Law of Large Numbers for Bayesian two-layer Neural Network trained with Variational Inference

Arnaud Descours, Tom Huix, Arnaud Guillin, Manon Michel, Éric Moulines, Boris Nectoux

Proceedings of Machine Learning Research vol 195:1–39, 2023 36th Annual Conference on Learning Theory

Law of Large Numbers for Bayesian two-layer Neural Network trained with Variational Inference

Arnaud Descours ARNAUD.DESCOURS@UCA.FR
Laboratoire de Mathématiques Blaise Pascal, UMR 6620, CNRS, Université Clermont Auvergne, Aubière, France.

Tom Huix TOM.HUIX@POLYTECHNIQUE.EDU
Centre de Mathématiques Appliquées, UMR 7641, Ecole polytechnique, France.

Arnaud Guillin ARNAUD.GUILLIN@UCA.FR
Laboratoire de Mathématiques Blaise Pascal, UMR 6620, CNRS, Université Clermont Auvergne, Aubière, France.

Manon Michel MANON.MICHEL@UCA.FR
Laboratoire de Mathématiques Blaise Pascal, UMR 6620, CNRS, Université Clermont Auvergne, Aubière, France.

Éric Moulines ERIC.MOULINES@POLYTECHNIQUE.EDU
Centre de Mathématiques Appliquées, UMR 7641, Ecole polytechnique, France.

Boris Nectoux BORIS.NECTOUX@UCA.FR
Laboratoire de Mathématiques Blaise Pascal, UMR 6620, CNRS, Université Clermont Auvergne, Aubière, France.

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We provide a rigorous analysis of training by variational inference (VI) of Bayesian neural networks in the two-layer and infinite-width case. We consider a regression problem with a regularized evidence lower bound (ELBO) which is decomposed into the expected log-likelihood of the data and the Kullback-Leibler (KL) divergence between the a priori distribution and the variational posterior. With an appropriate weighting of the KL, we prove a law of large numbers for three different training schemes: (i) the idealized case with exact estimation of a multiple Gaussian integral from the reparametrization trick, (ii) a minibatch scheme using Monte Carlo sampling, commonly known as *Bayes by Backprop*, and (iii) a new and computationally cheaper algorithm which we introduce as *Minimal VI*. An important result is that all methods converge to the same mean-field limit. Finally, we illustrate our results numerically and discuss the need for the derivation of a central limit theorem.

Keywords: Bayesian neural networks, variational inference, mean-field, law of large numbers, infinite-width neural networks.

1. Introduction

Deep Learning has led to a revolution in machine learning with impressive successes. However, some limitations of DL have been identified and, despite, many attempts, our understanding of DL is still limited. A long-standing problem is the assessment of predictive uncertainty: DL tends to be overconfident in its predictions Abdar et al. (2021), which is a problem in applications such as autonomous driving (McAllister et al., 2017; Michelmore et al., 2020), medical diagnosis (Kendall and Gal, 2017; Filos et al., 2019), or finance; cf Krzywinski and Altman (2013); Ghahramani (2015). Therefore, on the one hand, analytical efforts are being made to thoroughly investigate the performance of DL; and on the other hand, many approaches have been proposed to alleviate its shortcomings. The Bayesian paradigm is an attractive way to tackle predictive uncertainty, as it provides a framework for training uncertainty-aware neural networks (NNs) (e.g. Ghahramani (2015); Blundell et al. (2015); Gal and Ghahramani (2016)).

Thanks to a fully probabilistic approach, Bayesian Neural Networks (BNN) combine the impressive neural-network expressivity with the decision-theoretic approach of Bayesian inference, making them capable of providing predictive uncertainty; see Blundell et al. (2015); Michelmore et al. (2020); McAllister et al. (2017); Filos et al. (2019). However, Bayesian inference requires deriving the posterior distribution of the NN weights. This posterior distribution is typically not tractable. A classical approach is to sample the posterior distribution using Markov chain Monte Carlo methods (such as Hamilton-Monte-Carlo methods). There are however long-standing difficulties, such as the proper choice of the prior and fine-tuning of the sampler. Such difficulties often become prohibitive in large-dimensional cases,(Cobb and Jalaian, 2021). An alternative is to use variational inference, which has a long history (Hinton and Camp, 1993; MacKay, 1995; MacKay et al., 1995). Simpler methods that do not require exact computation of integrals over the variational posterior were then developed, e.g. first by Graves (2011) thanks to some approximation and then by Blundell et al. (2015) with the *Bayes by Backprop* approach. In the latter, the posterior distribution is approximated by a parametric distribution and a generalisation of the reparametrization trick used by Kingma and Welling (2014) leads to an unbiased estimator of the gradient of the ELBO; see also Gal and Ghahramani (2016); Louizos and Welling (2017); Khan et al. (2018). Despite the successful application of this approach, little is known about the overparameterized limit and appropriate weighting that must be assumed to obtain a nontrivial Bayesian posterior, see Izmailov et al. (2021). Recently, Huix et al. (2022) outlined the importance of balancing in ELBO the integrated log-likelihood term and the KL regularizer, to avoid both overfitting and dominance of the prior. However, a suitable limiting theory has yet to be established, as well as guarantees for the practical implementation of the stochastic gradient descent (SGD) used to estimate the parameters of the variational distribution.

Motivated by the need to provide a solid theoretical framework, asymptotic analysis of NN has gained much interest recently. The main focus has been on the gradient descent algorithm and its variants (Rotskoff and Vanden-Eijnden, 2018; Chizat and Bach, 2018; Mei et al., 2018; Sirignano and Spiliopoulos, 2020; Descours et al., 2022). In much of these works, a mean-field analysis is performed to characterize the limiting nonlinear evolution of the weights of a two-layer NN, allowing the derivation of a law of large numbers and a central limit theorem for the empirical distribution of neuron weights. A long-term goal of these works is to demonstrate convergence toward a global minimum of these limits for the mean field. Despite some progress in this direction, this is still an open and highly challenging problem; cf Chizat and Bach (2018); Chizat (2022);

Chizat et al. (2022). Nevertheless, this asymptotic analysis is also of interest in its own right, as we show here in the case of variational inference for Bayesian neural networks. Indeed, based on this asymptotic analysis, we develop an efficient and new variant of the stochastic gradient descent (SGD) algorithm for variational inference in BNN that computes only the information necessary to recover the limit behavior.

Our goal, then, is to work at the intersection of analytical efforts to gain theoretical guarantees and insights and of practical methods for a workable variational inference procedure. By adapting the framework developed by Descours et al. (2022), we produce a rigorous asymptotic analysis of BNN trained in a variational setting for a regression task. From the limit equation analysis, we first find that a proper regularisation of the Kullback-Leibler divergence term in relation with the integrated loss leads to their right asymptotic balance. Second, we prove the asymptotic equivalence of the idealized and Bayes-by-Backprop SGD schemes, as both preserve the same core contributions to the limit. Finally, we introduce a computationally more favourable scheme, directly stemming from the effective asymptotic contributions. This scheme is the true mean-field algorithmic approach, as only deriving from non-interacting terms.

More specifically, our contributions are the following:

- We first focus on the idealized SGD algorithm, where the variational expectations of the derivative of the loss from the reparametrization trick of Blundell et al. (2015) are computed exactly. More precisely, we prove that with the number of neurons $N \to +\infty$, the sequence of trajectories of the scaled empirical distributions of the parameters satisfies a law of large numbers. This is the purpose of Theorem 2. The proof is completely new: it establishes directly the limit in the topology inherited by the Wasserstein distance bypassing the highly technical Sobolev space arguments used in Descours et al. (2022).

The idealized SGD requires the computation of some integrals, which in practice prevents a direct application of this algorithm. However, we can prove its convergence to an explicit nonlinear process. These integrals are usually obtained by a Monte Carlo approximation, leading to the *Bayes-by-Backprop* SGD, see Blundell et al. (2015).

- We show for the *Bayes-by-Backprop* SGD (see Theorem 3) that the sequence of trajectories of the scaled empirical distributions of the parameters satisfies the same law of large numbers as that in Theorem 2, which justifies such an approximation procedure. Note that each step of the algorithm involves the simulation of $O(N)$ Gaussian random variables, which can make the associated gradient evaluation prohibitively expensive.

- A careful analysis of the structure of the limit equation (11) allows us to develop a new algorithm, called *Minimal-VI* SGD, which at each step generates only two Gaussian random variables and for which we prove the same limiting behavior. The key idea here is to keep only those contributions which affect the asymptotic behavior and which can be understood as the mean-field approximation from the uncorrelated degrees of freedom. This is all the more interesting since we observe numerically that the number weights N required to reach this asymptotic limit is quite small which makes this variant of immediate practical interest.

- We numerically investigate the convergence of the three methods to the common limit behavior on a toy example. We observe that the mean-field method is effective for a small number of neurons ($N = 300$). The differences between the methods are reflected in the variances.

The paper is organized as follows: Section 2 introduces the variational inference in BNN, as well as the SGD schemes commonly considered, namely the idealized and *Bayes-by-backprop* variants. Then, in Section 3 we establish our initial result, the LLN for the idealized SGD. In Section 4 we prove the LLN for the *Bayes-by-backprop* SGD and its variants. We show that both SGD schemes have the same limit behavior. Based on an analysis of the obtained limit equation, we present in Section 5 the new *minimal- VI*. Finally, in Section 6 we illustrate our findings using numerical experiments. The proofs of the mean-field limits, which are original and quite technically demanding, are gathered in the supplementary paper.

Related works. Law of Large Numbers (LLN) for mean-field interacting particle systems, have attracted a lot of attentions; see for example Hitsuda and Mitoma (1986); Sznitman (1991); Fernandez and Méléard (1997); Jourdain and Méléard (1998); Delarue et al. (2019); Del Moral and Guionnet (1999); Kurtz and Xiong (2004) and references therein. The use of mean-field particle systems to analyse two-layer neural networks with random initialization have been considered in Mei et al. (2018, 2019), which establish a LLN on the empirical measure of the weights at fixed times - we consider in this paper the trajectory convergence, i.e. the whole empirical measure process (time indexed) converges uniformly w.r.t. Skorohod topology. It enables not only to use the limiting PDE, for example to study the convergence of the weights towards the infimum of the loss function (see Chizat and Bach (2018) for preliminary results), but is is also crucial to establish the central limit theorem, see for example Descours et al. (2022). Rotskoff and Vanden-Eijnden (2018) give conditions for global convergence of GD for exact mean-square loss and online stochastic gradient descent (SGD) with mini-batches increasing in size with the number of weights N. A LLN for the entire trajectory of the empirical measure is also given in Sirignano and Spiliopoulos (2020) for a standard SGD. De Bortoli et al. (2020) establish the propagation of chaos for SGD with different step size schemes. Compared to the existing literature dealing with the SGD empirical risk minimization in two-layer neural networks, Descours et al. (2022) provide the first rigorous proof of the existence of the limit PDE, and in particular its uniqueness, in the LLN.

We are interested here in deriving a LLN but for Variational Inference (VI) of two-layer Bayesian Neural Networks (BNN), where we consider a regularized version of the Evidence Lower Bound (ELBO).

2. Variational inference in BNN: Notations and common SGD schemes

2.1. Variational inference and Evidence Lower Bound

Setting. Let X and Y be subsets of $\mathbf{R}^n$ ($n \geq 1$) and $\mathbf{R}$ respectively. For $N \geq 1$ and $\boldsymbol{w} = (w^1, \dots, w^N) \in (\mathbf{R}^d)^N$, let $f^N_{\boldsymbol{w}} : \mathsf{X} \to \mathbf{R}$ be the following two-layer neural network: for $x \in \mathsf{X}$,

$$f^N_{\boldsymbol{w}}(x) := \frac{1}{N} \sum_{i=1}^{N} s(w^i, x) \in \mathbf{R},$$

where $s : \mathbf{R}^d \times \mathsf{X} \to \mathbf{R}$ is the activation function. We work in a Bayesian setting, in which we seek a distribution of the latent variable $\boldsymbol{w}$ which represents the weights of the neural network. The standard problem in Bayesian inference over complex models is that the posterior distribution is hard to sample. To tackle this problem, we consider Variational Inference, in which we consider a family of distribution $\mathcal{Q}^N = \{q^N_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Xi^N\}$ (where Ξ is some parameter space) easy to sample. The

objective is to find the best $q_{\boldsymbol{\theta}}^N \in \mathcal{Q}^N$, the one closest in KL divergence (denoted $\mathscr{D}_{\mathrm{KL}}$) to the exact posterior. Because we cannot compute the KL, we optimize the evidence lower bound (ELBO), which is equivalent to the KL up to an additive constant.

Denoting by $\mathfrak{L} : \mathbf{R} \times \mathbf{R} \to \mathbf{R}_+$ the negative log-likelihood (by an abuse of language, we call this quantity the *loss*), the ELBO (see Blei et al. (2017)) is defined, for $\boldsymbol{\theta} \in \Xi^N$, $(x, y) \in \mathsf{X} \times \mathsf{Y}$, by

$$\mathrm{E}_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y) := -\int_{(\mathbf{R}^d)^N} \mathfrak{L}(y, f_{\boldsymbol{w}}^N(x)) q_{\boldsymbol{\theta}}^N(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} - \mathscr{D}_{\mathrm{KL}}(q_{\boldsymbol{\theta}}^N | P_0^N),$$

where P_0^N is some prior on the weights of the NN. The ELBO is decomposed into two terms: one corresponding to the Kullback-Leibler (KL) divergence between the variational density and the prior and the other to a marginal likelihood term. It was empirically found that the maximization of the ELBO function is prone to yield very poor inferences (Coker et al., 2022). It is argued in Coker et al. (2022) and Huix et al. (2022) that optimizing the ELBO leads as $N \to \infty$ to the collapse of the variational posterior to the prior. Huix et al. (2022) proposed to consider a regularized version of the ELBO, which consists in multiplying the KL term by a parameter which is scaled by the inverse of the number of neurons:

$$\mathrm{E}_{\mathrm{lbo}}^N(\boldsymbol{\theta}, x, y) := -\int_{(\mathbf{R}^d)^N} \mathfrak{L}(y, f_{\boldsymbol{w}}^N(x)) q_{\boldsymbol{\theta}}^N(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q_{\boldsymbol{\theta}}^N | P_0^N), \tag{1}$$

A first objective of this paper is to show that the proposed regularization leads to a stable asymptotic behavior and the effect of both the integrated loss and Kullback-Leibler terms on the limiting behavior are balanced in the limit $N \to \infty$. The maximization of $\mathrm{E}_{\mathrm{lbo}}^N$ is carried out using SGD.

The variational family $\mathcal{Q}^N$ we consider is a Gaussian family of distributions. More precisely, we assume that for any $\boldsymbol{\theta} = (\theta^1, \dots, \theta^N) \in \Xi^N$, the variational distribution $q_{\boldsymbol{\theta}}^N$ factorizes over the neurons: for all $\boldsymbol{w} = (w^1, \dots, w^N) \in (\mathbf{R}^d)^N$, $q_{\boldsymbol{\theta}}^N(\boldsymbol{w}) = \prod_{i=1}^N q_{\theta^i}^1(w^i)$, where $\theta = (m, \rho) \in \Xi := \mathbf{R}^d \times \mathbf{R}$ and q_θ^1 is the probability density function (pdf) of $\mathcal{N}(m, g(\rho)^2 I_d)$, with $g(\rho) = \log(1 + e^\rho)$, $\rho \in \mathbf{R}$.

In the following, we simply write $\mathbf{R}^{d+1}$ for $\mathbf{R}^d \times \mathbf{R}$. In addition, following the reparameterisation trick of Blundell et al. (2015), $q_\theta^1(w)\mathrm{d}w$ is the pushforward of a reference probability measure with density γ by Ψ_θ (see more precisely Assumption **A1**). In practice, γ is the pdf of $\mathcal{N}(0, I_d)$ and $\Psi_\theta(z) = m + g(\rho)z$. With these notations, (1) writes

$$\mathrm{E}_{\mathrm{lbo}}^N(\boldsymbol{\theta}, x, y) = -\int_{(\mathbf{R}^d)^N} \mathfrak{L}\Big(y, \frac{1}{N}\sum_{i=1}^N s(\Psi_{\theta^i}(z^i), x)\Big) \gamma(z^1) \dots \gamma(z^N) \mathrm{d}z_1 \dots \mathrm{d}z_N - \frac{1}{N} \mathscr{D}_{\mathrm{KL}}(q_{\boldsymbol{\theta}}^N | P_0^N).$$

Loss function and prior distribution. In this work, we focus on the regression problem, i.e. $\mathfrak{L}$ is the Mean Square Loss: for $y_1, y_2 \in \mathbf{R}$, $\mathfrak{L}(y_1, y_2) = \frac{1}{2}|y_1 - y_2|^2$. We also introduce the function $\phi : (\theta, z, x) \in \mathbf{R}^{d+1} \times \mathbf{R}^d \times \mathsf{X} \mapsto s(\Psi_\theta(z), x)$. On the other hand, we assume that the prior distribution P_0^N write, for all $\boldsymbol{w} \in (\mathbf{R}^d)^N$, $P_0^N(\boldsymbol{w}) = \prod_{i=1}^N P_0^1(w^i)$, where $P_0^1 : \mathbf{R}^d \to \mathbf{R}_+$ is the pdf of $\mathcal{N}(m_0, \sigma_0^2 I_d)$, and $\sigma_0 > 0$. Therefore $\mathscr{D}_{\mathrm{KL}}(q_{\boldsymbol{\theta}}^N | P_0^N) = \sum_{i=1}^N \mathscr{D}_{\mathrm{KL}}(q_{\theta^i} | P_0^1)$ and, for $\theta = (m, \rho) \in \mathbf{R}^{d+1}$,

$$\mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1) = \int_{\mathbf{R}^d} q_\theta^1(x) \log(q_\theta^1(x)/P_0^1(x)) \mathrm{d}x = \frac{\|m - m_0\|_2^2}{2\sigma_0^2} + \frac{d}{2}\Big(\frac{g(\rho)^2}{\sigma_0^2} - 1\Big) + \frac{d}{2}\log\Big(\frac{\sigma_0^2}{g(\rho)^2}\Big).$$

Note that $\mathscr{D}_{\mathrm{KL}}$ has at most a quadratic growth in m and ρ.

Note that we assume here a Gaussian prior to get an explicit expression of the Kullback-Leibler divergence. Most arguments extend to sufficiently regular densities and are essentially the same for exponential families, using conjugate families for the variational approximation.

2.2. Common SGD schemes in backpropagation in a variational setting

Idealized SGD. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Consider a data set $\{(x_k, y_k)\}_{k\geq 0}$ i.i.d. w.r.t. $\pi \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$, the space of probability measures over $\mathsf{X}\times\mathsf{Y}$. For $N \geq 1$ and given a learning rate $\eta > 0$, the maximization of $\theta \in \mathbf{R}^{d+1} \mapsto \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y)$ with a SGD algorithm writes as follows: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases} \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \nabla_{\boldsymbol{\theta}} \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}_k, x_k, y_k) \\ \boldsymbol{\theta}_0 \sim \mu_0^{\otimes N}, \end{cases} \tag{2}$$

where $\mu_0 \in \mathcal{P}(\mathbf{R}^{d+1})$ and $\boldsymbol{\theta}_k = (\theta^1_k, \dots, \theta^N_k)$. We now compute $\nabla_{\boldsymbol{\theta}} \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y)$.

First, under regularity assumptions on the function ϕ (which will be formulated later, see **A1** and **A3** below) and by assumption on $\mathfrak{L}$, we have for all $i \in \{1, \dots, N\}$ and all $(x, y) \in \mathsf{X}\times\mathsf{Y}$,

$$\begin{aligned} &\int_{(\mathbf{R}^d)^N} \nabla_{\theta^i} \mathfrak{L}\Big(y, \frac{1}{N}\sum_{j=1}^N \phi(\theta^j, z^j, x)\Big) \gamma(z^1)\dots\gamma(z^N) \mathrm{d}z^1 \dots \mathrm{d}z^N \\ &= -\frac{1}{N^2}\sum_{j=1}^N \int_{(\mathbf{R}^d)^N} (y - \phi(\theta^j, z^j, x)) \nabla_\theta \phi(\theta^i, z^i, x) \gamma(z^1)\dots\gamma(z^N)\mathrm{d}z^1\dots\mathrm{d}z^N \\ &= -\frac{1}{N^2}\Big[\sum_{j=1, j\neq i}^N (y - \langle \phi(\theta^j, \cdot, x), \gamma\rangle)\langle \nabla_\theta\phi(\theta^i, \cdot, x), \gamma\rangle + \langle (y - \phi(\theta^i, \cdot, x))\nabla_\theta\phi(\theta^i, \cdot, x), \gamma\rangle\Big], \end{aligned} \tag{3}$$

where we have used the notation $\langle U, \nu\rangle = \int_{\mathbf{R}^q} U(z)\nu(\mathrm{d}z)$ for any integrable function $U : \mathbf{R}^q \to \mathbf{R}$ w.r.t. a measure ν (with a slight abuse of notation, we denote by γ the measure $\gamma(z)\mathrm{d}z$). Second, for $\theta \in \mathbf{R}^{d+1}$, we have

$$\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\theta | P^1_0) = \begin{pmatrix} \nabla_m \mathscr{D}_{\mathrm{KL}}(q^1_\theta | P^1_0) \\ \partial_\rho \mathscr{D}_{\mathrm{KL}}(q^1_\theta | P^1_0) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_0^2}(m - m_0) \\ \frac{d}{\sigma_0^2} g'(\rho) g(\rho) - d\frac{g'(\rho)}{g(\rho)} \end{pmatrix}. \tag{4}$$

In conclusion, the SGD (2) writes: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases} \theta^i_{k+1} = \theta^i_k - \dfrac{\eta}{N^2}\displaystyle\sum_{j=1, j\neq i}^N \Big(\langle \phi(\theta^j_k, \cdot, x_k), \gamma\rangle - y_k\Big)\langle \nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\rangle \\ \qquad\qquad -\dfrac{\eta}{N^2}\Big\langle (\phi(\theta^i_k, \cdot, x_k) - y_k)\nabla_\theta\phi(\theta^i_k, \cdot, x_k), \gamma\Big\rangle - \dfrac{\eta}{N}\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k} | P^1_0) \\ \theta^i_0 \sim \mu_0. \end{cases} \tag{5}$$

We shall call this algorithm *idealised* SGD because it contains an intractable term given by the integral w.r.t. γ. This has motivated the development of methods where this integral is replaced by an

unbiased Monte Carlo estimator (see Blundell et al. (2015)) as detailed below.

***Bayes-by-Backprop* SGD**. The second SGD algorithm we study is based on an approximation, for $i \in \{1, \dots, N\}$, of $\int_{(\mathbf{R}^d)^N} (y - \phi(\theta^j, z^j, x)) \nabla_\theta \phi(\theta^i, z^i, x) \gamma(z^1) \dots \gamma(z^N) \mathrm{d}z^1 \dots \mathrm{d}z^N$ (see (3)) by

$$\frac{1}{B} \sum_{\ell=1}^{B} \big(y - \phi(\theta^j, \mathsf{Z}^{j,\ell}, x)\big) \nabla_\theta \phi(\theta^i, \mathsf{Z}^{i,\ell}, x) \tag{6}$$

where $B \in \mathbf{N}^*$ is a fixed integer and $(\mathsf{Z}^{q,\ell}, q \in \{i, j\}, 1 \le \ell \le B)$ is a i.i.d finite sequence of random variables distributed according to $\gamma(z)\mathrm{d}z$. In this case, for $N \ge 1$, given a dataset $(x_k, y_k)_{k \ge 0}$, the maximization of $\theta \in \mathbf{R}^{d+1} \mapsto \mathrm{E}^N_{\mathrm{lbo}}(\boldsymbol{\theta}, x, y)$ with a SGD algorithm is the following: for $k \ge 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases} \theta^i_{k+1} = \theta^i_k - \dfrac{\eta}{N^2 B} \displaystyle\sum_{j=1}^{N} \sum_{\ell=1}^{B} \big(\phi(\theta^j_k, \mathsf{Z}^{j,\ell}_k, x_k) - y_k\big) \nabla_\theta \phi(\theta^i_k, \mathsf{Z}^{i,\ell}_k, x_k) - \dfrac{\eta}{N} \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_k} | P^1_0) \\ \theta^i_0 = (m^i_0, \rho^i_0) \sim \mu_0, \end{cases} \tag{7}$$

where $\eta > 0$ and $(\mathsf{Z}^{j,\ell}_k, 1 \le j \le N, 1 \le \ell \le B, k \ge 0)$ is a i.i.d sequence of random variables distributed according to γ.

3. Law of large numbers for the idealized SGD

Assumptions and notations. When E is a metric space and $\mathscr{I} = \mathbf{R}_+$ or $\mathscr{I} = [0, T]$ ($T \ge 0$), we denote by $\mathcal{D}(\mathscr{I}, E)$ the Skorohod space of càdlàg functions on $\mathscr{I}$ taking values in E and $\mathcal{C}(\mathscr{I}, E)$ the space of continuous functions on $\mathscr{I}$ taking values in E. The evolution of the parameters $(\{\theta^i_k, i = 1, \dots, N\})_{k \ge 1}$ defined by (5) is tracked through their empirical distribution ν^N_k (for $k \ge 0$) and its scaled version μ^N_t (for $t \in \mathbf{R}_+$), which are defined as follows:

$$\nu^N_k := \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta^i_k} \quad \text{and} \quad \mu^N_t := \nu^N_{\lfloor Nt \rfloor}, \text{ where the } \theta^i_k\text{'s are defined (5).} \tag{8}$$

Fix $T > 0$. For all $N \ge 1$, $\mu^N := \{\mu^N_t, t \in [0, T]\}$ is a random element of $\mathcal{D}([0, T], \mathcal{P}(\mathbf{R}^{d+1}))$, where $\mathcal{P}(\mathbf{R}^{d+1})$ is endowed with the weak convergence topology. For $N \ge 1$ and $k \ge 1$, we introduce the following $\boldsymbol{\sigma}$-algebras:

$$\mathcal{F}^N_0 = \boldsymbol{\sigma}(\theta^i_0, 1 \le i \le N) \text{ and } \mathcal{F}^N_k = \boldsymbol{\sigma}(\theta^i_0, (x_q, y_q), 1 \le i \le N, 0 \le q \le k-1). \tag{9}$$

Recall $q^1_\theta : \mathbf{R}^d \to \mathbf{R}_+$ be the pdf of $\mathcal{N}(m, g(\rho)^2 I_d)$ ($\theta = (m, \rho) \in \mathbf{R}^{d+1}$). In this work, we assume the following.

A1. There exists a pdf $\gamma : \mathbf{R}^d \to \mathbf{R}_+$ such that for all $\theta \in \mathbf{R}^{d+1}$, $q^1_\theta \mathrm{d}x = \Psi_\theta \# \gamma \mathrm{d}x$, where $\{\Psi_\theta, \theta \in \mathbf{R}^{d+1}\}$ is a family of $\mathcal{C}^1$-diffeomorphisms over $\mathbf{R}^d$ such that for all $z \in \mathbf{R}^d$, $\theta \in \mathbf{R}^{d+1} \mapsto \Psi_\theta(z)$ is of class $\mathcal{C}^\infty$. Finally, there exists $\mathfrak{b} : \mathbf{R}^d \to \mathbf{R}_+$ such that for all multi-index $\alpha \in \mathbf{N}^{d+1}$ with $|\alpha| \ge 1$, there exists $C_\alpha > 0$, for all $z \in \mathbf{R}^d$ and $\theta = (\theta_1, \dots, \theta_{d+1}) \in \mathbf{R}^{d+1}$,

$$\big|\partial_\alpha \Psi_\theta(z)\big| \le C_\alpha \mathfrak{b}(z) \quad \text{with for all } q \ge 1, \ \langle \mathfrak{b}^q, \gamma \rangle < +\infty, \tag{10}$$

where $\partial_\alpha = \partial^{\alpha_1}_{\theta_1} \dots \partial^{\alpha_{d+1}}_{\theta_{d+1}}$ and $\partial^{\alpha_j}_{\theta_j}$ is the partial derivatives of order α_j w.r.t. to θ_j.

A2. The sequence $\{(x_k, y_k)\}_{k\geq 0}$ is i.i.d. w.r.t. $\pi \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$. The set $\mathsf{X}\times\mathsf{Y} \subset \mathbf{R}^d \times \mathbf{R}$ is compact. For all $k \geq 0$, $(x_k, y_k) \perp\!\!\!\perp \mathcal{F}_k^N$, where $\mathcal{F}_k^N$ is defined in (9).

A3. The activation function $s : \mathbf{R}^d \times \mathsf{X} \to \mathbf{R}$ belongs to $\mathcal{C}_b^\infty(\mathbf{R}^d \times \mathsf{X})$ (the space of smooth functions over $\mathbf{R}^d \times \mathsf{X}$ whose derivatives of all order are bounded).

A4. The initial parameters $(\theta_0^i)_{i=1}^N$ are i.i.d. w.r.t. $\mu_0 \in \mathcal{P}(\mathbf{R}^{d+1})$ which has compact support.

Note that **A1** is satisfied when γ is the pdf of $\mathcal{N}(0, I_d)$ and $\Psi_\theta(z) = m + g(\rho)z$, with $\mathfrak{b}(z) = 1 + |z|$. With these assumptions, for every fixed $T > 0$, the sequence $(\{\theta_k^i, i = 1, \dots, N\})_{k=0,\dots,\lfloor NT\rfloor}$ defined by (5) is a.s. bounded:

Lemma 1 (Uniform bound on the parameters) *Assume* **A1**→**A4**. *Then, there exists $C > 0$ such that a.s. for all $T > 0$, $N \geq 1$, $i \in \{1, \dots, N\}$, and $0 \leq k \leq \lfloor NT\rfloor$, $|\theta_k^i| \leq Ce^{[C(2+T)]T}$.*

Lemma 1 implies that a.s. for all $T > 0$ and $N \geq 1$, $\mu^N \in \mathcal{D}([0,T], \mathcal{P}(\Theta_T))$, where

$$\Theta_T = \{\theta \in \mathbf{R}^{d+1}, |\theta| \leq Ce^{[C(2+T)]T}\}.$$

Law of large numbers for $(\mu^N)_{N\geq 1}$ defined in (8). The first main result of this work is the following.

Theorem 2 *Assume* **A1**→**A4**. *Let $T > 0$. Then, the sequence $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,T], \mathcal{P}(\Theta_T))$ defined in (8) converges in probability to the unique deterministic solution $\bar{\mu} \in \mathcal{C}([0,T], \mathcal{P}(\Theta_T))$ to the following measure-valued evolution equation: $\forall f \in \mathcal{C}^\infty(\Theta_T)$ and $\forall t \in [0,T]$,*

$$\begin{aligned}\langle f, \bar{\mu}_t\rangle - \langle f, \mu_0\rangle = &-\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \big\langle \phi(\cdot,\cdot,x) - y, \bar{\mu}_s \otimes \gamma\big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \bar{\mu}_s \otimes \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y)\mathrm{d}s \\ &- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_s\big\rangle \mathrm{d}s. \end{aligned} \tag{11}$$

The proof of Theorem 2 is given in Appendix A. We stress here the most important steps and used techniques. In a first step, we derive an identity satisfied by $(\mu^N)_{N\geq 1}$, namely the pre-limit equation (28); see Sec. A.1. Then we show in Sec. A.2.2 that $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}([0,T], \mathcal{P}(\Theta_T))$. To do so, we check that the sequence $(\mu^N)_{N\geq 1}$ satisfies all the required assumptions of (Jakubowski, 1986, Theorem 3.1) when $E = \mathcal{P}(\Theta_T)$ there. In Sec. A.2.3 we prove that every limit point of $(\mu^N)_{N\geq 1}$ satisfies the limit equation (11). Then, in Section A.2.4, we prove that there is a unique solution of the measure-valued equation (11). To prove the uniqueness of the solution of (11), we use techniques developed in Piccoli et al. (2015) which are based on a representation formula for solution to measure-valued equations (Villani, 2003, Theorem 5.34) together with estimates in Wasserstein distances between two solutions of (11) derived in Piccoli and Rossi (2016). In Section A.2.4, we also conclude the proof of Theorem 2. Compared to (Descours et al., 2022, Theorem 1), the fact that $(\{\theta_k^i, i = 1, \dots, N\})_{k=0,\dots,\lfloor NT\rfloor}$ defined by (5) are a.s. bounded allows to use different and more straightforward arguments to prove (i) the relative compactness in $\mathcal{D}([0,T], \mathcal{P}(\Theta_T))$ of $(\mu^N)_{N\geq 1}$ (defined in (8)) (ii) the continuity property of the operator $\mathsf{m} \mapsto \mathbf{\Lambda}_t[f](\mathsf{m})$ defined in (35) w.r.t. the topology of $\mathcal{D}([0,T], \mathcal{P}(\Theta_T))$ and (iii) $(\mu^N)_{N\geq 1}$ has limit points in $\mathcal{C}([0,T], \mathcal{P}(\Theta_T))$. Step (ii) is necessary in order to pass to the limit $N \to +\infty$

in the pre-limit equation and Step (iii) is crucial since we prove that there is at most one solution of (11) in $\mathcal{C}([0,T],\mathcal{P}(\Theta_T))$. It is worthwhile to emphasize that, as $N\to\infty$, the effects of the integrated loss and of the KL terms are balanced, as conjectured in Huix et al. (2022).

To avoid further technicalities, we have chosen what may seem restrictive assumptions on the data or the activation function. Note however that it readily extends to unbounded set X, and also unbounded Y assuming that π as polynomial moments of sufficiently high order. Also, RELU (or more easily leaky RELU) may be considered by using weak derivatives (to consider the singularity at 0), and a priori moment bounds on the weights.

4. LLN for the *Bayes-by-Backprop* SGD

The sequence $\{\theta_k^i, i\in\{1,\dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ defined recursively by the algorithm (7) is in general not bounded, since $\nabla_\theta\phi(\theta,\mathsf{Z},x)$ is not necessarily bounded if $\mathsf{Z}\sim\gamma(s)\mathrm{d}z$. Therefore, we cannot expect Lemma 1 to hold for $\{\theta_k^i, i\in\{1,\dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ set by (7). Thus, the sequence $\{\theta_k^i, i\in\{1,\dots N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ is considered on the whole space $\mathbf{R}^{d+1}$.

Wasserstein spaces and results. For $N\geq 1$, and $k\geq 1$, we set

$$\mathcal{F}_k^N=\boldsymbol{\sigma}\Big(\theta_0^i,\mathsf{Z}_q^{j,\ell},(x_q,y_q),1\leq i,j\leq N,1\leq\ell\leq B,0\leq q\leq k-1\Big\}\Big). \tag{12}$$

In addition to **A1**→**A4** (where in **A2**, when $k\geq 1$, $\mathcal{F}_k^N$ is now the one defined in (12)), we assume:

A5. The sequences $(\mathsf{Z}_k^{j,\ell},1\leq j\leq N,1\leq\ell\leq B,k\geq 0)$ and $((x_k,y_k),k\geq 0)$ are independent. In addition, for $k\geq 0$, $\big((x_k,y_k),\mathsf{Z}_k^{j,\ell},1\leq j\leq N,1\leq\ell\leq B\big)\perp\!\!\!\perp\mathcal{F}_k^N$.

Note that the last statement of **A5** implies the last statement of **A2**. We introduce the scaled empirical distribution of the parameters of the algorithm (7), i.e. for $k\geq 0$ and $t\geq 0$:

$$\nu_k^N:=\frac{1}{N}\sum_{i=1}^N\delta_{\theta_k^i}\ \text{ and }\ \mu_t^N:=\nu_{\lfloor Nt\rfloor}^N,\ \text{where the }\theta_k^i\text{'s are defined (7).} \tag{13}$$

One can no longer rely on the existence of a compact subset $\Theta_T\subset\mathbf{R}^{d+1}$ such that a.s. $(\mu^N)_{N\geq 1}\subset\mathcal{D}([0,T],\mathcal{P}(\Theta_T))$, where $\mu^N=\{t\geq 0\mapsto\mu_t^N\}$ is defined in (13). For this reason, we will work in Wasserstein spaces $\mathcal{P}_q(\mathbf{R}^{d+1})$, $q\geq 0$, which, we recall, are defined by

$$\mathcal{P}_q(\mathbf{R}^{d+1})=\Big\{\nu\in\mathcal{P}(\mathbf{R}^{d+1}),\int_{\mathbf{R}^{d+1}}|\theta|^q\nu(\mathrm{d}\theta)<+\infty\Big\}. \tag{14}$$

These spaces are endowed with the Wasserstein metric W_q, see e.g. (Santambrogio, 2015, Chapter 5) for more materials on Wasserstein spaces. For all $q\geq 0$, $(\mu^N)_{N\geq 1}\subset\mathcal{D}(\mathbf{R}_+,\mathcal{P}_q(\mathbf{R}^{d+1}))$. The second main results of this work is a LLN for $(\mu^N)_{N\geq 1}$ defined in (13).

Theorem 3 *Assume* **A1**→**A5**. *Let* $\gamma_0>1+\frac{d+1}{2}$. *Then, the sequence* $(\mu^N)_{N\geq 1}$ *defined in* (13) *converges in probability in* $\mathcal{D}(\mathbf{R}_+,\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ *to a deterministic element* $\bar\mu\in\mathcal{D}(\mathbf{R}_+,\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$, *where* $\bar\mu\in\mathcal{C}(\mathbf{R}_+,\mathcal{P}_1(\mathbf{R}^{d+1}))$ *is the unique solution in* $\mathcal{C}(\mathbf{R}_+,\mathcal{P}_1(\mathbf{R}^{d+1}))$ *to the following measure-valued evolution equation:*$\forall f\in\mathcal{C}_b^\infty(\mathbf{R}^{d+1})$ *and* $\forall t\in\mathbf{R}_+$,

$$\begin{aligned}\langle f,\bar\mu_t\rangle-\langle f,\mu_0\rangle=&-\eta\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\big\langle\phi(\cdot,\cdot,x)-y,\bar\mu_s\otimes\gamma\big\rangle\big\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\bar\mu_s\otimes\gamma\big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\&-\eta\int_0^t\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\bar\mu_s\big\rangle\mathrm{d}s.\end{aligned} \tag{15}$$

Theorem 3 is proved in the appendix B. Since $\{\theta_k^i, i \in \{1, \dots N\}\}_{k=0,\dots,\lfloor NT \rfloor}$ defined by (7) is not bounded in general, we work in the space $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. The proof of Theorem 3 is more involved than that of Theorem 2, and generalizes the latter to the case where the parameters of the SGD algorithm are unbounded. We prove that $(\mu^N)_{N\geq 1}$ (defined in (13)) is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. To this end we now use (Jakubowski, 1986, Theorem 4.6). The compact containment, which is the purpose of Lemma 20, is not straightforward since $\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ is not compact contrary to Theorem 2 where we used the compactness of $\mathcal{P}(\Theta_T)$. More precisely, the compact containment of $(\mu^N)_{N\geq 1}$ relies on a characterization of the compact subsets of $\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ (see Proposition 18) and moment estimates on $\{\theta_k^i, i \in \{1, \dots N\}\}_{k=0,\dots,\lfloor NT \rfloor}$ (see Lemma 17). We also mention that contrary to what is done in the proof of Theorem 2, we do not show that every limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ is continuous in time but we still manage to prove that they all satisfy (15). Then, using the duality formula for the W_1-distance together with rough estimates on the jumps of $t \mapsto \langle f, \mu_t^N \rangle$ (for f uniformly Lipschitz over $\mathbf{R}^{d+1}$), we then show that every limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ belongs a.s. to $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$. Again this is important since we have uniqueness of (15) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$.

We conclude this section with the following important uniqueness result.

Proposition 4 *Under the assumptions of Theorems 2 and 3, the solution to* (11) *is independent of* T *and is equal to the solution to* (15).

This uniqueness result states that both idealized and *Bayes-by-backprop* SGD have the same limiting behavior. It is also noteworthy that the mini-batch B is held fixed B. The effect of batch size can be seen at the level of the central limit theorem, which we leave for future work.

5. The *Minimal-VI* SGD algorithm

The idea behing the *Bayes-by-Backprop* SGD stems from the fact that there are integrals wrt γ in the loss function that cannot be computed in practice and it is quite natural up to a reparameterization trick, to replace these integrals by a Monte Carlo approximation (with i.i.d. gaussian random variables). To devise a new cheaper algorithm based on the only terms impacting the asymptotic limit, we directly analyse the limit equation (11) and remark that it can be rewritten as, $\forall f \in \mathcal{C}^\infty(\Theta_T)$ and $\forall t \in [0, T]$,

$$\begin{aligned}
&\langle f, \bar{\mu}_t \rangle - \langle f, \mu_0 \rangle \\
&= -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}\times(\mathbf{R}^d)^2} \big\langle \phi(\cdot, z_1, x) - y, \bar{\mu}_s \big\rangle \big\langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, z_2, x), \bar{\mu}_s \big\rangle \gamma^{\otimes 2}(\mathrm{d}z_1 \mathrm{d}z_2) \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\
&\quad - \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_s \big\rangle \mathrm{d}s.
\end{aligned}$$

Thus, the integration over $\gamma^{\otimes 2}$ can be considered as that over π, i.e., we can consider them as two more data variables that only need to be sampled at each new step. In this case, the SGD (7) becomes: for $k \geq 0$ and $i \in \{1, \dots, N\}$,

$$\begin{cases}
\theta_{k+1}^i = \theta_k^i - \dfrac{\eta}{N^2} \displaystyle\sum_{j=1}^N \big(\phi(\theta_k^j, \mathsf{Z}_k^1, x_k) - y_k\big) \nabla_\theta \phi(\theta_k^i, \mathsf{Z}_k^2, x_k) - \dfrac{\eta}{N} \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_{\theta_k^i}^1 | P_0^1) \\
\theta_0^i = (m_0^i, \rho_0^i) \sim \mu_0,
\end{cases} \tag{16}$$

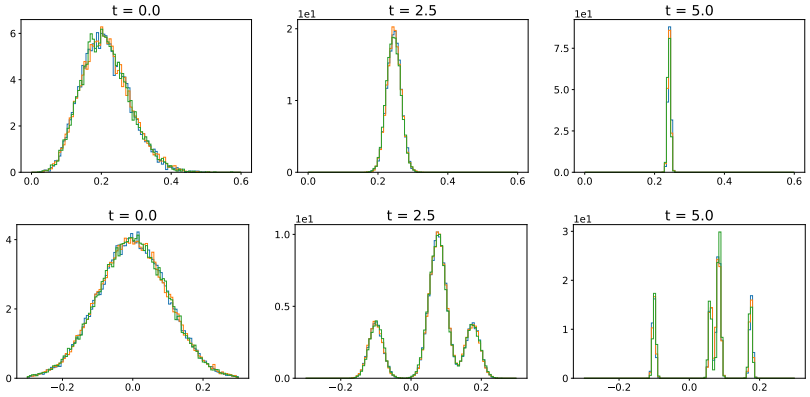

Figure 1: Histograms of $\{F(\theta^i_{\lfloor NT \rfloor}), i = 1, \dots, N\}$, at different times (initialization ($t = 0$), half ($t = 2.5$) and end of training ($T = 5$)), when $N = 10000$. First line: $F(\theta) = \|m\|_2$, where $\theta = (m, \rho) \in \mathbf{R}^d \times \mathbf{R}$. Second line: $F(\theta) = m \in \mathbf{R}^d$. Idealized (blue), *Bayes-by-Backprop* (orange) and *Minimal-VI* (green).

where $\eta > 0$ and $(\mathsf{Z}^p_k, p \in \{1, 2\}, k \geq 0)$ is a i.i.d sequence of random variables distributed according to $\gamma^{\otimes 2}$. We call this backpropagation scheme *minimal- VI SGD* which is much cheaper in terms of computational complexity, with the same limiting behavior as we now discuss.

We introduce the $\boldsymbol{\sigma}$-algebra for $N, k \geq 1$:

$$\mathcal{F}^N_k = \boldsymbol{\sigma}\Big(\theta^i_0, \mathsf{Z}^p_q, (x_q, y_q), 1 \leq i \leq N, p \in \{1, 2\}, 0 \leq q \leq k-1\Big\}\Big). \tag{17}$$

In addition to **A1**→**A4** (where in **A2**, $\mathcal{F}^N_k$ is now the one defined above in (17) when $k \geq 1$), the following assumption

A6. The sequences $(\mathsf{Z}^p_k, p \in \{1, 2\}, k \geq 0)$ and $((x_k, y_k), k \geq 0)$ are independent. In addition, for $k \geq 0$, $\big((x_k, y_k), \mathsf{Z}^p_k, p \in \{1, 2\}\big) \perp\!\!\!\perp \mathcal{F}^N_k$, where $\mathcal{F}^N_k$ is defined in (17).

Set for $k \geq 0$ and $t \geq 0$, $\nu^N_k := \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i_k}$ and $\mu^N_t := \nu^N_{\lfloor Nt \rfloor}$, where the θ^i_k's are defined in (16). The last main result of this work states that the sequence $(\mu^N)_{N\geq 1}$ satisfies the same law of large numbers when $N \to +\infty$ as the one satisfied by (13), whose proof will be omitted as it is the same as the one made for Theorem 3.

Theorem 5 *Assume* **A1**→**A4** *and* **A6**. *Then, the sequence of* $(\mu^N)_{N\geq 1}$ *satisfies all the statements of Theorem 3.*

6. Numerical experiments

In this section we illustrate the theorems 2, 3, and 5 using the following toy model. We set $d = 5$. Given $\theta^* \in \mathbf{R}^d$ (drawn from a normal distribution and scaled to the unit norm), we draw i.i.d observations as follows: Given $x \sim \mathcal{U}([-1, 1]^d)$, we draw $y = \tanh(x^\top \theta^*) + \epsilon$, where ϵ is

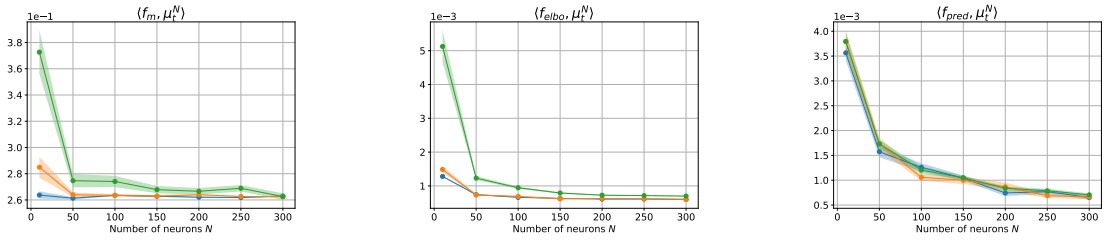

Figure 2: Convergence of $\langle f, \mu_T^N \rangle$ to $\langle f, \bar{\mu}_T \rangle$, for the idealized (blue), *Bayes-by-Backprop* (orange) and *Minimal-VI* (green) SGD algorithms over 50 realizations.

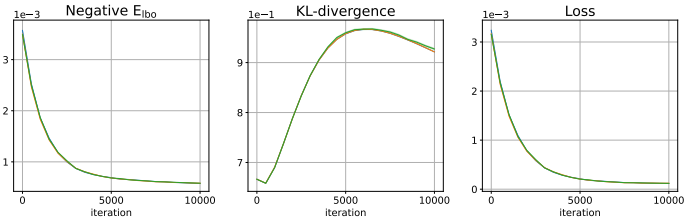

Figure 3: Decay of the negative ELBO (left) and its two components (KL (middle), loss (right)) during the training process done by the idealized (blue), *Bayes-by-Backprop* (orange) and *Minimal-VI* (green) SGD algorithms, for $N = 10000$.

zero mean with variance 10^{-4}. The initial distribution of parameters is centered around the prior: $\boldsymbol{\theta}_0 \sim (\mathcal{N}(m_0, 0.01 I_d) \times \mathcal{N}(g^{-1}(\sigma_0), 0.01))^{\otimes N}$, with $m_0 = 0$ and $\sigma_0 = 0.2$. Since the idealized algorithm cannot be implemented exactly, a mini-batch of size 100 is used as a proxy for the following comparisons of the different algorithms. For the algorithm (7) SGD we set $B = 1$.

Evolution and limit of the distribution Fig. 1 displays the histograms of $\{F(\theta^i_{\lfloor Nt \rfloor}), i = 1, \dots, N\}$ ($F(\theta) = \|m\|_2, g(\rho)$ or m, where $\theta = (m, \rho) \in \mathbf{R}^d \times \mathbf{R}$), for $N = 10000$, at initialization, halfway through training, and at the end of training. The empirical distributions illustrated by these histograms are very similar over the course of training. It can be seen that for $N = 10000$ the limit of the mean field is reached.

Convergence with respect to the numbers of neurons. We investigate here the speed of convergence of μ_t^N to $\bar{\mu}_t$ (as $N \to +\infty$), when tested against test functions f. More precisely, we fix a time T (end of training) and Figure 2 represents the empirical mean of $\langle f, \mu_T^N \rangle$ over 50 realizations. The test functions used for this experiment are $f_m(\theta) = \|m\|_2$, $f_{\mathrm{Elbo}(\theta)} = -\hat{\mathrm{E}}_{\mathrm{lbo}}(\theta)^N$ where $\hat{\mathrm{E}}_{\mathrm{lbo}}$ is the empirical $\mathrm{E}^N_{\mathrm{lbo}}$ (see (1)) computed with 100 samples of (x, y) and $(z^1, \dots, z^N)$. Finally, $f_{pred}(\theta) = \hat{\mathbb{E}}_x \left[\hat{\mathbb{V}}_{\boldsymbol{w} \sim q^N_{\boldsymbol{\theta}}} [f^N_{\boldsymbol{w}}(x)] \right]^{1/2}$ where $\hat{\mathbb{E}}$ and $\hat{\mathbb{V}}$ denote respectively the empirical mean and the empirical variance over 100 samples. All algorithms are converging to the same limit and are performing similarly even with a limited number of neurons ($N = 300$ in this example).

Convergence with respect to time. This section illustrates the training process of a BNN with a given number of neurons $N = 10000$. In Figure 3, we plot the negative ELBO on a test set and its two components, the loss and the KL-divergence terms. Figure 3 shows that the BNN is able to learn

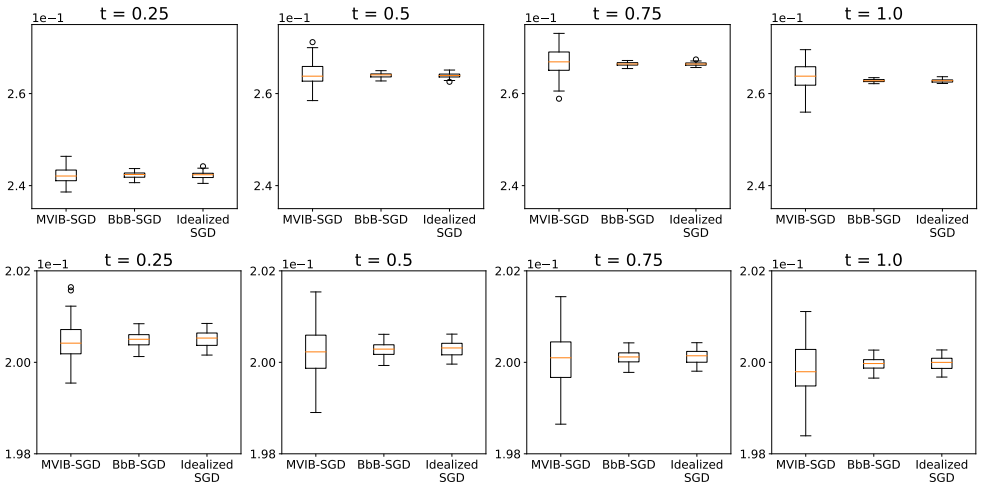

Figure 4: Boxplots for 50 runs of $\langle f, \mu_t^N \rangle$ for the three SGD schemes for $f(\theta) = \|m\|_2$ on the first line and $f(\theta) = g(\rho)$ on the second line. MVIB-SGD: *Minimal-VI* SGD. BbB-SGD: *Bayes-by-Backprop* SGD.

on this specific task and all algorithms exhibit a similar performance. It illustrates the trajectorial convergence of $\{\mu_t^N, t \in [0,T]\}_{N \geq 1}$ to $\{\bar{\mu}_t, t \in [0,T]\}$ as $N \to +\infty$.

Behavior around the limit $\bar{\mu}$. On Figure 4, we plot the boxplots of $\langle f, \mu_t^N \rangle$ for 50 realizations and $N = 10000$, at different times of the training. *Minimal-VI* scheme (which is computationally cheaper as explained in 5) exhibit a larger variance than the other algorithms.

7. Conclusion

By establishing the limit behavior of the idealized SGD for the variational inference of BNN with the weighting suggested by Huix et al. (2022), we have rigorously shown that the most-commonly used in practice *Bayes-by-Backprop* scheme indeed exhibits the same limit behavior. Furthermore, the analysis of the limit equation led us to validate the correct scaling of the KL divergence term in with respect to the loss. Notably, the mean-field limit dynamics has also helped us to devise a far less costly new SGD algorithm, the *Minimal-VI*. This scheme shares the same limit behavior, but only stems from the non-vanishing asymptotic contributions, hence the reduction of the computational cost. Aside from confirming the analytical results, the first simulations presented here show that the three algorithms, while having the same limit, may differ in terms of variance. Thus, deriving a CLT result and discussing the right trade-off between computational complexity and variance will be done in future work. Also, on a more general level regarding uncertainty quantification, an interesting question is to analyse the impact of the correct scaling of the KL divergence term on the error calibration and how to apply the same analysis in the context of deep ensembles.

Acknowledgments

A.D. is grateful for the support received from the Agence Nationale de la Recherche (ANR) of the French government through the program "Investissements d'Avenir" (16-IDEX-0001 CAP 20-25) A.G. is supported by the French ANR under the grant ANR-17-CE40-0030 (project *EFI*) and the Institut Universitaire de France. M.M. acknowledges the support of the the French ANR under the grant ANR-20-CE46-0007 (*SuSa* project). B.N. is supported by the grant IA20Nectoux from the Projet I-SITE Clermont CAP 20-25. E.M. and 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

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2nd edition, 1999.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/blundell15.html.

L. Chizat. Mean-field langevin dynamics: Exponential convergence and annealing, 2022. URL https://arxiv.org/abs/2202.01009.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

L. Chizat, M. Colombo, X. Fernandez-Real, and A. Figalli. Infinite-width limit of deep linear neural networks, 2022. URL https://arxiv.org/abs/2211.16980.

A. D. Cobb and B. Jalaian. Scaling hamiltonian monte carlo inference for bayesian neural networks with symmetric splitting. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 675–685. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/cobb21a.html.

Beau Coker, Wessel P. Bruinsma, David R. Burt, Weiwei Pan, and Finale Doshi-Velez. Wide mean-field bayesian neural networks ignore the data. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5276–5333. PMLR, 2022.

V. De Bortoli, A. Durmus, X. Fontaine, and U. Simsekli. Quantitative propagation of chaos for SGD in wide neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 278–288. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/02e74f10e0327ad868d138f2b4fdd6f0-Paper.pdf.

P. Del Moral and A. Guionnet. Central limit theorem for nonlinear filtering and interacting particle systems. *The Annals of Applied Probability*, 9(2):275–297, 1999.

F. Delarue, D. Lacker, and K. Ramanan. From the master equation to mean field game limit theory: a central limit theorem. *Electronic Journal of Probability*, 24:1–54, 2019.

A. Descours, A. Guillin, M. Michel, and B. Nectoux. Law of large numbers and central limit theorem for wide two-layer neural networks: the mini-batch and noisy case. *arXiv preprint arXiv:2207.12734*, 2022.

S. Ethier and T. Kurtz. *Markov Processes: Characterization and Convergence*, volume 282. John Wiley & Sons, 2009.

B. Fernandez and S. Méléard. A Hilbertian approach for fluctuations on the Mckean-Vlasov model. *Stochastic Processes and their Applications*, 71(1):33–53, 1997.

A. Filos, S. Farquhar, A. N. Gomez, T. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. De Kroon, and Y. Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

A. Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

Geoffrey Hinton and Drew Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, pages 5–13. ACM Press, 1993.

M. Hitsuda and I. Mitoma. Tightness problem and stochastic evolution equation arising from fluctuation phenomena for interacting diffusions. *Journal of Multivariate Analysis*, 19(2):311–328, 1986.

T. Huix, S. Majewski, A. Durmus, E. Moulines, and A. Korba. Variational inference of overparameterized bayesian neural networks: a theoretical and empirical study, 2022. URL https://arxiv.org/abs/2207.03859.

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. G. Wilson. What are bayesian neural network posteriors really like? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4629–4640. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/izmailov21a.html.

J. Jacod and A. Shiryaev. *Skorokhod Topology and Convergence of Processes*. Springer, 1987.

A. Jakubowski. On the skorokhod topology. In *Annales de l'IHP Probabilités et statistiques*, volume 22, pages 263–285, 1986.

B. Jourdain and S. Méléard. Propagation of chaos and fluctuations for a moderate model with smooth initial data. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 34(6): 727–766, 1998.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

M. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2611–2620. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/khan18a.html.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

M. Krzywinski and N. Altman. Importance of being uncertain. *Nature methods*, 10(9):809–811, 2013.

T. Kurtz and J. Xiong. A stochastic evolution equation arising from the fluctuations of a class of interacting particle systems. *Communications in Mathematical Sciences*, 2(3):325–358, 2004.

C. Louizos and M. Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2218–2227. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/louizos17a.html.

David JC MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469, 1995.

David JC MacKay et al. Ensemble learning and evidence maximization. In *Proc. Nips*, volume 10, page 4083. Citeseer, 1995.

R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, and A. Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In *IJCAI*, 2017.

S. Mei, A. Montanari, and P-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

R. Michelmore, M. Wicker, L. Laurenti, L. Cardelli, Y. Gal, and M. Kwiatkowska. Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7344–7350, 2020. doi: 10.1109/ICRA40945.2020.9196844.

V.M. Panaretos and Y. Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer Nature, 2020.

B. Piccoli and F. Rossi. On properties of the generalized Wasserstein distance. *Archive for Rational Mechanics and Analysis*, 222(3):1339–1365, 2016.

B. Piccoli, F. Rossi, and E. Trélat. Control to flocking of the kinetic Cucker–Smale model. *SIAM Journal on Mathematical Analysis*, 47(6):4685–4719, 2015.

G.M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *Preprint arXiv:1805.00915, to appear in Comm. Pure App. Math.*, 2018.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 55. Springer, 2015.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

A-S. Sznitman. Topics in propagation of chaos. In *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251. Springer, 1991. ISBN 978-3-540-46319-1.

C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

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

Appendix A. Proof of Theorem 2

For simplicity, we prove the theorem 2 when $T = 1$, and we denote Θ_1 simply by Θ. In this section we assume **A1**–**A4**.

A.1. Pre-limit equation (28) and error terms in (28)

A.1.1. Derivation of the pre-limit equation

The aim of this section is to establish the so-called pre-limit equation (28), which will be our starting point to derive Equation (11). Let $N \geq 1$, $k \in \{0, \dots, N\}$, and $f \in \mathcal{C}^\infty(\Theta)$. Recall that by Lemma 1 and since $0 \leq k \leq N$, a.s. $\theta_k^i \in \Theta$, and thus a.s. $f(\theta_k^i)$ is well-defined. The Taylor-Lagrange formula yields

$$\begin{aligned}\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle &= \frac{1}{N} \sum_{i=1}^N f(\theta_{k+1}^i) - f(\theta_k^i) \\ &= \frac{1}{N} \sum_{i=1}^N \nabla_\theta f(\theta_k^i) \cdot (\theta_{k+1}^i - \theta_k^i) + \frac{1}{2N} \sum_{i=1}^N (\theta_{k+1}^i - \theta_k^i)^T \nabla^2 f(\widehat{\theta}_k^i)(\theta_{k+1}^i - \theta_k^i),\end{aligned}$$

where, for all $i \in \{1, \dots, N\}$, $\widehat{\theta}_k^i \in (\theta_k^i, \theta_{k+1}^i) \subset \Theta$. Using (5), we then obtain

$$\begin{aligned}\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle = &- \frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N \big(\langle \phi(\theta_k^j, \cdot, x_k), \gamma \rangle - y_k \big) \big\langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x_k), \gamma \big\rangle \\ &- \frac{\eta}{N^2} \big\langle \big(\phi(\cdot, \cdot, x_k) - y_k \big) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x_k), \nu_k^N \otimes \gamma \big\rangle \\ &- \frac{\eta}{N} \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \nu_k^N \big\rangle + \mathbf{R}_k^N[f], \end{aligned} \tag{18}$$

where

$$\mathbf{R}_k^N[f] := \frac{1}{2N} \sum_{i=1}^N (\theta_{k+1}^i - \theta_k^i)^T \nabla^2 f(\widehat{\theta}_k^i)(\theta_{k+1}^i - \theta_k^i). \tag{19}$$

Let us define

$$\begin{aligned}\mathbf{D}_k^N[f] := &\mathbf{E}\Big[- \frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N \big(\langle \phi(\theta_k^j, \cdot, x_k), \gamma \rangle - y_k \big) \big\langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x_k), \gamma \big\rangle \Big| \mathcal{F}_k^N \Big] \\ &- \mathbf{E}\Big[\frac{\eta}{N^2} \big\langle (\phi(\cdot, \cdot, x_k) - y_k) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x_k), \nu_k^N \otimes \gamma \big\rangle \Big| \mathcal{F}_k^N \Big]. \end{aligned} \tag{20}$$

Note that using (45) and (47) together with the fact that $|\nabla_\theta f(\theta_k^i)| \leq \sup_{\theta \in \Theta} |\nabla_\theta f(\theta)|$, the integrant in (20) is integrable and thus $\mathbf{D}_k^N[f]$ is well defined. Using the fact that $(x_k, y_k) \perp\!\!\!\perp \mathcal{F}_k^N$ by **A2** and that $\{\theta_k^i, i = 1, \dots, N\}$ is $\mathcal{F}_k^N$-measurable by (5), we have:

$$\begin{aligned}\mathbf{D}_k^N[f] = &- \frac{\eta}{N^3} \sum_{i=1}^N \sum_{j=1, j\neq i}^N \int_{\mathsf{X}\times\mathsf{Y}} \big(\langle \phi(\theta_k^j, \cdot, x), \gamma \rangle - y \big) \big\langle \nabla_\theta f(\theta_k^i) \cdot \nabla_\theta \phi(\theta_k^i, \cdot, x), \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y) \\ &- \frac{\eta}{N^2} \int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot, \cdot, x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \nu_k^N \otimes \gamma \big\rangle \pi(\mathrm{d}x, \mathrm{d}y). \end{aligned} \tag{21}$$

Introduce also

$$\mathbf{M}_k^N[f] := -\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1,j\neq i}^N (\langle\phi(\theta_k^j,\cdot,x_k),\gamma\rangle - y_k)\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x_k),\gamma\rangle$$
$$-\frac{\eta}{N^2}\langle(\phi(\cdot,\cdot,x_k)-y_k)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x_k),\nu_k^N\otimes\gamma\rangle - \mathbf{D}_k^N[f].$$

Note that $\mathbf{E}\big[\mathbf{M}_k^N[f]|\mathcal{F}_k^N\big]=0$. Equation (18) then writes

$$\langle f,\nu_{k+1}^N\rangle - \langle f,\nu_k^N\rangle = \mathbf{D}_k^N[f]+\mathbf{M}_k^N[f]-\frac{\eta}{N}\big\langle\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1),\nu_k^N\big\rangle+\mathbf{R}_k^N[f]. \tag{22}$$

Notice also that

$$\begin{aligned}
\mathbf{D}_k^N[f] &= -\frac{\eta}{N^3}\sum_{i=1}^N\sum_{j=1}^N\int_{\mathsf{X}\times\mathsf{Y}}(\langle\phi(\theta_k^j,\cdot,x),\gamma\rangle-y)\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&\quad+\frac{\eta}{N^3}\sum_{i=1}^N\int_{\mathsf{X}\times\mathsf{Y}}(\langle\phi(\theta_k^i,\cdot,x),\gamma\rangle-y)\langle\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i,\cdot,x),\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&\quad-\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&=-\frac{\eta}{N}\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\nu_k^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&\quad+\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle(\langle\phi(\cdot,\cdot,x),\gamma\rangle-y)\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\nu_k^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\\
&\quad-\frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}}\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\nu_k^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y).
\end{aligned}\tag{23}$$

Now, we define for $t\in[0,1]$:

$$\mathbf{D}_t^N[f] := \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{D}_k^N[f],\quad \mathbf{R}_t^N[f]:=\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{R}_k^N[f],\quad\text{and}\quad \mathbf{M}_t^N[f]:=\sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{M}_k^N[f]. \tag{24}$$

We can rewrite $\mathbf{D}_t^N[f]$ has follows:

$$\mathbf{D}_t^N[f]=\sum_{k=0}^{\lfloor Nt\rfloor-1}\int_{\frac{k}{N}}^{\frac{k+1}{N}}N\mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s = N\int_0^t\mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s - N\int_{\frac{\lfloor Nt\rfloor}{N}}^t\mathbf{D}_{\lfloor Ns\rfloor}^N[f]\mathrm{d}s.$$

Since $\nu_{\lfloor Ns\rfloor}^N=\mu_s^N$ (by definition, see (8)), we have, using also (23) with $k=\lfloor Ns\rfloor$,

$$\begin{aligned}
\mathbf{D}_t^N[f] &= -\eta\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\langle\phi(\cdot,\cdot,x)-y,\mu_s^N\otimes\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad+\frac{\eta}{N}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle\langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle,\mu_s^N\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s\\
&\quad-\frac{\eta}{N}\int_0^t\int_{\mathsf{X}\times\mathsf{Y}}\Big\langle(\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\mu_s^N\otimes\gamma\Big\rangle\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s-\mathbf{V}_t^N[f],
\end{aligned}\tag{25}$$

where

$$\mathbf{V}_t^N[f] := -\eta \int_{\frac{\lfloor Nt\rfloor}{N}}^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$+ \frac{\eta}{N} \int_{\frac{\lfloor Nt\rfloor}{N}}^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$- \frac{\eta}{N} \int_{\frac{\lfloor Nt\rfloor}{N}}^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.$$

On the other hand, we also have for $t \in [0,1]$,

$$\sum_{k=0}^{\lfloor Nt\rfloor - 1} -\frac{\eta}{N} \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \nu_k^N \big\rangle = -\eta \int_0^{\frac{\lfloor Nt\rfloor}{N}} \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \big\rangle \mathrm{d}s. \tag{26}$$

We finally set:

$$\mathbf{W}_t^N[f] := -\mathbf{V}_t^N[f] + \eta \int_{\frac{\lfloor Nt\rfloor}{N}}^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \big\rangle \mathrm{d}s. \tag{27}$$

Since $\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle = \sum_{k=0}^{\lfloor Nt\rfloor - 1} \langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle$, we deduce from (22), (24), (25), (26) and (27), the so-called pre-limit equation satisfied by μ^N: for $N \geq 1$, $t \in [0,1]$, and $f \in \mathcal{C}^\infty(\Theta)$,

$$\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle = -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$- \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N \big\rangle \mathrm{d}s$$
$$+ \frac{\eta}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$- \frac{\eta}{N} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$+ \mathbf{M}_t^N[f] + \mathbf{W}_t^N[f] + \mathbf{R}_t^N[f]. \tag{28}$$

A.1.2. THE LAST FIVE TERMS IN (28) ARE ERROR TERMS

The purpose of this section is to show that the last five terms appearing in the r.h.s. of (28) are error terms when $N \to +\infty$. For $J \in \mathbf{N}^*$ and $f \in \mathcal{C}^J(\Theta)$, set $\|f\|_{\mathcal{C}^J(\Theta)} := \sum_{|k|\leq J} \|\partial_k f\|_{\infty,\Theta}$, where $\|g\|_{\infty,\Theta} = \sup_{\theta\in\Theta} |g(\theta)|$ for $g : \Theta \to \mathbf{R}^m$.

Lemma 6 (Error terms) *Assume* **A1**→**A4**. *Then, there exists $C > 0$ such that a.s. for all $f \in \mathcal{C}^\infty(\Theta)$ and $N \geq 1$,*

1. $\frac{\eta}{N} \int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \gamma\rangle, \mu_s^N \Big\rangle \Big| \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \leq C\|f\|_{\mathcal{C}^1(\Theta)}/N.$
2. $\frac{\eta}{N} \int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big| \Big\langle (\phi(\cdot,\cdot,x) - y) \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N \otimes \gamma \Big\rangle \Big| \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \leq C\|f\|_{\mathcal{C}^1(\Theta)}/N.$

3. $\sup_{t\in[0,1]} |\mathbf{W}_t^N[f]| + \sup_{t\in[0,1]} |\mathbf{R}_t^N[f]| \le C\|f\|_{\mathcal{C}^2(\Theta)}/N$.

Finally, $\sup_{t\in[0,1]} \mathbf{E}\big[|\mathbf{M}_t^N[f]|\big] \le C\|f\|_{\mathcal{C}^1(\Theta)}/\sqrt{N}$.

Proof All along the proof, $C > 0$ denotes a positive constant independent of $N \ge 1, k \in \{0,\dots,N-1\}, (s,t) \in [0,1]^2, (x,y) \in \mathsf{X}\times\mathsf{Y}, \theta \in \Theta, z \in \mathbf{R}^d$, and $f \in \mathcal{C}^\infty(\Theta)$ which can change from one occurrence to another. Using (47), the Cauchy-Schwarz inequality, and the fact that $\nabla_\theta f$ is bounded over Θ imply:

$$|\langle \nabla_\theta f(\theta)\cdot\nabla_\theta\phi(\theta,\cdot,x),\gamma\rangle| \le \langle |\nabla_\theta f(\theta)\cdot\nabla_\theta\phi(\theta,\cdot,x)|,\gamma\rangle \le C\|f\|_{\mathcal{C}^1(\Theta)}. \tag{29}$$

Combining (45) and (29), we obtain:

$$\int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big|\Big\langle \langle\phi(\cdot,\cdot,x)-y,\gamma\rangle\langle\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle, \mu_s^N\Big\rangle\Big|\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \le C\|f\|_{\mathcal{C}^1(\Theta)}$$

and

$$\int_0^1 \int_{\mathsf{X}\times\mathsf{Y}} \Big|\Big\langle (\phi(\cdot,\cdot,x)-y)\nabla_m f\cdot\nabla_m\phi(\cdot,\cdot,x), \mu_s^N\otimes\gamma\Big\rangle\Big|\pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \le C\|f\|_{\mathcal{C}^1(\Theta)},$$

which proves Items 1 and 2.

Let us now prove Item 3. By (45) and (29), $\sup_{t\in[0,1]} |\mathbf{V}_t^N[f]| \le C\|f\|_{\mathcal{C}^1(\Theta)}/N$. On the other hand, because $f \in \mathcal{C}^\infty(\Theta)$ and $\theta \mapsto \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)$ is continuous (see (4)) over Θ which is compact, it holds, $\|\nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)\|_{\infty,\Theta} < +\infty$. Hence, it holds:

$$\sup_{t\in[0,1]} \Big| \int_{\frac{\lfloor Nt\rfloor}{N}}^t \big\langle \nabla_\theta f\cdot\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mu_s^N\big\rangle \mathrm{d}s\Big| \le C\|f\|_{\mathcal{C}^1(\Theta)}/N.$$

Using (27), it then holds $\sup_{t\in[0,1]} |\mathbf{W}_t^N[f]| \le C\|f\|_{\mathcal{C}^1(\Theta)}/N$. Since $f \in \mathcal{C}^\infty(\Theta)$, we have, by (19), for $N \ge 1$ and $0 \le k \le N-1$, $|\mathbf{R}_k^N[f]| \le \|f\|_{\mathcal{C}^2(\Theta)} \frac{C}{N}\sum_{i=1}^N |\theta_{k+1}^i - \theta_k^i|^2$. By (48) and Lemma 1, $|\theta_{k+1}^i - \theta_k^i|^2 \le C/N^2$ and consequently, one has:

$$|\mathbf{R}_k^N[f]| \le C\|f\|_{\mathcal{C}^2(\Theta)}/N^2. \tag{30}$$

Hence, for all $t \in [0,1]$, $|\mathbf{R}_t^N[f]| \le C\|f\|_{\mathcal{C}^2(\Theta)}/N$. This proves Item 3.

Let us now prove the last item in Lemma 6. Let $t \in [0,1]$. We have, by (24),

$$|\mathbf{M}_t^N[f]|^2 = \sum_{k=0}^{\lfloor Nt\rfloor-1} \big|\mathbf{M}_k^N[f]\big|^2 + 2\sum_{k<j} \mathbf{M}_k^N[f]\,\mathbf{M}_j^N[f].$$

For all $0 \le k < j < \lfloor Nt\rfloor$, $\mathbf{M}_k^N[f]$ is $\mathcal{F}_j^N$-measurable (see (9)), and since $\mathbf{E}\big[\mathbf{M}_j^N[f]|\mathcal{F}_j^N\big] = 0$, one deduces that $\mathbf{E}\big[\,\mathbf{M}_k^N[f]\,\mathbf{M}_j^N[f]\,\big] = \mathbf{E}\big[\mathbf{M}_k^N[f]\,\mathbf{E}\big[\mathbf{M}_j^N[f]|\mathcal{F}_j^N\big]\,\big] = 0$. Hence, $\mathbf{E}[|\mathbf{M}_t^N[f]|^2] = \sum_{k=0}^{\lfloor Nt\rfloor-1}\mathbf{E}[|\mathbf{M}_k^N[f]|^2]$. By (45) and (29), one has a.s. for all $0 \le k \le N-1$,

$$|\mathbf{M}_k^N[f]| \le C\|f\|_{\mathcal{C}^1(\Theta)}/N. \tag{31}$$

Hence, $\mathbf{E}[|\mathbf{M}_t^N[f]|^2] \le C\|f\|_{\mathcal{C}^1(\Theta)}/N$, which proves the last inequality in Lemma 6. ■

A.2. Convergence to the limit equation as $N \to +\infty$

In this section we prove the relative compactness of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$. We then show that any of its limit points satisfies the limit equation (11).

A.2.1. WASSERSTEIN SPACES AND DUALITY FORMULA

In this section we recall some basic results which will be used throughout this work on the space $\mathcal{P}(\mathcal{S})$ when $(\mathcal{S},\mathsf{d})$ is a Polish space. First when endowed with the weak convergence topology, $\mathcal{P}(\mathcal{S})$ is a Polish space (Billingsley, 1999, Theorem 6.8). In addition, $\mathcal{P}_q(\mathcal{S}) = \{\nu \in \mathcal{P}(\mathcal{S}), \int_{\mathcal{S}} \mathsf{d}(w_0,w)^q \nu(\mathrm{d}w) < +\infty\}$, where $w_0 \in \mathcal{S}$ is arbitrary (note that this space was defined previously in (14) when $\mathcal{S} = \mathbf{R}^{d+1}$) when endowed with the W_q metric is also a Polish space (Villani, 2009, Theorem 6.18). Recall also the duality formula for the W_1-distance on $\mathcal{P}_1(\mathcal{S})$ (see e.g (Villani, 2009, Remark 6.5)):

$$\mathsf{W}_1(\mu,\nu) = \sup\Big\{\Big|\int_{\mathcal{S}} f(w)\mathrm{d}\mu(w) - \int_{\mathcal{S}} f(w)\nu(\mathrm{d}w)\Big|,\ \|f\|_{\mathrm{Lip}} \leq 1\Big\}. \tag{32}$$

Finally, when $\mathcal{K} \subset \mathbf{R}^{d+1}$ is compact, the convergence in W_q-distance is equivalent to the usual weak convergence on $\mathcal{P}(\mathcal{K})$ (see e.g. (Villani, 2009, Corollary 6.13)).

A.2.2. RELATIVE COMPACTNESS

The main result of this section is to prove that $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$, which is the purpose of Proposition 8 below. To this end, we need to prove that for all $f \in \mathcal{C}^\infty(\Theta)$, every sequence $(\langle f,\mu_t^N\rangle)_{N\geq 1}$ satisfies some regularity conditions, which is the purpose of the next result.

Lemma 7 (Regularity condition) *Assume* **A1**→**A4**. *Then there exists $C>0$ such that a.s. for all $f \in \mathcal{C}^\infty(\Theta)$, $0 \leq r < t \leq 1$, and $N \geq 1$:*

$$|\langle f,\mu_t^N\rangle - \langle f,\mu_r^N\rangle| \leq C\|f\|_{\mathcal{C}^2(\Theta)}\Big[|t-r| + \frac{|t-r|}{N} + \frac{1}{N}\Big]. \tag{33}$$

Proof Let $f \in \mathcal{C}^\infty(\Theta)$ and let $N \geq 1$ and $0 \leq r < t \leq 1$. In the following $C>0$ is a positive constant independent of $f \in \mathcal{C}^\infty(\Theta)$, $N \geq 1$, and $0 \leq r < t \leq 1$, which can change from one occurrence to another. From (28), we have

$$\begin{aligned}\langle f,\mu_t^N\rangle - \langle f,\mu_r^N\rangle = \mathbf{A}_{r,t}^N[f] - \eta\int_r^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mu_s^N\big\rangle \mathrm{d}s \\ + \mathbf{M}_t^N[f] - \mathbf{M}_r^N[f] + \mathbf{W}_t^N[f] - \mathbf{W}_r^N[f] + \mathbf{R}_t^N[f] - \mathbf{R}_r^N[f],\end{aligned} \tag{34}$$

where

$$\begin{aligned}\mathbf{A}_{r,t}^N[f] = -\eta\int_r^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N\otimes\gamma\rangle\langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \mu_s^N\otimes\gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y) \\ + \frac{\eta}{N}\int_r^t\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x)-y,\gamma\rangle\langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle, \mu_s^N\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y) \\ - \frac{\eta}{N}\int_r^t\int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x)-y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \mu_s^N\otimes\gamma\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y).\end{aligned}$$

By (45) and (29), $|\mathbf{A}^N_{r,t}[f]| \leq C\|f\|_{\mathcal{C}^1(\Theta)}\big[|t-r| + \frac{|t-r|}{N}\big]$. In addition, since $\theta \mapsto \mathscr{D}_{\mathrm{KL}}(q^1_\theta|P^1_0)$ is bounded over Θ (since it is smooth and Θ is compact),

$$\Big|\int_r^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_\cdot|P^1_0), \mu^N_s\rangle \mathrm{d}s\Big| \leq C\|f\|_{\mathcal{C}^1(\Theta)}|t-r|.$$

Furthermore, using (31),

$$|\mathbf{M}^N_t[f] - \mathbf{M}^N_r[f]| = \Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{M}^N_k[f]\Big| \leq (\lfloor Nt\rfloor - \lfloor Nr\rfloor)C\|f\|_{\mathcal{C}^1(\Theta)}/N.$$

Next, we have, by Item 3 in Lemma 6, $|\mathbf{W}^N_t[f] - \mathbf{W}^N_r[f]| \leq |\mathbf{W}^N_t[f]| + |\mathbf{W}^N_r[f]| \leq C\|f\|_{\mathcal{C}^2(\Theta)}/N$. Finally, by (30),

$$|\mathbf{R}^N_t[f] - \mathbf{R}^N_r[f]| = \Big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1} \mathbf{R}^N_k[f]\Big| \leq (\lfloor Nt\rfloor - \lfloor Nr\rfloor)C\|f\|_{\mathcal{C}^2(\Theta)}/N^2.$$

The proof of Proposition 7 is complete plugging all the previous estimates in (34). ■

Proposition 8 (Relative compactness) *Assume* **A1**→**A4**. *Then, the sequence* $(\mu^N)_{N\geq 1}$ *is relatively compact in* $\mathcal{D}([0,1],\mathcal{P}(\Theta))$.

Proof The proof consists in applying (Jakubowski, 1986, Theorem 3.1) with $E = \mathcal{P}(\Theta)$ endowed with the weak convergence topology. Set $\mathbb{F} = \{\mathfrak{L}_f, f \in \mathcal{C}^\infty(\Theta)\}$ where

$$\mathsf{L}_f : \nu \in \mathcal{P}(\Theta) \mapsto \langle f, \nu\rangle.$$

The class of continuous functions $\mathbb{F}$ on $\mathcal{P}(\Theta)$ satisfies Conditions (Jakubowski, 1986, (3.1) and (3.2) in Theorem 3.1).

On the other hand, the condition (Jakubowski, 1986, (3.3) in Theorem 3.1) is satisfied since $\mathcal{P}(\Theta)$ is compact because Θ is compact (see e.g. (Panaretos and Zemel, 2020, Corollary 2.2.5) together with (Villani, 2009, Corollary 6.13)).

It remains to verify Condition (3.4) of (Jakubowski, 1986, Theorem 3.1), i.e. that for all $f \in \mathcal{C}^\infty(\Theta)$, $(\langle f, \mu^N\rangle)_{N\geq 1}$ is relatively compact in $\mathcal{D}([0,1],\mathbf{R})$. To this end, we apply (Billingsley, 1999, Theorem 13.2). Condition (i) in (Billingsley, 1999, Theorem 13.2) is satisfied because $|\langle f, \mu^N_t\rangle| \leq \|f\|_{\infty,\Theta}$ for all $t \in [0,1]$ and $N \geq 1$. Let us now show that Condition (ii) in (Billingsley, 1999, Theorem 13.2) holds. For this purpose, we use Lemma 7. For $\delta, \beta > 0$ sufficiently small, it is possible to construct a subdivision $\{t_i\}_{i=0}^v$ of $[0,1]$ such that $t_0 = 0$, $t_v = 1$, $t_{i+1} - t_i = \delta + \beta$ for $i \in \{0,\ldots,v-2\}$ and $\delta + \beta \leq t_v - t_{v-1} \leq 2(\delta+\beta)$. According to the terminology introduced in (Billingsley, 1999, Section 12), $\{t_i\}_{i=0}^v$ is δ-sparse. Then, by Lemma 7, there exists $C > 0$ such that a.s. for all $\delta, \beta > 0$, all such subdivision $\{t_i\}_{i=0}^v$, $i \in \{0,\ldots,v-1\}$, and $N \geq 1$,

$$\sup_{t,r\in[t_i,t_{i+1}]} |\langle f, \mu^N_t\rangle - \langle f, \mu^N_r\rangle| \leq C\Big(|t_{i+1}-t_i| + \frac{|t_{i+1}-t_i|}{N} + \frac{1}{N}\Big) \leq C\Big(2(\delta+\beta) + \frac{2(\delta+\beta)}{N} + \frac{1}{N}\Big).$$

Thus, one has:

$$\inf_{\beta>0} \max_i \sup_{t,r\in[t_i,t_{i+1}]} |\langle f,\mu_t^N\rangle - \langle f,\mu_r^N\rangle| \le C\Big(2\delta + \frac{2\delta}{N} + \frac{1}{N}\Big).$$

Consequently, there exists $C>0$ such that a.s. for all $\delta>0$ small enough and $N\ge 1$,

$$w'_{\langle f,\mu^N\rangle}(\delta) := \inf_{\substack{\{t_i\}\\ \delta\text{-sparse}}} \max_i \sup_{t,r\in[t_i,t_{i+1}]} |\langle f,\mu_t^N\rangle - \langle f,\mu_r^N\rangle| \le C\Big(2\delta + \frac{2\delta}{N} + \frac{1}{N}\Big).$$

This implies $\lim_{\delta\to 0}\limsup_{N\to+\infty}\mathbf{E}[w'_{\langle f,\mu^N\rangle}(\delta)]=0$. By Markov's inequality, this proves Condition (ii) of (Billingsley, 1999, Theorem 13.2). Therefore, for all $f\in\mathcal{C}^\infty(\Theta)$, using also Prokhorov theorem, the sequence $(\langle f,\mu^N\rangle)_{N\ge1}\subset\mathcal{D}([0,1],\mathbf{R})$ is relatively compact. In conclusion, according to (Jakubowski, 1986, Theorem 3.1), $(\mu^N)_{N\ge1}\subset\mathcal{D}([0,1],\mathcal{P}(\Theta))$ is tight. ■

A.2.3. LIMIT POINTS SATISFY THE LIMIT EQUATION (11)

In this section we prove that every limit point of $(\mu^N)_{N\ge1}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$ satisfies (11).

Lemma 9 *Let $\mathsf{m},(\mathsf{m}^N)_{N\ge1}\subset\mathcal{D}([0,1],\mathcal{P}(\Theta))$ be such that $\mathsf{m}^N\to\mathsf{m}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$. Then, for all Lipschitz continuous function $f:\Theta\to\mathbf{R}$, we have $\langle f,\mathsf{m}^N\rangle\to\langle f,\mathsf{m}\rangle$ in $\mathcal{D}([0,1],\mathbf{R})$.*

Proof Let f be such a function. By (Billingsley, 1999, p.124), $\mathsf{m}^N\to\mathsf{m}$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$ iff there exist functions $\lambda_N:[0,1]\to[0,1]$ continuous, increasing onto itself such that $\sup_{t\in[0,1]}|\lambda_N(t)-t|\to_{N\to\infty}0$ and $\sup_{t\in[0,1]}\mathsf{W}_1(\mathsf{m}^N_{\lambda_N(t)},\mathsf{m}_t)\to_{N\to\infty}0$. Then $\langle f,\mathsf{m}^N\rangle\to\langle f,\mathsf{m}\rangle$ in $\mathcal{D}([0,1],\mathbf{R})$ since by (32), $\sup_{t\in[0,1]}|\langle f,\mathsf{m}^N_{\lambda_N(t)}\rangle-\langle f,\mathsf{m}_t\rangle|\le\|f\|_{\mathrm{Lip}}\sup_{t\in[0,1]}\mathsf{W}_1(\mathsf{m}^N_{\lambda_N(t)},\mathsf{m}_t)\to_{N\to\infty}0$. ■

Proposition 10 (Continuity of the limit points of $\langle f,\mu^N\rangle$) *Let $f\in\mathcal{C}^\infty(\Theta)$. Then, any limit point of $(\langle f,\mu^N\rangle)_{N\ge1}\subset\mathcal{D}([0,1],\mathbf{R})$ belong a.s. to $\mathcal{C}([0,1],\mathbf{R})$.*

Proof Fix $t\in(0,1]$. Letting $r\to t$ in (33), we obtain $|\langle f,\mu_t^N\rangle-\langle f,\mu_{t^-}^N\rangle|\le C/N$. Therefore $\sup_{t\in(0,1]}|\langle f,\mu_t^N\rangle-\langle f,\mu_{t^-}^N\rangle|\xrightarrow{\mathscr{D}}0$ as $N\to+\infty$. The result follows from (Billingsley, 1999, Theorem 13.4). ■

Proposition 11 (Continuity of the limit points of μ^N) *Let $\mu^*\in\mathcal{D}([0,1],\mathcal{P}(\Theta))$ be a limit point of $(\mu^N)_{N\ge1}\subset\mathcal{D}([0,1],\mathcal{P}(\Theta))$. Then, a.s. $\mu^*\in\mathcal{C}([0,1],\mathcal{P}(\Theta))$.*

Proof Up to extracting a subsequence, we assume that $\mu^N\xrightarrow{\mathscr{D}}\mu^*$. By Skorohod representation theorem, there exists another probability space $(\hat\Omega,\hat{\mathcal{F}},\hat{\mathbf{P}})$ on which are defined random elements $(\hat\mu^N)_{N\ge1}$ and $\hat\mu^*$, where,

$$\hat\mu^*\overset{\mathscr{D}}{=}\mu^*,\quad\text{and for all } N\ge1,\ \hat\mu^N\overset{\mathscr{D}}{=}\mu^N,$$

and such that $\hat{\mathbf{P}}$-a.s., $\hat{\mu}^N \to \hat{\mu}^*$ in $\mathcal{D}([0,1],\mathcal{P}(\Theta))$ as $N \to +\infty$. Fix $f \in \mathcal{C}^\infty(\Theta)$. We have, by Lemma 9,

$$\hat{\mathbf{P}}-\text{a.s.}, \ \langle f, \hat{\mu}^N\rangle \to_{N\to+\infty} \langle f, \hat{\mu}^*\rangle \ \text{ in } \ \mathcal{D}([0,1],\mathbf{R}).$$

In particular, $\langle f, \hat{\mu}^N\rangle \to_{N\to+\infty} \langle f, \hat{\mu}^*\rangle$ in distribution. By Proposition 10, there exists $\hat{\Omega}_f \subset \hat{\Omega}$ of $\hat{\mathbf{P}}$-mass 1 such that for all $\omega \in \hat{\Omega}_f$, $\langle f, \hat{\mu}^*(\omega)\rangle \in \mathcal{C}([0,1],\mathbf{R})$. Denote by $\mathscr{F}$ the class polynomial functions with rational coefficients. Since this class is countable, the set $\hat{\Omega}_{\mathscr{F}} := \cap_{f\in\mathscr{F}}\hat{\Omega}_f$ is of $\hat{\mathbf{P}}$-mass 1. Consider now an arbitrary $f \in \mathcal{C}(\Theta)$ and let us show that for all $\omega \in \hat{\Omega}_{\mathscr{F}}$, $\langle f, \hat{\mu}^*(\omega)\rangle \in \mathcal{C}([0,1],\mathbf{R})$. By the Stone-Weierstrass theorem, there exist $(f_n)_{n\geq 1} \subset \mathscr{F}$ such that $\|f_n - f\|_{\infty,\Theta} \to_{n\to+\infty} 0$. On $\hat{\Omega}_{\mathscr{F}}$, for all n, $t \in [0,1] \mapsto \langle f_n, \hat{\mu}_t^*\rangle$ is continuous and converges uniformly to $t \in [0,1] \mapsto \langle f, \hat{\mu}_t^*\rangle$. Hence, for all $\omega \in \hat{\Omega}_{\mathscr{F}}$ and $f \in \mathcal{C}(\Theta)$, $\langle f, \hat{\mu}^*(\omega)\rangle \in \mathcal{C}([0,1],\mathbf{R})$, i.e. for all $\omega \in \hat{\Omega}_{\mathscr{F}}$, $\hat{\mu}^*(\omega) \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$. This concludes the proof. ■

Now, we introduce, for $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, the function $\mathbf{\Lambda}_t[f] : \mathcal{D}([0,1],\mathcal{P}(\Theta)) \to \mathbf{R}_+$ defined by:

$$\begin{aligned}\mathbf{\Lambda}_t[f] : \mathsf{m} \mapsto \Big| &\langle f, \mathsf{m}_t\rangle - \langle f, \mu_0\rangle \\ &+ \eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\ &+ \eta \int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mathsf{m}_s\big\rangle \mathrm{d}s\Big|. \end{aligned} \tag{35}$$

We now study the continuity of $\mathbf{\Lambda}_t[f]$.

Lemma 12 *Let $(\mathsf{m}^N)_{N\geq 1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$ converge to $\mathsf{m} \in \mathcal{D}([0,1],\mathcal{P}(\Theta))$. Then, for all continuity point $t \in [0,1]$ of m and all $f \in \mathcal{C}^\infty(\Theta)$, we have $\mathbf{\Lambda}_t[f](\mathsf{m}^N) \to \mathbf{\Lambda}_t[f](\mathsf{m})$.*

Proof Let $f \in \mathcal{C}^\infty(\Theta)$ and denote by $\mathcal{C}(\mathsf{m}) \subset [0,1]$ the set of continuity points of m. Let $t \in \mathcal{C}(\mathsf{m})$. From (Billingsley, 1999, p. 124), we have, for all $s \in \mathcal{C}(\mathsf{m})$,

$$\mathsf{m}_s^N \to \mathsf{m}_s \ \text{ in } \ \mathcal{P}(\Theta). \tag{36}$$

Thus, $\langle f, \mathsf{m}_t^N\rangle \to_{N\to\infty} \langle f, \mathsf{m}_t\rangle$. For all $z \in \mathbf{R}^d$ and $(x,y) \in \mathsf{X}\times\mathsf{Y}$, **A1** and **A3** ensure that the functions $\theta \in \Theta \mapsto \phi(\theta,z,x) - y$ and $\theta \in \Theta \mapsto \nabla_\theta f(\theta)\cdot\nabla_\theta\phi(\theta,z,x)$ are continuous and also bounded because Θ is compact. Hence, for all $s \in [0,t] \cap \mathcal{C}(\mathsf{m})$, using (36),

$$\langle \phi(\cdot,z,x) - y, \mathsf{m}_s^N\rangle \to \langle \phi(\cdot,z,x) - y, \mathsf{m}_s\rangle \ \text{ and } \ \langle \nabla_\theta f \cdot \nabla_\theta\phi(\cdot,z,x), \mathsf{m}_s^N\rangle \to \langle \nabla_\theta f \cdot \nabla_\theta\phi(\cdot,z,x), \mathsf{m}_s\rangle$$

Since $[0,1]\backslash\mathcal{C}(\mathsf{m})$ is at most countable (see (Billingsley, 1999, p. 124)) we have that for a.e. $(s,z',z,x,y) \in [0,t]\times\mathbf{R}^d\times\mathbf{R}^d\times\mathsf{X}\times\mathsf{Y}$,

$$\langle \phi(\cdot,z',x) - y, \mathsf{m}_s^N\rangle\langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,z,x), \mathsf{m}_s^N\rangle \to \langle \phi(\cdot,z',x) - y, \mathsf{m}_s\rangle\langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,z,x), \mathsf{m}_s\rangle.$$

Since $\phi(\theta,z',x) - y$ is bounded and by (46), there exists $C > 0$ such that for all $(s,z',z,x,y) \in [0,t]\times\mathbf{R}^d\times\mathbf{R}^d\times\mathsf{X}\times\mathsf{Y}$, $\langle |\phi(\cdot,z',x) - y|, \mathsf{m}_s^N\rangle\langle |\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,z,x)|, \mathsf{m}_s^N\rangle \leq C\|\nabla_\theta f\|_{\infty,\Theta}\mathfrak{b}(z)$.

By the dominated convergence theorem, we then have:

$$\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s^N \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s^N \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s$$
$$\xrightarrow[N\to+\infty]{} \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mathsf{m}_s \otimes \gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mathsf{m}_s \otimes \gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s.$$

With the same arguments as above, one shows that $\int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mathsf{m}_s^N\rangle \mathrm{d}s \to \int_0^t \langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mathsf{m}_s\rangle \mathrm{d}s$. The proof of the lemma is complete. ■

Proposition 13 (Convergence to the limit equation) *Let $\mu^* \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$ be a limit point of $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,1], \mathcal{P}(\Theta))$. Then, a.s. μ^* satisfies (11).*

Proof Up to extracting a subsequence, we can assume that $\mu^N \xrightarrow{\mathscr{D}} \mu^*$ as $N \to +\infty$. Let $f \in \mathcal{C}^\infty(\Theta)$. The pre-limit equation (28) and Lemma 6 imply that a.s. for all $N \geq 1$ and $t \in [0,1]$, $\mathbf{\Lambda}_t[f](\mu^N) \leq C/N + \mathbf{M}_t^N[f]$. Hence, using the last statement in Lemma 6, it holds for all $t \in [0,1]$,

$$\lim_{N\to\infty} \mathbf{E}[\mathbf{\Lambda}_t[f](\mu^N)] = 0.$$

In particular, $\mathbf{\Lambda}_t[f](\mu^N) \xrightarrow{\mathscr{D}} 0$. Let us now show that $\mathbf{\Lambda}_t[f](\mu^N) \xrightarrow{\mathscr{D}} \mathbf{\Lambda}_t[f](\mu^*)$. Denoting by $\mathsf{D}(\mathbf{\Lambda}_t[f])$ the set of discontinuity points of $\mathbf{\Lambda}_t[f]$, we have, from Proposition 11 and Lemma 12, for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$,

$$\mathbf{P}(\mu^* \in \mathsf{D}(\mathbf{\Lambda}_t[f])) = 0.$$

By the continuous mapping theorem, $\mathbf{\Lambda}_t[f](\mu^N) \xrightarrow{\mathscr{D}} \mathbf{\Lambda}_t[f](\mu^*)$. By uniqueness of the limit in distribution, we have that for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, a.s. $\mathbf{\Lambda}_t[f](\mu^*) = 0$. Let us now prove that a.s. for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, $\Lambda_t[f](\mu^*) = 0$.

On the one hand, for all $f \in \mathcal{C}^\infty(\Theta)$ and $\mathsf{m} \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$, the function $t \mapsto \Lambda_t[f](\mathsf{m})$ is right-continuous. Since $[0,1]$ is separable, we have that for all $f \in \mathcal{C}^\infty(\Theta)$, a.s. for all $t \in [0,1]$, $\mathbf{\Lambda}_t[f](\mu^*) = 0$.

One the other hand $\mathcal{C}^\infty(\Theta)$ is separable (when endowed with the norm $\|f\|_{\mathcal{C}^\infty(\Theta)} = \sum_{k\geq 0} 2^{-k} \min(1, \sum_{|j|=k} \|\partial_j f\|_{\infty,\Theta})$) and the function $f \in \mathcal{C}^\infty(\Theta) \mapsto \mathbf{\Lambda}_t[f](\mathsf{m})$ is continuous (for fixed $t \in [0,1]$ and $\mathsf{m} \in \mathcal{D}([0,1], \mathcal{P}(\Theta))$) relatively to the topology induced by $\|f\|_{\mathcal{C}^\infty(\Theta)}$.

Hence, we obtain that a.s. for all $t \in [0,1]$ and $f \in \mathcal{C}^\infty(\Theta)$, $\mathbf{\Lambda}_t[f](\mu^*) = 0$. The proof of the proposition is thus complete. ■

A.2.4. UNIQUENESS AND END OF THE PROOF OF THEOREM 2

Proposition 14 *There exists a unique solution to (11) in $\mathcal{C}([0,1], \mathcal{P}(\Theta))$.*

Proof First of all, the fact that there is a solution to (11) is provided by Propositions 8, 11 and 13. The proof of the fact that there is a unique solution to (11) relies on the same arguments as those used in the proof of (Descours et al., 2022, Proposition 2.14).

For $\mu \in \mathcal{P}(\mathbf{R}^{d+1})$, we introduce $\boldsymbol{v}[\mu] : \mathbf{R}^{d+1} \to \mathbf{R}^{d+1}$ defined, for $\theta = (m, \rho) \in \mathbf{R}^{d+1}$, by

$$\boldsymbol{v}[\mu](\theta) = -\eta \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu \otimes \gamma \rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) - \eta \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1). \quad (37)$$

In addition, if $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ is solution to (11), it satisfies also (11) with test functions $f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$. Then, adopting the terminology of (Santambrogio, 2015, Section 4.1.2), any solution $\bar{\mu}$ to (11) is a *weak solution*[1] on $[0,T]$ of the measure-valued equation

$$\begin{cases} \partial_t \bar{\mu}_t = \mathrm{div}(\boldsymbol{v}[\bar{\mu}_t]\bar{\mu}_t) \\ \bar{\mu}_0 = \mu_0. \end{cases} \quad (38)$$

Let us now prove that:

1. There exists $C > 0$ such that for all $\mu \in \mathcal{P}(\mathbf{R}^{d+1})$ and $\theta \in \mathbf{R}^{d+1}$,
$$|\mathrm{J}_\theta \boldsymbol{v}[\mu](\theta)| \le C.$$
2. There exists $C > 0$ such that for all $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ solution to (11), $0 \le s, t \le 1$, and $\theta \in \mathbf{R}^{d+1}$,
$$|\boldsymbol{v}[\bar{\mu}_t](\theta) - \boldsymbol{v}[\bar{\mu}_s](\theta)| \le C|t-s|.$$
3. There exists $L' > 0$ such that for all $\mu, \nu \in \mathcal{P}_1(\mathbf{R}^{d+1})$,
$$\sup_{\theta \in \mathbf{R}^d} |\boldsymbol{v}[\mu](\theta) - \boldsymbol{v}[\nu](\theta)| \le L' \mathsf{W}_1(\mu, \nu).$$

Before proving the three items above, we quickly conclude the proof of the proposition. Items 1 and 2 above imply that $v(t,\theta) = \boldsymbol{v}[\bar{\mu}_t](\theta)$ is globally Lipschitz continuous over $[0,1] \times \mathbf{R}^{d+1}$ when $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ is a solution to (11). Since $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta)) \subset \mathcal{C}([0,1], \mathcal{P}(\mathbf{R}^{d+1}))$, this allows to use the representation theorem (Villani, 2003, Theorem 5.34) for the solution of (38) in $\mathcal{C}([0,1], \mathcal{P}(\mathbf{R}^{d+1}))$, i.e. it holds:

$$\forall t \in [0,1],\ \bar{\mu}_t = \phi_t \# \mu_0, \quad (39)$$

where ϕ_t is the flow generated by the vector field $\boldsymbol{v}[\bar{\mu}_t](\theta)$ over $\mathbf{R}^{d+1}$. Equation (39) and the fact that $\mathcal{C}([0,1], \mathcal{P}(\Theta)) \subset \mathcal{C}([0,1], \mathcal{P}_1(\mathbf{R}^{d+1}))$ together with Item 3 above and the same arguments as those used in the proof of (Descours et al., 2022, Proposition 2.14) (which we recall is based estimates in Wasserstein distances between two solutions of (11) derived in Piccoli and Rossi (2016)), one deduces that there is a unique solution to (11).

Let us prove Item 1. Recall $g(\rho) = \ln(1 + e^\rho)$. The functions

$$\rho \mapsto g''(\rho)g(\rho),\ \rho \mapsto g'(\rho),\ \rho \mapsto \frac{g'(\rho)}{g(\rho)},\ \text{ and } \rho \mapsto \frac{g''(\rho)}{g(\rho)}$$

1. We mention that according to (Santambrogio, 2015, Proposition 4.2), the two notions of solutions of (38) (namely the weak solution and the *distributional* solution) are equivalent.

are bounded on $\mathbf{R}$. Thus, in view of (4), $\|\mathrm{Hess}_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)\|_{\infty,\mathbf{R}^{d+1}} < +\infty$. On the other hand, by **A1** and **A3**, for $x \in \mathsf{X}$, $z \in \mathbf{R}^d$, $\theta \in \Theta \mapsto \phi(\theta, z, x)$ is smooth and there exists $C > 0$, for all $x \in \mathsf{X}$, $\theta \in \mathbf{R}^{d+1}$, $z \in \mathbf{R}^d$:

$$|\mathrm{Hess}_\theta \phi(\theta, z, x)| \leq C(\mathfrak{b}(z)^2 + \mathfrak{b}(z)).$$

This bound allows us to differentiate under the integral signs in (37) and proves that $|\mathrm{J}_\theta \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu \otimes \gamma\rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y)| \leq C$, where $C > 0$ is independent of $\mu \in \mathcal{P}(\Theta)$ and $\theta \in \Theta$. The proof of Item 1 is complete.

Let us prove Item 2. Let $\bar{\mu} \in \mathcal{C}([0,1], \mathcal{P}(\Theta))$ be a solution to (11), $0 \leq s \leq t \leq 1$, and $\theta \in \mathbf{R}^{d+1}$. We have

$$\boldsymbol{v}[\bar{\mu}_t](\theta) - \boldsymbol{v}[\bar{\mu}_s](\theta) = -\eta \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), (\bar{\mu}_t - \bar{\mu}_s) \otimes \gamma\rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y). \tag{40}$$

Let $z \in \mathbf{R}^d$ and $x \in \mathsf{X}$. By **A1** and **A3**, $\phi(\cdot, z, x) \in \mathcal{C}^\infty(\Theta)$. Therefore, by (11),

$$\begin{aligned}\langle \phi(\cdot, z, x), \bar{\mu}_t - \bar{\mu}_s\rangle = &-\eta \int_s^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x') - y, \bar{\mu}_r \otimes \gamma\rangle \langle \nabla_\theta \phi(\cdot, z, x) \cdot \nabla_\theta \phi(\cdot,\cdot,x'), \bar{\mu}_r \otimes \gamma\rangle \pi(\mathrm{d}x', \mathrm{d}y)\mathrm{d}r \\ &- \eta \int_s^t \langle \nabla_\theta \phi(\cdot, z, x) \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \bar{\mu}_r\rangle \mathrm{d}r\end{aligned}$$

We have $\|\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)\|_{\infty,\Theta} < +\infty$. Using also (46) and the fact that $\mathsf{X} \times \mathsf{Y}$ is a compact (see **A2**), it holds:

$$|\langle \phi(\cdot, z, x), \bar{\mu}_t - \bar{\mu}_s\rangle| \leq C\mathfrak{b}(z)|t - s|.$$

Hence, for all $x' \in \mathsf{X}$,

$$|\langle \phi(\cdot,\cdot,x'), (\bar{\mu}_t - \bar{\mu}_s) \otimes \gamma\rangle| \leq \langle |\langle \phi(\cdot,\cdot,x'), \bar{\mu}_t - \bar{\mu}_s\rangle|, \gamma\rangle \leq C|t - s|.$$

Thus, by (40) and (47), $|\boldsymbol{v}[\bar{\mu}_t](\theta) - \boldsymbol{v}[\bar{\mu}_s](\theta)| \leq C|t - s|$. This ends the proof of Item 2.

Let us now prove Item 3. Fix $\mu, \nu \in \mathcal{P}_1(\mathbf{R}^{d+1})$ and $\theta \in \mathbf{R}^{d+1}$. We have

$$\boldsymbol{v}[\mu](\theta) - \boldsymbol{v}[\nu](\theta) = -\eta \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x), (\mu - \nu) \otimes \gamma\rangle \langle \nabla_\theta \phi(\theta,\cdot,x), \gamma\rangle \pi(\mathrm{d}x, \mathrm{d}y) \tag{41}$$

For all $x \in \mathsf{X}$, using (32) and (46), it holds:

$$\begin{aligned}|\langle \phi(\cdot,\cdot,x), (\mu - \nu) \otimes \gamma\rangle| &\leq \int_{\mathbf{R}^d} |\langle \phi(\cdot, z, x), \mu\rangle - \langle \phi(\cdot, z, x), \nu\rangle| \gamma(z)\mathrm{d}z \\ &\leq C \int_{\mathbf{R}^d} \mathsf{W}_1(\mu, \nu)\mathfrak{b}(z)\gamma(z)\mathrm{d}z \leq C\mathsf{W}_1(\mu, \nu).\end{aligned}$$

Finally, using in addition (47) and (41), we deduce Item 3.

This ends the proof of the proposition. ∎

We are now ready to prove Theorem 2.

Proof [Proof of Theorem 2] Recall Lemma 1 ensures that a.s. $(\mu^N)_{N\geq 1} \subset \mathcal{D}([0,1],\mathcal{P}(\Theta))$. By Proposition 8, this sequence is relatively compact. Let $\mu^* \in \mathcal{D}([0,1],\mathcal{P}(\Theta))$ be a limit point. Along some subsequence N', it holds:

$$\mu^{N'} \xrightarrow{\mathscr{D}} \mu^*.$$

In addition, a.s. $\mu^* \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$ (by Proposition 11) and μ^* satisfies (11) (by Proposition 13). By Proposition 14, (11) admits a unique solution $\bar{\mu} \in \mathcal{C}([0,1],\mathcal{P}(\Theta))$. Hence, a.s. $\mu^* = \bar{\mu}$. Therefore,

$$\mu^{N'} \xrightarrow{\mathscr{D}} \bar{\mu}.$$

Since the sequence $(\mu^N)_{N\geq 1}$ admits a unique limit point, the whole sequence converges in distribution to $\bar{\mu}$. The convergence also holds in probability since $\bar{\mu}$ is deterministic. The proof of Theorem 2 is complete. ■

A.3. Proof of Lemma 1

In this section we prove Lemma 1. We start with the following simple result.

Lemma 15 *Let $T > 0$, $N \geq 1$, and $c_1 > 0$. Consider a sequence $(u_k)_{0\leq k\leq \lfloor NT\rfloor} \subset \mathbf{R}_+$ for which there exists v_0 such that $u_0 \leq v_0$ and for all $1 \leq k \leq \lfloor NT\rfloor$, $u_k \leq c_1(1 + \frac{1}{N}\sum_{\ell=0}^{k-1} u_\ell)$. Then, for all $0 \leq k \leq \lfloor NT\rfloor$, $u_k \leq v_0 e^{c_1 T}$.*

Proof Define $v_k = c_1(1+\frac{1}{N}\sum_{\ell=0}^{k-1} v_\ell)$. For all $0 \leq k \leq \lfloor NT\rfloor$, $u_k \leq v_k$ and $v_k = v_{k-1}(1+c_1/N)$. Hence $v_k = v_0\big(1+c_1/N\big)^k \leq v_0\big(1+c_1/N\big)^{\lfloor NT\rfloor} \leq v^0 e^{c_1 T}$. This ends the proof of the Lemma. ■

Proof [Proof of Lemma 1] Since $\rho \mapsto g'(\rho)$ and $\rho \mapsto g'(\rho)/g(\rho)$ are bounded continuous functions over $\mathbf{R}$, and since $|g(\rho)| \leq C(1+|\rho|)$, according to (4), there exists $c > 0$, for all $\theta \in \mathbf{R}^{d+1}$,

$$|\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1)| \leq c(1+|\theta|). \tag{42}$$

All along the proof, $C > 0$ is a constant independent of $N \geq 1$, $T > 0$, $i \in \{1,\dots,N\}$, $1 \leq k \leq \lfloor NT\rfloor$, $(x,y) \in \mathsf{X}\times\mathsf{Y}$, $\theta \in \mathbf{R}^{d+1}$, and $z \in \mathbf{R}^d$, which can change from one occurence to another. It holds:

$$|\theta_k^i| \leq |\theta_0^i| + \sum_{\ell=0}^{k-1} |\theta_{\ell+1}^i - \theta_\ell^i|. \tag{43}$$

Using (5), we have, for $0 \leq \ell \leq k-1$,

$$\begin{aligned}|\theta_{\ell+1}^i - \theta_\ell^i| \leq \frac{\eta}{N^2}\sum_{j=1,j\neq i}^{N} \Big|(\langle\phi(\theta_\ell^j,\cdot,x_\ell),\gamma\rangle - y_\ell)\langle\nabla_\theta\phi(\theta_\ell^i,\cdot,x_\ell),\gamma\rangle\Big| \\ + \frac{\eta}{N^2}\Big|\Big\langle(\phi(\theta_\ell^i,\cdot,x_\ell)-y_\ell)\nabla_\theta\phi(\theta_\ell^i,\cdot,x_\ell),\gamma\Big\rangle\Big| + \frac{\eta}{N}|\nabla_\theta\mathscr{D}_{\mathrm{KL}}(q_{\theta_\ell^i}^1|P_0^1)|.\end{aligned} \tag{44}$$

For all $\theta \in \mathbf{R}^{d+1}$, $z \in \mathbf{R}^d$, $(x,y) \in \mathsf{X}\times\mathsf{Y}$, we have, by **A2** and **A3**, since $\phi(\theta,z,x) = s(\Psi_\theta(z),x)$,

$$|\phi(\theta,z,x) - y| \leq C. \tag{45}$$

Moreover, we have $\nabla_\theta \phi(\theta, z, x) = \nabla_1 s(\Psi_\theta(z), x) \mathrm{J}_\theta \Psi_\theta(z)$ (here $\nabla_1 s$ refers to the gradient of s w.r.t. its first variable). By **A3**, $|\nabla_1 s(\Psi_\theta(z), x)| \le C$ and, hence, denoting by J_θ the Jacobian w.r.t. θ, using (10),

$$|\nabla_\theta \phi(\theta, z, x)| \le C |\mathrm{J}_\theta \Psi_\theta(z)| \le C \mathfrak{b}(z). \tag{46}$$

Therefore, by (10),

$$\langle |\nabla_\theta \phi(\theta, \cdot, x)|, \gamma \rangle \le C. \tag{47}$$

Hence, we obtain, using (44) and (42),

$$|\theta^i_{\ell+1} - \theta^i_\ell| \le \frac{\eta}{N^2} \sum_{j=1, j\neq i}^{N} C + \frac{\eta}{N^2} C + \frac{c\eta}{N}(1 + |\theta^i_\ell|) \le \frac{C}{N}(1 + |\theta^i_\ell|). \tag{48}$$

Using **A4**, there exists $K_0 > 0$ such that a.s. for all i, $|\theta^i_0| \le K_0$. Then, from (43) and (48), for $1 \le k \le \lfloor NT \rfloor$, it holds:

$$|\theta^i_k| \le K_0 + \frac{C}{N} \sum_{\ell=0}^{k-1} (1 + |\theta^i_\ell|) \le K_0 + CT + \frac{C}{N} \sum_{\ell=0}^{k-1} |\theta^i_\ell| \le C_{0,T}(1 + \frac{1}{N} \sum_{\ell=0}^{k-1} |\theta^i_\ell|),$$

with $C_{0,T} = \max(K_0 + CT, C) \le K_0 + C(1+T)$. Then, by Lemma 15 and **A4**, we have that for all $N \ge 1$, $i \in \{1, \dots, N\}$ and $0 \le k \le \lfloor NT \rfloor$, $|\theta^i_k| \le K_0 e^{[K_0 + C(1+T)]T}$. The proof of Lemma 1 is thus complete. ∎

Appendix B. Proof of Theorem 3

In this section, we assume **A1**→**A5** (where in **A2**, when $k \ge 1$, $\mathcal{F}^N_k$ is now the one defined in (12)) and the θ^i_k's (resp. μ^N) are those defined by (7) for $i \in \{1, \dots, N\}$ and $k \ge 0$ (resp. by (13) for $N \ge 1$).

B.1. Preliminary analysis and pre-limit equation

B.1.1. NOTATION AND WEIGHTED SOBOLEV EMBEDDINGS

For $J \in \mathbf{N}$ and $\beta \ge 0$, let $\mathcal{H}^{J,\beta}(\mathbf{R}^{d+1})$ be the closure of the set $\mathcal{C}^\infty_c(\mathbf{R}^{d+1})$ for the norm

$$\|f\|_{\mathcal{H}^{J,\beta}} := \Big(\sum_{|k| \le J} \int_{\mathbf{R}^{d+1}} \frac{|\partial_k f(\theta)|^2}{1 + |\theta|^{2\beta}} \mathrm{d}\theta \Big)^{1/2}.$$

The space $\mathcal{H}^{J,\beta}(\mathbf{R}^{d+1})$ is a separable Hilbert space and we denote its dual space by $\mathcal{H}^{-J,\beta}(\mathbf{R}^{d+1})$ (see e.g. Fernandez and Méléard (1997); Jourdain and Méléard (1998)). The associated scalar product on $\mathcal{H}^{J,\beta}(\mathbf{R}^{d+1})$ will be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}^{J,\beta}}$. For $\Phi \in \mathcal{H}^{-J,\beta}(\mathbf{R}^{d+1})$, we use the notation

$$\langle f, \Phi \rangle_{J,\beta} = \Phi[f], \ f \in \mathcal{H}^{J,\beta}(\mathbf{R}^{d+1}).$$

For ease of notation, and if no confusion is possible, we simply denote $\langle f, \Phi \rangle_{J,\beta}$ by $\langle f, \Phi \rangle$. The set $\mathcal{C}^{J,\beta}_0(\mathbf{R}^{d+1})$ (resp. $\mathcal{C}^{J,\beta}(\mathbf{R}^{d+1})$) is defined as the space of functions $f : \mathbf{R}^{d+1} \to \mathbf{R}$ with continuous partial derivatives up to order $J \in \mathbf{N}$ such that

$$\text{for all } |k| \le J, \ \lim_{|\theta| \to \infty} \frac{|\partial_k f(\theta)|}{1 + |\theta|^\beta} = 0 \ \ (\text{resp. } \sum_{|k| \le J} \sup_{\theta \in \mathbf{R}^{d+1}} \frac{|\partial_k f(\theta)|}{1 + |\theta|^\beta} < +\infty).$$

The spaces $\mathcal{C}^{J,\beta}(\mathbf{R}^{d+1})$ and $\mathcal{C}_0^{J,\beta}(\mathbf{R}^{d+1})$ is endowed with the norm

$$\|f\|_{\mathcal{C}^{J,\beta}} := \sum_{|k|\leq J} \sup_{\theta\in\mathbf{R}^{d+1}} \frac{|\partial_k f(\theta)|}{1+|\theta|^\beta}.$$

We note that

$$\theta \in \mathbf{R}^{d+1} \mapsto (1-\chi(\theta))|\theta|^\alpha \in \mathcal{H}^{J,\beta}(\mathbf{R}^{d+1}) \text{ if } \beta-\alpha > (d+1)/2, \tag{49}$$

where $\chi \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$ equals 1 near 0. We recall that from (Fernandez and Méléard, 1997, Section 2), for $m' > (d+1)/2$ and $\alpha, j \geq 0$, $\mathcal{H}^{m'+j,\alpha}(\mathbf{R}^{d+1}) \hookrightarrow \mathcal{C}_0^{j,\alpha}(\mathbf{R}^{d+1})$. In the following, we consider $\gamma_0, \gamma_1 \in \mathbf{R}$ and $L_0 \in \mathbf{N}$ such that

$$\gamma_1 > \gamma_0 > \frac{d+1}{2} + 1 \text{ and } L_0 > \frac{d+1}{2} + 1.$$

We finally recall the following standard result.

Proposition 16 *Let $q > p \geq 1$ and $C > 0$. The set $\mathscr{K}_C^q := \{\mu \in \mathcal{P}_p(\mathbf{R}^{d+1}), \int_{\mathbf{R}^{d+1}} |x|^q \mu(\mathrm{d}x) \leq C\}$ is compact.*

B.1.2. BOUND ON THE MOMENTS OF THE θ_k^i'S

We have the following uniform bound in $N \geq 1$ on the moments of the sequence $\{\theta_k^i, i \in \{1,\dots,N\}\}_{k=0,\dots,\lfloor NT\rfloor}$ defined by (7).

Lemma 17 *Assume* **A1**$\to$ **A5**. *For all $T > 0$ and $p \geq 1$, there exists $C > 0$ such that for all $N \geq 1$, $i \in \{1,\dots,N\}$ and $0 \leq k \leq \lfloor NT\rfloor$,*

$$\mathbf{E}[|\theta_k^i|^p] \leq C.$$

Proof Let $p \geq 1$. By **A4**, $\mathbf{E}[|\theta_0^i|^p] \leq C_p$ for all $i \in \{1,\dots,N\}$. Let $T > 0$. In the following $C > 0$ is a constant independent of $N \geq 1$, $i \in \{1,\dots,N\}$, and $1 \leq k \leq \lfloor NT\rfloor$. Using (7), the fact that ϕ is bounded, Y is bounded, and (46), we have, for $0 \leq n \leq k-1$,

$$\begin{aligned}
|\theta_{n+1}^i - \theta_n^i| &\leq \frac{C}{N^2 B}\sum_{j=1}^N \sum_{\ell=1}^B \mathfrak{b}(\mathsf{Z}_n^{i,\ell}) + \frac{C}{N}|\nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta_n^i}|P_0^1)| \\
&\leq \frac{C}{NB}\sum_{\ell=1}^B (1+\mathfrak{b}(\mathsf{Z}_n^{i,\ell})) + \frac{C}{N}(1+|\theta_n^i|),
\end{aligned} \tag{50}$$

where we have also used (42) for the last inequality. Let us recall the following convexity inequality: for $m, p \geq 1$ and $x_1,\dots,x_p \in \mathbf{R}_+$,

$$\Big(\sum_{n=1}^m x_n\Big)^p \leq m^{p-1}\sum_{n=1}^m x_n^p. \tag{51}$$

Using (43), **A1** with $q = p$, and the fact that $1 \leq k \leq \lfloor NT\rfloor$, one has setting $u_k = \mathbf{E}[|\theta_k^i|^p]$, $u_k \leq C(1+\frac{1}{N}\sum_{n=0}^{k-1} u_n)$. The result then follows from Lemma 15. ■

B.1.3. PRE-LIMIT EQUATION

In this section, we derive the pre-limit equation for μ^N defined by (13). For simplicity we will keep the same notations as those introduced in Section A.1.1, though these objects will now be defined with θ_k^i set by (7), and on $\mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$, for all integer $k \geq 0$, and all time $t \geq 0$. Let $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$. Then, set for $k \geq 0$,

$$\mathbf{D}_k^N[f] = -\frac{\eta}{N^3}\sum_{i=1}^N \sum_{j=1, j\neq i}^N \int_{\mathsf{X}\times\mathsf{Y}} \big(\langle \phi(\theta_k^j, \cdot, x), \gamma\rangle - y\big)\big\langle \nabla_\theta f(\theta_k^i)\cdot \nabla_\theta \phi(\theta_k^i, \cdot, x), \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y)$$
$$- \frac{\eta}{N^2}\int_{\mathsf{X}\times\mathsf{Y}} \big\langle (\phi(\cdot,\cdot,x) - y)\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \nu_k^N \otimes \gamma\big\rangle \pi(\mathrm{d}x, \mathrm{d}y).$$

Note that $\mathbf{D}_k^N$ above is the one defined in (21) but now on $\mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$ and with θ_k^i defined by (7). For $k \geq 0$, we set

$$\mathbf{M}_k^N[f] = -\frac{\eta}{N^3 B}\sum_{i,j=1}^N \sum_{\ell=1}^B (\phi(\theta_k^j, \mathsf{Z}_k^{j,\ell}, x_k) - y_k)\nabla_\theta f(\theta_k^i)\cdot\nabla_\theta\phi(\theta_k^i, \mathsf{Z}_k^{i,\ell}, x_k) - \mathbf{D}_k^N[f]. \tag{52}$$

By Lemma 17 together with (45) and (46), $\mathbf{M}_k^N[f]$ is integrable. Also, using **A5** and the fact that θ_k^j is $\mathcal{F}_k^N$-measurable (see (12)),

$$\mathbf{E}[\mathbf{M}_k^N[f] | \mathcal{F}_k^N] = 0.$$

Set $\mathbf{M}_t^N[f] = \sum_{k=0}^{\lfloor Nt\rfloor - 1} \mathbf{M}_k^N[f]$, $t \geq 0$. We now extend the definition of $\mathbf{W}_t^N[f]$ and $\mathbf{R}_k^N[f]$ in (27) and (19) to any time $t \geq 0$, $k \geq 0$, and $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$, and with θ_k^i set by (7). We then set

$$\mathbf{R}_t^N[f] = \sum_{k=0}^{\lfloor Nt\rfloor - 1} \mathbf{R}_k^N[f],\ t \geq 0.$$

With the same algebraic computations as those made in Section A.1.1, one obtains the following pre-limit equation: for $N \geq 1$, $t \geq 0$, and $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$,

$$\begin{aligned}
\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle &= -\eta \int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \langle \phi(\cdot,\cdot,x) - y, \mu_s^N\otimes\gamma\rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x), \mu_s^N\otimes\gamma\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\quad - \eta\int_0^t \big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mu_s^N\big\rangle \mathrm{d}s \\
&\quad + \frac{\eta}{N}\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle \langle \phi(\cdot,\cdot,x) - y, \gamma\rangle \langle \nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x),\gamma\rangle, \mu_s^N\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\quad - \frac{\eta}{N}\int_0^t \int_{\mathsf{X}\times\mathsf{Y}} \Big\langle (\phi(\cdot,\cdot,x) - y)\nabla_\theta f\cdot\nabla_\theta\phi(\cdot,\cdot,x), \mu_s^N\otimes\gamma\Big\rangle \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\quad + \mathbf{M}_t^N[f] + \mathbf{W}_t^N[f] + \mathbf{R}_t^N[f].
\end{aligned} \tag{53}$$

We will now show that the sequence $(\mu^N)_{N\geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$.

B.2. Relative compactness and convergence to the limit equation

B.2.1. RELATIVE COMPACTNESS IN $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$

In this section we prove the following result.

Proposition 18 *Assume* **A1**→**A5**. *Recall* $\gamma_0 > \frac{d+1}{2} + 1$. *Then, the sequence* $(\mu^N)_{N\geq 1}$ *is relatively compact in* $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$.

We start with the following lemma.

Lemma 19 *Assume* **A1**→ **A5**. *Then,* $\forall T > 0$ *and* $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$,

$$\sup_{N\geq 1} \mathbf{E}\Big[\sup_{t\in[0,T]} \langle f, \mu_t^N\rangle^2\Big] < +\infty.$$

Proof Let $T > 0$. In what follows, $C > 0$ is a constant independent of $f \in \mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$, $(s,t) \in [0,T]^2$, and $z \in \mathbf{R}^d$ which can change from one occurence to another. We have by **A4**, $\mathbf{E}[\langle f, \mu_0^N\rangle^2] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}$. By (53) and (45), it holds:

$$\begin{aligned}
\sup_{t\in[0,T]} \langle f, \mu_t^N\rangle^2 \leq C\Big[\|f\|^2_{\mathcal{C}^{2,\gamma_1}} &+ \int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \Big|\Big\langle \big\langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)|, \gamma\big\rangle, \mu_s^N\Big\rangle\Big|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&\int_0^T \Big|\Big\langle \big|\nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1)\big|, \mu_s^N\Big\rangle\Big|^2 \mathrm{d}s \\
&+ \frac{1}{N^2}\int_0^T \int_{\mathsf{X}\times\mathsf{Y}} \Big|\Big\langle \big\langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)|, \gamma\big\rangle, \mu_s^N\Big\rangle\Big|^2 \pi(\mathrm{d}x,\mathrm{d}y)\mathrm{d}s \\
&+ \sup_{t\in[0,T]} |\mathbf{M}_t^N[f]|^2 + \sup_{t\in[0,T]} |\mathbf{W}_t^N[f]|^2 + \sup_{t\in[0,T]} |\mathbf{R}_t^N[f]|^2.\Big]. \qquad (54)
\end{aligned}$$

We have using (46), for $s \in [0,T]$ and $z \in \mathbf{R}^d$,

$$|\nabla_\theta f(\theta^i_{\lfloor Ns\rfloor}) \cdot \nabla_\theta \phi(\theta^i_{\lfloor Ns\rfloor}, z, x)| \leq C\|f\|_{\mathcal{C}^{1,\gamma_1}} \mathfrak{b}(z)(1 + |\theta^i_{\lfloor Ns\rfloor}|^{\gamma_1}). \qquad (55)$$

Thus, using Lemma 17,

$$\mathbf{E}\Big[\Big\langle \big\langle |\nabla_\theta f \cdot \nabla_\theta \phi(\cdot,\cdot,x)|, \gamma\big\rangle, \mu_s^N\Big\rangle^2\Big] \leq C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}. \qquad (56)$$

Using (42), for $s \in [0,T]$, it holds:

$$\big|\nabla_\theta f(\theta^i_{\lfloor Ns\rfloor}) \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q^1_{\theta^i_{\lfloor Ns\rfloor}}|P_0^1)\big| \leq C\|f\|_{\mathcal{C}^{1,\gamma_1}}(1 + |\theta^i_{\lfloor Ns\rfloor}|^{\gamma_1+1}). \qquad (57)$$

Thus, using Lemma 17,

$$\mathbf{E}\Big[\Big|\Big\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mu_s^N\Big\rangle\Big|^2\Big] \leq C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}. \qquad (58)$$

On the other hand, we have using (51):

$$\sup_{t\in[0,T]} |\mathbf{M}_t^N[f]|^2 \leq \lfloor NT\rfloor \sum_{k=0}^{\lfloor NT\rfloor - 1} |\mathbf{M}_k^N[f]|^2. \qquad (59)$$

Recall (52). By (21), (51), **A1**, and (55), it holds:

$$|\mathbf{D}_k^N[f]|^2 \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}\big[\frac{1}{N^4}\sum_{i\neq j=1}^N(1+|\theta_k^i|^{2\gamma_1})+\frac{1}{N^4}(1+\langle|\cdot|^{2\gamma_1},\nu_k^N\rangle)\big] \le \frac{C}{N^2}\|f\|^2_{\mathcal{C}^{1,\gamma_1}}(1+|\theta_k^i|^{2\gamma_1})$$

and

$$|\mathbf{M}_k^N[f]|^2 \le \frac{C}{N^4B}\sum_{i,j=1}^N\sum_{\ell=1}^B\|f\|^2_{\mathcal{C}^{1,\gamma_1}}|\mathfrak{b}(\mathsf{Z}_k^{i,\ell})|^2(1+|\theta^i_{\lfloor Ns\rfloor}|^{2\gamma_1})+|\mathbf{D}_k^N[f]|^2.$$

By Lemma 17 and **A1**, one deduces that

$$\mathbf{E}[|\mathbf{M}_k^N[f]|^2] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N^2. \tag{60}$$

Going back to (59), we then have $\mathbf{E}[\sup_{t\in[0,T]}|\mathbf{M}_t^N[f]|^2] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}$. Using the same arguments as those used so far, one also deduces that for $t\in[0,T]$

$$\begin{aligned}\sup_{t\in[0,T]}|\mathbf{W}_t^N[f]|^2 &\le \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2}\sup_{t\in[0,T]}(1+\langle|\cdot|^{\gamma_1+1},\nu^N_{\lfloor Nt\rfloor}\rangle)^2\\ &= \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2}\max_{0\le k\le\lfloor NT\rfloor}(1+\langle|\cdot|^{\gamma_1+1},\nu_k^N\rangle)^2\\ &\le \frac{C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}}{N^2}\sum_{k=0}^{\lfloor NT\rfloor}(1+\langle|\cdot|^{\gamma_1+1},\nu_k^N\rangle)^2.\end{aligned}$$

and thus

$$\mathbf{E}\Big[\sup_{t\in[0,T]}|\mathbf{W}_t^N[f]|^2\Big] \le C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N. \tag{61}$$

Let us finally deal with the term involving $\mathbf{R}_t^N[f]$. One has using (51):

$$\sup_{t\in[0,T]}|\mathbf{R}_t^N[f]|^2 \le \lfloor NT\rfloor\sum_{k=0}^{\lfloor NT\rfloor-1}|\mathbf{R}_k[f]|^2.$$

For $0\le k\le\lfloor NT\rfloor-1$, we have, from (19),

$$\begin{aligned}|\mathbf{R}_k^N[f]|^2 &\le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}}{N}\sum_{i=1}^N|\theta^i_{k+1}-\theta^i_k|^4(1+|\hat\theta^i_k|^{\gamma_1})^2\\ &\le \frac{C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}}{N}\sum_{i=1}^N|\theta^i_{k+1}-\theta^i_k|^4(1+|\theta^i_{k+1}|^{2\gamma_1}+|\theta^i_k|^{2\gamma_1}).\end{aligned}$$

Using (50),

$$|\theta^i_{k+1}-\theta^i_k|^4 \le C\Big[\frac{1}{N^4}+\frac{|\theta^i_k|^4}{N^4}+\frac{1}{N^4B}\sum_{\ell=1}^B|\mathfrak{b}(\mathsf{Z}_k^{i,\ell})|^4\Big].$$

By Lemma 17 and **A1**, it then holds $\mathbf{E}[|\theta^i_{k+1}-\theta^i_k|^4(1+|\theta^i_{k+1}|^{2\gamma_1}+|\theta^i_k|^{2\gamma_1})] \leq C/N^4$. Hence, one deduces that

$$\mathbf{E}[\sup_{t\in[0,T]}|\mathbf{R}^N_t[f]|^2] \leq C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}/N^2. \tag{62}$$

This ends the proof of Lemma 19. ■

Lemma 20 (Compact containment for $(\mu^N)_{N\geq 1}$) *Assume* **A1**→**A5**. *Let* $0<\epsilon<\gamma_1-\gamma_0$. *For every* $T>0$,

$$\sup_{N\geq 1}\mathbf{E}\Big[\sup_{t\in[0,T]}\int_{\mathbf{R}^{d+1}}|x|^{\gamma_0+\epsilon}\mu^N_t(\mathrm{d}x)\Big] < +\infty. \tag{63}$$

Proof Apply Lemma 19 with $f:\theta\mapsto(1-\chi)|\theta|^{\gamma_0+\epsilon}\in\mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$. ■

Lemma 21 *Assume* **A1**→**A5**. *Let* $T>0$ *and* $f\in\mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$. *Then, there exists* $C>0$ *such that for all* $\delta>0$ *and* $0\leq r<t\leq T$ *such that* $t-r\leq\delta$, *one has for all* $N\geq 1$,

$$\mathbf{E}\big[|\langle f,\mu^N_t\rangle-\langle f,\mu^N_r\rangle|^2\big] \leq C(\delta^2+\delta/N+1/N).$$

Proof Using (53), Jensen's inequality, (45), (56), and (58), one has for $f\in\mathcal{C}^{2,\gamma_1}(\mathbf{R}^{d+1})$,

$$\begin{aligned}\mathbf{E}\big[|\langle f,\mu^N_t\rangle-\langle f,\mu^N_r\rangle|^2\big] \leq C\Big[&(t-r)^2(1+1/N^2)\|f\|^2_{\mathcal{C}^{1,\gamma_1}}+\mathbf{E}\big[\big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\mathbf{M}^N_k[f]\big|^2\big]\\ &+\mathbf{E}\big[\,\big|\mathbf{W}^N_t[f]-\mathbf{W}^N_r[f]\big|^2\,\big]+\mathbf{E}\big[\,\big|\mathbf{R}^N_t[f]-\mathbf{R}^N_r[f]\big|^2\,\big].\end{aligned} \tag{64}$$

We also have with the same arguments as those used just before (31)

$$\mathbf{E}\big[\big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\mathbf{M}^N_k[f]\big|^2\big] = \sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\mathbf{E}[|\mathbf{M}^N_k[f]|^2].$$

Using in addition (60), one has $\mathbf{E}\big[\big|\sum_{k=\lfloor Nr\rfloor}^{\lfloor Nt\rfloor-1}\mathbf{M}^N_k[f]\big|^2\big]\leq C(N\delta+1)\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N^2$. Note that with this argument, we also deduce that

$$\mathbf{E}[|\mathbf{M}^N_t[f]|^2]\leq C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N. \tag{65}$$

On the other hand, by (61) and (62), one has

$$\mathbf{E}\big[\,\big|\mathbf{W}^N_t[f]-\mathbf{W}^N_r[f]\big|^2\,\big]\leq C\|f\|^2_{\mathcal{C}^{1,\gamma_1}}/N \text{ and } \mathbf{E}\big[\,\big|\mathbf{R}^N_t[f]-\mathbf{R}^N_r[f]\big|^2\,\big]\leq C\|f\|^2_{\mathcal{C}^{2,\gamma_1}}/N^2.$$

One then plugs all the previous estimates in (64) to deduce the result of Lemma 21. ■

We are now in position to prove Proposition 18.

Proof [Proof of Proposition 18] The proof consists in applying (Jakubowski, 1986, Theorem 4.6) with $E = \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ and $\mathbb{F} = \{\mathsf{H}_f, f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})\}$ where

$$\mathsf{H}_f : \nu \in \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}) \mapsto \langle f, \nu \rangle.$$

The set $\mathbb{F}$ on $\mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1})$ satisfies Conditions (Jakubowski, 1986, (3.1) and (3.2) in Theorem 3.1). Condition (4.8) there follows from Proposition 16, Lemma 20, and Markov's inequality. Let us now show (Jakubowski, 1986, Condition (4.9)) is verified, i.e. that for all $f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$, the family $(\langle f, \mu^N \rangle)_{N \geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathbf{R})$. To do this, it suffices to use Lemma 21 and (Descours et al., 2022, Proposition A.1) (with $\mathcal{H}_1 = \mathcal{H}_2 = \mathbf{R}$ there). In conclusion, according to (Jakubowski, 1986, Theorem 4.6), the sequence $(\mu^N)_{N \geq 1} \subset \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ is relatively compact. ■

B.2.2. LIMIT POINTS SATISFY THE LIMIT EQUATION (15)

For $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbf{R}^{d+1})$ and $t \geq 0$, we introduce for $\mathsf{m} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$,

$$\begin{aligned} \mathbf{\Phi}_t[f] : \mathsf{m} \mapsto \Big| & \langle f, \mathsf{m}_t \rangle - \langle f, \mu_0 \rangle \\ & + \eta \int_0^t \int_{\mathsf{X} \times \mathsf{Y}} \langle \phi(\cdot, \cdot, x) - y, \mathsf{m}_s \otimes \gamma \rangle \langle \nabla_\theta f \cdot \nabla_\theta \phi(\cdot, \cdot, x), \mathsf{m}_s \otimes \gamma \rangle \pi(\mathrm{d}x, \mathrm{d}y) \mathrm{d}s \\ & + \eta \int_0^t \left\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \mathsf{m}_s \right\rangle \mathrm{d}s \Big|. \end{aligned} \tag{66}$$

Note that $\mathbf{\Phi}_t[f]$ is the function $\mathbf{\Lambda}_t[f]$ previously defined in (35) for test functions $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbf{R}^{d+1})$ and for $\mathsf{m} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$.

Lemma 22 *Assume* **A1**→**A5**. *Let $f \in \mathcal{C}^{1,\gamma_0-1}(\mathbf{R}^{d+1})$. Then $\mathbf{\Phi}_t[f]$ is well defined. In addition, if a sequence $(\mathsf{m}^N)_{N \geq 1}$ converges to m in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$, then, for all continuity point $t \geq 0$ of m, we have $\mathbf{\Phi}_t[f](\mathsf{m}^N) \to \mathbf{\Phi}_t[f](\mathsf{m})$.*

Proof Using **A1**, and because Y is bounded and the function ϕ is bounded, $\mathscr{G}_1^{x,y} : \theta \mapsto \langle \phi(\theta, \cdot, x) - y, \gamma \rangle \in \mathcal{C}_b^\infty(\mathbf{R}^{d+1})$. In addition, for all multi-index $\alpha \in \mathbf{N}^{d+1}$, there exists $C > 0$, for all $x, y \in \mathsf{X} \times \mathsf{Y}$ and all $\theta \in \mathbf{R}^{d+1}$, $|\partial_\alpha \mathscr{G}_1^{x,y}(\theta)| \leq C$. The same holds for the function $\mathscr{G}_2^x : \theta \in \mathbf{R}^{d+1} \mapsto \langle \nabla_\theta \phi(\theta, \cdot, x), \gamma \rangle$. Consequently, $\theta \mapsto \nabla_\theta f(\theta) \cdot \mathscr{G}_2^x(\theta) \in \mathcal{C}^{0,\gamma_0-1}(\mathbf{R}^{d+1}) \hookrightarrow \mathcal{C}^{0,\gamma_0}(\mathbf{R}^{d+1})$. Then, there exists $C > 0$ independent of $(x, y) \in \mathsf{X} \times \mathsf{Y}$ and $s \in [0, t]$ such that

$$|\langle \mathscr{G}_1^{x,y}, \mathsf{m}_s \rangle| \leq C,$$

and

$$|\langle \nabla_\theta f \cdot \mathscr{G}_2^x, \mathsf{m}_s \rangle| \leq C \|f\|_{\mathcal{C}^{1,\gamma_0-1}} \langle 1 + |.|^{\gamma_0}, \mathsf{m}_s \rangle.$$

Finally, the function $\theta \mapsto \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)$ is smooth (see (4)) and (42) extends to all its derivatives, i.e. for all multi-index $\alpha \in \mathbf{N}^{d+1}$, there exists $c > 0$, for all $\theta \in \mathbf{R}^{d+1}$,

$$|\partial_\alpha \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1 | P_0^1)| \leq c(1 + |\theta|).$$

Thus, $\nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\theta^1|P_0^1) \in \mathcal{C}^{0,\gamma_0}(\mathbf{R}^{d+1})$ and for some $C > 0$ independent of $s \in [0,t]$

$$|\langle \nabla_\theta f \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1|P_0^1), \mathsf{m}_s \rangle| \leq C\|f\|_{\mathcal{C}^{1,\gamma_0-1}} \langle 1 + |.|^{\gamma_0}, \mathsf{m}_s \rangle.$$

Since in addition $\sup_{s\in[0,t]} \langle 1 + |.|^{\gamma_0}, \mathsf{m}_s \rangle < +\infty$ (since $s \mapsto \langle 1 + |.|^{\gamma_0}, \mathsf{m}_s \rangle \in \mathcal{D}(\mathbf{R}_+, \mathbf{R})$), $\mathbf{\Phi}_t[f]$ is well defined. To prove the continuity property of $\mathbf{\Phi}_t[f]$ it then suffices to use the previous upper bounds together similar arguments as those used in the proof of Lemma 12 (see also Descours et al. (2022)). ■

Proposition 23 *Assume* **A1**→**A5**. *Let μ^* be a limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. Then, μ^* satisfies a.s. Equation (15).*

Proof Let us consider $f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$ and μ^* be a limit point of $(\mu^N)_{N\geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. Recall that by (Ethier and Kurtz, 2009, lemma 7.7 in Chapter 3), the complementary of the set

$$\mathcal{C}(\mu^*) = \{t \geq 0,\, \mathbf{P}(\mu^*_{t-} = \mu^*_t) = 1\}$$

is at most countable. Let $t_* \in \mathcal{C}(\mu^*)$. Then, by Lemma 22, one has that $\mathbf{P}(\mu^* \in \mathsf{D}(\mathbf{\Phi}_{t_*}[f])) = 0$. Thus, by the continuous mapping theorem, it holds

$$\mathbf{\Phi}_{t_*}[f](\mu^N) \xrightarrow{\mathscr{D}} \mathbf{\Phi}_{t_*}[f](\mu^*).$$

On the other hand, using (53) and the estimates (62), (61), (65), (56), and (58), it holds

$$\lim_{N\to\infty} \mathbf{E}[\mathbf{\Phi}_{t_*}[f](\mu^N)] = 0.$$

Consequently, for all $f \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$ and $t_* \in \mathcal{C}(\mu^*)$, it holds a.s. $\mathbf{\Phi}_{t_*}[f](\mu^*) = 0$. On the other hand, for all $\psi \in \mathcal{C}_c^\infty(\mathbf{R}^{d+1})$, $\mathsf{m} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$, and $s \geq 0$, the mappings

$$t \geq 0 \mapsto \mathbf{\Phi}_t[\psi](\mathsf{m})$$

is right continuous, and

$$f \in \mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1}) \mapsto \mathbf{\Phi}_s[f](\mathsf{m})$$

is continuous (because $\mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1}) \hookrightarrow \mathcal{C}_0^{1,\gamma_0-1}(\mathbf{R}^{d+1})$). In addition, $\mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1})$ admits a dense and countable subset of elements in $\mathcal{C}_c^\infty(\mathbf{R}^{d+1})$. Moreover, there exists a countable subset $\mathcal{T}_{\mu^*}$ of $\mathcal{C}(\mu^*)$ such that for all $t \geq 0$ and $\epsilon > 0$, there exists $s \in \mathcal{T}_{\mu^*}$, $s \in [t, t+\epsilon]$. We prove this claim. Since $\mathbb{R}_+$ is a metric space, $\mathcal{C}(\mu^*)$ is separable and thus admits a dense subset $\mathcal{O}_{\mu^*}$. Since $[t+\epsilon/4, t+3\epsilon/4] \cap \mathcal{C}(\mu^*) \neq \emptyset$, there exists $u \in [t+\epsilon/4, t+3\epsilon/4] \cap \mathcal{C}(\mu^*)$. Consider now $s \in \mathcal{O}_{\mu^*}$ such that $|s-u| \leq \epsilon/4$. It then holds $t \leq s \leq t+\epsilon$, proving the claim with $\mathcal{T}_{\mu^*} = \mathcal{O}_{\mu^*}$.

Hence, we have with a classical argument that a.s. for all $f \in \mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1})$ and $t \geq 0$, $\mathbf{\Lambda}_t[f](\mu^*) = 0$. Note also that $\mathcal{C}_b^\infty(\mathbf{R}^{d+1}) \subset \mathcal{H}^{L_0,\gamma_0-1}(\mathbf{R}^{d+1})$ since $2\gamma_0 > d+1$. This ends the proof of the proposition. ■

B.3. Uniqueness of the limit equation and end of the proof of Theorem 3

In this section, we prove that there is a unique solution to (15) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$. To this end, we first need to prove that every limit points of $(\mu^N)_{N\geq 1}$ a.s. belongs to $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$.

B.3.1. LIMIT POINTS BELONG TO $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$

Proposition 24 *Assume* **A1**→**A5**. *Let* $\mu^* \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ *be a limit point of* $(\mu^N)_{N\geq 1}$ *in* $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. *Then, a.s.* $\mu^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$.

Proof Note that since $\mathsf{W}_1 \leq \mathsf{W}_{\gamma_0}$, $\mu^{N'} \xrightarrow{\mathscr{D}} \mu^*$ also in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$, along some subsequence N'. According to (Jacod and Shiryaev, 1987, Proposition 3.26 in Chapter VI), $\mu^* \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ a.s. if for all $T > 0$, $\lim_{N\to+\infty} \mathbf{E}\big[\sup_{t\in[0,T]} \mathsf{W}_1(\mu^N_{t_-}, \mu^N_t)\big] = 0$. Using (32), this is equivalent to prove that

$$\lim_{N\to+\infty} \mathbf{E}\Big[\sup_{t\in[0,T]} \sup_{\|f\|_{\mathrm{Lip}}\leq 1} |\langle f, \mu^N_{t_-}\rangle - \langle f, \mu^N_t\rangle|\Big] = 0. \tag{67}$$

Let us consider $T > 0$ and a Lipschitz function $f : \mathbf{R}^{d+1} \to \mathbf{R}$ such that $\|f\|_{\mathrm{Lip}} \leq 1$. We have $\langle f, \mu^N_t\rangle = \langle f, \mu^N_0\rangle + \sum_{k=0}^{\lfloor Nt\rfloor -1} \langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle$ (with usual convention $\sum_0^{-1} = 0$). Thus the discontinuity points of $t \in [0,T] \mapsto \langle f, \mu^N_t\rangle$ lies exactly at $\{1/N, 2/N, \ldots, \lfloor NT\rfloor/N\}$ and

$$|\langle f, \mu^N_{t_-}\rangle - \langle f, \mu^N_t\rangle| \leq \max_{k=0,\ldots,\lfloor NT\rfloor -1} |\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle|, \ \ \forall t \in [0,T],\ f \text{ Lipschitz}. \tag{68}$$

Pick $k = 0, \ldots, \lfloor NT\rfloor - 1$. We have by (50),

$$|\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle| \leq \frac{1}{N}\sum_{i=1}^N |\theta^i_{k+1} - \theta^i_k| \leq \frac{C}{N}\sum_{i=1}^N \Big[\frac{1}{NB}\sum_{\ell=1}^B (1 + \mathfrak{b}(\mathsf{Z}^{i,\ell}_k)) + \frac{1}{N}(1 + |\theta^i_k|)\Big] =: d^N_k \tag{69}$$

Hence, it holds:

$$|d^N_k|^2 \leq \frac{C}{N}\sum_{i=1}^N \Big[\frac{1}{N^2B}\sum_{\ell=1}^B (1 + \mathfrak{b}^2(\mathsf{Z}^{i,\ell}_k)) + \frac{1}{N^2}(1 + |\theta^i_k|^2)\Big],$$

where thanks to Lemma 17 and **A1**, for all $k = 0, \ldots, \lfloor NT\rfloor - 1$, $\mathbf{E}[|d^N_k|^2] \leq C/N^2$ for some $C > 0$ independent of $N \geq 1$ and $k = 0, \ldots, \lfloor NT\rfloor - 1$. Thus, using (68) and (69),

$$\begin{aligned}
\mathbf{E}\Big[\sup_{t\in[0,T]} \sup_{\|f\|_{\mathrm{Lip}}\leq 1} |\langle f, \mu^N_{t_-}\rangle - \langle f, \mu^N_t\rangle|\Big] &\leq \mathbf{E}\Big[\sup_{\|f\|_{\mathrm{Lip}}\leq 1} \max_{k=0,\ldots,\lfloor NT\rfloor -1} |\langle f, \nu^N_{k+1}\rangle - \langle f, \nu^N_k\rangle|\Big] \\
&\leq \mathbf{E}\Big[\max_{k=0,\ldots,\lfloor NT\rfloor -1} d^N_k\Big] \\
&\leq \mathbf{E}\Big[\sqrt{\sum_{k=0}^{\lfloor NT\rfloor -1} |d^N_k|^2}\Big] \\
&\leq \sqrt{\mathbf{E}\Big[\sum_{k=0}^{\lfloor NT\rfloor -1} |d^N_k|^2\Big]} \leq \frac{C}{\sqrt{N}}.
\end{aligned}$$

This concludes the proof of Proposition 24. ∎

B.3.2. UNIQUENESS OF THE SOLUTION TO (15)

Proposition 25 *There is a unique solution* $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ *to* (15).

Proof First of all, the existence of a solution is provided by Propositions 18, 24 and 23. Let us now prove that there is a unique solution to (15) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$.

Recall the definition of $\boldsymbol{v}[\mu]$ in (37). We claim that for all $T > 0$ and all solution $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ of (15), there exists $C > 0$ such that

$$|\boldsymbol{v}[\bar{\mu}_t](\theta) - \boldsymbol{v}[\bar{\mu}_s](\theta)| \leq C|t-s|, \quad \text{for all } 0 \leq s \leq t \leq T \text{ and } \theta \in \mathbf{R}^{d+1}. \tag{70}$$

The proof of item (70) is the same as the one made for Item 2 in Proposition 14 since it holds using (42) and (46), for all $0 \leq s \leq t \leq T$ and $z \in \mathbf{R}^d$,

$$\begin{aligned}\Big| \int_s^t \langle \nabla_\theta \phi(\cdot, z, x) \cdot \nabla_\theta \mathscr{D}_{\mathrm{KL}}(q_\cdot^1 | P_0^1), \bar{\mu}_r \rangle \mathrm{d}r \Big| &\leq C\mathfrak{b}(z) \int_s^t \langle (1 + |\cdot|), \bar{\mu}_r \rangle \mathrm{d}r \\ &\leq C\mathfrak{b}(z) \max_{r \in [0,T]} \langle (1 + |\cdot|), \bar{\mu}_r \rangle |t-s|.\end{aligned}$$

We now conclude the proof of Proposition 25. Item 1 in the proof of Proposition 14 and (70) imply that $v(t, \theta) = \boldsymbol{v}[\bar{\mu}_t](\theta)$ is globally Lipschitz on $[0, T] \times \mathbf{R}^{d+1}$, for all $T > 0$, when $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$ is a solution of (15). Since in addition a solution $\bar{\mu}$ to (15) is a weak solution on $\mathbf{R}_+$ to (38) in $\mathcal{C}(\mathbf{R}_+, \mathcal{P}(\mathbf{R}^{d+1}))$, it holds by (Villani, 2003, Theorem 5.34):

$$\forall t \geq 0, \ \bar{\mu}_t = \phi_t \# \mu_0, \tag{71}$$

where ϕ_t is the flow generated by the vector field $\boldsymbol{v}[\bar{\mu}_t](\theta)$ over $\mathbf{R}^{d+1}$. Together with Item 3 in the proof of Proposition 14 and using the same arguments as those used in Step 3 of the proof of (Descours et al., 2022, Proposition 2.14), two solutions agrees on each $[0, T]$ for all $T > 0$. One then deduces the uniqueness of the solution to (11). The proof of Proposition 25 is complete. ■

We are now in position to end the proof of Theorem 3.

Proof [Proof of Theorem 3] By Proposition 18, $(\mu^N)_{N \geq 1}$ is relatively compact in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. Let $\mu^1, \mu^2 \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ be two limit points of this sequence. By Proposition 24, a.s. $\bar{\mu}^1, \bar{\mu}^2 \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1}))$. In addition, according to Proposition 23, μ^1 and μ^2 are a.s. solutions of (15). Denoting by $\bar{\mu} \in \mathcal{C}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ the unique solution to (15) (see Proposition 25), we have a.s.

$$\bar{\mu}^1 = \bar{\mu} \text{ and } \bar{\mu}^2 = \bar{\mu} \text{ in } \mathcal{C}(\mathbf{R}_+, \mathcal{P}_1(\mathbf{R}^{d+1})).$$

In particular $\bar{\mu} \in \mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ and $\bar{\mu}^j = \bar{\mu}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$, $j \in \{1, 2\}$. As a consequence, $\bar{\mu}$ is the unique limit point of $(\mu^N)_{N \geq 1}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$ and the whole sequence $(\mu^N)_{N \geq 1}$ converges to $\bar{\mu}$ in $\mathcal{D}(\mathbf{R}_+, \mathcal{P}_{\gamma_0}(\mathbf{R}^{d+1}))$. Since $\bar{\mu}$ is deterministic, the convergence also holds in probability. The proof of Theorem 3 is complete. ■

Let us now prove Proposition 4.

Proof [Proof of Proposition 4] Any solution to (11) in $\mathcal{C}([0, T], \mathcal{P}(\Theta_T))$ is a solution to (15) in $\mathcal{C}([0, T], \mathcal{P}_1(\mathbf{R}^{d+1}))$. The result follows from Proposition 25. ■