Mean Field Langevin Dynamics, Mean Field Neural Networks, and Mean Field Ising Models

by

Chandan Tankala

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Mathematics

Dissertation Committee:

David A. Levin, Chair

Krishnakumar Balasubramanian, Core Member

Benjamin Young, Core Member

Chris Sinclair, Core Member

Peter Ralph, Institutional Representative

University of Oregon

Summer 2025

DISSERTATION ABSTRACT

Chandan Tankala

Doctor of Philosophy in Mathematics

Title: Mean Field Langevin Dynamics, Mean Field Neural Networks, and Mean Field Ising Models

We study novel theoretical and algorithmic frameworks for sampling from complex probability distributions. We present three interconnected contributions that advance our understanding of Markov chain mixing, mean field optimization, and neural network training.

First, we introduce a virtual particle stochastic approximation algorithm for mean field Langevin dynamics. The key innovation is a two-particle system: real particles that form the output and virtual particles used for unbiased gradient estimation. This design achieves quadratic computational savings compared to standard particle methods while avoiding the technical machinery of propagation of chaos. We prove exponential convergence under standard regularity conditions and demonstrate the method's effectiveness on pairwise interaction energies common in physics and machine learning.

Second, we extend our framework to mean field neural networks, providing a computationally efficient algorithm for entropy-regularized training of two-layer networks. By leveraging the favorable geometry of proximal Gibbs distributions, we establish quantitative convergence guarantees without requiring the uniform-in-dimension bounds typical in prior work. This bridges the gap between mean field theory and practical neural network optimization.

Third, we analyze the mean-field tensor Ising model, which generalizes the classical Ising model to capture higher-order interactions beyond pairwise dependencies. Using discrete Ricci curvature theory — a departure from traditional coupling-based methods — we establish the first polynomial mixing time bounds for a Markov chain based on locally balanced proposals. Our approach reveals how geometric tools from optimal transport can provide new insights into sampling from high-dimensional spin systems.

This dissertation includes previously published co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Chandan Tankala

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
Texas A&M University, College Station, TX, USA
National Institute of Technology, Kurukshetra, HR, INDIA

DEGREES AWARDED:

Doctor of Philosophy, Mathematics, 2025, University of Oregon
Master of Science, Geophysics, 2012, Texas A&M University
Bachelor of Engineering, Engineering, 2009, National Institute of Technology

AREAS OF SPECIAL INTEREST:

Probability Theory
Machine Learning

PROFESSIONAL EXPERIENCE:

Geophysicist, British Petroleum, 2012-2015
Geophysicist intern, British Petroleum, 2011

PUBLICATIONS:

Chandan Tankala, Dheeraj M. Nagaraj, & Anant Raj (2025). Beyond propagation of chaos: A stochastic approximation for mean field optimization. *arXiv:2503.13115*

ACKNOWLEDGEMENTS

I am truly grateful to my advisor Professor Krishnakumar Balasubramanian for teaching me the art of formulating mathematical insight into complex problems. I am indebted to him for advising me on a wide range of topics in probability theory and statistics, and for honing my skills in specific topics. His constant support, encouragement, and strength will stay with me always. I thank Dr. David A. Levin for reading my dissertation and providing feedback on it. I have also been inspired by my committee members Professors Ben Young and Christopher Sinclair for their ability to write single-author publications, while advising multiple students and maintaining a vibrant teaching style.

During the last two years of my Ph.D., I have been fortunate to work with several exceptional people, including Krishna, and I am thankful for their friendship. I am especially grateful for my research collaboration with Dr. Dheeraj Nagaraj. His mathematical insight and courage to undertake hard problems intentionally are admirable.

I am also thankful to Professor Anant Raj for making the work environment fun, and to Professor Quan Zhou and Dr. Andrea Ottolini for various discussions about Markov chains. I have been struck by Quan's ability to connect seemingly disparate topics. Though I met Andrea towards the end of my Ph.D., I am grateful for his unwavering support, friendship, and remarkable enthusiasm when discussing complex mathematical problems. I also thank Professor Sourav Chatterjee for his support whenever I reached out.

Finally, I am deeply grateful to my parents for loving me, listening to my thoughts, investing in my education, and having my back during every single stressful situation.

I cannot thank Amma enough for supporting me in her own way of making my chatter sound joyful.

To Amma and Nanna

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Chapters II and III are based on published and co-authored material with Dr. Dheeraj M. Nagaraj, Google DeepMind and Dr. Anant Raj, Indian Institute of Science. Chapter IV is based on work which has not been published, for which I was advised by Dr. Krishnakumar Balasubramanian at University of California, Davis and Dr. Quan Zhou at Texas A&M University, College Station.

This dissertation studies sampling from discrete and continuous mean-field probability distributions, with applications to neural networks. For discrete distributions, we analyze the mixing time of Markov chains, while for continuous distributions we establish convergence rates for space and time discretized Langevin dynamics. Each chapter is self-contained with its own motivation, literature review, and notation. A unifying theme throughout is the application of functional inequalities to obtain quantitative convergence bounds for sampling algorithms.

1.1 Mean field Langevin dynamics

Mean field Langevin dynamics arise naturally when optimizing functionals over probability distributions, with applications spanning machine learning and Bayesian inference. This chapter develops a novel computational approach for sampling from the stationary distribution of such dynamics, avoiding traditional propagation of chaos arguments.

We consider functionals of the form $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defined on the space of probability measures with finite second moments:

$$\mathcal{E}(\mu) = \mathcal{F}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu)\,,$$

where $\mathcal{F}$ is an energy functional, $\mathcal{H}(\mu) = \int \mu(x)\log\mu(x)dx$ is the negative entropy, and $\sigma > 0$ controls regularization strength. The stationary measure π satisfies the fixed-point

equation

$$\pi(x) \propto \exp\left(-\frac{2}{\sigma^2}\frac{\delta \mathcal{F}}{\delta \mu}(\pi)(x)\right),$$

where $\frac{\delta \mathcal{F}}{\delta \mu}$ denotes the first variation.

The key innovation is our virtual particle stochastic approximation algorithm (Algorithm 1), defined in Chapter II, which maintains two types of particles:

- Real particles $\{X_k^{(i)}\}_{i=1}^n$ that form the output
- Virtual particles $\{Y_k^{(j)}\}_{j=k}^T$ used for gradient estimation

The algorithm uses an unbiased estimator $\hat{G}(x, Y, \xi)$ of the Wasserstein gradient $\nabla_{\mathcal{W}} \mathcal{F}(x, \mu)$, achieving computational complexity $O(nT + T^2)$ compared to $O(n^2 T)$ for standard particle methods.

First, we analyze a modified functional $\bar{\mathcal{E}}$. For a functional $\bar{\mathcal{F}}$ (possibly different from $\mathcal{F}$) with the same minimizer π, we define

$$\bar{\mathcal{E}}(\mu) := \bar{\mathcal{F}}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu) - \bar{\mathcal{F}}(\pi) - \frac{\sigma^2}{2}\mathcal{H}(\pi)\,.$$

Let $\mathcal{R}_T$ be the sigma algebra generated by $Y_0^{(0)}, Y_1^{(1)}, \ldots, Y_T^{(T)}$. Let $\mu_T | \mathcal{R}_{T-1}$ be the law of $X_T^{(1)}$ conditional on $\mathcal{R}_{T-1}$. This is a random probability measure measurable with respect to $\mathcal{R}_{T-1}$. The main convergence result is:

- **Theorem 1 (Chapter II)**: Under appropriate assumptions, Algorithm 1 produces n i.i.d. samples from $\mu_T | \mathcal{R}_{T-1}$ satisfying

  $$\mathbb{E}[\bar{\mathcal{E}}(\mu_T | \mathcal{R}_{T-1})] \leq e^{-\frac{\eta C_{\bar{\mathcal{E}}} T}{8}} \bar{\mathcal{E}}(\mu_0) + C\left[\frac{\gamma_3 \eta^2}{C_{\bar{\mathcal{E}}}} + \frac{\gamma_2 \eta}{C_{\bar{\mathcal{E}}}} + \frac{\gamma_1 \sqrt{\eta}}{C_{\bar{\mathcal{E}}}}\right],$$

  where $C_{\bar{\mathcal{E}}}$ is the Polyak-Łojasiewicz constant and $\gamma_1, \gamma_2, \gamma_3$ depend on problem parameters.

  For the important case of pairwise interaction energy

  $$\mathcal{E}(\mu) = \int V(x) d\mu(x) + \frac{1}{2}\int\int W(x-y) d\mu(x) d\mu(y) + \frac{\sigma^2}{2}\mathcal{H}(\mu)\,,$$

we establish explicit convergence rates under smoothness and log-Sobolev inequality assumptions in Theorem 2 of Chapter II. The analysis employs a novel descent lemma decomposing the error into discretization, stochastic, and linearization components, providing a systematic framework for analyzing stochastic approximations of Wasserstein gradient flows.

1.2 Mean field neural networks

Mean field analysis provides a powerful framework for understanding optimization dynamics in wide neural networks. This chapter extends the virtual particle methodology developed in the mean field Langevin dynamics chapter to the specific setting of two-layer neural networks, establishing convergence guarantees for entropy-regularized empirical risk minimization.

We consider a two-layer mean field neural network $f(\mu; z) = \int h(x, z)d\mu(x)$ with activation function $h : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ and parameter distribution $\mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$. Given empirical data $(z_i, w_i)_{i=1}^m$, we minimize the regularized square loss functional:

$$\mathcal{E}(\mu) = \frac{1}{m}\sum_{i=1}^{m}\left(\int h(z_i, x)d\mu(x) - w_i\right)^2 + \frac{\lambda}{2}\int \|x\|^2 d\mu(x) + \frac{\sigma^2}{2}\mathcal{H}(\mu),$$

where $\lambda > 0$ controls weight regularization and $\sigma > 0$ controls entropic regularization.

The Wasserstein gradient of this functional takes the explicit form

$$\nabla_{\mathcal{W}}\mathcal{F}(x; \mu) = \frac{2}{m}\sum_{i=1}^{m}\left(\int h(z_i, y)d\mu(y) - w_i\right)\nabla_x h(z_i, x) + \lambda x,$$

and the unique minimizer π satisfies the fixed-point equation

$$\pi(x) \propto \exp(-\frac{2}{\sigma^2}\delta\mathcal{F}(x, \pi))\,.$$

We employ the stochastic gradient estimator

$$\hat{G}(x, Y, I) = -(h(z_I, Y) - w_I)\nabla_x h(z_I, x) - \lambda x,$$

where I is sampled uniformly from $[m]$. Let $\mathcal{R}_T$ be the sigma algebra generated by $Y_0^{(0)}, Y_1^{(1)}, \ldots, Y_T^{(T)}, I_1, \ldots, I_T$ and $\mu_T|\mathcal{R}_{T-1}$ the law of $X_T^{(1)}$ conditional on $\mathcal{R}_{T-1}$.

The main result establishes the following rate of convergence of the functional $\mathcal{E}$ in expectation along a discrete trajectory of random probability measures $\mu_T|\mathcal{R}_{T-1}$.

– **Theorem 3 (Chapter III)**: Under appropriate assumptions, Algorithm 1 (defined in Chapter II) achieves

$$\mathbb{E}[\mathcal{E}(\mu_T|\mathcal{R}_{T-1})] - \mathcal{E}(\pi) \leq e^{-\frac{T\eta\sigma^2}{8C_{\text{LSI}}}}(\mathcal{E}(\mu_0) - \mathcal{E}(\pi)) + O(\eta + \sqrt{\eta}),$$

where σ is the regularization parameter of the entropy functional, $\eta > 0$ is the learning rate, C_{LSI} is the logarithmic-Sobolev inequality (LSI) constant of the probability measure π described above, and where the $O(\cdot)$ term depends on problem parameters.

This provides the first particle-efficient algorithm for mean field neural networks with quantitative convergence guarantees. The analysis leverages the general framework from mean field Langevin dynamics chapter while carefully verifying the required assumptions for the neural network setting, demonstrating that the proximal Gibbs distribution's favorable geometric properties.

1.3 Mean field Ising model

The mean-field tensor Ising model represents a natural generalization of the classical Ising model, designed to capture higher-order interactions beyond pairwise dependencies. This chapter investigates the mixing properties of Markov chains for sampling from this model using tools from discrete Ricci curvature theory.

We consider the mean-field p-tensor Ising model on the discrete space $\mathcal{X} = \{-1, +1\}^n$, defined by the probability distribution

$$\pi_{\beta,p}(x) := \frac{1}{2^n Z_n(\beta, p)} \exp\left(\frac{\beta}{n^{p-1}} H(x)\right),$$

where $p \geq 2$ is the tensor order, $\beta > 0$ is the inverse temperature parameter, and the Hamiltonian $H(x) = \sum_{1\leq i_1,\ldots,i_p\leq n} x_{i_1} \cdots x_{i_p}$ captures all p-way interactions.

The main contribution of this chapter is the analysis of a continuous-time Markov chain Q_S based on the square-root locally balanced proposal of Zanella (2020), defined by

$$Q_S(x,y) = \begin{cases} \sqrt{\frac{\pi_{\beta,p}(y)}{\pi_{\beta,p}(x)}} \text{ if } \|x-y\|_{\ell^1} = 2 \\ 0 \text{ otherwise} \end{cases} \tag{1.1}$$

Our proof approach employs the entropic Ricci curvature framework of Erbar and Maas (2012).

The central results are:

- **Theorem 6 (Chapter IV)**: For any fixed $p \geq 2$, if $\exp(2\beta(p+1))4\beta p(p-1) < 1$, the Markov chain Q_S, defined in (1.1), has Ricci curvature bounded below by $\text{Ric}(\mathcal{X}, Q_S, \pi_{\beta,p}) \geq 2(1-\exp(2\beta(p+1))4\beta p(p-1)) \exp\left(\frac{\beta}{2n^{p-1}}((n-2)^p - n^p)\right)$.
- **Theorem 7 (Chapter IV)**: For any fixed $p \geq 2$, if $\exp(2\beta(p+1))4\beta p(p-1) < 1$, then the mixing time, defined in Chapter IV, of the continuous time Markov chain Q_S, defined in (1.1), satisfies the following upper bound:

$$t_{\mathsf{mix}}(\varepsilon) \leq \frac{\exp(\beta p)}{2(1-\exp(2\beta p + 2\beta)4\beta p(p-1))} \log\left(\frac{2\exp(\beta p)n}{\sqrt{2}\varepsilon}\right).$$

The proof leverages Erbar, Henderson, Menz, and Tetali (2017) perturbative criterion for positive Ricci curvature requiring careful combinatorial analysis. This functional inequality approach provides polynomial mixing time bounds in the high-temperature regime, offering a new perspective on sampling from tensor-valued spin systems.

The next chapter develops the theory of optimal transport and functional inequalities to obtain quantitative bounds for an algorithm to sample from a continuous probability distribution.

CHAPTER II

MEAN FIELD LANGEVIN DYNAMICS

This work is from Tankala, Nagaraj, and Raj (2025), which has been accepted for publication at the Conference on Learning Theory (COLT), 2025. The conference has not published a camera-ready version of the paper yet. This topic and the main algorithm were suggested to me by Dr. Dheeraj M. Nagaraj, who, along with Dr. Anant Raj, advised me on which books and papers to read for optimal transport, and either directed my proofs or checked them. I was the main contributor to all the theorems, proving them, and writing the paper. Dr. Dheeraj M. Nagaraj and Dr. Anant Raj also contributed to writing the introduction section of the paper.

The structure of this chapter is the following.

1. The first section provides an overview of gradient flows in the space of probability measures and a literature review of mean field Langevin dynamics.

2. The second section presents previous work in sampling from mean field Langevin dynamics and the associated challenges.

3. The third section contains notation which will be used for this chapter.

4. The fourth section introduces our algorithm to sample from the stationary distribution of mean field Langevin dynamics.

5. The fifth section proves the rate of convergence of our algorithm.

2.1 Introduction

There is a strong connection between sampling and optimization problems because the problem of sampling can be cast an optimization problem in the space of probability distributions. Toward that end, optimizing a functional $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ over

the space of all probability distributions over $\mathbb{R}^d$ with finite second moments ($\mathcal{P}_2(\mathbb{R}^d)$) has gained immense interest in the recent years with applications in machine learning and Bayesian inference. One of the commonly used functionals is the Kullback-Leibler (KL) divergence to the target distribution as described in Durmus, Majewski, and Miasojedow (2019); Vempala and Wibisono (2019). Variational inference extends this notion of optimization to constrained optimization over the space of distributions to fit the given data as shown in Lambert, Chewi, Bach, Bonnabel, and Rigollet (2022); Yan, Wang, and Rigollet (2024); Yao and Yang (2022).

Given an energy functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, and regularization strength $\sigma > 0$, we consider functionals of the form $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defined as:

$$\mathcal{E}(\mu) = \mathcal{F}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu)\,, \tag{2.1}$$

where $\mathcal{H}(\mu)$ is the negative entropy defined as follows:

$$\mathcal{H}(\mu) = \begin{cases} \int \mu(x) \log \mu(x) dx & \text{if } \mu \ll \mathsf{Leb} \text{ and } d\mu(x) = \mu(x)dx \\ \infty & \text{otherwise}\,. \end{cases}$$

A common approach to optimization over $\mathcal{P}_2(\mathbb{R}^d)$ is *gradient flow* with respect to the Wasserstein metric. Given a functional $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, the Wasserstein gradient flow $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ is the solution to the differential equation

$$\frac{d}{dt}\mu_t = -\nabla_{\mathcal{W}}\mathcal{E}(\mu_t)\,,$$

where $\nabla_{\mathcal{W}}\mathcal{E}(\mu)$ is the Wasserstein gradient of the functional $\mathcal{E}$ at probability measure μ.

The well-known Langevin dynamics was shown to be the gradient flow of the Kullback-Leibler (KL) divergence to the target distribution in the seminal work of Jordan, Kinderlehrer, and Otto (1998). This framework can be extended to a broader class of functionals, including interaction energy and entropy as demonstrated in Ambrosio, Gigli, and Savaré (2008); McCann (1997).

We consider functionals $\mathcal{F}(\mu)$ of the form such that $\nabla_{\mathcal{W}}\mathcal{F}$ depends on μ. The Wasserstein gradient flow of the functional in (2.1) corresponds to the McKean-Vlasov SDE

$$dX_t = -\nabla_{\mathcal{W}}\mathcal{F}(X_t, \mu_t) + \sigma dB_t\,, \tag{2.2}$$

where $\mu_t = \mathsf{Law}(X_t)$, $(B_t)_{t\geq 0}$ is the standard Brownian motion on $\mathbb{R}^d$, if the above SDE converges. By Ren and Wang (2021)[Theorem 2.3], the McKean-Vlasov SDE converges to a unique stationary measure π, in 2-Wasserstein and relative entropy distances between μ_t and π, if the following conditions hold:

- Continuity: $\nabla_{\mathcal{W}}\mathcal{F}(x, \mu)$ is continuous in (x, μ)
- Lipschitz condition: There exists a constant $K > 0$ such that

$$\langle \nabla_{\mathcal{W}}\mathcal{F}(y, \nu) - \nabla_{\mathcal{W}}\mathcal{F}(x, \mu), x - y\rangle \leq K\left[|x-y|^2 + W_2(\mu,\nu)^2\right]$$

- Growth condition:

$$|\nabla_{\mathcal{W}}\mathcal{F}(0,\mu)| \leq c\left(1 + \sqrt{\int |x|^2 d\mu}\right) \text{ for some constant } c > 0$$

- For some constants $K_2 > K_1 \geq 0$:

$$\langle \nabla_{\mathcal{W}}\mathcal{F}(y, \nu) - \nabla_{\mathcal{W}}\mathcal{F}(x, \mu), x - y\rangle \leq K_1 W_2(\mu,\nu)^2 - K_2|x-y|^2$$

-

$$\text{Hess}_x[\nabla_{\mathcal{W}}\mathcal{F}(\mu)(x)] \geq K I_d\ ,\ K \in \mathbb{R}\,,$$

where Hess_x is the Hessian with respect to the variable x.

If the McKean-Vlasov has a unique stationary measure π, then it is given by the solution to the fixed point equation:

$$\pi(x) \propto \exp\left(-\frac{2}{\sigma^2}\delta\mathcal{F}(\pi)(x)\right)\,,$$

where $\delta\mathcal{F}$ is the first-variation of the functional $\mathcal{F}$. Note that π does not admit a closed form solution.

2.2 Challenges and our approach

While Langevin dynamics can be implemented algorithmically through time discretization of Itô stochastic differential equations, for the McKean-Vlasov type SDEs where the drift function depends on the distribution of the variable, an algorithmic implementation is not that straightforward because this distribution is not known.

The popular computational approximation in this context is the particle approximation where n instances (or particles) of the SDE are implemented computationally, with the distribution of the variables in the McKean-Vlasov SDE, as described by the equation (2.2), replaced by their empirical distribution. Theoretically, one way to show that this approximation optimizes the objective functional is to show that

1. the particle approximation converges rapidly to its nd dimensional stationary distribution and

2. a sample from the n particle stationary distribution gives a representative sample from the stationary distribution of mean field Langevin dynamics. The latter problem is referred to as *propagation of chaos*.

The propagation of chaos problem for McKean-Vlasov SDEs was originally studied by Sznitman (1991), which established convergence rates in the Wasserstein metric via coupling arguments. These bounds were first made uniform in time by Malrieu (2001, 2003) in the quadratic Wasserstein and relative entropy metrics. For the functional defined in equation (2.1)), these works obtain a bound on the error of order $O(k/n)$ in the squared quadratic Wasserstein distance and assume strong convexity for the external potential V and convexity for the interaction potential W. A uniform in time propagation

of chaos was recently shown by Chen, Ren, and Wang (2022) by assuming convexity of the mean-field functional, as opposed to imposing convexity conditions on the interaction potential. The error in the squared quadratic Wasserstein error bound was improved to $O((k/n)^2)$ in Lacker and Le Flem (2023) by assuming a uniform-in-n log-Sobolev inequality for the stationary distribution of the n-particle system and using the so-called recursive BBGKY proof technique. Kook, Zhang, Chewi, Erdogdu, and Li (2024) build on this proof technique and also obtain an error bound of order $O((k/n)^2)$ in the squared quadratic Wasserstein distance and KL-divergence under a slightly different assumption on a ratio involving log-Sobolev inequality, smoothness and diffusion constants, which is referred to as the "weak interaction" condition. Recently, Bou-Rabee and Schuh (2023) presented a non-linear Hamiltonian Monte Carlo algorithm and prove its rate of convergence in L^1-Wasserstein distance, without using the propagation of chaos arguments.

To show the convergence of n particle approximation, prior works such as Chen et al. (2022); Chewi, Nitanda, and Zhang (2024); Wang (2024) consider the finite-particle stationary distribution and establish Logarithmic Sobolev inequalities (LSI) for the nd dimensional system with a constant independent of n. This can be used establish computational complexity of standard sampling algorithms such as Langevin Monte Carlo (LMC), Underdamped Langevin Monte Carlo (ULMC), and Metropolis Adjusted Langevin Algorithm (MALA). However, this can be technically involved and obtaining guarantees for the LSI constant $C_{\mathsf{LSI},n}$ for the n-particle stationary distribution can be very hard and $C_{\mathsf{LSI},n}$ could be much worse than the LSI constant C_{LSI} for the mean field stationary distribution π. For instance, in Wang (2024)[Theorem 1], $C_{\mathsf{LSI},n} = O(dC_{\mathsf{LSI}}^3)$ (along with additional assumptions), and in Chewi et al. (2024)[Theorem 2], $C_{\mathsf{LSI},n} = O(e^d)$, while in Chewi et al. (2024)[Equation 2.2], C_{LSI} does not depend on d.

In our work, we use a particle approximation but not a propagation of chaos argument which makes our contribution novel. In Tankala et al. (2025), we consider a stochastic approximation of the mean field Langevin dynamics in (2.2), which can be implemented exactly, using an unbiased estimator for the Wasserstein gradient. The virtual particle stochastic approximation method, first proposed by Das and Nagaraj (2023) for Stein Variational Gradient Descent (SVGD), is the theoretical foundation of our analysis. Our approach, however, presents novel analytical challenges.

The main proof technique in this chapter is the usage of a functional inequality called the Log-Sobolev Inequality. Extending the virtual particle stochastic approximation framework presented in this work from Log-Sobolev Inequality (LSI) to Poincaré inequality presents fundamental theoretical challenges that would require substantial modifications to the analysis. With Poincaré inequality, the weaker concentration properties would necessitate stronger moment assumptions on the stochastic gradient estimator $\hat{G}(x, Y, \xi)$ (defined in the Algorithm subsection), potentially requiring bounds on $\mathbb{E}[\|\hat{G}(x, Y, \xi)\|^{2+\delta}]$ for some $\delta > 0$. For the pairwise interaction energy example, relaxing to Poincaré would invalidate the weak interaction condition (Assumption 6), while for mean field neural networks in Chapter III, the proximal Gibbs distribution analysis fundamentally requires the exponential concentration provided by LSI.

2.3 Notation

First, we introduce some notation. For any measure ρ over $\mathbb{R}^d$ and functions $f, g : \mathbb{R}^d \to \mathbb{R}^d$. Let

$$\langle f, g \rangle_{L_2(\rho)} := \int \rho(dx) \langle f(x), g(x) \rangle \, , \; L_2^2(\rho; f)^2 := \int \rho(dx) \|f(x)\|^2 \, ,$$

if f, g are square integrable with respect to ρ. For a vector field $f : \mathbb{R}^d \to \mathbb{R}^d$, its divergence is given by $\nabla \cdot f = \sum_{i=1}^d \frac{\partial f}{\partial x_i}$, and for a function $f : \mathbb{R} \to \mathbb{R}$, the Laplacian is defined as $\Delta f := \sum_{i=1}^d \frac{\partial^2 f}{\partial x_i^2}$. Let $\mathcal{P}_2(\mathbb{R}^d), \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ denote the space of probability

measures on $\mathbb{R}^d$ with finite second moment, and those that are absolutely continuous with respect to the Lebesgue measure. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we let $\mathsf{Var}(\mu)$ denote the trace of its covariance. The Wasserstein distance $\mathcal{W}_2(\mu, \nu)$ between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as:

$$\mathcal{W}_2^2(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|X - Y\|^2 d\gamma(x, y)\,,$$

where $\Gamma(\mu, \nu)$ is the set of all joint distributions over $\mathbb{R}^d \times \mathbb{R}^d$ such that the marginal distribution of X is μ and of Y is ν. The Fisher Divergence of a probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with respect to $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as: $\mathsf{FD}(\mu||\nu) := \int_{\mathbb{R}^d} \mu(x) \left\| \nabla \log \frac{\mu(x)}{\nu(x)} \right\|^2 dx$. The first variation of a functional $\mathcal{F}$ at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is denoted by $\delta_\mu \mathcal{F}(\mu)(x)$ or just $\delta\mathcal{F}(x, \mu)$, where $x \in \mathbb{R}^d$, and is defined as the quantity which satisfies the equality:

$$\left.\frac{d\mathcal{F}(\mu + \varepsilon(\mu' - \mu))}{d\varepsilon}\right|_{\varepsilon=0} = \int \delta_\mu \mathcal{F}(\mu)(x)(\mu' - \mu)(x)dx\,.$$

If the probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ have densities p, q respectively, then the total-variation distance between them is defined as: $\|\mu - \nu\|_{\mathsf{TV}} := \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)|dx$.

2.4 Algorithm

As noted in the second section, the drift in (2.2) being dependent on μ_t makes it hard to approximate the dynamics algorithmically. Therefore, generally the following particle approximation is used. Let $X_0^{(1)}, \ldots, X_0^{(n)} \overset{\text{i.i.d.}}{\sim} \mu_0$, $\hat{\mu}_k$ be the empirical distribution of $(X_k^{(i)})_{i\in[n]}$. Let η be the timestep size and $(Z_k^{(i)})_{k,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$. This interacting particle based approximation is then given by:

$$X_{k+1}^{(i)} = X_k^{(i)} - \eta \nabla_{\mathcal{W}} \mathcal{F}(X_k^{(i)}, \hat{\mu}_k) + \sigma\sqrt{\eta} Z_k^{(i)}\,; \quad \forall\, i \in [n]\,, \tag{2.3}$$

where the position of each particle $X_{k+1}^{(i)}$, for $i \in [n]$, at time $k + 1$ depends on all the particles via the empirical distribution $\hat{\mu}_k$ of the n particles.

In our algorithm we introduce a tractable particle based approximation of $\nabla_{\mathcal{W}} \mathcal{F}(\cdot, \mu_k)$, called the virtual particle stochastic approximation, a generalization of VP-SVGD originally introduced in Das and Nagaraj (2023). In particular, we introduce two types of particles called *real*

and *virtual*. The virtual particles are used to approximate the gradient but are not part of the final output of the algorithm. We denote $X_k^{(i)}, Y_k^{(i)}$ to be the real and virtual particles respectively at time k.

Next, we propose an unbiased estimator $\hat{G} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ of $\nabla_{\mathcal{W}}\mathcal{F}$, i.e., if $Y, \xi \sim \mu \times \nu$, then $\mathbb{E}[\hat{G}(x, Y, \xi)] = \nabla_{\mathcal{W}}\mathcal{F}(x, \mu)$ for every $x \in \mathbb{R}^d$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, and where ν is a fixed, known distribution.

Formally, our algorithm is defined as follows.

Algorithm 1: Virtual particle stochastic approximation

Input: Time steps T, number of samples n, timestep size η. Initial Distribution μ_0, distribution ν, estimator $\hat{G}$

Output: $X_T^{(1)}, \ldots, X_T^{(n)}$

$X_0^{(1)}, \ldots, X_0^{(n)}, Y_0^{(0)}, \ldots, Y_0^{(T)} \overset{\text{i.i.d.}}{\sim} \mu_0$

$\xi_1, \ldots, \xi_T \overset{\text{i.i.d.}}{\sim} \nu$

$k \leftarrow 0$

while $k \leq T$ **do**

  for $i = 1, .., n$ **do**

    $X_{k+1}^{(i)} = X_k^{(i)} + \eta\hat{G}(X_k^{(i)}, Y_k^{(k)}, \xi_k) + \sigma\sqrt{\eta}Z_k^{(i)} \quad Z_k^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$

  end

  for $j = k + 1, ...T$, **do**

    $Y_{k+1}^{(j)} = Y_k^{(j)} + \eta\hat{G}(Y_k^{(j)}, Y_k^{(k)}, \xi_k) + \sigma\sqrt{\eta}W_k^{(j)} \quad W_k^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$

  end

end

At timestep $k + 1$, the virtual particle $Y_k^{(k)}$ is used to estimate μ_k for all the real particles $X_k^{(i)}$, and the remaining virtual particles $Y_k^{(j)}$ are discarded. Let $\mathcal{R}_k$ be the sigma algebra of $Y_0^{(0)}, \ldots, Y_k^{(k)}, \xi_0, \ldots, \xi_k$.

Our Algorithm 1 produces n i.i.d. samples from $\mu_T|\mathcal{R}_{T-1}$, which is shown to converge to π in Theorem 1. The computational complexity is $O(nT + T^2)$, as opposed to the $O(n^2T)$ complexity incurred in the straight forward particle approximation (2.3). Our computational complexity follows from the fact that we use n real particles at each time step for T steps and at most T virtual particles at each time step for T steps. The computational complexity of the particle approximation in (2.3) follows from the fact every particle depends on all of the n particles at each time step for T steps.

In algorithm 1, we call the diagonal trajectory $Y_k^{(k)}$ for $0 \leq k \leq T$ as the "witness path". Notice that given a witness path, we can obtain a sample from $\mu_T|\mathcal{R}_{T-1}$ in T steps, without fixing n beforehand. Therefore, storing the witness path yields an approximate sampling algorithm from the global minimum π of $\mathcal{E}$ if the McKean-Vlasov SDE in (2.2) converges to π.

2.4.1 Applications. For a concrete example of application of Algorithm 1, we consider a specific form for the functional $\mathcal{E}$ defined in (**??**) and call it the *pairwise interaction energy* case.

Let $V, W : \mathbb{R}^d \to \mathbb{R}$ and $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$. Recall that V is commonly referred to as the external potential and W the interaction potential. Let W be even (i.e, $W(x) = W(-x)$). The functional $\mathcal{E}$ in this case is defined as:

$$\mathcal{E}(\mu) := \int V(x)d\mu(x) + \tfrac{1}{2}\int\int W(x-y)d\mu(x)d\mu(y) + \tfrac{\sigma^2}{2}\mathcal{H}(\mu)\,. \tag{2.4}$$

Here the Wasserstein gradient flow gives the following McKean-Vlasov dynamics

$$dX_t = -\nabla V(X_t)dt - (\nabla W * \mu_t)(X_t) + \sqrt{\sigma}dB_t\,,$$

where $\mu_t = \mathsf{Law}(X_t)$ and $*$ stands for convolution. From Ambrosio and Savaré (2007)[Proposition 4.13], for the potential energy functional $\mathcal{V}(\mu) := \int V(x)\mu(dx)$, the Wasserstein gradient is given by $\nabla_{\mathcal{W}}\mathcal{V}(\mu) = \nabla V$. Next, from Ambrosio and Savaré (2007)[Theorem 4.19], for the interaction energy functional $\mathscr{W}(\mu) := \frac{1}{2}\int\int W(x-y)\mu(dx)\mu(dy)$, the Wasserstein gradient is given by $\nabla_{\mathcal{W}}\mathscr{W}(\mu) = (\nabla W) * \mu$ if $(\nabla W) * \mu \in L^2(\mu;\mathbb{R}^d)$. Finally, from Ambrosio and Savaré (2007)[Theorem 4.16],

for the entropy functional $\mathcal{H}(\mu) := \int \log \mu d\mu$, the Wasserstein gradient is given by $\nabla_{\mathcal{W}}\mathcal{H}(\mu) = \nabla \log \mu$ if $\nabla \log \mu \in L^2(\mu; \mathbb{R}^d)$. Thus,

$$\nabla_{\mathcal{W}}\mathcal{E}(\mu) = \nabla V + (\nabla W) * \mu + \tfrac{\sigma^2}{2}\nabla \log \mu \,. \tag{2.5}$$

The unique minimizer of the functional $\mathcal{E}$ defined in (2.4) satisfies a fixed point equation and is given in Kook et al. (2024)[Equation 1.1] as:

$$\pi(x) \propto \exp\left(-\tfrac{2}{\sigma^2}V(x) - \tfrac{2}{\sigma^2}W * \pi(x)\right) . \tag{2.6}$$

2.5 Convergence analysis

In this section, we first establish a general convergence theory for Algorithm 1. We begin with a key *descent lemma* (Lemma 10) that bounds the evolution of the energy functional by decomposing the error into discretization, stochastic, and linearization terms. Using this, we prove our main result (Theorem 1) showing that Algorithm 1 produces i.i.d. samples converging to the minimizer of the energy functional, with rates independent of the number of particles n. We then demonstrate how this framework applies to the pairwise interaction energy case. Indeed, we verify the assumptions of Theorem 1 and establish explicit convergence rates for the pairwise interaction energy case while avoiding the need for separate propagation of chaos bounds.

2.5.1 General convergence:.

For some functional $\bar{\mathcal{F}} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, let π be the unique minimizer of the functional $\bar{\mathcal{F}} + \frac{\sigma^2}{2}\mathcal{H}$. Define $\bar{\mathcal{E}}(\mu) := \bar{\mathcal{F}}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu) - \bar{\mathcal{F}}(\pi) - \frac{\sigma^2}{2}\mathcal{H}(\pi)$ (not necessarily $\mathcal{E}$). The functional $\bar{\mathcal{E}}$ is introduced, rather than simply analyzing $\mathcal{E}$ since $\bar{\mathcal{E}}$ can have better contraction properties. This is indeed the case with pairwise interaction energy where the KL functional to the minimizer π has a contraction whenever π satisfies an LSI.

We consider Algorithm 1, with $\hat{G}$ such that whenever $(Y, \xi) \sim \mu \times \nu$, we have: $\mathbb{E}[\hat{G}(x, Y, \xi)] = -\nabla_{\mathcal{W}}\mathcal{F}(x, \mu)$ for every $x \in \mathbb{R}^d$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. We do not assume $\mathcal{F} \neq \bar{\mathcal{F}}$, which is important for the case of the interaction energy. We track the evolution of $\mathbb{E}\bar{\mathcal{E}}(\mu_k|\mathcal{R}_{k-1})$ along the discrete time trajectory $\mu_k|\mathcal{R}_{k-1}$ via continuous interpolations. We then specialize to the case of pairwise interaction energy (Equation (2.4)) allowing us to obtain convergence bounds of Algorithm 1 for this specific case.

To simplify notation, consider X_0, Y i.i.d from a distribution $\rho_0 \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ and $\xi \sim \nu$ independent of X_0, Y. For $t \in [0, \eta]$, we consider the random variable $X_t := X_0 + tu(X_0, Y, \xi) + \sigma B_t$, where $u : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a velocity field and B_t is the standard $\mathbb{R}^d$ Brownian motion independent of everything else. Assume that $\mathbb{E}_{Y,\xi}[u(x, Y, \xi)] = -\nabla_{\mathcal{W}} \mathcal{F}(x, \rho_0)$ for every $x \in \mathbb{R}^d$ and that $u(x, Y, \xi)$ has a finite second moment when $x \sim \rho_t(|Y, \xi)$ almost surely Y, ξ. Let $\rho_t(|Y, \xi) := \mathsf{Law}(X_t|Y, \xi)$. Here ρ_0, X_0, Y, ξ corresponds to $\mu_k|\mathcal{R}_{k-1}, X_k^{(1)}, Y_k^{(k)}, \xi_k$ respectively. The velocity field $u(x, Y, \xi)$ corresponds to $\hat{G}(x, Y, \xi)$. We use the notation u and $\hat{G}$ interchangeably. Under this correspondence, it is clear that $X_\eta|Y, \xi$ has the same distribution as $\mu_{k+1}|\mathcal{R}_k$. Following the proof of (Vempala & Wibisono, 2019, Lemma 3), we conclude that ρ_t satisfies the Fokker-Planck equation:

$$\begin{aligned}\frac{\partial \rho_t(x|Y,\xi)}{\partial t} &= -\nabla_x.(\rho_t(x|Y,\xi)\mathbb{E}[u(X_0, Y, \xi)|X_t = x, Y, \xi]) + \frac{\sigma^2}{2}\Delta_x \rho_t(x|Y,\xi) \\ &= -\nabla_x.(\rho_t(x|Y,\xi)v_t(x, Y, \xi))\,, \end{aligned} \tag{2.7}$$

where

$$v_t(x, Y, \xi) := \mathbb{E}[u(X_0, Y, \xi)|X_t = x, Y, \xi] - \frac{\sigma^2}{2}\nabla \log \rho_t(x|Y, \xi)\,,\ \forall x \in \mathbb{R}^d, t \in [0, \eta]\,. \tag{2.8}$$

Taking $\nabla_{\mathcal{W}} \bar{\mathcal{E}}(x, \rho_t(|Y, \xi)) = \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t(|Y, \xi)) + \frac{\sigma^2}{2}\nabla \log \rho_t(x|Y, \xi)$, the following lemma describes the evolution of the functional $\bar{\mathcal{E}}$ along the trajectory $\rho_t(|y, \xi)$.

Lemma 1. *$\rho_0 \times \nu$ almost surely (y, ξ), suppose that the energy functional $\bar{\mathcal{E}}(\rho_t(|y, \xi))$ satisfies:*

$$\nabla_{\mathcal{W}} \bar{\mathcal{E}}(, \rho_t(|y, \xi)) \in L^2(\rho_t(|y, \xi)); \quad \frac{d\bar{\mathcal{E}}(\rho_t(|y, \xi))}{dt} = \left\langle \nabla_{\mathcal{W}} \bar{\mathcal{E}}(\cdot, \rho_t(|y, \xi)), v_t(\cdot, y, \xi) \right\rangle_{L_2(\rho_t(|y,\xi))}$$

Then,

$$\frac{d\bar{\mathcal{E}}(\rho_t(|Y, \xi))}{dt} = -\int d\rho_t(x|Y, \xi)\|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(x, \rho_t(|Y, \xi))\|^2 + \mathsf{DE}_1(t) + \mathsf{DE}_2(t) + \mathsf{SE}(t) + \mathsf{LE}(t)\,, \tag{2.9}$$

where the discretization, stochastic, and linearlization errors are defined as

1. $\mathsf{DE}_1(t) := \left\langle \nabla_{\mathcal{W}} \bar{\mathcal{E}}(\cdot, \rho_t(|Y, \xi)), \mathbb{E}[u(X_0, Y, \xi)|X_t = \cdot, Y, \xi] - u(\cdot, Y, \xi) \right\rangle_{L_2(\rho_t(|Y,\xi))}$
2. $\mathsf{DE}_2(t) := \left\langle \nabla_{\mathcal{W}} \bar{\mathcal{E}}(\cdot, \rho_t(|Y, \xi)), \nabla_{\mathcal{W}} \mathcal{F}(\cdot, \rho_t(|Y, \xi)) - \nabla_{\mathcal{W}} \mathcal{F}(\cdot, \rho_0) \right\rangle_{L_2(\rho_t(|Y,\xi))}$

3. $\mathsf{SE}(t) := \left\langle \nabla_{\mathcal{W}} \bar{\mathcal{E}}(\cdot, \rho_t(|Y,\xi)), u(\cdot, Y, \xi) + \nabla_{\mathcal{W}} \mathcal{F}(\cdot, \rho_0) \right\rangle_{L_2(\rho_t(\cdot|Y,\xi))}$

4. $\mathsf{LE}(t) := \left\langle \nabla_{\mathcal{W}} \bar{\mathcal{E}}(\cdot, \rho_t(|Y,\xi)), \nabla_{\mathcal{W}} \bar{\mathcal{F}}(\cdot, \rho_t(|Y,\xi)) - \nabla_{\mathcal{W}} \mathcal{F}(\cdot, \rho_t(|Y,\xi) \right\rangle_{L_2(\rho_t(|Y,\xi))}$

Remark 1. *Here,* $\mathsf{DE}_1(t), \mathsf{DE}_2(t), \mathsf{SE}(t), \mathsf{LE}(t)$ *are all functions of* (Y, ξ)*. The discretization error* DE_1 *arises since* X_0 *is used instead of* X_t *in* $u(X_0, Y, \xi)$ *and* DE_2 *arises since* $-\mathbb{E}[u(x, Y, \xi)] = \nabla_{\mathcal{W}} \mathcal{F}(x, \rho_0) \neq \nabla_{\mathcal{W}} \mathcal{F}(x, \rho_t(|Y,\xi))$*.* SE *is the stochastic error, since we are using an estimator* $u(\cdot, Y, \xi)$ *of* $\nabla_{\mathcal{W}} \mathcal{F}(\cdot, \rho_0)$*.* LE *is the linearization error since we consider the evolution of* $\bar{\mathcal{E}}$ *where as the gradient descent is for* $\mathcal{E} \neq \bar{\mathcal{E}}$*. This is important for linearizing the non-linear Fokker-Planck equation, in the case of pairwise interaction energy.*

The next lemma bounds each of the error terms defined in the previous lemma in expectation.

Lemma 2. *In the setting of Lemma 1, let* π *be the unique global minimizer of* $\bar{\mathcal{E}}$*. Let* $(Y^*, \xi) \sim \pi \times \nu$*. Define* $(\sigma^*)^2 := \mathbb{E}_{x \sim \pi^*} \mathsf{Var}(u(x, Y^*, \xi))$*,* $(G_\pi)^2 := \mathbb{E}_{x \sim \pi} \|\nabla_{\mathcal{W}} \mathcal{F}(x, \pi)\|^2$ *and* $J_t := \sqrt{\mathbb{E}\|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(X_t, \rho_t(|Y,\xi))\|^2}$

1. *Suppose the function* $x \to u(x, y, \xi)$ *and* $y \to u(x, y, \xi)$ *are* L_u*-Lipschitz. Then,*

$$\mathbb{E}[\mathsf{DE}_1(t)] \leq L_u J_t \cdot \sqrt{2t^2 \left(2L_u^2 \mathcal{W}_2^2(\rho_0, \pi) + (\sigma^*)^2 + (G_\pi)^2\right) + \sigma^2 t d}\,. \tag{2.10}$$

2. *Suppose* $\|\nabla_{\mathcal{W}} \mathcal{F}(x, \mu) - \nabla_{\mathcal{W}} \mathcal{F}(x, \nu)\| \leq L_{\mathcal{F}} \mathcal{W}_2(\mu, \nu)$ *Then,*

$$\mathbb{E}[\mathsf{DE}_2(t)] \leq L_{\mathcal{F}} J_t \cdot \sqrt{2t^2 \left(2L_u^2 \mathcal{W}_2^2(\rho_0, \pi) + (\sigma^*)^2 + (G_\pi)^2\right) + \sigma^2 t d}\,. \tag{2.11}$$

3. *Assume that* $x \to u(x, y, \xi)$*,* $y \to u(x, y, \xi)$ *and* $x \to \nabla_{\mathcal{W}} \mathcal{F}(x, \mu)$ *are continuously differentiable and* L_u *Lipschitz. Define* $\Theta(x, y, \xi) := u(x, y, \xi) + \nabla_{\mathcal{W}} \mathcal{F}(x, \rho_0)$*. Assume* $\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(x_1, \mu) - \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x_2, \nu)\| \leq L_{\bar{u}} \|x_1 - x_2\| + L_{\bar{\mathcal{F}}} \mathcal{W}_2(\mu, \nu)$

$$\begin{aligned}\mathbb{E}[\mathsf{SE}(t)] &\leq \sigma \sqrt{2t} L_u d \sqrt{\mathbb{E}_{x \sim \rho_0}[\mathsf{Var}(u(x, Y, \xi)]} \\ &\quad + 2\sqrt{\mathbb{E}[\mathcal{W}_2^2(\rho_t, \rho_t(\cdot|Y,\xi))]} \left[(L_{\bar{u}} + L_{\bar{\mathcal{F}}}) \sqrt{\mathbb{E}\left\|\Theta(X_t, Y, \xi)\right\|^2} + L_u \sqrt{\mathbb{E}\left\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(X_t, \rho_t)\right\|^2} \right]\end{aligned} \tag{2.12}$$

4. Suppose $\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)\| \le L_l \mathcal{W}_2(\pi,\mu)$. *Then,*

$$\mathbb{E}[\mathsf{LE}(t)] \le L_l J_t \cdot \sqrt{\mathbb{E}\left[\mathcal{W}_2^2(\pi, \rho_t(|Y,\xi))\right]}\,. \tag{2.13}$$

Proof. First, we bound $\mathbb{E}_{Y,\xi}[\mathsf{DE}_1(t)]$. Moving the expectation out of the inner product, applying Cauchy-Schwarz inequality, using the assumption that the function $(x,y) \to u(x,y,\xi)$ is L_u-Lipschitz, and the fact that $X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t$, we get:

$$\begin{aligned}
\mathbb{E}_{Y,\xi}[\mathsf{DE}_1(t)] &= \mathbb{E}\langle \nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t, \rho_t(|Y,\xi)), u(X_0,Y,\xi) - u(X_t,Y,\xi)\rangle \\
&\le \mathbb{E}\left[\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t, \rho_t(|Y,\xi))\| \cdot \|u(X_0,Y,\xi) - u(X_t,Y,\xi)\|\right] \\
&\le L_u \mathbb{E}\left[\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t, \rho_t(|Y,\xi))\| \cdot \|X_t - X_0\|\right] \\
&= L_u \mathbb{E}\left[\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t, \rho_t(|Y,\xi))\| \cdot \|tu(X_0,Y,\xi) + \sigma B_t\|\right] \\
&\le L_u \sqrt{\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t, \rho_t(|Y,\xi))\|^2}\sqrt{\mathbb{E}\|tu(X_0,Y,\xi) + \sigma B_t\|^2} \\
&= L_u \sqrt{\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t, \rho_t(|Y,\xi))\|^2}\sqrt{t^2\mathbb{E}\|u(X_0,Y,\xi)\|^2 + \sigma^2 t d}\,.
\end{aligned} \tag{2.14}$$

Next, we bound $\mathbb{E}\|u(X_0,Y,\xi)\|^2$. Let $(X^*,Y^*) \sim \pi \times \pi$ be optimally coupled to $(X_0,Y) \sim \rho_0 \times \rho_0$ in the 2-Wasserstein distance and independent of ξ. Thus, by the triangle inequality, the inequality $(a+b)^2 \le 2a^2 + 2b^2$, by L_u-Lipschitz continuity of $(x,y) \to u(x,y,\xi)$, and the fact that $\mathcal{W}_2^2(\mu\times\mu, \nu\times\nu) \le 2\mathcal{W}_2^2(\mu,\nu)$, for any probability measures μ,ν, we have:

$$\begin{aligned}
\mathbb{E}\|u(X_0,Y,\xi)\|^2 &\le 2\mathbb{E}\|u(X_0,Y,\xi) - u(X^*,Y^*,\xi)\|^2 + 2\mathbb{E}\|u(X^*,Y^*,\xi)\|^2 \\
&\le 2L_u^2\mathbb{E}[\|X_0 - X^*\|^2 + \|Y - Y^*\|^2] + 2\mathbb{E}\|u(X^*,Y^*,\xi)\|^2 \\
&= 2L_u^2\mathcal{W}_2^2(\rho_0\times\rho_0, \pi\times\pi) + 2\mathbb{E}\|u(X^*,Y^*,\xi)\|^2 \\
&\le 4L_u^2\mathcal{W}_2^2(\rho_0,\pi) + 2\mathbb{E}\|u(X^*,Y^*,\xi)\|^2\,.
\end{aligned} \tag{2.15}$$

Next, using $\mathbb{E}[u(x,Y^*,\xi)] = -\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)$, for every x, we obtain:

$$\begin{aligned}
\mathbb{E}\|u(X^*,Y^*,\xi)\|^2 &= \mathbb{E}\|u(X^*,Y^*,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(X^*,\pi)\|^2 + \mathbb{E}\|\nabla_{\mathcal{W}}\mathcal{F}(X^*,\pi)\|^2 \\
&= (\sigma^*)^2 + (G_\pi)^2\,.
\end{aligned} \tag{2.16}$$

Now by using the bounds (2.15) and (2.16) in (2.14) proves (2.10).

Next, we bound $\mathbb{E}_{Y,\xi}[\mathsf{DE}_2(t)]$. Using the assumption $\|\nabla_{\mathcal{W}}\mathcal{F}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\nu)\| \leq L_{\mathcal{F}}\mathcal{W}_2(\mu,\nu)$, applying the Cauchy-Schwarz inequality and Jensen's inequality, we get:

$$\begin{aligned}\mathbb{E}_{Y,\xi}[\mathsf{DE}_2(t)] &\leq L_{\mathcal{F}}\mathbb{E}\left[\mathcal{W}_2(\rho_t(\cdot|Y,\xi),\rho_0)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t,\rho_t(|Y,\xi))\|\right] \\ &\leq L_{\mathcal{F}}\sqrt{\mathbb{E}[\mathcal{W}_2^2(\rho_t(\cdot|Y,\xi),\rho_0)]}\cdot\sqrt{\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t,\rho_t(|Y,\xi))\|^2}\,. \end{aligned} \tag{2.17}$$

Next, we bound $\mathbb{E}[\mathcal{W}_2^2(\rho_t(\cdot|Y,\xi),\rho_0)]$. Since $X_t|Y,\xi \sim \rho_t(\cdot|Y,\xi)$, $X_0 \sim \rho_0$ and $X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t$. Therefore, by the definition of the Wasserstein distance, we have:

$$\begin{aligned}\mathbb{E}[\mathcal{W}_2^2(\rho_t(\cdot|Y,\xi),\rho_0)] &\leq \mathbb{E}[\mathbb{E}[\|X_t - X_0\|^2|Y,\xi]] \\ &= \mathbb{E}\|tu(X_0,Y,\xi) + \sigma B_t\|^2 \\ &= t^2\mathbb{E}_{(X,Y,\xi)\sim\rho_0\times\rho_0\times\nu}\|u(X,Y,\xi)\|^2 + \sigma^2 td \\ &\leq 4t^2L_u^2\mathcal{W}_2^2(\rho_0,\pi) + 2t^2((\sigma^*)^2 + (G_\pi)^2) + \sigma^2 td\,, \end{aligned} \tag{2.18}$$

where the last inequality follows from (2.15) and (2.16). By plugging the bound in (2.18) into (2.17) proves (2.11).

Next, we bound $\mathbb{E}_{Y,\xi}[\mathsf{SE}(t)]$. Define $\Theta(x,y,\xi) := u(x,y,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)$. Note that since $\mathbb{E}[u(x,Y,\xi)] = -\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)$, for every x, we have:

$$\begin{aligned}\mathbb{E}[\nabla_x\cdot\Theta(x,Y,\xi)] &= \mathbb{E}[\nabla_x\cdot u(x,Y,\xi) + \nabla_x\cdot\nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)] \\ &= \nabla_x\cdot\mathbb{E}[u(x,Y,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)] = 0\,. \end{aligned} \tag{2.19}$$

Here, we have exchanged the integral and derivative using the dominated convergence theorem along with Lipchitz continuity of $u, \nabla_{\mathcal{W}}\mathcal{F}$. Using $\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(|y,\xi)) = \nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t(|y,\xi)) +$

$\frac{\sigma^2}{2}\nabla \log \rho_t(x|y,\xi)$ and integrating by parts, we get:

$$\begin{aligned}
\mathbb{E}_{Y,\xi}[\mathsf{SE}(t)] &= \mathbb{E}\left\langle \nabla_{\mathcal{W}}\bar{\mathcal{F}}(\cdot,\rho_t(|Y,\xi)), \Theta(\cdot,Y,\xi)\right\rangle_{L_2(\rho_t(|Y,\xi))} \\
&\quad + \frac{\sigma^2}{2}\mathbb{E}\int \rho_t(x|Y,\xi)\langle \nabla_x \log \rho_t(x|Y,\xi)\,,\, \Theta(x,Y,\xi)\rangle dx \\
&= \mathbb{E}\left\langle \nabla_{\mathcal{W}}\bar{\mathcal{F}}(\cdot,\rho_t(|Y,\xi)), \Theta(\cdot,Y,\xi)\right\rangle_{L_2(\rho_t(|Y,\xi))} \\
&\quad + \frac{\sigma^2}{2}\mathbb{E}\int \langle \nabla_x \rho_t(x|y,\xi)\,,\, \Theta(x,y,\xi)\rangle dx, \qquad (\text{since } \nabla \log p_t = \tfrac{\nabla p_t}{p_t}) \\
&= \mathbb{E}\left\langle \nabla_{\mathcal{W}}\bar{\mathcal{F}}(\cdot,\rho_t(|Y,\xi)), \Theta(\cdot,Y,\xi)\right\rangle_{L_2(\rho_t(|Y,\xi))} \\
&\quad - \frac{\sigma^2}{2}\mathbb{E}\int (\nabla_x \cdot \Theta(x,Y,\xi))\rho_t(x|Y,\xi)dx, \qquad (\text{integration by parts})\,. \qquad (2.20)
\end{aligned}$$

Now we bound the term $\mathbb{E}\int(\nabla_x \cdot \Theta(x,Y,\xi))\rho_t(x|Y,\xi)dx$. From (2.19) it follows that

$$\mathbb{E}\int (\nabla_x \cdot \Theta(x,Y,\xi))\rho_t(x|y,\xi)dx = \mathbb{E}\int (\nabla_x \cdot \Theta(x,Y,\xi))(\rho_t(x|Y,\xi) - \rho_t(x))dx\,.$$

Since the functions $x \to u(x,y,\xi)$ and $x \to \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)$ are continuously differentiable and L_u-Lipschitz continuous, $\Theta(x,y,\xi)$ is $2L_u$-Lipschitz continuous and continuously differentiable. Thus $|\nabla_x \cdot \Theta(x,y,\xi)| \le 2dL_u$ uniformly. By noting $\frac{1}{2}\int |\rho_t(x|y,\xi) - \rho_t(x)|dx$ is the total-variation distance between $\rho_t(\cdot|y,\xi)$ and ρ_t, and applying Pinsker's inequality, we get:

$$\begin{aligned}
\left|\mathbb{E}\int (\nabla \cdot \Theta(x,Y,\xi))(\rho_t(x|Y,\xi) - \rho_t(x))dx\right| &\le \mathbb{E}\int |\nabla \cdot \Theta(x,Y,\xi)| \cdot |\rho_t(x|Y,\xi) - \rho_t(x)|dx \\
&\le 4dL_u\mathbb{E}[\mathsf{TV}(\rho_t,\rho_t(|Y,\xi))] \\
&\le 4dL_u\mathbb{E}\sqrt{\frac{1}{2}\,\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t)} \\
&= \sqrt{8}dL_u\mathbb{E}\sqrt{\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t)}\,. \qquad (2.21)
\end{aligned}$$

Since $\rho_t = \mathbb{E}[\rho_t(\cdot \mid Y,\xi)]$, we consider Y',ξ' to be an i.i.d copy of (Y,ξ) respectively and to be independent of X_0. Since the KL divergence functional is convex jointly in its arguments, we have:

$$\begin{aligned}
\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t) &= \mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\mathbb{E}[\rho_t(\cdot \mid Y',\xi')]) \\
&\le \mathbb{E}\left[\mathrm{KL}(\rho_t(\cdot \mid Y,\xi)\,||\,\rho_t(\cdot \mid Y',\xi'))|Y,\xi\right]\,. \qquad (2.22)
\end{aligned}$$

Further conditioning on X_0 and taking an expectation yields $\rho_t(\cdot|Y,\xi) = \mathbb{E}[\rho_t(\cdot|X_0,Y,\xi)|Y,\xi]$. Another application of the joint convexity of the KL divergence functional in the above inequality gives

$$\mathbb{E}\left[\mathrm{KL}(\rho_t(\cdot|Y,\xi)\,||\,\rho_t(\cdot|Y',\xi'))|Y,\xi\right] \leq \mathbb{E}\left[\mathrm{KL}(\rho_t(\cdot\mid X_0,Y,\xi)\,||\,\rho_t(\cdot\mid X_0,Y',\xi'))|Y,\xi\right]\,. \quad (2.23)$$

Since $X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t$, we have $\rho_t(\cdot|X_0,Y,\xi) = \mathcal{N}(\mu_1,\sigma^2 tI)$, $\rho_t(\cdot|X_0,Y',\xi') = \mathcal{N}(\mu_2,\sigma^2 tI)$, where $\mu_1 := X_0 + tu(X_0,Y,\xi)$, $\mu_2 := X_0 + tu(X_0,Y',\xi')$. Using the formula for KL-divergence between multivariate normal distributions in Duchi (2007) yields:

$$\begin{aligned}\mathrm{KL}(\rho_t(\cdot\mid X_0,Y,\xi)\,||\,\rho_t(\cdot\mid X_0,Y',\xi')) &= \frac{\|\mu_1-\mu_2\|^2}{2\sigma^2 t}\\ &= \frac{t}{2\sigma^2}\|u(X_0,Y,\xi)-u(X_0,Y',\xi')\|^2\,. \end{aligned} \quad (2.24)$$

From the bounds in (2.22), (2.23), (2.24), and by applying Jensen's inequality to the outer expectation, we obtain

$$\begin{aligned}\mathbb{E}\sqrt{\mathrm{KL}(\rho_t(\cdot\mid Y,\xi)\,||\,\rho_t)} &\leq \mathbb{E}\sqrt{\frac{t}{2\sigma^2}\mathbb{E}\big[\|u(X_0,Y,\xi)-u(X_0,Y',\xi')\|^2\big|Y,\xi\big]}\\ &\leq \frac{\sqrt{t}}{\sigma}\sqrt{\frac{1}{2}\mathbb{E}\|u(X_0,Y,\xi)-u(X_0,Y',\xi')\|^2}\\ &\leq \frac{\sqrt{t}}{\sigma}\sqrt{\mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x,Y,\xi))]}\,. \end{aligned} \quad (2.25)$$

Then by plugging the bound in (2.25) into (2.21), we get:

$$\left|\mathbb{E}\int(\nabla_x\cdot\Theta(x,Y,\xi))(\rho_t(x|Y,\xi)-\rho_t(x))dx\right| \leq \frac{\sqrt{8}\sqrt{t}L_u d}{\sigma}\sqrt{\mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x,Y,\xi))]}\,. \quad (2.26)$$

Next, we bound the first term in (2.20). First, note that from (2.19), we have:

$$\mathbb{E}\int\langle\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\,,\,\Theta(x,Y,\xi)\rangle\rho_t(x)dx = 0\,.$$

Using this fact, we obtain:

$$\begin{aligned}
&\mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(X_t, \rho_t(|Y,\xi))\,,\, \Theta(X_t, Y, \xi)\rangle_{L_2(\rho_t(|Y,\xi))} \\
&= \mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(X_t, \rho_t(|Y,\xi))\,,\, \Theta(X_t, Y, \xi)\rangle_{L_2(\rho_t(|Y,\xi))} - \mathbb{E}\int \langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t)\,,\, \Theta(x, Y, \xi)\rangle \rho_t(x) dx \\
&= \mathbb{E}\int \langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t(|Y,\xi))\,,\, \Theta(x, Y, \xi)\rangle \rho_t(x|Y,\xi) dx \\
&\quad - \mathbb{E}\int \langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t)\,,\, \Theta(x, Y, \xi)\rangle \rho_t(x) dx\,.
\end{aligned}$$

For a given Y, ξ, let $Z_1 \sim \rho_t(\cdot|Y,\xi)$ and $Z_2 \sim \rho_t$ be optimally coupled in the 2-Wasserstein distance. By the Cauchy-Schwarz inequality we conclude that almost surely Y, ξ:

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t(|y,\xi))\,,\, \Theta(x, Y, \xi)\rangle - \mathbb{E}_{x\sim\rho_t}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t)\,,\, \Theta(x, Y, \xi)\rangle \\
&\quad = \mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(Z_1, \rho_t(\cdot|y,\xi))\,,\, \Theta(Z_1, Y, \xi)\rangle - \mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(Z_2, \rho_t)\,,\, \Theta(Z_2, Y, \xi)\rangle \\
&\quad = \mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(Z_1, \rho_t(\cdot|Y,\xi)) - \nabla_{\mathcal{W}} \bar{\mathcal{F}}(Z_2, \rho_t)\,,\, \Theta(Z_1, Y, \xi)\rangle \\
&\quad + \mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(Z_2, \rho_t)\,,\, \Theta(Z_1, y, \xi) - \Theta(Z_2, y, \xi)\rangle \\
&\quad \le \mathbb{E}\big\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(Z_1, \rho_t(\cdot|Y,\xi)) - \nabla_{\mathcal{W}} \bar{\mathcal{F}}(Z_2, \rho_t)\big\| \cdot \big\|\Theta(Z_1, Y, \xi)\big\| \\
&\quad + \mathbb{E}\big\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(Z_2, \rho_t)\big\| \cdot \big\|\Theta(Z_1, Y, \xi) - \Theta(Z_2, Y, \xi)\big\|\,.
\end{aligned} \tag{2.27}$$

Next, note that by the definition of $\Theta(x, y, \xi)$, the assumption that $(x, y) \to u(x, y, \xi)$ and $x \to \nabla_{\mathcal{W}} \mathcal{F}(x, \mu)$ are L_u-Lipschitz, and the assumption $\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \mu) - \nabla_{\mathcal{W}} \bar{\mathcal{F}}(y, \nu)\| \le L_{\bar{u}}\|x - y\| + L_{\bar{\mathcal{F}}}\mathcal{W}_2(\mu, \nu)$, we obtain:

$$\big\|\Theta(X_1, Y, \xi) - \Theta(X_2, Y, \xi)\big\| \le 2L_u\|X_1 - X_2\|\,,$$

$$\big\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(X_1, \rho_t(\cdot|Y,\xi)) - \nabla_{\mathcal{W}} \bar{\mathcal{F}}(X_2, \rho_t)\big\| \le L_{\bar{u}}\|X_1 - X_2\| + L_{\bar{\mathcal{F}}}\mathcal{W}_2(\rho_t, \rho_t(\cdot|Y,\xi))\,.$$

Using the above bounds in (2.27), applying the Cauchy-Schwarz inequality, and the Jensen's inequality yields the following almost surely Y, ξ:

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t(|Y,\xi))\,,\, \Theta(x, Y, \xi)\rangle - \mathbb{E}_{x\sim\rho_t}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t)\,,\, \Theta(x, Y, \xi)\rangle \\
&\quad \le \sqrt{2}(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\mathcal{W}_2(\rho_t, \rho_t(\cdot|Y,\xi))\sqrt{\mathbb{E}_{x\sim\rho_t(|Y,\xi)}\big\|\Theta(x, Y, \xi)\big\|^2} \\
&\quad + 2L_u\mathcal{W}_2(\rho_t, \rho_t(\cdot|Y,\xi))\sqrt{\mathbb{E}_{x\sim\rho_t}\big\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \rho_t)\big\|^2}\,.
\end{aligned}$$

Hence, by taking expectation with respect to Y, ξ and another application of the Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned}
&\mathbb{E}\langle \nabla_{\mathcal{W}} \bar{\mathcal{F}}(X_t, \rho_t(|Y,\xi))\,,\, \Theta(X_t, Y, \xi)\rangle_{L_2(\rho_t(|Y,\xi))} \\
&\quad \leq 2\sqrt{\mathbb{E}[\mathcal{W}_2^2(\rho_t, \rho_t(\cdot|Y,\xi))]}\left[(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\sqrt{\mathbb{E}\left\|\Theta(X_t, Y, \xi)\right\|^2} + L_u\sqrt{\mathbb{E}\left\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_t, \rho_t)\right\|^2}\right].
\end{aligned} \tag{2.28}$$

Finally, by multiplying the bound in (2.26) by $\sigma^2/2$, adding the resultant to (2.28), and using (2.20) we prove (2.12).

Finally, we bound $\mathbb{E}_{y,\xi}[\mathsf{LE}(t)]$. By applying the Cauchy-Schwartz inequality twice and the assumption $\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)\| \leq L_l \mathcal{W}_2(\pi,\mu)$, we have almost surely Y, ξ:

$$\begin{aligned}
\mathsf{LE}(t) &= \mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\langle \nabla_{\mathcal{W}}\bar{\mathcal{E}}(x, \rho_t(|Y,\xi)), \nabla_{\mathcal{W}}\bar{\mathcal{F}}(x, \rho_t(|Y,\xi)) - \nabla_{\mathcal{W}}\mathcal{F}(x, \rho_t(|Y,\xi)\rangle \\
&\leq \mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\left[\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x, \rho_t(|Y,\xi))\| \cdot \|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x, \rho_t(|Y,\xi)) - \nabla_{\mathcal{W}}\mathcal{F}(x, \rho_t(|Y,\xi)\|\right] \\
&\leq L_l \mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\left[\mathcal{W}_2(\pi, \rho_t(\cdot|Y,\xi))\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x, \rho_t(|Y,\xi))\|\right] \\
&\leq L_l\sqrt{\mathcal{W}_2^2(\pi, \rho_t(\cdot|Y,\xi))}\sqrt{\mathbb{E}_{x\sim\rho_t(\cdot|Y,\xi)}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x, \rho_t(|Y,\xi))\|^2}\,.
\end{aligned} \tag{2.29}$$

Next, by taking an expectation with respect to $(Y,\xi) \sim \rho_0 \times \nu$ and applying the Jensen's inequality, we obtain (2.13). □

Next, we make a few assumptions that aid in the proof of the descent lemma (Lemma 10). Assumption 7 concerns the regularity of the functional $\bar{\mathcal{E}}$ and its stochastic gradients, whereas Assumption 8 gives growth conditions and functional inequalities $\bar{\mathcal{E}}$. Assumption 9 bounds the fluctuations of the stochastic gradient. As we show in the specific example of pairwise interacting system, these are implied by standard assumptions in the literature, and allow us to establish state-of-the-art convergence bounds.

2.5.2 Assumptions.

Assumption 1 (Lipschitz continuity)**.** *For some $L_u, L_{\bar{u}}, L_{\bar{\mathcal{F}}}, L_{\mathcal{F}} > 0$, the function $x \to u(x, y, \xi)$ and $y \to u(x, y, \xi)$ are L_u-Lipschitz. For every $x, y \in \mathbb{R}^d$, $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:*

$$(i)\ \ \|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \mu) - \nabla_{\mathcal{W}} \bar{\mathcal{F}}(y, \nu)\| \leq L_{\bar{\mathcal{F}}} \mathcal{W}_2(\mu, \nu) + L_{\bar{u}} \|x - y\|$$

$$(ii)\ \ \|\nabla_{\mathcal{W}} \mathcal{F}(x, \mu) - \nabla_{\mathcal{W}} \mathcal{F}(x, \nu)\| \leq L_{\mathcal{F}} \mathcal{W}_2(\mu, \nu) + L_u \|x - y\|$$

$$(iii)\ \ \|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \mu) - \nabla_{\mathcal{W}} \mathcal{F}(x, \mu)\| \leq L_l \mathcal{W}_2(\pi, \mu)$$

Assumption 2. *For some $C_{\bar{\mathcal{E}}}, C_{\mathsf{LSI}}, C_{\mathrm{KL}} > 0$, the functional $\bar{\mathcal{E}}$ satisfies the:*

$(i)\ \ \|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(x, \mu)\|^2_{L^2(\mu)} \geq C_{\bar{\mathcal{E}}} \bar{\mathcal{E}}(\mu)$ *for all* $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ (Polyak-Łojasiewicz inequality)

$(ii)\ \ \mathrm{KL}(\mu \,||\, \pi) \leq \dfrac{C_{\mathsf{LSI}}}{2} \mathsf{FD}(\mu \,||\, \pi)$ *for all* $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ (Log-Sobolev inequality)

$(iii)\ \ \mathrm{KL}(\mu||\pi) \leq C_{\mathrm{KL}} \bar{\mathcal{E}}(\mu)$ *for all* $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ (KL-Growth)

Assumption 3. *If $Y, \xi \sim \rho_0 \times \nu$, then from some $C^{\mathsf{Var}}, C^{\mathsf{Var}}_{\nu} > 0$, and for all $x \in \mathbb{R}^d$:*

$$\mathbb{E}\|u(x, Y, \xi) + \nabla_{\mathcal{W}} \mathcal{F}(x; \rho_0)\|^2 \leq C^{\mathsf{Var}} \mathsf{Var}(\rho_0) + C^{\mathsf{Var}}_{\nu}\,.$$

Before presenting the descent lemma, we introduce some technical lemmas that pertain to topics including coupling arguments, bounds on variance of certain random variables, and functional inequalities.

2.5.3 Technical lemmas. Recall that the random variable X_t is defined as $X_t := X_0 + tu(X_0, Y, \xi) + \sigma B_t$, where $u : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is the velocity field. Also, note that $\rho_t := \mathsf{Law}(X_t)$ and $\rho_t(|Y, \xi) := \mathsf{Law}(X_t|Y, \xi)$. Let $\mathsf{Var}(\mu)$ denote the trace of the covariance matrix of the probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathsf{Var}(u(x, Y, \xi))$ the variance of the velocity field u with respect to the random variables Y, ξ, for every fixed $x \in \mathbb{R}^d$. Next, $\mathcal{W}_p(\mu, \nu)$ denotes the p-Wasserstein distance between the probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\nabla_{\mathcal{W}} \mathcal{F}$ the Wasserstein gradient of the functional $\mathcal{F}$. Recall that, for a functional $\bar{\mathcal{F}}$, we define $\bar{\mathcal{E}}(\mu) := \bar{\mathcal{F}}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu) - \bar{\mathcal{F}}(\pi) - \frac{\sigma^2}{2}\mathcal{H}(\pi)$, where π is the minimizer of the functional $\bar{\mathcal{F}} + \frac{\sigma^2}{2}\mathcal{H}$. The functional $\bar{\mathcal{E}}$ need not be the same as the functional $\mathcal{E}$. Let $C_{\bar{\mathcal{E}}}, C_{\mathsf{LSI}}, C_{\mathsf{KL}}$ be the constants corresponding to (Polyak-Łojasiewicz inequality), (Log-Sobolev inequality), (KL-Growth) in

Assumption 8, respectively and $L_{\bar{u}}, L_{\bar{\mathcal{F}}}$ the constants defined in Assumption 7. Furthermore, $\|\cdot\|$ denotes the L^2 norm of a function.

The following lemma bounds the expected Wasserstein distance between $\rho_t, \rho_t(|Y,\xi)$.

Lemma 3. *1.*

$$\mathbb{E}\mathcal{W}_2^2(\rho_t, \rho_t(|Y,\xi)) \le 2t^2 \mathbb{E}_{x\sim\rho_0} \mathsf{Var}(u(x,Y,\xi))\,.$$

2.

$$\mathbb{E}\mathcal{W}_2^2(\rho_0, \rho_t(|Y,\xi)) \le \sigma^2 td + t^2 \mathbb{E}\|u(X_0,Y,\xi)\|^2\,.$$

Proof. Notice that $\rho_t(\cdot) = \mathbb{E}\rho_t(\cdot|Y,\xi)$. Since Wasserstein distance is convex in each of its coordinates and $x \to x^2$ is increasing and convex over $\mathbb{R}^+$, we have:

$$\mathbb{E}\mathcal{W}_2^2(\rho_t, \rho_t(|Y,\xi)) \le \mathbb{E}\mathcal{W}_2^2(\rho_t(|Y',\xi'), \rho_t(|Y,\xi))\,,$$

where (Y',ξ') is an independent copy of (Y,ξ). We now couple $\rho_t(|Y',\xi')$ and $\rho_t(|Y,\xi)$ for a given Y,ξ as follows:

$$X_t' = X_0 + tu(X_0,Y',\xi') + \sigma B_t\,; \qquad X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t\,.$$

Therefore,

$$\mathbb{E}\mathcal{W}_2^2(\rho_t, \rho_t(|Y,\xi)) \le t^2\mathbb{E}\|u(X_0,Y',\xi') - u(X_0,Y,\xi)\|^2 = 2t^2\mathbb{E}_{x\sim\rho_0}\mathsf{Var}(u(x,Y,\xi)) \tag{2.30}$$

Finally, we couple ρ_0 and $\rho_t(|Y,\xi)$ by $X_0 \sim \rho_0$ and $X_t = X_0 + tu(X_0,Y,\xi) + \sigma B_t$. This implies, $\mathbb{E}\mathcal{W}_2^2(\rho_0, \rho_t(|Y,\xi)) \le \sigma^2 td + t^2\mathbb{E}\|u(X_0,Y,\xi)\|^2$. □

Lemma 4 (Vempala and Wibisono (2019)[Lemma 11)**.** *] Let ν be a probability measure over $\mathbb{R}^d$ with density $\nu(x) \propto e^{-F(x)}$ where F is L-smooth. Then,*

$$\mathbb{E}_{x\sim\nu}\|\nabla F(x)\|^2 \le dL$$

Lemma 5 (Otto-Villani Theorem Otto and Villani (2000))**.** *Let π satisfy LSI with constant C_{LSI}. Then π satisfies Talangrand's inequality T_p for any $p \in [1,2]$, i.e., for all $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$:*

$$\mathcal{W}_p^2(\mu,\pi) \le 2C_{\mathsf{LSI}}\,\mathrm{KL}(\mu,\pi)\,.$$

Lemma 6. *For any $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, we have:*

1.

$$\mathsf{Var}(\mu) \le 2\mathcal{W}_2^2(\mu, \pi) + 2\mathsf{Var}(\pi)\,.$$

2. Suppose Assumption 7 holds. Let $X \sim \mu$ and $X^* \sim \pi$. Then:

$$\mathbb{E}\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(X, \mu)\|^2 \le 3(L_{\bar{u}}^2 + L_{\bar{\mathcal{F}}}^2)\mathcal{W}_2^2(\mu, \pi) + 3\mathbb{E}\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(X^*, \pi)\|^2\,.$$

Proof. Let X, X' be i.i.d from μ. Let Y, Y' be i.i.d. from π such that Y, X are optimally coupled and Y', X' are optimally coupled. Now consider:

$$\begin{aligned}
\mathsf{Var}(\mu) &= \frac{1}{2}\mathbb{E}\|X - Y + Y - Y' + Y' - X'\|^2 \\
&\le \mathbb{E}\|X - Y + Y' - X'\|^2 + \mathbb{E}\|Y - Y'\|^2 \\
&= \mathbb{E}\|X - Y + Y' - X'\|^2 + 2\mathsf{Var}(\pi) \\
&= 2\mathbb{E}\|X - Y\|^2 - 2\|\mathbb{E}X - \mathbb{E}Y\|^2 + 2\mathsf{Var}(\pi) \\
&\le 2\mathcal{W}_2^2(\mu, \pi) + 2\mathsf{Var}(\pi)\,,
\end{aligned}$$

where in the penultimate step we have used the fact that $X - Y, X' - Y'$ are i.i.d. Next, let $X^* \sim \pi$ be optimally coupled to $X \sim \mu$. By Assumption 7 and triangle inequality, we have:

$$\begin{aligned}
\|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(X, \mu)\| &= \|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(X, \mu) - \nabla_{\mathcal{W}} \bar{\mathcal{F}}(X^*, \pi) + \nabla_{\mathcal{W}} \bar{\mathcal{F}}(X^*, \pi)\| \\
&\le L_{\bar{\mathcal{F}}}\mathcal{W}_2(\mu, \pi) + L_{\bar{u}}\|X - X^*\| + \|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(X^*, \pi)\|\,.
\end{aligned}$$

By squaring, applying $(a + b + c)^2 \le 3(a^2 + b^2 + c^2)$, and taking expectation proves the second statement of this lemma. □

Lemma 7. $V(u, \rho_0) := \mathbb{E}_{x \sim \rho_0}[\mathsf{Var}(u(x, Y, \xi))]$. *Under Assumptions 8 and 9, we have:*

$$V(u, \rho_0) \lesssim C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}} \bar{\mathcal{E}}(\rho_0) + C^{\mathsf{Var}} \mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}\,, \tag{2.31}$$

$$\mathsf{Var}(\rho_0) \lesssim C_{\mathsf{LSI}} C_{\mathrm{KL}} \bar{\mathcal{E}}(\rho_0) + \mathsf{Var}(\pi)\,. \tag{2.32}$$

Proof. By Assumption 9, we have:

$$
\begin{aligned}
V(u,\rho_0) &\le C^{\mathsf{Var}}\mathsf{Var}(\rho_0) + C_\nu^{\mathsf{Var}} \\
&\lesssim C^{\mathsf{Var}}\mathcal{W}_2^2(\rho_0,\pi) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}} && \text{(By Lemma 6)} \\
&\lesssim C^{\mathsf{Var}}C_{\mathsf{LSI}}\,\mathrm{KL}(\rho_0||\pi) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}} && \text{(By Lemma 5)} \\
&\lesssim C^{\mathsf{Var}}C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}} && \text{(By Assumption 8-(KL-Growth))}
\end{aligned}
$$

Applying a similar reasoning to the bounds in Lemma 6, we conclude the bound on $\mathsf{Var}(\rho_0)$ □

Lemma 8. *Consider a probability measure π over $\mathbb{R}^d$, with density $\pi(x) \propto \exp(-F(x))$, which satisfies the LSI with constant C_{LSI}. Then, $\mathsf{Var}(\pi) \le dC_{\mathsf{LSI}}$.*

Proof. By Bakry, Gentil, and Ledoux (2013a)[Definition 4.2.1], the probability measure π satisfies the Poincaré inequality with constant $C_{\mathsf{PI}} > 0$ if

$$
\mathsf{Var}_\pi(f) \le C_{\mathsf{PI}}\mathbb{E}_\pi[\|\nabla f\|^2]\,,
$$

for all $f : \mathbb{R}^d \to \mathbb{R}$ such that f is continuously differentiable and ∇f is square integrable with respect to π, . For any $i \in \{1,2,\ldots,d\}$ and $x = (x_1,x_2,\ldots,x_d)$, let $f(x) = x_i$. Then $\|\nabla f(x)\| = 1$, for all $x \in \mathbb{R}^d$. Thus, by Poincaré inequality, $\mathsf{Var}_\pi(f(X)) = \mathsf{Var}(X_i) \le C_{\mathsf{PI}}$. Now, $\mathsf{Var}(\pi) = \sum_{i=1}^d \mathsf{Var}(X_i) \le dC_{\mathsf{PI}}$. Finally, since π satisfying the LSI implies that it also satisfies the Poincaré inequality with the same constant C_{LSI}, we conclude that $\mathsf{Var}(\pi) \le dC_{\mathsf{LSI}}$. □

Lemma 9. *Consider a probability measure π over $\mathbb{R}^d$ with density $\pi(x) \propto \exp(-F(x))$ satisfies the Logarithmic Sobolev Inequality with constant C_{LSI}. Assume that F is L-smooth (i.e, ∇F is L-Lipschitz) and ∇F is square integrable with respect to π. Then,*

$$
C_{\mathsf{LSI}} \ge \frac{1}{L}\,.
$$

Proof. Let $x_\pi \in \mathbb{R}^d$ be the mean of π (a fixed vector). Without loss of generality, we assume $\pi(x) = e^{-F(x)}$. This requires adding a constant to $F(x)$ in the original definition, which does not

change ∇F. Using integration by parts, we have:

$$\begin{aligned} d &= \int \langle x - x_\pi, \nabla F(x) \rangle e^{-F(x)} dx \\ &\leq \sqrt{\mathsf{Var}(\pi)} \sqrt{\mathbb{E}_\pi \|\nabla F\|^2} && \text{(Cauchy-Schwarz Inequality)} \\ &\leq \sqrt{d C_{\mathsf{LSI}}} \sqrt{dL} && \text{(Lemma 8 and Lemma 4)} \end{aligned}$$

The claim follows from the above equation. □

With the assumptions and technical lemmas in place, we are now in a position to discuss the descent lemma.

2.5.4 Descent lemma.

Lemma 10 (Descent lemma)**.** *Suppose that Assumptions 7, 8 and 9 are satisfied.*

$$(\sigma^*)^2 := \mathbb{E}_{x\sim\pi} \mathsf{Var}(u(x, Y^*, \xi)) \,,\; (G_\pi)^2 := \mathbb{E}_{x\sim\pi} \|\nabla_{\mathcal{W}} \mathcal{F}(x, \pi)\|^2 \,,$$

$$(G_{\mathsf{mod}})^2 := \mathbb{E}_{x\sim\pi} \|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \pi)\|^2 \,.$$

Assume that the following inequalities are satisfied for some small enough $c_0 > 0$:

1. $C_{\bar{\mathcal{E}}} \geq 8 L_l^2 C_{\mathrm{KL}} C_{\mathsf{LSI}}$
2. $\eta < c_0 \min\left(\frac{1}{L_u}, \frac{1}{C_{\bar{\mathcal{E}}}}, \frac{C_{\bar{\mathcal{E}}}}{C_{\mathsf{LSI}} C_{\mathrm{KL}} L_u^2 (L_{\bar{u}} + L_{\bar{\mathcal{F}}})}, \frac{1}{L_u (L_{\bar{u}} + L_{\bar{\mathcal{F}}})} \sqrt{\frac{C_{\bar{\mathcal{E}}}}{C_{\mathsf{LSI}} C_{\mathrm{KL}}}}, \frac{C_{\bar{\mathcal{E}}}}{C^{\mathsf{Var}} C_{\mathrm{KL}} C_{\mathsf{LSI}} (L_{\bar{u}} + L_{\bar{\mathcal{F}}})} \right)$

Then, for some universal constant $C > 0$:

$$\mathbb{E}\bar{\mathcal{E}}(\rho_\eta) \leq e^{-\frac{\eta C_{\bar{\mathcal{E}}}}{8}} \mathbb{E}\bar{\mathcal{E}}(\rho_0) + C \left[\gamma_3 \eta^3 + \gamma_2 \eta^2 + \gamma_1 \eta^{\frac{3}{2}} \right] ,$$

where $\gamma_3 := (L_u^2 + L_{\mathcal{F}}^2)((\sigma^*)^2 + (G_\pi)^2) + \frac{L_u^2 G_{\mathsf{mod}}^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{C_{\bar{\mathcal{E}}}}$, $\gamma_2 := (L_u^2 + L_{\mathcal{F}}^2)\sigma^2 d + L_u G_{\mathsf{mod}} \sqrt{C^{\mathsf{Var}} \mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}} + (L_u + L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}} \mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}) + \frac{\sigma^2 L_u^2 d^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{C_{\bar{\mathcal{E}}}}$ *and* $\gamma_1 := \sigma d L_u \sqrt{C^{\mathsf{Var}} \mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}}$.

Before delving into the proof of Lemma 10, we note the following bound on the stochastic error $\mathbb{E}[\mathsf{SE}(t)]$ using Lemma 2 and the previously stated assumptions.

Lemma 11 (Stochastic error bound)**.** *Define* $V(u, \rho_0) := \mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x, Y, \xi))]$ *and* $G_{\mathsf{mod}}^2 := \mathbb{E}_{X^*\sim\pi} \|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(X^*, \pi)\|^2$*. Let Assumptions 7, 8 and 9 hold. Assume* $L_u t \leq 1$*. Then, for arbitrary* $\beta > 0$*, we have:*

$$\mathbb{E}[\mathsf{SE}(t)] \lesssim (\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\sqrt{(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}})} + t(L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}})$$
$$+ \frac{(\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{\beta} + (tC^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}} + \beta)\bar{\mathcal{E}}(\rho_0)$$
$$+ tL_u^2(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(|Y,\xi))\,.$$

Proof. We define the following to reflect the result of Lemma 2:

$$\Theta(x,y,\xi) := u(x,y,\xi) + \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)$$
$$T_4 := 2\sqrt{\mathbb{E}[\mathcal{W}_2^2(\rho_t,\rho_t(\cdot|Y,\xi))]}\left[(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\sqrt{\mathbb{E}\left\|\Theta(X_t,Y,\xi)\right\|^2} + L_u\sqrt{\mathbb{E}\left\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(X_t,\rho_t)\right\|^2}\right]$$
$$T_5 := \sigma\sqrt{2t}L_u d\sqrt{\mathbb{E}_{x\sim\rho_0}[\mathsf{Var}(u(x,Y,\xi)]}\,.$$

From Lemma 2, we have $\mathbb{E}[\mathsf{SE}(t)] \le T_4 + T_5$. We now bound each term of $T_4(t)$. By Lemma 3, we have:

$$\sqrt{\mathbb{E}\mathcal{W}_2^2(\rho_t,\rho_t(\cdot|Y,\xi))} \le \sqrt{2t^2V(u,\rho_0)}\,. \tag{2.33}$$

For the second term, given Y,ξ, let $Z_1 \sim \rho_t(\cdot|Y,\xi)$ and $Z_2 \sim \rho_t$ be optimally coupled in the 2-Wasserstein distance. By Assumption 7, we obtain:

$$\begin{aligned}\left\|\Theta(Z_1,Y,\xi)\right\| &\le \left\|\Theta(Z_1,Y,\xi) - \Theta(Z_2,y,\xi)\right\| + \left\|\Theta(Z_2,Y,\xi)\right\| \\ &\le 2L_u\|Z_1 - Z_2\| + \left\|\Theta(Z_2,Y,\xi)\right\|.\end{aligned}$$

Next, by squaring, applying the inequality $(a+b)^2 \le 2a^2 + 2b^2$, taking expectation, applying Lemma 3, we get almost surely Y,ξ:

$$\mathbb{E}[\left\|\Theta(Z_1,Y,\xi)\right\|^2|Y,\xi] \le 8L_u^2\mathcal{W}_2^2(\rho_0,\rho_t(\cdot|Y,\xi)) + 2\mathbb{E}[\left\|\Theta(Z_2,Y,\xi)\right\|^2|Y,\xi]\,.$$

Applying Lemma 3, Lemma 7 and Assumption 9, after noting that $X_t|Y,\xi \stackrel{d}{=} Z_1$ we obtain:

$$\begin{aligned}\mathbb{E}\left\|\Theta(X_t,Y,\xi)\right\|^2 &\leq 16L_u^2t^2V(u,\rho_0) + 2\mathbb{E}[\mathbb{E}[\left\|\Theta(Z_2,Y,\xi)\right\|^2|Y,\xi]] \\ &\leq 16L_u^2t^2V(u,\rho_0) + 2C^{\mathsf{Var}}\mathsf{Var}(\rho_0) + 2C_\nu^{\mathsf{Var}} \qquad \text{(By Assumption 9)} \\ &\lesssim C^{\mathsf{Var}}C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}\,. \qquad (2.34)\end{aligned}$$

Here we have used the assumption that $L_ut \leq 1$. The last step above follows by an application of Lemma 7. Now, additionally consider Assumptions 8 and 9. We apply Lemma 6, to conclude:

$$\begin{aligned}\mathbb{E}_{x\sim\rho_t}\left\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\right\|^2 &\leq 3(L_{\bar{\mathcal{F}}}^2 + L_{\bar{u}}^2)\mathcal{W}_2^2(\rho_t,\pi) + 3\mathbb{E}_{x\sim\pi}\left\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\pi)\right\|^2. \\ &\leq 3(L_{\bar{\mathcal{F}}}^2 + L_{\bar{u}}^2)\mathbb{E}\mathcal{W}_2^2(\rho_t(|Y,\xi),\pi) + 3\mathbb{E}_{x\sim\pi}\left\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\pi)\right\|^2.\end{aligned}$$

The last step follows from the usual convexity of $\mathcal{W}_2^2$. By applying the Talagrand's T_2-inequality implied by Lemma 5 and Assumption 8-(Log-Sobolev inequality) along with Assumption 8-(KL-Growth), we get:

$$\mathbb{E}_{x\sim\rho_t}\left\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\rho_t)\right\|^2 \lesssim (L_{\bar{\mathcal{F}}}^2 + L_{\bar{u}}^2)C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(|Y,\xi)) + \mathbb{E}_{x\sim\pi}\left\|\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\pi)\right\|^2. \qquad (2.35)$$

Let $\zeta(\rho_0) := \sqrt{C^{\mathsf{Var}}C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0) + C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}}$. We now use equations (2.35), (2.34) and (2.33), along with Lemma 7 to bound $V(u,\rho_0)$ and hence $\mathbb{E}[\mathsf{SE}(t)]$ as:

$$\begin{aligned}\mathbb{E}[\mathsf{SE}(t)] &\lesssim (\sigma L_u d\sqrt{t} + tL_uG_{\mathsf{mod}})\zeta(\rho_0) + t(L_{\bar{u}} + L_{\bar{F}})\zeta^2(\rho_0) \\ &\quad + tL_u(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\zeta(\rho_0)\sqrt{C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(|Y,\xi))}\,.\end{aligned}$$

Applying AM-GM inequality on $tL_u(L_{\bar{u}} + L_{\bar{\mathcal{F}}})\zeta(\rho_0)\sqrt{C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(|Y,\xi))}$, we conclude:

$$\begin{aligned}\mathbb{E}[\mathsf{SE}(t)] &\lesssim (\sigma L_u d\sqrt{t} + tL_uG_{\mathsf{mod}})\zeta(\rho_0) + t(L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}) \\ &\quad + tL_u^2(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}\mathbb{E}\bar{\mathcal{E}}(\rho_t(|Y,\xi)) + tC^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0)\,.\end{aligned}$$

Now, using the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for every $x, y \geq 0$ on $\zeta(\rho_0)$, we conclude:

$$\begin{aligned}(\sigma L_u d\sqrt{t} + tL_uG_{\mathsf{mod}})\zeta(\rho_0) &\leq (\sigma L_u d\sqrt{t} + tL_uG_{\mathsf{mod}})\sqrt{(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}})} \\ &\quad + (\sigma L_u d\sqrt{t} + tL_uG_{\mathsf{mod}})\sqrt{C^{\mathsf{Var}}C_{\mathsf{LSI}}C_{\mathrm{KL}}\bar{\mathcal{E}}(\rho_0)}\,.\end{aligned}$$

Plugging this into the bound for $\mathbb{E}\mathsf{SE}(t)$ above and applying the AM-GM inequality on $(\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\sqrt{C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}} \bar{\mathcal{E}}(\rho_0)}$, we conclude the result. □

Proof of Lemma 10. Let $t \in [0, \eta]$. First, we consider the error terms $\mathsf{DE}_1, \mathsf{DE}_2, \mathsf{SE}, \mathsf{LE}$ in (2.9) and obtain:

$$\begin{aligned}\frac{d}{dt}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) &= -\int dx \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 + \mathsf{DE}_1(t) + \mathsf{DE}_2(t) + \mathsf{SE}(t) + \mathsf{LE}(t) \\ &= -\frac{1}{2}\int dx \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 + \mathsf{res}(t)\,, \end{aligned} \tag{2.36}$$

where $\mathsf{res}(t) := \mathsf{DE}_1(t) + \mathsf{DE}_2(t) + \mathsf{SE}(t) + \mathsf{LE}(t) - \frac{1}{2}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx$. We begin by bounding the linearization error $\mathsf{LE}(t)$ using the assumptions stated in this lemma. From (2.29), the observation that π satisfies Talagrand's T_2-inequality by Lemma 5, Assumption 8-(KL-Growth), and the inequality $ab \le \frac{a^2}{4} + b^2$ for all $a, b \in \mathbb{R}$, we get:

$$\begin{aligned}\mathsf{LE}(t) &\le L_l\sqrt{2C_{\mathsf{LSI}}}\sqrt{\mathrm{KL}(\pi,\rho_t(\cdot|Y,\xi))}\sqrt{\mathbb{E}_{\rho_t(\cdot|Y,\xi)}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2} \\ &\le L_l\sqrt{2C_{\mathrm{KL}}C_{\mathsf{LSI}}}\sqrt{\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))}\sqrt{\mathbb{E}_{\rho_t(\cdot|Y,\xi)}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2} \\ &\le 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \frac{1}{4}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx \\ &\le 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \frac{1}{4}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx\,. \end{aligned}$$

By plugging the above inequality into (2.36) and applying Assumption 8-(Polyak-Łojasiewicz inequality) with $\mu = \rho_t(\cdot|Y,\xi)$, we get:

$$\begin{aligned}\frac{d}{dt}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) &\le -\frac{1}{2}\int dx \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 + 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_1(t) \\ &\le -\frac{C_{\bar{\mathcal{E}}}}{2}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_1(t) \\ &\le -C'\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_1(t)\,, \end{aligned} \tag{2.37}$$

where

$$\mathsf{res}_1(t) := \mathsf{DE}_1(t) + \mathsf{DE}_2(t) + \mathsf{SE}(t) - \frac{1}{4}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx\,,\ C' := \frac{C_{\bar{\mathcal{E}}}}{2} - 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}}\,.$$

From Equation (2.37) we get:

$$\frac{d}{dt}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) \le -C'\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{SE}(t) + \mathsf{res}_2(t)\,, \tag{2.38}$$

where

$$\mathsf{res}_2(t) := \mathsf{DE}_1(t) + \mathsf{DE}_2(t) - \frac{1}{4}\int \rho_t(x|Y,\xi)\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\rho_t(\cdot|Y,\xi))\|^2 dx\,.$$

The next step is to bound $\mathbb{E}[\mathsf{SE}(t)]$. To simplify notation, we define:

$$\begin{aligned}
A_1^{\mathsf{SE}}(t,\beta) &:= (\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})\sqrt{(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}})} + t(L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}(\pi) + C_\nu^{\mathsf{Var}}) \\
&\quad + \frac{(\sigma L_u d\sqrt{t} + tL_u G_{\mathsf{mod}})^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{\beta} + (tC^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}} + \beta)\bar{\mathcal{E}}(\rho_0) \\
A_2^{\mathsf{SE}}(t) &:= tL_u^2(L_{\bar{u}} + L_{\bar{\mathcal{F}}})C_{\mathsf{LSI}}C_{\mathrm{KL}}\,.
\end{aligned}$$

By Lemma 11, we conclude that for arbitrary $\beta > 0$, we have:

$$\mathbb{E}[\mathsf{SE}(t)] \lesssim A_1^{\mathsf{SE}}(t,\beta) + A_2^{\mathsf{SE}}(t)\mathbb{E}\bar{\mathcal{E}}(\rho_t(|Y,\xi))\,, \tag{2.39}$$

Using this notation in Equation (2.38), we obtain:

$$\begin{aligned}
\frac{d}{dt}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) &\le (-C' + A_2^{\mathsf{SE}}(t))\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{SE}(t) - A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_2(t) \\
&\le -C''(\eta)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{SE}(t) - A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_2(t)\,,
\end{aligned} \tag{2.40}$$

where $C''(\eta) := \frac{C_{\bar{\mathcal{E}}}}{2} - 2L_l^2 C_{\mathrm{KL}}C_{\mathsf{LSI}} - A_2^{\mathsf{SE}}(\eta)$. Note that by our assumptions on η and the parameter L_l in the statement of this lemma, for a small enough c_0, we must have $C''(\eta) \ge \frac{C_{\bar{\mathcal{E}}}}{4}$.

Next, by multiplying both sides of (2.40) by $e^{C''(\eta)t}$, we get:

$$\frac{d}{dt}\left[e^{C''(\eta)t}\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi))\right] \le e^{C''(\eta)t}\left[\mathsf{SE}(t) - A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_2(t)\right]\,.$$

Thus, by integrating on both sides and taking expectation (along with Fubini's theorem), we obtain:

$$\mathbb{E}\bar{\mathcal{E}}(\rho_\eta(\cdot|Y,\xi)) \le e^{-C''(\eta)\eta}\bar{\mathcal{E}}(\rho_0) + e^{-C''(\eta)\eta}\int_0^\eta \mathbb{E}\left[\mathsf{SE}(t) - A_2^{\mathsf{SE}}(t)\bar{\mathcal{E}}(\rho_t(\cdot|Y,\xi)) + \mathsf{res}_2(t)\right]e^{C''(\eta)t}dt\,. \tag{2.41}$$

To simplify notation in the subsequent steps of the proof, we define:

$$\begin{aligned}
T_1(t) &:= \sqrt{\mathbb{E}\|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(X_t,\rho_t(\cdot|Y,\xi))\|^2} \\
A(t) &:= L_u\sqrt{2t^2\left(2L_u^2\mathcal{W}_2^2(\rho_0,\pi) + (\sigma^*)^2 + (G_\pi)^2\right) + \sigma^2 td} \\
B(t) &:= L_{\mathcal{F}}\sqrt{2t^2\left(2L_u^2\mathcal{W}_2^2(\rho_0,\pi) + (\sigma^*)^2 + (G_\pi)^2\right) + \sigma^2 td}\,.
\end{aligned}$$

Next, by using the inequality $ab \leq \frac{a^2}{8} + 2b^2$ for all $a, b \in \mathbb{R}$, the bounds for $\mathsf{DE}_1(t) + \mathsf{DE}_2(t)$ using Lemma 2, we get:

$$\begin{aligned}
e^{-C''(\eta)\eta} \int_0^\eta \mathbb{E}\left[\mathsf{res}_2(t)\right] e^{C''(\eta)t} dt &= e^{-C''(\eta)\eta} \int_0^\eta e^{C''(\eta)t} \left[\mathbb{E}[\mathsf{DE}_1(t) + \mathsf{DE}_2(t)] - \tfrac{1}{4}\mathbb{E}T_1(t)^2\right] dt \\
&\leq e^{-C''(\eta)\eta} \int_0^\eta e^{C''(\eta)t} \left[\mathbb{E}T_1(t)(A(t) + B(t)) - \frac{1}{4}\mathbb{E}T_1(t)^2\right] dt \\
&\leq 2e^{-C''(\eta)\eta} \int_0^\eta e^{C''(\eta)t} \mathbb{E}(A(t)^2 + B(t)^2) dt \\
&\leq 2 \int_0^\eta \left[\mathbb{E}A(t)^2 + \mathbb{E}B(t)^2\right] dt \,. \qquad (2.42)
\end{aligned}$$

Next, we bound $\int_0^\eta \left[A(t)^2 + B(t)^2\right] dt$. To further simplify notation, we define:

$$T_2 := 2\left(2L_u^2 \mathcal{W}_2^2(\rho_0, \pi) + (\sigma^*)^2 + (G_\pi)^2\right) \,,\; T_3 := \sigma^2 d \,.$$

Therefore, integration yields:

$$\begin{aligned}
e^{-C''(\eta)\eta} \int_0^\eta \mathbb{E}\left[\mathsf{res}_2(t)\right] e^{C''(\eta)t} dt &\leq 2 \int_0^\eta \left[A(t)^2 + B(t)^2\right] dt \\
&\leq 2(L_u^2 + L_{\mathcal{F}}^2) \int_0^\eta (T_2 t^2 + T_3 t) dt \\
&= 2(L_u^2 + L_{\mathcal{F}}^2)(\tfrac{T_2\eta^3}{3} + \tfrac{T_3\eta^2}{2}) \\
&\lesssim (L_u^2 + L_{\mathcal{F}}^2)\left[\eta^3\left[L_u^2 \mathcal{W}_2^2(\rho_0, \pi) + (\sigma^*)^2 + (G_\pi)^2\right] + \sigma^2 d\eta^2\right] \,.
\end{aligned}$$

Now, we apply Assumptions 8-(KL-Growth) and Assumption 8-(Log-Sobolev inequality) along with Lemma 5 to upper bound $\mathcal{W}_2^2(\rho_0, \pi)$ in the equation above to conclude:

$$e^{-C''(\eta)\eta} \int_0^\eta \mathbb{E}\left[\mathsf{res}_2(t)\right] e^{C''(\eta)t} dt \lesssim (L_u^2 + L_{\mathcal{F}}^2)\left[\eta^3\left[L_u^2 C_{\mathsf{LSI}} C_{\mathrm{KL}} \bar{\mathcal{E}}(\rho_0) + (\sigma^*)^2 + (G_\pi)^2\right] + \sigma^2 d\eta^2\right] \,. \qquad (2.43)$$

Now we consider:

$$\begin{aligned}
& e^{-C''(\eta)\eta} \int_0^\eta e^{C''(\eta)t} \mathbb{E}[\mathsf{SE}(t) - A_2^{\mathsf{SE}}(t) \bar{\mathcal{E}}(\rho_t(\cdot|Y, \xi))] dt \\
& \lesssim e^{-C''(\eta)\eta} \int_0^\eta e^{C''(\eta)t} A_1^{\mathsf{SE}}(t, \beta) dt \qquad \text{(From Equation (2.39))} \\
& \leq \eta A_1^{\mathsf{SE}}(\eta, \beta) \,,
\end{aligned}$$

where we recall

$$A_1^{\mathsf{SE}}(\eta,\beta) = (\sigma L_u d\sqrt{\eta} + \eta L_u G_{\mathsf{mod}})\sqrt{(C^{\mathsf{Var}}\mathsf{Var}\pi + C_\nu^{\mathsf{Var}})} + \eta(L_{\bar{u}} + L_{\bar{\mathcal{F}}})(C^{\mathsf{Var}}\mathsf{Var}\pi + C_\nu^{\mathsf{Var}}) + \frac{(\sigma L_u d\sqrt{\eta} + \eta L_u G_{\mathsf{mod}})^2 C^{\mathsf{Var}} C_{\mathsf{LSI}} C_{\mathrm{KL}}}{\beta} + (\eta C^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}}) C_{\mathsf{LSI}} C_{\mathrm{KL}} + \beta)\bar{\mathcal{E}}(\rho_0)\,. \tag{2.44}$$

First, note that since for every $x \geq 0$, we have $1 - x \leq e^{-x} \leq 1 - x + \frac{x^2}{2}$ and as demonstrated above $C^{''}(\eta) \geq \frac{C_{\bar{\mathcal{E}}}}{4}$. With $\beta = c_0 C_{\bar{\mathcal{E}}}$ for some small enough constant c_0 and letting η small enough as noted in the statement of this lemma, we conclude:

$$e^{-C^{''}(\eta)\eta} + (L_u^2 + L_{\mathcal{F}}^2)\eta^3 L_u^2 C_{\mathsf{LSI}} C_{\mathrm{KL}} + (\eta^2 C^{\mathsf{Var}}(L_{\bar{u}} + L_{\bar{\mathcal{F}}}) C_{\mathsf{LSI}} C_{\mathrm{KL}} + \eta\beta) \leq e^{-\frac{\eta C_{\bar{\mathcal{E}}}}{8}}\,.$$

We now plug (2.43) and (2.44) along with the above bound into (2.41) to conclude:

$$\mathbb{E}\bar{\mathcal{E}}(\rho_\eta(\cdot|Y,\xi)) \leq e^{-\frac{\eta C_{\bar{\mathcal{E}}}}{8}}\bar{\mathcal{E}}(\rho_0) + C\gamma_3\eta^3 + C\gamma_2\eta^2 + C\gamma_1\eta^{\frac{3}{2}}\,.$$

□

Using the recursion established in Lemma 10 allows us to prove our main result.

2.5.5 Main theorem.

Theorem 1. *Consider Algorithm 1. Let π be the unique minimizer of $\bar{\mathcal{E}}$. Suppose Assumptions 7, 8, 9 hold with $G_\pi = u$. With $\gamma_1, \gamma_2, \gamma_3$ as defined in Lemma 10, the following holds:*

1. *Conditioned on $\mathcal{R}_{T-1}$, $X_T^{(1)}, \ldots, X_T^{(n)} \overset{i.i.d.}{\sim} \mu_T|\mathcal{R}_{T-1}$.*
2. $\mathbb{E}\bar{\mathcal{E}}(\mu_T|\mathcal{R}_{T-1}) \leq e^{-\frac{\eta C_{\bar{\mathcal{E}}} T}{8}}\bar{\mathcal{E}}(\mu_0) + C\left[\frac{\gamma_3\eta^2}{C_{\bar{\mathcal{E}}}} + \frac{\gamma_2\eta}{C_{\bar{\mathcal{E}}}} + \frac{\gamma_1\sqrt{\eta}}{C_{\bar{\mathcal{E}}}}\right]$.

We now turn to the pairwise interaction energy example discussed in the previous section.

2.5.6 Applications. *Let $V, W : \mathbb{R}^d \to \mathbb{R}$ and $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$. Let W be even (i.e, $W(x) = W(-x)$). Recall that the definition of the functional $\mathcal{F}$, its Wasserstein gradient, and its unique minimizer are in* (2.4), (2.5), (2.6) *respectively. We call $f : \mathbb{R}^d \to \mathbb{R}$ to be L-smooth if $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for every $x, y \in \mathbb{R}^d$.*

Assumption 4 (Smoothness)**.** *V is L_V smooth and W is L_W smooth.*

Assumption 5 (LSI). *π satisfies LSI with constant C_{LSI}, i.e., for all $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$:*

$$\mathrm{KL}(\mu \,||\, \pi) \leq \frac{C_{\mathsf{LSI}}}{2} \mathsf{FD}(\mu \,||\, \pi)\,.$$

The assumption $L_W \leq \frac{\sigma^2}{\sqrt{24} C_{\mathsf{LSI}}}$ is called " weak interaction" in Kook et al. (2024). Our assumption below is less restrictive in terms of multiplicative constants.

Assumption 6 (Weak Interaction). $L_W \leq \frac{\sigma^2}{4 C_{\mathsf{LSI}}}$.

We define the velocity field $\hat{G}(x,y) := -\nabla V(x) - \nabla W(x-y)\,,\ \forall\ x \in \mathbb{R}^d$ and $\bar{\mathcal{E}}(\mu) = \frac{\sigma^2}{2}\,\mathrm{KL}(\mu||\pi)$ (which corresponds to picking $\bar{\mathcal{F}}(\mu) = \int V(x) d\mu(x) + \int W(x-y) d\pi(y) d\mu(x)$), where π is the minimizer of Equation (2.4)*. The following Lemma establishes the general Assumptions required for Theorem 1 using the Assumptions 4 and 5.*

Lemma 12. *Under Assumptions 4 and 5, the general Assumption 7 is satisfied with $L_u = L_{\bar{u}} = L_V + L_W$, $L_{\mathcal{F}} = L_W$, $L_{\bar{\mathcal{F}}} = 0$ and $L_l = L_W$. The Assumption 8 is satisfied with $C_{\mathsf{LSI}} = C_{\mathsf{LSI}}$, $C_{\bar{\mathcal{E}}} = \frac{\sigma^2}{C_{\mathsf{LSI}}}$ and $C_{\mathrm{KL}} = \frac{2}{\sigma^2}$. The Assumption 9 is satisfied with $C^{\mathsf{Var}} = 2L_W^2$ and $C_\nu^{\mathsf{Var}} = 0$.*

Proof. Note that $\hat{G}$ satisfies the Lipschitz continuity property of Assumption 7 with $L_u = L_V + L_W$ because for every $x_1, x_2, y_1, y_2 \in \mathbb{R}^d$:

$$\|\hat{G}(x_1,y_1) - \hat{G}(x_2,y_2)\| \leq L_V \|x_1 - x_2\| + L_W(\|x_1 - x_2\| + \|y_1 - y_2\|)\,.$$

Next, the functional $\mathcal{F}$ satisfies the Lipschitz continuity property of Assumption 7, with $L_u = L_V + L_W$ and $L_{\mathcal{F}} = L_W$, because, if $Z_1 \sim \mu$ and $Z_2 \sim \nu$ are optimally coupled in the 1-

Wasserstein distance, then by definition of $\nabla_{\mathcal{W}}\mathcal{F}$ in (2.5), for every $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$
\begin{aligned}
&\|\nabla_{\mathcal{W}}\mathcal{F}(x,\mu) - \nabla_{\mathcal{W}}\mathcal{F}(y,\nu)\| \\
&\leq \|\nabla V(x) - \nabla V(y)\| + \|\int \nabla W(x-z)\mu(dz) - \int \nabla W(y-z)\nu(dz)\| \\
&\leq L_V\|x-y\| \\
&+ \|\int \nabla W(x-z)\mu(dz) - \int \nabla W(y-z)\mu(dz) + \int \nabla W(y-z)\mu(dz) - \int \nabla W(y-z)\nu(dz)\| \\
&= (L_V + L_W)\|x-y\| + \mathbb{E}_{(Z_1,Z_2)\sim\mu\times\nu}\|\nabla W(y-Z_1) - \nabla W(y-Z_2)\| \\
&= (L_V + L_W)\|x-y\| + L_W\mathcal{W}_1(\mu,\nu) \\
&\leq (L_V + L_W)\|x-y\| + L_W\mathcal{W}_2(\mu,\nu)\,.
\end{aligned}
$$

Similarly, we consider $\nabla_{\mathcal{W}}\bar{\mathcal{F}}(x,\mu) = \nabla V(x) + \nabla W * \pi(x)$ and conclude $L_{\bar{u}} = L_V + L_W$, $L_{\bar{\mathcal{F}}} = 0$ and $L_l = L_W$. Clearly, Assumption 8-(Log-Sobolev inequality) is satisfied with constant C_{LSI} since it is the same as Assumption 5. Notice that $\bar{\mathcal{E}}(\mu) := \frac{\sigma^2}{2}\,\mathrm{KL}(\mu||\pi)$, and that $\nabla_{\mathcal{W}}\,\mathrm{KL}(\mu||\pi) = \mathsf{FD}(\mu||\pi)$, we conclude that Assumption 5 implies Assumption 8-(Polyak-Łojasiewicz inequality) with $C_{\bar{\mathcal{E}}} = \frac{\sigma^2}{C_{\mathsf{LSI}}}$. The choice of $\bar{\mathcal{E}}$ also implies Assumption 8-(KL-Growth) with $C_{\mathrm{KL}} = \frac{2}{\sigma^2}$.

Now consider Assumption 9. Note that by Jensen's inequality:

$$
\begin{aligned}
\mathbb{E}_{Y\sim\rho_0}\|u(x,Y) + \nabla_{\mathcal{W}}\mathcal{F}(x;\rho_0)\|^2 &= \mathbb{E}_{Y\sim\rho_0}\|\nabla W(x-Y) - \nabla W * \rho_0(x)\|^2 \\
&\leq \mathbb{E}_{Y\sim\rho_0}\int \|\nabla W(x-Y) - \nabla W(x-z)\|^2\rho_0(dz) \\
&= L_W^2\mathbb{E}_{(Y,Z)\sim\rho_0\times\rho_0}\|Y-Z\|^2 \\
&= 2L_W^2\mathsf{Var}(\rho_0)\,.
\end{aligned}
$$

Hence, Assumption 9 is satisfied with $C^{\mathsf{Var}} = 2L_W^2$ and $C_\nu^{\mathsf{Var}} = 0$. □

The following theorem instantiates Theorem 1 to the pairwise interaction energy case.

Theorem 2. *Consider the Pairwise Interaction Energy in Equation* (2.4) *under Assumptions 4, 5 and 6. There exist universal constants $c_0, C > 0$ such that Algorithm 1 with $\hat{G}$ as above with $\eta < c_0 \min\left(\frac{C_{\mathsf{LSI}}}{\sigma^2}, \frac{\sigma^4}{C_{\mathsf{LSI}}^2(L_V+L_W)^3}\right)$ satisfies:*

$$\mathbb{E}\left[\mathrm{KL}(\mu_T|\mathcal{R}_{T-1}||\pi)\right] \leq e^{\left(-\frac{T\sigma^2\eta}{8C_{\mathsf{LSI}}}\right)}\,\mathrm{KL}(\mu_0||\pi) + C\frac{\sqrt{\eta}d^{3/2}(C_{\mathsf{LSI}})^{1/2}\sigma(L_V+L_W)}{4} \tag{2.45}$$
$$+ C\eta(L_V+L_W)^2d^2C_{\mathsf{LSI}}\,.$$

Proof. Using the notation in Theorem 1, we conclude that $\bar{\mathcal{F}}(x,\pi) = \mathcal{F}(x,\pi)$ for every $x \in \mathbb{R}^d$ and hence $G_\pi = G_{\mathsf{mod}}$. Since $\pi \propto \exp(-\frac{2\delta\mathcal{F}(x,\pi)}{\sigma^2})$ and $\nabla_{\mathcal{W}}\mathcal{F}(x,\pi) = \nabla_x\delta\mathcal{F}(x,\pi)$, we apply Lemma 4 to conclude that $(G_\pi)^2 := \mathbb{E}_{x\sim\pi}\|\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)\|^2 \leq \frac{d\sigma^2(L_V+L_W)}{2}$. Applying Lemma 8, we conclude $\mathsf{Var}(\pi) \leq C_{\mathsf{LSI}}d$. Using Assumption 9 instantiated to our case, we conclude $(\sigma^*)^2 \leq 2L_W^2C_{\mathsf{LSI}}d$. Under the assumption on η, all the requisite assumptions for Theorem 1 are satisfied (after simplifying with the fact that $C_{\mathsf{LSI}} \geq \frac{\sigma^2}{2(L_V+L_W)}$ from Lemma 9). Thus, we note that $\gamma_1, \gamma_2, \gamma_3$ in Theorem 1 can be instantiated as

$$\gamma_1 \lesssim \sigma d^{\frac{3}{2}}\sqrt{C_{\mathsf{LSI}}}L_W(L_V+L_W)\,;\,\gamma_2 \lesssim (L_V+L_W)^2\sigma^2d^2\,;\,\gamma_3 \lesssim (L_V+L_W)^3d\sigma^2\,.$$

Here, we have Assumption 6 that $L_W \leq \frac{\sigma^2}{4C_{\mathsf{LSI}}}$ to simplify the expressions. Invoking Theorem 1, using the fact that $\eta(L_V+L_W) < c_0$ by assumption in the statement of this theorem, and $C_{\mathsf{LSI}} \geq \frac{\sigma^2}{2(L_V+L_W)}$ from Lemma 9, we conclude the result.

□

The following remark states the number of timesteps T needed for $\mathrm{KL}(\mu_T|\mathcal{R}_{T-1})$ to converge in expectation.

Remark 2. *Given $\epsilon/3 \in (0, \mathrm{KL}(\mu_0||\pi) \wedge 1)$, as per Theorem 2, we can achieve $\mathbb{E}\left[\mathrm{KL}(\mu_T|\mathcal{R}_{T-1}) - \mathrm{KL}(\pi)\right] \leq \epsilon$ by picking:*

1. $\eta = \frac{8C_{\mathsf{LSI}}}{\sigma^2T}\log\left(\frac{3\,\mathrm{KL}(\mu_0||\pi)}{\epsilon}\right)$
2. $T \gtrsim \max\left(\frac{C_{\mathsf{LSI}}^2d^3(L_V+L_W)^2}{\epsilon^2}, \frac{C_{\mathsf{LSI}}^2d^2(L_V+L_W)^2}{\sigma^2\epsilon}, \frac{C_{\mathsf{LSI}}^3(L_V+L_W)^3}{\sigma^6}\right)\log\left(\frac{3\,\mathrm{KL}(\mu_0||\pi)}{\epsilon}\right)$

The next chapter builds on the theory developed in this chapter to the problem of sampling from the mean-field neural network.

CHAPTER III

MEAN FIELD NEURAL NETWORKS

This work is from Tankala et al. (2025), which has been accepted for publication at the Conference on Learning Theory (COLT), 2025. The conference has not published a camera-ready version of the paper yet. This topic and the main algorithm were suggested to me by Dr. Dheeraj M. Nagaraj, who, along with Dr. Anant Raj, advised me on which books and papers to read for optimal transport, and either directed my proofs or checked them. I was the main contributor to all the theorems, proving them, and writing the paper. Dr. Dheeraj M. Nagaraj and Dr. Anant Raj also contributed to writing the introduction section of the paper.

Our results obtained on this topic can be seen as an extension of results obtained in Mean field Langevin dynamics chapter. However, we present this as a separate chapter because of its significant interest in the statistics and machine learning community. In terms of notation, we use the same notation as that of the Mean field Langevin dynamics chapter for consistency.

The structure of this chapter is the following.

1. The first section provides a literature review of this topic.
2. The second section re-introduces notation from the mean field Langevin dynamics chapter. This is done for self-containment.
3. The third section presents the mean field neural network model we study.
4. The fourth section proves the rate of convergence of our algorithm.

3.1 Introduction

Mean-field analysis of neural networks emerged as a theoretical framework for understanding the optimization dynamics of wide neural networks. Early foundational work by Chizat and Bach (2018); Mei, Montanari, and Nguyen (2018); Nitanda and Suzuki (2017) established that gradient flow on infinite width two-layer neural networks converges to the global minimum under appropriate conditions, demonstrating that we can successfully study neural

networks in the infinite-dimensional space of parameter distributions by exploiting convexity. The connection with the mean-field Langevin dynamics arises with the addition of Gaussian noise to the gradient, corresponding to the entropy-regularized term in the objective function. Chizat (2022); Nitanda, Wu, and Suzuki (2022) were among the first to establish exponential convergence rates under certain log-Sobolev inequalities, which are verifiable in regularized risk minimization problems using two-layer neural networks. Subsequently, Suzuki, Nitanda, and Wu (2023); Suzuki, Wu, and Nitanda (2024) study uniform in time propagation of chaos result in the context of mean-field neural networks where the main ingredient is the proximal Gibbs distribution, which also satisfies a log-Sobolev inequality for convex losses with smooth and bounded activation functions.

3.2 Notation

For any measure ρ over $\mathbb{R}^d$ and functions $f, g : \mathbb{R}^d \to \mathbb{R}^d$. Let

$$\langle f, g \rangle_{L_2(\rho)} := \int \rho(dx) \langle f(x), g(x) \rangle \, , \; L_2^2(\rho; f)^2 := \int \rho(dx) \|f(x)\|^2 \, ,$$

if f, g are square integrable with respect to ρ. For a vector field $f : \mathbb{R}^d \to \mathbb{R}^d$, its divergence is given by $\nabla \cdot f = \sum_{i=1}^d \frac{\partial f}{\partial x_i}$, and for a function $f : \mathbb{R} \to \mathbb{R}$, the Laplacian is defined as $\Delta f := \sum_{i=1}^d \frac{\partial^2 f}{\partial x_i^2}$. Let $\mathcal{P}_2(\mathbb{R}^d), \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ with finite second moment, and those that are absolutely continuous with respect to the Lebesgue measure. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we let $\mathsf{Var}(\mu)$ denote the trace of its covariance. The Wasserstein distance $\mathcal{W}_2(\mu, \nu)$ between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as:

$$\mathcal{W}_2^2(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|X - Y\|^2 d\gamma(x, y) \, ,$$

where $\Gamma(\mu, \nu)$ is the set of all joint distributions over $\mathbb{R}^d \times \mathbb{R}^d$ such that the marginal distribution of X is μ and of Y is ν. The Fisher Divergence of a probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with respect to $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as: $\mathsf{FD}(\mu||\nu) := \int_{\mathbb{R}^d} \mu(x) \left\| \nabla \log \frac{\mu(x)}{\nu(x)} \right\|^2 dx$. The first variation of a functional $\mathcal{F}$ at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is denoted by $\delta_\mu \mathcal{F}(\mu)(x)$ or just $\delta \mathcal{F}(x, \mu)$, where $x \in \mathbb{R}^d$, and is

defined as the quantity which satisfies the equality:

$$\frac{d\mathcal{F}(\mu+\varepsilon(\mu'-\mu))}{d\varepsilon}\bigg|_{\varepsilon=0} = \int \delta_\mu \mathcal{F}(\mu)(x)(\mu'-\mu)(x)dx\,.$$

If the probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ have densities p, q respectively, then the total-variation distance between them is defined as: $\|\mu-\nu\|_{\mathsf{TV}} := \frac{1}{2}\int_{\mathbb{R}^d}|p(x)-q(x)|dx$.

3.3 Model description

Let $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$. Consider the activation function $h(x,z) : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ of a neural network. We consider the two layer mean-field neural network to be

$$f(\mu; z) := \int h(x,z)d\mu(x)\,.$$

Given data $(Z, W) \sim P$, where P is some probability distribution, then we consider the square loss functional:

$$\mathcal{E}(\mu) = \mathbb{E}_{(Z,W)}(f(\mu;Z)-W)^2 + \frac{\lambda}{2}\int \|x\|^2 d\mu(x) + \frac{\sigma^2}{2}\mathcal{H}(\mu)\,. \tag{3.1}$$

From Nitanda et al. (2022)[Section 2.3], the first variation of the functional $\mathcal{E}$ defined above is:

$$\delta\mathcal{E}(x;\mu) = \mathbb{E}_{(Z,W)}\left[2(f(\mu;Z)-W)h(x,Z)] + \frac{\lambda}{2}\|x\|^2\right] + \tfrac{\sigma^2}{2}(\log\mu+1)\,,$$

and since $\nabla_{\mathcal{W}}\mathcal{E} = \nabla\delta\mathcal{E}$, we have:

$$\nabla_{\mathcal{W}}\mathcal{E}(\mu) = 2\mathbb{E}_{(Z,W)}\left[(f(\mu;Z)-W)\nabla_x h(x,Z)\right] + \lambda x] + \tfrac{\sigma^2}{2}\nabla\log\mu\,. \tag{3.2}$$

The unique minimizer of the functional $\mathcal{E}$ defined in (3.1) is given in Nitanda et al. (2022)[Equation 15] as the solution of the fixed point equation:

$$\pi(x) \propto \exp\left(-\tfrac{2}{\sigma^2}\delta\mathcal{F}(x,\pi)\right)\,. \tag{3.3}$$

In (3.1), we assume P to be empirical distribution of m data samples $(z_1, w_1), \ldots, (z_m, w_m)$. Then, the functional in (3.1) simplifies to:

$$\mathcal{F}(\mu) = \frac{1}{m}\sum_{i=1}^{m}\left(\int h(z_i,x)d\mu(x) - w_i\right)^2 + \frac{\lambda}{2}\int\|x\|^2 d\mu(x)\,, \tag{3.4}$$

The Wasserstein gradient of this functional follows from (3.2) to be:

$$\nabla_{\mathcal{W}}\mathcal{F}(x;\mu) = \frac{2}{m}\sum_{i=1}^{m}\left(\int h(z_i,y)d\mu(y) - w_i\right)\nabla_x h(z_i,x) + \lambda x\,. \tag{3.5}$$

The unique minimizer of the functional $\mathcal{E}$ satisfies the equation (3.3). We set $\bar{\mathcal{E}}(\mu) = \mathcal{E}(\mu) - \mathcal{E}(\pi)$. We consider the proximal Gibbs distribution corresponding to μ, which is given by:

$$\pi_\mu(x) \propto \exp\left(-\frac{2\delta\mathcal{F}(x,\mu)}{\sigma^2}\right), \tag{3.6}$$

where $\delta\mathcal{F}$ is the first-variation of the functional $\mathcal{F}$. We let $\nu = \mathsf{Unif}([m])$.

Next, we again propose an unbiased estimator $\hat{G} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ of $\nabla_{\mathcal{W}}\mathcal{F}$ defined in (3.5), i.e., if $Y, \xi \sim \mu \times \nu$, then $\mathbb{E}[\hat{G}(x,Y,\xi)] = \nabla_{\mathcal{W}}\mathcal{F}(x,\mu)$ for every $x \in \mathbb{R}^d$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, and where ν is a fixed, known distribution.

We denote the random variable ξ, used in the definition of the Wasserstein gradient estimator $\hat{G}$, by I and choose $\hat{G}$ to be:

$$\hat{G}(x,y,i) := -(h(z_i,y) - w_i)\nabla_x h(z_i,x) - \lambda x\,,\ \forall\ x,y \in \mathbb{R}^d\,, i \in [m]\,.$$

3.4 Main results

First, we make some assumptions that relate to continuity and boundedness properties of the activation function h of the neural network, and a functional inequality for the proximal Gibbs distribution defined in (3.6) and the probability measure π defined in (3.3). The first three assumptions are from the mean field Langevin dynamics chapter and added here for self-containment.

3.4.1 Assumptions. We use u and $\hat{G}$ interchangeably. So $\hat{G}(x,y,\xi) = u(x,y,\xi)$. Following the idea in the mean field Langevin dynamics chapter, for some functional $\bar{\mathcal{F}} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, let π be the unique minimizer of the functional $\bar{\mathcal{F}} + \frac{\sigma^2}{2}\mathcal{H}$. Define $\bar{\mathcal{E}}(\mu) := \bar{\mathcal{F}}(\mu) + \frac{\sigma^2}{2}\mathcal{H}(\mu) - \bar{\mathcal{F}}(\pi) - \frac{\sigma^2}{2}\mathcal{H}(\pi)$.

Assumption 7 (Lipschitz continuity)**.** *For some $L_u, L_{\bar{u}}, L_{\bar{\mathcal{F}}}, L_{\mathcal{F}} > 0$, the function $x \to u(x, y, \xi)$ and $y \to u(x, y, \xi)$ are L_u-Lipschitz. For every $x, y \in \mathbb{R}^d$, $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:*

$$(i)\ \ \|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \mu) - \nabla_{\mathcal{W}} \bar{\mathcal{F}}(y, \nu)\| \leq L_{\bar{\mathcal{F}}} \mathcal{W}_2(\mu, \nu) + L_{\bar{u}} \|x - y\|$$

$$(ii)\ \ \|\nabla_{\mathcal{W}} \mathcal{F}(x, \mu) - \nabla_{\mathcal{W}} \mathcal{F}(x, \nu)\| \leq L_{\mathcal{F}} \mathcal{W}_2(\mu, \nu) + L_u \|x - y\|$$

$$(iii)\ \ \|\nabla_{\mathcal{W}} \bar{\mathcal{F}}(x, \mu) - \nabla_{\mathcal{W}} \mathcal{F}(x, \mu)\| \leq L_l \mathcal{W}_2(\pi, \mu)$$

Assumption 8. *For some $C_{\bar{\mathcal{E}}}, C_{\mathsf{LSI}}, C_{\mathrm{KL}} > 0$, the functional $\bar{\mathcal{E}}$ satisfies the:*

$(i)\ \ \|\nabla_{\mathcal{W}} \bar{\mathcal{E}}(x, \mu)\|^2_{L^2(\mu)} \geq C_{\bar{\mathcal{E}}} \bar{\mathcal{E}}(\mu)$ *for all* $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ (Polyak-Łojasiewicz inequality)

$(ii)\ \ \mathrm{KL}(\mu \,||\, \pi) \leq \dfrac{C_{\mathsf{LSI}}}{2} \mathsf{FD}(\mu \,||\, \pi)$ *for all* $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ (Log-Sobolev inequality)

$(iii)\ \ \mathrm{KL}(\mu||\pi) \leq C_{\mathrm{KL}} \bar{\mathcal{E}}(\mu)$ *for all* $\mu \in \mathcal{P}_{2,\mathsf{ac}}(\mathbb{R}^d)$ (KL-Growth)

Assumption 9. *If $Y, \xi \sim \rho_0 \times \nu$, then from some $C^{\mathsf{Var}}, C^{\mathsf{Var}}_{\nu} > 0$, and for all $x \in \mathbb{R}^d$:*

$$\mathbb{E}\|u(x, Y, \xi) + \nabla_{\mathcal{W}} \mathcal{F}(x; \rho_0)\|^2 \leq C^{\mathsf{Var}} \mathsf{Var}(\rho_0) + C^{\mathsf{Var}}_{\nu}\,.$$

Assumption 10 (Boundedness)**.** *For every $x, z \in \mathbb{R}^d$:*

$$\|h(z, x)\| \leq B; \quad \|z\| \leq R; \quad |w| \leq R; \quad \|\nabla_x h(z, x)\| \leq M\|z\|\,.$$

Assumption 11 (Lipschitz continuity)**.** *The function $x \to \nabla h(z, x)$ is $L\|z\|$-Lipschitz.*

Excluding special cases, the boundedness assumption on $h, \nabla_x h$ and Lipschitz assumption on $\nabla_x h$ are necessary to satisfy the general assumptions in prior works Nitanda (2024); Wang (2024) for the square loss. For instance, we refer to (Nitanda, 2024, Assumption 1).

Assumption 12 (LSI)**.** *For any $q \in \mathcal{P}_2(\mathbb{R}^d)$, the proximal Gibbs distribution π_q satisfies the LSI with constant C_{LSI}. π also satisfies LSI with the same constant.*

3.4.2 Technical lemmas.

The following lemma proves that Assumptions 10, 11 and 12 stated above are satisfied with the stated constants in the statement of the lemma. The fact that these assumptions are satisfied lets us use the theory developed in the mean field Langevin dynamics chapter.

Lemma 13. *Under Assumptions 10, 11 and 12, the general Assumption 7 is satisfied with* $L_u = L_{\bar{u}} = (B+R)LR + \lambda + M^2R^2$, $L_{\mathcal{F}} = L_{\bar{\mathcal{F}}} = M^2R^2$, *and* $L_l = 0$. *The general Assumption 8 is satisfied with* $C_{\mathsf{LSI}} = C_{\mathsf{LSI}}$, $C_{\bar{\mathcal{E}}} = \frac{\sigma^2}{C_{\mathsf{LSI}}}$ *and* $C_{\mathrm{KL}} = \frac{2}{\sigma^2}$. *The Assumption 9 is satisfied with* $C^{\mathsf{Var}} = 0$ *and* $C_{\nu}^{\mathsf{Var}} = 4M^2R^2(B+R)^2$.

Proof. Note that $u = \hat{G}$ satisfies the Lipschitz continuity property of $x \to u(x, y, \xi)$ and $y \to u(x, y, \xi)$ in Assumption 7 with $L_u = (B+R)(LR+N) + \lambda + M^2R^2$ because for every $x_1, x_2 \in \mathbb{R}^d$:

$$\begin{aligned}\|u(x_1, y, i) - u(x_2, y, i)\| &\leq (h(z_i, Y) - w_i)\|\nabla_x h(z_i, x_1) - \nabla_x h(z_i, x_2)\| + \lambda\|x_1 - x_2\| \\ &\leq (h(z_i, y) - w_i)(L\|z_i\|)\|x_1 - x_2\| + \lambda\|x_1 - x_2\| \\ &\leq ((B+R)LR + \lambda)\,\|x_1 - x_2\|\,,\end{aligned}$$

where the last two inequalities follow from Assumptions 10 and 11. Similarly considering $\|u(x, y_1, i) - u(x, y_2, i)\|$, we conclude the result. Similarly, we can take $L_{\mathcal{F}} = M^2R^2$ because, for every $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\begin{aligned}&\|\nabla_{\mathcal{W}}\mathcal{F}(x, \mu) - \nabla_{\mathcal{W}}\mathcal{F}(x, \nu)\| \\ &= \left\|\frac{1}{m}\sum_{i=1}^{m}\left[\left(\int h(z_i, w)d\mu(w) - \int h(z_i, w)d\mu(v)\right)\nabla_x h(z_i, x)\right]\right\| \\ &\leq \frac{1}{m}\sum_{i=1}^{m}\|\nabla_x h(z_i, x)\|M\|z_i\|\mathcal{W}_1(\mu, \nu) \\ &\leq M^2R^2\mathcal{W}_2(\mu, \nu)\,.\end{aligned}$$

Next, since the proximal Gibbs measure satisfies the LSI by Assumption 12, the LSI condition in Assumption 8 is satisfied with $C_{\mathsf{LSI}} = C_{\mathsf{LSI}}$. By (Nitanda et al., 2022, Proposition 1), we conclude that $C_{\mathrm{KL}} = \frac{2}{\sigma^2}$. Again, by (Nitanda et al., 2022, Proposition 1), we have:

$\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\mu) = \frac{\sigma^2}{2}\nabla_x \log(\frac{\mu(x)}{\pi_\mu(x)})$. Therefore, have:

$$\begin{aligned}\int \|\nabla_{\mathcal{W}}\bar{\mathcal{E}}(x,\mu)\|^2 d\mu(x) &= \frac{\sigma^4}{4}\mathsf{FD}(\mu||\pi_\mu) \\ &\geq \frac{\sigma^4}{2C_{\mathsf{LSI}}}\,\mathrm{KL}(\mu||\pi_\mu) && \text{(By Assumption 12)} \\ &\geq \frac{\sigma^2}{C_{\mathsf{LSI}}}\bar{\mathcal{E}}(\mu)\,. && \text{(By (Nitanda et al., 2022, Proposition 1))}\end{aligned}$$

Therefore, we conclude that Assumption 8-(Polyak-Łojasiewicz inequality) holds with $C_{\bar{\mathcal{E}}} = \frac{\sigma^2}{C_{\mathsf{LSI}}}$. Let $Y, I \sim \rho_0 \times \nu$ and let Y', I' be an independent copy of Y, I. By Assumptions 10, 11 we have $\|u(x,y,i) + \nabla_{\mathcal{W}}\mathcal{F}(x,\rho_0)\| \leq 2(B+R)MR$ almost surely. Thus, the mean-field neural network case satisfies Assumption 9 with $C^{\mathsf{Var}} = 0$ and $C_\nu^{\mathsf{Var}} = 4(B+R)^2M^2R^2$. □

We now apply the above lemma to prove the rate of convergence of Algorithm 1 described in the mean field Langevin dynamics chapter for the mean field neural networks problem.

3.4.3 Main results.

Theorem 3. *Consider the case of the Mean Field Neural Network with square loss in Equation* (3.1) *under Assumptions 11, 10 and 12. We consider Algorithm 1 with $\hat{G}$ as defined above and $\eta < c_0 \min\left(\frac{C_{\mathsf{LSI}}}{\sigma^2}, \frac{\sigma^4}{C_{\mathsf{LSI}}^2 L_u^3}\right)$ for some $c_0 > 0$ small enough and $L_u = (B+R)LR + \lambda + M^2R^2$. Then for some universal constant $C > 0$:*

$$\begin{aligned}&\mathbb{E}\mathcal{E}(\mu_T|\mathcal{R}_{T-1}) - \mathcal{E}(\pi) \\ &\leq e^{-\frac{T\eta\sigma^2}{8C_{\mathsf{LSI}}}}(\mathcal{E}(\mu_0) - \mathcal{E}(\pi)) + C\frac{C_{\mathsf{LSI}}}{\sigma^2}\left[\eta(\sigma^2L_u^2 d + L_u M^2R^2(B+R)^2) + \sqrt{\eta}\sigma d L_u MR(B+R)\right]\,.\end{aligned}$$

The following remark states the number of timesteps T needed for $\mathcal{E}(\mu_T|\mathcal{R}_{T-1})$ to converge to $\mathcal{E}(\pi)$ in expectation.

Remark 3. *Given $\epsilon/3 \in (0, \mathcal{E}(\mu_0) - \mathcal{E}(\pi))$, as per Theorem 3, we can achieve $\mathbb{E}\mathcal{E}(\mu_T|\mathcal{R}_{T-1}) - \mathcal{E}(\pi) \leq \epsilon$ by picking*

1. $\eta = \frac{8C_{\mathsf{LSI}}}{\sigma^2 T}\log\left(\frac{3(\mathcal{E}(\mu_0)-\mathcal{E}(\pi))}{\epsilon}\right)$
2. $T \gtrsim \max\left(\frac{C_{\mathsf{LSI}}^3 d^2 L_u^2 M^2R^2(B+R)^2}{\sigma^4\epsilon^2}, \frac{C_{\mathsf{LSI}}^2(\sigma^2L_u^2 d + L_u M^2R^2(B+R)^2)}{\sigma^4\epsilon}, \frac{L_u^3 C_{\mathsf{LSI}}^3}{\sigma^6}\right)\log\left(\frac{3(\mathcal{E}(\mu_0)-\mathcal{E}(\pi))}{\epsilon}\right)$

Proof of Theorem 3. Under the parameter correspondence established in Lemma 13, with our choice of η, we conclude that the conditions for Theorem 1 are satisfied (once considered with the fact that $C_{\mathsf{LSI}} \geq \frac{\sigma^2}{2L_u}$ from Lemma 9). Since $\pi \propto \exp(-\frac{2\delta\mathcal{F}(x,\pi)}{\sigma^2})$ and $\nabla_{\mathcal{W}}\mathcal{F}(x,\pi) = \nabla_x \delta\mathcal{F}(x,\pi)$, we apply Lemma 4 to conclude that $(G_\pi)^2 = G_{\mathsf{mod}}^2 := \mathbb{E}_{x\sim\pi}\|\nabla_{\mathcal{W}}\mathcal{F}(x,\pi)\|^2 \leq \frac{d\sigma^2 L_u}{2}$. Using Assumption 9 instantiated to our case, we conclude $(\sigma^*)^2 \leq 4M^2R^2(B+R)^2$.

Instantiating the quantities γ_1, γ_2 and γ_3 found in Theorem 1 to the case of mean field neural networks, we have:

$$\gamma_1 \lesssim \sigma d L_u M R(B+R)\,,\ \gamma_2 \lesssim \sigma^2 L_u^2 d + L_u M^2 R^2 (B+R)^2\,,\ \gamma_3 \lesssim \sigma^2 L_u^3 d + L_u^2 M^2 R^2 (B+R)^2\,.$$

We then apply Theorem 1 and simplify using the fact that $\eta L_u \leq c_0$, and $C_{\mathsf{LSI}} \geq \frac{\sigma^2}{2L_u}$ from Lemma 9 to conclude the result.

□

The next chapter is not directly related to Chapters II and III, but it shares the idea of using optimal transport theory and functional inequalities to prove the rate of convergence of a stochastic process on a discrete state space.

CHAPTER IV

MEAN-FIELD ISING MODEL

The topic of working on applying informed proposals for the tensor Ising model was suggested to me by Dr. Krishnakumar Balasubramanian, who, along with Dr. Quan Zhou, advised me on using functional inequalities as a proof technique. I was the primary contributor in identifying the papers related to Ricci curvature of Markov chains, understanding the proof technique, and writing all the lemmas and theorems.

The work presented in this chapter is part of a paper on a broader topic, which I hope to finish in 2025.

The structure of this chapter is the following:

1. The first section provides an introduction to the mean-field Tensor Ising model and discusses prior work on analyzing Markov chains used for sampling from this distribution.

2. The second section defines a notion of Ricci curvature for Markov chains.

3. The third section reviews literature where this curvature notion has been applied to prove convergence rates and establish functional inequalities.

4. The fourth section introduces our studied chain for sampling from the Tensor Ising model and proves its mixing time bound at sufficiently high temperature using Ricci curvature defined in the second section.

4.1 Definition of the model

Consider a discrete space $\mathcal{X} = \{-1, +1\}^n$. We define the mean-field p-tensor Ising model, similar to Samanta, Mukherjee, and Zhang (2024)[Equation 1.2] with zero external magnetic field, as a probability distribution on $\mathcal{X}$:

$$\pi_{\beta,p}(x) := \frac{1}{2^n Z_n(\beta, p)} \exp\left(\frac{\beta}{n^{p-1}} H(x)\right), \tag{4.1}$$

where $x \in \{-1,+1\}^n, p \in \mathbb{N}, p \geq 2, \beta > 0$ is the inverse temperature parameter, Z_n is a normalization constant, and the Hamiltonian H is defined as

$$H(x) := \sum_{1 \leq i_1, i_2, \ldots, i_p \leq n} x_{i_1} x_{i_2} \ldots x_{i_p} .$$

Note that in the above Hamiltonian we allow the indices $\{i_1, i_2, \ldots, i_p\}$ to have repetitions. This probability distribution is also commonly referred to as the Curie-Weiss Ising model in the literature.

Tensor Ising models represent a natural generalization of the classical 2-spin Ising model, designed to capture higher-order dependencies in complex relational data. This model has found applications in several domains in computational statistics. The most prominent statistical application is in social network analysis, particularly for modeling peer group effects in logistic regression frameworks Daskalakis, Dikkala, and Panageas (2020). In this context, Daskalakis et al. (2020) model binary outcomes on a network as a higher-order spin glass, where the behavior of an individual depends on a linear function of their own vector of covariates and some polynomial function of the behavior of others, capturing peer-group effects. This addresses the fundamental limitation that pairwise interactions are seldom observed in the real world and the decision of an individual is affected not just by pairwise communications, but by interactions with larger community tuples.

We introduce some notation in this paragraph. Consider a spin configuration $x \in \mathcal{X} = \{-1,+1\}^n$. Let $x(v) \in \{-1,+1\}$ denote the spin at index v of x, and let $\|x-y\|$ be the ℓ^1 distance between $x, y \in \mathcal{X}$. A classical way of simulating from the probability distribution in (4.1) is a Markov chain called *Glauber dynamics* or *Gibbs sampler*. The Glauber dynamics is defined as follows in Levin and Peres (2017)[Section 3.3.2] for a general discrete probability distribution. When this definition is applied to the probability distribution in (4.1), the Glauber dynamics moves from state $x \in \mathcal{X}$ as follows. Choose an index v uniformly at random from $[n] = \{1, 2, \ldots, n\}$. A new spin value at v is chosen according to probability measure $\pi_{\beta,p}$ conditioned on the set of spin configurations $y \in \mathcal{X}$ that are equal to x at all indices other than v. Therefore, the probability

transition matrix P_G of Glauber dynamics on $\mathcal{X}$ is:

$$P_G(x,y) = \begin{cases} \frac{1}{n} \cdot \frac{\pi_{\beta,p}(y)}{\pi_{\beta,p}(y)+\pi_{\beta,p}(x)} \text{ if } \|x-y\| \leq 2 \\ 0 \text{ otherwise} . \end{cases} \tag{4.2}$$

The mixing time of Glauber dynamics for the probability distribution in (4.1) with $p = 2$ has been studied in Levin, Luczak, and Peres (2010). In Samanta et al. (2024), the mixing time behavior of Glauber dynamics for a probability distribution which generalizes the probability distribution in (4.1) and where $p \geq 3$ is studied. This probability measure is defined in Samanta et al. (2024)[Equation 1.2] as:

$$\pi_{\beta,h,p}(x) = \frac{1}{2^n Z(\beta,h,p)} \exp\{n(\beta \bar{x}^p + h\bar{x})\} , \tag{4.3}$$

where $x \in \{-1,+1\}^n$, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ is the mean magnetization, $\beta \in (0,\infty)$ is again the inverse temperature parameter, $h \in \mathbb{R}$ is the external magnetic field, and Z is again the normalization constant.

The essence of the main theorem in Samanta et al. (2024) is the identification of three disjoint regions in the parameter space $(\beta, h) \in (0,\infty) \times \mathbb{R}$, which correspond to different mixing times of Glauber dynamics. These regions are defined based on the behavior of the function

$$H_{\beta,h,p}(x) := \beta x^p + hx - I(x) ,$$

where $I(x) = \frac{1}{2}[(1+x)\log(1+x) + (1-x)\log(1-x)]$. The regions are:

1. $\mathcal{R}_p$ (p-locally regular points) where $H_{\beta,h,p}$ has a unique local maximizer $m^* \in (-1,1)$ with $H''_{\beta,h,p}(m_*) < 0$ and no other stationary points

2. $\mathcal{C}_p$ (p-locally critical points) where $H_{\beta,h,p}$ has more than one local maximizer

3. $\mathcal{S}_p$ (p-special points) where $H_{\beta,h,p}$ has a unique local maximizer $m_* \in (-1,1)$ with $H''_{\beta,h,p}(m_*) = 0$

Finally, the main theorem in Samanta et al. (2024) is the following where it is proven that the Glauber dynamics has different mixing times in the regions defined above.

Theorem 4 (Samanta et al. (2024)[Theorem 1)**.** *For every* $\epsilon \in (0, 1/2)$*,* $p \geq 3$*, and* $(\beta, h) \in (0, \infty) \times \mathbb{R}$*:*

1. *If* $(\beta, h) \in \mathcal{R}_p$*, then* $t_{\text{mix}}(\epsilon) = \Theta_\epsilon(N \log N)$
2. *If* $(\beta, h) \in \mathcal{C}_p$*, then* $t_{\text{mix}}(\epsilon) \geq \exp(\Omega_\epsilon(N))$
3. *If* $(\beta, h) \in \mathcal{S}_p$*, then* $t_{\text{mix}}(\epsilon) = \Theta_\epsilon(N^{3/2})$

Since our result is about the mixing time of a Markov chain for the mean-field Tensor Ising model at high temperature, we focus on the proof technique employed in Theorem 4 for the same temperature regime, i.e., when $(\beta, h) \in \mathcal{R}_p$.

Proof sketch. The idea of the proof can be split into three parts. The first part analyzes the drift of a sufficient statistic. Let $X_t = (X_t^1, \ldots, X_t^n) \in \{-1, 1\}^n$ denote the spin configuration at time t of the Glauber dynamics for the mean-field Tensor Ising model. Define the mean magnetization

$$c_t := \bar{X}_t = n^{-1} \sum_{i=1}^{n} X_t^i \,.$$

It is shown that the Glauber dynamics exhibits the following drift in Samanta et al. (2024)[Lemma 3.1]:

$$\mathbb{E}[c_{t+1} - c_t \mid c_t = c] = n^{-1}(\lambda(c) - c)\,,$$

where $\lambda(c) = \tanh(p\beta c^{p-1} + h)$. Let c^* be a fixed point of the function λ. Building on the above drift analysis, it is then shown in Samanta et al. (2024)[Lemma 3.3] that starting from a mean magnetization c_0 which is between c^* and the fixed point of λ nearest to c^*, referred to as the *burn-in* analysis, establishes that for any $\epsilon > 0$, there exist constants $k, k' > 0$ such that,

$$\mathbb{P}\left(\bigcap_{t=kn}^{e^{k'\sqrt{n}}} \{c^* - \epsilon < c_t < c^* + \epsilon\}\right) \geq 1 - e^{-\Omega(\sqrt{n})}\,.$$

This shows that the mean magnetization enters an ϵ-neighborhood of c^* within $O(n)$ steps and remains there for exponentially long time with high probability.

The next step of the proof is about a path coupling argument. To analyze the mixing time, Samanta et al. (2024) construct a coupling of two copies (X_t, Y_t) of the Glauber dynamics with

possibly different initial configurations $X_0 = x$ and $Y_0 = y$. At each time step, both chains use: (i) the same uniformly random vertex $I \in [n] := \{1, \ldots, n\}$, and (ii) the same uniform random variable $U_{t+1} \in [0, 1]$. The spin updates at time $t + 1$ are given by

$$X_{t+1}^I = \begin{cases} 1 \text{ if } U_{t+1} \leq f(\bar{X}_t) \\ -1 \text{ otherwise} \end{cases}, \; Y_{t+1}^I = \begin{cases} 1 \text{ if } U_{t+1} \leq f(\bar{Y}_t) \\ -1 \text{ otherwise} \end{cases}.$$

where

$$f(t) := (1 + \tanh(p\beta t^{p-1} + h))/2\,,$$

and by keeping all the other coordinates of X_{t+1}, Y_{t+1} unchanged from X_t, Y_t. This coupling ensures that when $\bar{X}_t = \bar{Y}_t$, the chains update identically, and more generally, the probability of disagreement on the updated spin equals $|f(\bar{X}_t) - f(\bar{Y}_t)|$.

The next step of the proof is a result showing contraction of Hamming distance under the aforementioned coupling of Markov chains. Let $\rho(X_t, Y_t) = \sum_{i=1}^n \mathbf{1}\{X_t^i \neq Y_t^i\}$ denote the Hamming distance between configurations. It is then proven in Samanta et al. (2024)[Lemma 3.4] that under certain conditions for every possible pair $(x, y) \in \{-1, +1\}^n$ the coupling (X_t, Y_t) satisfies

$$\mathbb{E}[\rho(X_t, Y_t)] \leq \rho(x, y) \exp(-t\delta/n)\,,$$

where $\delta > 0$. If $\rho(x, y) = n$, then by Samanta et al. (2024)[Lemma 3.4], $\mathbb{E}[\rho(X_t, Y_t)] \leq ne^{-t\delta/n}$. Finally, to reduce the expected Hamming distance below ϵ, we require $t = (n/\delta)\log(n/\epsilon) = O(n \log n)$ steps. Combined with the $O(n)$ burn-in time in the first step of the proof sketch, this establishes the total mixing time $t_{\mathsf{mix}}(\epsilon) = O(n \log n)$ for the high-temperature regime. □

Note that at a high level, the above proof is based on probabilistic coupling and drift analysis based arguments. For the Markov chain we study, the proof technique is completely different and is based on a functional inequality.

4.2 Entropic Ricci curvature of Markov chains

Let $\mathcal{X}$ be a finite set and $Q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a continuous-time Markov chain with generator L that acts on functions $\psi : \mathcal{X} \to \mathbb{R}$ as:

$$L\psi(x) = \sum_{y \in \mathcal{X}} Q(x,y)[\psi(y) - \psi(x)] \,. \tag{4.4}$$

We assume Q is irreducible, i.e. for all $(x,y) \in \mathcal{X}$, there exist points $(x_1 = x, x_2, \ldots, x_n = y)$ such that $Q(x_i, x_{i+1}) > 0$ for $i \in \{1, 2, \ldots, n-1\}$. This condition implies that there exists a unique probability measure π on $\mathcal{X}$ that is stationary under the continuous time Markov chain Q. We also assume that Q is reversible with respect to π, i.e. Q satisfies the detailed balance condition:

$$Q(x,y)\pi(x) = Q(y,x)\pi(y) \quad \forall x, y \in \mathcal{X} \,. \tag{4.5}$$

Let $\mathcal{P}(\mathcal{X}) := \{\rho : \mathcal{X} \to \mathbb{R}_+ : \sum_x \rho(x)\pi(x) = 1\}$ denote the space of probability densities with respect to π and $\mathcal{P}_*(\mathcal{X})$ be the space of strictly positive probability densities with respect to π. Define θ to be the logarithmic mean as:

$$\theta(s,t) := \int_0^1 s^\alpha t^{1-\alpha} d\alpha \,. \tag{4.6}$$

Next, we define a metric $\mathcal{W}$ on $\mathcal{P}(\mathcal{X})$ introduced in Erbar et al. (2017)[Equation 2.2]:

$$\mathcal{W}(\rho_0, \rho_1)^2 := \inf_{\rho, \psi} \left\{ \frac{1}{2} \int_0^1 \sum_{x,y \in \mathcal{X}} |\psi_t(x) - \psi_t(y)|^2 \theta(\rho_t(x), \rho_t(y)) Q(x,y)\pi(x) dt \right\} , \tag{4.7}$$

where the infimum is over curves $\rho : [0,1] \to \mathcal{P}(\mathcal{X})$ and $\psi : [0,1] \to \mathbb{R}^{\mathcal{X}}$ satisfying the continuity equation as described in Erbar et al. (2017)[Equation 2.2]:

$$\begin{aligned} \frac{d}{dt}\rho_t(x) + \sum_{y \in \mathcal{X}} (\psi_t(y) - \psi_t(x)) \theta(\rho_t(x), \rho_t(y)) Q(x,y) = 0 \quad \forall\, x \in \mathcal{X} \,, \\ \rho(0) = \rho_0 \,,\ \rho(1) = \rho_1 \,. \end{aligned} \tag{4.8}$$

This Wasserstein metric was originally introduced by Maas (2011) for a discrete time Markov chain, as opposed to the continuous time Markov chain Q used in the above definition. More specifically, in Maas (2011), the definition of the Wasserstein metric is the same as (4.7) except that Q is replaced by K.

It is proven in Maas (2011)[Theorem 3.29] that the space $(\mathcal{P}_*(\mathcal{X}), \mathcal{W})$ is a Riemannian manifold. Let $D_t\rho$ denote the tangent vector field along a smooth curve $\rho : (0, \infty) \to \mathcal{P}_*(\mathcal{X})$. For a smooth functional $\varphi : \mathcal{P}_*(\mathcal{X}) \to \mathbb{R}$, let grad $\varphi(\rho)$ denote its gradient at ρ. The gradient flow of the functional φ is the equation defined as:

$$D_t\rho := -\text{grad}\,\varphi(\rho_t) \tag{4.9}$$

Next, we define the relative entropy functional as

$$\mathcal{H}(\rho) := \sum_{x\in\mathcal{X}} \pi(x)\rho(x)\log\rho(x)\,. \tag{4.10}$$

The idea behind defining the metric $\mathcal{W}$ in a way described in (4.7) in the work of Maas (2011) is that the discrete heat equation $\partial_t\rho = L\rho$ is the gradient flow of the relative entropy functional, i.e.:

$$D_t\rho = -\text{grad}\,\mathcal{H}(\rho_t)\,, \tag{4.11}$$

which is proven in Maas (2011)[Theorem 1.2].

It has also been proven in Erbar and Maas (2012) that every pair $\rho_0, \rho_1 \in \mathcal{P}(\mathcal{X})$ of densities can be joined by a constant speed $\mathcal{W}$-geodesic $(\rho_s)_{s\in[0,1]}$ where a constant speed $\mathcal{W}$-geodesic means $\mathcal{W}(\rho_s, \rho_t) = |s-t|\mathcal{W}(\rho_0, \rho_1)$ for all $s, t \in [0, 1]$.

Having defined the geometry of the spaces $\mathcal{P}(\mathcal{X}), \mathcal{P}_*(\mathcal{X})$, we now state the notion of Ricci curvature originally defined in Maas (2011) and used in Erbar and Maas (2012), Erbar et al. (2017).

Definition 1 (Erbar et al. (2017) Definition 2.1)**.** *The Markov triple $(\mathcal{X}, Q, \pi)$ has Ricci curvature bounded from below by $\kappa \in \mathbb{R}$, denoted Ric$(\mathcal{X}, Q, \pi) \geq \kappa$, if for any constant speed geodesic $\{\rho_t\}_{t\in[0,1]}$ in $(\mathcal{P}(\mathcal{X}), \mathcal{W})$:*

$$\mathcal{H}(\rho_t) \leq (1-t)\mathcal{H}(\rho_0) + t\mathcal{H}(\rho_1) - \frac{\kappa}{2}t(1-t)\mathcal{W}(\rho_0, \rho_1)^2\,. \tag{4.12}$$

An equivalent characterization of the above notion of Ricci curvature is given through a Bochner-type inequality which was originally introduced in Maas (2011) and used in Erbar and Maas (2012), Erbar et al. (2017). We now introduce some definitions from Erbar et al. (2017)[Section 2.2] to aid defining the Bochner-type inequality. For $\psi : \mathcal{X} \to \mathbb{R}$, define the

discrete gradient $\nabla\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by $\nabla\psi(x,y) = \psi(y) - \psi(x)$. Let θ_1, θ_2 be the derivatives of θ, as defined in (4.6), with respect to the first and second variables respectively. Define the inner product between gradients $\nabla\psi, \nabla\phi$ as:

$$\langle \nabla\psi, \nabla\phi \rangle_\rho := \frac{1}{2} \sum_{x,y \in \mathcal{X}} \nabla\psi(x,y)\nabla\phi(x,y)\theta(\rho(x),\rho(y))Q(x,y)\pi(x)\,, \tag{4.13}$$

and the Hessian of the functional $\mathcal{H}$ as:

$$\text{Hess}\,\mathcal{H}(\rho)[\nabla\psi] = \frac{1}{2} \sum_{x,y \in \mathcal{X}} \left[\frac{1}{2}\hat{L}_\rho(x,y)|\nabla\psi(x,y)|^2 - \theta(\rho(x),\rho(y))\nabla\psi(x,y)\nabla L\psi(x,y) \right] Q(x,y)\pi(x)\,, \tag{4.14}$$

where

$$\hat{L}_\rho(x,y) := \partial_1\theta(\rho(x),\rho(y))L\rho(x) + \partial_2\theta(\rho(x),\rho(y))L\rho(y)\,.$$

Now define $\mathcal{A}(\rho,\psi) := \|\nabla\psi\|_\rho^2$ and $\mathcal{B}(\rho,\psi) := \text{Hess}\,\mathcal{H}(\rho)[\nabla\psi]$. Then the equivalent characterization of Ricci curvature is given by:

Definition 2 (Erbar et al. (2017)[Proposition 2.2)**.** *$Ric(X,Q,\pi) \geq \kappa$ if and only if for every $\rho \in \mathcal{P}_*(\mathcal{X})$ and every $\psi \in \mathbb{R}^{\mathcal{X}}$, the inequality $\mathcal{B}(\rho,\psi) \geq \kappa\mathcal{A}(\rho,\psi)$ holds.*

4.3 Perturbative criterion for positive Ricci curvature

The main proof technique to obtain Ricci curvature bounds for the Markov chain we study is Erbar et al. (2017)[Theorem 3.9]. In this section, we describe how this theorem is related to the Bochner-type inequality described in the previous section and provide a sketch of its proof. The proof of Erbar et al. (2017)[Theorem 3.9] draws inspiration from Holley-Stroock perturbation condition.

First, we introduce some definitions. Let Q be a finite irreducible reversible continuous-time Markov chain on $\mathcal{X}$ with a unique stationary distribution π. A *mapping representation* in Erbar et al. (2017)[Definition 3.1] is a pair (G, c) where:

1. G is a set of bijective maps $\delta : \mathcal{X} \to \mathcal{X}$,
2. $c : \mathcal{X} \times G \to \mathbb{R}_+$ are jump rates,

such that:

1. The generator L of the continuous time Markov chain Q can be written as

$$L\psi(x) = \sum_{\delta \in G} \nabla_\delta \psi(x) c(x, \delta)\,, \quad \text{where } \nabla_\delta \psi(x) := \psi(\delta x) - \psi(x)\,. \tag{4.15}$$

2. For every $\delta \in G$, there exists a unique $\delta^{-1} \in G$ such that $\delta^{-1}(\delta(x)) = x$ for all x with $c(x, \delta) > 0$.

Note that $c(x, \delta) = Q(x, \delta x)$, so the detailed balance condition in (4.5) becomes

$$c(x, \delta)\pi(x) = c(\delta x, \delta^{-1})\pi(\delta x)\,. \tag{4.16}$$

We assume the mapping representation is commutative: $\delta \circ \eta = \eta \circ \delta$ for all $\delta, \eta \in G$.

Next, we state the main theorem that we use to obtain Ricci curvature bounds in the next section.

Theorem 5 (Erbar et al. (2017) Theorem 3.9)**.** *Assume θ to be the logarithmic mean, i.e.*

$$\theta(s, t) = \int_0^1 s^\alpha t^{1-\alpha} d\alpha\,.$$

Define the following quantities:

$$q(x, \delta, \eta) := c(x, \delta)c(x, \eta)\pi(x)\,, \tag{4.17}$$

$$q_*(x, \delta, \eta) := \min\{q(x, \delta, \eta), q(\delta x, \delta^{-1}, \eta), q(\eta x, \delta, \eta^{-1}), q(\delta\eta x, \delta^{-1}, \eta^{-1})\}\,, \tag{4.18}$$

for $\delta, \eta \in G$ with $\eta \neq \delta, \delta^{-1}$. If

$$\lambda := \min_{\substack{x \in \mathcal{X}, \delta \in G \\ c(x,\delta) > 0}} \left[c(x, \delta) - \mathbf{1}_{\delta \neq \delta^{-1}} c(\delta x, \delta) - \sum_{\eta : \eta \neq \delta, \delta^{-1}} \frac{(q - q_*)(\delta x, \delta^{-1}, \eta)}{c(x, \delta)\pi(x)} \right] \geq 0\,, \tag{4.19}$$

then Ric$(\mathcal{X}, Q, \pi) \geq 2\lambda$.

In the above theorem, the quantity $(q - q_*)(x, \delta, \eta) = q(x, \delta, \eta) - q_*(x, \delta, \eta)$ measures the deviation from homogeneity of the transition rates. When $c(\delta x, \eta) = c(x, \eta)$ for all x, δ, η (homogeneous case), we have $q = q_*$, hence $(q - q_*) = 0$.

Next, we provide an explanation about the connection between this theorem and the discrete Bochner's inequality defined in the previous section. In particular, we show in the

following sketch of the proof of Theorem 5 that the discrete Bochner's inequality $\mathcal{B}(\rho, \psi) \geq 2\lambda\mathcal{A}(\rho, \psi)$ is established by controlling the deviation from the homogeneous case.

Proof sketch. The proof decomposes $\mathcal{B}(\rho, \psi)$ into diagonal and off-diagonal contributions:

$$\mathcal{B}(\rho, \psi) = \sum_{x,\delta} \mathcal{B}(x, \delta, \delta) q(x, \delta, \delta) + \sum_{x,\delta \neq \eta} \mathcal{B}(x, \delta, \eta) q(x, \delta, \eta) \,. \tag{4.20}$$

Using the Bochner-Bakry-Émery method Erbar et al. (2017)[Proposition 3.4] and careful estimates Erbar et al. (2017)[Lemmas 3.6, 3.7], the proof shows:

1. The diagonal terms satisfy $\sum_{x,\delta} \mathcal{B}(x, \delta, \delta) c(x, \delta) \pi(x) \geq 2\mathcal{A}(\rho, \psi)$.

2. The off-diagonal terms can be controlled when $(q - q_*)$ is small.

Specifically, the proof establishes:

$$\mathcal{B}(\rho, \psi) \geq \sum_{x,\delta} \mathcal{B}(x, \delta, \delta) c(x, \delta) \pi(x) \min_{\substack{x,\delta \\ c(x,\delta)>0}} \left[c(x, \delta) - \mathbf{1}_{\delta \neq \delta^{-1}} c(\delta x, \delta) - \sum_{\eta \neq \delta, \delta^{-1}} \frac{(q - q_*)(\delta x, \delta^{-1}, \eta)}{c(x, \delta)\pi(x)} \right] . \tag{4.21}$$

Since the minimum in the above expression equals λ by (4.19) and the diagonal sum is at least $2\mathcal{A}(\rho, \psi)$ by Erbar et al. (2017)[Lemma 3.6], we obtain $\mathcal{B}(\rho, \psi) \geq 2\lambda\mathcal{A}(\rho, \psi)$.

□

Finally, we explain the connection between the above proof sketch and the Holley-Stroock perturbation condition. For the purpose of explanation, assume $\mathcal{X} = \{-1, +1\}^n$ is a spin system. Loosely speaking, the Holley-Stroock perturbation argument starts with a product probability measure on $\mathcal{X}$ and introduces interactions between spins such that the probability measure on $\mathcal{X}$ is not a product measure anymore. If the interactions are sufficiently weak, then functional inequalities like the logarithmic Sobolev inequality still hold. More concretely, note the Holley-Stroock perturbation result for the logarithmic-Sobolev inequality (LSI):

Proposition 1 (Bakry, Gentil, and Ledoux (2013b)[Proposition 5.1.6)**.** *]: Let μ be a probability measure that satisfies LSI with constant C^{μ}_{LSI}. Let ν be a probability measure with density e^k with*

respect to μ, then ν also satisfies LSI with constant

$$C_{\mathsf{LSI}}^{\nu} = e^{\mathsf{osc}(k)} C_{\mathsf{LSI}}^{\mu},$$

where $\mathsf{osc}(k) = \sup k - \inf k$.

Note that the condition $\lambda \geq 0$ in (4.19) ensures that the deviation from homogeneity, measured by $(q - q_*)$, is not too large relative to the transition rates $c(x, \delta)$. When this holds, positive Ricci curvature and hence the discrete Bochner's inequality holds despite the inhomogeneity of the rates. This transforms the abstract Bochner's characterization of Ricci curvature into a concrete, verifiable condition on the Markov chain's transition rates.

4.4 Main results

Zanella (2020) introduced a framework for designing Metropolis-Hastings proposals that incorporate information about the target distribution to propose local moves that have a higher target probability measure. In particular, a *locally balanced function* g is defined as a function that satisfies the property $g(t) = tg(1/t)$ in Zanella (2020). Next, consider a consider an irreducible continuous-time Markov chain Q on the state space $\mathcal{X} = \{-1, +1\}^n$, with stationary distribution π, and defined as:

$$Q(x, y) = \begin{cases} g(\pi(y)/\pi(x)) \text{ if } \|x - y\| = 2 \\ 0 \text{ otherwise}. \end{cases}$$

We consider the case where $g(t) = \sqrt{t}$. Note that this choice of g satisfies the locally balanced property. We denote the corresponding continuous-time Markov chain by Q_S to identify Q with the square-root locally balanced function, and choose the target distribution of Q_S to be $\pi_{\beta,p}$ as defined in (4.1). Our main results are Theorem 6 where we obtain a bound on the Ricci curvature of the Markov chain Q_S and Corollary **??** where we prove the mixing time of Q_S. We begin by stating a few lemmas which obtain bounds on certain counting objects and aid in proving Theorem 6.

Lemma 14. *Let n, p be positive integers with $n > 4, p \geq 2$, and $p < n$. Fix $i, j \in \{1, 2, \ldots, n\}$ where $i \neq j$. Let $S(n, p, i, j)$ denote the number of sequences $(i_1, i_2, \ldots, i_p)$, where each $i_k \in$*

$\{1, 2, \ldots, n\}$ *and both the indices* i, j *appear an odd number of times in* $(i_1, i_2, \ldots, i_p)$*. Then*

$$S(n, p, i, j) = \frac{n^p + (n-4)^p - 2(n-2)^p}{4}\,.$$

Additionally,

$$S(n, p, i, j) \le p(p-1)n^{p-2}\,.$$

Proof. We provide a probabilistic argument. We choose a sequence $(i_1, i_2, \ldots, i_p)$ uniformly at random from all n^p possible sequences. For each sequence $(i_1, i_2, \ldots, i_p)$, let the random variables X_i, X_j denote the number of occurrences of indices i, j in that sequence. Note that

$$\mathbf{1}\{X_i \text{ is odd and } X_j \text{ is odd}\} = \frac{1}{4}\left[1 + (-1)^{X_i+X_j} - (-1)^{X_i} - (-1)^{X_j}\right]\,.$$

By taking expectation above, we obtain:

$$\mathbb{P}(X_i \text{ is odd and } X_j \text{ is odd}) = \frac{1}{4}\left[1 + \mathbb{E}(-1)^{X_i+X_j} - \mathbb{E}(-1)^{X_i} - \mathbb{E}(-1)^{X_j}\right]\,.$$

Next, we compute $\mathbb{E}[(-1)^{X_i}]$. For each position in the sequence, the probability that it contains index i is $1/n$. Let Y_k be the indicator function for position k containing index i. So $X_i = \sum_{k=1}^{p} Y_k$. Since Y_k are independent random variables,

$$\mathbb{E}[(-1)^{X_i}] = \prod_{k=1}^{p} \mathbb{E}[(-1)^{Y_k}] = \left[(-1)\cdot\frac{1}{n} + 1\cdot\frac{n-1}{n}\right]^p = \left(\frac{n-2}{n}\right)^p\,.$$

By symmetry $\mathbb{E}[(-1)^{X_j}] = \left(\frac{n-2}{n}\right)^p$. Now we compute $\mathbb{E}[(-1)^{X_i+X_j}]$. Let Z_k be the indicator function for position k containing index j. For each position k, $Y_k + Z_k$ can be one or zero. So

$$\mathbb{E}[(-1)^{Y_k+Z_k}] = 1\cdot\mathbb{P}(Y_k + Z_k = 0) + (-1)\cdot\mathbb{P}(Y_k + Z_k = 1) = 1\cdot\frac{n-2}{n} + (-1)\cdot\frac{2}{n} = \frac{n-4}{n}\,.$$

Therefore, $\mathbb{E}[(-1)^{X_i+X_j}] = \left(\frac{n-4}{n}\right)^p$. This yields

$$S(n, p, i, j) = n^p\mathbb{P}(X_i \text{ is odd and } X_j \text{ is odd}) = \frac{n^p + (n-4)^p - 2(n-2)^p}{4}\,.$$

Next, we prove an upper bound for $S(n, p, i, j)$. Define $f(x) := x^p$, where $x \in \mathbb{R}$, and

$$g(t) := f(a+t) + f(a-t) - 2f(a)\,.$$

By Taylor's expansion of $g(t)$ around 0 along with Lagrange's remainder, we have that for all $h > 0$:

$$g(h) = g(0) + g'(0)h + \frac{h^2}{2}g''(\tau)\,,$$

for some $\tau \in (0, h)$. Note that by definition, $g(0) = g'(0) = 0$. Thus

$$g(h) = h^2\frac{f''(a+\tau) + f''(a-\tau)}{2}\,.$$

Since f'' is continuous for $p \geq 2$ by definition of f, by the intermediate value theorem there exists $\xi \in (a-h, a+h)$ such that:

$$g(h) = h^2 f''(\xi)\,.$$

By plugging $a = n-2$ and $h = 2$ into the above expression and using the definition of f, we obtain:

$$n^p + (n-4)^p - 2(n-2)^p = 4p(p-1)\xi^{p-2}\,,$$

for some $\xi \in (n-4, n)$. Finally, this means

$$n^p + (n-4)^p - 2(n-2)^p \leq 4p(p-1)n^{p-2}\,,$$

which finishes the proof of this lemma.

□

Lemma 15. *Let n, p be positive integers with $n > 2, p \geq 2$, and $p < n$. Fix $i, j \in \{1, 2, \ldots, n\}$ where $i \neq j$. Let $S'(n, p, i, j)$ denote the number of sequences $(i_1, i_2, \ldots, i_p)$, such that $i \in \{i_1, i_2, \ldots, i_p\}$ and $j \notin \{i_1, i_2, \ldots, i_p\}$ and index i appears an odd number of times. Then*

$$S'(n, p, i, j) = \frac{1}{2}\left[(n-1)^p - (n-3)^p\right]\,.$$

Additionally,

$$S'(n, p, i, j) \leq p(n-1)^{p-1}\,.$$

Proof. We again provide a probabilistic argument. We again choose a sequence $(i_1, i_2, \ldots, i_p)$ uniformly at random from all n^p possible sequences. For each sequence $(i_1, i_2, \ldots, i_p)$, let the random variables X_i, X_j denote the number of occurrences of indices i, j in that sequence.. We

find the probability $\mathbb{P}(X_i \bmod 2 = 1\,,\, X_j = 0)$ and then use $S'(n,p,i,j) = n^p \mathbb{P}(X_i \bmod 2 = 1\,,\, X_j = 0)$ to prove this lemma. Note that

$$\mathbf{1}\{X_i \bmod 2 = 1\} = \frac{1-(-1)^{X_i}}{2}\,.$$

Next, by taking expectation, we get:

$$\begin{aligned}\mathbb{P}(X_i \bmod 2 = 1\,,\, X_j = 0) &= \mathbb{E}[\mathbf{1}\{X_i \bmod 2 = 1\}\mathbf{1}\{X_j = 0\}] \\ &= \mathbb{E}\left[\frac{1-(-1)^{X_i}}{2}\mathbf{1}\{X_j = 0\}\right] \\ &= \frac{1}{2}\left[\mathbb{P}(X_j = 0) - \mathbb{E}[(-1)^{X_i}\mathbf{1}\{X_j = 0\}]\right]\,.\end{aligned}$$

Note that $\mathbb{P}(X_j = 0) = \frac{(n-1)^p}{n^p}$. To find the above second term, define Y_k to be the indicator for position k containing index i. Then:

$$\mathbb{E}[(-1)^{X_i} \mid X_j = 0] = \prod_{k=1}^{p} \mathbb{E}[(-1)^{Y_k} \mid X_j = 0] = \left((-1)\cdot\frac{1}{n-1} + \frac{n-2}{n-1}\right)^p = \frac{(n-3)^p}{(n-1)^p}\,.$$

This means $\mathbb{E}[(-1)^{X_i}\mathbf{1}\{X_j = 0\}] = \mathbb{E}[(-1)^{X_i} \mid X_j = 0]\mathbb{P}(X_j = 0) = \frac{(n-3)^p}{n^p}$. By substituting the above, we obtain:

$$\mathbb{P}(X_i \bmod 2 = 1\,,\, X_j = 0) = \frac{1}{2n^p}\left[(n-1)^p - (n-3)^p\right]\,,$$

which proves the expression for $S'(n,p,i,j)$. Using the Taylor's expansion of $(n-3)^p$ around $n-1$, we get:

$$(n-3)^p = ((n-1)-2)^p = \sum_{j=0}^{p} \frac{(-2)^j}{j!} p(p-1)\ldots(p-j+1)(n-1)^{p-j}\,,$$

and thus

$$\frac{1}{2}\left[(n-1)^p - (n-3)^p\right] = p(n-1)^{p-1} - p(p-1)(n-1)^{p-2} + \ldots\,,$$

which is an alternating series and the magnitudes of successive terms decrease. Therefore,

$$\frac{1}{2}\left[(n-1)^p - (n-3)^p\right] \le p(n-1)^{p-1}\,,$$

which finishes the proof of this lemma. □

Lemma 16. *Let n, p be positive integers with $n > 4, p \ge 2$, and $p < n$. Fix $i, j \in \{1, 2, \ldots, n\}$ where $i \neq j$. Let $S(n,p,i,j)$ denote the number of sequences $(i_1, i_2, \ldots, i_p)$, where each*

$i_k \in \{1, 2, \ldots, n\}$, *such that* $i \in (i_1, i_2, \ldots, i_p)$ *appears an even number of times and* $j \in (i_1, i_2, \ldots, i_p)$ *appears an odd number of times. Then*

$$S''(n, p, i, j) = \frac{n^p - (n-4)^p}{4}\,.$$

Additionally,

$$S''(n, p, i, j) \leq pn^{p-1}\,.$$

Proof. We again use a probabilistic approach. We choose a sequence $(i_1, i_2, \ldots, i_p)$ uniformly at random from all n^p possible sequences. For each sequence $(i_1, i_2, \ldots, i_p)$, let the random variables X_i, X_j denote the number of occurrences of indices i, j in that sequence. We find the probability $\mathbb{P}(X_i \bmod 2 = 0\,,\, X_j \bmod 2 = 1)$ and use $S''(n, p, i, j) = n^p \mathbb{P}(X_i \bmod 2 = 0\,,\, X_j \bmod 2 = 1)$ to prove this lemma. Note that

$$\mathbf{1}\{X_i \bmod 2 = 0\,,\, X_j \bmod 2 = 1\} = \frac{1}{4}\left[1 + (-1)^{X_i} - (-1)^{X_j} - (-1)^{X_i + X_j}\right]\,.$$

Taking expectation, we obtain:

$$\mathbb{P}(X_i \bmod 2 = 0\,,\, X_j \bmod 2 = 1) = \frac{1}{4}\left[1 + \mathbb{E}[(-1)^{X_i}] - \mathbb{E}[(-1)^{X_j}] - \mathbb{E}[(-1)^{X_i + X_j}]\right]\,.$$

Again, by defining Y_k, Z_k to be the indicators for position k containing index i, j respectively, we get:

$$\mathbb{E}[(-1)^{X_i}] = \prod_{k=1}^{p} \mathbb{E}[(-1)^{Y_k}] = \left((-1) \cdot \frac{1}{n} + \frac{n-1}{n}\right)^p = \left(\frac{n-2}{n}\right)^p\,.$$

By symmetry, $\mathbb{E}[(-1)^{X_j}] = \left(\frac{n-2}{n}\right)^p$. Next, similar to Lemma 14,

$$\mathbb{E}[(-1)^{X_i + X_j}] = \left(\frac{n-4}{n}\right)^p\,.$$

Substituting these expectations, we obtain:

$$\mathbb{P}(X_i \bmod 2 = 0\,,\, X_j \bmod 2 = 1) = \frac{n^p - (n-4)^p}{4n^p}\,,$$

and thus $S''(n, p, i, j) = \frac{n^p - (n-4)^p}{4}$. Finally, by the mean-value theorem, we get the upper bound on $S''(n, p, i, j)$. □

Theorem 6. *Consider the following continuous time Markov chain whose stationary distribution is the mean-field Tensor Ising model defined in* (4.1)*:*

$$Q_S(x,y) = \begin{cases} \sqrt{\frac{\pi_{\beta,p}(y)}{\pi_{\beta,p}(x)}} \text{ if } \|x-y\|_{\ell^1} = 2\,, \\ = 0 \text{ otherwise}\,. \end{cases}$$

Assume

$$\begin{aligned} \varepsilon_S(\beta) &:= (n-1)\exp\left(2\beta p + \frac{2\beta(n-1)^{p-1}}{(n)^{p-1}}\right)\left(\exp\left(\frac{2\beta p(p-1)}{n}\right)-1\right) \\ &< 1\,. \end{aligned} \tag{4.22}$$

Then the Ricci curvature of this Markov chain is bounded from below by:

$$\mathsf{Ric}(\mathcal{X}, Q_S, \pi_{\beta,p}) \geq 2(1-\varepsilon_S(\beta))\exp\left(\frac{\beta}{2n^{p-1}}\left((n-2)^p - n^p\right)\right)\,. \tag{4.23}$$

Proof. First, we define a mapping representation (G,c). Let $G = \{\delta_i : i = 1,2,\ldots,n\}$, where $\delta_i : \mathcal{X} \to \mathcal{X}$ is the map that flips the i^{th}-coordinate. Note that if we flip the spin at a coordinate twice we get the original spin, so $\delta_i^{-1} = \delta_i$ and $\delta_i\delta_j = \delta_j\delta_i$. Next, we define the transition rate c as:

$$\begin{aligned} c(x,\delta_i) &:= \sqrt{\frac{\pi_{\beta,p}(\delta_i x)}{\pi_{\beta,p}(x)}} \\ &= \exp\left(\frac{\beta}{2n^{p-1}}\left(H(\delta_i x) - H(x)\right)\right) \\ &= \exp\left(\frac{\beta}{2n^{p-1}}\Delta_i H(x)\right)\,, \end{aligned} \tag{4.24}$$

where $\Delta_i H(x) := H(\delta_i x) - H(x)$ is the difference in Hamiltonian when flipping spin i. Note that

$$\Delta_i H(x) = -x_i \sum_{\substack{1\leq i_1 i_2 \ldots i_p \leq n \\ i\in\{i_1,i_2,\ldots,i_p\}}} \left(1-(-1)^{\#\{k\in\{i_1,i_2,\ldots,i_p\}:\, i_k = i\}}\right) \prod_{j\,:\, i_j \neq i} x_{i_j}\,.$$

The next steps of the proof are designed to utilize Erbar et al. (2017)[Theorem 3.9]. First, we define the *probability flow*:

$$\begin{aligned} q(x,\delta_i,\delta_j) &:= c(x,\delta_i)c(x,\delta_j)\pi_{\beta,p}(x) \\ &= \exp\left(\frac{\beta}{2n^{p-1}}\left(H(\delta_i x) + H(\delta_j x)\right)\right)\,. \end{aligned} \tag{4.25}$$

Next, we define the term $q_*(x, \delta_i, \delta_j)$ as the minimum of four related q terms, similar to Erbar et al. (2017)[Equation 3.16]:

$$q_*(x, \delta_i, \delta_j) := \min\left\{q(x, \delta_i, \delta_j), q(\delta_i x, \delta_i^{-1}, \delta_j), q(\delta_j x, \delta_i, \delta_j^{-1}), q(\delta_i\delta_j x, \delta_i^{-1}, \delta_j^{-1})\right\}. \tag{4.26}$$

One way of interpreting the above four related q terms is that these correspond to the four corners of a "square" of states formed by applying the maps δ_i and δ_j (and their inverses) to x:

1. State x (original state)
2. State $\delta_i x$ (after flipping the spin in the i-th coordinate)
3. State $\delta_j x$ (after flipping the spin in the j-th coordinate)
4. State $\delta_i\delta_j x$ (after flipping the spins in the both the i-th and the j-th coordinate)

At each corner, we consider the probability flow of transitions along the two edges that lead to the opposite corner. Taking the minimum of these four terms identifies the "bottleneck" in traversing this square of states, which constrains how easily the system can move between these configurations.

Next, we do the following.

1. Compute the quantity:

$$\begin{aligned}\lambda &:= \min_{x\in\mathcal{X}\,,\,\delta_i\in G\,,\,c(x,\delta_i)>0}\left[c(x,\delta_i) - \mathbf{1}_{\delta_i\neq\delta_i^{-1}}c(\delta_i x,\delta_i) - \sum_{\delta_j\neq\delta_i\,,\,\delta_i^{-1}}\frac{(q-q_*)(\delta_i x,\delta_i^{-1},\delta_j)}{c(x,\delta_i)\pi_{\beta,p}(x)}\right]\\ &= \min_{x\in\mathcal{X}\,,\,\delta_i\in G\,,\,c(x,\delta_i)>0} c(x,\delta_i)\left[1 - \sum_{\delta_j\neq\delta_i}\frac{(q-q_*)(\delta_i x,\delta_i^{-1},\delta_j)}{q(x,\delta_i,\delta_i)}\right],\end{aligned} \tag{4.27}$$

   where the last step follows from the fact that $\delta_i = \delta_i^{-1}$ for all i and $c(x,\delta_i)^2\pi_{\beta,p}(x) = q(x,\delta_i,\delta_i)$.

2. Prove that $\lambda \geq 0$ so that Erbar et al. (2017)[Assumption 3.17] is satisfied and the statement of this theorem follows Erbar et al. (2017)[Theorem 3.9].

The critical step is estimating the term $(q - q_*)(\delta_i x, \delta_i^{-1}, \delta_j)$. Toward this end, we obtain an upper bound on $q(\delta_i x, \delta_i, \delta_j)$ and a lower bound on $q_*(\delta_i x, \delta_i, \delta_j)$. To simplify notation in the next steps, we define:

$$R_i := \#\{k \in \{i_1, i_2, \ldots, i_p\} : \ i_k = i\}\,.$$

By definition of $H(x)$ in (4.1), we get the following for $i \neq j$:

$$\begin{aligned}
H(x) = &\sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i_1,i_2,\ldots,i_p\neq i,j}} x_{i_1}x_{i_2}\ldots x_{i_p} + \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i\in\{i_1,i_2,\ldots,i_p\}\\ j\notin\{i_1,i_2,\ldots,i_p\}}} (x_i)^{R_i} \prod_{\substack{m:\, i_m\neq i\\ 1\le m\le p}} x_{i_m} + \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ j\in\{i_1,i_2,\ldots,i_p\}\\ i\notin\{i_1,i_2,\ldots,i_p\}}} (x_j)^{R_j} \prod_{\substack{m:\, i_m\neq j\\ 1\le m\le p}} x_{i_m}\\
&+ \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i,j\in\{i_1,i_2,\ldots,i_p\}}} (x_i)^{R_i}(x_j)^{R_j} \prod_{\substack{m:\, i_m\neq i,j\\ 1\le m\le p}} x_{i_m}\\
H(\delta_i x) = &\sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i_1,i_2,\ldots,i_p\neq i,j}} x_{i_1}x_{i_2}\ldots x_{i_p} + \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i\in\{i_1,i_2,\ldots,i_p\}\\ j\notin\{i_1,i_2,\ldots,i_p\}}} (-x_i)^{R_i} \prod_{\substack{m:\, i_m\neq i\\ 1\le m\le p}} x_{i_m} + \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ j\in\{i_1,i_2,\ldots,i_p\}\\ i\notin\{i_1,i_2,\ldots,i_p\}}} (x_j)^{R_j} \prod_{\substack{m:\, i_m\neq j\\ 1\le m\le p}} x_{i_m}\\
&+ \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i,j\in\{i_1,i_2,\ldots,i_p\}}} (-x_i)^{R_i}(x_j)^{R_j} \prod_{\substack{m:\, i_m\neq i,j\\ 1\le m\le p}} x_{i_m}\,.
\end{aligned} \tag{4.28}$$

By using the above two equations in (4.25), we obtain:

$$
\begin{aligned}
q(x,\delta_i,\delta_j) = \exp\left(\frac{2\beta}{2n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i_1,i_2,\ldots,i_p\neq i,j}} x_{i_1}x_{i_2}\ldots x_{i_p}\right)\times \\
\exp\left(\frac{\beta}{2n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i\in\{i_1,i_2,\ldots,i_p\}\\ j\notin\{i_1,i_2,\ldots,i_p\}}} \left((-x_i)^{R_i}+(x_i)^{R_i}\right) \prod_{\substack{m:\ i_m\neq i\\ 1\le m\le p}} x_{i_m}\right)\times \\
\exp\left(\frac{\beta}{2n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ j\in\{i_1,i_2,\ldots,i_p\}\\ i\notin\{i_1,i_2,\ldots,i_p\}}} \left((-x_j)^{R_j}+(x_j)^{R_j}\right) \prod_{\substack{m:\ i_m\neq j\\ 1\le m\le p}} x_{i_m}\right)\times \\
\exp\left(\frac{\beta}{2n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i,j\in\{i_1,i_2,\ldots,i_p\}}} \left((-x_i)^{R_i}(x_j)^{R_j}+(x_i)^{R_i}(-x_j)^{R_j}\right) \prod_{\substack{m:\ i_m\neq i,j\\ 1\le m\le p}} x_{i_m}\right).
\end{aligned}
\tag{4.29}
$$

Also note that for all $y\in\{x,\delta_i x,\delta_j x,\delta_i\delta_j x\}$:

$$
\begin{aligned}
q(y,\delta_i,\delta_j) = \exp\left(\frac{2\beta}{2n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i_1,i_2,\ldots,i_p\neq i,j}} x_{i_1}x_{i_2}\ldots x_{i_p}\right)\times \\
\exp\left(\frac{\beta}{2n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i\in\{i_1,i_2,\ldots,i_p\}\\ j\notin\{i_1,i_2,\ldots,i_p\}}} \left((-y_i)^{R_i}+(y_i)^{R_i}\right) \prod_{\substack{m:\ i_m\neq i\\ 1\le m\le p}} x_{i_m}\right)\times \\
\exp\left(\frac{\beta}{2n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ j\in\{i_1,i_2,\ldots,i_p\}\\ i\notin\{i_1,i_2,\ldots,i_p\}}} \left((-y_j)^{R_j}+(y_j)^{R_j}\right) \prod_{\substack{m:\ i_m\neq j\\ 1\le m\le p}} x_{i_m}\right)\times \\
\exp\left(\frac{\beta}{2n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i,j\in\{i_1,i_2,\ldots,i_p\}}} \left((-y_i)^{R_i}(y_j)^{R_j}+(y_i)^{R_i}(-y_j)^{R_j}\right) \prod_{\substack{m:\ i_m\neq i,j\\ 1\le m\le p}} x_{i_m}\right).
\end{aligned}
\tag{4.30}
$$

To simplify notation in the next steps, we introduce some new notation to separate cases when R_i, R_j are odd and even. When R_i, R_j are even, the above second and third in (4.30) simplify to the following. Define

$$E_2^i := \exp\left(\frac{\beta}{n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n \\ i\in\{i_1,i_2,\ldots,i_p\} \\ j\notin\{i_1,i_2,\ldots,i_p\} \\ R_i \bmod 2=0}} \prod_{\substack{m:\ i_m\neq i \\ 1\le m\le p}} x_{i_m}\right),\ E_3^j := \exp\left(\frac{\beta}{n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n \\ j\in\{i_1,i_2,\ldots,i_p\} \\ i\notin\{i_1,i_2,\ldots,i_p\} \\ R_j \bmod 2=0}} \prod_{\substack{m:\ i_m\neq j \\ 1\le m\le p}} x_{i_m}\right). \tag{4.31}$$

Similarly, note that when R_i, R_j are odd, the second and third terms in (4.30) simplify to $\exp(0) = 1$. For the fourth term in (4.30), note that when R_i is even and R_j is odd or vice-versa, then it simplifies to $\exp(0) = 1$. For the case where both R_i, R_j are even or odd, we define

$$\begin{aligned} EE_4^{ij} &:= \exp\left(\frac{\beta}{n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n \\ i,j\in\{i_1,i_2,\ldots,i_p\} \\ R_i \bmod 2\,, R_j \bmod 2=0}} \prod_{\substack{m:\ i_m\neq i,j \\ 1\le m\le p}} x_{i_m}\right) \\ OO_4^{ij}(y) &:= \exp\left(\frac{-\beta}{n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n \\ i,j\in\{i_1,i_2,\ldots,i_p\} \\ R_i \bmod 2\,, R_j \bmod 2=1}} (y_i y_j) \prod_{\substack{m:\ i_m\neq i,j \\ 1\le m\le p}} x_{i_m}\right). \end{aligned} \tag{4.32}$$

We also denote the first term in (4.30) as:

$$T_1 := \exp\left(\frac{\beta}{n^{p-1}} \sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n \\ i_1,i_2,\ldots,i_p\neq i,j}} x_{i_1} x_{i_2} \ldots x_{i_p}\right).$$

With this notation, $q(y, \delta_i, \delta_j)$ in (4.30) can be rewritten as:

$$q(y, \delta_i, \delta_j) = T_1 E_2^i E_3^j E E_4^{ij} O O_4^{ij}(y)\,. \tag{4.33}$$

Next, we analyze the term $q(x,\delta_i,\delta_i) = \exp\left(\frac{\beta}{n^{p-1}}H(\delta_i x)\right)$. We introduce more notation for further simplification.

$$\tilde{T}_2 = \exp\left(-\frac{\beta}{n^{p-1}}\sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i\in\{i_1,i_2,\ldots,i_p\}\\ j\notin\{i_1,i_2,\ldots,i_p\}\\ R_i \bmod 2=1}}\prod_{m=1}^{p} x_{i_m}\right),\ \tilde{T}_3 = \exp\left(\frac{\beta}{n^{p-1}}\sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ j\in\{i_1,i_2,\ldots,i_p\}\\ i\notin\{i_1,i_2,\ldots,i_p\}\\ R_j \bmod 2=1}}\prod_{m=1}^{p} x_{i_m}\right),$$

$$\tilde{T}_4 = \exp\left(\frac{\beta}{n^{p-1}}\sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i,j\in\{i_1,i_2,\ldots,i_p\}\\ R_i \bmod 2=0\,, R_j \bmod 2=1}}\prod_{\substack{m:\ i_m\neq i\\ 1\le m\le p}} x_{i_m}\right),\ \tilde{T}_4' = \exp\left(-\frac{\beta}{n^{p-1}}\sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i,j\in\{i_1,i_2,\ldots,i_p\}\\ R_i \bmod 2=1\,, R_j \bmod 2=0}}\prod_{\substack{m:\ i_m\neq j\\ 1\le m\le p}} x_{i_m}\right). \tag{4.34}$$

From the expression for $H(\delta_i x)$ in (4.29), by splitting the terms in $H(\delta_i x)$ over the cases where R_i, R_j are even or odd and by using the notation in (4.31), (4.32), (4.34), we obtain:

$$q(x,\delta_i,\delta_i) = T_1 E_2^i \tilde{T}_2 E_3^j \tilde{T}_3 E E_4^{ij} (OO_4^{ij}(\delta_i x))^{-1} \tilde{T}_4 \tilde{T}_4'\,. \tag{4.35}$$

Next, we find an upper bound for $(q-q_*)(\delta_i x,\delta_i,\delta_j)$ and a lower bound on $q(x,\delta_i,\delta_i)$ so that we get an upper bound on $\frac{(q-q_*)(\delta_i x,\delta_i,\delta_j)}{q(x,\delta_i,\delta_i)}$, which in turn can be plugged into (4.27). By observing that the terms $T_1, E_2^i, E_3^j, EE_4^{ij}$ do not depend on y, using the definition of $OO_4^{ij}(y)$ and the expression in (4.33), we get:

$$\begin{aligned}(q-q_*)(\delta_i x,\delta_i,\delta_j) &\le T_1 E_2^i E_3^j E E_4^{ij}\left(\exp\left(\frac{2\beta}{n^{p-1}}\sum_{\substack{1\le i_1,i_2,\ldots,i_p\le n\\ i,j\in\{i_1,i_2,\ldots,i_p\}\\ R_i \bmod 2\,, R_j \bmod 2=1}} 1\right)-1\right)\\ &\le T_1 E_2^i E_3^j E E_4^{ij}\left(\exp\frac{2\beta p(p-1)}{n}-1\right)\exp\left(-\frac{\beta p(p-1)}{n}\right),\end{aligned} \tag{4.36}$$

where the last inequality follows from Lemma 14. From (4.35), (4.36), we obtain:

$$\begin{aligned}\frac{(q-q_*)(\delta_i x,\delta_i,\delta_j)}{q(x,\delta_i,\delta_i)} &\le \tilde{T}_2\tilde{T}_3\tilde{T}_4\tilde{T}_4'\left(\exp\left(\frac{2\beta p(p-1)}{n}\right)-1\right)\\ &\le \exp\left(2\beta p+\frac{2\beta(n-1)^{p-1}}{(n)^{p-1}}\right)\left(\exp\left(\frac{2\beta p(p-1)}{n}\right)-1\right),\end{aligned}$$

where the first inequality follows from applying Lemma 14 to the factor $(OO_4^{ij}(\delta_i x))^{-1}$ in $q(x, \delta_i, \delta_i)$, and the second inequality follows from applying Lemma 15 to $\tilde{T}_2, \tilde{T}_3$ and Lemma 16 to $\tilde{T}_4, \tilde{T}_4'$. By plugging the above bound into (4.27), we get

$$\lambda = c_* \left[1 - (n-1) \exp\left(2\beta p + \frac{2\beta (n-1)^{p-1}}{(n)^{p-1}} \right) \left(\exp\left(\frac{2\beta p(p-1)}{n} \right) - 1 \right) \right] ,$$

where $c_* = \min_{x,i} c(x, \delta_i)$.

Finally, we find the minimum transition rate c_*. This corresponds to the transition rate obtained by flipping one spin in all positive ones or all negative ones spin configuration. Let $x^+ := (+1, +1, \ldots, +1)$ and $x_i^- = (+1, +1, \ldots, -1, \ldots, +1)$ where only the i-th spin is -1. Since $H(x^+) = n^p$ and by the Binomial theorem:

$$H(x_i^-) = \sum_{1 \le i_1, i_2, \ldots, i_p \le n} (-1)_i^R = \sum_{k=0}^{p} \binom{p}{k} (-1)^k (n-1)^{p-k} = (-1 + n - 1)^p = (n-2)^p ,$$

from (4.24), we obtain:

$$c_* = \exp\left(\frac{\beta}{2n^{p-1}} \left((n-2)^p - n^p \right) \right) . \tag{4.37}$$

By (4.22), we have that $\lambda \ge 0$. This means Erbar et al. (2017)[Equation 3.17] in Erbar et al. (2017)[Theorem 3.9] is satisfied. This fact along with the value of c_* shown above finishes the proof of this theorem by applying Erbar et al. (2017)[Theorem 3.9].

□

We introduce some definitions related to mixing before obtaining a bound on the mixing time of the continuous time Markov chain Q_S. First, consistent with Erbar and Maas (2012) and Erbar et al. (2017), define the set of probability densities on $\mathcal{X} = \{-1, +1\}^n$:

$$\mathcal{P}(\mathcal{X}) := \{ \rho : \mathcal{X} \to \mathbb{R}_+ \, : \, \sum_{x \in \mathcal{X}} \pi(x) \rho(x) = 1 \} . \tag{4.38}$$

Next, we define the total-variation distance between probability densities $\rho_0, \rho_1 \in \mathcal{P}(\mathcal{X})$, also consistent with Erbar and Maas (2012) and Erbar et al. (2017):

$$d_{\mathsf{TV}}(\rho_0, \rho_1) := \sum_{x \in \mathcal{X}} \pi(x) |\rho_0(x) - \rho_1(x)| . \tag{4.39}$$

Then, we define the mixing time as:

$$t_{\mathsf{mix}}(\varepsilon) = \inf\left\{t \geq 0 \; ; \; \max_{\rho_0 \in \mathcal{P}(\mathcal{X})} d_{\mathsf{TV}}(P_t\rho_0, \mathbf{1}) \leq \varepsilon\right\} .$$

We also introduce notation for Ricci curvature consistent with Erbar et al. (2017):

$$\kappa := \mathsf{Ric}(\mathcal{X}, Q_S, \pi_{\beta,p}) .$$

Theorem 7. *For any fixed integer $p \geq 2$, assume the inverse temperature parameter $\beta > 0$ of the mean-field Tensor Ising model defined in* (4.1) *is such that:*

$$\exp(2\beta(p+1))4\beta p(p-1) < 1 .$$

Then the mixing time of the continuous time Markov chain Q_S defined in Theorem 6 is given by

$$t_{\mathsf{mix}}(\varepsilon) \leq \frac{\exp(\beta p)}{2(1 - \exp(2\beta p + 2\beta)4\beta p(p-1))} \log\left(\frac{2\exp(\beta p)n}{\sqrt{2}\varepsilon}\right) .$$

Proof. Let $P_t = e^{tQ_S}$ be the continuous time Markov semigroup associated with the continuous time Markov chain Q_S. First, the total-variation distance can be bounded by $\mathcal{W}$ distance using Erbar and Maas (2012)[Proposition 2.12] as:

$$d_{\mathsf{TV}}(P_t\rho_0, \mathbf{1}) \leq \sqrt{2}\mathcal{W}(P_t\rho_0, \mathbf{1}) , \tag{4.40}$$

where $\rho_0 \in \mathcal{P}(\mathcal{X})$ is the initial probability density of the Markov chain Q_S. Next, since the Ricci curvature κ of the Markov chain Q_S is positive as shown in Theorem 6, we obtain from Erbar and Maas (2012)[Proposition 4.7] an exponential contraction in $\mathcal{W}$ distance defined in (4.7):

$$\mathcal{W}(P_t\rho_0, \mathbf{1}) \leq e^{-\kappa t}\mathcal{W}(\rho_0, \mathbf{1}) , \tag{4.41}$$

where $\kappa > 0$ is the Ricci curvature. From (4.40) and (4.41), the mixing time of Q_S can be bounded by

$$t_{\mathsf{mix}}(\varepsilon) \leq \frac{1}{\kappa} \log\left(\frac{D}{\sqrt{2}\varepsilon}\right) , \tag{4.42}$$

where $D = \max_{\rho_0 \in \mathcal{P}(\mathcal{X})} \mathcal{W}(\rho_0, \mathbf{1})$. Next, we find an upper bound on $1/\kappa$ using the bound on κ in Theorem 6. Since $e^x - 1 < 2x$ for $0 < x < 1$, we have that for $n > 2\beta p(p-1)$:

$$\exp\left(\frac{2\beta p(p-1)}{n}\right) - 1 < \frac{4\beta p(p-1)}{n} . \tag{4.43}$$

Also, for all $n > p \geq 2$:

$$\exp\left(2\beta p + \frac{2\beta(n-1)^{p-1}}{(n)^{p-1}}\right) < \exp(2\beta p + 2\beta)\,. \tag{4.44}$$

By plugging (4.43) and (4.44) into (4.22), we get the upper bound:

$$\begin{aligned}\varepsilon_S(\beta) &= (n-1)\exp\left(2\beta p + \frac{2\beta(n-1)^{p-1}}{(n)^{p-1}}\right)\left(\exp\left(\frac{2\beta p(p-1)}{n}\right) - 1\right) \\ &< \exp(2\beta p + 2\beta) 4\beta p(p-1)\frac{n-1}{n}\,.\end{aligned} \tag{4.45}$$

By the assumption $\exp(2\beta p + 2\beta)4\beta p(p-1) < 1$ in the statement of this corollary, we infer from the above bound that $\varepsilon_S(\beta) < 1$. Also, by the mean-value theorem, we have:

$$\exp\left(\frac{\beta}{2n^{p-1}}\left((n-2)^p - n^p\right)\right) > \exp(-\beta p)\,. \tag{4.46}$$

By plugging (4.45), (4.46) into (4.23), we obtain:

$$\kappa > 2(1 - \varepsilon_S(\beta)) > 2(1 - \exp(2\beta p + 2\beta)4\beta p(p-1))\exp(-\beta p)\,, \tag{4.47}$$

and by plugging the above bound into (4.42), we get:

$$t_{\mathsf{mix}}(\varepsilon) \leq \frac{\exp(\beta p)}{2(1 - \exp(2\beta p + 2\beta)4\beta p(p-1))}\log\left(\frac{D}{\sqrt{2}\varepsilon}\right)\,. \tag{4.48}$$

Next, we find an upper bound on $D = \max_{\rho_0 \in \mathcal{P}(\mathcal{X})} \mathcal{W}(\rho_0, \mathbf{1})$ using Erbar and Maas (2012)[Lemma 2.3]. By Erbar and Maas (2012)[Lemma 2.3], we have that for any $\rho_0 \in \mathcal{P}(\mathcal{X})$:

$$\mathcal{W}(\rho_0, \mathbf{1}) \leq \frac{v}{c_*} W_{2,g}(\rho_0, \mathbf{1})\,,$$

where the constant $v > 0$ is a constant, $c_* = \min_{x,i} c(x, \delta_i)$, and $W_{2,g}$ is the 2-Wasserstein distance with respect to the graph distance, as defined in Erbar and Maas (2012)[Equation 2.11]. Note that from Erbar and Maas (2012)[Lemma 2.13], the constant v satisfies $v < 2$ if θ is the logarithmic mean. Indeed, in Theorem 5, we assumed θ to be the logarithmic mean. Additionally, using the value of c_* obtained in (4.37) and its bound obtained in (4.46), yields

$$\mathcal{W}(\rho_0, \mathbf{1}) \leq 2\exp(\beta p) W_{2,g}(\rho_0, \mathbf{1})\,. \tag{4.49}$$

By definition, $Q_S(x, y) > 0$ if and only if $\|x - y\|_{\ell^1} = 2$, i.e. x, y differ in exactly one spin. Therefore, the graph distance induced by Q_S is the Hamming distance.

Next, by the convexity of the function $\rho \to W_{2,g}^2(\rho, \mathbf{1})$ and since $\rho_0 = \sum_{x\in\mathcal{X}} \rho_0(x)\mathbf{1}_x$, we obtain:

$$\begin{aligned} W_{2,g}^2(\rho_0, \mathbf{1}) &= W_{2,g}^2\left(\sum_{x\in\mathcal{X}} \rho_0(x)\mathbf{1}_x, \mathbf{1}\right) \\ &\leq \sum_{x\in\mathcal{X}} \rho_0(x)\pi(x) W_{2,g}^2\left(\frac{\mathbf{1}_x}{\pi(x)}, \mathbf{1}\right) \\ &= \sum_{x\in\mathcal{X}} \rho_0(x)\pi(x) \sum_{y\in\mathcal{X}} d_g(x,y)^2\pi(y)\,, \end{aligned} \tag{4.50}$$

where (4.50) follows from the definition of the distance $W_{2,g}$ in Erbar and Maas (2012)[Equation 2.11] and where d_g is the Hamming distance between $x, y \in \mathcal{X}$. Since $d_g(x,y) \leq n$, we get

$$W_{2,g}^2(\rho_0, \mathbf{1}) \leq \sum_{x\in\mathcal{X}} \rho_0(x)\pi(x)n^2 = n^2\,. \tag{4.51}$$

By plugging (4.51) into (4.49), we obtain:

$$\mathcal{W}(\rho_0, \mathbf{1}) \leq 2\exp(\beta p)n\,,$$

and therefore, $D = \max_{\rho_0\in\mathcal{P}(\mathcal{X})} \mathcal{W}(\rho_0, \mathbf{1}) \leq 2\exp(\beta p)n$. Finally, by plugging this bound on D into (4.48), we get:

$$t_{\mathsf{mix}}(\varepsilon) \leq \frac{\exp(\beta p)}{2(1-\exp(2\beta p + 2\beta)4\beta p(p-1))} \log\left(\frac{2\exp(\beta p)n}{\sqrt{2}\varepsilon}\right)\,.$$

This finishes the proof of this theorem. □

REFERENCES CITED

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

Ambrosio, L., & Savaré, G. (2007). Gradient flows of probability measures. In *Handbook of differential equations: evolutionary equations* (Vol. 3, pp. 1–136). Elsevier.

Bakry, D., Gentil, I., & Ledoux, M. (2013a). *Analysis and geometry of markov diffusion operators* (Vol. 348). Springer Science & Business Media.

Bakry, D., Gentil, I., & Ledoux, M. (2013b). *Analysis and geometry of markov diffusion operators* (Vol. 348). Springer Science & Business Media.

Bou-Rabee, N., & Schuh, K. (2023). Nonlinear hamiltonian monte carlo & its particle approximation. *arXiv preprint arXiv:2308.11491*.

Chen, F., Ren, Z., & Wang, S. (2022). Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv preprint arXiv:2212.03050*.

Chewi, S., Nitanda, A., & Zhang, M. S. (2024). Uniform-in-n log-sobolev inequality for the mean-field langevin dynamics with convex energy. *arXiv preprint arXiv:2409.10440*.

Chizat, L. (2022). Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*.

Chizat, L., & Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, *31*.

Das, A., & Nagaraj, D. (2023). Provably fast finite particle variants of svgd via virtual particle stochastic approximation. *Advances in Neural Information Processing Systems*, *36*, 49748–49760.

Daskalakis, C., Dikkala, N., & Panageas, I. (2020). Logistic regression with peer-group effects via inference in higher-order ising models. In *International conference on artificial intelligence and statistics* (pp. 3653–3663).

Duchi, J. (2007). Derivations for linear algebra and optimization. *Berkeley, California*, *3*(1), 2325–5870.

Durmus, A., Majewski, S., & Miasojedow, B. (2019). Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, *20*(73), 1–46.

Erbar, M., Henderson, C., Menz, G., & Tetali, P. (2017). Ricci curvature bounds for weakly interacting markov chains.

Erbar, M., & Maas, J. (2012). Ricci curvature of finite markov chains via convexity of the entropy. *Archive for Rational Mechanics and Analysis*, *206*, 997–1038.

Jordan, R., Kinderlehrer, D., & Otto, F. (1998). The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, *29*(1), 1–17.

Kook, Y., Zhang, M. S., Chewi, S., Erdogdu, M. A., & Li, M. B. (2024). Sampling from the mean-field stationary distribution. In *The thirty seventh annual conference on learning theory* (pp. 3099–3136).

Lacker, D., & Le Flem, L. (2023). Sharp uniform-in-time propagation of chaos. *Probability Theory and Related Fields*, *187*(1-2), 443–480.

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

Levin, D. A., Luczak, M. J., & Peres, Y. (2010). Glauber dynamics for the mean-field ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, *146*, 223–265.

Levin, D. A., & Peres, Y. (2017). *Markov chains and mixing times* (Vol. 107). American Mathematical Soc.

Maas, J. (2011). Gradient flows of the entropy for finite markov chains. *Journal of Functional Analysis*, *261*(8), 2250–2292.

Malrieu, F. (2001). Logarithmic sobolev inequalities for some nonlinear pde's. *Stochastic processes and their applications*, *95*(1), 109–132.

Malrieu, F. (2003). Convergence to equilibrium for granular media equations and their euler schemes. *The Annals of Applied Probability*, *13*(2), 540–560.

McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in mathematics*, *128*(1), 153–179.

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

Nitanda, A. (2024). Improved particle approximation error for mean field neural networks. *arXiv preprint arXiv:2405.15767*.

Nitanda, A., & Suzuki, T. (2017). Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*.

Nitanda, A., Wu, D., & Suzuki, T. (2022). Convex analysis of the mean field langevin dynamics. In *International conference on artificial intelligence and statistics* (pp. 9741–9757).

Otto, F., & Villani, C. (2000). Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, *173*(2), 361–400.

Ren, P., & Wang, F.-Y. (2021). Exponential convergence in entropy and wasserstein for mckean–vlasov sdes. *Nonlinear Analysis*, *206*, 112259.

Samanta, R. J., Mukherjee, S., & Zhang, J. (2024). Mixing phases of the glauber dynamics for the p-spin curie-weiss model. *arXiv preprint arXiv:2412.16952*.

Suzuki, T., Nitanda, A., & Wu, D. (2023). Uniform-in-time propagation of chaos for the mean-field gradient langevin dynamics. In *The eleventh international conference on learning representations.*

Suzuki, T., Wu, D., & Nitanda, A. (2024). Mean-field langevin dynamics: Time-space discretization, stochastic gradient, and variance reduction. *Advances in Neural Information Processing Systems*, *36*.

Sznitman, A.-S. (1991). Topics in propagation of chaos. *Ecole d'été de probabilités de Saint-Flour XIX—1989*, *1464*, 165–251.

Tankala, C., Nagaraj, D. M., & Raj, A. (2025). Beyond propagation of chaos: A stochastic algorithm for mean field optimization. *arXiv preprint arXiv:2503.13115*.

Vempala, S., & Wibisono, A. (2019). Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, *32*.

Wang, S. (2024). Uniform log-sobolev inequalities for mean field particles with flat-convex energy. *arXiv preprint arXiv:2408.03283*.

Yan, Y., Wang, K., & Rigollet, P. (2024). Learning gaussian mixtures using the wasserstein–fisher–rao gradient flow. *The Annals of Statistics*, *52*(4), 1774–1795.

Yao, R., & Yang, Y. (2022). Mean field variational inference via wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*.

Zanella, G. (2020). Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, *115*(530), 852–865.