Theoretical Analysis for Learning from Unlabeled Data for Interacting Particle Systems

Theory Supplement for LED_ips_nn

January 2025

Abstract

This document presents a comprehensive theoretical analysis for learning interaction and kinetic potentials from unlabeled ensemble data of interacting particle systems. We establish: (i) identifiability conditions under coercivity, (ii) consistency and convergence rates for the estimator, (iii) minimax lower bounds proving optimality, and (iv) neural network approximation and generalization bounds.

Contents

1 Theoretical Analysis

We develop a systematic theory for learning the potential functions (Φ, V) from unlabeled ensemble data. Our analysis addresses three fundamental questions: (i) *identifiability*—under what conditions can we uniquely recover (Φ, V)? (ii) *well-posedness*—is the inverse problem stable? (iii) *convergence rates*—how fast does the estimator converge as data increases?

1.1 Notation and Setup

Let $\mathcal{H}_\Phi$ and $\mathcal{H}_V$ be the hypothesis spaces for the interaction and kinetic potentials, respectively. We assume:

(A1) $\Phi \in \mathcal{H}_\Phi \subset C^2(\mathbb{R}^d)$ with $\Phi(x) = \Phi(-x)$ (symmetry).

(A2) $V \in \mathcal{H}_V \subset C^2(\mathbb{R}^d)$ with V confining: $\lim_{|x|\to\infty} V(x) = +\infty$.

(A3) The process $\{X_t^{1:N}\}$ has a unique invariant measure π on $\mathbb{R}^{Nd}$.

Define the population loss:

$$\mathcal{E}_\infty(\Phi, V) := \lim_{M,L\to\infty} \mathcal{E}_\mathcal{D}(\Phi, V) = \mathbb{E}\left[\mathcal{E}_{\mathbf{X}_t,\mathbf{X}_{t+\Delta t}}(\Phi, V)\right], \tag{1}$$

where the expectation is over the stationary distribution.

1.2 Derivation of the Loss Function

Proposition 1.1 (Energy Dissipation Identity)**.** *Let (Φ^*, V^*) be the true potentials. For any test potentials (Φ, V), the loss function satisfies:*

$$\mathcal{E}_{\mathbf{X}_t,\mathbf{X}_{t+\Delta t}}(\Phi, V) = \mathcal{E}_{\mathbf{X}_t,\mathbf{X}_{t+\Delta t}}(\Phi^*, V^*) + \mathcal{R}_{\mathbf{X}_t}(\Phi - \Phi^*, V - V^*) + o(\Delta t), \tag{2}$$

where $\mathcal{R}_{\mathbf{X}_t}(\delta\Phi, \delta V) \geq 0$ is the residual term, with equality iff $(\delta\Phi, \delta V) = (c, c')$ for constants c, c'.

Proof. Starting from the Itô formula applied to the energy functional:

$$E_t := \frac{1}{N}\sum_i V(X_t^i) + \frac{1}{2N^2}\sum_{i,j} \Phi(X_t^i - X_t^j),$$

we have

$$\begin{aligned} dE_t = &\frac{1}{N}\sum_i \nabla V(X_t^i) \cdot dX_t^i + \frac{1}{2N^2}\sum_{i,j} \nabla\Phi(X_t^i - X_t^j) \cdot (dX_t^i - dX_t^j) \\ &+ \frac{\sigma^2}{2N}\sum_i \Delta V(X_t^i) dt + \frac{\sigma^2}{4N^2}\sum_{i,j} \Delta\Phi(X_t^i - X_t^j) dt. \end{aligned}$$

Substituting the dynamics (**??**) and taking expectations:

$$\begin{aligned} \mathbb{E}\left[E_{t+\Delta t} - E_t\right] = &-\mathbb{E}\left[\frac{1}{N}\sum_i \left|\nabla V(X_t^i) + \frac{1}{N}\sum_j \nabla\Phi(X_t^i - X_t^j)\right|^2\right] \Delta t \\ &+ \frac{\sigma^2}{2}\mathbb{E}\left[\frac{1}{N}\sum_i \Delta V(X_t^i) + \frac{1}{N^2}\sum_{i,j} \Delta\Phi(X_t^i - X_t^j)\right] \Delta t + O(\Delta t^2). \end{aligned}$$

Rearranging gives the loss function structure. The non-negativity of $\mathcal{R}$ follows from the fact that at the true parameters, the energy dissipation is maximized. $\square$

1.3 Identifiability

Definition 1.2 (Identifiability)**.** *The pair* (Φ^*, V^*) *is* identifiable *from the data distribution if for any* $(\Phi, V) \in \mathcal{H}_\Phi \times \mathcal{H}_V$*:*

$$\mathcal{E}_\infty(\Phi, V) = \mathcal{E}_\infty(\Phi^*, V^*) \implies \Phi = \Phi^* + c_1, \quad V = V^* + c_2,$$

for some constants $c_1, c_2 \in \mathbb{R}$*.*

Remark 1.3. *Potentials are only identifiable up to additive constants since shifting both* Φ *and* V *by constants does not change the dynamics.*

Definition 1.4 (Coercivity Condition)**.** *The data distribution satisfies the* (Φ, V)-coercivity condition *with constant* $c_H > 0$ *if for all* $(\delta\Phi, \delta V) \in \mathcal{H}_\Phi \times \mathcal{H}_V$ *with* $\int \delta\Phi \, d\rho = \int \delta V \, d\nu = 0$*:*

$$\mathbb{E}\left[\frac{1}{N}\sum_i \left|\nabla\delta V(X_t^i) + \frac{1}{N}\sum_j \nabla\delta\Phi(X_t^i - X_t^j)\right|^2\right] \geq c_H\left(\|\nabla\delta V\|_{L^2_\nu}^2 + \|\nabla\delta\Phi\|_{L^2_\rho}^2\right), \tag{3}$$

where ν *is the marginal distribution of* X_t^i *and* ρ *is the distribution of* $X_t^i - X_t^j$*.*

Theorem 1.5 (Identifiability from Coercivity)**.** *Under assumptions (A1)-(A3), if the coercivity condition* (3) *holds with* $c_H > 0$*, then* (Φ^*, V^*) *is identifiable.*

Proof. Suppose $\mathcal{E}_\infty(\Phi, V) = \mathcal{E}_\infty(\Phi^*, V^*)$. Let $\delta\Phi = \Phi - \Phi^*$ and $\delta V = V - V^*$.

From Proposition 1.1, we have $\mathcal{R}(\delta\Phi, \delta V) = 0$. The residual can be written as:

$$\begin{aligned}\mathcal{R}(\delta\Phi, \delta V) &= \mathbb{E}\left[\frac{1}{N}\sum_i \left|\nabla\delta V(X_t^i) + \frac{1}{N}\sum_j \nabla\delta\Phi(X_t^i - X_t^j)\right|^2\right]\Delta t \\ &\geq c_H\left(\|\nabla\delta V\|_{L^2_\nu}^2 + \|\nabla\delta\Phi\|_{L^2_\rho}^2\right)\Delta t,\end{aligned}$$

by the coercivity condition. Thus $\mathcal{R} = 0$ implies $\nabla\delta V = 0$ and $\nabla\delta\Phi = 0$ in L^2, hence δV and $\delta\Phi$ are constants. □

1.4 Sufficient Conditions for Coercivity

We now provide verifiable sufficient conditions for coercivity.

Proposition 1.6 (Gradient Coercivity)**.** *Assume the particles* $\{X_t^i\}_{i=1}^N$ *are exchangeable and, conditional on* X_t^1*, the differences* $\{r_{1j} = X_t^j - X_t^1\}_{j=2}^N$ *are conditionally independent. If the marginal distribution* ρ *of* r_{ij} *satisfies:*

$$\mathrm{Var}(\nabla\Phi(r_{12}) \mid X_t^1) \geq c_0\|\nabla\Phi\|_{L^2_\rho}^2 \quad \textit{for all } \Phi \in \mathcal{H}_\Phi, \tag{4}$$

then the coercivity condition (3) *holds with* $c_H = \min(c_0, c_V)C_{a,N}$*, where* c_V *is the analogous constant for* V *and* $C_{a,N}$ *depends on* N*.*

Proof. The proof follows the strategy in [**?**]. By conditional independence and Lemma 1.7, we have:

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{j\neq 1}\nabla\delta\Phi(r_{1j})\right|^2 \middle| X_t^1\right] &\geq \sum_{j\neq 1}\operatorname{tr}\operatorname{Cov}(\nabla\delta\Phi(r_{1j}) \mid X_t^1) \\ &\geq (N-1)c_0\|\nabla\delta\Phi\|_{L^2_\rho}^2.\end{aligned}$$

The cross terms between $\nabla \delta V$ and $\nabla \delta \Phi$ are handled by noting that they contribute non-negatively to the variance. $\square$

Lemma 1.7 (Conditional Independence Lemma)**.** *Let $\{Y_j\}_{j=1}^n$ be $\mathbb{R}^d$-valued random variables that are conditionally independent given a σ-algebra $\mathcal{F}$. Then for any square-integrable functions $\{f_j\}$:*

$$\mathbb{E}\left[\left|\sum_{j=1}^{n} f_j(Y_j)\right|^2 \Big| \mathcal{F}\right] \geq \sum_{j=1}^{n} \operatorname{tr}\operatorname{Cov}(f_j(Y_j) \mid \mathcal{F}). \tag{5}$$

Proof. Expanding the square:

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{j} f_j(Y_j)\right|^2 \Big| \mathcal{F}\right] &= \sum_{j} \mathbb{E}\left[|f_j(Y_j)|^2 \mid \mathcal{F}\right] + \sum_{j \neq k} \mathbb{E}\left[f_j(Y_j) \mid \mathcal{F}\right] \cdot \mathbb{E}\left[f_k(Y_k) \mid \mathcal{F}\right] \\
&= \sum_{j} \operatorname{tr}\operatorname{Cov}(f_j(Y_j) \mid \mathcal{F}) + \left|\sum_{j} \mathbb{E}\left[f_j(Y_j) \mid \mathcal{F}\right]\right|^2 \\
&\geq \sum_{j} \operatorname{tr}\operatorname{Cov}(f_j(Y_j) \mid \mathcal{F}).
\end{aligned}$$

$\square$

1.5 Consistency and Convergence Rates

Theorem 1.8 (Consistency)**.** *Let $(\hat{\Phi}_n, \hat{V}_n)$ be the minimizer of $\mathcal{E}_{\mathcal{D}}$ over $\mathcal{H}_\Phi \times \mathcal{H}_V$ with data size $n = ML$. Under assumptions (A1)-(A3) and the coercivity condition, as $n \to \infty$:*

$$\|\nabla \hat{\Phi}_n - \nabla \Phi^*\|_{L^2_\rho} + \|\nabla \hat{V}_n - \nabla V^*\|_{L^2_\nu} \xrightarrow{P} 0.$$

Proof Sketch. The proof combines:

1. **Uniform convergence**: By the law of large numbers and uniform integrability, $\mathcal{E}_{\mathcal{D}}(\Phi, V) \to \mathcal{E}_\infty(\Phi, V)$ uniformly over compact subsets of $\mathcal{H}_\Phi \times \mathcal{H}_V$.

2. **Argmin continuity**: The functional $\mathcal{E}_\infty$ has a unique minimizer (up to constants) by identifiability.

3. **Coercivity lower bound**: The coercivity condition ensures that near-minimizers are close to the true solution.

The detailed proof follows the M-estimation framework in [**?**]. $\square$

Theorem 1.9 (Convergence Rate)**.** *Under the assumptions of Theorem 1.8, if additionally $\mathcal{H}_\Phi$ and $\mathcal{H}_V$ have finite VC dimension or are parametric families, then:*

$$\mathbb{E}\left[\|\nabla \hat{\Phi}_n - \nabla \Phi^*\|^2_{L^2_\rho} + \|\nabla \hat{V}_n - \nabla V^*\|^2_{L^2_\nu}\right] \leq \frac{C}{c_H^2} \cdot \frac{\dim(\mathcal{H})}{n}, \tag{6}$$

where $\dim(\mathcal{H})$ *is the effective dimension of the hypothesis space.*

1.6 Neural Network Approximation

For neural network estimators, we decompose the error into approximation and estimation components.

Theorem 1.10 (NN Approximation Error). *Let $\mathcal{F}_{NN}(W, D)$ denote the class of ReLU networks with width W and depth D. If $\Phi^*, V^* \in C^s(\mathbb{R}^d)$ for some $s > 0$, then there exist networks $\Phi_{NN}, V_{NN} \in \mathcal{F}_{NN}$ such that:*

$$\|\Phi_{NN} - \Phi^*\|_{C^2(K)} + \|V_{NN} - V^*\|_{C^2(K)} \leq C_K W^{-2s/d} (\log W)^{2s/d}, \tag{7}$$

for any compact $K \subset \mathbb{R}^d$, where C_K depends on K and the smoothness of Φ^*, V^*.

Theorem 1.11 (Total Error Bound). *The neural network estimator $(\hat{\Phi}_{NN}, \hat{V}_{NN})$ satisfies:*

$$\|\nabla\hat{\Phi}_{NN} - \nabla\Phi^*\|^2_{L^2_\rho} \leq \underbrace{C_1 W^{-2(s-1)/d}}_{\text{approximation}} + \underbrace{\frac{C_2 W D \log(WD)}{n}}_{\text{estimation}} + \underbrace{C_3 \Delta t}_{\text{discretization}} . \tag{8}$$

Optimal balance: $W \asymp n^{d/(2s+d-2)}$ gives rate $n^{-2(s-1)/(2s+d-2)}$.

1.7 Specific Examples

Example 1.12 (Gaussian Initial Distribution, $d = 1, 2, 3$). *When particles are initialized as i.i.d. Gaussian $X_0^i \sim \mathcal{N}(0, I_d)$, the coercivity constant can be explicitly computed:*

d	*1*	*2*	*3*
c_H	≥ 0.48	≥ 0.87	≥ 0.73

These bounds follow from Proposition 1.13 below.

Proposition 1.13 (Gaussian Coercivity). *For i.i.d. Gaussian particles with $X^i \sim \mathcal{N}(0, I_d)$, the kernel coercivity condition* (4) *holds with:*

$$c_0 \geq 1 - I(d, G_d),$$

where $I(d, G_d) = \sqrt{4/15}$ for $d = 1$, and is bounded by the integrals involving the spherical harmonics for $d \geq 2$.

Proof. The key is to bound $\mathbb{E}\left[\langle \nabla\Phi(r_{12}), \nabla\Phi(r_{13})\rangle\right]$ from above.

For radial kernels $\Phi(x) = \phi(|x|)$, we have $\nabla\Phi(x) = \phi'(|x|)\frac{x}{|x|}$. The correlation is:

$$\mathbb{E}\left[\nabla\Phi(r_{12}) \cdot \nabla\Phi(r_{13})\right] = \mathbb{E}\left[\phi'(|r_{12}|)\phi'(|r_{13}|)\frac{\langle r_{12}, r_{13}\rangle}{|r_{12}||r_{13}|}\right].$$

For $d = 1$: Using the explicit formula for the joint distribution of $(|r_{12}|, |r_{13}|, \text{sign}(r_{12})\text{sign}(r_{13}))$ under Gaussian, we compute:

$$I(1, G_1) = \frac{1}{\sqrt{3\pi}}\sqrt{2\pi - \frac{6\pi}{5}} = \sqrt{\frac{4}{15}} \approx 0.516.$$

Thus $c_0 \geq 1 - 0.516 = 0.484$.

For $d = 2, 3$: The computation involves integrals over spheres (see Appendix A.3). □

A Detailed Proofs

A.1 Proof of Proposition 1.1 (Energy Dissipation)

We provide the complete derivation of the trajectory-free loss function from energy dissipation principles.

Complete Proof of Proposition 1.1. Define the energy functional with test potentials (Φ, V):

$$E_t^{(\Phi,V)} := \int V \, d\mu_t^N + \frac{1}{2} \iint \Phi(x-y) \, d\mu_t^N(x) d\mu_t^N(y) = \frac{1}{N} \sum_i V(X_t^i) + \frac{1}{2N^2} \sum_{i,j} \Phi(X_t^i - X_t^j).$$

Step 1: Itô's formula. Apply Itô's formula to $E_t^{(\Phi,V)}$:

$$\begin{aligned} dE_t^{(\Phi,V)} &= \sum_i \frac{\partial E_t}{\partial X_t^i} \cdot dX_t^i + \frac{1}{2} \sum_i \operatorname{tr}\left(\frac{\partial^2 E_t}{\partial (X_t^i)^2}\right) \sigma^2 dt \\ &= \frac{1}{N} \sum_i \left[\nabla V(X_t^i) + \frac{1}{N} \sum_j \nabla_x \Phi(X_t^i - X_t^j)\right] \cdot dX_t^i \\ &\quad + \frac{\sigma^2}{2N} \sum_i \left[\Delta V(X_t^i) + \frac{1}{N} \sum_j \Delta \Phi(X_t^i - X_t^j)\right] dt. \end{aligned}$$

Step 2: Substitute dynamics. Using $dX_t^i = b^*(X_t^i, \mu_t^N)dt + \sigma dW_t^i$ where the true drift is:

$$b^*(x, \mu) = -\nabla V^*(x) - \nabla \Phi^* * \mu(x),$$

we get:

$$\begin{aligned} dE_t^{(\Phi,V)} = \frac{1}{N} \sum_i & \underbrace{\left[\nabla V(X_t^i) + \nabla \Phi * \mu_t^N(X_t^i)\right]}_{=:D_i^{(\Phi,V)}} \cdot \left[-\nabla V^*(X_t^i) - \nabla \Phi^* * \mu_t^N(X_t^i)\right] dt \\ &+ \frac{\sigma^2}{2N} \sum_i \left[\Delta V(X_t^i) + \Delta \Phi * \mu_t^N(X_t^i)\right] dt + \text{martingale}. \end{aligned}$$

Step 3: Decompose the drift product. Let $D_i = D_i^{(\Phi,V)}$ and $D_i^* = D_i^{(\Phi^*,V^*)}$. Then:

$$\begin{aligned} D_i \cdot (-D_i^*) &= -D_i \cdot D_i^* \\ &= -|D_i|^2 + D_i \cdot (D_i - D_i^*) \\ &= -|D_i|^2 + D_i \cdot \delta D_i, \end{aligned}$$

where $\delta D_i = D_i - D_i^* = \nabla \delta V(X_t^i) + \nabla \delta \Phi * \mu_t^N(X_t^i)$.

Step 4: Integrate and take expectation.

$$\begin{aligned} \mathbb{E}\left[E_{t+\Delta t}^{(\Phi,V)} - E_t^{(\Phi,V)}\right] &= -\mathbb{E}\left[\frac{1}{N}\sum_i |D_i|^2\right] \Delta t + \mathbb{E}\left[\frac{1}{N} \sum_i D_i \cdot \delta D_i\right] \Delta t \\ &\quad + \frac{\sigma^2}{2} \mathbb{E}\left[\frac{1}{N} \sum_i [\Delta V + \Delta \Phi * \mu_t^N](X_t^i)\right] \Delta t + O(\Delta t^2). \end{aligned}$$

Step 5: Rearrange to get the loss. The loss function is constructed so that minimizing it maximizes the energy dissipation rate. Rearranging:

$$\mathcal{E}(\Phi, V) := \mathbb{E}\left[\frac{1}{N}\sum_i |D_i|^2\right]\Delta t + \frac{\sigma^2}{2}\mathbb{E}\left[\frac{1}{N}\sum_i [\Delta V + \Delta\Phi * \mu_t^N](X_t^i)\right]\Delta t$$
$$- 2\mathbb{E}\left[E_{t+\Delta t}^{(\Phi,V)} - E_t^{(\Phi,V)}\right].$$

At the true parameters (Φ^*, V^*), this equals the energy dissipation rate plus the diffusion contribution, which is a known constant. The residual term is:

$$\mathcal{R}(\delta\Phi, \delta V) = \mathcal{E}(\Phi, V) - \mathcal{E}(\Phi^*, V^*) = \mathbb{E}\left[\frac{1}{N}\sum_i |\delta D_i|^2\right]\Delta t \geq 0.$$

□

A.2 Minimax Lower Bound

Theorem A.1 (Minimax Lower Bound). *Let $\mathcal{F}_s = \{\Phi \in C^s : \|\Phi\|_{C^s} \leq R\}$ be a Hölder ball. For any estimator $\hat{\Phi}$ based on $n = ML$ samples:*

$$\inf_{\hat{\Phi}} \sup_{\Phi^* \in \mathcal{F}_s} \mathbb{E}\left[\|\nabla\hat{\Phi} - \nabla\Phi^*\|_{L^2_\rho}^2\right] \geq c \cdot n^{-\frac{2(s-1)}{2s+d}}, \tag{9}$$

where $c > 0$ depends on R, d, s and the coercivity constant.

Proof. The proof uses Fano's inequality and the standard reduction to hypothesis testing.

Step 1: Construct a packing. Let $\{\Phi_1, \dots, \Phi_M\}$ be a maximal ϵ-packing of $\mathcal{F}_s$ in the $\|\nabla \cdot\|_{L^2_\rho}$ metric. By metric entropy bounds for Hölder balls:

$$\log M \geq c_1 \epsilon^{-d/(s-1)}.$$

Step 2: Bound the KL divergence. For two hypotheses Φ_i, Φ_j, the KL divergence between the induced distributions on n samples is:

$$D_{KL}(P_{\Phi_i}^{(n)} \| P_{\Phi_j}^{(n)}) \leq Cn\|\nabla\Phi_i - \nabla\Phi_j\|_{L^2_\rho}^2 \cdot (\Delta t)^2 \leq Cn\epsilon^2 \Delta t^2.$$

This follows from the fact that the loss function difference is quadratic in the gradient perturbation.

Step 3: Apply Fano's inequality. For reliable discrimination, we need:

$$\frac{1}{M}\sum_{i \neq j} D_{KL}(P_{\Phi_i}^{(n)} \| P_{\Phi_j}^{(n)}) \leq \alpha \log M,$$

for some $\alpha < 1$. This requires:

$$Cn\epsilon^2 \leq \alpha c_1 \epsilon^{-d/(s-1)},$$

which gives $\epsilon \geq c_2 n^{-(s-1)/(2s+d-2)}$.

Step 4: Conclude. Any estimator must have error at least $\epsilon^2 \geq cn^{-2(s-1)/(2s+d-2)}$ on some hypothesis. □

A.3 Gaussian Integral Computations

We compute the coercivity constants for Gaussian initial distributions.

Setup: Let $X^1, X^2, X^3 \sim_{iid} \mathcal{N}(0, I_d)$. Define $r_{12} = X^2 - X^1$ and $r_{13} = X^3 - X^1$.

Distribution of differences:

$$r_{12} \sim \mathcal{N}(0, 2I_d), \quad |r_{12}| \sim \rho_d(r) = C_d r^{d-1} e^{-r^2/4},$$

where $C_d = 1/(2^{d-1}\Gamma(d/2))$.

Joint distribution: The key quantity is:

$$\mathbb{E}\left[\frac{\langle r_{12}, r_{13}\rangle}{|r_{12}||r_{13}|}\phi(|r_{12}|)\phi(|r_{13}|)\right].$$

Dimension $d = 1$:

$$G_1(r, s) = |S^0| \int_{S^0} \xi \cdot e^{rs\xi/3} d\xi = 2(e^{rs/3} - e^{-rs/3}).$$

The bound becomes:

$$I(1, G_1) = \frac{1}{\sqrt{3}\pi}\left[\frac{1}{2}\int_{\mathbb{R}^2} e^{-5(r^2+s^2)/12+2rs/3} dr\, ds - \frac{1}{2}\int_{\mathbb{R}^2} e^{-5(r^2+s^2)/12} dr\, ds\right]^{1/2} = \sqrt{\frac{4}{15}}.$$

Dimension $d = 2$:

$$G_2(r, s) = |S^1||S^0| \int_0^1 \xi(1 - \xi^2)^{1/2}(e^{rs\xi/3} - e^{-rs\xi/3}) d\xi.$$

Numerical evaluation gives $I(2, G_2) \approx 0.127$, so $c_0 \geq 0.873$.

Dimension $d = 3$:

$$G_3(r, s) = |S^2||S^1| \int_0^1 \xi(1 - \xi^2)(e^{rs\xi/3} - e^{-rs\xi/3}) d\xi.$$

With $J_0(3) = 784\pi/125$, we get $I(3, G_3) \approx 0.266$, so $c_0 \geq 0.734$.

A.4 Neural Network Approximation Theory

Lemma A.2 (ReLU Network Approximation of Smooth Functions)**.** *Let $f \in C^s([0,1]^d)$ with $\|f\|_{C^s} \leq B$. For any $\epsilon > 0$, there exists a ReLU network f_{NN} with:*

- *Width $W = O(\epsilon^{-d/s}\log(1/\epsilon))$*
- *Depth $D = O(\log(1/\epsilon))$*

such that $\|f - f_{NN}\|_{C^0} \leq \epsilon$.

For C^2 approximation needed in our loss function:

$$\|f - f_{NN}\|_{C^2} \leq C\epsilon^{(s-2)/s},$$

using smoothed ReLU activations or sufficiently deep networks.

Proof Sketch. The proof uses:

1. Local polynomial approximation on a grid with spacing $h = \epsilon^{1/s}$.
2. ReLU networks can exactly represent piecewise linear functions.
3. Smooth activation (GELU, softplus) gives better derivative approximation.

See [**?**, **?**] for detailed constructions. □

Theorem A.3 (Rademacher Complexity of Neural Networks)**.** *Let $\mathcal{F}_{NN}(W, D, B)$ be ReLU networks with width W, depth D, and weight bound B. The Rademacher complexity satisfies:*

$$\mathcal{R}_n(\mathcal{F}_{NN}) \leq \frac{CB^D\sqrt{WD\log(WD)}}{\sqrt{n}}.$$

This leads to the estimation error bound in Theorem 1.11.

A.5 Time Discretization Error

Proposition A.4 (Discretization Error)**.** *The error from using discrete time observations with step Δt is:*

$$|\mathcal{E}_{\mathcal{D}}(\Phi, V) - \mathcal{E}_{cont}(\Phi, V)| \leq C_{Lip}(\|\nabla\Phi\|_\infty + \|\nabla V\|_\infty)^2\Delta t,$$

where $\mathcal{E}_{cont}$ is the continuous-time limit and C_{Lip} depends on the Lipschitz constants of the dynamics.

Proof. The discretization introduces error through:

1. Approximating $\int_t^{t+\Delta t}$ by $(\cdot)|_t \cdot \Delta t$.
2. Using μ_t^N instead of μ_s^N for $s \in [t, t + \Delta t]$.

By the mean value theorem and Lipschitz continuity of the flow:

$$|\mu_{t+\Delta t}^N - \mu_t^N|_{W_1} \leq C\Delta t,$$

where W_1 is the Wasserstein-1 distance. The error bound follows. □