting Scheme for Kantorovich–Fisher–Rao Gradient Flows". *SIAM Journal on Mathematical Analysis* 49.2, pp. 1100–1130. DOI: 10.1137/16M106666X.

Giard R., De Castro Y., and Marteau C. (2025). *Gaussian Mixture Model with unknown diagonal covariances via continuous sparse regularization*. arXiv: 2509.12889v5 [math.ST].

Giard R. and Denis R. (2026). *giardr/blasso_gmm_cpgd: v1.0*. Zenodo. Version v1.0. DOI: 10.5281/zenodo.22808928.

Kondratyev S., Monsaingeon L., and Vorotnikov D. (2016). "A new optimal transport distance on the space of finite Radon measures". *Advances in Differential Equations* 21.11/12, pp. 1117 –1164. DOI: 10.57262/ade/1476369298.

Liero M., Mielke A., and Savaré G. (2016). "Optimal Transport in Competition with Reaction: The Hellinger–Kantorovich Distance and Geodesic Curves". *SIAM Journal on Mathematical Analysis* 48.4, pp. 2869–2911. DOI: 10.1137/15M1041420.

Liero M., Mielke A., and Savaré G. (2018). "Optimal Entropy-Transport problems and a new Hellinger–Kantorovich distance between positive measures". *Inventiones mathematicae* 211.3, pp. 969–1117. DOI: 10.1007/s00222-017-0759-8.

Maugis C. and Michel B. (2011). "A non asymptotic penalized criterion for gaussian mixture model selection". *ESAIM: Probability and Statistics* 15, pp. 41–68. DOI: 10.1051/ps/2009004.

McLachlan G. J. and Krishnan T. (1997). *The EM algorithm and extensions*. DOI: 10.1002/9780470191613.

Nielsen F. (2020). "An Elementary Introduction to Information Geometry". *Entropy* 22.10. DOI: 10.3390/e22101100.

Poon C., Keriven N., and Peyré G. (2023). "The geometry of off-the-grid compressed sensing". *Foundations of Computational Mathematics* 23, pp. 241–327. DOI: 10.1007/s10208-021-09545-5.

Rotskoff G., Jelassi S., Bruna J., and Vanden-Eijnden E. (2019). "Neuron birth-death dynamics accelerates gradient descent and converges asymptotically". *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 5508–5517.

Riemannian Gradient Descent for Gaussian Mixture Models with unknown diagonal covariances

—

Appendix

A Technical lemmas and definitions

A.1 Geodesics, distances

Compatibility with the semi-distance For $A \subset \mathbb{R}^d \times [u_{\min}, +\infty)^d$ and $x_0 \in A$, we define the set containing all points from geodesics connecting x_0 with points of A,

$$\mathcal{G}_{x_0}(A) \coloneqq \{\gamma(y) \,:\, \gamma \text{ is a Fisher-Rao geodesic}\,,\, y \in [0,1]\,,\, \gamma(0) = x_0\,, \gamma(1) \in A\}\,.$$

For $R > 0$ and $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$, we write

$$B_{\mathfrak{d}}(x, R) = \{x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d \,:\, \mathfrak{d}(x, x') \leq R\}\,.$$

From [Giard et al., 2025, Lemma 5.2], we know that for $r > 0$ and $x_0 \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$, it holds that $\mathcal{G}_{x_0}(B_{\mathfrak{d}}(x_0, r)) \subset B_{\mathfrak{d}}(x_0, r)$ (the semi-distance is increasing on geodesics, *i.e.* $y \in [0,1] \mapsto \mathfrak{d}(x_0, \gamma_{x_0 \to x}(y))$ is increasing for any $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$).

Fisher-Rao distance [Giard et al., 2025, Lemmas H.2 and H.3] give a closed-form expression for $\mathfrak{d}_{\mathfrak{g}}$. This expression depends on τ.

A.2 Riemannian derivatives

For the sake of convenience, we reproduce definitions and properties from [Giard et al., 2025].

The non-zero Christoffel symbols associated with $\mathfrak{g}$ are

$$\begin{aligned}
\Gamma^{t_k}{}_{u_k t_k} = \Gamma^{t_k}{}_{t_k u_k} &= \frac{-2u_k}{2u_k^2 + \tau^2}\,, \\
\Gamma^{u_k}{}_{t_k t_k} &= \frac{1}{u_k}\,, \\
\Gamma^{u_k}{}_{u_k u_k} &= \frac{\tau^2 - 2u_k^2}{u_k(2u_k^2 + \tau^2)}
\end{aligned}$$

with $k = 1, \ldots, d$. We define

$$\Gamma^{t_k} = \begin{pmatrix} (\Gamma^{t_k}{}_{t_l t_m})_{1 \leq l,m \leq d} & (\Gamma^{t_k}{}_{t_l u_m})_{1 \leq l,m \leq d} \\ (\Gamma^{t_k}{}_{u_l t_m})_{1 \leq l,m \leq d} & (\Gamma^{t_k}{}_{u_l u_m})_{1 \leq l,m \leq d} \end{pmatrix} \quad \text{and} \quad \Gamma^{u_k} = \begin{pmatrix} (\Gamma^{u_k}{}_{t_l t_m})_{1 \leq l,m \leq d} & (\Gamma^{u_k}{}_{t_l u_m})_{1 \leq l,m \leq d} \\ (\Gamma^{u_k}{}_{u_l t_m})_{1 \leq l,m \leq d} & (\Gamma^{u_k}{}_{u_l u_m})_{1 \leq l,m \leq d} \end{pmatrix}. \tag{19}$$

Definition A.1 (Covariant derivatives, associated norms)**.** Let $\psi \in \mathcal{C}^2(\mathbb{R}^d \times [u_{\min}, +\infty)^d)$. Let $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$.

Covariant derivatives: Let $v, v' \in \mathbb{R}^{2d}$. We define

$$D_0[\psi](x) = \psi(x)\,, \quad D_1[\psi](x)[v] = v^T \nabla \psi(x) = \langle v, \nabla^{\mathfrak{g}} \psi(x) \rangle_x\,, \quad D_2[\psi](x)[v, v'] = v^T H^{\mathfrak{g}} \psi(x) v'\,.$$

Operator norms: For $j \in \{0, 1, 2\}$, we define the operator norm

$$\|D_j[\psi](x)\|_x \coloneqq \sup_{\substack{V = [v_1, \ldots, v_j] \in (\mathbb{R}^{2d})^j \\ \forall l = 1, \ldots, j\,,\, \|v_l\|_x \leq 1}} D_j[\psi](x)[V]\,.$$

Kernel derivatives and associated operators: Let $i, j \in \{0, 1, 2\}$. We define the covariant derivative of the kernel of order i with respect to the first variable x and of order j with respect to the second variable x' by

$$[Q] K^{(ij)}_{\text{norm}}(x, x')[V] = \langle D_i[\Psi](x)[Q], D_j[\Psi](x')[V] \rangle_{\mathbb{L}} \quad \forall\, Q \in (\mathbb{R}^{2d})^i\,, V \in (\mathbb{R}^{2d})^j\,.$$

The associated operator norm is

$$\left\|K^{(ij)}_{\mathrm{norm}}(x,x')\right\|_{x,x'} := \sup_{\substack{Q=[Q_1,\dots,Q_i]\in(\mathbb{R}^{2d})^i\,,\,V=[V_1,\dots,V_j]\in(\mathbb{R}^{2d})^j\\ \forall l\ \|Q_l\|_x\,,\,\|V_l\|_{x'}\le 1}} [Q]K^{(ij)}_{\mathrm{norm}}(x,x')[V]\,.$$

Simplified expressions for the operator norms are given in [Giard et al., 2025, Lemma J.1]. Note in particular that $\|\nabla^{\mathfrak{g}}\psi(x)\|_x = \left\|\mathfrak{g}_x^{-1/2}\nabla\psi(x)\right\|_2$.

We use global controls on the covariant derivatives of the kernel, and also on "euclidean-flavored" counterparts, as well as on the local curvature.

In the following lemma and its proof, we adopt the notation introduced in the proof of [Giard et al., 2025, Lemma J.2]. In particular, we use the decomposition of the kernel as a product: $K_{\mathrm{norm}}(x,x') = \prod_{k=1}^d K_{\mathrm{norm}}(x_k,x'_k)$ where, by abuse of notation, $K_{\mathrm{norm}}(x_k,x'_k)$ denotes the one-dimensional kernel evaluated at $x_k=(t_k,u_k)$ and $x'_k=(t'_k,u'_k)$. The notation b_k stands for either t_k or u_k, and $\mathfrak{g}_{b_kb_k} = \partial b_k\partial_{b'_k}K_{\mathrm{norm}}(x_k,x_k)$ denotes a diagonal entry of the metric tensor.

Note that any derivative of K_{norm} belongs to $\mathbb{L}$, and that for all $x,x'\in\mathbb{R}^d\times[u_{\min},+\infty)^d$, one has (see [Giard et al., 2025, Remark 5.1]) $\langle\partial_1\Psi\delta_x,\partial_2\Psi\delta_{x'}\rangle_{\mathbb{L}} = \partial_1\partial_2K_{\mathrm{norm}}(x,x')$, where ∂_1 (resp. ∂_2) denotes differentiation with respect to x (resp. x').

Finally, the notation $H^{\mathfrak{g}}\partial_{b_k}\Psi\delta_x$ stands for $H^{\mathfrak{g}}\left(x\mapsto\partial_{b_k}\Psi\delta_x\right)$. where ∂_1 (resp. ∂_2) denotes here any derivative w.r.t. x (resp. x').

Lemma A.1 (Controls on the kernel)**.** *Assume that $\tau\le u_{\min}$.*
Global controls on the covariant derivatives: For $i,j\in\{0,1,2\}$, we have

$$\sup_{x,x'\in\mathbb{R}^d\times[u_{\min},+\infty)^d}\left\|K^{(ij)}_{\mathrm{norm}}(x,x')\right\|_{x,x'}\le B_{ij}$$

with

$$\begin{aligned}
B_{00} &= 1\,,\\
B_{10}=B_{01} &= \sqrt{2d}\,,\\
B_{11} &= 2d\,,\\
B_{02}=B_{20} &= \sqrt{4d^2+10d}\,,\\
B_{12}=B_{21} &= \sqrt{2d}B_{02}\,,\\
B_{22} &= 28d^2\,.
\end{aligned}$$

Global controls on the "euclidean-flavored" derivatives: We also have

$$\sup_{x,x'\in\mathbb{R}^d\times[u_{\min},+\infty)^d]}\left\|\mathfrak{g}_x^{-1/2}\nabla_1^2K_{\mathrm{norm}}(x,x')\mathfrak{g}_x^{-1/2}\right\|_2\le 6d =: B^e_{02}$$

and

$$2d\sup_{\substack{x\in\mathbb{R}^d\times[u_{\min},+\infty)^d,\,\|v_1\|_2\le1,\|v_2\|_2\le1\\ b_k\in\{t_k,u_k\}}}\left\|v_1^T\,\mathfrak{g}_{b_kb_k}^{-1/2}\mathfrak{g}_x^{-1/2}H^{\mathfrak{g}}\partial_{b_k}\Psi\delta_x\mathfrak{g}_x^{-1/2}v_2\right\|^2_{\mathbb{L}}\le 1554d^3 =: B^e_{33}\,.$$

Local curvature: Let $x,x'\in\mathbb{R}^d\times[u_{\min},+\infty)^d$. If $\mathfrak{d}(x,x')\le\frac{0.3025}{\sqrt{d}}$, then

$$-K^{(20)}(x,x')[v,v]\ge\bar{\varepsilon}_2\|v\|_x^2\quad\forall\,v\in\mathbb{R}^{2d}\,,\quad\text{with }\bar{\varepsilon}_2=0.13139\,.$$

Proof. The bounds $B_{00},B_{10},B_{11},B_{02},B_{12}$ come from the global controls of the covariant derivatives given in [Giard et al., 2025, Lemma J.2]. The bound B_{22} can be found in the proof of [Giard et al., 2025, Lemma K.1]. To derive the global bounds B^e_{02} and B^e_{33}, we follow the strategy used in the proof of [Giard et al., 2025, Lemma J.2].
B^e_{02}: We refer to `cpgd_sympy.ipynb` in [Giard and Denis, 2026] for the formal computations used to derive the bounds presented here.

Using that the metric tensor $\mathfrak{g}_x$ is diagonal, we have

$$\sup_{x,x'\in\mathbb{R}^d\times[u_{\min},+\infty)^d]}\left\|\mathfrak{g}_x^{-1/2}\nabla_1^2K_{\mathrm{norm}}(x,x')\mathfrak{g}_{x'}^{-1/2}\right\|_2\le 2d\sup_{b_k,b'_l,x,x'\in\mathbb{R}^d\times[u_{\min},+\infty)^d]}\mathfrak{g}_{b_kb_k}^{-1/2}\mathfrak{g}_{b_lb_l}^{-1/2}|\partial_{b_kb_l}K_{\mathrm{norm}}(x,x')|\,.$$

If $b_k \in \{t_k, u_k\}$ and $b_l \in \{t_l, u_l\}$ with $k \neq l$ (*i.e.* the derivatives are not taken in the same dimension), then

$$\begin{aligned}\mathfrak{g}_{b_k b_k}^{-1/2}\mathfrak{g}_{b_l b_l}^{-1/2}|\partial_{b_k b_l}K_{\mathrm{norm}}(x,x')| &\leq \mathfrak{g}_{b_k b_k}^{-1/2}|\partial_{b_k}K_{\mathrm{norm}}(x_k,x_k)|\mathfrak{g}_{b_l b_l}^{-1/2}|\partial_{b_l}K_{\mathrm{norm}}(x_l,x_l)|\,,\\ &\leq 1\end{aligned}$$

as shown in the proof of [Giard et al., 2025, Lemma J.2] when deriving B_{10}. If $b_k = b_l = t_k$,

$$\begin{aligned}\mathfrak{g}_{t_k t_k}^{-1}|\partial_{t_k t_k}K_{\mathrm{norm}}(x,x')| &\leq \mathfrak{g}_{t_k t_k}^{-1}\sqrt{\partial_{t_k t_k}\partial_{t'_k t'_k}K_{\mathrm{norm}}(x_k,x_k)}\,,\\ &= \sqrt{3}\,.\end{aligned}$$

With the same approach, we obtain

$$\mathfrak{g}_{t_k t_k}^{-1/2}\mathfrak{g}_{u_k u_k}^{-1/2}|\partial_{t_k u_k}K_{\mathrm{norm}}(x,x')| \leq \sqrt{5}$$

and

$$\mathfrak{g}_{u_k u_k}^{-1}|\partial_{u_k u_k}K_{\mathrm{norm}}(x,x')| = \sqrt{9 - \frac{2\tau^2}{u_k^2} + \frac{\tau^4}{2u_k^4}} \leq 3\,,$$

where we used that $\tau \leq u_k$. So we can take $B_{02}^e = 6d$.
$\underline{B_{33}^e}$: We have

$$\begin{aligned}&2d \sup_{\substack{x\in\mathbb{R}^d\times[u_{\min},+\infty)^d,\,\|v_1\|_2\leq 1,\|v_2\|_2\leq 1\\ b_k\in\{t_k,u_k\}}} \sup \left\| v_1^T\,\mathfrak{g}_{b_k b_k}^{-1/2}\mathfrak{g}_x^{-1/2}H^{\mathfrak{g}}\partial_{b_k}\Psi\delta_x\mathfrak{g}_x^{-1/2}v_2\right\|_{\mathbb{L}}^2\\ &\leq (2d)^3 \sup_{x\in\mathbb{R}^d\times[u_{\min},+\infty)^d}\sup_{b_k,b_i,b_j}\mathfrak{g}_{b_k b_k}^{-1}\mathfrak{g}_{b_i b_i}^{-1}\mathfrak{g}_{b_j b_j}^{-1}\left\|(H^{\mathfrak{g}}\partial_{b_k}\Psi\delta_x)_{b_i,b_j}\right\|_{\mathbb{L}}^2\,.\end{aligned}$$

If b_k is not in the same dimension as either b_i or b_j,

$$\begin{aligned}\mathfrak{g}_{b_k b_k}^{-1}\mathfrak{g}_{b_i b_i}^{-1}\mathfrak{g}_{b_j b_j}^{-1}\left\|(H^{\mathfrak{g}}\partial_{b_k}\Psi\delta_x)_{b_i,b_j}\right\|_{\mathbb{L}}^2 &\leq \mathfrak{g}_{b_k b_k}^{-1}\partial_{b_k b'_k}K_{\mathrm{norm}}(x_k,x_k)\mathfrak{g}_{b_i b_i}^{-1}\mathfrak{g}_{b_j b_j}^{-1}\left\|(H^{\mathfrak{g}}\Psi\delta_x)_{b_i,b_j}\right\|_{\mathbb{L}}^2\,,\\ &\leq 7\end{aligned}$$

where for the last inequality we used the controls established in the proofs of [Giard et al., 2025, Lemma J.2, Lemma K.1] when deriving B_{11} and B_{22}.

If b_k is taken on the same dimension as b_j or b_i, but b_j and b_i are not on the same dimension (say that for instance b_k is on the same dimension as b_i),

$$\begin{aligned}\mathfrak{g}_{b_k b_k}^{-1}\mathfrak{g}_{b_i b_i}^{-1}\mathfrak{g}_{b_j b_j}^{-1}\left\|(H^{\mathfrak{g}}\partial_{b_k}\Psi\delta_x)_{b_i,b_j}\right\|_{\mathbb{L}}^2 &= \mathfrak{g}_{b_k b_k}^{-1}\mathfrak{g}_{b_i b_i}^{-1}\mathfrak{g}_{b_j b_j}^{-1}\left\|\partial_{b_k b_i b_j}\Psi\delta_x\right\|_{\mathbb{L}}^2\,,\\ &\leq \mathfrak{g}_{b_j b_j}^{-1}\partial_{b_j b_j}K_{\mathrm{norm}}(x_j,x_j)\mathfrak{g}_{b_k b_k}^{-1}\mathfrak{g}_{b_i b_i}^{-1}\partial_{b_k b_i}\partial_{b'_k b'_i}K_{\mathrm{norm}}(x_k,x_k)\,,\\ &\leq 9\,,\end{aligned}$$

as shown when establishing B_{02}^e.

If b_k, b_i, b_j are all taken on the same dimension, $\mathfrak{g}_{b_k b_k}^{-1}\mathfrak{g}_{b_i b_i}^{-1}\mathfrak{g}_{b_j b_j}^{-1}\left\|(H^{\mathfrak{g}}\partial_{b_k}\Psi\delta_x)_{b_i,b_j}\right\|_{\mathbb{L}}^2$ can take the forms

$$\begin{aligned}&\mathfrak{g}_{t_k t_k}^{-3}\left\|\partial_{t_k t_k t_k}\Psi\delta_x - \Gamma^{u_k}{}_{t_k t_k}\partial_{u_k t_k}\Psi\delta_x\right\|_{\mathbb{L}}^2\\ \text{or}\quad &\mathfrak{g}_{t_k t_k}^{-2}\mathfrak{g}_{u_k u_k}^{-1}\left\|\partial_{t_k t_k u_k}\Psi\delta_x - \Gamma^{t_k}{}_{t_k u_k}\partial_{t_k t_k}\Psi\delta_x\right\|_{\mathbb{L}}^2\\ \text{or}\quad &\mathfrak{g}_{t_k t_k}^{-1}\mathfrak{g}_{u_k u_k}^{-2}\left\|\partial_{t_k u_k u_k}\Psi\delta_x - \Gamma^{t_k}{}_{t_k u_k}\partial_{t_k u_k}\Psi\delta_x\right\|_{\mathbb{L}}^2\\ \text{or}\quad &\mathfrak{g}_{u_k u_k}^{-2}\mathfrak{g}_{t_k t_k}^{-1}\left\|\partial_{u_k u_k t_k}\Psi\delta_x - \Gamma^{u_k}{}_{u_k u_k}\partial_{u_k t_k}\Psi\delta_x\right\|_{\mathbb{L}}^2\\ \text{or}\quad &\mathfrak{g}_{u_k u_k}^{-3}\left\|\partial_{u_k u_k u_k}\Psi\delta_x - \Gamma^{u_k}{}_{u_k u_k}\partial_{u_k u_k}\Psi\delta_x\right\|_{\mathbb{L}}^2\,.\end{aligned}$$

These expressions are bounded respectively by 1, 21, $\frac{5\tau^4}{2u_k^4} - \frac{10\tau^2}{u_k^2} + 55 \leq 55$, 55, $\frac{\tau^8}{4u_k^8} + \frac{4\tau^6}{u_k^6} + \frac{35\tau^4}{u_k^4} - \frac{16\tau^2}{u_k^2} + 171 \leq 194.25$. We used that $\tau \leq u_{\min} \leq u_k$. Hence we can take $B_{33}^e = (2d)^3 \times 194.25$.
$\underline{\text{Local curvature:}}$ This control comes from [Giard et al., 2025, Theorem 5.3]. □

Lemma A.2 (Taylor expansions on the kernel)**.** *Let $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$. We have*

$$\|\Psi(\delta_x - \delta_{x'})\|_{\mathbb{L}}^2 \leq B_{02}\mathfrak{d}_{\mathfrak{g}}(x,x')^2\,.$$

Proof. Remark that $K_{\text{norm}}(x,x) = K_{\text{norm}}(x',x') = 1$ and that $\nabla_2 K_{\text{norm}}(x,x) = 0$. Writing $\gamma_{x\to x'}$ the Fisher-Rao geodesic such that $\gamma_{x\to x'}(0) = x$, $\gamma_{x\to x'}(1) = x'$, it comes that

$$\begin{aligned}\|\Psi(\delta_x - \delta_{x'})\|_{\mathbb{L}}^2 &= 2K_{\text{norm}}(x,x) - 2K_{\text{norm}}(x,x')\,,\\ &= -2\dot\gamma_{x\to x'}(0)^T \nabla_2 K_{\text{norm}}(x,x) - 2\int_0^1 (1-y)\dot\gamma_{x\to x'}(y)^T H_2^{\mathfrak{g}} K_{\text{norm}}(x,\gamma_{x\to x'}(y))\dot\gamma_{x\to x'}(y)\,\mathrm{d}y\,,\\ &= -2\int_0^1 (1-y)\dot\gamma_{x\to x'}(y)^T H_2^{\mathfrak{g}} K_{\text{norm}}(x,\gamma_{x\to x'}(y))\dot\gamma_{x\to x'}(y)\,\mathrm{d}y\,.\end{aligned}$$

From the control on the Riemannian Hessian (Lemma A.1), and as $\gamma_{x\to x'}$ stays in $\mathbb{R}^d \times [u_{\min}, +\infty)^d$ (Section A.1), we have

$$\left|\dot\gamma_{x\to x'}(y)^T H_2^{\mathfrak{g}} K_{\text{norm}}(x,\gamma_{x\to x'}(y))\dot\gamma_{x\to x'}(y)\right| \le B_{02}\mathfrak{d}_{\mathfrak{g}}(x,x')^2\,.$$

So $\|\Psi(\delta_x - \delta_{x'})\|_{\mathbb{L}}^2 \le B_{02}\mathfrak{d}_{\mathfrak{g}}(x,x')^2$. □

A.3 Bound on the TV norm of the iterates, control of the Fréchet derivative

Lemma A.3 (TV norm of the iterates)**.** *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d \times [u_{\min}, +\infty)^d$ and $C_{\mathcal{X}} \coloneqq \inf_{x,x'\in\mathcal{X}} K_{\text{norm}}(x,x')$. Let $\mu_\omega^1 = \sum_{j=1} \omega_j^1 \delta_{x_j^1}$ be the initialization of CPGD, with $\omega_j^1 \ge 0$ for $j = 1,\dots,p$. Let μ_ω^k be the kth iterate of the algorithm, with weights updated via $\omega_j^{k+1} = \omega_j^k e^{-\alpha J'_{\mu_\omega^k}(x_j^k)}$, and location-scale parameters verifying $x_j^k \in \mathcal{X}$ for all $j = 1,\dots,p$ and $k \ge 1$. Then μ_ω^k is a nonnegative measure. Moreover, for all $k \ge 1$,*

$$\left\|\mu_\omega^k\right\|_{\text{TV}} \le \max\left\{\frac{1}{C_{\mathcal{X}}}(\|y\|_{\mathbb{L}} - \kappa)e^{\alpha(\|y\|_{\mathbb{L}}-\kappa)}\,,\ \left\|\mu_\omega^1\right\|_{\text{TV}}\right\} \le \max\left\{\frac{1}{C_{\mathcal{X}}}\|y\|_{\mathbb{L}}\, e^{\alpha\|y\|_{\mathbb{L}}}\,,\ \left\|\mu_\omega^1\right\|_{\text{TV}}\right\} =: B_{\text{TV}}\,.$$

Proof. Note that as our kernel is positive ($\langle \Psi\delta_x, \Psi\delta_{x'}\rangle > 0$ for $x, x' \in \mathbb{R}^d \times (0,+\infty)^d$) and as $\mathcal{X}$ is compact, $C_{\mathcal{X}} > 0$. As $\omega_j^k \ge 0$ for all $j = 1,\dots,p$, it comes that, for $x \in \mathcal{X}$,

$$J'_{\mu_\omega^k}(x) = \left\langle \Psi\delta_x, \sum_{j=1}^p \omega_j^k \Psi\delta_{x_j^k} - y\right\rangle_{\mathbb{L}} + \kappa \ge C_{\mathcal{X}}\left\|\mu_\omega^k\right\|_{\text{TV}} - \|y\|_{\mathbb{L}} + \kappa\,. \tag{20}$$

Case $\left\|\mu_\omega^k\right\|_{\text{TV}} \ge \frac{\|y\|_{\mathbb{L}}-\kappa}{C_{\mathcal{X}}}$: From (20), as we assumed that $x_j^k \in \mathcal{X}$, it comes that for all $j = 1,\dots,p$, $J'_{\mu_\omega^k}(x_j^k) \ge 0$. Hence $\omega_j^{k+1} = \omega_j^k e^{-\alpha J'_{\mu_\omega^k}(x_j^k)} \le \omega_j^k$ and $\left\|\mu_\omega^{k+1}\right\|_{\text{TV}} \le \left\|\mu_\omega^k\right\|_{\text{TV}}$.
Case $\left\|\mu_\omega^k\right\|_{\text{TV}} \le \frac{\|y\|_{\mathbb{L}}-\kappa}{C_{\mathcal{X}}}$: Then $\|y\|_{\mathbb{L}} - \kappa \ge 0$ (because $\left\|\mu_\omega^k\right\|_{\text{TV}} \ge 0$) and from (20) we get $-J'_{\mu_\omega^k}(x_j^k) \le \|y\|_{\mathbb{L}} - \kappa$, leading to $\omega_j^{k+1} \le \omega_j^k e^{\alpha(\|y\|_{\mathbb{L}}-\kappa)}$. It comes that $\left\|\mu_\omega^{k+1}\right\|_{\text{TV}} \le \frac{1}{C_{\mathcal{X}}}(\|y\|_{\mathbb{L}} - \kappa)e^{\alpha(\|y\|_{\mathbb{L}}-\kappa)}$.

We showed that for all $k \ge 1$,

$$\left\|\mu_\omega^k\right\|_{\text{TV}} \le \max\left\{\frac{1}{C_{\mathcal{X}}}(\|y\|_{\mathbb{L}} - \kappa)e^{\alpha(\|y\|_{\mathbb{L}}-\kappa)}\,,\ \left\|\mu_\omega^1\right\|_{\text{TV}}\right\}.$$

□

Remark A.1 (Controls on the TV norm of the iterates without $C_{\mathcal{X}}$). Without exploiting the positivity of the kernel K_{norm}, one can still derive a bound on $\left\|\mu_\omega^k\right\|_{\text{TV}}$, although this bound depends on p. Indeed, without using that $C_{\mathcal{X}} > 0$, we have

$$J'_{\mu_\omega^k}(x) \ge \omega_j^k - \|y\|_{\mathbb{L}} + \kappa \quad \forall\, x \in \mathcal{X}\,,$$

from which, by distinguishing the cases $\omega_j^k \ge \|y\|_{\mathbb{L}} - \kappa$ and $\omega_j^k \le \|y\|_{\mathbb{L}} - \kappa$, we can obtain

$$\left\|\mu_\omega^k\right\|_{\text{TV}} \le \max\left\{p(\|y\|_{\mathbb{L}} - \kappa)e^{\alpha(\|y\|_{\mathbb{L}}-\kappa)}\,,\ \left\|\mu_\omega^1\right\|_{\text{TV}}\right\}.$$

Lemma A.4 (Control of J')**.** *For $\mu \in \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d)^+$ and $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$,*

$$-\|y\|_{\mathbb{L}} + \kappa \le J'_\mu(x) \le \|\mu\|_{\text{TV}} + \|y\|_{\mathbb{L}} + \kappa\,.$$

Moreover, for $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$,

$$\left\|\nabla^{\mathfrak{g}} J'_\mu(x)\right\|_x \le \sqrt{B_{11}}(\|\mu\|_{\text{TV}} + \|y\|_{\mathbb{L}})\,,$$

$$\left\|\mathfrak{g}_x^{-1/2} H^{\mathfrak{g}} J'_\mu(x) \mathfrak{g}_x^{-1/2}\right\|_2 \le \sqrt{B_{22}}(\|\mu\|_{\text{TV}} + \|y\|_{\mathbb{L}})$$

and

$$|J'_\mu(x) - J'_\mu(x')| \le \sqrt{B_{02}}\mathfrak{d}_{\mathfrak{g}}(x,x')(\|\mu\|_{\text{TV}} + \|y\|_{\mathbb{L}})\,.$$

Proof. Let $\mu \in \mathcal{M}(\mathbb{R}^d \times [u_{\min}, +\infty)^d)^+$. For $x = (t_1, \dots, t_d, u_1, \dots, u_d) \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$, again as our kernel is positive it comes that $J'_\mu(x) \geq -\|y\|_{\mathbb{L}} + \kappa$. Moreover, as $\|\Psi\delta_x\|_{\mathbb{L}} \leq 1$, we have $J'_\mu(x) \leq \|\mu\|_{\mathrm{TV}} + \|y\|_{\mathbb{L}} + \kappa$. In addition (*cf.* Definition A.1),

$$\begin{aligned}\left\|\nabla^{\mathfrak{g}} J'_\mu(x)\right\|_x &= \left\|\mathfrak{g}_x^{-1/2} \left\langle \nabla_x \Psi\delta_x, \Psi\mu - y\right\rangle_{\mathbb{L}}\right\|_2 , \\ &\leq \sqrt{B_{11}}(\|\mu\|_{\mathrm{TV}} + \|y\|_{\mathbb{L}})\end{aligned}$$

and

$$\begin{aligned}\left\|\mathfrak{g}_x^{-1/2} H^{\mathfrak{g}} J'_\mu(x) \mathfrak{g}_x^{-1/2}\right\|_2 &= \left\|\mathfrak{g}_x^{-1/2} \left\langle H_x^{\mathfrak{g}} \Psi\delta_x, \Psi\mu - y\right\rangle_{\mathbb{L}} \mathfrak{g}_x^{-1/2}\right\|_2 , \\ &\leq \sqrt{B_{22}}(\|\mu\|_{\mathrm{TV}} + \|y\|_{\mathbb{L}}) .\end{aligned}$$

Finally, using Lemma A.2, for $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$ we have

$$\begin{aligned}|J'_\mu(x) - J'_\mu(x')| &= |\left\langle \Psi(\delta_x - \delta_{x'}), \Psi\mu - y\right\rangle_{\mathbb{L}}| , \\ &\leq \|\Psi(\delta_x - \delta_{x'})\|_{\mathbb{L}} (\|\mu\|_{\mathrm{TV}} + \|y\|_{\mathbb{L}}) , \\ &\leq \sqrt{B_{02}} \mathfrak{d}_{\mathfrak{g}}(x, x')(\|\mu\|_{\mathrm{TV}} + \|y\|_{\mathbb{L}}) .\end{aligned}$$

□

B Proof of Proposition 3.1

Let $1 \leq k \leq K$. Our condition on $\mathcal{T}$ ensures that $\mu_\omega^k \in \mathcal{M}(\mathcal{X})^+$. We follow the proof of [De Castro et al., 2023, Theorem 1.2] (inspired by [Chizat, 2022, Section 4.2]).

In this proof, we denote c_a a constant depending on a, where a can stand for instance for $\|y\|_{\mathbb{L}}, \left\|\mu^1\right\|_{\mathrm{TV}}$. Its value may change between lines.

Remark that

$$J_W(\mu_\omega^k) - J_W(\mu_\omega^\star) = \int_{\mathcal{X}} J'_{\mu_\omega^k} \mathrm{d}(\mu_\omega^k - \mu_\omega^\star) - \frac{1}{2}\left\|\Psi(\mu_\omega^k - \mu_\omega^\star)\right\|_{\mathbb{L}}^2 \leq \int_{\mathcal{X}} J'_{\mu_\omega^k} \mathrm{d}(\mu_\omega^k - \mu_\omega^\star) .$$

We introduce an auxiliary measure $\mu_\epsilon^k = \sum_{j=1}^p \omega_j^\epsilon \delta_{x_j^k}$, which shares the same position–scale parameters as μ_ω^k, but with different weights $\omega_j^\epsilon \geq 0$ which are independent of k (remain fixed through the optimization procedure). We have

$$\begin{aligned}J_W(\mu_\omega^k) - J_W(\mu_\omega^\star) &\leq \int_{\mathcal{X}} J'_{\mu_\omega^k} \mathrm{d}(\mu_\omega^k - \mu_\varepsilon^k) + \int_{\mathcal{X}} J'_{\mu_\omega^k} \mathrm{d}(\mu_\varepsilon^k - \mu_\omega^\star) , \\ &= \underbrace{\sum_{j=1}^p (\omega_j^k - \omega_j^\epsilon) J'_{\mu_\omega^k}(x_j^k)}_{A} + \underbrace{\int_{\mathcal{X}} J'_{\mu_\omega^k} \mathrm{d}(\mu_\epsilon^k - \mu_\omega^\star)}_{B} .\end{aligned}$$

Control of A: We introduce the entropy between $\mu_1, \mu_2 \in \mathcal{M}(\mathcal{X})^+$:

$$\mathcal{H}(\mu_1, \mu_2) = -\int_{\mathcal{X}} \ln\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right) \mathrm{d}\mu_2 - \|\mu_2\|_{\mathrm{TV}} + \|\mu_1\|_{\mathrm{TV}} . \tag{21}$$

Remark that as $\omega_j^k > 0$ for all $k \geq 1$ and as $\omega_j^{k+1} = \omega_j^k e^{-\alpha J'_{\mu_\omega^k}(x_j^k)}$, $J'_{\mu_\omega^k}(x_j^k) = -\frac{1}{\alpha} \ln\left(\frac{\omega_j^{k+1}}{\omega_j^k}\right)$.

We also use the inequality $e^{-z} + z - 1 \leq c_{z_{\min}} z^2$ for $z \geq z_{\min}$, together with Lemma A.4 which gives

$\alpha J'_{\mu_\omega^k} \geq -\alpha_{\max} \|y\|_{\mathbb{L}}$ for $\alpha \leq \alpha_{\max}$. With the convention $\omega_j^\epsilon \ln\left(\frac{\omega}{\omega_j^\epsilon}\right) = 0$ when $\omega_j^\epsilon = 0$ and $\omega > 0$, it comes that

$$\begin{aligned}
A &= \sum_{j=1}^p \omega_j^k J'_{\mu_\omega^k}(x_j^k) - \omega_j^\epsilon J'_{\mu_\omega^k}(x_j^k)\,, \\
&= \sum_{j=1}^p \omega_j^k J'_{\mu_\omega^k}(x_j^k) + \frac{1}{\alpha}\omega_j^\epsilon \ln\left(\frac{\omega_j^{k+1}}{\omega_j^k}\right)\,, \\
&= \sum_{j=1}^p \omega_j^k J'_{\mu_\omega^k}(x_j^k) + \frac{1}{\alpha}(\omega_j^{k+1} - \omega_j^k) + \frac{1}{\alpha}\left(-\omega_j^\epsilon \ln\left(\frac{\omega_j^k}{\omega_j^\epsilon}\right) + \omega_j^k + \omega_j^\epsilon \ln\left(\frac{\omega_j^{k+1}}{\omega_j^\epsilon}\right) - \omega_j^{k+1}\right)\,, \\
&= \frac{1}{\alpha}\sum_{j=1}^p \omega_j^k \left(\alpha J'_{\mu_\omega^k}(x_j^k) + e^{-\alpha J'_{\mu_\omega^k}(x_j^k)} - 1\right) + \frac{1}{\alpha}(\mathcal{H}(\mu_\omega^k, \mu_\epsilon^k) - \mathcal{H}(\mu_\omega^{k+1}, \mu_\epsilon^{k+1}))\,, \\
&\leq \alpha c_{\alpha_{\max}, \|y\|_{\mathbb{L}}} \left\|\mu_\omega^k\right\|_{\mathrm{TV}} + \frac{1}{\alpha}(\mathcal{H}(\mu_\omega^k, \mu_\epsilon^k) - \mathcal{H}(\mu_\omega^{k+1}, \mu_\epsilon^{k+1}))\,, \\
&\leq \alpha c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} + \frac{1}{\alpha}(\mathcal{H}(\mu_\omega^k, \mu_\epsilon^k) - \mathcal{H}(\mu_\omega^{k+1}, \mu_\epsilon^{k+1}))\,, \qquad (22)
\end{aligned}$$

where for the last line we used Lemma A.3 to get

$$\begin{aligned}
\left\|\mu_\omega^k\right\|_{\mathrm{TV}} &\leq \max\left\{\frac{1}{C_{\mathcal{X}}}\|y\|_{\mathbb{L}} e^{\alpha_{\max}\|y\|_{\mathbb{L}}}, \left\|\mu_\omega^1\right\|_{\mathrm{TV}}\right\}\,, \\
&\leq c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}}\,.
\end{aligned}$$

Control of B: For f a continuous real-valued function on $\mathcal{X}$, we denote

$$Lip(f) = \sup_{x \neq x' \in \mathcal{X}} \frac{|f(x) - f(x')|}{\mathfrak{d}_{\mathfrak{g}}(x, x')}\,.$$

When f is bounded and has a finite Lipschitz constant, we also define the norm BL (Bounded-Lipschitz) by $\|f\|_{BL} = \|f\|_\infty + Lip(f)$, *cf.* [Chizat, 2022, Section 1.3]. The dual norm is $\|\mu\|_{BL}^* \coloneqq \sup_{\substack{f \in \mathcal{C}(\mathcal{X}) \\ \|f\|_{BL} \leq 1}} \int_{\mathcal{X}} f \,\mathrm{d}\mu$. Using Lemma A.4, for $x, x' \in \mathcal{X}$ we have

$$\begin{aligned}
|J'_{\mu_\omega^k}(x) - J'_{\mu_\omega^k}(x')| &\leq \sqrt{B_{02}}\mathfrak{d}_{\mathfrak{g}}(x, x')(\left\|\mu_\omega^k\right\|_{\mathrm{TV}} + \|y\|_{\mathbb{L}})\,, \\
&\leq c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \mathfrak{d}_{\mathfrak{g}}(x, x')
\end{aligned}$$

(the dependence on d coming from B_{02} is contained in the dependence on $\mathcal{X}$. From Lemmas A.4 and A.3 again, using that $\kappa \leq 1$ it also comes that $|J'_{\mu_\omega^k}| \leq c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}}$. So

$$\begin{aligned}
B &\leq \left\|J'_{\mu_\omega^k}\right\|_{BL} \left\|\mu_\epsilon^k - \mu_\omega^\star\right\|_{BL}^*\,, \\
&\leq c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \left\|\mu_\epsilon^k - \mu_\omega^\star\right\|_{BL}^*\,. \qquad (23)
\end{aligned}$$

Control of the Cesàro average: Gathering (22) and (23), we get

$$J_W(\mu_\omega^k) - J_W(\mu_\omega^\star) \leq \alpha c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} + \frac{1}{\alpha}(\mathcal{H}(\mu_\omega^k, \mu_\epsilon^k) - \mathcal{H}(\mu_\omega^{k+1}, \mu_\epsilon^{k+1})) + c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \left\|\mu_\epsilon^k - \mu_\omega^\star\right\|_{BL}^*\,.$$

As J_W is convex,

$$J_W\left(\frac{1}{K}\sum_{k=1}^K \mu_\omega^k\right) - J_W(\mu_\omega^\star) \leq \frac{1}{K}\sum_{k=1}^K (J_W(\mu_\omega^k) - J_W(\mu_\omega^\star))\,,$$

leading to

$$J_W\left(\frac{1}{K}\sum_{k=1}^K \mu_\omega^k\right) - J_W(\mu_\omega^\star) \leq \frac{\mathcal{H}(\mu_\omega^1, \mu_\epsilon^1)}{\alpha K} + \alpha c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} + \frac{c_{\alpha_{\max}, \|y\|_{\mathbb{L}}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}}}{K} \sum_{k=1}^K \left\|\mu_\epsilon^k - \mu_\omega^\star\right\|_{BL}^* \qquad (24)$$

using that $\mathcal{H} \geq 0$. The triangle inequality gives

$$\frac{1}{K}\sum_{k=1}^K \left\|\mu_\epsilon^k - \mu_\omega^\star\right\|_{BL}^* \leq \left\|\mu_\epsilon^1 - \mu_\omega^\star\right\|_{BL}^* + \frac{1}{K}\sum_{k=2}^K \sum_{l=1}^{k-1} \left\|\mu_\epsilon^{l+1} - \mu_\epsilon^l\right\|_{BL}^*\,.$$

Moreover for $l = 1, \dots, K-1$, according to the assumption on our update scheme $\mathcal{T}$,

$$\left\|\mu_\epsilon^{l+1} - \mu_\epsilon^l\right\|_{BL}^* \leq \sum_{j=1}^p \omega_j^\epsilon \mathfrak{d}_\mathfrak{g}(x_j^{l+1} - x_j^l) \leq \left\|\mu_\epsilon^1\right\|_{\mathrm{TV}} C_\mathcal{T} \beta \,. \tag{25}$$

Indeed, any f such that $\|f\|_{BL} \leq 1$ is 1-Lipschitz, hence leading to

$$\int_\mathcal{X} f \,\mathrm{d}(\mu_\epsilon^{l+1} - \mu_\epsilon^l) = \sum_{j=1}^p \omega_j^\epsilon (f(x_j^{l+1}) - f(x_j^l)) \leq \sum_{j=1}^p \omega_j^\epsilon \mathfrak{d}_\mathfrak{g}(x_j^{l+1} - x_j^l) \,.$$

Using (24) together with (25), we obtain

$$\begin{aligned} J_W\left(\frac{1}{K}\sum_{k=1}^K \mu_\omega^k\right) - J_W(\mu_\omega^\star) \leq \frac{\mathcal{H}(\mu_\omega^1, \mu_\epsilon^1)}{\alpha K} + \alpha c_{\alpha_{\max}, \|y\|_\mathbb{L}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \\ + c_{\alpha_{\max}, \|y\|_\mathbb{L}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \left\|\mu_\epsilon^1 - \mu_\omega^\star\right\|_{BL}^* + K\beta \left\|\mu_\epsilon^1\right\|_{\mathrm{TV}} c_{\alpha_{\max}, \|y\|_\mathbb{L}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} C_\mathcal{T} \,. \end{aligned} \tag{26}$$

Choice of $\{\omega_j^\epsilon\}_{j=1}^p$ and control of $\left\|\mu_\epsilon^1 - \mu_\omega^\star\right\|_{BL}^*$: Let $\{\mathcal{X}_j\}_{j=1}^p$ be disjoint regions of $\mathcal{X}$ such that $x_j^1 \in \mathcal{X}_j$ for $j = 1, \dots, p$, and let $\mathcal{X}_0 \coloneqq \mathcal{X} \setminus \left(\cup_{j=1}^p \mathcal{X}_j\right)$. We set

$$\omega_j^\epsilon = \mu_\omega^\star(\mathcal{X}_j) \,, \quad \forall\, j = 1, \dots, p \,. \tag{27}$$

Remark that $\left\|\mu_\epsilon^1\right\|_{\mathrm{TV}} \leq \|\mu_\omega^\star\|_{\mathrm{TV}}$.

Let f such that $\|f\|_{BL} \leq 1$. Then f is 1-Lipschitz and $\|f\|_\infty \leq 1$. We get

$$\begin{aligned} \int_\mathcal{X} f \,\mathrm{d}(\mu_\epsilon^1 - \mu_\omega^\star) &= \sum_{j=1}^p \int_{\mathcal{X}_j} f \,\mathrm{d}(\mu_\epsilon^1 - \mu_\omega^\star) - \int_{\mathcal{X}_0} f \,\mathrm{d}\mu_\omega^\star \,, \\ &\leq \sum_{j=1}^p \mu_\omega^\star(\mathcal{X}_j) \sup_{x \in \mathcal{X}_j} \mathfrak{d}_\mathfrak{g}(x_j^1, x) + \mu_\omega^\star(\mathcal{X}_0) \,, \\ &\leq \|\mu_\omega^\star\|_{\mathrm{TV}} \sup_{\substack{x \in \mathcal{X}_j \,, \\ j=1,\dots,p}} \mathfrak{d}_\mathfrak{g}(x_j^1, x) + \mu_\omega^\star(\mathcal{X}_0) \,, \end{aligned}$$

from which we deduce that

$$\left\|\mu_\epsilon^1 - \mu_\omega^\star\right\|_{BL}^* \leq \|\mu_\omega^\star\|_{\mathrm{TV}} \sup_{\substack{x \in \mathcal{X}_j \,, \\ j=1,\dots,p}} \mathfrak{d}_\mathfrak{g}(x_j^1, x) + \mu_\omega^\star(\mathcal{X}_0) \,. \tag{28}$$

Control of $\mathcal{H}(\mu_\omega^1, \mu_\epsilon^1)$: Recalling (21) and using (27), we get

$$\begin{aligned} \mathcal{H}(\mu_\omega^1, \mu_\epsilon^1) &= -\sum_{\substack{j \in \{1,\dots,p\} \,, \\ \mu_\omega^\star(\mathcal{X}_j) > 0}} \ln\left(\frac{\omega_j^1}{\mu_\omega^\star(\mathcal{X}_j)}\right) \mu_\omega^\star(\mathcal{X}_j) - \left\|\mu_\epsilon^1\right\|_{\mathrm{TV}} + \left\|\mu_\omega^1\right\|_{\mathrm{TV}} \,, \\ &= \sum_{\substack{j \in \{1,\dots,p\} \,, \\ \mu_\omega^\star(\mathcal{X}_j) > 0}} \left(-\ln\left(\frac{\omega_j^1}{\mu_\omega^\star(\mathcal{X}_j)}\right) \mu_\omega^\star(\mathcal{X}_j) - \mu_\omega^\star(\mathcal{X}_j)\right) + \left\|\mu_\omega^1\right\|_{\mathrm{TV}} \eqqcolon H_{\mu_\omega^1, \mu_\omega^\star} \,. \end{aligned} \tag{29}$$

Conclusion: Choosing $\alpha = \frac{1}{\sqrt{K}} \leq 1 \eqqcolon \alpha_{\max}$ and $0 \leq \beta \leq \frac{1}{K^{3/2} C_\mathcal{T}}$, gathering (26), (28) and (27), we get

$$\begin{aligned} J_W\left(\frac{1}{K}\sum_{k=1}^K \mu_\omega^k\right) - J_W(\mu_\omega^\star) &\leq \frac{H_{\mu_\omega^1, \mu_\omega^\star}}{\alpha K} + \alpha c_{\alpha_{\max}, \|y\|_\mathbb{L}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} + c_{\alpha_{\max}, \|y\|_\mathbb{L}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \left(\|\mu_\omega^\star\|_{\mathrm{TV}} \, r_{\mu_\omega^1, \mathcal{X}} + \mu_\omega^\star(\mathcal{X}_0)\right) \\ &\quad + K\beta \|\mu_\omega^\star\|_{\mathrm{TV}} C_\mathcal{T} c_{\alpha_{\max}, \|y\|_\mathbb{L}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \,, \\ &\lesssim_{\|\mu_\omega^\star\|_{\mathrm{TV}}, \|y\|_\mathbb{L}, \mathcal{X}, \|\mu_\omega^1\|_{\mathrm{TV}}} \frac{H_{\mu_\omega^1, \mu_\omega^\star} + 1}{\sqrt{K}} + r_{\mu_\omega^1, \mathcal{X}} + \mu_\omega^\star(\mathcal{X}_0) \,. \end{aligned}$$

C Proof of Lemma 3.1

Well-definedness of the updates We denote $\overline{\mathcal{X}^C} \coloneqq (\mathbb{R}^d \times [u_{\min}, +\infty)^d) \setminus \mathring{\mathcal{X}}$. Remark that

$$\mathfrak{d}_{\mathfrak{g}}(\overline{\mathcal{X}^C}, \tilde{\mathcal{X}}) \coloneqq \min_{x \in \overline{\mathcal{X}^C},\, x' \in \tilde{\mathcal{X}}} \mathfrak{d}_{\mathfrak{g}}(x, x') > 0\,,$$

since $\mathcal{X}, \tilde{\mathcal{X}}$ are compact sets such that $\tilde{\mathcal{X}} \subset \mathring{\mathcal{X}}$. This minimum distance depends on $\mathcal{X}, \tilde{\mathcal{X}}$, and τ (through the expression of $\mathfrak{d}_{\mathfrak{g}}$—*cf.* Appendix A.1). It therefore suffices to choose β small enough so that

$$\left\|\beta \nabla^{\mathfrak{g}} J'_\mu(x_j^k)\right\|_{x_j^k} \leq \mathfrak{d}_{\mathfrak{g}}(\overline{\mathcal{X}^C}, \tilde{\mathcal{X}})$$

whenever $x_j^k \in \tilde{\mathcal{X}}$. This guarantees that $x_j^{k+1} = \exp_{x_j^k}(-\beta \nabla^{\mathfrak{g}} J'_\mu(x_j^k))$ remains in $\mathcal{X}$.

Moreover, we get from Lemmas A.4 and A.3 that for $\mu \in \mathcal{M}(\mathcal{X})^+$ and $x \in \mathcal{X}$, $\left\|\nabla^{\mathfrak{g}} J'_\mu(x)\right\|_x \leq \sqrt{B_{11}}(\|\mu\|_{\mathrm{TV}} + \|y\|_{\mathbb{L}})$ and that $\left\|\mu_\omega^k\right\|_{\mathrm{TV}} \leq \max\left\{\frac{1}{C_{\mathcal{X}}} \|y\|_{\mathbb{L}}\, e^{\|y\|_{\mathbb{L}}}, \left\|\mu_\omega^1\right\|_{\mathrm{TV}}\right\}$ (we used that $\alpha \leq 1$).

Therefore, the required smallness of β depends only on $\mathfrak{d}_{\mathfrak{g}}(\overline{\mathcal{X}^C}, \tilde{\mathcal{X}}), C_{\mathcal{X}}, \|y\|_{\mathbb{L}}, \left\|\mu_\omega^1\right\|_{\mathrm{TV}}, B_{11}$; equivalently, on $\mathcal{X}, \tilde{\mathcal{X}}, \|y\|_{\mathbb{L}}, \left\|\mu_\omega^1\right\|_{\mathrm{TV}}, \tau$ (with the dependence on d through B_{11} encoded in $\mathcal{X}$).

Descent property Let $k \geq 1$. Using (5) we have

$$J_W(\mu_\omega^{k+1}) - J_W(\mu_\omega^k) = \int_{\mathcal{X}} J'_{\mu_\omega^k}(x)\, \mathrm{d}(\mu_\omega^{k+1} - \mu_\omega^k)(x) + \frac{1}{2} \left\|\Psi(\mu_\omega^{k+1} - \mu_\omega^k)\right\|_{\mathbb{L}}^2 .$$

Study of the second order term: We write $\tilde{\mu}_\omega^{k+1} = \sum_{j=1}^p \omega_j^k \delta_{x_j^{k+1}}$. We have

$$\begin{aligned}
\frac{1}{2} \left\|\Psi(\mu_\omega^{k+1} - \mu_\omega^k)\right\|_{\mathbb{L}}^2 &\leq \left\|\Psi(\mu_\omega^{k+1} - \tilde{\mu}_\omega^{k+1})\right\|_{\mathbb{L}}^2 + \left\|\Psi(\mu^k - \tilde{\mu}_\omega^{k+1})\right\|_{\mathbb{L}}^2 , \\
&= \underbrace{\left\|\sum_{j=1}^p \omega_j^k (e^{-\alpha J'_{\mu_\omega^k}(x_j^k)} - 1) \Psi \delta_{x_j^{k+1}}\right\|_{\mathbb{L}}^2}_{E_1} + \underbrace{\left\|\sum_{j=1}^p \omega_j^k \Psi(\delta_{x_j^{k+1}} - \delta_{x_j^k})\right\|_{\mathbb{L}}^2}_{E_2} .
\end{aligned}$$

As $J'_{\mu_\omega^k}(x_j^k) \leq B_{\mathrm{TV}} + \|y\|_{\mathbb{L}} + 1$ (we used Lemmas A.3 and A.4 with $\kappa \leq 1$), for α small enough depending only on $B_{\mathrm{TV}}, \|y\|_{\mathbb{L}}$, it holds that $|e^{-\alpha J'_{\mu_\omega^k}(x_j^k)} - 1| \leq 2\alpha |J'_{\mu_\omega^k}(x_j^k)|$. So

$$\begin{aligned}
E_1 &\leq \left\|\mu_\omega^k\right\|_{\mathrm{TV}} \sum_{j=1}^p \omega_j^k (e^{-\alpha J'_{\mu_\omega^k}(x_j^k)} - 1)^2 \left\|\Psi \delta_{x_j^{k+1}}\right\|_{\mathbb{L}}^2 , \\
&\leq 4 B_{\mathrm{TV}} \sum_{j=1}^p \omega_j^k \alpha^2 |J'_{\mu_\omega^k}(x_j^k)|^2 . \qquad (30)
\end{aligned}$$

Using Lemma A.2, we also have

$$\begin{aligned}
E_2 &\leq \left\|\mu_\omega^k\right\|_{\mathrm{TV}} \sum_{j=1}^p \omega_j^k \left\|\Psi(\delta_{x_j^{k+1}} - \delta_{x_j^k})\right\|_{\mathbb{L}}^2 , \\
&\leq B_{02} \left\|\mu_\omega^k\right\|_{\mathrm{TV}} \sum_{j=1}^p \omega_j^k \mathfrak{d}_{\mathfrak{g}}(x_j^k, x_j^{k+1})^2 , \\
&\leq B_{02} B_{\mathrm{TV}} \beta^2 \sum_{\substack{1 \leq j \leq p, \\ x_j^k \in \mathcal{X}}} \omega_j^k \left\|\nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x_j^k)\right\|_{x_j^k}^2 . \qquad (31)
\end{aligned}$$

Study of the first order term:

$$\begin{aligned}\int_{\mathcal{X}} J'_{\mu_\omega^k}(x)\,\mathrm{d}(\mu_\omega^{k+1}-\mu_\omega^k) &= \int_{\mathcal{X}} J'_{\mu_\omega^k}(x)\,\mathrm{d}(\mu_\omega^{k+1}-\tilde{\mu}_\omega^{k+1}) + \int_{\mathcal{X}} J'_{\mu_\omega^k}(x)\,\mathrm{d}(\tilde{\mu}_\omega^{k+1}-\mu_\omega^k)\,,\\ &= \underbrace{\sum_{j=1}^p(\omega_j^{k+1}-\omega_j^k)J'_{\mu_\omega^k}(x_j^k)}_{A} + \underbrace{\sum_{j=1}^p \omega_j^k\left(J'_{\mu_\omega^k}(x_j^{k+1})-J'_{\mu_\omega^k}(x_j^k)\right)}_{B}\\ &\quad + \underbrace{\sum_{j=1}^p(\omega_j^{k+1}-\omega_j^k)(J'_{\mu_\omega^k}(x_j^{k+1})-J'_{\mu_\omega^k}(x_j^k))}_{E_3}\,.\end{aligned}$$

E_3 is a second order term. Indeed, using Lemmas A.4 and A.3,

$$\begin{aligned}|E_3| &\le \sum_{j=1}^p \omega_j^k\left|e^{-\alpha J'_{\mu_\omega^k}(x_j^k)}-1\right||J'_{\mu_\omega^k}(x_j^{k+1})-J'_{\mu_\omega^k}(x_j^k)|\,,\\ &\le 2\alpha\beta\sqrt{B_{02}}(\|y\|_{\mathbb{L}}+\left\|\mu_\omega^k\right\|_{\mathrm{TV}})\sum_{\substack{1\le j\le p,\\ x_j^k\in\tilde{\mathcal{X}}}}\omega_j^k|J'_{\mu_\omega^k}(x_j^k)|\left\|\nabla^{\mathfrak{g}}J'_{\mu_\omega^k}(x_j^k)\right\|_{x_j^k}\,,\\ &\le \sqrt{B_{02}}(\|y\|_{\mathbb{L}}+B_{\mathrm{TV}})\sum_{\substack{1\le j\le p,\\ x_j^k\in\tilde{\mathcal{X}}}}\omega_j^k\left(\alpha^2|J'_{\mu_\omega^k}(x_j^k)|^2+\beta^2\left\|\nabla^{\mathfrak{g}}J'_{\mu_\omega^k}(x_j^k)\right\|_{x_j^k}^2\right)\,. \end{aligned} \tag{32}$$

Again using the bound on $J'_{\mu_\omega^k}(x_j^k)$ from Lemma A.4, we also have, for α small enough only depending on $B_{\mathrm{TV}}, \|y\|_{\mathbb{L}}$,

$$|e^{-\alpha J'_{\mu_\omega^k}(x_j^k)}-1+\alpha J'_{\mu_\omega^k}(x_j^k)| \le \alpha^2 J'_{\mu_\omega^k}(x_j^k)^2\,.$$

By dealing separately with the cases $J'_{\mu_\omega^k}(x_j^k)>0$ and $J'_{\mu_\omega^k}(x_j^k)\le 0$, we get

$$\begin{aligned}A &= \sum_{j=1}^p \omega_j^k(e^{-\alpha J'_{\mu_\omega^k}(x_j^k)}-1)J'_{\mu_\omega^k}(x_j^k)\,,\\ &\le -\alpha\sum_{j=1}^p\omega_j^k|J'_{\mu_\omega^k}(x_j^k)|^2+\alpha^2\sum_{j=1}^p\omega_j^k|J'_{\mu_\omega^k}(x_j^k)|^3\,,\\ &\le -\alpha\sum_{j=1}^p\omega_j^k|J'_{\mu_\omega^k}(x_j^k)|^2+\alpha^2(B_{\mathrm{TV}}+\|y\|_{\mathbb{L}}+1)\sum_{j=1}^p\omega_j^k|J'_{\mu_\omega^k}(x_j^k)|^2\,.\end{aligned} \tag{33}$$

To deal with B, we consider 2 cases: if $x_j^k\in\mathcal{X}\setminus\tilde{\mathcal{X}}$, $x_j^{k+1}=x_j^k$ and $\omega_j^k\left(J'_{\mu_\omega^k}(x_j^{k+1})-J'_{\mu_\omega^k}(x_j^{k+1})\right)=0$. If $x_j^k\in\tilde{\mathcal{X}}$, by definition of the update using the exponential map we have (using the control on $H^{\mathfrak{g}}J'_{\mu_\omega^k}$ from Lemma A.4)

$$\begin{aligned}J'_{\mu_\omega^k}(x_j^{k+1})-J'_{\mu_\omega^k}(x_j^k) &= \left\langle\dot{\gamma}_{x_j^k\to x_j^{k+1}}(0)^T,\nabla^{\mathfrak{g}}J'_{\mu_\omega^k}(x_j^k)\right\rangle_{x_j^k}\\ &\quad+\int_0^1(1-y)\dot{\gamma}_{x_j^k\to x_j^{k+1}}(y)^T H^{\mathfrak{g}}J'_{\mu_\omega^k}(\gamma_{x_j^k\to x_j^{k+1}}(y))\dot{\gamma}_{x_j^k\to x_j^{k+1}}(y)\,\mathrm{d}y\,,\\ &\le -\beta\left\|\nabla^{\mathfrak{g}}J'_{\mu_\omega^k}(x_j^k)\right\|_{x_j^k}^2+\frac{1}{2}\beta^2\left\|\nabla^{\mathfrak{g}}J'_{\mu_\omega^k}(x_j^k)\right\|_{x_j^k}^2\sqrt{B_{22}}(\|y\|_{\mathbb{L}}+B_{\mathrm{TV}})\,.\end{aligned}$$

Finally,

$$\begin{aligned}B &= \sum_{\substack{1\le j\le p,\\ x_j^k\in\tilde{\mathcal{X}}}}\omega_j^k\left(J'_{\mu_\omega^k}(x_j^{k+1})-J_{\mu_\omega^k}(x_j^k)\right)\,,\\ &\le -\beta\sum_{\substack{1\le j\le p,\\ x_j^k\in\tilde{\mathcal{X}}}}\omega_j\left\|\nabla^{\mathfrak{g}}J'_{\mu_\omega^k}(x_j^k)\right\|_{x_j^k}^2+\frac{1}{2}\beta^2\sum_{\substack{1\le j\le p,\\ x_j^k\in\tilde{\mathcal{X}}}}\omega_j\left\|\nabla^{\mathfrak{g}}J'_{\mu_\omega^k}(x_j^k)\right\|_{x_j^k}^2\sqrt{B_{22}}(\|y\|_{\mathbb{L}}+B_{\mathrm{TV}})\,.\end{aligned} \tag{34}$$

Choice of α, β and end of the proof: Combining (30), (31), (32), (33), (34), we get

$$
\begin{aligned}
J_W(\mu_\omega^{k+1}) - J_W(\mu_\omega^k) \leq & -\alpha \sum_{j=1}^{p} \omega_j^k |J'_{\mu_\omega^k}(x_j^k)|^2 - \beta \sum_{\substack{1 \leq j \leq p, \\ x_j^k \in \tilde{\mathcal{X}}}} \omega_j^k \left\| \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x_j^k) \right\|_{x_j^k}^2 \\
& + \alpha^2 (5B_{\mathrm{TV}} + \|y\|_{\mathbb{L}} + 1 + \sqrt{B_{02}}(\|y\|_{\mathbb{L}} + B_{\mathrm{TV}})) \sum_{j=1}^{p} \omega_j^k |J'_{\mu_\omega^k}(x_j^k)|^2 \\
& + \beta^2 \left(\frac{1}{2}\sqrt{B_{22}}(\|y\|_{\mathbb{L}} + B_{\mathrm{TV}}) + B_{02} B_{\mathrm{TV}} + \sqrt{B_{02}}(\|y\|_{\mathbb{L}} + B_{\mathrm{TV}}) \right) \\
& \times \sum_{\substack{1 \leq j \leq p, \\ x_j^k \in \tilde{\mathcal{X}}}} \omega_j^k \left\| \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x_j^k) \right\|_{x_j^k}^2 .
\end{aligned}
$$

We can choose α, β small enough only depending on $\mathcal{X}, \tilde{\mathcal{X}}, \|y\|_{\mathbb{L}}, \tau, \left\|\mu_\omega^1\right\|_{\mathrm{TV}}$ (the dependence on d is captured by the dependence on $\mathcal{X}$) such that the gradients steps are well-defined and ensuring that for all $k \geq 1$,

$$
J_W(\mu_\omega^{k+1}) - J_W(\mu_\omega^k) \leq -\frac{1}{2} \int_{\tilde{\mathcal{X}}} \left(\alpha |J'_{\mu_\omega^k}(x)|^2 + \beta \left\| \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x) \right\|_x^2 \right) \mathrm{d}\mu_\omega^k(x) - \frac{1}{2} \int_{\mathcal{X} \setminus \tilde{\mathcal{X}}} \alpha |J'_{\mu_\omega^k}(x)|^2 \, \mathrm{d}\mu_\omega^k(x) .
$$

D Proof of Theorem 3.1

We rewrite here the calculations of the proof of [Chizat, 2022, Corollary 3.5] with the geometric objects induced by our particular geometry, including geodesics and Riemannian derivatives. In particular, we make use of the specific retraction (RGD), corresponding to Riemannian gradient descent.

Roadmap of the proof In the next subsections, we assume that Assumption 1 holds. We fix the parameters α and β as in Lemma 3.1, so that for all $k \geq 1$,

$$
J_W(\mu_\omega^{k+1}) - J_W(\mu_\omega^k) \leq -\frac{1}{2} \min\{\alpha, \beta\} \left(\int_{\tilde{\mathcal{X}}} \left(|J'_{\mu_\omega^k}(x)|^2 + \left\| \nabla^{\mathfrak{g}} J'_{\mu_\omega^k}(x) \right\|_x^2 \right) \mathrm{d}\mu_\omega^k(x) + \int_{\mathcal{X} \setminus \tilde{\mathcal{X}}} |J'_{\mu_\omega^k}(x)|^2 \, \mathrm{d}\mu_\omega^k(x) \right) . \tag{35}
$$

This estimate on the objective decrease serves as the starting point for the convergence analysis of the algorithm.

When the solution $\mu_\omega^\star$ is discrete and of the form $\sum_{j=1}^{p^\star} \omega_j^\star \delta_{x_j^\star}$, we define the associated near and far regions, for $r > 0$, as

$$
\mathcal{X}_j^{\mathrm{near}}(r) \coloneqq \{x \in \mathcal{X} \,:\, \mathrm{d}(x, x_j^\star) \leq r\} , \qquad \mathcal{X}^{\mathrm{far}}(r) \coloneqq \mathcal{X} \setminus \mathcal{X}^{\mathrm{near}}(r) \qquad \text{with} \quad \mathcal{X}^{\mathrm{near}}(r) = \bigcup_j \mathcal{X}_j^{\mathrm{near}}(r) .
$$

We choose r small enough so that $(\mathcal{X}_j^{\mathrm{near}}(r))_{1 \leq j \leq p^\star}$ are disjoint and contained in $\tilde{\mathcal{X}}$ (Lemma D.3). In particular, this implies that $\mathcal{X} \setminus \tilde{\mathcal{X}} \subset \mathcal{X}^{\mathrm{far}}(r)$. As a consequence, (35) yields

$$
J_W(\mu_\omega^{k+1}) - J_W(\mu_\omega^k) \leq -\frac{1}{2} \min\{\alpha, \beta\} G_{\mu_\omega^k} \tag{36}
$$

where for $\mu \in \mathcal{M}(\mathcal{X})^+$,

$$
G_\mu \coloneqq \int_{\mathcal{X}^{\mathrm{near}}(r)} \left(|J'_\mu(x)|^2 + \left\| \nabla^{\mathfrak{g}} J'_\mu(x) \right\|_x^2 \right) \mathrm{d}\mu(x) + \int_{\mathcal{X}^{\mathrm{far}}(r)} |J'_\mu(x)|^2 \, \mathrm{d}\mu(x) . \tag{37}
$$

Using Assumption 1, we derive quantitative controls on the certificate $\eta^\star$ over the near and far regions (see Appendix D.1).

In Appendix D.2, we establish a lower bound on the objective decrease G_μ for measures $\mu \in \mathcal{M}(\mathcal{X})^+$ such that $\|\mu\|_{\mathrm{TV}} \leq B_{\mathrm{TV}}$.

In Appendix D.3, we derive an upper bound on the optimality gap $J_W(\mu) - J_W(\mu_\omega^\star)$.

Finally, in Appendix D.4, we combine these estimates. For r sufficiently small, and for μ such that $J_W(\mu)$ is close enough to $J_W(\mu_\omega^\star)$, we show that $G_\mu \gtrsim J_W(\mu) - J_W(\mu_\omega^\star)$.

Recalling (36), this relation leads to exponential convergence, provided that the initialization is chosen sufficiently close to the solution (see Appendix D.5).

Throughout the proof, the notations A, B, and E_i may refer to different quantities depending on the subsection.

D.1 Controls of the certificate over near and far regions

D.1.1 Control of the Fisher-Rao metric

Lemma D.1 (Bounding the semi-distance by the Fisher–Rao distance). *For all $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$ such that $\mathfrak{d}(x, x') \leq \frac{0.3025}{\sqrt{d}}$, it holds that*

$$\mathfrak{d}_{\mathfrak{g}}(x, x')^2 \leq \frac{1}{\bar{\varepsilon}_2} \mathfrak{d}(x, x')^2 \quad \textit{with} \quad \bar{\varepsilon}_2 = 0.13139$$

along with

$$\mathfrak{d}_{\mathfrak{g}}(x, x')^2 \geq \frac{1}{\bar{\varepsilon}_3} \mathfrak{d}(x, x')^2 \quad \textit{with} \quad \bar{\varepsilon}_3 = 2.84\,.$$

The radius $\frac{0.3025}{\sqrt{d}}$ and the constants $\bar{\varepsilon}_2$, $\bar{\varepsilon}_3$ are the local positive curvature parameters of K_{norm} obtained in [Giard et al., 2025, Theorem 5.3 and Lemma 5.4].

Proof. Let $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$ such that $\mathfrak{d}(x, x') \leq \frac{0.3025}{\sqrt{d}}$. For $y \in [0, 1]$, $\gamma_{x \to x'}(y) \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$ and $\mathfrak{d}(\gamma_{x \to x'}(y), x) \leq \mathfrak{d}(x, x') \leq \frac{0.3025}{\sqrt{d}}$ (*cf.* Appendix A.1). We then use the control on the local curvature (Lemma A.1) to deduce that $-K^{(20)}(\gamma_{x \to x'}(y), x)[v, v] \geq \bar{\varepsilon}_2 \|v\|^2_{\gamma_{x \to x'}(y)}$ for $v \in \mathbb{R}^{2d}$.

So

$$\begin{aligned}
1 - e^{-\frac{1}{2}\mathfrak{d}(x,x')^2} &= 1 - K_{\mathrm{norm}}(x', x)\,, \\
&= -\int_0^1 (1-y) K^{(20)}(\gamma_{x \to x'}(y), x)[\dot{\gamma}_{x \to x'}(y), \dot{\gamma}_{x \to x'}(y)]\,\mathrm{d}y\,, \\
&\geq \int_0^1 (1-y)\bar{\varepsilon}_2 \|\dot{\gamma}_{x \to x'}(y)\|^2_{\gamma_{x \to x'}(y)}\,\mathrm{d}y\,, \\
&= \frac{\bar{\varepsilon}_2}{2} \mathfrak{d}_{\mathfrak{g}}(x, x')^2\,.
\end{aligned}$$

As $z^2 \geq 2(1 - e^{-\frac{1}{2}z^2})$, we deduce that

$$\mathfrak{d}_{\mathfrak{g}}(x, x')^2 \leq \frac{1}{\bar{\varepsilon}_2} \mathfrak{d}(x, x')^2\,.$$

The second inequality comes from [Giard et al., 2025, Lemma 5.4]. □

Lemma D.2 (Local variation of the metric tensor). *For $x, x' \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$ such that $\mathfrak{d}(x, x') \leq 1$,*

$$\left\| \mathfrak{g}_{x'}^{1/2} \mathfrak{g}_x^{-1/2} \right\|_2 \leq \sqrt{6} e =: C_{\mathfrak{g}}\,.$$

Proof. Recall that

$$\mathfrak{g}_x = \mathrm{diag}\left(\frac{1}{2u_1^2 + \tau^2}, \ldots, \frac{1}{2u_d^2 + \tau^2}, \frac{2u_1^2}{(2u_1^2 + \tau^2)^2}, \ldots, \frac{2u_d^2}{(2u_d^2 + \tau^2)^2} \right)\,.$$

So the eigenvalues of $\mathfrak{g}_{x'}^{1/2} \mathfrak{g}_x^{-1/2}$ are bounded by $\lambda \coloneqq \max_i \left\{ \frac{\sqrt{2(u_i)^2+\tau^2}}{\sqrt{2(u_i')^2+\tau^2}}, \frac{(2(u_i)^2+\tau^2)\sqrt{2(u_i')^2}}{(2(u_i')^2+\tau^2)\sqrt{2u_i^2}} \right\}$. As $\tau \leq u_{\min}$, $\lambda \leq \sqrt{\frac{3}{2}} \max_i \frac{\sqrt{2(u_i)^2+\tau^2}}{\sqrt{2(u_i')^2+\tau^2}}$. From [Giard et al., 2025, Equation (34)], as $\frac{(u_i')^2+u_i^2+\tau^2}{\sqrt{2(u_i')^2+\tau^2}\sqrt{2u_i^2+\tau^2}} \leq e$ we have

$$\frac{\sqrt{2(u_i)^2+\tau^2}}{\sqrt{2(u_i')^2+\tau^2}} \leq e + \sqrt{e^2 - 1} \leq 2e\,.$$

□

D.1.2 Implications of Assumption 1

Lemma D.3 (Control of the certificate on near and far regions). *Under Assumption 1, we can choose $0 \leq r_{\max} \leq \frac{0.3025}{\sqrt{d}}$ small enough, depending on $d, \mathcal{X}, \tilde{\mathcal{X}}, \tau, \mu_\omega^\star$, such that*

1. *$(\mathcal{X}_j^{\mathrm{near}}(r_{\max}))_{j=1}^{p^\star}$ are non-intersecting regions contained in $\tilde{\mathcal{X}}$,*
2. *$\eta^\star(x) \leq 1 - \varepsilon_0(r_{\max})$ for all $x \in \mathcal{X}^{\mathrm{far}}(r_{\max})$, for $\varepsilon_0(r_{\max}) > 0$ depending on $r_{\max}$, $\eta^\star$, $\mathcal{X}$,*

3. $-vH^{\mathfrak{g}}\eta^\star(x)v \geq \frac{\varepsilon_2}{2}\|v\|_x^2$ *for all* $v \in \mathbb{R}^{2d}$, $x \in \mathcal{G}_{x_j^\star}(B_{\mathscr{d}}(x_j^\star, r_{\max}))$, *implying that* $\eta^\star(x) \leq 1 - \frac{\varepsilon_2}{4}\mathfrak{d}_{\mathfrak{g}}(x, x_j^\star)^2$ *for all* $x \in \mathcal{X}_j^{\mathrm{near}}(r_{\max})$,

4. *For* $0 < r \leq \min\left\{r_{\max}, \sqrt{\frac{4\varepsilon_0(r_{\max})\bar{\varepsilon}_3}{\varepsilon_2}}\right\} =: r'_{\max}$ *and* $x \in \mathcal{X}^{\mathrm{far}}(r)$, $\eta^\star(x) \leq 1 - \varepsilon'_0 r^2$ *(with* $\varepsilon'_0 := \frac{\varepsilon_2}{4\bar{\varepsilon}_3}$*).*

Proof. We exploit the information about the location of the solution provided by Assumption 1, the smoothness of $\mathscr{d}$, the fact that $\mathscr{d}(x, x') = 0 \iff x = x'$, the compactness of $\tilde{\mathcal{X}}$, and that $x_j^\star \in \mathring{\tilde{\mathcal{X}}}$ to establish the first item, for $r_{\max}$ small enough.

The items 2 and 3 come from Assumption 1 together with the smoothness of $\eta^\star$ and the compactness of $\mathcal{X}$. With item 3, we also have, for $x \in \mathcal{X}_j^{\mathrm{near}}(r_{\max})$,

$$\begin{aligned}
\eta^\star(x) &= \eta^\star(x_j^\star) + \dot{\gamma}_{x_j^\star \to x}(0)^T \nabla\eta^\star(x_j^\star) + \int_0^1 (1-y)\dot{\gamma}_{x_j^\star\to x}(y)^T H^{\mathfrak{g}}\eta^\star(\gamma_{x_j^\star\to x}(y))\dot{\gamma}_{x_j^\star\to x}(y)\,\mathrm{d}y\,,\\
&= 1 + \int_0^1 (1-y)\dot{\gamma}_{x_j^\star\to x}(y)^T H^{\mathfrak{g}}\eta^\star(\gamma_{x_j^\star\to x}(y))\dot{\gamma}_{x_j^\star\to x}(y)\,\mathrm{d}y\,,\\
&\leq 1 - \frac{\varepsilon_2}{4}\mathfrak{d}_{\mathfrak{g}}(x, x_j^\star)^2\,.
\end{aligned}$$

The last item can be deduced using the previous ones, along with Lemma D.1, taking $r_{\max} \leq \frac{0.3025}{\sqrt{d}}$. For $x \in \mathcal{X}^{\mathrm{far}}(r_{\max})$, $\eta^\star(x) \leq 1 - \varepsilon_0(r_{\max})$. For $x \in \mathcal{X}^{\mathrm{far}}(r) \setminus \mathcal{X}^{\mathrm{far}}(r_{\max})$, there exists $j \in \{1, \dots p^\star\}$ such that $x \in \mathcal{X}_j^{\mathrm{near}}(r_{\max})$. So as $\mathscr{d}(x, x_j^\star) > r$,

$$\eta^\star(x) \leq 1 - \frac{\varepsilon_2}{4}\mathfrak{d}_{\mathfrak{g}}(x, x_j^\star)^2 \leq 1 - \frac{\varepsilon_2}{4\bar{\varepsilon}_3}\mathscr{d}(x, x_j^\star)^2 \leq 1 - \frac{\varepsilon_2}{4\bar{\varepsilon}_3}r^2\,.$$

□

In the next subsections, we consider $0 < r \leq r'_{\max}$ small enough so that the conclusions of Lemma D.3 apply. In particular, (36) holds. The radius r will be chosen even smaller later in the proof.

Identification of the error terms Throughout the proof, we rely on a basic control of the optimality gap $J_W(\mu) - J_W(\mu_\omega^\star)$ (for $\mu \in \mathcal{M}(\mathcal{X})^+$), which makes it possible to isolate the relevant error terms. Combining (5) with Lemma D.3, and using $J'_{\mu_\omega^\star} = \kappa(1 - \eta^\star)$, we obtain

$$\begin{aligned}
J_W(\mu) - J_W(\mu_\omega^\star) &= \int_{\mathcal{X}} J'_{\mu_\omega^\star}(x)\,\mathrm{d}\mu(x) + \frac{1}{2}\left\|\Psi(\mu - \mu_\omega^\star)\right\|_{\mathbb{L}}^2\,,\\
&= \sum_{j=1}^{p^\star} \kappa \int_{\mathcal{X}_j^{\mathrm{near}}(r)} (1-\eta^\star)\,\mathrm{d}\mu + \kappa\int_{\mathcal{X}^{\mathrm{far}}(r)}(1-\eta^\star)\,\mathrm{d}\mu + \frac{1}{2}\left\|\Psi(\mu - \mu_\omega^\star)\right\|_{\mathbb{L}}^2\,,\\
&\geq \kappa\frac{\varepsilon_2}{4}\sum_{j=1}^{p^\star}\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_j^\star, x)^2\,\mathrm{d}\mu(x) + \kappa\varepsilon'_0 r^2 \mu(\mathcal{X}^{\mathrm{far}}(r)) + \frac{1}{2}\left\|\Psi(\mu - \mu_\omega^\star)\right\|_{\mathbb{L}}^2\,. \qquad (38)
\end{aligned}$$

D.2 Lower bound on the objective decrease

Let $\mu \in \mathcal{M}^+(\mathcal{X})$ such that $\|\mu\|_{\mathrm{TV}} \leq B_{\mathrm{TV}}$ (*cf.* Lemma A.3). The goal of this section is to provide a lower bound on G_μ (see (37)). There are 3 terms to consider: $\int_{\mathcal{X}^{\mathrm{near}}(r)} |J'_\mu(x)|^2\,\mathrm{d}\mu(x)$, $\int_{\mathcal{X}^{\mathrm{near}}(r)} \left\|\nabla^{\mathfrak{g}} J'_\mu(x)\right\|_x^2\,\mathrm{d}\mu(x)$ and $\int_{\mathcal{X}^{\mathrm{far}}(r)} |J'_\mu(x)|^2\,\mathrm{d}\mu(x)$.

For $x \in \mathcal{X}$, we control $|J'_\mu(x)|$ and $\left\|\nabla^{\mathfrak{g}} J'_\mu(x)\right\|_x$ using the decomposition

$$\begin{aligned}
J'_\mu(x) &= J'_{\mu_\omega^\star}(x) + \langle \Psi\delta_x, \Psi(\mu - \mu_\omega^\star)\rangle_{\mathbb{L}}\,,\\
&= J'_{\mu_\omega^\star}(x) + \int K_{\mathrm{norm}}(x, x')\,\mathrm{d}(\mu - \mu_\omega^\star)(x')\,. \qquad (39)
\end{aligned}$$

We study these terms with Taylor expansions on $J_{\mu_\omega^\star}$ and K_{norm}. Throughout the proof, we will use the identity $J'_{\mu_\omega^\star} = \kappa(1 - \eta^\star)$ together with the controls on $\eta^\star$ provided by Assumption 1 and Lemma D.3. In particular,

$$J'_{\mu_\omega^\star}(x_j^\star) = 0 \quad \text{and} \quad \nabla J'_{\mu_\omega^\star}(x_j^\star) = 0 \quad \forall\, j = 1, \dots, p^\star\,. \qquad (40)$$

We will also rely on the controls on J'_μ given in Lemma A.4.

Definitions for the main terms For any $i, j \in \{1, \dots, p^\star\}$, we denote $\tilde{b}_{x_j} = \int_{\mathcal{X}_j^{\mathrm{near}}} \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) \, \mathrm{d}\mu(x)$, $\tilde{b}_{\omega_j} = \mu(\mathcal{X}_j^{\mathrm{near}}(r)) - \omega_j^\star$, $\tilde{b} = \begin{pmatrix} \tilde{b}_\omega \\ \tilde{b}_x \end{pmatrix}$ and $\tilde{K}(x_j^\star, x_i^\star) = K_{\mathrm{norm}}(x_j^\star, x_i^\star)$. We multiply the derivatives of K_{norm} by the metric, defining

$$\begin{aligned}
\tilde{K}_1(x_j^\star, x_i^\star) &\coloneqq \mathfrak{g}_{x_j^\star}^{-1/2} \nabla_1 K_{\mathrm{norm}}(x_j^\star, x_i^\star) \, , \\
\tilde{K}_2(x_j^\star, x_i^\star) &\coloneqq \mathfrak{g}_{x_i^\star}^{-1/2} \nabla_2 K_{\mathrm{norm}}(x_j^\star, x_i^\star) \, , \\
\tilde{K}_{12}(x_j^\star, x_i^\star) &\coloneqq \mathfrak{g}_{x_j^\star}^{-1/2} \nabla_2 \nabla_1 K_{\mathrm{norm}}(x_j^\star, x_i^\star) \mathfrak{g}_{x_i^\star}^{-1/2} \, .
\end{aligned}$$

We also write $\tilde{H}_j = \mathfrak{g}_{x_j^\star}^{-1/2} H^{\mathfrak{g}} J'_{\mu_\omega^\star}(x_j^\star) \mathfrak{g}_{x_j^\star}^{-1/2}$.

Remark that as $\dot{\gamma}_{x_j^\star \to x_j^\star} = 0$, $\tilde{b}_{x_j} = \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) \, \mathrm{d}(\mu - \mu_\omega^\star)(x)$.

D.2.1 Lower bound on $\int_{\mathcal{X}^{\mathrm{near}}(r)} |J'_\mu|^2 \, \mathrm{d}\mu$

We first derive an upper bound on $\int_{\mathcal{X}^{\mathrm{near}}(r)} |J'_{\mu_\omega^\star}|^2 \, \mathrm{d}\mu$, then a lower bound on

$$\int_{\mathcal{X}^{\mathrm{near}}(r)} \left| \int_{\mathcal{X}} K_{\mathrm{norm}}(x, x') \, \mathrm{d}(\mu - \mu_\omega^\star)(x') \right|^2 \mathrm{d}\mu(x) \, ,$$

and finally conclude using (39).

Control of $J'_{\mu_\omega^\star}(x)$: Let $x \in \mathcal{X}_j^{\mathrm{near}}(r)$. Using (40), we obtain

$$\begin{aligned}
E_1(x) \coloneqq J'_{\mu_\omega^\star}(x) &= J'_{\mu_\omega^\star}(x_j^\star) + \dot{\gamma}_{x_j^\star \to x}(0)^T \nabla J'_{\mu_\omega^\star}(x_j^\star) + \int_0^1 (1-y) \dot{\gamma}_{x_j^\star \to x}(y)^T H^{\mathfrak{g}} J'_{\mu_\omega^\star}(\gamma_{x_j^\star \to x}(y)) \dot{\gamma}_{x_j^\star \to x}(y) \, \mathrm{d}y \, , \\
&= \int_0^1 (1-y) \dot{\gamma}_{x_j^\star \to x}(y)^T H^{\mathfrak{g}} J'_{\mu_\omega^\star}(\gamma_{x_j^\star \to x}(y)) \dot{\gamma}_{x_j^\star \to x}(y) \, \mathrm{d}y \, .
\end{aligned}$$

Applying Lemma A.4, we deduce that

$$\begin{aligned}
\int_{X^{\mathrm{near}}(r)} |E_1(x)|^2 \, \mathrm{d}\mu(x) &\leq \sum_{j=1}^{p^\star} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left(\mathfrak{d}_{\mathfrak{g}}(x_j^\star, x)^2 \frac{\sqrt{B_{22}}}{2} (\|\mu_\omega^\star\|_{\mathrm{TV}} + \|y\|_{\mathbb{L}}) \right)^2 \mathrm{d}\mu(x) \, , \\
&\lesssim_{d, \|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}} r^2 \sum_{j=1}^{p^\star} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_j^\star, x)^2 \, \mathrm{d}\mu(x) \, ,
\end{aligned}$$

where we used that, for $x \in \mathcal{X}_j^{\mathrm{near}}(r)$,

$$\mathfrak{d}_{\mathfrak{g}}(x_j^\star, x)^2 \leq \frac{1}{\bar{\varepsilon}_2} d(x_j^\star, x)^2 \leq \frac{r^2}{\bar{\varepsilon}_2}$$

by Lemma D.1.

Control of $\int K_{\mathrm{norm}}(x,x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x')$: Moreover, for any $x\in\mathcal{X}_j^{\mathrm{near}}(r)$,

$$
\begin{aligned}
&\int K_{\mathrm{norm}}(x,x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x') \\
&\qquad= \int K_{\mathrm{norm}}(x_j^\star,x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x') + \int_0^1 \dot\gamma_{x_j^\star\to x}(y)^T \int \nabla_1 K_{\mathrm{norm}}(\gamma_{x_j^\star\to x}(y),x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x')\,\mathrm{d}y\,, \\
&\qquad= \sum_{i=1}^{p^\star}\int_{\mathcal{X}_i^{\mathrm{near}}(r)} K_{\mathrm{norm}}(x_j^\star,x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x') \\
&\qquad\quad+ \int_0^1 \dot\gamma_{x_j^\star\to x}(y)^T \int \nabla_1 K_{\mathrm{norm}}(\gamma_{x_j^\star\to x}(y),x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x')\,\mathrm{d}y + \int_{\mathcal{X}^{\mathrm{far}}(r)} K_{\mathrm{norm}}(x_j^\star,x')\,\mathrm{d}\mu(x')\,, \\
&\qquad= \sum_{i=1}^{p^\star}\tilde b_{\omega_i}\tilde K(x_j^\star,x_i^\star) + \sum_{i=1}^{p^\star}\tilde K_2(x_j^\star,x_i^\star)^T\tilde b_{x_i^\star} \\
&\qquad\quad+ \sum_{i=1}^{p^\star}\int_{\mathcal{X}_i^{\mathrm{near}}(r)}\int_0^1 (1-y)\dot\gamma_{x_i^\star\to x'}(y)^T H_2^{\mathfrak{g}} K_{\mathrm{norm}}(x_j^\star,\gamma_{x_i^\star\to x'}(y))\dot\gamma_{x_i^\star\to x'}(y)\,\mathrm{d}y\,\mathrm{d}(\mu-\mu_\omega^\star)(x') \\
&\qquad\quad+ \int_0^1 \dot\gamma_{x_j^\star\to x}(y)^T \int \nabla_1 K_{\mathrm{norm}}(\gamma_{x_j^\star\to x}(y),x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x')\,\mathrm{d}y + \int_{\mathcal{X}^{\mathrm{far}}(r)} K_{\mathrm{norm}}(x_j^\star,x')\,\mathrm{d}\mu(x')\,, \\
&\qquad=: \sum_{i=1}^{p^\star}\tilde b_{\omega_i}\tilde K(x_j^\star,x_i^\star) + \sum_{i=1}^{p^\star}\tilde K_2(x_j^\star,x_i^\star)^T\tilde b_{x_i} + E_2(x)
\end{aligned}
$$

where

$$
|E_2(x)| \le \frac{B_{02}}{2}\sum_{i=1}^{p^\star}\int_{\mathcal{X}_i^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_i^\star,x')^2\,\mathrm{d}\mu(x') + \mathfrak{d}_{\mathfrak{g}}(x_j^\star,x)\sqrt{B_{11}}\,\|\Psi(\mu-\mu_\omega^\star)\|_{\mathbb{L}} + \mu(\mathcal{X}^{\mathrm{far}}(r))\,.
$$

Note that we used $\dot\gamma_{x_i^\star\to x_i^\star}=0$ to get

$$
\begin{aligned}
&\left|\sum_{i=1}^{p^\star}\int_{\mathcal{X}_i^{\mathrm{near}}(r)}\int_0^1 (1-y)\dot\gamma_{x_i^\star\to x'}(y)^T H_2^{\mathfrak{g}} K_{\mathrm{norm}}(x,\gamma_{x_i^\star\to x'}(y))\dot\gamma_{x_i^\star\to x'}(y)\,\mathrm{d}y\,\mathrm{d}(\mu-\mu_\omega^\star)(x')\right| \\
&= \left|\sum_{i=1}^{p^\star}\int_{\mathcal{X}_i^{\mathrm{near}}(r)}\int_0^1 (1-y)\dot\gamma_{x_i^\star\to x'}(y)^T H_2^{\mathfrak{g}} K_{\mathrm{norm}}(x,\gamma_{x_i^\star\to x'}(y))\dot\gamma_{x_i^\star\to x'}(y)\,\mathrm{d}y\,\mathrm{d}\mu(x')\right| \\
&\le \frac{B_{02}}{2}\sum_{i=1}^{p^\star}\int_{\mathcal{X}_i^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_i^\star,x')^2\,\mathrm{d}\mu(x')\,.
\end{aligned}
$$

So according to (38),

$$
\begin{aligned}
\int_{X^{\mathrm{near}}(r)} |E_2(x)|^2\,\mathrm{d}\mu(x) &\lesssim_{B_{\mathrm{TV}},d} \left(\sum_i\int_{\mathcal{X}_i^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_i^\star,x')^2\,\mathrm{d}\mu(x')\right)^2 + \sum_j\int_{\mathcal{X}_i^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_j^\star,x)^2\,\mathrm{d}\mu(x)\,\|\Psi(\mu-\mu_\omega^\star)\|_{\mathbb{L}}^2 \\
&\qquad+ \mu(\mathcal{X}^{\mathrm{far}}(r))^2\,, \\
&\lesssim_{B_{\mathrm{TV}},d,\varepsilon_0',\varepsilon_2} \frac{1}{\kappa^2 r^4}(J_W(\mu)-J_W(\mu_\omega^\star))^2\,.
\end{aligned}
$$

Conclusion for $\int_{\mathcal{X}^{\mathrm{near}}(r)} |J'_\mu(x)|^2\,\mathrm{d}\mu(x)$: Using that $(a+b)^2\ge\frac12 a^2-b^2$ and collecting the error terms $\int_{X^{\mathrm{near}}(r)}|E_1(x)|^2\,\mathrm{d}\mu(x)$ and $\int_{X^{\mathrm{near}}(r)}|E_2(x)|^2\,\mathrm{d}\mu(x)$ into b, we obtain

$$
\begin{aligned}
&\frac12\sum_{j=1}^{p^\star}\mu(\mathcal{X}_j^{\mathrm{near}}(r))\left(\sum_{i=1}^{p^\star}\tilde b_{\omega_i}\tilde K(x_j^\star,x_i^\star) + \sum_{i=1}^{p^\star}\tilde K_2(x_j^\star,x_i^\star)^T\tilde b_{x_i}\right)^2 - \int_{\mathcal{X}^{\mathrm{near}}(r)}|J'_\mu(x)|^2\,\mathrm{d}\mu(x) \\
&\qquad\lesssim_{d,B_{\mathrm{TV}},\|y\|_{\mathbb{L}},\|\mu_\omega^\star\|_{\mathrm{TV}},\varepsilon_2,\varepsilon_0'} \frac{1}{\kappa^2 r^4}(J_W(\mu)-J_W(\mu_\omega^\star))^2 + r^2\sum_j\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_j^\star,x)^2\,\mathrm{d}\mu(x)\,.
\end{aligned}
\tag{41}
$$

D.2.2 Lower bound on $\int_{\mathcal{X}^{\mathrm{far}}(r)} |J'_\mu|^2 \,\mathrm{d}\mu$

According to Lemma D.3 and recalling that $J_{\mu^\star_\omega} = \kappa(1-\eta^\star)$, we have $|J'_{\mu^\star_\omega}(x)| \geq \kappa\varepsilon'_0 r^2$ for $x \in \mathcal{X}^{\mathrm{far}}(r)$. Hence

$$|J'_\mu(x)| = |J'_{\mu^\star_\omega}(x) + \langle \Psi\delta_x, \Psi(\mu-\mu^\star_\omega)\rangle_{\mathbb{L}}| \geq \kappa\varepsilon'_0 r^2 - \|\Psi(\mu-\mu^\star_\omega)\|_{\mathbb{L}} \,.$$

Using again the inequality $(a+b)^2 \geq \frac{1}{2}a^2 - b^2$ with (38),

$$\begin{aligned}\int_{\mathcal{X}^{\mathrm{far}}(r)} |J'_\mu(x)|^2 \,\mathrm{d}\mu(x) &\geq \left(\frac{1}{2}\kappa^2\varepsilon_0'^2 r^4 - \|\Psi(\mu-\mu^\star_\omega)\|^2_{\mathbb{L}}\right)\mu(\mathcal{X}^{\mathrm{far}}(r))\,,\\ &\geq \frac{1}{2}\kappa^2\varepsilon_0'^2 r^4 \mu(\mathcal{X}^{\mathrm{far}}(r)) - \frac{1}{\kappa^2\varepsilon_0'^2 r^2}(J_W(\mu) - J_W(\mu^\star_\omega))^2\,,\end{aligned}$$

leading to

$$\frac{1}{2}\kappa^2\varepsilon_0'^2 r^4 \mu(\mathcal{X}^{\mathrm{far}}(r)) - \int_{\mathcal{X}^{\mathrm{far}}(r)} |J'_\mu(x)|^2 \,\mathrm{d}\mu(x) \lesssim_{\varepsilon'_0} \frac{1}{\kappa^2 r^2}(J_W(\mu) - J_W(\mu^\star_\omega))^2 \,. \tag{42}$$

D.2.3 Lower bound on $\int_{\mathcal{X}^{\mathrm{near}}(r)} \left\|\nabla^{\mathfrak{g}} J'_\mu(x)\right\|_x^2 \,\mathrm{d}\mu(x)$

Dealing with Riemannian gradients: For $x \in \mathbb{R}^d \times [u_{\min}, +\infty)^d$ such that $\mathfrak{d}(x^\star_j, x) \leq r$, Lemma D.2 gives

$$\left\|v\mathfrak{g}_{x^\star_j}^{-1/2}\right\|_2 = \left\|v\mathfrak{g}_x^{-1/2}\mathfrak{g}_x^{1/2}\mathfrak{g}_{x^\star_j}^{-1/2}\right\|_2 \leq \left\|v\mathfrak{g}_x^{-1}\right\|_x \left\|\mathfrak{g}_x^{1/2}\mathfrak{g}_{x^\star_j}^{-1/2}\right\|_2 \leq C_{\mathfrak{g}} \left\|v\mathfrak{g}_x^{-1}\right\|_x \quad \forall\, v \in \mathbb{R}^d \,.$$

So for $x \in \mathcal{X}^{\mathrm{near}}_j(r)$ $(j = 1, \dots, p^\star)$,

$$\left\|\nabla^{\mathfrak{g}} J'_\mu(x)\right\|_x = \left\|\mathfrak{g}_x^{-1}\nabla J'_\mu(x)\right\|_x \geq \frac{1}{C_{\mathfrak{g}}}\left\|\mathfrak{g}_{x^\star_j}^{-1/2}\nabla J'_\mu(x)\right\|_2 \,. \tag{43}$$

We use this inequality together with $\mathfrak{d}(x^\star_j, \gamma_{x^\star_j\to x}(y)) \leq \mathfrak{d}(x^\star_j, x) \leq r$ for $y \in [0,1]$ (see Appendix A.1).

Decomposition of $\mathfrak{g}_{x^\star_j}^{-1/2}\nabla J'_\mu(x)$: Let $x \in \mathcal{X}^{\mathrm{near}}_j(r)$. As $\nabla^2 J'_{\mu^\star_\omega}(x^\star_j) = H^{\mathfrak{g}} J'_{\mu^\star_\omega}(x^\star_j)$ (because $\nabla J'_{\mu^\star_\omega}(x^\star_j) = 0$), using (39) we get the decomposition

$$\begin{aligned}\mathfrak{g}_{x^\star_j}^{-1/2}\nabla J'_\mu(x) &= \mathfrak{g}_{x^\star_j}^{-1/2}\nabla^2 J'_{\mu^\star_\omega}(x^\star_j)\dot\gamma_{x^\star_j\to x}(0)\\ &\quad + \underbrace{\mathfrak{g}_{x^\star_j}^{-1/2}\int \nabla_1 K_{\mathrm{norm}}(x^\star_j, x')\,\mathrm{d}(\mu-\mu^\star_\omega)(x')}_{A(x)}\\ &\quad + \underbrace{\mathfrak{g}_{x^\star_j}^{-1/2}\int\int_0^1 \nabla_1^2 K_{\mathrm{norm}}(\gamma_{x^\star_j\to x}(y), x')\dot\gamma_{x^\star_j\to x}(y)\,\mathrm{d}(\mu-\mu^\star_\omega)(x')}_{E_1(x)} + E_2(x)\,,\\ &= \mathfrak{g}_{x^\star_j}^{-1/2} H^{\mathfrak{g}} J'_{\mu^\star_\omega}(x^\star_j)\dot\gamma_{x^\star_j\to x}(0) + A(x) + E_1(x) + E_2(x)\end{aligned} \tag{44}$$

where

$$E_2(x) \coloneqq \begin{pmatrix} \mathfrak{g}_{t^\star_{j,1}t^\star_{j,1}}^{-1/2}\int_0^1 (1-y)\dot\gamma_{x^\star_j\to x}(y)^T H^{\mathfrak{g}}\partial_{t_1} J'_{\mu^\star_\omega}(\gamma_{x^\star_j\to x}(y))\dot\gamma_{x^\star_j\to x}(y)\,\mathrm{d}y \\ \vdots \\ \mathfrak{g}_{t^\star_{j,d}t^\star_{j,d}}^{-1/2}\int_0^1 (1-y)\dot\gamma_{x^\star_j\to x}(y)^T H^{\mathfrak{g}}\partial_{t_d} J'_{\mu^\star_\omega}(\gamma_{x^\star_j\to x}(y))\dot\gamma_{x^\star_j\to x}(y) \\ \mathfrak{g}_{u^\star_{j,1}u^\star_{j,1}}^{-1/2}\int_0^1 (1-y)\dot\gamma_{x^\star_j\to x}(y)^T H^{\mathfrak{g}}\partial_{u_1} J'_{\mu^\star_\omega}(\gamma_{x^\star_j\to x}(y))\dot\gamma_{x^\star_j\to x}(y)\,\mathrm{d}y \\ \vdots \\ \mathfrak{g}_{u^\star_{j,d}u^\star_{j,d}}^{-1/2}\int_0^1 (1-y)\dot\gamma_{x^\star_j\to x}(y)^T H^{\mathfrak{g}}\partial_{u_d} J'_{\mu^\star_\omega}(\gamma_{x^\star_j\to x}(y))\dot\gamma_{x^\star_j\to x}(y) \end{pmatrix}.$$

E_1 and E_2 are negligible. In fact,

$$\begin{aligned}\|E_1(x)\|_2 &\leq \sup_y \left\|\mathfrak{g}_{x^\star_j}^{-1/2}\mathfrak{g}_{\gamma_{x^\star_j\to x}(y)}^{1/2}\right\|_2 \left\|\mathfrak{g}_{\gamma_{x^\star_j\to x}(y)}^{-1/2}\left\langle\nabla^2\Psi\delta_{\gamma_{x^\star_j\to x}(y)}, \Psi(\mu-\mu^\star_\omega)\right\rangle \mathfrak{g}_{\gamma_{x^\star_j\to x}(y)}^{-1/2}\right\|_2 \left\|\dot\gamma_{x^\star_j\to x}(y)^T\mathfrak{g}_{\gamma_{x^\star_j\to x}(y)}^{1/2}\right\|_2 ,\\ &\leq C_{\mathfrak{g}}\mathfrak{d}_{\mathfrak{g}}(x, x^\star_j) B^e_{02}\|\Psi(\mu-\mu^\star_\omega)\|_{\mathbb{L}} \,.\end{aligned}$$

We also have

$$\begin{aligned}\|E_2(x)\|_2 &\le \frac{1}{2}\mathfrak{d}_{\mathfrak{g}}(x,x_j^\star)^2 \sup_y \left\|\mathfrak{g}^{1/2}_{\gamma_{x_j^\star\to x}(y)}\mathfrak{g}_{x_j^\star}^{-1/2}\right\|_2 \sqrt{2d} \\ &\qquad \times \sup_{x\in\mathbb{R}^d\times[u_{\min},+\infty)^d\,,\, b_k\in\{t_k,u_k\}} \left\|\mathfrak{g}_{b_kb_k}^{-1/2}\mathfrak{g}_x^{-1/2}H^{\mathfrak{g}}\partial_{b_k}J'_{\mu_\omega^\star}(x)\mathfrak{g}_x^{-1/2}\right\|_2 , \\ &\le \frac{1}{2}\sqrt{B_{33}^e}C_{\mathfrak{g}}(\|\mu_\omega^\star\|_{\mathrm{TV}}+\|y\|_{\mathbb{L}})\mathfrak{d}_{\mathfrak{g}}(x,x_j^\star)^2\,, \\ &\le \frac{1}{2}\sqrt{B_{33}^e}C_{\mathfrak{g}}(\|\mu_\omega^\star\|_{\mathrm{TV}}+\|y\|_{\mathbb{L}})\tfrac{1}{\sqrt{\varepsilon_2}}r\,\mathfrak{d}_{\mathfrak{g}}(x,x_j^\star)\,.\end{aligned}$$

Moreover,

$$\begin{aligned}A(x) &= \int \mathfrak{g}_{x_j^\star}^{-1/2}\nabla_1 K_{\mathrm{norm}}(x_j^\star,x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x')\,, \\ &= \sum_i \tilde{b}_{\omega_i}\tilde{K}_1(x_j^\star,x_i^\star)+\sum_i \tilde{K}_{12}(x_j^\star,x_i^\star)\tilde{b}_{x_i} \\ &\quad + \underbrace{\sum_i \int_{X_i^{\mathrm{near}}(r)}\int_0^1 (1-y)\mathfrak{g}_{x_j^\star}^{-1/2}H_2^{\mathfrak{g}}\nabla_1 K_{\mathrm{norm}}(x_j^\star,\gamma_{x_i^\star\to x'}(y))[\dot\gamma_{x_i^\star\to x'}(y),\dot\gamma_{x_i^\star\to x'}(y)]\,\mathrm{d}y\,\mathrm{d}(\mu-\mu_\omega^\star)(x')}_{E_3(x)} \\ &\quad + \underbrace{\int_{\mathcal{X}^{\mathrm{far}}(r)}\mathfrak{g}_{x_j^\star}^{-1/2}\nabla_1 K_{\mathrm{norm}}(x_j^\star,x')\,\mathrm{d}\mu(x')}_{E_4(x)}\,.\end{aligned}$$

E_3 and E_4 are negligible: $\|E_4(x)\|_2 \le \mu(\mathcal{X}^{\mathrm{far}}(r))B_{10}$ and using again that $\int_{X_i^{\mathrm{near}}(r)}\mathfrak{d}_{\mathfrak{g}}(x_i^\star,x')\,\mathrm{d}\mu_\omega^\star(x')=0$, we get

$$\|E_3(x)\|_2 \le \frac{B_{12}}{2}\sum_i\int_{X_i^{\mathrm{near}}(r)}\mathfrak{d}_{\mathfrak{g}}(x_i^\star,x')^2\,\mathrm{d}\mu(x')\,.$$

Gathering the error terms, we have

$$\begin{aligned}&\int_{\mathcal{X}^{\mathrm{near}}(r)}\|E_1(x)+E_2(x)+E_3(x)+E_4(x)\|_2^2\,\mathrm{d}\mu(x) \\ &\quad\lesssim_{d,B_{\mathrm{TV}},\|y\|_{\mathbb{L}},\|\mu_\omega^\star\|_{\mathrm{TV}}} \left(\sum_i\int_{X_i^{\mathrm{near}}(r)}\mathfrak{d}_{\mathfrak{g}}(x_i^\star,x')^2\,\mathrm{d}\mu(x')\right)\|\Psi(\mu-\mu_\omega^\star)\|^2 + r^2\sum_j\int_{X_j^{\mathrm{near}}(r)}\mathfrak{d}_{\mathfrak{g}}(x_j^\star,x')^2\,\mathrm{d}\mu(x') \\ &\qquad + \left(\sum_i\int_{X_i^{\mathrm{near}}(r)}\mathfrak{d}_{\mathfrak{g}}(x_i^\star,x')^2\,\mathrm{d}\mu(x')\right)^2 + \mu(\mathcal{X}^{\mathrm{far}}(r))^2\,, \\ &\quad\lesssim_{d,B_{\mathrm{TV}},\|y\|_{\mathbb{L}},\|\mu_\omega^\star\|_{\mathrm{TV}},\varepsilon_2,\varepsilon_0'} \frac{1}{\kappa^2r^4}(J_W(\mu)-J_W(\mu_\omega^\star))^2 + r^2\sum_j\int_{\mathcal{X}_j^{\mathrm{near}}(r)}\mathfrak{d}_{\mathfrak{g}}(x_j^\star,x)^2\,\mathrm{d}\mu(x)\,, \qquad (45)\end{aligned}$$

where we used (38) to obtain the last inequality.
Assembling the bounds: From (43), (44) and (45), it follows that

$$\begin{aligned}&\frac{1}{2C_{\mathfrak{g}}^2}\sum_j\int_{\mathcal{X}_j^{\mathrm{near}}(r)}\left\|\tilde{H}_j\mathfrak{g}_{x_j^\star}^{1/2}\dot\gamma_{x_j^\star\to x}(0)+\sum_i\tilde{b}_{\omega_i}\tilde{K}_1(x_j^\star,x_i^\star)+\sum_i\tilde{K}_{12}(x_j^\star,x_i^\star)\tilde{b}_{x_i}\right\|_2^2\mathrm{d}\mu(x) \\ &\quad -\int_{\mathcal{X}^{\mathrm{near}}(r)}\left\|\nabla^{\mathfrak{g}}J'_\mu(x)\right\|_x^2\,\mathrm{d}\mu(x) \\ &\qquad\lesssim_{d,B_{\mathrm{TV}},\|y\|_{\mathbb{L}},\|\mu_\omega^\star\|_{\mathrm{TV}},\varepsilon_2,\varepsilon_0'} \frac{1}{\kappa^2r^4}(J_W(\mu)-J_W(\mu_\omega^\star))^2 + r^2\sum_j\int_{\mathcal{X}_j^{\mathrm{near}}(r)}\mathfrak{d}_{\mathfrak{g}}(x_j^\star,x)^2\,\mathrm{d}\mu(x)\,.\end{aligned}$$

Finally, remark that for some v not depending on x, a bias–variance decomposition gives, for j such that

$\mu(\mathcal{X}_j^{\mathrm{near}}(r)) > 0$,

$$\begin{aligned}
&\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \tilde{H}_j \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) + v \right\|_2^2 \mathrm{d}\mu(x) \\
&\qquad = \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \tilde{H}_j \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} + v \right\|_2^2 \mathrm{d}\mu(x) + \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \tilde{H}_j \left(\mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right) \right\|_2^2 \mathrm{d}\mu(x) \,, \\
&\qquad = \mu(\mathcal{X}_j^{\mathrm{near}}(r)) \left\| \tilde{H}_j \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} + v \right\|_2^2 + \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \tilde{H}_j \left(\mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right) \right\|_2^2 \mathrm{d}\mu(x) \,.
\end{aligned}$$

As $\tilde{H}_j$ is a symmetric matrix of positive definite type, with eigenvalues greater than $\kappa\varepsilon_2$ (by Assumption 1), it comes that

$$\begin{aligned}
&\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \tilde{H}_j \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) + v \right\|_2^2 \mathrm{d}\mu(x) \\
&\qquad \geq \mu(\mathcal{X}_j^{\mathrm{near}}(r)) \left\| \tilde{H}_j \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} + v \right\|_2^2 + \kappa^2 \varepsilon_2^2 \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) \,.
\end{aligned}$$

So applying this inequality with $v = \sum_i \tilde{b}_{\omega_i} \tilde{K}_1(x_j^\star, x_i^\star) + \sum_i \tilde{K}_{12}(x_j^\star, x_i^\star) \tilde{b}_{x_i}$, we get

$$\begin{aligned}
&\frac{1}{2C_{\mathfrak{g}}^2} \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r)) > 0} \mu(\mathcal{X}_j^{\mathrm{near}}(r)) \left\| \tilde{H}_j \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} + \sum_i \tilde{b}_{\omega_i} \tilde{K}_1(x_j^\star, x_i^\star) + \sum_i \tilde{K}_{12}(x_j^\star, x_i^\star) \tilde{b}_{x_i} \right\|_2^2 \\
&\quad + \frac{\kappa^2 \varepsilon_2^2}{2C_{\mathfrak{g}}^2} \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r)) > 0} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{-1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) - \int_{\mathcal{X}^{\mathrm{near}}(r)} \left\| \nabla^{\mathfrak{g}} J'_\mu(x) \right\|_x^2 \mathrm{d}\mu(x) \\
&\quad \lesssim_{d, B_{\mathrm{TV}}, \|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}, \varepsilon_2, \varepsilon_0'} \frac{1}{\kappa^2 r^4} (J_W(\mu) - J_W(\mu_\omega^\star))^2 + r^2 \sum_j \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_j^\star, x)^2 \, \mathrm{d}\mu(x) \,.
\end{aligned} \tag{46}$$

D.2.4 Conclusion for the lower bound, choice of r

Remark that

$$\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_j^\star, x)^2 \, \mathrm{d}\mu(x) = \mathbb{1}_{\mu(\mathcal{X}_j^{\mathrm{near}}(r)) > 0} \left(\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) + \frac{\left\| \tilde{b}_{x_j} \right\|_2^2}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right) . \tag{47}$$

We now have all the ingredients to control the objective decrease G_μ (37). Assembling the bounds (41), (42), (46) and using (47), we get

$$
\begin{aligned}
G_\mu \gtrsim_{\varepsilon_0',\varepsilon_2} & \ \kappa^2 r^4 \mu(\mathcal{X}^{\mathrm{far}}(r)) + \sum_{j=1}^{p^\star} \mu(\mathcal{X}_j^{\mathrm{near}}(r)) \left(\sum_{i=1}^{p^\star} \tilde{b}_{\omega_i} \tilde{K}(x_j^\star, x_i^\star) + \sum_{i=1}^{p^\star} \tilde{K}_2(x_j^\star, x_i^\star)^T \tilde{b}_{x_i} \right)^2 \\
& + \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \mu(\mathcal{X}_j^{\mathrm{near}}(r)) \left\| \tilde{H}_j \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} + \sum_i \tilde{b}_{\omega_i} \tilde{K}_1(x_j^\star, x_i^\star) + \sum_i \tilde{K}_{12}(x_j^\star, x_i^\star) \tilde{b}_{x_i} \right\|_2^2 \\
& + \kappa^2 \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{-1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) \\
& - C_{lower} \frac{1}{\kappa^2 r^4} (J_W(\mu) - J_W(\mu_\omega^\star))^2 \\
& - C_{lower} r^2 \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \left(\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) + \frac{\left\| \tilde{b}_{x_j} \right\|_2^2}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right), \\
\gtrsim_{\varepsilon_0',\varepsilon_2} & \ \kappa^2 r^4 \mu(\mathcal{X}^{\mathrm{far}}(r)) + \min_j \mu(\mathcal{X}_j^{\mathrm{near}}(r)) \left\| \left(\tilde{\Upsilon} + \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \frac{1}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \tilde{H}_j)_{j=1,\dots,p^\star} \end{pmatrix} \right) \begin{pmatrix} \tilde{b}_\omega \\ \tilde{b}_x \end{pmatrix} \right\|_2^2 \\
& + \kappa^2 \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{-1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) \\
& - C_{lower} \frac{1}{\kappa^2 r^4} (J_W(\mu) - J_W(\mu_\omega^\star))^2 \\
& - C_{lower} r^2 \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \left(\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) + \frac{\left\| \tilde{b}_{x_j} \right\|_2^2}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right),
\end{aligned}
$$

with $C_{lower} > 0$ only depending on $d, B_{\mathrm{TV}}, \|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}, \varepsilon_2, \varepsilon_0'$, and where the matrix $\tilde{\Upsilon}$ has been introduced in (9).

Now, remark that $\tilde{\Upsilon}$ and $\begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \frac{1}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \tilde{H}_j)_{j=1,\dots,p} \end{pmatrix}$ are symmetric positive matrices. Moreover, $\tilde{\Upsilon} \succeq C_{\tilde{\Upsilon}} I$. Remark also that for $j = 1, \dots, p^\star$, as $\omega_j^\star \geq \omega_\star$ (*cf.* Assumption 1), $\min_j \mu(X_j^{\mathrm{near}}(r)) \geq \omega_\star - \left\| \tilde{b}_\omega \right\|_\infty \geq \omega_\star - \left\| \tilde{b}_\omega \right\|_2$. So

$$
\begin{aligned}
G_\mu \gtrsim_{\varepsilon_0',\varepsilon_2,C_{\tilde{\Upsilon}},\omega_\star} & \ \kappa^2 r^4 \mu(\mathcal{X}^{\mathrm{far}}(r)) + \tilde{b}^T \left(\tilde{\Upsilon} + \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \frac{1}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \tilde{H}_j)_{j=1,\dots,p^\star} \end{pmatrix} \right) \tilde{b} \\
& + \kappa^2 \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{-1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) \\
& - C'_{lower} \left\| \tilde{b}_\omega \right\|_2 \tilde{b}^T \left(\tilde{\Upsilon} + \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \frac{1}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \tilde{H}_j)_{j=1,\dots,p^\star} \end{pmatrix} \right) \tilde{b} \\
& - C'_{lower} \frac{1}{\kappa^2 r^4} (J_W(\mu) - J_W(\mu_\omega^\star))^2 \\
& - C'_{lower} r^2 \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \left(\int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x) + \frac{\left\| \tilde{b}_{x_j} \right\|_2^2}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right).
\end{aligned}
$$

with $C'_{lower} > 0$ only depending on $d, B_{\mathrm{TV}}, \|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}, \varepsilon_2, \varepsilon_0', C_{\tilde{\Upsilon}}, \omega_\star$.

Choice of r According to Lemma D.3, $\tilde{H}_j$ has eigenvalues greater than $\kappa\varepsilon_2$, leading to the inequality

$$\tilde{b}_x^T \operatorname{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\text{near}}(r))>0}\frac{1}{\mu(\mathcal{X}_j^{\text{near}}(r))}\tilde{H}_j)_{j=1,\dots,p^\star}\tilde{b}_x \geq \kappa\varepsilon_2 \sum_{j,\mu(\mathcal{X}_j^{\text{near}}(r))>0} \frac{\left\|\tilde{b}_{x_j}\right\|_2^2}{\mu(\mathcal{X}_j^{\text{near}}(r))}\,. \tag{48}$$

Setting $C_r = \sqrt{\frac{1}{2C'_{lower}}\min\{\varepsilon_2\,,\,1\}}$ and $r = \min\{C_r\kappa, r'_{\max}\}$, we obtain

$$\begin{aligned}
G_\mu \gtrsim_{\varepsilon'_0,\varepsilon_2,C_{\tilde{\Upsilon}},\omega_\star} \kappa^2 r^4 \mu(\mathcal{X}^{\text{far}}(r)) &+ \tilde{b}^T\left(\tilde{\Upsilon} + \begin{pmatrix} 0 & 0 \\ 0 & \operatorname{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\text{near}}(r))>0}\frac{1}{\mu(\mathcal{X}_j^{\text{near}}(r))}\tilde{H}_j)_{j=1,\dots,p^\star}\end{pmatrix}\right)\tilde{b} \\
&+ \kappa^2 \sum_{j,\mu(\mathcal{X}_j^{\text{near}}(r))>0}\int_{\mathcal{X}_j^{\text{near}}(r)}\left\|\mathfrak{g}_{x_j^\star}^{-1/2}\dot{\gamma}_{x_j^\star\to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\text{near}}(r))}\right\|_2^2 \mathrm{d}\mu(x) \\
&- C''_{lower}\left\|\tilde{b}_\omega\right\|_2\tilde{b}^T\left(\tilde{\Upsilon} + \begin{pmatrix} 0 & 0 \\ 0 & \operatorname{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\text{near}}(r))>0}\frac{1}{\mu(\mathcal{X}_j^{\text{near}}(r))}\tilde{H}_j)_{j=1,\dots,p^\star}\end{pmatrix}\right)\tilde{b} \\
&- C''_{lower}\frac{1}{\kappa^2 r^4}(J_W(\mu) - J_W(\mu_\omega^\star))^2\,.
\end{aligned} \tag{49}$$

with $C''_{lower} > 0$ only depending on $d, B_{\text{TV}}, \|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\text{TV}}, \varepsilon_2, \varepsilon'_0, C_{\tilde{\Upsilon}}, \omega_\star$. Moreover, the constant C_r depends only on $d, B_{\text{TV}}, \|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\text{TV}}, \varepsilon_2, \varepsilon'_0, C_{\tilde{\Upsilon}}, \omega_\star$.

D.3 Upper bound on the optimality gap

Let $\mu \in \mathcal{M}^+(\mathcal{X})$ such that $\|\mu\|_{\text{TV}} \leq B_{\text{TV}}$. We aim to upper bound $J_W(\mu) - J_W(\mu_\omega^\star)$. Recall that (see (5))

$$J_W(\mu) - J_W(\mu_\omega^\star) = \underbrace{\int_{\mathcal{X}} J'_{\mu_\omega^\star}\,\mathrm{d}\mu}_{A} + \underbrace{\frac{1}{2}\left\|\Psi(\mu - \mu_\omega^\star)\right\|_{\mathbb{L}}^2}_{B}\,. \tag{50}$$

Control of A: Remark that $J_{\mu_\omega^\star} \leq \kappa + \|\mu_\omega^\star\|_{\text{TV}} + \|y\|_{\mathbb{L}}$ (Lemma A.4). As $J'_{\mu_\omega^\star}(x_j^\star) = 0$ and $\nabla J'_{\mu_\omega^\star}(x_j^\star) = 0$ (see (40)), using a Taylor expansion we have

$$\begin{aligned}
A &= \sum_{j=1}^{p^\star}\int_{\mathcal{X}_j^{\text{near}}(r)} J'_{\mu_\omega^\star}(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{X}^{\text{far}}(r)} J'_{\mu_\omega^\star}(x)\,\mathrm{d}\mu(x)\,, \qquad (51)\\
&= \sum_{j=1}^{p^\star}\int_{\mathcal{X}_j^{\text{near}}(r)}\int_0^1 (1-y)\dot{\gamma}_{x_j^\star\to x}(y)^T\left\langle H^{\mathfrak{g}}\Psi\delta_{\gamma_{x_j^\star\to x}(y)}, \Psi\mu_\omega^\star - y\right\rangle_{\mathbb{L}}\dot{\gamma}_{x_j^\star\to x}(y)\,\mathrm{d}y\,\mathrm{d}\mu(x) + \int_{\mathcal{X}^{\text{far}}(r)} J'_{\mu_\omega^\star}(x)\,\mathrm{d}\mu(x)\,, \\
&\leq \frac{1}{2}\sqrt{B_{22}}(\|y\|_{\mathbb{L}} + \|\mu_\omega^\star\|_{\text{TV}})\sum_{j=1}^{p^\star}\int_{X_j^{\text{near}}(r)}\mathfrak{d}_{\mathfrak{g}}(x_j^\star, x)^2\,\mathrm{d}\mu(x) + (\kappa + \|\mu_\omega^\star\|_{\text{TV}} + \|y\|_{\mathbb{L}})\mu(\mathcal{X}^{\text{far}}(r))\,, \\
&\lesssim_{d,\varepsilon_2,\|y\|_{\mathbb{L}},\|\mu_\omega^\star\|_{\text{TV}}} \sum_{j,\mu(\mathcal{X}_j^{\text{near}}(r))>0}\left(\frac{1}{\kappa}\frac{1}{\mu(\mathcal{X}_j^{\text{near}}(r))}\tilde{b}_{x_j}^T\tilde{H}_j\tilde{b}_{x_j} + \int_{\mathcal{X}_j^{\text{near}}(r)}\left\|\mathfrak{g}_x^{1/2}\dot{\gamma}_{x_j^\star\to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\text{near}}(r))}\right\|_2^2\mathrm{d}\mu(x)\right) \\
&\quad + \mu(\mathcal{X}^{\text{far}}(r))\,,
\end{aligned}$$

where we used Equations (47) and (48) to get the last inequality.
Control of B: We also have

$$
\begin{aligned}
B &= \frac{1}{2}\left\langle \Psi\mu, \Psi(\mu-\mu_\omega^\star)\right\rangle_{\mathbb{L}} - \frac{1}{2}\left\langle \Psi\mu_\omega^\star, \Psi(\mu-\mu_\omega^\star)\right\rangle_{\mathbb{L}}\,,\\
&= \frac{1}{2}\sum_j \left(\int_{\mathcal{X}} \tilde{b}_{\omega_j} K_{\mathrm{norm}}(x_j^\star, x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x') + \int_{\mathcal{X}} \nabla_1 K_{\mathrm{norm}}(x_j^\star, x')^T \tilde{b}_{x_j}\,\mathrm{d}(\mu-\mu_\omega^\star)(x')\right)\\
&\quad \underbrace{+ \frac{1}{2}\sum_j \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \int_{\mathcal{X}} \int_0^1 (1-y)\dot{\gamma}_{x_j^\star\to x}(y)^T H_1^{\mathfrak{g}} K_{\mathrm{norm}}(\gamma_{x_j^\star\to x}(y), x')\dot{\gamma}_{x_j^\star\to x}(y)\,\mathrm{d}y\,\mathrm{d}(\mu-\mu_\omega^\star)(x')\,\mathrm{d}\mu(x)}_{E_1}\\
&\quad \underbrace{+ \frac{1}{2}\int_{\mathcal{X}^{\mathrm{far}}(r)} \int_{\mathcal{X}} K_{\mathrm{norm}}(x, x')\,\mathrm{d}(\mu-\mu_\omega^\star)(x')\,\mathrm{d}\mu(x)}_{E_2}\,,\\
&= \frac{1}{2}\sum_j \tilde{b}_{\omega_j} \sum_i \left(\tilde{b}_{\omega_i}\tilde{K}(x_j^\star, x_i^\star) + \nabla_2\tilde{K}(x_j^\star, x_i^\star)^T\tilde{b}_{x_i}\right)\\
&\quad + \frac{1}{2}\sum_j\sum_i \left(\tilde{b}_{\omega_i}\nabla_1\tilde{K}(x_j^\star, x_i^\star)^T\tilde{b}_{x_j} + \tilde{b}_{x_i}^T\nabla_2\nabla_1\tilde{K}(x_j^\star, x_i^\star)^T\tilde{b}_{x_j}\right)\\
&\quad \underbrace{+ \frac{1}{2}\sum_j\sum_i \tilde{b}_{\omega_j}\int_{\mathcal{X}_i^{\mathrm{near}}(r)} \int_0^1 (1-y)\dot{\gamma}_{x_i^\star\to x}(y)^T H_2^{\mathfrak{g}} K_{\mathrm{norm}}(x_j^\star, \gamma_{x_i^\star\to x'}(y))\dot{\gamma}_{x_i^\star\to x}(y)\,\mathrm{d}y\,\mathrm{d}\mu(x')}_{E_3}\\
&\quad \underbrace{+ \frac{1}{2}\sum_j \int_{\mathcal{X}^{\mathrm{far}}(r)} \tilde{b}_{\omega_j} K_{\mathrm{norm}}(x_j^\star, x')\,\mathrm{d}\mu(x')}_{E_4}\\
&\quad \underbrace{+ \frac{1}{2}\sum_j\sum_i \int_{\mathcal{X}_i^{\mathrm{near}}(r)} \int_0^1 (1-y)[\tilde{b}_{x_j}]K_{\mathrm{norm}}^{(12)}(x_j^\star, \gamma_{x_i^\star\to x'}(y))[\dot{\gamma}_{x_i^\star\to x}(y), \dot{\gamma}_{x_i^\star\to x}(y)]\,\mathrm{d}y\,\mathrm{d}\mu(x')}_{E_5}\\
&\quad \underbrace{+ \frac{1}{2}\sum_j \int_{\mathcal{X}^{\mathrm{far}}(r)} \nabla_1 K_{\mathrm{norm}}(x_j^\star, x')^T\tilde{b}_{x_j}\,\mathrm{d}\mu(x')}_{E_6}\\
&\quad + E_1 + E_2\,,\\
&= \frac{1}{2}\tilde{b}^T\tilde{\Upsilon}\tilde{b} + E_1 + E_2 + E_3 + E_4 + E_5 + E_6\,,
\end{aligned}
\tag{53}
$$

where $\tilde{\Upsilon}$ is defined by (9). Using controls from Lemma A.1 and (38), we get

$$
\begin{aligned}
\sum_{k=1}^6 |E_i| &\le \frac{1}{4}\sqrt{B_{22}}\,\|\Psi(\mu-\mu_\omega^\star)\|_{\mathbb{L}} \sum_j \int_{\mathcal{X}_j^{\mathrm{near}}} \mathfrak{d}_{\mathfrak{g}}(x_j^\star, x)^2\,\mathrm{d}\mu(x) + \frac{1}{2}\mu(\mathcal{X}^{\mathrm{far}}(r))\,\|\Psi(\mu-\mu_\omega^\star)\|_{\mathbb{L}}\\
&\quad + \frac{B_{02}}{4}\sum_j |\tilde{b}_{\omega_j}| \sum_i \int_{\mathcal{X}_i^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_i^\star, x')^2\,\mathrm{d}\mu(x') + \frac{1}{2}\mu(\mathcal{X}^{\mathrm{far}}(r))\sum_j |\tilde{b}_{\omega_j}|\\
&\quad + \frac{B_{12}}{4}\sum_j \left\|\tilde{b}_{x_j}\right\|_2 \sum_i \int_{\mathcal{X}_i^{\mathrm{near}}(r)} \mathfrak{d}_{\mathfrak{g}}(x_i^\star, x')^2\,\mathrm{d}\mu(x') + \frac{B_{10}}{2}\mu(\mathcal{X}^{\mathrm{far}}(r))\sum_j \left\|\tilde{b}_{x_j}\right\|_2\,,\\
&\lesssim_{d,\varepsilon_2,\varepsilon_0'} \frac{1}{\kappa r^2}(J_W(\mu) - J_W(\mu_\omega^\star))^{3/2} + \frac{1}{\kappa r^2}(J_W(\mu) - J_W(\mu_\omega^\star))\left(\sum_j |\tilde{b}_{\omega_j}| + \left\|\tilde{b}_{x_j}\right\|_2\right)\,,\\
&\lesssim_{d,\varepsilon_2,\varepsilon_0'} \frac{1}{\kappa r^2}(J_W(\mu) - J_W(\mu_\omega^\star))^{3/2} + \frac{1}{\kappa r^2}p^\star(J_W(\mu) - J_W(\mu_\omega^\star))\left\|\tilde{b}\right\|_2\,.
\end{aligned}
\tag{54}
$$

As $\tilde{\Upsilon} \succeq C_{\tilde{\Upsilon}} I$, we have $\left\|\tilde{b}\right\|_2^2 \le \frac{1}{C_{\tilde{\Upsilon}}}\tilde{b}^T\tilde{\Upsilon}\tilde{b}$. So

$$
\begin{aligned}
B &\lesssim_{d,\varepsilon_2,\varepsilon_0',p^\star} \tilde{b}^T\tilde{\Upsilon}\tilde{b} + \frac{1}{\kappa r^2}(J_W(\mu) - J_W(\mu_\omega^\star))^{3/2} + \frac{1}{\kappa^2 r^4}(J_W(\mu) - J_W(\mu_\omega^\star))^2 + \left\|\tilde{b}\right\|_2^2\,,\\
&\lesssim_{d,\varepsilon_2,\varepsilon_0',p^\star,C_{\tilde{\Upsilon}}} \tilde{b}^T\tilde{\Upsilon}\tilde{b} + \frac{1}{\kappa r^2}(J_W(\mu) - J_W(\mu_\omega^\star))^{3/2} + \frac{1}{\kappa^2 r^4}(J_W(\mu) - J_W(\mu_\omega^\star))^2\,.
\end{aligned}
\tag{55}
$$

Conclusion for the upper bound: Assembling (50), (52) and (55), we get

$$J_W(\mu) - J_W(\mu_\omega^\star) \lesssim_{\|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}, d, \varepsilon_2, \varepsilon_0', p^\star, C_{\tilde{\Upsilon}}} \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_x^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x)$$
$$+ \mu(\mathcal{X}^{\mathrm{far}}(r)) + \frac{1}{\kappa} \tilde{b}^T \left(\tilde{\Upsilon} + \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \frac{1}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \tilde{H}_j)_{j=1,\dots,p^\star} \end{pmatrix} \right) \tilde{b} \tag{56}$$
$$+ \frac{1}{\kappa r^2} (J_W(\mu) - J_W(\mu_\omega^\star))^{3/2} + \frac{1}{\kappa^2 r^4} (J_W(\mu) - J_W(\mu_\omega^\star))^2 .$$

Bound on $\left\| \tilde{b} \right\|_2$: We revisit the bound on $J_W(\mu) - J_W(\mu_\omega^\star)$. Remark that $A \geq 0$, because $\eta^\star \leq 1$ (implying that $J_{\mu_\omega^\star} \geq 0$). Recall also that $\tilde{b}^T \tilde{\Upsilon} \tilde{b} \geq C_{\tilde{\Upsilon}} \left\| \tilde{b} \right\|_2^2$. Hence, using (50), (53) and (54), we get

$$J_W(\mu) - J_W(\mu_\omega^\star) \geq \frac{1}{2} C_{\tilde{\Upsilon}} \left\| \tilde{b} \right\|_2^2 - \sum_{k=1}^{6} |E_i| ,$$
$$\geq \frac{1}{2} C_{\tilde{\Upsilon}} \left\| \tilde{b} \right\|_2^2 - c_{d, \varepsilon_2, \varepsilon_0', p^\star} \left(\frac{1}{\kappa r^2} (J_W(\mu) - J_W(\mu_\omega^\star))^{3/2} + \frac{1}{\kappa r^2} (J_W(\mu) - J_W(\mu_\omega^\star)) \left\| \tilde{b} \right\|_2 \right) ,$$

with c_a denoting a positive constant depending only on a. It comes that

$$\frac{1}{2} C_{\tilde{\Upsilon}} \left(\left\| \tilde{b} \right\|_2 - \frac{c_{d, \varepsilon_2, \varepsilon_0', p^\star}}{C_{\tilde{\Upsilon}} \kappa r^2} (J_W(\mu) - J_W(\mu_\omega^\star)) \right)^2 \leq J_W(\mu) - J_W(\mu_\omega^\star) + c_{d, \varepsilon_2, \varepsilon_0', p^\star} \frac{1}{\kappa r^2} (J_W(\mu) - J_W(\mu_\omega^\star))^{3/2}$$
$$+ \frac{1}{2} \frac{c_{d, \varepsilon_2, \varepsilon_0', p^\star}^2}{C_{\tilde{\Upsilon}} \kappa^2 r^4} (J_W(\mu) - J_W(\mu_\omega^\star))^2$$

If μ is such that $J_W(\mu) - J_W(\mu_\omega^\star) \leq \kappa^2 r^4$, we get

$$\left\| \tilde{b} \right\|_2^2 \lesssim_{d, \varepsilon_2, \varepsilon_0', p^\star, C_{\tilde{\Upsilon}}} J_W(\mu) - J_W(\mu_\omega^\star) . \tag{57}$$

D.4 Assembling the bounds

Choice of J_0': We recall Equations (49), (56) and (57). We deduce that there exists $0 < J_0'' \leq 1$ small enough, depending on $C_{lower}'', d, \varepsilon_2, \varepsilon_0', p^\star, C_{\tilde{\Upsilon}}, \|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}$, such that if $J_W(\mu) - J_W(\mu_\omega^\star) \leq J_0'' \kappa^2 r^4$, then it holds that

$$G_\mu \gtrsim_{\varepsilon_0', \varepsilon_2, C_{\tilde{\Upsilon}}, \omega_\star} \kappa^2 r^4 \mu(\mathcal{X}^{\mathrm{far}}(r)) + \tilde{b}^T \left(\tilde{\Upsilon} + \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \frac{1}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \tilde{H}_j)_{j=1,\dots,p^\star} \end{pmatrix} \right) \tilde{b}$$
$$+ \kappa^2 \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_{x_j^\star}^{-1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x)$$
$$- C_{lower}''' \frac{1}{\kappa^2 r^4} (J_W(\mu) - J_W(\mu_\omega^\star))^2$$

(where C_{lower}''' depends on $d, B_{\mathrm{TV}}, \|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}, \varepsilon_2, \varepsilon_0', C_{\tilde{\Upsilon}}, \omega_\star$) together with

$$J_W(\mu) - J_W(\mu_\omega^\star) \lesssim_{\|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}, d, \varepsilon_2, \varepsilon_0', p^\star, C_{\tilde{\Upsilon}}} \sum_{j, \mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \int_{\mathcal{X}_j^{\mathrm{near}}(r)} \left\| \mathfrak{g}_x^{1/2} \dot{\gamma}_{x_j^\star \to x}(0) - \frac{\tilde{b}_{x_j}}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \right\|_2^2 \mathrm{d}\mu(x)$$
$$+ \mu(\mathcal{X}^{\mathrm{far}}(r)) + \frac{1}{\kappa} \tilde{b}^T \left(\tilde{\Upsilon} + \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(\mathbb{1}_{\mu(\mathcal{X}_j^{\mathrm{near}}(r))>0} \frac{1}{\mu(\mathcal{X}_j^{\mathrm{near}}(r))} \tilde{H}_j)_{j=1,\dots,p^\star} \end{pmatrix} \right) \tilde{b} .$$

This implies the existence of $R_0'' > 0$ depending on $\|y\|_{\mathbb{L}}, \|\mu_\omega^\star\|_{\mathrm{TV}}, d, \varepsilon_2, \varepsilon_0', p^\star, C_{\tilde{\Upsilon}}, \omega_\star$ such that

$$G_\mu \geq R_0'' \kappa^2 r^4 (J_W(\mu) - J_W(\mu_\omega^\star)) - c_{\varepsilon_2, \varepsilon_0', C_{\tilde{\Upsilon}}, \omega_\star} C_{lower}''' \frac{1}{\kappa^2 r^4} (J_W(\mu) - J_W(\mu_\omega^\star))^2 .$$

Hence there exists $J_0' > 0$ depending on J_0'', C'''_{lower}, ε_0', ε_2, $C_{\tilde{\Upsilon}}$, $\omega_\star$, and $R_0' > 0$ depending on $\|y\|_{\mathbb{L}}$, $\|\mu_\omega^\star\|_{\mathrm{TV}}$, d, ε_2, ε_0', $p^\star$, $C_{\tilde{\Upsilon}}$, $\omega_\star$, such that

$$J_W(\mu) - J_W(\mu_\omega^\star) \leq J_0' \kappa^4 r^8 \quad \Longrightarrow \quad G_\mu \geq R_0' \kappa^2 r^4 (J_W(\mu) - J_W(\mu_\omega^\star)) \,.$$

Tracking dependence on parameters: We recall the definition of the constant factors, and the choice of r made in Appendix D.2: $r = \min\{C_r \kappa, r'_{\max}\}$ with $C_r > 0$ depending on d, $\|y\|_{\mathbb{L}}$, $\|\mu_\omega^\star\|_{\mathrm{TV}}$, ε_2, ε_0', $C_{\tilde{\Upsilon}}$, $\omega_\star$, $\mathcal{X}$, $\left\|\mu_\omega^1\right\|_{\mathrm{TV}}$, and $r'_{\max}$ depending on d, $\mathcal{X}$, $\tilde{\mathcal{X}}$, τ, $\mu_\omega^\star$, $\eta^\star$. The constant $J_0' > 0$ depends on $p^\star$, d, $\mathcal{X}$, $\left\|\mu_\omega^1\right\|_{\mathrm{TV}}$, $\|y\|_{\mathbb{L}}$, $\|\mu_\omega^\star\|_{\mathrm{TV}}$, ε_2, ε_0', $C_{\tilde{\Upsilon}}$, $\omega_\star$. We denote $R_0 \coloneqq R_0' \min\{r'_{\max}, C_r\}^4$. This constant depends on $\mathcal{X}$, $\tilde{\mathcal{X}}$, τ, $\mu_\omega^\star$, $\eta^\star$, $\|y\|_{\mathbb{L}}$, $\left\|\mu_\omega^1\right\|_{\mathrm{TV}}$. (Remark that $\eta^\star$, $\mu_\omega^\star$ contain the dependence on $C_{\tilde{\Upsilon}}$, ε_0', ε_2, $\omega_\star$, $p^\star$, $\|\mu_\omega^\star\|_{\mathrm{TV}}$, and that $\mathcal{X}$ contains the dependence on d.) We also define $J_0 \coloneqq J_0' \min\{r'_{\max}, C_r\}^8$. It also depends on $\mathcal{X}$, $\tilde{\mathcal{X}}$, τ, $\mu_\omega^\star$, $\eta^\star$, $\|y\|_{\mathbb{L}}$, $\left\|\mu_\omega^1\right\|_{\mathrm{TV}}$.

If μ satisfies $J_W(\mu) - J_W(\mu_\omega^\star) \leq J_0 \kappa^{12}$, then

$$G_\mu \geq R_0 \kappa^6 (J_W(\mu) - J_W(\mu_\omega^\star)) \,. \tag{58}$$

D.5 Exponential convergence

If μ_ω^1 verifies

$$J_W(\mu_\omega^1) - J_W(\mu_\omega^\star) \leq J_0 \kappa^{12} \,, \tag{59}$$

we retrieve the result of [Chizat, 2022, Corollary 3.5]. First note that $\left\|\mu_\omega^k\right\|_{\mathrm{TV}} \leq B_{\mathrm{TV}}$ for all $k \geq 1$ (Lemma A.3).

Then, as $k \in \mathbb{N}^* \mapsto J_W(\mu_\omega^k)$ is decreasing, it also holds that $J_W(\mu_\omega^k) - J_W(\mu_\omega^\star) \leq J_0 \kappa^{12}$ and we get

$$J_W(\mu_\omega^{k+1}) - J_W(\mu_\omega^k) \leq -\frac{1}{2} \min\{\alpha, \beta\} G_{\mu_\omega^k} \leq -\frac{1}{2} \min\{\alpha, \beta\} R_0 \kappa^6 (J_W(\mu_\omega^k) - J_W(\mu_\omega^\star)) \,,$$

where we used (58). Note that the contraction factor is automatically admissible: writing $c \coloneqq \frac{1}{2} \min\{\alpha, \beta\} R_0 \kappa^6$ and $g_k \coloneqq J_W(\mu_\omega^k) - J_W(\mu_\omega^\star)$, the above reads $g_{k+1} \leq (1-c) g_k$, while $g_{k+1} \geq 0$ since $\mu_\omega^\star$ minimizes J_W over $\mathcal{M}(\mathcal{X})^+$. Hence $c \leq 1$ as soon as $g_k > 0$, so that $1 - c \in [0, 1)$ and the geometric bound below is never vacuous. So for $k \geq 1$,

$$J_W(\mu_\omega^k) - J_W(\mu_\omega^\star) \leq \left(1 - \frac{1}{2} \min\{\alpha, \beta\} R_0 \kappa^6\right)^{k-1} (J_W(\mu_\omega^1) - J_W(\mu_\omega^\star)) \,. \tag{60}$$

Remark D.1 (Refining the dependence on κ). To obtain the lower bound for G_μ, we used several times the inequality $(a+b)^2 \geq \frac{1}{2} a^2 - b^2$. A parameterized version of this inequality holds: for $\delta \in (0,1)$, $(a+b)^2 \geq (1-\delta) a^2 - \frac{1-\delta}{\delta} b^2$. By optimizing δ where we use this inequality, we can improve both the condition on the initialization and the convergence rate. Specifically, this allows one to reduce the exponent associated with κ in Equations (59) and (60).
We do not present this approach here in order to keep the exposition simple, since it would require choosing δ based on results given later in the proof.

E Proof of Theorem 3.2

This proof relies on calculations established in a previous work [Giard et al., 2025]. More precisely, we import results from the following parts: Lemma K.1, Theorem 5.1, Theorem 6.1, Proposition K.1 and its proof, Lemma K.4, Proposition I.1 and its proof.

Non degeneracy of the statistical target We suppose that $y = \Psi \mu_\omega^0 + b$, where $b \in \mathbb{L}$ and $\mu_\omega^0 \coloneqq \sum_{j=1}^s \omega_j^0 \delta_{x_j^0}$, with $\omega_j^0 > 0$ and $x_j^0 \in \mathring{\tilde{\mathcal{X}}}$, and $s \geq 2$.

We then use [Giard et al., 2025, Lemma K.1]. If the minimal distance between the particles of μ_ω^0 is sufficiently large, *i.e.* if $\min_{i \neq j} d(x_j^0, x_i^0) \geq \Delta_\tau$ where Δ_τ depends on $d, u_{\min}, u_{\max}, \tau, s$ (see [Giard et al., 2025, Theorem 5.1] for the precise value of Δ_τ), then μ_ω^0 satisfies the NDSC.

We can apply [Giard et al., 2025, Theorem 6.1]: there exists $\kappa_0 > 0$ (depending on $\mathcal{X}, \tau, \mu_\omega^0$), $\gamma_0 > 0$ (depending on d) such that for all $0 < \kappa \leq \kappa_0$ and if $\|b\|_{\mathbb{L}} \leq \gamma_0 \kappa$, then the solution $\mu_\omega^\star$ is unique and s-sparse: $\mu_\omega^\star = \sum_{j=1}^s \omega_j^\star \delta_{x_j^\star}$ with $\omega_j^\star > 0$ (*i.e.* $p^\star = s$).

Non degeneracy of the solution Under these assumptions, the proof of [Giard et al., 2025, Proposition K.1] shows that $\eta^\star$ verifies items 1,2,3,4 of Assumption 1 with $\varepsilon_2 = 0.06158$.

Control of the solution We have $\omega_\star = \min_j \omega_j^\star > 0$ (*cf*. [Giard et al., 2025, Lemma K.4]). This minimum weight depends a priori on the problem ($\mathcal{P}_\kappa$). We can obtain finer results with [Giard et al., 2025, Theorem 6.1] It comes that for $1 \leq j \leq s$,

$$|\omega_j^0 - \omega_j^\star| \lesssim_{\mathcal{X},\mu^0,\tau} \kappa \quad \text{and} \quad \mathrm{d}(x_j^0, x_j^\star) \lesssim_{\mathcal{X},\mu^0,\tau} \kappa\,.$$

If κ is chosen sufficiently small, and if the noise is small w.r.t. κ, we have $\omega^\star = \min_j \omega_j^\star$ close to $\min_j \omega_j^0$,and $x_j^\star$ close to x_j^0 and in particular, $x_j^\star \in \mathring{\tilde{\mathcal{X}}}$. This provides a condition ensuring that the statement regarding the location of the solution in Assumption 1 is satisfied. More explicitly, there exists C'_κ depending on $\mathcal{X}, \tilde{\mathcal{X}}, \tau, \mu_\omega^0$ such that if $\kappa \leq \min\{\kappa_0, C'_\kappa\}$ and $\|b\|_{\mathbb{L}} \leq \gamma_0\kappa$, then $\min_j \omega_j^\star \geq \frac{1}{2}\min_j \omega_j^0$, and for all $j \in \{1,\dots,s\}$, $x_j^\star \in \mathring{\tilde{\mathcal{X}}}$.

Positive definiteness of $\tilde{\Upsilon}$ We know that

$$\tilde{\Upsilon}_0 \coloneqq \begin{pmatrix} (K_{\mathrm{norm}}(x_j^0, x_i^0))_{j,i=1,\dots,s} & (\nabla_2 K_{\mathrm{norm}}(x_j^0, x_i^0)^T \mathfrak{g}_{x_i^0}^{-1/2})_{j,i=1,\dots,s} \\ (\mathfrak{g}_{x_j^0}^{-1/2}\nabla_1 K_{\mathrm{norm}}(x_j^0, x_i^0))_{j,i=1,\dots,s} & (\mathfrak{g}_{x_j^0}^{-1/2}\nabla_1\nabla_2 K_{\mathrm{norm}}(x_j^0, x_i^0)\mathfrak{g}_{x_i^0}^{-1/2})_{j,i=1,\dots,s} \end{pmatrix}$$

is positive definite, from the separation assumptions made to obtain the NDSC. We refer to the proof of [Giard et al., 2025, Proposition I.1]: $\tilde{\Upsilon}_0 \succeq \frac{1}{2}I$.

As the minimal eigenvalue of $\tilde{\Upsilon}_0$ is continuous in the parameters $\{x_j^0\}_{j=1}^s$, and as $\tilde{\Upsilon}$ is the matrix built from the same expression at the points $\{x_j^\star\}_{j=1}^s$, it suffices for each $x_j^\star$ to be sufficiently close to x_j^0 for the minimal eigenvalue of $\tilde{\Upsilon}$ to remain above $\frac{1}{4}$. In particular, there exists $0 < C_\kappa \leq \min\{C'_\kappa, \kappa_0\}$ depending on $\mathcal{X}, \tilde{\mathcal{X}}, \tau, \mu_\omega^0$ such that if $0 < \kappa \leq C_\kappa$ and $\|b\|_{\mathbb{L}} \leq \gamma_0\kappa$, then $\tilde{\Upsilon} \succeq \frac{1}{4}I$.

F Proof of the bound on the optimality gap

We prove the result stated by (18) in the Discussion section (Section 5).

Let $\mu_\omega^\star$ be the minimizer of J_W on $\mathcal{M}(\mathcal{X})^+$. Let $\mu \in \mathcal{M}(\mathcal{X})^+$. We want to give an upper bound on $J_W(\mu) - J_W(\mu_\omega^\star)$ that does not depend on $\mu_\omega^\star$, but that is still more informative than the value of J_W at point μ.

Let $\alpha > 0$ such that $\inf_{x\in\mathcal{X}} \alpha(J'_\mu(x) - \kappa) \geq -\kappa$. Let $h = \alpha(\Psi\mu - y)$. As $\mu_\omega^\star$ is nonnegative, it holds that

$$-\langle h, \Psi\mu_\omega^\star\rangle - \kappa\,\|\mu_\omega^\star\|_{\mathrm{TV}} \leq 0\,. \tag{61}$$

Furthermore, as $\mu_\omega^\star$ is a minimizer, the KKT conditions hold and in particular $\int_{\mathcal{X}} J'_{\mu_\omega^\star}\,\mathrm{d}\mu_\omega^\star = 0$, which can be rewritten as

$$\langle \Psi\mu_\omega^\star, \Psi\mu_\omega^\star - y\rangle_{\mathbb{L}} + \kappa\,\|\mu_\omega^\star\|_{\mathrm{TV}} = 0\,.$$

Using (61), it follows that $\langle \Psi\mu_\omega^\star, \Psi\mu_\omega^\star - y\rangle_{\mathbb{L}} \leq \langle h, \Psi\mu_\omega^\star\rangle$, implying that $\|\Psi\mu_\omega^\star\|_{\mathbb{L}}^2 \leq \|h + y\|_{\mathbb{L}}^2$. Hence

$$\begin{aligned} J_W(\mu) - J_W(\mu_\omega^\star) &= J_W(\mu) + \frac{1}{2}\|\Psi\mu_\omega^\star\|_{\mathbb{L}}^2 - \frac{1}{2}\|y\|_{\mathbb{L}}^2 - \langle \Psi\mu_\omega^\star, \Psi\mu_\omega^\star - y\rangle_{\mathbb{L}} - \kappa\,\|\mu_\omega^\star\|_{\mathrm{TV}}\,, \\ &= J_W(\mu) + \frac{1}{2}\|\Psi\mu_\omega^\star\|_{\mathbb{L}}^2 - \frac{1}{2}\|y\|_{\mathbb{L}}^2\,, \\ &\leq J_W(\mu) + \frac{1}{2}\|h + y\|_{\mathbb{L}}^2 - \frac{1}{2}\|y\|_{\mathbb{L}}^2\,, \\ &= J_W(\mu) + \frac{1}{2}\|h\|_{\mathbb{L}}^2 + \langle h, y\rangle_{\mathbb{L}}\,. \end{aligned}$$

Since

$$J'_\mu(x) = \langle \Psi\mu - y, \Psi\delta_x\rangle_{\mathbb{L}} + \kappa \geq -\|\Psi\mu - y\|_{\mathbb{L}} + \kappa \quad \forall\, x \in \mathcal{X}\,,$$

the condition on α is always satisfied by choosing $\alpha = \frac{\kappa}{\max\{\|\Psi\mu - y\|_{\mathbb{L}}, \kappa\}}$.