Quantitative Local Convergence of Mean-Field Stein Variational Gradient Flow

Lénaïc Chizat*
Department of Mathematics
National University of Singapore
10 Lower Kent Ridge Road
Singapore 119076
lenaic.chizat@nus.edu.sg

Maria Colombo*
Institute of Mathematics
EPFL
Station 8
1015 Lausanne, Switzerland
maria.colombo@epfl.ch

Roberto Colombo*
Institute of Mathematics
EPFL
Station 8
1015 Lausanne, Switzerland
roberto.colombo@epfl.ch

Xavier Fernández-Real*
Institute of Mathematics
EPFL
Station 8
1015 Lausanne, Switzerland
xavier.fernandez-real@epfl.ch

Abstract

Stein Variational Gradient Descent (SVGD) is a deterministic interacting-particle method for sampling from a target probability measure given access to its score function. In the mean-field and continuous-time limit, it is known that the flow converges weakly toward the target, but no quantitative rate is known for the last iterate. In this paper, we establish quantitative local convergence in strong norms for this dynamics, when the interaction kernel is of Riesz type on the d-dimensional torus. Specifically, assuming that the initial density and the target are smooth and close in L^2-norm, we obtain explicit polynomial convergence rates in L^2-norm that depend on the dimension and on the regularity parameters of the kernel, the initialization and the target. We further show that these rates are sharp in certain regimes, and support the theory with numerical experiments. In the edge case of kernels with a Coulomb singularity, we recover the global exponential convergence result established in prior work. Our analysis is inspired by recent results on Wasserstein gradient flows of kernel mean discrepancies.

1 Introduction

Stein Variational Gradient Descent (SVGD), introduced in [LW16], is a deterministic interacting-particle method for sampling from a target probability measure $\pi \propto e^{-V}$, requiring only access to ∇V. In the mean-field and continuous-time limit, the distribution of particles converges to a flow (ρ_t) in the space of probability measures that solves a variant of the Fokker-Planck equation with a velocity field smoothed by weighted convolution with a positive definite kernel [LLN19]. This flow can be interpreted as the gradient flow of the relative entropy $\mathcal{H}(\cdot|\pi)$ with respect to a "kernelized" Wasserstein metric [Liu17].

The goal of this paper is to investigate the convergence of (ρ_t) towards π. To this end, we focus on the model case of Riesz kernels of order s on the d-dimensional torus $\mathbb{T}^d$. This is a family of translation-invariant kernels whose Fourier coefficients decay as $|\xi|^{-2s}$. The parameter s hence directly controls

*Authors are listed in alphabetical order.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the "smoothing strength" of the interaction; in particular, continuous kernels correspond to $s > d/2$, C^1 kernels to $s > (d+1)/2$, and C^2 kernels to $s > (d+2)/2$.

What is known: qualitative weak convergence The starting point of convergence analyses is the energy dissipation formula [Liu17]

$$\frac{d}{dt}\mathcal{H}(\rho_t|\pi) = -I_s(\rho_t|\pi), \tag{1.1}$$

where $I_s \geq 0$ is the so-called kernel Stein discrepancy (KSD), see (2.3). In our setting, $I_s(\rho|\pi)$ is comparable to the squared negative Sobolev norm $\|\rho - \pi\|^2_{H^{1-s}}$ (see Lemma 4.2), which vanishes iff $\rho = \pi$, so π is the only equilibrium of the dynamics among finite-entropy solutions satisfying (1.1). Since the flow (ρ_t) is also weakly compact, one may expect that ρ_t converges weakly to π, which was indeed shown in [KSA+20, Prop. 2] for smooth potentials and kernels. In Proposition 2.1, we give an alternative proof of this qualitative weak convergence under slightly weaker regularity assumptions.

What is not known: quantitative strong convergence Such a result gives, however, limited information about the finite-time behavior. For instance, one cannot exclude a priori that $I_s(\rho_t|\pi)$ remains small for an arbitrarily long time and then increases significantly. Several works overcome this difficulty by considering the averaged iterate $\bar{\rho}_t = \frac{1}{t}\int_0^t \rho_u du$ which, by (1.1) and convexity of $I_s(\cdot|\pi)$, satisfies the quantitative rate

$$I_s(\bar{\rho}_t|\pi) \leq \mathcal{H}(\rho_0|\pi)/t. \tag{1.2}$$

The best iterate ρ_t^* over $[0,t]$ also satisfies (1.2). However, $\bar{\rho}_t$ and ρ_t^* are impractical to compute and rarely used. Another limitation of this class of results is that I_s is a weak notion of discrepancy, especially in high dimension, and it depends on the choice of kernel, making it difficult to compare the behavior for various kernels. Moreover, there is no indication that the rate $O(1/t)$ is tight. We therefore ask the following question:

Can one obtain sharp quantitative convergence rates for the mean-field SVGD flow in strong norms?

Main results Our main contribution (Theorem 2.2) answers this question in a local sense. Namely, for all $s > 1$, assuming that the initial density and the target are smooth and close to each other in L^2-norm, we obtain explicit polynomial convergence rates in L^2-norm that depend on s, d, and on the regularity parameters of the initialization and the target. The polynomial rate in Theorem 2.2 is sharp when the target π is uniform. We discuss in Remark 2.4 an improvement of the rate for general targets, which we conjecture to be sharp in all regimes where it applies.

The edge case $s = 1$ corresponds to kernels with a Coulomb singularity on the diagonal[2]. In this case, both the energy and the energy dissipation are comparable to $\|\rho_t - \pi\|^2_{L^2}$, so by (1.1) and Grönwall's lemma, one deduces global convergence in L^2 at an exponential rate. This result was already established in the inspiring work [CSW24] for kernels with a Coulomb singularity and suitable tails on $\mathbb{R}^d$. In Theorem 2.5, we recover a compact analogue of their result with a short proof, and in addition, we show the global well-posedness of the flow in L^∞ via a maximum principle.

Note, however, that the practical implementation of SVGD requires, at least, that the gradient of the kernel is locally bounded, which holds iff $s \geq (d+1)/2$. So in dimension $d \geq 2$, the practically relevant regime is that of Theorem 2.2, where the energy dissipation involves a norm weaker than the energy itself, preventing a global Łojasiewicz-type inequality. We overcome this difficulty by showing refined propagation of regularity estimates; see Section 2.3 for the proof idea. Our analysis is inspired by recent results on Wasserstein gradient flows of kernel mean discrepancies [CCCFR26].

From a practical viewpoint, Theorem 2.2 sheds new light on the behavior of SVGD. It is the first result that yields quantitative guarantees depending (sharply) on properties of the kernel and the target[3]. In particular, the rates from Remark 2.4 suggest that a good choice of kernel should have s as small as possible while maintaining stability of the discrete scheme, e.g. $s = (d+1)/2$. This corresponds to kernels behaving like $-\mathrm{dist}(x,y)$ near the diagonal. In particular, Gaussian-like

[2]Namely, the kernel explodes near the diagonal as $\mathrm{dist}(x,y)^{-d+2}$ when $d \geq 3$.

[3]This is, to a certain extent, analogous to the rates expressed under source/capacity conditions in the literature on kernel regression with gradient descent [YRC07]. A major difference, however, is that the dynamical system we study is non-linear.

kernels, for which the Fourier coefficients decay super-polynomially, are expected to lead to very slow convergence rates in L^2, because higher frequency components of $\rho_t - \pi$ dissipate extremely slowly. This is a spectral heuristic, outside the scope of our theorems.

Related work The influential works [LW16, Liu17] introduced SVGD along with its mean-field limit, energy decay formula (1.1) and gradient flow interpretation. The rigorous mean-field limit was studied in [LLN19] and several works have investigated its properties [DNS23, NR23, CLGL+20]. Other works have focused on the discrete-time dynamics [SSR22, SR22] and/or the particle dynamics [CS25, BBG24, HBBG26]. In those settings, the state-of-the-art analyses rely on (1.2) or its discrete variants. We note also that variants of SVGD with noise have been proposed in [PBS24, GI18]: the noise tends to facilitate convergence because it makes the dynamics more similar to the overdamped Langevin dynamics, for which unconditional exponential convergence holds. Finally, some undesirable behaviors for SVGD in high dimension are known, such as a "mode collapse" phenomenon [BEG+21]. Our work highlights another problem: unless the regularity of the target increases with the dimension, the convergence guarantee in L^2 degrades exponentially with the dimension, by the conjectured sharpness of the rate in Remark 2.4.

Organisation In Section 2, we detail our setting, state the main results, and discuss their implications. Numerical experiments in synthetic settings are in Section 3, and Section 4 contains the proof, with some lemmas postponed to the supplementary materials. We conclude in Section 5.

2 Setting and main results

2.1 Setting

Let $d \geq 1$, and let $\mathbb{T}^d$ be the d-dimensional torus. For any $s \geq 1$, we denote by K_s the zero-mean Riesz kernel of order s solving $(-\Delta)^s K_s = \delta_0 - 1$ in the sense of distributions. We refer to [CCCFR26, Appendix A.1] for a detailed description of K_s.

Let $V : \mathbb{T}^d \to \mathbb{R}$ be a sufficiently regular potential, and consider the associated probability measure

$$\pi = \frac{e^{-V}}{Z} \in \mathcal{P}(\mathbb{T}^d), \qquad Z := \int_{\mathbb{T}^d} e^{-V}. \tag{2.1}$$

We study solutions to the following equation, which we refer to as the Mean-Field Stein Variational Gradient Flow (mean-field SVGF):

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho v) &= 0 & \text{in } (0,T) \times \mathbb{T}^d, \\ v_t &= -K_s * \left(\rho_t \nabla \log\left(\frac{\rho_t}{\pi}\right)\right) & \forall t \in (0,T), \\ \rho_0 &= \bar{\rho}. \end{cases} \tag{2.2}$$

Let $\mathcal{H}(\rho|\pi) := \int_{\mathbb{T}^d} \log\left(\frac{\rho}{\pi}\right) \rho$ denote the relative entropy of ρ with respect to π. Then, a solution ρ_t of (2.2) formally satisfies the following dissipation identity:

$$\frac{d}{dt}\mathcal{H}(\rho_t|\pi) = -\int_{\mathbb{T}^d}\int_{\mathbb{T}^d} K_s(x-y)\nabla\frac{\rho_t}{\pi}(x) \cdot \nabla\frac{\rho_t}{\pi}(y)\pi(x)\pi(y) = -I_s(\rho_t|\pi), \tag{2.3}$$

where the last equality defines $I_s(\rho|\pi)$, known as the kernel Stein discrepancy (KSD) or as the Stein relative Fisher information. As detailed in Lemma 4.2, it holds $I_s(\rho|\pi) \approx \|\rho - \pi\|^2_{\dot{H}^{1-s}}$ and therefore I_s vanishes iff $\rho = \pi$. In particular, π is the unique equilibrium point of the dynamics among finite-entropy solutions satisfying (2.3).

We begin with the qualitative weak convergence of the dynamics, which is an adaptation of [KSA+20, Prop. 2] to our setting. While [KSA+20] obtains this result by differentiating $\mathcal{H}(\cdot|\pi)$ twice along the flow, and therefore requires the potential V and the kernel K of regularity C^2, we use a different proof technique, exploiting the gradient flow structure, which requires less regularity.

We write $\mathcal{P}(\mathbb{T}^d)$ for the probability measures on $\mathbb{T}^d$. The space of curves $C([0,T);\mathcal{P}(\mathbb{T}^d))$ is understood with the weak topology of measures, and solutions of the continuity equation are understood in the sense of distributions.

Proposition 2.1. *Let $s > \frac{d}{2}$, $V \in C^1(\mathbb{T}^d)$, and let $\rho \in C([0,\infty); \mathcal{P}(\mathbb{T}^d))$ be a global solution of (2.2) with $\mathcal{H}(\bar{\rho}|\pi) < \infty$, satisfying the energy dissipation inequality*

$$\mathcal{H}(\rho_T|\pi) + \int_0^T I_s(\rho_t|\pi)\, dt \le \mathcal{H}(\bar{\rho}|\pi) \qquad \forall T \ge 0.$$

Then ρ_t converges weakly to π as $t \to \infty$.

Before stating our main results, we introduce some essential notation and briefly discuss local well-posedness of the system (2.2). The Fourier coefficients of a general distribution f on the d-torus $\mathbb{T}^d$ are given by $\widehat{f}_k := \langle f, e^{2\pi i k\cdot x}\rangle$, where $k \in \mathbb{Z}^d$. For any $\beta \in \mathbb{R}$, inhomogeneous and homogeneous β-Sobolev norms are defined, respectively, as follows:

$$\|f\|_{H^\beta}^2 := \sum_{k\in\mathbb{Z}^d} (1 + |2\pi k|^2)^\beta |\widehat{f}_k|^2, \qquad \|f\|_{\dot{H}^\beta}^2 := \sum_{k\in\mathbb{Z}^d\setminus\{0\}} |2\pi k|^{2\beta} |\widehat{f}_k|^2. \tag{2.4}$$

We denote by $H^\beta(\mathbb{T}^d)$ the space of distributions f for which $\|f\|_{H^\beta} < \infty$. For negative β, high frequencies receive smaller weights. If X is a Banach space of functions on $\mathbb{T}^d$, $L^\infty_{\mathrm{loc}}([0,T); X)$ denotes the set of curves in X with norm bounded on every compact time interval in $[0,T)$.

By the arguments detailed in [CCCFR26, Section 2], one gets in particular the following well-posedness (existence and uniqueness/stability) results for weak solutions to (2.2). For $s \ge 1$ and $\gamma > d/2$, suppose that $V \in H^{\gamma+s}(\mathbb{T}^d)$. Then, for every $\bar{\rho} \in \mathcal{P} \cap H^\gamma(\mathbb{T}^d)$, there is a unique solution $\rho \in C([0,T); \mathcal{P}(\mathbb{T}^d)) \cap L^\infty_{\mathrm{loc}}([0,T); H^\gamma(\mathbb{T}^d))$ defined up to some maximal time $T > 0$. Moreover, if $T < \infty$, then $\limsup_{t\to T^-} \|\rho_t\|_{H^\gamma} = \infty$. When $s = 1$, for every $\bar{\rho} \in L^\infty \cap \mathcal{P}(\mathbb{T}^d)$ there is a unique maximal locally bounded solution $\rho \in C([0,T); \mathcal{P}(\mathbb{T}^d)) \cap L^\infty_{\mathrm{loc}}([0,T); L^\infty(\mathbb{T}^d))$ and $T < \infty$ implies $\limsup_{t\to T^-} \|\rho_t\|_{L^\infty} = \infty$. Solutions are stable with respect to variation of the data, and propagate higher regularity: smooth data generate smooth solutions.

2.2 Statement of the main results

Our main result is the following quantitative local convergence in the case $s > 1$:

Theorem 2.2 (Local polynomial convergence for $s > 1$)**.** *Let $s > 1$ and $\gamma > \max\{\frac{d}{2}, s-1\}$. Let $V \in H^{\gamma+s}(\mathbb{T}^d)$ and $\pi \in \mathcal{P} \cap H^{\gamma+s}(\mathbb{T}^d)$ be defined as in (2.1). Let $\bar{\rho} \in \mathcal{P} \cap H^\gamma(\mathbb{T}^d)$, and let $\rho \in C([0,T); \mathcal{P}(\mathbb{T}^d)) \cap L^\infty_{\mathrm{loc}}([0,T); H^\gamma(\mathbb{T}^d))$ be the solution of (2.2) extended up to the maximal time of existence $T > 0$. There are constants $C > 0$ and $\delta > 0$ depending only on d, s, γ, $\|V\|_{H^{\gamma+s}}$, and $\|\bar{\rho}\|_{H^\gamma}$ such that, if $\mathcal{H}(\bar{\rho}|\pi) \le \delta$, then $T = \infty$, and*

$$\mathcal{H}(\rho_t|\pi) \le \left(\mathcal{H}(\bar{\rho}|\pi)^{-\frac{s-1}{\gamma}} + \frac{t}{C}\right)^{-\frac{\gamma}{s-1}} \qquad \forall t \ge 0. \tag{2.5}$$

Moreover, for every $\beta \in [0,\gamma]$ we have

$$\|\rho_t - \pi\|_{\dot{H}^\beta}^2 \lesssim \|\bar{\rho} - \pi\|_{\dot{H}^\gamma}^{2\frac{\beta}{\gamma}} \left(\mathcal{H}(\bar{\rho}|\pi)^{-\frac{s-1}{\gamma}} + \frac{t}{C}\right)^{-\frac{\gamma-\beta}{s-1}} \qquad \forall t \ge 0. \tag{2.6}$$

Some comments are in order. First, the Sobolev bounds are fixed before choosing the neighborhood of π. Indeed, Sobolev embedding gives an upper bound Λ for $\bar{\rho}$ and π, while $\lambda := \min \pi > 0$; Lemma 4.1 then gives $(2\Lambda)^{-1}\|\bar{\rho} - \pi\|_{L^2}^2 \le \mathcal{H}(\bar{\rho}|\pi) \le \lambda^{-1}\|\bar{\rho} - \pi\|_{L^2}^2$. Thus entropy-smallness and L^2-smallness are equivalent within this bounded regularity class. Second, each estimate is for a fixed finite γ: both C and the admissible threshold δ depend on γ and the corresponding Sobolev bounds. Even for smooth data, letting $\gamma \to \infty$ does not establish super-polynomial convergence for a fixed initialization.

The proof has three ingredients: comparison of the KSD $I_s(\rho_t|\pi)$ with the squared negative Sobolev norm $\|\rho_t - \pi\|_{\dot{H}^{1-s}}^2$, interpolation to obtain a local Łojasiewicz inequality for the dissipated relative entropy $\mathcal{H}(\rho_t|\pi)$ under uniform boundedness of a higher Sobolev norm, and propagation of the high Sobolev norm to close a bootstrap argument. We give further details on the proof and its relation to the KMD flow in Section 2.3.

Remark 2.3 (Sharpness of the rate for uniform targets). Under a control of the initial datum in H^γ, the exponent $-\gamma/(s-1)$ is already sharp for the uniform target $\pi = 1$. As in [CCCFR26, Rmk. 1.6], we check this for the zero-mean quantity σ_t that solves the linearized equation

$$\partial_t \sigma_t = -(-\Delta)^{1-s}\sigma_t,$$

obtained by writing $\rho = 1 + \sigma$ in (2.2) and retaining only the terms linear in σ. The Fourier coefficients of σ_t evolve as

$$\widehat{\sigma}_k(t) = \widehat{\sigma}_k(0)\exp\left(-|2\pi k|^{2-2s}t\right).$$

Taking $\sigma_0^n(x) = \sqrt{2}(2\pi n)^{-\gamma}\cos(2\pi n x_1)$ gives $\|\sigma_0^n\|_{\dot{H}^\gamma} = 1$ and, at time $t = n^{2s-2}$,

$$\|\sigma_t^n\|_{L^2}^2 \gtrsim n^{-2\gamma} \simeq t^{-\frac{\gamma}{s-1}}.$$

Thus neither the squared L^2-norm nor, by Lemma 4.1, the relative entropy can decay faster for the linearized flow, uniformly over a bounded $\dot{H}^\gamma$ class.

Remark 2.4 (Conjecture for tight rates). By arguing as in [CCCFR26, Remark 4.8], one can replace the assumption $V \in H^{\gamma+s}$ in Theorem 2.2 by $V \in H^{\gamma+\varepsilon}$ for some $\varepsilon > 0$, at least when γ and s are integers (we expect the same to be true for fractional exponents, but the proof would require a further technical generalization of the Kato-Ponce estimate in Lemma A.2). This suggests that the rate is governed by the least regular of V and $\bar{\rho}$: if we define $\gamma^* = \sup\{\gamma : V, \bar{\rho} \in H^\gamma\}$ and if $\max\{d/2, s-1\} < \gamma^* < \infty$, then for any sufficiently small $\varepsilon > 0$, and sufficiently small initial entropy depending on ε and the corresponding Sobolev bounds, it holds

$$\mathcal{H}(\rho_t|\pi) \le \left(\mathcal{H}(\bar{\rho}|\pi)^{-\frac{s-1}{\gamma^*-\varepsilon}} + \frac{t}{C_\varepsilon}\right)^{-\frac{\gamma^*-\varepsilon}{s-1}}. \tag{2.7}$$

Motivated by the previous remark and numerical experiments (see Figure 1b where this rate with $\varepsilon = 0$ is reported in dotted lines), we conjecture that the $-\gamma^*/(s-1)$ exponent is sharp.

As discussed in the introduction, the energy dissipation (2.3) directly implies global exponential convergence in the case $s = 1$. This result was previously obtained in [CSW24] on $\mathbb{R}^d$, where the control on tail behavior makes the analysis more difficult. Here we give a short proof in the compact case, and in addition we show the global well-posedness of the flow in L^∞ via a maximum principle:

Theorem 2.5 (Global exponential convergence for $s = 1$)**.** *Let $s = 1$ and $\gamma > \frac{d}{2}$, $\gamma \ge 1$. Let $V \in H^{\gamma+1}(\mathbb{T}^d)$ and $\pi \in \mathcal{P} \cap H^\gamma(\mathbb{T}^d)$ be defined as in (2.1). Then, there exists $\alpha > 0$ depending only on d and $\|V\|_{H^{\gamma+1}}$ such that for every $\bar{\rho} \in \mathcal{P} \cap L^\infty(\mathbb{T}^d)$, there is a unique global-in-time, uniformly bounded solution $\rho \in C([0,\infty);\mathcal{P}(\mathbb{T}^d)) \cap L^\infty([0,\infty)\times\mathbb{T}^d)$ of (2.2), and*

$$\mathcal{H}(\rho_t|\pi) \le \mathcal{H}(\bar{\rho}|\pi)e^{-\alpha t} \qquad \forall t \ge 0. \tag{2.8}$$

Under the assumptions of Theorem 2.5, there is a unique solution $\rho \in C([0,T);\mathcal{P}(\mathbb{T}^d)) \cap L^\infty_{\mathrm{loc}}([0,T);L^\infty(\mathbb{T}^d))$ defined up to some maximal time $T > 0$. In Lemma D.1 we prove that this solution is uniformly bounded, which implies, in particular, $T = \infty$.

Extension to other kernels Although Theorem 2.2 is formulated for Riesz kernels, the proof only uses structural properties of its Fourier multiplier. In particular, the argument is stable under replacing $(-\Delta)^{-s}$ by a symmetric convolution operator $\mathcal{K}f = K * f$ with the same order, the same L^p multiplier bounds, and the same leading dissipative part up to lower-order errors, as shown in Theorem E.1. As an illustration, the conclusions of Theorem 2.2 also hold for the "negative distance-like" kernel

$$K(x) = (1 - 4|x|)_+^{d+2}, \tag{2.9}$$

in the regime $s = (d+1)/2$, with $d \ge 2$; see Remark E.2. This kernel can be conveniently evaluated in any dimension, in contrast to Riesz kernels, which have no closed form on $\mathbb{T}^d$. The same spectral criterion also covers periodized Matérn kernels; see Remark E.3.

2.3 Link with Wasserstein gradient flows of KMD and ideas of the proof

Our analysis is closely related to the recent work [CCCFR26] on Wasserstein gradient flows of Kernel Mean Discrepancy (KMD, or MMD) functionals. In that setting, given a target measure ν on a compact manifold M, one considers

$$\mathscr{E}^{\nu}(\mu) := \frac{1}{2}\int_M \int_M K(x,y)d(\mu-\nu)(x)d(\mu-\nu)(y),$$

and its Wasserstein gradient flow

$$\partial_t \mu_t = \operatorname{div}\left(\mu_t \nabla \mathcal{K}\left(\mu_t - \nu\right)\right) \quad \text{in } (0,T)\times M, \quad \text{where} \quad \mathcal{K}(\eta)(x) := \int_M K(x,y)d\eta(y). \tag{2.10}$$

For the Riesz kernels on $\mathbb{T}^d$, $\mathcal{K} = (-\Delta)^{-s}$ and $\mathscr{E}^{\nu}(\mu) = \frac{1}{2}\|\mu-\nu\|^2_{\dot{H}^{-s}}$. The equation studied here has the same active-scalar structure (see (4.1) below). In fact, when V is constant, our dynamics coincide exactly with the KMD flow (2.10) with target $\nu = 1$; otherwise, the additional drift term $-D^{-2s}((\rho_t - \pi)\nabla V)$ appears.

There is, however, an important difference in the Lyapunov structure. In the KMD flow, the decreasing energy is the weak Sobolev discrepancy $\mathscr{E}^{\nu}(\mu_t) = \frac{1}{2}\|\mu_t - \nu\|^2_{\dot{H}^{-s}}$. In the SVGD flow, the decreasing functional is the relative entropy, and the dissipation is the kernel Stein discrepancy, which under the fixed lower bound on π satisfies (see Lemma 4.2)

$$I_s(\rho|\pi) = \left\| \pi \nabla \frac{\rho-\pi}{\pi} \right\|^2_{\dot{H}^{-s}} \approx \|\rho-\pi\|^2_{\dot{H}^{1-s}}.$$

This coercivity depends only on the reference density π, and not on a lower bound for the evolving density ρ_t. This is a major difference with respect to the KMD dissipation, where coercivity is tied to lower bounds on μ_t.

The quantitative decay follows from a local Łojasiewicz inequality along the flow. Combining the previous comparison with the entropy bound (4.2) in terms of $\|\rho-\pi\|_{L^2}$, and using Sobolev interpolation, for $s > 1$ we obtain the following: if $\sigma = \rho - \pi \in H^{\gamma}$ for some $\gamma > \max\{d/2, s-1\}$, then

$$I_s(\rho|\pi) \gtrsim \mathcal{H}(\rho|\pi)^{1+\frac{s-1}{\gamma}} \left(\|\rho-\pi\|^2_{\dot{H}^{\gamma}}\right)^{-\frac{s-1}{\gamma}}. \tag{2.11}$$

When $s = 1$, the high norm H^{γ} drops out: $I_1(\rho|\pi) \gtrsim \|\rho-\pi\|^2_{L^2} \gtrsim \mathcal{H}(\rho|\pi)$, and the entropy dissipation identity immediately gives exponential decay by Grönwall. The maximum principle in Lemma D.1 supplies a uniform L^{∞} bound; the convergence mechanism itself does not require imposing a positive lower bound on ρ_t. The proof of the maximum principle for SVGD must control the nonlocal drift, whose divergence has no fixed sign at a maximum of ρ_t; the elementary maximum principle for the KMD flow does not apply directly.

For $s > 1$, the Łojasiewicz inequality (2.11) is effective only if the high Sobolev norm H^{γ} stays controlled. This is obtained using the strategy of [CCCFR26]: the high Sobolev norm is not expected to dissipate monotonically, but its growth can be controlled. Specifically, Lemma 4.3 propagates $\|\sigma_t\|_{\dot{H}^{\gamma}}$ up to the integral of a Lipschitz norm of the velocity field, and Lemma 4.4 bounds this integral by a small power of the initial entropy. By a bootstrap argument, this shows that the solution remains for all times in a region where a Łojasiewicz inequality holds, yielding the polynomial entropy decay of Theorem 2.2. The Sobolev decay follows by interpolation.

3 Numerical experiments

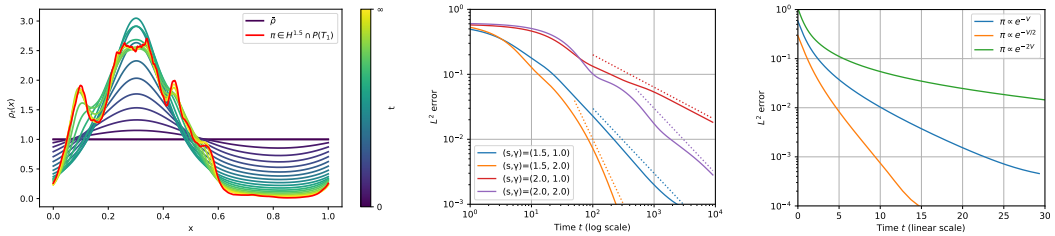
We perform numerical experiments in two settings[4]. On the 1D torus (see Figure 1), we solve the mean-field PDE (2.2) directly. Since the equation can be written in divergence form, we use an upwind finite-volume scheme that preserves mass and positivity, and the velocity field is computed using Fourier multipliers. We use a grid of 2048 points and adaptive CFL step sizes.

On the 2D torus (see Figure 2), we instead implement the fully discrete SVGD particle iterations. For computational efficiency, we use the "negative-distance" kernel (2.9) rather than a Riesz kernel,

[4]Code repository: https://github.com/lchizat/2026-svgd-local-rate

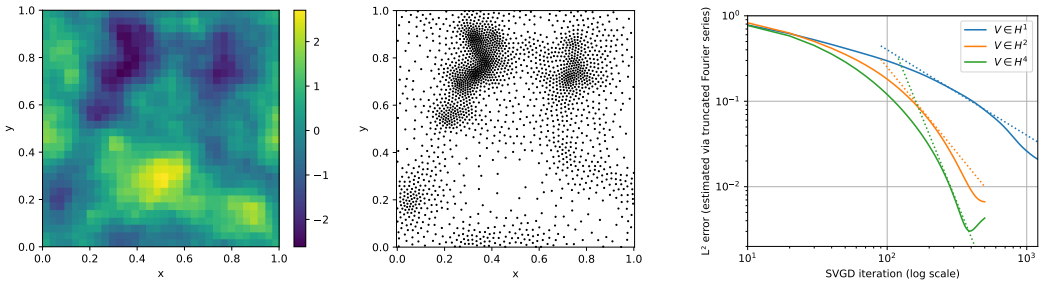
whose evaluation is more costly since it does not have a closed-form expression. This kernel has the same spectral decay as a Riesz kernel with parameter $s = (d+1)/2$ and our results apply; see Appendix E. The L^2-error is estimated through a truncated Fourier expansion. In this setting, the Fourier truncation error, together with finite-particle fluctuations, makes the convergence rates harder to read off cleanly but we can still observe our predictions at a more qualitative level[5]. We use 2000 particles, a step size of 0.05, and the 16×16 lowest Fourier modes for the L^2-error estimation.

In both settings, we consider uniform initialization $\bar{\rho} = 1$, and we generate a random potential V of the desired Sobolev regularity by sampling its Fourier coefficients with centered Gaussian weights with the appropriate variance decay. This allows us to directly control the largest value of γ^* such that $V \in H^{\gamma^*}$ that appears in the convergence rate (2.7).



(a) Evolution of the flow ($s = 2$) (b) Polynomial rates for $s > 1$ (c) Exponential rates for $s = 1$

Figure 1: Mean-field SVGF in 1D, solved with a finite-volume method (upwind scheme). (a) Evolution of the density for $s = 2$ and target in $H^{1.5}$. (b) For $s > 1$, the experimental errors (solid lines) and (the square root of) the predicted rates of (2.7) (dotted lines) are in good agreement. (c) For $s = 1$, convergence is exponential, with a rate depending on the potential V, as in Theorem 2.5.



(a) Random potential $V \in H^1$ (b) Output of SVGD (c) Convergence rates for $s = 1.5$

Figure 2: SVGD in 2D solved via the interacting particle system, with a distance-like kernel ($s = 1.5$). (a) A realization of a Gaussian process V with H^1 regularity. (b) Output of SVGD for this potential V. (c) Decay of the estimated L^2-error for various target regularities, compared with the (square-root of the) predicted rates of (2.7) (dotted lines). The agreement is weaker than in Figure 1, in particular because of the difficulty of accurately estimating the mean-field L^2-error using particles.

4 Proof of the main results

4.1 Basic estimates on relative entropy and kernel Stein discrepancy

In the sequel we will use the notation $D^\beta = (-\Delta)^{\beta/2}$ to denote the multiplier operator defined by the formula

$$D^\beta f = \sum_{k \in \mathbb{Z}^d \setminus \{0\}} |2\pi k|^\beta \widehat{f}_k e^{2\pi i k \cdot x}.$$

[5]We stop plotting the curves once discretization errors visibly dominate. These errors may, however, already affect the slopes before that point, since they are not uniform across Fourier modes. In particular, the curves artificially accelerate when the leading residual error is carried by high-frequency modes that are suppressed by the truncated Fourier estimator of the L^2-norm.

The operators D^β are self-adjoint, and commute with standard derivatives. Moreover, homogeneous Sobolev norms can be written as

$$\|f\|_{\dot{H}^\beta} = \|D^\beta f\|_{L^2}.$$

For any $s \geq 1$, the Riesz kernel K_s is defined by the expansion

$$K_s := \sum_{k \in \mathbb{Z}^d \setminus \{0\}} |2\pi k|^{-2s} e^{2\pi i k \cdot x}.$$

In particular, the convolution operator $K_s *$ corresponds precisely to $D^{-2s} = (-\Delta)^{-s}$.

The velocity field generated by a solution ρ_t of (2.2) can be rewritten in one of the following equivalent forms:

$$v_t = -K_s * \left(\rho_t \nabla \log\left(\frac{\rho_t}{\pi}\right)\right) = -D^{-2s}(\pi \nabla \frac{\rho_t - \pi}{\pi}) = -\nabla D^{-2s}(\rho_t - \pi) - D^{-2s}((\rho_t - \pi)\nabla V), \tag{4.1}$$

where the last identity was derived using that ∇ commutes with D^{-2s}, and $-\frac{\nabla \pi}{\pi} = \nabla V$.

The first ingredient is the elementary comparability between relative entropy and the squared L^2-norm of the difference between two densities enjoying suitable uniform bounds:

Lemma 4.1. *Let $\pi \in \mathcal{P}(\mathbb{T}^d)$ be such that $0 < \lambda \leq \pi \ll \mathcal{L}^d$. Then, the following bound holds for all $\rho \in \mathcal{P}(\mathbb{T}^d)$:*

$$\mathcal{H}(\rho|\pi) \leq \frac{1}{\lambda}\|\rho - \pi\|_{L^2}^2. \tag{4.2}$$

If in addition $\rho, \pi \leq \Lambda$, then

$$\mathcal{H}(\rho|\pi) \geq \frac{1}{2\Lambda}\|\rho - \pi\|_{L^2}^2. \tag{4.3}$$

Proof. Since ρ, π are both probability measures, we may write

$$\mathcal{H}(\rho|\pi) = \int_{\mathbb{T}^d} \left(\log\left(\frac{\rho}{\pi}\right)\frac{\rho}{\pi} - \frac{\rho}{\pi} + 1\right)\pi.$$

The entropy function $\varphi(r) := \log(r)r - r + 1$ satisfies $\varphi(r) \leq (r-1)^2$. Therefore

$$\mathcal{H}(\rho|\pi) \leq \int_{\mathbb{T}^d} \left(\frac{\rho}{\pi} - 1\right)^2 \pi \leq \frac{1}{\lambda}\|\rho - \pi\|_{L^2}^2.$$

Moreover, it holds $\varphi(r) \geq \frac{1}{2\max\{r,1\}}(r-1)^2$. Therefore, if $\rho, \pi \leq \Lambda$,

$$\mathcal{H}(\rho|\pi) \geq \int_{\mathbb{T}^d} \frac{1}{2\max\{\frac{\rho}{\pi}, 1\}}\left(\frac{\rho}{\pi} - 1\right)^2 \pi = \int_{\mathbb{T}^d} \frac{(\rho - \pi)^2}{2\max\{\rho, \pi\}} \geq \frac{1}{2\Lambda}\|\rho - \pi\|_{L^2}^2. \qquad \square$$

We next identify the kernel Stein discrepancy with the homogeneous Sobolev norm naturally associated with the Riesz kernel.

Lemma 4.2. *Let $s \geq 1$, and let $\pi \in \mathcal{P} \cap H^\alpha(\mathbb{T}^d)$ with $\alpha > \frac{d}{2}$ and $\alpha \geq s$. Suppose that $\pi \geq \lambda > 0$. Then, the following holds for every $\rho \in \mathcal{P}(\mathbb{T}^d)$:*

$$I_s(\rho|\pi) \approx_{d,s,\alpha,\lambda,\|\pi\|_{H^\alpha}} \|\rho - \pi\|_{\dot{H}^{1-s}}^2. \tag{4.4}$$

In particular, if $s > 1$ and $\rho - \pi \in H^\gamma(\mathbb{T}^d)$ for some $\gamma > 0$, then

$$I_s(\rho|\pi) \gtrsim_{d,s,\alpha,\gamma,\lambda,\|\pi\|_{H^\alpha}} \mathcal{H}(\rho|\pi)^{1+\frac{s-1}{\gamma}} \left(\|\rho - \pi\|_{\dot{H}^\gamma}^2\right)^{-\frac{s-1}{\gamma}}. \tag{4.5}$$

Proof. We consider the zero-mean measure $\sigma := \rho - \pi$. Note that, by the definition of KSD, applying Lemma A.1 twice with multipliers π and π^{-1}, and zero-mean functions $\nabla\frac{\sigma}{\pi}$, σ, respectively, we deduce

$$I_s(\rho|\pi) = \|\pi\nabla\frac{\sigma}{\pi}\|_{\dot{H}^{-s}}^2 \approx_{d,s,\alpha,\lambda,\|\pi\|_{H^\alpha}} \|\nabla\frac{\sigma}{\pi}\|_{\dot{H}^{-s}}^2 = \|\frac{\sigma}{\pi}\|_{\dot{H}^{1-s}}^2 \approx_{d,s,\alpha,\lambda,\|\pi\|_{H^\alpha}} \|\sigma\|_{\dot{H}^{1-s}}^2,$$

which gives (4.4). Assuming $\rho - \pi \in H^\gamma(\mathbb{T}^d)$ for some $\gamma > 0$, (4.2) and Sobolev interpolation of homogeneous norms give

$$\mathcal{H}(\rho|\pi) \lesssim_\lambda \|\sigma\|_{L^2}^2 \leq \|\sigma\|_{\dot{H}^{1-s}}^{\frac{2\gamma}{\gamma+s-1}} \|\sigma\|_{\dot{H}^\gamma}^{\frac{2(s-1)}{\gamma+s-1}} \lesssim_{d,s,\alpha,\gamma,\lambda,\|\pi\|_{H^\alpha}} I_s(\rho|\pi)^{\frac{\gamma}{\gamma+s-1}} \|\sigma\|_{\dot{H}^\gamma}^{\frac{2(s-1)}{\gamma+s-1}},$$

which rewrites as (4.5). $\square$

4.2 Controls on propagation of regularity

Let us now state the key technical estimates which ensure that the flow does not lose regularity too fast. Their proofs can be found in Appendix C. The main a priori estimate is the following high-order energy inequality, which isolates the dissipative linear part from the transport terms.

Lemma 4.3. *Let $s > 1$ and $\gamma > \frac{d}{2}$. Let $V \in H^{\gamma+s}(\mathbb{T}^d)$ and $\pi \in \mathcal{P} \cap H^{\gamma+s}(\mathbb{T}^d)$ be defined as in (2.1). Let $\rho \in C([0,T);\mathcal{P}(\mathbb{T}^d)) \cap L^\infty_{\mathrm{loc}}([0,T);H^\gamma(\mathbb{T}^d))$ be a solution of (2.2) defined up to some time $T > 0$, and call $\sigma_t = \rho_t - \pi$. Then, there is a constant $C > 0$ depending only on d, s, γ, and $\|V\|_{H^{\gamma+s}}$ such that the following estimate holds for all $t \in (0,T)$:*

$$\begin{aligned} &\|\sigma_t\|^2_{\dot{H}^\gamma} + 2\big(\min_{\mathbb{T}^d} \pi\big) \int_0^t \|\sigma_r\|^2_{\dot{H}^{\gamma-s+1}} \exp\left(C\int_r^t L(u)\,du\right) dr \\ &\leq \|\sigma_0\|^2_{\dot{H}^\gamma} \exp\left(C\int_0^t L(u)\,du\right) + C\int_0^t \|\sigma_r\|_{\dot{H}^{\gamma-s}} \|\sigma_r\|_{\dot{H}^{\gamma-s+1}} \exp\left(C\int_r^t L(u)\,du\right) dr, \end{aligned} \tag{4.6}$$

where we set

$$L(t) := \|\nabla^2 D^{-2s}\sigma_t\|_{L^\infty} + \|\nabla D^{-2s}(\sigma_t \nabla V)\|_{L^\infty}. \tag{4.7}$$

It remains to control the cumulative effect of the Lipschitz norm $L(t)$.

Lemma 4.4. *Under the same assumptions and using the same notation as in Lemma 4.3, assume further that $\gamma > s - 1$ and $\|\sigma_t\|^2_{\dot{H}^\gamma} \leq M$ for all $t \in [0,T)$. Then, there is $\beta = \beta(d,s,\gamma) > 0$ such that*

$$\int_0^T L(t)\,dt \lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}},M} \mathcal{H}(\bar{\rho}|\pi)^\beta.$$

4.3 Proof of the main results

We are now ready to prove our main result, Theorem 2.2:

Proof of Theorem 2.2. If $\bar{\rho} = \pi$, then the solution is stationary. We may therefore assume $\mathcal{H}(\bar{\rho}|\pi) > 0$. Let $T > 0$ be the maximal time of existence in $C([0,T);\mathcal{P}(\mathbb{T}^d)) \cap L^\infty_{\mathrm{loc}}([0,T);H^\gamma(\mathbb{T}^d))$ of the solution ρ_t, and let $\sigma_t := \rho_t - \pi$. Set $M := \max\{1, 2\|\bar{\rho} - \pi\|^2_{\dot{H}^\gamma}\}$ and consider the first time $\tau > 0$ such that $\|\sigma_t\|^2_{\dot{H}^\gamma}$ exceeds M, i.e.

$$\tau := \sup\left\{t \in (0,T] : \|\sigma_r\|^2_{\dot{H}^\gamma} < M,\ \forall r \in [0,t)\right\} > 0.$$

Combining (2.3) with (4.5), from the definition of τ we deduce that

$$\frac{d}{dt}\mathcal{H}(\rho_t|\pi) \lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}},M} -\mathcal{H}(\rho_t|\pi)^{1+\frac{s-1}{\gamma}} \qquad \forall t \in (0,\tau),$$

which integrated yields

$$\mathcal{H}(\rho_t|\pi) \leq \left(\mathcal{H}(\bar{\rho}|\pi)^{-\frac{s-1}{\gamma}} + \frac{t}{C}\right)^{-\frac{\gamma}{s-1}} \qquad \forall t \in [0,\tau),$$

where $C = C(d,s,\gamma,\|V\|_{H^{\gamma+s}},M) > 0$. The conclusion of the Theorem would follow if we could prove that $\tau = T$. In fact, in this case, by the definition of τ, the solution would be uniformly bounded in $H^\gamma(\mathbb{T}^d)$ up to the maximal time of existence, thus forcing $\tau = T = \infty$.

First observe that for any given $\varepsilon > 0$, Sobolev interpolation of homogeneous norms, together with Young's inequality, yields

$$\|\sigma_t\|_{\dot{H}^{\gamma-s}} \|\sigma_t\|_{\dot{H}^{\gamma-s+1}} \leq \varepsilon \|\sigma_t\|^2_{\dot{H}^{\gamma-s+1}} + C(\gamma,s,\varepsilon)\|\sigma_t\|^2_{\dot{H}^{1-s}}.$$

Choosing ε sufficiently small depending only on d, s, γ and $\|V\|_{H^{\gamma+s}}$ we may then absorb the term with coefficient ε into the dissipative term on the left-hand side of (4.6), up to adding a multiple of $\|\sigma_t\|^2_{\dot{H}^{1-s}}$. In addition, with $L(t)$ as in (4.7), by Lemma 4.4, for some $\beta = \beta(d,s,\gamma) > 0$ we have

$$\int_0^\tau L(t)\,dt \lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}},M} \mathcal{H}(\bar{\rho}|\pi)^\beta. \tag{4.8}$$

The integral estimate (4.6), together with (4.4), then yields, for all $t \in [0, \tau)$,

$$\begin{aligned}\|\sigma_t\|_{\dot{H}^\gamma}^2 &\le \|\bar{\rho} - \pi\|_{\dot{H}^\gamma}^2 \exp\left(C \int_0^t L(r)\, dr\right) + C \int_0^t I_s(\rho_r|\pi) \exp\left(C \int_r^t L(u)\, du\right) dr \\ &\le \left(\|\bar{\rho} - \pi\|_{\dot{H}^\gamma}^2 + C \int_0^t I_s(\rho_r|\pi)\, dr\right) \exp\left(C \int_0^t L(r)\, dr\right) \\ &\le \left(\frac{M}{2} + C\mathcal{H}(\bar{\rho}|\pi)\right) \exp\left(C\mathcal{H}(\bar{\rho}|\pi)^\beta\right), \end{aligned} \tag{4.9}$$

where in the third step we used (4.8), the dissipation identity (2.3), and the definition of M. If $\mathcal{H}(\bar{\rho}|\pi)$ is chosen sufficiently small, then the right-hand side of the expression above is strictly less than $M \ge 1$. By the definition of τ, this implies that $\tau = T$, concluding the proof of global existence and (2.5). In (4.9) we also obtained that $\sup_{t\ge 0}\|\sigma_t\|_{\dot{H}^\gamma}^2 \lesssim \|\bar{\rho} - \pi\|_{\dot{H}^\gamma}^2$. Since, by Lemma 4.1, the uniform H^γ bound, and Sobolev embedding, $\mathcal{H}(\rho_t|\pi) \approx \|\sigma_t\|_{L^2}^2$, (2.6) follows directly from Sobolev interpolation. □

In the edge case $s = 1$, once global well-posedness is established, exponential convergence follows directly as follows.

Proof of Theorem 2.5. The solution is global thanks to Lemma D.1. Then, by (2.3), (4.4), and (4.2), we find

$$\frac{d}{dt}\mathcal{H}(\rho_t|\pi) = -I_s(\rho_t|\pi) \lesssim_{d,\gamma,\|V\|_{H^\gamma}} -\|\rho_t - \pi\|_{L^2}^2 \le -\big(\min_{\mathbb{T}^d} \pi\big)\mathcal{H}(\rho_t|\pi) \qquad \forall t > 0.$$

Therefore, (2.8) follows from Grönwall's inequality. □

5 Conclusion and limitations

We have provided a sharp local convergence analysis of SVGD in the continuous-time mean-field limit. Our results clarify the role of the kernel through its spectral decay and highlight how the regularity of the target affects the convergence rates.

Several limitations of our analysis call for further investigation:

- **From local to global.** Except in the case $s = 1$, our analysis is local: it requires the initial density and target to be sufficiently close in L^2, with fixed Sobolev bounds. In general, it remains unclear whether global L^2 convergence holds. In fact, even global well-posedness remains open in the range $s \in (1, \frac{d}{2} + 1)$. Weak convergence alone does not imply entry into the local regime. Loss of regularity could also change the asymptotic rate, although (1.2) remains available for global finite-entropy solutions. A possible route toward global results is to inject noise into the dynamics, as in [PBS24, GI18]. Even an arbitrarily small amount of noise may yield uniform regularity estimates, allowing the solution to eventually enter the local regime studied here.
- **From compact to non-compact domains.** We have worked on the torus in order to isolate the spectral mechanism behind the convergence rates, a viewpoint that has been relatively underexploited in the SVGD literature. Extending the analysis to $\mathbb{R}^d$ is a natural next step. On the torus, $\min \pi > 0$ enters both the entropy–L^2 comparison and the multiplier estimates for the KSD. On $\mathbb{R}^d$, even Gaussian targets have no positive uniform lower bound, so these arguments require weighted estimates and control of tails. The noncompact Coulomb analysis of [CSW24] addresses this additional difficulty.
- **From continuum to discrete dynamics.** As discussed in the introduction, existing analyses of practical SVGD algorithms typically rely on analogues of the weak energy estimate (1.2). It would be interesting to adapt the sharper approach developed here to discrete particle settings. A particle approximation resolves primarily low frequencies. We expect the continuum rates to describe a regime before finite particle and Fourier truncation errors dominate, with sufficiently small time steps. Quantifying this regime and its dependence on particle number and step size remains open.

Acknowledgments and Disclosure of Funding

Funding. M.C., R.C., and X.F. are supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number MB22.00034 through the project TENSE. R.C. and X.F. are also supported by the Swiss National Science Foundation (SNF grant PZ00P2_208930). X.F. is further supported by the AEI project PID2021-125021NA-I00 (Spain).

Competing interests. The authors declare no competing interests.

References

[Ask73] Richard Askey. Radial characteristic functions. *Mathematical Research Center, University of Wisconsin–Madison, Technical Summary Report*, (1262), 1973.

[BBG24] Sayan Banerjee, Krishnakumar Balasubramanian, and Promit Ghosal. Improved finite-particle convergence rates for Stein variational gradient descent. *arXiv preprint arXiv:2409.08469*, 2024.

[BEG+21] Jimmy Ba, Murat A Erdogdu, Marzyeh Ghassemi, Shengyang Sun, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Understanding the variance collapse of SVGD in high dimensions. In *International Conference on Learning Representations*, 2021.

[BM18] Haïm Brezis and Petru Mironescu. Gagliardo–Nirenberg inequalities and non–inequalities: the full story. *Ann. Inst. H. Poincaré C Anal. Non Linéaire*, 35(5):1355–1376, 2018.

[CCCFR26] Lénaïc Chizat, Maria Colombo, Roberto Colombo, and Xavier Fernández-Real. Quantitative convergence of Wasserstein gradient flows of Kernel Mean Discrepancies. *arXiv preprint arXiv:2603.01977*, 2026.

[CGW23] José A Carrillo, Rishabh S Gvalani, and Jeremy S-H Wu. An invariance principle for gradient flows in the space of probability measures. *Journal of Differential Equations*, 345:233–284, 2023.

[CLGL+20] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109, 2020.

[CS25] José Carrillo and Jakub Skrzeczkowski. Convergence and stability results for the particle system in the Stein gradient descent method. *Mathematics of Computation*, 94(354):1793–1814, 2025.

[CSW24] José A Carrillo, Jakub Skrzeczkowski, and Jethro Warnett. The Stein-log-Sobolev inequality and the exponential rate of convergence for the continuous Stein variational gradient descent method. *arXiv preprint arXiv:2412.10295*, 2024.

[DNS23] Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023.

[GI18] Victor Gallego and David Rios Insua. Stochastic gradient MCMC with repulsive forces. *arXiv preprint arXiv:1812.00071*, 2018.

[HBBG26] Ye He, Krishnakumar Balasubramanian, Sayan Banerjee, and Promit Ghosal. Finite-particle rates for regularized Stein variational gradient descent. *arXiv preprint arXiv:2602.05172*, 2026.

[KSA+20] Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:4672–4682, 2020.

[Li19] Dong Li. On Kato–Ponce and fractional Leibniz. *Rev. Mat. Iberoam.*, 35(1):23–100, 2019.

[Liu17] Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in Neural Information Processing Systems*, 30, 2017.

[LLN19] Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019.

[LW16] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

[NR23] Nikolas Nüsken and DR Michiel Renger. Stein variational gradient descent: Many-particle and long-time asymptotics. *Foundations of Data Science*, 5(3):286–320, 2023.

[PBS24] Victor Priser, Pascal Bianchi, and Adil Salim. Long-time asymptotics of noisy SVGD outside the population limit. *arXiv preprint arXiv:2406.11929*, 2024.

[SR22] Lukang Sun and Peter Richtárik. A note on the convergence of mirrored Stein variational gradient descent under (L_0, L_1)-Smoothness condition. *arXiv preprint arXiv:2206.09709*, 2022.

[SSR22] Adil Salim, Lukang Sun, and Peter Richtarik. A convergence theory for SVGD in the population limit under Talagrand's inequality T_1. In *International Conference on Machine Learning*, pages 19139–19152. PMLR, 2022.

[YRC07] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive approximation*, 26(2):289–315, 2007.

Appendix

Contents of the appendix:

- Appendix A: Auxiliary estimates
- Appendix B: Qualitative convergence for Stein Variational Gradient Descent
- Appendix C: Proof of the propagation of regularity estimates
- Appendix D: Maximum principle for $s = 1$
- Appendix E: More general kernels

A Auxiliary estimates

We shall repeatedly use the following multiplier estimate to insert smooth fixed weights in homogeneous Sobolev norms.

Lemma A.1. *Let $\alpha, \beta \in \mathbb{R}$ be such that $\alpha > \frac{d}{2}$ and $\alpha \geq |\beta|$. Let $f \in H^\beta(\mathbb{T}^d)$ be such that $\widehat{f}_0 = 0$, and let $g \in H^\alpha(\mathbb{T}^d)$. Then, the following bound holds:*

$$\|gf\|_{\dot{H}^\beta} \lesssim_{d,\alpha,\beta,\|g\|_{H^\alpha}} \|f\|_{\dot{H}^\beta}.$$

Assume in addition that $g \geq \lambda > 0$. Then

$$\|f\|_{\dot{H}^\beta} \lesssim_{d,\alpha,\beta,\lambda,\|g\|_{H^\alpha}} \|gf\|_{\dot{H}^\beta}.$$

Proof. The first claim follows from the following chain of inequalities:

$$\|gf\|_{\dot{H}^\beta} \lesssim_{d,\beta} \|gf\|_{H^\beta} \lesssim_{d,\alpha,\beta,\|g\|_{H^\alpha}} \|f\|_{H^\beta} \lesssim_{d,\beta} \|f\|_{\dot{H}^\beta},$$

where in the penultimate step we used the fact that multiplication by g is a bounded operator[6] in $H^\beta(\mathbb{T}^d)$, while in the last step we used that f has zero average.

[6] Indeed, $\|fg\|_{L^2} \leq \|g\|_{L^\infty}\|f\|_{L^2} \lesssim_{d,\alpha} \|g\|_{H^\alpha}\|f\|_{L^2}$, by the Sobolev embedding, since $\alpha > \frac{d}{2}$. Moreover, $\|fg\|_{H^\alpha} \lesssim_{d,\alpha} \|g\|_{H^\alpha}\|f\|_{H^\alpha}$, again because of $\alpha > \frac{d}{2}$. By interpolation, the multiplication by g is then bounded in $H^\beta(\mathbb{T}^d)$, for all $\beta \in [0, \alpha]$, and by duality, for all $\beta \in [-\alpha, 0]$ as well.

Now let us assume that $g \geq \lambda > 0$, so that $\frac{1}{g} \in H^\alpha(\mathbb{T}^d)$, and $\|\frac{1}{g}\|_{H^\alpha} \leq C(d, \alpha, \lambda, \|g\|_{H^\alpha})$. We have

$$\|f\|_{\dot{H}^\beta} = \|g^{-1}gf\|_{\dot{H}^\beta} \lesssim_{d,\beta} \|g^{-1}gf\|_{H^\beta} \lesssim_{d,\alpha,\beta,\lambda,\|g\|_{H^\alpha}} \|gf\|_{H^\beta} \lesssim |\widehat{(gf)}_0| + \|gf\|_{\dot{H}^\beta},$$

where we used that multiplication by $\frac{1}{g}$ is a bounded operator on $H^\beta(\mathbb{T}^d)$. We only need to bound the average of gf by $\|gf\|_{\dot{H}^\beta}$. To do that, observe that, since $\widehat{f}_0 = 0$, then

$$0 = \widehat{(g^{-1}gf)}_0 = \langle gf - \widehat{(gf)}_0, g^{-1}\rangle + \widehat{(g^{-1})}_0\widehat{(gf)}_0.$$

As a consequence

$$|\widehat{(gf)}_0| \leq \frac{1}{\widehat{(g^{-1})}_0}|\langle gf - \widehat{(gf)}_0, g^{-1}\rangle| \leq \frac{1}{\widehat{(g^{-1})}_0}\|g^{-1}\|_{\dot{H}^{-\beta}}\|gf\|_{\dot{H}^\beta} \lesssim_{d,\alpha,\beta,\lambda,\|g\|_{H^\alpha}} \|gf\|_{\dot{H}^\beta},$$

as claimed. □

The following lemma, which contains Kato-Ponce type commutator estimates on the d-dimensional torus, was deduced in [CCCFR26] from the corresponding result on Euclidean space (see [Li19]). Recall the notation $D^\beta := (-\Delta)^{\beta/2}$, for all $\beta \in \mathbb{R}$.

Lemma A.2 ([CCCFR26, Lemma A.6, Remark A.7])**.** *Let $\gamma > 0$ and $2 \leq p_1, q_1, p_2, q_2 \leq \infty$ be such that $1/2 = 1/p_i + 1/q_i$, $i = 1, 2$. Then, for every $f, g \in C^\infty(\mathbb{T}^d)$ the following holds:*

i) If $\gamma > 1$:

$$\|D^\gamma(fg) - fD^\gamma g\|_{L^2} \lesssim_{d,\gamma,p_1,q_1,p_2,q_2} \|\nabla f\|_{L^{p_1}}\|D^{\gamma-1}g\|_{L^{q_1}} + \|D^\gamma f\|_{L^{p_2}}\|g\|_{L^{q_2}}. \quad \text{(A.1)}$$

ii) If $\gamma = 1$:

$$\|D^\gamma(fg) - fD^\gamma g\|_{L^2} \lesssim_{d,p_1,q_1} \|\nabla f\|_{L^{p_1}}\|g\|_{L^{q_1}}. \quad \text{(A.2)}$$

iii) If $\gamma \in (0, 1)$:

$$\|D^\gamma(fg) - fD^\gamma g\|_{L^2} \lesssim_{d,\gamma,p_1,q_1} \|D^\gamma f\|_{L^{p_1}}\|g\|_{L^{q_1}}. \quad \text{(A.3)}$$

B Qualitative convergence for Stein Variational Gradient Descent

B.1 An abstract LaSalle principle

The following proof is inspired by the LaSalle invariance principle for gradient flows in the Wasserstein space shown in [CGW23].

Let M be a compact metric space. Let $\delta : \mathcal{P}(M) \times \mathcal{P}(M) \to [0, \infty]$, $\mathcal{E} : \mathcal{P}(M) \to [0, \infty]$, and $\mathcal{D} : \mathcal{P}(M) \to [0, \infty]$ be given functionals representing a pseudo-distance, an energy, and a dissipation functional, respectively.

For a given weakly continuous curve $[0, \infty) \ni t \mapsto \rho_t \in \mathcal{P}(M)$, consider the following assumptions:

(1) *Energy dissipation.* The curve satisfies $\mathcal{E}(\rho_0) < \infty$, and

$$\mathcal{E}(\rho_T) + \int_0^T \mathcal{D}(\rho_t)\, dt \leq \mathcal{E}(\rho_0) \qquad \forall T \geq 0. \quad \text{(B.1)}$$

(2) *Motion estimate.* The following bound holds along the curve:

$$\delta(\rho_a, \rho_b) \leq \int_a^b \mathcal{D}(\rho_t)^{1/2}\, dt \qquad \forall a, b, \quad 0 \leq a \leq b < \infty. \quad \text{(B.2)}$$

(3) *Coercivity of the pseudo-distance.* Whenever $\mu_n, \nu_n, \nu \in \mathcal{P}(M)$ are such that $\nu_n \rightharpoonup \nu$ and $\delta(\mu_n, \nu_n) \to 0$, then $\mu_n \rightharpoonup \nu$.

(4) *Lower semicontinuity of the dissipation functional.* The map $\mathcal{D}$ is lower semicontinuous with respect to the weak topology of $\mathcal{P}(M)$.

(5) *Unique equilibrium state.* There is $\pi \in \mathcal{P}(M)$ such that

$$\mathcal{D}(\mu) = 0 \qquad \Longleftrightarrow \qquad \mu = \pi. \tag{B.3}$$

Proposition B.1 (Abstract LaSalle principle)**.** *Assume* (1)–(5)*. Then*

$$\rho_t \rightharpoonup \pi \qquad \text{as} \quad t \to \infty.$$

Proof. Since $\mathcal{E}(\rho_0) < \infty$ and $\mathcal{E} \geq 0$, the energy-dissipation estimate (B.1) implies

$$\int_0^\infty \mathcal{D}(\rho_t)\, dt < \infty. \tag{B.4}$$

Let $t_n \to \infty$. Since M is compact, $\mathcal{P}(M)$ is weakly compact. Therefore, after extracting a subsequence, not relabeled, there exists $\rho_\infty \in \mathcal{P}(M)$ such that $\rho_{t_n} \rightharpoonup \rho_\infty$. We wish to prove that $\rho_\infty = \pi$. From (B.4) and $\mathcal{D} \geq 0$, we deduce that

$$\int_{t_n}^{t_n+1} \mathcal{D}(\rho_t)\, dt \to 0. \tag{B.5}$$

For every fixed $r \in [0,1]$, the motion estimate (B.2) and Cauchy-Schwarz give

$$\delta(\rho_{t_n+r}, \rho_{t_n}) \leq \int_{t_n}^{t_n+r} \mathcal{D}(\rho_t)^{1/2}\, dt \leq \left(\int_{t_n}^{t_n+1} \mathcal{D}(\rho_t)\, dt \right)^{1/2} \to 0. \tag{B.6}$$

By the coercivity assumption (3), we thus obtain

$$\rho_{t_n+r} \rightharpoonup \rho_\infty \qquad \forall r \in [0,1]. \tag{B.7}$$

The lower semicontinuity of $\mathcal{D}$, together with Fatou's lemma, yields

$$\mathcal{D}(\rho_\infty) \leq \int_0^1 \liminf_{n\to\infty} \mathcal{D}(\rho_{t_n+r})\, dr \leq \liminf_{n\to\infty} \int_{t_n}^{t_n+1} \mathcal{D}(\rho_t)\, dt = 0. \tag{B.8}$$

Thus $\mathcal{D}(\rho_\infty) = 0$. Finally, by (5), we get $\rho_\infty = \pi$, concluding the proof. □

B.2 Proof of qualitative convergence

The abstract LaSalle principle above applies to a broad class of kernels, provided the associated dissipation and motion estimates satisfy assumptions (1)–(5). We now verify these assumptions for the Riesz kernel K_s used in (2.2), with the choices $\mathcal{E} := \mathcal{H}(\cdot|\pi)$, $\mathcal{D} = I_s(\cdot|\pi)$ and $\delta = W_1$, the 1-Wasserstein distance, which metrizes weak convergence in $\mathcal{P}(\mathbb{T}^d)$.

Lemma B.2. *Let $s > \frac{d}{2}$, $V \in C^1(\mathbb{T}^d)$, and let $\rho \in C([0,T); \mathcal{P}(\mathbb{T}^d))$ be a solution of (2.2) defined up to some positive time T. Then, the following estimate holds:*

$$W_1(\rho_a, \rho_b) \lesssim_{d,s} \int_a^b I_s(\rho_t|\pi)^{1/2}\, dt \qquad \forall a, b, \quad 0 \leq a \leq b < T. \tag{B.9}$$

Proof. Let $f : \mathbb{T}^d \to \mathbb{R}$ be a smooth function with $\|\nabla f\|_{L^\infty} \leq 1$. Since $s > \frac{d}{2}$, we can embed $H^s(\mathbb{T}^d)$ into $C^0(\mathbb{T}^d)$. Therefore, the following bound holds:

$$\|\rho_t \nabla f\|_{\dot{H}^{-s}} = \sup_{\hat{g}_0 = 0, \|g\|_{\dot{H}^s} \leq 1} \langle \rho_t \nabla f, g \rangle \leq \sup_{\hat{g}_0 = 0, \|g\|_{\dot{H}^s} \leq 1} \|g\|_{C^0} \lesssim_{d,s} 1.$$

Using equation (2.2), for almost every t, we may bound

$$\frac{d}{dt} \int_{\mathbb{T}^d} f\, d\rho_t = -\int_{\mathbb{T}^d} K_s * (\pi \nabla \frac{\rho_t}{\pi}) \cdot \nabla f\, d\rho_t \leq \|\pi \nabla \frac{\rho_t}{\pi}\|_{\dot{H}^{-s}} \|\rho_t \nabla f\|_{\dot{H}^{-s}} \lesssim_{d,s} I_s(\rho_t|\pi)^{1/2}.$$

Integrating this inequality from a to b, we obtain

$$\left| \int_{\mathbb{T}^d} f\, d(\rho_b - \rho_a) \right| \lesssim_{d,s} \int_a^b I_s(\rho_t|\pi)^{1/2}\, dt.$$

Then, (B.9) follows from the arbitrariness of f and the Kantorovich duality formula for the 1-Wasserstein distance[7]. □

[7] Namely, $W_1(\mu,\nu) = \sup\left\{ \int_{\mathbb{T}^d} f\, d(\mu - \nu) : \|\nabla f\|_{L^\infty} \leq 1 \right\}$ for all $\mu, \nu \in \mathcal{P}(\mathbb{T}^d)$.

Lemma B.3. *Let $s \geq 1$, and assume that $V \in C^1(\mathbb{T}^d)$. Then, $I_s(\cdot|\pi)$ is lower semicontinuous in $\mathcal{P}(\mathbb{T}^d)$ with respect to the weak topology. Moreover, $I_s(\rho|\pi) = 0$ if and only if $\rho = \pi$.*

Proof. The first claim follows from the fact that $I_s(\cdot|\pi)^{1/2}$ can be written as the supremum of a family of weakly continuous functions defined in the space of probability measures $\mathcal{P}(\mathbb{T}^d)$:

$$I_s(\rho|\pi)^{1/2} = \|\pi\nabla\frac{\rho}{\pi}\|_{\dot{H}^{-s}} = \sup_{g\in C^\infty,\, \widehat{g}_0=0,\, \|g\|_{\dot{H}^s}\leq 1} \langle \pi\nabla\frac{\rho}{\pi}, g\rangle = \sup_{g\in C^\infty,\, \widehat{g}_0=0,\, \|g\|_{\dot{H}^s}\leq 1} \langle \rho, \frac{1}{\pi}\operatorname{div}(\pi g)\rangle,$$

where $\frac{1}{\pi}\operatorname{div}(\pi g)$ is continuous due to the assumption $V \in C^1(\mathbb{T}^d)$. To prove the second claim, observe that if $I_s(\rho|\pi) = 0$, then $\pi\nabla\frac{\rho}{\pi}$ coincides with a constant vector $c \in \mathbb{R}^d$. Since $\pi > 0$ and $\nabla\frac{\rho}{\pi}$ has zero average, this gives $c = 0$, which in turn implies that $\frac{\rho}{\pi}$ is a constant. Since ρ and π both have unit mass, we deduce $\rho = \pi$. □

Proof of Proposition 2.1. The proof follows from the abstract result in Proposition B.1, with the choices $\mathcal{E} := \mathcal{H}(\cdot|\pi)$ and $\mathcal{D} = I_s(\cdot|\pi)$, and $\delta = W_1$, taking into account Lemmas B.2 and B.3. □

C Proof of the propagation of regularity estimates

In this section, we prove the propagation of regularity estimates stated in Section 4. We recall their full statements, for convenience.

Lemma C.1. *Let $s > 1$ and $\gamma > \frac{d}{2}$. Let $V \in H^{\gamma+s}(\mathbb{T}^d)$ and $\pi \in \mathcal{P} \cap H^{\gamma+s}(\mathbb{T}^d)$ be defined as in (2.1). Let $\rho \in C([0,T);\mathcal{P}(\mathbb{T}^d)) \cap L^\infty_{\mathrm{loc}}([0,T);H^\gamma(\mathbb{T}^d))$ be a solution of (2.2) defined up to some time $T > 0$, and call $\sigma_t = \rho_t - \pi$. Then, there is a constant $C > 0$ depending only on d, s, γ, and $\|V\|_{H^{\gamma+s}}$ such that the following estimate holds for all $t \in (0,T)$:*

$$\begin{aligned}
&\|\sigma_t\|^2_{\dot{H}^\gamma} + 2\big(\min_{\mathbb{T}^d}\pi\big)\int_0^t \|\sigma_r\|^2_{\dot{H}^{\gamma-s+1}} \exp\left(C\int_r^t L(u)\,du\right) dr \\
&\leq \|\sigma_0\|^2_{\dot{H}^\gamma} \exp\left(C\int_0^t L(u)\,du\right) + C\int_0^t \|\sigma_r\|_{\dot{H}^{\gamma-s}}\|\sigma_r\|_{\dot{H}^{\gamma-s+1}} \exp\left(C\int_r^t L(u)\,du\right) dr,
\end{aligned} \tag{C.1}$$

where we set

$$L(t) := \|\nabla^2 D^{-2s}\sigma_t\|_{L^\infty} + \|\nabla D^{-2s}(\sigma_t\nabla V)\|_{L^\infty}. \tag{C.2}$$

Proof. We assume that the data and hence the solution are smooth. The passage to general solutions in the stated regularity class can be justified with an approximation argument, by passing to the limit in the integral energy estimate as in [CCCFR26, Section 2].

Throughout the proof, we use the notation $D^\beta = (-\Delta)^{\beta/2}$ for all $\beta \in \mathbb{R}$. Using (2.2), we see that σ_t solves the equation

$$\partial_t\sigma_t = \operatorname{div}\left(\pi D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right) + \operatorname{div}\left(\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right),$$

where the first term on the right-hand side originates from the linearized equation, while the second contains the non-linear parts. Therefore, we derive

$$\begin{aligned}
\frac{d}{dt}\frac{\|\sigma_t\|^2_{\dot{H}^\gamma}}{2} &= \int_{\mathbb{T}^d} D^\gamma\partial_t\sigma_t D^\gamma\sigma_t \\
&= \underbrace{\int_{\mathbb{T}^d} D^\gamma \operatorname{div}\left(\pi D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right) D^\gamma\sigma_t}_{\mathrm{Lin}} + \underbrace{\int_{\mathbb{T}^d} D^\gamma \operatorname{div}\left(\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right) D^\gamma\sigma_t}_{\mathrm{Nonlin}}.
\end{aligned}$$

We estimate the two integrals Lin and Nonlin above separately.

Step 1: (The linear term). Using $\pi\nabla\frac{\sigma_t}{\pi} = \nabla\sigma_t + \sigma_t\nabla V$, integration by parts, and self-adjointness of fractional derivatives, we may further split Lin into three terms:

$$\begin{aligned}
\mathrm{Lin} &= \int_{\mathbb{T}^d} D^\gamma \operatorname{div}\left(\pi\nabla D^{-2s}\sigma_t\right) D^\gamma\sigma_t + \int_{\mathbb{T}^d} D^\gamma \operatorname{div}\left(\pi D^{-2s}(\sigma_t\nabla V)\right) D^\gamma\sigma_t \\
&= -\int_{\mathbb{T}^d} D^{\gamma+s}\left(\pi\nabla D^{-2s}\sigma_t\right)\cdot\nabla D^{\gamma-s}\sigma_t - \int_{\mathbb{T}^d} D^{\gamma+s}\left(\pi D^{-2s}(\sigma_t\nabla V)\right)\cdot\nabla D^{\gamma-s}\sigma_t \\
&= -\int_{\mathbb{T}^d}\pi|\nabla D^{\gamma-s}\sigma_t|^2 - \int_{\mathbb{T}^d}\left[D^{\gamma+s}\left(\pi\nabla D^{-2s}\sigma_t\right) - \pi D^{\gamma+s}\nabla D^{-2s}\sigma_t\right]\cdot\nabla D^{\gamma-s}\sigma_t \\
&\quad - \int_{\mathbb{T}^d} D^{\gamma+s}\left(\pi D^{-2s}(\sigma_t\nabla V)\right)\cdot\nabla D^{\gamma-s}\sigma_t \\
&= \mathrm{Lin}_1 + \mathrm{Lin}_2 + \mathrm{Lin}_3.
\end{aligned}$$

Now, for Lin_1, we simply bound π below with its minimum value, and get

$$\mathrm{Lin}_1 \le -\big(\min_{\mathbb{T}^d}\pi\big)\int_{\mathbb{T}^d}|\nabla D^{\gamma-s}\sigma_t|^2 = -\big(\min_{\mathbb{T}^d}\pi\big)\|\sigma_t\|^2_{\dot H^{\gamma-s+1}}. \tag{C.3}$$

For Lin_2, using Cauchy-Schwarz, the Kato-Ponce commutator estimate (A.1), and the Sobolev embedding $H^{\gamma+s-1}(\mathbb{T}^d)\hookrightarrow L^\infty(\mathbb{T}^d)$, we derive

$$\begin{aligned}
\mathrm{Lin}_2 &\le \left\|D^{\gamma+s}\left(\pi\nabla D^{-2s}\sigma_t\right) - \pi D^{\gamma+s}\nabla D^{-2s}\sigma_t\right\|_{L^2}\|\nabla D^{\gamma-s}\sigma_t\|_{L^2} \\
&\lesssim_{d,s,\gamma}\left(\|\nabla\pi\|_{L^\infty}\|D^{\gamma+s-1}\nabla D^{-2s}\sigma_t\|_{L^2} + \|D^{\gamma+s}\pi\|_{L^2}\|\nabla D^{-2s}\sigma_t\|_{L^\infty}\right)\|\sigma_t\|_{\dot H^{\gamma-s+1}} \\
&\lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}}} \|\sigma_t\|_{\dot H^{\gamma-s}}\|\sigma_t\|_{\dot H^{\gamma-s+1}}.
\end{aligned} \tag{C.4}$$

Finally, for Lin_3, we use Cauchy-Schwarz and Lemma A.1:

$$\mathrm{Lin}_3 \le \|D^{\gamma+s}\left(\pi D^{-2s}(\sigma_t\nabla V)\right)\|_{L^2}\|\nabla D^{\gamma-s}\sigma_t\|_{L^2} \lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}}} \|\sigma_t\|_{\dot H^{\gamma-s}}\|\sigma_t\|_{\dot H^{\gamma-s+1}}. \tag{C.5}$$

Step 2: (The non-linear term). After integrating by parts, we may split Nonlin into three terms:

$$\begin{aligned}
\mathrm{Nonlin} &= -\int_{\mathbb{T}^d} D^\gamma\left(\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right)\cdot\nabla D^\gamma\sigma_t \\
&= -\int_{\mathbb{T}^d} D^\gamma\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\cdot\nabla D^\gamma\sigma_t \\
&\quad + \int_{\mathbb{T}^d}\left[D\left(D^\gamma\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right) - DD^\gamma\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right]\cdot\nabla D^{\gamma-1}\sigma_t \\
&\quad - \int_{\mathbb{T}^d}\left[D^{\gamma+1}\left(\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right) - D^{\gamma+1}\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right]\cdot\nabla D^{\gamma-1}\sigma_t \\
&= \mathrm{Nonlin}_1 + \mathrm{Nonlin}_2 + \mathrm{Nonlin}_3.
\end{aligned}$$

For Nonlin_1, we use the identity $2D^\gamma\sigma_t\nabla D^\gamma\sigma_t = \nabla(D^\gamma\sigma_t)^2$ together with integration by parts, and get

$$\mathrm{Nonlin}_1 = \frac{1}{2}\int_{\mathbb{T}^d}\operatorname{div}\left(D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right)(D^\gamma\sigma_t)^2 \lesssim_d \|\nabla D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^\infty}\|\sigma_t\|^2_{\dot H^\gamma}. \tag{C.6}$$

For Nonlin_2, we use Cauchy-Schwarz together with (A.2), and obtain

$$\begin{aligned}
\mathrm{Nonlin}_2 &\le \|D\left(D^\gamma\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right) - DD^\gamma\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^2}\|\nabla D^{\gamma-1}\sigma_t\|_{L^2} \\
&\lesssim_d \|\nabla D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^\infty}\|\sigma_t\|^2_{\dot H^\gamma}.
\end{aligned} \tag{C.7}$$

Finally, let us consider Nonlin_3. Using Cauchy-Schwarz along with (A.1), we get

$$\begin{aligned}
\mathrm{Nonlin}_3 &\le \|D^{\gamma+1}\left(\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\right) - D^{\gamma+1}\sigma_t D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^2}\|\nabla D^{\gamma-1}\sigma_t\|_{L^2} \\
&\lesssim_{d,s,\gamma}\left(\|\nabla D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^\infty}\|D^\gamma\sigma_t\|_{L^2} + \|\sigma_t\|_{L^p}\|D^{\gamma+1}D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^q}\right)\|\sigma_t\|_{\dot H^\gamma},
\end{aligned}$$

where $p = 2\frac{\gamma+2s-2}{2s-2}$ and $q = 2\frac{\gamma+2s-2}{\gamma}$ are such that $\frac{1}{p} + \frac{1}{q} = \frac{1}{2}$. By the fractional Gagliardo-Nirenberg inequality [BM18], we may bound

$$\begin{aligned}
&\|\sigma_t\|_{L^p}\|D^{\gamma+1}D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^q} \\
&\quad\lesssim_{d,s,\gamma} \|\nabla^2 D^{-2s}\sigma_t\|_{L^\infty}^{\frac{2}{q}}\|D^\gamma\sigma_t\|_{L^2}^{\frac{2}{p}}\|\nabla D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^\infty}^{\frac{2}{p}}\|D^{\gamma-1}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^2}^{\frac{2}{q}} \\
&\quad\lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}}} \left(\|\nabla^2 D^{-2s}\sigma_t\|_{L^\infty} + \|\nabla D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^\infty}\right)\|\sigma_t\|_{\dot{H}^\gamma},
\end{aligned}$$

where in the last step we used Young's inequality, and Lemma A.1. As a consequence, we find

$$\text{Nonlin}_3 \lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}}} \left(\|\nabla^2 D^{-2s}\sigma_t\|_{L^\infty} + \|\nabla D^{-2s}(\pi\nabla\frac{\sigma_t}{\pi})\|_{L^\infty}\right)\|\sigma_t\|_{\dot{H}^\gamma}^2. \tag{C.8}$$

Combining (C.3),(C.4),(C.5),(C.6),(C.7), and (C.8), we obtain

$$\frac{d}{dt}\|\sigma_t\|_{\dot{H}^\gamma}^2 \le CL(t)\|\sigma_t\|_{\dot{H}^\gamma}^2 - 2\big(\min_{\mathbb{T}^d}\pi\big)\|\sigma_t\|_{\dot{H}^{\gamma-s+1}}^2 + C\|\sigma_t\|_{\dot{H}^{\gamma-s}}\|\sigma_t\|_{\dot{H}^{\gamma-s+1}} \qquad \forall t\in(0,T)$$

Multiplying this differential inequality by $\exp(-C\int_0^t L(r)\,dr)$, and integrating in dr from 0 to t we obtain the integral energy estimate (4.6). □

It remains to control the cumulative effect of the Lipschitz norm appearing in the high-order estimate.

Lemma C.2. *Under the same assumptions and using the same notation as in Lemma 4.3, assume further that $\gamma > s-1$ and $\|\sigma_t\|_{\dot{H}^\gamma}^2 \le M$ for all $t\in[0,T)$. Then, there is $\beta = \beta(d,s,\gamma) > 0$ such that*

$$\int_0^T L(t)\,dt \lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}},M} \mathcal{H}(\bar\rho|\pi)^\beta.$$

Proof. Let $r := \max\{2-2s+\frac{\gamma}{2}+\frac{d}{4}, 1-s\}$. Since $\gamma > \frac{d}{2}$, we have $r > 2-2s+\frac{d}{2}$. Therefore, by the Sobolev embedding and Lemma A.1, we may bound

$$L(t) = \|\nabla^2 D^{-2s}\sigma_t\|_{L^\infty} + \|\nabla D^{-2s}(\sigma_t\nabla V)\|_{L^\infty} \lesssim_{d,\gamma,s,\|V\|_{H^{\gamma+s}}} \|\sigma_t\|_{\dot{H}^r}.$$

Notice that we have $r\in[1-s,\gamma)$, hence we may use Sobolev interpolation of homogeneous norms to get

$$\|\sigma_t\|_{\dot{H}^r} \le \|\sigma_t\|_{\dot{H}^{1-s}}^{1-\theta}\|\sigma_t\|_{\dot{H}^\gamma}^{\theta} \qquad \theta := \frac{r+s-1}{\gamma+s-1}.$$

Then, using (4.4) and (4.5) together with $\|\sigma_t\|_{\dot{H}^\gamma}^2 \le M$, we find

$$\begin{aligned}
L(t) &\lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}},M} I_s(\rho_t|\pi)I_s(\rho_t|\pi)^{-\frac{1+\theta}{2}} \\
&\lesssim_{d,s,\gamma,\|V\|_{H^{\gamma+s}},M} \left(-\frac{d}{dt}\mathcal{H}(\rho_t|\pi)\right)\mathcal{H}(\rho_t|\pi)^{-\frac{1+\theta}{2}(1+\frac{s-1}{\gamma})} = -\frac{1}{\beta}\frac{d}{dt}\mathcal{H}(\rho_t|\pi)^\beta,
\end{aligned}$$

where $\beta = \min\{\frac{2\gamma-d}{8\gamma}, \frac{\gamma-s+1}{2\gamma}\} > 0$. The result follows by integration. □

D Maximum principle for $s = 1$

In the endpoint case $s = 1$, the equation has an additional maximum-principle structure, which gives global control of bounded solutions. This allows us to ensure that the equation is globally well-posed.

Lemma D.1 (Maximum principle for $s = 1$)**.** *Let $s = 1$, and $\gamma > \frac{d}{2}$, $\gamma \ge 1$. Let $V \in H^{\gamma+1}(\mathbb{T}^d)$ and $\pi \in \mathcal{P}\cap H^\gamma(\mathbb{T}^d)$ be defined as in (2.1). Let $\bar\rho \in \mathcal{P}\cap L^\infty(\mathbb{T}^d)$ and $\rho \in C([0,T);\mathcal{P}(\mathbb{T}^d))\cap L^\infty_{\text{loc}}([0,T);L^\infty(\mathbb{T}^d))$ be a solution of (2.2). Then ρ can be extended to all positive times, and there is a constant $M > 0$ depending only on $d,\gamma,\|V\|_{H^{\gamma+1}}$, and $\mathcal{H}(\bar\rho|\pi)$ such that*

$$\|\rho_t\|_{L^\infty} \le \max\{M, \|\bar\rho\|_{L^\infty}\} \qquad \forall t\ge 0. \tag{D.1}$$

Proof. Arguing by approximation, we may assume that the solution is smooth. We call $\sigma_t = \rho_t - \pi$ and $v_t = -\nabla K_1 * \sigma_t - K_1 * (\sigma_t \nabla V)$ the velocity field. The equation can be rewritten as

$$\partial_t \rho_t + \nabla \rho_t \cdot v_t = -\rho_t \operatorname{div} v_t = -\rho_t \left(\sigma_t - \operatorname{div}(-\Delta)^{-1}(\sigma_t \nabla V)\right).$$

Let $r > d$ be such that the embedding $H^{\gamma-1}(\mathbb{T}^d) \hookrightarrow L^r(\mathbb{T}^d)$ holds (the existence of such r is guaranteed by the assumptions on γ). Let $p \in (d, r)$ and $q \in (p, \infty)$ be such that $\frac{1}{p} = \frac{1}{r} + \frac{1}{q}$. Here p, r, q can be chosen depending only on d and γ. Then, by Calderón-Zygmund estimates, Sobolev embedding, and Hölder's inequality, we find

$$\begin{aligned}
\|\operatorname{div}(-\Delta)^{-1}(\sigma_t \nabla V)\|_{L^\infty} &\lesssim_{d,\gamma} \|\sigma_t \nabla V\|_{L^p} \le \|\nabla V\|_{L^r} \|\sigma_t\|_{L^q} \\
&\lesssim_{d,\gamma,\|V\|_{H^\gamma}} \|\sigma_t\|_{L^1}^{\frac{1}{q}} \|\sigma_t\|_{L^\infty}^{\frac{q-1}{q}} \\
&\lesssim_{d,\gamma,\|V\|_{H^\gamma}} \mathcal{H}(\rho_t|\pi)^{\frac{1}{2q}} \|\sigma_t\|_{L^\infty}^{\frac{q-1}{q}} \le \mathcal{H}(\bar\rho|\pi)^{\frac{1}{2q}} \left(\max_{\mathbb{T}^d} \rho_t + \max_{\mathbb{T}^d} \pi\right)^{\frac{q-1}{q}},
\end{aligned}$$

where we also used Pinsker's inequality $\|\rho_t - \pi\|_{L^1} \le \sqrt{2\mathcal{H}(\rho_t|\pi)}$ and the fact that the relative entropy is dissipated by (2.3). For $t > 0$, if $x_0 \in \mathbb{T}^d$ is a maximum point of ρ_t, then $\nabla \rho_t(x_0) = 0$ and

$$\begin{aligned}
\partial_t \rho_t(x_0) &= -\rho_t(x_0) \left(\rho_t(x_0) - \pi(x_0) - \operatorname{div}(-\Delta)^{-1}(\sigma_t \nabla V)(x_0)\right) \\
&\le -\max_{\mathbb{T}^d} \rho_t \left(\max_{\mathbb{T}^d} \rho_t - \max_{\mathbb{T}^d} \pi - C \left(\max_{\mathbb{T}^d} \rho_t + \max_{\mathbb{T}^d} \pi\right)^{\frac{q-1}{q}}\right).
\end{aligned}$$

The expression in parentheses on the right-hand side is positive if $\max_{\mathbb{T}^d} \rho_t \ge M$, with M depending only on $d, \gamma, \|V\|_{H^\gamma}$, and $\mathcal{H}(\bar\rho|\pi)$. Since for almost every $t > 0$, $\frac{d}{dt} \max_{\mathbb{T}^d} \rho_t = \partial_t \rho_t(x_0)$, where x_0 is a maximizer of ρ_t, (D.1) follows. □

E More general kernels

Here we describe how to change the proof of quantitative convergence in the case where K_s is replaced by a more general kernel K with the same spectral behavior.

Theorem E.1 (General kernels)**.** *Let $s > 1$, $K : \mathbb{T}^d \to \mathbb{R}$ be a symmetric kernel and $\mathcal{K}(f) := K * f$ the corresponding convolution operator. Assume that the following spectral conditions hold:*

(i) $\widehat{K}_0 \ge 0$ *and* $\widehat{K}_k \approx |k|^{-2s}$ *for every* $k \in \mathbb{Z}^d \setminus \{0\}$.

(ii) The operator $T := (-\Delta)^s \mathcal{K}$ *satisfies the following estimates:*

$$\|Tf\|_{L^p} \approx_p \|f\|_{L^p} \qquad \forall f \in L^p(\mathbb{T}^d), \quad \widehat{f}_0 = 0, \qquad \forall p \in (1, \infty).$$

(iii) There is a constant $\bar c > 0$ *such that*

$$\left\|\left(\mathcal{K} - \bar c(-\Delta)^{-s}\right)(f)\right\|_{L^2} \lesssim \|f\|_{\dot H^{-2s-1}} \qquad \forall f \in L^2(\mathbb{T}^d), \quad \widehat{f}_0 = 0.$$

Then the same conclusions of Theorem 2.2 hold true for solutions of (2.2) if we replace the Riesz kernel K_s by K, with constants also depending on $\widehat{K}_0$ and the bounds in (i)–(iii).

Remark E.2. This theorem can be applied, in particular, when $d \ge 2$, to the negative distance-like kernel K introduced in (2.9), for which $s = (d+1)/2$. Indeed, since K is supported in $B_{1/4}$, it may equivalently be regarded as a function on $\mathbb{R}^d$. We can then study the relevant properties of its Fourier transform $\widehat{K}$, which is real-valued, radial, and smooth. To prove points (i) and (iii), one may first apply Askey's theorem [Ask73] to get $\widehat{K} > 0$. The asymptotic behavior is then determined by the 1-homogeneous singularity of the kernel at the origin, yielding $\widehat{K}(\xi) = C_d|\xi|^{-d-1} + O(|\xi|^{-d-2})$ as $\xi \to \infty$, which gives the desired spectral comparability. Finally, point (ii) follows by applying the standard Mikhlin multiplier theorem to the symbol $c(\xi) := |\xi|^{d+1}\widehat{K}(\xi)$, once one proves that $|\partial^\beta c(\xi)| \lesssim |\xi|^{-|\beta|}$, for all $|\beta| \le \lfloor \frac{d}{2} \rfloor + 1$. These estimates hold thanks to the regularity imposed to the kernel away from the origin by raising $(1 - 4|x|)$ to the sufficiently large power $d + 2$.

Remark E.3. Let $\nu, \ell > 0$, set $a := \sqrt{2\nu}/\ell$, $s := \nu + d/2$, and assume that $s > 1$. On $\mathbb{R}^d$, consider the Matérn function

$$M_{\nu,\ell}(z) := \frac{2^{1-\nu}}{\Gamma(\nu)}(a|z|)^\nu K_\nu(a|z|), \qquad M_{\nu,\ell}(0) := 1,$$

where K_ν denotes the modified Bessel function of the second kind. Then its periodization $K(x) := \sum_{m\in\mathbb{Z}^d} M_{\nu,\ell}(x+m)$ satisfies the assumptions of Theorem E.1. Indeed, the sum converges by the exponential decay of $M_{\nu,\ell}$ at infinity. The integral representation

$$M_{\nu,\ell}(z) = \frac{1}{\Gamma(\nu)}\int_0^\infty t^{\nu-1}\exp\left(-t - \frac{a^2|z|^2}{4t}\right)dt$$

and the Fourier transform of a Gaussian give

$$\widehat{K}_k = A_{d,\nu,\ell}\big(a^2 + |2\pi k|^2\big)^{-s}, \qquad A_{d,\nu,\ell} := \frac{2^d\pi^{d/2}\Gamma(s)a^{2\nu}}{\Gamma(\nu)} > 0, \qquad k \in \mathbb{Z}^d. \tag{E.1}$$

In particular, $\widehat{K}_0 > 0$ and $\widehat{K}_k \approx |k|^{-2s}$ for $k \neq 0$, which proves (i). Moreover,

$$\widehat{K}_k = A_{d,\nu,\ell}|2\pi k|^{-2s} + O(|k|^{-2s-2}) \qquad \text{as } |k| \to \infty.$$

By Parseval's identity this proves (iii), with $\bar{c} = A_{d,\nu,\ell}$. Finally, the symbol of $(-\Delta)^s\mathcal{K}$ on nonzero modes is

$$m(\xi) = A_{d,\nu,\ell}\left(\frac{|2\pi\xi|^2}{a^2 + |2\pi\xi|^2}\right)^s.$$

Away from the origin m and m^{-1} satisfy the Mikhlin bounds $|\partial^\alpha m(\xi)| + |\partial^\alpha(m^{-1})(\xi)| \lesssim_\alpha |\xi|^{-|\alpha|}$. Therefore, applying the periodic Mikhlin multiplier theorem yields (ii).

In order to prove Theorem E.1, one can proceed as in the proof of Theorem 2.2, with the modifications described below.

The equation. Let $\mathcal{K} = K*$ be the convolution operator associated to the kernel K. We consider a solution of

$$\partial_t\rho = \operatorname{div}\left(\rho\mathcal{K}(\rho\nabla\log\left(\frac{\rho}{\pi}\right))\right), \qquad \rho_0 = \bar{\rho}.$$

Denoting $\sigma_t = \rho_t - \pi$, the velocity field can be rewritten in the following equivalent ways:

$$v_t := -\mathcal{K}(\rho_t\nabla\log\left(\frac{\rho_t}{\pi}\right)) = -\mathcal{K}(\pi\nabla\frac{\sigma_t}{\pi}) = -\nabla\mathcal{K}(\sigma_t) - \mathcal{K}(\sigma_t\nabla V).$$

Energy dissipation identity (2.3). We get

$$\frac{d}{dt}\mathcal{H}(\rho_t|\pi) = -\langle\mathcal{K}(\pi\nabla\frac{\sigma_t}{\pi}), \pi\nabla\frac{\sigma_t}{\pi}\rangle =: -I(\rho_t|\pi). \tag{E.2}$$

Lemma 4.2. By assumption (i), the following equivalence holds:

$$I(\rho_t|\pi) \approx \widehat{K}_0\left|\widehat{(\sigma_t\nabla V)}_0\right|^2 + \|\pi\nabla\frac{\sigma_t}{\pi}\|^2_{\dot{H}^{-s}}.$$

Therefore, since $\widehat{K}_0 \geq 0$, we deduce

$$I(\rho_t|\pi) \gtrsim \|\sigma_t\|^2_{\dot{H}^{1-s}} \gtrsim \mathcal{H}(\rho_t|\pi)^{1+\frac{s-1}{\gamma}}\left(\|\sigma_t\|^2_{\dot{H}^\gamma}\right)^{-\frac{s-1}{\gamma}}.$$

Lemma 4.3. Working with smooth solutions, we may prove the following estimate, for a sufficiently large constant $C > 0$:

$$\frac{d}{dt}\|\sigma_t\|^2_{\dot{H}^\gamma} \leq CL(t)\|\sigma_t\|^2_{\dot{H}^\gamma} - \frac{1}{C}\|\sigma_t\|^2_{\dot{H}^{\gamma-s+1}} + C\|\sigma_t\|_{\dot{H}^{\gamma-s}}\|\sigma_t\|_{\dot{H}^{\gamma-s+1}}, \tag{E.3}$$

where we set

$$L(t) := \|\nabla^2\mathcal{K}(\sigma_t)\|_{L^\infty} + \|\nabla\mathcal{K}(\sigma_t\nabla V)\|_{L^\infty}. \tag{E.4}$$

The proof is almost identical to that of Lemma 4.3, after replacing D^{-2s} with $\mathcal{K}$, and using their spectral comparability on zero-mean functions. The only notable difference appears in the estimate of the term Lin arising from the linearized equation. There, the negative term Lin_1 is now replaced by

$$\begin{aligned}-\int_{\mathbb{T}^d} \pi D^{\gamma+s}\nabla\mathcal{K}(\sigma_t)\cdot\nabla D^{\gamma-s}\sigma_t &= -\bar{c}\int_{\mathbb{T}^d}\pi|\nabla D^{\gamma-s}\sigma_t|^2 \\ &\quad -\int_{\mathbb{T}^d}\pi D^{\gamma+s}\nabla\left(\mathcal{K}-\bar{c}D^{-2s}\right)(\sigma_t)\cdot\nabla D^{\gamma-s}\sigma_t \\ &=: \mathrm{Lin}_{1,1}+\mathrm{Lin}_{1,2},\end{aligned}$$

where we added and subtracted $\bar{c}D^{-2s}$ to $\mathcal{K}$, with $\bar{c}>0$ the constant given by assumption (iii). Then, $\mathrm{Lin}_{1,1}$ is estimated by

$$\mathrm{Lin}_{1,1}\le -\bar{c}\big(\min_{\mathbb{T}^d}\pi\big)\int_{\mathbb{T}^d}|\nabla D^{\gamma-s}\sigma_t|^2 = -\bar{c}\big(\min_{\mathbb{T}^d}\pi\big)\|\sigma_t\|^2_{\dot{H}^{\gamma-s+1}}.$$

On the other hand, since $\mathcal{K}-\bar{c}D^{-2s}$ gains one spectral order by assumption (iii), we may bound $\mathrm{Lin}_{1,2}$ by

$$\mathrm{Lin}_{1,2}\lesssim \|\sigma_t\|_{\dot{H}^{\gamma-s}}\|\sigma_t\|_{\dot{H}^{\gamma-s+1}}.$$

Lemma 4.4. The result is exactly the same using L from (E.4). Indeed, by assumption (i), the estimate $L(t)\lesssim\|\sigma_t\|_{\dot{H}^r}$ still holds for the same exponent r introduced in the proof of Lemma 4.4.

Conclusion. Having derived the same ingredients for the case of more general kernels, the proof of Theorem E.1 follows exactly as that of Theorem 2.2.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: they describe the results proved in the paper and situate them in the literature.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: These are discussed in the conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide assumptions and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: the experimental code is elementary and parameter used are given in the main text. The full code will be shared online upon publication.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We do not use data. The code is elementary and will be shared online upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: They are described in the numerical section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The randomness effects are not important in our settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The time of execution of the code is around 5 minutes on a standard laptop.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research does not violate the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a theory paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: the paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects.**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.