Supplementary Materials

A More notation

We introduce some additional notation to be used in the Appendix. Denote $\boldsymbol{y}^* = (f^*(x_1), \cdots, f^*(x_n))^\top$ as the the vector of underlying function's functional values at sample points. Let $\mathbb{I}_r(\boldsymbol{x}) = \mathbb{I}\{\boldsymbol{w}_r^\top \boldsymbol{x} \geq 0\}$ and

$$\boldsymbol{z}(\boldsymbol{x}) = \frac{1}{\sqrt{m}} \begin{pmatrix} a_1 \mathbb{I}_1(\boldsymbol{x})\boldsymbol{x} \\ \vdots \\ a_m \mathbb{I}_m(\boldsymbol{x})\boldsymbol{x} \end{pmatrix} \in \mathbb{R}^{md \times 1}. \tag{A.1}$$

Thus, $\boldsymbol{Z}(k) = (\boldsymbol{z}(\boldsymbol{x}_1), ..., \boldsymbol{z}(\boldsymbol{x}_n))|_{\boldsymbol{W}=\boldsymbol{W}(k)}$. When the context is clear, we omit the dimension and write $\boldsymbol{I}_d$ as $\boldsymbol{I}$.

B Proof of Lemma 3.1

We will use the following lemma, which states the Mercer decomposition of h as in (3.2).

Lemma B.1 (Mercer decomposition of NTK h)**.** For any $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{S}^{d-1}$, we have the following decomposition of the NTK,

$$h(\boldsymbol{s}, \boldsymbol{t}) = \sum_{k=0}^{\infty} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(\boldsymbol{s}) Y_{k,j}(\boldsymbol{t}),$$

where $Y_{k,j}$, $j = 1, ..., N(d,k)$ are spherical harmonic polynomials of degree k, and the non-negative eigenvalues μ_k satisfy $\mu_k \asymp k^{-d}$, and $\mu_k = 0$ if $k = 2j + 1$ for $k \geq 2$.

The proof of Lemma B.1 is similar to the proof of Proposition 5 in Bietti and Mairal [2019]. The difference is that the Proposition 5 in Bietti and Mairal [2019] considers the kernel function

$$h_1(\boldsymbol{s}, \boldsymbol{t}) = 4h(\boldsymbol{s}, \boldsymbol{t}) + \frac{\sqrt{1 - (\boldsymbol{s}^\top \boldsymbol{t})^2}}{\pi},$$

and we only need to consider the kernel function $h(\boldsymbol{s}, \boldsymbol{t})$. A generalization of Proposition 5 in Bietti and Mairal [2019] can be found in Theorem 3.5 of Cao et al. [2019].

Note that in the proof of Lemma B.1,

$$N(d,j) = \frac{2j + d - 2}{j} \begin{pmatrix} j + d - 3 \\ d - 2 \end{pmatrix} = \frac{\Gamma(j + d - 2)}{\Gamma(d-1)\Gamma(j)},$$

where Γ is the Gamma function. By the Stirling approximation, we have $\Gamma(x) \approx \sqrt{2\pi} x^{x-1/2} e^{-x}$. Therefore, we have the number $N(d,j)$ is equivalent to j^{d-2}. Thus, by Lemma B.1, the j-th eigenvalue λ_j can be denoted by

$$\lambda_j = \mu_l, \text{ for } \sum_{i=1}^{l-1} N(d, 2i) \leq j < \sum_{i=1}^{l} N(d, 2i),$$

which can be approximated by $\lambda_j \asymp \mu_l$, for $(2l-2)^{d-1} \leq j < (2l)^{d-1}$. By Lemma B.1, we have $\mu_l \asymp l^{-d}$, which implies $\lambda_j \asymp j^{-\frac{d}{d-1}}$.

C Proof of Theorem 3.2

Let $\mathcal{G}$ be a metric space equipped with a metric d_g. The δ-covering number of the metric space $(\mathcal{G}, d_g)$, denoted by $N(\delta, \mathcal{G}, d_g)$, is the minimum integer N so that there exist N distinct balls in $(\mathcal{G}, d_g)$ with radius δ, and the union of these balls covers $\mathcal{G}$. Let $H(\delta, \mathcal{G}, d_g) = \log N(\delta, \mathcal{G}, d_g)$ be the entropy of the metric space $(\mathcal{G}, d_g)$. We first present an upper bound on the entropy of the metric space $(\mathcal{N}, \|\cdot\|_\infty)$, where the proof can be found in Appendix F.

Lemma C.1. Let $\mathcal{N}$ be the reproducing kernel Hilbert space generated by the NTK h defined in (3.2), equipped with norm $\|\cdot\|_{\mathcal{N}}$. The entropy $H(\delta, \mathcal{N}(1), \|\cdot\|_\infty)$ can be bounded by

$$H(\delta, \mathcal{N}(1), \|\cdot\|_\infty) \leq A_0 \delta^{-\frac{2(d-1)}{d}}, \tag{C.1}$$

where $\mathcal{N}(1) = \{f : f \in \mathcal{N}, \|f\|_{\mathcal{N}} \leq 1\}$, and $A_0 > 0$ is a constant not depending on δ.

For the regression problem, consider a general penalized least-square estimator

$$\widehat{f} := \operatorname*{argmin}_{f \in \mathcal{N}} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda_n^2 I^v(f) \right),$$

where $\lambda_n > 0$ is the smoothing parameter and $I : \mathcal{N} \to [0, \infty)$ is a pseudo-norm measuring the complexity. We use the RKHS norm $\|f\|_{\mathcal{N}}$ in our case. Let $\|\cdot\|_n$ denote the empirical norm. The following lemma establishes the rate of convergence for the estimator $\widehat{f}$.

Lemma C.2 (Lemma 10.2 in van de Geer [2000])**.** Assume Gaussian noises and entropy bound $H(\delta, \mathcal{N}(1), \|\cdot\|_n) \leq A\delta^{-\alpha}$ for some constants $A > 0$ and $0 < \alpha < 2$. If $v \geq \frac{2\alpha}{2+\alpha}$, $I(f^*) > 0$ and

$$\lambda_n^{-1} = O_{\mathbb{P}}\left(n^{1/(2+\alpha)}\right) I^{(2v-2\alpha+v\alpha)/2(2+\alpha)}(f^*).$$

Then we have

$$\left\| \widehat{f} - f^* \right\|_n = O_{\mathbb{P}}(\lambda_n) I^{v/2}(f^*)$$

and $I(\widehat{f}) = O_{\mathbb{P}}(1) I(f^*)$.

To bound the difference between empirical norm and L_2 norm, we utilize the following lemma. For a class of functions $\mathcal{F}$, define for $z > 0$

$$J_\infty(z, \mathcal{F}) := C_0 \inf_{\delta > 0} \left[z \int_{\delta/4}^{1} \sqrt{\mathcal{H}_\infty(uz/2, \mathcal{F})} du + \sqrt{n} \delta z \right].$$

Lemma C.3 (Theorem 2.2 in van de Geer [2014])**.** Let

$$R := \sup_{f \in \mathcal{F}} \|f\|_2, \ K := \sup_{f \in \mathcal{F}} \|f\|_\infty$$

Then, for all $t > 0$, with probability at least $1 - \exp[-t]$,

$$\sup_{f \in \mathcal{F}} \left| \|f\|_n^2 - \|f\|_2^2 \right| / C_1 \leq \frac{2RJ_\infty(K, \mathcal{F}) + RK\sqrt{t}}{\sqrt{n}} + \frac{4J_\infty^2(K, \mathcal{F}) + K^2 t}{n}$$

where $C_1 > 0$ is some constant not depending on n.

Proof of Theorem 3.2. Consider our estimator $\widehat{f}$ as in (3.4), in which case, $v = 2$ and $I(f)$ is the RKHS norm of f. Since $\|f\|_n \leq \|f\|_\infty$, Lemma C.1 indicates that $\alpha = 2(d-1)/d < 2$. By choosing $\lambda_n \asymp n^{-d/(4d-2)}$, which corresponds to $\mu \asymp n^{(d-1)/(2d-1)}$ in (3.3), Lemma C.2 yields that

$$\left\| \widehat{f} - f^* \right\|_n^2 = O_{\mathbb{P}}(n^{-d/(2d-1)}) \quad \text{and} \quad \left\| \widehat{f} \right\|_{\mathcal{N}}^2 = O_{\mathbb{P}}(1).$$

Now we use Lemma C.3 to obtain a bound on $\left\|\widehat{f}-f^*\right\|_2$. First consider $\{f-f^*: f\in\mathcal{N}(1)\}$, where $\mathcal{N}(1)=\{f\in\mathcal{N}, \|f\|_{\mathcal{N}}\leq 1\}$. Thus, we have $K, R=O(1)$. By the entropy bound in Lemma C.1, we have $J_\infty(z,\mathcal{N}(1))\leq 2C_0 z^{1/d}$. Therefore, Lemma C.3 yields

$$\sup_{f\in\mathcal{N}(1)}\left|\|f-f^*\|_n^2-\|f-f^*\|_2^2\right|=O_{\mathbb{P}}\left(\sqrt{\frac{1}{n}}\right).$$

Combined with $\left\|\widehat{f}-f^*\right\|_n^2=O_{\mathbb{P}}(n^{-d/(2d-1)})$, we can conclude that for any $t>0$ large enough, $\left\|\widehat{f}-f^*\right\|_2^2=O(\sqrt{t/n})$ with probability at least $1-\exp(-t)$. Utilizing Lemma C.3 again with $R=O(\sqrt{t/n})$ we have for some $C>0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{G}(R)}\left|\|f-f^*\|_n^2-\|f-f^*\|_2^2\right|\leq\frac{Ct}{n}\right)\geq 1-e^{-t},$$

where $\mathcal{G}(R):=\{f\in\mathcal{N}(1):\|f-f^*\|_2\leq R\}$. Notice that $\widehat{f}\in\mathcal{G}(R)$ with probability at least $1-\exp(-t)$. Therefore, $\left\|\widehat{f}-f^*\right\|_2^2=O(n^{-d/(2d-1)}+t/n)$ with probability at least $1-2\exp(-t)$. □

D Proofs of main theorems in Section 4

For brevity, let $\widehat{f}_k=f_{\boldsymbol{W}(k),\boldsymbol{a}}$. For two positive semidefinite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we write $\boldsymbol{A}\geq\boldsymbol{B}$ to denote that $\boldsymbol{A}-\boldsymbol{B}$ is positive semidefinite and $\boldsymbol{A}>\boldsymbol{B}$ to denote that $\boldsymbol{A}-\boldsymbol{B}$ is positive definite. This partial order of positive semidefinite matrices is also known as Loewner order. We focus on the L_2 loss of our estimator $\widehat{f}_k$ after k GD updates. Let $\widetilde{f}$ denote the kernel regression solution with kernel $h(\cdot,\cdot)$ that interpolates all $\{(\boldsymbol{x}_i,f^*(\boldsymbol{x}_i))\}_{i=1}^n$, i.e.,

$$g(\boldsymbol{x})=h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}^*. \tag{D.1}$$

We first provide some lemmas used in this section. The proofs of lemmas are presented in Appendix F. Lemma D.1 states some basic inequalities that are also used in the proof of Theorem 5.1. Lemma D.2 provides the convergence rate of interpolant using NTK. Lemmas D.3 can be found in Arora et al. [2019]. Lemma D.4 is implied by the proof in Arora et al. [2019]. Lemma D.5 provides some bounds on the related quantities used in the proofs of Theorems 4.1 and 5.2. Lemma D.6 provide some properties of Loewner order.

Lemma D.1. Let μ be as in Theorem 3.2. Then we have

$$\begin{aligned}
h(\boldsymbol{s},\boldsymbol{s})-h(\boldsymbol{s},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\boldsymbol{s})\geq 0,\\
\int_{\boldsymbol{x}\in\Omega}h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-2}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x}=&O_{\mathbb{P}}(n^{-\frac{d}{2d-1}}),\\
\int_{\boldsymbol{x}\in\Omega}h(\boldsymbol{x},\boldsymbol{x})-h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x}=&O_{\mathbb{P}}(n^{-\frac{1}{2d-1}}),
\end{aligned}$$

where $h(\boldsymbol{x},\boldsymbol{X})=(h(\boldsymbol{x},\boldsymbol{x}_1),...,h(\boldsymbol{x},\boldsymbol{x}_n))$ and $h(\boldsymbol{X},\boldsymbol{x})=h(\boldsymbol{x},\boldsymbol{X})^\top$.

Lemma D.2. Assume the true function $f^*\in\mathcal{N}$ with finite RKHS norm, then $g(\boldsymbol{x})$ defined (D.1) satisfies

$$\|g-f^*\|_2=O_{\mathbb{P}}\left(n^{-1/2}\right).$$

Lemma D.3 (Lemma C.1 in Arora et al. [2019])**.** If $\lambda_0=\lambda_{min}(\boldsymbol{H}^\infty)>0$, $m=\Omega\left(\frac{n^6}{\lambda_0^4\tau^2\delta^3}\right)$ and $\eta=O\left(\frac{\lambda_0}{n^2}\right)$, with probability at least $1-\delta$ over the random initialization, we have

$$\|\boldsymbol{w}_r(k)-\boldsymbol{w}_r(0)\|_2\leq R_0,\quad\forall\ r\in[m],\forall\ k\geq 0,$$

where $R_0=\frac{4\sqrt{n}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2}{\sqrt{m}\lambda_0}$.

Lemma D.4 (Arora et al. [2019])**.** Denote $u_i(k) = f_{\boldsymbol{W}(k),\boldsymbol{a}}(\boldsymbol{x}_i)$ to be the network's prediction on the i-th input and let $\boldsymbol{u}(k) = (u_1(k), ..., u_n(k))^\top \in \mathbb{R}^n$ denote all n predictions on the points $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ at iteration k. We have

$$\boldsymbol{u}(k) - \boldsymbol{y} = (\boldsymbol{I} - \eta \boldsymbol{H}^\infty)^k (\boldsymbol{u}(0) - \boldsymbol{y}) + \boldsymbol{e}(k)$$

where

$$\|\boldsymbol{e}(k)\|_2 = O\left(k \left(1 - \frac{\eta\lambda_0}{4}\right)^{k-1} \frac{\eta n^{5/2} \|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^2}{\sqrt{m}\lambda_0 \tau \delta} \right).$$

Lemma D.5. With probability at least $1 - \delta$, we have

(a) $\|\boldsymbol{Z}(k) - \boldsymbol{Z}(0)\|_F = O\left(\frac{n^{3/4}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^{1/2}}{\sqrt{m^{1/2}\lambda_0\tau\delta}}\right)$;

(b) $\|\boldsymbol{H}(0) - \boldsymbol{H}^\infty\|_F = O\left(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}\right)$;

(c) $\left\|\boldsymbol{z}_0(\cdot)^\top \boldsymbol{Z}(0) - h(\cdot, \boldsymbol{X})\right\|_2 = O\left(\frac{\sqrt{n}\sqrt{\log(n/\delta)}}{\sqrt{m}}\right)$;

(d) $\left\|z_0(\cdot)^\top \mathrm{vec}(\boldsymbol{W}(0))\right\|_2 = O\left(\tau\sqrt{\log(1/\delta)}\right)$.

Lemma D.6 (Properties of Loewner order)**.** For two positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

(a). Suppose $\boldsymbol{A}$ is non-singular, then $\boldsymbol{A} \geq \boldsymbol{B} \Longleftrightarrow \lambda_{max}(\boldsymbol{B}\boldsymbol{A}^{-1}) \leq 1$ and $\boldsymbol{A} > \boldsymbol{B} \Longleftrightarrow \lambda_{\max}(\boldsymbol{B}\boldsymbol{A}^{-1}) > 1$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of the input matrix.

(b). Suppose $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{Q}$ are positive definite, $\boldsymbol{A}$ and $\boldsymbol{B}$ are exchangeable, then $\boldsymbol{A} \geq \boldsymbol{B} \Longrightarrow \boldsymbol{A}\boldsymbol{Q}\boldsymbol{A} \geq \boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}$.

D.1 Proof of Theorem 4.1

For notational simplification, we use $\widehat{f}_k = f_{\boldsymbol{W}(k),\boldsymbol{a}}$. Define

$$\widetilde{f}_k(\boldsymbol{x}) = \mathrm{vec}(\boldsymbol{W}(k))^\top \boldsymbol{z}_0(\boldsymbol{x}), \tag{D.2}$$

where $\boldsymbol{z}_0(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x})|_{\boldsymbol{W}=\boldsymbol{W}(0)}$. Then we can write the following decomposition

$$\widehat{f}_k - f^* = (\widehat{f}_k - \widetilde{f}_k) + (\widetilde{f}_k - g) + (g - f^*) = \Delta_1 + \Delta_2 + \Delta_3, \tag{D.3}$$

where g is as in (D.1).

Before the proof, we provide a road map of this proof. We first show that $\|\Delta_1\|_2$ and $\|\Delta_3\|_2$ are small. We then show the term $\|\Delta_2\|_2$ can be large if the iteration number is too small or too large. Intuitively, if the iteration number if too small, the resulting estimator $\widetilde{f}_k$ is not well-trained. On the other hand, if the iteration number is too large, then the resulting estimator $\widetilde{f}_k$ could be over-fitted. In either case, the error term $\|\Delta_2\|_2$ is large.

It follows from Lemma D.2 that

$$\|\Delta_3\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{1}{n}}\right). \tag{D.4}$$

For Δ_1, under the assumptions of Lemma D.3, with high probability, we have $\|\boldsymbol{w}_r(k) - \boldsymbol{w}_r(0)\|_2 \leq R_0$. Thus, for fixed $\boldsymbol{x}$, we have

$$|\boldsymbol{w}_r(k)^\top \boldsymbol{x} - \boldsymbol{w}_r(0)^\top \boldsymbol{x}| \leq \|\boldsymbol{w}_r(k) - \boldsymbol{w}_r(0)\|_2 \|\boldsymbol{x}\|_2 \leq R_0.$$

Define event

$$B_r(\boldsymbol{x}) = \{|\boldsymbol{w}_r(0)^\top \boldsymbol{x}| \leq R_0\}, \forall r \in [m].$$

If $\mathbb{I}\{B_r(\boldsymbol{x})\} = 0$, then we have $\mathbb{I}_{r,k}(\boldsymbol{x}) = \mathbb{I}_{r,0}(\boldsymbol{x})$, where $\mathbb{I}_{r,k}(\boldsymbol{x}) = \mathbb{I}\{\boldsymbol{w}_r(k)^\top \boldsymbol{x} \geq 0\}$. Therefore, for any fixed $\boldsymbol{x}$, we have

$$
\begin{aligned}
|\widehat{f}_k(\boldsymbol{x}) - \widetilde{f}_k(\boldsymbol{x})| &= \left| \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r (\mathbb{I}_{r,k}(\boldsymbol{x}) - \mathbb{I}_{r,0}(\boldsymbol{x})) \boldsymbol{w}_r(k)^\top \boldsymbol{x} \right| \\
&= \left| \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathbb{I}\{B_r(\boldsymbol{x})\} (\mathbb{I}_{r,k}(\boldsymbol{x}) - \mathbb{I}_{r,0}(\boldsymbol{x})) \boldsymbol{w}_r(k)^\top \boldsymbol{x} \right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}\{B_r(\boldsymbol{x})\} |\boldsymbol{w}_r(k)^\top \boldsymbol{x}| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}\{B_r(\boldsymbol{x})\} \left(|\boldsymbol{w}_r(0)^\top \boldsymbol{x}| + |\boldsymbol{w}_r(k)^\top \boldsymbol{x} - \boldsymbol{w}_r(0)^\top \boldsymbol{x}| \right) \\
&\leq \frac{2R_0}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}\{B_r(x)\}
\end{aligned}
$$

Recall that $\|\boldsymbol{x}\|_2 = 1$, which implies that $\boldsymbol{w}_r(0)^\top \boldsymbol{x}$ is distributed as $N(0, \tau^2)$. Therefore, we have

$$
\mathbb{E}[\mathbb{I}\{B_r(x)\}] = \mathbb{P}\left(|\boldsymbol{w}_r(0)^\top \boldsymbol{x}| \leq R_0 \right) = \int_{-R_0}^{R_0} \frac{1}{\sqrt{2\pi}\tau} \exp\left\{ -\frac{u^2}{2\tau^2} \right\} du \leq \frac{2R_0}{\sqrt{2\pi}\tau}.
$$

By Markov's inequality, with probability at least $1 - \delta$, we have

$$
\sum_{r=1}^{m} \mathbb{I}\{B_r(x)\} \leq \frac{2mR_0}{\sqrt{2\pi}\tau\delta}.
$$

Thus, we have

$$
\|\Delta_1\|_2 \leq \frac{2R_0}{\sqrt{m}} \left\| \sum_{r=1}^{m} \mathbb{I}\{B_r(\cdot)\} \right\|_2 \leq \frac{4\sqrt{m}R_0^2}{\sqrt{2\pi}\tau\delta} = O\left(\frac{n \|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^2}{\sqrt{m}\tau\lambda_0^2\delta} \right). \tag{D.5}
$$

Next, we evaluate Δ_2. Recall that the GD update rule is

$$
\mathrm{vec}(\boldsymbol{W}(j+1)) = \mathrm{vec}(\boldsymbol{W}(j)) - \eta \boldsymbol{Z}(j)(\boldsymbol{u}(j) - \boldsymbol{y}), j \geq 0.
$$

Applying Lemma D.4, we can get

$$
\begin{aligned}
&\mathrm{vec}(\boldsymbol{W}(k)) - \mathrm{vec}(\boldsymbol{W}(0)) \\
=& \sum_{j=0}^{k-1} (\mathrm{vec}(\boldsymbol{W}(j+1)) - \mathrm{vec}(\boldsymbol{W}(j))) \\
=& -\sum_{j=0}^{k-1} \eta \boldsymbol{Z}(j)(\boldsymbol{u}(j) - \boldsymbol{y}) \\
=& \sum_{j=0}^{k-1} \eta \boldsymbol{Z}(j)(\boldsymbol{I} - \eta \boldsymbol{H}^\infty)^j (\boldsymbol{y} - \boldsymbol{u}(0)) - \sum_{j=0}^{k-1} \eta \boldsymbol{Z}(j) \boldsymbol{e}(j) \\
=& \sum_{j=0}^{k-1} \eta \boldsymbol{Z}(0)(\boldsymbol{I} - \eta \boldsymbol{H}^\infty)^j (\boldsymbol{y} - \boldsymbol{u}(0)) + \sum_{j=0}^{k-1} \eta (\boldsymbol{Z}(j) - \boldsymbol{Z}(0))(\boldsymbol{I} - \eta \boldsymbol{H}^\infty)^j (\boldsymbol{y} - \boldsymbol{u}(0)) - \sum_{j=0}^{k-1} \eta \boldsymbol{Z}(j) \boldsymbol{e}(j) \\
=& \sum_{j=0}^{k-1} \eta \boldsymbol{Z}(0)(\boldsymbol{I} - \eta \boldsymbol{H}^\infty)^j (\boldsymbol{y} - \boldsymbol{u}(0)) + \zeta(k).
\end{aligned}
$$

For the first term of $\zeta(k)$, applying Lemma D.5 (a), with probability at least $1-\delta$, we get

$$\begin{aligned}
&\left\|\sum_{j=0}^{k-1}\eta(\boldsymbol{Z}(j)-\boldsymbol{Z}(0))(\boldsymbol{I}-\eta\boldsymbol{H}^{\infty})^{j}(\boldsymbol{y}-\boldsymbol{u}(0))\right\|_2\\
\leq&\sum_{j=0}^{k-1}O\left(\frac{n^{3/4}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^{1/2}}{\sqrt{m^{1/2}\lambda_0\tau\delta}}\right)\eta\left\|\boldsymbol{I}-\eta\boldsymbol{H}^{\infty}\right\|_2^{j}\left\|(\boldsymbol{y}-\boldsymbol{u}(0))\right\|_2\\
\leq&O\left(\frac{n^{3/4}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^{3/2}}{\sqrt{m^{1/2}\lambda_0\tau\delta}}\right)\sum_{j=0}^{k-1}\eta(1-\eta\lambda_0)^{j}\\
=&O\left(\frac{n^{3/4}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^{3/2}}{m^{1/4}\tau^{1/2}\lambda_0^{3/2}\delta^{1/2}}\right).
\end{aligned}$$

Denote that $z_i(j)=z(\boldsymbol{x}_i)|_{\boldsymbol{W}=\boldsymbol{W}(j)}$. By (A.1), we have $\|\boldsymbol{z}_i(j)\|_2\leq 1$. Thus,

$$\left\|\boldsymbol{Z}(j)\right\|_F=\left(\sum_{i=1}^{n}\left\|\boldsymbol{z}_i(j)\right\|_2^2\right)^{\frac{1}{2}}\leq\sqrt{n}\quad,\forall\ j\geq 0. \tag{D.6}$$

For the second term of $\zeta(k)$, we have

$$\begin{aligned}
&\left\|\sum_{j=0}^{k-1}\eta\boldsymbol{Z}(j)\boldsymbol{e}(j)\right\|_2\\
\leq&\sum_{j=0}^{k-1}\eta\left\|\boldsymbol{Z}(j)\right\|_F\left\|\boldsymbol{e}(j)\right\|_2\\
\leq&\sum_{j=0}^{k-1}\eta\sqrt{n}O\left(j\left(1-\frac{\eta\lambda_0}{4}\right)^{j-1}\frac{\eta n^{5/2}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^2}{\sqrt{m}\tau\lambda_0\delta}\right)\\
=&O\left(\frac{n^{3}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^2}{\sqrt{m}\lambda_0^{3}\tau\delta}\right).
\end{aligned}$$

Therefore,

$$\left\|\zeta(k)\right\|_2=O\left(\frac{n^{3/4}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^{3/2}}{m^{1/4}\tau^{1/2}\lambda_0^{3/2}\delta^{1/2}}\right)+O\left(\frac{n^{3}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^2}{\sqrt{m}\lambda_0^{3}\tau\delta}\right). \tag{D.7}$$

Define $\boldsymbol{G}_k=\sum_{j=0}^{k-1}\eta(\boldsymbol{I}-\eta\boldsymbol{H}^{\infty})^{j}$. Recalling that $\boldsymbol{y}=\boldsymbol{y}^*+\boldsymbol{\epsilon}$, for fixed $\boldsymbol{x}$, we have

$$\begin{aligned}
\widetilde{f}_k(\boldsymbol{x})-g(\boldsymbol{x})=&\boldsymbol{z}_0(\boldsymbol{x})^{\top}\mathrm{vec}(\boldsymbol{W}(k))-h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\\
=&\boldsymbol{z}_0(\boldsymbol{x})^{\top}\left[\boldsymbol{Z}(0)\boldsymbol{G}_k(\boldsymbol{y}-\boldsymbol{u}(0))+\zeta(k)+\mathrm{vec}(\boldsymbol{W}(0))\right]\\
=&\left[h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{G}_k-(\boldsymbol{H}^{\infty})^{-1})\boldsymbol{y}^*+h(\boldsymbol{x},\boldsymbol{X})\boldsymbol{G}_k\boldsymbol{\epsilon}\right]+\left[\boldsymbol{z}_0(\boldsymbol{x})^{\top}\boldsymbol{Z}(0)-h(\boldsymbol{x},\boldsymbol{X})\right]\boldsymbol{G}_k\boldsymbol{y}\\
&+\left[\boldsymbol{z}_0(\boldsymbol{x})^{\top}\mathrm{vec}(\boldsymbol{W}(0))+\boldsymbol{z}_0(\boldsymbol{x})^{\top}\zeta(k)-\boldsymbol{z}_0(\boldsymbol{x})^{\top}\boldsymbol{Z}(0)\boldsymbol{G}_k\boldsymbol{u}(0)\right]\\
=&\Delta_{21}(\boldsymbol{x})+\Delta_{22}(\boldsymbol{x})+\Delta_{23}(\boldsymbol{x}).
\end{aligned} \tag{D.8}$$

Using Lemma D.5 (c), we can bound Δ_{22} as

$$\begin{aligned}
\left\|\Delta_{22}\right\|_2\leq&\left\|\boldsymbol{z}_0(\boldsymbol{x})^{\top}\boldsymbol{Z}(0)-h(\boldsymbol{x},\boldsymbol{X})\right\|_2\left\|\boldsymbol{G}_k\boldsymbol{y}\right\|_2\\
\leq&O\left(\frac{\sqrt{n}\sqrt{\log(n/\delta)}}{\sqrt{m}}\right)\left\|(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}\right\|_2\\
=&O\left(\frac{\sqrt{n}\sqrt{\log(n/\delta)}\left\|\boldsymbol{y}\right\|_2}{\sqrt{m}\lambda_0}\right).
\end{aligned} \tag{D.9}$$

Since the i-th coordinate of $\boldsymbol{u}(0)$ is

$$u_i(0) = \boldsymbol{z}_0(\boldsymbol{x}_i)^\top \mathrm{vec}(\boldsymbol{W}(0)) = \sum_{r=1}^{m} a_r \boldsymbol{w}(0)^\top \boldsymbol{x}_i \mathbb{I}\{\boldsymbol{w}(0)^\top \boldsymbol{x}_i\},$$

where $a_r \sim \mathrm{unif}\{1,-1\}$ and $\boldsymbol{w}(0)^\top \boldsymbol{x}_i \sim N(0,\tau^2)$, it is easy to prove that $u_i(0)$ has zero mean and variance τ^2. This implies $\mathbb{E}[\|\boldsymbol{u}(0)\|_2^2] = O(n\tau^2)$. By Markov's inequality, with probability at least $1-\delta$, we have $\|\boldsymbol{u}(0)\|_2 = O\left(\frac{\sqrt{n}\tau}{\delta}\right)$. Similar to (D.6), we can obtain $\|\boldsymbol{Z}(0)\|_F = O(\sqrt{n})$. Thus,

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top \boldsymbol{Z}(0)\boldsymbol{G}_k\boldsymbol{u}(0)| \le \|\boldsymbol{z}_0(\boldsymbol{x})\|_2 \|\boldsymbol{Z}(0)\|_F \|\boldsymbol{G}_k\boldsymbol{u}(0)\|_2 \le \sqrt{n}\left\|(\boldsymbol{H}^\infty)^{-1}\boldsymbol{u}(0)\right\|_2 = O\left(\frac{n\tau}{\lambda_0\delta}\right). \tag{D.10}$$

Combining Lemma D.5 (d), (D.7) and (D.10), we obtain

$$\begin{aligned}
\|\Delta_{23}\|_2 \le& \left\|\boldsymbol{z}_0(\cdot)^\top \mathrm{vec}(\boldsymbol{W}(0))\right\|_2 + \|\boldsymbol{z}_0(\cdot)\|_2 \|\zeta(k)\|_2 + \left\|\boldsymbol{z}_0(\cdot)^\top \boldsymbol{Z}(0)\boldsymbol{G}_k\boldsymbol{u}(0)\right\|_2 \\
=& O\left(\tau\sqrt{\log(1/\delta)}\right) + O\left(\frac{n^{3/4}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^{3/2}}{m^{1/4}\tau^{1/2}\lambda_0^{3/2}\delta^{1/2}}\right) + O\left(\frac{n^3\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^2}{\sqrt{m}\lambda_0^3\tau\delta}\right) + O\left(\frac{n\tau}{\lambda_0\delta}\right) \\
=& O\left(\frac{n^{3/4}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^{3/2}}{m^{1/4}\tau^{1/2}\lambda_0^{3/2}\delta^{1/2}}\right) + O\left(\frac{n^3\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^2}{\sqrt{m}\lambda_0^3\tau\delta}\right) + O\left(\frac{n\tau}{\lambda_0\delta}\right).
\end{aligned} \tag{D.11}$$

By (D.3) and (D.8), we can rewrite $\widehat{f}_k - f^*$ as

$$\widehat{f}_k - f^* = \Delta_{21} + (\Delta_1 + \Delta_3 + \Delta_{22} + \Delta_{23}) := \Delta_{21} + \Xi,$$

Next we show that the expected value of $\|\Xi\|_2^2$ over noise, $\mathbb{E}_{\boldsymbol{\epsilon}}\|\Xi\|_2^2$, is small. Note that we have

$$\mathbb{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{y}\|_2^2 = \mathbb{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{y}^* + \boldsymbol{\epsilon}\|_2^2 \le 2\boldsymbol{y}^{*\top}\boldsymbol{y}^* + 2\mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} = O(n). \tag{D.12}$$

By Markov's inequality, with probability $1-\delta$ over random initialization, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2 \le& \left(\mathbb{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^2\right)^{\frac{1}{2}} \\
\le& \left(\frac{3\mathbb{E}_{\boldsymbol{W}(0),\boldsymbol{a}}\left[\boldsymbol{u}(0)^\top\boldsymbol{u}(0) + \boldsymbol{y}^{*\top}\boldsymbol{y}^* + \mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}\right]}{\delta}\right)^{\frac{1}{2}} \\
=& O\left(\sqrt{\frac{n(1+\tau^2)}{\delta}}\right) = O\left(\sqrt{\frac{n}{\delta}}\right),
\end{aligned} \tag{D.13}$$

where the last equality of D.13 is because $\tau^2 \lesssim 1$. By (D.4), (D.5), (D.9), (D.11), (D.12) and (D.13), $\mathbb{E}_{\boldsymbol{\epsilon}}\|\Xi\|_2^2$ can be upper bounded as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\|\Xi\|_2^2 \le& 4\mathbb{E}_{\boldsymbol{\epsilon}}(\|\Delta_1\|_2^2 + \|\Delta_3\|_2^2 + \|\Delta_{22}\|_2^2 + \|\Delta_{23}\|_2^2) \\
=& \mathbb{E}_{\boldsymbol{\epsilon}}\left[O\left(\frac{n^2\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^4}{m\tau^2\lambda_0^4\delta^2}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{n\log(n/\delta)\|\boldsymbol{y}\|_2^2}{m\lambda_0^2}\right)\right] + 4\mathbb{E}_{\boldsymbol{\epsilon}}\|\Delta_{23}\|_2^2 \\
\le& O\left(\frac{n^4}{m\tau^2\lambda_0^4\delta^4}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{n^2\log(n/\delta)}{m\lambda_0^2\delta}\right) + O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right) + \\
&+ \mathbb{E}_{\boldsymbol{\epsilon}}\left[O\left(\frac{n^{3/2}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^3}{m^{1/2}\tau\lambda_0^3\delta}\right) + O\left(\frac{n^6\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^4}{m\tau^2\lambda_0^6\delta^2}\right)\right] \\
=& O\left(\frac{n^4}{m\tau^2\lambda_0^4\delta^4}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{n^2\log(n/\delta)}{m\lambda_0^2\delta}\right) + O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right) \\
&+ O\left(\frac{n^3}{\sqrt{m}\tau\lambda_0^3\delta^{5/2}}\right) + O\left(\frac{n^8}{m\tau^2\lambda_0^6\delta^4}\right) \\
=& O\left(\frac{1}{n}\right) + O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right) + \frac{\mathrm{poly}\left(n, \frac{1}{\lambda_0}, \frac{1}{\delta}\right)}{m^{\frac{1}{2}}\tau}.
\end{aligned}$$

In the following, we will evaluate Δ_{21} and discuss how the iteration number k would affect the L_2 estimation error $\left\|\widehat{f}_k - f^*\right\|_2^2$.

Case 1: The iteration number k cannot be too small By taking expectation of $\|\Delta_{21}\|_2^2$ over the noise, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}} \|\Delta_{21}\|_2^2 &= \int_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x},\boldsymbol{X})\big[(\boldsymbol{H}^\infty)^{-1} - \boldsymbol{G}_k)\boldsymbol{y}^*\boldsymbol{y}^{*\top}((\boldsymbol{H}^\infty)^{-1} - \boldsymbol{G}_k) + \boldsymbol{G}_k^2\big]h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{M}_k(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x},\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{M}_k =&(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k\boldsymbol{S}(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k + (\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k)^2 \\ =&[(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k - (\boldsymbol{S}+\boldsymbol{I})^{-1}](\boldsymbol{S}+\boldsymbol{I})[(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k - (\boldsymbol{S}+\boldsymbol{I})^{-1}] + \boldsymbol{I} - (\boldsymbol{S}+\boldsymbol{I})^{-1}\end{aligned} \tag{D.14}$$

and $\boldsymbol{S} = \boldsymbol{y}^*\boldsymbol{y}^{*\top}$. If $k \geq C_0\left(\frac{\log n}{\eta\lambda_0}\right)$ for some constant $C_0 > 1$, we have

$$(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k \leq (1-\eta\lambda_0)^k\boldsymbol{I} \leq \exp\{-\eta\lambda_0 k\}\boldsymbol{I} \leq \exp\{-C_0\log n\}\boldsymbol{I} = \frac{1}{n^{C_0}}\boldsymbol{I},$$

Since $1 + \|\boldsymbol{y}^*\|_2^2 \leq C_1 n$ for some constant C_1, we have

$$\lambda_{\max}\left(\frac{1}{n^{C_0}}(\boldsymbol{S}+\boldsymbol{I})\right) = \frac{1+\|\boldsymbol{y}^*\|_2^2}{n^{C_0}} \leq \frac{C_1}{n^{C_0-1}} < 1.$$

By Lemma D.6 (a), we have

$$(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k \leq \frac{1}{n^{C_0}}\boldsymbol{I} < (\boldsymbol{S}+\boldsymbol{I})^{-1}.$$

Therefore, we have

$$(\boldsymbol{S}+\boldsymbol{I})^{-1} - (\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k \geq (\boldsymbol{S}+\boldsymbol{I})^{-1} - \frac{1}{n^{C_0}}\boldsymbol{I},$$

where $(\boldsymbol{S}+\boldsymbol{I})^{-1} - (\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k$ and $(\boldsymbol{S}+\boldsymbol{I})^{-1} - n^{-C_0}\boldsymbol{I}$ are positive definite matrices. It is also obvious that the two matrices are exchangeable. By Lemma D.6 (b) and (D.14), we have

$$\boldsymbol{M}_k \geq \left(1 - \frac{1}{n^{C_0}}\right)^2\boldsymbol{I} + \frac{1}{n^{2C_0}}\boldsymbol{S}.$$

Then we have

$$\mathbb{E}_{\boldsymbol{\epsilon}} \|\Delta_{21}\|_2^2 \geq \left(1 - \frac{1}{n^{C_0}}\right)^2 I_1 + \frac{1}{n^{2C_0}} I_2 \geq c_0 I_1$$

where $c_0 \in (0,1)$ is a constant,

$$I_1 = \int h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-2}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x}, \quad \text{and} \quad I_2 = \int [h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}^*]^2 d\boldsymbol{x}.$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\widehat{f}_k - f^*\right\|_2^2 =& \mathbb{E}_{\boldsymbol{\epsilon}} \|\Delta_{21} + \Xi\|_2^2 \\ \geq& \frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}} \|\Delta_{21}\|_2^2 - \mathbb{E}_{\boldsymbol{\epsilon}} \|\Xi\|_2^2 \\ \geq& \frac{c_0}{2} I_1 - O\left(\frac{1}{n}\right) - O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right) - \frac{\text{poly}\left(n, \frac{1}{\lambda_0}, \frac{1}{\delta}\right)}{m^{\frac{1}{2}}\tau}.\end{aligned} \tag{D.15}$$

Let $\tau \leq C_3 \frac{\lambda_0 \delta}{n} \left\|(\boldsymbol{H}^\infty)^{-1} h(\boldsymbol{X}, \cdot)\right\|_2$ for some constant $C_3 > 0$ such that the third term of (D.15) is bounded by $\frac{c_0}{4} \left\|(\boldsymbol{H}^\infty)^{-1} h(\boldsymbol{X}, \cdot)\right\|_2^2$. Therefore, $\mathbb{E}_{\boldsymbol{\epsilon}} \left\|\widehat{f}_k - f^*\right\|_2^2$ can be lower bounded as

$$\mathbb{E}_{\boldsymbol{\epsilon}} \left\|\widehat{f}_k - f^*\right\|_2^2 \geq C_1^* \left\|(\boldsymbol{H}^\infty)^{-1} h(\boldsymbol{X}, \cdot)\right\|_2^2 - O\left(\frac{1}{n}\right), \tag{D.16}$$

where $C_1^* > 0$ is a constant. Note that I_1 is $\mathbb{E}_\epsilon \left\|\widehat{f}_\infty - g^*\right\|_2^2$, where $g^* \equiv 0$ and $\widehat{f}_\infty$ is the interpolated estimator of g^*, as in Theorem 4.2. Therefore, by Theorem 4.2, there exists a constant c_1 such that $\mathbb{E}_\epsilon \left\|\widehat{f}_\infty - g^*\right\|_2^2 \geq c_1$, which implies $I_1 \geq c_1$. Taking n large enough such that the second term in (D.16) is smaller than $C_1^* c_1$, we finish the proof of the case that k is large.

Case 2: The iteration number k cannot be too large We can rewrite Δ_{21} as

$$\begin{aligned}
\Delta_{21} =& h(\boldsymbol{x}, \boldsymbol{X}) \boldsymbol{G}_k (\boldsymbol{y}^* + \boldsymbol{\epsilon}) - h(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^* \\
=& \Delta_{21}^* - h(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*.
\end{aligned}$$

Since

$$\boldsymbol{G}_k = \sum_{j=0}^{k-1} \eta (\boldsymbol{I} - \eta \boldsymbol{H}^\infty)^j = \sum_{j=0}^{k-1} \eta \sum_{i=1}^{n} (1 - \eta \lambda_i)^j \boldsymbol{v}_i \boldsymbol{v}_i^\top \leq \eta k \boldsymbol{I},$$

we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} \left\|\Delta_{21}^*\right\|_2^2 =& \int_{\boldsymbol{x} \in \Omega} h(\boldsymbol{x}, \boldsymbol{X}) \boldsymbol{G}_k (\boldsymbol{S} + \boldsymbol{I}) \boldsymbol{G}_k h(\boldsymbol{X}, \boldsymbol{x}) d\boldsymbol{x} \\
\leq& \eta^2 k^2 \int_{\boldsymbol{x} \in \Omega} h(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{S} + \boldsymbol{I}) h(\boldsymbol{X}, \boldsymbol{x}) d\boldsymbol{x} \\
=& \eta^2 k^2 \left(\int_{\boldsymbol{x} \in \Omega} \left[h(\boldsymbol{x}, \boldsymbol{X}) \boldsymbol{y}^*\right]^2 d\boldsymbol{x} + \left\|h(\cdot, \boldsymbol{X})\right\|_2^2 \right) \\
=& O\left(\eta^2 k^2 n^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} \left\|\widehat{f}_k - f^*\right\|_2^2 =& \mathbb{E}_{\boldsymbol{\epsilon}} \left\|\Delta_{21}^* + \Xi - h(\cdot, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*\right\|_2^2 \\
\geq& \frac{1}{2} \left\|h(\cdot, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*\right\|_2^2 - \mathbb{E}_{\boldsymbol{\epsilon}} \left\|\Delta_{21}^* + \Xi\right\|_2^2 \\
\geq& \frac{1}{2} \left\|h(\cdot, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*\right\|_2^2 - 2\mathbb{E}_{\boldsymbol{\epsilon}} \left\|\Delta_{21}^*\right\|_2^2 - 2\mathbb{E}_{\boldsymbol{\epsilon}} \left\|\Xi\right\|_2^2 \\
\geq& \frac{1}{2} \left\|h(\cdot, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*\right\|_2^2 - O\left(\eta^2 k^2 n^2\right) \\
& - O\left(\frac{1}{n}\right) - O\left(\frac{n^2 \tau^2}{\lambda_0^2 \delta^2}\right) - \frac{\mathrm{poly}\left(n, \frac{1}{\lambda_0}, \frac{1}{\delta}\right)}{m^{\frac{1}{2}} \tau}.
\end{aligned} \tag{D.17}$$

Let $k \leq C_1 \left(\frac{1}{\eta n}\right)$ for some constant $C_1 > 0$ such that the the second term of (D.17) can be bounded by $\frac{1}{8} \left\|h(\cdot, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*\right\|_2^2$. Let $\tau \leq C_2 \left(\frac{\delta \lambda_0}{n}\right)$ for some constant $C_2 > 0$ such that the fourth term in (D.17) can be bounded by $\frac{1}{8} \left\|h(\cdot, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*\right\|_2^2$. Note that we can also choose m such that the fifth term in (D.17) is bounded by $\frac{1}{8} \left\|h(\cdot, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*\right\|_2^2$. Therefore, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} \left\|\widehat{f}_k - f^*\right\|_2^2 \geq& C_2^* \left\|h(\cdot, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1} \boldsymbol{y}^*\right\|_2^2 - O\left(\frac{1}{n}\right) \\
\geq& C_3^* \left\|f^*\right\|_2^2 - O\left(\frac{1}{n}\right),
\end{aligned} \tag{D.18}$$

where the last inequality is because of Lemma D.2, and $C_2^* > 0$ is a constant. By taking n large enough such that the second term in (D.18) is smaller than $C_3^* \left\|f^*\right\|_2^2 / 2$, we finish the proof.

D.2 Proof of Theorem 4.2

Let's first introduce the GD update for the kernel ridge regression. By the representer theorem [Kimeldorf and Wahba, 1971], the kernel estimator can be written as

$$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \omega_i h(\boldsymbol{x}, \boldsymbol{x}_i) := h(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{\omega},$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ is the coefficient vector. Consider using the squared loss

$$\Phi(\boldsymbol{\omega}) = \frac{1}{2}\sum_{i=1}^{n} (\widehat{f}(\boldsymbol{x}_i) - y_i)^2.$$

Let $\boldsymbol{\omega}_k$ be the $\boldsymbol{\omega}$ at the k-th GD iteration and choose $\boldsymbol{\omega}_0 = \mathbf{0}$. Then, the GD update rule for estimating $\boldsymbol{\omega}$ can be expressed as

$$\boldsymbol{\omega}_{k+1} = \boldsymbol{\omega}_k - \eta\left((\boldsymbol{H}^{\infty})^2\boldsymbol{\omega} - \boldsymbol{H}^{\infty}\boldsymbol{y}\right) \tag{D.19}$$

In the formulation of the stopping rule, two quantities play an important role: first, the running sum of the step sizes $\alpha_j := \sum_{i=0}^{j} \eta_i$, and secondly, the eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_n \geq 0$ of the empirical kernel matrix H^{∞}, which are computable from the data. Recall the definition of the optimal stopping time k^* as in (4.2). The following lemma establishes the L_2 estimation results for $\widehat{f}_{k^*}$ for kernels with polynomial eigendecay.

Lemma D.7 (Corollary 1 in Raskutti et al. [2014])**.** Suppose that variables $\{\boldsymbol{x}_i\}_{i=1}^n$ are sampled i.i.d. and the kernel class $\mathcal{N}$ satisfies the polynomial eigenvalue decay $\lambda_j \lesssim j^{-2\nu}$ for some $\nu > 1/2$. Then there is a universal constant C such that

$$\mathbb{E}\left\|\widehat{f}_{k^*} - f^*\right\|_2^2 \leq C\left(\frac{\sigma^2}{n}\right)^{\frac{2\nu}{2\nu+1}}.$$

Moreover, if $\lambda_j \asymp j^{-2\nu}$ for all $j = 1, 2, \ldots$, then for all iterations $k = 1, 2, \ldots$,

$$\mathbb{E}\left\|\widehat{f}_{k^*} - f^*\right\|_2^2 \geq \frac{\sigma^2}{4}\min\left\{1,\ \frac{(\alpha_k)^{\frac{1}{2\nu}}}{n}\right\}.$$

By Lemma 3.1, apply Lemma D.7 with $2\nu = d/(d-1)$ and the running sum of the step sizes $\alpha_k = k\eta$ gives the convergence rate.

Moreover, if $k \to \infty$, i.e., interpolation of training data, the lower bound result in Lemma D.7 implies $\mathbb{E}\left\|f_{\widehat{T}} - f^*\right\|_2^2 \gtrsim \sigma^2$ that doesn't converge to 0.

E Proofs of main theorems in Section 5

E.1 Proof of Theorem 5.1

Let $\boldsymbol{u}_D(l) = (u_{D,1}(l), ..., u_{D,n}(l))^{\top} \in \mathbb{R}^n$ be the predictions on the points $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ using the modified GD at the k-th iteration. The idea of the proof is to establish a relationship between $\boldsymbol{y} - \boldsymbol{u}_D(l)$ and $\boldsymbol{y} - \boldsymbol{u}_D(l+1)$ for all $l = 0, 1, ...$, so that we can obtain a relationship between $\boldsymbol{u}_D(l+1)$ and $\boldsymbol{u}_D(0)$. Based on this relationship, we can show that $\boldsymbol{u}_D(l+1)$ is close to $\boldsymbol{H}^{\infty}(C\mu I + \boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}$, which is $\widehat{f}$.

Consider event

$$A_{ir} = \{\exists \boldsymbol{w} \in \mathbb{R}^d : \left\|\boldsymbol{w} - (1-\eta_2\mu)^k \boldsymbol{w}_r(0)\right\|_2 \leq R, \mathbb{I}\{\boldsymbol{x}_i^{\top}\boldsymbol{w}_r(0) \geq 0\} \neq \mathbb{I}\{\boldsymbol{x}_i^{\top}\boldsymbol{w} \geq 0\}\},$$

where R will be determined later. Set $S_i = \{r \in [m] : \mathbb{I}\{A_{ir}\} = 0\}$ and $S_i^{\perp} = [m] \backslash S_i$. Then A_{ir} happens if and only if $|\boldsymbol{w}_r(0)^{\top}\boldsymbol{x}_i| < R/(1-\eta_2\mu)^k$. By concentration inequality of Gaussian, we have $\mathbb{P}(A_{ir}) = \mathbb{P}(|\boldsymbol{w}_r(0)^{\top}\boldsymbol{x}_i| <$

$R/(1-\eta_2\mu)^k \leq \frac{2R}{\sqrt{2\pi}\tau(1-\eta_2\mu)^k}$. Thus, it follows the union bound inequality that with probability at least $1-\delta$ we have

$$\sum_{i=1}^{n}|S_i^{\perp}| \leq \frac{CmnR}{\delta(1-\eta_2\mu)^k}, \tag{E.1}$$

where C is a positive constant.

We first study the difference between two predictions $\boldsymbol{u}_D(l+1)$ and $\boldsymbol{u}_D(l)$. For any $i \in [n]$, we have

$$\begin{aligned}
u_{D,i}(l+1) - (1-\eta_2\mu)u_{D,i}(l) =& \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\
=& \frac{1}{\sqrt{m}}\sum_{r\in S_i^{\perp}} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\
&+ \frac{1}{\sqrt{m}}\sum_{r\in S_i} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\
=& I_{1,i}(l) + I_{2,i}(l).
\end{aligned} \tag{E.2}$$

The first term $I_{1,i}(l)$ can be bounded by

$$\begin{aligned}
I_{1,i}(l) =& \frac{1}{\sqrt{m}}\sum_{r\in S_i^{\perp}} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\
\leq& \frac{1}{\sqrt{m}}\sum_{r\in S_i^{\perp}} \left|(\boldsymbol{w}_{D,r}(l+1) - (1-\eta_2\mu)\boldsymbol{w}_{D,r}(l))^\top \boldsymbol{x}_i\right| \\
\leq& \frac{1}{\sqrt{m}}\sum_{r\in S_i^{\perp}} \left\|\boldsymbol{w}_{D,r}(l+1) - (1-\eta_2\mu)\boldsymbol{w}_{D,r}(l)\right\|_2 \\
=& \frac{1}{\sqrt{m}}\sum_{r\in S_i^{\perp}} \left\|\frac{\eta_1}{\sqrt{m}}a_r\sum_{j=1}^{n}(u_{D,j}(l)-y_j)\mathbb{I}_{r,j}(l)\boldsymbol{x}_j\right\|_2 \\
\leq& \frac{\eta_1}{m}\sum_{r\in S_i^{\perp}}\sum_{j=1}^{n}|u_{D,j}(l)-y_j| \\
\leq& \frac{\eta_1\sqrt{n}|S_i^{\perp}|}{m}\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2.
\end{aligned} \tag{E.3}$$

In (E.3), the second and the last inequalities are by the Cauchy-Schwarz inequality. The second term $I_{2,i}(l)$ can be bounded by

$$\begin{aligned}
I_{2,i}(l) =& \frac{1}{\sqrt{m}}\sum_{r\in S_i} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\
=& \frac{1}{\sqrt{m}}\sum_{r\in S_i} a_r\mathbb{I}_{r,i}(l)(\boldsymbol{w}_{D,r}(l+1) - (1-\eta_2\mu)\boldsymbol{w}_{D,r}(l))^\top \boldsymbol{x}_i \\
=& -\frac{1}{\sqrt{m}}\sum_{r\in S_i} a_r\mathbb{I}_{r,i}(l)\left(\frac{\eta_1}{\sqrt{m}}a_r\sum_{j=1}^{n}(u_{D,j}(l)-y_j)\mathbb{I}_{r,j}(l)\boldsymbol{x}_j\right)^\top \boldsymbol{x}_i \\
=& -\frac{\eta_1}{m}\sum_{j=1}^{n}(u_{D,j}(l)-y_j)\boldsymbol{x}_j^\top \boldsymbol{x}_i \sum_{r\in S_i}\mathbb{I}_{r,i}(l)\mathbb{I}_{r,j}(l) \\
=& -\eta_1\sum_{j=1}^{n}(u_{D,j}(l)-y_j)\boldsymbol{H}_{ij}(l) + I_{3,i}(l),
\end{aligned} \tag{E.4}$$

where

$$I_{3,i}(l) = \frac{\eta_1}{m}\sum_{j=1}^{n}(u_{D,j}(l) - y_j)\boldsymbol{x}_j^\top \boldsymbol{x}_i \sum_{r\in S_i^\perp} \mathbb{I}_{r,i}(l)\mathbb{I}_{r,j}(l).$$

The term $I_{3,i}(l)$ in (E.4) can be bounded by

$$\begin{aligned}
|I_{3,i}(l)| \leq& \left|\frac{\eta_1}{m}\sum_{j=1}^{n}(u_{D,j}(l) - y_j)\boldsymbol{x}_j^\top \boldsymbol{x}_i \sum_{r\in S_i^\perp} \mathbb{I}_{r,i}(l)\mathbb{I}_{r,j}(l)\right| \\
\leq& \frac{\eta_1}{m}|S_i^\perp|\sum_{j=1}^{n}|u_{D,j}(l) - y_j| \\
\leq& \frac{\eta_1\sqrt{n}|S_i^\perp|}{m}\left\|\boldsymbol{u}_D(l) - \boldsymbol{y}\right\|_2. \qquad \text{(E.5)}
\end{aligned}$$

Plugging (E.3) and (E.4) into (E.2), we have

$$u_{D,i}(l+1) - (1-\eta_2\mu)u_{D,i}(l) = -\eta_1\sum_{j=1}^{n}(u_{D,j}(l) - y_j)\boldsymbol{H}_{ij}(l) + I_{1,i}(l) + I_{3,i}(l),$$

which leads to

$$\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l) = -\eta_1\boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) + \boldsymbol{I}(l), \qquad \text{(E.6)}$$

where $\boldsymbol{I}(l) = (I_{1,1}(l) + I_{3,1}(l), ..., I_{1,n}(l) + I_{3,n}(l))^\top$. By the triangle inequality, we have

$$\left\|\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)\right\|_2 \leq \left\|\eta_1\boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y})\right\|_2 + \left\|\boldsymbol{I}(l)\right\|_2. \qquad \text{(E.7)}$$

By (E.1), (E.3), and (E.5), the term $\left\|\boldsymbol{I}(l)\right\|_2$ in (E.7) can be bounded by

$$\begin{aligned}
\left\|\boldsymbol{I}(l)\right\|_2 \leq& \sum_{i=1}^{n}|I_{3,i}(l)| + |I_{1,i}(l)| \leq \sum_{i=1}^{n}\frac{2\eta_1\sqrt{n}|S_i^\perp|}{m}\left\|\boldsymbol{u}_D(l) - \boldsymbol{y}\right\|_2 \\
\leq& \frac{2\eta_1\sqrt{n}}{m}\frac{CmnR}{\delta(1-\eta_2\mu)^k}\left\|\boldsymbol{u}_D(l) - \boldsymbol{y}\right\|_2 = \frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\left\|\boldsymbol{u}_D(l) - \boldsymbol{y}\right\|_2. \qquad \text{(E.8)}
\end{aligned}$$

Gershgorin's theorem [Varga, 2010] implies

$$\lambda_{\max}(H(l)) \leq \max_j \sum_{i=1}^{n} H_{ij}(l) \leq n.$$

Therefore, the term $\left\|\eta_1\boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y})\right\|_2$ in (E.7) can be bounded by

$$\left\|\eta_1\boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y})\right\|_2 \leq \eta_1\lambda_{\max}(H(l))\left\|\boldsymbol{u}_D(l) - \boldsymbol{y}\right\|_2 \leq \eta_1 n\left\|\boldsymbol{u}_D(l) - \boldsymbol{y}\right\|_2. \qquad \text{(E.9)}$$

By (E.7) and (E.8), $\left\|\boldsymbol{y} - \boldsymbol{u}_D(l+1)\right\|_2$ can be bounded by

$$\begin{aligned}
\left\|\boldsymbol{y} - \boldsymbol{u}_D(l+1)\right\|_2^2 =& \left\|\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l)\right\|_2^2 - 2(\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l))^\top(\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)) \\
&+ \left\|\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)\right\|_2^2 \\
=& \left\|\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l)\right\|_2^2 + 2\eta_1(\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l))^\top\boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) \\
&- 2\eta_1(\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l))^\top\boldsymbol{I}(l) + \left\|\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)\right\|_2^2 \\
=& T_1 + T_2 + T_3 + T_4. \qquad \text{(E.10)}
\end{aligned}$$

The first term T_1 can be bounded by

$$\begin{aligned}
T_1 =& \left\|\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l)\right\|_2^2 \\
=& \eta_2^2\mu^2\left\|\boldsymbol{y}\right\|_2^2 + (1-\eta_2\mu)^2\left\|\boldsymbol{y} - \boldsymbol{u}_D(l)\right\|_2^2 + 2\eta_2\mu(1-\eta_2\mu)\boldsymbol{y}^\top(\boldsymbol{y} - \boldsymbol{u}_D(l)) \\
\leq& (\eta_2^2\mu^2 + \eta_2\mu)\left\|\boldsymbol{y}\right\|_2^2 + (1+\eta_2\mu)(1-\eta_2\mu)^2\left\|\boldsymbol{y} - \boldsymbol{u}_D(l)\right\|_2^2. \qquad \text{(E.11)}
\end{aligned}$$

The second term T_2 can be bounded by

$$\begin{aligned}T_2 =&2\eta_1(\boldsymbol{y}-(1-\eta_2\mu)\boldsymbol{u}_D(l))^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l)-\boldsymbol{y})\\ =&2\eta_1(1-\eta_2\mu)(\boldsymbol{y}-\boldsymbol{u}_D(l))^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l)-\boldsymbol{y})+2\eta_1\eta_2\mu\boldsymbol{y}^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l)-\boldsymbol{y})\\ =&-2\eta_1(1-\eta_2\mu)(\boldsymbol{y}-\boldsymbol{u}_D(l))^\top \boldsymbol{H}(l)(\boldsymbol{y}-\boldsymbol{u}_D(l))+2\eta_1\eta_2\mu\boldsymbol{y}^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l)-\boldsymbol{y})\\ \le&4\eta_1\eta_2\mu n\left\|\boldsymbol{y}\right\|_2^2+4\eta_1\eta_2\mu n\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2. \end{aligned}\tag{E.12}$$

Using (E.8), the third term T_3 can be bounded by

$$\begin{aligned}T_3 =&-2\eta_1(\boldsymbol{y}-(1-\eta_2\mu)\boldsymbol{u}_D(l))^\top \boldsymbol{I}(l)\\ =&-2\eta_1(1-\eta_2\mu)(\boldsymbol{y}-\boldsymbol{u}_D(l))^\top \boldsymbol{I}(l)+2\eta_1\eta_2\mu\boldsymbol{y}^\top \boldsymbol{I}(l)\\ \le&2\eta_1(1-\eta_2\mu)\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2+4\eta_1\eta_2\mu\left\|\boldsymbol{y}\right\|_2^2+4\eta_1\eta_2\mu\left\|\boldsymbol{I}(l)\right\|_2^2\\ \le&2\eta_1(1-\eta_2\mu)\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2+4\eta_1\eta_2\mu\left\|\boldsymbol{y}\right\|_2^2+4\eta_1\eta_2\mu\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2. \end{aligned}\tag{E.13}$$

The fourth term T_4 can be bounded by

$$\begin{aligned}T_4 =&\left\|\boldsymbol{u}_D(l+1)-(1-\eta_2\mu)\boldsymbol{u}_D(l)\right\|_2^2\\ \le&2\left\|\eta_1\boldsymbol{H}(l)(\boldsymbol{u}_D(l)-\boldsymbol{y})\right\|_2^2+2\left\|\boldsymbol{I}(l)\right\|_2^2\\ \le&2\eta_1^2n^2\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2+2\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2. \end{aligned}\tag{E.14}$$

Plugging (E.11) - (E.14) into (E.10), we have

$$\begin{aligned}&\left\|\boldsymbol{y}-\boldsymbol{u}_D(l+1)\right\|_2^2\\ \le&(\eta_2^2\mu^2+\eta_2\mu)\left\|\boldsymbol{y}\right\|_2^2+(1+\eta_2\mu)(1-\eta_2\mu)^2\left\|\boldsymbol{y}-\boldsymbol{u}_D(l)\right\|_2^2+4\eta_1\eta_2\mu n\left\|\boldsymbol{y}\right\|_2^2+4\eta_1\eta_2\mu n\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2\\ &+2\eta_1(1-\eta_2\mu)\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2+4\eta_1\eta_2\mu\left\|\boldsymbol{y}\right\|_2^2+4\eta_1\eta_2\mu\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2\\ &+2\eta_1^2n^2\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2+2\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2\\ =&a_1\left\|\boldsymbol{y}\right\|_2^2+a_2\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2, \end{aligned}\tag{E.15}$$

where

$$\begin{aligned}a_1 =&(\eta_2^2\mu^2+\eta_2\mu)+4\eta_1\eta_2\mu n+4\eta_1\eta_2\mu\le 2\eta_2\mu+8\eta_1\eta_2\mu n,\\ a_2 =&(1+\eta_2\mu)(1-\eta_2\mu)^2+4\eta_1\eta_2\mu n+2\eta_1(1-\eta_2\mu)\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\\ &+4\eta_1\eta_2\mu\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2+2\eta_1^2n^2+2\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2\\ \le&1-\left(\eta_2\mu-4\eta_1\eta_2\mu n-2\eta_1\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}-2\eta_1^2n^2\right)\\ =&1-\nu_0. \end{aligned}$$

By the conditions imposed on η_1,η_2,μ,m, the dominating terms in a_1 and ν_0 are both $\eta_2\mu$. Thus $a_1=o(1/n)$, $\nu_0=o(1/n)$ and $a_1/\nu_0=O(1)$. Using (E.15) iteratively, we have

$$\begin{aligned}\left\|\boldsymbol{y}-\boldsymbol{u}_D(l+1)\right\|_2^2\le&a_1\left\|\boldsymbol{y}\right\|_2^2+a_2\left\|\boldsymbol{u}_D(l)-\boldsymbol{y}\right\|_2^2\\ \le&\ldots\le\sum_{i=0}^{l}(1-\nu_0)^i(a_1\left\|\boldsymbol{y}\right\|_2^2)+(1-\nu_0)^{l+1}\left\|\boldsymbol{y}-\boldsymbol{u}_D(0)\right\|_2^2 \end{aligned}\tag{E.16}$$

$$\le\frac{a_1\left\|\boldsymbol{y}\right\|_2^2}{\nu_0}+(1-\nu_0)^{l+1}\left\|\boldsymbol{y}-\boldsymbol{u}_D(0)\right\|_2^2.\tag{E.17}$$

By the modified GD rule, we have

$$\boldsymbol{w}_{D,r}(l+1)-(1-\eta_2\mu)\boldsymbol{w}_{D,r}(l)=-\frac{\eta_1}{\sqrt{m}}a_r\sum_{j=1}^{n}(u_{D,j}(l)-y_j)\mathbb{I}_{r,j}(l)\boldsymbol{x}_j,$$

which implies

$$\|\boldsymbol{w}_{D,r}(l+1)-(1-\eta_2\mu)\boldsymbol{w}_{D,r}(l)\|_2\leq\frac{\eta_1\sqrt{n}}{\sqrt{m}}\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2\leq\frac{C\eta_1 n}{\sqrt{m}} \tag{E.18}$$

for some constant C. Using (E.18) iteratively yields

$$\begin{aligned}
&\left\|\boldsymbol{w}_{D,r}(l+1)-(1-\eta_2\mu)^{l+1}\boldsymbol{w}_{D,r}(0)\right\|_2\\
\leq&\|\boldsymbol{w}_{D,r}(l+1)-(1-\eta_2\mu)\boldsymbol{w}_{D,r}(l)\|_2+\left\|(1-\eta_2\mu)\boldsymbol{w}_{D,r}(0)-(1-\eta_2\mu)^{l+1}\boldsymbol{w}_{D,r}(l)\right\|_2\\
\leq&\frac{C\eta_1 n}{\sqrt{m}}+(1-\eta_2\mu)\left\|\boldsymbol{w}_{D,r}(l)-(1-\eta_2\mu)^{l}\boldsymbol{w}_{D,r}(0)\right\|_2\\
\leq&\ldots\leq\sum_{i=0}^{l}(1-\eta_2\mu)^i\frac{C\eta_1 n}{\sqrt{m}}\leq\frac{C\eta_1 n}{\eta_2\mu\sqrt{m}}.
\end{aligned} \tag{E.19}$$

By similar approach as in the proof of Lemma C.2 of Du et al. [2018], we can show that with probability at least $1-\delta$ with respect to random initialization,

$$\|\boldsymbol{Z}(l)-\boldsymbol{Z}(0)\|_F^2\leq\frac{2nR}{\sqrt{2\pi}\tau\delta(1-\eta_2\mu)^k}+\frac{n}{m}=O\left(\frac{\eta_1 n^2}{(1-\eta_2\mu)^k\eta_2\mu\sqrt{m}\delta^{3/2}\tau}\right),\forall l\in[k],$$

and

$$\|\boldsymbol{H}(l)-\boldsymbol{H}(0)\|_F\leq\frac{4n^2R}{\sqrt{2\pi}\tau}+\frac{2n^2\delta}{m}=O\left(\frac{\eta_1 n^3}{(1-\eta_2\mu)^k\eta_2\mu\sqrt{m}\delta^{3/2}\tau}\right),\forall l\in[k].$$

By Lemma C.3 of Du et al. [2018], we have with probability at least $1-\delta$ with respect to random initialization,

$$\|\boldsymbol{H}(0)-\boldsymbol{H}^{\infty}\|_F=O\left(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}\right). \tag{E.20}$$

By (E.6), we have

$$\begin{aligned}
\boldsymbol{u}_D(l+1)-(1-\eta_2\mu)\boldsymbol{u}_D(l)=&-\eta_1\boldsymbol{H}(l)(\boldsymbol{u}_D(l)-\boldsymbol{y})+\boldsymbol{I}(l)\\
=&-\eta_1\boldsymbol{H}^{\infty}(\boldsymbol{u}_D(l)-\boldsymbol{y})+\boldsymbol{I}(l)-\eta_1(\boldsymbol{H}(l)-\boldsymbol{H}^{\infty})(\boldsymbol{u}_D(l)-\boldsymbol{y}),
\end{aligned}$$

which yields

$$\boldsymbol{u}_D(l+1)-B=\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^{\infty}\right)(\boldsymbol{u}_D(l)-B)+\boldsymbol{I}(l)-\eta_1(\boldsymbol{H}(l)-\boldsymbol{H}^{\infty})(\boldsymbol{u}_D(l)-\boldsymbol{y}), \tag{E.21}$$

where

$$B=(\eta_2\mu I+\eta_1\boldsymbol{H}^{\infty})^{-1}\eta_1\boldsymbol{H}^{\infty}\boldsymbol{y}=\eta_1\boldsymbol{H}^{\infty}(\eta_2\mu I+\eta_1\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}. \tag{E.22}$$

Iteratively using (E.21), we have

$$\begin{aligned}
\boldsymbol{u}_D(l+1)-B=&\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^{\infty}\right)^{l+1}(\boldsymbol{u}_D(0)-B)\\
&+\sum_{i=0}^{l}\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^{\infty}\right)^{i}(\boldsymbol{I}(l-i)-\eta_1(\boldsymbol{H}(l-i)-\boldsymbol{H}^{\infty})(\boldsymbol{u}_D(l-i)-\boldsymbol{y}))\\
=&\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^{\infty}\right)^{l+1}(\boldsymbol{u}_D(0)-B)+e_l,
\end{aligned} \tag{E.23}$$

where

$$e_l = \sum_{i=0}^{l} \left((1-\eta_2\mu)I - \eta_1 \boldsymbol{H}^{\infty}\right)^i \left(\boldsymbol{I}(l-i) - \eta_1(\boldsymbol{H}(l-i) - \boldsymbol{H}^{\infty})(\boldsymbol{u}_D(l-i) - \boldsymbol{y})\right). \tag{E.24}$$

The term e_l can be bounded by

$$\begin{aligned}
\|e_l\|_2 =& \left\| \sum_{i=0}^{l} \left((1-\eta_2\mu)I - \eta_1 \boldsymbol{H}^{\infty}\right)^i \left(\boldsymbol{I}(l-i) - \eta_1(\boldsymbol{H}(l-i) - \boldsymbol{H}^{\infty})(\boldsymbol{u}_D(l-i) - \boldsymbol{y})\right) \right\|_2 \\
\leq& \sum_{i=0}^{l} \|(1-\eta_2\mu)I - \eta_1 \boldsymbol{H}^{\infty}\|_2^i \left(\|\boldsymbol{I}(l-i)\|_2 + \eta_1 \|\boldsymbol{H}(l-i) - \boldsymbol{H}^{\infty}\|_2 \|\boldsymbol{u}_D(l-i) - \boldsymbol{y}\|_2\right) \\
\leq& \sum_{i=0}^{l} (1-\eta_2\mu)^i O\bigg(\frac{2C\eta_1^2 n^{5/2}}{\eta_2\mu\sqrt{m}\delta^{3/2}(1-\eta_2\mu)^k} + \frac{\eta_1^2 n^{7/2}}{(1-\eta_2\mu)^k \eta_2\mu\sqrt{m}\delta^2\tau}\bigg) \\
=& O\bigg(\frac{\eta_1^2 n^{7/2}}{\eta_2^2\mu^2\sqrt{m}\delta^2(1-\eta_2\mu)^k\tau}\bigg).
\end{aligned} \tag{E.25}$$

By (E.23) and taking $l = k-1$, with probability at least $1-\delta$ with respect to the random initialization, the difference $\boldsymbol{u}_D(k) - B$ can be bounded by

$$\begin{aligned}
\|\boldsymbol{u}_D(k) - B\|_2 \leq& \left\| \left((1-\eta_2\mu)I - \eta_1 \boldsymbol{H}^{\infty}\right)^k (\boldsymbol{u}_D(0) - B) \right\|_2 + \|e_k\|_2 \\
=& O\left(\sqrt{n}(1-\eta_2\mu - \eta_1\lambda_0)^k + \frac{n^{7/2}}{\mu^2\sqrt{m}\delta^2(1-\eta_2\mu)^k\tau}\right) \\
=& O\left(\sqrt{n}(1-\eta_2\mu)^k + \frac{n^{7/2}}{\mu^2\sqrt{m}\delta^2(1-\eta_2\mu)^k\tau}\right).
\end{aligned}$$

This implies that

$$\|\boldsymbol{u}_D(k) - B\|_2 = O_{\mathbb{P}}\left(\sqrt{n}(1-\eta_2\mu)^k + \frac{n^{7/2}}{\mu^2\sqrt{m}(1-\eta_2\mu)^k\tau}\right).$$

By choosing $m = \mathrm{poly}(n, 1/\tau, 1/\lambda_0)$ such that $\frac{n^{7/2}}{\mu^2\sqrt{m}(1-\eta_2\mu)^k\tau} \leq \sqrt{n}(1-\eta_2\mu)^k$, we finish the proof of (5.3).

Now consider $\mathrm{vec}(\boldsymbol{W}_D(l+1))$. Direct calculation shows that

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{W}_D(l+1)) =& (1-\eta_2\mu)\mathrm{vec}(\boldsymbol{W}_D(l)) - \eta_1 \boldsymbol{Z}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) \\
=& (1-\eta_2\mu)\mathrm{vec}(\boldsymbol{W}_D(l)) - \eta_1 \boldsymbol{Z}(0)(\boldsymbol{u}_D(l) - \boldsymbol{y}) - \eta_1(\boldsymbol{Z}(l) - \boldsymbol{Z}(0))(\boldsymbol{u}_D(l) - \boldsymbol{y}) \\
=& (1-\eta_2\mu)^{l+1}\mathrm{vec}(\boldsymbol{W}_D(0)) - \eta_1 \boldsymbol{Z}(0) \sum_{i=0}^{l} (1-\eta_2\mu)^i (\boldsymbol{u}_D(l-i) - \boldsymbol{y}) \\
& - \sum_{i=0}^{l} (1-\eta_2\mu)^i \eta_1 (\boldsymbol{Z}(l) - \boldsymbol{Z}(0))(\boldsymbol{u}_D(l) - \boldsymbol{y}).
\end{aligned} \tag{E.26}$$

Plugging

$$\boldsymbol{u}_D(l+1) = \left((1-\eta_2\mu)I - \eta_1 \boldsymbol{H}^{\infty}\right)^{l+1} (\boldsymbol{u}_D(0) - B) + e_l + B$$

into (E.26), we have

$$\begin{aligned}
&\operatorname{vec}(\boldsymbol{W}_D(l+1)) - (1-\eta_2\mu)^{l+1}\operatorname{vec}(\boldsymbol{W}_D(0)) \\
=&-\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i}(\boldsymbol{u}_D(0)-B) \\
&-\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i(e_{l-i-1}+B-\boldsymbol{y}) - \sum_{i=0}^{l}(1-\eta_2\mu)^i\eta_1(\boldsymbol{Z}(l)-\boldsymbol{Z}(0))(\boldsymbol{u}_D(l)-\boldsymbol{y}) \\
=&\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i}\eta_1\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y} \\
&-\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i}\boldsymbol{u}_D(0) \\
&-\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i e_{l-i-1} - \eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i(B-\boldsymbol{y}) \\
&-\sum_{i=0}^{l}(1-\eta_2\mu)^i\eta_1(\boldsymbol{Z}(l)-\boldsymbol{Z}(0))(\boldsymbol{u}_D(l)-\boldsymbol{y}) \\
=&E_1-E_2+E_3-T_5-E_4. \qquad \text{(E.27)}
\end{aligned}$$

Let

$$\begin{aligned}
\boldsymbol{T}_l =&\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i} \\
=&(1-\eta_2\mu)^l\sum_{i=0}^{l}\left(I-\frac{\eta_1}{(1-\eta_2\mu)}\boldsymbol{H}^\infty\right)^i \qquad \text{(E.28)}
\end{aligned}$$

and

$$\boldsymbol{a}_1 =\eta_1\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}. \qquad \text{(E.29)}$$

The first term E_1 can be bounded by

$$\begin{aligned}
\|E_1\|_2^2 =&\|\eta_1\boldsymbol{Z}(0)\boldsymbol{T}_l\boldsymbol{a}_1\|_2^2 \\
=&\eta_1^2\boldsymbol{a}_1^\top\boldsymbol{T}_l\boldsymbol{Z}(0)^\top\boldsymbol{Z}(0)\boldsymbol{T}_l\boldsymbol{a}_1 \\
=&\eta_1^2\boldsymbol{a}_1^\top\boldsymbol{T}_l\boldsymbol{H}^\infty\boldsymbol{T}_l\boldsymbol{a}_1+\eta_1^2\boldsymbol{a}_1^\top\boldsymbol{T}_l(\boldsymbol{H}(0)-\boldsymbol{H}^\infty)\boldsymbol{T}_l\boldsymbol{a}_1 \\
=&\eta_1^2\boldsymbol{a}_1^\top\boldsymbol{T}_l\boldsymbol{H}^\infty\boldsymbol{T}_l\boldsymbol{a}_1+\eta_1^2O\left(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}\right)\boldsymbol{a}_1^\top\boldsymbol{T}_l^2\boldsymbol{a}_1. \qquad \text{(E.30)}
\end{aligned}$$

By (E.28), we have

$$\boldsymbol{T}_l =(1-\eta_2\mu)^l\sum_{j=1}^{n}\frac{1-(1-\frac{\eta_1}{(1-\eta_2\mu)}\lambda_j)^{l+1}}{\frac{\eta_1}{(1-\eta_2\mu)}\lambda_j}\boldsymbol{v}_j\boldsymbol{v}_j^\top \preceq \frac{(1-\eta_2\mu)^l}{\eta_1\lambda_0}\boldsymbol{I},$$

and

$$\boldsymbol{T}_l\boldsymbol{H}^\infty\boldsymbol{T}_l =(1-\eta_2\mu)^{2l}\sum_{j=1}^{n}\left(\frac{1-(1-\frac{\eta_1}{(1-\eta_2\mu)}\lambda_j)^{2l+2}}{\frac{\eta_1}{(1-\eta_2\mu)}\lambda_j}\right)^2\lambda_j\boldsymbol{v}_j\boldsymbol{v}_j^\top \preceq \frac{(1-\eta_2\mu)^{l+1}}{\eta_1^2}(\boldsymbol{H}^\infty)^{-1}.$$

Therefore,

$$\begin{aligned}
\eta_1^2\boldsymbol{a}_1^\top\boldsymbol{T}_l\boldsymbol{H}^\infty\boldsymbol{T}_l\boldsymbol{a}_1 \leq&(1-\eta_2\mu)^{2l+2}\boldsymbol{a}_1^\top(\boldsymbol{H}^\infty)^{-1}\boldsymbol{a}_1, \\
\eta_1^2O\left(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}\right)\boldsymbol{a}_1^\top\boldsymbol{T}_l^2\boldsymbol{a}_1 \leq&O\left(\frac{n^2(1-\eta_2\mu)^{2l}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right).
\end{aligned}$$

Together with (E.30), we have

$$\|E_1\|_2^2 = (1-\eta_2\mu)^{2l+2}\boldsymbol{a}_1^\top(\boldsymbol{H}^\infty)^{-1}\boldsymbol{a}_1 + O\left(\frac{n^2(1-\eta_2\mu)^{2l}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right). \tag{E.31}$$

By similar approach, the second term E_2 can be bounded by

$$\begin{aligned}
\|E_2\|_2^2 =& \left\|\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i}\boldsymbol{u}_D(0)\right\|_2^2\\
=&\eta_1^2\boldsymbol{u}_D(0)^\top\boldsymbol{T}_1(l)\boldsymbol{Z}(0)^\top\boldsymbol{Z}(0)\boldsymbol{T}_1(l)\boldsymbol{u}_D(0)\\
=&\eta_1^2\boldsymbol{u}_D(0)^\top\boldsymbol{T}_1(l)\boldsymbol{H}^\infty\boldsymbol{T}_1(l)\boldsymbol{u}_D(0)+\eta_1^2\boldsymbol{u}_D(0)^\top\boldsymbol{T}_1(l)(\boldsymbol{H}(0)-\boldsymbol{H}^\infty)\boldsymbol{T}_1(l)\boldsymbol{u}_D(0)\\
=&(1-\eta_2\mu)^{2l+2}\boldsymbol{u}_D(0)^\top(\boldsymbol{H}^\infty)^{-1}\boldsymbol{u}_D(0)+O\left(\frac{n^2(1-\eta_2\mu)^{2l}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right).
\end{aligned} \tag{E.32}$$

By (E.25), the third term E_3 can be bounded by

$$\begin{aligned}
\|E_3\|_2^2 =& \left\|\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i e_{l-i-1}\right\|_2^2\\
=&\eta_1^2\left(\sum_{i=0}^{l}(1-\eta_2\mu)^i e_{l-i-1}\right)^\top\boldsymbol{H}(0)\left(\sum_{i=0}^{l}(1-\eta_2\mu)^i e_{l-i-1}\right)\\
=&O\left(\frac{\eta_1^6 n^8}{\eta_2^6\mu^6 m\delta^4(1-\eta_2\mu)^{2k}\tau^2}\right).
\end{aligned} \tag{E.33}$$

The fourth term E_4 can be bounded by

$$\begin{aligned}
\|E_4\|_2^2 =& \left\|\sum_{i=0}^{l}(1-\eta_2\mu)^i\eta_1(\boldsymbol{Z}(l)-\boldsymbol{Z}(0))(\boldsymbol{u}_D(l)-\boldsymbol{y})\right\|_2^2\\
=&O\left(\frac{\eta_1^3 n^3}{(1-\eta_2\mu)^k\eta_2^3\mu^3\sqrt{m}\delta^{3/2}\tau}\right).
\end{aligned} \tag{E.34}$$

Note that

$$\begin{aligned}
B-\boldsymbol{y} =&\eta_1\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}-\boldsymbol{y}\\
=&(\eta_1\boldsymbol{H}^\infty-\eta_2\mu I-\eta_1\boldsymbol{H}^\infty)(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}\\
=&-\eta_2\mu(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}.
\end{aligned}$$

Therefore, the remaining term T_5 can be bounded by

$$\begin{aligned}
\|T_5\|_2^2 =& \left\|\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i(B-\boldsymbol{y})\right\|_2^2\\
\leq&\eta_1^2\boldsymbol{y}^\top(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}\\
\leq&\boldsymbol{y}^\top(\eta_2\mu/\eta_1 I+\boldsymbol{H}^\infty)^{-1}\boldsymbol{H}^\infty(\eta_2\mu/\eta_1 I+\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}.
\end{aligned}$$

By the assumption that $\eta_2 \asymp \eta_1$, the term T_5 can be further bounded by

$$\|T_5\|_2^2 \leq \boldsymbol{y}^\top(C\mu I+\boldsymbol{H}^\infty)^{-1}\boldsymbol{H}^\infty(C\mu I+\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}. \tag{E.35}$$

The right-hand side of (E.35) is $\left\|\widehat{f}\right\|_{\mathcal{N}}^2$, where $\widehat{f}$ is defined in (3.4). The term $\left\|\widehat{f}\right\|_{\mathcal{N}}^2$ can be bounded by some constant as in Theorem 3.2. This also implies

$$\boldsymbol{a}_1^\top(\boldsymbol{H}^\infty)^{-1}\boldsymbol{a}_1 = \eta_1^2\boldsymbol{y}^\top(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y} = O(1). \tag{E.36}$$

Note also that

$$\boldsymbol{u}_D(0)^\top(\boldsymbol{H}^\infty)^{-1}\boldsymbol{u}_D(0) = O\left(\frac{n\tau^2}{\lambda_0}\right). \tag{E.37}$$

By the assumptions of Theorem 5.1, plugging (E.30)-(E.37) into (E.27), and taking the iteration number at k, we can conclude that

$$\begin{aligned}
&\left\|\mathrm{vec}(\boldsymbol{W}_D(k)) - (1-\eta_2\mu)^k\mathrm{vec}(\boldsymbol{W}_D(0))\right\|_2^2\\
=&O((1-\eta_2\mu)^{2k}) + O\left(\frac{n^2(1-\eta_2\mu)^{2k-2}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right)\\
&+ O\left(\frac{n\tau^2}{\lambda_0}(1-\eta_2\mu)^{2k}\right) + O\left(\frac{n^2(1-\eta_2\mu)^{2k-2}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right)\\
&+ O\left(\frac{n^8}{\mu^6 m\delta^4(1-\eta_2\mu)^{2k}\tau^2}\right) + O\left(\frac{n^3}{(1-\eta_2\mu)^k\mu^3\sqrt{m}\delta^{3/2}\tau}\right) + O(1)\\
=&O(1),
\end{aligned} \tag{E.38}$$

where the last equality is because we can select some polynomials such that all the terms in (E.38) except the $O(1)$ term converge to zero, and $\exp(-2\eta_2\mu k) \le (1-\eta_2\mu)^k \le \exp(-\eta_2\mu k)$ for sufficiently large n. This finishes the proof of (5.4) in Theorem 5.1.

E.2 Proof of Theorem 5.2

For notational simplification, we use $\widehat{f}_k = f_{\boldsymbol{W}(k),\boldsymbol{a}}$. Similar to the proof of Theorem 4.1, we define

$$\widetilde{f}_k(\boldsymbol{x}) = \mathrm{vec}(\boldsymbol{W}_D(k))^\top\boldsymbol{z}_0(\boldsymbol{x}), \tag{E.39}$$

where $\boldsymbol{z}_0(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x})|_{\boldsymbol{W}_D=\boldsymbol{W}_D(0)}$. Then we can write the following decomposition

$$\begin{aligned}
\widehat{f}_k(\boldsymbol{x}) - f^*(\boldsymbol{x}) =&(\widehat{f}_k(\boldsymbol{x}) - \widetilde{f}_k(\boldsymbol{x})) + (\widetilde{f}_k(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x})) + (\widehat{f}(\boldsymbol{x}) - f^*(\boldsymbol{x}))\\
=&\Delta_1(\boldsymbol{x}) + \Delta_2(\boldsymbol{x}) + \Delta_3(\boldsymbol{x}),
\end{aligned} \tag{E.40}$$

where $\widehat{f}$ is as in (3.4). In the rest of the proof, we show $\Delta_1(\boldsymbol{x})$, $\Delta_2(\boldsymbol{x})$, and $\Delta_3(\boldsymbol{x})$ are all small.

It follows from Theorem 3.2 that

$$\|\Delta_3\|_2^2 = O_{\mathbb{P}}\left(n^{-\frac{d}{2d-1}}\right). \tag{E.41}$$

Next, we consider Δ_1. From (E.19), it can be seen that

$$\left\|\boldsymbol{w}_{D,r}(k) - (1-\eta_2\mu)^k\boldsymbol{w}_{D,r}(0)\right\|_2 \le \frac{C\eta_1 n}{\eta_2\mu\sqrt{m}}. \tag{E.42}$$

Define event

$$B_{D,r}(\boldsymbol{x}) = \{|(1-\eta_2\mu)^k\boldsymbol{w}_{D,r}(0)^\top\boldsymbol{x}| \le R_1\}, \forall r \in [m],$$

where $R_1 = \frac{C\eta_1 n}{\eta_2\mu\sqrt{m}}$. If $\mathbb{I}\{B_{D,r}(\boldsymbol{x})\} = 0$, then we have $\mathbb{I}_{r,k}(\boldsymbol{x}) = \mathbb{I}_{r,0}(\boldsymbol{x})$, where $\mathbb{I}_{r,k}(\boldsymbol{x}) = \mathbb{I}\{\boldsymbol{w}_{D,r}(k)^\top\boldsymbol{x} \ge 0\}$.

Therefore, for any fixed $\boldsymbol{x}$,

$$\begin{aligned}
|\Delta_1(\boldsymbol{x})| &= |\widehat{f}_k(\boldsymbol{x}) - \widetilde{f}_k(\boldsymbol{x})| \\
&= \left| \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r (\mathbb{I}_{r,k}(\boldsymbol{x}) - \mathbb{I}_{r,0}(\boldsymbol{x})) \boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x} \right| \\
&= \left| \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathbb{I}\{B_{D,r}(\boldsymbol{x})\} (\mathbb{I}_{r,k}(\boldsymbol{x}) - \mathbb{I}_{r,0}(\boldsymbol{x})) \boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x} \right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}\{B_{D,r}(\boldsymbol{x})\} |\boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x}| \\
&\leq \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}\{B_{D,r}(\boldsymbol{x})\} \left(|(1-\eta_2\mu)^k \boldsymbol{w}_{D,r}(0)^\top \boldsymbol{x}| + |\boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x} - (1-\eta_2\mu)^k \boldsymbol{w}_r(0)^\top \boldsymbol{x}| \right) \\
&\leq \frac{2R_1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}\{B_{D,r}(x)\}.
\end{aligned}$$

Note that $\|\boldsymbol{x}\|_2 = 1$, which implies that $\boldsymbol{w}_{D,r}(0)^\top \boldsymbol{x}$ is distributed as $N(0, \tau^2)$. Therefore, we have

$$\begin{aligned}
\mathbb{E}[\mathbb{I}\{B_{D,r}(x)\}] &= \mathbb{P}\left(|(1-\eta_2\mu)^k \boldsymbol{w}_{D,r}(0)^\top \boldsymbol{x}| \leq R_1 \right) \\
&= \int_{-R_1/(1-\eta_2\mu)^k}^{R_1/(1-\eta_2\mu)^k} \frac{1}{\sqrt{2\pi}\tau} \exp\left\{ -\frac{u^2}{2\tau^2} \right\} du \leq \frac{2R_1}{\sqrt{2\pi}(1-\eta_2\mu)^k \tau}.
\end{aligned}$$

By Markov's inequality, with probability at least $1-\delta$, we have

$$\sum_{r=1}^{m} \mathbb{I}\{B_{D,r}(x)\} \leq \frac{2mR_1}{\sqrt{2\pi}(1-\eta_2\mu)^k \tau \delta}.$$

Thus, we have with probability at least $1-\delta$,

$$\|\Delta_1\|_2 \leq \frac{2R_1}{\sqrt{m}} \left\| \sum_{r=1}^{m} \mathbb{I}\{B_{D,r}(\cdot)\} \right\|_2 \leq \frac{4\sqrt{m}R_1^2}{\sqrt{2\pi}(1-\eta_2\mu)^k \tau \delta} = O\left(\frac{n^2}{\sqrt{m}\lambda_0^2 \delta^2 (1-\eta_2\mu)^k \tau} \right),$$

which implies

$$\|\Delta_1\|_2 = O_{\mathbb{P}}\left(\frac{n^2}{\sqrt{m}\lambda_0^2 (1-\eta_2\mu)^k \tau} \right). \tag{E.43}$$

Now we bound Δ_2. Note that Define $\boldsymbol{G}_k = \sum_{j=0}^{k-1} \eta(\boldsymbol{I} - \eta \boldsymbol{H}^\infty)^j$. Recalling that $\boldsymbol{y} = \boldsymbol{y}^* + \boldsymbol{\epsilon}$, for fixed $\boldsymbol{x}$, we have

$$\begin{aligned}
\Delta_2(\boldsymbol{x}) =& \widetilde{f}_k(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x}) \\
=& \boldsymbol{z}_0(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{W}_D(k)) - h(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1} \boldsymbol{y} \\
=& \boldsymbol{z}_0(\boldsymbol{x})^\top E_1 - \boldsymbol{z}_0(\boldsymbol{x})^\top E_2 + \boldsymbol{z}_0(\boldsymbol{x})^\top E_3 - \boldsymbol{z}_0(\boldsymbol{x})^\top T_5 - \boldsymbol{z}_0(\boldsymbol{x})^\top E_4 \\
&+ (1-\eta_2\mu)^k \boldsymbol{z}_0(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{W}_D(0)) - h(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1} \boldsymbol{y},
\end{aligned} \tag{E.44}$$

where E_1, E_2, E_3, T_5, E_4 are as in (E.27). Noting that $\|z_0(\boldsymbol{x})\|_2 = O_{\mathbb{P}}(1)$, we have that

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top E_1|^2 \leq \|\boldsymbol{z}_0(\boldsymbol{x})\|_2^2 \|E_1\|_2^2 = O_{\mathbb{P}}((1-\eta_2\mu)^{2k}) + O_{\mathbb{P}}\left(\frac{n^2(1-\eta_2\mu)^{2k-2}\sqrt{\log(n)}}{\sqrt{m}\lambda_0^2} \right), \tag{E.45}$$

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top E_2|^2 \leq \|\boldsymbol{z}_0(\boldsymbol{x})\|_2^2 \|E_2\|_2^2 = O_{\mathbb{P}}\left(\frac{n\tau^2}{\lambda_0}(1-\eta_2\mu)^{2k} \right) + O_{\mathbb{P}}\left(\frac{n^2(1-\eta_2\mu)^{2k-2}\sqrt{\log(n)}}{\sqrt{m}\lambda_0^2} \right), \tag{E.46}$$

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top E_3|^2 \leq \|\boldsymbol{z}_0(\boldsymbol{x})\|_2^2 \|E_3\|_2^2 = O_{\mathbb{P}}\left(\frac{\eta_1^6 n^8}{\eta_2^6 \mu^6 m (1-\eta_2\mu)^{2k} \tau^2} \right), \tag{E.47}$$

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top E_4|^2 \leq \|\boldsymbol{z}_0(\boldsymbol{x})\|_2^2 \|E_4\|_2^2 = O_{\mathbb{P}}\left(\frac{n^3}{(1-\eta_2\mu)^k \mu^3 \sqrt{m} \delta^{3/2} \tau} \right), \tag{E.48}$$

where (E.45) is because of (E.31) and (E.36), (E.46) is because of (E.32) and (E.37), (E.47) is because of (E.33), and (E.48) is because of (E.34). By Lemma D.5 (d), the term $(1-\eta_2\mu)^k \boldsymbol{z}_0(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{W}_D(0))$ in (E.44) can be bounded by

$$\left\|(1-\eta_2\mu)^k \boldsymbol{z}_0(\cdot)^\top \mathrm{vec}(\boldsymbol{W}_D(0))\right\|_2 = O_{\mathbb{P}}((1-\eta_2\mu)^k\tau). \tag{E.49}$$

Define

$$B = \eta_1 \boldsymbol{H}^\infty (\eta_2\mu I + \eta_1 \boldsymbol{H}^\infty)^{-1}\boldsymbol{y}.$$

Note that

$$\begin{aligned} B-\boldsymbol{y} =& \eta_1 \boldsymbol{H}^\infty (\eta_2\mu I + \eta_1 \boldsymbol{H}^\infty)^{-1}\boldsymbol{y} - \boldsymbol{y} \\ =& (\eta_1 \boldsymbol{H}^\infty - \eta_2\mu I - \eta_1 \boldsymbol{H}^\infty)(\eta_2\mu I + \eta_1 \boldsymbol{H}^\infty)^{-1}\boldsymbol{y} \\ =& -\eta_2\mu(\eta_2\mu I + \eta_1 \boldsymbol{H}^\infty)^{-1}\boldsymbol{y}. \end{aligned}$$

Therefore, the remaining term in (E.44) $-\boldsymbol{z}_0(\boldsymbol{x})^\top T_5 - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}$ can be bounded by

$$\begin{aligned} & -\boldsymbol{z}_0(\boldsymbol{x})^\top T_5 - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y} \\ =& -\boldsymbol{z}_0(\boldsymbol{x})^\top \boldsymbol{Z}(0) \sum_{i=0}^{k-1} \eta_1(1-\eta_2\mu)^i (B-\boldsymbol{y}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y} \\ =& -\boldsymbol{z}_0(\boldsymbol{x})^\top \boldsymbol{Z}(0)\eta_1 \frac{1-(1-\eta_2\mu)^k}{\eta_2\mu}(B-\boldsymbol{y}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y} \\ =& \boldsymbol{z}_0(\boldsymbol{x})^\top \boldsymbol{Z}(0)\eta_1(1-(1-\eta_2\mu)^k)(\eta_2\mu I + \eta_1 \boldsymbol{H}^\infty)^{-1}\boldsymbol{y} - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y} \\ =& (\boldsymbol{z}_0(\boldsymbol{x})^\top \boldsymbol{Z}(0) - h(\boldsymbol{x},\boldsymbol{X}))(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y} - \eta_1(1-\eta_2\mu)^k \boldsymbol{z}_0(\boldsymbol{x})^\top \boldsymbol{Z}(0)(\eta_2\mu I + \eta_1 \boldsymbol{H}^\infty)^{-1}\boldsymbol{y}. \end{aligned} \tag{E.50}$$

The first term in (E.50) can be bounded by

$$\begin{aligned} & \left\|(\boldsymbol{z}_0(\cdot)^\top \boldsymbol{Z}(0) - h(\cdot,\boldsymbol{X}))(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2 \\ \le& \left\|(\boldsymbol{z}_0(\cdot)^\top \boldsymbol{Z}(0) - h(\cdot,\boldsymbol{X}))\right\|_2 \left\|(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2 \\ =& O_{\mathbb{P}}\left(\frac{n\sqrt{\log(n)}\eta_1}{\sqrt{m}\eta_2\mu}\right), \end{aligned} \tag{E.51}$$

where we utilize

$$\left\|(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2^2 = \boldsymbol{y}^\top(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-2}\boldsymbol{y} \le \frac{\eta_1^2}{\eta_2^2\mu^2}\|\boldsymbol{y}\|_2^2 = O_{\mathbb{P}}\left(\frac{\eta_1^2}{\eta_2^2\mu^2}n\right),$$

and Lemma D.5 (c).

The second term in (E.50) can be bounded by

$$\begin{aligned} & \left\|(1-\eta_2\mu)^k \boldsymbol{z}_0(\cdot)^\top \boldsymbol{Z}(0)(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2 \\ \le& (1-\eta_2\mu)^k \left\|(\boldsymbol{z}_0(\cdot)^\top \boldsymbol{Z}(0) - h(\cdot,\boldsymbol{X}))(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2 \\ & + (1-\eta_2\mu)^k \left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2 \\ \le& O_{\mathbb{P}}\left(\frac{n\sqrt{\log(n)}\eta_1}{\sqrt{m}\eta_2\mu}\right) + (1-\eta_2\mu)^k \left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_{\mathcal{N}} \\ =& O_{\mathbb{P}}((1-\eta_2\mu)^k), \end{aligned} \tag{E.52}$$

where the second inequality is because of (E.51) and the last equality is because of Theorem 3.2 and the assumption $\eta_1 \asymp \eta_2$. Plugging (E.45)-(E.52) to (E.44), we can conclude that

$$\|\Delta_2\|_2 = o_{\mathbb{P}}(n^{-\frac{d}{2d-1}}), \tag{E.53}$$

by choosing k and m as in Theorem 5.2. Combining (E.43), (E.53), and (E.41) finishes the proof.

F Proof of lemmas in the Appendix

F.1 Proof of Lemma B.1

The proof of Lemma B.1 mainly from Appendix C of Bietti and Mairal [2019] and Appendix D of Bach [2017], with some modification.

We first review some background of spherical harmonic analysis [Atkinson and Han, 2012, Costas and Christopher, 2014]. Let $Y_{k,j}$ be the spherical harmonics of degree k on $\mathcal{S}^{d-1}$, where $N(p.k) = \frac{2k+d-2}{k}\begin{pmatrix} k+d-3 \\ d-2 \end{pmatrix}$. Then $Y_{k,j}$ is an orthonormal basis of $L_2(\mathcal{S}^{p-1}, d\xi)$, where $d\xi$ is the uniform measure on the sphere. Then we have

$$\sum_{j=1}^{N(d,k)} Y_{k,j}(\boldsymbol{s})Y_{k,j}(\boldsymbol{t}) = N(d,k)P_k(\boldsymbol{s}^\top \boldsymbol{t}), \tag{F.1}$$

where P_k is the k-th Legendre polynomial in dimension d, given by

$$P_k(t) =(-1/2)^k \frac{\Gamma(\frac{d-1}{2})}{\Gamma(k+\frac{d-1}{2})}(1-t^2)^{(3-d)/2}\left(\frac{d}{dt}\right)^k (1-t^2)^{k+(d-3)/2}. \tag{F.2}$$

The polynomials P_k are orthogonal in $L_2([-1,1])d\nu$, where the measure $d\nu = (1-t^2)^{(d-3)/2}dt$ with Lebesgue measure dt, and

$$\int_{[-1,1]} P_k^2(t)(1-t^2)^{(d-3)/2}dt = \frac{w_{d-1}}{w_{d-2}}\frac{1}{N(d,k)}, \tag{F.3}$$

where $w_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$. Furthermore, it can be shown that [Atkinson and Han, 2012]

$$tP_k(t) = \frac{k}{2k+d-2}P_{k-1}(t) + \frac{k+d-2}{2k+d-2}P_{k+1}(t), \tag{F.4}$$

for $k \geq 1$, and for $j = 0$ we have $tP_0(t) = P_1(t)$. This implies that for large k enough, we have

$$\mu_k = \frac{k}{2k+d-2}\mu_{0,k-1} + \frac{k+d-2}{2k+d-2}\mu_{0,k+1},$$

where $\mu_{0,k-1}$ and $\mu_{0,k+1}$ are as in Lemma 17 of Bietti and Mairal [2019]. By Lemma 17 of Bietti and Mairal [2019], we have $\mu_{0,k} \asymp k^{-d}$ for large k, if $k = 1 \bmod 2$. This finish the proof of Lemma B.1.

F.2 Proof of Lemma C.1

By Theorem 1 of Brauchart and Dick [2013] and Lemma B.1, we can see that the function space $\mathcal{N}$ is a subspace of the Sobolev space $H^s(\mathcal{S}^{d-1})$. Therefore, the entropy of $\mathcal{N}(1)$ can be bounded if the entropy of $H^{d/2}(\mathcal{S}^{d-1})(1)$ can be bounded. By Theorem 1.2 of Wang et al. [2014], we have that the k-th entropy number $e_k(T)$ can be bounded by $k^{-d/(2(d-1))}$. This implies that

$$H(\delta, \mathcal{N}(1), \|\cdot\|_{L_\infty}) \leq A\delta^{-\frac{2(d-1)}{d}}.$$

F.3 Proof of Lemma D.1

The first inequality follows the fact that h is positive definite, which implies the inverse of

$$\begin{pmatrix} h(\boldsymbol{s},\boldsymbol{s}) & h(\boldsymbol{X},\boldsymbol{s}) \\ h(\boldsymbol{s},\boldsymbol{X}) & \boldsymbol{h}^\infty \end{pmatrix}$$

is positive definite. By block matrix inverse, we have the first inequality in Lemma D.1 holds.

The second inequality and third inequality are direct results of Theorem 3.2 implies

$$\begin{aligned}&\mathbb{E}_{\epsilon,\boldsymbol{X}}(\|\widehat{g}_n - g^*\|_2^2)\\ =&\int_{\mathbb{S}^{d-1}} (g^*(\boldsymbol{x}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-1}\boldsymbol{y}^*)^2 + h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-2}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x} = O_{\mathbb{P}}(n^{-\frac{d}{2d-1}})\end{aligned}$$

for any function g^* with $\|g^*\|_{\mathcal{N}} \leq 1$. Then we have

$$\int_{\mathbb{S}^{d-1}} h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-2}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x} = O_{\mathbb{P}}(n^{-\frac{d}{2d-1}}),$$

which finishes the proof of the second equality. Let $g^*(\boldsymbol{x}) = h(\boldsymbol{s},\boldsymbol{x})$, then we have

$$\int_{\mathbb{S}^{d-1}} (h(\boldsymbol{s},\boldsymbol{x}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))^2 d\boldsymbol{x} = O_{\mathbb{P}}(n^{-\frac{d}{2d-1}}).$$

By the interpolation inequality, we have

$$\begin{aligned}&h(\boldsymbol{s},\boldsymbol{s}) - h(\boldsymbol{s},\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))\\ \leq&\left\|h(\boldsymbol{s},\cdot) - h(\cdot,\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))\right\|_\infty\\ \leq&C\left\|h(\boldsymbol{s},\cdot) - h(\cdot,\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))\right\|_2^{1-\frac{d-1}{d}}\left\|h(\boldsymbol{s},\cdot) - h(\cdot,\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s})\right\|_{\mathcal{N}}^{\frac{d-1}{d}}\\ =&O_{\mathbb{P}}(n^{-\frac{1}{2d-1}})(h(\boldsymbol{s},\boldsymbol{s}) + h(\boldsymbol{s},\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-1}\boldsymbol{H}^\infty(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))^{\frac{d-1}{d}}\\ \leq&O_{\mathbb{P}}(n^{-\frac{1}{2d-1}})(h(\boldsymbol{s},\boldsymbol{s}) + h(\boldsymbol{s},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\boldsymbol{s}))^{\frac{d-1}{d}} = O_{\mathbb{P}}(n^{-\frac{1}{2d-1}}),\end{aligned}$$

where the last inequality follows the first inequality of Lemma D.1.

F.4 Proof of Lemma D.2

Given that g and f^* have the same value at all $\boldsymbol{x}_i$'s, the empirical norm $\|g - f^*\|_n = 0$. Notice that both g and f^* are in the RKHS generated by the NTK h, denoted by $\mathcal{N}$. Utilizing Lemma C.1 and C.3 similarly as in the proof of Theorem 3.2, we have $R, K = O(1)$ and $J_\infty(z,\mathcal{N}) \lesssim z^{1/d}$, which leads to

$$\sup_{h\in\mathcal{G}(R)}\left|\|h\|_n^2 - \|h\|_2^2\right| = O_{\mathbb{P}}\left(\sqrt{\frac{1}{n}}\right),$$

where $\mathcal{G}(R) := \{g \in \mathcal{N}(1) : \|g - g^*\|_2 \leq R\}$. Therefore, we can conclude that $\|g - f^*\|_2 = O_{\mathbb{P}}(n^{-1/2})$.

F.5 Proof of Lemma D.5

The proof of (a) and (b) can be found in Arora et al. [2019].

For (c), the i-th coordinates of $\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0)$ and $h(\boldsymbol{x},\boldsymbol{X})$ are

$$\frac{1}{m}\sum_{r=1}^m \boldsymbol{x}^\top\boldsymbol{x}_i\mathbb{I}\{\boldsymbol{w}_r^\top(0)\boldsymbol{x} \geq 0\}\mathbb{I}\{\boldsymbol{w}_r^\top(0)\boldsymbol{x}_i \geq 0\}, \quad \text{and} \quad \mathbb{E}_{\boldsymbol{w}\sim N(0,\boldsymbol{I})}[\boldsymbol{x}^\top\boldsymbol{x}_i\mathbb{I}\{\boldsymbol{w}^\top\boldsymbol{x} \geq 0\}\mathbb{I}\{\boldsymbol{w}^\top\boldsymbol{x}_i \geq 0\}],$$

respectively. $\forall i \in [n]$, $(\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0))_i$ is the average of m i.i.d. random variables, which have expectation $h_i(\boldsymbol{x},\boldsymbol{X})$ and bounded in $[0,1]$. For any fixed $\boldsymbol{x}$, by Hoeffding's inequality, with probability at least $1-\delta^*$,

$$|(\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0))_i - h_i(\boldsymbol{x},\boldsymbol{X})| \leq \sqrt{\frac{\log(2/\delta^*)}{2m}}$$

holds. By defining $\delta = n\delta^*$ and applying a union bound over all $i \in [n]$, with probability at least $1-\delta$, we have

$$\left\|\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0) - h(\boldsymbol{x},\boldsymbol{X})\right\|_2^2 = O\left(n\frac{\log(2n/\delta)}{2m}\right)$$

For (d), since

$$\boldsymbol{z}_0(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{W}(0)) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathbb{I}\{\boldsymbol{w}_r(0)^\top \boldsymbol{x} \geq 0\} \boldsymbol{w}_r(0)^\top \boldsymbol{x}$$

Define random variables V_r, $r \in [m]$ as

$$V_r = a_r \mathbb{I}\{\boldsymbol{w}_r(0)^\top \boldsymbol{x} \geq 0\} \boldsymbol{w}_r(0)^\top \boldsymbol{x}$$

Since

$$\boldsymbol{w}_r(0)^\top \boldsymbol{x} \ \sim \ N(0, \tau^2) \quad \text{and} \quad a_r \ \sim \ \mathrm{unif}\{1, -1\}.$$

It's easy to prove that V_r, $r \in [m]$ are i.i.d. with mean 0 and sub-Gaussian parameter τ. By Hoeffding's inequality, at fixed bx, with probability at least $1 - \delta$, we have

$$\left| \frac{1}{\sqrt{m}} \sum_{r=1}^{m} V_r \right| \leq \sqrt{2}\tau \sqrt{\log(2/\delta)}.$$

Thus $\left\| \boldsymbol{z}_0(\cdot)^\top \mathrm{vec}(\boldsymbol{W}(0)) \right\|_2 = O\left(\tau \sqrt{\log(1/\delta)}\right)$.

G More details and results for numerical experiments

Neural network setup The neural network used in all experiments is a 2-layer ReLU neural network with $m = 500$ nodes in each hidden layer. All the weighs are initialized with the Glorot uniform initializer, also called as Xavier uniform initializer [Glorot and Bengio, 2010], which is the default choice in the TensorFlow Keras Sequential module. All the weights are trained by RMSProp [Hinton et al.] optimizer with the default setting, e.g. learning rate of 0.001, etc. All ONN experiments are conducted using TensorFlow 2 with Python API.

G.1 Simulated Data

The learning rate for NTK+ES is $\eta = 0.01$ and the GD update rule is as specified in (D.19). In the ℓ_2-regularized methods, the tuning parameter μ for each task is chosen by cross validation. The validation dataset is of size 100 that is also noiseless and follows the same generating mechanism as the test dataset. For NTK+ℓ_2, we use a grid search of interval $[0, 1]$ with $\mu = 0.01, 0.02, \ldots, 1$ and for ONN+ℓ_2, the μ candidates are $0.1, 0.2, \ldots, 10$. In both cases, we observe that the optimal μ increases with the noise level σ. For f_2^*, we plot the chosen μ and k^* for NTK+ℓ_2 and NTK+ES respectively vs. σ. For each σ value, the reported value is the average of 100 replications. The results are shown in Figure 3.

Figure 1 clearly demonstrates that ONN and NTK do not recover the true function well. As is explained in the paper, without regularization, overfitting the training data is harmful for the L_2 estimation. To illustrate this point, we show the trained estimators of f_2^* for all the methods in Figure 4 when $\sigma = 0.1$.

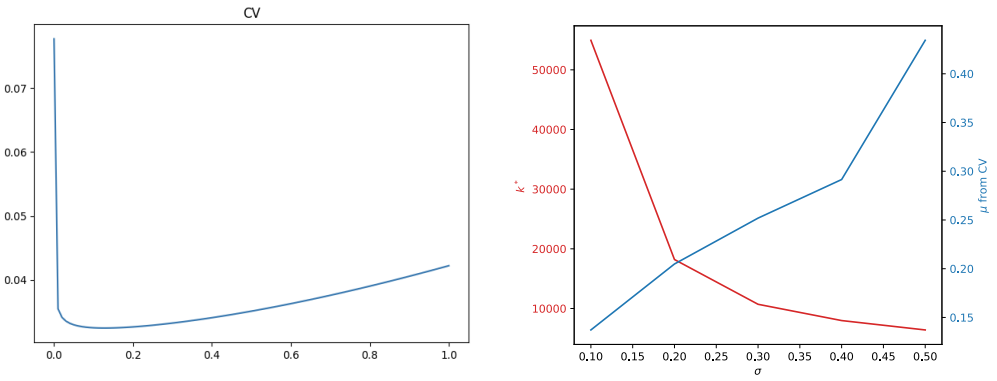

Figure 3: Left: Cross-validation of μ in NTK+ℓ_2 for fitting f_2^* when $\sigma = 0.1$. The horizontal axis is values of μ (100 points from 0.01 to 1) and the vertical axis is the validation mean squared error. The cross-validated μ in this case is 0.13. Right: Optimal stopping time k^* in NTK+ES and cross-validated μ in NTK+ℓ_2 for fitting f_2^* are shown vs. σ. The optimal GD stopping time decrease with noise level while the best μ increases with σ.

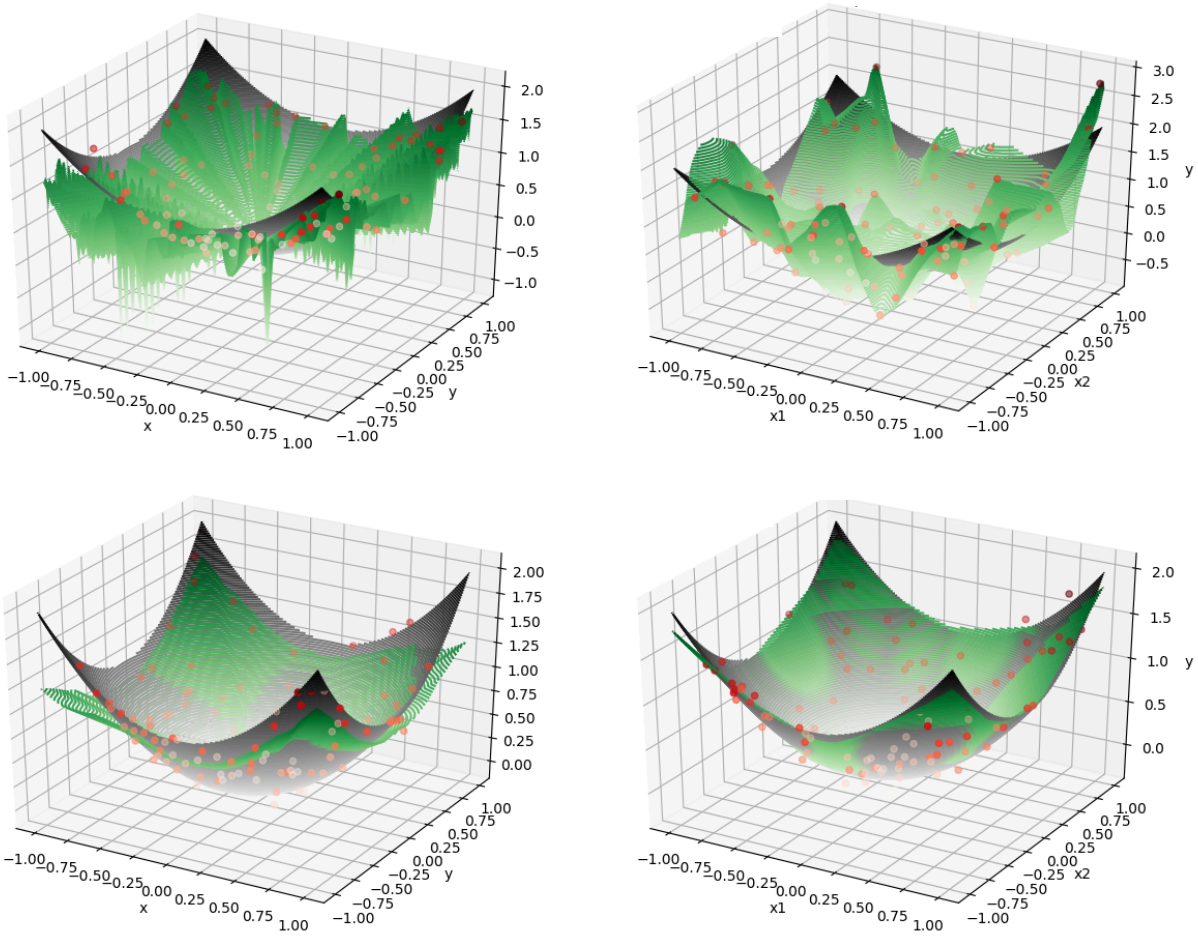

Figure 4: Visualizations for the trained estimators of NTK (top left), NTK+ℓ_2 (bottom left), ONN (top right) and ONN+ℓ_2 (bottom right). Training data are plotted as red dots. The green surface is the estimator and the grey surface is the true function f_2^*. Both surfaces are approximated by grid points $(i/100, j/100)$ for i, j from -100 to 100. As can be seen in the top row, without regularization, the estimators overfit training data. The fitted estimators are very rough and don't recover the true function well.

G.2 MNIST

For images 5 and 8, the training and test split are the default.[3] We change label 5 and 8 to -1 and 1 respectively. No further pre-processing is done to the dataset. For NTK+ES, the learning rate is $\eta = 0.0001$ and the GD update rule is as specified in (D.19). To account for the high data dimension, we divide the NTK matrix $\boldsymbol{H}^{\infty}$ by d. For the ONN+ℓ_2 and NTK+ℓ_2, we choose μ by cross-validation and the candidates are $\mu = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000$ for ONN+ℓ_2 and $\mu = 1, 2, 3, \ldots, 100$ for NTK+ℓ_2. The training/validation split is 80%/20% for cross-validation so the actual training data size is 9107 for all methods (ONN, NTK and NTK+ES do not use the validation dataset). The cross-validated μ for ONN+ℓ_2 and optimal stopping time k^* for NTK+ES are shown in Figure 5, together with the cross-validation results specifically for $\sigma = 1$.

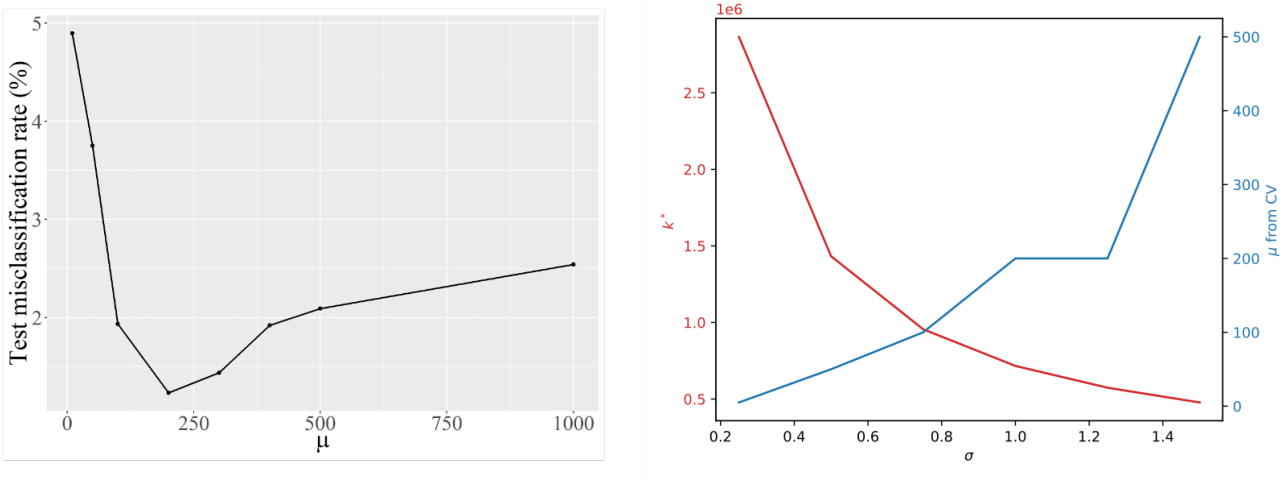

Figure 5: Left: Cross-validation result for μ in ONN+ℓ_2 when $\sigma = 1$ (with extra μ candidates of 300 and 400). In the range of $\mu = 5$ to $\mu = 1000$, we can clearly see a V-shape and the best μ in this case is 200. Right: Optimal stopping time k^* in NTK+ES and cross-validated μ in ONN+ℓ_2 for MNIST dataset are shown vs. σ. The optimal stopping time decreases with noise level while the best μ increases with σ.

[3] http://yann.lecun.com/exdb/mnist/