Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics

Taiji Suzuki
The University of Tokyo, Tokyo, Japan
RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
taiji@mist.i.u-tokyo.ac.jp

Abstract

We introduce a new theoretical framework to analyze deep learning optimization with connection to its generalization error. Existing frameworks such as mean field theory and neural tangent kernel theory for neural network optimization analysis typically require taking limit of infinite width of the network to show its global convergence. This potentially makes it difficult to directly deal with finite width network; especially in the neural tangent kernel regime, we cannot reveal favorable properties of neural networks beyond kernel methods. To realize more natural analysis, we consider a completely different approach in which we formulate the parameter training as a transportation map estimation and show its global convergence via the theory of the *infinite dimensional Langevin dynamics*. This enables us to analyze narrow and wide networks in a unifying manner. Moreover, we give generalization gap and excess risk bounds for the solution obtained by the dynamics. The excess risk bound achieves the so-called fast learning rate. In particular, we show an exponential convergence for a classification problem and a minimax optimal rate for a regression problem.

1 Introduction

Despite the extensive empirical success of deep learning, there are several missing issues in theoretical understanding of its optimization and generalizations. Even though there are several theoretical analyses on its generalization error and representation ability [46, 8, 2, 67, 56], they are not necessarily well connected with an optimization procedure. The biggest difficulty in neural network optimization lies in its non-convexity. Recently, this difficulty of non-convexity is partly resolved by considering infinite width limit of networks as performed in *mean field theory* [58, 40] and *Neural Tangent Kernel* (NTK) [32, 22]. These analyses deal with different scaling of parameters for taking the limit of the width, but they share a similar spirit that an appropriate gradient descent direction can be found in an over-parameterized setting until convergence.

The mean field analysis formulates the neural network training as a gradient flow in the space of probability measures over the weights. The gradient flow corresponding to a deterministic dynamics of the weights can be analyzed as an interacting particle system [47, 18, 53, 54]. On the other hand, a stochastic dynamics of an interacting particle system can be formulated as McKean–Vlasov dynamics, and convergence to the global optimal is ensured by the ergodicity of this dynamics [40, 41]. Intuitively, inducing stochastic noise makes the solution easier to get out of local optimal and facilitates convergence to the global optimal.

The second regime, NTK, deals with larger scaling than the mean field regime, and the gradient descent dynamics is approximated by that in the tangent space at the initial solution [32, 23, 1, 22, 3].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

That is, in the wide limit of the neural network, the gradient descent can be seen as that in an reproducing kernel Hilbert space (RKHS) corresponding to the neural tangent kernel, which resolves the difficulty of non-convexity. Actually, it is shown that the gradient descent converges to the zero error solution exponentially fast for a sufficiently large width network [23, 1, 22]. In addition to the optimization, its generalization error has been also extensively studied in the NTK regime [23, 1, 22, 76, 16, 17, 79, 50, 48, 34]. On the other hand, [29] pointed out that non-convexity of a deep neural network model is essential to show superiority of deep learning over linear estimators such as kernel methods as in the analysis of [65, 30, 66]. Therefore, the NTK regime would not be appropriate to show superiority of deep learning over other methods such as kernel methods.

The above mentioned researches opened up new directions for analyzing deep learning optimization. However, all of them require that the width should diverge as the sample size goes up to show the global convergence and obtain generalization error bounds. On the other hand, a convergence guarantee for "fixed width" training is still difficult and we have not obtained a satisfactory result that can bridge both of under-parameterized and over-parameterized settings in a *unifying manner*. One way to tackle non-convexity in a finite width situation would be stochastic gradient Langevin dynamics (SGLD) [77, 51, 24]. This would be useful to show the global convergence for the non-convex optimization in deep leaning. However, the convergence rate depends exponentially to the dimensionality, which is not realistic to analyzing neural network training that typically requires huge parameter size.

Our contribution: In this paper, we resolve these difficulties such as (i) diverging width against sample size and (ii) curse of dimensionality for analyzing Langevin dynamics in neural network training by formulating the neural network training as a *transport map* estimation problem of the parameters. By doing so, we can deal with finite width and infinite width in a unifying manner. We also give a generalization error bound for the solution obtained by our optimization formulation and further show that it achieves *fast learning rate* in a well-specified setting. The preferable generalization error heavily relies on similarity between a *nonparametric Bayesian Gaussian process estimator* and the Langevin dynamics. More details are summarized as follows:

- **(formulation)** We formulate neural network training as a transportation map learning of weights (parameters) and solve this problem by infinite dimensional gradient Langevin dynamics in RKHS [20, 45]. This formulation has a wide range of applications including two layer neural network, ResNet, Wasserstein optimal transportation map estimation and so on.
- **(optimization)** Based on this formulation, we show its global convergence for finite width and infinite width in a unifying manner. We give its size independent convergence rate.
- **(generalization)** We derive the generalization error bound of the estimator obtained by our optimization framework. We also derive the fast learning rate in a student-teacher setup. Especially, we show exponential convergence for classification.

2 Problem setting and model: Training parameter transportation map

In this section, we give the problem setting and notations that will be used in the theoretical analysis. Basically, we consider the standard supervised leaning where data consists of input-output pairs $z = (x, y)$ where $x \in \mathbb{R}^d$ is an input and $y \in \mathbb{R}$ is an output (or label). We may also consider a unsupervised learning setting, but just for the presentation simplicity, we consider a supervised learning. Suppose that we are given n i.i.d. observations $D_n = (x_i, y_i)_{i=1}^n$ distributed from a probability distribution P, the marginal distributions of which with respect to x and y are denoted by P_X and P_Y respectively. We denote $\mathcal{X} = \mathrm{supp}(P_X)$. To measure the performance of a trained function f, we use a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ $((y, f) \mapsto \ell(y, f))$ and define the expected risk and the empirical risk as $\mathcal{L}(f) := \mathrm{E}_{Y,X}[\ell(Y, f(X))]$ and $\widehat{\mathcal{L}}(f) := \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i))$ respectively. As in the standard deep learning, we optimize the training risk $\widehat{\mathcal{L}}$. Our theoretical interest is to bound the following errors for an estimator $\widehat{f}$:

$$\text{Excess risk: } \mathcal{L}(\widehat{f}) - \inf_{f:\text{measurable}} \mathcal{L}(f), \qquad \text{Generalization gap: } \mathcal{L}(\widehat{f}) - \widehat{\mathcal{L}}(\widehat{f}).$$

In a typical situation, the generalization gap is bounded as $O(1/\sqrt{n})$ via VC-theory type analysis [43], for example. On the other hand, the excess risk can be faster than $O(1/\sqrt{n})$, which is known as a *fast learning rate* [42, 5, 35, 27]. The population L_2-norm with respect to P is denoted by

$\|f\|_{L_2} := \sqrt{\mathrm{E}_{Z\sim P}[f(Z)^2]}$ and the sup-norm on the domain of the input distribution P_X is denoted by $\|f\|_\infty := \sup_{x\in\mathrm{supp}(P_X)} |f(x)|$.

2.1 Introductory setting: mean field training of two layer neural network

Here, we explain the motivation of our theoretical framework by introducing mean field analysis of two layer neural networks. Let us consider the following two layer neural network model:

$$f_\Theta(x) = \tfrac{1}{M}\textstyle\sum_{m=1}^M a_m \sigma(w_m^\top x). \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a smooth activation function, $(a_m)_{m=1}^M \subset \mathbb{R}$ is the set of weights in the second layer which we assume is fixed for simplicity, and $\Theta = (w_m)_{m=1}^M \subset \mathbb{R}^d$ is the set of weights in the first layer. We aim to minimize the following regularized empirical risk with respect to Θ and analyze the dynamics of gradient descent updates:

$$\min_\Theta \ \widehat{\mathcal{L}}(f_\Theta) + \tfrac{\lambda}{2M}\textstyle\sum_{m=1}^M \|w_m\|^2.$$

The stochastic gradient descent (SGD) update for optimizing $\widehat{\mathcal{L}}(f_\Theta)$ with respect to Θ is reduced to

$$w_m^{(t+1)} = w_m^{(t)} - \eta(\tfrac{\lambda}{M} w_m^{(t)} + \nabla_{w_m}\widehat{\mathcal{L}}(f_{\Theta^{(t)}})) + \sqrt{2\eta/\beta}\,\epsilon_t^{(m)}, \tag{2}$$

where $\nabla_{w_m}\widehat{\mathcal{L}}(f_{\Theta^{(t)}}) = \frac{a_m}{M}\frac{1}{n}\sum_{i=1}^n x_i \sigma'(w_m^{(t)\top} x_i)\ell'(y_i, f_{\Theta^{(t)}}(x_i))$ and $\epsilon_t^{(m)}$ is an i.i.d. Gaussian noise mimicking the deviation of the stochastic gradient. Here, $\eta > 0$ is a step size and $\beta > 0$ is an inverse temperature parameter. This could be time discretized version of the following continuous time stochastic differential equation (SDE):

$$\mathrm{d}w_m(t) = -\big(\tfrac{\lambda}{M} w_m(t) + \nabla_{w_m(t)}\widehat{\mathcal{L}}(f_{\Theta^{(t)}})\big)\mathrm{d}t + \sqrt{2\eta/\beta}\,\mathrm{d}B_t^{(m)},$$

where $(B_t^{(m)})_t$ is a d-dimensional Brownian motion. In the mean field analysis, this optimization process is casted to an optimization of probability distribution over the parameters [40, 41, 47, 18] based on the following integral representation of neural networks:

$$f_\rho(x) := \int_{\mathbb{R}^d} a\sigma(w^\top x)\mathrm{d}\rho(w), \tag{3}$$

where ρ is a Borel probability measure defined on the parameter space $\mathbb{R}^d$ and the parameter in the second layer is fixed to a constant $a \in \mathbb{R}$ just for presentation simplicity. The time evolution of the distribution ρ is deduced from the optimization dynamics with respect to each "particle" given by

$$\mathrm{d}W(t) = -\Big(\lambda W(t) + a\frac{1}{n}\sum_{i=1}^n x_i\sigma'(W(t)^\top x_i)\ell'(y_i, f_{\rho_t}(x_i))\Big)\mathrm{d}t + \sqrt{\beta^{-1}}\mathrm{d}B_t,$$

where ρ_t is the probability law of $W(t) \in \mathbb{R}^d$ with an initial distribution $W(0) \sim \rho_0$, which is one of the *McKean-Vlasov* processes. We can see that this equation is space-time continuous limit of the update Eq. (2). Importantly, ρ_t admits a density function π_t obeying the so-called continuity equation [40, 41]. The usual finite width network is regarded as a finite sum approximation of the integral representation (Eq. (3)). As a consequence, the convergence analysis needs to take limit of infinite width to approximate the absolutely continuous distribution ρ_t. Hence, a finite width dynamics is outside the scope of mean field analysis. This is due to the fact that an independent noise is injected to each particle regardless its location; the diffusion B_t is independently and identically applied to each realized path $\{W(t) \mid t \geq 0\}$ (interaction between particles is induced only through gradient). However, in a real neural network training, the noise induced by stochastic gradient has high correlation between each node. Thus, we need a different approach.

Lift of McKean-Vlasov process Our core idea is to "lift" the stochastic process $W(t)$ as a process of a function with the initial value $W(0)$. For each $W(0) = w_0$, the particle's location at time t is determined by $W(t) = W(t, w_0)$. This means that the process generates a function $w_0 \mapsto W(t, w_0)$ with respect to the initial solution w_0. By considering the stochastic process of this function itself directly, the dynamics is transformed to an *infinite dimensional stochastic differential equation*, which has been studied especially in the stochastic partial differential equation [20]. In other words, we try to estimate a map from the initial parameters to the solution at time t instead of analyzing each particle's behavior.

From this perspective, we can directly regularize the smoothness of the trajectory, especially, we can incorporate a smoothed noise of the dynamics by utilizing a spatially correlated Gaussian process in the space of functions on parameters. Let $W_t(w) = W(t, w)$ and we regard W_t as a member of $L_2(\rho_0)$ space. Then, f_{ρ_t} can be rewritten by

$$f_{W_t}(x) := \int_{\mathbb{R}^d} a\sigma(W_t(w)^\top x)\mathrm{d}\rho_0(w) = \int_{\mathbb{R}^d} a\sigma(w^\top x)\mathrm{d}W_t\sharp\rho_0(w), \tag{4}$$

where $W_t\sharp\rho_0$ is the pushforward of the measure ρ_0 by the map W_t, i.e., $f\sharp\mu(B) := \mu \circ f^{-1}(B) = \mu(f^{-1}(B))$ for a Borel measurable map $f : \mathbb{R}^d \to \mathbb{R}^d$, a Borel measure μ, and a Borel set $B \subset \mathbb{R}^d$. By using this notation, the stochastic process we consider can be written as

$$\mathrm{d}W_t = -(AW_t + \nabla_W \widehat{\mathcal{L}}(f_{W_t}))\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\xi_t, \tag{5}$$

where $A : L_2(\rho_0) \to L_2(\rho_0)$ is an unbounded linear operator corresponding to a regularization (which will be explained later in more details), $\nabla_W \widehat{\mathcal{L}}(f_W)$ is the Frechet derivative of $\widehat{\mathcal{L}}(f_W)$ with respect to W in the space of $L_2(\rho_0)$, in our setting, which is given by $\nabla_W \widehat{\mathcal{L}}(f_W)(w) = a\frac{1}{n}\sum_{i=1}^n x_i\sigma'(W(w)^\top x_i)\ell'(y_i, f_W(x_i))$. $(\xi_t)_t$ is a *cylindric Brownian motion* in $L_2(\rho_0)$ [20], which is an infinite dimensional Brownian motion and will be defined rigorously later on. In practical deep learning, the regularization term AW_t is induced by several mechanism such as weight decay [37], dropout [60, 74], batch-normalization [31]. As a result, the regularization term AW_t introduces spatial correlation between particles unlike the McKean-Vlasov process.

Then, training two layer neural networks is formulated as optimizing the map $W : w \in \mathbb{R}^d \mapsto W(w) \in \mathbb{R}^d$ with the initial condition $W_0 = \mathbb{I}$ (identity map). This dynamics is well analyzed and guaranteed to converge to at least a stationary distribution (a.k.a., invariant measure) under mild assumptions [19, 39, 59, 33, 57, 28] which is useful to show convergence to a (near) global optimal.

Remark 1. *We would like to emphasize that our formulation admits a finite width neural network training by setting the initial distribution ρ_0 as a discrete distribution $\rho_0 = \frac{1}{M}\sum_{m=1}^M \delta_{w_m}$ for a Dirac measure δ_{w_m} which has probability 1 on a point w_m. In this situation, optimizing the map W_t corresponds to optimizing the finite width model* (1) *because $\rho_t = W_t\sharp\rho_0 = \frac{1}{M}\sum_{m=1}^M \delta_{W_t(w_m)}$ which is still a discrete distribution throughout entire $t \in \mathbb{R}_+$. This is remarkably different from both mean field analysis and NTK analysis that essentially take infinite width limits: mean field analysis in [40, 41] requires $M = \Omega(e^T)$ for a time horizon T and NTK requires $M = \Omega(\mathrm{poly}(n))$ [79].*

General formulation of our optimization problem Here, we describe mathematical details of optimizing the transportation map in a more general setting and give a practical algorithm of the corresponding GLD. We assume that the map $W_t(\cdot)$ is included in a separable Hilbert space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$ and an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ (in the previous section, $\mathcal{H} = L_2(\rho_0)$). The Hilbert space $\mathcal{H}$ consists of functions whose domain is a set $\mathcal{W}$ and whose range is $\widetilde{\mathcal{W}}$ (in the previous example, $\mathcal{W} = \mathbb{R}^d$ amd $\widetilde{\mathcal{W}} = \mathbb{R}^d$). Since a function $w \in \mathcal{H}$ has no smoothness condition in typical settings, we consider a more "regulated" subspace of $\mathcal{H}$. Such a subspace is denoted by $\mathcal{H}_K$ and given by $\mathcal{H}_K := \left\{\sum_{k=0}^\infty \alpha_k e_k \mid \sum_{k=0}^\infty \alpha_k^2/\mu_k < \infty\right\}$, where $(e_k)_{k=0}^\infty$ is an orthonormal basis of $\mathcal{H}$ and $(\mu_k)_{k=0}^\infty$ is a non-increasing non-negative sequence. We equip an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ to the space $\mathcal{H}_K$ defined by $\langle f, g\rangle_{\mathcal{H}_K} = \sum_{k=0}^\infty \alpha_k\beta_k/\mu_k$ for $f = \sum_{k=0}^\infty \alpha_k e_k \in \mathcal{H}_K$ and $g = \sum_{k=0}^\infty \beta_k e_k \in \mathcal{H}_K$. Correspondingly, the norm $\|\cdot\|_{\mathcal{H}_K}$ is defined from the inner product. When $\mathcal{H} = L_2(\rho_0)$, $\mathcal{H}_K$ becomes a *reproducing kernel Hilbert space* (RKHS) corresponding to a kernel function $K(x, y) = \sum_{k=0}^\infty \mu_k e_k(x)e_k(y)$ where $x, y \in \mathbb{R}^d$ under an appropriate convergence condition. That is, we have the reproducing property $\langle K(x,\cdot), W\rangle_{\mathcal{H}_K} = W(x)$ for each $W \in \mathcal{H}_K$. Based on the norm $\|\cdot\|_{\mathcal{H}_K}$, we define an unbounded linear operator $A : \mathcal{H} \to \mathcal{H}$ as $Af = \lambda\sum_{k=0}^\infty \frac{\alpha_k}{\mu_k}e_k$, for $f = \sum_{k=0}^\infty \alpha_k e_k \in \mathcal{H}$. We note that $Af = \frac{\lambda}{2}\nabla_f\|f\|_{\mathcal{H}_K}^2$ which is a Frechet derivative of $\lambda\|\cdot\|_{\mathcal{H}_K}^2$ in $\mathcal{H}$ (which is the derivative of the RKHS norm, if $\mathcal{H}_K$ is an RKHS). We assume that for each $W \in \mathcal{H}$, there exits a function $f_W : \mathbb{R}^d \to \mathbb{R}$ as in Eq. (4), and we basically aim to minimize the regularized empirical risk

$$\widehat{\mathcal{L}}(f_W) + \tfrac{\lambda}{2}\|W\|_{\mathcal{H}_K}^2.$$

By abuse of notation, we denote by $\widehat{\mathcal{L}}(W)$ indicating $\widehat{\mathcal{L}}(f_W)$. To execute this non-convex optimization, we use the GLD in the infinite dimensional Hilbert space $\mathcal{H}$ as introduced in Eq. (5). Here, $(\xi_t)_{t\geq 0}$ in Eq. (5) is the cylindrical Brownian motion defined as $\xi_t = \sum_{k\geq 0} B_t^{(k)} e_k$ where $(B_t^{(k)})_{t\geq 0}$ is a real valued standard Brownian motion and they are independently identical for $k = 0, 1, 2, \dots$ [1]. Since this is defined on a continuous time domain, we introduce a discrete time *implicit Euler scheme* for practical implementation:

$$W_{k+1}{=}W_k{-}\eta(AW_{k+1}{+}\nabla_W\widehat{\mathcal{L}}(W_k)){+}\sqrt{\tfrac{2\eta}{\beta}}\epsilon_k \Leftrightarrow W_{k+1}{=}S_\eta\Big(W_k{-}\eta\nabla_W\widehat{\mathcal{L}}(W_k){+}\sqrt{\tfrac{2\eta}{\beta}}\epsilon_k\Big), \quad (6)$$

where $\eta > 0$ is the step size and $S_\eta = (\mathbb{I} + \eta A)^{-1}$. We can see that the "regularization effect" AW induces the spacial smoothness of the noise of the gradient. It is known [14] that under some assumption (Assumption 1 below is sufficient), the process (5) has a unique invariant measure π_∞ given by

$$\frac{\mathrm{d}\pi_\infty}{\mathrm{d}\nu_\beta}(W) \propto \exp(-\beta\widehat{\mathcal{L}}(W)),$$

where ν_β is the Gaussian measure in $\mathcal{H}$ with mean 0 and covariance $(\beta A)^{-1}$ (see Da Prato & Zabczyk [20] for the rigorous definition of the Gaussian measure on a Hilbert space and related topics about existence of invariant measure). In a special situation where $\beta = n,\ \lambda = 1/n$ and $\beta\widehat{\mathcal{L}}(W)$ is a log-likelihood function of some model, this invariant measure is nothing but the *Bayes posterior distribution* for a Gaussian process prior corresponding to the RKHS $\mathcal{H}_K$. Remarkably, this formulation can be applied to several problems other than training two layer neural networks:

- Ordinary nonparametric regression model: $\mathcal{W} = \mathbb{R}^d$, $\widetilde{\mathcal{W}} = \mathbb{R}$ and $f_W(x) = W(x)$.
- Two layer neural networks (continuous topology): $\mathcal{W} = \widetilde{\mathcal{W}} = \mathbb{R}^d$ and $f_W = \int_{\mathbb{R}^d} a(w)\sigma(W(w)^\top x)\mathrm{d}\rho_0(w)$.
- Two layer neural networks (discrete topology): $\mathcal{W} = \{1, 2, 3, \dots\}$, $\widetilde{\mathcal{W}} = \mathbb{R}^d$ and $f_W = \sum_{m=1}^\infty a_m\sigma(W(m)^\top x)$.
- Two layer neural networks (discrete topology): $\mathcal{W} = \{1, 2, 3, \dots\}$, $\widetilde{\mathcal{W}} = \mathbb{R}^d$ and $f_W = \sum_{m=1}^\infty a_m\sigma(W(m)^\top x)$.
- Deep neural networks (continuous topology): $\mathcal{W} = \mathbb{R}^d \times \{1, \dots, L\}$, $\widetilde{\mathcal{W}} = \mathbb{R}^d$ and

  $$f_W(x) = u^\top \left(\int_{\mathbb{R}^d} a_{w,L}\sigma(W(w,L)^\top \cdot)\mathrm{d}\rho_0(w)\right) \circ \cdots \circ \left(\int_{\mathbb{R}^d} a_{w,1}\sigma(W(w,1)^\top x)\mathrm{d}\rho_0(w)\right),$$

  where $u \in \mathbb{R}^d$ and $a_{w,\ell} \in \mathbb{R}^d$ for $w \in \mathbb{R}^d$ and $\ell \in \{1, \dots, L\}$.
- ResNet: $\mathcal{W} = \mathbb{R}^d \times \{1, \dots, T\}$, $\widetilde{\mathcal{W}} = \mathbb{R}^d$ and

  $$f_W(x){=}u^\top \left(\mathbb{I}{+}\int_{\mathbb{R}^d} a_{w,T}\sigma(W(w,T)^\top \cdot)\mathrm{d}\rho_0(w)\right) \circ \cdots \circ \left(\mathbb{I}{+}\int_{\mathbb{R}^d} a_{w,1}\sigma(W(w,1)^\top x)\mathrm{d}\rho_0(w)\right),$$

  where $u \in \mathbb{R}^d$ and $a_{w,t} \in \mathbb{R}^d$ for $w \in \mathbb{R}^d$ and $t \in \{1, \dots, T\}$.
- Wasserstein optimal transportation map: $\mathcal{W} = \widetilde{\mathcal{W}} = \mathbb{R}^d$ and $f_W(x) = W(x)$. For random variables X and Y obeying distributions P and Q respectively: $\mathcal{W}^2(P, Q) = \min_{W: Q = f_W \sharp P} \mathrm{E}_{X\sim P}[\|X - f_W(X)\|^2]$.

3 Optimization error bound of transportation map learning

To show convergence of the dynamics (6), we utilize the recent result given by [45]. Let $\|W\|_\varepsilon := \left(\sum_{k\geq 0}(\mu_k)^{2\varepsilon}\langle W, e_k\rangle_{\mathcal{H}}^2\right)^{1/2}$ and $P_N W := \sum_{k=0}^{N-1}\langle W, e_k\rangle_{\mathcal{H}} e_k$ for $W \in \mathcal{H}$ where $(e_k)_k$ is the orthonormal system of $\mathcal{H}$. Accordingly, let $\mathcal{H}_N$ be the image of P_N: $\mathcal{H}_N = P_N\mathcal{H}$.

Assumption 1.

(i) (Eigenvalue condition) *There exists a constant c_μ such that $\mu_k \leq c_\mu(k+1)^{-2}$.*

[1] More natural modeling would be that the regularization A and the covariance of ξ_t depend on the current solution W_t, but we consider this simplest model for technical tractability.

(ii) (Boundedness and Smoothness) *There exist $B, M > 0$ such that the gradient of the empirical risk is bounded by B and is M-Lipschitz continuous with $\alpha \in (1/4, 1)$ almost surely:*

$$\|\nabla \widehat{\mathcal{L}}(W)\|_{\mathcal{H}} \leq B \ (\forall W \in \mathcal{H}), \quad \|\nabla \widehat{\mathcal{L}}(W) - \nabla \widehat{\mathcal{L}}(W')\|_{\mathcal{H}} \leq L\|W - W'\|_{\alpha} \ (\forall W, W' \in \mathcal{H}).$$

(iii) (Third order smoothness [13, Assumption 2.7]) *Let $\widehat{\mathcal{L}}_N : \mathcal{H}_N \to \mathbb{R}$ be $\widehat{\mathcal{L}}_N = \widehat{\mathcal{L}}(P_N W)$. $\widehat{\mathcal{L}}$ is three times differentiable, and there exists $\alpha' \in [0, 1), C_{\alpha'} \in (0, \infty)$ such that for all $N \in \mathbb{N}$ and $\forall W, h, k \in \mathcal{H}_N$, $\|\nabla^3 \widehat{\mathcal{L}}_N(W) \cdot (h, k)\|_{\alpha'} \leq C_{\alpha'}\|h\|_{\mathcal{H}}\|k\|_{\mathcal{H}}$, $\|\nabla^3 \widehat{\mathcal{L}}_N(W) \cdot (h, k)\|_{\mathcal{H}} \leq C_{\alpha'}\|h\|_{-\alpha'}\|k\|_{\mathcal{H}}$ (a.s.), where $\nabla^3 \widehat{\mathcal{L}}_N(W)$ is the third-order derivative, we identify it with third-order linear form, and we also write $\nabla^3 \widehat{\mathcal{L}}_N(W) \cdot (h, k)$ for the Riesz representor of $l \in \mathcal{H} \mapsto \nabla^3 \widehat{\mathcal{L}}_N(W) \cdot (h, k, l)$.*

The first condition controls the strength of the regularization term. The second condition ensures the smoothness of the loss function that yields the *disspativity* condition of the objective combined with the regularization term. That is, the solution of the gradient Langevin dynamics can remain a bounded region with high probability. The Lipschitz continuity of the gradient is a bit strong condition because the right hand side appears a weaker norm $\|\cdot\|_{\alpha}$ than the canonical norm $\|\cdot\|_{\mathcal{H}}$. However, this gives the geometric ergodicity (exponential convergence to the stationary distribution) of the discrete time dynamics. The third condition is more technical assumption. This condition is used for bounding the continuous time dynamics and discrete time dynamics. Intuitively, a smoother loss function makes the two dynamics closer. In particular, $\eta^{1/2-a}$ term appearing in the following bound can be shown by this condition.

Then, we can show the following weak convergence rate. Let π_k be the probability measure on $\mathcal{H}$ corresponding to the distribution of W_k.

Proposition 1. *Assume Assumption 1 holds and $\beta > \eta$. Suppose that $\exists \bar{R} > 0$, $0 \leq \ell(Y, f_W(X)) \leq \bar{R}$ for any $W \in \mathcal{H}$ (a.s.). Let $\rho = \frac{1}{1+\lambda\eta/\mu_0}$ and $b = \frac{\mu_0}{\lambda}B + \frac{c_\mu}{\beta\lambda}$. Then, for $\Lambda^*_\eta = \frac{\min\left(\frac{\lambda}{2\mu_0}, \frac{1}{2}\right)}{4\log(\kappa(\bar{V}+1)/(1-\delta))}\delta$ and $C_{W_0} = \kappa[\bar{V} + 1] + \frac{\sqrt{2}(\bar{R}+b)}{\sqrt{\delta}}$ where $0 < \delta < 1$ satisfying $\delta = \Omega(\exp(-\Theta(\mathrm{poly}(\lambda^{-1})\beta)))$, $\bar{b} = \max\{b, 1\}$, $\kappa = \bar{b}+1$ and $\bar{V} = 4\bar{b}/(\sqrt{(1+\rho^{1/\eta})/2} - \rho^{1/\eta})$ (where $\bar{V} = 4\bar{b}/(\sqrt{(1+\exp(-\frac{\lambda}{\mu_1}))/2} - \exp(-\frac{\lambda}{\mu_1}))$ for $\eta = 0$), and for any $0 < a < 1/4$, the following convergence bound holds for almost sure observation D_n: for either $L = \mathcal{L}$ or $L = \widehat{\mathcal{L}}$,*

$$|\mathrm{E}_{W_k \sim \pi_k}[L(W_k)] - \mathrm{E}_{W \sim \pi_\infty}[L(W)]| \leq C_1 \left[C_{W_0} \exp(-\Lambda^*_\eta \eta k) + \frac{\sqrt{\beta}}{\Lambda^*_0}\eta^{1/2-a} \right] =: \Xi_k, \tag{7}$$

where C_1 is a constant depending only on $c_\mu, B, L, C_{\alpha'}, a, \bar{R}$ (independent of η, k, β, λ).

We utilized the theories of [45] as the core technique to show this proposition. Its complete proof is given in Appendix A. We can see that as k goes to infinity the first term of the right hand side converges exponentially, and as the step size η goes to 0, the second term converges arbitrary close to the rate of $\sqrt{\eta}$. It is known that the convergence rate with respect to η is optimal [15]. Therefore, if we choose sufficiently small η and sufficiently large k, we can sample W_k that obeys nearly the invariant measure π_∞. As we will see later, sample from π_∞ has a nice property in terms of generalization. As we have remarked in Remark 1, the convergence is guaranteed even for the finite width neural network setting, i.e., ρ_0 is a discrete distribution in the model (4). This is much advantageous against existing framework such as mean field analysis and NTK.

The above proposition gives a bound on the expectation of the loss of the solution W_k instead of a high probability bound. However, due to the geometric ergodicity of the dynamics, by running the algorithm for sufficiently large steps, we can show that the probability that there *does not* appear W_k in the trajectory that has a loss such that $L(W_k) - \mathrm{E}_{W \sim \pi_\infty}[L(W)] \leq O(\Xi_k)$ approaches 0 with exponential rate. Since this direction requires much more involved mathematics, we consider a simpler one as described above.

4 Generalization error analysis

Generalization gap bound Here, we analyze the generalization error of the solution of W_k obtained by the dynamics (6).

Theorem 1. *Assume Assumption 1 holds with $\beta > \eta$, and assume that the loss function is bounded, i.e., there exits $\bar{R} > 0$ such that $\forall W \in \mathcal{H},\ 0 \le \ell(Y, f_W(X)) \le \bar{R}$ (a.s.). Then, for any $1 > \delta > 0$, with probability $1-\delta$, the generalization error is bounded by*

$$\mathrm{E}_{W_k}[\mathcal{L}(W_k)] \le \mathrm{E}_{W_k}[\widehat{\mathcal{L}}(W_k)] + \frac{\bar{R}^2}{\sqrt{n}}\left[2\left(1+\frac{2\beta}{\sqrt{n}}\right) + \log\left(\frac{1+e^{\bar{R}^2/2}}{\delta}\right)\right] + 2\Xi_k.$$

The proof is given in Appendix B. To prove this, we used a PAC-Bayes stability bound [52]. From this theorem, we have that the generalization error is bounded by $O(1/\sqrt{n})$ and the optimization error Ξ_k. The $O(1/\sqrt{n})$ term is the generalization gap for the stationary distribution, and as k goes to infinity, the total generalization gap converges to this one. [44] also showed a PAC-Bayesian stability bound for a finite dimensional Langevin dynamics (roughly speaking, their bound is $O(\sqrt{\beta B^2/(n\lambda)})$), but their proof technique is quite different from ours. Our proof analyzes the generalization error under the stationary distribution of the dynamics and bounds the gap between the stationary distribution and the current solution, while [44] evaluated the bound by "accumulating" the error through the updates without analyzing the stationary distribution.

Excess risk bound: fast learning rate Next, we bound the excess risk. Unlike the $O(1/\sqrt{n})$ convergence rate of the generalization gap bound, we can derive a fast learning rate which is faster than $O(1/\sqrt{n})$ in a setting of realizable case, i.e., a student-teacher model, for the excess risk instead of the generalization gap. As a concrete example, we keep the following two layer neural network model in our mind. For a map $W : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, let a "clipped map" $\bar{W}$ be $\bar{W}(w) := R \times \tanh(W(w)/R)$, where $R \ge 1$ is a constant and $\tanh$ is applied elementwise. Then, the following two layer neural network model falls into our analysis:

$$f_W(x) := \textstyle\int_{\mathbb{R}\times\mathbb{R}^d} \bar{W}_2(a)\sigma(\bar{W}_1(w)^\top x)\mathrm{d}\rho_0(a,w) \tag{8}$$

for a measurable map $W = (W_1, W_2) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d \times \mathbb{R}$ and an activation function σ that is 1-Lipschitz continuous and included in a *Hölder class* $\mathcal{C}^3(\mathbb{R})$. Here, we used the clipping operation only for a technical reason because the current convergence analysis of the infinite dimensional Langevin dynamics requires a boundedness condition. This could be removed if we could show its convergence under more relaxed conditions. The fast learning rate analysis is not restricted to the two layer model, but it can be applied as long as the following statement is satisfied (e.g., ResNet).

Lemma 1. *For the model* (8)*, if $\|x\| \le D$ for any $x \in \mathrm{supp}(P_X)$, then it holds that $\|f_W - f_{W'}\|_\infty \le (1+RD)\|W - W'\|_{L_2(\rho_0)}$ where $\|W - W'\|_{L_2(\rho_0)}^2 := \int \|W((a,w)) - W'((a,w))\|^2 \mathrm{d}\rho_0(a,w)$.*

The proof is given in Appendix C. This lemma indicates that to estimate a function f_{W^*}, its estimation error can be bounded by the estimation error of the parameter W. To ensure the smooth gradient assumption (Assumption 1-(ii)) and precisely characterize the estimation accuracy by the model complexity, we consider an RKHS with "smoothness" parameter γ as the model of W. Let $T_K : \mathcal{H} \to \mathcal{H}$ be a linear bounded operator such that $\langle T_K h, h'\rangle_{\mathcal{H}} = \sum_{k=0}^\infty \mu_k \alpha_k \alpha'_k$ for $h = \sum_k \alpha_k e_k$ and $h' = \sum_k \alpha'_k e_k$. Let the range of power of T_K be $\mathcal{H}_{K^\gamma} = \{f = T_K^{\gamma/2} h \mid h \in \mathcal{H}\}$ for $\gamma > 0$ which is equipped with the inner product $\langle h, h'\rangle_{\mathcal{H}_{K^\gamma}} = \sum_{k=0}^\infty \mu_k^{-\gamma}\alpha_k\alpha'_k$. We can see that $\gamma = 1$ corresponds to $\mathcal{H}_K$ and γ controls the "complexity" of $\mathcal{H}_{K^\gamma}$, that is, if $\gamma < 1$, then $\mathcal{H}_K \hookrightarrow \mathcal{H}_{K^\gamma}$, and otherwise, $\mathcal{H}_{K^\gamma} \hookrightarrow \mathcal{H}_K$. We consider a problem of optimizing $\widehat{\mathcal{L}}(f_W)$ or $\mathcal{L}(f_W)$ with respect to W in the model $\mathcal{H}_{K^\gamma}$. To so so, by noticing that any $g \in \mathcal{H}_{K^\gamma}$ can be written as $g = T_K^{\gamma/2} W$ for $W \in \mathcal{H}$, we write the empirical and population risk with respect to $W \in \mathcal{H}$ as $\widehat{\mathcal{L}}(W) = \widehat{\mathcal{L}}(f_{T_K^{\gamma/2}W}),\ \mathcal{L}(W) = \mathcal{L}(f_{T_K^{\gamma/2}W})$. Let $f^* \in \mathrm{argmin}_f\, \mathcal{L}(f)$ where min is taken over all measurable functions and we assume the existence of the minimizer.

Assumption 2 (Bernstein condition and predictor condition [73, 7])**.** *The Bernstein condition is satisfied: there exist $C_B > 0$ and $s \in (0,1]$ such that for any f_W $(W \in \mathcal{H})$,*

$$\mathrm{E}[(\ell(Y, f_W(X)) - \ell(Y, f^*(X)))^2] \le C_B(\mathcal{L}(f_W) - \mathcal{L}(f^*))^s.$$

Moreover, we assume that, for any $h : \mathbb{R}^d \to \mathbb{R}$ and $x \in \mathrm{supp}(P_X)$, it holds that

$$\mathrm{E}_{Y|X=x}\left[\exp\left(-\tfrac{\beta}{n}(\ell(Y, h(x)) - \ell(Y, f^*(x)))\right)\right] \le 1.$$

The first assumption is called *Bernstein condition*. We can show that this condition is satisfied by the logistic loss and the squared loss with bounded f_W and f^* (Theorem 3). The second assumption is called *predictor condition* [73] and can be satisfied if ℓ is a log-likelihood function and the model is correctly specified (that is, the true conditional probability density (or probability mass) $p(y|x)$ is expressed as $p(y|x) \simeq \exp(-\ell(y, f^*(x)))$). To extend the theory to misspecified situations, we need the second assumption. For example, if we use a squared loss in a regression problem whereas the label noise is *not* Gaussian, then it is a misspecified situation but if the noise has a light tail (such as sub-Gaussian), then the assumption can be satisfied [73].

Our analysis is valid even if f^* cannot be represented by f_W for $W \in \mathcal{H}$. This model misspecification can be incorporated as bias-variance trade-off in the excess risk bound. This trade-off can be captured by the following *concentration function*. Let $\mathcal{H}_{\tilde{K}} = \mathcal{H}_{K^{\gamma+1}}$, and the Gaussian process law of $T_K^{\gamma/2}W$ for $W \sim \nu_\beta$ be $\tilde{\nu}_\beta$. Then, define the concentration function as

$$\phi_{\beta,\lambda}(\epsilon) := \inf_{h \in \mathcal{H}_{\tilde{K}} : \mathcal{L}(h) - \mathcal{L}(f^*) \le \epsilon^2} \beta\lambda \|h\|_{\mathcal{H}_{\tilde{K}}}^2 - \log \tilde{\nu}_\beta(\{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \le \epsilon\}) + \log(2),$$

where, if there does not exist $h \in \mathcal{H}_{\tilde{K}}$ satisfying the condition in inf, then we set $\phi_{\beta,\lambda}(\epsilon) = \infty$.

Theorem 2. *Assume that Assumption 2 holds, $\|x\| \le D$ ($\forall x \in \mathcal{X}$), $\gamma > 1/2$, $\beta > \eta$ and $\beta \le n$. Assume that the loss function $\ell(y, \cdot)$ is included in $\mathcal{C}^3(\mathbb{R})$ for any $y \in \mathrm{supp}(P_Y)$ and there exists $B > 0$ such that $|\frac{\partial^k}{\partial u^k}\ell(y,u)| \le B$ ($\forall u \in \mathbb{R}$ s.t. $|u| \le R$, $\forall y \in \mathrm{supp}(P_Y)$, $k = 1,2,3$). Assume also that $0 \le \ell(Y, f(X)) \le \bar{R}$ (a.s.) for any $f = f_W$ ($W \in \mathcal{H}$) and $f = f^*$, and $\bar{\ell}_x(u) := \mathrm{E}_{Y|X=x}[\ell(Y,u)]$ satisfies $|\frac{\mathrm{d}\bar{\ell}_x}{\mathrm{d}u}(u) - \frac{\mathrm{d}\bar{\ell}_x}{\mathrm{d}u}(u')| \le L|u-u'|$ ($\forall u, u' \in \mathbb{R}, \forall x \in \mathcal{X}$) for a constant $L > 0$. Let $\tilde{\alpha} := 1/\{2(\gamma+1)\}$ and θ be an arbitrary real number satisfying $0 < \theta < 1 - \tilde{\alpha}$. We define $\epsilon^* := \inf\{\epsilon > 0 : \phi_{\beta,\lambda}(\epsilon) \le \beta\epsilon^2\} \vee n^{-\frac{1}{2-s}}$. Then, the expected excess risk is bounded as*

$$\mathrm{E}_{D^n}\left[\mathrm{E}_{W_k}[\mathcal{L}(W_k)] - \mathcal{L}(f^*)\right] \le C\Big[\epsilon^{*2} \vee \big(\tfrac{\beta}{n}\epsilon^{*2} + n^{-\frac{1}{1+\tilde{\alpha}/\theta}}(\lambda\beta)^{\frac{2\tilde{\alpha}/\theta}{1+\tilde{\alpha}/\theta}}\big)^{\frac{1}{2-s}} \vee \frac{1}{n}\Big] + \Xi_k, \qquad (9)$$

where C is a constant independent of $n, \beta, \lambda, \eta, k$.

The proof is given in Appendix D.2. It is proven by using the technique of nonparametric Bayes contraction rate analysis [25, 71, 72]. However, we cannot adapt these existing techniques because (i) the loss function is not necessarily the log-likelihood function, (ii) the inverse temperature is generally different from the sample size. In that sense, our proof is novel to derive an excess risk for (i) a misspecified model, and (ii) a randomized estimator with a general inverse temperature parameter.

The bound is about expectation of the excess risk instead of high probability bound. However, a high probability bound is also provided in the proof and the expectation bound is derived from the high probability bound.

If the bias is not zero, i.e., $\inf_{W \in \mathcal{H}} \mathcal{L}(W) - \mathcal{L}(f^*) = \delta_0 > 0$, then we may choose $\epsilon^{*2} = \Theta(\delta_0)$ because $\phi_{\beta,\lambda}(\epsilon)$ is finite for $\epsilon^2 > \delta_0$ and infinite for $\epsilon^2 < \delta_0$. Thus, a misspecified setting is covered.

(i) Example of fast rate: Regression Here, we apply our general result to a nonparametric regression problem by the neural network model. We consider the following nonparametric regression model: $y_i = f_{W^*}(x_i) + \epsilon_i$, for $W^* \in \mathcal{H}$ where ϵ_i is an i.i.d. noise with mean 0 and $|\epsilon_i| \le C < \infty$ (a.s.). To estimate f_{W^*}, we employ the squared loss $\ell(y,f) = (y-f)^2$. Then, we can easily confirm that f^* is achieved by f_{W^*} via a simple calculation: $\mathrm{argmin}_f\, \mathcal{L}(f) = f_{W^*}$. Moreover, for the squared loss, $s = 1$ is satisfied as remarked just after Assumption 2. Moreover, we further assume that $W^* \in \mathcal{H}_{K^{\theta(\gamma+1)}}$ for $\theta < 1 - \tilde{\alpha}$. Then, the "bias" and "variance" terms can be evaluated as $\inf_{h \in \mathcal{H}_{\tilde{K}} : \mathcal{L}(h) - \mathcal{L}(f^*) \le \epsilon^2} \lambda\beta \|h\|_{\mathcal{H}_{\tilde{K}}}^2 \lesssim \lambda\beta\epsilon^{-\frac{2(1-\theta)}{\theta}}$ and $-\log \tilde{\nu}_\beta(\{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \le \epsilon\}) \lesssim (\epsilon/(\lambda\beta)^{1/2})^{-\frac{2\tilde{\alpha}}{1-\tilde{\alpha}}}$. Accordingly, we can show the following excess risk bound:

$$\mathrm{E}_{D^n}\left[\mathrm{E}_{W_k}[\mathcal{L}(W_k)] - \mathcal{L}(f^*)\right] \lesssim \max\left\{(\lambda\beta)^{\frac{2\tilde{\alpha}/\theta}{1+\tilde{\alpha}/\theta}} n^{-\frac{1}{1+\tilde{\alpha}/\theta}}, \lambda^{-\tilde{\alpha}}\beta^{-1}, \lambda^{\theta}, 1/n\right\} + \Xi_k, \qquad (10)$$

(see Appendix D.4 for the derivation). In particular, if $\beta = \lambda^{-1} = n$, then this convergence rate can be rewritten as $\max\{n^{-\frac{1}{1+\tilde{\alpha}/\theta}}, n^{-\theta}\} = n^{-\theta}$ ($\because \theta < 1 - \tilde{\alpha}$), which can be faster than $1/\sqrt{n}$ and is controlled by the "difficulty" of the problem $\tilde{\alpha}$ and θ.

Remark 2. *As an example, if the RKHS $\mathcal{H}_K$ is a Sobolev space $W_2^{a+d/2}(\mathbb{R}^d)$ with regularity parameter $a + d/2$ (more precisely, each output $W_i(\cdot)$ is a member of a Sobolev space) and $\mathcal{H}$ is $L_2(\rho_0)$, then we can set $\tilde{\alpha} = \frac{d}{2a+d}$. If the true parameter W^* is included in another Sobolev space $W_2^b(\mathbb{R}^d)$ for $b \leq a$, then we may choose $\theta = 2b/(2a + d)$ and the convergence rate is bounded by $n^{-2b/(2a+d)}$, which coincides with the posterior contraction rate of Gaussian process estimator derived in [72]. It is known that, if $a = b$, this achieves the* minimax optimal rate *[78].*

(ii) Example of fast rate: Classification (exponential convergence) Here, we consider a binary classification problem $y \in \{\pm 1\}$. We employ the logistic loss function $\ell(y, f) = \log(1+\exp(-yf))$ for $y \in \{\pm 1\}$ and $f \in \mathbb{R}$. Corresponding to the loss function, we define the expected loss conditioned by $X = x$ as $h(u|x) = \mathrm{E}[\ell(Y, u)|X = x]$. Note that $h(0|x) = \log(2)$. We assume that the strength of noise of this binary classification problem is low as follows.

Assumption 3 (Strong low noise condition)**.** *Let $h^*(x) := \inf_{u \in \mathbb{R}} h(u|x)$. Assume that there exists $\delta > 0$ such that $h^*(x) \leq \log(2) - \delta$ $(\forall x \in \mathcal{X})$. Moreover, there exists $W^* \in \mathcal{H}$ such that $f^* = f_{W^*}$, that is, $\sup_{x \in \mathrm{supp}(P_X)} |h(f_{W^*}(x)|x) - h^*(x)| = 0$.*

The first assumption is satisfied if the label probability is away from the even probability $1/2$: $|P(Y|X = x) - 1/2| > \Omega(\sqrt{\delta})$. This condition means that the class label has less noisy than completely random labeling. In that sense, we call this assumption the *strong low noise condition*, which has been analyzed in [36, 4, 49]. A weaker low noise condition was introduced by [70] as Tsybakov's low noise condition. The second assumption can be relaxed to the existence of W only for some $\epsilon > c_0\delta$ with sufficiently small c_0, but we don't pursuit this direction for simplicity.

Assumption 4. *Assume $\mathcal{X}(= \mathrm{supp}(P_X)) \subset [0, 1]^d$ and $\mathcal{X}$ is a minimally smooth domain in a sense of [61]. P_X has a density $p(x)$ which is lower bounded as $p(x) \geq c_0$ $(\forall x \in \mathrm{supp}(P_X))$ on its support. For $2m > d$ and $m \geq 3$, the activation function satisfies $\sigma \in \mathcal{C}^m(\mathbb{R})$ and f^* is included in the Sobolev space $W_2^m(\mathcal{X})$ defined on $\mathcal{X}$ (see [21] for its definition).*

The following theorem gives an upper-bound of the probability of "perfect classification" for the estimator. More specifically, it shows the error probability converges in an *exponential rate*.

Theorem 3. *Under Assumptions 3 and 4, the convergence in Theorem 2 holds for $s = 1$. Let $g^*(x) = \mathrm{sign}(P(Y = 1|X = x) - 1/2)$ be the Bayes classifier. If the sample size n is sufficiently large and λ, β are appropriately chosen, then the classification error converges exponentially with respect to β and k:*

$$\mathrm{E}[\pi_k(\{W_k \in \mathcal{H} \mid P_X(\mathrm{sign}(f_{W_k}(X)) = g^*(X)) \neq 1\})] \lesssim \frac{\Xi_k}{\delta^{2m/(2m-d)}} + \exp(-c'\beta\delta^{\frac{2m}{2m-d}}).$$

The proof is given in Appendix D.3. This theorem states that if we choose the step size η sufficiently small, then the error probability converges exponentially as k and β increase. Even if the first term of the right hand side is larger than the second term, we can make this as small as the second term by running the algorithm several times and picking up the best one with respect to validation error if $\Xi_k \ll 1$ (see Appendix D.3 for this discussion).

5 Conclusion

In this paper, we have formulated the deep learning training as a transportation map estimation and analyzed its convergence and generalization error through the infinite dimensional Langevin dynamics. Unlike exiting analysis, our formulation can incorporate spatial correlation of noise and achieve global convergence without taking the limit of infinite width. The generalization analysis reveals the dynamics achieves a stable estimator with $O(1/\sqrt{n})$ convergence of generalization error and shows fast learning rate of the excess risk. Finally, we have shown a convergence rate of excess risk for regression and classification. The rate for regression recovers the minimax optimal rate known in Bayesian nonparametrics and that for classification achieves exponential convergence under the strong low noise condition.

Broader impact

Benefit Since deep learning is used in several applications across broad range of areas, our theoretical analysis about optimization of deep learning would influence wide range of areas in terms of understanding of the algorithmic behavior. One of the biggest criticisms on deep learning is its poor explainability and interpretability. Our work on optimization analysis of deep learning can much improve explainability and would facilitate its usage. This is quite important step toward trustworthy machine learning.

Potential risk On the other hand, this is purely theoretical work and thus would not directly bring on severe ethical issues. However, misunderstanding of theoretical work would cause misuse of its statement to conduct an intensional opinion making. To avoid such a potential risk, we made our best effort to minimize technical ambiguity in our paper presentation.

Acknowledgment

I would like to thank Atsushi Nitanda for insightful comments. TS was partially supported by JSPS KAKENHI (18K19793,18H03201, and 20H00576), Japan Digital Design, and JST CREST.

References

[1] Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of International Conference on Machine Learning*, pp. 242–252, 2019.

[2] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263. PMLR, 2018.

[3] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 2019.

[4] J.-Y. Audibert, A. B. Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

[5] P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1487–1537, 2005.

[6] P. Bartlett, D. J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

[7] P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006.

[8] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6241–6250, 2017.

[9] C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.

[10] C. Borell. The Brunn-Minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2): 207–216, 1975.

[11] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[12] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical process. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 334:495–500, 2002.

[13] C.-E. Bréhier and M. Kopec. Approximation of the invariant law of SPDEs: error analysis using a Poisson equation for a full-discretization scheme. *IMA Journal of Numerical Analysis*, 37(3):1375–1410, 07 2016.

[14] C.-E. Bréhier. Lecture notes: Invariant distributions for parabolic SPDEs and their numerical approximations, November 2017. HAL ID: cel-01633504.

[15] C.-E. Bréhier. Influence of the regularity of the test functions for weak convergence in numerical discretization of spdes. *Journal of Complexity*, 56:101424, 2020.

[16] Y. Cao and Q. Gu. A generalization theory of gradient descent for learning over-parameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*, 2019.

[17] Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*, 2019.

[18] L. Chizat and F. Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[19] G. Da Prato and J. Zabczyk. Non-explosion, boundedness and ergodicity for stochastic semilinear equations. *Journal of Differential Equations*, 98:181–195, 1992.

[20] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2 edition, 2014.

[21] R. A. DeVore and R. C. Sharpley. Besov spaces on domains in $\mathbb{R}^d$. *Transactions of the American Mathematical Society*, 335(2):843–864, 1993.

[22] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685, 2019.

[23] S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations 7*, 2019.

[24] M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems 31*, pp. 9671–9680. 2018.

[25] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

[26] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

[27] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

[28] M. Hairer. Exponential mixing properties of stochastic PDEs through asymptotic coupling. *Probability Theory and Related Fields*, 124(3):345–380, 2002.

[29] S. Hayakawa and T. Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020.

[30] M. Imaizumi and K. Fukumizu. Deep neural networks learn non-smooth functions effectively. In K. Chaudhuri and M. Sugiyama (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 869–878. PMLR, 16–18 Apr 2019.

[31] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[32] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 8580–8589, 2018.

[33] S. Jacquot and G. Royer. Ergodicité d'une classe d'équations aux dérivées partielles stochastiques. *C. R. Acad. Sci. Paris Sér. I Math.*, 320(2):231–236, 1995.

[34] Z. Ji and M. Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. *arXiv preprint arXiv:1909.12292*, 2019.

[35] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006.

[36] V. Koltchinskii and O. Beznosova. Exponential convergence rates in classification. In P. Auer and R. Meir (eds.), *Learning Theory*, pp. 295–307, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[37] A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

[38] R. Latała and D. Matlak. *Royen's Proof of the Gaussian Correlation Inequality*, pp. 265–275. Springer International Publishing, 2017.

[39] B. Maslowski. Strong Feller property for semilinear stochastic evolution equations and applications. In *Stochastic systems and optimization (Warsaw, 1988)*, volume 136 of *Lect. Notes Control Inf. Sci.*, pp. 210–224. Springer, Berlin, 1989.

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

[41] S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In A. Beygelzimer and D. Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2388–2464, Phoenix, USA, 25–28 Jun 2019. PMLR.

[42] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48:1977–1991, 2002.

[43] M. Mohri, A. Rostamizadeh, and A. Talwalkar. Foundations of machine learning. 2012.

[44] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 605–638. PMLR, 2018.

[45] B. Muzellec, K. Sato, M. Massias, and T. Suzuki. Dimension-free convergence rates for gradient Langevin dynamics in RKHS. *arXiv preprint 2003.00306*, 2020.

[46] B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Proceedings of The 28th Conference on Learning Theory*, pp. 1376–1401, Montreal Quebec, 2015.

[47] A. Nitanda and T. Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

[48] A. Nitanda and T. Suzuki. Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

[49] A. Nitanda and T. Suzuki. Stochastic gradient descent with exponential convergence rates of expected classification errors. In K. Chaudhuri and M. Sugiyama (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1417–1426. PMLR, 16–18 Apr 2019.

[50] S. Oymak and M. Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[51] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. volume 65 of *Proceedings of Machine Learning Research*, pp. 1674–1703. PMLR, 2017.

[52] O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. In *Advances in Neural Information Processing Systems 34*. 2020. to appear.

[53] G. Rotskoff and E. Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7146–7155. Curran Associates, Inc., 2018.

[54] G. M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

[55] T. Royen. A simple proof of the gaussian correlation conjecture extended to multivariate gamma distributions. *arXiv preprint arXiv:1408.1028*, 2014.

[56] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4), 2020.

[57] T. Shardlow. Geometric ergodicity for stochastic PDEs. *Stochastic Analysis and Applications*, 17(5):857–869, 1999.

[58] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks. *arXiv preprint arXiv:1805.01053*, 2018.

[59] R. Sowers. Large deviations for the invariant measure of a reaction-diffusion equation with non-Gaussian perturbations. *Probability Theory and Related Fields*, 92(3):393–421, 1992.

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[61] E. M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, 1970.

[62] I. Steinwart. Convergence types and rates in generic karhunen-loève expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395, 2019.

[63] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[64] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory*, pp. 79–93, 2009.

[65] T. Suzuki. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

[66] T. Suzuki and A. Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. *arXiv preprint arXiv:1910.12799*, 2019.

[67] T. Suzuki, H. Abe, and T. Nishimura. Compression based bound for non-compressed network: Unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020.

[68] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

[69] H. Triebel. *Theory of Function Spaces*. Monographs in Mathematics. Birkhäuser Verlag, 1983.

[70] A. B. Tsybakov et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[71] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

[72] A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

[73] T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

[74] S. Wager, S. Wang, and P. S. Liang. Dropout training as adaptive regularization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 351–359. Curran Associates, Inc., 2013.

[75] M. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[76] E. Weinan, C. Ma, and L. Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics*, pp. 1–24, 2019.

[77] M. Welling and Y.-W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, pp. 681–688, 2011.

[78] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

[79] D. Zou and Q. Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2053–2062, 2019.

——Appendix——

A Proof of Proposition 1

We apply the result [45]. Let $\{Z_n\}_{n\in\mathbb{N}}$ be a dynamics obeying

$$Z_{n+1} = S_\eta Z_n + \sqrt{\eta/\beta} S_\eta \varepsilon_n,$$

with $Z_0 = 0$. Let $k(p) := \sup_{n\geq 0} \mathrm{E}(\|Z_n\|^p)$ for $p > 0$, then it is known that $k(p) < \infty$ for any $p > 0$. Let $\{Z_n\}_{n\in\mathbb{N}}$ solve $Z_0 = 0$ and with $\beta > \eta$. Then, we can show that $k(p) := \sup_{n\geq 0} \mathrm{E}(\|Z_n\|^p) < \infty$ $(\forall p > 0)$[45]. Using $k(p)$, we define $b' = \frac{\mu_0}{\lambda} B + k(1)$. We will show that $k(1) \leq \frac{c_\mu}{\beta\lambda}$. Then, we can see that $b' \leq b$. Now, we show $k(1) \leq \frac{c_\mu}{\beta\lambda}$. First, note that

$$Z_n = \sqrt{\frac{\eta}{\beta}} \sum_{\ell=0}^{n-1} S_\eta^{n-\ell} \epsilon_\ell.$$

Therefore, we have

$$\begin{aligned}
\mathrm{E}[\|Z_n\|^2] &= \frac{\eta}{\beta} \sum_{\ell=0}^{n-1} \mathrm{Tr}[S_\eta^{2(n-\ell)}] = \frac{\eta}{\beta} \mathrm{Tr}\left[(S_\eta^2 - S_\eta^{2n})(I - S_\eta^2)^{-1}\right] \leq \frac{\eta}{\beta} \mathrm{Tr}\left[S_\eta^2 (I - S_\eta^2)^{-1}\right] \\
&= \frac{\eta}{\beta} \sum_{k=0}^{\infty} \left(\frac{1}{1+\eta\lambda/\mu_k}\right)^2 \left(1 - \frac{1}{(1+\eta\lambda/\mu_k)^2}\right)^{-1} = \frac{\eta}{\beta} \sum_{k=0}^{\infty} \frac{1}{(1+\eta\lambda/\mu_k)^2 - 1} \\
&\leq \frac{\eta}{\beta} \sum_{k=0}^{\infty} \frac{1}{2\eta\lambda/\mu_k} \leq \frac{1}{2\beta\lambda} \sum_{k=0}^{\infty} c_\mu (k+1)^{-2} \leq \frac{c_\mu}{\beta\lambda}.
\end{aligned}$$

Then, Jensen's inequality yields $k(1) = \mathrm{E}[\|Z_n\|_{\mathcal{H}}] \leq \sqrt{\mathrm{E}[\|Z_n\|_{\mathcal{H}}^2]} \leq \sqrt{\frac{c_\mu}{\beta\lambda}}$.

Let $\phi : \mathcal{H} \to \mathbb{R}$ be a test function satisfying $|\phi(\cdot)| \leq V(\cdot)$ and $\|\phi(x) - \phi(y)\| \leq M\|x - y\|$ $(x, y \in \mathcal{H})$ for $M > 0$. Then, [45] showed that there exists a unique invariant measure μ_η and the following exponential convergence of the expectation of ϕ holds:

$$|\mathrm{E}_{x_0}[\phi(X_n)] - \mathrm{E}[\phi(X^{\mu_\eta})]| \leq C_{x_0} \exp(-\Lambda_\eta^*(\eta n - 1)). \tag{11}$$

where

$$\Lambda_\eta^* = \frac{\min\left(\frac{\lambda}{2\mu_0}, \frac{1}{2}\right)}{4\log(\kappa(\bar{V}+1)/(1-\delta))}\delta,\ C_{W_0} = \kappa[\bar{V}+1] + \frac{\sqrt{2}(\bar{R}+b)}{\sqrt{\delta}}$$

with $0 < \delta < 1$ satisfying $\delta = \Omega(\exp(-C'\mathrm{poly}(\lambda^{-1})\beta))$, $\bar{b} = \max\{b, 1\}$, $\kappa = \bar{b} + 1$ and $\bar{V} = \frac{4\bar{b}}{\sqrt{(1+\rho^{1/\eta})/2 - \rho^{1/\eta}}}$. To show this we note that β in [45] is 2β using β in this paper. The definition of δ is not explicitly shown in [45] (in particular, λ is omitted), but we can recover our definition from the proof. Moreover, [45] assumed that there exists $\lambda_0, C_{\alpha,2} \in (0, \infty)$ such that

$$\begin{aligned}
&\|\nabla\widehat{\mathcal{L}}(W) - \nabla\widehat{\mathcal{L}}(W')\|_{\mathcal{H}} \leq L\|W - W'\|_{\mathcal{H}}\ \ (\forall W, W' \in \mathcal{H}), \\
&|\nabla^2\widehat{\mathcal{L}}(W) \cdot (h, k)| \leq C_{\alpha,2}\|h\|_{\mathcal{H}}\|k\|_\alpha\ \ (\forall W, h, k \in \mathcal{H}),
\end{aligned}$$

instead of our assumption $\|\nabla\widehat{\mathcal{L}}(W) - \nabla\widehat{\mathcal{L}}(W')\|_{\mathcal{H}} \leq L\|W - W'\|_\alpha$. However, we can see that their proof is valid even under our assumption.

Let C_b^2 be a set of functions $f : \mathcal{H} \to \mathbb{R}$ that is continuously twice differentiable with bounded derivatives. Under the same setting above, [45] also showed that, for any $0 < \kappa < 1/2$, $0 < \eta_0$, there exists a constant C such that, if the test function ϕ satisfies $\phi \in C_b^2$, then for any $0 < \eta < \eta_0$, it holds that

$$|\mathrm{E}[\phi(X^{\mu_\eta})] - \mathrm{E}[\phi(X^{\pi_\infty})]| \leq C\frac{\|\phi\|_{0,2}}{\Lambda_0^*} c_\beta \eta^{1/2-\kappa}, \tag{12}$$

where $\|\phi\|_{0,2} := \max\{\|\phi\|_\infty, \sup_{x\in\mathcal{H}} \|\nabla\phi(x)\|_{\mathcal{H}}, \sup_{x\in\mathcal{H}} \|\nabla^2\phi(x)\|_{\mathcal{B}(\mathcal{H})}\}$ for $\phi \in C_b^2$ where $\|\cdot\|_{\mathcal{B}(\mathcal{H})}$ is the norm as a linear operator.

Thus, if we let $\phi(\cdot) = \widehat{\mathcal{L}}(\cdot)/\bar{R}$, then ϕ satisfies the assumption with $M = B/\bar{R}$. Therefore, we obtain that

$$|\mathrm{E}_{x_0}[\widehat{\mathcal{L}}(X_n)] - \mathrm{E}[\widehat{\mathcal{L}}(X^{\pi_\infty})]| \le \bar{R}\left[C_{x_0}\exp(-\Lambda^*_\eta(\eta n - 1)) + C\frac{\|\phi\|_{0,2}}{\Lambda^*_0}c_\beta \eta^{1/2-\kappa}\right].$$

This gives the assertion.

Finally, we would like to note that since the assumption is satisfied almost surely, $\mathcal{L}$ also satisfies the assumption instead of $\widehat{\mathcal{L}}$. That means the same convergence rate holds also for $\phi(W) = \mathcal{L}$.

B Proof of Theorem 1

Proof. [52] proved that, for any probability measure Q which is absolutely continuous to π_∞, it holds that

$$\mathrm{E}_{W\sim Q}[\mathcal{L}(W)] \le \mathrm{E}_{W\sim Q}[\widehat{\mathcal{L}}(W)] + \frac{1}{\sqrt{n}}\mathrm{KL}(Q||\pi_\infty) + \frac{\bar{R}^2}{\sqrt{n}}\left[2\left(1+\frac{2\beta}{\sqrt{n}}\right) + \log\left(\frac{1+e^{\bar{R}^2/2}}{\delta}\right)\right], \tag{13}$$

with probability $1-\delta$.

On the other hand, Proposition 1 gives that

$$|\mathrm{E}_{W_k\sim\pi_k}[\mathcal{L}(W_k)] - \mathrm{E}_{W\sim\pi_\infty}[\mathcal{L}(W)]| \le \Xi_k,$$
$$|\mathrm{E}_{W_k\sim\pi_k}[\widehat{\mathcal{L}}(W_k)] - \mathrm{E}_{W\sim\pi_\infty}[\widehat{\mathcal{L}}(W)]| \le \Xi_k.$$

Then, by substituting $Q = \pi_\infty$ into Eq. (13) and applying the two inequalities above, we obtain the assertion. □

C Proof of Lemma 1

By the definition of f_W, we have

$$\begin{aligned}
&f_W(x) - f_{W'}(x)\\
&\le \int \left[(\bar{W}_2(a) - \bar{W}_2'(a))\sigma(\bar{W}_1(w)^\top x) + \bar{W}_2'(a)(\sigma(\bar{W}_1(w)^\top x) - \sigma(\bar{W}_1'(w)^\top x))\right]\mathrm{d}\rho_0(a,w)\\
&\le \sqrt{\int (\bar{W}_2(a) - \bar{W}_2'(a))^2 \mathrm{d}\rho_0(a,w)} + \sqrt{\int (\bar{W}_2'(a))^2(\sigma(\bar{W}_1(w)^\top x) - \sigma(\bar{W}_1'(w)^\top x))^2\mathrm{d}\rho_0(a,w)}\\
&\le \sqrt{\int (W_2(a) - W_2'(a))^2 \mathrm{d}\rho_0(a,w)} + R\sqrt{\int (\sigma(\bar{W}_1(w)^\top x) - \sigma(\bar{W}_1'(w)^\top x))^2\mathrm{d}\rho_0(a,w)},
\end{aligned}$$

where we used $|\sigma(x)| \le 0$, 1-Lipschitz continuity of the clipping operation and $|\bar{W}_2'(a)| \le R$. By noticing that $\|x\| \le D$ and σ and the clipping operation $W(w) \mapsto \bar{W}(w)$ are 1-Lipschitz continuous, the right hand side can be further bounded by

$$\begin{aligned}
&\sqrt{\int (W_2(a) - W_2'(a))^2 \mathrm{d}\rho_0(a,w)} + R\sqrt{\int (\bar{W}_1(w)^\top x - \bar{W}_1'(w)^\top x)^2\mathrm{d}\rho_0(a,w)}\\
&\le \sqrt{\int (W_2(a) - W_2'(a))^2 \mathrm{d}\rho_0(a,w)} + RD\sqrt{\int \|\bar{W}_1(w) - \bar{W}_1'(w)\|^2\mathrm{d}\rho_0(a,w)}\\
&\le \sqrt{\int (W_2(a) - W_2'(a))^2 \mathrm{d}\rho_0(a,w)} + RD\sqrt{\int \|W_1(w) - W_1'(w)\|^2\mathrm{d}\rho_0(a,w)}
\end{aligned}$$

$$\leq (1+RD)\sqrt{\int (W_2(a) - W_2'(a))^2 + \|W_1(w) - W_1'(w)\|^2 \mathrm{d}\rho_0(a,w)}$$

$$\leq (1+RD)\sqrt{\int \|W((a,w)) - W'((a,w))\|^2 \mathrm{d}\rho_0(a,w)} = (1+RD)\|W - W'\|_{L_2(\rho_0)}.$$

This gives the assertion.

D Proof of fast rate of excess risk bounds

D.1 Gaussian correlation inequality

Lemma 2 (Gaussian correlation inequality)**.** *Let $\tilde{\mathcal{H}}$ be a separable Hilbert space equipped with the complete orthonormal system $(e_i)_{i=1}^{\infty}$, and suppose that ν is a Gaussian measure in $\tilde{\mathcal{H}}$ with mean 0 and covariance $\Sigma = \mathrm{diag}\,(\mu_1, \mu_2, \dots)$ with respect to CONS $(e_i)_i$ where $\sum_{i=1}^n \mu_i^2 < \infty$, that is, ν is the distribution corresponding to $\sum_{i=1}^{\infty} \xi_i \sqrt{\mu_i} e_i$ for $\xi_i \sim N(0,1)$ (i.id.). Let $\mathcal{A}^1 = \{\sum_{i=1}^{\infty} \alpha_i e_i \in \tilde{\mathcal{H}} \mid \sum_{i=1}^{\infty} a_i \alpha_i^2 \leq 1, \alpha_i \in \mathbb{R}\}$ for $a_i \geq 0$ $(i = 1, 2, \dots)$ and $\mathcal{A}^2 = \{\sum_{i=1}^{\infty} \alpha_i e_i \in \tilde{\mathcal{H}} \mid \sum_{i=1}^{\infty} b_i \alpha_i^2 \leq 1, \alpha_i \in \mathbb{R}\}$ for $b_i \geq 0$ $(i = 1, 2, \dots)$. Then, we have*

$$\nu(\mathcal{A}^1 \cap \mathcal{A}^2) \geq \nu(\mathcal{A}^1)\nu(\mathcal{A}^2).$$

Proof. Let $\mathcal{A}_n^1$ an $\mathcal{A}_n^2$ be the cylinder set that "truncates" $\mathcal{A}^1$ an $\mathcal{A}^2$ up to index n: $\mathcal{A}_n^1 = \{\sum_{i=1}^{\infty} \alpha_i e_i \in \mathcal{H} \mid \sum_{i=1}^{n} a_i \alpha_i^2 \leq 1\}$ and $\mathcal{A}_n^2 = \{\sum_{i=1}^{\infty} \alpha_i e_i \in \mathcal{H} \mid |\sum_{i=1}^{n} b_i \alpha_i^2 \leq 1\}$. By the Gaussian correlation inequality [55, 38], it holds that

$$\nu(\mathcal{A}_n^1 \cap \mathcal{A}_n^2) \geq \nu(\mathcal{A}_n^1)\nu(\mathcal{A}_n^2).$$

Note that we can apply the Gaussian correlation inequality for a finite dimensional Gaussian measure. Next, we extend this inequality to the infinite dimensional space. Since $(\mathcal{A}_n^1)_n$ is a monotonically decreasing sequence, i.e., $\mathcal{A}_n^1 \subseteq \mathcal{A}_m^1$ for $m < n$, and $\cap_{n=1}^{\infty} \mathcal{A}_n^1 = \mathcal{A}^1$, the continuity of probability measure gives that $\lim_{n\to\infty} \nu(\mathcal{A}_n^1) = \nu(\mathcal{A}^1)$. Similarly, it holds that $\lim_{n\to\infty} \nu(\mathcal{A}_n^2) = \nu(\mathcal{A}^2)$.

Since $\mathcal{A}^2 \subset \mathcal{A}_n^1$ and $\mathcal{A}^2 \subset \mathcal{A}_n^2$, it holds that $\nu(\mathcal{A}^1 \cap \mathcal{A}^2) \leq \nu(\mathcal{A}_n^1 \cap \mathcal{A}_n^2)$. On the other hand, we also have

$$\begin{aligned}
\nu(\mathcal{A}_n^1 \cap \mathcal{A}_n^2) &= \nu((\mathcal{A}^1 \cup (\mathcal{A}_n^1 \backslash \mathcal{A}^1)) \cap (\mathcal{A}^2 \cup (\mathcal{A}_n^2 \backslash \mathcal{A}^2))) \\
&\leq \nu((\mathcal{A}^1 \cap \mathcal{A}^2) \cup (\mathcal{A}_n^1 \backslash \mathcal{A}^1) \cup (\mathcal{A}_n^2 \backslash \mathcal{A}^2)) \\
&\leq \nu(\mathcal{A}^1 \cap \mathcal{A}^2) + \nu(\mathcal{A}_n^1 \backslash \mathcal{A}^1) + \nu(\mathcal{A}_n^2 \backslash \mathcal{A}^2) \\
&\to \nu(\mathcal{A}^1 \cap \mathcal{A}^2).
\end{aligned}$$

Therefore, we have that

$$\lim_{n\to\infty} \nu(\mathcal{A}_n^1 \cap \mathcal{A}_n^2) = \nu(\mathcal{A}^1 \cap \mathcal{A}^2).$$

Combining all these arguments, we finally have that

$$\nu(\mathcal{A}^1 \cap \mathcal{A}^2) \geq \nu(\mathcal{A}^1)\nu(\mathcal{A}^2).$$

□

D.2 Proof of general excess risk bound (Theorem 2)

Proof. Since $\gamma > 1/2$, σ and $\ell(y, \cdot)$ are in $\mathcal{C}^3(\mathbb{R})$ where ℓ has a bounded partial derivative on a bounded domain, we can easily verify that the empirical risk satisfies Assumption 1 by noticing the clipping operation in the model.

For $0 < \theta < 1$, let $\mathcal{H}_{\tilde{K}^\theta} := \mathcal{H}_{K^{\theta(\gamma+1)}}$. It is known that if the natural inclusion $I_{\tilde{K},\tilde{K}^\theta} : \mathcal{H}_{\tilde{K}} \to \mathcal{H}_{\tilde{K}^\theta}$ is Hilbert-Schmidt, then the sample path of $\tilde{\nu}_\beta$ is included in $\mathcal{H}_{\tilde{K}^\theta}$ probability 1 (Theorem 5.2 of [62]). In our case, since $\mu_k \lesssim 1/k^2$, the eigenvalues $(\mu_k(\tilde{K}))_{k=1}^{\infty}$ of $\tilde{K}$ satisfies

$\mu_k(\tilde{K}) \lesssim 1/k^{2(\gamma+1)}$. Theorem 5.2 of [62] also states that $I_{\tilde{K},\tilde{K}^\theta}$ is Hilbert-Schmidt if and only if $\sum_{k=0}^\infty \mu_k(\tilde{K})^{1-\theta} < \infty$. Therefore, by setting $\tilde{\alpha} := 1/\{2(\gamma+1)\}$, $\theta < 1-\tilde{\alpha}$ is sufficient for this property. From now on, we assume that $\theta < 1-\tilde{\alpha}$. For notational simplicity, let $\lambda_\beta := \lambda\beta$.

By definition, we have

$$\mathrm{E}_{W\sim\tilde{\nu}_\beta}[\|T_K^{\gamma/2}W\|^2_{\mathcal{H}_{\tilde{K}^\theta}}] = \mathrm{Tr}[T_K^{\gamma/2}(\lambda_\beta^{-1}T_K)T_K^{\gamma/2}T_K^{-\theta(\gamma+1)}] = \lambda_\beta^{-1}\mathrm{Tr}[T_K^{(\gamma+1)(1-\theta)}]$$

Note that the assumption $\theta < 1-\tilde{\alpha}$ ensures the right hand side is finite. Therefore, we obtain that, for $\bar{R}_\theta > 0$,

$$\begin{aligned}\tilde{\nu}_\beta(\{h \in \mathcal{H} \mid \|h\|_{\mathcal{H}_{\tilde{K}^\theta}} \geq \lambda_\beta^{-1/2}\bar{R}_\theta\}) &\leq \frac{\mathrm{E}_{W\sim\tilde{\nu}_\beta}[\|T_K^{\gamma/2}W\|^2_{\mathcal{H}_{\tilde{K}^\theta}}]}{\lambda_\beta^{-1}\bar{R}_\theta^2} \leq \frac{\lambda_\beta^{-1}\mathrm{Tr}[T_K^{(\gamma+1)(1-\theta)}]}{\lambda_\beta^{-1}\bar{R}_\theta^2} \\ &= \frac{\mathrm{Tr}[T_K^{(\gamma+1)(1-\theta)}]}{\bar{R}_\theta^2}.\end{aligned}$$

Hence, by setting $\bar{R}_\theta = \sqrt{2\mathrm{Tr}[T_K^{(\gamma+1)(1-\theta)}]}$, we can guarantee that $\tilde{\nu}_\beta(\{h \in \mathcal{H} \mid \|h\|_{\mathcal{H}_{\tilde{K}^\theta}} \leq \lambda_\beta^{-1/2}\bar{R}_\theta\}) \geq 1/2$.

Let $B_\epsilon = (\epsilon\mathcal{B}_\mathcal{H}) \cap (\lambda_\beta^{-1/2}\bar{R}_\theta\mathcal{B}_{\mathcal{H}_{\tilde{K}^\theta}})$. We define

$$\phi^{(0)}_{\beta,\lambda}(\epsilon) := -\log\tilde{\nu}_\beta(\{W \in \mathcal{H} : W \in B_\epsilon\}).$$

For any $\delta > 0$, pick up $h^* \in \mathcal{H}_{\tilde{K}}$ that satisfies

$$\begin{aligned}&\lambda_\beta\|h^*\|^2_{\mathcal{H}_{\tilde{K}}} \leq (1+\delta)\inf_{h\in\mathcal{H}_{\tilde{K}}:\mathcal{L}(h)-\mathcal{L}(f^*)\leq\epsilon^2}\lambda_\beta\|h\|^2_{\mathcal{H}_{\tilde{K}}} \\ &\mathcal{L}(f_{h^*}) - \mathcal{L}(f^*) \leq \epsilon^2.\end{aligned}$$

Then, by Borel's inequality, it holds that

$$-\log\tilde{\nu}_\beta(\{h \in \mathcal{H} : \|h-h^*\|_\mathcal{H} \leq \epsilon\}) \leq (1+\delta)\phi_{\beta,\lambda}(\epsilon).$$

By the smoothness of the expected loss function $\bar{\ell}_x$, it holds that, for any $f, g \in L_2(P_X)$,

$$\begin{aligned}&\mathcal{L}(f) + \langle g-f, \nabla_f\mathcal{L}(f)\rangle_{L_2} + \frac{L}{2}\|g-f\|^2_{L_2} \\ =&\mathrm{E}_X\left[\bar{\ell}_X(f(X)) + (g(X)-f(X))\bar{\ell}'_X(f(X)) + \frac{L}{2}(g(X)-f(X))^2\right] \\ \geq&\mathrm{E}_X\left[\bar{\ell}_X(g(X))\right] = \mathcal{L}(g),\end{aligned}$$

where $\nabla_f\mathcal{L}$ is the Fréchet derivative in $L_2(P_X)$ (note that this inequality holds even though $\bar{\ell}_x(\cdot)$ is not a convex function). By substituting $g = f^*$ and $f = f^* + \frac{1}{L}\nabla_f\mathcal{L}(f_{h^*})$, we obtain

$$\begin{aligned}&\mathcal{L}(f^*) + \frac{1}{2L}\|\nabla_f\mathcal{L}(f_{h^*})\|^2_{L_2} \leq \mathcal{L}(f_{h^*}) \\ \Rightarrow\quad &\|\nabla_f\mathcal{L}(f_{h^*})\|^2_{L_2} \leq 2L(\mathcal{L}(f_{h^*}) - \mathcal{L}(f^*)) \leq 2L\epsilon^2.\end{aligned}$$

Therefore, for any $h \in \mathcal{H}$ such that $\|h-h^*\|_\mathcal{H} \leq \epsilon$, it holds that

$$\begin{aligned}\mathcal{L}(h) &\leq \mathcal{L}(f_{h^*}) + \langle f_h - f_{h^*}, \nabla\mathcal{L}(f_{h^*})\rangle_{L_2} + \frac{L}{2}\|f_h - f_{h^*}\|^2_{L_2} \\ &\leq \mathcal{L}(f_{h^*}) + \frac{1}{2}\|f_h - f_{h^*}\|^2_{L_2} + \frac{1}{2}\|\nabla\mathcal{L}(f_{h^*})\|^2_{L_2} + \frac{L}{2}\|f_h - f_{h^*}\|^2_\infty \\ &\leq \mathcal{L}(f_{h^*}) + \frac{1}{2}\|f_h - f_{h^*}\|^2_\infty + \frac{1}{2}2L\epsilon^2 + \frac{L}{2}\|f_h - f_{h^*}\|^2_\infty \\ &\leq \epsilon^2 + \frac{1+L}{2}(1+RD)^2\|h-h^*\|^2_\mathcal{H} + L\epsilon^2\end{aligned}$$

$$\leq \left(1 + \frac{(1+L)(1+RD)^2}{2} + L\right)\epsilon^2 =: C_{(L,R,D)}\epsilon^2.$$

This yields that

$$\begin{aligned}
&-\log \tilde{\nu}_\beta(\{h \in \mathcal{H} : \mathcal{L}(h) - \mathcal{L}(f^*) \leq C_{(L,R,D)}\epsilon^2\}) \\
&\leq -\log \tilde{\nu}_\beta(\{h \in \mathcal{H} : \mathcal{L}(h) - \mathcal{L}(f^*) \leq C_{(L,R,D)}\epsilon^2, \|h - h^*\|_{\mathcal{H}} \leq \epsilon\}) \\
&\leq (1+\delta)\phi_{\beta,\lambda}(\epsilon).
\end{aligned}$$

Since δ is arbitrary, we obtain that

$$-\log \tilde{\nu}_\beta(\{h \in \mathcal{H} : \mathcal{L}(h) - \mathcal{L}(f^*) \leq C_{(L,R,D)}\epsilon^2\}) \leq \phi_{\beta,\lambda}(\epsilon). \tag{14}$$

By the Gaussian correlation inequality (Lemma 2), we have that

$$\begin{aligned}
\phi^{(0)}_{\beta,\lambda}(\epsilon) &= -\log \tilde{\nu}_\beta(\{W \in \mathcal{H} \mid W \in B_\epsilon\}) \\
&\leq -\log[\tilde{\nu}_\beta(\{h \in \mathcal{H} \mid \|h\|_{\mathcal{H}} \leq \epsilon\}) \times \tilde{\nu}_\beta(\{h \in \mathcal{H} \mid \|h\|_{\mathcal{H}_{\tilde{K}^\theta}} \leq \lambda_\beta^{-1/2}\bar{R}_\theta\})] \\
&\leq -\log[\tilde{\nu}_\beta(\{h \in \mathcal{H} \mid \|h\|_{\mathcal{H}} \leq \epsilon\}) \times 1/2] \\
&= -\log \tilde{\nu}_\beta(\{h \in \mathcal{H} \mid \|h\|_{\mathcal{H}} \leq \epsilon\}) + \log(2).
\end{aligned}$$

Therefore, we can see that

$$\inf_{h \in \mathcal{H}_{\tilde{K}} : \mathcal{L}(h) - \mathcal{L}(f^*) \leq \epsilon^2/2} \lambda_\beta \|h\|^2_{\mathcal{H}_{\tilde{K}}} + \phi^{(0)}_{\beta,\lambda}(\epsilon) \leq \phi_{\beta,\lambda}(\epsilon).$$

Here we define ϵ^* as

$$\epsilon^* := \max\{\inf\{\epsilon > 0 : \phi_{\beta,\lambda}(\epsilon) \leq \beta\epsilon^2\}, n^{-\frac{1}{2(2-s)}}\}.$$

Note that, since $\phi_{\beta,\lambda}$ is monotonically non-increasing, ϵ^* satisfies

$$\phi_{\beta,\lambda}(\epsilon^*) \leq \beta{\epsilon^*}^2. \tag{15}$$

Let $r > 1$, and $M_r := -2\Phi^{-1}(e^{-\beta{\epsilon^*}^2 r})$ where Φ is the cumulative distribution function of the standard normal distribution. Since $\Phi^{-1}(y) \geq -\sqrt{5/2\log(1/y)}$ for every $y \in (0, 1/2)$ and Eq. (15) implies $\log(2) \leq \beta{\epsilon^*}^2$ yielding $e^{-\beta{\epsilon^*}^2 r} < e^{-\beta{\epsilon^*}^2} \leq e^{-\log(2)} = 1/2$, M_r is bounded by

$$-M_r/2 \geq -\sqrt{\frac{5}{2}\log(e^{\beta{\epsilon^*}^2 r})} = -\sqrt{\frac{5}{2}\beta{\epsilon^*}^2 r} \Rightarrow M_r \leq \sqrt{10\beta{\epsilon^*}^2 r}. \tag{16}$$

Let

$$\mathcal{F}_r := B_{\epsilon^*} + M_r\lambda_\beta^{-1/2}\mathcal{B}_{\mathcal{H}_{\tilde{K}}}.$$

By Borell's inequality (Theorem 3.1 of [10]), the prior probability mass of $\mathcal{F}_r$ is lower bounded by

$$\tilde{\nu}_\beta(\mathcal{F}_r) \geq \Phi(M_r + \alpha_r)$$

where $\alpha_r \in \mathbb{R}$ is determined by

$$\alpha_r = \Phi^{-1}(\tilde{\nu}_\beta(B_{\epsilon^*})) = \Phi^{-1}(e^{-\phi^{(0)}_{\beta,\lambda}(\epsilon^*)}).$$

Since $\phi^{(0)}_{\beta,\lambda}(\epsilon^*) \leq \phi_{\beta,\lambda}(\epsilon^*) + \log(2) \leq \beta{\epsilon^*}^2 \leq \beta{\epsilon^*}^2 r$ (Eq. (15)), we have $\Phi(\alpha_r) \geq e^{-\beta{\epsilon^*}^2 r}$ by the definition of α_r, which implies

$$\alpha_r \geq \Phi^{-1}(e^{-\beta{\epsilon^*}^2 r}) = -\frac{1}{2}M_r,$$

where the last equality is given by the definition of M_r. Therefore,

$$\tilde{\nu}_\beta(\mathcal{F}_r) \geq \Phi(M_r + \alpha_r) \geq \Phi(M_r/2) \geq 1 - \exp(-\beta{\epsilon^*}^2 r). \tag{17}$$

By the proof of Theorem 2.1 in [71], we obtain that the metric entropy of $\mathcal{F}_r$ is bounded by [2]

$$\log \mathcal{N}(\mathcal{F}_r, \|\cdot\|_{\mathcal{H}}, \epsilon^*) \le \frac{1}{2}M_r^2 + \phi_{\beta,\lambda}^{(0)}(\epsilon^*),$$

and, more strongly, there exist $h_1, \dots, h_{N_{\epsilon^*}} \in \mathcal{F}_r$ for $N_{\epsilon^*} := \mathcal{N}(\mathcal{F}_r, \epsilon^*, \|\cdot\|_{\mathcal{H}})$ such that

$$\mathcal{F}_r \subset B_{\epsilon^*} + \{h_1, \dots, h_{N_{\epsilon^*}}\}.$$

This indicates that, even for smaller $\epsilon' \le \epsilon^*$, it holds that

$$\log \mathcal{N}(\mathcal{F}_r, \|\cdot\|_{\mathcal{H}}, \epsilon') \le N_{\epsilon^*} + \log \mathcal{N}(B_{\epsilon^*}, \|\cdot\|_{\mathcal{H}}, \epsilon').$$

Then, if we let $\tilde{\phi}(\epsilon') = \log \mathcal{N}((\lambda_\beta^{-1/2}\bar{R}_\theta)\mathcal{B}_{\mathcal{H}_{\tilde{K}^\theta}}, \|\cdot\|_{\mathcal{H}}, \epsilon')$, then for $B_{\epsilon^*} \subset (\lambda_\beta^{-1/2}\bar{R}_\theta)\mathcal{B}_{\mathcal{H}_{\tilde{K}^\theta}}$ by its definition, Eq. (15) and Eq. (16) give

$$\log \mathcal{N}(\mathcal{F}_r, \|\cdot\|_{\mathcal{H}}, \epsilon') \le N_{\epsilon^*} + \tilde{\phi}(\epsilon') \le \frac{1}{2}M_r^2 + \phi_{\beta,\lambda}^{(0)}(\epsilon^*) + \tilde{\phi}(\epsilon') \le (5r+1)\beta\epsilon^{*2} + \tilde{\phi}(\epsilon'). \quad (18)$$

Note that if $\epsilon' > \epsilon^*$, then by using the fact $N_{\epsilon'} \le N_{\epsilon^*}$, we can see that this inequality still holds: $\log \mathcal{N}(\mathcal{F}_r, \epsilon', \|\cdot\|_{\mathcal{H}}) = N_{\epsilon'} \le N_{\epsilon^*} \le N_{\epsilon^*} + \tilde{\phi}(\epsilon')$. The covering number of $\mathcal{B}_{\mathcal{H}_{\tilde{K}^\theta}}$ can be evaluated using the decay rate of the spectrum $(\mu_k(\tilde{K}^\theta))_k$ [64]. Indeed, $\mu_k(\tilde{K}^\theta) \lesssim k^{-\frac{\theta}{\tilde{\alpha}}}$ implies $\tilde{\phi}(\epsilon') \lesssim (\epsilon'/\lambda_\beta^{1/2})^{-2\tilde{\alpha}/\theta}$ [64, Theorem 15]. Moreover, the small ball probability $\phi_{\beta,\lambda}^{(0)}(\epsilon)$ can be evaluated using the covering number. First, notice that $\phi_{\beta,\lambda}^{(0)}(\epsilon) = -\log \tilde{\nu}_\beta(\{W \in \mathcal{H} \mid W \in B_\epsilon\}) \le -\log \tilde{\nu}_\beta(\{W \in \mathcal{H} \mid \|W\|_{\mathcal{H}} \le \epsilon\}) =: \varphi(\epsilon)$, and then [26] showed that

$$\varphi(2\epsilon) \lesssim \log \mathcal{N}\left(\lambda_\beta^{-1/2}\mathcal{B}_{\mathcal{H}_{\tilde{K}}}, \|\cdot\|_{\mathcal{H}}, \frac{\epsilon}{\sqrt{2\varphi(\epsilon)}}\right) \lesssim \varphi(\epsilon).$$

Here, since the entropy number in the middle is evaluated as $\log \mathcal{N}\left(\lambda_\beta^{-1/2}\mathcal{B}_{\mathcal{H}_{\tilde{K}}}, \|\cdot\|_{\mathcal{H}}, \epsilon\right) \lesssim (\epsilon/\lambda_\beta^{1/2})^{-2\tilde{\alpha}}$, we obtain

$$\phi_{\beta,\lambda}^{(0)}(\epsilon) \le \varphi(\epsilon) \lesssim \left(\frac{\epsilon}{\lambda_\beta^{1/2}}\right)^{-\frac{2\tilde{\alpha}}{1-\tilde{\alpha}}}. \quad (19)$$

In this setting, we will show that, for any $r > 1$, there exists an event $\mathcal{E}_r$ with respect to data generation D_n and exists $u^* > 0$ such that

$$\begin{aligned}
&\mathrm{A}:\ P(\mathcal{E}_r^c) \le 2e^{-c'\min\{\beta\epsilon^{*2}, n\epsilon^{*2(2-s)}\}r} \ \text{ for a constant } c' > 0,\\
&\mathrm{B}:\ \widehat{\mathcal{L}}(W) - \widehat{\mathcal{L}}(f^*) \ge \frac{1}{2}[(\mathcal{L}(h) - \mathcal{L}(f^*)) - u^* r] \quad (\forall h \in \mathcal{F}_r)\\
&\qquad\qquad \text{under the event } \mathcal{E}_r,\\
&\mathrm{C}:\ \mathrm{E}[\pi_\infty(\mathcal{F}_r^c)\mathbf{1}_{\mathcal{E}}] \le 2\exp\left[-\tfrac{1}{2}(r-2)\beta\epsilon^{*2}\right],\\
&\mathrm{D}:\ \mathrm{E}[\pi_\infty(\{W \in \mathcal{F}_r : \mathcal{L}(W) - \mathcal{L}(f^*) \ge 3ru^*\})\mathbf{1}_{\mathcal{E}_r}] \le \exp\left[-\tfrac{1}{2}(r-2)\beta\epsilon^{*2}\right].
\end{aligned}$$

From now on, we will define u^* and $\mathcal{E}_r$ and prove the conditions $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$ one by one.

Step 1: Definitions of u^* and $\mathcal{E}_r$, and proof of A and B.

For notational simplicity, we write $\ell(f, Z)$ to indicate $\ell(Y, f(X))$ for $Z = (X, Y)$. By Talangrand's concentration inequality [68, 12], we have

$$P\left(\sup_{h \in \mathcal{F}_r} \frac{|\mathcal{L}(h) - \mathcal{L}(f^*) - (\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))|}{\mathcal{L}(h) - \mathcal{L}(f^*) + u}\right.$$

[2]For a metric space $\tilde{\mathcal{F}}$ equipped with a metric $\tilde{d}$, the ϵ-covering number $\mathcal{N}(\tilde{\mathcal{F}}, \tilde{d}, \epsilon)$ is defined as the minimum number of balls with radius ϵ (measured by the metric $\tilde{d}$) to cover the metric space $\tilde{\mathcal{F}}$.

$$\geq 2\mathrm{E}\left[\sup_{h\in\mathcal{F}_r}\frac{|\mathcal{L}(h)-\mathcal{L}(f^*)-(\widehat{\mathcal{L}}(h)-\widehat{\mathcal{L}}(f^*))|}{\mathcal{L}(h)-\mathcal{L}(f^*)+u}\right]+\sqrt{\frac{2t}{n}V}+\frac{2tU}{n}\Bigg)$$
$$\leq \exp(-t),$$

for any $t \geq 1$, where

$$V=\sup_{h\in\mathcal{F}_r}\frac{\mathrm{E}[(\ell(h,Z)-\ell(f^*,Z)-\mathrm{E}[\ell(h,Z)-\ell(f^*,Z)])^2]}{(\mathcal{L}(h)-\mathcal{L}(f^*)+u)^2},$$
$$U=\sup_{h\in\mathcal{F}_r}\frac{\|\ell(h,\cdot)-\ell(f^*,\cdot)-\mathrm{E}[\ell(h,Z)-\ell(f^*,Z)]\|_\infty}{\mathcal{L}(h)-\mathcal{L}(f^*)+u}.$$

By the Bernstein condition, it holds that

$$\mathrm{Var}[\ell(f_h,Z)-\ell(f^*,Z)]\leq \mathrm{E}[(\ell(f_h,Z)-\ell(f^*,Z))^2]\leq C_B(\mathcal{L}(h)-\mathcal{L}(f^*))^s,$$

which gives

$$V\leq C_B\sup_{h\in\mathcal{F}_r}\frac{(\mathcal{L}(h)-\mathcal{L}(f^*))^s}{(\mathcal{L}(h)-\mathcal{L}(f^*)+u)^2}\leq\frac{C_B}{u^{2-s}}.$$

By the boundedness assumption of the loss function, we can see that

$$U\leq\frac{\bar{R}}{u}.$$

Hence, we have that

$$P\Bigg(\sup_{h\in\mathcal{F}_r}\frac{|\mathcal{L}(h)-\mathcal{L}(f^*)-(\widehat{\mathcal{L}}(h)-\widehat{\mathcal{L}}(f^*))|}{\mathcal{L}(h)-\mathcal{L}(f^*)+u}$$
$$\geq 2\mathrm{E}\left[\sup_{h\in\mathcal{F}_r}\frac{|\mathcal{L}(h)-\mathcal{L}(f^*)-(\widehat{\mathcal{L}}(h)-\widehat{\mathcal{L}}(f^*))|}{\mathcal{L}(h)-\mathcal{L}(f^*)+u}\right]+\sqrt{\frac{2C_B}{nu^{2-s}}t}+\frac{2\bar{R}}{nu}t\Bigg)$$
$$\leq\exp(-t), \tag{20}$$

for any $t\geq 1$.

Hereafter, we bound the expectation of the supremum of the ratio type empirical process: $\mathrm{E}\left[\sup_{h\in\mathcal{F}_r}\frac{|\mathcal{L}(h)-\mathcal{L}(f^*)-(\widehat{\mathcal{L}}(h)-\widehat{\mathcal{L}}(f^*))|}{\mathcal{L}(h)-\mathcal{L}(f^*)+u}\right]$. Let the empirical L_2-norm be $\|h\|_n:=\sqrt{\frac{1}{n}\sum_{i=1}^n h(z_i)^2}$. By the usual Rademacher complexity and covering number argument (Lemma 11.4 of [11], Theorem 5.22 of [75] and Lemma A.5 of [6] for example), the non-ratio-type empirical process can be bounded as

$$\mathrm{E}\left[\sup_{h\in\mathcal{F}_r:\mathcal{L}(h)-\mathcal{L}(f^*)\leq u}|\mathcal{L}(h)-\mathcal{L}(f^*)-(\widehat{\mathcal{L}}(h)-\widehat{\mathcal{L}}(f^*))|\right]$$
$$\leq 2\mathrm{E}\left[\sup_{h\in\mathcal{F}_r:\mathcal{L}(h)-\mathcal{L}(f^*)\leq u}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i(\ell(f_h,z_i)-\ell(f^*,z_i)-(\mathcal{L}(h)-\mathcal{L}(f^*)))\right|\right]$$
$$\leq C\mathrm{E}\left[\inf_{a>0}\left\{a+\int_a^{\hat{r}(u)}\sqrt{\frac{\log\mathcal{N}(\{\ell(f_h,\cdot)-\ell(f^*,\cdot)\mid h\in\mathcal{F}_r,\ \mathcal{L}(h)-\mathcal{L}(f^*)\leq u\},\|\cdot\|_n,\epsilon')}{n}}\mathrm{d}\epsilon'\right\}\right],$$

where $\hat{r}(u):=\sup\{\|\ell(f_h,\cdot)-\ell(f^*,\cdot)\|_n\mid h\in\mathcal{F}_r,\ \mathcal{L}(h)-\mathcal{L}(f^*)\leq u\}$ and C is a universal constant. The Dudley integral in the right hand side can be bounded by

$$\int_a^{\hat{r}(u)}\sqrt{\frac{\log\mathcal{N}(\{\ell(f_h,\cdot)-\ell(f^*,\cdot)\mid h\in\mathcal{F}_r,\ \mathcal{L}(h)-\mathcal{L}(f^*)\leq u\},\|\cdot\|_n,\epsilon')}{n}}\mathrm{d}\epsilon'$$
$$\stackrel{(1)}{\leq}\int_a^{\hat{r}(u)}\sqrt{\frac{\log\mathcal{N}(\{f_h\mid h\in\mathcal{F}_r\},\|\cdot\|_\infty,\epsilon'/B)}{n}}\mathrm{d}\epsilon'$$
$$\stackrel{(2)}{\leq}\int_a^{\hat{r}(u)}\sqrt{\frac{\log\mathcal{N}(\mathcal{F}_r,\|\cdot\|_{\mathcal{H}},\epsilon'/(B(1+RD)))}{n}}\mathrm{d}\epsilon'$$

$$\overset{(3)}{\leq} \int_a^{\hat{r}(u)} \sqrt{\frac{N_{\epsilon^*} + \tilde{\phi}(\epsilon'/(B(1+RD)))}{n}} \mathrm{d}\epsilon',$$

where we used the bounded gradient condition on the loss function to show (1), used Lemma 1 to show (2), and used Eq. (18) to show (3). If we let $a = 1/n$, then we have

$$\begin{aligned}
&\mathrm{E}\left[\sup_{h \in \mathcal{F}_r : \mathcal{L}(h) - \mathcal{L}(f^*) \leq u} |\mathcal{L}(h) - \mathcal{L}(f^*) - (\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))|\right] \\
\leq & C\mathrm{E}\left[\frac{1}{n} + \int_{1/n}^{\hat{r}(u)} \sqrt{\frac{N_{\epsilon^*} + \tilde{\phi}(\epsilon'/(B(1+RD)))}{n}} \mathrm{d}\epsilon'\right] =: \psi_{r,\epsilon^*}(u).
\end{aligned}$$

Here, we assume that there exists an upper bound $\bar{\psi}_{r,\epsilon^*}(u)$ of $\psi_{r,\epsilon^*}(u)$ that satisfies

$$\bar{\psi}_{r,\epsilon^*}(4u) \leq 2\bar{\psi}_{r,\epsilon^*}(u) \quad (u > 0), \tag{21a}$$

$$\frac{\bar{\psi}_{r,\epsilon^*}(ur)}{ur} \leq \frac{\bar{\psi}_{1,\epsilon^*}(u)}{u} \quad (u > 0,\ r \geq 1). \tag{21b}$$

We will show these conditions in Step 5. Then, the so called *peeling device* gives

$$\mathrm{E}\left[\sup_{h \in \mathcal{F}_r} \frac{|\mathcal{L}(h) - \mathcal{L}(f^*) - (\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))|}{\mathcal{L}(h) - \mathcal{L}(f^*) + u}\right] \leq \frac{4\bar{\psi}_{r,\epsilon^*}(u)}{u}.$$

(Theorem 7.7 and Eq. (7.17) of [63]). Therefore, Eq. (20) can yields that

$$P\left(\sup_{h \in \mathcal{F}_r} \frac{|\mathcal{L}(h) - \mathcal{L}(f^*) - (\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))|}{\mathcal{L}(h) - \mathcal{L}(f^*) + u} \geq 8\frac{\bar{\psi}_{r,\epsilon^*}(u)}{u} + \sqrt{\frac{2C_B}{nu^{2-s}}t} + \frac{2\bar{R}}{nu}t\right) \leq \exp(-t).$$

Here, for $t_1 = \beta {\epsilon^*}^2$, let u^* be any real number satisfying

$$u^* \geq \max\left\{{\epsilon^*}^2, \inf\left\{u > 0 : 8\frac{\bar{\psi}_{1,\epsilon^*}(u)}{u} + \sqrt{\frac{2C_B}{nu^{2-s}}t_1} + \frac{2\bar{R}}{nu}t_1 \leq \frac{1}{2}\right\}\right\}.$$

For more general $t = t_1 r = \beta {\epsilon^*}^2 r$, since $\frac{\bar{\psi}_{r,\epsilon^*}(u^* r)}{u^* r} \leq \frac{\bar{\psi}_{1,\epsilon^*}(u^*)}{u^*}$, combining with the fact that $\sqrt{\frac{2C_B}{n(u^* r)^{2-s}}t_1 r} \leq \sqrt{\frac{2C_B}{n(u^*)^{2-s}}t_1}$ and $\frac{2\bar{R}}{nu^* r}t_1 r = \frac{2\bar{R}}{nu^*}t_1$, it also holds that

$$8\frac{\bar{\psi}_{r,\epsilon^*}(u^* r)}{u^* r} + \sqrt{\frac{2C_B}{n(u^* r)^{2-s}}t_1 r} + \frac{2\bar{R}}{n(u^* r)}t_1 r \leq \frac{1}{2},$$

for $r \geq 1$. Therefore, the following inequality holds:

$$\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*) \geq \frac{1}{2}[(\mathcal{L}(h) - \mathcal{L}(f^*)) - u^* r],$$

uniformly over all $h \in \mathcal{F}_r$ with probability $1 - e^{-\beta {\epsilon^*}^2 r}$. We denote this event by $\mathcal{E}_1$.

Eq. (14) gives that

$$-\log \tilde{\nu}_\beta(\{h \in \mathcal{H} : \mathcal{L}(h) - \mathcal{L}(f^*) \leq C_{(L,R,D)}{\epsilon^*}^2\}) \leq \phi_{\beta,\lambda}(\epsilon^*).$$

Let the conditional probability measure of $\tilde{\nu}_\beta$ conditioned on the set $A_{\epsilon^*} := \{h \in \mathcal{H} : \mathcal{L}(h) - \mathcal{L}(f^*) \leq C_{(L,R,D)}{\epsilon^*}^2\}$ be

$$\tilde{\nu}_\beta(B|A_{\epsilon^*}) := \frac{\tilde{\nu}_\beta(B \cap A_{\epsilon^*})}{\tilde{\nu}_\beta(A_{\epsilon^*})},$$

for a measurable set $B \subset \mathcal{H}$. Let $\bar{\ell}(Z) := \int \ell(h, Z)\tilde{\nu}_\beta(\mathrm{d}h|A_{\epsilon^*})$. Then, we have that

$$\int \exp\left(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))\right) \mathrm{d}\tilde{\nu}_\beta(h|A_{\epsilon^*}) \geq \exp\left(-\int \beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))\mathrm{d}\tilde{\nu}_\beta(h|A_{\epsilon^*})\right)$$

$$= \exp\left[-\beta\left(\frac{1}{n}\sum_{i=1}^n \bar{\ell}(z_i) - \mathrm{E}[\bar{\ell}(Z)]\right)\right].$$

Now, by the Bernstein's inequality,

$$P\left(\frac{1}{n}\sum_{i=1}^n \bar{\ell}(z_i) - \mathrm{E}[\bar{\ell}(Z)] \geq \sqrt{\frac{2C_B(C_{(L,R,D)}\epsilon^{*2})^s t}{n}} + \frac{\bar{R}t}{n}\right) \leq e^{-t},$$

for $t \geq 0$. Here, let $t = \frac{1}{8}\min\{\frac{1}{2C_B C_{(L,R,D)}^s}, \frac{1}{\bar{R}}\} n \min\{\epsilon^{*2(2-s)}, \epsilon^{*2}\} r$, then it holds that

$$P\left(\frac{1}{n}\sum_{i=1}^n \bar{\ell}(z_i) - \mathrm{E}[\bar{\ell}(Z)] \geq \frac{1}{2}\epsilon^{*2} r\right) \leq e^{-\frac{1}{8}\min\{\frac{1}{2C_B C_{(L,R,D)}^s}, \frac{1}{\bar{R}}\} n \min\{\epsilon^{*2(2-s)}, \epsilon^{*2}\} r}.$$

Therefore, this and the definition of $\tilde{\nu}_\beta(\cdot|A_{\epsilon^*})$ give that

$$\begin{aligned}
\int \exp\left(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))\right) \mathrm{d}\tilde{\nu}_\beta(h) &\geq \exp\left(-\tfrac{1}{2}\beta\epsilon^{*2} r\right) \tilde{\nu}_\beta(A_{\epsilon^*}) \\
&\geq \exp\left(-\tfrac{1}{2}\beta\epsilon^{*2} r - \beta\epsilon^{*2}\right) \quad (\because \text{Eqs. (14) and (15)}) \\
&\geq \exp\left(-(\tfrac{r}{2}+1)\beta\epsilon^{*2}\right). \qquad (22)
\end{aligned}$$

with probability $1 - \exp[-\frac{1}{8}\min\{\frac{1}{2C_B C_{(L,R,D)}^s}, \frac{1}{\bar{R}}\} n \min\{\epsilon^{*2(2-s)}, \epsilon^{*2}\} r]$. We define this event as $\mathcal{E}_2$.

Combining $\mathcal{E}_1$ and $\mathcal{E}_2$, we define $\mathcal{E}_r = \mathcal{E}_1 \cap \mathcal{E}_2$, then $P(\mathcal{E}_r) \geq 1 - e^{-\beta\epsilon^{*2} r} - e^{-\frac{1}{8}\min\{\frac{1}{2C_B C_{(L,R,D)}^s}, \frac{1}{\bar{R}}\} n \min\{\epsilon^{*2(2-s)}, \epsilon^{*2}\} r} \geq 1 - 2e^{-c' \min\{\beta\epsilon^{*2}, n\epsilon^{*2(2-s)}\} r}$ for a constant $c' > 0$, where we used $\beta \leq n$.

Step 2: Proof of C**.**

Next, we evaluate the condition C. Eq. (22) gives that, on the event $\mathcal{E}_r$, the Bayes posterior probability of $\mathcal{F}_r^c$ is upper bounded by

$$\begin{aligned}
\pi_\infty(\mathcal{F}_r^c) &= \frac{\int_{h\in\mathcal{F}_r^c} \exp\left(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))\right) \mathrm{d}\tilde{\nu}_\beta(h)}{\int \exp\left(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))\right) \mathrm{d}\tilde{\nu}_\beta(h)} \\
&\leq \exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2}\right) \int_{h\in\mathcal{F}_r^c} \exp\left(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))\right) \mathrm{d}\tilde{\nu}_\beta(h).
\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}
\mathrm{E}[\pi_\infty(\mathcal{F}_r^c)\mathbf{1}_{\mathcal{E}_r}] &\leq \mathrm{E}\left[\mathbf{1}_{\mathcal{E}_r} \exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2}\right) \int_{h\in\mathcal{F}_r^c} \exp(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*)))\mathrm{d}\tilde{\nu}_\beta(h)\right] \\
&\leq \exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2}\right) \int_{h\in\mathcal{F}_r^c} \mathrm{E}\left[\exp\left(-n\frac{\beta}{n}(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*))\right)\right] \mathrm{d}\tilde{\nu}_\beta(h) \\
&= \exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2}\right) \int_{h\in\mathcal{F}_r^c} \prod_{i=1}^n \mathrm{E}_{Z_i}\left[\exp\left(-\frac{\beta}{n}(\ell(h,Z_i) - \ell(f^*,Z_i))\right)\right] \mathrm{d}\tilde{\nu}_\beta(h) \\
&\leq \exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2}\right) \tilde{\nu}_\beta(\mathcal{F}_r^c) \quad (\because \text{predictor condition of Assumption 2}) \\
&\leq 2\exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2} - \beta\epsilon^{*2} r\right) \quad (\because \text{Eq. (17)}) \\
&= 2\exp\left(-\tfrac{1}{2}(r-2)\beta\epsilon^{*2}\right).
\end{aligned}$$

Step 3: Proof of D**.**

Next, we prove the condition D. Similarly to C, we have that

$$
\begin{aligned}
&\mathrm{E}[\pi_\infty(\{W \in \mathcal{F}_r : \mathcal{L}(W) - \mathcal{L}(f^*) \geq 3ru^*\})\mathbf{1}_{\mathcal{E}_r}] \\
&= \mathrm{E}\left[\frac{\int_{\mathcal{L}(h)-\mathcal{L}(f^*)\geq 3ru^*} \exp(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*)))\mathrm{d}\tilde{\nu}_\beta(h)}{\int \exp(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*)))\mathrm{d}\tilde{\nu}_\beta(h)}\mathbf{1}_{\mathcal{E}_r}\right] \\
&\leq \mathrm{E}\left[\mathbf{1}_{\mathcal{E}_r} \exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2}\right)\int_{\mathcal{L}(h)-\mathcal{L}(f^*)\geq 3ru^*} \exp(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*)))\mathrm{d}\tilde{\nu}_\beta(h)\right] \\
&\leq \mathrm{E}\left[\mathbf{1}_{\mathcal{E}_r} \exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2}\right)\int_{\widehat{\mathcal{L}}(h)-\widehat{\mathcal{L}}(f^*)\geq ru^*} \exp(-\beta(\widehat{\mathcal{L}}(h) - \widehat{\mathcal{L}}(f^*)))\mathrm{d}\tilde{\nu}_\beta(h)\right] \\
&\quad (\because \text{ condition B is satisfied on } \mathcal{E}_r) \\
&\leq \exp\left((\tfrac{r}{2}+1)\beta\epsilon^{*2} - r\beta u^*\right) \\
&\leq \exp\left[-\tfrac{1}{2}(r-2)\beta\epsilon^{*2}\right].
\end{aligned}
$$

Step 4: Integrating all bounds of $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$**.**

Finally, we integrate all bounds to obtain an excess risk bound.

$$
\begin{aligned}
&\mathrm{E}_{D^n}\left[\int \mathcal{L}(W) - \mathcal{L}(f^*)\mathrm{d}\pi_\infty(W)\right] \\
=&\mathrm{E}_{D^n}\left[\int_0^\infty \pi_\infty(\{W \in \mathcal{H} : \mathcal{L}(W) - \mathcal{L}(f^*) > t\})\mathrm{d}t\right] \\
=&\int_0^\infty \mathrm{E}_{D^n}\left[\pi_\infty(\{W \in \mathcal{H} : \mathcal{L}(W) - \mathcal{L}(f^*) > t\})\right]\mathrm{d}t \quad \text{(by Fubuni's theorem)} \\
=&3u^* + \int_1^\infty 3u^*\mathrm{E}_{D^n}\left[\pi_\infty(\{W \in \mathcal{H} : \mathcal{L}(W) - \mathcal{L}(f^*) > 3ru^*\})\right]\mathrm{d}r \\
\leq&3u^* + 3u^*\int_1^\infty \mathrm{E}_{D^n}\left\{\mathbf{1}_{\mathcal{E}_r^c} + \mathbf{1}_{\mathcal{E}_r}[\pi_\infty(\mathcal{F}_r^c) + \pi_\infty(\{W \in \mathcal{F}_r : \mathcal{L}(W) - \mathcal{L}(f^*) > 3ru^*\})]\right\}\mathrm{d}r \\
\leq&3u^* + 3u^*\int_1^\infty \min\left\{2e^{-c'\min\{\beta\epsilon^{*2}, n\epsilon^{*2(2-s)}\}r} + 3\exp\left(-\tfrac{1}{2}(r-2)\beta\epsilon^{*2}\right), 1\right\}\mathrm{d}r \\
\lesssim& u^*,
\end{aligned} \tag{23}
$$

where in the last inequality, we used $\beta\epsilon^{*2} \geq \log(2)$ by Eq. (15) and $n\epsilon^{*2(2-s)} \geq 1$ by the definition of ϵ^*.

Step 5: Evaluation of u^***.**

Based on the arguments above, our goal is reduced to evaluating u^*. We note that

$$
\begin{aligned}
\mathrm{E}[\hat{r}(u)^2] &\leq u^s + \bar{R}\psi_{r,\epsilon^*}(u) \leq u^s + \bar{R}\mathrm{E}\left[\int_0^{\hat{r}(u)}\sqrt{\frac{\beta\epsilon^{*2}r + \left(\epsilon'/\lambda_\beta^{1/2}(B(1+RD))^{-1}\right)^{-2\tilde{\alpha}/\theta}}{n}}\mathrm{d}\epsilon'\right] \\
&\lesssim u^s + \sqrt{\frac{\beta\epsilon^{*2}r}{n}\mathrm{E}[\hat{r}^2(u)]} + \frac{1}{\sqrt{n}}\lambda_\beta^{\tilde{\alpha}/\theta}\mathrm{E}[(\hat{r}(u))^{1-\tilde{\alpha}/\theta}] \\
&\leq u^s + \sqrt{\frac{\beta\epsilon^{*2}r}{n}\mathrm{E}[\hat{r}^2(u)]} + \frac{1}{\sqrt{n}}\lambda_\beta^{\tilde{\alpha}/\theta}(\mathrm{E}[\hat{r}(u)^2])^{(1-\tilde{\alpha}/\theta)/2}.
\end{aligned}
$$

Therefore, we have that

$$
\mathrm{E}[\hat{r}(u)^2] \lesssim u^s \vee \frac{\beta}{n}\epsilon^{*2}r \vee n^{-\frac{1}{1+\tilde{\alpha}/\theta}}\lambda_\beta^{\frac{2\tilde{\alpha}/\theta}{1+\tilde{\alpha}/\theta}}.
$$

This gives that

$$
\begin{aligned}
\psi_{r,\epsilon^*}(u) \lesssim & \mathrm{E}\left[\int_0^{\hat{r}(u)} \sqrt{\frac{\beta\epsilon^{*2} r + \left(\epsilon'/\lambda_\beta^{1/2}(B(1+RD))^{-1}\right)^{-2\tilde{\alpha}/\theta}}{n}} \mathrm{d}\epsilon'\right] \\
\lesssim & \sqrt{\frac{\beta\epsilon^{*2} r}{n}} \mathrm{E}[\hat{r}(u)^2] + \frac{1}{\sqrt{n}} \lambda_\beta^{\tilde{\alpha}/\theta} (\mathrm{E}[\hat{r}(u)^2])^{(1-\tilde{\alpha}/\theta)/2} \\
\lesssim & \frac{\beta}{n}\epsilon^{*2} r + n^{-\frac{1}{1+\tilde{\alpha}/\theta}} \lambda_\beta^{\frac{2\tilde{\alpha}/\theta}{1+\tilde{\alpha}/\theta}} + u^{s/2} \sqrt{\frac{\beta\epsilon^{*2} r}{n} + n^{-\frac{1}{1+\tilde{\alpha}/\theta}} \lambda_\beta^{\frac{2\tilde{\alpha}/\theta}{1+\tilde{\alpha}/\theta}}} \quad (\because \text{Young's inequality}).
\end{aligned}
$$

Here, we let the upper bound in the right hand side as $\bar{\psi}_{r,\epsilon^*}$, then we can easily show the condition (21), that is, $\bar{\psi}_{r,\epsilon^*}(4u) \leq 2\bar{\psi}_{r,\epsilon^*}(u)$ and $\frac{\bar{\psi}_{r,\epsilon^*}(ur)}{ur} \leq \frac{\bar{\psi}_{1,\epsilon^*}(u)}{u}$. Finally, by the definition of u^*, we obtain that

$$
u^* \lesssim \epsilon^{*2} \vee \left(\frac{\beta}{n}\epsilon^{*2} + n^{-\frac{1}{1+\tilde{\alpha}/\theta}} \lambda_\beta^{\frac{2\tilde{\alpha}/\theta}{1+\tilde{\alpha}/\theta}}\right)^{\frac{1}{2-s}} \vee \frac{1}{n}.
$$

This yields the assertion. □

D.3 Proof of fast rate for classification (Theorem 3)

Proof. Let the convergence rate in the right hand side of Eq. (9) in Theorem 2 be u^*.

Since both $\bar{W}_2(a)$ and $\bar{W}_1(w)$ are bounded and the activation function σ is included in the Hölder class $\mathcal{C}^m(\mathbb{R})$, the model $\{f_W \mid W \in \mathcal{H}\}$ is also included in the Hölder class $\mathcal{C}^m(\mathcal{X})$ with regularity m and especially it is included in the Sobolev space $W_2^m(\mathcal{X})$:

$$
f_W \in W_2^m(\mathcal{X}).
$$

Moreover, since the logistic loss is C^∞-class and its derivative up to m-th order is upper bounded, the function $x \mapsto \ell(f_W(x), y)$ is also included in $W^m(\mathcal{X})$ for all $y \in \{\pm 1\}$. Therefore, $\hat{h}_W(x) := \mathrm{E}_{Y|x}[\ell(f_W(x), Y)](= h(f_W(x)|x))$ is also included in $W^m(\mathcal{X})$. Moreover, $\|\hat{h}_W\|_{W_2^m(\mathcal{X})} \leq C$ uniformly over all $W \in \mathcal{H}$.

If X_0 and X_1 are a pair of quasi-normed spaces which are continuously embedded in a linear Hausdorff space $\mathcal{G}$, their K-functional is defined for any $f \in X_0 + X_1$ by

$$
K(f, t; X_0, X_1) := \inf_{f = f_0 + f_1} \|f_0\|_{X_0} + \|f_1\|_{X_1}.
$$

For each $0 < \theta < 1$, $0 < p \leq \infty$, the *interpolation space* $[X_0, X_1]_{q,\theta}$ is the set of all functions $f \in X_0 + X_1$ for which

$$
\|f\|_{[X_0,X_1]_{q,\theta}} := \left(\int_0^\infty (t^{-\theta} K(f, t; X_0, X_1))^q \frac{\mathrm{d}t}{t}\right)^{1/q}
$$

is finite. For $q = \infty$, the right hand side is properly modified in a usual manner. As shown by [21], it holds that

$$
[L_2(\Omega), W_2^m(\mathcal{X})]_{1,d/2m} = B_{2,1}^{d/2}(\mathcal{X}),
$$

where $L_2(\mathcal{X})$ is the L_2-space on $\mathcal{X}$ with respect to the Lebesgue measure and $B_{2,1}^{d/2}(\mathcal{X})$ is the Besov space defined on $\mathcal{X}$ (see [21] for its definition). Note that $d/2m < 1$ by the assumption. From this property, combined the extension theorem of [21] and the embedding property of the Besov space [69], we have that $B_{2,1}^{d/2}(\mathcal{X}) \hookrightarrow L_\infty(\mathcal{X})$. Under this condition, it is known that the following inequality holds

$$
\begin{aligned}
\|\hat{h}_W - h^*\|_\infty \leq & C\|\hat{h}_W - h^*\|_{L_2(\mathcal{X})}^{1-\frac{d}{2m}} \|\hat{h}_W - h^*\|_{W_2^m(\mathcal{X})}^{\frac{d}{2m}} \\
\leq & C c_0^{-(1-d/2m)} \|\hat{h}_W - h^*\|_{L_2(P_X)}^{1-\frac{d}{2m}} \|\hat{h}_W - h^*\|_{W_2^m(\mathcal{X})}^{\frac{d}{2m}},
\end{aligned}
$$

(see [9, 64]). Combining this with the assumption $h^* \in W_2^m(\mathcal{X})$ and the fact $\|\hat{h}_W\|_{W_2^m(\mathcal{X})} \le C$, if $\|\hat{h}_W - h^*\|_{L_2(P_X)} \le \epsilon$ for sufficiently small ϵ, we have an L_∞-norm bound as $\|\hat{h}_W - h^*\|_\infty \le C'\epsilon^{1-\frac{d}{2m}}$. Thus, if we choose ϵ so that $\epsilon^{1-\frac{d}{2m}} = \Theta(\delta)$ and let W satisfy $\|\hat{h}_W - h^*\|_{L_2(P_X)} \le \epsilon$, then we can have

$$\|\hat{h}_W - h^*\|_\infty < \delta/2.$$

Then, by the assumption that $h^*(x) \le \log(2) - \delta$, it holds that

$$\hat{h}_W(x) < \log(2) - \delta/2 \quad \text{(a.s.)},$$

which indicates that

$$P_X(\mathrm{sign}(f_W(X)) = g^*(X)) = 1.$$

Therefore, we only need to bound the quantity $\|\hat{h}_{W_k} - h^*\|^2_{L_2(P_X)}$ for $W_k \sim \pi_k$.

Here, we show that the Bernstein condition (Assumption 2) is satisfied with $s = 1$ under Assumptions 3. By Assumptions 3 and $\|f_W\|_\infty \le R$ for any $W \in \mathcal{H}$ by the definition of the clipping operator, it holds that $\|f^*\|_\infty \le R$. Therefore, Lemma 3 yields the Bernstein condition with $s = 1$ and $C_B = 4 + 3R$. Therefore, $\|\hat{h}_W - h^*\|^2_{L_2(P_X)}$ can be bounded as

$$\begin{aligned}
&\|\hat{h}_W - h^*\|^2_{L_2(P_X)} \\
\le& \mathrm{E}_Z[(\ell(f_W, Z) - \ell(f^*, Z))^2] \\
\le& C_B(\mathcal{L}(W) - \mathcal{L}(f^*))^s = C_B(\mathcal{L}(W) - \mathcal{L}(f^*)),
\end{aligned}$$

where we used Jensen's inequality in the first inequality, and we applied $s = 1$ in the last equality. First, we consider the stationary distribution. For any $\epsilon' > u^*$, we have already shown in the proof of Theorem 2 (See Eq. (23)) that

$$\begin{aligned}
&\mathrm{E}_{D^n}\left[\pi_\infty(\{W \in \mathcal{H} \mid \mathcal{L}(W) - \mathcal{L}(f^*) \ge \epsilon'\})\right] \\
\le& C\exp(-c\beta u^* \times (\epsilon'/u^*)) = C\exp(-c\beta\epsilon'). \qquad (24)
\end{aligned}$$

Next, we consider the intermediate solution W_k. Suppose that the sample size n is sufficiently large and λ is appropriately chosen with sufficiently large β so that $u^* \ll \delta^{2m/(2m-d)}$[3]. The probability of misclassification is bounded by

$$\begin{aligned}
&\mathrm{E}[\pi_k(\{W_k \in \mathcal{H} \mid P_X(\mathrm{sign}(f_{W_k}(X)) = \mathrm{sign}(f^*(X))) \neq 1\})] \\
\le& \mathrm{E}\left[P_{W_k\sim\pi_k}[\mathcal{L}(W_k) - \mathcal{L}(f^*) \ge \epsilon/C_B]\right] \\
=& \mathrm{E}\left[P_{W_k\sim\pi_k, W\sim\pi_\infty}[\mathcal{L}(W_k) - \mathcal{L}(W) - (\mathcal{L}(f^*) - \mathcal{L}(W)) \ge \epsilon/C_B]\right] \\
\le& \mathrm{E}\left[P_{W_k\sim\pi_k, W\sim\pi_\infty}[\mathcal{L}(W_k) - \mathcal{L}(W) \ge \epsilon/(2C_B)]\right] + \mathrm{E}\left[P_{W\sim\pi_\infty}[\mathcal{L}(W) - \mathcal{L}(f^*) \ge \epsilon/(2C_B)]\right] \\
\le& \mathrm{E}\left[\mathrm{E}_{W_k\sim\pi_k, W\sim\pi_\infty}[\mathcal{L}(W_k) - \mathcal{L}(W)]/(\epsilon/(2C_B))\right] + \mathrm{E}\left[P_{W\sim\pi_\infty}[\mathcal{L}(W) - \mathcal{L}(f^*) \ge \epsilon/(2C_B)]\right] \\
\lesssim& \frac{\Xi_k}{\delta^{2m/(2m-d)}} + \exp(-c'\beta\delta^{2m/(2m-d)}),
\end{aligned}$$

where we used $\epsilon = \Theta(\delta^{2m/(2m-d)})$ and Eq. (24) in the last inequality. Therefore, for a fixed δ, we can obtain the Bayes classifier with high probability by setting η sufficiently small and taking sufficiently large k.

Making the first term as large as the second term.

We see that the first term in the right hand side is coming from the bound of $\mathrm{E}\left[P_{W_k\sim\pi_k, W\sim\pi_\infty}[\mathcal{L}(W_k) - \mathcal{L}(W) \ge \epsilon/(2C_B)]\right]$. To bound this, we used the following bound:

$$\begin{aligned}
P_{W_k\sim\pi_k, W\sim\pi_\infty}[\mathcal{L}(W_k) - \mathcal{L}(W) \ge \epsilon/(2C_B)] &\le \mathrm{E}_{W_k\sim\pi_k, W\sim\pi_\infty}[\mathcal{L}(W_k) - \mathcal{L}(W)]/(\epsilon/(2C_B)) \\
&\lesssim \frac{\Xi_k}{\delta^{2m/(2m-d)}},
\end{aligned}$$

[3] This is a more precise meaning of the sentence "the sample size n is sufficiently large and λ is appropriately chosen" in the statement.

almost surely. Therefore, if Ξ_k is sufficiently small such that $\frac{\Xi_k}{\delta^{2m/(2m-d)}} \ll 1$, then we have

$$P_{W_k \sim \pi_k, W \sim \pi_\infty}[\mathcal{L}(W_k) - \mathcal{L}(W) \geq \epsilon/(2C_B)] \leq 1/2.$$

Therefore, by running the algorithm S-times and picking up the beset W_k in terms of the validation error (write it as $W_k^{(S)}$), then we have that

$$P_{W_k \sim \pi_k, W \sim \pi_\infty}[\mathcal{L}(W_k^{(S)}) - \mathcal{L}(W) \geq \epsilon/(2C_B)] \leq 1/2^S.$$

Thus, for sufficiently large S such that the right hand side can be smaller than the second term $\exp(-c'\beta\delta^{2m/(2m-d)})$, we have that

$$\mathrm{E}\left\{P_{W_k^{(S)}}\left[P_X(\mathrm{sign}(f_{W_k^{(S)}}(X)) = \mathrm{sign}(f^*(X))) \neq 1 \;\middle|\; D^n\right]\right\} \lesssim \exp(-c'\beta\delta^{2m/(2m-d)}).$$

□

Lemma 3. *Suppose that $\|f^*\|_\infty \leq R$ and $\sup_W \|f_W\|_\infty \leq R$. Then, the logistic loss satisfies the Bernstein condition with $s = 1$ and $C_B = 4 + 3R$.*

Proof. Since $\|f^*\|_\infty \leq R$, it holds that $|h^*(x)| = |h(f^*(x)|x)|$ is also bounded by R (a.s.). Here, we fix $x \in \mathcal{X}$ and write $p = P(Y = 1|X = x)$. By the optimality of $f^*(x)$, we have that $p = \frac{1}{1+\exp(-f^*(x))}$. Accordingly, we denote $q = \frac{1}{1+\exp(-f_W(x))}$ for any $W \in \mathcal{H}$.

Then, what we need to show is that

$$p\left[\log\left(\frac{p}{q}\right)\right]^2 + (1-p)\left[\log\left(\frac{1-p}{1-q}\right)\right]^2 \leq C_B\left\{p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)\right\}. \quad (25)$$

The right hand side can be rewritten as

$$p\left[\log\left(\frac{p}{q}\right) + \frac{1}{p}(q-p)\right] + (1-p)\left[\log\left(\frac{1-p}{1-q}\right) - \frac{1}{1-p}(q-p)\right],$$

and by noticing the convexity of $-\log(\cdot)$, each term of the right hand side is non-negative. We show the inequality (25) by showing

$$\left[\log\left(\frac{p}{q}\right)\right]^2 \leq C_B\left[\log\left(\frac{p}{q}\right) + \frac{1}{p}(q-p)\right], \quad (26)$$

$$\left[\log\left(\frac{1-p}{1-q}\right)\right]^2 \leq C_B\left[\log\left(\frac{1-p}{1-q}\right) - \frac{1}{1-p}(q-p)\right]. \quad (27)$$

Without loss of generality, we may assume $p \leq 1/2$.

Step 1: Proof of Eq. (26). We show the inequality by considering the following four settings (i) $p/2 \leq q \leq p$, (ii) $q < p/2$, (iii) $p \leq q \leq 2p$, (iv) $2p < q$. Let $f_1(q) = \log(p/q) + \frac{1}{p}(q-p)$ and $f_2(q) = [\log(p/q)]^2$.

(i) ($p/2 \leq q \leq p$) Since $f_1(q)$ is a convex function satisfying $\frac{\mathrm{d}^2}{\mathrm{d}q^2}f_1(q) = 1/q^2 \geq 1/p^2$ ($\forall q \leq p$), $f_1(p) = 0$ and $f_1(q) \geq 0$, it holds that $f_1(q) \geq \frac{1}{p^2}(q-p)^2$ for all $q \leq p$. On the other hand, $0 \leq \log(p/q) \leq \frac{2}{p}(p-q)$ for $p/2 \leq q \leq p$, it holds that $f_2(q) \leq \frac{4}{p^2}(q-p)^2$ ($p/2 \leq \forall q \leq p$). These inequalities yield

$$4f_1(q) \geq f_2(q) \;\; (p/2 \leq \forall q \leq p).$$

(ii) ($q < p/2$). Since $f_1(q) \geq \frac{1}{p^2}(p-q)^2$ ($\forall q \leq p$) and $-\frac{1}{p}(q-p) \leq 2\frac{1}{p^2}(q-p)^2 \leq 2f_1(q)$ ($\forall q \leq p/2$), we have that

$$-\frac{1}{3}\log(p/q) \leq \frac{1}{p}(q-p) \leq 0 \quad (\forall q \leq p/2).$$

Therefore, by the definition of f_1, we have

$$f_1(q) \geq \frac{2}{3}\log(p/q) \geq \frac{2}{3\log(p/q)}[\log(p/q)]^2 \geq \frac{2}{3\log(1+\exp(R))}f_2(q),$$

where we used $p \le 1$ and $q = \frac{1}{1+\exp(-f_W(x))} \ge \frac{1}{1+\exp(R)}$.

(iii) ($p \le q \le 2p$). In this setting, the convexity of $-\log(\cdot)$ gives $0 \ge \log(p/q) = -\log(q/p) \ge \frac{1}{p}(p-q)$. Therefore, it holds that $f_2(q) \le \frac{1}{p^2}(p-q)^2$. On the other hand, since $f_2''(q) = \frac{1}{q^2} \ge \frac{1}{4p^2}$, it holds that $f_1(x) \ge \frac{1}{4p^2}(p-q)^2$. Therefore, we have

$$4f_1(q) \ge f_2(q) \ \ (p \le \forall q \le 2p).$$

(iv) ($2p < q$). By the convexity of $-\log(q)$, we have that

$$\begin{aligned}
-\frac{1}{\log(2)}\log(q/p) &\ge \frac{1}{\log(2)}[-\log(q)+\log(p)] \ge \frac{1}{\log(2)}[-\log(2p) - \frac{1}{2p}(q-2p) + \log(p)] \\
&= \frac{1}{\log(2)}[-\log(2) - \frac{1}{2p}(q-2p)] = -1 - \frac{1}{2\log(2)p}(q-2p) \\
&\ge -\frac{1}{p}(q-2p) - 1 = -\frac{1}{p}(q-p) \quad (\forall q > 2p),
\end{aligned}$$

where we used $2\log(2) \ge 1$. This yields that

$$f_1(q) \ge \left(1 - \frac{1}{\log(2)}\right)\log(p/q) = \frac{1-\log(2)}{\log(2)}\log(q/p) \ge 0 \ \ (\forall q > 2p).$$

(Remember that $\log(2) < 1$). Therefore,

$$\begin{aligned}
f_1(q) &\ge \frac{1-\log(2)}{\log(2)\log(q/p)}[\log(q/p)]^2 \ge \frac{1-\log(2)}{\log(2)\log(1+\exp(R))}[\log(q/p)]^2 \\
&= \frac{1-\log(2)}{\log(2)\log(1+\exp(R))} f_2(q) \quad (\forall q > 2p),
\end{aligned}$$

where we used $q \le 1$ and $p \ge \frac{1}{1+\exp(R)}$.

Step 2: Proof of Eq. (27). This is shown completely in the same manner with the proof of Eq. (26) by setting $p \leftarrow 1-p$ and $q \leftarrow 1-q$.

Step 3. Combining the results of Step 1 and Step 3, we have the equations (26) and (27) with

$$C_B = \max\left\{4, \frac{3}{2}\log(1+\exp(R)), \frac{\log(2)}{1-\log(2)}\log(1+\exp(R))\right\} \le 4 + 3R,$$

where we used $\frac{3}{2} \le \frac{\log(2)}{1-\log(2)} \le 3$, $\log(1+\exp(R)) \le \log(2) + R$ by the Lipschitz continuity of $x \mapsto \log(1+\exp(x))$, and $\frac{\log^2(2)}{1-\log(2)} \le 2$. Therefore, by resetting $C_B = 4+3R$, we obtain the assertion. □

D.4 Derivation of the fast rate of regression (Eq. (10))

Since f^* is realized by f_{W^*}, $\|f_W\|_\infty \le R$ for any $W \in \mathcal{H}$ and $|\epsilon_i| \le C$, we have that

$$\ell(Y, f_W(X)) = (Y - f_W(X))^2 = (f_{W^*} + \epsilon - f_W(X))^2 \le 2[(f_{W^*}(X) - f_W(X))^2 + \epsilon^2] \le 2(4R^2 + C^2).$$

Therefore, the assumption $0 \le \ell(Y, f_W(X)) \le \bar{R}$ ($\forall W \in \mathcal{H}$) (a.s.) is obtained by $\bar{R} = 2(4R^2 + C^2)$. Other assumptions in

Write $\mathcal{H}_{\tilde{K}^\theta} := \mathcal{H}_{K^{\theta(\gamma+1)}}$. As we have stated in the main text, we can show the "bias" and "variance" terms can be bounded as

$$\begin{aligned}
\inf_{h \in \mathcal{H}_{\tilde{K}}: \mathcal{L}(h) - \mathcal{L}(f^*) \le \epsilon^2} \lambda_\beta \|h\|_{\mathcal{H}_{\tilde{K}}}^2 &\lesssim \lambda_\beta \epsilon^{-\frac{2(1-\theta)}{\theta}}, \\
-\log \tilde{\nu}_\beta(\{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \le \epsilon\}) &\lesssim (\epsilon/\lambda_\beta^{1/2})^{-\frac{2\tilde{\alpha}}{1-\tilde{\alpha}}}.
\end{aligned}$$

The variance term has been already evaluated in Eq. (19). Now, we evaluate the bias term. By the definitions of $\mathcal{H}_{\tilde{K}}$, $W^* \in \mathcal{H}_{\tilde{K}}$ means that there exists $(a_k)_{k=0}^\infty$ such that

$$W^* = \sum_{k=0}^\infty \sqrt{\mu_k^{\theta(\gamma+1)}} a_k e_k, \ \text{ and } \ \sum_{k=0}^\infty a_k^2 < \infty.$$

Here, we denote $Q := \sum_{k=0}^{\infty} a_k^2$. Now, let $\tilde{W} = \sum_{k=0}^{N} \sqrt{\mu_k^{\theta(\gamma+1)}} a_k e_k$ for some $N \in \mathbb{N}$ as an approximator of W^*. Then, its norm in $\mathcal{H}_{\tilde{K}}$ can be evaluated as

$$\|\tilde{W}\|_{\mathcal{H}_{\tilde{K}}}^2 = \sum_{k=0}^{N} \mu_k^{-(\gamma+1)} \mu_k^{\theta(\gamma+1)} a_k^2 = \sum_{k=0}^{N} \mu_k^{(\theta-1)(\gamma+1)} a_k^2. \tag{28}$$

We evaluate the discrepancy between W^* and $\tilde{W}$ and evaluate its norm in $\mathcal{H}$. Since $W^* - \tilde{W} = \sum_{k=N+1}^{\infty} \sqrt{\mu_k^{\theta(\gamma+1)}} a_k e_k$, its $\mathcal{H}$-norm is given by

$$\|W^* - \tilde{W}\|_{\mathcal{H}}^2 = \sum_{k=N+1}^{\infty} \mu_k^{\theta(\gamma+1)} a_k^2. \tag{29}$$

Note that $\mathcal{L}(f_{\tilde{W}}) - \mathcal{L}(f^*) = \|f_{\tilde{W}} - f^*\|_{L_2(P_X)}^2 \le (1+RD)^2 \|\tilde{W} - W^*\|_{\mathcal{H}}^2$ by Lemma 1. Therefore, to ensure $\mathcal{L}(f_{\tilde{W}}) - \mathcal{L}(f^*) \le \epsilon^2$, it suffices to let $(1+RD)^2 \|\tilde{W} - W^*\|_{\mathcal{H}}^2 \le \epsilon^2$. By Eq. (29), this means that $\sum_{k=N+1}^{\infty} \mu_k^{\theta(\gamma+1)} a_k^2 \le \epsilon^2/(1+RD)^2$. Here, note that

$$\|W^* - \tilde{W}\|_{\mathcal{H}}^2 = \sum_{k=N+1}^{\infty} \mu_k^{\theta(\gamma+1)} a_k^2 \le \mu_{N+1}^{\theta(\gamma+1)} \sum_{k=N+1}^{\infty} a_k^2 \le c_\mu^{\theta(\gamma+1)} (N+2)^{-2\theta(\gamma+1)} Q.$$

Hence, by setting $N \propto \epsilon^{-1/[\theta(\gamma+1)]}$, we can let $\mathcal{L}(f_{\tilde{W}}) - \mathcal{L}(f^*) \le \epsilon^2$. In this setting of N, by noticing $(k+1)^{-2(\theta-1)(\gamma+1)} = (k+1)^{2(1-\theta)(\gamma+1)}$ is monotonically increasing with respect to k, Eq. (28) gives that

$$\begin{aligned}\|\tilde{W}\|_{\mathcal{H}_{\tilde{K}}}^2 &\le \sum_{k=0}^{N} c_\mu^{(\theta-1)(\gamma+1)} (k+1)^{-2(\theta-1)(\gamma+1)} a_k^2 \\ &\le c_\mu^{(\theta-1)(\gamma+1)} (N+1)^{2(1-\theta)(\gamma+1)} Q \lesssim \epsilon^{-2(1-\theta)/\theta},\end{aligned}$$

which gives the bias term bound.

Combining the bias and variance terms, we may choose ϵ^* as the infimum of ϵ such that $\lambda_\beta \epsilon^{-\frac{2(1-\theta)}{\theta}} + (\epsilon/\lambda_\beta^{1/2})^{-\frac{2\tilde{\alpha}}{1-\tilde{\alpha}}} \le \beta\epsilon^2$. That is, we have that

$$\epsilon^{*2} \lesssim \max\left\{\lambda_\beta^{-\tilde{\alpha}} \beta^{-(1-\tilde{\alpha})}, \left(\frac{\lambda_\beta}{\beta}\right)^{\theta}, n^{-\frac{1}{2-s}}\right\} = \max\left\{\lambda^{-\tilde{\alpha}} \beta^{-1}, \lambda^{\theta}, n^{-\frac{1}{2-s}}\right\}.$$