Mean-field Analysis on Two-layer Neural Networks from a Kernel Perspective

Shokichi Takakura [1 2 *] **Taiji Suzuki** [1 2]

Abstract

In this paper, we study the feature learning ability of two-layer neural networks in the mean-field regime through the lens of kernel methods. To focus on the dynamics of the kernel induced by the first layer, we utilize a two-timescale limit, where the second layer moves much faster than the first layer. In this limit, the learning problem is reduced to the minimization problem over the intrinsic kernel. Then, we show the global convergence of the mean-field Langevin dynamics and derive time and particle discretization error. We also demonstrate that two-layer neural networks can learn a union of multiple reproducing kernel Hilbert spaces more efficiently than any kernel methods, and neural networks aquire data-dependent kernel which aligns with the target function. In addition, we develop a label noise procedure, which converges to the global optimum and show that the degrees of freedom appears as an implicit regularization.

1. Introduction

Although deep learning has achieved great success in various fields, the theoretical understanding is still limited. Several works studied the relation between deep learning and kernel medhods, which are well-studied in the machine learning community. A line of work has shown that the training dynamics of infinite-width neural networks can be approximated by linearized dynamics and the corresponding kernel is called neural tangent kernel (NTK) (Jacot et al., 2018; Arora et al., 2019b). Furthermore, generalizability of neural networks is shown to be characterized by the spectral properties of the NTK (Arora et al., 2019a; Nitanda & Suzuki, 2020). However, the NTK regime is reffered as a lazy regime and cannot explain the feature learning ability to adapt the intrinsic structure of the data since neural networks behave as a static kernel machine in the NTK regime. On the other hand, several works have shown the superiority of the neural networks to the kernel methods in terms of the sample complexity (Barron, 1993; Yehudai & Shamir, 2019; Hayakawa & Suzuki, 2020). Thus, as shown in several empirical studies (Atanasov et al., 2021; Baratin et al., 2021), neural networks must acquire the data-dependent kernel by gradient descent. However, it is challenging to establish a beyond NTK results on the feature learning of neural networks with gradient-based algorithm due to the non-convexity of the optimization landscape.

One promising approach is the mean-field analysis (Mei et al., 2018; Hu et al., 2020), which is an infinite-width limit of the neural networks in a different scaling than the NTK regime. In the mean-field regime, the optimization of 2-layer neural networks, which is non-convex in general, is reduced to the convex optimization problem over the distribution on the parameters. Exploiting the convexity of the problem, several works (Nitanda & Suzuki, 2017; Mei et al., 2018; Chizat & Bach, 2018) have shown the convergence to the global optimum. Recently, quantitative optimiztion guarantees has been established for the mean-field Langevin dynamics (MFLD) which can be regarded as a continuous limit of a noisy gradient descent (Chizat, 2022; Nitanda et al., 2022). Moreover, very recently, uniform-in-time results on the particle discretization error have been obtained (Chen et al., 2023; Suzuki et al., 2022; 2023a). This allows us to extend results effectively from infinite-width neural networks to finite-width neural networks.

Although the mean-field limit allows us to analyze the feature learning in neural networks, the connection between mean-field neural networks and its corresponding kernel is still unclear. To establish the connection to the previous works (Jacot et al., 2018; Suzuki, 2018; Ma & Wu, 2022) on the relationship between neural networks and kernel methods, we address the following question:
Is it possible to learn the optimal kernel through the MFLD? Futhermore, can this kernel align with the target function by excluding the effect of noise?

To analyze the dynamics of the kernel inside the neural net-

[*]The current affiliation is LY Corporation. This work was done when ST was affiliated with the University of Tokyo and AIP-RIKEN. [1]Department of Mathematical Informatics, the University of Tokyo, Tokyo, Japan [2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. Correspondence to: Shokichi Takakura <stakakur@lycorp.co.jp>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

works, we adopt a two-timescale limit (Marion & Berthier, 2023), which separates the dynamics of the first layer and the second layer. Then, we establish the connection between neural networks training and kernel learning (Bach et al., 2004), which involves selecting the optimal kernel for the data. We provide the global convergence gurantee of the MFLD by showing the convexity of the objective functional and derive the time and particle discretization error. Then, we prove that neural networks can aquire *data-dependent* kernel and achieve better sample complexity than any linear estimators including kernel methods for a union of multiple RKHSs. We also investigate the alignment with the target function and the degrees of freedom of the acquired kernel, which measures the complexity of the kernel, and develop the label noise procedure which provably reduces the degrees of freedom by just adding the label noise. Finally, we verify our theoretical findings by numerical experiments. Our contribution can be summarized as follows:

- We prove the convexity of the objective functional with respect to the first layer distribution and the global convergence of the MFLD in two-timescale limit in spite of the complex dependency of the second layer on the distribution of the first layer. We also derive the time and particle discretization error of the MFLD.
- We show that neural networks can adapt the intrinsic structure of the target function and achieve a better sample complexity than kernel methods for a variant of Baron space (Ma & Wu, 2022), which is a union of multiple RKHSs.
- We study the training dynamics of the kernel induced by the first layer and show that the alignment is increased during the training and achieve $\Omega(1)$ alignment while the kernel alignment at the initialization is $O(1/\sqrt{d})$ for a single-index model, where d is the input dimension. We also show that the presence of the intrinsic noise induces a bias towards the large degrees of freedom. To alleviate this issue, we propose the label noise procedure to reduce the degrees of freedom and prove the linear convergence to the global optimum.

1.1. Related Works

Relation between Neural Networks and Kernel Methods Suzuki (2018) derived the generalization error bound for deep learning models using the notion of the degrees of freedom in the kernel literature. Ma & Wu (2022) characterized the function class, which two-layer neural networks can approximate, by a union of multiple RKHSs. However, they did not give any optimization gurantee. Atanasov et al. (2021) pointed out the connection between training of neural networks and kernel learning, but their analysis is limited to linear neural networks with whitened data.

Mean-field Analysis Chen et al. (2020) conducted NTK-type analysis using mean-field limit. However, their analysis relies on the closeness of the first-layer distribution to the initial distribution. Several works have shown that the superiority of the mean-field neural networks to kernel methods including NTK with global optimization guarantee. For example, Suzuki et al. (2023b) derived a linear sample complexity with respect to the input dimension d for k-sparse parity problems although they requires exponential time for optimization. On the other hand, kernel methods require $\Omega(d^k)$ samples. In addition, Mahankali et al. (2023); Abbe et al. (2022) showed the superiority of the mean-field neural networks to the kernel methods for even quartic polymonial. However, these works fix the second layer during the training to ensure the boundedness of each neuron. Unlike these studies, we consider the joint training of the first and second layer, and focus on the relationship between neural networks and kernel machines and its implication to the feature learning ability of neural networks. We remark that Abbe et al. (2022) considered two-layer neural networks in the mean-field regime with the learnable second layer and showed the superiority to kernel methods, but their analysis utilized two-phase training, where the second layer is fixed in phase 1 and the first layer is fixed in phase 2.

Two-timescale Limit Two-timescale limit is introduced to the analysis for training of neural networks in Marion & Berthier (2023). They provided the global convegence gurantee for simplified neural networks but their analysis is limited to the single input setting and the relation to the kernel learning was not discussed. Bietti et al. (2023) leveraged the two-timescale limit to analyze the training of a non-parametric model with a linear feature extractor and show the saddle-to-saddle dynamics. However, their model differs from neural networks.

Feature Learning in Two-layer Neural Networks Aside from the mean-field analysis, there exists a line of work which studies the feature learning in (finite-width) two-layer neural networks (Damian et al., 2022; Mousavi-Hosseini et al., 2022). For instance, Damian et al. (2022) show that the random feature model with the first layer parameter updated by one-step gradient descent can learn a certain subset of p-degree polynomial with $O(d^2)$ samples while kernel methods require $\Omega(d^p)$ samples. However, most of these works consider two-stage optimization procedure, where the first layer is trained before proceeding to train the second layer.

Implicit Bias of Label Noise Implicit bias of label noise has been intensively studied recently (Damian et al., 2021; Li et al., 2021; Vivien et al., 2022). For example, Li et al. (2021) developed a theoretical framework to analyze the implicit bias of label noise in small noise and learning rate

limit and prove that the label noise induces bias towards flat minima. On the other hand, we elucidate the implicit regularization of label noise on the kernel inside the neural networks.

1.2. Notations

We write the expectation with respect to $X \sim \mu$ by $\mathbb{E}_{X\sim\mu}[\cdot]$ or $\mathbb{E}_{\mu}[\cdot]$. KL denotes the Kullback-Leibler divergence $\mathrm{KL}(\nu \mid \mu) = \int \log\left(\frac{\mu(w)}{\nu(w)}\right)\mathrm{d}\mu(w)$ and Ent denotes the negative entropy $\mathrm{Ent}(\mu) = \mathbb{E}_{\mu}[\log\mu]$. $N(v, \Sigma)$ denotes the Gaussian distribution with mean v and covariance Σ and $\upsilon(S)$ for $S \subset \mathbb{R}^d$ denotes the uniform distribution on S. $\mathcal{P}$ denotes the set of probability measures on $\mathbb{R}^{d'}$ with finite second moment. For a matrix A, $\|A\|_{\mathrm{op}}$ denotes the operator norm with respect to $\|\cdot\|_2$ and $\|A\|_{\mathrm{F}}$ denotes the Frobenius norm. For an operator $A : L^2(\mu) \to L^2(\mu)$, $\|A\|_{\mathrm{op}}$ denotes the operator norm with respect to $\|\cdot\|_{L^2(\mu)}$. For $l : \mathbb{R}^2 \to \mathbb{R}$, $\partial_1 l$ denotes the partial derivative with respect to the first argument. For a symmetric matrix A, $\lambda_{\min}(A)$ denotes the minimum eigenvalue of A. With a slight abuse of notation, we use $f(X)$ for $f : \mathbb{R}^d \to \mathbb{R}$ and $X = [x^{(1)}, \dots, x^{(n)}]^\top \in \mathbb{R}^{n\times d}$ to denote $[f(x^{(1)}), \dots, f(x^{(n)})]^\top$.

2. Problem Settings

2.1. Mean-field and Two-timescale Limit in Two-layer Neural Networks

Given input $x \in \mathbb{R}^d$, let us consider the following two-layer neural network model:

$$f(x; a, \{w_i\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^N a_i h(x; w_i),$$

where $a_i \in \mathbb{R}$, $w_i \in \mathbb{R}^{d'}$ and $h(x; w_i)$ is the activation function with parameter w_i.

Mean-field limit of the above model is defined as an integral over neurons:

$$f(x; P) := \int a h(x; w) P(\mathrm{d}a, \mathrm{d}w), \tag{1}$$

where P is a probability distribution over the parameters of the first and second layers. However, in this formulation, the first and the second layer are entangled, and thus it is difficult to characterize the feature learning, which takes place in the first layer. To alleviate this issue, we consider the following formulation:

$$f(x; a, \mu) := \int a(w) h(x; w)\mathrm{d}\mu(w),$$

where $a(w) = \int a P(\mathrm{d}a \mid w)$ and $\mu(w) = \int P(a, w)\mathrm{d}a$ is the marginal distribution of w. Similar formulation can be found in Fang et al. (2019). This formulation explicitly separates the first and the second layer, which allows us to focus on the feature learning dynamics in the first layer. More generally, we consider the multi-task learning settings. That is, $f(x; a, \mu) : \mathbb{R}^d \to \mathbb{R}^T$ is defined by

$$f_i(x; a, \mu) := \int a^{(i)}(w) h(x; w)\mathrm{d}\mu(w),$$

where $a : \mathbb{R}^{d'} \to \mathbb{R}^T$ is the second layer and $a^{(i)}$ is the i-th component of a. Note that the first layer $\mu(w)$ is shared among tasks.

Let ρ be the true or empirical distribution of the pair of input and output $(x, y) \in \mathbb{R}^{d+T}$, and ρ_X be the marginal distribution of x. Then, for $\lambda > 0$, the (regularized) risk is defined by

$$L(a, \mu) = \frac{1}{T}\sum_{i=1}^T \mathbb{E}_{\rho}[l_i(f_i(x; a, \mu), y)],$$
$$F(a, \mu) = L(a, \mu) + \lambda \mathbb{E}_{\mu}[r(a(w), w)],$$

where l_i is the loss function for the i-th task, and r is the regularization term. In this paper, we consider l^2-regularization $r(a, w) = \frac{\lambda_a}{2T}\sum_{i=1}^T {a^{(i)}}^2 + \frac{\lambda_w}{2}\|w\|_2^2$, where $a = (a^{(i)})_{i=1}^T$ and $\lambda_a, \lambda_w > 0$. We define $\bar{\lambda}_a = \lambda\lambda_a$, $\bar{\lambda}_w = \lambda\lambda_w$, and $\nu = N(0, I/\lambda_w)$ for notational simplicity.

To separate the dynamics of the first and second layer, we introduce the two-timescale limit (Marion & Berthier, 2023), where the second layer moves much faster than the first layer. In this limit, the first layer $a^{(i)}$ converges instantaneously to the unique optimal solution of $\min_a E_\rho[l(f(x; a, \mu), y)] + \frac{\bar{\lambda}_a}{2}\|a\|^2_{L^2(\mu)}$ since $F(a, \mu)$ is strongly convex with respect to a. As shown in the next section, $\|a\|^2_{L^2(\mu)}$ corresponds to the RKHS norm for the kernel induced by the first layer. Since the optimal second layer is a functional of the first layer distribution μ, we write a_μ for the optimal solution. Then, the learning problem is reduced to the minimization of the limiting functional $G(\mu) = F(a_\mu, \mu)$. We also define $U(\mu)$ by $U(\mu) := L(a_\mu, \mu) + \frac{\bar{\lambda}}{2T}\mathbb{E}_\mu[\|a_\mu(w)\|_2^2]$

Throughout the paper, we assume that $h(x; w_i)$ satisfies Assumption 2.1. For example, $\tanh(u \cdot x + b)(w = (u, b))$ satisfies the assumption if $\mathbb{E}_{\rho_X}[\|x\|_2^2], \mathbb{E}_{\rho_X}[\|x\|_2^4]$ are finite.

Assumption 2.1. $h(x; w)$ is twice differentiable with respect to w and there exist constants $c_R, c_L > 0$ such that $\sup_w |h(x; w)| \le 1, \mathbb{E}_{\rho_X}[\sup_w \|\nabla_w h(x; w)\|_2^2] \le c_R^2, \mathbb{E}_{\rho_X}[\sup_w \left\|\nabla_w^2 h(x; w)\right\|_{\mathrm{op}}^2] \le c_L^2$.

2.2. Kernel Induced by the First Layer

Let us define the kernel induced by the first layer as follows:

$$k_\mu(x, x') = \int h(x; w) h(x'; w)\mathrm{d}\mu(w).$$

Obviously, this is a symmetric positive definite kernel. It is well-known that there exists a unique RKHS $\mathcal{H}_\mu$ corresponding the kernel k_μ. Furthermore, the RKHS norm $\|f\|_{\mathcal{H}_\mu}$ is equal to the minimum of $\|a\|_{L^2(\mu)}$ over all a such that $f(x;a,\mu) = \int a(w)h(x;w)\mathrm{d}\mu(w)$ (Bach, 2017). Then, the learning problem of the second layer is equivalent to the following optimization problem:

$$\min_{f_i \in \mathcal{H}_\mu} \sum_{i=1}^T \mathbb{E}_\rho[l_i(f_i(x),y)] + \sum_{i=1}^T \frac{\bar{\lambda}_a}{2}\|f_i\|_{\mathcal{H}_\mu}^2. \quad (2)$$

Therefore, learning first layer is equivalent to kernel learning (Bach et al., 2004), which choosing suitable RKHS $\mathcal{H}_\mu$.

2.3. Mean-field Langevin Dynamics

We optimize the limiting functional $G(\mu)$ by the mean-field Langevin dynamics (MFLD):

$$\mathrm{d}w_t = -\nabla_w \frac{\delta G(\mu)}{\delta \mu}\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}B_t,$$

where $w_0 \sim \mu_0 := \nu$, $(B_t)_{t\geq 0}$ is the d'-dimensional Brownian motion, and $\frac{\delta G(\mu)}{\delta \mu}$ is the first variation of $G(\mu)$. The Fokker-Planck equation of the above SDE is given by

$$\partial_t \mu_t = \lambda \Delta \mu_t + \nabla \cdot \left[\mu_t \nabla \frac{\delta G(\mu)}{\delta \mu}\right],$$

where μ_t is the distribution of w_t. It is known that the MFLD is a Wasserstein gradient flow which minimizes the entropy-regularized functional: $\mathcal{G}(\mu) := G(\mu) + \lambda \operatorname{Ent}(\mu)$.

To implement the MFLD, we need time and particle discretization (Chen et al., 2023; Suzuki et al., 2022; 2023a). For a set of N particles $W = \{w_i\}_{i=1}^N$, we define the empirical distribution $\mu_W = \frac{1}{N}\sum_{i=1}^N \delta_{w_i}$. Let $W_k = \{w_i^{(k)}\}_{i=1}^N$ be a set of N particles at the k-th iteration, and $\mu_k^{(N)}$ be a distribution of W_k on $\mathbb{R}^{d'\times N}$. Then, at each step, we update the particles as follows:

$$w_i^{(k+1)} = w_i^{(k)} - \eta \nabla \frac{\delta G(\mu_{W_k})}{\delta \mu}(w_i^{(k)}) + \sqrt{2\eta\lambda}\xi_i^{(k)},$$

where $\eta > 0$ is the step size and $\xi_i^{(k)} \sim N(0, I)$ are i.i.d. Gaussian noise. This can be regarded as a noisy gradient descent. For more detailed discussion, see Hu et al. (2019); Suzuki et al. (2023a) for example.

3. Convergence Analysis

As shown in Nitanda et al. (2022); Chizat (2022), the convergence of the MFLD depends on the convexity of the functional and the properties of the proximal Gibbs distribution, which is defined as $p_\mu(w) \propto \exp\left(-\frac{1}{\lambda}\frac{\delta G(\mu)}{\delta \mu}(w)\right)$.

In the training of neural networks, the convexity is usually ensured by the linearity of f with respect to distribution as in Eq. (1). On the other hand, in the two-timescale limit, $f(x;\mu) := f(x;a_\mu,\mu)$ is not linear with respect to μ because the second layers a_μ depend on μ in a non-linear way. However, we can prove that the functional $G(\mu)$ is convex and its first variation can be written in a simple form if $\{l_i\}_{i=1}^T$ are convex.

Theorem 3.1. *Assume that the losses $\{l_i\}_{i=1}^T$ are convex. Then, the limiting functional $G(\mu)$ is convex. That is, it holds that*

$$G(\mu_1) + \int \frac{\delta G(\mu_1)}{\delta \mu}(w)(\mu_2(w) - \mu_1(w))\mathrm{d}w \leq G(\mu_2)$$

for any $\mu_1, \mu_2 \in \mathcal{P}$. In addition, the first variation of $G(\mu)$ is given by

$$\frac{\delta G}{\delta \mu}(\mu)(w) = \lambda\left(-\frac{\lambda_a}{2T}\|a_\mu(w)\|_2^2 + \frac{\lambda_w}{2}\|w\|_2^2\right). \quad (3)$$

See Appendix B.1 for the proof. We remark that the convexity holds for general regularization term r which is strongly convex with respect to a.

The convergence rate of the MFLD depends on the constant in the log-Sobolev inequality for the proximal distribution p_μ.

Definition 3.2. We say a probability distribution μ satisfies log-Sobolev inequality with constant $\alpha > 0$ if for all smooth function $g : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_\mu[g^2] < \infty$,

$$\mathbb{E}_\mu[g^2 \log g^2] - \mathbb{E}_\mu[g^2]\log \mathbb{E}_\mu[g^2] \leq \frac{2}{\alpha}\mathbb{E}_\mu[\|\nabla g\|^2].$$

To derive the LSI constant α, we assume either of the following conditions on the loss functions:

Assumption 3.3. l_i is squared loss. That is, $l_i(z,y) = \frac{1}{2}(z-y)^2$. In addition, $|y_i| \leq c_l$ a.s. for some constant $c_l > 0$.

Assumption 3.4. l_i is convex and twice differentiable with respect to the first argument and $|\partial_1 l_i(z,y)| \leq c_l, \left|\partial_1^2 l_i(z,y)\right| \leq 1$ for any $z, y \in \mathbb{R}$.

The latter assumption is satisfied by several loss functions such as logistic loss. Then, using the formula for the first variation, we can derive the LSI constant applying the Holley-Stroock argument (Holley & Stroock, 1987) for bounded perturbation.

Lemma 3.5. *Assume that each l_i satisfies Assumption 3.3 or 3.4. Then, the proximal distribution p_μ for any $\mu \in \mathcal{P}$ satisfies LSI with constant $\alpha = \lambda_w \exp\left(-2\frac{\lambda_a c_l^2}{\lambda_a^2}\right)$.*

The proof can be found in Appendix B.2. Unfortunately, the LSI constant is extremely small if λ is small, which reads to

exponential computational complexity with respect to $1/\lambda$. We remark that similar dependency also appears in some previous works (Chizat, 2022; Nitanda et al., 2022; Suzuki et al., 2023a), which consider models with the fixed second layer.

Remark 3.6. In the standard formulation (1), the first variation of $F(P) = \mathbb{E}_\rho[l(f(x;P),y)] + \lambda \mathbb{E}_P[r(a,w)]$ is given by $\frac{\delta F}{\delta P} = \mathbb{E}_\rho[\partial_1 l(f(x;P),y)ah(x;w)] + \lambda r(a,w)$. Then, $\mathbb{E}_\rho[\partial_1 l(f(x;P),y)ah(x;w)]$ is not bounded nor Lipschitz continuous with respect to (a,w), even if $\partial_1 l$ and h is bounded. Therefore, without two-timescale limit, it is difficult to obtain a LSI constant even for the single output setting. Indeed, previous works fix the second layer or clip the output using some bounded function (Chizat, 2022; Nitanda et al., 2022; Suzuki et al., 2023a) to ensure the boundedness or Lipschitz continuity of the output of neurons.

Combining above results, we can show the linear convergence of the MFLD.

Theorem 3.7. *Let μ^* be the minimizer of $\mathcal{G}(\mu)$ and $\mathcal{G}^N(\mu_k^{(N)}) = N\mathbb{E}_{W\sim\mu_k^{(N)}}[G(\mu_W)] + \lambda \operatorname{Ent}(\mu_k^{(N)})$. Then, for the constant α in Lemma 3.5, μ_t satisfies*

$$\mathcal{G}(\mu_t) - \mathcal{G}(\mu^*) \leq \exp(-2\alpha\lambda t)(\mathcal{G}(\mu_0) - \mathcal{G}(\mu^*))$$

for any $0 \leq t$. Furthermore, for any $\eta < \min\{1/4, 1/(4\lambda\alpha)\}$, $\mu_k^{(N)}$ satisfies

$$\frac{1}{N}\mathcal{G}^N(\mu_k^{(N)}) - \mathcal{G}(\mu^*) \leq \exp(-\alpha\lambda\eta k/2)(\mathcal{G}^N(\mu_0^{(N)}) - \mathcal{G}(\mu^*)) + \bar{\delta}_{\eta,N},$$

where $\bar{\delta}_{\eta,N} = O(\frac{1}{N} + \eta)$.

See Appendix B.3 for the proof and the concrete expression of discretization error $\bar{\delta}_{\eta,N}$. The proof is based on the framework in Suzuki et al. (2023a). To obtain discretization error, we prove some additional conditions on the smoothness of the objective functional via the optimality condition on a_μ. This is far from trivial due to the non-linear dependency of a_μ on μ. Note that we cannot apply the arguments in Chizat (2022); Nitanda et al. (2022); Suzuki et al. (2023a) without two-timescale limit since they assume the boundedness or Lipschitz continuity of each neuron. See also Remark 3.6 for detailed discussion.

Furthermore, the convergence of the loss function in the discretized setting can be transferred to the convergence of the function value of the neural networks as shown in the following proposition.

Proposition 3.8. *Assume that $h(x;w)$ is c_R-Lipschitz continuous with respect to w for any $x \in S$, where S is some subset of $\mathbb{R}^d$. Let $\Delta = \frac{c_R^2}{\lambda\alpha}(\mathcal{G}^N(\mu_k^{(N)}) - N\mathcal{G}(\mu^*)) + \frac{c_R^2 \mathcal{G}(\mu_0)}{\lambda_w}$. Then, we have*

$$\mathbb{E}_{W_k\sim\mu_k^{(N)}}\left[\sup_{\substack{(x,y)\\ \in S\times S}} \left|k_{\mu_{W_k}}(x,y) - k_{\mu^*}(x,y)\right|^2\right] = O\left(\frac{\Delta}{N}\right).$$

In addition, if l_i satisfies Assumption 3.3, then we have

$$\mathbb{E}_{W_k\sim\mu_k^{(N)}}\left[\sup_{x\in S}(f_i(x;\mu_{W_k}) - f_i(x;\mu^*))^2\right] = O\left(\frac{c_0^2(\bar{\lambda}_a^2+1)}{\bar{\lambda}_a^4}\cdot\frac{\Delta}{N}\right).$$

See Appendix B.4 for the proof. Note that this result is not covered by Lemma 2 in Suzuki et al. (2023b) since their analysis relies on the Lipschitz continuity of each neuron. In the following sections, we consider infinite-width neural networks trained by the MFLD for simplicity, but the results can be transferred via this proposition to the finite-width neural networks trained by the discretized MFLD.

4. Generalization Error for Barron Spaces and Superiority to Kernel Methods

In this section, we provide the separation of the generalization error between neural networks and kernel methods which cannot adapt the intrinsic structure of the target function.

Let $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ be training data sampled from the true distribution in an i.i.d. manner. We define $X = [x^{(1)}, \ldots, x^{(n)}]^\top \in \mathbb{R}^{n\times d}$, $Y_i = [y_i^{(1)}, \ldots, y_i^{(n)}]^\top \in \mathbb{R}^n$ and $\hat{\Sigma}_\mu = \mathbb{E}_\mu[h(X;w)h(X;w)^\top]$. In the following, we write the true distribution by ρ and the empirical distribution by $\hat{\rho}$. In addition, to distinguish the empirical risk and population risk, we write $U_\rho(\mu), \mathcal{G}_\rho(\mu)$ for the (regularized) population risk and $U_{\hat{\rho}}(\mu), \mathcal{G}_{\hat{\rho}}(\mu)$ for the empirical risk. Then, we assume the following.

Assumption 4.1. the output y_i for each task is generated by $y_i = f_i^\circ(x) + \varepsilon_i$, where $f_i^\circ : \mathbb{R}^d \to \mathbb{R}$ is the target function and ε_i is the noise, which follows $\upsilon([-\sigma,\sigma])$ independently for some $\sigma \geq 0$.

To see the benefit of the feature learning or kernel learning, we consider the following function class.

Definition 4.2 (KL-restricted Barron space)**.** Let $\mathcal{P}_M = \{\mu \in \mathcal{P} \mid \mathrm{KL}(\nu \mid \mu) \leq M\}$ for some $M > 0$. Then, we define the KL-restricted Barron space as

$$\mathcal{B}_M = \{f(x;a,\mu) \mid \mu \in \mathcal{P}_M, a \in L^2(\mu)\},$$

and the corresponding norm as

$$\|g\|_{\mathcal{B}_M} = \inf_{\mu\in\mathcal{P}_M, a\in L^2(\mu)} \left\{\|a\|_{L^2(\mu)} \mid g(x) = f(x;a,\mu)\right\}.$$

This can be seen as a variant of Barron space in E et al. (2019); Ma & Wu (2022). Similar function classes are also considered in Bach (2017) but they consider Frank-Wolfe type optimization algorithm, which is different from usual gradient descent. We remark that Barron space can be regarded as a union of RKHS: $\mathcal{B}_M = \bigcup_{\mu \in \mathcal{P}_M} \mathcal{H}_\mu$ and the norm $\|f\|_{\mathcal{B}_M}$ is equal to the minimum of $\|f\|_{\mathcal{H}_\mu}$ over all $\mu \in \mathcal{P}_M$ (Ma & Wu, 2022).

To obtain the generalization guarantee, we utilize the Rademacher complexity. The Rademacher complexity of a function class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}^T$ is defined by

$$\mathfrak{R}(\mathcal{F}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \sigma_{it} f_t(x_i) \right],$$

where σ_{it} is an i.i.d. Rademacher random variable ($P(\sigma_{it} = 1) = P(\sigma_{it} = -1) = 1/2$). Then, we have the following bound for the mean-field neural networks.

Lemma 4.3. *Assume that $h(x; w)$ satisfies Assumption 2.1. Define a class of the mean-field neural networks by*

$$\mathcal{F}_{R,M} = \left\{ f(x; a, \mu) \mid \mathrm{KL}(\nu \mid \mu) \le M, \|a\|^2_{L^2(\mu)} \le R \right\}.$$

Then, the Rademacher complexity of $\mathcal{F}_{R,M}$ is bounded by

$$\mathfrak{R}(\mathcal{F}_{R,M}) \le \sqrt{\frac{R(4M + 2T \log 2)}{nT}} = O\left(\frac{R(M+T)}{nT} \right).$$

This is a generalization of the result in Chen et al. (2020). See Appendix C.2 for the proof.

Then, we can derive the generalization error bound for the empirical risk minimizer $\hat{\mu} = \operatorname{argmin}_{\mu} \mathcal{G}_{\hat{\rho}}(\mu)$. Note that the empirical risk minimizer $\hat{\mu}$ can be obtained by the MFLD as shown in Theorem 3.7.

Theorem 4.4. *Assume that Assumption 2.1 and 4.1 holds with $\sigma = 0$, $T = 1$, and $f_1^\circ \in \mathcal{B}_M, \|f_1^\circ\|_{\mathcal{B}_M} \le R$ for given $M, R > 0$. In addition, let $\lambda = 1/\sqrt{n}$, and $\lambda_a = 2M/R$. Then, with probability at least $1 - \delta$ over the choice of training examples, it holds that*

$$\|f(\cdot; \hat{\mu}) - f^\circ\|^2_{L^2(\rho_X)} = O\left((R+1) \sqrt{\frac{M + 1 + \log 1/\delta}{n}} \right).$$

See Appendix C.3 for the proof. Therefore, if $R = O(1)$, the mean-field neural networks can learn the Barron space with $n = O(M)$ samples.

Next, we show the lower bound of the estimation error for kernel methods. For a given kernel k, a kernel method returns a function of the form $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ for $\alpha_i \in \mathbb{R}$. This type of estimator is called linear estimator. The following theorem gives the lower bound of the sample complexity for any linear estimators.

Theorem 4.5. *Fix $m \in \mathbb{N}$ and let $d \ge \max\{2, m\}$ and ρ_X be the uniform distribution on $\{-1, 1\}^d$ and $h(x; w) = \tanh(u \cdot x + b)$, where $w = (u, b) \in \mathbb{R}^{d+1}$. In addition, let $H_n \subset L^2(\rho_X)$ be a set of functions of the form $\sum_{i=1}^n \alpha_i h_i(x)$ and $d(f, H_n) = \inf_{\hat{f} \in H_n} \|f - \hat{f}\|_{L^2(\rho_X)}$. Then, there exist constants $c_1, c_2 > 0$ which is independent of d, such that, for every choice of fixed basis functions $h_1(x), \dots, h_n(x)$, it holds that*

$$\sup_{f \in \mathcal{B}_M, \|f\|^2_{\mathcal{B}_M} \le R} d(f, H_n) \ge \frac{1}{4}$$

if $n \le N/2$ and $M = c_1 d \log d$, $R = c_2$ where $N = \binom{d}{m} = \Omega(d^m)$.

The proof can be found in Appendix C.4. The key observation for the proof is that we can construct a function in the Barron space that approximates a single index model with certain regularity by taking a measure which concentrates on a line toward a certain direction. This theorem implies that any kernel estimator with $n = o(d^m)$ cannot learn the Barron space with $M = \Omega(d \log d)$. This is in contrast to the mean-field neural networks which can learn the Barron space with $n = O(d \log d)$ samples as shown in Theorem 4.4. This is because the kernel methods cannot adapt the underlying RKHS under the target function. Therefore, feature learning or kernel learning is essential to obtain good generalization results.

5. Properties of the Kernel Induced by the First Layer

In the previous section, we proved that feature learning is essential to obtain good generalization results and two-layer neural networks trained by the MFLD can excel over kernel methods. In this section, we study the properties of the kernel trained via the MFLD. We show that in regression problem, the kernel induced by the first layer moves to increase kernel and parameter alignment. We also proved that the presence of the noise ε induces bias towards the large degrees of freedom. To overcome this issue, we provide the label noise procedure, which provably converges to the global minima of the objective functional with the degrees of freedom regularization.

5.1. Kernel and Parameter Alignment

For simplicity, we consider the single output setting $T = 1$ and define $f^\circ = f_1^\circ$. In addition, we consider $\tanh$ activation $h(x; w) = \tanh(u \cdot x + b)$ ($w = (u, b)$) and assume that $\rho_X = N(0, I)$. To measure the adaptation of the kernel to the target function, we define the kernel alignment Cristianini et al. (2001), which is commonly used to measure the similarity between kernel and labels.

Definition 5.1. For $\mu \in \mathcal{P}$, the empirical and population kernel alignment is defined by

$$\hat{A}(\mu) = \frac{f^\circ(X)^\top \hat{\Sigma}_\mu f^\circ(X)}{\|f^\circ(X)\|_2^2 \left\|\hat{\Sigma}_\mu\right\|_{\mathrm{F}}},$$
$$A(\mu) = \frac{\mathbb{E}_{x\sim\rho_X, x'\sim\rho_X}[f^\circ(x) k_\mu(x,x') f^\circ(x')]}{\mathbb{E}_{\rho_X}[f^\circ(x)^2]\sqrt{\mathbb{E}_{x\sim\rho_X, x'\sim\rho_X}[k(x,x')^2]}}.$$

Note that $\hat{A}(\mu)$ and $A(\mu)$ satisfy $0 \le \hat{A}(\mu), A(\mu) \le 1$ and larger $A(\mu)$ means that the kernel is aligned with the target function.

In this section, we consider the regression problem with squared loss. If $\sigma = 0$, the limiting functional $U_{\hat{\rho}}(\mu)$ has the following explicit formula.

$$U_{\hat{\rho}}(\mu) = \frac{\bar{\lambda}_a}{2} f^\circ(X)^\top (\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1} f^\circ(X).$$

From Jensen's inequality, we have

$$\hat{A}(\mu) \ge \frac{\bar{\lambda}_a \|f^\circ(X)\|_2^2}{2U_{\hat{\rho}}(\mu) n} - \bar{\lambda}_a.$$

See Lemma D.2 for the detailed derivation. Therefore, the minimization of $U_{\hat{\rho}}(\mu)$ is equivalent to the maximization of the lower bound of the kernel alignment.

To derive a concrete expression of the kernel alignment, we assume that the target function f° is a single-index model , which is a common structural assumption on the target function (Bietti et al., 2022).

Assumption 5.2. There exist $\tilde{f} : \mathbb{R} \to \mathbb{R}$, $u^\circ \in \mathbb{R}^d$ ($\|u^\circ\|_2 = 1$) such that $\tilde{f}$ is differentiable, $\|\tilde{f}'\|_\infty$, $\|\tilde{f}\|_\infty \le 1$, $\mathbb{E}_{z\sim N(0,1)}[\tilde{f}(z)] = 0$, and $f^\circ(x) = \tilde{f}(u^\circ \cdot x)$.

We also define the parameter alignment, which measures the similarity between the first layer parameters and the intrinsic direction of the target function.

Definition 5.3. For $\mu \in \mathcal{P}$, the parameter alignment is defined as

$$P(\mu) = \mathbb{E}_{(u,b)\sim\mu}\left[\frac{(u^\top u^\circ)^2}{\|u\|^2}\right].$$

Here, we define $\frac{(u^\top u^\circ)^2}{\|u\|^2} = 0$ for $u = 0$.

This is the expected cosine similarity between parameters and the target direction, and thus $0 \le P(\mu) \le 1$. Note that larger $P(\mu)$ means that the first layer parameters are aligned with the intrinsic direction of the target function.

Then, we have the following result on the kernel for empirical risk minimizer $\hat{\mu}$.

Theorem 5.4. *Assume that Assumption 5.2 holds. Then, there exists universal constants c_3, c_4, c_5 satisfying the following: Let $\hat{\mu}$ be the minimizer of $\mathcal{G}_{\hat{\rho}}(\mu)$ with $n \ge c_3(d\log d + \log 1/\delta)$, $\lambda = c_4/(d\log d)$, and $\lambda_a = c_5 d\log d$ for $0 < \delta < 1$ and $d \ge 2$. Then, the kernel and parameter alignment for the initial distribution μ_0 and the empirical risk minimizer $\hat{\mu}$ satisfies*

$$A(\mu_0) = O(1/\sqrt{d}),\ A(\hat{\mu}) = \Omega(1),$$
$$P(\mu_0) = O(1/d),\ P(\hat{\mu}) = \Omega(1),$$

with probability at least $1-\delta$ over the choice of samples.

See Appendix D.2 for the proof. In high-dimensional setting $d \gg 1$, $A(\hat{\mu}), P(\hat{\mu}) = \Omega(1)$ is a significant improvement over $P(\mu_0) = O(1/d)$, $A(\mu_0) = O(1/\sqrt{d})$ at the initialization. For the parameter alignment, similar results are shown in Mousavi-Hosseini et al. (2022), but they train only the first layer and use the norm of the irrelevant directions as a measure of the alignment. On the other hand, we consider the joint learning of the first and second layers and use the cosine similarity as a measure of the alignment. In addition, Atanasov et al. (2021) studied the kernel alignment of NTK, but their analysis is limited to linear neural networks. Furthermore, Ba et al. (2022); Wang et al. (2024) studied the alignment of the conjugate kernel of two-layer neural networks after one-step gradient descent, but their frameworks cannot deal with full training dynamics.

5.2. Degrees of Freedom and Label Noise

To measure the complexity of the acquired kernel, we define the (empirical) degrees of freedom by

$$d_\lambda(\mu) = \mathrm{tr}\left[\hat{\Sigma}_\mu(\hat{\Sigma}_\mu + n\lambda I)^{-1}\right]$$

for $\lambda > 0$. This quantity is the effective dimension of the kernel k_μ and plays a crucial role in the analysis of kernel regression (Caponnetto & De Vito, 2007). In addition, it is known that the degrees of freedom is related to the compressibility of neural networks (Suzuki et al., 2020).

Under Assumption 4.1, each label Y_i can be decomposed as $Y_i = f_i^\circ(X) + \varepsilon_i$, where ε_i is the observation noise. Then, taking the expectation of $U_{\hat{\rho}}(\mu)$ with respect to ε yields $\mathbb{E}_\varepsilon[U_{\hat{\rho}}(\mu)] = B - V + \text{const.}$, where $B = \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^T \mathbb{E}_\varepsilon[f_i^\circ(X)^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1} f_i^\circ(X)]$ and $V = \frac{\bar{\lambda}_a \sigma^2}{6n} d_{\bar{\lambda}_a}(\mu)$. See Lemma D.3 for the derivation. Here, B is the bias term, which corresponds to the alignment with the target function as shown in the previous section, and V is the variance term, which corresponds to the degrees of freedom. Since $-V$ appears in $\mathbb{E}_\varepsilon[U_{\hat{\rho}}(\mu)]$, minimizing $U_{\hat{\rho}}(\mu)$ leads to the larger variance and the degrees of freedom. We verify this phenomenon in Section 6.

To obtain good prediction performance, we need to minimize $B+V$ and control the bias-variance tradeoff. Here, we

consider the following objective functional with the degrees of freedom regularization:

$$\mathcal{L}(\mu) := \mathcal{G}_{\hat{\rho}}(\mu) + \frac{\bar{\lambda}_a \tilde{\sigma}^2}{6n} d_{\bar{\lambda}_a}(\mu).$$

Here, $\tilde{\sigma} \geq 0$ controls the strength of the regularization. Since this regularization is proportional to the variance V, minimizing $\mathcal{L}(\mu)$ would lead to smaller variance and better generalization. To obtain the minimizer of the above functional $\mathcal{L}(\mu)$, we provide the label noise procedure, where we add independent *label noise* to the training data for implicit regularization. In the discretized MFLD update, we add independent label noise $\tilde{\varepsilon}_i \sim \upsilon([-\tilde{\sigma}, \tilde{\sigma}]^n)$ to Y_i at each time step. We use the noisy label $\tilde{Y}_i := Y_i + \tilde{\varepsilon}_i$ to train the second layer and obtain $\tilde{a}_\mu^{(i)} := \operatorname{argmin} \frac{1}{nT} \sum_{i=1}^T \|\tilde{Y}_i - f(X; a, \mu)\|_2^2 + \frac{\bar{\lambda}_a}{2} \mathbb{E}_\mu[a(w)^2]$. Here, the noisy limiting functional $G_{\tilde{\varepsilon}}(\mu)$ is defined as

$$G_{\tilde{\varepsilon}}(\mu) := \frac{1}{nT} \sum_{i=1}^T \|Y_i - f(X; \tilde{a}_\mu, \mu)\|_2^2 + \frac{\bar{\lambda}_a}{2} \mathbb{E}_\mu \left[\|\tilde{a}_\mu(w)\|_2^2 \right] + \frac{\bar{\lambda}_w}{2} \mathbb{E}_\mu[\|w\|_2^2].$$

Note that we use the clean label Y_i to define $G_{\tilde{\varepsilon}}(\mu)$ instead of $\tilde{Y}_i$. Then, we update the first layer by the following discretized MFLD.

$$w_i^{(k+1)} = w_i^{(k)} - \eta \nabla \frac{\delta G_{\tilde{\varepsilon}^{(k)}}(\mu_{W_k})}{\delta \mu}(w_i^{(k)}) + \sqrt{2\eta\lambda}\xi_i^{(k)},$$

where $\tilde{\varepsilon}^{(k)}$ is an independent noise at the k-th iteration. In fact, the expectation of $G_{\tilde{\varepsilon}}(\mu) + \lambda \operatorname{Ent}(\mu)$ with respect to $\tilde{\varepsilon}$ is equal to $\mathcal{L}(\mu)$ and the above procedure can be seen as the stochastic MFLD for minimizing $\mathcal{L}(\mu)$. Indeed, the following theorem holds.

Theorem 5.5. *Let* $\mu^* = \operatorname{argmin}_\mu \mathcal{L}(\mu)$. *Then, for* $\eta < \min(1/4, 1/(4\alpha\lambda))$ *and* $0 \leq \tilde{\sigma}^2/3 \leq \lambda_{\min}\left(\frac{1}{T}\sum_{i=1}^T Y_i Y_i^\top\right)$, *we have*

$$\frac{1}{N}\mathbb{E}[\mathcal{L}^N(\mu_k^N)] - \mathcal{L}(\mu^*) \leq \exp(-\alpha\lambda\eta k/2)(\mathbb{E}[\mathcal{L}^N(\mu_0^N)] - \mathcal{L}(\mu^*)) + \bar{\delta}'_{\eta,N},$$

where $\bar{\delta}'_{\eta,N} = O(\eta + \frac{1}{N})$. *Here, the expectation is taken with respect to the randomness of the label noise.*

See Appendix D.3 for the proof. Intuitively, the degrees of freedom represents a metric for quantifying the adaptability to noise and the first layer performs the robust feature learning where the second layer is difficult to fit the label noise. Suzuki & Suzuki (2023) has shown the Bayes optimality of two-layer linear neural networks which minimizes the empirical risk with the degrees of freedom regularization. However, they ignore the optimization aspect and directly assume that the optimal solution can be obtained. Note that the condition on $\tilde{\sigma}^2$ is needed to ensure the convexity of the objective and the multi-learning setting is necessary to set $\tilde{\sigma} > 0$ since $\frac{1}{T}\sum_{i=1}^T Y_i Y_i^\top$ must be full rank. However, as shown in Section 6, the label noise procedure is effective even for the single output setting.

6. Numerical Experiments

To validate our theoretical results, we conduct numerical experiments with synthetic data. Specifically, we consider $f^\circ(x) = x_1 x_2$ for $d = 15$. Then, the samples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ are independently generated so that $x^{(i)}$ follows $N(0, I)$ and $y^{(i)} = f^\circ(x^{(i)}) + \varepsilon^{(i)}$, where $\varepsilon^{(i)} \sim \upsilon([\sigma, \sigma])$. We consider a finite width neural network with the width $m = 2000$. We trained the network via noisy gradient descent with $\eta = 0.2, \lambda = 0.004, \lambda_w = 0.25, \lambda_a = 0.25$ until $T = 10000$. The results are averaged over 5 different random seeds.

First, we investigated the training dynamics of the kernel by changing the intrinsic noise σ. As shown in Figure 1, kernel moves to increase the kernel alignment and the degrees of freedom. In addition, the intrinsic noise increases the degrees of freedom, which is consistent with our arguments in Section 5.2.

Next, we demonstrated the effectiveness of the label noise procedure. Fig. 2 shows the evolution of the degrees of freedom and the test loss during the training for different $\tilde{\sigma}$. As expected, the label noise procedure reduces the degrees of freedom. Moreover, the test loss is also improved, which implies that the degrees of freedom is a good regularization for the generalization error.

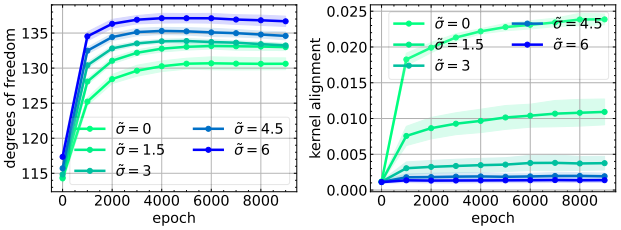

Figure 1. Evolution of the kernel alignment and the degrees of freedom of neural network optimized by the MFLD

7. Conclusion

In this paper, we studied the feature learning ability of two-layer neural networks in the mean-field regime via kernel learning formulation. For that purpose, we proposed to use the two-timescale limit to analyze the training dynamics of the mean-field neural networks. Then, we provided the linear convergence guarantee to the global optimum by

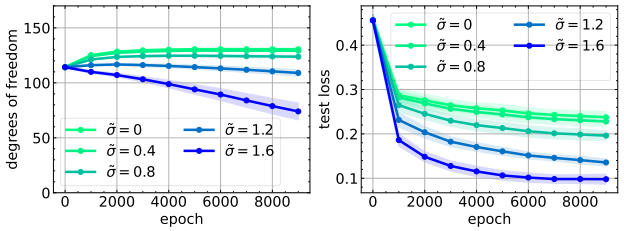

Figure 2. Evolution of the degrees of freedom and the test error of the label noise procedure

showing the convexity of the limiting functional and derive the discretization error. We also studied the generalization ability of the empirical risk minimizer and proved that the feature learning is essential to obtain good generalization results for a union of multiple RKHSs. Then, we showed that the kernel induced by the first layer moves to increase kernel and parameter alignment and the intrinsic noise in labels induces bias towards the large degrees of freedom. Finally, we proposed the label noise procedure to reduce the degrees of freedom and provided the global convergence guarantee.

Acknowledgements

ST was partially supported by JST CREST (JPMJCR2015). TS was partially supported by JSPS KAKENHI (20H00576) and JST CREST (JPMJCR2115).

Impact Statement

This paper presents theoretical analysis on machine learning. Thus, the potential impact to security or trustworhy issues to the society would be limited.

References

Abbe, E., Adsera, E. B., and Misiakiewicz, T. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pp. 4782–4887. PMLR, 2022.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 322–332. PMLR, May 2019a. URL https://proceedings.mlr.press/v97/arora19a.html. ISSN: 2640-3498.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On Exact Computation with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

Atanasov, A., Bordelon, B., and Pehlevan, C. Neural Networks as Kernel Learners: The Silent Alignment Effect. In *International Conference on Learning Representations*, October 2021. URL https://openreview.net/forum?id=1NvflqAdoom.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35: 37932–37946, 2022.

Bach, F. On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017. ISSN 1533-7928. URL http://jmlr.org/papers/v18/15-178.html.

Bach, F. R., Lanckriet, G. R. G., and Jordan, M. I. Multiple kernel learning, conic duality, and the SMO algorithm. In *Twenty-first international conference on Machine learning - ICML '04*, pp. 6, Banff, Alberta, Canada, 2004. ACM Press. doi: 10.1145/1015330.1015424. URL http://portal.acm.org/citation.cfm?doid=1015330.1015424.

Bakry, D. and Émery, M. Diffusions hypercontractives. *Séminaire de probabilités de Strasbourg*, 19:177–206, 1985. URL https://eudml.org/doc/113511. Publisher: Springer - Lecture Notes in Mathematics.

Baratin, A., George, T., Laurent, C., Hjelm, R. D., Lajoie, G., Vincent, P., and Lacoste-Julien, S. Implicit Regularization via Neural Feature Alignment. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 2269–2277. PMLR, March 2021. URL https://proceedings.mlr.press/v130/baratin21a.html. ISSN: 2640-3498.

Barron, A. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993. ISSN 0018-9448, 1557-9654. doi: 10.1109/18.256500. URL https://ieeexplore.ieee.org/document/256500/.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. In Goos, G., Hartmanis, J., Van Leeuwen, J., Helmbold, D., and Williamson, B. (eds.), *Computational Learning Theory*, volume 2111, pp. 224–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-42343-0 978-3-540-44581-4. doi: 10.1007/3-540-44581-1_15. URL http://link.springer.

com/10.1007/3-540-44581-1_15. Series Title: Lecture Notes in Computer Science.

Bietti, A., Bruna, J., Sanford, C., and Song, M. J. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems*, May 2022. URL https://openreview.net/forum?id=wt7cd9m2cz2.

Bietti, A., Bruna, J., and Pillaud-Vivien, L. On Learning Gaussian Multi-index Models with Gradient Flow, November 2023. URL http://arxiv.org/abs/2310.19793. arXiv:2310.19793 [cs, math, stat].

Caponnetto, A. and De Vito, E. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3): 331–368, July 2007. ISSN 1615-3375, 1615-3383. doi: 10.1007/s10208-006-0196-8. URL http://link.springer.com/10.1007/s10208-006-0196-8.

Chen, F., Ren, Z., and Wang, S. Uniform-in-time propagation of chaos for mean field Langevin dynamics, November 2023. URL http://arxiv.org/abs/2212.03050. arXiv:2212.03050 [math, stat].

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. A Generalized Neural Tangent Kernel Analysis for Two-layer Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13363–13373. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9afe487de556e59e6db6c862adfe25a4-Abstract.html.

Chizat, L. Mean-Field Langevin Dynamics : Exponential Convergence and Annealing. *Transactions on Machine Learning Research*, May 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=BDqzLH1gEm.

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

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. On Kernel-Target Alignment. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/hash/1f71e393b3809197ed66df836fe833e5-Abstract.html.

Damian, A., Ma, T., and Lee, J. D. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Damian, A., Lee, J., and Soltanolkotabi, M. Neural Networks can Learn Representations with Gradient Descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 5413–5452. PMLR, June 2022. URL https://proceedings.mlr.press/v178/damian22a.html. ISSN: 2640-3498.

E, W., Ma, C., and Wu, L. *A priori* estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019. ISSN 15396746, 19450796. doi: 10.4310/CMS.2019.v17.n5.a11. URL https://www.intlpress.com/site/pub/pages/journals/items/cms/content/vols/0017/0005/a011/.

Fang, C., Dong, H., and Zhang, T. Over Parameterized Two-level Neural Networks Can Learn Near Optimal Feature Representations, October 2019. URL http://arxiv.org/abs/1910.11508. arXiv:1910.11508 [cs, math, stat].

Hayakawa, S. and Suzuki, T. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, March 2020. ISSN 0893-6080. doi: 10.1016/j.neunet.2019.12.014. URL https://www.sciencedirect.com/science/article/pii/S089360801930406X.

Holley, R. and Stroock, D. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5):1159–1194, March 1987. ISSN 1572-9613. doi: 10.1007/BF01011161. URL https://doi.org/10.1007/BF01011161.

Hsu, D. Dimension lower bounds for linear approaches to function approximation. *Daniel Hsu's homepage*, 2021.

Hu, K., Ren, Z., Siska, D., and Szpruch, L. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019.

Hu, K., Ren, Z., Siska, D., and Szpruch, L. Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks, December 2020. URL http://arxiv.org/abs/1905.07769. arXiv:1905.07769 [math, stat].

Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Li, Z., Wang, T., and Arora, S. What Happens after SGD Reaches Zero Loss? –A Mathematical Framework. In *International Conference on Learning Representations*, October 2021. URL https://openreview.net/forum?id=siCt4xZn5Ve.

Ma, C. and Wu, L. The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022. URL https://link.springer.com/article/10.1007/s00365-021-09549-y. Publisher: Springer.

Mahankali, A. V., HaoChen, J. Z., Dong, K., Glasgow, M., and Ma, T. Beyond NTK with Vanilla Gradient Descent: A Mean-Field Analysis of Neural Networks with Polynomial Width, Samples, and Time. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Marion, P. and Berthier, R. Leveraging the two timescale regime to demonstrate convergence of neural networks, October 2023. URL http://arxiv.org/abs/2304.09576. arXiv:2304.09576 [cs, math, stat].

Maurer, A. A Vector-Contraction Inequality for Rademacher Complexities. In Ortner, R., Simon, H. U., and Zilles, S. (eds.), *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pp. 3–17, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7. doi: 10.1007/978-3-319-46379-7_1.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33), August 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1806579115. URL https://pnas.org/doi/full/10.1073/pnas.1806579115.

Mousavi-Hosseini, A., Park, S., Girotti, M., Mitliagkas, I., and Erdogdu, M. A. Neural Networks Efficiently Learn Low-Dimensional Representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2022.

Nitanda, A. and Suzuki, T. Stochastic Particle Gradient Descent for Infinite Ensembles, December 2017. URL http://arxiv.org/abs/1712.05438. arXiv:1712.05438 [cs, math, stat].

Nitanda, A. and Suzuki, T. Optimal Rates for Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime. In *International Conference on Learning Representations*, 2020.

Nitanda, A., Wu, D., and Suzuki, T. Convex Analysis of the Mean Field Langevin Dynamics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 9741–9757. PMLR, May 2022. URL https://proceedings.mlr.press/v151/nitanda22a.html. ISSN: 2640-3498.

Suzuki, K. and Suzuki, T. Optimal criterion for feature learning of two-layer linear neural network in high dimensional interpolation regime. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=Jc0FssXh2R.

Suzuki, T. Fast generalization error bound of deep learning from a kernel perspective. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 1397–1406. PMLR, March 2018. URL https://proceedings.mlr.press/v84/suzuki18a.html. ISSN: 2640-3498.

Suzuki, T., Abe, H., Murata, T., Horiuchi, S., Ito, K., Wachi, T., Hirai, S., Yukishima, M., and Nishimura, T. Spectral Pruning: Compressing Deep Neural Networks via Spectral Analysis and its Generalization Error. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2839–2846, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/393. URL https://www.ijcai.org/proceedings/2020/393.

Suzuki, T., Nitanda, A., and Wu, D. Uniform-in-time propagation of chaos for the mean-field gradient Langevin dynamics. In *The Eleventh International Conference on Learning Representations*, September 2022. URL https://openreview.net/forum?id=_JScUk9TBUn.

Suzuki, T., Wu, D., and Nitanda, A. Convergence of mean-field Langevin dynamics: time-space discretization, stochastic gradient, and variance reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023a. URL https://openreview.net/forum?id=9STYRIVx6u.

Suzuki, T., Wu, D., Oko, K., and Nitanda, A. Feature learning via mean-field Langevin dynamics: classifying sparse parities and beyond. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023b. URL https://https://openreview.net/forum?id=tj86aGVNb3.

Vivien, L. P., Reygner, J., and Flammarion, N. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 2127–2159. PMLR, June 2022. URL https://proceedings.mlr.press/v178/vivien22a.html. ISSN: 2640-3498.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, Z., Wu, D., and Fan, Z. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. *arXiv preprint arXiv:2402.10127*, 2024.

Yehudai, G. and Shamir, O. On the Power and Limitations of Random Features for Understanding Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/5481b2f34a74e427a2818014b8e103b0-Abstract.html.

A. Auxiliary Lemmas

Lemma A.1 (Holley & Stroock (1987))**.** *Assume that a probability distribution $p(w)$ satisfies the LSI with a constant $\alpha > 0$. For a bounded perturbation $B(w) : \mathbb{R}^{d'} \to \mathbb{R}$, define $p'(w) = p(w)\exp(B(w))/\mathbb{E}_p[\exp(B(w))]$. Then, p' satisfies the LSI with a constant $\frac{\alpha}{\exp\left(4\|B\|_\infty\right)}$.*

Lemma A.2. *The optimal i-th second layer $a_\mu^{(i)}$ satisfies*

$$a_\mu^{(i)}(w) = -\frac{1}{\bar{\lambda}_a}\mathbb{E}_\rho[\partial_1 l_i(f_i(x; a_\mu, \mu), y_i)h(x; w)].$$

for any $w \in \mathbb{R}^{d'}$.

Proof. From the optimality condition on $a_\mu^{(i)}$, it holds that

$$\frac{\partial L}{\partial a_\mu^{(i)}}(a^{(i)}, \mu)(w) + \lambda\partial_{a_i} r(a_\mu(w), w)\mu(w) = \mathbb{E}[\partial_1 l_i(f_i(x; a_\mu^{(i)}, \mu), y_i)h(x; w)]\mu(w) + \bar{\lambda}_a a_\mu^{(i)}(w)\mu(w) = 0.$$

Thus, we have

$$a_\mu^{(i)}(w) = -\frac{1}{\bar{\lambda}_a}\mathbb{E}_\rho[\partial_1 l_i(f_i(x; a_\mu, \mu), y)h(x; w)],$$

which completes the proof. □

Lemma A.3. *Define $T : L^2(\mu) \to L^2(\rho_X)$ by*

$$T(a) = \int a(w)h(x; w)\mathrm{d}\mu(w)$$

and its adjoint operator $T^* : L^2(\rho_X) \to L^2(\mu)$ by

$$T^*(f) = \int f(x)h(x; w)\mathrm{d}\rho(x).$$

For l^2-loss, the optimal i-th second layer $a_\mu^{(i)}$ has the following explicit formula.

$$\begin{aligned} a_\mu^{(i)}(w) &= (T^*T + \bar{\lambda}_a \,\mathrm{Id})^{-1}T^* f_i^\circ \\ &= T^*(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{f}_i^\circ, \end{aligned}$$

where $f_i^\circ(x') := \mathbb{E}_\rho[y_i \mid x = x']$ is the conditional expectation of y_i given x. In addition, if ρ_X is the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, then, $a_\mu^{(i)}$ is written by

$$a_\mu^{(i)}(w) = h(X; w)^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}Y_i.$$

Proof. For l^2-loss, the optimality condition on a is given by

$$\mathbb{E}_\rho[(T(a_\mu^{(i)}) - y)h(x; w)] + \bar{\lambda}_a a_\mu^{(i)}(w) = 0.$$

Using $\mathbb{E}_\rho[y_i h(x; w)] = \mathbb{E}_{\rho_X}[f_i^\circ(x)h(x; w)] = T^* f_i^\circ$, we have

$$T^*T(a_\mu^{(i)}) + \bar{\lambda}_a a_\mu^{(i)} = T^* f_i^\circ.$$

Since $\bar{\lambda}_a > 0$ and $(T^*T + \bar{\lambda}_a \,\mathrm{Id})$ is invertible, we arrive at

$$a_\mu^{(i)} = (T^*T + \bar{\lambda}_a \,\mathrm{Id})^{-1}T^* f_i^\circ.$$

Since $(T^*T + \bar{\lambda}_a \,\mathrm{Id})T^* = T^*(TT^* + \bar{\lambda}_a \,\mathrm{Id})$, we have $T^*(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1} = (T^*T + \bar{\lambda}_a \,\mathrm{Id})^{-1}T^*$, and thus $a_\mu^{(i)} = T^*(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}f_i^\circ$. □

Lemma A.4. *Assume that l_i satisfies Assumption 3.4 or 3.3. Then, for any $w \in \mathbb{R}^{d'}$, $a_\mu^{(i)}(w)$ satisfies the following conditions:*

- $\left|a_\mu^{(i)}(w)\right| \leq \frac{c_l}{\bar{\lambda}_1} =: B_a$
- $\left\|\nabla_w a_\mu^{(i)}(w)\right\|_2 \leq \frac{c_l c_R}{\bar{\lambda}_1} =: R_a$
- $\left\|\nabla_w^2 a_\mu^{(i)}(w)\right\|_{\mathrm{op}} \leq \frac{c_l c_L}{\bar{\lambda}_1} =: L_a$

Proof. In the case of $|\partial_1 l_i(z,y)| \leq c_l$, Lemma A.2 yields

$$\begin{aligned}\left|a_\mu^{(i)}(w)\right| &= \frac{1}{\bar{\lambda}_a}|\mathbb{E}_\rho[\partial_1 l_i(f_i(x;a_\mu,\mu),y)h(x;w)]| \\ &= \frac{c_l}{\bar{\lambda}_a}.\end{aligned}$$

In the case of $l(z,y) = \frac{1}{2}(z-y)^2$ and $|y_i| \leq c_l$, Lemma A.3 yields

$$\begin{aligned}\left|a_\mu^{(i)}(w)\right| &= \left|T^*(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i\right| \\ &= \left|\int h(x;w)[(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i](x)\mathrm{d}\rho(x)\right| \\ &\leq \|h(x;w)\|_{L^2(\rho)}\left\|(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i\right\|_{L^2(\rho)} \\ &\leq \left\|(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i\right\|_{L^2(\rho)}.\end{aligned}$$

Here, the last inequality follows from $|h(x;w)| \leq 1$. Since the operator norm of $(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}$ is bounded by $1/\bar{\lambda}_a$, we have

$$\begin{aligned}\left\|(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i\right\|_{L^2(\rho)} &\leq \frac{1}{\bar{\lambda}_a}\|\bar{y}_i\|_{L^2(\rho)} \\ &\leq \frac{c_l}{\bar{\lambda}_a}.\end{aligned}$$

Thus, we have the first assertion.

In a similar way, we have

$$\begin{aligned}\left\|\nabla_w a_\mu^{(i)}(w)\right\|_2 &= \frac{1}{\bar{\lambda}_a}\|\mathbb{E}_\rho[\partial_1 l_i(f_i(x;a_\mu,\mu),y)\nabla_w h(x;w)]\| \\ &\leq \frac{c_l}{\bar{\lambda}_a}\mathbb{E}[\|\nabla_w h(x;w)\|] \\ &\leq \frac{c_l c_R}{\bar{\lambda}_a},\end{aligned}$$

in the case of $|\partial_1 l_i(z,y)| \leq c_l$. In addition, for a squared loss, we have

$$\begin{aligned}\left\|\nabla a_\mu^{(i)}(w)\right\|_2 &= \left\|\int h(x;w)[(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i](x)\mathrm{d}\rho(x)\right\|_2 \\ &\leq \left\|\|\nabla_w h(x;w)\|_2\right\|_{L^2(\rho)}\left\|(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i\right\|_{L^2(\rho)} \\ &\leq \frac{c_l c_R}{\bar{\lambda}_a}.\end{aligned}$$

Thus, we have the second assertion.

Furthermore, we have

$$\begin{aligned}\left\|\nabla_w^2 a_\mu^{(i)}(w)\right\|_{\mathrm{op}} &= \frac{1}{\bar{\lambda}_a}\left\|\mathbb{E}_\rho[\partial_1 l_i(f_i(x;a_\mu,\mu),y)\nabla_w^2 h(x;w)]\right\|_{\mathrm{op}} \\ &\leq \frac{c_l}{\bar{\lambda}_a}\mathbb{E}[\left\|\nabla_w^2 h(x;w)\right\|_{\mathrm{op}}] \\ &\leq \frac{c_l c_L}{\bar{\lambda}_a},\end{aligned}$$

in the case of $|\partial_1 l_i(z,y)| \leq c_l$. On the other hand, in the case of l_i is a squared loss, we have

$$\begin{aligned}\left\|\nabla_w^2 a_\mu^{(i)}(w)\right\|_{\mathrm{op}} &= \left\|\int \nabla_w^2 h(x;w)[(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i](x)\mathrm{d}\rho(x)\right\|_{\mathrm{op}} \\ &\leq \left\|\left\|\nabla_w^2 h(x;w)\right\|_{\mathrm{op}}\right\|_{L^2(\rho)}\left\|(TT^* + \bar{\lambda}_a \,\mathrm{Id})^{-1}\bar{y}_i\right\|_{L^2(\rho)} \\ &\leq \frac{c_l c_L}{\bar{\lambda}_a}.\end{aligned}$$

This completes the proof. □

Lemma A.5. *Assume that each l_i satisfies Assumption 3.4 or 3.3. Then, we have*

$$\begin{aligned}\left|\frac{\delta U}{\delta \mu}(\mu)(w)\right| &\leq \frac{\bar{\lambda}_a}{2}B_a^2 \\ \left\|\nabla_w \frac{\delta U}{\delta \mu}(\mu)(w)\right\|_2 &\leq \bar{\lambda}_a R_a B_a, \\ \left\|\nabla_w \nabla_w^\top \frac{\delta U}{\delta \mu}(\mu)(w)\right\|_{\mathrm{op}} &\leq \bar{\lambda}_a(R_a^2 + B_a L_a).\end{aligned}$$

for any $w \in \mathbb{R}^{d'}$.

Proof. From Theorem 3.1, we have

$$\frac{\delta U(\mu)}{\delta \mu}(w) = -\frac{\bar{\lambda}_a}{2T}\sum_{i=1}^T a_\mu^{(i)}(w)^2.$$

Thus, Lemma A.4 yields

$$\begin{aligned}\left|\frac{\delta U}{\delta \mu}(\mu)(w)\right| &\leq \frac{\bar{\lambda}_a}{2}B_a^2, \\ \left\|\nabla_w \frac{\delta U}{\delta \mu}(\mu)(w)\right\|_2 &\leq \bar{\lambda}_a B_a R_a, \\ \left\|\nabla_w \nabla_w^\top \frac{\delta U}{\delta \mu}(\mu)(w)\right\|_{\mathrm{op}} &\leq \bar{\lambda}_a(R_a^2 + B_a L_a),\end{aligned}$$

which completes the proof. □

Lemma A.6. *Assume that each l_i satisfies Assumption 3.4 or 3.3. We have*

$$\begin{aligned}\left|\frac{\delta^2 U}{\delta \mu^2}(\mu)(w,w')\right| &\leq B_a^2 \\ \left\|\nabla_w \frac{\delta^2 U}{\delta \mu^2}(\mu)(w,w')\right\|_2 &\leq 2R_a B_a, \\ \left\|\nabla_w \nabla_{w'} \frac{\delta^2 U}{\delta \mu^2}(\mu)(w,w')\right\|_{\mathrm{op}} &\leq 4R_a^2.\end{aligned}$$

for any $w, w' \in \mathbb{R}^{d'}$.

Proof. Let $\bar{l}_i''(x') := \mathbb{E}_\rho[\partial_1^2 l_i(f(x), y_i) \mid x = x']$. Define $\Lambda, \Lambda^{1/2} : L^2(\rho) \to L^2(\rho)$ by

$$\Lambda_i(f)(x) = f(x)\bar{l}_i''(x),$$
$$\Lambda_i^{1/2}(f)(x) = f(x)\bar{l}_i''(x)^{1/2},$$

and $A^{(i)}(w) \in L^2(\rho)$ by

$$A^{(i)}(w)(x) = a_\mu^{(i)}(w)h(x; w)\bar{l}''(x)^{1/2}$$

for a given $w \in \mathbb{R}^{d'}$. Note that $\Lambda, \Lambda^{1/2}.A$ are well-defined since $\bar{l}_i''(x) \geq 0$ from the convexity of l_i with respect to the first argument.

The second variation of $U(\mu)$ is given by

$$\frac{\delta}{\delta\mu}\frac{\delta U(\mu)}{\delta\mu}(w, w') = -\frac{\bar{\lambda}_a}{T}\sum_{i=1}^{T} a_\mu^{(i)}(w)\frac{\delta a_\mu^{(i)}(w)}{\delta\mu}(w').$$

Taking the first variation of the both sides of the optimality condition on $a_\mu^{(i)}(w)$ for a given w, we have

$$\begin{aligned}
&\mathbb{E}_\rho[l_i''(f(x; a_\mu, \mu), y)\left(a_\mu^{(i)}(w')h(x; w') + \int \frac{\delta a_\mu^{(i)}(w'')}{\delta\mu}(w')h(x; w'')\mathrm{d}\mu(w'')\right)h(x; w)] + \bar{\lambda}_a\frac{\delta a_\mu^{(i)}(w)}{\delta\mu}(w') \\
&= \left[(T^*\Lambda_i T + \bar{\lambda}_a \,\mathrm{Id})\frac{\delta a_\mu^{(i)}(\cdot)}{\delta\mu}(w')\right](w) + \left[T^*\Lambda_i^{1/2}A^{(i)}(w')\right](w) \\
&= 0.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\frac{\delta}{\delta\mu}\frac{\delta U(\mu)}{\delta\mu} &= -\frac{\bar{\lambda}_a}{T}\sum_{i=1}^{T} a_\mu^{(i)}(w)\frac{\delta a_\mu^{(i)}(w)}{\delta\mu}(w') \\
&= -\frac{\bar{\lambda}_a}{T}\sum_{i=1}^{T} a_\mu^{(i)}(w)(T^*\Lambda_i T + \bar{\lambda}_a \,\mathrm{Id})^{-1}\left[T^*\Lambda_i^{1/2}A^{(i)}(w')\right](w) \\
&= -\frac{\bar{\lambda}_a}{T}\sum_{i=1}^{T} a_\mu^{(i)}(w)\left[T^*\Lambda_i^{1/2}(\Lambda^{1/2}TT^*\Lambda_i^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}A^{(i)}(w')\right](w) \\
&= -\frac{\bar{\lambda}_a}{T}\sum_{i=1}^{T} a_\mu^{(i)}(w)\int\left[(\Lambda_i^{1/2}TT^*\Lambda_i^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}A^{(i)}(w')\right](x)h(x; w)\bar{l}''(x)^{1/2}\mathrm{d}\rho(x) \\
&= -\frac{\bar{\lambda}_a}{T}\sum_{i=1}^{T}\int\left[(\Lambda_i^{1/2}TT^*\Lambda_i^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}A(w')\right](x)A(w)(x)\mathrm{d}\rho(x) \\
&= -\frac{\bar{\lambda}_a}{T}\sum_{i=1}^{T}\langle A^{(i)}(w), (\Lambda_i^{1/2}TT^*\Lambda_i^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}A^{(i)}(w')\rangle
\end{aligned}$$

The second equality follows from the equality $A(A^*A + \mathrm{Id})^{-1} = (AA^* + \mathrm{Id})^{-1}A$ for any operator A such that $(A^*A + \mathrm{Id}), (AA^* + \mathrm{Id})$ are invertible.

First, we bound $\left\|\left\|\nabla_w A^{(i)}(w)\right\|_2\right\|_{L^2(\rho)}$ and $\left\|A^{(i)}(w)\right\|_{L^2(\rho)}$. From Lemma A.4, we have

$$\begin{aligned}
\left\|\left\|\nabla_w A^{(i)}(w)\right\|_2\right\|_{L^2(\rho)} &= \left\|\left\|\nabla_w a_\mu^{(i)}(w) h(x;w) + a_\mu^{(i)}(w)\nabla_w h(x;w)\right\|_2 \bar{l}''(x)^{1/2}\right\|_{L^2(\rho)} \\
&\le \left\|\left\|\nabla_w a_\mu^{(i)}(w)\right\|_2 \bar{l}''(x)^{1/2}\right\|_{L^2(\rho)} + \left\|\left|a_\mu^{(i)}(w)\right| \|\nabla h(x;w)\|_2 \bar{l}''(x)^{1/2}\right\|_{L^2(\rho)} \\
&\le 2R_a, \\
\left\|A^{(i)}(w)\right\|_{L^2(\rho)} &= \left\|a_\mu^{(i)}(w) h(x;w)\bar{l}''(x)^{1/2}\right\|_{L^2(\rho)} \\
&\le B_a
\end{aligned}$$

since we have assumed that $0 \le \partial_1^2 l_i(x,y) \le 1$. The first assertion follows immediately from the above inequality and $\left\|(\Lambda_i^{1/2} T T^* \Lambda_i^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}\right\|_{\mathrm{op}} \le 1/\bar{\lambda}_a$.

Then, $\left\|\nabla_w \frac{\delta^2 U(\mu)}{\delta \mu^2}(w,w')\right\|$ is bounded as follows:

$$\begin{aligned}
\left\|\nabla_w \frac{\delta^2 U(\mu)}{\delta \mu^2}(w,w')\right\| &\le \frac{\bar{\lambda}_a}{T}\sum_{i=1}^T \int \mathrm{d}\rho(x) \left\|\nabla_w A^{(i)}(w)(x)\right\|_2 \left|[(\Lambda^{1/2} T T^* \Lambda^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1} A^{(i)}(w')](x)\right| \\
&\le \frac{\bar{\lambda}_a}{T}\sum_{i=1}^T \left\|\left\|\nabla_w A^{(i)}(w)(x)\right\|_2\right\|_{L^2(\rho)} \left\|(\Lambda^{1/2} T T^* \Lambda^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1} A^{(i)}(w')\right\|_{L^2(\rho)} \\
&\le \frac{\bar{\lambda}_a}{T}\sum_{i=1}^T \left\|\left\|\nabla_w A^{(i)}(w)(x)\right\|_2\right\|_{L^2(\rho)} \left\|(\Lambda^{1/2} T T^* \Lambda^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}\right\|_{\mathrm{op}} \left\|A^{(i)}(w')\right\|_{L^2(\rho)} \\
&\le 2R_a B_a
\end{aligned}$$

Finally, $\left\|\nabla_w \nabla_{w'}^\top \frac{\delta^2 U(\mu)}{\delta \mu^2}(w,w')\right\|_2$ is bounded as follows:

$$\begin{aligned}
\left\|\nabla_w \nabla_{w'}^\top \frac{\delta^2 U(\mu)}{\delta \mu^2}(w,w')\right\|_{\mathrm{op}} &\le \frac{\bar{\lambda}_a}{T}\sum_{i=1}^T \mathrm{tr}\Big[\int\int \mathrm{d}\rho(x)\mathrm{d}\rho(x')\nabla_w A^{(i)}(w)(x) \\
&\qquad (\Lambda^{1/2} T T^* \Lambda^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}(x,x')\nabla_{w'}^\top A^{(i)}(w')(x')\Big] \\
&\le \frac{\bar{\lambda}_a}{T}\sum_{i=1}^T \int\int \mathrm{d}\rho(x)\mathrm{d}\rho(x') \left\|\nabla_w A^{(i)}(w)(x)\right\| \\
&\qquad (\Lambda^{1/2} T T^* \Lambda^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}(x,x') \left\|\nabla_{w'}^\top A^{(i)}(w')(x')\right\| \\
&\le \frac{\bar{\lambda}_a}{T}\sum_{i=1}^T \left\|\left\|\nabla_w A^{(i)}(w)(x)\right\|\right\|_{L^2(\rho)}^2 \left\|(\Lambda^{1/2} T T^* \Lambda^{1/2} + \bar{\lambda}_a \,\mathrm{Id})^{-1}\right\|_{L^2(\rho)} \\
&\le 4R_a^2.
\end{aligned}$$

□

B. Proofs for Section 3

B.1. Proof of Theorem 3.1

First, we derive the expression of the first variation of $G(\mu)$. The envelope theorem implies that

$$\frac{\delta G(\mu)}{\delta \mu}(w) = \frac{\partial F(a_\mu, \mu)}{\partial \mu}(w) = \frac{\partial L(a_\mu, \mu)}{\partial \mu}(w) + \lambda r(a_\mu(w), w).$$

In addition, the first variations of L w.r.t. μ and a are given by

$$\frac{\partial L(a_\mu, \mu)}{\partial \mu} = \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_\rho[\partial_1 l_i(f_i(x; a, \mu), y)h(x; w)]a_\mu^{(i)}(w),$$
$$\frac{\partial L(a_\mu, \mu)}{\partial a^{(i)}} = \frac{1}{T}\mathbb{E}_\rho[\partial_1 l_i(f_i(x; a, \mu), y)h(x; w)]\mu(w),$$

respectively. Therefore, we have

$$\frac{\partial L(a_\mu, \mu)}{\partial \mu}(w) = \sum_{i=1}^{T} \frac{a_\mu^{(i)}(w)}{\mu(w)} \frac{\partial L(a_\mu, \mu)}{\partial a^{(i)}}(w).$$

The first-order optimality condition on a_μ yields

$$\frac{\partial L(a_\mu, \mu)}{\partial a^{(i)}}(w) = -\lambda \frac{\partial r(a_\mu(w), w)}{\partial a^{(i)}}\mu(w), \tag{4}$$

which implies

$$\begin{aligned}\frac{\partial L(a_\mu, \mu)}{\partial \mu}(w) &= -\lambda \sum_{i=1}^{T} \frac{\partial r(a_\mu(w), w)}{\partial a^{(i)}} a_\mu^{(i)}(w) \\ &= -\lambda \langle \boldsymbol{\nabla}_a r(a_\mu(w), w), a_\mu(w)\rangle.\end{aligned}$$

Combining above arguments, we arrive at

$$\begin{aligned}\frac{\delta G(\mu)}{\delta \mu}(w) &= -\lambda \langle \boldsymbol{\nabla}_a r(a_\mu(w), w), a_\mu(w)\rangle + \lambda r(a_\mu(w), w) \\ &= \lambda\left(-\frac{\lambda_a}{2T}\|a_\mu(w)\|_2^2 + \frac{\lambda_w}{2}\|w\|_2^2\right).\end{aligned} \tag{5}$$

Next, we prove the convexity of $G(\mu)$. From the convexity of $L(a, \mu)$ w.r.t. a, we have

$$L(a_{\mu_1}, \mu_1) + \int \sum_{i=1}^{T} \frac{\partial L(a_{\mu_1}, \mu_1)}{\partial a^{(i)}}(w)\left(\frac{\mu_2(w)}{\mu_1(w)}a_{\mu_2}^{(i)}(w) - a_{\mu_1}^{(i)}(w)\right)\mathrm{d}w \le L\left(\frac{\mu_2(w)}{\mu_1(w)}a_{\mu_2}, \mu_1\right) = L(a_{\mu_2}, \mu_2)$$

for any $\mu_1, \mu_2 \in \mathcal{P}$. Therefore, it holds that

$$\begin{aligned}&G(\mu_1) + \int \sum_{i=1}^{T} \frac{\partial L(a_{\mu_1}, \mu_1)}{\partial a^{(i)}}(w)\left(\frac{\mu_2(w)}{\mu_1(w)}a_{\mu_2}^{(i)}(w) - a_{\mu_1}^{(i)}(w)\right)\mathrm{d}w \\ &+ \lambda r(a_{\mu_2}(w), w)\mu_2(w) - \lambda r(a_{\mu_1}(w), w)\mu_1(w)\mathrm{d}w \\ &\le G(\mu_2).\end{aligned}$$

Thus, it is sufficient to show that

$$\begin{aligned}\frac{\delta G(\mu_1)}{\delta \mu}(w)(\mu_2(w) - \mu_1(w)) \le &\sum_{i=1}^{T} \frac{\partial L(a_{\mu_1}, \mu_1)}{\partial a^{(i)}}(w)\left(\frac{\mu_2(w)}{\mu_1(w)}a_\mu^{(i)}(w) - a_{\mu_1}^{(i)}(w)\right)\mathrm{d}w \\ &+ \lambda r(a_{\mu_2}(w), w)\mu_2(w) - \lambda r(a_{\mu_1}(w), w)\mu_1(w)\end{aligned}$$

for any w. To simplify the notation, we denote the LHS by $\rho_1(w)$ and the RHS by $\rho_2(w)$. Substituting Eq. (5) to $\rho_1(w)$, we have

$$\rho_1(w) = \lambda[-\langle \boldsymbol{\nabla}_a r(a_{\mu_1}(w), w), a_{\mu_1}(w)\rangle + r(a_{\mu_1}(w), w)]\,(\mu_2(w) - \mu_1(w)).$$

On the other hand, substituting Eq. (4) to $\rho_2(w)$, we have

$$\rho_2(w) = -\lambda\langle \nabla_a r(a_{\mu_1}(w), w)(\mu_2(w)a_{\mu_2}(w) - a_{\mu_1}(w)\mu_1(w))\rangle + \lambda r(a_{\mu_2}(w), w)\mu_2(w) - \lambda r(a_{\mu_1}(w), w)\mu_1(w).$$

Therefore,

$$\begin{aligned}\rho_2(w) - \rho_1(w) &= \lambda\mu_2(w)[\langle -\nabla_a r(a_{\mu_1}(w), w)(a_{\mu_2}(w) - a_{\mu_1}(w))\rangle + r(a_{\mu_2}(w), w) - r(a_{\mu_1}(w), w)] \\ &\geq 0.\end{aligned}$$

The last inequality follows from the convexity of $r(a, w)$ w.r.t. a. This completes the proof.

B.2. Proof of Lemma 3.5

Using Eq. (3), we have

$$p_\mu(w) \propto \exp\left(-\frac{1}{\lambda}\frac{\delta G}{\delta \mu}\right) = \exp\left(\frac{\lambda_a}{2T}\|a(w)\|_2^2 - \frac{\lambda_w}{2}\|w\|_2^2\right).$$

The distribution $p_0(w) \propto \exp\left(-\frac{\lambda_w}{2}\|w\|_2^2\right)$ satisfies the LSI with constant $\alpha = \lambda_w$ since $\frac{\lambda_w}{2}\|w\|_2^2$ is λ_w-strongly convex (Bakry & Émery, 1985). Thus, Lemma A.1 implies that p_μ satisfies the LSI with constant $\alpha = \frac{\lambda_w}{\exp\left(4\|B\|_\infty\right)}$, where $B(w) = \frac{\lambda_a}{2T}\|a_\mu(w)\|_2^2$. From Lemma A.4, we have $\left|a_\mu^{(i)}(w)\right| \leq \frac{c_l}{\lambda_a}$ for any w. Therefore, $\|B(w)\|_\infty \leq \frac{c_l^2\lambda_a}{2\lambda_a^2}$. This completes the proof.

B.3. Proof of Theorem 3.7

The continuous time result follows from Lemma 3.5, Theorem 3.1, and the result in Nitanda et al. (2022).

For the discretized time result, we follow the framework in Suzuki et al. (2023a). First, we prove the following lemma.

Lemma B.1. *For any $w \in \mathbb{R}^{d'}$, $U(\mu)$ satisfies the following conditions:*

- $\left\|\nabla\frac{\delta U}{\delta\mu}(\mu)(w) - \nabla\frac{\delta U}{\delta\mu}(\mu')(w')\right\| \leq L_U(W_2(\mu,\mu') + \|w - w'\|)$, $\left|\frac{\delta^2 U(\mu)}{\delta\mu^2}(w, w')\right| \leq L_U$ *for* $L_U = 4R_a^2 + \bar{\lambda}_a(B_aL_a + R_a^2) + B_a^2$.
- $\left\|\nabla\frac{\delta U}{\delta\mu}(\mu)(w)\right\| \leq R_U$ *for* $R_U = \bar{\lambda}_a B_a R_a$.

Proof. From Lemma A.5 and A.6, we have

$$\begin{aligned}\left\|\nabla_w^2\frac{\delta U(\mu)}{\delta\mu}(w)\right\|_{\mathrm{op}} &\leq \bar{\lambda}_a(B_aL_a + R_a^2), \\ \left|\frac{\delta^2 U}{\delta\mu^2}(w, w')\right| &\leq B_a^2\end{aligned}$$

for any $w, w' \in \mathbb{R}^{d'}$. Thus, $\frac{\delta U(\mu)}{\delta\mu}(w)$ is $\bar{\lambda}_a(B_aL_a + R_a^2)$-smooth and it holds that

$$\left\|\nabla\frac{\delta U}{\delta\mu}(\mu)(w) - \nabla\frac{\delta U}{\delta\mu}(\mu)(w')\right\|_2 \leq \bar{\lambda}_a(B_aL_a + R_a^2)\|w - w'\|_2.$$

Let $\mu_t = t\mu + (1-t)\mu'$. Then, we have

$$\begin{aligned}\left\|\nabla\frac{\delta U}{\delta\mu}(\mu)(w) - \nabla\frac{\delta U}{\delta\mu}(\mu')(w)\right\|_2 &\leq \int_0^1 \left\|\nabla_w\frac{\mathrm{d}}{\mathrm{d}t}\frac{\delta U}{\delta\mu}(\mu_t)(w)\right\|_2 \mathrm{d}t \\ &= \int_0^1 \left\|\int \nabla_w\frac{\delta^2 U}{\delta\mu^2}(\mu_t)(w, w')(\mu'(w') - \mu(w'))\mathrm{d}w'\right\|_2 \mathrm{d}t\end{aligned}$$

From the definition of W_1, for any $\varepsilon > 0$, there exists a coupling π of μ and μ' such that

$$\int \|w - w'\|_2 \mathrm{d}\pi(w, w') \leq W_1(\mu, \mu') + \varepsilon \leq W_2(\mu, \mu') + \varepsilon.$$

Thus, we have

$$\begin{aligned}
\left\| \int \nabla_w \frac{\delta^2 U}{\delta \mu^2}(\mu_t)(w, w')(\mu'(w') - \mu(w'))\mathrm{d}w' \right\|_2 &= \left\| \int \left(\nabla_w \frac{\delta^2 U}{\delta \mu^2}(\mu_t)(w, w'') - \nabla_w \frac{\delta^2 U}{\delta \mu^2}(\mu_t)(w, w') \right) \mathrm{d}\pi(w, w') \right\|_2 \\
&\leq \int \left\| \nabla_w \frac{\delta^2 U}{\delta \mu^2}(\mu_t)(w, w'') - \nabla_w \frac{\delta^2 U}{\delta \mu^2}(\mu_t)(w, w') \right\|_2 \mathrm{d}\pi(w, w') \\
&\leq \int 4R_a^2 \|w'' - w'\|_2 \mathrm{d}\pi(w, w') \\
&\leq 4R_a^2 W_2(\mu, \mu') + 4R_a^2 \varepsilon.
\end{aligned}$$

The second inequality follows from the Lipschitz continuity of $\nabla_w \frac{\delta^2 U}{\delta \mu^2}(\mu_t)(w, w')$ from Lemma A.6. Since ε is arbitrary, we have

$$\begin{aligned}
\left\| \nabla \frac{\delta U}{\delta \mu}(\mu)(w) - \nabla \frac{\delta U}{\delta \mu}(\mu')(w) \right\| &\leq \int_0^1 \left\| \int \nabla_w \frac{\delta^2 U}{\delta \mu^2}(\mu_t)(w, w')(\mu'(w') - \mu(w'))\mathrm{d}w' \right\| \mathrm{d}t \\
&\leq 4R_a^2 W_2(\mu, \mu').
\end{aligned}$$

This completes the proof. □

Combining the above lemma and Theorem 3 in Suzuki et al. (2023a), we have for any $\lambda\alpha\eta \leq 1/2$ and $\eta \leq 1/4$,

$$\frac{1}{N}\mathbb{E}[\mathcal{G}^N(\mu_k^{(N)})] - \mathcal{G}(\mu^*) \leq \exp(-\lambda\alpha\eta k)\left(\frac{1}{N}\mathbb{E}[\mathcal{G}^N(\mu_k^{(N)})] - \mathcal{G}(\mu^*) \right) + \frac{2}{\lambda\alpha}\bar{L}^2 C_1(\lambda\eta + \eta^2) + \frac{2C_\lambda}{\lambda\alpha N},$$

where $\bar{R}^2 = \mathbb{E}\left[\left\| w_i^{(0)} \right\|_2 \right] + \frac{1}{\bar{\lambda}_w}\left[\left(\frac{1}{4} + \frac{1}{\bar{\lambda}_w} \right) R_U^2 + \lambda d' \right]$, $\bar{L} = L_U + \bar{\lambda}_w$, $C_1 = 8[R_U^2 + \bar{\lambda}_w \bar{R}^2 + d']$, and $C_\lambda = 2\lambda L_U \alpha + 2\lambda^2 L_U^2 \bar{R}^2$.

B.4. Proof of Proposition 3.8

As shown in Lemma 3 in Suzuki et al. (2023a), we have

$$W_2^2(\mu_k^{(N)}, \mu^{*N}) \leq \frac{2}{\lambda\alpha}(\mathcal{G}^N(\mu_k^{(N)}) - N\mathcal{G}(\mu^*)).$$

Let γ be a coupling of $\mu_k^{(N)}$ and μ^{*N}. Then, for $(W, W^*) \sim \gamma$, we have

$$(k(x, y)_{\mu_W} - k(x, y)_{\mu^*})^2 \leq 2(k_{\mu_W}(x, y) - k_{\mu_{W^*}}(x, y))^2 + 2(k_{\mu_{W^*}}(x, y) - k_{\mu^*}(x, y))^2.$$

Let $k(x, y; w) := h(x; w)h(y; w)$. This is $2c_R$-Lipschitz continuous with respect to w for any $(x, y) \in S \times S$. Then, for the first term in the right hand side, we have

$$\begin{aligned}
(k_{\mu_W}(x, y) - k_{\mu_{W^*}}(x, y))^2 &\leq \left(\frac{1}{N}\sum_{i=1}^N k(x, y; w_i) - k(x, y; w_i^*) \right)^2 \\
&\leq \frac{1}{N}\sum_{i=1}^N (k(x, y; w_i) - k(x, y; w_i^*))^2 \\
&\leq \frac{1}{N}\sum_{i=1}^N 4c_R^2 \|w_i - w_i^*\|_2^2 \\
&\leq \frac{1}{N} 4c_R^2 \|W - W^*\|_{\mathrm{F}}^2
\end{aligned}$$

for any $(x, y) \in S \times S$. For the second term, we have

$$P\left(\sup_{(x,y)\in S\times S}|k_{\mu_{W^*}}(x,y)-k_{\mu^*}(x,y)| \geq 2\mathbb{E}_{\sigma,W^*}\left[\sup_{(x,y)\in S\times S}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i k(x,y;w_i)\right|+\sqrt{\frac{2t}{N}}\right]\right) \leq \exp(-t)$$

for any $t > 0$ by the same argument in Lemma 2 in Suzuki et al. (2023b). From the relation between Gaussian complexity and Rademacher complexity, and contraction inequality in Bartlett & Mendelson (2001), we have

$$\begin{aligned}\mathbb{E}\left[\sup_{(x,y)\in S\times S}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i k(x,y;w_i)\right|\right] &\leq cc_R\mathbb{E}\left[\sup_{(x,y)\in S\times S}\left|\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d'}\varepsilon_{ij}w_{ij}\right|\right]\\ &\leq \mathbb{E}\left[\frac{cc_R}{N}\sqrt{\sum_{i=1}^{N}\|w_i\|_2^2}\right]\\ &\leq \frac{cc_R}{\sqrt{N}}\sqrt{\mathbb{E}_\mu[\|w\|_2^2]},\end{aligned}$$

where ε_{ij} is a Gaussian random variable with mean 0 and variance 1, and c is a universal constant. From the optimality of μ^*, we have $\frac{\bar{\lambda}_w}{2}\mathbb{E}_{\mu^*}[\|w\|_2^2] \leq \mathcal{G}(\mu^*) \leq \mathcal{G}(\mu_0)$. Thus, we have

$$\mathbb{E}\left[\sup_{(x,y)\in S\times S}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i k(x,y;w_i)\right|\right] \leq \frac{cc_R}{\sqrt{N}}\sqrt{\frac{2\mathcal{G}(\mu_0)}{\bar{\lambda}_w}}$$

and

$$P\left[\left(\sup_{(x,x')\in S\times S}|k_{\mu_{W^*}}(x,y)-k_{\mu^*}(x,y)|\right)^2-\frac{4c^2\mathcal{G}(\mu_0)c_R^2}{\bar{\lambda}_w N}\geq\frac{4t}{N}\right]\leq\exp(-t)$$

Thus, it holds that

$$\begin{aligned}&\mathbb{E}\left[\left(\sup_{(x,x')\in S\times S}|k_{\mu_{W^*}}(x,y)-k_{\mu^*}(x,y)|\right)^2\right]-\frac{4c^2\mathcal{G}(\mu_0)c_R^2}{\bar{\lambda}_w N}\\ &\leq\int_0^\infty P\left[\left(\sup_{(x,x')\in S\times S}|k_{\mu_{W^*}}(x,y)-k_{\mu^*}(x,y)|\right)^2-\frac{4c^2\mathcal{G}(\mu_0)c_R^2}{\bar{\lambda}_w N}\geq\tau\right]\mathrm{d}\tau\\ &\leq\int_0^\infty\exp(-N\tau/4)\mathrm{d}t=\frac{4}{N}.\end{aligned}$$

Combining above arguments, we arrive at

$$\mathbb{E}_\gamma\left[\sup_{(x,y)\in S\times S}(k(x,y)_{\mu_W}-k(x,y)_{\mu^*})^2\right]\leq\frac{8}{N}c_R^2\mathbb{E}_\gamma[\|W-W^*\|_{\mathrm{F}}^2]+\frac{4}{N}+\frac{4c^2\mathcal{G}(\mu_0)c_R^2}{\bar{\lambda}_w N}.$$

By taking the infimum of the coupling γ, we have

$$\mathbb{E}_{W\sim\mu_k^{(N)}}\left[\sup_{(x,y)\in S\times S}(k(x,y)_{\mu_W}-k(x,y)_{\mu^*})^2\right]\leq\frac{2}{N}c_R^2W_2^2(\mu_k^{(N)},\mu^{*N})+\frac{4}{N}+\frac{4c^2\mathcal{G}(\mu_0)c_R^2}{\bar{\lambda}_w N}=O\left(\frac{\Delta}{N}\right).$$

For any $f \in L^2(\rho_X)$, we have

$$\begin{aligned}|(\Sigma_{\mu_W}-\Sigma_{\mu^*})(f)(x)| &= \int(k_{\mu_W}(x,x')-k_{\mu^*}(x,x'))f(x')\mathrm{d}\rho_X(x')\\ &\leq\sqrt{\int(k_{\mu_W}(x,x')-k_{\mu^*}(x,x'))^2\mathrm{d}\rho_X(x')}\|f\|_{L^2(\rho_X)},\end{aligned}$$

which yields

$$\|(\Sigma_{\mu_W} - \Sigma_{\mu^*})(f)\|_{L^2(\rho_X)}^2 \leq \|f\|_{L^2(\rho_X)}^2 \int (k_{\mu_W}(x,x') - k_{\mu^*}(x,x'))^2 \mathrm{d}\rho_X(x')\mathrm{d}\rho_X(x).$$

This implies $\|\Sigma_{\mu_W} - \Sigma_{\mu^*}\|_{\mathrm{op}}^2 \leq \mathbb{E}_{x,x'}[(k_{\mu_W}(x,x') - k_{\mu^*}(x,x'))^2] \leq \Delta/N$. Since $f_i(x;\mu)$ is the optimal solution of $\min_{f\in\mathcal{H}_\mu} \mathbb{E}_\rho[l_i(f(x),y)] + \frac{\bar{\lambda}_a}{2}\|f\|_{\mathcal{H}_\mu}$, where l_i is the squared loss, $f_i(x;\mu) = \Sigma_\mu(\Sigma_\mu + \bar{\lambda}_a \mathrm{Id})^{-1}\bar{y} = \int k(x,x')\alpha(x')\mathrm{d}\rho_X(x')$, where $\alpha_\mu(x) = (\Sigma_\mu + \bar{\lambda}_a \mathrm{Id})^{-1}\bar{y}$. From the identity $A^{-1} - A'^{-1} = -A^{-1}(A - A')A'^{-1}$ for any invertible operator A, A', we have

$$\begin{aligned}
\|\alpha_{\mu_W} - \alpha_{\mu^*}\|_{\rho_X}^2 &= \left\|(\Sigma_{\mu_W} + \bar{\lambda}_a \mathrm{Id})^{-1}(\Sigma_{\mu_W} - \Sigma_{\mu^*})(\Sigma_{\mu^*} + \bar{\lambda}_a \mathrm{Id})^{-1}\bar{y}\right\|_{L^2(\rho_X)}^2 \\
&\leq \frac{c_0^2\|\Sigma_{\mu_W} - \Sigma_{\mu^*}\|_{\mathrm{op}}^2}{\bar{\lambda}_a^4}.
\end{aligned}$$

Thus, for any $x \in S$, we have

$$\begin{aligned}
(f_i(x;\mu_W) - f_i(x;\mu^*))^2 &\leq 2\left(\int k_{\mu'_W}(x,x')\alpha_{\mu_W}(x') - k_{\mu^*}(x,x')\alpha_{\mu_W}(x')\mathrm{d}\rho_X(x')\right)^2 \\
&\quad + 2\left(\int k_{\mu^*}(x,x')\alpha_{\mu_W}(x') - k_{\mu^*}(x,x')\alpha_{\mu^*}(x')\mathrm{d}\rho_X(x')\right)^2 \\
&\leq 2\sup_{x,x'\in S\times S}(k_{\mu_W}(x,x') - k_{\mu^*}(x,x'))^2\|\alpha_{\mu_W}\|_{L^2(\rho_X)}^2 \\
&\quad + 2\|\alpha_{\mu_W} - \alpha_{\mu^*}\|_{\rho_X}^2 \\
&\leq \frac{2c_0^2\sup_{x,x'\in S\times S}(k_{\mu_W}(x,x') - k_{\mu^*}(x,x'))^2}{\bar{\lambda}_a^2} + \frac{2c_0^2\|\Sigma_{\mu_W} - \Sigma_{\mu^*}\|_{\mathrm{op}}^2}{\bar{\lambda}_a^4}.
\end{aligned}$$

By taking the supremum over x and expectation over W, we obtain the result.

C. Proofs for Section 4

C.1. Lemmas for Section 4

Lemma C.1. *For a given $\delta, \tau > 0$, we have*

$$|\tanh(\tau z) - (1[z \geq 0] - 1[z < 0])| \leq 2e^{-2\tau|z|}$$

for any $z \in \mathbb{R}$.

Proof. From the definition of $\tanh$, we have, for any $z \geq 0$,

$$\begin{aligned}
|\tanh(\tau z) - (1[z \geq 0] - 1[z < 0])| &= 1 - \frac{e^{\tau z} - e^{-\tau z}}{e^{\tau z} + e^{-\tau z}} \\
&= \frac{2e^{-\tau z}}{e^{\tau z} + e^{-\tau z}} \\
&\leq 2e^{-2\tau z}.
\end{aligned}$$

Similarly, for $z < 0$, we have

$$\begin{aligned}
|\tanh(\tau z) - (1[z \geq 0] - 1[z < 0])| &= 1 + \frac{e^{\tau z} - e^{-\tau z}}{e^{\tau z} + e^{-\tau z}} \\
&= \frac{2e^{\tau z}}{e^{\tau z} + e^{-\tau z}} \\
&\leq 2e^{2\tau z}.
\end{aligned}$$

Combining above arguments, we obtain the result. □

Lemma C.2. *Let ρ_X be the uniform distribution on $[0,1]^d$ and $\mathcal{S} := \left\{\sin(2\pi w \cdot x) \mid w \in \{0,1\}^d, \|w\|_1 = k\right\}$ be a subset of $L^2(\rho_X)$. Furthermore, for any fixed basis functions $\{h_j\}_{j=1}^n \subset L^2(\rho_X)$, let H_n be a span of $\{h_j\}_{j=1}^n$. Then, for any $\varepsilon \in [0,1]$, we have*

$$\sup_{\psi \in \mathcal{S}} d(\psi, H_n) \geq \varepsilon$$

if $n \leq N(1-\varepsilon)$, where $N = |\mathcal{S}| = \binom{d}{k}$.

Proof. Assume that $d(\psi, H_n) < \varepsilon$ for any $\psi \in \mathcal{S}$. From Theorem 1 in Hsu (2021), we have $n > N(1-\varepsilon)$ since $\mathcal{S}$ is an orthonormal system in $L^2(\rho_X)$ and $|\mathcal{S}| = N$. This contradicts $n \leq N(1-\varepsilon)$, which completes the proof. □

Lemma C.3. *For $\varepsilon, r, r_x > 0$, let $\lambda_w = 1$, $h(x; w) = \tanh(x \cdot u + b)$ $(w = (u, b))$. Then, for any $u^\circ \in \mathbb{R}^d$ and $\tilde{f} : \mathbb{R} \to \mathbb{R}$ which is 1-Lipschitz continuous and differentiable almost everywhere, there exists μ, a such that $\mathrm{KL}(\nu \mid \mu) = O\left(\frac{r^2}{\varepsilon^2} + d \log \frac{d r r_x}{\varepsilon}\right)$, $\|a\|_{L^2(\mu)} = r$ and*

$$\left| \tilde{f}(w^\circ \cdot x) - \left[f(x; a, \mu) + \frac{1}{2}(\tilde{f}(r) + \tilde{f}(-r)) \right] \right| \leq \varepsilon$$

for any $x \in \mathbb{R}^d$ such that $|u^\circ \cdot x| \leq r$ and $\|x\| \leq r_x \sqrt{d}$.

Proof. Let $a(u,b) = r\tilde{a}(b/\tau)$ for $\tilde{a}(b) : \mathbb{R} \to \mathbb{R}, \|\tilde{a}\|_\infty \leq 1$ and $\mu(w) = \mu(u,b) := \mu(u)\mu(b)$, where $\mu(u) = N(\tau u^\circ, \sigma^2 I), \mu(b) = \upsilon([-\tau r, \tau r])$ for $\tau, \sigma > 0$. In addition, let $\bar{g}(x) = \mathbb{E}_{b \sim \mu_\tau}[r\tilde{a}(b/\tau)\tanh(\tau x \cdot w^\circ + b)]$. Then, we have

$$\begin{aligned}
|\bar{g}(x) - f(x; a, \mu)| &\leq \int |r\tilde{a}(b/\tau)||\tanh(x \cdot u + b) - \tanh(\tau x \cdot u^\circ + b)| \mathrm{d}\mu(\tilde{u}, b) \\
&\leq \int |r\tilde{a}(b/\tau)||x \cdot u - \tau x u^\circ| \mathrm{d}\mu(\tilde{u}, b) \\
&\leq r \sqrt{\int |x \cdot (u - \tau u^\circ)|^2 \mathrm{d}\mu(u)} \\
&\leq r \sqrt{\int \|x\|^2 \|u - \tau u^\circ\|^2 \mathrm{d}\mu(u)} \\
&\leq r r_x \sqrt{d} \sigma.
\end{aligned}$$

Let $\tilde{g}(x; a) := \int_{-k}^{0} \frac{1}{2}\tilde{a}(b')(1[u^\circ \cdot x + b' \geq 0] - 1[u^\circ \cdot x + b' < 0]) \mathrm{d}b'$. Since

$$\begin{aligned}
\bar{g}(x; a) &= \int r\tilde{a}(b/\tau)\tanh(\tau x \cdot u^\circ + b) \mathrm{d}\mu(b) \\
&= \int_{-\tau r}^{\tau r} \frac{1}{2\tau}\tilde{a}(b/\tau)\tanh(\tau x \cdot u^\circ + b) \mathrm{d}b \\
&= \int_{-r}^{r} \frac{1}{2}\tilde{a}(b')\tanh(\tau(x \cdot u^\circ + b')) \mathrm{d}b',
\end{aligned}$$

it holds that

$$\begin{aligned}
|\bar{g}(x) - \tilde{g}(x)| &\leq \int_{-r}^{r} \frac{1}{2}|\tilde{a}(b)||\tanh(\tau(x \cdot u^\circ + b')) - (1[x \cdot u^\circ + b' \geq 0] - 1[x \cdot u^\circ + b' < 0])| \mathrm{d}b' \\
&\leq \int_{-\infty}^{\infty} \frac{1}{2}|\tanh(\tau(x \cdot u^\circ + b')) - (1[x \cdot u^\circ + b' \geq 0] - 1[x \cdot u^\circ + b' < 0])| \mathrm{d}b' \\
&\leq \int_{0}^{\infty} e^{-2\tau z} \mathrm{d}z \\
&= 1/(2\tau)
\end{aligned}$$

where we used Lemma C.1 for the last inequality. Since $-r \le u^\circ \cdot x \le r$, we have

$$\begin{aligned}\tilde{g}(x) &= \frac{1}{2}\left[\int_{-r}^{r} \tilde{a}(b')1[u^\circ \cdot x + b' \ge 0]\mathrm{d}b' - \int_{-r}^{r} \tilde{a}(b')1[u^\circ \cdot x + b' < 0]\mathrm{d}b'\right] \\ &= \frac{1}{2}\left[\int_{-u^\circ \cdot x}^{r} \tilde{a}(b')\mathrm{d}b' - \int_{-r}^{-u^\circ \cdot x} \tilde{a}(b')\mathrm{d}b'\right].\end{aligned}$$

By letting

$$\tilde{a}(b) = \begin{cases} \tilde{f}'(-b) & \text{if } b \in [-r, r] \\ 0 & \text{otherwise} \end{cases},$$

we obtain

$$\begin{aligned}\tilde{g}(x;\tilde{a}) &= \frac{1}{2}\left[\int_{-w^\circ \cdot x}^{r} \tilde{f}'(-b')\mathrm{d}b' - \int_{-r}^{-w^\circ \cdot x} \tilde{f}'(-b')\mathrm{d}b'\right] \\ &= \frac{1}{2}\left[\int_{-r}^{w^\circ \cdot x} \tilde{f}'(b')\mathrm{d}b' - \int_{w^\circ \cdot x}^{r} \tilde{f}'(b')\mathrm{d}b'\right] \\ &= \tilde{f}(w^\circ \cdot x) - \frac{1}{2}\Big[\tilde{f}(r) + \tilde{f}(-r)\Big].\end{aligned}$$

Combining above results, we have

$$\begin{aligned}\left|\tilde{f}(w^\circ \cdot x) - f(x;a,\mu)\right| &\le \left|\tilde{f}(w^\circ \cdot x) - \bar{g}(x)\right| + |\bar{g}(x) - \tilde{g}(x)| \\ &\le \frac{r r_x \sqrt{d}\sigma}{2} + \frac{1}{2\tau} \\ &\le \varepsilon\end{aligned}$$

by letting $\tau = 1/\varepsilon, \sigma = \varepsilon/(r r_x \sqrt{d})$.

Finally, we show that $f(x;a,\mu)$ is in $\mathcal{B}_M$ and $\|f\|_{\mathcal{B}_M} \le R$. Since u, b is independent each other when $(u,b) \sim \mu$, we have

$$\mathrm{KL}(\nu \mid \mu) = \mathrm{KL}\left(N(0,I) \mid N(\tau u^\circ, \sigma^2 I)\right) + \mathrm{KL}\left(N(0,1) \mid u([-r\tau, r\tau])\right).$$

For the first term, we have

$$\begin{aligned}\mathrm{KL}\left(N(0,I) \mid N(\tau u^\circ, \sigma^2 I)\right) &= \frac{1}{2}\left[d\log\frac{1}{\sigma^2} - d + \|\tau u^\circ\|^2 + d\sigma^2\right] \\ &\le \frac{1}{2}\left[d\log\frac{r^2 r_x^2 d}{\varepsilon} - d + \frac{r^2}{\varepsilon^2} + \frac{\varepsilon}{r^2 r_x^2}\right] \\ &= O\left(d\log\frac{d r r_x}{\varepsilon} + \frac{r^2}{\varepsilon^2}\right)\end{aligned}$$

For the second term, we have

$$\begin{aligned}\mathrm{KL}\left(N(0,1) \mid u([-r\tau, r\tau])\right) &\le \int_{-r\tau}^{r\tau} \log\frac{1/(2r\tau)}{\frac{1}{\sqrt{2\pi}}e^{-b^2/2}} 1/(2r\tau)\mathrm{d}b \\ &= \frac{(r\tau)^2}{6} + \frac{1}{2}\log(2\pi) - \log(2r\tau) \\ &= O\left(\frac{r^2}{\varepsilon^2}\right)\end{aligned}$$

Thus, it follows that

$$\mathrm{KL}(\nu \mid \mu) = O\left(\frac{r^2}{\varepsilon^2} + d \log \frac{drr_x}{\varepsilon}\right)$$

In addition, $\|a\|_\infty \leq r$ yields $\|a\|_{L^2(\mu)} \leq r$. This completes the proof. □

C.2. Proof of Lemma 4.3

Since Rademacher complexity is smaller than Gaussian complexity (Wainwright, 2019), it suffices to bound the Gaussian complexity $\mathfrak{G}(\mathcal{F}) := \mathbb{E}_{\varepsilon_{it} \sim N(0,1)}\left[\sup_{f \in \mathcal{F}} \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \varepsilon_{it} f_t(x_i)\right]$. Let $Z_t(w) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_{it} h(x_i; w)$. Note that $Z_t(w)$ follows a Gaussian distribution with mean 0 and variance $\sigma(w)^2 := \frac{1}{n} \sum_{i=1}^n h(x_i; w)^2 \leq 1$ independently. Then, we have

$$\begin{aligned}
\mathfrak{G}(\mathcal{F}_{M,R}) &:= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_{R,M}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \varepsilon_{it} f(x_i)\right] \\
&= \mathbb{E}_\varepsilon \left[\sup_{\mathrm{KL}(\nu|\mu) \leq M} \sup_{\frac{1}{T}\sum_{t=1}^T \left\|a^{(t)}\right\|_{L^2(\mu)}^2 \leq R} \frac{1}{nT} \sum_{i=1}^T \sum_{i=1}^n \varepsilon_{it} \int a^{(t)}(w) h(x_i; w) \mathrm{d}\mu(w)\right] \\
&= \frac{1}{\sqrt{n}} \mathbb{E}_\varepsilon \left[\sup_{\mathrm{KL}(\nu|\mu) \leq M} \sup_{\frac{1}{T}\sum_{t=1}^T \left\|a^{(t)}\right\|_{L^2(\mu)}^2 \leq R} \frac{1}{T} \sum_{t=1}^T \int a^{(t)}(w) Z_t(w) \mathrm{d}\mu(w)\right] \\
&\leq \frac{1}{\sqrt{n}} \mathbb{E}_\varepsilon \left[\sup_{\mathrm{KL}(\nu|\mu) \leq M} \sup_{\frac{1}{T}\sum_{t=1}^T \left\|a^{(t)}\right\|_{L^2(\mu)}^2 \leq R} \frac{1}{T} \sum_{t=1}^T \left\|a^{(t)}(w)\right\|_{L^2(\mu)} \|Z_t(w)\|_{L^2(\mu)}\right] \\
&\leq \frac{1}{\sqrt{n}} \mathbb{E}_\varepsilon \left[\sup_{\mathrm{KL}(\nu|\mu) \leq M} \sup_{\frac{1}{T}\sum_{t=1}^T \left\|a^{(t)}\right\|_{L^2(\mu)}^2 \leq R} \sqrt{\frac{1}{T} \sum_{t=1}^T \left\|a^{(t)}(w)\right\|_{L^2(\mu)}^2} \sqrt{\frac{1}{T} \sum_{t=1}^T \|Z_t(w)\|_{L^2(\mu)}^2}\right] \\
&= \sqrt{\frac{R}{n}} \mathbb{E}_\varepsilon \left[\sup_{\mathrm{KL}(\nu|\mu) \leq M} \sqrt{\frac{1}{T} \sum_{t=1}^T \|Z_t(w)\|_{L^2(\mu)}^2}\right] \\
&\leq \sqrt{\frac{R}{n}} \sqrt{\mathbb{E}_\varepsilon \left[\sup_{\mathrm{KL}(\nu|\mu) \leq M} \int \frac{1}{T} \sum_{t=1}^T Z_t(w)^2 \mathrm{d}\mu(w)\right]}
\end{aligned} \tag{6}$$

For the last inequality, we used the fact that $\sqrt{\cdot}$ is monotonically increasing and Jensen's inequality. From the Donsker-Varadhan duality formula of the KL-divergence, we have

$$\begin{aligned}
\frac{1}{\gamma} \mathbb{E}_\varepsilon \left[\sup_{\mathrm{KL}(\nu|\mu) \leq M} \gamma \int \frac{1}{T} \sum_{t=1}^T Z_t(w)^2 \mathrm{d}\mu(w)\right] &\leq \frac{1}{\gamma} \left\{M + \mathbb{E}_\varepsilon \left[\log \int \exp\left(\frac{\gamma}{T} \sum_{t=1}^T Z_t(w)^2\right) \mathrm{d}\nu(w)\right]\right\} \\
&\leq \frac{1}{\gamma} \left\{M + \log \int \mathbb{E}_\varepsilon \left[\exp\left(\frac{\gamma}{T} \sum_{t=1}^T Z_t(w)^2\right)\right] \mathrm{d}\nu(w)\right\} \\
&\leq \frac{1}{\gamma} \left\{M + \log \int \mathbb{E}_\varepsilon \prod_{t=1}^T \left[\exp\left(\frac{\gamma}{T} Z_t(w)^2\right)\right] \mathrm{d}\nu(w)\right\}
\end{aligned} \tag{7}$$

for any $\gamma > 0$. Since $Z_t(w) \sim N(0, \sigma(w)^2)$, we have

$$\begin{aligned}
\mathbb{E}_{Z\sim N(0,\sigma(w)^2)}\Big[\exp\Big(\frac{\gamma}{T}Z^2\Big)\Big] &= \frac{1}{\sqrt{2\pi\sigma(w)^2}}\int e^{\frac{\gamma}{T}Z^2} e^{-\frac{Z^2}{2\sigma(w)^2}}\mathrm{d}Z \\
&= \frac{1}{\sqrt{2\pi\sigma(w)^2}}\int e^{-\left[\frac{1}{2\sigma(w)^2}-\frac{\gamma}{T}\right]Z^2}\mathrm{d}Z \\
&= \frac{1}{\sqrt{2\pi\sigma(w)^2}}\sqrt{\frac{\pi}{\frac{1}{2\sigma(w)^2}-\frac{\gamma}{T}}} \\
&= \sqrt{\frac{1}{1-2\frac{\gamma}{T}\sigma(w)^2}}.
\end{aligned}$$

By letting $\gamma = \frac{T}{4}$, we have

$$\mathbb{E}_Z\big[\exp\big(\gamma Z^2\big)\big] = \sqrt{\frac{1}{1-\sigma(w)^2/2}} \le \sqrt{2}. \tag{8}$$

since $\sigma(w)^2 \le 1$. Combining Eq. (6), (7), and (8), we have

$$\begin{aligned}
\mathfrak{G}(\mathcal{F}_{R,M}) &\le \sqrt{\frac{R}{n}}\sqrt{\mathbb{E}_{\varepsilon_i}\left[\sup_{\mathrm{KL}(\nu|\mu)\le M}\int\frac{1}{T}\sum_{t=1}^{T}Z_t(w)^2\mathrm{d}\mu(w)\right]} \\
&\le \sqrt{\frac{R}{n}}\sqrt{\frac{1}{\gamma}\left\{M + \log\int\prod_{t=1}^{T}\mathbb{E}_\varepsilon[\exp(\gamma Z_t(w)^2)]\mathrm{d}\nu(w)\right\}} \\
&\le \sqrt{\frac{R}{n}}\sqrt{\frac{1}{\gamma}\left\{M + \log\int\sqrt{2}^T\mathrm{d}\nu(w)\right\}} \\
&\le \sqrt{\frac{R}{n}}\sqrt{4\Big\{M/T + \log\sqrt{2}\Big\}}.
\end{aligned}$$

This completes the proof.

C.3. Proof of Theorem 4.4

From the definition of $\mathcal{B}_M$, there exists $\mu^\circ \in \mathcal{P}_M$ and $a^\circ \in L^2(\mu)(\|a^\circ\|^2_{L^2(\mu)} \le R)$ such that $f^\circ(x) = f(x; a^\circ, \mu^\circ)$. Let $\hat{a} = a_{\hat{\mu}}$. Then, from the optimality of $\hat{\mu}$ and $\hat{a}$, we have

$$\begin{aligned}
\mathcal{G}(\hat{\mu}) &= L(\hat{a}, \hat{\mu}) + \lambda\left(\frac{\lambda_a}{2}\|\hat{a}\|^2_{L^2(\hat{\mu})} + \mathrm{KL}(\nu \mid \hat{\mu})\right) \\
&\le \mathcal{G}(\mu^\circ) \\
&= L(a^\circ, \mu^\circ) + \lambda\left(\frac{\lambda_a}{2}\|a_{\mu^\circ}\|^2_{L^2(\mu^\circ)} + \mathrm{KL}(\nu \mid \mu^\circ)\right) \\
&\le 2\lambda M
\end{aligned}$$

since we set $\lambda_a = 2M/R$. Thus, it holds that

$$\begin{aligned}
\|a_{\hat{\mu}}\|^2_{L^2(\hat{\mu})} &\le 2R, \\
\mathrm{KL}(\nu \mid \hat{\mu}) &\le 2M.
\end{aligned}$$

From Lemma 4.3, Rademacher complexity of $\mathcal{F}_{2R,2M}$ is bounded as follows:

$$\mathfrak{R}(\mathcal{F}_{2R,2M}) = O\left(\sqrt{\frac{R(M+1)}{n}}\right).$$

For any $f = \int a(w)h(x;w)\mathrm{d}\mu(w) \in \mathcal{F}_{2R,2M}$, we have

$$\begin{aligned}\|f\|_\infty &\leq \int |a(w)|\mathrm{d}\mu(w)\\ &\leq \|a\|_{L^2(\mu)}\\ &\leq \sqrt{2R}.\end{aligned}$$

Thus, for any $f \in \mathcal{F}_{2R,2M}$, $l(f(x),y) = l(f(x),f^\circ(x)) \leq 4R$ and $|l'(f(x),y)| \leq 2\sqrt{2R}$. Let $\mathcal{F}' = \{(x,y) \mapsto l(f(x),y) \mid f \in \mathcal{F}_{2R,2M}\}$. Utilizing the standard uniform bound (Wainwright, 2019), for any $\delta \in [0,1]$, we have

$$\sup_{g\in\mathcal{F}'}\left\{\mathbb{E}_\rho[g(x,y)] - \frac{1}{n}\sum_{i=1}^n g(x^{(i)},y^{(i)})\right\} \leq 2\mathfrak{R}(\mathcal{F}') + 12R\sqrt{\frac{\log 2/\delta}{2n}},$$

with probability at least $1-\delta$ over the choice of n i.i.d. samples $\left\{(x^{(i)},y^{(i)})\right\}_{i=1}^n \sim \rho$. From the contraction lemma (Maurer, 2016), we have

$$\begin{aligned}\mathfrak{R}(\mathcal{F}') &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_{2R,2M}}\sum_{i=1}^n \sigma_i l(f(x_i),y_i)\right]\\ &\leq 2\sqrt{2R}\mathfrak{R}(\mathcal{F}_{2R,2M})\\ &= O\left(R\sqrt{\frac{(M+1)}{n}}\right),\end{aligned}$$

since $l(\cdot,y_i)$ is $2\sqrt{2R}$-Lipschitz continuous in $[-\sqrt{2R},\sqrt{2R}]$. Combining above arguments, we arrive at

$$\begin{aligned}\bar{L}(a_{\hat{\mu}},\mu) &\leq L(a_{\hat{\mu}},\hat{\mu}) + 2\mathfrak{R}(\mathcal{F}') + 12R\sqrt{\frac{\log 2/\delta}{2n}}\\ &= O\left(\sqrt{\frac{M}{n}} + R\sqrt{\frac{(M+1)}{n}} + R\sqrt{\frac{\log 1/\delta}{n}}\right)\\ &= O\left((R+1)\left(\sqrt{\frac{(M+1)}{n}} + \sqrt{\frac{\log 1/\delta}{n}}\right)\right)\\ &= O\left((R+1)\sqrt{\frac{(M+1)+\log 1/\delta}{n}}\right),\end{aligned}$$

since we set $\lambda = 1/\sqrt{n}$. This completes the proof.

C.4. Proof of Theorem 4.5

Let $\mathcal{S} := \left\{\sin(2\pi u\cdot x) \mid u \in \{0,1\}^d, \|u\|_1 = k\right\}$ be a subset of $L^2(\rho_X)$. Note that $\mathcal{S}$ is an orthonormal system in $L^2(\rho_X)$. Assume that

$$\sup_{f\in\mathcal{B}_M, \|f\|_{\mathcal{B}_M}^2 \leq R} d(f,H_n) < 1/4.$$

Then, from Lemma C.3, for any $\psi = \sin(2\pi u\cdot x) \in \mathcal{S}$ ($\|u\|_1 = k$), there exists a,μ such that $\mathrm{KL}(\nu \mid \mu) = O(d\log dk + k^2)$, $\|a\|_{L^2(\mu)} = k$, and

$$|\psi(x) - f(x)| \leq 1/4$$

for any $x \in [0,1]^d$ since $|u\cdot x| \leq \|u\|_1\|x\|_\infty \leq k$ and $\sin(2\pi k) = \sin(-2\pi k) = 0$. Therefore, we have

$$\begin{aligned}d(\psi,H_n) &\leq \|\psi - f\|_{L^2(\rho_X)} + d(f,H_n)\\ &< 1/4 + 1/4 = 1/2.\end{aligned}$$

This contradicts Lemma C.2. Thus, we obtain the result.

D. Proofs for Section 5

D.1. Lemmas for Section 5

Lemma D.1. *Assume that l_i is the l^2-loss and Assumption 4.1 with $\sigma = 0$. Then, we have*

$$U_{\hat{\rho}}(\mu) = \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^{T} Y_i^\top (\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} Y_i,$$
$$U_{\rho}(\mu) = \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^{T} \langle f_i^\circ, (\Sigma + \bar{\lambda}_a \,\mathrm{Id})^{-1} f_i^\circ \rangle.$$

Proof. From Lemma A.3, we have

$$\begin{aligned}
U_{\hat{\rho}}(\mu) &= \frac{1}{2nT}\sum_{i=1}^{T} \left\| Y_i - \hat{\Sigma}(\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} Y_i \right\|_2^2 + \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^{T} Y_i^\top (\hat{\Sigma} + n\lambda_a I)^{-1}\hat{\Sigma}(\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} Y_i \\
&= \frac{n\bar{\lambda}_a^2}{2T}\sum_{i=1}^{T} \left\| (\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} Y_i \right\|_2^2 + \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^{T} Y_i^\top (\hat{\Sigma} + n\bar{\lambda}_a I)^{-1}\hat{\Sigma}(\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} Y_i \\
&= \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^{T} Y_i^\top (\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} Y_i, \\
U_{\rho}(\mu) &= \frac{1}{2nT}\sum_{i=1}^{T} \left\| f_i^\circ - \Sigma(\Sigma + \bar{\lambda}_a \,\mathrm{Id})^{-1} f_i^\circ \right\|_{L^2(\rho_X)}^2 + \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^{T} \langle f_i^\circ, (\Sigma + \lambda_a \,\mathrm{Id})^{-1}\Sigma(\Sigma + \bar{\lambda}_a \,\mathrm{Id})^{-1} f_i^\circ \rangle \\
&= \frac{\bar{\lambda}_a^2}{2T}\sum_{i=1}^{T} \left\| (\Sigma + \bar{\lambda}_a \,\mathrm{Id})^{-1} f_i^\circ \right\|_{L^2(\rho_X)}^2 + \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^{T} \langle f_i^\circ, (\Sigma + \lambda_a \,\mathrm{Id})^{-1}\Sigma(\Sigma + \bar{\lambda}_a \,\mathrm{Id})^{-1} f_i^\circ \rangle \\
&= \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^{T} \langle f_i^\circ, (\hat{\Sigma} + \bar{\lambda}_a \,\mathrm{Id})^{-1} f_i^\circ \rangle.
\end{aligned}$$

□

Lemma D.2. *Assume that l_i is the l^2-loss and Assumption 4.1 with $\sigma = 0$. Let $\hat{A}'(\mu) = \frac{f^\circ(X)^\top \hat{\Sigma}_\mu f^\circ(X)}{\|f^\circ(X)\|_2^2}$ and $A'(\mu) = \frac{E[f^\circ(x)k(x,x')f^\circ(x')]}{\|f^\circ\|_{L^2(\rho_X)}^2}$. Then, we have*

$$\hat{A}(\mu) \geq \hat{A}'(\mu) \geq \frac{\bar{\lambda}_a \|f^\circ(X)\|_{L^2(\rho)}^2}{2nU_{\hat{\rho}}(\hat{\mu})} - \bar{\lambda}_a,$$
$$A(\mu) \geq A'(\mu) \geq \frac{\bar{\lambda}_a \|f^\circ\|_{L^2(\rho)}^2}{2U(\mu)} - \bar{\lambda}_a.$$

Proof. Since $\left\|\hat{\Sigma}_\mu\right\| \leq n$ and $E[k(x,x')^2] \leq 1$ from $|h(x;w)| \leq 1$, we have $\hat{A}(\mu) \geq \hat{A}'(\mu)$ and $A(\mu) \geq A'(\mu)$. Let $T = \sum_{i=1}^{\infty} \mu_i e_i f_i$ be the singular value decomposition of T, where e_m, f_m are the orthonormal basis of $L^2(\rho), L^2(\mu)$, respectively. Then, there exists $(\alpha_i)_{i\geq 1}$ such that $\sum_i \alpha_i^2 = 1$ and $f^\circ = \|f^\circ\|_{L^2(\rho)}\alpha_i e_i$. Utilizing this expression, we have

$$\begin{aligned}
U_{\rho}(\mu) &= \frac{\bar{\lambda}_a}{2}\|f^\circ\|_{L^2(\rho)}^2 \sum_i \frac{\alpha_i^2}{\mu_i + \bar{\lambda}_a} \\
&\geq \frac{\bar{\lambda}_a}{2}\|f^\circ\|_{L^2(\rho)}^2 \frac{1}{\sum_i \alpha_i^2(\mu_i + \bar{\lambda}_a)} \\
&\geq \frac{\bar{\lambda}_a}{2}\|f^\circ\|_{L^2(\rho)}^2 \frac{1}{A'(\mu) + \bar{\lambda}_a}.
\end{aligned}$$

The second inequality follows from the convexity of $1/x$ and Jensen's inequality. By the same argument, we have

$$U_\rho(\mu) \geq \frac{\bar{\lambda}_a}{2}\|f^\circ\|_{L^2(\rho)}^2 \frac{1}{n\hat{A}'(\mu) + n\bar{\lambda}_a}.$$

By transposition of the above inequalities, we obtain the results. □

Lemma D.3. *Assume that l_i is the l^2-loss and Assumption 4.1 holds. Then, we have*

$$\mathbb{E}_\varepsilon[U_{\hat{\rho}}(\mu)] = \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^T f_i^\circ(X)^\top(\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} f_i^\circ(X) + \frac{\bar{\lambda}_a\sigma^2}{6n}(n - d_{\bar{\lambda}_a}(\mu)).$$

Proof.

$$\begin{aligned}
\mathbb{E}_\varepsilon[U_{\hat{\rho}}(\mu)] &= \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^T \mathbb{E}_\varepsilon[(f_i^\circ(X) + \varepsilon_i)^\top(\hat{\Sigma} + n\bar{\lambda}_a I)^{-1}(f_i^\circ(X) + \varepsilon_i)] \\
&= \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^T f_i^\circ(X)^\top(\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} f_i^\circ(X) + \frac{\bar{\lambda}_a\sigma^2}{6}\operatorname{tr}\Big((\hat{\Sigma} + n\bar{\lambda}_a I)^{-1}\Big) \\
&= \frac{\bar{\lambda}_a}{2T}\sum_{i=1}^T f_i^\circ(X)^\top(\hat{\Sigma} + n\bar{\lambda}_a I)^{-1} f_i^\circ(X) + \frac{\bar{\lambda}_a\sigma^2}{6n}(n - d_{\bar{\lambda}_a}(\mu)).
\end{aligned}$$

□

D.2. Proof of Theorem 5.4

For $r > 0$, let

$$\tilde{f}_r(z) = \begin{cases} \tilde{f}(z) & \text{if } |z| \leq r \\ \tilde{f}(r) - \operatorname{sgn}(\tilde{f}(r))(z - r) & \text{if } r \leq z \leq r + \left|\tilde{f}(r)\right| \\ \tilde{f}(-r) - \operatorname{sgn}(\tilde{f}(-r))(r - z) & \text{if } -r - \left|\tilde{f}(-r)\right| \leq z \leq -r \\ 0 & \text{otherwise} \end{cases}.$$

This is continuous, differentiable almost everywhere and its derivative satisfies $\left|\tilde{f}_r'(x)\right| \leq 1$ a.s. From Lemma C.3, there exists $f := \int a(w)h(x;w)\mathrm{d}\mu(w)$ such that $\mathrm{KL}(\nu \mid \mu) = O(r^2/\varepsilon^2 + d\log\frac{dr r_x}{\varepsilon})$, $\|a\|_{L^2(\mu)} = O(r)$ and

$$\left|\tilde{f}_r(u^\circ \cdot x) - f(x)\right| \leq \varepsilon$$

for any $x \in \mathbb{R}^d$ such that $|u^\circ \cdot x| \leq r$, $\|x\|_2 \leq r_x\sqrt{d}$. Since $f^\circ(x) = \tilde{f}(u^\circ \cdot x) = \tilde{f}_r(u^\circ \cdot x)$ for $|u^\circ \cdot x| \leq r$, we have

$$|f^\circ(x) - f(x)| \leq \varepsilon$$

for any $x \in \mathbb{R}^d$ such that $\|x\| \leq r_x\sqrt{d}$ and $|u \cdot x| \leq r$. From the tail bound on Gaussian and chi-squared distribution (Wainwright, 2019), we have

$$P\left(\frac{1}{d}\|x\|^2 \geq \frac{r_x^2}{d}, |u \cdot x| \geq r\right) \leq 2\exp\big(-r^2/2\big) + 2\exp\big(-d(r_x^2 - 1)/8\big).$$

Thus, $\|f^\circ - f\|_{L^2(\rho_X)}$ is evaluated as follows:

$$\begin{aligned}
\|f^\circ - f\|_{L^2(\rho_X)}^2 &\leq (\|f^\circ\|_\infty + \|f\|_\infty)^2 P(\|x\| \geq \sqrt{d}r_x, |w \cdot x| \geq r) + \int_{\|x\| \leq \sqrt{d}, |w\cdot x| \leq r} (f^\circ(x) - f(x))^2 \mathrm{d}\rho_X(x) \\
&\leq (\|a\|_{L^2(\mu)} + 1)^2 P\left(\frac{1}{n}\|x\|^2 \geq \frac{r_x^2}{n}, |u \cdot x| \geq r\right) + \varepsilon^2 \\
&\leq 1/16 := \bar{\varepsilon}^2
\end{aligned}$$

by setting ε to a sufficiently small constant and r, r_x sufficiently large constants which are independent of d. Thus, there exists $M = O(d \log d), R = O(1)$ such that $\|f^\circ - f\|^2_{L^2(\rho_X)}$, where $f \in \mathcal{F}_{R,M}$. By the same argument in the proof of Theorem 4.4, we have

$$L_{\hat{\rho}}(a, \mu) \leq L_\rho(a, \mu) + O\left(R\sqrt{\frac{(M+1) + \log 1/\delta}{n}}\right)$$

with probability at least $1 - \delta$ over the choice of training data. Thus, by setting $n \geq c_3'(d \log d + \log 1/\delta)$ for sufficiently large c_3', we have

$$L_{\hat{\rho}}(a, \mu) \leq L_\rho(a, \mu) + \bar{\varepsilon}^2 \leq 2\bar{\varepsilon}^2$$

From the optimality of $\hat{\mu}$ and $\hat{a} = a_{\hat{\mu}}$, we have

$$\begin{aligned}
\mathcal{G}_{\hat{\rho}}(\hat{\mu}) &= L(\hat{a}, \hat{\mu}) + \frac{\bar{\lambda}_a}{2}\|\hat{a}\|^2_{L^2(\hat{\mu})} + \lambda \,\mathrm{KL}(\nu \mid \hat{\mu}) \\
&\leq \mathcal{G}(a, \mu) \\
&\leq 2\bar{\varepsilon}^2 + 2\lambda M \\
&\leq 3\bar{\varepsilon}^2,
\end{aligned}$$

by setting $\lambda = \bar{\epsilon}^2/2M$ and $\lambda_a = M/R$. Thus, it holds that

$$\begin{aligned}
\|\hat{a}\|^2_{L^2(\hat{\mu})} &\leq 12R, \\
\mathrm{KL}(\nu \mid \hat{\mu}) &\leq 6M.
\end{aligned}$$

Then, by the same reasoning as in the proof of Theorem 4.4, we have

$$\begin{aligned}
U_\rho(\hat{\mu}) &\leq E_\rho[l(f(x; \hat{\mu}), y)] + \frac{\bar{\lambda}_a}{2}\|a_{\hat{\mu}}\|^2_{L^2(\mu)} \\
&\leq E_{\hat{\rho}}[l(f(x; \hat{\mu}), y)] + \frac{\bar{\lambda}_a}{2}\|a_{\hat{\mu}}\|^2_{L^2(\mu)} + O\left((R+1)\sqrt{\frac{M+1+\log 1/\delta}{n}}\right) \\
&\leq \mathcal{G}_{\hat{\rho}}(\hat{\mu}) + O\left((R+1)\sqrt{\frac{M+1+\log 1/\delta}{n}}\right) \\
&\leq 3\bar{\varepsilon}^2 + O\left((R+1)\sqrt{\frac{M+1+\log 1/\delta}{n}}\right) \\
&\leq 4\bar{\varepsilon}^2
\end{aligned}$$

by setting $n = c_3(d \log d + \log 1/\delta)$ for a sufficiently large constant $c_3 \geq c_3'$ with probability at least $1 - \delta$ over the choice of training data. Therefore, it holds that

$$\begin{aligned}
A(\hat{\mu}) \geq A'(\hat{\mu}) &\geq \frac{\bar{\lambda}_a}{2U(\hat{\mu})} - \bar{\lambda}_a \\
&\geq \frac{\bar{\varepsilon}^2/(2R)}{8\bar{\varepsilon}^2} - \bar{\varepsilon}^2/2R \\
&\geq \frac{1 - 8\bar{\epsilon}^2}{16R} \\
&= \Omega(1).
\end{aligned}$$

Let $\{u_i\}_{i=1}^d$ $(u_1 = u)$ be an orthonormal basis of $\mathbb{R}^d$. Then, the symmetry of $\mu_0 = \nu$ implies that $\int \frac{\langle u_i, w\rangle^2}{\|w\|_2^2} \mathrm{d}\mu_0(w)$ is equal for any i. Since $\sum_{i=1}^d \int \frac{\langle u_i, w\rangle^2}{\|w\|_2^2} \mathrm{d}\mu_0(w) = 1$, we have $\int \frac{\langle u, w\rangle^2}{\|w\|_2^2} \mathrm{d}\mu_0(w) = 1/d$. In addition let $f_i(x) = \tilde{f}(u_i \cdot x)$. Then, we

have $\mathbb{E}_\rho[f_i(x)f_j(x)] = 0$ for $i \neq j$. Thus, we have

$$\begin{aligned}\mathbb{E}_{\rho_X}\left[\sum_{i=1}^d f_i(x)k(x,x')f_i(x')\right] &\leq \sqrt{\mathbb{E}_{\rho_X}\left[\left(\sum_{i=1}^d f_i(x)f_i(x')\right)^2\right]}\sqrt{\mathbb{E}_\rho[k(x,x')^2]} \\ &= \sqrt{\sum_{i=1}^d \mathbb{E}_\rho[f_i(x)^2 f_i(x')^2]}\sqrt{\mathbb{E}_\rho[k(x,x')^2]} \\ &= \sqrt{d}\|f^\circ(x)\|^2_{L^2(\rho_X)}\sqrt{\mathbb{E}_\rho[k(x,x')^2]}.\end{aligned}$$

Since ρ_X and μ_0 are rotationally invariant, it holds that

$$\begin{aligned}\mathbb{E}_\rho\left[\sum_{i=1}^d f_i(x)k_{\mu_0}(x,x')f_i(x')\right] &= d\mathbb{E}_\rho[f_1(x)k_{\mu_0}(x,x')f_1(x')] \\ &= d\mathbb{E}_\rho[f^\circ(x)k_{\mu_0}(x,x')f^\circ(x')].\end{aligned}$$

Thus, the kernel alignment at the initialization can be evaluated as follows:

$$\begin{aligned}A(\mu_0) &= \frac{\mathbb{E}_\rho[f^\circ(x)k_{\mu_0}(x,x')f^\circ(x')]}{\|f^\circ(x)\|^2_{L^2(\rho_X)}\sqrt{\mathbb{E}_\rho[k(x,x')^2]}} \\ &\leq \frac{1}{\sqrt{d}}.\end{aligned}$$

Let $u = u^\parallel + u^\perp$ and $x = x^\parallel + x^\perp$ be the orthogonal decomposition of u and x with respect to the space spanned by the rows of u°. Then, we have

$$\begin{aligned}\mathbb{E}_\rho[yh(x;w)] &= \int \tilde{f}(u^\circ \cdot x)\tilde{h}(x\cdot u + b)\mathrm{d}\rho_X(x) \\ &= \int \tilde{f}(u^\circ \cdot x^\parallel)\tilde{h}(u\cdot x^\parallel + u\cdot x^\perp + b)\mathrm{d}\rho_X(x).\end{aligned}$$

Here, $x^\parallel$ and $x^\perp$ follows the normal distribution $N(0, I_k), N(0, I_{d-k})$ independently. In addition, $u\cdot x^\perp$ follows $N(0, \left\|u^\perp\right\|)$. Therefore, we have

$$\mathbb{E}_\rho[yh(x;w)] = \int \tilde{f}(u^\circ \cdot x^\parallel)\tilde{h}_{\|u^\perp\|}(u^\parallel \cdot x^\parallel + \tau)\mathrm{d}\nu_k(x^\parallel).$$

where $\tilde{h}_\tau(x) = \mathbb{E}_{z\sim\nu}[\tilde{h}(x+\tau z)]$. For $\tilde{h}_\tau$, we have the following lemma.

Lemma D.4. *For any $\tau \in \mathbb{R}$, we have the following results*

- *$\tilde{h}_\tau$ is L_τ-Lipschitz continuous with $L_\tau = \min\left\{1, \frac{4}{\pi\tau}\right\}$.*

Proof. From the definition of $\tilde{h}_\tau$, we have

$$\begin{aligned}\tilde{h}'_\tau(x) &= \mathbb{E}_{z\sim N(0,1)}[\tanh'(x+\tau z)] \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\tanh'(x+\tau z)e^{-z^2/2}\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\tanh'(\tau z)e^{-(z-x/\tau)^2/2}\mathrm{d}z.\end{aligned}$$

Thus, $\tilde{h}''_\tau(x)$ is given by

$$\begin{aligned}\tilde{h}''_\tau(x) &= \frac{1}{\sqrt{2\pi\tau}}\int_{-\infty}^{\infty}(z - x/\tau)\tanh'(\tau z)e^{-(z-x/\tau)^2/2}\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi\tau}}\int_{-\infty}^{\infty} z\tanh'(x+\tau z)e^{-z^2/2}\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi\tau}}\int_{0}^{\infty}\frac{ze^{-z^2/2}}{\cosh^2(x+\tau z)} - \frac{ze^{-z^2/2}}{\cosh^2(x-\tau z)}\mathrm{d}z.\end{aligned}$$

From the symmetry of $\cosh$, we have

$$\frac{1}{\cosh^2(x+\tau z)} - \frac{1}{\cosh^2(x-\tau z)} \leq 0$$

for any $x \geq 0$. Thus, we have $\tilde{h}''_\tau(x) \leq 0$ for any $x \geq 0$, that is, $\tilde{h}_\tau(x)$ is monotonically decreasing in $[0, \infty]$. From the symmetry of $\tilde{h}'_\tau(x)$ and $\tilde{h}'_\tau(x) \geq 0$ for any x, $\tilde{h}_\tau(0) \geq \left|\tilde{h}_\tau(x)\right|$ for any x. Thus, it suffices to evaluate $\tilde{h}'_\tau(0)$. Here, we have

$$\begin{aligned}\tilde{h}'_\tau(0) &= \mathbb{E}_{z\sim\nu}[\tilde{h}'(bz)] \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{4}{(e^{bz}+e^{-bz})^2}\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z \\ &\leq \frac{8}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-2bz}\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z \\ &\leq \frac{8}{\sqrt{2\pi}}e^{2\tau^2}\int_{0}^{\infty}\exp\left(-\frac{(z+2\tau)^2}{2}\right)\mathrm{d}z \\ &\leq \frac{8}{\sqrt{2\pi}}e^{2\tau^2}\int_{2\tau}^{\infty}\exp\left(-\frac{z'^2}{2}\right)\mathrm{d}z' \\ &\leq \frac{8}{\sqrt{2\pi}}e^{2\tau^2}\int_{2\tau}^{\infty}\exp\left(-\frac{z'^2}{2}\right)\mathrm{d}z' \\ &\leq \frac{8}{\sqrt{2\pi}}e^{2\tau^2}P(Z' > 2\tau) \\ &\leq \frac{8}{\sqrt{2\pi}}e^{2\tau^2}\sqrt{\frac{1}{2\pi}}\frac{e^{-2\tau^2}}{2\tau} \\ &= \frac{2}{\pi\tau}.\end{aligned}$$

For the last inequality, we used Mill's inequality. Obviously, $\tilde{h}'_\tau(0) \leq 1$. This completes the proof. □

Utilizing the above lemma, we have

$$
\begin{aligned}
|\mathbb{E}_\rho[f^\circ(x)h(x;w)]|^2 &\leq \left| \int \tilde{f}(u^\circ \cdot x^{\|}) \tilde{h}_{\|u^\perp\|}(u^{\|} \cdot x^{\|} + \tau) \mathrm{d}\nu_k(x^{\|}) \right|^2 \\
&\leq \left| \int \tilde{f}(u^\circ \cdot x^{\|}) \Big[\tilde{h}_\tau(\tau) + (\tilde{h}_\tau(u^{\|} \cdot x^{\|} + \tau) - \tilde{h}_\tau(\tau)) \Big] \mathrm{d}\nu_k(x^{\|}) \right|^2 \\
&\leq \left| \int \tilde{f}(u^\circ \cdot x^{\|}) (\tilde{h}_\tau(u^{\|} \cdot x^{\|} + \tau) - \tilde{h}_\tau(\tau)) \mathrm{d}\nu_k(x^{\|}) \right|^2 \\
&\leq \|f^\circ\|^2_{L^2(\rho_X)} \int (\tilde{h}_\tau(u^{\|} \cdot x^{\|} + \tau) - \tilde{h}_\tau(\tau))^2 \mathrm{d}\nu(z_1) \\
&\leq \|f^\circ\|^2_{L^2(\rho_X)} \int (L_\tau u^{\|} \cdot x^{\|})^2 \mathrm{d}\nu_k(x^{\|}) \\
&\leq \|f^\circ\|^2_{L^2(\rho_X)} \int \left(L_\tau \left\| u^{\|} \right\| z \right)^2 \mathrm{d}\nu(z) \\
&\leq \|f^\circ\|^2_{L^2(\rho_X)} L_\tau^2 \left\| u^{\|} \right\|^2 \\
&\leq \|f^\circ\|^2_{L^2(\rho_X)} \frac{8 \left\| u^{\|} \right\|^2}{\pi^2 \|u^\perp\|^2}.
\end{aligned}
$$

From the boundedness of h, we have $|\mathbb{E}_\rho[f^\circ(x)h(x;w)]|^2 \leq \|f^\circ\|^2_{L^2(\rho_X)}$. Combining above arguments, we have

$$
|\mathbb{E}_\rho[f^\circ(x)h(x;w)]|^2 \leq \|f^\circ\|^2_{L^2(\rho_X)} \min \left\{ 1, \frac{8 \left\| w^{\|} \right\|^2}{\pi^2 \|w^\perp\|^2} \right\}
$$

By the way, for any a, b, we have

$$
\min \left\{ 1, \frac{a^2}{b^2} \right\} \leq 2 \frac{a^2}{a^2 + b^2}
$$

Therefore, it holds that

$$
\mathbb{E}_\rho[f^\circ(x)h(x;w)]^2 \leq \|f^\circ\|_{L^2(\rho_X)} \frac{16}{\pi^2} \frac{\left\| w^{\|} \right\|^2}{\|w\|^2}
$$

Recall that $A'(\mu) = \frac{E_\rho[f^\circ(x) E_\mu[h(x;w)h(x';w)] f^\circ(x')]}{\|f^\circ\|^2_{L^2(\rho)}} = \frac{E_\mu[E_\rho[f^\circ(x)h(x;w)]^2]}{\|f^\circ\|^2_{L^2(\rho)}}$. Thus, from the definition of $P(\mu)$, we have

$$
\begin{aligned}
P(\hat{\mu}) &= \mathbb{E} \left[\frac{\left\| w^{\|} \right\|^2}{\|w\|^2} \right] \\
&= \Omega(A'(\hat{\mu})) = \Omega(1).
\end{aligned}
$$

D.3. Proof of Theorem 5.5

Tha label noise procedure can be regarded as a SGD-MFLD in (Suzuki et al., 2023a). Thus, as in the proof of Theorem 3.7, we follow the framework in Suzuki et al. (2023a).

First, we show the convexity of $\bar{U}(\mu) := E_{\tilde{\varepsilon}}[U_{\tilde{\varepsilon}}(\mu)]$. The functional $U_{\tilde{\varepsilon}}(\mu)$ can be written as

$$\begin{aligned}
U_{\tilde{\varepsilon}}(\mu) &= \frac{1}{T}\sum_{i=1}^{T}\left[\frac{1}{2n}\left\|Y_i - \hat{\Sigma}_\mu(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}(Y_i+\tilde{\varepsilon}_i)\right\|_2^2 + \frac{\bar{\lambda}_a}{2}(Y_i+\tilde{\varepsilon}_i)^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}\hat{\Sigma}_\mu(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}(Y_i+\tilde{\varepsilon}_i)\right] \\
&= \frac{1}{T}\sum_{i=1}^{T}\left[\frac{\bar{\lambda}_a}{2}Y_i^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}Y_i + \frac{1}{2n}\tilde{\varepsilon}_i^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}\hat{\Sigma}_\mu\tilde{\varepsilon}_i\right] \\
&= \frac{1}{T}\sum_{i=1}^{T}\left[\frac{\bar{\lambda}_a}{2}Y_i^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}Y_i - \frac{\bar{\lambda}_a}{2}\tilde{\varepsilon}_i^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}\tilde{\varepsilon}_i + \frac{\|\tilde{\varepsilon}_i\|^2}{2n}\right].
\end{aligned}$$

Taking the expectation of the above equation with respect to $\tilde{\varepsilon}_i$, we have

$$\begin{aligned}
\mathbb{E}_\varepsilon[U_{\tilde{\varepsilon}}(\mu)] &= \frac{1}{T}\sum_{i=1}^{T}\frac{\bar{\lambda}_a}{2}Y_i^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}Y_i - \frac{\tilde{\sigma}^2\bar{\lambda}_a}{6}\operatorname{tr}\left[(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}\right] + \frac{\tilde{\sigma}^2}{6} \\
&= \frac{\bar{\lambda}_a}{6}\operatorname{tr}\left[(\hat{\Sigma}_\mu + n\bar{\lambda}_a)^{-1}\tilde{Y}\right] + \frac{\tilde{\sigma}^2}{6},
\end{aligned}$$

where $\tilde{Y} = \frac{1}{T}\sum_{i=1}^{T} Y_iY_i^\top - \tilde{\sigma}^2 I/3$. From the assumption on $\tilde{\sigma}^2$, $\tilde{Y}$ is positive semi-definite and thus, $\bar{U}(\mu)$ is convex.

Next, we derive an LSI constant for $\tilde{p}_\mu \propto \exp\left(-\frac{1}{\lambda}\frac{\delta\bar{F}(\mu)}{\delta\mu}\right)$. The first variation of $\bar{F}(\mu)$ is given by

$$\frac{\delta\bar{F}(\mu)}{\delta\mu}(w) = -\frac{\bar{\lambda}_a}{2}\operatorname{tr}\left[(\hat{\Sigma}_\mu + n\bar{\lambda}_a)^{-1}h(X;w)h(X;w)^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}\tilde{Y}\right] + \frac{\bar{\lambda}_w}{2}\|w\|_2^2.$$

Since $\left\|\tilde{Y}\right\|_2 \le \frac{1}{T}\sum_{i=1}^{T}\|Y_i\|_2 \le nc_l^2$, we have

$$\left|\frac{\bar{\lambda}_a}{2}\operatorname{tr}\left[(\hat{\Sigma}_\mu + n\bar{\lambda}_a)^{-1}h(X;w)h(X;w)^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}\tilde{Y}\right]\right| \le \frac{\left\|\tilde{Y}\right\|_2}{2n^2\bar{\lambda}_a}\|h(X;w)\|_2^2 \le \frac{c_l^2}{\bar{\lambda}_a}.$$

Thus, $\bar{p}_\mu$ satisfies the LSI with the same constant as in Lemma 3.5.

Let $V(\mu) := \frac{1}{2T}\sum_{i=1}^{T}\tilde{\varepsilon}_i^\top(\hat{\Sigma}_\mu + n\bar{\lambda}_a I)^{-1}\tilde{\varepsilon}_i$. Then, $V(\mu)$ is equal to $U(\mu)$ if $Y_i = \varepsilon_i$ for any $i \in [T]$. Thus, by the same argument as in the proof of Theorem 3.7, we have

- $\left\|\nabla\frac{\delta V}{\delta\mu}(\mu)(w) - \nabla\frac{\delta V}{\delta\mu}(\mu')(w')\right\| \le L_V(W_2(\mu,\mu') + \|w-w'\|)$, $\left|\frac{\delta^2 U(\mu)}{\delta\mu^2}(w,w')\right| \le L_V$ for $L_U = 4R_a^2 + \bar{\lambda}_a(B_a'L_a' + R_a'^2) + B_a'^2$.
- $\left\|\nabla\frac{\delta V}{\delta\mu}(\mu)(w)\right\| \le R_V$ for $R_V = \bar{\lambda}_a B_a' R_a'$.

where we define R_a', B_a', L_a' by replacing c_l in R_a, B_a, L_a with $\tilde{\sigma}$, respectively. Furthermore, it holds that

- $\left\|\nabla_w\frac{\delta V}{\delta\mu}(\mu)(w)\right\|_2 \le R_V$.
- $\left\|\nabla_w\nabla_w^\top\frac{\delta V}{\delta\mu}(\mu)(w)\right\|_{\mathrm{op}} \le L_V$, $\left\|\nabla_w\nabla_{w'}^\top\nabla_w^\top\frac{\delta^2 V}{\delta\mu^2}(\mu)(w,w')\right\|_{\mathrm{op}} \le L_V$.

Let $L_{UV} = L_U + L_V, R_{UV} = R_U + R_V$. Then, it holds That

- $\left\|\nabla\frac{\delta U_{\tilde{\varepsilon}}}{\delta\mu}(\mu)(w) - \nabla\frac{\delta V}{\delta\mu}(\mu')(w')\right\| \le L_{UV}(W_2(\mu,\mu') + \|w-w'\|)$, $\left|\frac{\delta^2 U(\mu)}{\delta\mu^2}(w,w')\right| \le L_{UV}$,

- $\left\|\nabla \frac{\delta U_{\tilde{\varepsilon}}}{\delta \mu}(\mu)(w)\right\| \leq R_{UV}$,
- $\left\|\nabla_w \frac{\delta U_{\tilde{\varepsilon}}}{\delta \mu}(\mu)(w)\right\|_2 \leq R_{UV}$,
- $\left\|\nabla_w \nabla_w^\top \frac{\delta U_{\tilde{\varepsilon}}}{\delta \mu}(\mu)(w)\right\|_{\mathrm{op}} \leq L_{UV}$, $\left\|\nabla_w \nabla_{w'}^\top \frac{\delta^2 U_{\tilde{\varepsilon}}}{\delta \mu^2}(\mu)(w,w')\right\|_{\mathrm{op}} \leq L_{UV}$.

since $U_{\tilde{\varepsilon}}(\mu) = U(\mu) - V(\mu) + \text{const}$. Combining above arguments, Theorem 3 in Suzuki et al. (2023a) yields

$$\frac{1}{N}\mathbb{E}[\mathcal{L}^N(\mu_k^{(N)})] - \mathcal{L}(\mu^*) \leq \exp(-\lambda\alpha\eta k)\left(\frac{1}{N}\mathbb{E}[\mathcal{L}^N(\mu_k^{(N)})] - \mathcal{L}(\mu^*)\right) + \frac{2}{\lambda\alpha}\bar{L}^2 C_1(\lambda\eta + \eta^2) + \frac{4}{\lambda\alpha\eta}\bar{\Upsilon} + \frac{2C_\lambda}{\lambda\alpha N},$$

where $\bar{R}^2 = \mathbb{E}\left[\left\|w_i^{(0)}\right\|_2\right] + \frac{1}{\lambda_w}\left[\left(\frac{1}{4} + \frac{1}{\lambda_w}\right)R_{UV}^2 + \lambda d'\right]$, $\bar{L} = L_{UV} + \bar{\lambda}_w$, $C_1 = 8[R_{UV}^2 + \bar{\lambda}_w\bar{R}^2 + d']$, $C_\lambda = 2\lambda L_{UV}\alpha + 2\lambda^2 L_{UV}^2\bar{R}^2$, and

$$\bar{\Upsilon} := 4\eta\delta_\eta + [R_{UV} + \lambda_w\bar{R} + (L_{UV} + \bar{\lambda}_w)^2](1 + \sqrt{\lambda/\eta})\eta^2 R_{UV}^2 + (R_{UV} + \bar{\lambda}_w\bar{R})R_{UV}(1 + \sqrt{\lambda/\eta})\eta^3 R_{UV}^2.$$

This completes the proof.