Algorithm 1 Meta-Learning in RKHS

Require: $p(\mathcal{T})$: distribution over tasks, randomly initialized neural network parameters $\boldsymbol{\theta}$.
  while not done **do**
    Sample a batch of tasks $\{\mathcal{T}_m\}_{m=1}^B \sim p(\mathcal{T})$
    for all $\mathcal{T}_m$ **do**
      Sample a batch of data points $\mathcal{D}_m$ **or** Sample two batches of data points $\mathcal{D}_m^{tr}$, $\mathcal{D}_m^{test}$.
    end for
    Evaluate the energy functional by equation 4 with $\{\mathcal{D}_m\}_{m=1}^B$ **or** Evaluate the energy functional by equation 7 with $\{\mathcal{D}_m^{tr}, \mathcal{D}_m^{test}\}_{m=1}^B$. Minimize the energy functional w.r.t $\boldsymbol{\theta}$.
  end while

A ALGORITHMS

Our proposed algorithms for meta-learning in the RKHS are summarized in Algorithm 1.

B PROOF OF THEOREM 1

Theorem 1 *If $f_{\boldsymbol{\theta}}$ is a neural network with parameter $\boldsymbol{\theta} \in R^P$ and $\mathcal{H}$ is the Reproducing Kernel Hilbert Space (RKHS) induced by $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the Neural Tangent Kernel (NTK) of $f_{\boldsymbol{\theta}}$, then with initialization $f^0 = f_{\boldsymbol{\theta}^0}$, the gradient flow of $\mathcal{E}(f^t)$ coincides with the function evolution of $f_{\boldsymbol{\theta}^t}$ induced by the gradient flow of $E(\boldsymbol{\theta}^t)$.*

Proof Without loss of generality, we can rewrite $\mathcal{E}(f) = \mathbb{E}_{\mathcal{T}_m}\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}[C(f(\mathbf{x}_m),\mathbf{y}_m)]\}$ with some function $C(\cdot,\cdot)$.

For a neural network $f_{\boldsymbol{\theta}}$ with parameter $\boldsymbol{\theta} \in R^P$, the gradient flow of E in R^P is

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}^t} E(\boldsymbol{\theta}^t).$$

We have

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} &= -\nabla_{\boldsymbol{\theta}^t}(\mathcal{E}\circ F)(\boldsymbol{\theta}^t) \\
&= -\mathbb{E}_{\mathcal{T}_m}\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}[\nabla_{\boldsymbol{\theta}^t} C(f_{\boldsymbol{\theta}^t}(\mathbf{x}_m),\mathbf{y}_m)]\} \\
&= -\mathbb{E}_{\mathcal{T}_m}\left\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\left[\frac{\partial C(f^t_{\boldsymbol{\theta}}(\mathbf{x}_m),\mathbf{y}_m)}{\partial f^t_{\boldsymbol{\theta}}(\mathbf{x}_m)}\frac{\partial f^t_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}^t}\right]\right\}.
\end{aligned}$$

We know that the dynamics of $f_{\boldsymbol{\theta}^t}$ is

$$\begin{aligned}
\frac{\mathrm{d}f_{\boldsymbol{\theta}^t}}{\mathrm{d}t} &= \frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t}\frac{\partial f_{\boldsymbol{\theta}^t}}{\partial \boldsymbol{\theta}^t}^\intercal \\
&= -\mathbb{E}_{\mathcal{T}_m}\left\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\left[\frac{\partial C(f_{\boldsymbol{\theta}^t}(\mathbf{x}_m),\mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}^t}(\mathbf{x}_m)}\frac{\partial f_{\boldsymbol{\theta}^t}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}^t}\right]\right\}\frac{\partial f_{\boldsymbol{\theta}^t}}{\partial \boldsymbol{\theta}^t}^\intercal \\
&= -\mathbb{E}_{\mathcal{T}_m}\left\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\left[\frac{\partial C(f_{\boldsymbol{\theta}^t}(\mathbf{x}_m),\mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}^t}(\mathbf{x}_m)}\frac{\partial f_{\boldsymbol{\theta}^t}(\mathbf{x})}{\partial \boldsymbol{\theta}^t}\frac{\partial f_{\boldsymbol{\theta}^t}}{\partial \boldsymbol{\theta}^t}^\intercal\right]\right\} \\
&= -\mathbb{E}_{\mathcal{T}_m}\left\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\left[\frac{\partial C(f_{\boldsymbol{\theta}^t}(\mathbf{x}_m),\mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}^t}(\mathbf{x}_m)}\boldsymbol{\Theta}^t(\mathbf{x}_m,\cdot)\right]\right\},
\end{aligned} \tag{8}$$

where $\boldsymbol{\Theta}^t$ is the Neural Tangent Kernel of neural network $f_{\boldsymbol{\theta}^t}$ (Jacot et al., 2018).

If $\mathcal{H}^t$ is the Reproducing Kernel Hilbert Space induced by a kernel $\boldsymbol{\Theta}^t$ and $V_{\mathbf{x}_m} : \mathcal{H} \to R$ is the evaluation functional at $\mathbf{x}_m$, which is defined as

$$V_{\mathbf{x}_m}(f) = f(\mathbf{x}_m),$$

then for an arbitrary function g and a small perturbation ϵ, we have

$$\begin{aligned}
\langle \nabla_f V_{\mathbf{x}_m}(f), g\rangle &= \lim_{\epsilon\to 0} \frac{V_{\mathbf{x}_m}(f+\epsilon g) - V_{\mathbf{x}_m}(f)}{\epsilon} \\
\langle \nabla_f V_{\mathbf{x}_m}(f), g\rangle &= \lim_{\epsilon\to 0} \frac{f(\mathbf{x}_m) + \epsilon g(\mathbf{x}_m) - f(\mathbf{x}_m)}{\epsilon} \\
\langle \nabla_f V_{\mathbf{x}_m}(f), g\rangle &= g(\mathbf{x}_m) \\
\langle \nabla_f V_{\mathbf{x}_m}(f), g\rangle &= \langle \boldsymbol{\Theta}^t(\mathbf{x}_m, \cdot), g\rangle \\
\nabla_f V_{\mathbf{x}_m}(f) &= \boldsymbol{\Theta}^t(\mathbf{x}_m, \cdot) \\
\nabla_f f(\mathbf{x}_m) &= \boldsymbol{\Theta}^t(\mathbf{x}_m, \cdot).
\end{aligned}$$

With an initial function $f^0 = f_{\boldsymbol{\theta}^0} \in \mathcal{H}$, the gradient flow of $\mathcal{E}$ in $\mathcal{H}$ is

$$\frac{\mathrm{d}f^t}{\mathrm{d}t} = -\nabla_{f^t}\mathcal{E}(f^t).$$

We have

$$\begin{aligned}
\frac{\mathrm{d}f^t}{\mathrm{d}t} &= -\mathbb{E}_{\mathcal{T}_m}\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\left[\nabla_{f^t} C(f^t(\mathbf{x}_m), \mathbf{y}_m)\right]\} \\
&= -\mathbb{E}_{\mathcal{T}_m}\left\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\left[\frac{\partial C(f^t(\mathbf{x}_m), \mathbf{y}_m)}{\partial f^t(\mathbf{x}_m)} \nabla_{f^t} f^t(\mathbf{x}_m)\right]\right\} \\
&= -\mathbb{E}_{\mathcal{T}_m}\left\{\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\left[\frac{\partial C(f^t(\mathbf{x}_m), \mathbf{y}_m)}{\partial f^t(\mathbf{x}_m)} \boldsymbol{\Theta}^t(\mathbf{x}_m, \cdot)\right]\right\}. \qquad (9)
\end{aligned}$$

We can complete the proof by comparing equation 8 and equation 9. ■

C Proof of Theorem 2

Theorem 2 *If $f_{\boldsymbol{\theta}}$ is a neural network with parameter $\boldsymbol{\theta}$ and $\mathcal{H}$ is the Reproducing Kernel Hilbert Space (RKHS) induced by $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the Neural Tangent Kernel (NTK) of $f_{\boldsymbol{\theta}}$, then*

$$\mathcal{M}_1 = \widetilde{\mathcal{E}}(\alpha, f_{\boldsymbol{\theta}}), \text{ and } \beta_0 = \alpha\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2 = \alpha\|\nabla_{f_{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|_{\mathcal{H}}^2.$$

Proof Without loss of generality, we rewrite $\mathcal{L}_m(f_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left[C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m)\right]$.

In regression task, we have

$$C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m) = \frac{1}{2}\left\|f_{\boldsymbol{\theta}}(\mathbf{x}_m) - \mathbf{y}_m\right\|^2$$

. In classification task, we have

$$C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m) = \mathbf{y}_m \log(f_{\boldsymbol{\theta}}(\mathbf{x}_m))^\intercal,$$

where log is element-wise logarithm operation.

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2 \\
&= \nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})^\intercal \\
&= \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left[C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m)\right]\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left[C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m)\right]^\intercal \\
&= \mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left[\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}}\right]\mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left[\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}}^\intercal\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}^\intercal\right] \\
&= \mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left\{\mathbb{E}_{\mathbf{x}'_m,\mathbf{y}'_m}\left[\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}}\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}'_m)}{\partial \boldsymbol{\theta}}^\intercal\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}'_m), \mathbf{y}'_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}'_m)}^\intercal\right]\right\} \\
&= \mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left\{\mathbb{E}_{\mathbf{x}'_m,\mathbf{y}'_m}\left[\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}_m), \mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}\boldsymbol{\Theta}(\mathbf{x}_m, \mathbf{x}'_m)\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}'_m), \mathbf{y}'_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}'_m)}^\intercal\right]\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \left\langle \mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left[\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}_m),\mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}\boldsymbol{\Theta}(\mathbf{x}_m,\cdot)\right], \mathbb{E}_{\mathbf{x}'_m,\mathbf{y}'_m}\left[\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}'_m),\mathbf{y}'_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}'_m)}\boldsymbol{\Theta}(\mathbf{x}'_m,\cdot)\right] \right\rangle_{\mathcal{H}} \\
&= \left\langle \mathbb{E}_{\mathbf{x}_m,\mathbf{y}_m}\left[\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}_m),\mathbf{y}_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}\nabla_{f_{\boldsymbol{\theta}}} f_{\boldsymbol{\theta}}(\mathbf{x}_m)\right], \mathbb{E}_{\mathbf{x}'_m,\mathbf{y}'_m}\left[\frac{\partial C(f_{\boldsymbol{\theta}}(\mathbf{x}'_m),\mathbf{y}'_m)}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}'_m)}\nabla_{f_{\boldsymbol{\theta}}} f_{\boldsymbol{\theta}}(\mathbf{x}'_m)\right] \right\rangle_{\mathcal{H}} \\
&= \left\langle \nabla_{f_{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\boldsymbol{\theta}}), \nabla_{f_{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\boldsymbol{\theta}}) \right\rangle_{\mathcal{H}} \\
&= \|\nabla_{f_{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2_{\mathcal{H}},
\end{aligned}
$$

where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is the inner product in Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$. In the above equations, we use the definition of Neural Tangent Kernel (NTK), the property of inner product in RKHS, the definition of evaluation functional and its gradient in RKHS.

Recall that

$$\widetilde{\mathcal{E}}(\alpha, f_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - \alpha\|\nabla_{f_{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2_{\mathcal{H}}\right]$$

and

$$\mathcal{M}_k = \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - \sum_{i=0}^{k-1}\beta_i\right],$$

where $\beta_i = \alpha\nabla_{\boldsymbol{\theta}_i}\mathcal{L}_m(f_{\boldsymbol{\theta}_i})\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})^\intercal$ and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}, \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \alpha\nabla_{\boldsymbol{\theta}_i}\mathcal{L}(f_{\boldsymbol{\theta}_i}, \mathcal{D}_m^{tr})$. The result is straightforward now.

■

D PROOF OF THEOREM 3

The proof techniques we use are similar to some previous works such as (Arora et al., 2019; Allen-Zhu et al., 2019). We summaries some of the differences. Different from previous works that typically assume a neural network is Gaussian initialized, we do not have such an assumption as we are trying to learn a good meta-initialization in the meta-learning setting. Previous works try to investigate the behavior of models during training, while we focus on revealing the connection between different meta-learning algorithms. Previous work focuses on single-task regression/classification problems, while we focus on meta-learning problem.

Theorem 3 *Let $f_{\boldsymbol{\theta}}$ be a fully-connected neural network with L hidden layers and ReLU activation function, $s_1, ..., s_{L+1}$ be the spectral norm of the weight matrices, $s = \max_h s_h$, and α be the learning rate of gradient descent. If $\alpha \leq O(qr)$ with $q = \min(1/(Ls^L), L^{-1/(L+1)})$ and $r = \min(s^{-L}, s)$, then the following holds*

$$|\widetilde{\mathcal{E}}(k\alpha, f_{\boldsymbol{\theta}}) - \mathcal{M}_k| \leq O\Big(\frac{1}{L}\Big).$$

Proof We first prove the case of $k = 2$, i.e. applying a two-step gradient descent adaptation in MAML.

We need to prove the following theorem first.

Theorem 6 *Let $f_{\boldsymbol{\theta}}$ be a fully-connected neural network with L hidden layers, and $\mathbf{x}$ be a data sample. Represent the neural network by $f_{\boldsymbol{\theta}}(\mathbf{x}) = \sigma(\sigma(...\sigma(\mathbf{x}W^1)...W^{L-1})W^L)W^{L+1}$, where $W^1, ..., W^{L+1}$ denote the weight matrices, and σ is the ReLU activation function. Let $s_1, ..., s_{L+1}$ be the spectral norm of weight matrices, and $s = \max_h s_h$. Let α be the learning rate of gradient descent, and $\boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})$ be the resulting value after one step of gradient descent, and $\|\cdot\|_{\mathcal{F}}$ be the Frobenius norm.*

If $\alpha \leq O(qs^{-L})$, where $q = \min(1/(Ls^L), L^{-1/(L+1)})$, then

$$\Big\|\frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}}\Big\|_{\mathcal{F}} \leq O(\frac{1}{s\sqrt{L+1}}).$$

Remark 1 *Theorem 6 states that for a neural network with L hidden layers, if the learning rate of gradient descent is bounded, then the norm of derivative w.r.t all the parameters will not change*

too much, although there are $O(Lm^2)$ parameters, where m denotes the maximum width of hidden layers. We use row vector instead of column vector for consistency, while it does not affect our results.

For simplicity, we will write $g^h(\mathbf{x})$ as g^h. The bias terms in the neural network are introduced by adding an additional coordinate thus omitted in Theorem 6. Without loss of generality, we can assume $\| \mathbf{x} \| \leq 1$, which can be done by data normalization in pre-processing.

Let $g^h(\mathbf{x}) = \sigma(\sigma(...\sigma(\mathbf{x}W^1)...W^{h-1})W^h)$ be the activation at h^{th} hidden layer and $g^0(\mathbf{x}) = \mathbf{x}, g^{L+1} = f_{\boldsymbol{\theta}}(\mathbf{x})$. Define diagonal matrices D^h, where $D^h_{(i,i)} = \mathbf{1}\{g^{h-1}W^h \geq 0\}$ and

$$b^h = \begin{cases} \mathbf{I}_{d_y}, & \text{if } h = L+1 \\ b^{h+1}(W^{h+1})^\intercal D^h, & \text{otherwise} \end{cases}$$

where $\mathbf{I}_{d_y}$ is a $d_y \times d_y$ identity matrix. We first prove the following Lemma.

Lemma 7 *Given a neural network as stated in Theorem 6, let $\| \cdot \|_2$ denote the spectral norm, $\triangle W^h = \tilde{W}^h - W^h$ denote some perturbation on weight matrices, $\tilde{g}^h(\mathbf{x})$ denote the resulting value after perturbation, and $\triangle g^h(\mathbf{x}) = \tilde{g}^h(\mathbf{x}) - g^h(\mathbf{x})$. If $s \geq 1$ and $\|\triangle W^h\|_2 \leq O(s^{-L}/L)$ for all h, then*

$$\|\triangle g^h\| \leq O(\frac{1}{Ls^{L-h+1}});$$

If $s < 1$ and $\|\triangle W^h\|_2 \leq O(q)$ for all h, where $q = \min(1/(Ls^L), L^{-1/(L+1)})$ and $r = \max(q, s)$, then

$$\|\triangle g^h\| \leq O(r^{h-1}q) = \begin{cases} O(\frac{1}{Ls^{L-h+1}}), & \text{if } 1/(Ls^L) \leq L^{-1/(L+1)} \\ O(L^{-h/(L+1)}), & \text{if } 1/(Ls^L) > L^{-1/(L+1)}. \end{cases}$$

Proof Proof of Lemma 7 is based on induction.

We first prove the case of $s \geq 1$. Note that $g^0 = \mathbf{x}$, thus $\triangle g^0 = 0 \leq O(\frac{1}{Ls^{L-0+1}})$ always holds.

For $\triangle g^1$, we have

$$\begin{aligned} \|\triangle g^1\| &= \|\sigma(\mathbf{x}\tilde{W}^1) - \sigma(\mathbf{x}W^1)\| \\ &\leq \| \mathbf{x}\tilde{W}^1 - \mathbf{x}W^1\|, \quad \text{due to the property of ReLU activation} \\ &\leq \| \mathbf{x} \| \|\triangle W^1\|_2 \\ &\leq O(\frac{1}{Ls^L}). \end{aligned}$$

Thus, the hypothesis holds for $\triangle g^1$.

Now, assume that the hypothesis holds for $\triangle g^h$, then we have

$$\begin{aligned} \|\triangle g^{h+1}\| &= \|\sigma(\tilde{g}^h\tilde{W}^{h+1}) - \sigma(g^hW^{h+1})\| \\ &\leq \|\tilde{g}^h\tilde{W}^{h+1} - g^hW^{h+1}\|, \quad \text{due to the property of ReLU activation} \\ &\leq \|\tilde{g}^hW^{h+1} + \tilde{g}^h\triangle W^{h+1} - g^hW^{h+1}\| \\ &\leq \|\triangle g^h\|\|W^{h+1}\|_2 + \|\tilde{g}^h\|\|\triangle W^{h+1}\|_2 \\ &\leq O(s)\|\triangle g^h\| + \|g^h + \triangle g^h\|\|\triangle W^{h+1}\|_2 \\ &\leq O(s)\|\triangle g^h\| + O(s^h)\|\triangle W^{h+1}\|_2 + \|\triangle g^h\|\|\triangle W^{h+1}\|_2 \\ &\leq O(s)O(\frac{1}{Ls^{L-h+1}}) + O(s^h)O(\frac{1}{Ls^L}) + O(\frac{1}{Ls^{L-h+1}})O(\frac{1}{Ls^L}) \\ &\leq O(\frac{1}{Ls^{L-h}}). \end{aligned}$$

The last three inequalities come from the fact that $g^h = \sigma(\sigma(...\sigma(\mathbf{x}W^1)...W^{h-1})W^h) \leq O(s^h)$ and $s \geq 1$. Thus, we have proved the Lemma in the case $s \geq 1$.

Now, we prove the first part of the case of $s < 1$, i.e. $\|\triangle g^h\| \leq O(r^{h-1}q)$. Because $\triangle g^0 = 0$, thus the hypothesis for $\triangle g^0$ always holds.

For $\triangle g^1$, we have

$$\begin{aligned}\|\triangle g^1\| &= \|\sigma(\mathbf{x}\tilde{W}^1) - \sigma(\mathbf{x}W^1)\| \\ &\leq \|\mathbf{x}\tilde{W}^1 - \mathbf{x}W^1\| \\ &\leq \|\mathbf{x}\|\|\triangle W^1\|_2 \\ &\leq O(q).\end{aligned}$$

Thus, the hypothesis holds for $\triangle g^1$.

Now, we assume that the hypothesis holds for $\triangle g^h$. Then, we have

$$\begin{aligned}\|\triangle g^{h+1}\| &= \|\sigma(\tilde{g}^h\tilde{W}^{h+1}) - \sigma(g^hW^{h+1})\| \\ &\leq \|\tilde{g}^h\tilde{W}^{h+1} - g^hW^{h+1}\| \\ &\leq \|\tilde{g}^hW^{h+1} + \tilde{g}^h\triangle W^{h+1} - g^hW^{h+1}\| \\ &\leq \|\triangle g^h\|\|W^{h+1}\|_2 + \|\tilde{g}^h\|\|\triangle W^{h+1}\|_2 \\ &\leq O(s)\|\triangle g^h\| + \|g^h + \triangle g^h\|\|\triangle W^{h+1}\|_2 \\ &\leq O(s)O(r^{h-1}q) + O(s^h)q + qO(r^{h-1}q) \\ &\leq O(r^hq).\end{aligned}$$

The last inequality comes from the fact that $r = \max(q, s)$ and $s^h < s < 1$.

Next we consider the second part of the case of $s < 1$.

If $1/(Ls^L) \leq L^{-1/(L+1)}$, we know that $q = 1/(Ls^L)$ and

$$\begin{aligned}1/(Ls^L) &\leq L^{-1/(L+1)} \\ L^{1/(L+1)} &\leq Ls^L \\ L^{-L/(L+1)} &\leq s^L \\ L^{-1} &\leq s^{L+1} \\ L^{-1}s^{-L} &\leq s,\end{aligned}$$

which means $q \leq s$, thus $r = s$. Then, we have

$$\|\triangle g^h\| = O(r^{h-1}q) = O(s^{h-1}q) = O(s^{h-1}L^{-1}s^{-L}) = O(\frac{1}{Ls^{L-h+1}}).$$

If $1/(Ls^L) > L^{-1/(L+1)}$, we know that $q = L^{-1/(L+1)}$ and $q > s$; then, $r = q$ and

$$\|\triangle g^h\| = O(r^{h-1}q) = O(q^{h-1}q) = O(q^h) = O(L^{-h/(L+1)}).$$

Thus, we can conclude that Lemma 7 also holds for the case of $s < 1$, which completes the proof. ■

We now prove a similar Lemma for $\triangle b^h$.

Lemma 8 *Given a neural network as stated in Theorem 6, let $\|\cdot\|_2$ denote the spectral norm, $\triangle W^h = \tilde{W}^h - W^h$ denote some perturbation on weight matrices, $\tilde{b}^h$ denote the resulting value after perturbation, and $\triangle b^h = \tilde{b}^h - b^h$.*

If $s \geq 1$ and $\|\triangle W^h\|_2 \leq O(s^{-L}/L)$ for all h, then

$$\|\triangle b^h\| \leq O(\frac{1}{Ls^h});$$

If $s < 1$ and $\|\triangle W^h\|_2 \leq O(q)$ for all h, where $q = \min(1/(Ls^L), L^{-1/(L+1)})$, then

$$\|\triangle b^h\| \leq \begin{cases} O(L^{-1}s^{-h}), & \text{if } 1/(Ls^L) \leq L^{-1/(L+1)} \\ O(L^{(h-L-1)/(L+1)}), & \text{if } 1/(Ls^L) > L^{-1/(L+1)}. \end{cases}$$

Proof Recall that

$$b^h = \begin{cases} \mathbf{I}_{d_y}, & \text{if } h = L+1 \\ b^{h+1}(W^{h+1})^\intercal D^h, & \text{otherwise} \end{cases}$$

where $\mathbf{I}_{d_y}$ is a $d_y \times d_y$ identity matrix and $D^h_{(i,i)} = \mathbf{1}\{g^{h-1}W^h \geq 0\}$. It is easy to see that $\|b^h\| \leq O(s^{L-h+1})$, because $\|D^h\|_2 \leq 1$ and $\|W^h\|_2 \leq s$.

We first prove the case of $s \geq 1$. We know that $\triangle b^{L+1} = 0 \leq O(s^{-L-1}/L)$ always holds.

For $h \leq L$, we can re-write b^h as

$$b^h = \mathbf{I}_{d_y}(W^{L+1})^\intercal D^L (W^L)^\intercal D^{L-1} ... (W^{h+1})^\intercal D^h.$$

Then, we have

$$b^h (g^h)^\intercal = \mathbf{I}_{d_y}(W^{L+1})^\intercal D^L (W^L)^\intercal D^{L-1} ... (W^{h+1})^\intercal D^h (g^h)^\intercal. \tag{10}$$

Because of the fact that

$$f_{\boldsymbol{\theta}} = g^{L+1} = \mathbf{x} W^1 D^1 W^2 D^2 ... D^L W^{L+1} = g^h W^{h+1} D^{h+1} ... D^L W^{L+1}$$

and $g^h = g^h D^h$, $D^h = (D^h)^\intercal$. We can re-write equation 10 as

$$b^h (g^h)^\intercal = f^\intercal_{\boldsymbol{\theta}}.$$

Thus,

$$\|\tilde{b}^h (\tilde{g}^h)^\intercal - b^h (g^h)^\intercal\| = \|\boldsymbol{f}_{\tilde{\boldsymbol{\theta}}} - f_{\boldsymbol{\theta}}\| = \|\triangle g^{L+1}\| \leq O(\frac{1}{L})$$

by Lemma 7. Consequently, we have

$$\|\tilde{b}^h (\tilde{g}^h)^\intercal - b^h (g^h)^\intercal\| = \|\triangle b^h (g^h)^\intercal + \triangle b^h \triangle (g^h)^\intercal + \tilde{b}^h \triangle (g^h)^\intercal\| \leq O(\frac{1}{L}).$$

Since $\|g^h\| \leq O(s^h)$, we know that

$$\|\triangle b^h\| \leq O(\frac{1}{Ls^h}), \quad \|\triangle b^h\| \leq O(s^{L-h+1})$$

always hold. Since $L \geq 1, s \geq 1$, we simply have $\|\triangle b^h\| \leq O(\frac{1}{Ls^h})$.

Now, we prove the case of $s < 1$. Similarly, we have

$$\|\tilde{b}^h (\tilde{g}^h)^\intercal - b^h (g^h)^\intercal\| = \|\boldsymbol{f}_{\tilde{\boldsymbol{\theta}}} - f_{\boldsymbol{\theta}}\| = \|\triangle g^{L+1}\| \leq O(\frac{1}{L}).$$

Similarly, we must have

$$\|\triangle b^h\| \leq O(\frac{1}{Ls^h}), \quad \|\triangle b^h\| \leq O(\frac{1}{Lr^{h-1}q}),$$

where $q = \min(1/(Ls^L), L^{-1/(L+1)})$ and $r = \max(q, s)$ by Lemma 7.

If $1/(Ls^L) \leq L^{-1/(L+1)}$, then $s^{L+1} \geq 1/L$. We thus have

$$O(\frac{1}{Lr^{h-1}q}) = O(\frac{Ls^{L-h+1}}{L}) = O(\frac{s^{L+1}}{s^h}) \geq O(\frac{1}{Ls^h}).$$

Hence, we get $\|\triangle b^h\| \leq O(\frac{1}{Ls^h})$.

If $1/(Ls^L) > L^{-1/(L+1)}$, then $s^{L+1} < 1/L$. We have

$$O(\frac{1}{Lr^{h-1}q}) = O(L^{-1} \cdot L^{h/(L+1)}) \leq O(L^{-1} \cdot s^{-h}) = O(\frac{1}{Ls^h}).$$

Thus, we get $\|\triangle b^h\| \leq O(L^{(h-L-1)/(L+1)})$. ■

Lemma 9 *Given a neural network as stated in Theorem 6, let $\|\cdot\|_{\mathcal{F}}$ be the Frobenius norm, $W^1, ..., W^{L+1}$ be the weight matrices in the neural network, $\triangle W^h = \tilde{W}^h - W^h$ be the perturbation on weight matrices, $\boldsymbol{\theta}^h$ be the parameter vector containing all the elements in W^h, $\triangle \boldsymbol{\theta}^h = \tilde{\boldsymbol{\theta}}^h - \boldsymbol{\theta}^h$ be the perturbation on parameter vectors, and $\boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})$ be the resulting value after perturbation.*

If $s \geq 1$ and $\|\triangle W^h\|_2 \leq O(s^{-L}/L)$ for all h, for any weight matrices the following holds

$$\Big\|\frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}^h} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}^h}\Big\|_{\mathcal{F}} \leq O(\frac{1}{sL});$$

If $s < 1$ and $\|\triangle W^h\|_2 \leq O(q)$ for all h, where $q = \min(1/(Ls^L), L^{-1/(L+1)})$, for any weight matrices the following holds

$$\Big\|\frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}^h} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}^h}\Big\|_{\mathcal{F}} \leq O(\frac{1}{sL}).$$

Proof We first prove the case of $d_y = 1$, i.e. the output of neural network is 1-dimensional.

In this case, we know that

$$\Big\|\frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}^h} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}^h}\Big\|_{\mathcal{F}} = \Big\|\frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{W}^h} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h}\Big\|_{\mathcal{F}} = \Big\|\triangle \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h}\Big\|_{\mathcal{F}}$$

and the derivative to W^h is

$$\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h} = (b^h)^\intercal g^{h-1}.$$

Then, we have

$$\begin{aligned}
\Big\|\triangle \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h}\Big\|_{\mathcal{F}} &= \|(\tilde{b}^h)^\intercal \tilde{g}^{h-1} - (b^h)^\intercal g^{h-1}\|_{\mathcal{F}} \\
&= \|(\tilde{b}^h)^\intercal g^{h-1} - (b^h)^\intercal g^{h-1} + (\tilde{b}^h)^\intercal \triangle g^{h-1}\|_{\mathcal{F}} \\
&\leq \|(\triangle b^h)^\intercal g^{h-1}\|_{\mathcal{F}} + \|(b^h + \triangle b^h)^\intercal \triangle g^{h-1}\|_{\mathcal{F}}.
\end{aligned}$$

Recall the fact that $g^h \leq O(s^h)$ and $b^h \leq O(s^{L+1-h})$.

When $s \geq 1$, from Lemma 7 and Lemma 8 we know that

$$\|\triangle g^h\| \leq O(\frac{1}{Ls^{L-h+1}}), \quad \|\triangle b^h\| \leq O(\frac{1}{Ls^h}).$$

Then, we have

$$\begin{aligned}
\Big\|\triangle \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h}\Big\|_{\mathcal{F}} &\leq O(s^{h-1})O(\frac{1}{Ls^h}) + O(s^{L+1-h})O(\frac{1}{Ls^{L-h+2}}) + O(\frac{1}{Ls^{L-h+2}})O(\frac{1}{Ls^h}) \\
&\leq O(\frac{1}{sL}).
\end{aligned}$$

When $s < 1$, from Lemma 7 and Lemma 8 we know that

$$\|\triangle g^h\| \leq \begin{cases} O(\frac{1}{Ls^{L-h+1}}), & \text{if } 1/(Ls^L) \leq L^{-1/(L+1)} \\ O(L^{-h/(L+1)}), & \text{if } 1/(Ls^L) > L^{-1/(L+1)} \end{cases}$$

and

$$\|\triangle b^h\| \leq \begin{cases} O(L^{-1}s^{-h}), & \text{if } 1/(Ls^L) \leq L^{-1/(L+1)} \\ O(L^{(h-L-1)/(L+1)}), & \text{if } 1/(Ls^L) > L^{-1/(L+1)}. \end{cases}$$

If $1/(Ls^L) \leq L^{-1/(L+1)}$, we have

$$\Big\|\triangle \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h}\Big\|_{\mathcal{F}} \leq O(s^{h-1})O(\frac{1}{Ls^h}) + O(s^{L-h+1})O(\frac{1}{Ls^{L-h+2}}) + O(\frac{1}{Ls^{L-h+2}})O(\frac{1}{Ls^h}).$$

Since $1/(Ls^L) \leq L^{-1/(L+1)}$ implies $L^{-1} \leq s^{L+1}$ (from proof of Lemma 7), we have

$$\frac{1}{Ls^h} \leq s^{L-h+1}.$$

Then we can conclude that

$$\left\| \triangle \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h} \right\|_{\mathcal{F}} \leq O(\frac{1}{sL}).$$

If $1/(Ls^L) > L^{-1/(L+1)}$, we have

$$\begin{aligned}\left\| \triangle \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h} \right\|_{\mathcal{F}} \leq & O(s^{h-1})O(L^{(h-L-1)/(L+1)}) + O(s^{L+1-h})O(L^{-(h-1)/(L+1)}) \\ & + O(L^{-(h-1)/(L+1)})O(L^{(h-L-1)/(L+1)}).\end{aligned}$$

Since $1/(Ls^L) > L^{-1/(L+1)}$ implies $L^{-1} > s^{L+1}$ (from proof of Lemma 7), we have

$$L^{(h-L-1)/(L+1)} > s^{L-h+1}, \quad \frac{1}{L^{(h-1)/(L+1)}} > s^{h-1}.$$

Then we have

$$\left\| \triangle \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h} \right\|_{\mathcal{F}} \leq O(\frac{1}{L}) \leq O(\frac{1}{sL}), \text{ because } s < 1.$$

We have proved the Lemma for the case of $d_y = 1$.

For the case of $d_y > 1$, we know that

$$\left\| \frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}^h} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}^h} \right\|_{\mathcal{F}}^2 = \sum_{i=1}^{d_y} \left\| \frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}},i}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}^h} - \frac{\partial f_{\boldsymbol{\theta},i}(\mathbf{x})}{\partial \boldsymbol{\theta}^h} \right\|_{\mathcal{F}}^2 \leq O(\frac{d_y}{s^2 L^2}),$$

where $f_{\boldsymbol{\theta},i}(\mathbf{x})$ is the i^{th} dimension of $f_{\boldsymbol{\theta}}(\mathbf{x})$. The last inequality directly comes from the 1-dimensional case.

Since d_y is a constant, we ignore it. Then, we have

$$\left\| \frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}^h} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}^h} \right\|_{\mathcal{F}} \leq O(\frac{1}{sL}),$$

which completes the proof. ■

Now we can prove Theorem 6, if $\tilde{W}^h$ is obtained by one step gradient descent starting from W^h, $\tilde{\boldsymbol{\theta}}$ is obtained by one step gradient descent starting from $\boldsymbol{\theta}$, and learning rate is α. Then, for any weight matrix we have

$$\begin{aligned}\|\triangle W^h\|_2 &= \|\alpha \nabla_{W^h} \mathcal{L}(\boldsymbol{\theta})\|_2 \\ &\leq \|\alpha \nabla_{W^h} \mathcal{L}(\boldsymbol{\theta})\|_{\mathcal{F}} \\ &= \|\alpha \nabla_{\boldsymbol{\theta}^h} \mathcal{L}(\boldsymbol{\theta})\|_{\mathcal{F}} \\ &= \alpha \left\| \frac{\sum_{i=1}^n}{n} \left[f_{\boldsymbol{\theta}}(\mathbf{x}_i) - \mathbf{y}_i \right] \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \boldsymbol{\theta}^h}^{\intercal} \right\|_{\mathcal{F}} \\ &\leq \frac{\alpha \sum_{i=1}^n}{n} c_i \left[\sum_j^{d_y} \left\| \frac{\partial f_{\boldsymbol{\theta},j}(\mathbf{x}_i)}{\partial W^h} \right\|_{\mathcal{F}}^2 \right]^{1/2} \\ &\leq \frac{\alpha \sum_{i=1}^n}{n} c_i \sqrt{d_y} O(s^{L-h+1}) O(s^{h-1}) \\ &\leq \alpha O(s^L),\end{aligned}$$

where $c_i = \|f_{\boldsymbol{\theta}}(\mathbf{x}_i) - \mathbf{y}_i\|$ are some constants.

If $\alpha \leq O(s^{-2L}/L)$ when $s \geq 1$, then for any weight matrix we have

$$\|\triangle W^h\|_2 \leq \alpha O(s^L) \leq O(s^{-L}/L).$$

If $\alpha \leq O(qs^{-L})$ where $q = \min(1/(Ls^L), L^{-1/(L+1)})$ when $s < 1$, then for any weight matrix we have

$$\|\triangle W^h\|_2 \leq \alpha O(s^L) \leq O(q).$$

By Lemma 9, we can conclude that

$$\Big\|\frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{W}^h} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h}\Big\|_{\mathcal{F}} \leq O(\frac{1}{sL}).$$

Then, we have

$$\begin{aligned}\Big\|\frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}}\Big\|_{\mathcal{F}} &= \left[\sum_{h=1}^{L+1}\Big\|\frac{\partial \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})}{\partial \tilde{\boldsymbol{\theta}}^h} - \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}^h}\Big\|_{\mathcal{F}}^2\right]^{1/2} \\ &\leq O(\frac{1}{s\sqrt{L+1}}).\end{aligned}$$

When $s \geq 1$, we know that

$$s^{-L} \leq 1 \leq L^{L/(L+1)}.$$

Then, we have

$$\frac{1}{Ls^L} \leq \frac{1}{L} \leq L^{-1/(L+1)}.$$

Thus, we know $1/(Ls^L) = \min(1/(Ls^L), L^{-1/(L+1)})$ when $s \geq 1$.

For the case of $s \geq 1$, we can rewrite $\alpha \leq O(s^{-2L}/L) = O(qs^{-L})$, where $q = \min(1/(Ls^L), L^{-1/(L+1)})$, which completes the proof of Theorem 6.

Now, we prove Theorem 3 with $k = 2$, i.e. two-step gradient descent adaptation. We know that

$$\beta_1 = \alpha\nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\tilde{\boldsymbol{\theta}}})\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})^\intercal, \|\nabla_{f_{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|_{\mathcal{H}}^2 = \|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2.$$

Thus, we have

$$\begin{aligned}&\big|\beta_1 - \alpha\|\nabla_{f_{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|_{\mathcal{H}}^2\big| \\ =&\big|\alpha\nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\tilde{\boldsymbol{\theta}}})\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})^\intercal - \alpha\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})^\intercal\big| \\ =&\alpha\|\nabla_{\tilde{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\tilde{\boldsymbol{\theta}}}) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\| \\ =&\alpha\Big\|\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\Big\{\left[f_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_m) - \mathbf{y}_m\right]\frac{\partial f_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_m)}{\partial \tilde{\boldsymbol{\theta}}}^\intercal - \left[f_{\boldsymbol{\theta}}(\mathbf{x}_m) - \mathbf{y}_m\right]\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}}^\intercal\Big\}\Big\|\big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\big\| \\ =&\alpha\Big\|\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\Big\{\left[f_{\boldsymbol{\theta}}(\mathbf{x}_m) - \mathbf{y}_m + \triangle f_{\boldsymbol{\theta}}(\mathbf{x}_m)\right]\left[\frac{\partial f_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_m)}{\partial \tilde{\boldsymbol{\theta}}} + \triangle\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}}\right]^\intercal \\ &- \left[f_{\boldsymbol{\theta}}(\mathbf{x}_m) - \mathbf{y}_m\right]\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}}^\intercal\Big\}\Big\|\big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\big\| \\ =&\alpha\Big\|\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\Big\{\triangle f_{\boldsymbol{\theta}}(\mathbf{x}_m)\left[\frac{\partial f_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_m)}{\partial \tilde{\boldsymbol{\theta}}} + \triangle\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}}\right]^\intercal \\ &+ \left[f_{\boldsymbol{\theta}}(\mathbf{x}_m) - \mathbf{y}_m\right]\triangle\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}}^\intercal\Big\}\Big\|\big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\big\| \\ \leq&\alpha\left[O(\frac{1}{L})O(s^L\sqrt{L}) + O(\frac{1}{L})O(\frac{1}{s\sqrt{L}}) + O(\frac{1}{s\sqrt{L}})\right]\big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\big\| \\ \leq&\alpha\left[O(\frac{s^L}{\sqrt{L}}) + O(\frac{1}{s\sqrt{L}})\right]\big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\big\|, \text{ because } L \geq 1 \\ \leq&\left[O(\frac{qrs^L}{\sqrt{L}}) + O(\frac{qr}{s\sqrt{L}})\right]\big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\big\|, \text{ where } q = \min(1/(Ls^L), L^{-1/(L+1)}), r = \min(s^{-L}, s) \\ \leq&O(\frac{q}{\sqrt{L}})\big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\big\|.\end{aligned}$$

In the case of $d_y = 1$, we have

$$\Big\| \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h} \Big\|_{\mathcal{F}} = (b^h)^\intercal g^{h-1} \leq O(s^L),$$

which has already been shown in the proof of Lemma 9. Then, we have

$$\big\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_m(f_{\boldsymbol{\theta}})\big\| = O\Bigg(\sqrt{\sum_{h=1}^{L+1} \Big\| \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}^h} \Big\|^2}\Bigg) = O\Bigg(\sqrt{\sum_{h=1}^{L+1} \Big\| \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial W^h} \Big\|_{\mathcal{F}}^2}\Bigg) \leq O(s^L \sqrt{L+1}).$$

In the case of $d_y \geq 1$, the bound is simply scaled by a constant of $\sqrt{d_y}$.

Thus we have

$$\big|\beta_1 - \alpha \|\nabla_{f_{\boldsymbol{\theta}}} \mathcal{L}_m(f_{\boldsymbol{\theta}})\|_{\mathcal{H}}^2\big| \leq O(\frac{q}{\sqrt{L}}) \big\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_m(f_{\boldsymbol{\theta}})\big\| \leq O(q s^L) \leq O(\frac{1}{L})$$

because $q = \min(1/(Ls^L), L^{-1/(L+1)})$, which completes the proof for the case of $k = 2$.

For the case of $k > 2$, we only need to make sure that the bound on learning rate always holds. Fortunately, since k is a finite constant, according to what we have already showed in the proof of previous lemmas, every step of gradient descent will not change the spectral norm of the weight matrix too much: $\|\triangle W^h\|_2 \leq O(s^{-L}/L)$ for all h if $s \geq 1$, and $\|\triangle W^h\|_2 \leq O(q)$ for all h if $s < 1$, where $q = \min(1/(Ls^L), L^{-1/(L+1)})$. Thus, we may assume that the bound on learning rate always holds during the adaptation. Using triangle inequality to generalize the results from $k = 2$ to $k > 2$, i.e. for all $1 \leq i \leq k-1$, we have

$$\big|\beta_i - \alpha \|\nabla_{f_{\boldsymbol{\theta}}} \mathcal{L}_m(f_{\boldsymbol{\theta}})\|_{\mathcal{H}}^2\big| \leq O\Big(\frac{1}{L}\Big).$$

Recall that

$$\widetilde{\mathcal{E}}(\alpha, f_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathcal{T}_m} \left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - \alpha \|\nabla_{f_{\boldsymbol{\theta}}} \mathcal{L}_m(f_{\boldsymbol{\theta}})\|_{\mathcal{H}}^2\right]$$

and

$$\mathcal{M}_k = \mathbb{E}_{\mathcal{T}_m} \left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - \sum_{i=0}^{k-1} \beta_i\right],$$

where $\beta_i = \alpha \nabla_{\boldsymbol{\theta}_i} \mathcal{L}_m(f_{\boldsymbol{\theta}_i}) \nabla_{\boldsymbol{\theta}} \mathcal{L}_m(f_{\boldsymbol{\theta}})^\intercal$ and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}, \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \alpha \nabla_{\boldsymbol{\theta}_i} \mathcal{L}(f_{\boldsymbol{\theta}_i}, \mathcal{D}_m^{tr})$. The result is straightforward now. ■

E PROOF OF THEOREM 4

Theorem 4 *Let $f_{\boldsymbol{\theta}}$ be a convolutional neural network with $L - l$ convolutional layers and l fully-connected layers and with ReLU activation function, and d_x be the input dimension. Denote by W^h the parameter **vector** of the convolutional layer for $h \leq L - l$, and the weight **matrices** of the fully connected layers for $L - l + 1 < h \leq L + 1$. $\|\cdot\|_2$ means both the spectral norm of a matrix and the Euclidean norm of a vector. Define*

$$s_h = \begin{cases} \sqrt{d_x}\|W^h\|_2, & \text{if } h = 1, ..., L - l \\ \|W^h\|_2, & \text{if } L - l + 1 < h \leq L + 1 \end{cases}$$

and let $s = \max_h s_h$ and α be the learning rate of gradient descent. If $\alpha \leq O(qr)$ with $q = \min(1/(Ls^L), L^{-1/(L+1)})$ and $r = \min(s^{-L}, s)$, the following holds

$$|\mathcal{M}_k - \widetilde{\mathcal{E}}(k\alpha, f_{\boldsymbol{\theta}})| \leq O\Big(\frac{1}{L}\Big).$$

Proof We prove Theorem 4 by first transforming the convolutional neural network into an equivalent fully connected neural network and then applying Theorem 3.

First of all, we assume that there are c_h channels in h^{th} convolutional layer's output $g^h(\mathbf{x})$, where $h = 0, ..., L-l$. For fully-connected layers, define $c_{L-l} = ... = c_{L+1} = 1$. We may represent the dimensionality of input data by $\mathbf{x} \in R^{d_x c_0}$. Instead of using matrices, we represent the output of every convolutional layer by a $d_x c_h$ length vector $g^h = \left[g_1^h, g_2^h, ..., g_{d_x}^h\right]$, where every $g_i^h = \left[g_{i,1}^h, g_{i,2}^h, ..., g_{i,c_h}^h\right]$ is a c_h length vector contains value of different channels at the same position.

We assume that for every element $g_{i,j}^h$ of g_i^h, its value is completely determined by elements of set Q_i^{h-1}, where Q_i^{h-1} contains kc_{h-1} elements with fixed positions in g^{h-1} for a given i. In other words, every element of the output of a convolutional layer is determined by some elements with fixed positions from output of the previous layer. This is exactly how convolutional layer works in deep learning.

If we use $g_{Q_i^{h-1}}^{h-1}$ to represent the concatenation of $g_{a,b}^{h-1} \in Q_i^{h-1}$, then $g_{Q_i^{h-1}}^{h-1}$ is a kc_{h-1} length vector, where k is the kernel size. Then we have

$$g_i^h = \sigma(g_{Q_i^{h-1}}^{h-1} U_i^h)$$

where $U_{i,j}^h \in R^{kc_{h-1} \times c_h}$ is a $kc_{h-1} \times c_h$ matrix.

For notation simplicity, one can define a matrix $U^h \in R^{d_x c_{h-1} \times d_x c_h}$, where every column of U^h only has kc_{h-1} non-zero elements, and it satisfies

$$g^h = \sigma(g^{h-1} U^h)$$

By the property of convolutional layer, we know the following facts:

- One can represent U^h by $U^h = \left[V_1^h, V_2^h, ..., V_{d_x}^h\right]$ where $V_i^h \in R^{d_x c_{h-1} \times c_h}$ is sub-matrix of U^h;
- Every V_i^h contains the same set of elements as W^h, while these elements are located at different positions;
- Every V_i^h can be obtained by any other V_j^h by swapping rows;

Let's define $U^{L-l} = W^{L-l}, ..., U^{L+1} = W^{L+1}$ for the fully-connected layer and output layer. Then we can represent the neural network just as in Theorem 3 by $f_{\boldsymbol{\theta}}(\mathbf{x}) = \sigma(\sigma(...\sigma(\mathbf{x}\, U^1)...U^{L-1})U^L)U^{L+1}$, and $\mathbf{x} \in R^{d_x c_0}$.

Now let t_h be the spectral norm of U^h, and $t = \max_h t_h$. By Theorem 3, we know that we want $\alpha \leq O(qr)$, where $q = \min(1/(Ls^L), L^{-1/(L+1)}), r = \min(s^{-L}, s)$.

Because every V_i^h contains the same set of elements, we know that every V_i^h has the same Frobenius norm. Because every V_i^h can be obtained by any other V_j^h by swapping rows, we know that every V_i^h has the same rank.

We know that

$$\frac{1}{\sqrt{r}}\|V_1^h\|_{\mathcal{F}} \leq \|V_1^h\|_2 \leq \|U^h\|_2 \leq \|U^h\|_{\mathcal{F}} = \sqrt{d_x}\|V_1^h\|_{\mathcal{F}} = \sqrt{d_x}\|W^h\|_2$$

where $\|\cdot\|_{\mathcal{F}}$ denotes Frobenius norm, r denotes the rank of V_1^h. The last equality holds because matrix V_1^h and vector W^h have the same set of elements.

Let's define

$$s_h = \begin{cases} \sqrt{d_x}\|W^h\|_2, & \text{if } h = 1, ..., L-l \\ \|W^h\|_2, & \text{if } L-l+1 < h \leq L+1 \end{cases}$$

and $s = \max_h s_h$.

From above we know that $t_h = \Theta(s_h)$, because $s_h/\sqrt{d_x r} \leq t_h \leq s_h$. So we also have $t = \Theta(s)$. Then the conclusion is straightforward. ■

F REVISION OF THEOREM 3 AND THEOREM 4 IN CLASSIFICATION CASE

We now show how to obtain similar results of Theorem 3 and Theorem 4 in classification problem, where cross-entropy loss is used instead of squared loss. We need two more restrictions in the classification case:

1. There exist matrix A and B such that $g^L A \leq \text{softmax}(g^L W^{L+1}) \leq g^L B$ for all data points, where softmax is the softmax operation at the last layer.
2. For any data point $\mathbf{x}$ whose belongs to c^{th} class, there exists a constant $\epsilon > 0$ such that $f_{\boldsymbol{\theta},c}(\mathbf{x}) \geq \epsilon$, i.e. the output of neural network has a lower bound on the true class position.

The proof is actually similar to the proof in regression case. We briefly talk about the differences here.

Firstly, in the classification case, softmax function is used at the last layer. By the first restriction, we can get rid of softmax function by introducing new matrices, which further leads to bound of the learning rate as in regression case.

Secondly, if the loss function is the cross-entropy loss, we have:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_m(f_{\boldsymbol{\theta}}) = \mathbb{E}_{(\mathbf{x}_m, \mathbf{y}_m)} \left[\frac{1}{f_{\boldsymbol{\theta},c_m}(\mathbf{x}_m)} \frac{\partial f_{\boldsymbol{\theta},c_m}(\mathbf{x}_m)}{\partial \boldsymbol{\theta}} \right]$$

where c_m denotes the class of $\mathbf{x}_m$, e.g. if $\mathbf{x}_m$ belongs to the third class, then $c_m = 3$. $f_{\boldsymbol{\theta},c_m}(\mathbf{x}_m)$ denotes the c_m^{th} dimensional element of $f_{\boldsymbol{\theta}}(\mathbf{x}_m)$. We want a lower bound of $f_{\boldsymbol{\theta},c}(\mathbf{x})$ exists, so that the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}_m(f_{\boldsymbol{\theta}})$ can be further bounded.

Then we can prove similar theorems just follow the steps in regression case.

G PROOF OF THEOREM 5

Theorem 5 *Let $f_{\boldsymbol{\theta}}$ be a neural network with L hidden layers, with each layer being either fully-connected or convolutional. Assume that $\|\mathcal{L}\|_\infty < \infty$. Then, $error(T) = |\widetilde{\mathcal{E}}(T, f_{\boldsymbol{\theta}}) - \overline{\mathcal{E}}(T, f_{\boldsymbol{\theta}})|$ is a non-decreasing function of T. Furthermore, for arbitrary $T > 0$ we have:*

$$error(T) \leq O\big(T^{2L+3}\big).$$

Proof Recall that $\overline{\mathcal{E}}(t, f_{\boldsymbol{\theta}})$ is defined based on $f^t_{m,\boldsymbol{\theta}}$, which is the resulting function whose parameters evolve according to the gradient flow $\dfrac{\mathrm{d}\boldsymbol{\theta}^t_m}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}^t_m} \mathcal{L}(f^t_{m,\boldsymbol{\theta}}, \mathcal{D}^{tr}_m)$.

We actually have the following (Santambrogio, 2016):

$$\|\triangle \boldsymbol{\theta}\| = \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^t\| \leq O(\sqrt{t}).$$

For simplicity and clearness, we use $\triangle$ to denote the change of any vectors and matrices. Thus, we know that

$$\|\triangle W^h\|_2 \leq \|\triangle W^h\|_{\mathcal{F}} \leq \|\triangle \boldsymbol{\theta}\| \leq O(\sqrt{t}).$$

Just like the proofs of Lemma 7, Lemma 8 and Lemma 9, we show that

$$\|\triangle g^h\| \leq O(t^{h/2}), \|\triangle b^h\| \leq O(t^{(L-h+1)/2}), \left\| \triangle \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}} \right\|_{\mathcal{F}} \leq O(t^{(L+1)/2}\sqrt{L+1})$$

by mathematical inductions; we skip the details here. Note that different from some previous theorem, here we focus on time t, and thus hide the effect of the spectral norms by treating them as constants.

Then, we have

$$
\begin{aligned}
&\left\|\triangle\Big(\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\Big)\right\| \\
=&\|\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\| \\
=&\left\|\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\Big\{\left[f^t_{m,\boldsymbol{\theta}}(\mathbf{x}_m)-\mathbf{y}_m\right]\frac{\partial f^t_{m,\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial\boldsymbol{\theta}^t}^{\intercal} - \left[f_{\boldsymbol{\theta}}(\mathbf{x}_m)-\mathbf{y}_m\right]\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial\boldsymbol{\theta}}^{\intercal}\Big\}\right\| \\
=&\left\|\mathbb{E}_{(\mathbf{x}_m,\mathbf{y}_m)}\Big\{\triangle f_{\boldsymbol{\theta}}(\mathbf{x}_m)\left[\frac{\partial f^t_{m,\boldsymbol{\theta}}(\mathbf{x}_m)(\mathbf{x}_m)}{\partial\boldsymbol{\theta}^t}+\triangle\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial\boldsymbol{\theta}}\right]^{\intercal} + \left[f_{\boldsymbol{\theta}}(\mathbf{x}_m)-\mathbf{y}_m\right]\triangle\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_m)}{\partial\boldsymbol{\theta}}^{\intercal}\Big\}\right\| \\
\leq&O(t^{L+1}\sqrt{L+1}).
\end{aligned}
$$

Recall that:

$$
\begin{aligned}
\overline{\mathcal{E}}(T,f_{\boldsymbol{\theta}}) &= \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f^T_{m,\boldsymbol{\theta}})\right] \\
&= \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) + \int_0^T \nabla_t\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})\mathrm{d}t\right] \\
&= \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) + \int_0^T \frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})\mathrm{d}t\right] \\
&= \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - \int_0^T \left\|\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})\right\|^2\mathrm{d}t\right]
\end{aligned}
$$

and

$$
\widetilde{\mathcal{E}}(T,f_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - T\|\nabla_{f_{\boldsymbol{\theta}}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2_{\mathcal{H}}\right] = \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - T\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2\right].
$$

Because

$$
\begin{aligned}
&\widetilde{\mathcal{E}}(T,f_{\boldsymbol{\theta}}) - \overline{\mathcal{E}}(T,f_{\boldsymbol{\theta}}) \\
=&\int_0^T \left\|\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})\right\|^2\mathrm{d}t - T\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2 \\
=&\int_0^T \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}}) + \triangle\Big(\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\Big)\right\|^2\mathrm{d}t - T\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\|^2 \\
=&\int_0^T 2\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\triangle\Big(\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\Big)^{\intercal} + \left\|\triangle\Big(\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\Big)\right\|^2\mathrm{d}t,
\end{aligned}
$$

we have

$$
error(T) = |\widetilde{\mathcal{E}}(T,f_{\boldsymbol{\theta}}) - \overline{\mathcal{E}}(T,f_{\boldsymbol{\theta}})| \leq O\Big(\frac{L+1}{2L+3}T^{2L+3}\Big) = O\big(T^{2L+3}\big)
$$

by simple calculation.

On the other hand, observe that

$$
\bar{\mathcal{E}}(T,f_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathcal{T}_m}\left[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - \int_0^T \left\|\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})\right\|^2_{\mathcal{H}}dt\right],
$$

$$
\tilde{\mathcal{E}}(T,f_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathcal{T}_m}[\mathcal{L}_m(f_{\boldsymbol{\theta}}) - T\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\right\|^2_{\mathcal{H}}].
$$

We let

$$
G(\tau) = \int_0^\tau \left\|\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})\right\|^2 dt,
$$

and assume that $\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\theta})$ is continuous at $t=0$. Then, we have $G'(\tau) = \|\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{\boldsymbol{\theta}})\|^2$.

$$
\begin{aligned}
\left\|\overline{\mathcal{E}}(T,f_{\boldsymbol{\theta}}) - \tilde{\mathcal{E}}(T,f_{\boldsymbol{\theta}})\right\| =&\left\|\mathbb{E}_{\mathcal{T}_m}\left[\int_0^T \left\|\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})\right\|^2 dt - T\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_m(f_{\boldsymbol{\theta}})\right\|^2_{\mathcal{H}}\right]\right\| \\
=&\left\|\mathbb{E}_{\mathcal{T}_m}(G(T) - T\cdot G'(0))\right\|,
\end{aligned}
$$

where $TG'(0) = G(0) + TG'(0)$ (note that $G(0) = 0$) is a first order approximation to $G(T)$ at $\tau = 0$. When $T = 1$, $G(T) - TG'(0)$ can be taken as a local truncation error (i.e., the error that occurs in one step of a numerical approximation). When T increases, the difference is no better than the global truncation error (in T steps):

$$\begin{aligned}\left\|G(T) - \sum_{i=0}^{T}(i-(i-1))G'(i)\right\| &= \left\|\sum_{i=0}^{T}\int_{i}^{i+1}\left\|\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})\right\|^2 - \left\|\nabla_{\boldsymbol{\theta}^i}\mathcal{L}_m(f^{t=i}_{m,\boldsymbol{\theta}})\right\|^2 dt\right\| \\ &\approx \left\|\sum_{i=0}^{T}\int_{i}^{i+1} 2\cdot\triangle^i_t\mathcal{L}_m(f_{m,\boldsymbol{\theta}})\cdot\nabla\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})dt\right\|,\end{aligned}$$

where $\triangle^i_t\mathcal{L}_m(f_{m,\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}^i}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})$ as shown previously , i is the i-th time step, and $G'(i)$ is the gradient of G at time step i. Now we can see that $\left\|\overline{\mathcal{E}}(T, f_{\boldsymbol{\theta}}) - \tilde{\mathcal{E}}(T, f_{\boldsymbol{\theta}})\right\|$ highly relates to the difference between $\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})$ at different time steps (i.e. $\triangle^i_t\mathcal{L}_m(f_{m,\boldsymbol{\theta}})$), $\nabla_{\boldsymbol{\theta}^t}\mathcal{L}_m(f^t_{m,\boldsymbol{\theta}})$ and T . The first two terms relate to how flat or sharp the hyperplane of $\mathcal{L}_m(f_{m,\boldsymbol{\theta}})$ is near $t = 0$. We can wrap it as a constant $C_0(\mathcal{L}, t = 0)$. Then, the error is at least $C_0(\mathcal{L}, t = 0)\cdot O(T)$. For the hyperplane smooth enough, we can further get a first order approximation of $\triangle^i_t\mathcal{L}_m(f_{m,\boldsymbol{\theta}})$ and yield $C(\mathcal{L}, t = 0)O(T^2)$, where $C(\mathcal{L}, t = 0)$ can be analogized as the second order derivative of $\mathcal{L}$. ■

H Some Experimental Details

H.1 Implementation of Classification for Meta-RKHS-II

As we mentioned earlier, our proposed energy functional with closed form adaptation can not be directly applied to classification problem. We handle this challenge following Arora et al. (2019). For a d_y class classification problem, every data $\mathbf{x}$ is associated with a R^{d_y} one-hot vector $\mathbf{y}$ as its label. For C classes classification problem, its encoding is C dimensional vector and we use $-1/C$ and $(C-1)/C$ as its correct and incorrect entries encoding. In the prediction, Y^{tr} is replaced by the encoding of training data. $f_{\boldsymbol{\theta}}(\mathbf{x})$ is replaced by $f_{\boldsymbol{\theta}}(\mathbf{x})^\intercal[1, ..., 1] \in R^{n\times d_y}$ for dimension consistency. During the testing time, we compute the encoding of the test data point, and choose the position with largest value as its predicted class.

I Extra Experimental Results

I.1 Comparison with RBF kernel

One interesting question is, without introducing extra model components or networks, what will the results of other kernel be? We provide the results of using RBF (Gaussian) kernel here: 42.1 ± 1.9 (5-way 1-shot) and 54.9 ± 1.1 (5-way 5-shot) on Mini-ImageNet, 32.4 ± 2.0 (5-way 1-shot) and 38.2 ± 0.9 (5-way 5-shot) on FC-100, which are worse than the NTK based Meta-RKHS-II, showing the superiority of using NTK.

I.2 More Results on out-of-distribution Generalization

We provide some more results on out-of-distribution generalization experiments here. From the results we can find that the proposed methods is more robust and can generalize to different datasets better.

Table 5: Meta testing on different out-of-distribution datasets with model trained on FC-100.

	5 WAY 1 SHOT		5 WAY 5 SHOT	
ALGORITHM	CUB	VGG FLOWER	CUB	VGG FLOWER
MAML	$31.58 \pm 1.89\%$	$50.82 \pm 1.94\%$	$41.72 \pm 1.29\%$	$65.19 \pm 1.36\%$
FOMAML	$32.34 \pm 1.57\%$	$49.90 \pm 1.78\%$	$41.96 \pm 1.53\%$	$66.87 \pm 1.45\%$
REPTILE	$33.56 \pm 1.40\%$	$46.77 \pm 1.81\%$	$42.79 \pm 1.38\%$	$67.97 \pm 0.71\%$
IMAML	$32.49 \pm 1.52\%$	$49.96 \pm 1.98\%$	$38.92 \pm 1.62\%$	$59.80 \pm 1.82\%$
BAYESIAN TAML(SOTA)	$31.82 \pm 0.49\%$	$49.58 \pm 0.55\%$	$43.97 \pm 0.57\%$	$67.36 \pm 0.53\%$
META-RKHS-I	$34.12 \pm 1.34\%$	$48.81 \pm 1.89\%$	$43.31 \pm 1.43\%$	$69.02 \pm 0.62\%$
META-RKHS-II	$\mathbf{36.35 \pm 1.07\%}$	$\mathbf{59.75 \pm 1.23\%}$	$\mathbf{49.92 \pm 0.68\%}$	$\mathbf{76.32 \pm 0.58\%}$

I.3 More Results on Adversarial Attack

We now show some more extra results on adversarial attack in the following figures. Consistent to the results in main text, we can find that our proposed methods are more robust to adversarial attacks.

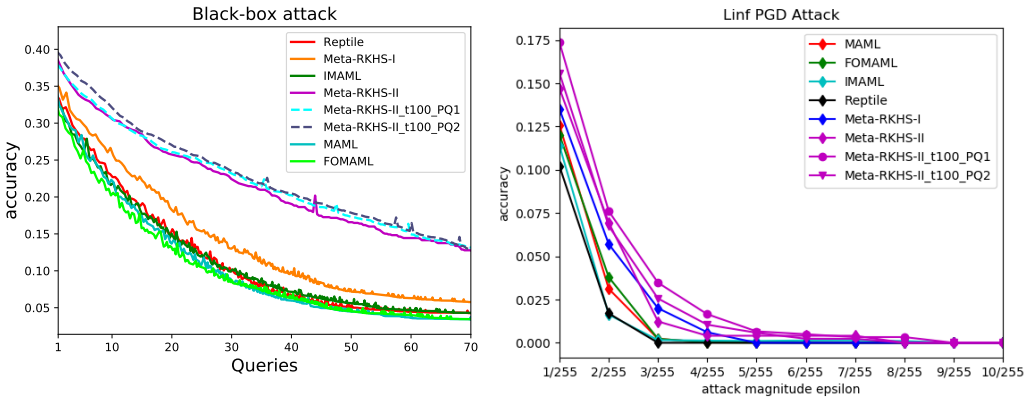
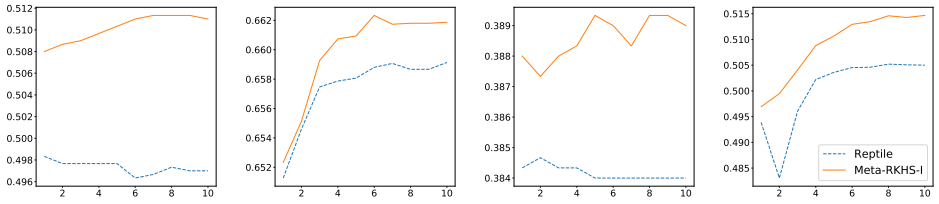


Figure 6: FC-100 5-way 5-shot Black-box attacks (left) and 5-way 1-shot PGD ℓ_∞ norm attack (right).

I.4 Impact of Gradient Norm in Meta-RKHS-I

In this experiment, we compare between our proposed Meta-RKHS-I and Reptile. We evaluate the trained models with different adaptation steps in testing-time. The comparison is shown in Figure 7. As we can see, our Meta-RKHS-I always gets better results than Reptile, which supports our idea that the learned function should be close to task-specific optimal and have large functional gradient norm. These two conditions together lead to the ability of fast adaptation.



(a) Mini-ImageNet, 5 Way 1 Shots (b) Mini-ImageNet, 5 Way 5 Shots (c) FC-100, 5 Way 1 Shots (d) FC-100, 5 Way 5 Shots

Figure 7: Reptile (dashed) vs. Meta-RKHS-I (solid) with different testing adaptation steps (x-axis).

I.5 Impact of Network Architecture for Different Meta-Learning Models

In this section, we compare different meta-learning models with feature channels of 100 and 200 of the CNN network structure with 4 or 5 CNN layers respectively.

Table 6: Few-shot classification results on Mini-ImageNet with different number of feature channels of 4 convolution layers.

	100		200	
ALGORITHM	5 WAY 1 SHOT	5 WAY 5 SHOTS	5 WAY 1 SHOT	5 WAY 5 SHOTS
MAML	$49.50 \pm 1.58\%$	$64.31 \pm 1.07\%$	$48.91 \pm 1.69\%$	$63.96 \pm 0.82\%$
FOMAML	$48.69 \pm 1.62\%$	$63.73 \pm 0.76\%$	$48.55 \pm 1.86\%$	$63.18 \pm 0.96\%$
IMAML	$49.30 \pm 1.94\%$	$62.89 \pm 0.95\%$	$48.23 \pm 1.58\%$	$62.25 \pm 0.83\%$
REPTILE	$50.20 \pm 1.69\%$	$64.12 \pm 0.92\%$	$48.72 \pm 1.97\%$	$63.67 \pm 0.79\%$
META-RKHS-I	$51.23 \pm 1.79\%$	$66.69 \pm 0.73\%$	$\mathbf{51.54 \pm 1.64}\%$	$\mathbf{65.92 \pm 0.92}\%$
META-RKHS-II	$\mathbf{51.37 \pm 2.31}\%$	$\mathbf{66.97 \pm 0.98}\%$	$50.96 \pm 2.15\%$	$65.21 \pm 0.87\%$

Table 7: Few-shot classification results on Mini-ImageNet with different number of feature channels of 5 convolution layers.

	100		200	
ALGORITHM	5 WAY 1 SHOT	5 WAY 5 SHOTS	5 WAY 1 SHOT	5 WAY 5 SHOTS
MAML	$49.87 \pm 1.65\%$	$65.78 \pm 1.18\%$	$48.62 \pm 1.82\%$	$63.25 \pm 0.75\%$
FOMAML	$48.93 \pm 1.71\%$	$64.37 \pm 0.80\%$	$48.27 \pm 1.74\%$	$62.95 \pm 0.83\%$
IMAML	$48.03 \pm 1.76\%$	$62.15 \pm 0.83\%$	$47.52 \pm 1.73\%$	$61.77 \pm 0.89\%$
REPTILE	$50.62 \pm 1.83\%$	$64.53 \pm 0.97\%$	$49.33 \pm 1.89\%$	$63.26 \pm 0.70\%$
META-RKHS-I	$\mathbf{52.45 \pm 1.88}\%$	$66.07 \pm 0.69\%$	$\mathbf{51.37 \pm 1.92}\%$	$\mathbf{65.39 \pm 0.98}\%$
META-RKHS-II	$50.92 \pm 2.16\%$	$\mathbf{66.45 \pm 0.91}\%$	$50.43 \pm 2.42\%$	$64.17 \pm 1.06\%$