Mean-Field Nonparametric Estimation of Interacting Particle Systems

Rentian Yao RENTIAN2@ILLINOIS.EDU
Xiaohui Chen XHCHEN@ILLINOIS.EDU
Yun Yang YY84@ILLINOIS.EDU
University of Illinois at Urbana-Champaign

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

This paper concerns the nonparametric estimation problem of the distribution-state dependent drift vector field in an interacting N-particle system. Observing single-trajectory data for each particle, we derive the mean-field rate of convergence for the maximum likelihood estimator (MLE), which depends on both Gaussian complexity and Rademacher complexity of the function class. In particular, when the function class contains d-variate α-Hölder smooth functions, our rate of convergence is minimax optimal on the order of $N^{-\frac{\alpha}{d+2\alpha}}$. Combining with a Fourier analytical deconvolution argument, we derive the consistency of MLE for the external force and interaction kernel in the McKean-Vlasov equation.

Keywords: interacting particle system, maximum likelihood estimation, Mckean-Vlasov equation, mean-field regime, learning interaction kernel.

1. Introduction

Recent years have seen increasing research interest and progress in learning dynamical pattern of a large interacting particle system (IPS). Motivating applications on modeling collective behaviors come from statistical physics (D'Orsogna et al., 2006), mathematical biology (Mogilner and Edelstein-Keshet, 1999; Topaz et al., 2006), social science (Motsch and Tadmor, 2014), stochastic control (Buckdahn et al., 2017), mean-field games (Carmona and Delarue, 2018), and more recently computational statistics on high-dimensional sampling (Liu, 2017; Lu et al., 2019b) and machine learning for neural networks (Mei et al., 2018, 2019; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020a,b). Due to the large number of particles with interactions, such dynamical systems are high-dimensional and often non-linear. In this paper, we consider a general interacting N-particle system described by the stochastic differential equations (SDEs)

$$\mathrm{d}X_t^i = b^*(t, \mu_t^N, X_t^i)\,\mathrm{d}t + \sigma^*(t, X_t^i)\,\mathrm{d}W_t^i, \quad 1 \le i \le N, \tag{1}$$

where $(W_t^1)_{t\ge 0}, \ldots, (W_t^N)_{t\ge 0}$ are independent Brownian motions on the d-dimensional Euclidean space $\mathbb{R}^d$, $\mu_t^N = N^{-1}\sum_{i=1}^N \delta_{X_t^i}$ is the empirical law of the interacting particles, and the initialization $X_0^1, \ldots, X_0^N$ are i.i.d. $\mathbb{R}^d$-valued random variables with a common law μ_0, independent of $(W_t^i)_{t\ge 0}$. Here in the non-linear diffusion process (1), letting $\mathcal{P}(\mathbb{R}^d)$ be the space of all probability measures on $\mathbb{R}^d$, the vector field $b^* : \mathbb{R}_+ \times \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ is a distribution-state dependent drift vector field to be estimated and σ^* is a known diffusion function (or volatility coefficient) quantifying the magnitude of the self-energy of the particle. For simplicity, we focus on systems with time-homogeneous and space-(one-)periodic drift vector field $b^*(t, \nu, x) =: b^*(\nu, x)$ satisfying $b^*(\nu, x + m) = b^*(\nu, x)$ for every $m \in \mathbb{Z}^d$, and constant diffusion function $\sigma^*(t, x) \equiv 1$. The

periodic model effectively confines the SDEs to a compact state space as the d-dimensional torus $\mathbb{T}^d$, and is commonly adopted in the SDE analysis to avoid boundary issues (van Waaij and van Zanten, 2016; Pokern et al., 2013; Nickl and Ray, 2020). Suppose that we observe continuous-time single-trajectory data for each particle $\mathcal{X}_T = \{(X_t^1, \dots, X_t^N) : 0 \le t \le T\}$ in a finite time horizon $T > 0$. Our goal is to derive a statistically valid procedure to estimate the vector field b^* in a large IPS based on the data $\mathcal{X}_T$.

1.1. System governed by external-interaction force

In the periodic setting, the values of the process (X_t^i) modulo $\mathbb{Z}^d$ contain all relevant statistical information, so we can identify the law of (X_t^i) with a uniquely defined probability measure on $\mathbb{T}^d$ (cf. Section 2.2 in Nickl and Ray (2020) for further details). Under such identification, one important class of IPS with a time-homogeneous drift vector field can be represented as

$$b^*(\nu, x) = \int_{\mathbb{T}^d} \tilde{b}^*(x, y)\, \mathrm{d}\nu(y), \quad \text{with} \quad \tilde{b}^*(x, y) = G^*(x) + F^*(x - y), \tag{2}$$

for $\nu \in \mathcal{P}(\mathbb{T}^d)$ and continuous $F^*, G^* : \mathbb{T}^d \to \mathbb{R}^d$. In this case, one can interpret G^* as an external force to the global system characterizing the drift tendency of particles and F^* as an interaction kernel between particles. Then the IPS in (1) can be reformulated as

$$\mathrm{d}X_t^i = G^*(X_t^i)\, \mathrm{d}t + \frac{1}{N} \sum_{j=1}^{N} F^*(X_t^i - X_t^j)\, \mathrm{d}t + \mathrm{d}W_t^i.$$

In statistical mechanics, microscopic behaviors of N random particles are usually related to explain some observed macroscopic physical quantities (e.g., temperature distributions) in the sense that the evolution of the empirical law μ_t^N of the particles converges to a non-random mean-field limit μ_t as $N \to \infty$ and the probability measure flow $\mu_t(x) := \mu(x, t)$ solves the *McKean-Vlasov equation* (McKean, 1966)

$$\partial_t \mu = \Delta \mu + \mathrm{div}\left(\mu \left[G^* + \int_{\mathbb{T}^d} F^*(\cdot - y) \mu_t(\mathrm{d}y) \right] \right), \tag{3}$$

which is a non-linear parabolic partial differential equation (PDE). For this special class of IPS, a further goal is to study the identifiability of (F^*, G^*) and consistency of the derived estimators.

1.2. Related work

It is a classical result that N-particle interacting system (1) admits a unique strong solution, when both b^* and σ^* are Lipschitz continuous and the solution converges to its mean-field limit McKean-Vlasov stochastic differential equation (MVSDE) both in pathwise and weakly under the same Lipschitz condition (Carmona, 2016; Carmona and Delarue, 2018). The latter is usually known as propagation of chaos (Sznitman, 1991). Another inspiring work from Lacker (2018) showed that the convergence can be proved in a much stronger topology (τ-topology), when volatility coefficient σ^* involves no interaction term.

Several works about learning the interaction kernel of interacting particle system have be done lately. Bongini et al. (2017) proposed an estimator by minimizing the discrete error functional, whose convergence rate is usually no faster than $N^{-1/d}$. This reflects the phenomenon of curse-of-dimension. Lu et al. (2019a) constructed the least square estimator for interaction kernel, which

enjoys an optimal rate of convergence under mild conditions. These two works were done under a noiseless setting, i.e., the system evolves according to an ordinary differential equation and initial conditions of agents are i.i.d. As for the stochastic system, Li et al. (2021) studied the learnability (identifiability) of interaction kernel by maximum likelihood estimator (MLE) under the coercivity condition, and Lang and Lu (2021) provided a complete characterization of learnability. Della Maestra and Hoffmann (2021) investigated a nonparametric estimation of the drift coefficient, and the interaction kernel can be separated by applying Fourier transform for deconvolution. The convergence result is provided under a fixed time horizon, meaning that time T is fixed in their asymptotic result. Another nonparametric estimation algorithm based on least squares was proposed by Lang and Lu (2020).

Estimating parameters of interacting systems by maximum likelihood can date back to 1990. Kasonga (1990) proved the consistency and asymptotic normality of MLE for linear parametrized interacting systems. As for MVSDE, Wen et al. (2016) discussed the consistency of MLE in a broad class of MVSDE, based on a single trajectory $(x_t)_{0\le t\le T}$. Liu and Qiao (2020) extended it to path-dependent case with non-Lipschitz coefficients. Both of these works focused on the asymptotic behaviour when $T \to \infty$. Sharrock et al. (2021) studied the case with N realizations of MVSDE, and the case of N interacting particle systems, under which consistency of MLE was proved when $N \to \infty$ and an online parameter estimation method was also discussed. Chen (2021) showed that MLE has optimal rate of convergence in mean-field limit and long-time dynamics, when assuming linear interactions and no external force.

1.3. Our contributions

We provide a rigorous non-asymptotic analysis of MLE of drift coefficient restricted on a general class of functions with certain smoothness condition. Della Maestra and Hoffmann (2021) proposed a kernel based estimation procedure for the same estimation problem. However, unlike our method, the behaviour of estimation based on kernel method rely heavily on tuning the bandwidth and their analysis does not involve uniform laws of dependent variables. Moreover, the MLE framework provides a unified and principled strategy that naturally incorporates finer structures such as (2) in modelling the drift vector field b^*. In comparison, the kernel method requires further specialized steps for separating interaction force F^* from the external force G^* after the estimation of b^*. As a consequence, we do not need to explicitly specify the deconvolution operator ($\mathcal{L}$ in Assumption 2) and only need to assume its existence in our consistency analysis, making the MLE approach more robust to changes in problem characteristics and less sensitive to parameter tuning.

In our study, there are several obstacles while analyzing the MLE, some of which make our analysis technically more involved than that for the kernel method. Firstly, observations in $\mathcal{X}_T$ are not i.i.d. because of interaction among particles from the drift $b^*(\mu_t^N, \cdot)$. To decouple the dependence, we follow Della Maestra and Hoffmann (2021) by using Girsanov's theorem to construct a new measure, under which the trajectory of particles becomes i.i.d. However, this change of measure will introduce some additional decoupling errors in our analysis of the MLE that is not present in the analysis of the kernel method (Della Maestra and Hoffmann, 2021). Dealing with these decoupling errors requires substantial efforts and is technically highly non-trivial. Secondly, we derive a new and specialized maximal inequality (cf. Lemma 6) for handling the supreme of an unbounded process involving the Itô integral that appears in our analysis. The derived maximal inequality is general and interesting in its own right, and can be applied to other problems involving diffusion processes beyond our current setting. Thirdly, a standard union bound argument cannot

be applied to deal with the decoupling error (see the discussion after equation (15) for a precise meaning) in terms of the supreme of a random process expressed as the average of correlated Itô integrals that naturally appears when analyzing the MLE. To address this issue, we develop a concentration inequality for U-statistics involving Itô integrals (cf. Lemma 17), which is then combined with chaining and leads to a new maximal inequality for U-processes (cf. Lemma 10). This refined maximal inequality helps us derive a better rate in our problem than using existing general versions of the inequality.

1.4. Notation

Let $\mathbb{Z}$ ($\mathbb{N}$) denote the set of all (non-negative) integers. For any arbitrary functions $f : \mathbb{T}^d \to \mathbb{R}^d$, the Fourier series $(f)_k$ of f is defined as

$$(f_i)_k = \int_{\mathbb{T}^d} f_i(x) e^{-2\pi i k\cdot x}\,\mathrm{d}x, \quad 1 \le i \le d,\, k \in \mathbb{Z}^d,$$

where we let $f = (f_1, \cdots, f_d)^T$ and $(f)_k = ((f_1)_k, \cdots, (f_d)_k)^T$ are d-dimensional column vectors. Properties of Fourier analysis on torus can be found in Chapter 3 of Grafakos (2008).

For $k = (k_1, \cdots, k_d)^T \in \mathbb{Z}^d$, let $|k| = k_1 + \cdots + k_d$ be the sum of all elements of k, and $D^k = \partial_{k_1 \cdots k_d}$ is a $|k|$-th order partial derivative. We use $\|\cdot\|$ for l_2-norm of a vector, and $\|\cdot\|_2$ for $L^2(\mathbb{T}^d)$-norm of a (vector-valued) function, i.e., $\|f\|_2^2 = \int_{\mathbb{T}^d} \|f(x)\|^2\,\mathrm{d}x$. For a Lipschitz function f, we denote $\|f\|_{\mathrm{Lip}}$ is the smallest constant $C > 0$ such that $\|f(x) - f(y)\| \le \|x - y\|$ for all $x, y \in \mathbb{T}^d$. Let $\|\cdot\|_{H^1}$ be the Sobolev norm defined as $\|f\|_{H^1}^2 = \|f\|_2^2 + \sum_{i=1}^d \|\nabla f_i\|_2^2$. In addition, for a function $b(\nu, x)$, we define seminorms $\|b\|_E^2 := \int_0^T \int_{\mathbb{T}^d} \|b(\mu_t, x)\|^2\,\mathrm{d}\mu_t(x)\mathrm{d}t$, and $\|b\|_X^2 := N^{-1}\sum_{i=1}^N \int_0^T \left\|b(\mu_t, X_t^i)\right\|^2\,\mathrm{d}t$, and let the norm $\|\tilde{\cdot}\|_{\mathrm{Lip}}$ of any $b(\nu,\cdot) = \int_{\mathbb{T}^d} \tilde{b}(\cdot, y)\,\mathrm{d}\nu(y)$ be $\|\tilde{b}\|_{\mathrm{Lip}}$.

For $0 < \beta < \infty$, let ψ_β be the function on $[0,\infty)$ defined by $\psi_\beta(x) = e^{x^\beta} - 1$, and for a real-valued random variable ξ, define $\|\xi\|_{\psi_\beta} = \inf\{C > 0 : \mathbb{E}[\psi_\beta(|\xi|/C)] \le 1\}$. For $\beta \in [1,\infty)$, $\|\cdot\|_{\psi_\beta}$ is an *Orlicz norm*, while for $\beta \in (0,1)$, $\|\cdot\|_{\psi_\beta}$ is not a norm but a quasi-norm, i.e., there exists a constant C_β depending only on β such that $\|\xi_1 + \xi_2\|_{\psi_\beta} \le C_\beta(\|\xi_1\|_{\psi_\beta} + \|\xi_2\|_{\psi_\beta})$. Indeed, there is a norm equivalent to $\|\cdot\|_{\psi_\beta}$ obtained by linearizing ψ_β in a neighborhood of the origin; cf. Lemma C.2 in Chen and Kato (2019).

For a function class $\mathcal{H}$, define the shifted class $\mathcal{H}^* := \mathcal{H} - h^*$ for some $h^* \in \mathcal{H}$. The function class $\mathcal{H}^*$ is *star-shaped* (or equivalently $\mathcal{H}$ is star-shaped around h^*) if for any $h \in \mathcal{H}$ and $\alpha \in [0,1]$, the function $\alpha h \in \mathcal{H}^*$; cf. Chapter 13 of Wainwright (2019). We use $N(\varepsilon, \mathcal{H}, \|\cdot\|)$ to denote the ε-covering number for the function class $\mathcal{H}$ under the metric induced by the norm $\|\cdot\|$.

2. Constrained Maximum Likelihood Estimation

Let $\mathcal{C} = \mathcal{C}([0,T], (\mathbb{T}^d)^N)$ be the set of all continuous functions on $(\mathbb{T}^d)^N$, and $\{\mathcal{F}_t : 0 \le t \le T\}$ be the filtration generated by our observation $\mathcal{X}_T$. According to Girsanov's theorem (Theorem 1.12 in Kutoyants and Kutojanc (2004)), the log-likelihood ratio function for the continuous time trajectory data $\mathcal{X}_T$ takes the form as

$$L_T(b) := \log \frac{\mathrm{d}\mathbb{P}_b^N}{\mathrm{d}\mathbb{P}_0^N} = \sum_{i=1}^N \int_0^T \left\langle b(\mu_t^N, X_t^i), \mathrm{d}X_t^i \right\rangle - \frac{1}{2}\sum_{i=1}^N \int_0^T \left\|b(\mu_t^N, X_t^i)\right\|^2 \mathrm{d}t, \tag{4}$$

where $\frac{\mathrm{d}\mathbb{P}_b^N}{\mathrm{d}\mathbb{P}_0^N}$ denotes the Radon-Nikodym derivative of the probability measure $\mathbb{P}_b^N$ associated with $\mathcal{X}_T$ from model $\mathrm{d}X_t^i = b(t, \mu_t^N, X_t^i)\,\mathrm{d}t + \mathrm{d}W_t^i$, $1 \le i \le N$, relative to the base measure $\mathbb{P}_0^N$.

When the drift vector field b^* is driven by the external-interaction force in (2), it is natural to consider the maximum likelihood estimator (MLE) for b^* in the function class

$$\mathcal{H} = \left\{ b : \mathcal{P}(\mathbb{T}^d) \times \mathbb{T}^d \to \mathbb{R}^d \,\middle|\, \exists\, F, G \in \tilde{\mathcal{H}}, b(\nu, x) = G(x) + \int_{\mathbb{T}^d} F(x-y)\,\mathrm{d}\nu(y) \right\},$$

where $\tilde{\mathcal{H}}$ is a uniformly bounded function class whose elements map from $\mathbb{T}^d$ to $\mathbb{R}^d$ with certain smoothness (cf. assumptions in Theorem 1 below). Note that for $b \in \mathcal{H}$, we can equivalently compute the MLE $\widehat{b}_N := \operatorname{argmax}_{b\in\mathcal{H}} L_T(b)$ by first obtaining the MLE of F^* and G^* as in (2)

$$(\widehat{F}_N, \widehat{G}_N) = \operatorname*{argmax}_{F,G\in\tilde{\mathcal{H}}} \tilde{L}_T(F,G), \quad \text{subject to} \int_{\mathbb{T}^d} F(x)\,\mathrm{d}x = 0, \tag{5}$$

where
$$\tilde{L}_T(F,G) = -\frac{1}{2}\sum_{i=1}^N \int_0^T \left\| G(X_t^i) + \frac{1}{N}\sum_{j=1}^N F(X_t^i - X_t^j) \right\|^2 \mathrm{d}t + \sum_{i=1}^N \int_0^T \left\langle G(X_t^i) + \frac{1}{N}\sum_{j=1}^N F(X_t^i - X_t^j),\, \mathrm{d}X_t^i \right\rangle,$$

and then setting

$$\widehat{b}_N(\nu, x) = \widehat{G}_N(x) + \int_{\mathbb{T}^d} \widehat{F}_N(x-y)\,\mathrm{d}\nu(y), \quad \forall\, \nu \in \mathcal{P}(\mathbb{T}^d).$$

Note that for any solution $(\widehat{F}_N, \widehat{G}_N)$ of (5) and a constant $C \neq 0$, $(\widehat{F}_N + C, \widehat{G}_N - C)$ is also a solution. Therefore, we impose an additional restriction $\int_{\mathbb{T}^d} F^*(x)\,\mathrm{d}x = 0$ for the sake of identifiability of the interaction kernel. This also explains the extra constraint $\int_{\mathbb{T}^d} F(x)\,\mathrm{d}x = 0$ imposed in the estimation procedure (5).

3. Convergence Rate of Constrained MLE

In this section, we first derive a general rate of convergence for the constrained MLE $\hat{b}_N$ based on the entropy method which will lead to a computable and simple bound when specialized to α-smooth Hölder function class. In the latter case, we will show that the constrained MLE achieves the minimax optimal rate in Section 3.1. As a consequence, we derive the consistency $\widehat{G}_N$ and $\widehat{F}_N$ in Section 3.2 for the external-interaction force model (2).

Assumption 1 *The class $\tilde{\mathcal{H}}$ is* pointwise measurable*, i.e., it contains a countable subset $\mathcal{G}$ such that for every $h \in \mathcal{H}$ there exists a sequence $g_m \in \mathcal{G}$ such that $g_m \to h$ pointwise.*

Assumption 1 is made to avoid measurability issues (van der Vaart and Wellner, 2013) since it guarantees that the supremum of a suitable empirical process indexed by $\tilde{\mathcal{H}}$ is a measurable map. Define the *localized* function class

$$\mathcal{H}_u^* = \{ f \in \mathcal{H}^* : \|f\|_E \le u \}.$$

If $B = \sup_{f\in\mathcal{H}_u^*} \|f\|_\infty < \infty$, let $H(\varepsilon, \mathcal{H}_u^*)$ be the cardinality of the smallest set $S \subset \mathcal{H}_u^*$ such that $\forall\, g \in \mathcal{H}_u^*$ there exists $f \in S$ satisfying $\|f-g\|_E \le \varepsilon u$ and $\|f-g\|_\infty \le \varepsilon B$. We shall notice that $H(\varepsilon, \mathcal{H}^*)$ requires covering w.r.t. both $\|\cdot\|_\infty$ and $\|\cdot\|_E$, while $N(\varepsilon, \mathcal{H}^*, \|\cdot\|)$ only requires covering w.r.t. a certain norm $\|\cdot\|$. We further define several entropy integrals that control the complexity of our nonparametric estimation problem:

$$J_1(u) := \sqrt{T}B \int_0^{\frac{1}{2}} \log\big(1 + H(\varepsilon, \mathcal{H}_u^*)\big)\,\mathrm{d}\varepsilon + u\int_0^{\frac{1}{2}} \sqrt{\log\big(1 + H(\varepsilon, \mathcal{H}_u^*)\big)}\,\mathrm{d}\varepsilon,$$
$$J_2(L) := \int_0^{\frac{L}{2}} \log\big(1 + N(\varepsilon, \mathcal{H}^*, \|\cdot\|_{\mathrm{Lip}})\big)\,\mathrm{d}\varepsilon, \quad J_3(B) := \int_0^{\frac{B}{2}} \sqrt{\log\big(1 + N(u, \mathcal{H}^*, \|\cdot\|_\infty)\big)}\,\mathrm{d}u,$$
$$J_4(B) := \int_0^{\frac{B}{2}} \big[\log(1 + N(u, \mathcal{H}^*, \|\cdot\|_\infty))\big]^{\frac{3}{2}}\,\mathrm{d}u, \quad J_5(r) := \int_0^{\frac{r}{2}} \sqrt{\log N(s/\sqrt{T}, \mathcal{H}^*, \|\cdot\|_\infty)}\,\mathrm{d}s.$$

We assume that $J_1(u), J_2(L), J_3(B), J_4(B), J_5(r)$ are finite for some function class parameters on $\mathcal{H}^*$ and some localization parameter u.

Theorem 1 (Rate of convergence of constrained MLE) *Suppose the function class $\tilde{\mathcal{H}}$ satisfies Assumption 1 such that $\|F\|_\infty \le B$ and $\|F\|_{Lip} \le L$ for all $F \in \tilde{\mathcal{H}}$. Assume there exist positive constants δ_N and r_N satisfying*

$$\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g\in\mathcal{H}_{\delta_N}^*} \left| \frac{1}{N}\sum_{i=1}^N \int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle \right| \le 2\delta_N^2 \quad \textit{and} \quad J_5(r_N) \le \frac{\sqrt{N} r_N^2}{6CB\sqrt{T}},$$

where $(\overline{W}_t^i)_{t\ge 0}$ is the (transformed) Brownian motion defined in (12) and $\overline{\mathbb{P}}^N$ is the associated probability measure. If $b^* \in \mathcal{H}$, then

$$\|\widehat{b}_N - b^*\|_E \le 48\Big(\delta_N + r_N + \sqrt{\frac{\log N}{N}}\Big) \tag{6}$$

with probability at least

$$\begin{aligned} 1 - \bigg[\kappa_1 \exp\Big\{-\frac{\kappa_2 N\delta_N^2}{3}\Big\} + 3\kappa_1 \exp\Big\{-\frac{\kappa_2 N\delta_N^2}{C\log N J_1(1)}\Big\}\bigg] - 2\kappa_1 \exp\Big\{-\frac{\kappa_2 \log N}{4TK_1\|\tilde{b}^*\|_{Lip}^2}\Big\} \\ - 2\kappa_1 \exp\Big\{-\frac{\kappa_2 \log N}{CLK_1 J_2(L)}\Big\} - 2\kappa_1 \exp\Big\{-\frac{\kappa_2 N(\log N)^2}{CJ_3^2(B)}\Big\} \\ - 2\kappa_1 \exp\Big\{-\Big(\frac{\kappa_2 \log N}{CJ_4(B)\log\log N}\Big)^{\frac{2}{3}}\Big\} - \kappa_1 \exp\Big\{-\frac{\kappa_2 N r_N^2}{36C^2B^2T}\Big\}. \end{aligned} \tag{7}$$

Here $\kappa_1, \kappa_2, K_1, C$ are some positive constants.

There are some interesting remarks for Theorem 1 in order.

1. Theorem 1 remains true for drift vector field b^* with anisotropic interaction force, namely $b^*(\nu, x) = \int_{\mathbb{T}^d} \tilde{b}^*(x, y)\,\mathrm{d}\nu(y)$ with $\tilde{b}^* = G^*(x) + F^*(x, y)$. In this case, we need to consider functions in $\tilde{\mathcal{H}}$ map from $\mathbb{T}^{2d}$ to $\mathbb{R}^d$.

2. δ_N corresponds to the Gaussian complexity of function class $\mathcal{H}^*$ in a discrete setting, while r_N is an upper bound of Rademacher complexity of $\mathcal{H}^*$. Intuitively, the estimation problem will be harder if the function class is more complex, and thus δ_N and r_N should affect the rate of convergence in certain extent. It is quite common that the rate of convergence of nonparametric estimation depends on both δ_N and r_N, see e.g. Corollary 14.15 in Wainwright (2019) as an example.
3. Usually we use chaining method (Wainwright, 2019; Geer et al., 2000) to bound the expectation term in order to derive an explicit form of δ_N. Since we can show the sum of i.i.d. Itô integral is sub-exponential, one direct method is to use ψ_1-norm to bound the expectation (see Lemma 19). Another possible way to apply chaining is based on Bernstein-Orlicz norm (van de Geer and Lederer, 2013). In this case, bracketing number rather than covering number will be needed.
4. In practice, we can only observe discretely sampled trajectory data. Thus it is important to understand the impact of discretization. Note that the drift coefficient of an SDE is related to the mean function of the corresponding stochastic process. It is known that, in functional data analysis, convergence rate of estimation of mean function will not be affected if sampling frequency is large enough (Cai and Yuan, 2011). This is called phase *transition phenomenon*. We conjecture that a similar phase transition phenomenon may occur in our setting. In parametric setting, if parameters have linear effects on the interaction function, the problem is studied by Bishwal et al. (2011) and the convergence rate remains optimal as $O(n^{-1/2})$. A rigorous analysis in nonparametric setting is open and we leave it as the future work.
5. As is common in the nonparametric regression literature, the constrained MLE is usually adopted due to its technical simplicity; while the penalized MLE (e.g. by adding a squared RKHS norm penalty) is used for practical computation since they are dual to each other: with proper choice of the regularization parameter, they lead to the same solution. With discretely sampled trajectory data, the penalized MLE can be equivalently formulated as a finite dimensional optimization problem due to representer's theorem (kernel trick).

Our proof of Theorem 1 is quite involved. The proof builds on a number of recently developed technical tools such as change of measure via Girsanov's theorem for decoupling the IPS, concentration inequalities for unbounded empirical processes and degenerate U-processes, localization technique for sum of i.i.d. Itô integrals. We shall sketch the main structure of the argument in Section 4 and defer the complete proof details to Appendix.

3.1. Application to Hölder smooth function class

Consider the case where $\tilde{\mathcal{H}} = C_1^{\alpha}\big([0,1]^d\big)$ with fixed smooth parameter α and bounded α-Hölder norm. Here, the α-smooth Hölder function class is the set of all functions f such that

$$\|f\|_{\alpha} := \max_{|k|\le\lfloor\alpha\rfloor} \sup_x \left|D^k f(x)\right| + \max_{|k|=\lfloor\alpha\rfloor} \sup_{x\neq y} \frac{\left|D^k f(x) - D^k f(y)\right|}{\|x-y\|^{\alpha-\lfloor\alpha\rfloor}} \le M$$

with some fixed $M > 0$. In this case, assumptions of boundedness in Theorem 1 hold for $B = L = M$.

From Theorem 2.7.1 in van der Vaart and Wellner (2013) we know

$$\log N(\varepsilon, \tilde{\mathcal{H}}, \|\cdot\|_{\infty}) \lesssim \varepsilon^{-\frac{d}{\alpha}}.$$

By Theorem 1 we obtain the following rate of convergence in Corollary 2. The detailed proof is deferred to Appendix. Our theory works for Sobolev (Besov) space with bounded norm as well, and an analogy of the following corollary can be derived by a similar argument.

Corollary 2 (Hölder smooth drift estimation error) *Suppose the α-Hölder smooth function class $\tilde{\mathcal{H}}$ satisfies $\alpha > 3d/2$, then there are positive constants C and C' such that*

$$\left\|\widehat{b}_N - b^*\right\|_E \lesssim N^{-\frac{\alpha}{d+2\alpha}} \tag{8}$$

with probability at least $1 - C\exp\left\{-C'(\frac{\log N}{\log\log N})^{2/3}\right\}$.

Note that rate of convergence for the MLE derived in (8) attains the minimax rate of estimating the drift term in IPS (Della Maestra and Hoffmann, 2021). When $\alpha \leq 3d/2$, the term $J_4(B)$ in our general result Theorem 1 (also cf. Lemma 10 for definition) is not finite. However, we still can derive a rate of convergence for $\hat{b}_N$ by refining the definition of J_4 as the one in van de Geer (2014) to avoid the integrability issue. The price we pay is that the tail probability converges to zero more slowly, which would translate to a sub-optimal rate of convergence of $\widehat{b}_N$.

3.2. Estimating interaction kernel in Vlasov model

In this section, we specialize our theory to the external-interaction force system, and study the convergence behaviour of interaction kernel F^*. This is an inverse problem and can be done in a similar argument as in Della Maestra and Hoffmann (2021) by Fourier transform. Della Maestra and Hoffmann (2021) *explicitly* use Fourier transform to construct an estimator based on deconvolving a kernel density estimator for μ_t from an estimator for $F^* * \mu_t$, where $*$ denotes the function convolution operator. In comparison, we *implicitly* use Fourier transform to derive a stability estimate for translating an error bound on $\widehat{b}_N$ to that on $\widehat{F}_N$ in the analysis. More specifically, recall that

$$\widehat{b}_N(\mu_t, x) - b^*(\mu_t, x) = \widehat{G}_N(x) - G^*(x) + \int_{\mathbb{T}^d} \left(\widehat{F}_N(x-y) - F^*(x-y)\right) \mathrm{d}\mu_t(y). \tag{9}$$

Let $L^2([0,T])$ denote the space of all square-integrable functions on $[0,T]$ and view $\mu(x,t) = \mu_t(x)$ as a function of (x,t). For any linear operator $\mathcal{L} : L^2([0,T]) \to \mathbb{R}$, $f \mapsto \mathcal{L}f := \int_0^T f(t)w(t)\,\mathrm{d}t$, where w is a bounded measurable function on $[0,T]$ such that $\int_0^T w(t)\,\mathrm{d}t = 0$, we obtain by applying $\mathcal{L}$ to both sides of (9),

$$\mathcal{L}\big[(\widehat{b}_N - b^*)(\mu, x)\big] = \mathcal{L}\big[(\widehat{F}_N - F^*) * \mu(x)\big] = \big((\widehat{F}_N - F^*) * \mathcal{L}\mu\big)(x), \tag{10}$$

where we have used the property that $\mathcal{L}g = 0$ for any t-independent function g. Since the goal is to relate $\|\widehat{F}_N - F^*\|_2$ to $\|\widehat{b}_N - b^*\|_E$, we may apply Fourier transform to both sides of (10) to deconvolute $\widehat{F}_N - b^*$ and $\mathcal{L}\mu$, leading to

$$\Big(\mathcal{L}\big[(\widehat{b}_N - b^*)(\mu, \cdot)\big]\Big)_k = (\widehat{F}_N - F^*)_k \cdot (\mathcal{L}\mu)_k, \quad \forall\, k \in \mathbb{Z}^d. \tag{11}$$

Note that by the definition of $\mathcal{L}$, we have $(\mathcal{L}\mu)_0 = \mathcal{L}\int_{\mathbb{T}^d} \mathrm{d}\mu(x) = \mathcal{L}1 = 0$, so equation (11) only determines the Fourier coefficient of $(\widehat{F}_N - F^*)_k$ for $k \neq 0$. However, $(\widehat{F}_N - F^*)_0$ can be uniquely determined by our additional identifiability constraint $\int_{\mathbb{T}^d} \widehat{F}_N(x)\,\mathrm{d}x = \int_{\mathbb{T}^d} F^*(x)\,\mathrm{d}x = 0$.

To ensure that $\widehat{F}_N - F^*$ remains small when $\widehat{b}_N$ is close to b^*, we need the following assumption motivated by identity (11).

Assumption 2 *There exists a bounded measurable function $w(t)$ on $[0,T]$ such that $\int_0^T w(t)\,\mathrm{d}t = 0$ and $(\mathcal{L}\mu)_k = \int_0^T (\mu_t)_k\, w(t)\,\mathrm{d}t \neq 0$ for all nonzero $k \in \mathbb{Z}^d$.*

Assumption 2 guarantees the identifiability of interaction force F^* (and therefore external force G^*) from the drift vector field b^*, and requires the system to be away from stationarity. To see this, consider the ideal setting where we exactly know the true b^* and $\{\mu_t : 0 \le t \le T\}$, and want to uniquely recover F^* from (2). Suppose system (1) already attains stationarity, i.e. $\mu_t = \mu^*$ for all $t \in [0,T]$ with μ^* denoting the stationary distribution of the system which solves equation (3) with $\partial_t \mu = 0$. Then it is impossible to separate out the time-homogeneous interaction term $F^* * \mu^*$ from the drift vector field

$$b^*(\mu^*, x) = G^*(x) + \int_{\mathbb{T}^d} F^*(x-y)\,\mathrm{d}\mu^*(y),$$

since for any function $g : \mathbb{T}^d \to \mathbb{R}$, the new pair $(G', F') = (G^* - g * \mu^*,\, F^* + g)$ induces the same drift b^*. However, this setting violates Assumption 2 since $(\mathcal{L}\mu)_k = 0$ for all $k \in \mathbb{Z}^d$. In other words, the interaction force F^* can only be recovered from the transient behaviour of the system, and Assumption 2 is one mathematical description implying the system to be away from stationarity.

The linearly dependence structure (11) in the frequency domain suggests that the estimation error of $\widehat{F}_N$ depends on the accuracy of $\widehat{b}_N$ and the behaviour of $(\mathcal{L}\mu)_k$. It is common that an inverse problem related to a convolution equation, such as equation (10), tends to be numerically unstable, since it will be ill-posed when the Fourier coefficients $\{(\mathcal{L}\mu)_k\}_{k=1}^{\infty}$ of $\mathcal{L}\mu$ decay too fast (Isakov, 2006). By quantifying the stability of solution to (10) around the true interaction F^*, we arrive at the following corollary.

Corollary 3 (Interaction kernel estimation error) *Let η_N be the smallest integer satisfying $\eta_N(\delta_N + r_N + \frac{\log N}{N}) \le C_1 \inf_{0<\|k\|<\eta_N} \|(\mathcal{L}\mu)_k\|$. If F^* and G^* belong to $\tilde{\mathcal{H}}$ and both have finite H_1-norms, then*

$$\left\|\widehat{F}_N - F^*\right\|_2 \le C_2\, \eta_N^{-1}$$

holds with at least probability given by (7). Here constants $(C_1, , C_2)$ are independent of N.

Remark 4 *In the proof of Corollary 3, we only used the condition that the Sobolev norm $\|F^*\|_{H^1}$ is finite. If we further use the condition that F^* is α-Hölder continuous with $\alpha > 1$, then the error bound can be improved to $\eta_N^{-\alpha}$ where $\eta_N^{\alpha}(\delta_N + r_N + \frac{\log N}{N}) \lesssim \inf_{0<\|k\|<\eta_N} \|(\mathcal{L}\mu)_k\|$. In particular, if Assumption 2 holds, then Corollary 3 implies $\widehat{F}_N$ to be a consistent estimator of F^* as $N \to \infty$.*

4. Proof of Theorem 1

By decomposing the sample space into $\mathcal{E} = \{\left\|\widehat{\Delta}_N\right\|_E \le \delta_N\}$ and $\mathcal{E}^c$, Theorem 1 is true on the event $\mathcal{E}$. So, we only need the proof on the event $\mathcal{E}^c = \{\left\|\widehat{\Delta}_N\right\|_E > \delta_N\}$.

Step 1: decoupling. The first technical difficulty is to decouple the interaction effect between particles that would cause the dependence of particle trajectories. Let $\mathbb{P}^N$ be the probability measure on $(\mathcal{C}, \{\mathcal{F}_t\}_{t=0}^T)$ induced by solution of the true data generating mechanism (1). Motivated by the

change of measure argument in Lacker (2018) and Della Maestra and Hoffmann (2021), we construct a new measure $\overline{\mathbb{P}}^N$ on $(\mathcal{C}, \{\mathcal{F}\}_{t=0}^T)$, under which $(X_t^i)_{t=0}^T$ and $(X_t^j)_{t=0}^T$ are independent for all $1 \le i \ne j \le N$. This is possible thanks to Girsanov's Theorem. Specifically, define a process $\{Z_t\}_{t=0}^T$ as

$$Z_t = \exp\left\{ \sum_{i=1}^N \left[\int_0^t \langle b^*(\mu_s, X_s^i) - b^*(\mu_s^N, X_s^i), \mathrm{d}W_s^i \rangle - \frac{1}{2} \int_0^t \|b^*(\mu_s, X_s^i) - b^*(\mu_s^N, X_s^i)\|^2 \mathrm{d}s \right] \right\},$$

and let $\mathrm{d}\overline{\mathbb{P}}^N = Z_T \, \mathrm{d}\mathbb{P}^N$. By Girsanov's Theorem,

$$\overline{W}_t^i := W_t^i - \int_0^t \left[b^*(\mu_s, X_s^i) - b^*(\mu_s^N, X_s^i) \right] \mathrm{d}s, \quad 0 \le t \le T \tag{12}$$

are i.i.d. Brownian motions on $\mathbb{T}^d$ under $\overline{\mathbb{P}}^N$. With the transformation (12), the original IPS (1) turns into a system of *independent* SDEs given by

$$\mathrm{d}X_t^i = b^*(\mu_t, X_t^i) \, \mathrm{d}t + \mathrm{d}\overline{W}_t^i \tag{13}$$

with i.i.d. initialization $\mathcal{L}(X_0^1, \dots, X_0^N) = \otimes_{i=1}^N \mu_0$. Our subsequent strategy for analyzing the log-likelihood ratio is to control the probability of some "bad events" under $\overline{\mathbb{P}}^N$, and then to use the following Lemma 5 to convert it back to the probability under $\mathbb{P}^N$. The proof of Lemma 5 can be found in Theorem 18 in Della Maestra and Hoffmann (2021).

Lemma 5 (Change of measure equivalence) *There are positive constants κ_1 and κ_2 such that for any $\mathcal{F}_T$-measurable event $\mathcal{B}$,*

$$\mathbb{P}^N(\mathcal{B}) \le \kappa_1 \overline{\mathbb{P}}^N(\mathcal{B})^{\kappa_2}.$$

Step 2: basic inequality. By definition of $\widehat{b}_N$, we have $L_T(\widehat{b}_N) \ge L_T(b^*)$ and thus

$$\begin{aligned} \sum_{i=1}^N \int_0^T \langle (\widehat{b}_N - b^*)(\mu_t^N, X_t^i), \, b^*(\mu_t, X_t^i) \, \mathrm{d}t + \mathrm{d}\overline{W}_t^i \rangle \\ - \frac{1}{2} \sum_{i=1}^N \int_0^T \left\{ \|\widehat{b}_N(\mu_t^N, X_t^i)\|^2 - \|b^*(\mu_t^N, X_t^i)\|^2 \right\} \mathrm{d}t \ge 0. \end{aligned}$$

Denote $\widehat{\Delta}_N = \widehat{b}_N - b^*$, we can derive the basic inequality

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \int_0^T \langle \widehat{\Delta}_N(\mu_t^N, X_t^i), \, \mathrm{d}\overline{W}_t^i \rangle \ge \frac{1}{2N} \sum_{i=1}^N \int_0^T \|\widehat{b}_N(\mu_t^N, X_t^i) - b^*(\mu_t, X_t^i)\|^2 \, \mathrm{d}t \\ - \frac{1}{2N} \sum_{i=1}^N \int_0^T \|b^*(\mu_t^N, X_t^i) - b^*(\mu_t, X_t^i)\|^2 \, \mathrm{d}t. \end{aligned} \tag{14}$$

Furthermore, by noticing that

$$\begin{aligned} &\widehat{b}_N(\mu_t^N, X_t^i) - b^*(\mu_t, X_t^i) \\ &= \left[\widehat{\Delta}_N(\mu_t^N, X_t^i) - \widehat{\Delta}_N(\mu_t, X_t^i) \right] + \left[b^*(\mu_t^N, X_t^i) - b^*(\mu_t, X_t^i) \right] + \left[\widehat{b}_N(\mu_t, X_t^i) - b^*(\mu_t, X_t^i) \right], \end{aligned}$$

we can obtain by using the Cauchy-Schwartz inequality and basic inequality (14) that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\int_0^T \langle \widehat{\Delta}_N(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle \geq \frac{1}{6N}\sum_{i=1}^{N}\int_0^T \left\|\widehat{\Delta}_N(\mu_t, X_t^i)\right\|^2 \mathrm{d}t \\ - \frac{1}{N}\sum_{i=1}^{N}\int_0^T \left\|b^*(\mu_t^N, X_t^i) - b^*(\mu_t, X_t^i)\right\|^2 \mathrm{d}t \\ - \frac{1}{2N}\sum_{i=1}^{N}\int_0^T \left\|\widehat{\Delta}_N(\mu_t^N, X_t^i) - \widehat{\Delta}_N(\mu_t, X_t^i)\right\|^2 \mathrm{d}t \\ - \frac{1}{N}\sum_{i=1}^{N}\int_0^T \langle \widehat{\Delta}_N(\mu_t^N, X_t^i) - \widehat{\Delta}_N(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle.\end{aligned} \tag{15}$$

We will call the last three terms on the right hand side of the above display as *decoupling errors* since they characterize the degree of dependence among the particle trajectories due to the presence of empirical law μ_t^N in the drift term b^* of SDE (1). To bound the left-hand side of (15), we need the following Lemma 6.

Lemma 6 (Localization) *For the star-shaped set $\mathcal{H}^*$, define*

$$\mathcal{A}(u) = \left\{\exists\, g \in \mathcal{H}^*, \|g\|_E \geq u : \left|\frac{1}{N}\sum_{i=1}^{N}\int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle\right| \geq 4u\|g\|_E\right\},$$

and critical radius δ_N as the smallest number $u > 0$ such that

$$\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g\in\mathcal{H}_u^*}\left|\frac{1}{N}\sum_{i=1}^{N}\int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle\right| \leq 2u^2.$$

Recall $H(\varepsilon, \mathcal{H}_u^*)$ is the cardinality of the smallest set $S \subset \mathcal{H}_u^*$, such that $\forall\, g \in \mathcal{H}_u^*$, there exists $f \in S$ satisfying $\|f - g\|_E \leq \varepsilon u$ and $\|f - g\|_\infty \leq \varepsilon B$. Then, for every $u \geq \delta_N$ we have

$$\overline{\mathbb{P}}^N\big(\mathcal{A}(u)\big) \leq \exp\left\{-\frac{N\delta_N^2}{3}\right\} + 3\exp\left\{-\frac{Nu\delta_N}{C\log N J_1(1)}\right\}.$$

Thus, on the event $\mathcal{A}(\delta_N)^c$ where δ_N is the critical radius of our nonparametric estimation problem, we have

$$\frac{1}{6N}\sum_{i=1}^{N}\int_0^T \left\|\widehat{\Delta}_N(\mu_t, X_t^i)\right\|^2 \mathrm{d}t \leq 4\delta_N\|\hat{\Delta}_N\|_E + T_1 + 0.5T_2 + T_3, \tag{16}$$

where the three error terms

$$\begin{aligned} T_1 &= \frac{1}{N}\sum_{i=1}^{N}\int_0^T \left\|b^*(\mu_t^N, X_t^i) - b^*(\mu_t, X_t^i)\right\|^2 \mathrm{d}t, \\ T_2 &= \sup_{g\in\mathcal{H}^*}\frac{1}{N}\sum_{i=1}^{N}\int_0^T \left\|g(\mu_t^N, X_t^i) - g(\mu_t, X_t^i)\right\|^2 \mathrm{d}t, \\ T_3 &= \frac{1}{N}\sum_{i=1}^{N}\int_0^T \langle \widehat{\Delta}_N(\mu_t^N, X_t^i) - \widehat{\Delta}_N(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle. \end{aligned}$$

are expected to decay to zero as $N \to \infty$ since μ_t^N is expected to be close to μ_t. However, a rigorous analysis of these three error terms requires substantial efforts due to their complicated structures and the need of a uniform control over $\mathcal{H}^*$, which will be sketched below.

Step 3: bound T_1. To bound T_1, we use the following lemma. Recall that $\|f\|_{\text{Lip}}$ denotes the Lipschitz constant of f.

Lemma 7 (Decoupling error bound) *For any $u > 0$, $N \geq 2$ and $g \in \mathcal{H}^*$, we have*

$$\overline{\mathbb{P}}^N\left(\frac{1}{N}\sum_{i=1}^N \int_0^T \left\|g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i)\right\|^2 \mathrm{d}t > u^2\right) \leq 2e^{-\frac{Nu^2}{4TK_1\|\tilde{g}\|_{Lip}^2}}.$$

Applying Lemma 7 with $g = b^*$ and $u^2 = \log(N)/N$, we can conclude that $T_1 \leq \frac{\log N}{N}$ holds with probability at least $1 - 2\exp\big(-\frac{\log n}{4TK_1\|\tilde{b^*}\|_{\text{Lip}}^2}\big)$. Lemma 7 tells us that the decoupling error is sub-Gaussian with parameter of order $O(N^{-1})$, although the summands are not independent. In fact, we can decompose

$$\begin{aligned} g(\mu_t^N, X_t^i) - g(\mu_t, X_t^i) = \frac{1}{N}\left[\tilde{g}(X_t^i, X_t^i) - \int_{\mathbb{T}^d} \tilde{g}(X_t^i, y)\,\mathrm{d}\mu_t(y)\right] \\ + \frac{1}{N}\sum_{j=1, j\neq i}^N \left[\tilde{g}(X_t^i, X_t^j) - \int_{\mathbb{T}^d} \tilde{g}(X_t^i, y)\,\mathrm{d}\mu_t(y)\right]. \end{aligned}$$

Notice that the second term above is a summation of $(N-1)$ i.i.d. centered random variables under $\overline{\mathbb{P}}^N$ conditioning on X_t^i. This is where the sub-Gaussianity comes from. Intuitively, the second decoupling error T_2 (or third line of (15)) should have the same order $O(N^{-1})$ by some discretization or chaining arguments. This will be done in Lemma 8.

Step 4: bound T_2. To bound T_2, we use the following lemma.

Lemma 8 (Uniform laws of dependent variables) *We have*

$$\overline{\mathbb{P}}^N\left(\sup_{g\in\mathcal{H}^*} \frac{1}{N}\sum_{i=1}^N \int_0^T \left\|g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i)\right\|^2 \mathrm{d}t > u\right) \leq 2e^{-\frac{Nu}{CLK_1J_2(L)}}$$

for some constant $C > 0$.

Remark 9 *In order for $J_2(L)$ to be finite, we need functions in $\mathcal{H}^*$ to have higher-order smoothness than just being Lipschitz continuous, so that the covering number with respect to $\|\cdot\|_{Lip}$ is finite.*

Step 5: bound T_3. The last decoupling error T_3 in (16) (or last line of (15)) is much more involved to control. This is because the Ito integral is expected to only have the order of the square root of its quadratic variation. Notice that for each $g(\nu, x) = \int_{\mathbb{T}^d} \tilde{g}(x, y)\,\mathrm{d}\nu(y)$, we have

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^N \int_0^T \left\langle g(\mu_t^N, X_t^i) - g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\right\rangle \\ &= \frac{1}{N^2}\sum_{i=1}^N \int_0^T \left\langle \xi_{i,i}^g(t),\, \mathrm{d}\overline{W}_t^i\right\rangle + \frac{1}{N^2}\sum_{1\leq i\neq j\leq N} \int_0^T \left\langle \xi_{i,j}^g(t),\, \mathrm{d}\overline{W}_t^i\right\rangle =: V_N(g) + U_N(g), \end{aligned} \tag{17}$$

where we used the shorthand $\xi_{i,j}^g(t) = \tilde{g}(X_t^i, X_t^j) - \int_{\mathbb{T}^d} \tilde{g}(X_t^i, y)\,\mathrm{d}\mu_t(y)$ for notation simplicity. The first term $V_N(g)$ in (17) is a sum of i.i.d. random variable, which is expected to have order $O(N^{-\frac{3}{2}})$. The second term $U_N(g)$ can be viewed as a U-statistics (after proper symmetrization), with kernel function $g^\dagger(\overline{W}^i, \overline{W}^j) := \int_0^T \langle \xi_{i,j}^g(t),\, \mathrm{d}\overline{W}_t^i \rangle$. It is easy to verify that $\mathbb{E}_{\overline{\mathbb{P}}^N}\big[g^\dagger(\overline{W}^i, \overline{W}^j) \,|\, \overline{W}^i\big] = \mathbb{E}_{\overline{\mathbb{P}}^N}\big[g^\dagger(\overline{W}^i, \overline{W}^j) \,|\, \overline{W}^j\big] = 0$. So $g^\dagger$ is degenerate (Definition 3.5.1 in de la Pena and Giné (1999)), indicating that the U-statistics should have order around $O(N^{-1})$ asymptotically (see Chapter 3 of Lee (1990)). Above discussions can be rigorously stated as Lemma 16 and Lemma 17 under a non-asymptotic setting. Then the last term T_3 in (16) can be bounded by the following lemma, which gives an optimal upper bound $O(N^{-1})$ of the decoupling error up to some log factor.

Lemma 10 (Uniform laws of dependent Itô integrals) *We have*

$$\sup_{g\in\mathcal{H}^*} \frac{1}{N}\sum_{i=1}^N \int_0^T \langle g(\mu_t^N, X_t^i) - g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i \rangle \le \frac{2\log N}{N}$$

with probability at least

$$1 - 2\exp\Big\{-\frac{N(\log N)^2}{CJ_3^2(B)}\Big\} - 2\exp\Big\{-\Big(\frac{\log N}{CJ_4(B)\log\log N}\Big)^{\frac{2}{3}}\Big\}.$$

Step 6: conclude. To finish the bound for the estimation error under the $\|\cdot\|_E$ norm, we need the following norm equivalence between $\|\cdot\|_E$ and its empirical version $\|\cdot\|_X$.

Lemma 11 (Equivalence of norms) *There is a constant $C > 0$ such that*

$$\overline{\mathbb{P}}^N\Big(\sup_{g\in\mathcal{H}^*}\Big|\|g\|_X^2 - \|g\|_E^2\Big| > \frac{1}{2}\|g\|_E^2 + \frac{u^2}{2}\Big) < \exp\Big\{-\frac{Nu^2}{36C^2B^2T}\Big\}$$

for all $u \ge r_N$. Here $r_N > 0$ is a constant satisfying $J_5(r_N) \le \frac{\sqrt{N}r_N^2}{6CB\sqrt{T}}$.

Return to the proof. Since $\|\widehat{\Delta}_N\|_E > \delta_N$, we can obtain by combining all pieces with (16) that

$$4\delta_N\|\widehat{\Delta}_N\|_E \overset{\text{(i)}}{\ge} \frac{1}{6}\|\widehat{\Delta}_N\|_X^2 - \frac{\log N}{N} - \frac{\log N}{2N} - \frac{2\log N}{N} \overset{\text{(ii)}}{\ge} \frac{1}{12}\|\widehat{\Delta}_N\|_E^2 - \frac{r_N^2}{2} - \frac{7\log N}{N}$$

with at least probability given in (7). Here, (i) is by Lemma 6 (taking $u = \delta_N$), Lemma 7 (taking $u^2 = \frac{\log N}{N}$), Lemma 8 (taking $u = \frac{\log N}{N}$), and Lemma 10, and (ii) is by Lemma 11 (taking $u = r_N$). This implies the desired bound (6).

Acknowledgments

Xiaohui Chen was partially supported by NSF CAREER Award DMS-1752614. Yun Yang was partially supported by NSF DMS-1907316.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

Martin T Barlow and Marc Yor. Semi-martingale inequalities via the garsia-rodemich-rumsey lemma, and applications to local times. *Journal of functional Analysis*, 49(2):198–229, 1982.

Jaya Prakash Narayan Bishwal et al. Estimation in interacting diffusions: Continuous and discrete sampling. *Applied Mathematics*, 2(9):1154–1158, 2011.

Mattia Bongini, Massimo Fornasier, Markus Hansen, and Mauro Maggioni. Inferring interaction rules from observations of evolutive systems i: The variational approach. *Mathematical Models and Methods in Applied Sciences*, 27(05):909–951, 2017.

Rainer Buckdahn, Juan Li, and Jin Ma. A mean-field stochastic control problem with partial observations. *The Annals of Applied Probability*, 27(5):3201 – 3245, 2017. doi: 10.1214/17-AAP1280. URL https://doi.org/10.1214/17-AAP1280.

T Tony Cai and Ming Yuan. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The annals of statistics*, 39(5):2330–2355, 2011.

René Carmona. *Lectures on BSDEs, stochastic control, and stochastic differential games with financial applications*. SIAM, 2016.

René Carmona and Francois Delarue. *Probabilistic Theory of Mean Field Games with Applications I*. Springer International Publishing, 2018.

Xiaohui Chen. Maximum likelihood estimation of potential energy in interacting particle systems from single-trajectory data. *Electronic Communications in Probability*, 26:1–13, 2021.

Xiaohui Chen and Kengo Kato. Randomized incomplete u-statistics in high dimensions. *The Annals of Statistics*, 47(6):3127–3156, 2019.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Vıctor H de la Pena and Evarist Giné. Decoupling. probability and its applications, 1999.

Laetitia Della Maestra and Marc Hoffmann. Nonparametric estimation for interacting particle systems: Mckean–vlasov models. *Probability Theory and Related Fields*, pages 1–63, 2021.

M. R. D'Orsogna, Y. L. Chuang, A. L. Bertozzi, and L. S. Chayes. Self-propelled particles with soft-core interactions: Patterns, stability, and collapse. *Phys. Rev. Lett.*, 96:104302, Mar 2006. doi: 10.1103/PhysRevLett.96.104302. URL https://link.aps.org/doi/10.1103/PhysRevLett.96.104302.

Sara A Geer, Sara van de Geer, and D Williams. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for u-statistics. In *High Dimensional Probability II*, pages 13–38. Springer, 2000.

Loukas Grafakos. *Classical fourier analysis*, volume 2. Springer, 2008.

Victor Isakov. *Inverse problems for partial differential equations*, volume 127. Springer, 2006.

Raphael A Kasonga. Maximum likelihood theory for large interacting systems. *SIAM Journal on Applied Mathematics*, 50(3):865–875, 1990.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

Yury A Kutoyants and Jurij A Kutojanc. *Statistical inference for ergodic diffusion processes*. Springer Science & Business Media, 2004.

Daniel Lacker. On a strong form of propagation of chaos for mckean-vlasov equations. *Electronic Communications in Probability*, 23:1–11, 2018.

Quanjun Lang and Fei Lu. Learning interaction kernels in mean-field equations of 1st-order systems of interacting particles. *arXiv preprint arXiv:2010.15694*, 2020.

Quanjun Lang and Fei Lu. Identifiability of interaction kernels in mean-field equations of interacting particles. *arXiv preprint arXiv:2106.05565*, 2021.

A J Lee. *U-statistics: Theory and Practice*. Routledge, 1990.

Zhongyang Li, Fei Lu, Mauro Maggioni, Sui Tang, and Cheng Zhang. On the identifiability of interaction functions in systems of interacting particles. *Stochastic Processes and their Applications*, 132:135–163, 2021.

Meiqi Liu and Huijie Qiao. Parameter estimation of path-dependent mckean-vlasov stochastic differential equations. *arXiv preprint arXiv:2004.09580*, 2020.

Qiang Liu. Stein variational gradient descent as gradient flow. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/17ed8abedc255908be746d245e50263a-Paper.pdf.

Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences*, 116 (29):14424–14433, 2019a.

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, 2019b. doi: 10.1137/18M1187611. URL https://doi.org/10.1137/18M1187611.

Pascal Massart. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

H. P. McKean. A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911, 1966. ISSN 0027-8424. doi: 10.1073/pnas.56.6.1907. URL https://www.pnas.org/content/56/6/1907.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/content/115/33/E7665.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *32nd Annual Conference on Learning Theory*, 2019.

Alexander Mogilner and Leah Edelstein-Keshet. A non-local model for a swarm. *Journal of Mathematical Biology*, 38(6):534–570, 1999. doi: 10.1007/s002850050158. URL https://doi.org/10.1007/s002850050158.

Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014. doi: 10.1137/120901866. URL https://doi.org/10.1137/120901866.

Richard Nickl and Kolyan Ray. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *The Annals of Statistics*, 48(3):1383–1408, 2020.

Yvo Pokern, Andrew M Stuart, and J Harry van Zanten. Posterior consistency via precision operators for bayesian nonparametric drift estimation in sdes. *Stochastic Processes and their Applications*, 123(2):603–628, 2013.

Louis Sharrock, Nikolas Kantas, Panos Parpas, and Grigorios A Pavliotis. Parameter estimation for the mckean-vlasov stochastic differential equation. *arXiv preprint arXiv:2106.13751*, 2021.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020a. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2019.06.003. URL https://www.sciencedirect.com/science/article/pii/S0304414918306197.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020b. doi: 10.1137/18M1192184. URL https://doi.org/10.1137/18M1192184.

Alain-Sol Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg. ISBN 978-3-540-46319-1.

Chad M. Topaz, Andrea L. Bertozzi, and Mark A. Lewis. A nonlocal continuum model for biological aggregation. *Bulletin of Mathematical Biology*, 68(7):1601, 2006. doi: 10.1007/s11538-006-9088-6. URL https://doi.org/10.1007/s11538-006-9088-6.

Sara van de Geer. On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electronic Journal of Statistics*, 8(1):543–574, 2014.

Sara van de Geer and Johannes Lederer. The bernstein–orlicz norm and deviation inequalities. *Probability theory and related fields*, 157(1):225–250, 2013.

Aad Van Der Vaart and Jon A Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5(2011):192, 2011.

Ada van der Vaart and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Jan van Waaij and Harry van Zanten. Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electronic Journal of Statistics*, 10(1):628–645, 2016.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Jianghui Wen, Xiangjun Wang, Shuhua Mao, and Xinping Xiao. Maximum likelihood estimation of mckean–vlasov stochastic differential equation and its application. *Applied Mathematics and Computation*, 274:237–246, 2016.

Appendix

Appendix A. Proof of Main Results

Proof [Proof of Lemma 6] Let

$$\mathcal{Z}_N(u) = \sup_{g\in\mathcal{H}_u^*} \left| \frac{1}{N}\sum_{i=1}^{N} \int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle \right|$$

First, we shall show that

$$\mathcal{A}(u) \subset \left\{\mathcal{Z}_N(u) \geq C_3 u^2\right\}. \tag{18}$$

In fact, assume there is $g \in \mathcal{H}^*$, such that $\|g\|_E \geq u$ and

$$\left| \frac{1}{N}\sum_{i=1}^{N} \int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle \right| \geq C_3 u \|g\|_E. \tag{19}$$

Let $h = \frac{u}{\|g\|_E} g$. Since $\|g\|_E \geq u$ and $\mathcal{H}^*$ is star-shaped, we know $h \in \mathcal{H}^*$ and $\|h\|_E = u$. Then (19) can be reformulated as

$$\left| \frac{1}{N}\sum_{i=1}^{N} \int_0^T \langle h(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle \right| \geq C_3 u^2.$$

Now we have finished (18). By a deviation inequality (Theorem 4 in Adamczak (2008)),

$$\begin{aligned}
&\overline{\mathbb{P}}^N\Big(\mathcal{Z}_N(u) \geq 1.5\mathbb{E}_{\overline{\mathbb{P}}^N}\mathcal{Z}_N(u) + t\Big) \\
&\leq \exp\left\{-\frac{t^2}{3\sigma^2}\right\} + 3\exp\left\{-\frac{t}{C\big\|\max_i \sup_{g\in\mathcal{H}_u^*} \frac{1}{N}\int_0^T \langle g(\mu_t, X_t^i), \mathrm{d}\overline{W}_t^i\rangle\big\|_{\psi_1}}\right\},
\end{aligned}$$

with

$$\sigma^2 = \sup_{g\in\mathcal{H}_u^*} \frac{1}{N^2}\sum_{i=1}^{N} \mathbb{E}_{\overline{\mathbb{P}}^N}\left(\int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle\right)^2 \leq \sup_{g\in\mathcal{H}_u^*} \frac{1}{N}\|g\|_E^2 \leq \frac{u^2}{N}.$$

Then take $t = u\delta_N$ in the above deviation inequality, and we have

$$\begin{aligned}
&\overline{\mathbb{P}}^N\Big(\mathcal{Z}_N(u) \geq 1.5\mathbb{E}_{\overline{\mathbb{P}}^N}\mathcal{Z}_N(u) + u\delta_N\Big) \\
&\leq \exp\left\{-\frac{N\delta_N^2}{3}\right\} + 3\exp\left\{-\frac{Nu\delta_N}{C\big\|\max_i \sup_{g\in\mathcal{H}_u^*} \int_0^T \langle g(\mu_t, X_t^i), \mathrm{d}\overline{W}_t^i\rangle\big\|_{\psi_1}}\right\}.
\end{aligned} \tag{20}$$

Notice that we have

$$\begin{aligned}
&\left\| \max_{1\le i\le N} \sup_{g\in\mathcal{H}_u^*} \int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle \right\|_{\psi_1} \\
&\overset{\text{(i)}}{\lesssim} \log N \max_{1\le i\le N} \left\| \sup_{g\in\mathcal{H}_u^*} \int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle \right\|_{\psi_1} \\
&\le \log N \left\| \sup_{g\in\mathcal{H}_u^*} \int_0^T \langle g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i\rangle \right\|_{\psi_1} \\
&\overset{\text{(ii)}}{\lesssim} \log N \left[\sqrt{T} B \int_0^{\frac{1}{2}} \log\left(1 + H(\varepsilon, \mathcal{H}^*)\right) \mathrm{d}\varepsilon + \int_0^{\frac{1}{2}} \sqrt{\log\left(1 + H(\varepsilon, \mathcal{H}_u^*)\right)}\, \mathrm{d}\varepsilon \right] \\
&= \log N \cdot J_1(1).
\end{aligned}$$

Here (i) is by Lemma 8.2 in Kosorok (2008), and (ii) is by taking $N = 1$ in Lemma 19. Since $u \ge \delta_N$, by Lemma 13 we know

$$\frac{\mathbb{E}_{\overline{\mathbb{P}}^N} \mathcal{Z}_N(u)}{u} \le \frac{\mathbb{E}_{\overline{\mathbb{P}}^N} \mathcal{Z}_N(\delta_N)}{\delta_N} \le 2\delta_N,$$

i.e. $\mathbb{E}_{\overline{\mathbb{P}}^N} \mathcal{Z}_N(u) \le 2u\delta_N$. Therefore,

$$\begin{aligned}
&\overline{\mathbb{P}}^N\big(\mathcal{A}(u)\big) \le \overline{\mathbb{P}}^N\big(\mathcal{Z}_N(u) \ge 4u^2\big) \le \overline{\mathbb{P}}^N\big(\mathcal{Z}_N(u) \ge 4u\delta_N\big) \\
&\le \exp\left\{ -\frac{N\delta_N^2}{3} \right\} + 3\exp\left\{ -\frac{Nu\delta_N}{C \log N J_1(u)} \right\}.
\end{aligned}$$

We finish the proof. ■

Proof [Proof of Lemma 7] Note that

$$\begin{aligned}
&\mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{1}{N\lambda^2} \sum_{i=1}^N \int_0^T \left\| g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i) \right\|^2 \mathrm{d}t \right\} \\
&\overset{\text{(i)}}{\le} \mathbb{E}_{\overline{\mathbb{P}}^N} \int_0^T \exp\left\{ \frac{T}{N\lambda^2} \sum_{i=1}^N \left\| g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i) \right\|^2 \right\} \frac{\mathrm{d}t}{T} \\
&\overset{\text{(ii)}}{\le} \int_0^T \mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{T}{\lambda^2} \left\| g(\mu_t, X_t^N) - g(\mu_t^N, X_t^N) \right\|^2 \right\} \frac{\mathrm{d}t}{T} \\
&\le \sup_{0\le t\le T} \mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{T}{\lambda^2} \left\| g(\mu_t, X_t^N) - g(\mu_t^N, X_t^N) \right\|^2 \right\}.
\end{aligned}$$

Here (i) is by Jensen's inequality, and (ii) is by Hölder's inequality for multivariables, which implies $\mathbb{E}_{\overline{\mathbb{P}}^N} Y_1 \cdots Y_N \le \mathbb{E}_{\overline{\mathbb{P}}^N} Y_N^N$, since $Y_i = \exp\left\{ \frac{T}{N\lambda^2} \|g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i)\|^2 \right\}$ are identically

distributed (no need to be independent) under $\overline{\mathbb{P}}^N$ for $i = 1, 2, \cdots, N$. By Taylor's expansion,

$$\begin{aligned}
&\mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{\frac{T}{\lambda^2}\left\|g(\mu_t, X_t^N) - g(\mu_t^N, X_t^N)\right\|^2\right\} \\
&\leq 1 + \sum_{p\geq 1} \frac{T^p}{p!\lambda^{2p}} \mathbb{E}_{\overline{\mathbb{P}}^N}\left\|g(\mu_t, X_t^N) - g(\mu_t^N, X_t^N)\right\|^{2p} \\
&\overset{\text{(i)}}{\leq} 1 + \sum_{p\geq 1} \frac{T^p}{p!\lambda^{2p}} \cdot \frac{p! K_1^p}{(N-1)^p} \|\tilde{g}\|_{\text{Lip}}^{2p} \\
&= \frac{1}{1 - \frac{TK_1\|\tilde{g}\|_{\text{Lip}}^2}{(N-1)\lambda^2}}
\end{aligned}$$

for all $\lambda^2 > TK_1\|\tilde{g}\|_{\text{Lip}}^2/(N-1)$. Here (i) is by Lemma 14. Therefore, we get

$$\mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{\frac{1}{N\lambda^2}\sum_{i=1}^N \int_0^T \left\|g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i)\right\|^2 \mathrm{d}t\right\} \leq \frac{1}{1 - \frac{TK_1\|\tilde{g}\|_{\text{Lip}}^2}{(N-1)\lambda^2}}.$$

This implies

$$\left\|\left(\frac{1}{N}\sum_{i=1}^N \int_0^T \left\|g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i)\right\|^2 \mathrm{d}t\right)^{\frac{1}{2}}\right\|_{\psi_2} \leq 2\|\tilde{g}\|_{\text{Lip}}\sqrt{\frac{TK_1}{N}}. \tag{21}$$

Therefore,

$$\overline{\mathbb{P}}^N\left(\frac{1}{N}\sum_{i=1}^N \int_0^T \left\|g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i)\right\|^2 \mathrm{d}t > u^2\right) \leq 2e^{-\frac{Nu^2}{4TK_1\|\tilde{g}\|_{\text{Lip}}^2}}$$

■

Proof [Proof of Lemma 8] For simplicity, let

$$Z_g := \frac{1}{N}\sum_{i=1}^N \int_0^T \left\|g(\mu_t, X_t^i) - g(\mu_t^N, X_t^i)\right\|^2 \mathrm{d}t.$$

Notice that

$$\begin{aligned}
&\left|Z_f - Z_g\right| \\
&= \left|\frac{1}{N}\sum_{i=1}^N \int_0^T \left\langle (f-g)(\mu_t, X_t^i) - (f-g)(\mu_t^N, X_t^i), (f+g)(\mu_t, X_t^i) - (f+g)(\mu_t, X_t^i)\right\rangle \mathrm{d}t\right| \\
&\leq \sqrt{Z_{f-g}} \cdot \sqrt{Z_{f+g}}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
\|Z_f - Z_g\|_{\psi_1} &\overset{\text{(i)}}{\leq} \left\|\sqrt{Z_{f-g}}\right\|_{\psi_2} \cdot \left\|\sqrt{Z_{f+g}}\right\|_{\psi_2} \\
&\overset{\text{(ii)}}{\lesssim} \left\|\tilde{f} - \tilde{g}\right\|_{\text{Lip}}\sqrt{\frac{K_1}{N}} \cdot \left\|\tilde{f} + \tilde{g}\right\|_{\text{Lip}}\sqrt{\frac{K_1}{N}} \\
&\lesssim \frac{LK_1}{N}\left\|\tilde{f} - \tilde{g}\right\|_{\text{Lip}}.
\end{aligned}$$

Here, (i) is by Lemma 15 and (ii) is by (21) in the proof of Lemma 7. Then, by the standard chaining argument, we know

$$\left\| \sup_{g\in\mathcal{H}^*} Z_g \right\|_{\psi_1} \lesssim \frac{LK_1}{N} \int_0^{\frac{L}{2}} \log\left(1 + N(\varepsilon, \mathcal{H}^*, \|\cdot\|_{\text{Lip}})\right) \mathrm{d}\varepsilon = \frac{LK_1}{N} J_2(L).$$

Therefore, we have

$$\overline{\mathbb{P}}^N\Big(\sup_{g\in\mathcal{H}^*} Z_g > u \Big) \le 2e^{-\frac{Nu}{CLK_1 J_2(L)}}$$

for some constant $C > 0$. ■

Proof [Proof of Lemma 10] Recall that

$$\begin{aligned}
&\sup_{g\in\mathcal{H}^*} \frac{1}{N} \sum_{i=1}^N \int_0^T \langle g(\mu_t^N, X_t^i) - g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i \rangle \\
\le\ &\sup_{g\in\mathcal{H}^*} \frac{1}{N^2} \sum_{i=1}^N \int_0^T \langle \xi_{i,i}^g(t),\, \mathrm{d}\overline{W}_t^i \rangle + \sup_{g\in\mathcal{H}^*} \frac{1}{N^2} \sum_{1\le i\ne j\le N} \int_0^T \langle \xi_{i,j}^g(t)\, \mathrm{d}\overline{W}_t^i \rangle \\
=\ &\sup_{g\in\mathcal{H}^*} V_N(g) + \sup_{g\in\mathcal{H}^*} U_N(g).
\end{aligned}$$

We will start from $\sup_{g\in\mathcal{H}^*} V_N(g)$. By Lemma 16 and standard chaining argument,

$$\left\| \sup_{g\in\mathcal{H}^*} V_N(g) \right\|_{\psi_2} \lesssim \frac{1}{N\sqrt{N}} \int_0^{\frac{B}{2}} \sqrt{\log\left(1 + N(u, \mathcal{H}^*, \|\cdot\|_\infty)\right)}\, \mathrm{d}u = \frac{J_3(B)}{N\sqrt{N}}.$$

This implies that

$$\overline{\mathbb{P}}^N\left(\sup_{g\in\mathcal{H}^*} V_N(g) > u \right) \le 2e^{-\frac{N^3u^2}{CJ_3^2(B)}} \tag{22}$$

for some constant C. Next, we will bound $\sup_{g\in\mathcal{H}^*} U_N(g)$. By Lemma 17,

$$\|U_N(g)\|_{\psi_{2/3}} \lesssim \frac{B \log\log N}{N}.$$

Notice that we cannot directly use chaining method with respect to $\psi_{\frac{2}{3}}$-norm, since ψ_β is not convex when $0 < \beta < 1$. Luckily, by Lemma C.2 in Chen and Kato (2019), there is a convex function $\tilde{\psi}_{\frac{2}{3}}$, such that $\|\cdot\|_{\psi_{2/3}}$ and $\|\cdot\|_{\tilde{\psi}_{2/3}}$ are equivalent. In fact, a possible construction is

$$\tilde{\psi}_{\frac{2}{3}}(x) = \begin{cases} \frac{\psi_{2/3}(u)}{u} x & x \le u \\ \psi_{\frac{2}{3}}(x) & x > u, \end{cases}$$

where u is the unique solution to $3(1 - e^{-u}) = 2u$. Thus

$$\|U_N(g)\|_{\tilde{\psi}_{2/3}} \lesssim \frac{B \log\log N}{N}.$$

We can then verify that $\tilde{\psi}_{2/3}^{-1} \lesssim \psi_{2/3}^{-1} = [\log(1+x)]^{3/2}$. Again, by a chaining argument,

$$\begin{aligned}
\left\| \sup_{g\in\mathcal{H}^*} U_N(g) \right\|_{\psi_{2/3}} &\lesssim \left\| \sup_{g\in\mathcal{H}^*} U_N(g) \right\|_{\tilde{\psi}_{2/3}} \\
&\lesssim \frac{\log\log N}{N} \int_0^{\frac{B}{2}} \tilde{\psi}_{\frac{2}{3}}^{-1}\big(N(u,\mathcal{H}^*,\|\cdot\|_\infty)\big)\,\mathrm{d}u \\
&\lesssim \frac{\log\log N}{N} \int_0^{\frac{B}{2}} \big[\log(1+N(u,\mathcal{H}^*,\|\cdot\|_\infty))\big]^{\frac{3}{2}}\,\mathrm{d}u \\
&= \frac{\log\log N}{N} J_4(B).
\end{aligned}$$

It implies that

$$\overline{\mathbb{P}}^N\left(\sup_{g\in\mathcal{H}^*} V_N(g) > u \right) < 2\exp\left\{ -\Big(\frac{Nu}{CJ_4(B)\log\log N}\Big)^{\frac{2}{3}} \right\} \tag{23}$$

for some constant $C>0$. Taking $u = \frac{\log N}{N}$ in (22) and (23), we have

$$\sup_{g\in\mathcal{H}^*} V_N(g) + \sup_{g\in\mathcal{H}^*} U_N(g) \le \frac{2\log N}{N}$$

with probability at least

$$1 - 2\exp\left\{ -\frac{N(\log N)^2}{CJ_3^2(B)} \right\} - 2\exp\left\{ -\Big(\frac{\log N}{CJ_4(B)\log\log N}\Big)^{\frac{2}{3}} \right\}.$$

■

Proof [Proof of Lemma 11] Recall that

$$\|g\|_X^2 = \frac{1}{N}\sum_{i=1}^N \int_0^T \big\|g(\mu_t, X_t^i)\big\|^2\,\mathrm{d}t, \qquad \|g\|_E^2 = \mathbb{E}_{\overline{\mathbb{P}}^N}\|g\|_X^2.$$

First, we will show that

$$\overline{\mathbb{P}}^N\left(\exists\, g\in\mathcal{H}^* : \Big|\|g\|_X^2 - \|g\|_E^2\Big| > \frac{1}{2}\|g\|_E^2 + \frac{u^2}{2} \right) \le \overline{\mathbb{P}}^N\left(\sup_{g\in\mathcal{H}_u^*} \Big|\|g\|_X^2 - \|g\|_E^2\Big| > \frac{u^2}{2} \right). \tag{24}$$

In fact, let $g\in\mathcal{H}^*$ such that

$$\Big|\|g\|_X^2 - \|g\|_E^2\Big| > \frac{1}{2}\|g\|_E^2 + \frac{u^2}{2}.$$

If $\|g\|_E \le u$, then it is natural to have $\big|\|g\|_X^2 - \|g\|_E^2\big| > u^2/2$. Otherwise, if $\|g\|_E < u$, consider $h = \frac{u}{\|g\|_E}\cdot g$. We know $h\in\mathcal{H}^*$ since $\mathcal{H}^*$ is star-shaped and $\|h\|_E = u$, and

$$\Big|\|h\|_X^2 - \|h\|_E^2\Big| = \frac{u^2}{\|g\|_E^2}\Big|\|g\|_X^2 - \|g\|_E^2\Big| > \frac{u^2}{2}.$$

Thus (24) is true, and we only need to bound the right-hand side. Since

$$\left\| \frac{1}{N} \int_0^T \left\| g(\mu_t, X_t^i) \right\|^2 \mathrm{d}t \right\|_\infty \leq \frac{16B^2T}{N},$$

and

$$\begin{aligned} \sup_{g \in \mathcal{H}_u^*} \sum_{i=1}^N \mathbb{V}_{\overline{\mathbb{P}}^N} \left(\frac{1}{N} \int_0^T \left\| g(\mu_t, X_t^i) \right\|^2 \mathrm{d}t \right) &\leq \sup_{g \in \mathcal{H}_u^*} \sum_{i=1}^N \mathbb{E}_{\overline{\mathbb{P}}^N} \left(\frac{1}{N} \int_0^T \left\| g(\mu_t, X_t^i) \right\|^2 \mathrm{d}t \right)^2 \\ &= \sup_{g \in \mathcal{H}_u^*} N^{-1} \|g\|_E^2 \\ &\leq \frac{4B^2Tu^2}{N}. \end{aligned}$$

By Talagrand's inequality (Massart, 2000),

$$\begin{aligned} \overline{\mathbb{P}}^N \Bigg(\sup_{g \in \mathcal{H}_u^*} \left| \|g\|_X^2 - \|g\|_E^2 \right| \geq 2\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} \left| \|g\|_X^2 - \|g\|_E^2 \right| \\ + \frac{4Bu\sqrt{2Tx}}{\sqrt{N}} + \frac{560B^2Tx}{N} \Bigg) \leq e^{-x}, \end{aligned} \tag{25}$$

Now, let us bound the expectation term in the probability. By a standard symmetrization argument, it can be bounded by the Rademacher complexity of $(\mathcal{H}_u^*)^2$, i.e.

$$\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} \left| \|g\|_X^2 - \|g\|_E^2 \right| \leq 2\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \int_0^T \left\| g(\mu_t, X_t^i) \right\|^2 \mathrm{d}t \right|,$$

where ε_i are i.i.d. Rademacher random variables. Let

$$W_g := \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \int_0^T \left\| g(\mu_t, X_t^i) \right\|^2 \mathrm{d}t,$$

and

$$\widehat{R}_u^2 := \sup_{g \in \mathcal{H}_u^*} \frac{1}{N} \sum_{i=1}^N \int_0^T \left\| g(\mu_t, X_t^i) \right\|^2 \mathrm{d}t = \sup_{g \in \mathcal{H}_u^*} \|g\|_X^2.$$

First, we will show that

$$\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} |W_g| \leq 32\sqrt{T} B J_5 \left(\sqrt{\mathbb{E}_{\overline{\mathbb{P}}^N} \widehat{R}_u^2} \right). \tag{26}$$

Here, u indicates that we only consider the covering number of $\mathcal{H}_u^*$ rather than $\mathcal{H}^*$. Notice that for every f and $g \in \mathcal{H}_u^*$,

$$\begin{aligned}
\mathbb{E}^{\varepsilon}_{\overline{\mathbb{P}}^N} e^{\lambda(W_f - W_g)} &= \mathbb{E}^{\varepsilon}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{\lambda}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \bigg(\int_0^T \left\| f(\mu_t, X_t^i) \right\|^2 - \left\| g(\mu_t, X_t^i) \right\|^2 \mathrm{d}t \bigg) \right\} \\
&\leq \exp\left\{ \frac{\lambda^2}{2N} \sum_{i=1}^N \bigg(\int_0^T \left\| f(\mu_t, X_t^i) \right\|^2 - \left\| g(\mu_t, X_t^i) \right\|^2 \mathrm{d}t \bigg)^2 \right\} \\
&\leq \exp\left\{ \frac{\lambda^2}{2N} \sum_{i=1}^N \int_0^T \left\| (f-g)(\mu_t, X_t^i) \right\|^2 \mathrm{d}t \cdot \int_0^T \left\| (f+g)(\mu_t, X_t^i) \right\|^2 \mathrm{d}t \right\} \\
&\leq \exp\left\{ 8T \|f-g\|_X^2 B^2 \lambda^2 \right\}.
\end{aligned}$$

So $W_f - W_g$ is sub-Gaussian with parameter $16TB^2\|f-g\|_X^2$. By a standard chaining argument, we have

$$\begin{aligned}
\mathbb{E}^{\varepsilon}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} |W_g| &\leq 32\sqrt{T}B \int_0^{\widehat{R}_u/2} \sqrt{\log N(s, \mathcal{H}_u^*, \|\cdot\|_X)}\, \mathrm{d}s \\
&\leq 32\sqrt{T}B \int_0^{\widehat{R}_u/2} \sqrt{\log N(s/\sqrt{T}, \mathcal{H}_u^*, \|\cdot\|_\infty)}\, \mathrm{d}s \\
&= 32\sqrt{T}B J_5(\widehat{R}_u).
\end{aligned}$$

Thus, we have

$$\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} |W_g| \leq 32\sqrt{T}B \mathbb{E}_{\overline{\mathbb{P}}^N} J_5(\widehat{R}_u) \leq 32\sqrt{T}B J_5\Big(\sqrt{\mathbb{E}_{\overline{\mathbb{P}}^N} \widehat{R}_u^2} \Big).$$

The last inequality is because $z \mapsto J_5(\sqrt{z})$ is a concave function.

Next, We will estimate $J_5\Big(\sqrt{\mathbb{E}_{\overline{\mathbb{P}}^N} \widehat{R}_u^2} \Big)$. Recall that we have shown

$$\begin{aligned}
\mathbb{E}_{\overline{\mathbb{P}}^N} \widehat{R}_u^2 - u^2 &\leq \frac{2}{\sqrt{N}} \mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} |W_g| \\
&\leq \frac{64\sqrt{T}B}{\sqrt{N}} J_5\Big(\sqrt{\mathbb{E}_{\overline{\mathbb{P}}^N} \widehat{R}_u^2} \Big),
\end{aligned}$$

i.e.

$$\mathbb{E}_{\overline{\mathbb{P}}^N} \widehat{R}_u^2 \leq u^2 + \frac{64\sqrt{T}B}{\sqrt{N}} J_5\Big(\sqrt{\mathbb{E}_{\overline{\mathbb{P}}^N} \widehat{R}_u^2} \Big).$$

Applying Lemma 2.1 in Van Der Vaart and Wellner (2011) to $J_5(\cdot)$, we get

$$J_5\Big(\sqrt{\mathbb{E}_{\overline{\mathbb{P}}^N} \widehat{R}_u^2} \Big) \lesssim J_5(u) \bigg(1 + J_5(u) \frac{64\sqrt{T}B}{\sqrt{N}u^2} \bigg).$$

Taking it back to (26) leads to

$$\begin{aligned}\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g\in\mathcal{H}_u^*} |W_g| &\lesssim \sqrt{T}BJ_5(u)\left(1 + J_5(u)\frac{64\sqrt{T}B}{\sqrt{N}u^2}\right)\\ &\lesssim B\sqrt{T}J_5(u) + \frac{B^2TJ_5^2(u)}{\sqrt{N}u^2}\end{aligned}$$

Combining with (25), we get

$$\overline{\mathbb{P}}^N\left(C^{-1}\sup_{g\in\mathcal{H}_u^*}\left|\|g\|_X^2 - \|g\|_E^2\right| \geq \frac{\left[J_5(u) + u\sqrt{x}\right]B\sqrt{T}}{\sqrt{N}} + \frac{\left[xu^2 + J_5^2(u)\right]B^2T}{Nu^2}\right) \leq e^{-x}$$

for some constant $C > 0$. Take $x = Nu^2/36C^2B^2T$, and notice that

$$\frac{J_5(u)}{u} \leq \frac{J_r(r_N; r_N)}{r_N} \leq \frac{\sqrt{N}r_N}{6CB\sqrt{T}}.$$

Then, we get

$$\overline{\mathbb{P}}^N\left(\sup_{g\in\mathcal{H}_u^*}\left|\|g\|_X^2 - \|g\|_E^2\right| \geq \frac{ur_N + u^2}{6} + \frac{u^2 + r_N^2}{9}\right) \leq \exp\left\{-\frac{Nu^2}{36C^2B^2T^2}\right\}.$$

Again, since $r_N \leq u$ we have

$$\overline{\mathbb{P}}^N\left(\sup_{g\in\mathcal{H}_u^*}\left|\|g\|_X^2 - \|g\|_E^2\right| > \frac{u^2}{2}\right) \leq \exp\left\{-\frac{Nu^2}{36C^2B^2T}\right\}.$$

■

Remark 12 *Usually, there are several ways to bound the Rademacher complexity of* $(\mathcal{H}_u^*)^2$. *A direct way is using Ledoux–Talagrand contraction inequality, see e.g. the proof of Lemma 14.9 in Wainwright (2019). The method we use here is similar to the one in van de Geer (2014).*

Appendix B. Proof of Technical Lemmas

Lemma 13 *For any* $0 < u_1 \leq u_2$, *we have*

$$\frac{\mathbb{E}_{\overline{\mathbb{P}}^N}\mathcal{Z}_N(u_2)}{u_2} \leq \frac{\mathbb{E}_{\overline{\mathbb{P}}^N}\mathcal{Z}_N(u_1)}{u_1}.$$

Proof Recall that

$$\mathcal{Z}_N(u) = \sup_{g\in\mathcal{H}_u^*}\left|\frac{1}{N}\sum_{i=1}^N\int_0^T \langle g(\mu_t, X_t^i), \mathrm{d}\overline{W}_t^i\right|$$

and $\mathcal{H}_u^*$ is star-shaped for any u. Then for any $g \in \mathcal{H}_{u_2}^*$, let $\tilde{g} = \frac{u_1}{u_2} g \in \mathcal{H}_{u_1}^*$ since $0 < \frac{u_1}{u_2} \leq 1$. So

$$\begin{aligned}
\frac{u_1}{u_2}\mathbb{E}_{\overline{\mathbb{P}}^N}\mathcal{Z}_N(u_2) &= \mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g\in\mathcal{H}_{u_2}^*} \left| \frac{1}{N}\sum_{i=1}^N \int_0^T \left\langle \frac{u_1}{u_2} g(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i \right\rangle \right| \\
&\leq \mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{\tilde{g}\in\mathcal{H}_{u_1}^*} \left| \frac{1}{N}\sum_{i=1}^N \int_0^T \left\langle \tilde{g}(\mu_t, X_t^i),\, \mathrm{d}\overline{W}_t^i \right\rangle \right| \\
&= \mathbb{E}_{\overline{\mathbb{P}}^N}\mathcal{Z}_N(u_1).
\end{aligned}$$

■

Lemma 14 (Lemma 22 in Della Maestra and Hoffmann (2021)) *For any $g \in \mathcal{H}^*$ we have*

$$\mathbb{E}_{\overline{\mathbb{P}}^N}\left\| g(\mu_t, X_t^N) - g(\mu_t^N, X_t^N) \right\|^{2p} \leq \frac{p! K_1^p}{(N-1)^p} \|\tilde{g}\|_{Lip}^{2p}.$$

Here $K_1 \lesssim 1 + d^2$ is a constant.

Lemma 15 (ψ_α-norm of product) *For any $\alpha > 0$ and random variables X and Y, we have*

$$\|XY\|_{\psi_\alpha} \leq \|X\|_{\psi_{2\alpha}} \cdot \|Y\|_{\psi_{2\alpha}}.$$

Proof By Cauchy–Schwartz's inequality,

$$\begin{aligned}
\mathbb{E}\exp\left\{ \left(\frac{|XY|}{\lambda_x \lambda_y} \right)^\alpha \right\} &\leq \mathbb{E}\exp\left\{ \frac{1}{2}\left(\frac{|X|}{\lambda_x} \right)^{2\alpha} + \frac{1}{2}\left(\frac{|Y|}{\lambda_y} \right)^{2\alpha} \right\} \\
&= \sqrt{\mathbb{E}\exp\left\{ \left(\frac{|X|}{\lambda_x} \right)^{2\alpha} \right\}} \cdot \sqrt{\mathbb{E}\exp\left\{ \left(\frac{|Y|}{\lambda_y} \right)^{2\alpha} \right\}} \leq 2,
\end{aligned}$$

if we take $\lambda_x = \|X\|_{\psi_{2\alpha}}$ and $\lambda_y = \|Y\|_{\psi_{2\alpha}}$. Therefore, by definition

$$\|XY\|_{\psi_\alpha} \leq \lambda_x \lambda_y = \|X\|_{\psi_{2\alpha}} \cdot \|Y\|_{\psi_{2\alpha}}.$$

■

Lemma 16 *For every $g \in \mathcal{H}^*$, let*

$$V_N(g) := \frac{1}{N^2}\sum_{i=1}^N \int_0^T \langle \xi_{i,i}^g(t),\, \mathrm{d}\overline{W}_t^i \rangle.$$

Then $V_n(g)$ is sub-Gaussian with parameter CB^2N^{-3} for some positive constant C (may depends on T), i.e.,

$$\overline{\mathbb{P}}^N\Big(V_N(g) > u \Big) \leq e^{-\frac{N^3u^2}{2CB^2}}, \qquad \forall\ u > 0.$$

This implies that

$$\|V_N(g)\|_{\psi_2} \leq \frac{2B\sqrt{C}}{N\sqrt{N}}.$$

Proof For any integer $p \geq 1$, by Burkholder–Davis–Gundy's inequality,

$$\mathbb{E}_{\overline{\mathbb{P}}^N}\big|V_N(g)\big|^{2p} \leq \frac{C^p p^p}{N^{4p}} \mathbb{E}_{\overline{\mathbb{P}}^N}\bigg(\sum_{i=1}^N \int_0^T \Big\|\xi_{i,i}^g(t)\Big\|^2 \,\mathrm{d}t\bigg)^p \leq \frac{C^p B^{2p} p^p}{N^{3p}} \leq p! \cdot \frac{C^p B^{2p}}{N^{3p}}.$$

The last inequality is by $(p/e)^p \leq p!$. Also, we know $V_N(g)$ is mean-zero. The high probability bound just follows from Appendix in Della Maestra and Hoffmann (2021). The bound of ψ_2-norm can be derived by the argument in Lemma 17. ■

Lemma 17 (Concentration of degenerate U-statistics) *For every $g \in \mathcal{H}^*$, let $g^\dagger$ be the degenerate kernel defined in* **Step 5** *in the proof of Theorem 6 and define the U-statistics*

$$U_N(g) = \frac{1}{N^2} \sum_{1 \leq i \neq j \leq N} g^\dagger(\overline{W}^i, \overline{W}^j)$$

Then, there is a constant $C > 0$ (may depends on T) such that

$$\overline{\mathbb{P}}^N\Big(\big|U_N(g)\big| > u\Big) \leq \exp\Big\{-\Big(\frac{uN}{CB\log\log N}\Big)^{\frac{2}{3}}\Big\}, \qquad \forall\, u \geq \frac{CB(\log N)^2}{N}.$$

This implies that

$$\|U_N(g)\|_{\psi_{2/3}} \leq \frac{3CB\log\log N}{N}.$$

Proof Let $\{\overline{W}^{il} : 1 \leq i \leq N, l = 1, 2\}$ be independent copies of standard Brownian motions under $\overline{\mathbb{P}}^N$. By Theorem 3.1.1 in de la Pena and Giné (1999),

$$\big\|U_N(g)\big\|_{L_p(\overline{\mathbb{P}}^N)} \lesssim \big\|U_N^D(g)\big\|_{L_p(\overline{\mathbb{P}}^N)} := \bigg\|\frac{1}{N^2} \sum_{1 \leq i \neq j \leq N} g^\dagger(\overline{W}^{i1}, \overline{W}^{j2})\bigg\|_{L_p(\overline{\mathbb{P}}^N)}.$$

For simplicity, let

$$g_{ij}^\dagger = N^{-2} g^\dagger(\overline{W}^{i1}, \overline{W}^{j2}) = \frac{1}{N^2} \int_0^T \langle \xi_{i,j}^{g,D}(t),\, \mathrm{d}\overline{W}_t^{i1} \rangle,$$

where $\xi_{i,j}^{g,D}$ is the decoupling version of $\xi_{i,j}^g$ defined as

$$\xi_{i,j}^{g,D}(t) = \tilde{g}(X_t^{i,1}, X_t^{j,2}) - \int_{\mathbb{T}^d} \tilde{g}(X_t^{i,1}, y)\,\mathrm{d}\mu_t(y).$$

Here $X^{i,l}$ is the solution of (13) corresponding to the Brownian motion $\overline{W}^{i,l}$. For $p \geq 2$, follow the proof of Theorem 3.2 in Giné et al. (2000), we can get

$$\mathbb{E}_{\overline{\mathbb{P}}^N}\bigg|\sum_{1 \leq i \neq j \leq N} g_{ij}^\dagger\bigg|^p \leq C^p \mathbb{E}_{\overline{\mathbb{P}}^N}^2 \bigg(p^{\frac{p}{2}}\bigg[\sum_{i=1}^N \mathbb{E}_{\overline{\mathbb{P}}^N}^1\Big(\sum_{j=1, j\neq i}^N g_{ij}^\dagger\Big)^2\bigg]^{\frac{p}{2}} + p^p \mathbb{E}_{\overline{\mathbb{P}}^N}^1 \sum_{i=1}^N \Big|\sum_{j=1, j\neq i}^N g_{ij}^\dagger\Big|^p\bigg), \tag{27}$$

by conditioning on $\{\overline{W}^{j,2} : 1 \le j \le N\}$ first. Applying Proposition 3.1 in Giné et al. (2000), the first term can be bounded by

$$\begin{aligned}
p^{\frac{p}{2}}\mathbb{E}^2_{\overline{\mathbb{P}}^N}\Big[\sum_{i=1}^N \mathbb{E}^1_{\overline{\mathbb{P}}^N}\Big(\sum_{j=1,j\neq i}^N g_{ij}^\dagger\Big)^2\Big]^{\frac{p}{2}} &\le p^{\frac{p}{2}}\Big(\sum_{1\le i\neq j\le N}\mathbb{E}_{\overline{\mathbb{P}}^N}\big(g_{ij}^\dagger\big)^2\Big)^{\frac{p}{2}} \\
&\quad + p^p\mathbb{E}\Big(\sup\Big\{\mathbb{E}^2_{\overline{\mathbb{P}}^N}\sum_{j=1}^N\Big(\mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i=1,i\neq j}^N g_{ij}^\dagger h_i(\overline{W}^{i1})\Big)^2 : \mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i=1}^N h_i^2(\overline{W}^{i1})\le 1\Big\}\Big)^{\frac{p}{2}} \\
&\quad + p^{\frac{3p}{2}}\mathbb{E}^2_{\overline{\mathbb{P}}^N}\max_{1\le j\le N}\Big(\mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i=1}^N\big(g_{ij}^\dagger\big)^2\Big)^{\frac{p}{2}}.
\end{aligned}$$

Now, let us bound these three terms. Notice that

$$\Big(\sum_{1\le i\neq j\le N}\mathbb{E}_{\overline{\mathbb{P}}^N}\big(g_{ij}^\dagger\big)^2\Big)^{\frac{p}{2}} = \frac{1}{N^{2p}}\Big(\sum_{1\le i\neq j\le N}\mathbb{E}_{\overline{\mathbb{P}}^N}\int_0^T\Big\|\xi_{i,j}^{g,D}(t)\Big\|^2\mathrm{d}t\Big)^{\frac{p}{2}} \le \Big(\frac{CB}{N}\Big)^p.$$

Next, we have

$$\begin{aligned}
&\sup\Big\{\mathbb{E}^2_{\overline{\mathbb{P}}^N}\sum_{j=1}^N\Big(\mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i=1,i\neq j}^N g_{ij}^\dagger h_i(\overline{W}^{i1})\Big)^2 : \mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i=1}^N h_i^2(\overline{W}^{i1})\le 1\Big\} \\
&\le \sup\Big\{\mathbb{E}^2_{\overline{\mathbb{P}}^N}\sum_{j=1}^N\Big(\mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i\neq j}\big(g_{ij}^\dagger\big)^2\Big)\Big(\mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i\neq j}h_i^2(\overline{W}^{i1})\Big) : \mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i=1}^N h_i^2(\overline{W}^{i1})\le 1\Big\} \\
&\le \mathbb{E}^2_{\overline{\mathbb{P}}^N}\sum_{j=1}^N\mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i\neq j}\big(g_{ij}^\dagger\big)^2 \\
&\lesssim N^{-2}B^2.
\end{aligned}$$

Lastly,

$$\begin{aligned}
\mathbb{E}^2_{\overline{\mathbb{P}}^N}\max_{1\le j\le N}\Big(\mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i=1}^N\big(g_{ij}^\dagger\big)^2\Big)^{\frac{p}{2}} &\le \sum_{j=1}^N\mathbb{E}^2_{\overline{\mathbb{P}}^N}\Big(\mathbb{E}^1_{\overline{\mathbb{P}}^N}\sum_{i=1}^N\big(g_{ij}^\dagger\big)^2\Big)^{\frac{p}{2}} \\
&= \frac{1}{N^{2p}}\sum_{j=1}^N\mathbb{E}^2_{\overline{\mathbb{P}}^N}\Big(\sum_{i=1}^N\mathbb{E}^1_{\overline{\mathbb{P}}^N}\int_0^T\Big\|\xi_{i,j}^{g,D}(t)\Big\|^2\mathrm{d}t\Big)^{\frac{p}{2}} \\
&\le \frac{C^pB^p}{N^{\frac{3p}{2}-1}}.
\end{aligned}$$

Rather than bound the second term of (27) as Giné et al. (2000) did, we directly bound it by Burkholder–Davis–Gundy's inequality, see e.g., Barlow and Yor (1982),

$$\begin{aligned} p^p \mathbb{E}_{\overline{\mathbb{P}}^N} \sum_{i=1}^N \Big| \sum_{j=1, j\neq i}^N g_{ij}^\dagger \Big|^p &= \frac{p^p}{N^{2p}} \sum_{i=1}^N \mathbb{E}_{\overline{\mathbb{P}}^N} \bigg| \int_0^T \Big\langle \sum_{j=1, j\neq i}^N \xi_{i,j}^{g,D}, \, \mathrm{d}\overline{W}_t^{i1} \Big\rangle \bigg|^p \\ &\leq \frac{p^p}{N^{2p}} \sum_{i=1}^N C^p p^{\frac{p}{2}} \mathbb{E}_{\overline{\mathbb{P}}^N} \bigg(\int_0^T \Big\| \sum_{j=1, j\neq i}^N \xi_{i,j}^{g,D} \Big\|^2 \mathrm{d}t \bigg)^{\frac{p}{2}} \\ &\leq \bigg(\frac{CBp^{\frac{3}{2}}}{N^{1-\frac{1}{p}}} \bigg)^p. \end{aligned}$$

Combining all the pieces together, we have

$$\big\| U_N(g) \big\|_{L_p(\overline{\mathbb{P}}^N)}^p \leq C^p B^p \bigg[\frac{p^{\frac{p}{2}}}{N^p} + \frac{p^p}{N^p} + \frac{p^{\frac{3p}{2}}}{N^{\frac{3p}{2}-1}} + \frac{p^{\frac{3p}{2}}}{N^{p-1}} \bigg] \leq C^p B^p p^{3p/2} N^{1-p}.$$

By Markov's inequality, we have

$$\overline{\mathbb{P}}^N \Big(\big| U_N(g) \big| > u \Big) \leq \frac{C^p B^p p^{3p/2}}{u^p N^{p-1}} \leq e^{-p},$$

by tuning the parameter $p > 2$ such that $CBp^{3/2} < uN^{1-1/p}$. By choosing

$$p = \bigg(\frac{uN}{eCB \log\log N} \bigg)^{\frac{2}{3}}, \qquad \text{when } u \geq \frac{CBe(\log N)^2}{N},$$

we get

$$\overline{\mathbb{P}}^N \Big(\big| U_N(g) \big| > u \Big) \leq \exp\Big\{ - \Big(\frac{uN}{CB \log\log N} \Big)^{\frac{2}{3}} \Big\}, \qquad \forall\, u \geq \frac{CB(\log N)^2}{N}$$

for some constant C.

Lastly, we will prove the statement about the Orlicz norm of $U_N(g)$. Take

$$\lambda = \frac{3CB \log\log N}{N},$$

and we have

$$\begin{aligned} \mathbb{E}_{\overline{\mathbb{P}}^N} \exp\bigg\{ \Big(\frac{|U_N(g)|}{\lambda} \Big)^{\frac{2}{3}} \bigg\} &= \int_0^\infty \overline{\mathbb{P}}^N \Big(\big| U_N(g) \big| > \lambda (\log u)^{\frac{3}{2}} \Big) \, \mathrm{d}u \\ &\leq 1 + \int_1^\infty \exp\Big\{ - \Big(\frac{\lambda (\log u)^{3/2} N}{CB \log\log N} \Big)^{\frac{2}{3}} \Big\} \, \mathrm{d}u \\ &= 1 + \int_1^\infty \exp\Big\{ - \Big(\frac{\lambda N}{CB \log\log N} \Big)^{\frac{2}{3}} \log u \Big\} \, \mathrm{d}u \\ &= 1 - \bigg[1 - \Big(\frac{\lambda N}{CB \log\log N} \Big)^{\frac{2}{3}} \bigg]^{-1} \\ &< 2. \end{aligned}$$

We finish the proof. ■

Lemma 18 (sub-exponential increments) *Let*

$$Y_g = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \int_0^T \big\langle g(\mu_t, X_t^i),\, \mathrm{d}W_t^i \big\rangle.$$

Then for any f, and $g \in \mathcal{H}^*$, $Y_f - Y_g$ is sub-exponential with parameters satisfying

$$\mathbb{E}_{\overline{\mathbb{P}}^N} e^{\lambda(Y_f - Y_g)} \le \exp\left\{ \frac{\lambda^2 \|f-g\|_E^2/2}{1 - \sqrt{T/2N}\lambda \|f-g\|_\infty} \right\}.$$

Moreover, the Bernstein's bound holds as well, i.e.,

$$\overline{\mathbb{P}}^N\big(|Y_f - Y_g| > u\big) < 2\exp\Big\{ - \frac{u^2}{2\big(\|f-g\|_E^2 + \sqrt{T/2N}\|f-g\|_\infty u\big)} \Big\}.$$

Proof Note that

$$\begin{aligned}
&\mathbb{E}_{\overline{\mathbb{P}}^N} e^{\lambda(Y_f - Y_g)} \\
&= \left(\mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{\lambda}{\sqrt{N}} \int_0^T \big\langle (f-g)(\mu_t, X_t),\, \mathrm{d}\overline{W}_t \big\rangle \right\} \right)^N \\
&\overset{\text{(i)}}{\le} \left(\mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{2\lambda}{\sqrt{N}} \int_0^T \big\langle (f-g)(\mu_t, X_t),\, \mathrm{d}\overline{W}_t \big\rangle - \frac{2\lambda^2}{N} \int_0^T \|(f-g)(\mu_t, X_t)\|^2 \,\mathrm{d}t \right\} \right)^{\frac{N}{2}} \\
&\qquad \cdot \left(\mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{2\lambda^2}{N} \int_0^T \|(f-g)(\mu_t, X_t)\|^2 \,\mathrm{d}t \right\} \right)^{\frac{N}{2}} \\
&\overset{\text{(ii)}}{=} \left(\mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{2\lambda^2}{N} \int_0^T \|(f-g)(\mu_t, X_t)\|^2 \,\mathrm{d}t \right\} \right)^{\frac{N}{2}}.
\end{aligned}$$

Here, (i) is by Cauchy–Schwartz's inequality, and (ii) is by the fact that

$$\exp\left\{ \frac{2\lambda}{\sqrt{N}} \int_0^s \big\langle (f-g)(\mu_t, X_t),\, \mathrm{d}\overline{W}_t \big\rangle - \frac{1}{2}\Big(\frac{2\lambda}{\sqrt{N}}\Big)^2 \int_0^s \|(f-g)(\mu_t, X_t)\|^2 \,\mathrm{d}t \right\}, \quad 0 \le s \le T$$

is an exponential martingale under $\overline{\mathbb{P}}^N$. Now, by Taylor's formula

$$\begin{aligned}
&\mathbb{E}_{\overline{\mathbb{P}}^N} \exp\left\{ \frac{2\lambda^2}{N} \int_0^T \|(f-g)(\mu_t, X_t)\|^2 \,\mathrm{d}t \right\} \\
&= 1 + \sum_{k=1}^{\infty} \frac{1}{k!} \Big(\frac{2\lambda^2}{N}\Big)^k \mathbb{E}_{\overline{\mathbb{P}}^N} \left(\int_0^T \|(f-g)(\mu_t, X_t)\|^2 \,\mathrm{d}t \right)^k \\
&\le 1 + \sum_{k=1}^{\infty} \Big(\frac{2\lambda^2}{N}\Big)^k \big(T\|f-g\|_\infty^2\big)^{k-1} \|f-g\|_E^2 \\
&= 1 + \frac{2\lambda^2 \|f-g\|_E^2/N}{1 - 2T\lambda^2\|f-g\|_\infty^2} \\
&\le \exp\left\{ \frac{2\lambda^2 \|f-g\|_E^2/N}{1 - 2T\lambda^2\|f-g\|_\infty^2/N} \right\} \\
&\le \exp\left\{ \frac{2\lambda^2 \|f-g\|_E^2/N}{1 - \sqrt{2T/N}\lambda\|f-g\|_\infty} \right\}
\end{aligned}$$

for N large enough. So

$$\mathbb{E}_{\overline{\mathbb{P}}^N} e^{\lambda(Y_f - Y_g)} \le \exp\left\{ \frac{\lambda^2 \|f-g\|_E^2/2}{1 - \sqrt{T/2N}\lambda \|f-g\|_\infty} \right\}.$$

Then the Bernstein's bound holds by Proposition 2.10 in Wainwright (2019). ■

Lemma 19 (estimation of ψ_1-norm)

$$\left\| \sup_{g \in \mathcal{H}_u^*} |Y_g| \right\|_{\psi_1} \lesssim \sqrt{\frac{TB^2}{2N}} \int_0^{\frac{1}{2}} \log\left(1 + H(\varepsilon, \mathcal{H}_u^*)\right) \mathrm{d}\varepsilon + u \int_0^{\frac{1}{2}} \sqrt{\log\left(1 + H(\varepsilon, \mathcal{H}_u^*)\right)}\, \mathrm{d}\varepsilon.$$

Proof We only need to consider the case where $\mathcal{H}_u^*$ is a finite set, and the whole proof can be extended to a separable $\mathcal{H}_u^*$ ($\mathcal{H}^*$ is separable indeed) by a standard argument. Let $S_k \subset \mathcal{H}_u^*$ such that $|S_k| = H(2^{-k}, \mathcal{H}_u^*)$ for all integer $k \ge 0$. We specify $S_0 = \{0\}$ and $S_K = \mathcal{H}_u^*$. Such K exists since $\mathcal{H}_u^*$ is finite. Let $\pi_k(g)$ be an element in S_k, such that

$$\|g - \pi_k(g)\|_E \le 2^{-k} u, \qquad \|g - \pi_k(g)\|_\infty \le 2^{-k} B.$$

For any $g \in \mathcal{H}_u^*$ and $1 \le k \le K$, let $g^K = g$ and $g^{k-1} = \pi_{k-1}(g^k)$. Then, we can decompose $\sup_{g \in \mathcal{H}_u^*} Y_g$ by

$$\sup_{g \in \mathcal{H}_u^*} |Y_g| = \sup_{g \in \mathcal{H}_u^*} |Y_g - Y_0| \le \sum_{k=1}^K \sup_{g \in \mathcal{H}_u^*} \left|Y_{g^k} - Y_{g^{k-1}}\right| \le \sum_{k=1}^K \sup_{g \in S_k} \left|Y_g - Y_{\pi_{k-1}(g)}\right|.$$

By Lemma 8.3 in Kosorok (2008), Lemma 18, and triangular inequality of ψ_1-norm, we can get

$$\begin{aligned}
\left\| \sup_{g \in \mathcal{H}_u^*} |Y_g| \right\|_{\psi_1} &\le \sum_{k=1}^K \left\| \sup_{g \in S_k} \left|Y_g - Y_{\pi_{k-1}(g)}\right| \right\|_{\psi_1} \\
&\lesssim \sum_{k=1}^K \left[\sqrt{\frac{T}{2N}} 2^{-k} B \cdot \log\left(1 + H(2^{-k}, \mathcal{H}_u^*)\right) + 2^{-k} u \cdot \sqrt{\log\left(1 + H(2^{-k}, \mathcal{H}_u^*)\right)} \right] \\
&\lesssim \sqrt{\frac{TB^2}{2N}} \int_0^{\frac{1}{2}} \log\left(1 + H(\varepsilon, \mathcal{H}_u^*)\right) \mathrm{d}\varepsilon + u \int_0^{\frac{1}{2}} \sqrt{\log\left(1 + H(\varepsilon, \mathcal{H}_u^*)\right)}\, \mathrm{d}\varepsilon.
\end{aligned}$$

■

Appendix C. Proof of Applications

Proof [proof of Corollary 2] First, let us calculate the order of r_N. Notice that if $\tilde{f}, \tilde{g} \in \tilde{\mathcal{H}}^*$, s.t. $\left\|\tilde{f} - \tilde{g}\right\|_\infty < \varepsilon$, we have $\|f - g\|_\infty < \varepsilon$. This implies $\log N(\varepsilon, \mathcal{H}^*, \|\cdot\|_\infty) \le 2 \log N(\varepsilon/2, \tilde{\mathcal{H}}^*, \|\cdot\|_\infty)$. Therefore

$$J_5(r_N) \lesssim \int_0^{\frac{r_N}{2}} \sqrt{\varepsilon^{-\frac{d}{\alpha}}}\, \mathrm{d}\varepsilon \lesssim r_N^{1 - \frac{d}{2\alpha}}.$$

By letting $r_N^{1-\frac{d}{2\alpha}} \lesssim \sqrt{N} r_N^2$, we get $r_N \lesssim N^{-\frac{\alpha}{d+2\alpha}}$. Next, we shall calculate the order of δ_N. In fact,

$$\begin{aligned}
&\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} \left| \frac{1}{N} \sum_{i=1}^N \int_0^T \left\langle g(\mu_t, X_t^i), \, \mathrm{d}\overline{W}_t^i \right\rangle \right| \\
\lesssim & \left\| \sup_{g \in \mathcal{H}_u^*} \left| \frac{1}{N} \sum_{i=1}^N \int_0^T \left\langle g(\mu_t, X_t^i), \, \mathrm{d}\overline{W}_t^i \right\rangle \right| \right\|_{\psi_1} \\
\lesssim & \sqrt{\frac{TB^2}{2N^2}} \int_0^{\frac{1}{2}} \log\left(1 + H(\varepsilon, \mathcal{H}_u^*)\right) \mathrm{d}\varepsilon + \frac{u}{\sqrt{N}} \int_0^{\frac{1}{2}} \sqrt{\log\left(1 + H(\varepsilon, \mathcal{H}_u^*)\right)} \, \mathrm{d}\varepsilon.
\end{aligned}$$

The last inequality is by Lemma 19. Note that

$$\|f - g\|_E^2 = \int_0^T \int_{\mathbb{T}^d} \|(f-g)(\mu_t, x)\|^2 \, \mathrm{d}\mu_t(x) \mathrm{d}t \le T \left\| \tilde{f} - \tilde{g} \right\|_\infty^2,$$

and

$$\|f - g\|_\infty \le \left\| \tilde{f} - \tilde{g} \right\|_\infty.$$

Therefore $\log H(\varepsilon, \mathcal{H}_u^*) \le 2 \log N\big((2\sqrt{T}+2)^{-1}\varepsilon u, \tilde{\mathcal{H}}^*, \|\cdot\|_\infty\big) \lesssim (\varepsilon u)^{-d/\alpha}$. So we know

$$\begin{aligned}
\mathbb{E}_{\overline{\mathbb{P}}^N} \sup_{g \in \mathcal{H}_u^*} \left| \frac{1}{N} \sum_{i=1}^N \int_0^T \left\langle g(\mu_t, X_t^i), \, \mathrm{d}\overline{W}_t^i \right\rangle \right| &\lesssim \frac{1}{N} \int_0^{\frac{1}{2}} (\varepsilon u)^{-\frac{d}{\alpha}} \, \mathrm{d}\varepsilon + \frac{u}{\sqrt{N}} \int_0^{\frac{1}{2}} (\varepsilon u)^{-\frac{d}{2\alpha}} \, \mathrm{d}\varepsilon \\
&\lesssim \frac{u^{-\frac{d}{\alpha}}}{N} + \frac{u^{1-\frac{d}{2\alpha}}}{\sqrt{N}}.
\end{aligned}$$

By Letting

$$\frac{\delta_N^{-\frac{d}{\alpha}}}{N} + \frac{\delta_N^{1-\frac{d}{2\alpha}}}{\sqrt{N}} \lesssim \delta_N^2$$

we get $\delta_N \lesssim N^{-\frac{\alpha}{d+2\alpha}}$. By Theorem 1, we know $\left\| \widehat{b}_N - b^* \right\|_E \lesssim N^{-\frac{\alpha}{d+2\alpha}}$ with probability $1 - C\exp\left\{ - C'(\frac{\log N}{\log\log N})^{2/3} \right\}$ for some positive constant C and C' that may depends on T. ■

Lemma 20 *Under Assumption 2 for all $C > 0$,*

$$\left\| \widehat{F}_N - F^* \right\|_2 \le \frac{\left\| \mathcal{L}(\widehat{b}_N - b^*)(\mu, \cdot) \right\|_2}{\inf_{0 < \|k\| < C} \|(\mathcal{L}\mu)_k\|} + \frac{\|\widehat{F}_N - F^*\|_{H^1}}{2\pi C}.$$

Proof [Proof of Lemma 20] By Plancherel's identity (Proposition 3.1.16 in Grafakos (2008))

$$\begin{aligned}
\left\| \widehat{F}_N - F^* \right\|_2^2 &= \sum_{k \in \mathbb{Z}^d} \left\| (\widehat{F}_N - F^*)_k \right\|^2 \\
&= \sum_{0 < \|k\| < C} \left\| (\widehat{F}_N - F^*)_k \right\|^2 + \sum_{\|k\| \ge C} \left\| (\widehat{F}_N - F^*)_k \right\|^2.
\end{aligned}$$

The last equality is by the restriction $\int_{\mathbb{T}^d} \widehat{F}_N(x)\,\mathrm{d}x = \int_{\mathbb{T}^d} F^*(x)\,\mathrm{d}x = 0$, which implies that $(\widehat{F}_N - F^*)_0 = 0$. By (11) and Assumption 2 that $(\mathcal{L}\mu)_k \neq 0$ for all $k \neq 0$,

$$\begin{aligned}\sum_{0<\|k\|<C} \left\|(\widehat{F}_N - F^*)_k\right\|^2 &= \sum_{0<\|k\|<C} \left\| \frac{\big(\mathcal{L}(\widehat{b}_N - b^*)(\cdot,\mu)\big)_k}{(\mathcal{L}\mu)_k} \right\|^2 \\ &\leq \inf_{0<\|k\|<C} |(\mathcal{L}\mu)_k|^{-2} \cdot \sum_{k\in\mathbb{Z}^d} \left\| \big(\mathcal{L}(\widehat{b}_N - b^*)(\cdot,\mu)\big)_k \right\|^2 \\ &= \inf_{0<\|k\|<C} |(\mathcal{L}\mu)_k|^{-2} \cdot \left\| \mathcal{L}(\widehat{b}_N - b^*)(\cdot,\mu) \right\|_2^2.\end{aligned}$$

For the second term, recall that for any function $f : \mathbb{T}^d \to \mathbb{R}^d$ and $k \in \mathbb{Z}^d$

$$\begin{aligned}\sum_{l=1}^d \|(\nabla f_l)_k\|^2 = \sum_{l,j=1}^d \left| \Big(\frac{\partial f_l}{\partial x_j}\Big)_k \right|^2 &= \sum_{l,j=1}^d \left| 2\pi i k_j (f_l)_k \right|^2 \\ &= 4\pi^2 \sum_{l=1}^d \left|(f_l)_k\right|^2 \sum_{j=1}^d k_j^2 \\ &= 4\pi^2 \|k\|^2 \|(f)_k\|^2.\end{aligned}$$

Take $f = \widehat{F}_N - F^*$, and we get

$$\begin{aligned}\sum_{\|k\|\geq C} \left\|(\widehat{F}_N - F^*)_k\right\|^2 &\leq \frac{1}{(2\pi C)^2} \cdot \sum_{\|k\|>C} (2\pi)^2 |k|^2 \left\|(\widehat{F}_N - F^*)_k\right\|^2 \\ &= \frac{1}{(2\pi C)^2} \sum_{\|k\|>C} \sum_{l=1}^d \left\| \big(D^j(\widehat{F}_N - F^*)\big)_k \right\|^2 \\ &\leq \frac{1}{(2\pi C)^2} \sum_{l=1}^d \sum_{k\in\mathbb{Z}^d} \left\| \big(\nabla(\widehat{F}_N - F^*)_l\big)_k \right\|^2 \\ &= \frac{1}{(2\pi C)^2} \sum_{l=1}^d \left\| \nabla(\widehat{F}_N - F^*)_l \right\|_2^2 \\ &\leq (2\pi C)^{-2} \left\| \widehat{F}_N - F^* \right\|_{H^1}^2.\end{aligned}$$

Combining these pieces together yields the result. ■

Proof [Proof of Corollary 3] We can assume $\mu_0(x) > 0$ for all x. Otherwise, we can choose $w(t)$ and $\rho(t)$ satisfying $\int_s^T w(t)\,\mathrm{d}\rho(t) = 0$ for some time $s > 0$, and apply the operator $\mathcal{L}$ with new w

and ρ.

$$\begin{aligned}\left\|\mathcal{L}(\widehat{b}_N - b^*)(\mu,\cdot)\right\|_2^2 &= \int_{\mathbb{T}^d}\left\|\int_0^T w(t)(\widehat{b}_N - b^*)(\mu_t,x)\,\mathrm{d}t\right\|^2\mathrm{d}x\\ &\le \int_{\mathbb{T}^d}\int_0^T w^2(t)\left\|(\widehat{b}_N - b^*)(\mu_t,x)\right\|^2\mathrm{d}t\,\mathrm{d}x\\ &\lesssim \int_0^T\int_{\mathbb{T}^d}\left\|(\widehat{b}_N - b^*)(\mu_t,x)\right\|^2\mathrm{d}x\,\mathrm{d}t\\ &\lesssim \int_0^T\int_{\mathbb{T}^d}\left\|(\widehat{b}_N - b^*)(\mu_t,x)\right\|^2\mathrm{d}\mu_t(x)\,\mathrm{d}t.\end{aligned}$$

The last inequality is by the fact that $\mu_t(x)$ is bounded away from zero, since $[0,T]\times\mathbb{T}^d$ is compact. Then by Lemma 20, with high probability we have

$$\left\|\widehat{F}_N - F^*\right\|_2 \le \frac{\delta_N + r_N + \log N/N}{\inf_{0<\|k\|<C}\|(\mathcal{L}\mu)_k\|} + \frac{1}{\pi C}\sup_{F\in\tilde{\mathcal{H}}}\|F\|_{H^1}.$$

By taking $C = \eta_N$, we finish the proof. ■