Ergodicity of the underdamped mean-field Langevin dynamics

Anna Kazeykina, Zhenjie Ren, Xiaolu Tan, Junjian Yang

Ergodicity of the underdamped mean-field Langevin dynamics

Anna Kazeykina * Zhenjie Ren † Xiaolu Tan ‡ Junjian Yang §

July 23, 2020

Abstract

We study the long time behavior of an underdamped mean-field Langevin (MFL) equation, and provide a general convergence as well as an exponential convergence rate result under different conditions. The results on the MFL equation can be applied to study the convergence of the Hamiltonian gradient descent algorithm for the overparametrized optimization. We then provide a numerical example of the algorithm to train a generative adversarial networks (GAN).

1 Introduction

In this paper we study the ergodicity of the following underdamped mean-field Langevin (MFL) equation:

$$dX_t = V_t dt, \quad dV_t = -\big(D_m F(\mathcal{L}(X_t), X_t) + \gamma V_t\big)dt + \sigma dW_t, \tag{1.1}$$

where $\mathcal{L}(X_t)$ represents the law of X_t, $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is a function on $\mathcal{P}(\mathbb{R}^n)$ (the space of all probability measures on $\mathbb{R}^n$), $D_m F$ is its intrinsic derivative (recalled in Section 2.1), and W is an n-dimensional standard Brownian motion. Note that the marginal distribution $m_t = \mathcal{L}(X_t, V_t)$ satisfies the nonlinear kinetic Fokker-Planck equation

$$\partial_t m = -v \cdot \nabla_x m + \nabla_v \cdot \big((D_m F(m^X, x) + \gamma v)m\big) + \frac{1}{2}\sigma^2 \Delta_v m, \tag{1.2}$$

where m^X denotes the marginal distribution of m on X_t.

Ignoring the mean-field interaction, the standard underdamped Langevin dynamics was first introduced in statistical physics to describe the motion of a particle with position X and velocity V in a potential field $\nabla_x F$ subject to damping and random collisions, see e.g. [20, 31, 38]. It is well known that under mild conditions the Langevin dynamics has a unique invariant measure on $\mathbb{R}^n \times \mathbb{R}^n$ with the density:

$$m_\infty(x, v) = Ce^{-\frac{2}{\sigma^2}\left(F(x) + \frac{1}{2}|v|^2\right)}, \tag{1.3}$$

where C is the normalization constant. This observation brings up the interest in developing Hamiltonian Monte Carlo methods, based on various discrete time analogues to the underdamped Langevin dynamics, for sampling according to the distributions in form of (1.3), see e.g. Lelièvre, Rousset and Stoltz [32], Neal [37]. Nowadays this interest resurges in the community of machine learning. Notably, the underdamped Langevin dynamics has been empirically observed to converge more quickly to the invariant measure compared to the overdamped

*Laboratoire de Mathématique d'Orsay, Université Paris-Sud, CNRS, Université Paris-Saclay. anna.kazeykina@math.u-psud.fr

†CEREMADE, Université Paris-Dauphine, PSL. ren@ceremade.dauphine.fr.

‡Department of Mathematics, The Chinese University of Hong Kong. xiaolu.tan@cuhk.edu.hk.

§FAM, Fakultät für Mathematik und Geoinformation, Vienna University of Technology, A-1040 Vienna, Austria. junjian.yang@tuwien.ac.at

Langevin dynamics (of which the related MCMC was studied in e.g. Dalalyan [12], Durmus and Moulines [16]), and it was theoretically justified by Cheng, Chatterji, Bartlett and Jordan in [9] for some particular choice of coefficients.

More recently, the mean-field Langevin dynamics draws increasing attention among the attempts to rigorously prove the trainability of neural networks, in particular the two-layer networks (with one hidden layer). It becomes popular (see e.g. Chizat and Bach [10], Mei, Montanari and Nguyen [35], Rotskoff and Vanden-Eijnden [41], Hu, Ren, Šiška and Szpruch [28]) to relax the optimization over the weights of the two-layer network, namely,

$$\inf_{c,A,b} \int \Big| y - \sum_i c_i \varphi(A_i z + b_i) \Big|^2 \mu(dy, dz), \quad \text{with the distribution } \mu \text{ of the data } z \text{ and the label } y,$$

by the optimization over the probability measures:

$$\inf_{m \in \mathcal{P}(\mathbb{R}^n)} \int \big| y - \mathbb{E}^m[c\varphi(Az+b)] \big|^2 \mu(dy, dz), \text{ where } m \text{ is the law of the r.v. } X := (c, A, b).$$

Denote by $F(m) := \int \big| y - \mathbb{E}^m[c\varphi(Az+b)] \big|^2 \mu(dy, dz)$. In Mei, Montanari and Nguyen [35] and Hu, Ren, Šiška and Szpruch [28] the authors further add an entropic regularization to the minimization:

$$\inf_{m \in \mathcal{P}} F(m) + \frac{\sigma^2}{2} H(m), \tag{1.4}$$

where H is the relative entropy with respect to Lebesgue measure. It follows by a variational calculus, see e.g. [28], that the necessary first order condition of the minimization above reads

$$D_m F(m^*, x) + \frac{\sigma^2}{2} \nabla_x \ln m^*(x) = 0.$$

Moreover, since F defined above is convex, this is also a sufficient condition for m^* being a minimizer. It has been proved that such m^* can be characterized as the invariant measure of the overdamped mean-field Langevin dynamics:

$$dX_t = -D_m F(\mathcal{L}(X_t), X_t) dt + \sigma dW_t.$$

Also it has been shown that the marginal laws m_t converge towards m^* in Wasserstein metric. Notably, the (stochastic) gradient descent algorithm used in training the neural networks can be viewed as a numerical discretization scheme for the overdamped MFL dynamics. Similar mean-field analysis has been done to deep networks, optimal controls and games, see e.g. Hu, Kazeykina and Ren [27], Jabir, Šiška and Szpruch [29], Conforti, Kazeykina and Ren [11], Domingo-Enrich, Jelassi and Mensch [15], Šiška and Szpruch [42], Lu, Ma, Lu, Lu and Ying [33].

This paper is devoted to study the analog to the underdamped MFL dynamics. When considering the optimization (1.4), one may in addition introduce a velocity variable V and regularize the problem as

$$\inf_{m \in \mathcal{P}} \mathfrak{F}(m), \quad \text{with} \quad \mathfrak{F}(m) := F(m^X) + \frac{1}{2}\mathbb{E}^m\Big[|V|^2\Big] + \frac{\sigma^2}{2} H(m), \tag{1.5}$$

where m becomes the joint distribution of (X, V), and m^X represents its marginal distribution on X. By the same variational calculus as above, the first order condition reads

$$D_m F(m^{*,X}, x) + \frac{\sigma^2}{2} \nabla_x \ln m^*(x, v) = 0 \quad \text{and} \quad v + \frac{\sigma^2}{2} \nabla_v \ln m^*(x, v) = 0. \tag{1.6}$$

We are going to identify the minimizer m^* as the unique invariant measure of the underdamped MFL dynamics (1.1) in two cases: (i) F is convex; (ii) F is possibly non-convex but satisfies further technical conditions. Moreover, in case (i) we prove the marginal laws m_t of (1.1) converge to m^* under very mild conditions. In case (ii) we show that the convergence is exponentially quick, and notably the convergence rate is dimension-free under a Wassestein distance.

Related works The underdamped Langevin dynamics, even in case without mean-field interaction, is degenerate, so the classical approaches cannot be applied straightforwardly to show the (exponential) ergodicity. In [45,46], Villani introduced the term "hypocoercivity" and prove the exponential convergence of m_t in $H^1_{m_\infty}$. A more direct approach was later developed in Dolbeault, Mouhot and Schmeiser [13, 14], and it triggered many results on kinetic equations. Note that both Villani's and DMS's results on the exponential convergence rate highly depends on the dimension, and (therefore) does not apply to the case with mean-field interaction. It is noteworthy that in the recent paper by Cao, Lu and Wang [7], they developed a new estimate on the convergence rate based on the variational method proposed by Armstrong and Mourrat [2]. There are few articles in the literature studying the ergodicity of underdamped Langevin dynamics using more probabilistic arguments, see e.g. Wu [47], Rotskoff and Rey-Bellet and Thomas [40], Talay [44], Bakry, Cattiaux and Guillin [3]. These works are mostly based on Lyapunov conditions and the rates they obtained also depends on the dimension. In the recent work by Guillin, Liu, Wu and Zhang [23], it has been shown for the first time that the underdamped Langevin equation with non-convex potential is exponentially ergodic in $H^1_{m_\infty}$. Their argument combines Villani's hypocoercivity with the uniform functional inequality and Lyapunov conditions. To complete the brief literature review, we would draw special attention to the coupling argument applied in Bolley, Guillin and Malrieu [5] and Eberle, Guillin and Zimmer [18], which found transparent convergence rates in sense of Wasserstein-type distance.

Theoretical novelty Most of the articles concerning the ergodicity of underdamped Langevin dynamics obtain the convergence rates depending on the dimension, and in particular very few allow both non-convex potential and the mean-field interaction. One exception would be the paper of Guillin, Liu, Wu and Zhang [23], but it focuses on a particular convolution-type interaction and their assumption of uniform functional inequality is quite demanding. As mentioned, in the paper we address the ergodicity of the underdamped MFL dynamics in two cases.

In case that F is convex (on the probability measure space), we provide a general ergodicity result, which mainly relies on the observation in Theorem 2.6, namely, the value of the function $\mathfrak{F}$ defined in (1.5) decreases along the dynamics of the MFL and the derivative $\frac{d\mathfrak{F}(m_t)}{dt}$ can be explicitly computed. This can viewed as an analog of the gradient flow for the overdamped Langevin equation, initiated in the seminal paper by Jordan, Kinderlehrer, and Otto [30], see also the monograph by Ambrosio, Gigli and Savaré [1]. Due to the degeneracy and the mean-field interaction of the underdamped MFL process, the proof for the claim is non-trivial. Base on this observation and using an argument similar to that in Mei, Montanari and Nguyen [35], Hu, Ren, Šiška and Szpruch [28], we prove (in Lemma 4.9) that all cluster points of (m_t) should satisfy

$$v + \frac{\sigma^2}{2}\nabla_v \ln m^*(x,v) = 0.$$

Finally, by intriguing LaSalle's invariance principle for the dynamic system, we show that m^* must satisfy the first order condition (1.6). Since F is convex, (1.6) is sufficient to identify m^* as the unique minimizer of $\mathfrak{F}$. To our knowledge, this approach for proving the ergodicity of the underdamped MFL dynamics is original and the result holds true under very mild condition.

In case that F is possibly non-convex but satisfies further technical conditions, we adopt the reflection-synchronous coupling technique that initiated in Eberle, Guillin and Zimmer [18, 19] to obtain an exponential contraction result. Note that [18] is not concerned with mean-field interaction and the rate found there is dimension dependent. In our context, we design a new Lyapunov function in a quadratic form (see Section 4.4.2) to obtain the contraction when the coupled particles are far away, and as a result obtain a dimension-free convergence rate. The construction of the quadratic form shares some flavor with the argument in Bolley, Guillin and Malrieu [5]. Notably, our construction helps to capture the optimal rate in the area of interest

(see Remark 4.15), so may be more intrinsic.

The rest of the paper is organized as follows. In Section 2 we annonce the main results. Before entering the detailed proofs, we study a numerical example concerning the nowadays popular generative adversarial networks (GAN). The main theorems in Section 2 guides us to propose a theoretical convergent algorithm for the GAN, and the numerical test in Section 3 shows a satisfactory result. Finally we report the proofs in Section 4.

2 Ergodicity of the mean-field Langevin dynamics

2.1 Preliminaries

Let $\mathcal{P}(\mathbb{R}^n)$ denote the space of the probability measures on $\mathbb{R}^n$, and by $\mathcal{P}_p(\mathbb{R}^n)$ the subspace of those with finite p-th moment. Without further specifying, in this paper the continuity on $\mathcal{P}(\mathbb{R}^n)$ is with respect to the weak topology, while the continuity on $\mathcal{P}_p(\mathbb{R}^n)$ is in the sense of $\mathcal{W}_p$ (p-Wasserstein) distance.

A function $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is said to belong to $\mathcal{C}^1$, if there exists a jointly continuous function $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^n) \times \mathbb{R}^n \to \mathbb{R}$ such that

$$F(m') - F(m) = \int_0^1 \int_{\mathbb{R}^n} \frac{\delta F}{\delta m}\big((1-u)m + um', x\big)\ (m' - m)(dx)du.$$

When $\frac{\delta F}{\delta m}$ is continuously differentiable in x, we denote by $D_m F(m, x) := \nabla_x \frac{\delta F}{\delta m}(m, x)$ the intrinsic derivative of F. We say function $F \in \mathcal{C}_b^\infty$ if all the derivatives $\partial^i_{x_1,\cdots,x_k} D^k_m F(m, x_1, \cdots, x_k)$ exist and are bounded.

Let (X, V) denote the canonical variable on $\mathbb{R}^n \times \mathbb{R}^n$. For $m \in \mathcal{P}(\mathbb{R}^n \times R^n)$, we denote by $m^X := \mathcal{L}^m(X) = m \circ X^{-1}$ the marginal law of the variable X under m. Denote by $H(m)$ the relative entropy of the measure $m \in \mathcal{P}(\mathbb{R}^n \times R^n)$ with respect to the Lebesgue measure, that is,

$$H(m) \ := \ \mathbb{E}^m\big[\ln\big(\rho_m(X, V)\big)\big] = \int_{\mathbb{R}^n \times \mathbb{R}^n} \ln\big(\rho_m(x, v)\big)\rho_m(x, v)dxdv,$$

if m has a density function $\rho_m : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$; or $H(m) := \infty$ if m is not absolutely continuous.

2.2 Optimization with entropy regularizer

Throughout the paper, we fix a potential function $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$, and study the following optimization problem:

$$\inf_{m \in \mathcal{P}(\mathbb{R}^\times \mathbb{R}^n)} \mathfrak{F}(m), \quad \text{with} \quad \mathfrak{F}(m) \ := \ F(m^X) + \frac{1}{2}\mathbb{E}^m\big[|V|^2\big] + \frac{\sigma^2}{2\gamma}H(m). \tag{2.1}$$

Assumption 2.1. *The potential function $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is given by*

$$F(m) = F_\circ(m) + \mathbb{E}^m[f(X)],$$

where $F_\circ \in \mathcal{C}_b^\infty$, $f : \mathbb{R}^n \to \mathbb{R}$ belongs to C^∞ with bounded derivatives of all orders larger or equal to 2, and

$$|f(x)| \geq \lambda|x|^2, \quad \textit{for some } \lambda > 0. \tag{2.2}$$

The following result is due to a variational calculus argument, see e.g. Hu, Ren, Šiška and Szpruch [28] for a proof.

Lemma 2.2. *Let Assumption 2.1 hold true. If* $m = \operatorname{argmin}_{\mu\in\mathcal{P}} \mathfrak{F}(\mu)$*, then* m *admits a density and there exists a constant* $C \in \mathbb{R}$*, such that*

$$\frac{\delta F}{\delta m}(m^X, x) + \frac{|v|^2}{2} + \frac{\sigma^2}{2\gamma} \ln m(x,v) = C, \quad \text{for all } (x,v) \in \mathbb{R}^{2n}, \tag{2.3}$$

or equivalently

$$m(x,v) = C \exp\Big(-\frac{2\gamma}{\sigma^2}\Big(\frac{\delta F}{\delta m}(m^X, x) + \frac{|v|^2}{2}\Big)\Big). \tag{2.4}$$

Moreover, if F *is convex then* (2.3) *is sufficient for* $m = \operatorname{argmin}_{\mu\in\mathcal{P}} \mathfrak{F}(\mu)$*.*

Remark 2.3. *In case* $F_\circ \equiv 0$*, note that* $m \mapsto F(m) = \mathbb{E}^m[f(X)]$ *is linear and hence convex. Moreover, one has* $\frac{\delta F}{\delta m}(m^X, x) = f(x)$*, and the result above reduces to the classical result in the variational calculus literature.*

Remark 2.4. *To intuitively understand the first order condition* (2.3)*, we may analyze the simple case without the terms* $\frac{1}{2}\mathbb{E}^m\big[|V|^2\big]$ *and* $\frac{\sigma^2}{2\gamma}H(m)$ *in* $\mathfrak{F}(m)$*. Given a convex function* $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$*, we have for* $m^\varepsilon = (1-\varepsilon)m + \varepsilon m'$ *that*

$$\begin{aligned} F(m') - F(m) &\geq \frac{1}{\varepsilon}\Big(F(m^\varepsilon) - F(m)\Big) \\ &= \frac{1}{\varepsilon}\int_0^\varepsilon \int_{\mathbb{R}^n} \frac{\delta F}{\delta m}\big((1-u)m + um', x\big)\ (m'-m)(dx)du \\ &\longrightarrow \int_{\mathbb{R}^n} \frac{\delta F}{\delta m}\big(m, x\big)(m'-m)(dx)du, \quad \text{as } \varepsilon \to 0. \end{aligned}$$

Therefore, $\frac{\delta F}{\delta m}\big(m, x\big) = C$ *is sufficient for* m *being a minimizer of* F*.*

2.3 Lyapunov function and ergodicity

From the first order optimization condition (2.3), one can check by direct computation that a solution to the optimization problem (2.1) is also an invariant measure of the mean-field Langevin dynamic (1.1), which we recall below:

$$dX_t = V_t dt, \qquad dV_t = -\big(D_m F(\mathcal{L}(X_t), X_t) + \gamma V_t\big)dt + \sigma dW_t. \tag{2.5}$$

Assumption 2.5. *The initial distribution of the MFL equation has finite* p*-th moment for all* $p \geq 0$*, i.e.* $\mathbb{E}[|X_0|^p + |V_0|^p] < \infty$*.*

Under Assumption 2.1 and 2.5, it is well known that the MFL equation (2.5) admits a unique strong solution $(X_t, V_t)_{t\geq 0}$, see e.g. Sznitman [43]. In this paper we first prove that the function $\mathfrak{F}$ defined in (2.1) acts as a Lyapunov function for the marginal flow of the MFL dynamics (2.5) in the following sense.

Theorem 2.6. *Let Assumptions 2.1 and 2.5 hold true, denote* $m_t := \mathcal{L}(X_t, V_t)$ *for all* $t \geq 0$*. Then, for all* $t > s > 0$*,*

$$\mathfrak{F}(m_t) - \mathfrak{F}(m_s) = -\int_s^t \gamma\mathbb{E}\Big[\Big|V_r + \frac{\sigma^2}{2\gamma}\nabla_v \ln m_r(X_r, V_r)\Big|^2\Big]dr.$$

With the help of the Lyapunov function $\mathfrak{F}$, we may prove the convergence of the marginal laws of (2.5) towards to the minimizer $\underline{m} := \operatorname{argmin}_{m\in\mathcal{P}} \mathfrak{F}(m)$, provided that the function F is convex.

Theorem 2.7. *Let Assumptions 2.1 and 2.5 hold true. Suppose in addition that the function F is convex. Then the MFL dynamic* (2.5) *has a unique invariant measure $\underline{m}$, which is also the unique minimizer of* (2.1)*, and moreover*

$$\lim_{t\to\infty} \mathcal{W}_1(m_t, \underline{m}) = 0,$$

Remark 2.8. *The ergodicity of the diffusions with mean-field interaction is a long-standing problem. One may taste the non-triviality through the following simple example. Consider the process*

$$dX_t = (-X_t + \alpha\mathbb{E}[X_t])dt + dW_t.$$

It is not hard to show that the process X has a unique invariant measure $\mathcal{N}(0, 1/2)$ when $\alpha < 1$ and has none of them when $\alpha > 1$. Therefore a structural condition is inevitable to ensure the existence of a unique invariant measure and the convergence of the marginal distributions towards it. Theorem 2.7 shows that F being convex on the probability measure space is sufficient for the underdamped MFL dynamics to be ergodic. It is a sound analogue of Theorem 2.11 in Hu, Ren, Šiška and Szpruch [28], where it has been proved that the convexity of the potential function ensures the ergodicity of the overdamped MFL dynamics.

2.4 Exponential ergodicity given small mean-field dependence

We next study the case where F is possibly non-convex, and are going to obtain an exponential convergence rate if the invariant measure exists.

Assumption 2.9. *Assume that the function $F_\circ \in \mathcal{C}^1$ and $D_m F_\circ$ exists and is Lipschitz continuous. Further assume that for any $\varepsilon > 0$ there exists $K > 0$ such that for all $m, m' \in \mathcal{P}(\mathbb{R}^n)$*

$$\big|D_m F_\circ(m, x) - D_m F_\circ(m', x')\big| \le \varepsilon|x - x'| \quad \text{whenever } |x - x'| \ge K,$$

and the function $f(x) = \frac{\lambda}{2}|x|^2$ with some $\lambda > 0$.

Note that that $D_m\mathbb{E}^m[f(X)](m, x) = \nabla_x f(x) = \lambda x$.

Example 2.10. *The function $F_\circ$, of which the intrinsic derivative $D_m F_\circ$ is uniformly bounded, satisfies Assumption 2.9.*

Given $(X, V), (X', V') \in \mathbb{R}^{2d}$, we denote

$$P := V - V' + \gamma(X - X'), \quad r := |X - X'|, \quad u := |P|, \quad z := (X - X') \cdot P,$$

and define the function

$$\psi(X - X', V - V') := \big(1 + \beta G(X - X', P)\big)h(\eta u + r), \tag{2.6}$$

where the positive constants β, η, the quadratic form G and the function $h : \mathbb{R} \to \mathbb{R}$ will be determined later. Finally define the semi-metric:

$$\mathcal{W}_\psi(m, m') = \inf\left\{\int \psi(x - x', v - v')d\pi(x, v, x', v') : \pi \text{ is a coupling of } m, m' \in \mathcal{P}(\mathbb{R}^{2n})\right\}.$$

Theorem 2.11. *Let Assumption 2.9 hold true. Further assume that*

$$|D_m F_\circ(m, x) - D_m F_\circ(m', x)| \le \iota \mathcal{W}_1(m, m').$$

Then for $\iota > 0$ small enough, we have

$$\mathcal{W}_\psi(m_t, m_t') \le e^{-ct}\mathcal{W}_\psi(m_0, m_0'),$$

for a constant $c > 0$ defined below in (4.41)*. In parituclar, the rate c does not depend on the dimension n.*

Remark 2.12. *The proof of Theorem 2.11 are mainly based on the reflection-synchronous coupling technique developed by Eberle, Guillin and Zimmer in [18]. Note that the contraction obtained in [18] holds true under the assumptions more general than Assumption 2.9, but the convergence rate there is dimension-dependent. We manage to make this tradeoff by considering a new Lyapunov function (see Section 4.4.2) and a new semi-metric, allowing to obtain the exponential ergodicity in the case with small mean-field dependence. Notice also that Guillin, Liu, Wu and Zhang proved the exponential ergodicity in [23] for the underdamped Langevin dynamics with a convolution-type interactions, by a completely different approach based on Villani's hypocoercivity and the uniform functional inequality.*

Remark 2.13. *Since the function ψ is not concave, the semi-metric $\mathcal{W}_\psi$ is not necessarily a metric, and therefore the contraction proved above does not imply the existence of the invariant measure, but only describes the convergence rate whenever the invariant measure exists, in particular when F is convex.*

3 Application to GAN

Recently there is a strong interest in generating samplings according to a distribution only empirically known using the so-called generative adversarial networks (GAN). From a mathematical perspective, the GAN can be viewed as a (zero-sum) game between two players: the generator and the discriminator, and can be trained through an overdamped Langevin process, see e.g. Conforti, Kazeykina and Ren [11], Domingo-Enrich, Jelassi, Mensch, Rotskoff and Bruna [15]. On the other hand, it has been empirically observed and theoretically proved (in case with convex potentials) by Cheng, Chatterji, Bartlett and Jordan in [9] that the simulation of the underdamped Langevin process converges more quickly than that of the overdamped Langevin dynamics. Therefore, in this section we shall implement an algorithm to train the GAN through the underdamped mean-field Langevin dynamics.

Denote by μ the empirically known distribution. The generator aims at generating samplings of a random variable Y so that its distribution ℓ is eventually close to μ. Meanwhile, the discriminator trains a parametrized function $y \mapsto \Phi(m^X, y)$ in the form:

$$\Phi(m^X, y) = \mathbb{E}^{m^X}[C\phi(Ay + b)], \tag{3.1}$$

where ϕ is a fixed activation function and the random variable $X := (C, A, b)$ satisfies the law m^X, to distinguish the distributions μ and ℓ. Such parametrization is expressive enough to represent all continuous functions on a compact set according to the Universal Representation Theorem, see e.g. Hornik [26]. Since we are going to make use of the underdamped MFL process to train the discriminator, we also introduce the random variable V, so that together (X, V) satisfies a distribution m.

Define

$$\mathfrak{F}(m, \ell) := \int \Phi(m^X, y)(\ell - \mu)(dy) - \frac{\eta}{2}\mathbb{E}^m[|V|^2] + \frac{\sigma_0^2}{2}H(\ell) - \frac{\sigma_1^2}{2\gamma}H(m).$$

Remark 3.1. *To make the following discussion mathematically rigorous, one need to :*

- *truncate the variable C in (3.1) to make Φ bounded;*
- *add the small ridge rigularization to the function $\mathfrak{F}$:*

$$\tilde{\mathfrak{F}}(m, \ell) := \mathfrak{F}(m, \ell) + \frac{\lambda_0}{2}\int |y|^2 \ell(dy) - \frac{\lambda_1}{2}\mathbb{E}^m[|X|^2].$$

For the notational simplicity, we omit these technical details in this section.

Consider the zero-sum game between the two players:

$$\begin{cases} \text{generator}: & \inf_\ell \mathfrak{F}(m,\ell) \\ \text{discriminator}: & \sup_m \mathfrak{F}(m,\ell) \end{cases}.$$

One may view

$$d(\ell,\mu) := \sup_m \left\{ \int \Phi(m^X, y)(\ell-\mu)(dy) - \frac{\eta}{2}\mathbb{E}^m[|V|^2] - \frac{\sigma_1^2}{2\gamma} H(m) \right\}$$

as a distance between the distributions ℓ and μ. Then by solving the zero-sum game above, one may achieve as a part of the equilibrium: $\ell^* \in \mathrm{argmin}_\ell\{d(\ell,\mu) + \frac{\sigma_0^2}{2}H(\ell)\}$, which is intuitively close to μ whereas σ_0 is small.

In order to compute the equilibrium of the game, we observe as in Conforti, Kazeykina and Ren [11] that given the choice m of the discriminator, the optimal response of the generative can be computed explicitly: it has the density

$$\ell^*[m](y) = C(m) e^{-\frac{2}{\sigma_0^2}\Phi(m^X, y)}, \tag{3.2}$$

where $C(m)$ is the normalization constant depending on m. Then computing the value of the zero-sum game becomes an optimization over m:

$$\sup_m \inf_\ell \mathfrak{F}(m,\ell) = \sup_m \mathfrak{F}(m, \ell^*[m]).$$

As the main result of this paper goes, the optimizer of the problem above can be characterized by the invariant measure of the underdamped MFL dynamics

$$dX_t = \eta V_t dt, \qquad dV_t = -\big(D_m F(\mathcal{L}(X_t), X_t) + \gamma V_t\big)dt + \sigma_1 dW_t, \tag{3.3}$$

with the potential function:

$$F(m) := -\int \Phi(m^X, y)(\ell^*[m] - \mu)(dy) - \frac{\sigma_0^2}{2} H(\ell^*[m]).$$

Together with (3.2), we may calculate and obtain

$$D_m F(m,x) = \int D_m \Phi(m^X, y, x)(\mu - \ell^*[m])(dy).$$

Next we shall support the theoretical result with a simple numerical test. We set μ as the empirical law of 2000 samples of the distribution $\frac{1}{2}\mathcal{N}(-1,1) + \frac{1}{2}\mathcal{N}(4,1)$, and the coefficients of game as:

$$\phi(z) = \max\{-10, \min\{10, z\}\}, \quad \sigma_0 = 0.1, \quad \sigma_1 = 1, \quad \gamma = 1, \quad \lambda_0 = 0.01, \quad \lambda_1 = 0.1.$$

In order to compute the optimal response of the generator $\ell^*[m]$, we use the Gaussian random walk Metropolis Hasting algorithm, with the optimal scaling proposed in Gelman, Roberts and Gilks [22]. Further, as the numerical scheme for the underdamped Langevin process (3.3), we adopt the well-known splitting procedure, the Brünger-Brooks-Karplus integrator [6], see also Section 2.2.3.2 of Lelièvre, Rousset and Stoltz [32]. Also, in the numerical implementation, the marginal law $\mathcal{L}(X_t)$ in (3.3) is replaced by the empirical law of 2000 samples of X_t. Along the training (the underdamped MFL dynamics), we record the potential energy:

$$\int \Phi(m^X, y)(\mu - \ell^*[m])(dy),$$

as well as the kinetic energy $\frac{\eta}{2}\mathbb{E}^m[|V|^2]$, and we stop the iteration once the potential energy stays considerably small. The result of the numerical test is shown in the following Figure 1. Observe that the total energy is almost monotonously decreasing, as foreseen by Theorem 2.6, and that the samplings generated by the GAN is visibly close to the μ given.

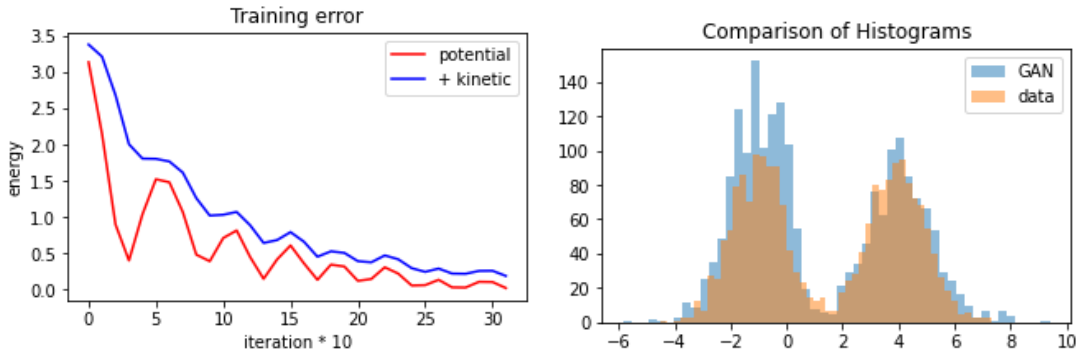

Figure 1: Training errors and GAN samplings

4 Proofs

4.1 Some fine properties of the marginal distributions of the SDE

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an abstract probability space, equipped with a n-dimensional standard Brownian motion W. Let $T > 0$, $b : [0,T] \times \mathbb{R}^{2n} \to \mathbb{R}^n$ be a continuous function such that, for some constant $C > 0$,

$$|b(t,x,v) - b(t,x',v')| \le C(|x-x'| + |v-v'|), \text{ for all } (t,x,v,x',v') \in [0,T] \times \mathbb{R}^{2n} \times \mathbb{R}^{2n},$$

and $\sigma > 0$ be a positive constant, we consider the stochastic differential equation (SDE):

$$dX_t = V_t\, dt, \qquad dV_t = b(t, X_t, V_t)\, dt + \sigma dW_t, \tag{4.1}$$

where the initial condition (X_0, V_0) satisfies

$$\mathbb{E}[|X_0|^2 + |V_0|^2] < \infty.$$

The above SDE has a unique strong solution (X, V), and the marginal distribution $m_t := \mathcal{L}(X_t, V_t)$ satisfies the corresponding Fokker-Planck equation:

$$\partial_t m + v \cdot \nabla_x m + \nabla_v \cdot \big(bm\big) - \frac{1}{2}\sigma^2 \Delta_v m \;=\; 0. \tag{4.2}$$

In this subsection we are going to prove some properties of the density function $\rho_t(x,v)$ of m_t.

Existence of positive densities Let us fix a time horizon $T > 0$, let $C([0,T], \mathbb{R}^n)$ be the space of all $\mathbb{R}^n$–valued continuous paths on $[0,T]$. Denote by $\overline{\Omega} := C([0,T], \mathbb{R}^n) \times C([0,T], \mathbb{R}^n)$ the canonical space, with canonical process $(\overline{X}, \overline{V}) = (\overline{X}_t, \overline{V}_t)_{0 \le t \le 1}$ and canonical filtration $\overline{\mathbb{F}} = (\overline{\mathcal{F}}_t)_{t \in [0,T]}$ defined by $\overline{\mathcal{F}}_t := \sigma(\overline{X}_s, \overline{V}_s \;: s \le t)$. Let $\overline{\mathbb{P}}$ be a (Borel) probability measure on $\overline{\Omega}$, under which

$$\overline{X}_t := \overline{X}_0 + \int_0^t \overline{V}_s ds, \qquad \big(\sigma^{-1}\overline{V}_t\big)_{t \ge 0} \text{ is a Brownian motion,} \tag{4.3}$$

and $\overline{\mathbb{P}} \circ (\overline{X}_0, \overline{V}_0)^{-1} = \mathbb{P} \circ (X_0, V_0)^{-1}$.

Then under the measure $\overline{\mathbb{P}}$, $(\overline{X}_0, \overline{V}_0)$ is independent of $(\overline{X}_t - \overline{X}_0 - \overline{V}_0 t, \overline{V}_t - \overline{V}_0)$, and the latter follows a Gaussian distribution with mean value 0 and $2n \times 2n$ variance matrix

$$\sigma^2 \begin{pmatrix} t^3 I_n/3 & t^2 I_n/2 \\ t^2 I_n/2 & t I_n \end{pmatrix}. \tag{4.4}$$

Let $\overline{\mathbb{Q}} := \mathbb{P} \circ (X, V)^{-1}$ be the image measure of the solution (X, V) to the SDE (4.1), so that

$$d\overline{X}_t = \overline{V}_t dt, \quad d\overline{V}_t = b(t, \overline{X}_t, \overline{V}_t)dt + \sigma d\overline{W}_t, \quad \overline{\mathbb{Q}}\text{–a.s.}, \tag{4.5}$$

with a $\overline{\mathbb{Q}}$-Brownian motion $\overline{W}$. We are going to prove that $\overline{\mathbb{Q}}$ is equivalent to $\overline{\mathbb{P}}$ and

$$\frac{d\overline{\mathbb{Q}}}{d\overline{\mathbb{P}}}\Big|_{\overline{\mathcal{F}}_T} = Z_T, \quad \text{with } Z_t := \exp\Big(\int_0^t \sigma^{-2}\bar{b}_s \cdot d\overline{V}_s - \frac{1}{2}\int_0^t \big|\sigma^{-1}\bar{b}_s\big|^2 ds\Big), \tag{4.6}$$

where $\bar{b}_s := b(s, \overline{X}_s, \overline{V}_s)$.

Lemma 4.1. *The strictly positive random variable Z_T is a density under $\overline{\mathbb{P}}$, i.e. $\mathbb{E}^{\overline{\mathbb{P}}}[Z_T] = 1$.*

Proof We follow the arguments in [28, Lemma A.1] by Hu, Ren, Šiška and Szpruch. For simplification of the notation, we consider the case $\sigma = 1$. The general case follows by exactly the same arguments or simply by considering the corresponding SDE on $(\sigma^{-1}X, \sigma^{-1}V)$.

Let us denote $\bar{b}_t := b(t, \overline{X}_t, \overline{V}_t)$, $\widehat{Y}_t := Z_t(|\overline{X}_t|^2 + |\overline{V}_t|^2)$ and $f_\varepsilon(x) := \frac{x}{1+\varepsilon x}$. By Itô formula, one has

$$d\widehat{Y}_t \;=\; Z_t\big(2\overline{X}_t \cdot \overline{V}_t + 2\bar{b}_t \cdot \overline{V}_t + n\big)dt + Z_t\big(2V_t + \bar{b}_t(|\overline{X}_t|^2 + |\overline{V}_t|^2)\big) \cdot d\overline{V}_t, \quad \overline{\mathbb{P}}\text{–a.s.}$$

and

$$\begin{aligned} d\mathbb{E}^{\overline{\mathbb{P}}}\big[f_\varepsilon(\widehat{Y}_t)\big] \;&=\; \mathbb{E}\Big[\frac{Z_t\big(2\overline{X}_t \cdot \overline{V}_t + 2\bar{b}_t \cdot \overline{V}_t + n\big)}{(1+\varepsilon\widehat{Y}_t)^2} - \frac{\varepsilon Z_t^2\big|2\overline{V}_t + \bar{b}_t\big(|\overline{X}_t|^2 + |\overline{V}_t|^2\big)\big|^2}{(1+\varepsilon\widehat{Y}_t)^3}\Big]dt \\ &\le\; C\mathbb{E}\Big[\frac{Z_t\big(|\overline{X}_t|^2 + |\overline{V}_t|^2\big) + Z_t}{1+\varepsilon\widehat{Y}_t}\Big]dt, \end{aligned}$$

where we use the fact that $b(t, x, v)$ is of linear growth in (x, v), and $C > 0$ is a constant independent of ε.

Next, notice that $Z = (Z_t)_{0\le t\le T}$ is a positive local martingale under $\overline{\mathbb{P}}$, and hence a $\overline{\mathbb{P}}$-super martingale, so that $\mathbb{E}^{\overline{\mathbb{P}}}[Z_t] \le 1$ for all $t \in [0, T]$. Then

$$d\mathbb{E}^{\overline{\mathbb{P}}}\big[f_\varepsilon(\widehat{Y}_t)\big] \;\le\; C\Big(\mathbb{E}^{\overline{\mathbb{P}}}\big[f_\varepsilon(\widehat{Y}_t)\big] + 1\Big)dt \quad\Longrightarrow\quad \sup_{t\in[0,T]} \mathbb{E}^{\overline{\mathbb{P}}}\big[f_\varepsilon(\widehat{Y}_t)\big] \;<\; \infty.$$

Letting $\varepsilon \searrow 0$, it follows by Fatou Lemma that

$$\sup_{t\in[0,T]} \mathbb{E}^{\overline{\mathbb{P}}}\big[\widehat{Y}_t\big] \;=\; \sup_{t\in[0,T]} \mathbb{E}^{\overline{\mathbb{P}}}\big[Z_t\big(|\overline{X}_t|^2 + |\overline{V}_t|^2\big)\big] \;<\; \infty. \tag{4.7}$$

By the Itô formula, one obtains that, for all $t \in [0, T]$,

$$d\frac{Z_t}{1+\varepsilon Z_t} = \frac{Z_t\bar{b}_t}{(1+\varepsilon Z_t)^2} \cdot d\overline{W}_t - \frac{\varepsilon Z_t^2|\bar{b}_t|^2}{(1+\varepsilon Z_t)^3}dt$$

Taking expectation on both sides, we get

$$\mathbb{E}^{\overline{\mathbb{P}}}\Big[\frac{Z_t}{1+\varepsilon Z_t}\Big] - \frac{1}{1+\varepsilon} = -\mathbb{E}^{\overline{\mathbb{P}}}\Big[\int_0^t \frac{\varepsilon Z_s^2|\bar{b}_s|^2}{(1+\varepsilon Z_s)^3}ds\Big].$$

Together with the estimate (4.7), it follows from the monotone convergence and the dominated convergence theorem that $\mathbb{E}^{\overline{\mathbb{P}}}\big[Z_t\big] = 1$ for all $t \in [0, T]$. □

Lemma 4.2 (Existence of positive density)**.** *Let (X, V) be the solution of* (4.1)*. Then for all $t \in (0, T]$, (X_t, V_t) has a strictly positive density function, denoted by ρ_t.*

Proof Notice that under $\overline{\mathbb{P}}$, $(\overline{X}, \overline{V})$ can be written as the sum of a square integrable r.v. and an independent Gaussian r.v. with variance (4.4), then $\overline{\mathbb{P}} \circ (\overline{X}_t, \overline{V}_t)^{-1}$ has strictly positive and smooth density function. Besides, $\overline{\mathbb{Q}}$ is equivalent to $\overline{\mathbb{P}}$, with strictly positive density $d\overline{\mathbb{Q}}/d\overline{\mathbb{P}} = Z_T$, it follows that $\mathbb{P} \circ (X_t, V_t)^{-1} = \overline{\mathbb{Q}} \circ (\overline{X}_t, \overline{V}_t)^{-1}$ has also a strictly positive density function. □

Estimates on the densities We next provide an estimate on $\nabla_v(\ln \rho_t(x,v))$, which is crucial for proving Theorem 2.6.

Lemma 4.3 (Moment estimate)**.** *Suppose that* $\mathbb{E}\big[|X_0|^{2p}+|V_0|^{2p}\big]<\infty$ *for* $p\geq 1$*, then*

$$\mathbb{E}\Big[\sup_{0\leq t\leq T}\big(|X_t|^{2p}+|V_t|^{2p}\big)\Big]<\infty. \tag{4.8}$$

Consequently, the relative entropy between $\overline{\mathbb{Q}}$ *and* $\overline{\mathbb{P}}$ *is finite, i.e.*

$$H\big(\overline{\mathbb{Q}}|\overline{\mathbb{P}}\big) \;:=\; \mathbb{E}^{\overline{\mathbb{Q}}}\Big[\log\Big(\frac{d\overline{\mathbb{Q}}}{d\overline{\mathbb{P}}}\Big)\Big] \;=\; \mathbb{E}\Big[\frac{1}{2}\int_0^T\big|\sigma^{-1}b(t,X_t,V_t)\big|^2dt\Big]<\infty. \tag{4.9}$$

Proof Let us first consider (4.8). As b is of linear growth in (x,v), it is standard to apply Itô formula on $|X_t|^{2p}+|V_t|^{2p}$, and use BDG inequality and then Grownwall lemma to obtain (4.8).

Next, since $\mathbb{E}\big[|X_0|^2+|V_0|^2\big]<\infty$, it follows by (4.6) and (4.8) that

$$H\big(\overline{\mathbb{Q}}|\overline{\mathbb{P}}\big) \;=\; \mathbb{E}^{\overline{\mathbb{Q}}}\Big[\frac{1}{2}\int_0^T\big|\sigma^{-1}b(t,\overline{X}_t,\overline{V}_t)\big|^2dt\Big] \;=\; \mathbb{E}\Big[\frac{1}{2}\int_0^T\big|\sigma^{-1}b(t,X_t,V_t)\big|^2dt\Big]<\infty.$$

□

Let us introduce the time reverse process $(\widetilde{X},\widetilde{V})$ and time reverse probability measures $\widetilde{\mathbb{P}}$ and $\widetilde{\mathbb{Q}}$ on the canonical space $\overline{\Omega}$ by

$$\widetilde{X}_t:=\overline{X}_{T-t},\quad \widetilde{V}_t:=\overline{V}_{T-t},\quad \text{and}\quad \widetilde{\mathbb{P}}:=\overline{\mathbb{P}}\circ(\widetilde{X},\widetilde{V})^{-1},\quad \widetilde{\mathbb{Q}}:=\overline{\mathbb{Q}}\circ(\widetilde{X},\widetilde{V})^{-1}.$$

Lemma 4.4. *The density function* $\rho_t(x,v)$ *is absolutely continuous in* v*, and it holds that*

$$\mathbb{E}\Big[\int_t^T\big|\nabla_v\ln\big(\rho_s(X_s,V_s)\big)\big|^2ds\Big]<\infty,\quad \textit{for all}\quad t>0. \tag{4.10}$$

Proof This proof is largely based on the time-reversal argument in Föllmer [21, Lemma 3.1 and Theorem 3.10], where the author sought a similar estimate for a non-degenerate diffusion. For simplicity of notations, let us assume $\sigma=1$.

Step 1. We first prove that, $(\overline{X},\overline{V})$ is an Itô process under $\widetilde{\mathbb{Q}}$, and there exists a $\overline{\mathbb{F}}$–predictable process $\tilde{b}=(\tilde{b}_s)_{0\leq s\leq T}$ such that

$$\mathbb{E}^{\widetilde{\mathbb{Q}}}\Big[\int_0^{T-t}\big|\tilde{b}_s\big|^2ds\Big]<\infty,\quad \text{and } \overline{V}_t=\overline{V}_0+\int_0^t\tilde{b}_sds+\widetilde{W}_t,\ \text{for all } t>0, \tag{4.11}$$

with a $(\overline{\mathbb{F}},\widetilde{\mathbb{Q}})$–Brownian motion $\widetilde{W}$.

Let $\overline{\mathbb{P}}_{x_0,v_0}$ be the conditional probability of $\overline{\mathbb{P}}$ given $\overline{X}_0=x_0,\overline{V}_0=v_0$,

$$\overline{\mathbb{P}}_{x_0,v_0}[\cdot]=\overline{\mathbb{P}}\big[\,\cdot\,\big|\overline{X}_0=x_0,\overline{V}_0=v_0\big],\quad\text{and}\quad \widetilde{\mathbb{P}}_{x_0,v_0}:=\overline{\mathbb{P}}_{x_0,v_0}\circ(\overline{X},\overline{V})^{-1}.$$

Recall the dynamic of $(\overline{X},\overline{V})$ under $\overline{\mathbb{P}}$ in (4.3) and note that the marginal distribution of $(\overline{X}_t,\overline{V}_t)$ under $\overline{\mathbb{P}}_{x_0,v_0}$ is Gaussian, in particular, its density function $\rho_t^{x_0,v_0}(x,v)$ is smooth. It follows from Theorem 2.1 of Haussmann and Pardoux [24]) (or Theorem 2.3 of Millet, Nualart and Sanz [36]) that $\overline{V}$ is still a diffusion process w.r.t. $(\overline{\mathbb{F}},\widetilde{\mathbb{P}}_{x_0,v_0})$, and

$$\overline{V}_t-\overline{V}_0-\int_0^t\nabla_v\ln\rho_{T-s}^{x_0,v_0}(\overline{X}_s,\overline{V}_s)ds \text{ is a } (\overline{\mathbb{F}},\widetilde{\mathbb{P}}_{x_0,v_0})\text{–Brownian motion,}$$

where by direct computation we know

$$\nabla_v \ln \rho^{x_0,v_0}_{T-s}(\overline{X}_s, \overline{V}_s) \;=\; \frac{6(x_0+(T-s)v_0-\overline{X}_s)}{(T-s)^2} + \frac{4(v_0-\overline{V}_s)}{T-s} \;=:\; \tilde{c}_s(x_0,v_0).$$

Therefore,

$$\widetilde{W}^1_t := \overline{V}_t - \overline{V}_0 - \int_0^t \tilde{c}_s(\overline{X}_T, \overline{V}_T)ds \quad \text{is a } (\overline{\mathbb{F}}^*, \widetilde{\mathbb{P}})\text{–Brownian motion,}$$

where the enlarged filtration $\overline{\mathbb{F}}^* = (\overline{\mathcal{F}}^*_t)_{0\le t\le T}$ is defined by

$$\overline{\mathcal{F}}^*_t := \sigma\big(\overline{X}_T, \overline{V}_T, \overline{X}_s, \overline{V}_s \;: s\in[0,t]\big).$$

By the moment estimate (4.8), we have

$$\mathbb{E}^{\widetilde{\mathbb{Q}}}\Big[\int_0^{T-t} |\tilde{c}_s(\overline{X}_T,\overline{V}_T)|^2 ds\Big] = \mathbb{E}^{\overline{\mathbb{Q}}}\Big[\int_t^T |\tilde{c}_s(\overline{X}_0,\overline{V}_0)|^2 ds\Big] < \infty, \text{ for } t>0.$$

Next note that the relative entropy satisfies

$$H(\widetilde{\mathbb{Q}}|\widetilde{\mathbb{P}}) = H(\overline{\mathbb{Q}}|\overline{\mathbb{P}}) < \infty.$$

Therefore, there exists a $\overline{\mathbb{F}}^*$–predictable process $\tilde{a}$ such that $\mathbb{E}^{\widetilde{\mathbb{Q}}}\Big[\int_0^T |\tilde{a}_t|^2 dt\Big] < \infty$ and

$$\widetilde{W}^2_t := \widetilde{W}^1_t - \int_0^t \tilde{a}_s ds = \overline{V}_t - \overline{V}_0 - \int_0^t \big(\tilde{a}_s + \tilde{c}_s(\overline{X}_T,\overline{V}_T)\big)ds \quad \text{is a } (\overline{\mathbb{F}}^*, \widetilde{\mathbb{P}})\text{–Brownian motion.}$$

Finally we prove Claim (4.11), by letting $\tilde{b}_t$ denote an optional version of the process $\mathbb{E}^{\widetilde{\mathbb{Q}}}\big[\tilde{a}_t + \tilde{c}_t(\overline{X}_T,\overline{V}_T)\big|\overline{\mathcal{F}}_t\big]$.

Step 2. Let $R : \overline{\Omega} \to \overline{\Omega}$ be the reverse operator defined by $R(\bar{\omega}) = (\bar{\omega}_{T-t})_{0\le t\le T}$. Then for every fixed $t<T$ and $\varphi \in C_c(\mathbb{R}^{2n})$, one has

$$\mathbb{E}^{\overline{\mathbb{Q}}}\Big[\big(\tilde{b}_{T-t}\circ R\big)\varphi(\overline{X}_t,\overline{V}_t)\Big] \;=\; -\lim_{h\searrow 0}\frac{1}{h}\mathbb{E}^{\overline{\mathbb{Q}}}\Big[\big(\overline{V}_t - \overline{V}_{t-h}\big)\varphi\big(\overline{X}_t,\overline{V}_t\big)\Big].$$

Recall the dynamic of $(\overline{X},\overline{V})$ under $\overline{\mathbb{Q}}$ in (4.5), and thus

$$\begin{aligned}\varphi\big(\overline{X}_t,\overline{V}_t\big) \;&=\; \varphi\big(\overline{X}_{t-h},\overline{V}_{t-h}\big) + \int_{t-h}^t \nabla_x\varphi(\overline{X}_s,\overline{V}_s)\cdot V_s ds \\ &\quad + \int_{t-h}^t \nabla_v\varphi(\overline{X}_s,\overline{V}_s)\cdot d\overline{V}_s \;+\; \frac{1}{2}\int_{t-h}^t \Delta_v\varphi(\overline{X}_s,\overline{V}_s)ds, \quad \overline{\mathbb{Q}}\text{-a.s.}\end{aligned}$$

Denoting

$$\bar{b}_t \;:=\; b(t,\overline{X}_t,\overline{V}_t), \text{ which clearly satisfies that } \mathbb{E}^{\overline{\mathbb{Q}}}\Big[\int_0^T \big|\bar{b}_t\big|^2 dt\Big] < \infty, \tag{4.12}$$

we have

$$\mathbb{E}^{\overline{\mathbb{Q}}}\Big[\big(\tilde{b}_{T-t}\circ R\big)\varphi(\overline{X}_t,\overline{V}_t)\Big] = -\mathbb{E}^{\overline{\mathbb{Q}}}\Big[\bar{b}_t\varphi(\overline{X}_t,\overline{V}_t)\Big] - \mathbb{E}^{\overline{\mathbb{Q}}}\Big[\nabla_v\varphi(\overline{X}_s,\overline{V}_s)\Big].$$

Therefore, denoting by $\nabla_v\rho_t(x,v)$ the weak derivative of ρ in sense of distribution, one has

$$\int_{\mathbb{R}^{2n}} \nabla_v\rho_t(x,v)\varphi(x,v)dxdv = -\mathbb{E}^{\overline{\mathbb{Q}}}\Big[\nabla_v\varphi(\overline{X}_s,\overline{V}_s)\Big] = \mathbb{E}^{\overline{\mathbb{Q}}}\Big[\big(\tilde{b}_{T-t}\circ R + \bar{b}_t\big)\varphi(\overline{X}_t,\overline{V}_t)\Big].$$

As $\varphi \in C_c(\mathbb{R}^{2n})$ is arbitrary, this implies that, for a.e. (x,v),

$$\nabla_v \rho_t(x,v) \;=\; \rho_t(x,v)\mathbb{E}^{\overline{\mathbb{Q}}}\Big[\big(\tilde{b}_{T-t}\circ R + \bar{b}_t\big)\Big|\overline{X}_t = x, \overline{V}_t = v\Big].$$

Finally, it follows from the moment estimates in (4.11) and (4.12) that

$$\mathbb{E}^{\overline{\mathbb{Q}}}\Big[\int_{t_0}^{t_1}\big|\nabla_v \ln\big(\rho_t(\overline{X}_t,\overline{V}_t)\big)\big|^2 dt\Big] \;=\; \mathbb{E}^{\overline{\mathbb{Q}}}\Big[\int_{t_0}^{t_1}\Big|\frac{\nabla_v\rho_t}{\rho_t}(\overline{X}_t,\overline{V}_t)\Big|^2 dt\Big] \;<\; \infty.$$

We hence conclude the proof by the fact that $\mathbb{P}\circ(X,V)^{-1} = \overline{\mathbb{Q}}\circ(\overline{X},\overline{V})^{-1}$. □

From (4.11), we already know that $\overline{V}$ is a diffusion process w.r.t. $(\overline{\mathbb{F}},\widetilde{\mathbb{Q}})$. With the integrability result (4.10), we can say more on its dynamics.

Lemma 4.5. *The reverse process $(\widetilde{X},\widetilde{V})$ is a diffusion process under $\overline{\mathbb{Q}}$, or equivalently, the canonical process $(\overline{X},\overline{V})$ is a diffusion process under the reverse probability $\widetilde{\mathbb{Q}}$. Moreover, $\widetilde{\mathbb{Q}}$ is a weak solution to the SDE:*

$$d\overline{X}_t = -\overline{V}_t dt, \quad d\overline{V}_t = \big(-b(t,\overline{X}_t,\overline{V}_t) + \sigma^2\nabla_v \ln\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big)dt + \sigma d\widetilde{W}_t,\ \widetilde{\mathbb{Q}}\text{–a.s.}, \qquad (4.13)$$

where $\widetilde{W}$ is a $(\overline{\mathbb{F}},\widetilde{\mathbb{Q}})$–Brownian motion.

Proof It follows from the Cauchy-Schwarz inequality and (4.10) that

$$\int_t^T\int_{\mathbb{R}^{2n}}|\nabla_v\rho_s(x,v)|dxdv \le \left(\int_t^T\int_{\mathbb{R}^{2n}}\frac{|\nabla_v\rho_s(x,v)|^2}{\rho_s(x,v)^2}\rho_s(x,v)dxdv\right)^{\frac{1}{2}} < \infty, \quad \text{for all } T > t > 0.$$

Together with the Lipschitz assumption on the coefficient $b(t,x,v)$, the desired result is a direct consequence of Theorem 2.1 of Haussmann and Pardoux [24], or Theorem 2.3 of Millet, Nualart and Sanz [36]. □

Finally, we provide a sufficient condition on b to ensure that the density function ρ of (X,V) is a smooth function.

Lemma 4.6 (Regularity of the density)**.** *Assume in addition that $b \in C^\infty((0,T)\times\mathbb{R}^{2n})$ with all derivatives of order k bounded for all $k \ge 1$. Then the function $(t,x,v)\mapsto\rho_t(x,v)$ belongs to $C^\infty((0,T)\times\mathbb{R}^{2n})$.*

Proof Under the additional regularity conditions on b, it is easy to check that the coefficients of SDE (4.1) satisfies the Hörmander's conditions, and hence the density function $\rho \in C^\infty((0,T)\times\mathbb{R}^{2n})$ (see e.g. Bally [4, Theorem 5.1, Remark 5.2]). □

Application to the MFL equation (2.5) We will apply the above technical results to the MFL equation (2.5). Let (X,V) be the unique solution of (2.5), and $m_t^X := \mathcal{L}(X_t)$, then (X,V) is also the unique solution of SDE (4.1) with drift coefficient function

$$b(t,x,v) \;:=\; D_m F_\circ(m_t^X,x) + \nabla_x f(x) + \gamma v. \qquad (4.14)$$

Proposition 4.7. (i) *Let Assumption 2.1 hold true, then $b(t,x,v)$ is a continuous function, uniformly Lipschitz in (x,v).*

(ii) *Suppose in addition that Assumption 2.5 holds true, then $b \in C^\infty((0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n)$ and and for each $k \ge 1$, its derivative of order k is bounded.*

Proof (i) For a diffusion process (X,V), it is clear that $t \mapsto m_t^X$ is continuous, then $(t,x,v) \mapsto b(t,x,v) := D_m F(m_t^X, x) + \gamma v$ is continuous. Moreover, it is clear that b is globally Lipschitz in (x,v) under Assumption 2.1.

(ii) Let us denote

$$b_\circ(t,x) \;:=\; D_m F_\circ(m_t^X, x).$$

We claim that for the coefficient function b defined in (4.14), for all $k \geq 0$, one has

$$\partial_t^k b_\circ(t,x) \;=\; \mathbb{E}\Big[\sum_{i=0}^{k}\sum_{j=0}^{k-i} \varphi_{i,j}^n\big(m_t^X, X_t, V_t, x\big) X_t^i V_t^j\Big], \tag{4.15}$$

where $\varphi_{i,j}^n$ are smooth functions with bounded derivatives of any order.

Further, it follows by Lemma 4.3 that, under additional conditions in Assumption 2.5, one has $\mathbb{E}\big[\sup_{0\leq t\leq T}(|X_t|^p + |V_t|^p)\big] < \infty$ for all $T > 0$ and $p \geq 1$. Therefore, one has $b_\circ \in C^\infty((0,\infty)\times\mathbb{R}^n)$ and hence $b \in C^\infty((0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n)$.

It is enough to prove (4.15). Recall (see e.g. Carmona and Delarue [8]) that for a smooth function $\varphi : \mathcal{P}(\mathbb{R}^n)\times\mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}$, one has the Itô formula

$$\begin{aligned} d\varphi(m_t^X, X_t, V_t) \;&=\; \mathbb{E}\big[D_m\varphi(m_t^X)(X_t)\cdot V_t\big]dt + \nabla_x\varphi(m_t^X, X_t, V_t)\cdot V_t dt \\ &\quad -\nabla_v\varphi(m_t^X, X_t, V_t)\big(D_m F(m_t^X, X_t), X_t) + \gamma V_t\big)dt + \frac{1}{2}\sigma^2\Delta_v\varphi(m_t^X, X_t, V_t)dt \\ &\quad +\nabla_v\varphi(m_t^X, X_t, V_t)\cdot\sigma dW_t. \end{aligned}$$

Then we can easily conclude the proof of (4.15) by the induction argument. □

4.2 Proof of Theorem 2.6

Let us fix $T > 0$, and consider the reverse probability $\widetilde{\mathbb{Q}}$ given before Lemma 4.4 with coefficient function b in (4.14). Recall also the dynamic of $(\overline{X},\overline{V})$ under $\widetilde{\mathbb{Q}}$ in (4.13). Applying Itô formula on $\ln\big(\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big)$, and then using the Fokker-Planck equation (1.2), it follows that

$$\begin{aligned} &d\ln\big(\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big) \\ =\;& \Big(-\frac{\partial_t\rho_{T-t}}{\rho_{T-t}}(\overline{X}_t,\overline{V}_t) - \nabla_x\ln\big(\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big)\cdot\overline{V}_t + \frac{1}{2}\sigma^2\Delta_v\ln\big(\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big) \\ &\qquad +\nabla_v\ln\big(\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big)\cdot\big(-b(t,\overline{X}_t,\overline{V}_t) + \sigma^2\nabla_v\ln\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big)\Big)dt \\ &+\nabla_v\ln\big(\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big)\cdot\sigma dW_t \\ =\;& \Big(-n\gamma + \frac{1}{2}\Big|\frac{\sigma\nabla_v\rho_{T-t}}{\rho_{T-t}}(\overline{X}_t,\overline{V}_t)\Big|^2\Big)dt + \nabla_v\ln\big(\rho_{T-t}(\overline{X}_t,\overline{V}_t)\big)\cdot\sigma d\widetilde{W}_t, \quad \widetilde{\mathbb{Q}}\text{–a.s.} \end{aligned}$$

By (4.10), it follows that for $t > 0$

$$dH(m_t) \;=\; d\mathbb{E}^{\widetilde{\mathbb{Q}}}\Big[\ln\big(\rho_t(X_{T-t},V_{T-t})\big)\Big] \;=\; \Big(-n\gamma + \frac{1}{2}\mathbb{E}\Big[\big|\sigma\nabla_v\ln\big(\rho_t(X_t,X_t)\big)\big|^2\Big]\Big)dt. \tag{4.16}$$

On the other hand, recall that

$$F(m) = F_\circ(m) + \mathbb{E}^m[f(X)], \quad\text{and}\quad D_m F\big(\mathcal{L}(X_t)\big) = D_m F_\circ\big(\mathcal{L}(X_t)\big) + \nabla f. \tag{4.17}$$

By a direct computation, one has

$$dF_\circ\big(\mathcal{L}(X_t)\big) = \mathbb{E}\big[D_m F_\circ\big(\mathcal{L}(X_t), X_t\big)\cdot V_t\big]dt. \tag{4.18}$$

By Itô formula and (4.17), one has

$$\begin{aligned} d\Big(f(X_t)+\frac{1}{2}|V_t|^2\Big) &= \Big(\nabla f(X_t)\cdot V_t - V_t\cdot\big(D_mF(\mathcal{L}(X_t),X_t)+\gamma V_t\big)+\frac{1}{2}\sigma^2 n\Big)dt + V_t\cdot\sigma dW_t \\ &= \Big(-D_mF_\circ(\mathcal{L}(X_t),X_t)\cdot V_t-\gamma|V_t|^2+\frac{1}{2}\sigma^2 n\Big)dt + V_t\cdot\sigma dW_t. \end{aligned} \tag{4.19}$$

Combining (4.16), (4.18) and (4.19), we obtain

$$\begin{aligned} d\mathfrak{F}(m_t) &= d\Big(F\big(\mathcal{L}(X_t)\big)+\frac{1}{2}\mathbb{E}\big[|V_t|^2\big]+\frac{\sigma^2}{2\gamma}H(m_t)\Big) \\ &= \mathbb{E}\Big[-\gamma|V_t|^2+\sigma^2 n-\frac{\sigma^4}{4\gamma}\big|\nabla_v\ln\big(\rho_t(X_t,V_t)\big)\big|^2\Big]dt. \end{aligned} \tag{4.20}$$

Further, by Lemmas 4.3 and 4.4, it is clear that $\mathbb{E}\big[\big|\nabla_v\ln\big(\rho_t(X_t,V_t)\big)\cdot V_t\big|\big]<\infty$ and by integration by parts we have

$$\begin{aligned} \mathbb{E}\Big[\nabla_v\ln\big(\rho_t(X_t,V_t)\big)\cdot V_t\Big] &= \frac{1}{2}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\Big(\nabla_v\rho_t(x,v)\cdot\nabla_v|v|^2\Big)dxdv \\ &= -\frac{1}{2}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\Big(\rho_t(x,v)\Delta_v|v|^2\Big)dxdv \;=\; -n. \end{aligned}$$

Together with (4.20), it follows

$$d\mathfrak{F}(m_t) = -\gamma\mathbb{E}\left[\left|V_t+\frac{\sigma^2}{2\gamma}\nabla_v\ln\big(\rho_t(X_t,V_t)\big)\right|^2\right]dt.$$

□

4.3 Proof of Theorem 2.7

Let $(m_t)_{t\in\mathbb{R}^+}$ be the flow of marginal laws of the solution to (2.5), given an initial law $m_0\in\mathcal{P}_2^{2n}$. Define a dynamic system $S(t)[m_0]:=m_t$. We shall consider the so-called w-limit set:

$$w(m_0):=\Big\{\mu\in\mathcal{P}_2^{2n}:\quad\text{there exist } t_k\to\infty \text{ such that } \mathcal{W}_1\Big(S(t_k)[m_0],\mu\Big)\to 0\Big\}$$

We recall LaSalle's invariance principle.

Proposition 4.8. *[Invariance Principle] Let Assumption 2.1 and 2.5 hold true. Then the set $w(m_0)$ is nonempty, $\mathcal{W}_1$-compact and invariant, that is,*

1. *for any $\mu\in w(m_0)$, we have $S(t)[\mu]\in w(m_0)$ for all $t\geq 0$.*
2. *for any $\mu\in w(m_0)$ and all $t\geq 0$, there exits $\mu'\in w(m_0)$ such that $S(t)[\mu']=\mu$.*

Proof Under the upholding assumptions, it follows from Lemma 4.3 that $t\mapsto S(t)$ is continuous with respect to the $\mathcal{W}_1$-topology. On the other side, due to Theorem 2.6 and the fact that the relative entropy $H\geq 0$, we know that $\{F(m_t)+\frac{1}{2}\mathbb{E}[|V_t|^2]\}_{t\geq 0}$ is bounded. Together with (2.2), we obtain

$$\sup_{t\geq 0}\mathbb{E}\big[|X_t|^2+|V_t|^2\big]<\infty.$$

Therefore $\big(S(t)[m_0]\big)_{t\geq 0}=(m_t)_{t\geq 0}$ live in a $\mathcal{W}_1$-compact subset of $\mathcal{P}_2^{2n}$. The desired result follows from the invariance principle, see e.g. Henry [25, Theorem 4.3.3]. □

Lemma 4.9. *Let Assumption 2.1 and 2.5 hold true. Then, every $m^* \in w(m_0)$ has a density and we have*

$$v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0, \quad \text{Leb}^{2n} - a.s. \tag{4.21}$$

Proof Let $m^* \in w(m_0)$ and denote by $(m_{t_k})_{k\in\mathbb{N}}$ the subesquence converging to m^* in $\mathcal{W}_1$.

Step 1. We first prove that there exists a sequence $\delta_i \to 0$ such that

$$\liminf_{k\to\infty} \mathbb{E}\Big[\big|V_{t_k+\delta_i} + \frac{\sigma^2}{2\gamma}\nabla_v \ln\big(m_{t_k+\delta_i}(X_{t_k+\delta_i}, V_{t_k+\delta_i})\big)\big|^2\Big] = 0, \quad \text{for all } i \in \mathbb{N}. \tag{4.22}$$

Suppose the contrary. Then we would have for some $\delta > 0$

$$\begin{aligned} 0 &< \int_0^\delta \liminf_{k\to\infty} \mathbb{E}\Big[\big|V_{t_k+s} + \frac{\sigma^2}{2\gamma}\nabla_v \ln\big(m_{t_k+s}(X_{t_k+s}, V_{t_k+s})\big)\big|^2\Big] ds \\ &\le \liminf_{k\to\infty} \int_0^\delta \mathbb{E}\Big[\big|V_{t_k+s} + \frac{\sigma^2}{2\gamma}\nabla_v \ln\big(m_{t_k+s}(X_{t_k+s}, V_{t_k+s})\big)\big|^2\Big] ds, \end{aligned}$$

where the last inequality is due to Fatou's lemma. This is a contradiction against Theorem 2.6 and the fact that $\mathfrak{F}$ is bounded from below.

Step 2. Denote by $t_k^i := t_k + \delta_i$ and $m_t^* := S(t)[m^*]$. Note that

$$\lim_{k\to\infty} \mathcal{W}_1\Big(m_{t_k}, m^*\Big) = 0 \quad \Longrightarrow \quad \lim_{k\to\infty} \mathcal{W}_1\Big(m_{t_k^i}, m^*_{\delta_i}\Big) = \lim_{k\to\infty} \mathcal{W}_1\Big(S(\delta_i)[m_{t_k}], S(\delta_i)[m^*]\Big) = 0.$$

Now fix $i \in \mathbb{N}$. Due to Theorem 2.6 and the fact that $\{F(m_t) + \frac{1}{2}\mathbb{E}[|V_t|^2]\}_{t\ge 0}$ is bounded from below, the set $\{H(m_{t_k^i})\}_{k\in\mathbb{N}}$ is uniformly bounded. Therefore the densities $(m_{t_k^i})_{k\in\mathbb{N}}$ are uniformly integrable with respect to Lebesgue measure, and thus m^* has a density. Note that

$$\begin{aligned} &\mathbb{E}\Big[\big|V_{t_k^i} + \frac{\sigma^2}{2\gamma}\nabla_v \ln\big(m_{t_k^i}(X_{t_k^i}, V_{t_k^i})\big)\big|^2\Big] \\ = \;& \frac{\sigma^4}{4\gamma^2}\int_{\mathbb{R}^{2n}} \frac{\Big|\nabla_v\big(m_{t_k^i}(x,v)e^{\frac{\gamma}{\sigma^2}|v|^2}\big)\Big|^2}{m_{t_k^i}(x,v)e^{\frac{\gamma}{\sigma^2}|v|^2}} e^{-\frac{\gamma}{\sigma^2}|v|^2} dxdv \end{aligned}$$

Denote by $\mu_v^* := \mathcal{N}(0, \frac{\sigma^2}{2\gamma} I_n)$ and define the function $h_k^i(x,v) := m_{t_k^i}(x,v)e^{\frac{\gamma}{\sigma^2}|v|^2}$. By logarithmic Sobolev inequality for the Gaussian distribution we obtain

$$\int \left(\int h_k^i \ln h_k^i d\mu_v^* - \int h_k^i d\mu_v^* \ln \int h_k^i d\mu_v^*\right) dx \le C \int \frac{|\nabla_v h_k^i|^2}{h_k^i} d\mu_v^* dx.$$

Together with (4.22) we obtain

$$\begin{aligned} 0 &= \lim_{k\to\infty} \mathbb{E}\Big[\big|V_{t_k^i} + \frac{\sigma^2}{2\gamma}\nabla_v \ln\big(m_{t_k^i}(X_{t_k^i}, V_{t_k^i})\big)\big|^2\Big] \\ &\ge C \limsup_{k\to\infty} \int \left(\int h_k^i \ln h_k^i d\mu_v^* - \int h_k^i d\mu_v^* \ln \int h_k^i d\mu_v^*\right) dx. \end{aligned} \tag{4.23}$$

Since $\int h_k^i d\mu_v^* = \int m_{t_k^i} dv = m_{t_k^i}^X$, we further have

$$\begin{aligned} 0 &\geq C \limsup_{k\to\infty} \int \left(m_{t_k^i} \ln h_k^i - m_{t_k^i} \ln m_{t_k^i}^X \right) dvdx \\ &= C \limsup_{k\to\infty} \int \left(m_{t_k^i} \ln \frac{m_{t_k^i}}{m_{t_k^i}^X e^{-\frac{\gamma}{\sigma^2}|v|^2}} \right) dvdx \\ &= C \limsup_{k\to\infty} H\left(m_{t_k^i} \Big| m_{t_k^i}^X \times \mathcal{N}(0, \frac{\sigma^2}{2\gamma} I_n) \right) \\ &\geq CH\left(m_{\delta_i}^* \Big| m_{\delta_i}^{*,X} \times \mathcal{N}(0, \frac{\sigma^2}{2\gamma} I_n) \right). \end{aligned}$$

The last inequality is due to the lower semi-continuity of the relative entropy in weak topology. Finally, since $\lim_{i\to\infty} \mathcal{W}_1(m_{\delta_i}^*, m^*) = 0$, we get $H\left(m^* \Big| m^{*,X} \times \mathcal{N}(0, \frac{\sigma^2}{2\gamma} I_n) \right) = 0$ and thus

$$m^* = m^{*,X} \times \mathcal{N}(0, \frac{\sigma^2}{2\gamma} I_n).$$

This immediately implies (4.21). □

Lemma 4.10. *Let Assumption 2.1 and 2.5 hold true. Then, each $m^* \in w(m_0)$ is equivalent to Lebesgue measure.*

Proof By the invariant principle we may find a probability measure $m^\circ \in w(m_0)$ such that $m^* = S(t)[m^\circ]$ for a fixed $t > 0$. Then the desired result follows from Lemma 4.2. □

Note that the necessary condition (4.21) for $m^* \in w(m_0)$ is not enough to identify $m^* = \underline{m}$. We are going to trigger the invariance principle to complete the proof of Theorem 2.7.

Proof of Theorem 2.7. Let $m^* \in w(m_0)$ and define $m_t^* := S(t)[m^*]$ for all $t \geq 0$. Denote by $(X_t^*, V_t^*)_{t\geq 0}$ the solution to the MFL equation (2.5) with initial distribution m^*. Take a test function $h \in C^1(\mathbb{R}^n)$ with compact support. It follows from Itô's formula that

$$dV_t^* h(X_t^*) = \Big(-h(X_t^*)\big(D_m F(m_{t,X}^*, X_t^*) + \gamma V_t^*\big) + \big(\nabla_x h(X_t^*) \cdot V_t^*\big) V_t^* \Big) dt + \sigma h(X_t^*) dW_t. \quad (4.24)$$

By the invariance principle, we have $m_t^* \in w(m_0)$ for all $t \geq 0$, and by Lemma 4.9 we have

$$v + \frac{\sigma^2}{2\gamma} \nabla_v \ln m_t^*(x, v) = 0, \quad \text{Leb}^{2n} - a.s.$$

So there exists a measurable function $(t, x) \mapsto \hat{m}_t(x)$ such that $m_t^*(x, v) = e^{-\frac{\gamma}{\sigma^2}|v|^2} \hat{m}_t(x)$. In particular, we observe that for each $t \geq 0$, the random variables X_t^*, V_t^* are independent and V_t^* follows the Gaussian distribution $\mathcal{N}(0, \frac{\sigma^2}{2\gamma} I_n)$. Taking expectation on both sides of (4.24), we obtain

$$\begin{aligned} 0 &= \mathbb{E}\left[-h(X_t^*)\big(D_m F(m_{t,X}^*, X_t^*) + \gamma V_t^*\big) + \big(\nabla_x h(X_t^*) \cdot V_t^*\big) V_t^* \right] \quad \text{for a.s. } t \\ &= \mathbb{E}\left[-h(X_t^*) D_m F(m_{t,X}^*, X_t^*) + \frac{\sigma^2}{2\gamma} \nabla_x h(X_t^*) \right]. \end{aligned} \quad (4.25)$$

Observe that

$$\mathbb{E}\left[\nabla_x h(X_t^*)\right] = C_t \int_{\mathbb{R}^n} \nabla_x h(x) \hat{m}_t(x) dx = -C_t \int_{\mathbb{R}^n} h(x) \nabla_x \hat{m}_t(x) dx,$$

where C_t is the normalization constant such that $C_t\hat{m}_t$ is a density function, and $\nabla_x\hat{m}_t$ is the weak derivative in sense of distribution. Together with (4.25) we have

$$\int_{\mathbb{R}^n} h(x)\Big(- D_mF(m^*_{t,X},x)\hat{m}_t(x) - \frac{\sigma^2}{2\gamma}\nabla_x\hat{m}_t(x)\Big)dx = 0.$$

Since h is arbitrary, we have

$$D_mF(m^*_{t,X},x) + \frac{\sigma^2}{2\gamma}\nabla_x \ln \hat{m}_t(x) = 0, \quad m^*_t\text{-a.s. for a.s. } t.$$

By Lemma 4.10, m^*_t is equivalent to Lebesgue measure, and thus we have

$$\begin{cases} D_mF(m^*_{t,X},x) + \frac{\sigma^2}{2\gamma}\nabla_x \ln m^*_t(x,v) = 0, \\ v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*_t(x,v) = 0, \end{cases} \quad \text{for all } (x,v) \in \mathbb{R}^+ \times \mathbb{R}^{2n}, \text{ for a.s. } t.$$

Since F is convex, by Lemma 2.2 we have $m^*_t = \operatorname{argmin}_{m\in\mathcal{P}^{2n}} \mathfrak{F}(m) =: \underline{m}$ for a.s. t. Taking into account that $\lim_{t\to 0}\mathcal{W}_1(m^*_t, m^*) = 0$, we obtain $m^* = \underline{m}$. Finally, since m^* is arbitrary, we conclude that $w(m_0) = \{\underline{m}\}$ is a singleton, and thus $\lim_{t\to\infty}\mathcal{W}_1(m_t,\underline{m}) = 0$. □

4.4 Exponential ergodicity given small mean-field dependence

Under Assumption 2.1 and Assumption 2.9, we consider the following equation:

$$\begin{cases} dX_t = V_t dt, \\ dV_t = -\big(b(m^X_t, X_t) + \lambda X_t + \gamma V_t\big)dt + \sigma dW_t, \end{cases} \tag{4.26}$$

where $b : \mathcal{P}(\mathbb{R}^n)\times\mathbb{R}^n \to \mathbb{R}$ is Lipschitz in the variable x

$$\big|b(m,x) - b(m,x')\big| \le L^x|x-x'|,$$

and for any $\varepsilon > 0$ there exists $M > 0$ such that for any $m, m' \in \mathcal{P}(\mathbb{R}^n)$

$$\big|b(m,x) - b(m',x')\big| \le \varepsilon|x-x'|, \quad \text{whenever } |x-x'| \ge M,$$

and for each $x \in \mathbb{R}^n$

$$\big|b(m,x) - b(m',x)\big| \le \iota\mathcal{W}_1(m,m'), \quad \text{with some sufficiently small } \iota > 0. \tag{4.27}$$

4.4.1 Reflection-Synchronous Coupling

We are going to show the contraction result in Theorem 2.11 via the coupling technique. Let (X,V) and (X',V') be the two solutions of (4.26) driven by the Brownian motions W and W', respectively. Define $\delta X = X - X'$ and $\delta V = V - V'$. We introduce the change of variable

$$P_t := \delta V_t + \gamma\delta X_t.$$

Then, the processes δX and P satisfy the following stochastic differential equations

$$\begin{aligned} d\delta X_t &= (P_t - \gamma\delta X_t)dt, \\ dP_t &= -\big(\delta b_t + \lambda\delta X_t\big)dt + \sigma d\delta W_t, \end{aligned}$$

where $\delta W = W - W'$ and $\delta b_t := b\big(m^X_t, X_t\big) - b\big(m^{X'}_t, X'_t\big)$.

Remark 4.11. *We shall apply the reflection-synchronous coupling following the blueprint in Eberle, Guillin and Zimmer [18], of which the main idea is to separate the space $\mathbb{R}^n \times \mathbb{R}^n$ into two parts:* (i). *$(\delta X_t, P_t)$ locates in a compact set;* (ii). *$|\delta X_t| + \eta|P_t|$ is big enough, where the constant η is to be determined. As in [18] we are going to apply the reflection coupling on the area* (i) *and the synchronous coupling on the area* (ii). *However, note that in [18] the argument for the contraction on the area* (ii) *relies on a Lyapunov function, which can no longer play its role in the mean-field context. Therefore, we are going to construct another function (the function G in* (2.6)*) which decays exponentially on the area* (ii).

Recall the definitions

$$r_t := |\delta X_t|, \quad u_t := |P_t|, \quad z_t := \delta X_t \cdot P_t.$$

Let $\xi > 0$. For technical reason we shall also apply the synchronous coupling on the area $u_t < \xi$, and eventually we will let $\xi \downarrow 0$. In order to couple the two processes $(X, V), (X', V')$, we consider two Lipschitz continuous functions $\mathrm{rc}, \mathrm{sc} : \mathbb{R}^{2n} \to [0, 1]$ such that $\mathrm{rc}_t^2 + \mathrm{sc}_t^2 \equiv 1$,

$$\mathrm{rc}(\delta X_t, P_t) = \begin{cases} 0, & \text{when } \ u_t = 0 \ \text{ or } \ r_t + \eta u_t \geq 2M + \xi, \\ 1, & \text{when } \ u_t \geq \xi \ \text{ and } \ r_t + \eta u_t \leq 2M. \end{cases}$$

The values of the constants η, $M \in (0, \infty)$ will be determined later. Define

$$e_t^x := \begin{cases} \frac{\delta X_t}{|\delta X_t|}, & \text{if } \delta X_t \neq 0, \\ 0, & \text{if } \delta X_t = 0, \end{cases} \quad \text{and} \quad e_t^p := \begin{cases} \frac{P_t}{|P_t|}, & \text{if } P_t \neq 0, \\ 0, & \text{if } P_t = 0. \end{cases}.$$

With two independent Brownian motions W^{rc} and W^{sc} we consider the following coupling

$$\begin{cases} dW_t = \mathrm{rc}(\delta X_t, P_t) dW_t^{\mathrm{rc}} + \mathrm{sc}(\delta X_t, P_t) dW_t^{\mathrm{sc}}, \\ dW_t' = \mathrm{rc}(\delta X_t, P_t)\big(I_n - 2e_t^p (e_t^p)^\top\big) dW_t^{\mathrm{rc}} + \mathrm{sc}(\delta X_t, P_t) dW_t^{\mathrm{sc}}, \end{cases}$$

in particular we have $d\delta W_t = 2\mathrm{rc}(\delta X_t, P_t) e_t^p (e_t^p)^\top dW_t^{\mathrm{rc}}$. By Lévy characterization, the process $B_t := (e_t^p)^\top W_t^{\mathrm{rc}}$ is a one-dimensional Brownian motion. For the sake of simplicity, denote $\mathrm{rc}_t := \mathrm{rc}(\delta X_t, P_t)$. We notice that the Lipschitz continuity of the functions rc, sc ensures the existence and uniqueness of the coupling process.

To conclude, with the reflection-synchronous coupling, the processes δX and P satisfy the following stochastic differential equations

$$\begin{aligned} d\delta X_t &= (P_t - \gamma \delta X_t) dt, \\ dP_t &= -\big(\delta b_t + \lambda \delta X_t\big) dt + 2\sigma \mathrm{rc}_t e_t^p dB_t. \end{aligned} \tag{4.28}$$

4.4.2 The auxiliary function

As reported in Remark 4.11, the main novelty of our contraction result is to construct a function exponentially decaying along the process (4.28) whereas $r + \eta u$ is big enough. In this subsection we are going to construct the auxiliary function according to the different settings.

First, it follows from (4.28) and Itô's formula that

$$\begin{aligned} dr_t &= \frac{1}{|\delta X_t|} \delta X_t \cdot d\delta X_t = \big(e_t^x \cdot P_t - \gamma r_t\big) dt, \\ du_t &= -\big(\delta b_t \cdot e_t^p + \lambda e_t^p \cdot \delta X_t\big) dt + 2\sigma \mathrm{rc}_t dB_t. \end{aligned}$$

Also we have

$$d \begin{pmatrix} z_t \\ r_t^2 \\ u_t^2 \end{pmatrix} = A \begin{pmatrix} z_t \\ r_t^2 \\ u_t^2 \end{pmatrix} dt + \begin{pmatrix} -\delta b_t \cdot \delta X_t \\ 0 \\ -2\delta b_t \cdot P_t + 4\mathrm{rc}_t^2 \sigma^2 \end{pmatrix} dt + \begin{pmatrix} 2\frac{z_t}{u_t} \\ 0 \\ 4u_t \end{pmatrix} \sigma \mathrm{rc}_t dB_t \tag{4.29}$$

with the matrix

$$A := \begin{pmatrix} -\gamma & -\lambda & 1 \\ 2 & -2\gamma & 0 \\ -2\lambda & 0 & 0 \end{pmatrix}.$$

Remark 4.12. *As we will show later, the value of δb_t is small whereas $r_t + \eta u_t$ is big enough. Therefore, the coupling system is nearly linear and its contraction rate mainly depends on the matrix A.*

The eigenvalues of A solve the equation:

$$0 = (\zeta+\gamma)(\zeta+2\gamma)\xi + 2\lambda(\zeta+2\gamma) + 2\lambda\zeta = (\zeta+\gamma)(\zeta^2+2\gamma\zeta+4\lambda).$$

We divide the discussion into two cases, based on the different values of λ and γ.

(a) If $\lambda < \frac{\gamma^2}{4}$, the matrix has three different negative eigenvalues

$$\zeta = -\gamma, \quad \zeta = -\gamma + \sqrt{\gamma^2 - 4\lambda}, \quad \zeta = -\gamma - \sqrt{\gamma^2 - 4\lambda},$$

in particular, it can be diagonalized. More precisely, we have $QA = \Lambda Q$ with the transformation matrix

$$Q := \begin{pmatrix} -\gamma & \lambda & 1 \\ -\gamma + \sqrt{\gamma^2-4\lambda} & \frac{1}{2}\left(\gamma^2 - 2\lambda - \gamma\sqrt{\gamma^2-4\lambda}\right) & 1 \\ -\gamma - \sqrt{\gamma^2-4\lambda} & \frac{1}{2}\left(\gamma^2 - 2\lambda + \gamma\sqrt{\gamma^2-4\lambda}\right) & 1 \end{pmatrix}$$

and the diagonal matrix $\Lambda = \operatorname{diag}\big(-\gamma, -\gamma+\sqrt{\gamma^2-4\lambda}, -\gamma-\sqrt{\gamma^2-4\lambda}\big)$. Multiply Q on both sides of (4.29) and obtain

$$dQ\begin{pmatrix} z_t \\ r_t^2 \\ u_t^2 \end{pmatrix} = \Lambda Q \begin{pmatrix} z_t \\ r_t^2 \\ u_t^2 \end{pmatrix} dt + Q \begin{pmatrix} -\delta b_t \cdot \delta X_t \\ 0 \\ -2\delta b_t \cdot P_t + 4\mathrm{rc}_t^2\sigma^2 \end{pmatrix} dt + Q \begin{pmatrix} 2\frac{z_t}{u_t} \\ 0 \\ 4u_t \end{pmatrix} \sigma \mathrm{rc}_t dB_t. \quad (4.30)$$

Further note that

$$\begin{aligned} \big(-\gamma+\sqrt{\gamma^2-4\lambda}\big)z_t + \frac{1}{2}\left(\gamma^2-2\lambda-\gamma\sqrt{\gamma^2-4\lambda}\right)r_t^2 + u_t^2 &= \left|\frac{\gamma-\sqrt{\gamma^2-4\lambda}}{2}\delta X_t - P_t\right|^2, \\ \big(-\gamma-\sqrt{\gamma^2-4\lambda}\big)z_t + \frac{1}{2}\left(\gamma^2-2\lambda+\gamma\sqrt{\gamma^2-4\lambda}\right)r_t^2 + u_t^2 &= \left|\frac{\gamma+\sqrt{\gamma^2-4\lambda}}{2}\delta X_t - P_t\right|^2. \end{aligned} \quad (4.31)$$

Now define the function G:

$$G(\delta X_t, P_t) := \left|\frac{\gamma-\sqrt{\gamma^2-4\lambda}}{2}\delta X_t - P_t\right|^2 + \left|\frac{\gamma+\sqrt{\gamma^2-4\lambda}}{2}\delta X_t - P_t\right|^2.$$

Denote by $G_t := G(\delta X_t, P_t)$. Together with (4.30), (4.31), we obtain

$$dG_t \le -\big(\gamma-\sqrt{\gamma^2-4\lambda}\big)G_t dt + \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}^\top Q \left\{ \begin{pmatrix} -\delta b_t \cdot \delta X_t \\ 0 \\ -2\delta b_t \cdot P_t + 4\mathrm{rc}_t^2\sigma^2 \end{pmatrix} dt + \begin{pmatrix} 2\frac{z_t}{u_t} \\ 0 \\ 4u_t \end{pmatrix} \sigma \mathrm{rc}_t dB_t \right\}.$$

(b) If $\lambda > \frac{\gamma^2}{4}$, the eigenvalues of A are

$$\zeta = -\gamma, \quad \zeta = -\gamma + i\sqrt{4\lambda-\gamma^2}, \quad \zeta = -\gamma - i\sqrt{4\lambda-\gamma^2}.$$

We have $QA = \Lambda Q$ with the transformation matrix

$$Q := \begin{pmatrix} -\gamma & \lambda & 1 \\ 4\lambda & -\lambda\gamma & -\gamma \\ 0 & \lambda\sqrt{4\lambda-\gamma^2} & -\sqrt{4\lambda-\gamma^2} \end{pmatrix}$$

and the standard form

$$\Lambda := \begin{pmatrix} -\gamma & 0 & 0 \\ 0 & -\gamma & -\sqrt{4\lambda-\gamma^2} \\ 0 & \sqrt{4\lambda-\gamma^2} & -\gamma \end{pmatrix}.$$

Multiplying Q on both sides of (4.29), we again obtain (4.30). Now note that

$$-\gamma z_t + \lambda r_t^2 + u_t^2 = \left|\frac{\gamma}{2}\delta X_t - P_t\right|^2 + \left(\lambda - \frac{\gamma^2}{4}\right)|\delta X_t|^2 =: G(\delta X_t, P_t).$$

Together with (4.30), we obtain

$$dG_t = -\gamma G_t dt + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}^\top Q \left\{ \begin{pmatrix} -\delta b_t \cdot \delta X_t \\ 0 \\ -2\delta b_t \cdot P_t + 4\mathrm{rc}_t^2\sigma^2 \end{pmatrix} dt + \begin{pmatrix} 2\frac{z_t}{u_t} \\ 0 \\ 4u_t \end{pmatrix} \sigma \mathrm{rc}_t dB_t \right\}.$$

By defining

$$\overline{\gamma} := \begin{cases} \gamma - \sqrt{\gamma^2 - 4\lambda}, & \text{if } \gamma^2 > 4\lambda \\ \gamma, & \text{if } \gamma^2 < 4\lambda \end{cases}, \qquad \overline{Q} := \begin{cases} \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} Q, & \text{if } \gamma^2 > 4\lambda \\ \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} Q, & \text{if } \gamma^2 < 4\lambda \end{cases},$$

we have

$$dG_t \le -\overline{\gamma} G_t dt + \overline{Q} \left\{ \begin{pmatrix} -\delta b_t \cdot \delta X_t \\ 0 \\ -2\delta b_t \cdot P_t + 4\mathrm{rc}_t^2\sigma^2 \end{pmatrix} dt + \begin{pmatrix} 2\frac{z_t}{u_t} \\ 0 \\ 4u_t \end{pmatrix} \sigma \mathrm{rc}_t dB_t \right\}. \tag{4.32}$$

Finally, notice that in each case the function G is a quadratic form and is coercive, that is,

Lemma 4.13. *There exists $\lambda_G > 0$ such that $G_t \ge \lambda_G(r_t^2 + u_t^2)$.*

Proof In both cases, the functions G can be written in the form: $G_t = \left|\Sigma \begin{pmatrix} \delta X_t \\ P_t \end{pmatrix}\right|^2$, where the matrices Σ are of full rank in both cases. Denote by λ_G the smallest eigenvalue of the matrix $\Sigma^\top\Sigma$. Clearly $\lambda_G > 0$. Then we have $G_t \ge \lambda_G(r_t^2 + u_t^2)$. □

Remark 4.14. *Careful readers have noticed that we did not discuss the case $\lambda = \frac{\gamma^2}{4}$. Indeed, in this case one may extract $\varepsilon > 0$ from λ and define the new $\tilde{\lambda} := \lambda - \varepsilon < \frac{\gamma^2}{4}$. Provided that ε is small enough, it will not cause trouble to the following analysis.*

Remark 4.15. *In case $b = 0$, the contraction result can directly follow from the synchronous coupling, i.e. $\mathrm{rc}_t \equiv 0$. Since $\lambda_G(r_t^2 + u_t^2) \le G_t \le C_G(r_t^2 + u_t^2)$, it follows from (4.32) that*

$$\mathcal{W}_2(m_t, m_t') \le \sqrt{\frac{C_G}{\lambda_G}} e^{-\frac{\overline{\gamma}}{2}t} \mathcal{W}_2(m_0, m_0').$$

On the other hand, it follows from Theorem 6.4 of Pavliotis [39] that the spectral gap of the operator

$$-\mathfrak{L} := -v \cdot \nabla_x + \lambda x \cdot \nabla_v - \gamma(\Delta_v - v \cdot \nabla_v)$$

is also equal to $\frac{\overline{\gamma}}{2}$. It justifies that using the quadratic forms G constructed above, we may capture the optimal contraction rate on the area of interest.

4.4.3 Proof of contraction

Lemma 4.16. *Let $c \in \mathbb{R}$, $\eta, \beta \in (0,\infty)$, and suppose that $h : [0,\infty) \to [0,\infty)$ is continuous, non-decreasing, concave, and C^2 except for finitely many points. Define*

$$\psi_t := (1+\beta G_t)h(\ell_t), \quad \textit{with} \quad \ell_t := r_t + \eta u_t.$$

Then,

$$e^{ct}\psi_t \le \psi_0 + \int_0^t e^{cs} K_s ds + M_t, \quad t \ge 0, \tag{4.33}$$

where M is a continuous martingale, and

$$\begin{aligned} K_t = {} & (1+\beta G_t)h'(\ell_t)\big\{\eta\big|\delta b_t\big| + u_t + \big(\eta\lambda - \gamma\big)r_t\big\} + (1+\beta G_t)2h''(\ell_t)\eta^2\sigma^2 \mathrm{rc}_t^2 \\ & + 4\beta\eta\sigma^2\mathrm{rc}_t^2 h'(\ell_t)\big\|\overline{Q}\big\|(r_t+2u_t) + c\psi_t - \overline{\gamma}\beta G_t h(\ell_t) \\ & + \beta h(\ell_t)\left|\overline{Q}\begin{pmatrix} -\delta b_t \cdot \delta X_t \\ 0 \\ -2\delta b_t \cdot P_t + 4\mathrm{rc}_t^2\sigma^2 \end{pmatrix}\right|. \end{aligned} \tag{4.34}$$

Proof Since by assumption, h is concave and piecewise C^2, we can now apply the Itô-Tanaka formula to $h(\ell_t)$. Denote by h' and h'' the left-sided first derivative and the almost everywhere defined second derivative. The generalized second derivative of h is a signed measure μ_h such that $\mu_h(d\ell) \le h''(\ell)d\ell$. We obtain

$$\begin{aligned} dh(\ell_t) &= h'(\ell_t)(dr_t + \eta du_t) + \frac{1}{2}h''(\ell_t)d\langle u\rangle_t \\ &= h'(\ell_t)\Big\{e_t^x \cdot P_t - \gamma r_t - \eta\delta b_t \cdot e_t^p - \eta\lambda e_t^p \cdot \delta X_t\Big\}dt + 2h''(\ell_t)\eta^2\sigma^2\mathrm{rc}_t^2 dt + 2h'(\ell_t)\eta\sigma\mathrm{rc}_t dB_t \\ &\le h'(\ell_t)\Big\{\eta|\delta b_t| + u_t + \big(\eta\lambda-\gamma\big)r_t\Big\}dt + 2h''(\ell_t)\eta^2\sigma^2\mathrm{rc}_t^2 dt + 2h'(\ell_t)\eta\sigma\mathrm{rc}_t dB_t. \end{aligned}$$

Calculate the quadratic variation

$$d\langle h(\ell), G\rangle_t = 2h'(\ell_t)\eta\sigma\mathrm{rc}_t\overline{Q}\begin{pmatrix} 2\frac{z_t}{u_t} \\ 0 \\ 4u_t \end{pmatrix}\sigma\mathrm{rc}_t dt \le 4\eta\sigma^2\mathrm{rc}_t^2 h'(\ell_t)\big\|\overline{Q}\big\|(r_t+2u_t)dt.$$

Finally, again by Itô's formula, we obtain

$$d\big(e^{ct}\psi_t\big) = e^{ct}\big((1+\beta G_t)dh(\ell_t) + \beta h(\ell_t)dG_t + \beta d\langle h(\ell), G\rangle_t + c\psi_t dt\big) \le e^{ct}\big(K_t dt + d\widetilde{M}_t\big),$$

with

$$d\widetilde{M}_t = (1+\beta G_t)2h'(\ell_t)\eta\sigma\mathrm{rc}_t dB_t + \beta h(\ell_t)\overline{Q}\begin{pmatrix} 2\frac{z_t}{u_t} \\ 0 \\ 4u_t \end{pmatrix}\sigma\mathrm{rc}_t dB_t$$

and the process K defined in (4.34). The assertion follows by taking $M_t = e^{ct}\widetilde{M}_t$. □

In order to make ψ_t a contraction under expectation, it remains to choose the coefficients η, β, h so that $\mathbb{E}[K_t] \le 0$.

Choice of coefficients Recall λ_G in Lemma 4.13. We fix a constant

$$\varepsilon_0 < \frac{\overline{\gamma}\lambda_G}{7\|\overline{Q}\|}. \tag{4.35}$$

Recall that there exists $M > 0$ such that for all $m, m' \in \mathcal{P}(\mathbb{R}^n)$

$$\big|b(m,x) - b(m,x')\big| \le \varepsilon_0 |x - x'| \quad \text{whenever } |x - x'| \ge M. \tag{4.36}$$

Using ε_0, M above, we choose the coefficient $\eta > 0$ such that

$$\eta < \frac{\gamma \wedge \varepsilon_0}{L^x + \lambda + 4\|\overline{Q}\|\sigma^2} \wedge \frac{M\sqrt{\varepsilon_0}}{\sigma}, \quad \text{and define} \quad \theta := \frac{1}{\eta} + 8\|\overline{Q}\|\sigma^2, \tag{4.37}$$

where L^x is the Lipschitz constant of the function b in x. Now we are ready to introduce the function

$$h(\ell) = \int_0^{2M\wedge\ell} \varphi(s) g(s) ds,$$

$$\text{where} \quad \varphi(s) = \exp\left(-\frac{\theta}{\eta^2\sigma^2}\frac{s^2}{4}\right) \quad \text{and} \quad g(s) = 1 - \frac{1}{2}\frac{\int_0^s \frac{\Phi(r)}{\varphi(r)} dr}{\int_0^{2M} \frac{\Phi(r)}{\varphi(r)} dr} \quad \text{with} \quad \Phi(r) = \int_0^r \varphi(x) dx.$$

Remark 4.17. *The function h and its similar variations are repeated used in Eberle [17], Eberle, Guillin and Zimmer [18, 19], Luo and Wang [34] to measure the contraction under the reflection coupling. In particular, the functions φ, g and h have the following properties:*

- *φ is decreasing,*
$$\varphi_{\min} := \min_{0\le s\le 2M} \varphi(s) = \exp\left(-\frac{\theta}{\eta^2\sigma^2} M^2\right).$$
- *g is decreasing, $g(0) = 1$ and $g(s) \ge g(2M) = \frac{1}{2}$ for $r \in [0, 2M]$.*
- *h is non-decreasing, concave, $h(0) = 0$, $h'(0) = 1$, $h'(0) = 1$, $h'(2M) = \varphi(2M)g(2M) = \frac{\varphi_{\min}}{2} > 0$ and h is constant on $[2M, \infty)$*
$$h(\ell) \le \ell, \quad \frac{\Phi(\ell)}{2} \le h(\ell) \le \Phi(\ell), \quad \ell \le 2M,$$
and
$$\theta\ell h'(\ell) + 2\eta^2\sigma^2 h''(\ell) \le -\overline{\kappa}_M h(\ell), \quad \ell \le 2M, \quad \textit{with} \quad \overline{\kappa}_M := \frac{\eta^2\sigma^2}{\int_0^{2M} \frac{\Phi(r)}{\varphi(r)} dr}. \tag{4.38}$$

For the later use we further define a constant $\kappa_M > 0$ such that

$$\kappa_M \le \overline{\kappa}_M \wedge \frac{\varphi_{\min}}{2}\Big(\gamma - \eta\big(L^x + \lambda + 4\|\overline{Q}\|\sigma^2\big)\Big). \tag{4.39}$$

Note that in (4.37) we choose η to be small so that $\eta\big(L^x + \lambda + 4\|\overline{Q}\|\sigma^2\big) < \gamma$. Next introduce the constants

$$C_1 := 4\|\overline{Q}\|L^x M^2\Big(1 + \tfrac{2}{\eta}\Big) + 4\|\overline{Q}\|\sigma^2,$$

and choose the coefficient $\beta > 0$ such that

$$\beta < \frac{\kappa_M}{C_1} \wedge 1. \tag{4.40}$$

Finally we may find a constant C_0 such that $r + u \le C_0\psi$ and thus $\mathcal{W}_1 \le C_0\mathcal{W}_\psi$. For the later use, define

$$C_2 := 2\|\overline{Q}\|M\Big(1 + \tfrac{2}{\eta}\Big)C_0.$$

Lemma 4.18. *With the choice of the coefficients η, β, h above, we have $\mathbb{E}[K_t] \le C\xi$ for*

$$c := \min\left\{\kappa_M - C_1\beta - \Big((1+\beta C_M)\eta + \beta h(2M)C_2\Big)\iota,\ \Big(\overline{\gamma} - \frac{7\|\overline{Q}\|}{\lambda_G}\varepsilon_0\Big)\frac{2\beta\lambda_G M^2}{1+2\beta\lambda_G M^2}\right\}. \quad (4.41)$$

Proof We divide $(r,u) \in \mathbb{R}_+ \times \mathbb{R}_+$ into two regions:
(i). $\ell_t = r_t + \eta u_t \le 2M$: It is due to $r_t + \eta u_t \le 2M$ that

$$G_t \le C_G(r_t^2 + u_t^2) \le 4M^2 C_G\Big(1 + \tfrac{1}{\eta^2}\Big) =: C_M.$$

It is due to the Lipschitz assumption (4.27) and the fact $\mathcal{W}_1 \le C_0\mathcal{W}_\psi$ that

$$\begin{aligned}
\left|\overline{Q}\begin{pmatrix} -\delta b_t \cdot \delta X_t \\ 0 \\ -2\delta b_t \cdot P_t + 4\sigma^2 \mathrm{rc}_t^2 \end{pmatrix}\right| &\le \|\overline{Q}\|\big(|\delta b_t| r_t + 2|\delta b_t| u_t + 4\sigma^2 \mathrm{rc}_t^2\big) \\
&\le \|\overline{Q}\|\Big(\big(C_0\iota\mathcal{W}_\psi(m_t, m_t') + L^x r_t\big)(r_t + 2u_t) + 4\sigma^2\Big) \\
&\le 2\|\overline{Q}\|M\Big(1+\tfrac{2}{\eta}\Big)C_0\iota\mathcal{W}_\psi(m_t,m_t') + 4\|\overline{Q}\|L^x M^2\Big(1+\tfrac{2}{\eta}\Big) + 4\|\overline{Q}\|\sigma^2 \\
&= C_2\iota\mathcal{W}_\psi(m_t,m_t') + C_1.
\end{aligned}$$

Together with (4.34) we obtain

$$\begin{aligned}
K_t &\le (1+\beta G_t)h'(\ell_t)\left\{\eta\iota\mathcal{W}_\psi(m_t,m_t') + \big(\eta(L^x+\lambda) - \gamma\big)r_t + u_t\right\} \\
&\quad + (1+\beta G_t)2h''(\ell_t)\eta^2\sigma^2\mathrm{rc}_t^2 + 4\beta\|\overline{Q}\|\eta\sigma^2\mathrm{rc}_t^2 h'(\ell_t)(r_t+2u_t) \\
&\quad + c\psi_t + \beta h(\ell_t)\big(C_2\iota\mathcal{W}_\psi(m_t,m_t') + C_1\big) \\
&\le \big((1+\beta C_M)\eta + \beta h(2M)C_2\big)\iota\mathcal{W}_\psi(m_t,m_t') + C_1\beta h(\ell_t) + c\psi_t \\
&\quad + (1+\beta G_t)h'(\ell_t)\left\{\eta\left(L^x + \lambda + \frac{4\beta\|\overline{Q}\|\sigma^2\mathrm{rc}_t^2}{1+\beta G_t}\right) - \gamma\right\} r_t + I_t \\
&\text{with} \qquad I_t := (1+\beta G_t)h'(\ell_t)\left(1 + \frac{8\beta\|\overline{Q}\|\eta\sigma^2\mathrm{rc}_t^2}{1+\beta G_t}\right)u_t + (1+\beta G_t)2h''(\ell_t)\eta^2\sigma^2\mathrm{rc}_t^2
\end{aligned}$$

Recall θ defined in (4.37). Since $\beta < 1$, we have

$$\frac{1}{\eta} + \frac{8\beta\|\overline{Q}\|\sigma^2\mathrm{rc}_t^2}{1+\beta G_t} \le \frac{1}{\eta} + 8\|\overline{Q}\|\sigma^2 = \theta.$$

Further recall that h satisfies the inequality (4.38) and the constant κ_M defined in (4.39). Since $h''(\ell) \le 0$, $h'(\ell) \le 1$, $h(\ell) \le \ell$ and $\mathrm{rc}_t = 1$ whenever $u_t \ge \xi$, we obtain

$$\begin{aligned}
I_t &\le (1+\beta G_t)\theta\eta u_t h'(\ell_t) + (1+\beta G_t)2h''(\ell_t)\eta^2\sigma^2\mathrm{rc}_t^2 \\
&\le (1+\beta G_t)\Big(\theta\ell_t h'(\ell_t) + 2\eta^2\sigma^2 h''(\ell_t)\Big)1_{\{u_t\ge\xi\}} + (1+\beta C_M)\theta\eta\xi 1_{\{u_t\le\xi\}} \\
&\le -(1+\beta G_t)\overline{\kappa}_M h(\ell_t)1_{\{u_t\ge\xi\}} + (1+\beta C_M)\theta\eta\xi \\
&\le -(1+\beta G_t)\kappa_M h(\ell_t) + (1+\beta G_t)\kappa_M h(\ell_t)1_{\{u_t\le\xi\}} + (1+\beta C_M)\theta\eta\xi \\
&\le -(1+\beta G_t)\kappa_M h(\ell_t) + (1+\beta G_t)\kappa_M r_t + (1+\beta C_M)(\kappa_M+\theta)\eta\xi.
\end{aligned}$$

Hence,

$$\begin{aligned}
K_t &\le \big((1+\beta C_M)\eta + \beta h(2M)C_2\big)\iota\mathcal{W}_\psi(m_t,m_t') + C_1\beta h(\ell_t) + c\psi_t \\
&\quad - (1+\beta G_t)\kappa_M h(\ell_t) + (1+\beta G_t)\kappa_M r_t + (1+\beta C_M)(\kappa_M+\theta)\eta\xi \\
&\quad + (1+\beta G_t)h'(\ell_t)\left\{\eta\left(L^x+\lambda+\frac{4\beta\|\overline{Q}\|\sigma^2\mathrm{rc}_t^2}{1+\beta G_t}\right)-\gamma\right\}r_t \\
&\le \big((1+\beta C_M)\eta + \beta h(2M)C_2\big)\iota\mathcal{W}_\psi(m_t,m_t') + C_1\beta h(\ell_t) + c\psi_t - \kappa_M\psi_t \\
&\quad + (1+\beta G_t)h'(\ell_t)\left\{\frac{\kappa_M}{h'(\ell_t)}+\eta\left(L^x+\lambda+4\|\overline{Q}\|\sigma^2\right)-\gamma\right\}r_t \\
&\quad + (1+\beta C_M)(\kappa_M+\theta)\eta\xi.
\end{aligned}$$

Due to the choice of η in (4.37) and κ_M in (4.39), the factor of r_t above is non-positive, i.e.

$$\frac{\kappa_M}{h'(\ell_t)}+\eta\left(L^x+\lambda+4\|\overline{Q}\|\sigma^2\right)-\gamma \le \frac{2\kappa_M}{\varphi_{\min}}+\eta\left(L^x+\lambda+4\|\overline{Q}\|\sigma^2\right)-\gamma \le 0.$$

Therefore, we obtain

$$K_t \le \big((1+\beta C_M)\eta+\beta h(2M)C_2\big)\iota\mathcal{W}_\psi(m_t,m_t') + \big(C_1\beta+c-\kappa_M\big)\psi_t + (1+\beta C_M)(\kappa_M+\theta)\eta\xi.$$

Since $\mathcal{W}_\psi(m_t,m_t') \le \mathbb{E}[\psi_t]$ and taking expectation on both sides we obtain that

$$\begin{aligned}
\mathbb{E}[K_t] &\le \Big(\big((1+\beta C_M)\eta+\beta h(2M)C_2\big)\iota + (C_1\beta+c-\kappa_M)\Big)\mathbb{E}[\psi_t] + (1+\beta C_M)(\kappa_M+\theta)\eta\xi \\
&\le (1+\beta C_M)(\kappa_M+\theta)\eta\xi,
\end{aligned}$$

where the last inequality is due to the definition of c in (4.41).

(ii). $\ell_t = r_t + \eta u_t \ge 2M$: In this region, $h(\ell_t)$ is constant, $h'(\ell_t) = h''(\ell_t) = 0$. Therefore,

$$K_t = c\psi_t - \overline{\gamma}\beta G_t h(2M) + \beta h(2M)\left|\overline{Q}\begin{pmatrix}-\delta b_t\cdot\delta X_t\\ 0\\ -2\delta b_t\cdot P_t + 4\sigma^2\mathrm{rc}_t^2\end{pmatrix}\right|. \tag{4.42}$$

Further we can divide this region into two parts:

$$\{(r,u):\ \eta u + r \ge 2M\} \ \subseteq\ \{r\ge M\}\cup\{u\ge \eta^{-1}(r\vee M)\}.$$

Recall that by the choice of η in (4.37) we have $\sigma^2 \le \varepsilon_0 M^2\big(1\vee\frac{1}{\eta}\big)^2$ and $\varepsilon_0 \ge L^x\eta$. Together with (4.36) we obtain

$$|\delta b_t| \le \varepsilon_0 r_t, \quad \mathrm{rc}_t^2\sigma^2 \le \varepsilon_0 r_t^2, \quad \text{on}\ \ \{r\ge M\}$$

as well as

$$|\delta b_t| \le L^x\eta u_t \le \varepsilon_0 u_t, \quad \mathrm{rc}_t^2\sigma^2 \le \varepsilon_0 u_t^2, \quad \text{on}\ \ \{u\ge\eta^{-1}(r\vee M)\}.$$

Combining the two estimates above, we get

$$|\delta b_t| \le \varepsilon_0(r_t\vee u_t).$$

and therefore

$$\left|\overline{Q}\begin{pmatrix}-\delta b_t\cdot\delta X\\ 0\\ -2\delta b_t\cdot P_t + 4\mathrm{rc}_t^2\sigma^2\end{pmatrix}\right| \le 7\|\overline{Q}\|\varepsilon_0(r_t^2+u_t^2) \le \frac{7\|\overline{Q}\|\varepsilon_0}{\lambda_G}G_t, \tag{4.43}$$

where for the last inequality we use the coercivity in Lemma 4.13. Also due to $r_t + \eta u_t \geq 2M$ and $\eta \leq 1$ we have

$$\frac{\beta G_t}{1+\beta G_t} \geq \frac{2\beta\lambda_G M^2}{1+2\beta\lambda_G M^2}.$$

Together with (4.42) and (4.43) we obtain

$$\begin{aligned} K_t &\leq c\psi_t - \overline{\gamma}\beta G_t h(2M) + \beta h(2M)\frac{7\|\overline{Q}\|\varepsilon_0}{\lambda_G} G_t = c\psi_t - \overline{\gamma}\frac{\beta G_t}{1+\beta G_t}\psi_t + \frac{7\|\overline{Q}\|\varepsilon_0}{\lambda_G}\frac{\beta G_t}{1+\beta G_t}\psi_t \\ &= \psi_t\left(c - \left(\overline{\gamma} - \frac{7\|\overline{Q}\|}{\lambda_G}\varepsilon_0\right)\frac{\beta G_t}{1+\beta G_t}\right) \leq \psi_t\left(c - \left(\overline{\gamma} - \frac{7\|\overline{Q}\|}{\lambda_G}\varepsilon_0\right)\frac{2\beta\lambda_G M^2}{1+2\beta\lambda_G M^2}\right) \leq 0, \end{aligned}$$

where the second last inequality is due to the choice of ε_0 in (4.35) and the last one is due to c defined in (4.41). □

Proof of Theorem 2.11. Let Γ be a coupling of two probability measures m_0 and m_0' on $\mathbb{R}^{2n}$ such that $\mathcal{W}_\psi(m_0, m_0') = \int \psi d\Gamma$. We consider the coupling process $\big((X,V),(X',V')\big)$ introduced above with initial law $\big((X_0,V_0),(X_0',V_0')\big) \sim \Gamma$. By taking expectation on both sides of (4.33), we obtain

$$\mathbb{E}[e^{ct}\psi_t] \leq \mathbb{E}[\psi_0] + \int_0^t e^{cs}\mathbb{E}[K_s]ds \leq \mathbb{E}[\psi_0] + Cc^{-1}(e^{ct}-1)\xi.$$

for any $\xi > 0$ and $t \geq 0$. Note that $\mathbb{E}[\psi_0] = \int \psi d\Gamma = \mathcal{W}_\psi(m_0, m_0')$. Therefore

$$\mathcal{W}_\psi(m_t, m_t') \leq \mathbb{E}[\psi_t] \leq e^{-ct}\mathcal{W}_\psi(m_0, m_0') + Cc^{-1}\big(1-e^{-ct}\big)\xi \to e^{-ct}\mathcal{W}_\psi(m_0, m_0'), \quad \text{as} \quad \xi \to 0.$$

Finally note that by the choice of β in (4.40), we have $c > 0$ according to (4.41) provided that ι is small enough. □

References

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer, 2008.

[2] S. Armstrong and J.-C. Mourrat, *Variational methods for the kinetic fokker-planck equation*, Preprint arXiv:1902.04037, (2019).

[3] D. Bakry, P. Cattiaux, and A. Guillin, *Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré.*, J. Funct. Anal., 254 (2008), pp. 727–759.

[4] V. Bally, *On the connection between the Malliavin covariance matrix and Hörmander's condition*, Journal of functional analysis, 96 (1991), pp. 219–255.

[5] F. Bolley, A. Guillin, and F. Malrieu, *Trend to equilibrium and particle approximation for a weakly selfconsistent Vlasov-Fokker-Planck equation*, M2AN Math. Model. Numer. Anal., 44 (2010), pp. 867–884.

[6] A. Brünger, C. Brooks III, and M. Karplus, *Stochastic boundary conditions for molecular dynamics simulations of ST2 water*, Chem. Phys. Lett., 105 (1984), pp. 495–500.

[7] Y. Cao, J. Lu, and L. Wang, *On explicit l2-convergence rate estimate for underdamped langevin dynamics*, Preprint arXiv:1908.04746, (2019).

[8] R. Carmona and F. Delarue, *Probabilistic Theory of Mean Field Games with Applications II*, Springer, 2018.

[9] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, *Underdamped Langevin MCMC: A non-asymptotic analysis*, Proceedings of Machine Learning research, 75 (2018), pp. 1–24.

[10] L. Chizat and F. Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, in Advances in neural information processing systems, 2018, pp. 3040–3050.

[11] G. Conforti, A. Kazeykina, and Z. Ren, *Game on Random Environement, Mean-field Langevin System and Neural networks*, Preprint arXiv:2004.02457, (2020).

[12] A. S. Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, Journal of the Royal Statistical Society: Series B, 79 (2017), pp. 651–676.

[13] J. Dolbeault, C. Mouhot, and C. Schmeiser, *Hypocoercivity for kinetic equations with linear relaxation terms*, Comptes Rendus Mathematique, 347 (2009), pp. 511–516.

[14] J. Dolbeault, C. Mouhot, and C. Schmeiser, *Hypocoercivity for linear kinetic equations conserving mass*, Transactions of the American Mathematical Society, 367 (2015), pp. 3807–3828.

[15] C. Domingo-Enrich, S. Jelassi, A. Mensch, G. M. Rotskoff, and J. Bruna, *A mean-field analysis of two-player zero-sum games*, Preprint arXiv:2002.06277, (2020).

[16] A. Durmus and E. Moulines, *Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm*, Preprint arXiv:1605.01559, (2016).

[17] A. Eberle, *Reflection couplings and contraction rates for diffusions*, Probability Theory and Related Fields, 166 (2016), pp. 851–886.

[18] A. Eberle, A. Guillin, and R. Zimmer, *Couplings and quantitative contraction rates for Langevin dynamics*, Ann. Probab., 47 (2019), pp. 1982–2010.

[19] A. Eberle, A. Guillin, and R. Zimmer, *Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes*, Transactions of the American Mathematical Society, 371 (2019), pp. 7135–7173.

[20] A. Einstein, *Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen*, Annalen der Physik, 322 (1905), pp. 549–560.

[21] H. Föllmer, *Time reversal on Wiener space*, in Stochastic Processes - Mathematics and Physics, S. A. Albeverio, P. Blanchard, and L. Streit, eds., Springer, 1986, pp. 119–129.

[22] A. Gelman, G. Roberts, and W. Gilks, *Efficient Metropolis Jumping Rules*, Bayesian Statistics, (1996), pp. 599–607.

[23] A. Guillin, W. Liu, L. Wu, and C. Zhang, *The kinetic Fokker-Planck equation with mean field interaction*, Preprint arXiv:1912.02594, (2019).

[24] U. G. Haussmann and E. Pardoux, *Time reversal of diffusions*, The Annals of Probability, (1986), pp. 1188–1205.

[25] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer, 1981.

[26] K. Hornik, *Approximation capabilities of multilayer feedforward networks*, Neural Networks, 4 (1991), pp. 251–257.

[27] K. Hu, A. Kazeykina, and Z. Ren, *Mean-field Langevin System, Optimal Control and Deep Neural Networks*, Preprint arXiv:1909.07278, (2019).

[28] K. Hu, Z. Ren, D. Siska, and L. Szpruch, *Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks*, Preprint arXiv:1905.07769, (2019).

[29] J.-F. Jabir, D. Siska, and L. Szpruch, *Mean-Field Neural ODEs via Relaxed Optimal Control*, Preprint arXiv:1912.05475, (2019).

[30] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker–Planck equation*, SIAM J. Math. Anal., 29 (1998), pp. 1–17.

[31] P. Langevin, *Sur la théorie du mouvement brownien*, CR Acad. Sci. Paris, 146 (1908), pp. 530–533.

[32] T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations*, Imperial College Press, 2010, https://doi.org/10.1142/p579.

[33] Y. Lu, C. Ma, Y. Lu, J. Lu, and L. Ying, *A Mean-field Analysis of Deep ResNet and Beyond: Towards Provable Optimization Via Overparameterization From Depth*, Preprint arXiv:2003.05508, (2020).

[34] D. Luo and J. Wang, *Exponential convergence in L^p-Wasserstein distance for diffusion processes without uniformly dissipative drift*, Mathematische Nachrichten, 289 (2016), pp. 1909–1926.

[35] 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 (2018), pp. E7665–E7671.

[36] A. Millet, D. Nualart, and M. Sanz, *Integration by parts and time reversal for diffusion processes*, The Annals of Probability, (1989), pp. 208–238.

[37] R. M. Neal, *MCMC using Hamiltonian dynamics*, Handbook of Markov chain Monte Carlo, Boca Raton: CRC Press, 2011.

[38] E. Nelson, *Dynamical theories of Brownian motion*, vol. 2, Princeton University Press, 1967.

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

[40] L. Rey-Bellet and L. E. Thomas, *Exponential convergence to non-equilibrium stationary states in classical statistical mechanics*, Comm. Math. Phys., 225 (2002), pp. 305–329.

[41] G. M. Rotskoff and E. Vanden-Eijnden, *Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error*, arXiv:1805.00915, (2018).

[42] D. Siska and L. Szpruch, *Gradient Flows for Regularized Stochastic Control Problems*, Preprint arXiv:2006.05956, (2020).

[43] A.-S. Sznitman, *Topics in propagation of chaos*, École d'Été de Probabilités de Saint-Flour XIX, Lecture Notes in Math., 1464 (1991), pp. 165–251.

[44] D. Talay, *Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme*, Markov Process. Related Fields, 8 (2002), pp. 163–198.

[45] C. Villani, *Hypocoercive diffusion operators*, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8), 10 (2007), pp. 257–275.

[46] C. Villani, *Hypocoercivity*, Mem. Amer. Math. Soc., 202 (2009), pp. iv–141.

[47] L. Wu, *Large and moderate deviations and exponential convergence for stochastic damping Hamiltonian systems*, Stochastic Processes and their Applications, 91 (2001), pp. 205–238.