Theoretical Guarantees for Variational Inference with Fixed-Variance Mixture of Gaussians

Tom Huix [1] **Anna Korba** [2] **Alain Durmus** [1] **Eric Moulines** [1]

Abstract

Variational inference (VI) is a popular approach in Bayesian inference, that looks for the best approximation of the posterior distribution within a parametric family, minimizing a loss that is typically the (reverse) Kullback-Leibler (KL) divergence. Despite its empirical success, the theoretical properties of VI have only received attention recently, and mostly when the parametric family is the one of Gaussians. This work aims to contribute to the theoretical study of VI in the non-Gaussian case by investigating the setting of Mixture of Gaussians with fixed covariance and constant weights. In this view, VI over this specific family can be casted as the minimization of a Mollified relative entropy, i.e. the KL between the convolution (with respect to a Gaussian kernel) of an atomic measure supported on Diracs, and the target distribution. The support of the atomic measure corresponds to the localization of the Gaussian components. Hence, solving variational inference becomes equivalent to optimizing the positions of the Diracs (the particles), which can be done through gradient descent and takes the form of an interacting particle system. We study two sources of error of variational inference in this context when optimizing the mollified relative entropy. The first one is an optimization result, that is a descent lemma establishing that the algorithm decreases the objective at each iteration. The second one is an approximation error, that upper bounds the objective between an optimal finite mixture and the target distribution.

[1]CMAP, Ecole polytechnique [2]ENSAE, CREST, IP Paris. Correspondence to: Tom Huix <tom.huix@polytechnique.edu>, Anna Korba <anna.korba@ensae.fr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

A fundamental problem in computational statistics and machine learning is to compute integrals with respect to some target probability distribution $\mu^\star$ on $\mathbb{R}^d$ whose density is known only up to a normalization constant. For instance in Bayesian inference, $\mu^\star$ is the posterior distribution over the parameters of complex models. The general goal of sampling methods is thus to provide an approximate distribution for which the integrals are easily computed. A large number of methods have been developed to tackle this problem. The classical approach is to sample the posterior using Markov Chain Monte Carlo (MCMC) algorithms, in which a Markov chain designed to converge to $\mu^\star$ is simulated for a sufficiently long time (Roberts & Rosenthal, 2004). These methods use the discrete measure over past iterates of the algorithm as an approximation of the posterior to compute integrals of interest. However, MCMC algorithms are generally computationally expensive, and it is an open problem to diagnose their convergence in practice (Moins et al., 2023). Variational inference (VI) (Blei et al., 2017) has emerged as a powerful and versatile alternative in Bayesian inference. By framing the problem as an optimization task, VI aims to find an approximate candidate distribution within a parametric family of distributions $\mathcal{C}$ that minimizes the (reverse) Kullback-Leibler (KL) divergence to the target:

$$\hat{\nu} := \operatorname*{argmin}_{\mu \in \mathcal{C}} \mathrm{KL}(\mu|\mu^\star), \tag{1}$$

where $\mathrm{KL}(\mu|\mu^\star) = \int \log(d\mu/d\mu^\star) d\mu$ if μ is absolutely continuous with respect to $\mu^\star$ denoting $d\mu/d\mu^\star$ its Radon-Nikodym density, and $+\infty$ else; and $\hat{\nu}$ is referred to as the optimal approximation within the variational family.

While VI methods can only return an approximation of the target, they are much more tractable in the large scale setting, since they benefit from efficient optimization methods, e.g. parallelization or stochastic optimization (Zhang et al., 2018). Hence, VI has proven effective in numerous applications and is a popular paradigm especially in high-dimensional scenarios. Still, the understanding of its theoretical properties remains a challenging and active area of research. Fundamentally, there are two sources of errors in VI: the *approximation* error that quantifies how far $\hat{\nu}$ is from $\mu^\star$, and the *optimization* error that comes from the

optimization of the objective in (1) to approach $\hat{\nu}$.

Even among the recent literature on theoretical guarantees for VI, most efforts have been concentrated in the case where $\mathcal{C}$ is the set of non-degenerate Gaussian distributions. Recently, Katsevich & Rigollet (2023) studied the approximation quality (in total variation) of the approximate posterior $\hat{\nu}$, i.e., minimizers of the objective (1), and show that it better estimates the true mean and covariance of the posterior than the well-known Laplace approximation (Helin & Kretschmann, 2022). Regarding the optimization of (1), still restricted to Gaussians, several recent works leverage the geometry of Wasserstein gradient flows, more precisely the equivalence between Bures-Wasserstein gradient flows on the space of probability distributions and Euclidean flows on the space of parameters of the variational approximation. They derive novel algorithms with convergence guarantees e.g. through gradient-descent (Lambert et al., 2022) or forward-backward (Diao et al., 2023; Domke et al., 2023) time discretizations; and precise connections with Black-Box Variational Inference (BBVI) (Yi & Liu, 2023).

However, to the best of our knowledge, the study of approximation and computational guarantees when $\mathcal{C}$ is a set of mixture of Gaussians has not been tackled yet. Mixture models are a widely used class of probabilistic models that capture complex and multi-modal data distributions by combining simpler components. Moreover, they are dense in the space of probability distributions with p bounded moments in the Wasserstein-p metric (Delon & Desolneux, 2020, Lemma 3.1).

In this study, we propose to consider a simplified setting where the Gaussian components have equal weights and share the same diagonal covariance. This regime breaks down the complexity of the problem, and is still theoretically challenging, but remains a practically relevant scenario. In this setting, variational inference aims to optimize the locations of the means of the Gaussian mixture to approximate the target distribution.

Contributions. In this paper, we derive theoretical guarantees for variational inference for some mixture of Gaussians family. We leverage the framework of Wasserstein gradient flows as well as the smoothness of the optimization objective to derive a descent lemma, showing that the objective decreases at each discrete time iteration. Regarding the approximation quality of Gaussian mixtures in (reverse) KL divergence, we use a similar technique than (Li & Barron, 1999) that established exact rates for the (forward) KL divergence, and we obtain upper bounds on the approximation error of VI in that setting.

This paper is organized as follows. Section 2 provides the relevant background on optimization over the space of probability distributions and introduces the mollified relative entropy that is the objective functional minimized in our context. In Section 3 we derive a descent lemma, establishing that the Wasserstein gradient descent algorithm decreases the objective at each iteration. In Section 4 we focus on the approximation error that quantifies how well minimizers of the VI objective approach the target distribution, for a given number of mixture components. In Section 5 we connect our results with relevant works in the Variational Inference literature.

Notations. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability distributions on $\mathbb{R}^d$ with bounded second moments. Given a Lebesgue measurable map $T : X \to X$ and $\mu \in \mathcal{P}_2(X)$, $T_{\#}\mu$ is the pushforward measure of μ by T. For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $L^2(\mu)$ is the space of functions $f : \mathbb{R}^d \to \mathbb{R}^d$ such that $\int \|f\|^2 d\mu < \infty$. We denote by $\|\cdot\|_{L^2(\mu)}$ and $\langle\cdot,\cdot\rangle_{L^2(\mu)}$ respectively the norm and the inner product of the Hilbert space $L^2(\mu)$. We consider, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance $W_2(\mu,\nu) = \inf_{s\in\mathcal{S}(\mu,\nu)} \int \|x-y\|^2 ds(x,y)$, where $\mathcal{S}(\mu,\nu)$ is the set of couplings between μ and ν. The metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is called the Wasserstein space. We use $C^k(\mathbb{R}^d)$ to denote continuously k-differentiable functions and $C^\infty(\mathbb{R}^d)$ to indicate the smooth functions. The space of continuous k-differentiable functions with compact support on X is $C_c^k(\mathbb{R}^d)$. If $\psi : \mathbb{R}^d \to \mathbb{R}^p$ is differentiable, we denote by $\mathrm{J}\psi : \mathbb{R}^d \to \mathbb{R}^{p\times d}$ its Jacobian. If $p = 1$, we denote by $\nabla\psi$ the gradient of ψ. Moreover, if $\nabla\psi$ is differentiable, the Jacobian of $\nabla\psi$ is the Hessian of ψ denoted by $\mathrm{H}\psi$. If $p = d$, $\boldsymbol{\nabla}\cdot\psi$ denotes the divergence of ψ. We also denote by $\Delta\psi$ the Laplacian of ψ, where $\Delta\psi = \boldsymbol{\nabla}\cdot\nabla\psi$. The Hilbert-Schmidt norm is denoted $\|\cdot\|_{HS}$.

In the following, we assume that $\mu^\star$ admits a density proportional to $\exp(-V)$ with respect to the Lebesgue measure over $\mathbb{R}^d$.

2. The mollified relative entropy

Writing $\mu^\star = e^{-V}/Z$ with Z the unknown normalization constant, the (reverse) Kullback-Leibler divergence (or relative entropy) can be written as

$$\begin{aligned}\mathrm{KL}(\mu|\mu^\star) &= \int V d\mu + \int \log(\mu) d\mu + \log(Z) \\ &:= \mathcal{G}_V(\mu) + \mathcal{U}(\mu) + \log(Z),\end{aligned}$$

for μ absolutely continuous with respect to $\mu^\star$, and $+\infty$ else. Hence, it decomposes as the sum of a potential energy $\mathcal{G}_V$, i.e. a linear functional, and the negative entropy $\mathcal{U}$, up to an additive constant that is fixed in the optimization problem.

We now consider the minimization problem of Variational Inference (1) for mixture of Gaussians. We will study a specific setting where the variational family is the set of

mixture of n Gaussians with equally weighted components, and where these components have the same diagonal covariance $\epsilon^2 \mathrm{I}_d$, for some $n \in \mathbb{N}^*, \epsilon > 0$.

$$\mathcal{C}_n = \left\{ \frac{1}{n} \sum_{i=1}^{n} q_i,\ q_i = \mathcal{N}(x_i, \epsilon^2 \mathrm{I}_d),\ x_i \in \mathbb{R}^d \right\},$$

where I_d denotes the d-dimensional identity matrix. In our setting, only the positions (the means) of the mixture components will be optimized. Hence, searching for the optimal distribution in the variational family approximating the target μ^* consists in finding the optimal locations of the Gaussian components in $\mathbb{R}^d$. We will denote k_ϵ the normalized Gaussian kernel, i.e. $k_\epsilon(x) = \exp\big(-\|x\|^2/(2\epsilon^2)\big) Z_\epsilon^{-1}$, where $\int k_\epsilon(x)dx = 1$ and $Z_\epsilon \propto (\epsilon^2)^{d/2}$. It is a specific example of mollifiers, i.e. smooth approximations of the Dirac delta at the origin, as introduced in (Friedrichs, 1944). For μ a given probability distribution on $\mathbb{R}^d$, we denote by $k_\epsilon \star \mu$ its convolution with the Gaussian kernel that writes $k_\epsilon \star \mu = \int k_\epsilon(\cdot - x)d\mu(x)$. Equipped with these notations, we can write $\mathcal{C}_n = \big\{k_\epsilon \star \mu_n,\ \mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x^i},\ x^1, \dots, x^n \in \mathbb{R}^d\big\}$.

Irrespective of the number of components n, VI with Gaussian mixtures whose components share the same variance can be written more generally as minimizing (1) restricted to the family $\mathcal{C} = \big\{k_\epsilon \star \mu,\ \mu \in \mathcal{P}(\mathbb{R}^d)\big\}$. The latter problem can be then reformulated as the optimization over $\mathcal{P}(\mathbb{R}^d)$ of the following objective functional, that we will refer to as the *mollified relative entropy* (or mollified KL):

$$\begin{aligned}\mathcal{F}_\epsilon(\mu) &= \int V d(k_\epsilon \star \mu) + \int \log(k_\epsilon \star \mu) d(k_\epsilon \star \mu) \\ &:= \mathcal{G}_{V_\epsilon}(\mu) + \mathcal{U}_\epsilon(\mu), \end{aligned} \tag{2}$$

where $\mathcal{G}_{V_\epsilon}$ is a potential energy with respect to a convoluted potential $V_\epsilon = k_\epsilon \star V$ (using the associativity of the convolution operation), and $\mathcal{U}_\epsilon(\mu) = \mathcal{U}(k_\epsilon \star \mu)$ is a functional that we will refer to as the mollified negative entropy. In contrast with the negative entropy defined above, the mollified one is well-defined for discrete measures.

2.1. Algorithm

We now discuss the optimization of the mollified relative entropy, starting from the continuous time dynamics to the practical discrete-time particle scheme.

A Wasserstein gradient flow of $\mathcal{F}_\epsilon$ (Ambrosio et al., 2008) can be described by the following continuity equation:

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot (\mu_t \nabla_{W_2} \mathcal{F}_\epsilon(\mu_t)),\ \nabla_{W_2} \mathcal{F}_\epsilon(\mu_t) := \nabla \mathcal{F}'_\epsilon(\mu_t), \tag{3}$$

where $\mathcal{F}'_\epsilon$ denotes the first variation of $\mathcal{F}_\epsilon$. Recall that if it exists, the first variation of a functional $\mathcal{F}$ at ν is the function $\mathcal{F}'(\nu) : \mathbb{R}^d \to \mathbb{R}$ s. t. for $\nu, \mu \in \mathcal{P}(\mathbb{R}^d)$: $\lim_{\epsilon \to 0} {}^1\!/\!_\epsilon [\mathcal{F}(\nu + \epsilon(\mu - \nu)) - \mathcal{F}(\nu)] = \int \mathcal{F}'(\nu)(x)(d\mu(x) - d\nu(x))$. Wasserstein gradient flows are paths of steepest descent with respect to the W_2 metric, and can be seen as analog to Euclidean gradient flows on the space of probability distributions (Santambrogio, 2017).

Starting from some initial distribution $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$, and for some given step-size $\gamma > 0$, a forward (or explicit) time-discretization of (3) corresponds to the Wasserstein gradient descent algorithm, and can be written at each discrete time iteration $l \in \mathbb{N}$ as:

$$\mu_{l+1} = (\mathrm{Id} - \gamma \nabla \mathcal{F}'_\epsilon(\mu_l))_{\#} \mu_l \tag{4}$$

where Id is the identity map in $L^2(\mu_l)$.

For discrete measures $\mu_n = {}^1\!/\!_n \sum_{i=1}^n \delta_{x^i}$, we can define the finite-dimensional objective $F(X^n) := \mathcal{F}_\epsilon(\mu_n)$ where $X^n = (x^1, \dots, x^n)$, since the functional $\mathcal{F}_\epsilon$ is well defined for discrete measures. The Wasserstein gradient descent dynamics of $\mathcal{F}_\epsilon$ (4) then correspond to standard gradient descent of the (finite-dimensional) function F, i.e., gradient descent on the position of the particles. In that setting, we recall that particles correspond to the means of the Gaussian components of the mixture. The gradient of F is readily obtained as

$$\begin{aligned}\nabla_{x^j} F(X^n) = &\int_{\mathbb{R}^d} \nabla V(y) k_\epsilon(y - x^j) dy \\ &+ \int_{\mathbb{R}^d} \frac{\sum_{i=1}^n \nabla k_\epsilon(y - x^i)}{\sum_{i=1}^n k_\epsilon(y - x^i)} k_\epsilon(y - x^j) dy. \end{aligned} \tag{5}$$

Notice that the gradient above involves integrals over $\mathbb{R}^d$. However, using a Gaussian kernel k_ϵ, since $\nabla k_\epsilon(x) = -\frac{x}{\epsilon^2} k_\epsilon(x)$, these integrals can be easily approximated through Monte Carlo using Gaussian samples. A particle version of (4), e.g., starting with μ_0 discrete, can then be written as the following gradient descent iterates:

$$x^j_{l+1} = x^j_l - \gamma \nabla_{x^j_l} F(X^n_l) \tag{6}$$

for $j = 1, \dots, n$ and where $X^n_l = (x^1_l, \dots, x^n_l)$. Hence, minimizing $\mathcal{F}_\epsilon$ on discrete measures results in a a particle system that interact through the gradient of the objective. The reader may refer to Appendix A for the detailed computations leading to the particle scheme. Notice that it recovers the scheme mentioned in (Lambert et al., 2022, Section 5) where the covariance of the mixture components are fixed, see Appendix B for a detailed discussion.

Remark 1. *Notice that the Wasserstein gradient at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ of the mollified KL in Equation (3), $\nabla \mathcal{F}'_\epsilon(\mu_t) : \mathbb{R}^d \to \mathbb{R}^d$ writes for any $w \in \mathbb{R}^d$:*

$$\nabla \mathcal{F}'_\epsilon(\mu_t)(w) = k_\epsilon \star \nabla V(w) + k_\epsilon \star \nabla \log(k_\epsilon \star \mu)(w), \tag{7}$$

see Appendix A. Hence, it differs from the Wasserstein gradient of the (standard) KL w.r.t. $\mu^\star \propto e^{-V}$*, i.e.* $\mathrm{KL}(\cdot|\mu^\star)$ *evaluated at the convoluted distribution that writes as* $\nabla \log\left(k_\epsilon \star \mu / \mu^\star\right)$*, see (Wibisono, 2018, Section 3.1.3).*

2.2. Non-smoothness of the KL

In Euclidean optimization, it is standard that the convergence of gradient descent is guaranteed when the objective function is convex and smooth, which relates to a lower bound and upper bound on the Hessian of the objective when the latter is twice differentiable (Garrigos & Gower, 2023). Analogously, when optimizing a functional on the Wasserstein space, lower and upper bounds on the Hessian characterize respectively convexity and smoothness on the functional $\mathcal{F}$ with respect to the Wasserstein-2 geometry (see Villani (2009, Proposition 16.2)). The Wasserstein space has a Riemannian geometry (Otto, 2001), where one can define for any μ the tangent space $\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla\psi, \ \ \psi \in C_c^\infty(\mathbb{R}^d)\}} \subset L^2(\mu)$ (Ambrosio et al., 2008, Definition 8.4.1). The W_2 Hessian of a functional $\mathcal{F}$, denoted $H\mathcal{F}_{|\mu}$ is an operator over $\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$ verifying $\langle H\mathcal{F}_{|\mu} v_t, v_t\rangle_{L^2(\mu)} = \frac{d^2}{dt^2}\Big|_{t=0} \mathcal{F}(\rho_t)$ if $t \mapsto \rho_t$ is a geodesic starting at μ with vector field $t \mapsto v_t$. Considering $\psi \in C_c^\infty(\mathbb{R}^d)$ and the path ρ_t from μ to $(I + \nabla\psi)_\# \mu$ given by: $\rho_t = (I + t\nabla\psi)_\# \mu$, for all $t \in [0,1]$, the Hessian of $\mathcal{F}$ at μ, $H\mathcal{F}_{|\mu}$, is defined as a symmetric bilinear form on $C_c^\infty(\mathbb{R}^d)$ associated with the quadratic form $\mathrm{Hess}_\mu \mathcal{F}(\psi,\psi) := \frac{d^2}{dt^2}\Big|_{t=0} \mathcal{F}(\rho_t)$.

We now recall the formula of the Wasserstein Hessian of the (standard) Kullback-Leibler divergence (or relative entropy).

Proposition 2. *(Villani, 2021, Section 9.1.2). Assume that* $\mu^\star$ *has a density* $\mu^\star \propto e^{-V}$ *where the potential* $V : X \to \mathbb{R}$ *is* $C^2(\mathbb{R}^d)$*. The Hessian of* $\mathrm{KL}(\cdot|\mu^\star)$ *at* μ *is given, for any* $\psi \in C_c^\infty(\mathbb{R}^d)$*, by:*

$$\begin{aligned}&\mathrm{Hess}_\mu \,\mathrm{KL}(\psi,\psi)\\ &= \int \left[\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|_{HS}^2\right] \mathrm{d}\mu(x)\\ &\qquad = \mathrm{Hess}_\mu \,\mathcal{G}_V(\psi,\psi) + \mathrm{Hess}_\mu \,\mathcal{U}(\psi,\psi), \quad (8)\end{aligned}$$

where H_V *is the Hessian of* V*.*

The proof of Proposition 2 is provided in Appendix E.1 for completeness. The reader may also refer to (Korba et al., 2021; Duncan et al., 2023) for similar computations on Wasserstein Hessians.

The KL divergence inherits the convexity of the target potential V in the Wasserstein geometry. Indeed, if $\mathrm{H}_V \succeq \lambda \,\mathrm{I}_d$, then $\mathrm{KL}(\cdot|\mu^\star)$ is λ-displacement convex, i.e. it is λ-convex along Wasserstein-2 geodesics, the underlying geometry for Wasserstein gradient flows. Yet, the Kullback-Leibler divergence is not a smooth objective in the Wasserstein sense, since its (Wasserstein) Hessian is not upper bounded even if the potential V is smooth. Indeed, assume $\mathrm{H}_V \preceq M \,\mathrm{I}_d$, i.e., the potential of the target distribution is M-smooth. This enables to control the first term in (8) by $M\|\nabla\psi\|^2_{L^2(\mu)}$, but the second term due to the negative entropy cannot be controlled similarly for any ψ (Wibisono, 2018; Korba et al., 2020).

Hence in this context, it is not possible to prove a descent lemma along (Wasserstein) gradient descent for the KL, unless restricting to smooth directions (Korba et al., 2020). The non-smoothness of the KL is also the reason why many algorithms aiming to minimize the KL in the Wasserstein geometry rely on splitting-schemes such as the forward-backward algorithm, to perform a gradient descent (explicit) step on the potential energy part, and a JKO (implicit) step on the entropy part (Salim et al., 2020; Diao et al., 2023; Domke et al., 2023). In contrast, we will leverage the fact that the *mollified* KL enjoys some smoothness properties that will allow us to derive a descent lemma in Section 3, at the price of loosing some convexity.

Still, we next show that $\mathcal{F}_\epsilon$ recovers displacement convexity (of the standard KL) as $\epsilon \to 0$, since its Hessian recovers the one of the KL.

Proposition 3. *Let* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$*. For any* $\psi \in C_c^\infty(\mathbb{R}^d)$*, the Wasserstein Hessian of* $\mathcal{F}_\epsilon$ *converges to the one of the regular KL, i.e:*

$$\mathrm{Hess}_\mu \,\mathcal{F}_\epsilon(\psi,\psi) \xrightarrow[\varepsilon\to 0]{} \mathrm{Hess}_\mu \,\mathrm{KL}(\psi,\psi). \quad (9)$$

The proof of Proposition 3 can be found in Appendix E.2; the main technical difficulties arise when dealing with the negative entropy term. This result shows that as $\epsilon \to 0$, one can recover the geometric properties of the KL.

Proposition 3 serves as an auxiliary finding within our study, not directly influencing other results, yet it enables us to illustrate key conceptual distinctions. Specifically, it demonstrates that while the standard Kullback-Leibler (KL) divergence is convex in the Wasserstein geometry for log-concave targets—exhibiting even strong convexity for targets that are strongly log-concave—it loses this convexity when mollified, although it gains smoothness with a positive ϵ. This transition is typically delineated through lower and upper bounds on the Hessians within the Wasserstein framework. Getting a non-asymptotic, quantitative bounds on the Hessian of the mollified KL in terms of ϵ is the subject of future work. Such research could potentially offer insights into how small ϵ may be selected relative to the strong convexity constant of the target potential, ensuring the optimization objective maintains convexity.

3. Optimization Guarantees

We now turn to the analysis of the optimization error for VI in our setting, i.e. the optimization of $\mathcal{F}_\epsilon$. Under a smoothness assumption on the target potential, as well as moment conditions on the trajectory, one can obtain a descent lemma for the Wasserstein gradient descent iterates.

Assumption 1. The potential V is L-smooth, i.e. for any $x, y \in \mathbb{R}^d$, $\|\nabla V(x) - \nabla V(y)\| \leq L\|x - y\|$.

Assumption 2. μ_0 is supported on n Diracs, and the second moments of $(\mu_l)_{l\geq 0}$ are bounded by $h > 0$ along gradient descent iterations, i.e. $\int \|x\|^2 d\mu_l(x) < h\ , \forall l \geq 0$.

Bounded moment assumptions such as these are commonly used in stochastic optimization, for instance in some analysis of the stochastic gradient descent (Moulines & Bach, 2011). We also verified empirically this assumption in a specific setting outlined afterwards. The target $\mu^\star$ is a mixture of 100 Gaussians that we approximate with a mixture of 10 Gaussians. Then we run (6) (equivalently (4)) for 1000 iterations. The expectations in (5) with respect to the Gaussian kernel are estimated by Monte Carlo with 100 samples. Figure 1 displays the second moments of the particle distributions along iterations, for various dimensions. The 95% confidence interval displayed in Figure 1 is calculated based on 50 runs, and represents the randomness corresponding to Monte Carlo approximations, initialization of the target and initialization of our mixture. Our experiment shows that Assumption 2 holds for any dimension, i.e., the second moment of the particles distribution is bounded along the (discrete-time) flow. Further details on the setup are provided in Appendix F. We now turn to one of our main results regarding the optimization of the mollified KL.

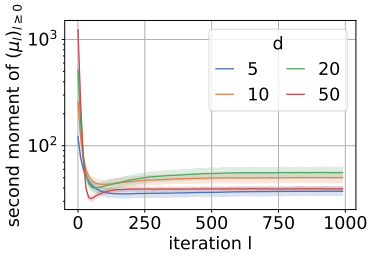

Figure 1. Second moment along Wasserstein gradient descent iterations.

Proposition 4. *Suppose Assumption 1 and Assumption 2 hold. Consider the sequence of iterates of Wasserstein gradient descent of $\mathcal{F}_\epsilon$ defined by (4). Then, the following inequality holds:*

$$\mathcal{F}_\epsilon(\mu_{l+1}) - \mathcal{F}_\epsilon(\mu_l) \leq -\gamma\left(1 - \frac{\gamma}{2}M\right)\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}.$$

where $M = L + K_{\epsilon,n,h}$, and $K_{\epsilon,n,h}$ is a constant depending on ϵ, n, h.

Hence, for a small enough step-size γ, the latter proposition shows that the objective decreases at each iteration. We now provide a proof for this result, using similar techniques as (Arbel et al., 2019; Korba et al., 2020). The main technical difficulties are left in the appendix and are related to showing the descent for the mollified entropy part, see Appendix D for details.

Proof of Proposition 4. Consider a path between μ_l and μ_{l+1} of the form $\rho_t = (\psi_t)_\# \mu_l$ with $\psi_t = (\mathrm{Id} + t\nabla\mathcal{F}'_\epsilon(\mu_l))$. We have $\frac{\partial \rho_t}{\partial t} = \boldsymbol{\nabla}\cdot(\rho_t v_t)$ with $v_t = -\nabla\mathcal{F}'_\epsilon(\mu_l)\circ\psi_t^{-1}$. The latter continuity equation holds in the sense of distributions (Ambrosio et al., 2008, Chapter 8) and holds for discrete measures. The function $t \mapsto \mathcal{F}_\epsilon(\rho_t)$ is differentiable and hence absolutely continuous. Therefore one can write:

$$\mathcal{F}_\epsilon(\rho_\gamma) = \mathcal{F}_\epsilon(\rho_0) + \gamma\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{F}_\epsilon(\rho_t) + \int_0^\gamma\left[\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_\epsilon(\rho_t) - \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{F}_\epsilon(\rho_t)\right]dt. \quad (10)$$

Moreover, using the chain rule in the Wasserstein space, we have successively:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_\epsilon(\rho_t) = \langle\nabla\mathcal{F}'_\epsilon(\rho_t), v_t\rangle_{L^2(\rho_t)}, \quad \text{and } \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{F}_\epsilon(\rho_t) = -\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}. \quad (11)$$

Then, since $\mathcal{F}_\epsilon = \mathcal{U}_\epsilon + \mathcal{G}_{V_\epsilon}$, we have first under Assumption 1 that $k_\epsilon \star V$ is L-smooth and by Proposition 11 that:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}_{V_\epsilon}(\rho_t) - \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}_{V_\epsilon}(\rho_t)\Big|_{t=0} \leq L\ t\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)}, \quad (12)$$

and by Proposition 12 and Assumption 2:

$$\frac{d}{dt}\mathcal{U}_\epsilon(\rho_t) - \frac{d}{dt}\mathcal{U}_\epsilon(\rho_t)\Big|_{t=0} \leq K_{\epsilon,n,h}t\|\phi\|^2_{L^2(\mu_l)},$$

where $K_{\epsilon,n,h} = 1/\epsilon^2 + 2\sqrt{hn}/\epsilon^3 + \sqrt{n}/\epsilon^2 + n\sqrt{h}/2\epsilon^3$. Hence, the result follows directly by applying the above expressions to Equation (10) where $M = L + K_{\epsilon,n,h}$. □

As a corollary, we obtain the convergence of the average of squared gradient norms along iterations.

Corollary 5. *Let $c_\gamma = \gamma(1 - \frac{\gamma M}{2})$. Under the assumptions of Proposition 4, one has*

$$\frac{1}{L}\sum_{l=1}^{L}\|\nabla\mathcal{F}'_\epsilon(\mu_l)\|^2_{L^2(\mu_l)} \leq \frac{\mathcal{F}_\epsilon(\mu_0)}{2c_\gamma L}. \quad (13)$$

In contrast with the KL that is non-smooth as explained in Section 2.2, the mollified KL is smooth, which is why

we can prove the descent lemma in Proposition 4 and the rate on average gradients. The descent lemma and its corollary imply that the sequence of squared gradient norms is summable and hence converges to zero.

We illustrate the validity of the rate derived in Corollary 5 with simple experiments. The variational family there is a family of Gaussian mixtures with 10 components, while the target is a Gaussian mixture with 100 components. Figure 2 shows the convergence of the cumulative sum $\frac{1}{L}\sum_{l=1}^{L}\|\nabla\mathcal{F}_\epsilon'(\mu_l)\|^2_{L^2(\mu_l)}$ along iterations, for various dimensions and in log scale. Similarly to the previous experiment, the expectations involved in the gradient descent schemes are estimated using Monte Carlo with 100 samples. The 95% confidence interval displayed in Figure 2 is computed based on 50 runs, representing the randomness due to Monte Carlo approximations, randomization of the target and ininital distribution for the scheme. The term $\|\nabla\mathcal{F}_\epsilon'(\mu_l)\|^2_{L^2(\mu_l)}$ also involves expectations that are estimated by Monte Carlo with 1000 samples. Figure 2 illustrates that the cumulative sum is indeed of order $\frac{1}{L}$ as stated by Corollary 5. A detailed description of the experimental setup can be found in Appendix F.

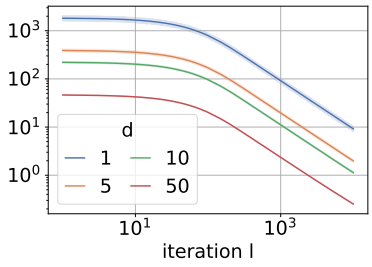

Figure 2. Illustration of the rate of $\frac{1}{L}\sum_{l=1}^{L}\|\nabla\mathcal{F}_\epsilon'(\mu_l)\|^2_{L^2(\mu_l)}$ derived in Corollary 5

Remark 6. *Non-convex rates similar to our Corollary 5 have been obtained for Langevin Monte-Carlo (Balasubramanian et al., 2022) or Stein Variational Gradient Descent (SVGD) algorithm (Korba et al., 2020) leveraging similar techniques and smoothness of the potential. However, since Langevin Monte Carlo and SVGD optimizes the (standard) KL divergence, the squared gradient norm correspond to the Fisher Divergence and Kernel Stein Discrepancy respectively, that are valid probability divergences. In our setting, Corollary 5 implies the following. If μ_l converges weakly to some distribution μ_∞ (up to a subsequence) as $l \to \infty$, the Wasserstein gradient of $\mathcal{F}_\epsilon$ given in Equation (7) is zero on the support of μ_∞, assuming $\mu \mapsto \|\nabla\mathcal{F}_\epsilon'(\mu)\|^2_{L^2(\mu)}$ is lower semi continuous with respect to the weak topology of measures. This can be rewritten $\nabla k_\epsilon \star (\log(k_\epsilon \star \mu_\infty) - \log(\mu^\star)) = 0$ μ_∞-a.e, i.e. $k_\epsilon \star (\log(k_\epsilon \star \mu_\infty) - \log(\mu^\star)) = c$ $\mu_\infty-$a.e. for some constant c.*

4. Approximation Guarantees

In this section, we investigate the approximation accuracy of a finite mixture of Gaussians to the posterior, i.e. minimizers of the objective functional $\mathcal{F}_\epsilon$ (assuming we are able to find these minimizers, e.g. after optimization). We obtain non-asymptotic rates with respect to the number of components in the mixture. For ease of notation, we will denote by $k_\epsilon^x := k_\epsilon(\cdot - x)$ for any $x \in \mathbb{R}^d$. We first consider the following assumption on the target distribution.

Assumption 3. The target posterior distribution $\mu^\star$ has a mixture representation form, i.e. there exists P on $\mathbb{R}^d$ such that

$$\mu^\star = \int_\Theta k_\epsilon^w dP(w).$$

Notice that Assumption 3 is a relatively weak assumption, as mixture of Gaussians are dense in the space of probability distributions (Delon & Desolneux, 2020). We now state our second main result.

Theorem 7. *Suppose Assumption 3 holds and define*

$$C^2_{\mu^\star} = \int \frac{\int (k_\epsilon^m(x))^2 dP(m)}{\int k_\epsilon^w(x) dP(w)} dx. \tag{14}$$

Define $\mathcal{C}_n = \left\{k_\epsilon \star \mu_n,\ \mu_n \in \mathcal{P}_n(\mathbb{R}^d)\right\}$, where $\mathcal{P}_n(\mathbb{R}^d)$ is the set of discrete probability distributions supported on n Dirac masses. Then,

$$\min_{\mu_n \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{KL}(k_\epsilon \star \mu_n | \mu^\star) \leq C^2_{\mu^\star} \frac{\log(n)+1}{n}.$$

Our result is novel and quantifies the approximation quality of the family of mixtures of n Gaussian distributions (with equal weight and constant covariance) in the (reverse) Kullback-Leibler sense.

A major limitation in the use of Gaussian distributions in VI arises from the inherent simplicity of this family. In particular, the unimodality of the Gaussian distribution becomes a critical stumbling block when the target distribution is multimodal. A notable exception exists in the work of Katsevich & Rigollet (2023), which provides an error bound for cases where the target is a posterior distribution in the Bayesian inference context. As the sample size goes to infinity, the Bernstein Von-Mises theorem shows that the posterior distribution asymptotically converges to a Gaussian distribution, thereby lending some predictability to the approximation error in this specific scenario. In stark contrast, Theorem 7 offers a more versatile result, applicable to any target distribution, including those encountered in Bayesian inference with a fixed sample size. It shows that increasing the number of components in a Gaussian mixture can significantly mitigate the limitations of Gaussian VI. As we expand the mixture, the approximation error not only

decreases, it converges to zero. This result highlights the potential of complexifying the variational family to achieve more accurate approximations of the target distribution.

The proof of Theorem 7 follows the steps of (Li & Barron, 1999), that proved similar guarantees for the forward KL (akin to likelihood maximization), while we focus on the reverse KL, i.e. the one considered in variational inference. Hence our proof requires non-trivial different inequalities and intermediate lemmas that are deferred to Appendix C.

Proof of Theorem 7. We will prove the previous result by induction. We denote by ν_n the minimizer of the Kullback-Leibler divergence to the target within this family, i.e.,

$$\nu_n := \operatorname*{argmin}_{\mu_n \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{KL}(k_\epsilon \star \mu_n | \mu^\star),$$

and $D_n = \mathrm{KL}(k_\epsilon \star \mu_n | \mu^\star)$. For any $m \in \mathbb{R}^d$, we consider the distribution $\rho^m_{n+1} \in \mathcal{C}_{n+1}$ defined as

$$\rho^m_{n+1} = (1-\alpha)(k_\epsilon \star \mu_n) + \alpha k^m_\epsilon$$

where $\alpha = 1/n+1$. Therefore we have $D_{n+1} = \mathrm{KL}(k_\epsilon \star \mu_{n+1} | \mu^\star) \leq \mathrm{KL}(\rho^m_{n+1} | \mu^\star)$. By definition of the Kullback-Leibler divergence, denoting $f(x) = x \log x$, we have

$$\mathrm{KL}(\rho^m_{n+1} | \mu^\star) = \int f(r_{n+1}) d\mu^\star \ ,$$

where we define r_{n+1} and r_0 as:

$$r_{n+1} := \frac{\rho^m_{n+1}}{\mu^\star} = (1-\alpha)\frac{(k_\epsilon \star \mu_n)}{\mu^\star} + \alpha \frac{k^m_\epsilon}{\mu^\star} := r_0 + \alpha \frac{k^m_\epsilon}{\mu^\star}.$$

Define $B(x) = (x \log x - x + 1)/(x-1)^2$ for $x \in [0, +\infty[$. Note that $r_{n+1}(x) \geq r_0(x)$ for any x, then using that B is decreasing (see Lemma 8), we have $B(r_{n+1}(x)) \leq B(r_0(x))$. It follows that

$$\begin{aligned}
&r_{n+1} \log(r_{n+1}) \\
&\quad \leq r_{n+1} - 1 + B(r_0)(r_{n+1} - 1)^2 \\
&\quad = r_0 + \alpha \frac{k^m_\epsilon}{\mu^\star} - 1 + B(r_0)\left(r_0 + \alpha \frac{k^m_\epsilon}{\mu^\star} - 1\right)^2 \\
&\quad = r_0 + \alpha \frac{k^m_\epsilon}{\mu^\star} - 1 \\
&\qquad + B(r_0)\Big\{(r_0-1)^2 + \left(\alpha \frac{k^m_\epsilon}{\mu^\star}\right)^2 + 2\alpha(r_0 - 1)\frac{k^m_\epsilon}{\mu^\star}\Big\} \\
&\quad = \alpha \frac{k^m_\epsilon}{\mu^\star} + r_0 \log(r_0) + \left(\alpha \frac{k^m_\epsilon}{\mu^\star}\right)^2 B(r_0) \\
&\qquad + 2\alpha B(r_0)(r_0 - 1)\frac{k^m_\epsilon}{\mu^\star}. \qquad (15)
\end{aligned}$$

Moreover, we have the following inequality:

$$\begin{aligned}
D_{n+1} &= \int D_{n+1} dP(m) \\
&\leq \int \mathrm{KL}(\rho^m_{n+1} | \mu) dP(m) \\
&= \alpha + \int r_0(x) \log(r_0(x)) d\mu^\star(x) \\
&\quad + \alpha^2 \iint \frac{k^m_\epsilon(x)^2}{\mu^\star(x)^2} B(r_0(x)) d\mu^\star(x) dP(m) \\
&\quad + 2\alpha \int B(r_0(x))(r_0(x) - 1) d\mu^\star(x),
\end{aligned}$$

where we used Equation (15) in the last equality. We now focus on bounding each term on the r.h.s. of the previous inequality. By definition of r_0, the second term can be rewritten

$$\int r_0 \log(r_0) d\mu^\star = (1-\alpha)\log(1-\alpha) + (1-\alpha) D_n.$$

We now turn to the third term. For any $x \in \mathbb{R}^+$, since B is monotone decreasing, $B(r_0(x)) \leq B(0) = 1$. Under Assumption 3, it follows that

$$\begin{aligned}
&\int \int \frac{k^m_\epsilon(x)^2}{\mu^\star(x)^2} B(r_0(x)) d\mu^\star(x) dP(m) \\
&\qquad \leq \int \int \frac{k^m_\epsilon(x)^2}{\mu^\star(x)} dP(m) dx = C^2_{\mu^\star}.
\end{aligned}$$

Finally let's focus on the last term. We have $B(x)(x-1) \leq \sqrt{x} - 1$ using Lemma 9. Denoting $H^2(f,g) = 1 - \int \sqrt{f(x)g(x)} dx \in [0,1]$ the squared Hellinger distance between f and g, we have

$$\begin{aligned}
\int B(r_0)(r_0 - 1) d\mu^\star \quad &\leq \int (\sqrt{r_0} - 1) d\mu^\star \\
= \sqrt{1-\alpha}(1 - H^2(k_\epsilon \star \mu_n, \mu^\star)) - 1 \quad &\leq \sqrt{1-\alpha} - 1.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&D_{n+1} \leq \alpha + (1-\alpha)\log(1-\alpha) + (1-\alpha) D_n + \\
&\alpha^2 C^2_{\mu^\star} + 2\alpha(\sqrt{1-\alpha} - 1) \leq (1-\alpha) D_n + \alpha^2 C^2_{\mu^\star},
\end{aligned}$$

where the last inequality uses that $-\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha\sqrt{1-\alpha} \leq 0$ (see Lemma 10).

Now, recall that $\alpha = 1/(n+1)$. Denoting $U_n = n D_n$, our previous computations imply that $U_{n+1} \leq U_n + C^2_{\mu^\star}/n+1$, which by telescoping yields $U_n - U_0 \leq C^2_{\mu^\star} H_n$, where H_n denotes the harmonic number and is upper bounded by $1 + \log(n)$. The rate on D_n follows. □

Our result is analog to the one of Li & Barron (1999) that bounds the forward Kullback-Leibler divergence to the target. Indeed under Assumption 3, their Theorem 1 states

that

$$\operatorname*{argmin}_{\mu_n \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{KL}(\mu^\star | k_\epsilon \star \mu_n) \leq \frac{C_{\mu^\star}^2 h}{n} \tag{16}$$

where $h = 4\log(3\sqrt{e}+a)$ is a constant depending on ϵ since $a = \sup_{m_1,m_2\in\mathbb{R}^d} \log\left(k_\epsilon^{m_1}(x)/k_\epsilon^{m_2}(x)\right)$. In our case, the constant in the rate does not involve h as we do not rely on the same functions (our $B \leq 1$ instead of $B \leq h$ in their case).

When Assumption 3 does not hold, they also show in Theorem 2 that for every $g_P = \int k_\epsilon(\cdot - w)dP(w)$,

$$\operatorname*{argmin}_{\mu_n \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{KL}(\mu^\star | k_\epsilon \star \mu_n) \leq \mathrm{KL}(\mu^\star | g_P) + \frac{C_{\mu^\star,P}^2 h}{n} \tag{17}$$

where $C_{\mu^\star,P}^2 = \int \frac{\int k_\epsilon^m(x)^2 dP(m)}{(\int k_\epsilon^w(x) dP(w))^2} d\mu^\star(x)$. However, they can easily obtain this result as a consequence of their first theorem along with the linearity of the forward KL. In constrast, the reverse KL does not verify linearity nor triangular inequality hence we cannot obtain readily such a generalization.

Notice that the forward KL rate obtained in (Li & Barron, 1999) is of order $1/n$, outpacing the one we attained. This is due to our chosen variational family, which is a Gaussian mixture with fixed weights. However, considering non-fixed weights (i.e. non-equally weighted mixtures) allows us to set $\alpha = 2/(n+1)$, thus achieving the exact same rate as (Li & Barron, 1999) for the reverse KL.

Since the Total Variation can be written as an Integral probability metric over measurable functions $f : \mathbb{R}^d \to [-1, 1]$, we deduce from Pinsker's inequality and Theorem 7 that a minimizer μ_n of $\mathrm{KL}(k_\epsilon \star \cdot | \mu^\star)$ achieves the following bound for the integral approximation error among this set of functions:

$$\left| \int f d(k_\epsilon \star \mu_n) - \int f d\mu^\star \right| \leq \sqrt{\frac{C_{\mu^\star}^2 (\log(n) + 1)}{2n}}.$$

The latter is then comparable to the integral approximation error of MCMC methods which is known to be of order $\mathcal{O}(n^{-\frac{1}{2}})$ when using n particles (Łatuszyński et al., 2013).

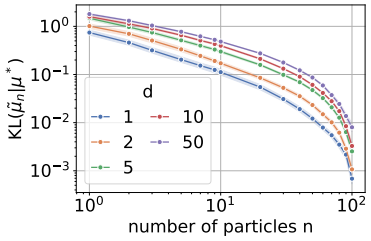

Figure 3. Illustration of the rates of Theorem 7, where $\nu_n = \operatorname{argmin}_{\nu\in\mathcal{C}_n} \mathrm{KL}(\nu|\mu^\star)$ is approximated by $\tilde{\nu}_n$.

We finally test numerically the validity of Theorem 7 in a simple setting. The target distribution considered is a Gaussian mixture with 100 components. We denote by $(x_i^\star)_{i\leq 100}$ the mean of these components. For any $n \in [1, 100]$, the objective is to solve (1) and find $\nu_n := \operatorname{argmin}_{\nu\in\mathcal{C}_n} \mathrm{KL}(\nu|\mu^\star)$, where $\mathcal{Q}_n$ represents the family of Gaussian mixtures with n components. This minimizer is approximated by selecting only the first n components $(x_i^\star)_{i\leq n}$ of $\mu^\star$, and we denote $\tilde{\nu}_n$ the resulting approximate distribution. Note that in that specific setting, the variational family $\mathcal{C}_n$ and the target distribution $\mu^\star$ share the same standard deviation. Figure 3 shows the convergence rate of $\mathrm{KL}(\tilde{\nu}_n|\mu^\star)$ with respect to the number of components n, for various dimensions. The objective is estimated by Monte Carlo with 1000 samples. The 95% confidence interval displayed in Figure 3 is approximated based on 100 samples, representing the randomness corresponding to Monte Carlo approximation of the KL, and the initialization of the target. Figure 3 illustrates that the Kullback-Leibler divergence between $\tilde{\nu}_n$ and the $\mu^\star$ is indeed decreasing linearly with n. This result proves the validity of the rates derived in Theorem 7 for this specific setting. A full description of the experimental setup can be found in Appendix F.

5. Related work

In this section we discuss relevant related work.

Theoretical guarantees for Variational Inference. For the variational inference optimization problem in (1), frequently employed constraint sets $\mathcal{C}$ in existing literature encompass the set of non-degenerate Gaussian distributions, location-scale families, mixtures of Gaussian components, and the set of product measures. In the Gaussian setting, (Lambert et al., 2022; Diao et al., 2023) have been the first to leverage the geometry of Wasserstein gradient flows to study the convergence properties of variational inference, and provide convergence rates when the target $\mu^* \propto e^{-V}$ has a smooth and strongly convex potential V. In Mean-Field Variational inference (MFVI), the space $\mathcal{C}$ in (1) is taken to be the class of product measures over $\mathbb{R}^d$, written $\mathcal{P}(\mathbb{R})^{\otimes d}$. Several works have proposed algorithms in this context via wasserstein gradient flows (Yao & Yang, 2022; Lacker, 2023). (Jiang et al., 2023) consider a smaller subset of $\mathcal{C}$, namely a polyhedral subset for which they can derive optimization and approximation guarantees. However the previous work do not tackle mixture of Gaussians for the variational family.

Mollified Relative entropies. A closely related line of work to this paper is the one of (Carrillo et al., 2019; Craig et al., 2023a;b; Carrillo et al., 2024) that study Wasserstein gradient flows of mollified relative entropies and the associated particle systems, that are of particular interest in the literature of partial differential equations and kinetic theory. In (Carrillo et al., 2019), the authors mention the molli-

fied negative entropy $\mathcal{U}_\epsilon$ that we define in (2) as a regularization of the negative entropy $\mathcal{U}(\mu) = \int \log(\mu) d\mu$ (or entropy of order 1), but they do not study it. Instead, they focus on a closely related functional, defined as $\tilde{\mathcal{U}}_\epsilon(\mu) = \int \log(k_\epsilon \star \mu) d\mu$ (i.e. with only one convolution inside the logarithm, while $\mathcal{U}_\epsilon$ involves two convolutions). While they mention the possible choice of $\mathcal{U}_\epsilon$ as a regularization of the entropy $\mathcal{U}$, they choose to study the alternative regularization $\tilde{\mathcal{U}}_\epsilon(\mu)$ for numerical reasons, as the Wasserstein gradient of the latter functional writes as an integral over the distribution of the particles, while the one of $\mathcal{U}_\epsilon$ (hence $\mathcal{F}_\epsilon$) writes as an integral over the whole space w.r.t. Lebesgue measure, as explained in Section 2.1. Hence their results on λ-convexity[1] of the functional or the particle system differ from our setting. (Craig et al., 2023a) focus on a mollified chi-square divergence that corresponds to a weighted second order entropy; (Li et al., 2022) studies another mollified approximation of the chi-square divergence. (Carrillo et al., 2024) also study λ-convexity of entropies but only for entropies of order strictly greater than 1. Finally (Craig et al., 2023b) study functionals of the form $\int f_\epsilon(k_\epsilon \star \mu) d\mathcal{L}$ as approximations of $\int f(\mu) d\mathcal{L}$ where $\mathcal{L}$ denotes the Lebesgue measure. In their case f_ϵ is a specific function depending on f and ϵ, which excludes $\mathcal{U}_\epsilon$ and thus also differs from our setting.

VI on mixtures. Several works have tackled VI on mixtures on a computational aspect. Gershman et al. (2012) optimize (with L-BFGS, that is a quasi Newton method) an approximate ELBO (recall that the ELBO is the reverse KL we consider up to an additive constant), using several consecutive approximations of ELBO terms for the case of mixture of Gaussians. In the end, their optimization objective differs a lot from the original KL objective from VI, that is a valid divergence between probability distributions - in contrast with their objective. Arenz et al. (2018) adopt an Expectation-Maximization (EM) approach. As noted in (Aubin-Frankowski et al., 2022; Kunstner et al., 2021) EM can be seen as mirror descent scheme on the KL. Also, this algorithm can be seen as an Euler discretization of the gradient flow of the KL in the Fisher-Rao geometry (Domingo-Enrich & Pooladian, 2023; Chopin et al., 2024). The parallel can be seen from eq (5) or (8) in (Arenz et al., 2018), that take a similar form as eq (6) in (Chopin et al., 2024), i.e. a geometric update on the distributions, i.e. that act directly on updating densities (in a "vertical" manner), equivalently weights. In contrast, we focus on gradient descent dynamics, that correspond to a time discretization of the KL gradient flow in the Wasserstein geometry. This correspond to "horizontal" updates, where particles are displaced at each iteration. Lin et al. (2019) use natural gradient updates for VI in the natural parameter space (e.g. means for Gaussians). However, from (Raskutti & Mukherjee, 2015; Kunstner et al., 2021), it is known that this is equivalent to mirror descent on the exponential family parameters, which again is related to Fisher-Rao dynamics on the space of probability distributions (see eq (13) in (Chopin et al., 2024)).

[1] λ-convexity for $\lambda \geq 0$ recovers displacement convexity, while $\lambda \leq 0$ recovers smoothness.

6. Conclusion

The goal of this paper is to improve our theoretical understanding of variational inference algorithms in the non-Gaussian case. We consider here a specific family of distributions, a mixture of Gaussians with constant covariance and equally weighted components, that enables us to derive novel results for the approximation and optimization error for Variational Inference. We derive theoretical guarantees regarding gradient descent of the objective (i.e. a descent lemma proving that the objective decreases at each iteration) leveraging smoothness of the objective and the Wasserstein geometry. We also derive novel approximation results for minimizers of the objective.

In our study, we chose to simplify our exploration of Variational Inference (VI) within the context of Gaussian Mixtures by assuming uniform weights for each Gaussian component and by fixing the covariances. Extending our findings to more complex scenarios, where the weights of each Gaussian are dynamically optimized and the covariances are variable, represents a significant challenge that goes beyond the scope of our current research. For instance, the task of optimizing the weights attached to each Gaussian component introduces a shift from the Wasserstein dynamics, which are central to our current discussion, to Fisher-Rao dynamics. Achieving a counterpart to our current optimization result Proposition 4 under these conditions would not only require the adoption of alternative proof techniques but also a deep dive into the intricacies of Fisher-Rao dynamics, which diverge significantly from those of Wasserstein. Furthermore, the optimization of covariance matrices introduces another level of complexity. Such an endeavor requires a unique analytical framework, primarily due to the constraints imposed by the requirement that these matrices be positive definite. While this aspect of the analysis is crucial for a comprehensive understanding of VI in Gaussian mixtures, it requires a specialized approach that our current methodology does not cover. The exploration of dynamic weight optimization and variable covariance matrices within the context of Gaussian Mixtures in VI presents a rich avenue for future work.

Impact statement

Variational inference is a crucial tool for modern and large-scale Bayesian inference, as it can approximate complex

posterior distributions in a computationally efficient manner. However, its theoretical properties are poorly understood outside the setting where the variational family is a set of Gaussians. Studying the theoretical foundations of variational inference is essential for understanding the method's limitations, and guiding users in making informed choices about model assumptions and optimization strategies.

Acknowledgements

The authors are grateful for Irène Waldspurger, José Carrillo and Marc Lambert for helpful discussions. A.K and E.M acknowledge the support of ANR Chaire AI, SCAI project. T.H acknowledge the support of ANR-CHIA-002, "Statistics, computation and Artificial Intelligence". Part of the work has been developed under the auspice of the Lagrange Center for Mathematics and Calculus. This work was granted access to the HPC resources of IDRIS under the allocation AD011013313R2 made by GENCI (Grand Equipement National de Calcul Intensif).

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum mean discrepancy gradient flow. *Advances in neural information processing systems*, 2019.

Arenz, O., Neumann, G., and Zhong, M. Efficient gradient-free variational inference using policy search. In *International conference on machine learning*, pp. 234–243. PMLR, 2018.

Aubin-Frankowski, P.-C., Korba, A., and Léger, F. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.

Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, S. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pp. 2896–2923. PMLR, 2022.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Carrillo, J. A., Craig, K., and Patacchini, F. S. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):1–53, 2019.

Carrillo, J. A., Esposito, A., and Wu, J. S.-H. Nonlocal approximation of nonlinear diffusion equations. *Calculus of Variations and Partial Differential Equations*, 63(4): 100, 2024.

Chopin, N., Crucinio, F. R., and Korba, A. A connection between tempering and entropic mirror descent. *International Conference of Machine Learning*, 2024.

Craig, K., Elamvazhuthi, K., Haberland, M., and Turanova, O. A blob method for inhomogeneous diffusion with applications to multi-agent control and sampling. *Mathematics of Computation*, 2023a.

Craig, K., Jacobs, M., and Turanova, O. Nonlocal approximation of slow and fast diffusion. *arXiv preprint arXiv:2312.11438*, 2023b.

Delon, J. and Desolneux, A. A Wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Diao, M. Z., Balasubramanian, K., Chewi, S., and Salim, A. Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein space. In *International Conference on Machine Learning*, pp. 7960–7991. PMLR, 2023.

Domingo-Enrich, C. and Pooladian, A.-A. An explicit expansion of the Kullback-Leibler divergence along its Fisher-Rao gradient flow. *Transactions of Machine Learning Research*, 2023.

Domke, J., Garrigos, G., and Gower, R. Provable convergence guarantees for black-box variational inference. *Advances in neural information processing systems*, 2023.

Duncan, A., Nüsken, N., and Szpruch, L. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 2023.

Friedrichs, K. O. The identity of weak and strong extensions of differential operators. *Transactions of the American Mathematical Society*, 55:132–151, 1944.

Garrigos, G. and Gower, R. M. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Gershman, S., Hoffman, M., and Blei, D. Nonparametric variational inference. *International Conference of Machine Learning*, 2012.

Helin, T. and Kretschmann, R. Non-asymptotic error estimates for the laplace approximation in bayesian inverse problems. *Numerische Mathematik*, 150(2):521–549, 2022.

Jiang, Y., Chewi, S., and Pooladian, A.-A. Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *arXiv preprint arXiv:2312.02849*, 2023.

Katsevich, A. and Rigollet, P. On the approximation accuracy of gaussian variational inference. *arXiv preprint arXiv:2301.02168*, 2023.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:4672–4682, 2020.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. Kernel Stein discrepancy descent. In *International Conference on Machine Learning*, pp. 5719–5730. PMLR, 2021.

Kunstner, F., Kumar, R., and Schmidt, M. Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 3295–3303. PMLR, 2021.

Lacker, D. Independent projections of diffusions: Gradient flows for variational inference and optimal mean field approximations. *arXiv preprint arXiv:2309.13332*, 2023.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35: 14434–14447, 2022.

Łatuszyński, K., Miasojedow, B., and Niemiro, W. Nonasymptotic bounds on the estimation error of mcmc algorithms. *Bernoulli*, 19(5A):2033–2066, 2013.

Li, J. and Barron, A. Mixture density estimation. *Advances in neural information processing systems*, 12, 1999.

Li, L., Liu, Q., Korba, A., Yurochkin, M., and Solomon, J. Sampling with mollified interaction energy descent. *arXiv preprint arXiv:2210.13400*, 2022.

Lin, W., Khan, M. E., and Schmidt, M. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *International Conference on Machine Learning*, pp. 3992–4002. PMLR, 2019.

Moins, T., Arbel, J., Dutfoy, A., and Girard, S. On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic. *Bayesian Analysis*, 1(1):1–26, 2023.

Moulines, E. and Bach, F. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Taylor & Francis*, 2001.

Pavliotis, G. A. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

Raskutti, G. and Mukherjee, S. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

Roberts, G. O. and Rosenthal, J. S. General state space Markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

Salim, A., Korba, A., and Luise, G. The Wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366, 2020.

Santambrogio, F. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

Topsøe, F. Some bounds for the logarithmic function. *Inequality theory and applications*, 4:137, 2007.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Wibisono, A. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pp. 2093–3027. PMLR, 2018.

Yao, R. and Yang, Y. Mean field variational inference via Wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*, 2022.

Yi, M. and Liu, S. Bridging the gap between variational inference and wasserstein gradient flows. *arXiv preprint arXiv:2310.20090*, 2023.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

The appendix is organized as follows. Appendix A details the computations for the Wasserstein gradients and the particle scheme corresponding to the optimization of the mollified relative entropy. Appendix B discusses the connection with the algorithm and framework presented in (Lambert et al., 2022). Appendix C contains the intermediate lemmas needed for the proof of Theorem 7. Appendix E contains the proofs of the Propositions regarding Wasserstein Hessians. Appendix F outlines the setup used for the numerical experiments.

A. Particle implementation of the gradient flow

A solution of (3) is implemented by the Mac-Kean Vlasov process:

$$\dot{m}_t = -(\nabla \mathcal{U}'_\epsilon(\mu_t)(m_t) + \nabla \mathcal{G}'_{V_\epsilon}(\mu_t)(m_t)). \tag{18}$$

Here we detail the computation of the vector field in (18) and its particle implementation.

For the negative entropy part, we can rewrite $\mathcal{U}_\epsilon(\mu) = \int U(k_\epsilon \star \mu(\theta))d\theta$ where $U : x \mapsto x\log(x)$. We have that

$$\mathcal{U}'_\epsilon(\mu)(\cdot) = k_\epsilon \star (U' \circ (k_\epsilon \star \mu))(\cdot) = \int_{\mathbb{R}^d} k_\epsilon(\theta - \cdot)U'\left(\int k_\epsilon(\theta - y)d\mu(y)\right)d\theta$$

where $U' : x \mapsto \log(x) + 1$. Hence, computing $\mathcal{U}'_\epsilon$ requires an integration over $\mathbb{R}^d$. Then, we have, since k_ϵ is smooth, using an integration by parts with $\nabla U'(x) = \nabla \log(x)$ and symmetry of k_ϵ, for any $w \in \mathbb{R}^d$ we have:

$$\begin{aligned}
\nabla_w \mathcal{U}'_\epsilon(\mu)(w) = \nabla_w k_\epsilon \star (U' \circ (k_\epsilon \star \mu))(w) &= \int_{\mathbb{R}^d} \nabla_w k_\epsilon(\theta - w)U'\left(\int k_\epsilon(\theta - y)d\mu(y)\right)d\theta \\
&= -\int_{\mathbb{R}^d} \nabla_\theta k_\epsilon(\theta - w)U'\left(\int k_\epsilon(\theta - y)d\mu(y)\right)d\theta \\
&= +\int_{\mathbb{R}^d} k_\epsilon(\theta - w)\nabla_\theta U'\left(\int k_\epsilon(\theta - y)d\mu(y)\right)d\theta \\
&= \int_{\mathbb{R}^d} k_\epsilon(\theta - w)\nabla_\theta \log\left(\int k_\epsilon(\theta - y)d\mu(y)\right)d\theta \\
&= \int_{\mathbb{R}^d} k_\epsilon(\theta - w)\frac{\int \nabla k_\epsilon(\theta - y)d\mu(y)}{\int k_\epsilon(\theta - y)d\mu(y)}d\theta.
\end{aligned} \tag{19}$$

Finally, if μ_t is an atomic measure of the form $\mu_t = \frac{1}{N}\sum_{i=1}^N \delta_{m_t^{(i)}}$, then a particle implementation of (21) reduces to solving a system of ordinary differential equations for the locations of the Dirac masses:

$$\dot{m}_t^{(j)} = -\int_{\mathbb{R}^d} \nabla V(y)k_\epsilon(y - m_t^{(j)})dy - \int_{\mathbb{R}^d} \frac{\sum_{i=1}^N \nabla k_\epsilon(y - m_t^{(i)})}{\sum_{i=1}^N k_\epsilon(y - m_t^{(i)})}k_\epsilon(y - m_t^{(j)})dy. \tag{20}$$

For the potential energy part, we can rewrite

$$\begin{aligned}
\mathcal{G}_{V_\epsilon}(\mu) = \int_{\mathbb{R}^d} V(\theta)d(k_\epsilon \star \mu)(\theta) = \int_{\mathbb{R}^d} V(\theta)\int k_\epsilon(\theta - m)d\mu(m)d\theta \\
= \iint_{\mathbb{R}^d} k_\epsilon(\theta - m)V(\theta)d\theta d\mu(m) := \int_{\mathbb{R}^d} V_\epsilon(m)d\mu(m),
\end{aligned}$$

where $V_\epsilon(m) = \int_{\mathbb{R}^d} k_\epsilon(\theta - m)V(\theta)d\theta = k_\epsilon \star V(m)$. Hence, we have successively for any $w \in \mathbb{R}^d$

$$\begin{aligned}
\mathcal{G}'_{V_\epsilon}(\mu)(w) &= V_\epsilon(w), \\
\nabla_w \mathcal{G}'_{V_\epsilon}(\mu)(w) = \nabla_w V_\epsilon(w) &= \int_{\mathbb{R}^d} \nabla_w k_\epsilon(\theta - w)V(\theta)d\theta = -\int_{\mathbb{R}^d} \nabla_\theta k_\epsilon(\theta - w)V(\theta)d\theta = \int_{\mathbb{R}^d} k_\epsilon(\theta - \cdot)\nabla V(\theta)d\theta
\end{aligned}$$

using again an integration by parts. Hence, (18) becomes:

$$\dot{m}_t = -\int_{\mathbb{R}^d} k_\epsilon(\theta - m_t)\nabla V(\theta)d\theta - \int_{\mathbb{R}^d} k_\epsilon(\theta - m_t)\frac{\int \nabla k_\epsilon(\theta - y)d\mu_t(y)}{\int k_\epsilon(\theta - y)d\mu_t(y)}d\theta. \tag{21}$$

B. Mixture of Gaussians optimization

Lambert et al. (2022) consider a Gaussian approximation of the Langevin diffusion given by Saarka's heuristic, i.e. $X_t \sim \mu_t$ where μ_t is the solution of Fokker-Planck equation is replaced by $Y_t \sim \mathcal{N}(m_t, \Sigma_t)$ where

$$\begin{aligned}\dot{m}_t &= -\mathbb{E}[\nabla V(Y_t)] \\ \dot{\Sigma}_t &= 2\,\mathrm{Id} - \mathbb{E}[\nabla V(Y_t) \otimes (Y_t - m_t) + (Y_t - m_t) \otimes \nabla V(Y_t)]\end{aligned}$$

They prove that the law of Y_t is the gradient flow of the KL on the Bures-Wasserstein manifold $\mathrm{BW}(\mathbb{R}^d) \cong \mathbb{R}^d \times S_d^{++}$ (the space of Gaussians equipped with the W_2 distance); which is a submanifold of $\mathcal{P}_2(\mathbb{R}^d)$. It can be seen as "Projected WGF" where the Wasserstein gradient of the KL is projected onto the tangent space of the submanifold; another way to view it is to see that its the GF of the KL on the Bures-Wasserstein manifold.

Then, they propose to write a Gaussian mixture ρ on $\mathbb{R}^d$ as $\rho_\nu(\theta) = \int_{\mathrm{BW}(\mathbb{R}^d)} p(\theta) d\nu(p)$[2] where ν is a measure over $\mathrm{BW}(\mathbb{R}^d)$; hence MOG is isomorphic to $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$. Then the WGF of $\nu \mapsto \mathrm{KL}(\rho_\nu|\mu^\star)$, ie the GF of this functional over $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$ is implemented through a particle system $\nu_t = \frac{1}{N}\sum_{i=1}^N \delta_{(m_t^{(i)}, \Sigma_t^{(i)})}$:

$$\dot{m}_t^{(i)} = -\mathbb{E}\left[\nabla \log\left(\frac{\rho_{\nu_t}}{\mu^\star}\right)\left(Y_t^{(i)}\right)\right] \tag{22}$$

$$\dot{\Sigma}_t^{(i)} = -\mathbb{E}\left[\nabla^2 \log\left(\frac{\rho_{\nu_t}}{\mu^\star}\right)\left(Y_t^{(i)}\right)\right]\Sigma_t^{(i)} - \Sigma_t^{(i)}\mathbb{E}\left[\nabla^2 \log\left(\frac{\rho_{\nu_t}}{\mu^\star}\right)\left(Y_t^{(i)}\right)\right] \tag{23}$$

where $Y_t^{(i)} \sim \mathcal{N}(m_t^{(i)}, \Sigma_t^{(i)})$.

In contrast, in this work we restrict ourselves to Gaussian mixtures ρ that write $\rho_\mu = \int_{\mathbb{R}^d} k_\epsilon(\theta - y) d\mu(y)$ where μ is a measure over $\mathbb{R}^d$. Then the WGF of $\mu \mapsto \mathrm{KL}(\rho_\mu|\pi)$, i.e. the GF of this functional over $\mathcal{P}_2(\mathbb{R}^d)$ is equivalent to the update above from (Lambert et al., 2022). Indeed if we fix $\nu = \mu \otimes \delta_{\epsilon\,\mathrm{Id}}$, a Gaussian mixture writes $\rho_\nu(\theta) = \rho_\mu(\theta) = \int_{\mathbb{R}^d} k_\epsilon(\theta - y) d\mu(y)$. In this case we consider the particle system $\mu_t = \frac{1}{N}\sum_{i=1}^N \delta_{m_t^{(i)}}$, we have $\rho_{\mu_t}(\theta) = \frac{1}{N}\sum_{i=1}^N k_\epsilon(\theta - m_t^{(i)})$. The update (22) becomes:

$$\begin{aligned}\dot{m}_t^{(j)} &= -\mathbb{E}\left[\nabla \log\left(\frac{\rho_{\nu_t}}{\pi}\right)\left(Y_t^{(j)}\right)\right] \\ &= -\mathbb{E}[\nabla V(Y_t^{(j)})] - \mathbb{E}\left[\nabla \log\left(\rho_{\mu_t}\right)\left(Y_t^{(j)}\right)\right] \\ &= -\mathbb{E}[\nabla V(Y_t^{(j)})] - \mathbb{E}\left[\frac{\sum_{i=1}^N \nabla k_\epsilon(Y_t^{(j)} - m_t^{(i)})}{\sum_{i=1}^N k_\epsilon(Y_t^{(j)} - m_t^{(i)})}\right] \\ &= -\int \nabla V(y) k_\epsilon(y - m_t^{(j)}) dy - \int \frac{\sum_{i=1}^N \nabla k_\epsilon(y - m_t^{(i)})}{\sum_{i=1}^N k_\epsilon(y - m_t^{(i)})} k_\epsilon(y - m_t^{(j)}) dy\end{aligned}$$

since $Y_t^{(j)} \sim \mathcal{N}(m_t^{(j)}, \epsilon\,\mathrm{Id})$ has density $k_\epsilon(\cdot - m_t^{(j)})$ hence we obtain the same update as (20).

C. Lemmas for the proof of Theorem 7

Lemma 8. *For any $x \in \mathbb{R}^+$, the function defined on $[0, +\infty[$ by*

$$B(x) = \frac{x \log x - x + 1}{(x-1)^2} \text{ if } x > 0,$$

and $B(0) = 1$ is monotone decreasing in r.

[2] We can rewrite it as $\int_{\mathbb{R}^d \times S_d^{++}} p_{y,\Sigma}(\theta) d\nu(y, \Sigma)$

Proof. Firstly, the derivative of f is given by

$$B'(x) = \frac{2 - \frac{x+1}{x-1}\log(x)}{(x-1)^2}.$$

Recall some inequalities of the log function derived in (Topsøe, 2007):

$$\forall x \in [1, +\infty[\ , \quad \frac{2(x-1)}{x+1} \leq \log(x),$$
$$\forall x \in [0, 1]\ , \quad \log(x) \leq \frac{2(x-1)}{x+1}.$$

Consequently, combining those two inequalities and multiplying by $1/(x-1)$ which is positive on $[1, +\infty[$ and negative on $[0, 1[$ we obtain for any $x \in [0, +\infty[$

$$\frac{\log(x)}{x-1} \geq 2(x-1)$$

It implies that the derivative $f'(x)$ is always negative and f is monotone decreasing. □

Lemma 9. *For any $x \in \mathbb{R}^+$ we have*

$$C(x) = B(x)(x-1) = \frac{x\log(x) - x + 1}{x-1} \leq \sqrt{x} - 1$$

Proof. Recall the inequalities derived in (Topsøe, 2007)

1. for any $x \in [1, +\infty[$, $\log(x) \leq \frac{x-1}{\sqrt{x}}$
2. for any $x \in [0, 1]$, $\log(x) \geq \frac{x-1}{\sqrt{x}}$.

Combining those inequalities and multiplying by $1/(x-1)$ which is positive on $[1, +\infty[$ and negative on $[0, 1[$, we obtain for any $x \in [0, \infty[$,

$$\frac{\log(x)}{x-1} \leq \frac{1}{\sqrt{x}}.$$

Moreover,

$$C(x) - \sqrt{x} - 1 = \frac{x\log(x)}{x-1} - \sqrt{x}.$$

Consequently, by multiplying the previous inequality by x, we obtain that $C(x) - (\sqrt{x} + 1) \leq 0$ □

Lemma 10. *For any $\alpha \in [0, 1]$, we have*

$$-\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha\sqrt{1-\alpha} \leq 0.$$

Proof. Let's start by applying the classical inequality $\forall x > -1$, $\log(1+x) \leq x$ at $x = -\alpha$, we obtain $\log(1-\alpha) \leq -\alpha$. Hence,

$$\begin{aligned} -\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha\sqrt{1-\alpha} &\leq -\alpha - \alpha(1-\alpha) + 2\alpha\sqrt{1-\alpha} \\ &= \alpha(2\sqrt{1-\alpha} - 2 + \alpha) \\ &:= \alpha g(\alpha) \end{aligned}$$

Moreover,

$$g(\alpha) = 2\sqrt{1-\alpha} - 2 + \alpha \text{ and } g'(\alpha) = \frac{-1}{\sqrt{1-\alpha}} - 1 \leq 0,$$

hence g is decreasing. Consequently,

$$-\alpha + (1-\alpha)\log(1-\alpha) + 2\alpha\sqrt{1-\alpha} \leq \alpha g(0) \leq 0$$ □.

D. Proof of Proposition 4

We first deal with the potential energy term. Notice that under Assumption 1, V_ϵ is also L-smooth, since for any $x, y \in \mathbb{R}^d$

$$\|\nabla V_\epsilon(x) - \nabla V_\epsilon(y)\| \leq \int k_\epsilon(\theta)\|\nabla V(x-\theta) - \nabla V(y-\theta)\| d\theta \leq L\|x-y\| \tag{24}$$

since $\int k_\epsilon(\theta)d\theta = 1$. Hence we have the following.

Proposition 11. *Let ρ in $\mathcal{P}_2(\mathbb{R}^d)$ and $\rho_t = T_{t\#}\rho$ where $T_t = \mathrm{Id} + t\phi$.*

$$\frac{d}{dt}\mathcal{G}_{V_\epsilon}(\rho_t) - \frac{d}{dt}\mathcal{G}_{V_\epsilon}(\rho_t)\Big|_{t=0} \leq Lt\|\phi\|^2_{L^2(\rho)}.$$

Proof. By the chain rule in Wasserstein space we have

$$\frac{d}{dt}\mathcal{G}_{V_\epsilon}(\rho_t) = \langle \nabla \mathcal{G}'_{V_\epsilon}(\rho_t), v_t\rangle_{L^2(\rho_t)} = \langle \nabla V_\epsilon, v_t\rangle_{L^2(\rho_t)}.$$

Hence, using the transfer Lemma and Cauchy-Schwarz successively,

$$\begin{aligned}\frac{d}{dt}\mathcal{G}_{V_\epsilon}(\rho_t) - \frac{d}{dt}\mathcal{G}_{V_\epsilon}(\rho_t)\Big|_{t=0} =& \langle \nabla V_\epsilon, v_t\rangle_{L^2(\rho_t)} - \langle \nabla V_\epsilon, \phi\rangle_{L^2(\rho)} = \langle \nabla V_\epsilon \circ T_t - \nabla V_\epsilon, \phi\rangle_{L^2(\rho)} \\ &\leq \mathbb{E}_{w\sim\rho}[L\|T_t(x) - x\|\|\phi(w)\|] \leq Lt\|\phi\|^2_{L^2(\rho)}\,. \end{aligned}$$ □

We now turn to the mollified entropy term that is the most challenging.

Proposition 12. *Let ρ denote a mixture of n Diracs and $\rho_t = T_{t\#}\rho$ where $T_t = \mathrm{Id} + t\phi$. We have:*

$$\frac{d}{dt}\mathcal{U}_\epsilon(\rho_t) - \frac{d}{dt}\mathcal{U}_\epsilon(\rho_t)\Big|_{t=0} \leq \left(\frac{1}{\epsilon^2} + \frac{\sqrt{m_2(\rho)n}}{\epsilon^3} + \frac{\sqrt{n}}{\epsilon^2} + \frac{n\sqrt{m_2(\rho)}}{2\epsilon^3}\right) t\|\phi\|^2_{L^2(\rho)}$$

where $m_2(\rho)$ denotes the second moment of ρ.

Proof. By the chain rule in Wasserstein space, we have

$$\frac{d}{dt}\mathcal{U}_\epsilon(\rho_t) = \langle \nabla \mathcal{U}'_\epsilon(\rho_t), v_t\rangle_{L^2(\rho_t)}.$$

Consequently,

$$\begin{aligned}\frac{d}{dt}\mathcal{U}_\epsilon(\rho_t) - \frac{d}{dt}\mathcal{U}_\epsilon(\rho_t)\Big|_{t=0} =& \langle \nabla \mathcal{U}'_\epsilon(\rho_t), v_t\rangle_{L^2(\rho_t)} - \langle \nabla \mathcal{U}'_\epsilon(\rho), \phi\rangle_{L^2(\rho)} \\ &= \langle \nabla \mathcal{U}'_\epsilon(T_t\#\rho) \circ T_t - \nabla \mathcal{U}'_\epsilon(\rho), \phi\rangle_{L^2(\rho)} \\ &\leq \mathbb{E}_{w\sim\rho}[\|\nabla \mathcal{U}'_\epsilon(T_t\#\rho)(T_t(w)) - \nabla \mathcal{U}'_\epsilon(\rho)(w)\|\|\phi(w)\|] \end{aligned} \tag{25}$$

where in the second line we have used the transfer Lemma and in the last inequality Cauchy-Schwarz. Now, let's focus on the term $\|\nabla \mathcal{U}'_\epsilon(T_t\#\rho)(T_t(w)) - \nabla \mathcal{U}'_\epsilon(\rho)(w)\|$, that we will decompose as

$$\begin{aligned}\nabla \mathcal{U}'_\epsilon(T_t\#\rho) \circ T_t - \nabla \mathcal{U}'_\epsilon(\rho) &= \nabla \mathcal{U}'_\epsilon(T_t\#\rho)(T_t(w)) - \nabla \mathcal{U}'_\epsilon(\rho)(T_t(w)) + \nabla \mathcal{U}'_\epsilon(\rho)(T_t(w)) - \nabla \mathcal{U}'_\epsilon(\rho)(w) \\ &:= \mathcal{B}_{T_t(w)}(\rho_t, \rho) + \mathcal{A}_\rho(T_t(w), w). \end{aligned} \tag{26}$$

In the rest of the proof, we will show the Lipschitzness on w for $\mathcal{A}$ and in ρ for $\mathcal{B}$.

Using Proposition 13 and Proposition 14, we have

$$\begin{aligned}\frac{d}{dt}\mathcal{U}_\epsilon(\rho_t) - \frac{d}{dt}\mathcal{U}_\epsilon(\rho_t)\Big|_{t=0} &\leq \mathbb{E}_{w\sim\rho}\left[(\|\mathcal{A}_\rho(T_t(w), w)\| + \|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\|)\|\phi(w)\|\right] \\ &\leq \left(\frac{1}{\epsilon^2} + \frac{\sqrt{m_2(\rho)n}}{2\epsilon^3}\right) t\|\phi\|^2_{L^2(\rho)} + \left(\frac{\sqrt{n}}{\epsilon^2} + \frac{\sqrt{nm_2(\rho)}}{2\epsilon^3} + \frac{n\sqrt{m_2(\rho)}}{2\epsilon^3}\right) t\|\phi\|^2_{L^2(\rho)} \\ &\leq \left(\frac{1}{\epsilon^2} + \frac{2\sqrt{m_2(\rho)n}}{\epsilon^3} + \frac{\sqrt{n}}{\epsilon^2} + \frac{n\sqrt{m_2(\rho)}}{2\epsilon^3}\right) t\|\phi\|^2_{L^2(\rho)}. \end{aligned}$$ □

Proposition 13. *Let ρ denote a mixture of n Diracs and $\rho_t = T_{t\#}\rho$ where $T_t = \mathrm{Id} + t\phi$. It holds that*

$$\|\mathcal{A}_\rho(T_t(w), w)\| \leq \left(\frac{1}{\epsilon^2} + \frac{\sqrt{2m_2(\rho)n}}{\epsilon^3}\right)\, t\|\phi(w)\| \tag{27}$$

Proof. Recalling the definition of $\nabla\mathcal{U}'_\epsilon$ in (19), we obtain

$$\nabla\mathcal{U}'_\epsilon(\rho)(w) = \int k_\epsilon(\theta - w)\frac{\int \nabla k_\epsilon(\theta - y)d\rho(y)}{\int k_\epsilon(\theta - y)d\rho(y)}d\theta = \frac{1}{\epsilon^2}\int k_\epsilon(\theta - w)\frac{\int y k_\epsilon(\theta - y)d\rho(y)}{\int k_\epsilon(\theta - y)d\rho(y)}d\theta - \frac{w}{\epsilon^2}.$$

Then from the definition of $\mathcal{A}$ in (26) we have

$$\begin{aligned}\|\mathcal{A}_\rho(T_t(w), w)\| &= \frac{1}{\epsilon^2}\left\|\int (k_\epsilon(\theta - T_t(w)) + k_\epsilon(\theta - w))\frac{\int y k_\epsilon(\theta - y)d\rho(y)}{k_\epsilon \star \rho(\theta)}d\theta - T_t(w) + w\right\| \\ &\leq \frac{1}{\epsilon^2}\int |k_\epsilon(\theta - T_t(w)) - k_\epsilon(\theta - w)|\frac{\int \|y\| k_\epsilon(\theta - y)d\rho(y)}{k_\epsilon \star \rho(\theta)}d\theta + \frac{t\|\phi(w)\|}{\epsilon^2}\end{aligned}$$

Moreover, recall that ρ is a mixture of n Diracs. Therefore, we have

$$\begin{aligned}\int \|y\| k_\epsilon(\theta - y)d\rho(y) &\leq \sqrt{\int \|y\|^2 d\rho(y)}\sqrt{\int k_\epsilon(\theta - y)^2 d\rho(y)} = \sqrt{m_2(\rho)}\sqrt{\frac{1}{n}\sum_{i=1}^n k_\epsilon(\theta - y_i)^2} \\ &\leq \sqrt{m_2(\rho)}\frac{1}{\sqrt{n}}\sum_{i=1}^n k_\epsilon(\theta - y_i) \leq \sqrt{m_2(\rho)n}\; k_\epsilon \star \rho(\theta). \end{aligned} \tag{28}$$

Consequently,

$$\begin{aligned}\|\mathcal{A}_\rho(T_t(w), w)\| &\leq \frac{\sqrt{m_2(\rho)n}}{\epsilon^2}\, 2\,\mathrm{TV}(\mathcal{N}(T_t(w), \epsilon^2\,\mathrm{Id}),\, \mathcal{N}(w, \epsilon^2\,\mathrm{Id}) + \frac{t\|\phi(w)\|}{\epsilon^2} \\ &\leq \frac{\sqrt{m_2(\rho)n}}{\epsilon^2}\,\sqrt{2\,\mathrm{KL}(\mathcal{N}(T_t(w), \epsilon^2\,\mathrm{Id}),\, \mathcal{N}(w, \epsilon^2\,\mathrm{Id})} + \frac{t\|\phi(w)\|}{\epsilon^2} \\ &= \left(\frac{1}{\epsilon^2} + \frac{\sqrt{2m_2(\rho)n}}{\epsilon^3}\right)\, t\|\phi(w)\|. \end{aligned}$$

□

Proposition 14. *Let ρ denote a mixture of n Diracs and $\rho_t = T_{t\#}\rho$ where $T_t = \mathrm{Id} + t\phi$. We have:*

$$\mathbb{E}_{w\sim\rho}[\|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\|\|\phi(w)\|] \leq \left(\frac{\sqrt{n}}{\epsilon^2} + \frac{\sqrt{nm_2(\rho)}}{\epsilon^3} + \frac{n\sqrt{m_2(\rho)}}{\epsilon^3}\right)\, t\|\phi\|^2_{L^2(\rho)}.$$

Proof. Recalling the definition of Equation (19), we have

$$\begin{aligned}\mathbb{E}_{w\sim\rho}[\|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\|\|\phi(w)\|] &= \int \left\|\frac{1}{\epsilon^2}\int k_\epsilon(\theta - T_t(w))\frac{\int y k_\epsilon(\theta - y)d\rho_t(y)}{k_\epsilon \star \rho_t(\theta)} - \frac{\int y k_\epsilon(\theta - y)d\rho(y)}{k_\epsilon \star \rho(\theta)}\right\| .\|\phi(w)\| d\theta d\rho(w) \\ &\leq \frac{1}{\epsilon^2}\int \Big(\int k_\epsilon(\theta - T_t(w))\|\phi(w)\| d\rho(w)\Big)\left\|\frac{\int y k_\epsilon(\theta - y)d\rho_t(y)}{k_\epsilon \star \rho_t(\theta)} - \frac{\int y k_\epsilon(\theta - y)d\rho(y)}{k_\epsilon \star \rho(\theta)}\right\| d\theta. \end{aligned} \tag{29}$$

Then, we can use Cauchy Schwarz inequality and that ρ_t is supported on n Diracs to obtain

$$\int \|\phi(w)\| k_\epsilon(\theta - T_t(w))d\rho(w) \leq \|\phi\|_{L^2(\rho)}\sqrt{\int k_\epsilon(\theta - w)^2 d\rho_t(w)} = \sqrt{n}\|\phi\|_{L^2(\rho)}\; k_\epsilon \star \rho_t(\theta). \tag{30}$$

Moreover, recall that

$$\begin{aligned}
&\Big\| \frac{\int y k_\epsilon(\theta - y) d\rho_t(y)}{k_\epsilon \star \rho_t(\theta)} - \frac{\int y k_\epsilon(\theta - y) d\rho(y)}{k_\epsilon \star \rho(\theta)} \Big\| \\
&\quad \leq \frac{\Big\| \int y k_\epsilon(\theta - y) d\rho_t(y) - \int y k_\epsilon(\theta - y) d\rho(y) \Big\|}{k_\epsilon \star \rho_t(\theta)} + \Big\| \int y k_\epsilon(\theta - y) d\rho(y) \Big\| \Big| \frac{1}{k_\epsilon \star \rho_t(\theta)} - \frac{1}{k_\epsilon \star \rho(\theta)} \Big| \\
&\quad \leq \frac{\int \|T_t(y) k_\epsilon(\theta - T_t(y)) - y k_\epsilon(\theta - y)\| d\rho(y)}{k_\epsilon \star \rho_t(\theta)} + \int \|y\| k_\epsilon(\theta - y) d\rho(y) \Big| \frac{k_\epsilon \star \rho(\theta) - k_\epsilon \star \rho_t(\theta)}{k_\epsilon \star \rho(\theta)\ k_\epsilon \star \rho_t(\theta)} \Big| \\
&\quad := \mathcal{C}_1(\theta) + \mathcal{C}_2(\theta). \quad (31)
\end{aligned}$$

We can now combine inequalities 29, 30 and 31 to obtain

$$\mathbb{E}_{w\sim\rho}[\|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\| \|\phi(w)\|] \leq \frac{\sqrt{n}\|\phi\|_{L^2(\rho)}}{\epsilon^2} \int k_\epsilon \star \rho_t(\theta)(\mathcal{C}_1(\theta) + \mathcal{C}_2(\theta)) d\theta.$$

We first focus on the $\mathcal{C}_1$ term:

$$\begin{aligned}
\int k_\epsilon \star \rho_t(\theta) \mathcal{C}_1(\theta) d\theta &= \int \int \|T_t(y) k_\epsilon(\theta - T_t(y)) - y k_\epsilon(\theta - y)\| d\rho(y) d\theta \\
&\leq \int \int \|T_t(y) - y\| k_\epsilon(\theta - T_t(y)) + \|y\| \ |k_\epsilon(\theta - T_t(y)) - k_\epsilon(\theta - y)| d\rho(y) d\theta \\
&\leq t\mathbb{E}_{y\sim\rho}[\|\phi(y)\|] + \int \|y\| 2\,\mathrm{TV}(\mathcal{N}(T_t(y), \epsilon^2\,\mathrm{Id}), \mathcal{N}(y, \epsilon^2\,\mathrm{Id})) d\rho(y) \\
&\leq t\|\phi\|_{L^2(\rho)} + \int \frac{t\|y\|\ \|\phi(y)\|}{\epsilon} d\rho(y) \\
&\leq t\|\phi\|_{L^2(\rho)} + \frac{t\|\phi\|_{L^2(\rho)}}{2\epsilon} \sqrt{\int \|y\|^2 d\rho(y)} \\
&= \left(1 + \frac{\sqrt{m_2(\rho)}}{\epsilon}\right)\ t\|\phi\|_{L^2(\rho)}.
\end{aligned}$$

Finally, we focus on the $\mathcal{C}_2$ term. We obtain using the same computations as in (28):

$$\begin{aligned}
\int k_\epsilon \star \rho_t(\theta) \mathcal{C}_2(\theta) d\theta &= \int \frac{\int \|y\| k_\epsilon(\theta - y) d\rho(y)}{k_\epsilon \star \rho(\theta)} \ |k_\epsilon \star \rho(\theta) - k_\epsilon \star \rho_t(\theta)| d\theta \\
&\leq \sqrt{m_2(\rho) n} \int |k_\epsilon \star \rho(\theta) - k_\epsilon \star \rho_t(\theta)| d\theta \\
&\leq \sqrt{m_2(\rho) n} \int \int |k_\epsilon(\theta - y) - k_\epsilon(\theta - T_t(y))| d\rho(y) d\theta \\
&\leq \sqrt{m_2(\rho) n} \int \frac{t\|\phi\|_{L^2(\rho)}}{\epsilon} d\rho(y) d\theta \\
&\leq \frac{t\sqrt{m_2(\rho) n}}{\epsilon} \|\phi\|_{L^2(\rho)}.
\end{aligned}$$

Combining the previous inequalities, we obtain

$$\mathbb{E}_{w\sim\rho}[\|\mathcal{B}_{T_t(w)}(\rho_t, \rho)\| \|\phi(w)\|] \leq \left(\frac{\sqrt{n}}{\epsilon^2} + \frac{\sqrt{n m_2(\rho)}}{\epsilon^3} + \frac{n\sqrt{m_2(\rho)}}{\epsilon^3} \right)\ t\|\phi\|^2_{L^2(\rho)}. \qquad \square$$

E. Wasserstein Hessians of relative entropies

E.1. Proof of Proposition 2

Proof. Let $\mu_t = (\mathrm{Id} + t\nabla\psi)_\# \mu$ where $\psi \in C_c^\infty(\mathbb{R}^d)$. Let μ_t, μ^* be the densities of μ_t and μ^* respectively. We denote by $\phi_t = \mathrm{Id} + tg$ where $g = \nabla\psi$. Hence we have $\mathrm{J}\phi_t = \mathrm{Id} + t\mathrm{J}g$. Time derivatives are denoted as $\phi_t' = \frac{d\phi_t}{dt}$. Notice that $(\mathrm{J}\phi_t)' = \mathrm{J}\phi_t' = \mathrm{J}g = \mathrm{Hess}\,\psi$.

For any f-divergence,

$$h_\mu(t) = \int f\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\mu^*(x)\mathrm{d}x = \int \tilde{f}\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\mu_t(x)\mathrm{d}x.$$

where $\tilde{f}(t) = f(t)/t$. By the transfer lemma and change of variables formula, we have

$$h_\mu(t) = \int \tilde{f}\left(\frac{\mu(x)}{\mu^*(\phi_t(x))|\mathrm{J}\phi_t(x)|}\right)\mathrm{d}\mu(x).$$

Let us rewrite

$$h_\mu(t) = \int \tilde{f}\left(\mu(x)e^{n_t(x)}\right)\mathrm{d}\mu(x), \quad \text{where} \quad n_t(x) = V(\phi_t(x)) - \log|\mathrm{J}\phi_t(x)|.$$

We have consecutively:

$$\begin{aligned}
h'_\mu(t) &= \int \tilde{f}'\left(\mu(x)e^{n_t(x)}\right)\mu(x)n'_t(x)e^{n_t(x)}\mathrm{d}\mu(x)\\
h''_\mu(t) &= \int \Big[\tilde{f}''(\mu(x)e^{n_t(x)})\left(\mu(x)n'_t(x)e^{n_t(x)}\right)^2\\
&\qquad \tilde{f}'(\mu(x)e^{n_t(x)})\left(n''_t(x) + n'_t(x)^2\right)\mu(x)e^{n_t(x)}\Big]\mathrm{d}x
\end{aligned}$$

where

$$\begin{aligned}
n'_t(x) &= \langle \nabla V(\phi_t(x)), \phi'_t(x)\rangle - \mathrm{Tr}\left((\mathrm{J}\phi_t(x))^{-1}\mathrm{J}\phi'_t(x)\right),\\
n''_t(x) &= \langle \mathrm{H}_V(\phi_t(x))\phi'_t(x), \phi'_t(x)\rangle + \mathrm{Tr}\left((\mathrm{J}\phi_t(x))^{-1}\mathrm{J}\phi'_t(x))^2\right),
\end{aligned}$$

since $\phi''_t = 0$. At time $t = 0$, we have

$$\begin{aligned}
n_0(x) &= V(x) = -\log(\mu^*(x))\\
n'_0(x) &= \langle \nabla V(x), \nabla\psi(x)\rangle - \Delta\psi(x),\\
n''_0(x) &= \langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|^2_{HS}
\end{aligned}$$

since $\mathrm{Tr}(\mathrm{H}\psi) = \Delta\Psi$ and $\mathrm{Tr}\big((\mathrm{H}\psi)^2\big) = \|\mathrm{H}\psi\|^2_{HS}$. Notice that $n'_0(x) = \mathcal{L}_{\mu^*}\psi(x)$ where $\mathcal{L}_{\mu^*} : \psi \mapsto \langle \nabla V, \nabla\psi\rangle - \Delta\psi$ denotes the (negative) generator of the standard Langevin diffusion with stationary distribution μ^* with density $\mu^* \propto e^{-V}$, see Pavliotis (2014, Section 4.5).

Now we get at time $t = 0$:

$$\begin{aligned}
h''_\mu(0) = \int \Bigg[&\left(\tilde{f}''\left(\frac{\mu(x)}{\mu^*(x)}\right)\left(\frac{\mu(x)}{\mu^*(x)}\right)^2 + \tilde{f}'\left(\frac{\mu(x)}{\mu^*(x)}\right)\left(\frac{\mu(x)}{\mu^*(x)}\right)\right)\left(\mathcal{L}_{\mu^*}\psi(x)\right)^2\\
&+\tilde{f}'\left(\frac{\mu(x)}{\mu^*(x)}\right)\left(\frac{\mu(x)}{\mu^*(x)}\right)\left(\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|^2_{HS}\right)\Bigg]\mu(x)\mathrm{d}x.
\end{aligned}$$

Hence if V is convex, and that $\min(\tilde{f}'(t), t\tilde{f}'(t) + t^2\tilde{f}''(t)) \geq 0$, then $h''_\mu(0) \geq 0$. Now let $f(t) = t\log t - t$, then $h_\mu(t) = \mathrm{KL}(\mu_t|\mu^*) - 1$. Then, $\tilde{f}(t) = \log(t) - 1$; $\tilde{f}'(t) = 1/t$, $\tilde{f}''(t) = -1/t^2$, hence $t\tilde{f}'(t) + t^2\tilde{f}''(t) = 0$ and we obtain more precisely:

$$\mathrm{Hess}_\mu \mathrm{KL}(\psi, \psi) = \int \left[\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|^2_{HS}\right]\mu(x)\mathrm{d}x. \qquad \square$$

E.2. Hessian of the mollified relative entropy

Recall that $\mathcal{F}_\epsilon(\mu) = \mathcal{G}_{V_\epsilon}(\mu) + \mathcal{U}_\epsilon(\mu)$. Hence, for any $\psi \in C^\infty_c(\mathbb{R}^d)$, $\mathrm{Hess}_\mu \mathcal{F}_\epsilon(\psi, \psi) = \mathrm{Hess}_\mu \mathcal{G}_{V_\epsilon}(\psi, \psi) + \mathrm{Hess}_\mu \mathcal{U}_\epsilon(\psi, \psi)$. We directly have for the potential energy part that

$$\frac{d^2\mathcal{G}_{V_\epsilon}(\rho_t)}{dt^2}\Big|_{t=0} = \int \langle \mathrm{H}_{V_\epsilon}(x)\nabla\psi(x), \nabla\psi(x)\rangle d\mu(x). \tag{32}$$

using again our computation from Appendix E.1. Since $\mathrm{H}_{V_\epsilon} = k_\epsilon \star \mathrm{H}_V$ and k_ϵ converges to a Dirac at origin as ϵ goes to zero, we get $\mathrm{Hess}_\mu \mathcal{G}_{V_\epsilon}(\psi,\psi) \xrightarrow[\varepsilon\to 0]{} \int \langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle d\mu(x)$.

We now turn to the mollified entropy part. We rewrite it along a geodesic $(\rho_t, v_t)_{t\in[0,1]}$ as

$$\mathcal{U}_\epsilon(\rho_t) = \int \log(k_\epsilon \star \rho_t) d(k_\epsilon \star \rho_t) = \int_\theta U(k_\epsilon \star \rho_t(\theta)) d\mathcal{L}_d(\theta),$$

denoting $U : x \mapsto x\log(x)$. The first time derivative of $t \mapsto \mathcal{U}_\epsilon(\rho_t)$ is:

$$\frac{d\mathcal{U}_\epsilon(\rho_t)}{dt} = \int U'(k_\epsilon \star \rho_t(\theta)) \frac{\mathrm{d}}{\mathrm{d}t} k_\epsilon \star \rho_t(\theta) \mathrm{d}\theta \tag{33}$$

$$= \int (1 + \log(k_\epsilon \star \rho_t(\theta)) \int \langle \nabla k_\epsilon(\theta - x), v_t(x)\rangle \mathrm{d}\rho_t(x)\mathrm{d}\theta \tag{34}$$

Since by an integration by parts,

$$\frac{\mathrm{d}}{\mathrm{d}t} k_\epsilon \star \rho_t(\theta)\mathrm{d}\theta = \int k_\epsilon(\theta - x)\frac{\partial \rho_t(x)}{\partial t}\mathrm{d}x = \int \nabla k_\epsilon(\theta - x)\rho_t(x)v_t(x)\mathrm{d}x.$$

From Equation (33) we obtain

$$\frac{d^2\mathcal{U}_\epsilon(\mu_t)}{dt^2} = \int \left[U^{''}(k_\epsilon \star \rho_t(\theta)) \left(\frac{dk_\epsilon \star \rho_t(\theta)}{dt}\right)^2 + U^{'}(k_\epsilon \star \rho_t(\theta)) \frac{d^2 k_\epsilon \star \rho_t(\theta)}{dt^2}\right] \mathrm{d}\theta$$

$$= \int_\theta \left[(k_\epsilon \star \rho_t(\theta))^{-1} \left(\frac{dk_\epsilon \star \rho_t(\theta)}{dt}\right)^2 + (1 + \log(k_\epsilon \star \rho_t(\theta))) \frac{d^2 k_\epsilon \star \rho_t(\theta)}{dt^2}\right] \mathrm{d}\theta. \tag{35}$$

The first term in (37) is always positive but the second may not because of the logarithmic term. However, as $\epsilon \to 0$, we recover the geodesic convexity of the negative entropy, as stated in the following proposition.

Proposition 15. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Let $\psi \in C_c^\infty(\mathbb{R}^d)$. As $\epsilon \to 0$, the Wasserstein Hessian of the regularized entropy $\mathcal{U}_\epsilon$ converges to the one of the regular negative entropy $\mathcal{U}(\mu) = \int \log(\mu)\mathrm{d}\mu$, i.e:*

$$\mathrm{Hess}_\mu \mathcal{U}_\epsilon(\psi,\psi) \xrightarrow[\varepsilon\to 0]{} \mathrm{Hess}_\mu \mathcal{U}(\psi,\psi) = \int \|\mathrm{H}\psi(x)\|^2_{HS}\mathrm{d}\mu(x). \tag{36}$$

Proof. For each term, we will first take the limit as $t \to 0$ to recover the definition of the Hessian at μ (limiting distribution of ρ_t as t goes to 0), then $\epsilon \to 0$ to recover the case of the standard (non-regularized) relative entropy. Denote $h^\epsilon_\mu(t,\theta) = k_\epsilon \star \rho_t(\theta) = \int k_\epsilon(\theta - x)d\rho_t(x) = \int k_\epsilon(\theta - \phi_t(x))d\mu(x)$ by the transfer lemma. We have

$$\frac{d^2\mathcal{U}_\epsilon(\mu_t)}{dt^2} = \int \left[U^{''}(h^\epsilon_\mu(t,\theta)) \left(\frac{dh^\epsilon_\mu(t,\theta)}{dt}\right)^2 + U^{'}(h^\epsilon_\mu(t,\theta)) \frac{d^2 h^\epsilon_\mu(t,\theta)}{dt^2}\right] \mathrm{d}\theta$$

$$= \int_\theta \left[(h^\epsilon_\mu(t,\theta))^{-1} \left(\frac{dh^\epsilon_\mu(t,\theta)}{dt}\right)^2 + \left(1 + \log\left(h^\epsilon_\mu(t,\theta)\right)\right) \frac{d^2 h^\epsilon_\mu(t,\theta)}{dt^2}\right] \mathrm{d}\theta. \tag{37}$$

We firstly have

$$h^\epsilon_\mu(t,\theta) = k_\epsilon \star \rho_t(\theta) \xrightarrow[t\to 0]{} k_\epsilon \star \mu(\theta) \xrightarrow[\varepsilon\to 0]{} \mu(\theta).$$

Recall that the continuity equation along Wasserstein geodesics write:

$$\frac{\partial \rho_t(x)}{\partial t} + \boldsymbol{\nabla} \cdot \left(\rho_t(x)\nabla\psi \circ \phi_t^{-1}(x)\right)) = 0. \tag{38}$$

Then, using $\boldsymbol{\nabla} \cdot (aB) = \langle \nabla a, B\rangle + a\boldsymbol{\nabla} \cdot (B)$, the first time derivative of $t \mapsto h^\epsilon_\mu(t,\theta)$ writes

$$\frac{dh^\varepsilon_\mu(t,\theta)}{dt} = \int k_\epsilon(\theta - x)\frac{\partial \rho_t(x)}{\partial t}\mathrm{d}x \tag{39}$$

$$= -\int k_\epsilon(\theta - x)\boldsymbol{\nabla} \cdot \left(\rho_t(x)\nabla\psi(\phi_t^{-1}(x))\right)\mathrm{d}x \tag{40}$$

$$\xrightarrow[t\to 0]{} -\int k_\epsilon(\theta - x)\boldsymbol{\nabla} \cdot (\mu(x)\nabla\psi(x))\mathrm{d}x, \tag{41}$$

Hence for the first term in (37) we have:

$$\int (h_\mu^\epsilon(t,\theta))^{-1}\left(\frac{dh_\mu^\varepsilon(t,\theta)}{dt}\right)^2 \mathrm{d}\theta \tag{42}$$

$$\xrightarrow[t\to 0]{} \int (k_\epsilon \star \mu(\theta))^{-1}\left(\int k_\epsilon(\theta-x)\boldsymbol{\nabla}\cdot(\mu(x)\nabla\psi(x))\mathrm{d}x\right)^2 \mathrm{d}\theta \tag{43}$$

$$\xrightarrow[\varepsilon\to 0]{} \int \mu(\theta)^{-1}\boldsymbol{\nabla}\cdot(\mu(\theta)\nabla\psi(\theta))^2\mathrm{d}\theta \tag{44}$$

$$= \int \mu(\theta)^{-1}\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle^2 d\theta + \int \Delta\psi(\theta)^2 d\mu(\theta) + 2\int \Delta\psi(\theta)\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle\mathrm{d}\theta \tag{45}$$

$$= (a)+(b)+(c). \tag{46}$$

We now turn to second term in (37). Using (39), the second time derivative of h_μ^ϵ writes:

$$\begin{aligned}\frac{d^2h_\mu^\varepsilon(t,\theta)}{dt^2} &= -\int k_\epsilon(\theta-x)\boldsymbol{\nabla}\cdot\left(\frac{d}{dt}(\rho_t(x)\nabla\psi(\phi_t^{-1}(x)))\right)\mathrm{d}x\\ &= -\int k_\epsilon(\theta-x)\boldsymbol{\nabla}\cdot\left(\frac{\partial\rho_t(x)}{\partial t}\nabla\psi(\phi_t^{-1}(x))\right)dx - \int k_\epsilon(\theta-x)\boldsymbol{\nabla}\cdot\left(\rho_t(x)\frac{d\nabla\psi(\phi_t^{-1}(x))}{dt}\right)\mathrm{d}x\\ &= (d)+(e).\end{aligned}$$

Then using $\frac{\partial\rho_t(x)}{\partial t} = -\boldsymbol{\nabla}\cdot\big(\rho_t(x)\nabla\psi(\phi_t^{-1}(x))\big) = -\langle\nabla\rho_t(x),\nabla\psi(\phi_t^{-1}(x))\rangle - \rho_t(x)\Delta\psi(\phi_t^{-1}(x))$, we have

$$\begin{aligned}(d) &= \int k_\epsilon(\theta-x)\boldsymbol{\nabla}\cdot\big(\langle\nabla\rho_t(x),\nabla\psi(\phi_t^{-1}(x))\rangle + \rho_t(x)\Delta\psi(\phi_t^{-1}(x))\big)\nabla(\psi(\phi_t^{-1}(x)))\mathrm{d}x\\ &\xrightarrow[t\to 0]{} \int k_\epsilon(\theta-x)\boldsymbol{\nabla}\cdot(\langle\nabla\mu(x),\nabla\psi(x)\rangle + \mu(x)\Delta\psi(x))\nabla(\psi(x))\mathrm{d}x\\ &\xrightarrow[\epsilon\to 0]{} \boldsymbol{\nabla}\cdot\Big(\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle\nabla\psi(\theta)\Big) + \boldsymbol{\nabla}\cdot\Big(\mu(\theta)\Delta\psi(\theta)\nabla(\psi(\theta))\Big)\end{aligned}$$

Now, using $\phi_t^{-1} \approx \mathrm{Id} - t\nabla\psi$ for $t\approx 0$:

$$\begin{aligned}(e) = -\int k_\epsilon(\theta-x)\boldsymbol{\nabla}\cdot\left(\rho_t(x)\frac{d}{dt}(\nabla\psi(\phi_t^{-1}(x)))\right)dx&\\ \xrightarrow[t\to 0]{} \int k_\epsilon(\theta-x)\boldsymbol{\nabla}\cdot(\mu(x)\mathrm{H}\psi(x)\nabla\psi(x))\mathrm{d}x &\xrightarrow[\epsilon\to 0]{} \boldsymbol{\nabla}\cdot(\mu(\theta)\mathrm{H}\psi(\theta)\nabla\psi(\theta))\mathrm{d}x.\end{aligned}$$

Finally, for the second term in (37) we have

$$\begin{aligned}&\int (1+\log(h_\mu^\epsilon(t,\theta)))\frac{d^2h_\mu^\epsilon(t,\theta))}{dt^2}d\theta\\ &\qquad\xrightarrow[t,\epsilon\to 0]{} \int (1+\log(\mu(\theta)))\Big\{\boldsymbol{\nabla}\cdot(\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle\nabla\psi(\theta) + \mu(\theta)\Delta\psi(\theta)\nabla\psi(\theta) + \mu(\theta)\mathrm{H}\psi(\theta)\nabla\psi(\theta))\Big\}\mathrm{d}\theta\\ &\qquad= -\int\langle\nabla\log(\mu(\theta)),\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle\nabla\psi(\theta) + \mu(\theta)\Delta\psi(\theta)\nabla\psi(\theta) + \mu(\theta)\mathrm{H}\psi(\theta)\nabla\psi(\theta)\rangle\mathrm{d}\theta\\ &\qquad= -\int\mu(\theta)^{-1}\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle^2\mathrm{d}\theta - \int\Delta\psi(\theta)\langle\nabla\mu(\theta),\nabla\psi(\theta)\rangle\mathrm{d}\theta - \int\langle\nabla\mu(\theta),\mathrm{H}\psi(\theta)\nabla\psi(\theta)\rangle\mathrm{d}\theta\\ &\qquad= -(a) - \frac{1}{2}(c) - \int\langle\nabla\mu(\theta),\mathrm{H}\psi(\theta)\nabla\psi(\theta)\rangle\mathrm{d}\theta.\end{aligned}$$

Moreover, by an integration by parts, using the divergence of matrix vector product $\boldsymbol{\nabla} \cdot (Ab) = \boldsymbol{\nabla} \cdot (A)b + \mathrm{Tr}(A\nabla b)$:

$$\begin{aligned}
-\int \langle \nabla\mu(\theta), \mathrm{H}\psi(\theta)\nabla\psi(\theta)\rangle \mathrm{d}\theta &= \int \boldsymbol{\nabla} \cdot (\mathrm{H}\psi(\theta)\nabla\psi(\theta))\mathrm{d}\mu(\theta) \\
&= \int \langle \boldsymbol{\nabla} \cdot \mathrm{H}\psi(\theta), \nabla\psi(\theta)\rangle + \mathrm{Tr}(\mathrm{H}\psi(\theta)\mathrm{H}\psi(\theta))^\top \mathrm{d}\mu(\theta) \\
&= \int \langle \nabla(\Delta\psi(\theta)), \nabla\psi(\theta)\rangle + \|\mathrm{H}\psi(\theta)\|_F^2 \mathrm{d}\mu(\theta),
\end{aligned}$$

where

$$\begin{aligned}
\int \langle \nabla(\Delta\psi(\theta)), \nabla\psi(\theta)\rangle d\mu(\theta) &= -\int \Delta\psi(\theta)\boldsymbol{\nabla} \cdot (\mu(\theta)\nabla\psi(\theta))\mathrm{d}\theta \\
&= -\int \Delta\psi(\theta)(\langle \nabla\mu(\theta), \nabla\psi(\theta)\rangle - \mu(\theta)\Delta\psi(\theta))\mathrm{d}\theta \\
&= -\frac{1}{2}(c) - (b).
\end{aligned}$$

Consequently,

$$\int (1 + \log\big(h_\mu^\epsilon(t,\theta)\big))\frac{d^2 h_\mu^\epsilon(t,\theta)}{dt^2} d\theta \xrightarrow[t,\epsilon\to 0]{} -(a) - \frac{1}{2}(c) - \frac{1}{2}(c) - (b) + \int \|\mathrm{H}\psi(\theta)\|_F^2 \mathrm{d}\mu(\theta). \tag{47}$$

Finally combining (46) and (47) we get the result. □

F. Experimental setting

The target distribution The target distribution $\mu^\star$ is chosen to be a Gaussian mixture with 100 components:

$$\mu^\star = \frac{1}{100}\sum_{i=1}^{100} \mathcal{N}(x_i^\star, \epsilon^2 \mathrm{I}_d)$$

The components $(x_i^\star)_{i\leq 100}$ are randomly sampled from a normal distribution $\mathcal{N}(0, \sigma^2 \mathrm{I}_d)$, where $\sigma = 5$ in all experiments. The standard deviation of the target is set to $\epsilon = \epsilon_0\sqrt{d}$, where $\epsilon_0 = 1$ in our setting. This standard deviation scales with $\sqrt{d}$ because the term $\|x_i^\star\|_2$ also scales with $\sqrt{d}$. Without this scaling, the term $\mathcal{N}(x_i^\star, \epsilon^2 \mathrm{I}_d)$ would be very close to a Dirac mass in high dimensions.

Variational family The variational family used for the experiments is the family of Gaussian mixtures with 10 components:

$$\mathcal{C}_n = \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}(x_i, \epsilon^2 \mathrm{I}_d),\ x_i \in \mathbb{R}^d \right\}$$

At the beginning of the training, the mean of each component $(x_i)_{i\leq 10}$ is randomly initialized, sampled from a normal distribution $\mathcal{N}(0, \zeta^2 \mathrm{I}_d)$, where $\zeta = 15$ in all experiments. Note that these components are initialized further than the parameters $(x_i^\star)_{i\leq 100}$ of the target $\mu^\star$, ie, $\zeta \geq \sigma$. It seems that this setting allows to slightly improve the performances and the mode coverage of the algorithm. For simplicity, the variational family shares the same standard deviation ϵ than the target.

Training parameters The step-size is set as $\gamma = \gamma_0 \cdot d$, where $\gamma_0 = 0.01$. According to Proposition 4, the step-size should satisfy $\gamma \leq 2/M$ to ensure a decrease in the objective of each iteration, where the constant M scales inversely with d. Therefore, we opted for γ to scale with d accordingly.

Monte Carlo approximation of the cumulative mean Let μ a Gaussian mixture with n components. We denote by $(x_i)_{i\leq n}$ the mean of those components. Therefore, the term $\|\nabla\mathcal{F}_\epsilon'(\mu)\|_{L^2(\mu)}^2$ can be approximated by Monte Carlo with B samples using

$$\begin{aligned}
\|\nabla \mathcal{F}_\epsilon'(\mu)\|^2_{L^2(\mu)} &= \int \|\nabla \mathcal{F}_\epsilon'(\mu)(w)\|_2^2 d\mu(w) \\
&= \frac{1}{n}\sum_{i=1}^n \|\nabla \mathcal{F}_\epsilon'(\mu)(x_i)\|_2^2 \\
&= \frac{1}{n}\sum_{i=1}^n \| \int \nabla \log\left(\frac{\mu(y)}{\mu^\star(y)}\right) dk_\epsilon^{x_i}(y)\|_2^2 \\
&\approx \frac{1}{n}\sum_{i=1}^n \| \frac{1}{B}\sum_{j=1}^B \nabla \log\left(\frac{\mu(y_j^i)}{\mu^\star(y_j^i)}\right) \|_2^2 \,,
\end{aligned}$$

where $y_j^i \sim \mathcal{N}(x_i, \epsilon^2 \mathrm{I}_d)$.

Monte Carlo approximation of the KL Let ν_n a Gaussian mixture with n components. We denote by $(x_i)_{i\leq n}$ the mean of those components. Therefore, the Kullback-Leibler divergence between ν_n and the target $\mu^\star$ can be approximated by Monte Carlo with B samples using

$$\begin{aligned}
\mathrm{KL}(\nu_n, \mu^\star) &= \int \log\left(\frac{\nu_n(y)}{\mu^\star(y)}\right) d\mu(y) \\
&= \frac{1}{n}\sum_{i=1}^n \int \log\left(\frac{\nu_n(y)}{\mu^\star(y)}\right) dk_\epsilon^{x_i}(y) \\
&\approx \frac{1}{B\cdot n}\sum_{i=1}^n\sum_{j=1}^B \log\left(\frac{\nu_n(y_j^i)}{\mu^\star(y_j^i)}\right) \,,
\end{aligned}$$

where $y_j^i \sim \mathcal{N}(x_i, \epsilon^2 \mathrm{I}_d)$.