Overview of the Appendix

The appendix is divided into the following sections,

- Sec. A: Describes the neural network setup and parameterization.
- Sec. B: Presents theoretical results and proof regarding the **global convergence** of gradient-based meta-learning.
- Sec. C: Derives the expression of MAML output.
- Sec. D: Derives the equivalence between MAML and kernel regressions.
- Sec. E: Presents more details of **experiments** in Sec. 6.

A. Neural Network Setup

In this paper, we consider a fully-connected feed-forward network with L hidden layers. Each hidden layer has width l_i, for $i = 1, ..., L$. The readout layer (i.e. output layer) has width $l_{L+1} = k$. At each layer i, for arbitrary input $x \in \mathbb{R}^d$, we denote the pre-activation and post-activation functions by $h^i(x), z^i(x) \in \mathbb{R}^{l_i}$. The relations between layers in this network are

$$\begin{cases} h^{i+1} &= z^i W^{i+1} + b^{i+1} \\ z^{i+1} &= \sigma\left(h^{i+1}\right) \end{cases} \text{ and } \begin{cases} W^i_{\mu,\nu} &= \omega^i_{\mu\nu} \sim \mathcal{N}(0, \frac{\sigma_\omega}{\sqrt{l_i}}) \\ b^i_\nu &= \beta^i_\nu \sim \mathcal{N}(0, \sigma_b) \end{cases}, \tag{16}$$

where $W^{i+1} \in \mathbb{R}^{l_i \times l_{i+1}}$ and $b^{i+1} \in \mathbb{R}^{l_{i+1}}$ are the weight and bias of the layer, $\omega^l_{\mu\nu}$ and b^l_ν are trainable variables drawn i.i.d. from zero-mean Gaussian distributions at initialization (i.e., $\frac{\sigma^2_\omega}{l_i}$ and σ^2_b are variances for weight and bias, and σ is a point-wise activation function.

B. Proof of Global Convergence for Gradient-Based Meta-Learning with Deep Neural Networks

In this section, we will prove the global convergence for gradient-based meta-learning with over-parameterized neural nets. To prove the global convergence theorem, we introduce several key lemmas first, i.e., Lemma 1, 2, 3. Specifically, the subsections of this section are formulated as follows.

- Sec. B.1: Present several helper lemmas with proof.
- Sec. B.2: Provides the proof of Lemma 1.
- Sec. B.3: Provides the proof of Lemma 2.
- Sec. B.4: Provides the proof of Lemma 3.
- Sec. B.5: Proves the global convergence theorem for MAML, i.e., Theorem 3 (restated version of Theorem 1).

Notice that in this section, we consider the standard parameterization scheme of neural networks shown in (16).

The global convergence theorem, Theorem 1, depends on several assumptions and lemmas. The assumptions are listed below. After that, we present the lemmas and the global convergence theorem, with proofs in Appendix B.1,B.3,B.4 and B.5. For Corollary 3.1, we append its proof to Appendix C.

Assumption 1 (Bounded Input Norm). *$\forall X \in \mathcal{X}$, for any sample $x \in X$, $\|x\|_2 \leq 1$. Similarly, $\forall X' \in \mathcal{X}'$, for any sample $x' \in X'$, $\|x'\|_2 \leq 1$. (This is equivalent to a input normalization operation, which is common in data preprocessing.)*

Assumption 2 (Non-Degeneracy). *The meta-training set $(\mathcal{X}, \mathcal{Y})$ and the meta-test set $(\mathcal{X}', \mathcal{Y}')$ are both contained in some compact set. Also, $\mathcal{X}$ and $\mathcal{X}'$ are both non-degenerate, i.e. $\forall X, \widetilde{X} \in \mathcal{X}$, $X \neq \widetilde{X}$, and $\forall X', \widetilde{X}' \in \mathcal{X}'$, $X' \neq \widetilde{X}'$.*

Assumption 3 (Same Width for Hidden Layers). *All hidden layers share the same width, l, i.e., $l_1 = l_2 = \cdots = l_L = l$.*

Assumption 4 (Full-Rank). *The kernel Φ defined in Lemma 3 is full-rank.*

These assumptions are common, and one can find similar counterparts of them in the literature for supervised learning [4, 37]. In particular, notice that Assumption 3 is just for simplicity purpose without loss generality. In fact, one can directly set $l = \min_{i \in [L]} l_i$ as the minimum width across hidden layers, and all theoretical results in this paper still hold true [37].

As defined in the main text, θ is used to represent the neural net parameters. For convenience, we define some short-hand notations:

$$f_t(\cdot) = f_{\theta_t}(\cdot) \tag{17}$$

$$F_t(\cdot) = F_{\theta_t}(\cdot) \tag{18}$$

$$f(\theta) = f_\theta(\mathcal{X}) = ((f_\theta(X_i))_{i=1}^N \tag{19}$$

$$F(\theta) = F_\theta(\mathcal{X}, \mathcal{X}', \mathcal{Y}') = ((F_\theta(X_i, X_i', Y_i'))_{i=1}^N \tag{20}$$

$$g(\theta) = F_\theta(\mathcal{X}, \mathcal{X}', \mathcal{Y}') - \mathcal{Y} \tag{21}$$

$$J(\theta) = \nabla_\theta F(\theta) = \nabla_\theta F_\theta(\mathcal{X}, \mathcal{X}', \mathcal{Y}') \tag{22}$$

and

$$\mathcal{L}(\theta_t) = \ell(F(\theta_t), \mathcal{Y}) = \frac{1}{2}\|g(\theta_t)\|_2^2 \tag{23}$$

$$\hat{\Phi}_t = \frac{1}{l} \nabla F_{\theta_t}(\mathcal{X}, \mathcal{X}', \mathcal{Y}') \nabla F_{\theta_t}(\mathcal{X}, \mathcal{X}', \mathcal{Y}') = \frac{1}{l} J(\theta) J(\theta)^\top \tag{24}$$

where we use the ℓ_2 loss function $\ell(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|_2^2$ in the definition of training loss $\mathcal{L}(\theta_t)$ in (23), and the $\hat{\Phi}_t$ in (24) is based on the definition[12] of $\hat{\Phi}_\theta(\cdot, \star)$ in Sec. 4.1.

Below, Lemma 1 proves the Jacobian J is locally Lipschitz, Lemma 2 proves the training loss at initialization is bounded, and Lemma 3 proves $\hat{\Phi}_0$ converges in probability to a deterministic kernel matrix with bounded positive eigenvalues. Finally, with these lemmas, we can prove the global convergence of MAML in Theorem 3.

Lemma 1 (**Local Lipschitzness of Jacobian**). *For arbitraily small $\delta > 0$, then there exists $K > 0$ and $l^* > 0$ such that: $\forall\ C > 0$ and $l > l^*$, the following inequalities hold true with probability at least $1 - \delta$ over random initialization,*

$$\forall \theta,\ \bar{\theta} \in B(\theta_0, Cl^{-\frac{1}{2}}), \begin{cases} \frac{1}{\sqrt{l}}\|J(\theta) - J(\bar{\theta})\|_F & \leq K\|\theta - \bar{\theta}\|_2 \\ \\ \frac{1}{\sqrt{l}}\|J(\theta)\|_F & \leq K \end{cases} \tag{25}$$

where B is a neighborhood defined as

$$B(\theta_0, R) := \{\theta : \|\theta - \theta_0\|_2 < R\}. \tag{26}$$

Proof. See Appendix B.1. □

Lemma 2 (**Bounded Initial Loss**). *For arbitrarily small $\delta_0 > 0$, there are constants $R_0 > 0$ and $l^* > 0$ such that as long as the width $l > l^*$, with probability at least $(1 - \delta_0)$ over random initialization,*

$$\|g(\theta_0)\|_2 = \|F_{\theta_0}(\mathcal{X}, \mathcal{X}', \mathcal{Y}') - \mathcal{Y}\|_2 \leq R_0, \tag{27}$$

which is also equivalent to

$$\mathcal{L}(\theta_0) = \frac{1}{2}\|g(\theta_0)\|_2^2 \leq \frac{1}{2}R_0^2.$$

Proof. See Appendix B.3. □

Lemma 3 (**Kernel Convergence**). *Suppose the learning rates η and λ suffiently small. As the network width l approaches infinity, $\hat{\Phi}_0 = J(\theta_0)J(\theta_0)^\top$ converges in probability to a deterministic kernel matrix Φ (i.e., $\Phi = \lim_{l\to\infty} \hat{\Phi}_0$), which is independent of θ_0 and can be analytically calculated. Furthermore, the eigenvalues of Φ is bounded as, $0 < \sigma_{min}(\Phi) \leq \sigma_{max}(\Phi) < \infty$.*

[12] There is a typo in the definition of $\hat{\Phi}_\theta(\cdot, \star)$ in Sec. 4.1: a missing factor $\frac{1}{l}$. The correct definition should be $\hat{\Phi}_\theta(\cdot, \star) = \frac{1}{l}\nabla_\theta F_\theta(\cdot)\nabla_\theta F_\theta(\star)^\top$. Similarly, the definition of Φ in Theorem 1 also missis this factor: the correct version is $\Phi = \frac{1}{l}\lim_{l\to\infty} J(\theta_0)J(\theta_0)^\top$

Proof. See Appendix B.4. □

Note the update rule of gradient descent on θ_t with learning rate η can be expressed as

$$\theta_{t+1} = \theta_t - \eta J(\theta_t)^\top g(\theta_t). \tag{28}$$

The following theorem proves the global convergence of MAML under the update rule of gradient descent.

Theorem 3 (**Global Convergence** *(Theorem 1 restated)*)**.** *Denote $\sigma_{min} = \sigma_{min}(\Phi)$ and $\sigma_{max} = \sigma_{max}(\Phi)$. For any $\delta_0 > 0$ and $\eta_0 < \frac{2}{\sigma_{max}+\sigma_{min}}$, there exist $R_0 > 0$, $\Lambda \in \mathbb{N}$, $K > 1$, and $\lambda_0 > 0$, such that: for width $l \geq \Lambda$, running gradient descent with learning rates $\eta = \frac{\eta_0}{l}$ and $\lambda < \frac{\lambda_0}{l}$ over random initialization, the following inequalities hold true with probability at least $(1 - \delta_0)$:*

$$\sum_{j=1}^{t} \|\theta_j - \theta_{j-1}\|_2 \leq \frac{3KR_0}{\sigma_{min}} l^{-\frac{1}{2}} \tag{29}$$

$$\sup_t \|\hat{\Phi}_0 - \hat{\Phi}_t\|_F \leq \frac{6K^3R_0}{\sigma_{\min}} l^{-\frac{1}{2}} \tag{30}$$

and

$$g(\theta_t) = \|F(\theta_t) - \mathcal{Y}\|_2 \leq \left(1 - \frac{\eta_0 \sigma_{\min}}{3}\right)^t R_0 \, , \tag{31}$$

which leads to

$$\mathcal{L}(\theta_t) = \frac{1}{2}\|F(\theta_t) - \mathcal{Y}\|_2^2 \leq \left(1 - \frac{\eta_0 \sigma_{\min}}{3}\right)^{2t} \frac{R_0^2}{2} \, , \tag{32}$$

indicating the training loss converges to zero at a linear rate.

Proof. See Appendix B.5. □

In the results of Theorem 3 above, (29) considers the optimization trajectory of network parameters, and show the parameters move locally during training. (30) indicates the kernel matrix $\hat{\Phi}_t$ changes slowly. Finally, (32) demonstrates that the training loss of MAML decays exponentially to zero as the training time evolves, indicating convergence to global optima at a linear rate.

B.1. Helper Lemmas

Lemma 4. *As the width $l \to \infty$, for any vector $\mathbf{a} \in \mathbb{R}^{m \times 1}$ that $\|\mathbf{a}\|_F \leq C$ with some constant $C > 0$, we have*

$$\|\nabla_\theta \hat{\Theta}_\theta(x, X') \cdot \mathbf{a}\|_F \to 0 \tag{33}$$

where θ is randomly intialized parameters.

Proof. Notice that

$$\hat{\Theta}_\theta(x, X') = \frac{1}{l} \underbrace{\nabla_\theta f_\theta(\overbrace{x}^{\in\mathbb{R}^d})}_{\in\mathbb{R}^{1\times D}} \cdot \underbrace{\nabla_\theta f_\theta(\overbrace{X'}^{\in\mathbb{R}^{m\times d}})^\top}_{\in\mathbb{R}^{D\times m}} \in \mathbb{R}^{1\times m} \tag{34}$$

with gradient as

$$\nabla_\theta \hat{\Theta}_\theta(x, X') = \frac{1}{l} \underbrace{\nabla_\theta^2 f_\theta(x)}_{\in\mathbb{R}^{1\times D\times D}} \cdot \underbrace{\nabla_\theta f_\theta(X')^\top}_{\in\mathbb{R}^{D\times m}} + \frac{1}{l} \underbrace{\nabla_\theta f_\theta(x)}_{\in\mathbb{R}^{1\times D}} \cdot \underbrace{\nabla_\theta^2 f_\theta(X')^\top}_{\in\mathbb{R}^{D\times m\times D}} \in \mathbb{R}^{1\times m\times D} \tag{35}$$

where we apply a *dot product* in the *first two dimensions* of 3-tensors and matrices to obtain matrices.

Then, it is obvious that our goal is to bound the Frobenius Norm of

$$\nabla_\theta \hat{\Theta}_\theta(x, X') \cdot \mathbf{a} = \left(\frac{1}{l}\nabla_\theta^2 f_\theta(x) \cdot \nabla_\theta f_\theta(X')^\top\right) \cdot \mathbf{a} + \left(\frac{1}{l}\nabla_\theta f_\theta(x) \cdot \nabla_\theta^2 f_\theta(X')^\top\right) \cdot \mathbf{a} \tag{36}$$

Below, we prove that as the width $l \to \infty$, the first and second terms of (36) both have vanishing Frobenius norms, which finally leads to the proof of (33).

- *First Term of (36).* Obviously, reshaping $\nabla^2_\theta f_\theta(x) \in \mathbb{R}^{1\times D\times D}$ as a $\mathbb{R}^{D\times D}$ matrix does not change the Frobenius norm $\|\frac{1}{l}\nabla^2_\theta f_\theta(x) \cdot \underbrace{\nabla_\theta f_\theta(X')^\top}_{\in\mathbb{R}^{D\times m}}\|_F$ (in other words, $\|\frac{1}{l}\underbrace{\nabla^2_\theta f_\theta(x)}_{\in\mathbb{R}^{1\times D\times D}} \cdot \underbrace{\nabla_\theta f_\theta(X')^\top}_{\in\mathbb{R}^{D\times m}}\|_F = \|\frac{1}{l}\underbrace{\nabla^2_\theta f_\theta(x)}_{\in\mathbb{R}^{D\times D}} \cdot \underbrace{\nabla_\theta f_\theta(X')^\top}_{\in\mathbb{R}^{D\times m}}\|_F$).

  By combining the following three facts,

  1. $\|\frac{1}{\sqrt{l}}\underbrace{\nabla^2_\theta f_\theta(x)}_{\in\mathbb{R}^{D\times D}}\|_{op} \to 0$ indicated by [27],
  2. the matrix algebraic fact $\|HB\|_F \leq \|H\|_{op}\|B\|_F$,
  3. the bound $\|\frac{1}{\sqrt{l}}\nabla_\theta f_\theta(\cdot)\|_F < constant$ from [37],

  one can easily show that the first term of (35) has vanishing Frobenius norm, i.e.,

$$\|\frac{1}{l}\nabla^2_\theta f_\theta(x) \cdot \nabla_\theta f_\theta(X')^\top\|_F \to 0 \tag{37}$$

  Then, obviously,

$$\| \left(\frac{1}{l}\nabla^2_\theta f_\theta(x) \cdot \nabla_\theta f_\theta(X')^\top\right) \cdot \mathbf{a}\|_F \leq \|\frac{1}{l}\nabla^2_\theta f_\theta(x) \cdot \nabla_\theta f_\theta(X')^\top\|_F \|\mathbf{a}\|_F \to 0 \tag{38}$$

- *Second Term of (36).* From [27], we know that

$$\| \underbrace{\frac{1}{\sqrt{l}}\nabla^2_\theta f_\theta(X')^\top}_{\in\mathbb{R}^{D\times m\times D}} \cdot \underbrace{\mathbf{a}}_{\in\mathbb{R}^{m\times 1}} \|_{op} \to 0 \tag{39}$$

  Then, similar to the derivation of (37), we have

$$\| \left(\frac{1}{l}\nabla_\theta f_\theta(x) \cdot \nabla^2_\theta f_\theta(X')^\top\right) \cdot \mathbf{a}\|_F \leq \overbrace{\|\frac{1}{\sqrt{l}}\nabla_\theta f_\theta(x)\|_F}^{\leq constant} \cdot \overbrace{\|\nabla^2_\theta f_\theta(X')^\top \cdot \mathbf{a}\|_{op}}^{\to 0} \to 0 \tag{40}$$

- Finally, combining (38) and (40), we obtain (33) by

$$\begin{aligned} \|\nabla_\theta \hat{\Theta}_\theta(x, X') \cdot \mathbf{a}\|_F &\leq \| \left(\frac{1}{l}\nabla^2_\theta f_\theta(x) \cdot \nabla_\theta f_\theta(X')^\top\right) \cdot \mathbf{a}\|_F \\ &\quad + \| \left(\frac{1}{l}\nabla_\theta f_\theta(x) \cdot \nabla^2_\theta f_\theta(X')^\top\right) \cdot \mathbf{a}\|_F \\ &\to 0 \end{aligned} \tag{41}$$

□

Lemma 5. *Given any task $\mathcal{T} = (X, Y, X', Y')$ and randomly initialized parameters θ, as the width $l \to \infty$, for any $x \in X$, where $x \in \mathbb{R}^d$ and $X \in \mathbb{R}^{n\times d}$, we have*

$$\|\nabla_\theta \left(\hat{\Theta}_\theta(x, X')\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})\right)(f_\theta(X') - Y')\|_F \to 0\,, \tag{42}$$

and furthermore,

$$\|\nabla_\theta \left(\hat{\Theta}_\theta(X, X')\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})\right)(f_\theta(X') - Y')\|_F \to 0\,. \tag{43}$$

Proof of Lemma 5.

Overview. In this proof, we consider the expression

$$\begin{aligned}
&\nabla_\theta \left(\hat{\Theta}_\theta(x, X')\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})\right)(f_\theta(X') - Y') && (44)\\
= \quad &\nabla_\theta \left(\hat{\Theta}_\theta(x, X')\right)\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})(f_\theta(X') - Y') && (45)\\
&+ \hat{\Theta}_\theta(x, X')\left(\nabla_\theta\hat{\Theta}_\theta^{-1}\right)(I - e^{-\lambda\hat{\Theta}_\theta\tau})(f_\theta(X') - Y') && (46)\\
&+ \hat{\Theta}_\theta(x, X')\hat{\Theta}_\theta^{-1}\left(\nabla_\theta(I - e^{-\lambda\hat{\Theta}_\theta\tau})\right)(f_\theta(X') - Y'), && (47)
\end{aligned}$$

and we prove the terms of (45), (46) and (47) all have vanishing Frobenius norm. Thus, (44) also has vanishing Frobenius norm in the infinite width limit, which is exactly the statement of (42). This indicates that (43) also has a vanishing Frobenius norm, since $\hat{\Theta}_\theta(X, X')$ can be seen as a stack of n copies of $\hat{\Theta}_\theta(x, X')$, where n is a finite constant.

Step I. Each factor of (44) has bounded Frobenius norm.

- $\|\hat{\Theta}_\theta(x, X')\|_F$. It has been shown that $\|\frac{1}{\sqrt{l}}\nabla_\theta f(\cdot)\|_F \leq constant$ in [37], thus we have $\|\hat{\Theta}_\theta(x, X')\|_F = \|\frac{1}{l}\nabla_\theta f(x)\nabla_\theta f(X')^\top\|_F \leq \|\frac{1}{\sqrt{l}}\nabla_\theta f(x)\|_F\|\frac{1}{\sqrt{l}}\nabla_\theta f(X')\|_F \leq constant.$

- $\|\hat{\Theta}_\theta^{-1}\|_F$. It has been shown that $\hat{\Theta}_\theta$ is positive definite with positive least eigenvalue [4, 26], thus $\|\hat{\Theta}_\theta^{-1}\|_F \leq constant$.

- $\|I - e^{-\lambda\hat{\Theta}_\theta\tau}\|_F$. [10] shows that largest eigenvalues of $\hat{\Theta}_\theta$ are of $O(L)$, and we know $\hat{\Theta}_\theta$ is positive definite [4, 26], thus it is obvious the eigenvalues of $I - e^{-\lambda\hat{\Theta}_\theta\tau}$ fall in the set $\{z \mid 0 < z < 1\}$. Therefore, certainly we have $\|I - e^{-\lambda\hat{\Theta}_\theta\tau}\|_F \leq constant$.

- $\|f_\theta(X') - Y'\|_F$. [37] shows that $\|f_\theta(X') - Y'\|_2 \leq constant$, which indicates that $\|f_\theta(X') - Y'\|_F \leq constant$.

In conclusion, we have shown

$$\|\hat{\Theta}_\theta(x, X')\|_F, \|\hat{\Theta}_\theta^{-1}\|_F, \|I - e^{-\lambda\hat{\Theta}_\theta\tau}\|_F, \|f_\theta(X') - Y'\|_F \leq constant \tag{48}$$

Step II. Bound (45).

Without loss of generality, let us consider the neural net output dimension $k = 1$ in this proof, i.e., $f_\theta : \mathbb{R}^d \mapsto \mathbb{R}$. (Note: with $k > 1$, the only difference is that $\nabla_\theta f(X') \in \mathbb{R}^{mk\times D}$, which has no impact on the proof). Then, we have

$$\hat{\Theta}_\theta(x, X') = \frac{1}{l}\underbrace{\nabla_\theta f_\theta(\overbrace{x}^{\in\mathbb{R}^d})}_{\in\mathbb{R}^{1\times D}} \cdot \underbrace{\nabla_\theta f_\theta(\overbrace{X'}^{\in\mathbb{R}^{m\times d}})^\top}_{\in\mathbb{R}^{D\times m}} \in \mathbb{R}^{1\times m} \tag{49}$$

with gradient as

$$\nabla_\theta\hat{\Theta}_\theta(x, X') = \frac{1}{l}\underbrace{\nabla_\theta^2 f_\theta(x)}_{\in\mathbb{R}^{1\times D\times D}} \cdot \underbrace{\nabla_\theta f_\theta(X')^\top}_{\in\mathbb{R}^{D\times m}} + \frac{1}{l}\underbrace{\nabla_\theta f_\theta(x)}_{\in\mathbb{R}^{1\times D}} \cdot \underbrace{\nabla_\theta^2 f_\theta(X')^\top}_{\in\mathbb{R}^{D\times m\times D}} \in \mathbb{R}^{1\times m\times D} \tag{50}$$

where we apply a *dot product* in the *first two dimensions* of 3-tensors and matrices to obtain matrices.

Based on (48), we know that

$$\|\overbrace{\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})(f_\theta(X') - Y')}^{\in\mathbb{R}^{m\times 1}}\|_F \leq \|\hat{\Theta}_\theta^{-1}\|_F\|I - e^{-\lambda\hat{\Theta}_\theta\tau}\|_F\|f_\theta(X') - Y'\|_F \leq constant\ .$$

Then, applying (33), we have

$$\|\nabla_\theta\left(\hat{\Theta}_\theta(x, X')\right) \cdot \hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})(f_\theta(X') - Y')\|_F \to 0 \tag{51}$$

Step III. Bound (46) and (47)

- Bound (46): $\hat{\Theta}_\theta(x, X') \left(\nabla_\theta \hat{\Theta}_\theta^{-1}\right) (I - e^{-\lambda \hat{\Theta}_\theta \tau})(f_\theta(X') - Y')$.

  Clearly, $\underbrace{\nabla_\theta \hat{\Theta}_\theta^{-1}}_{m\times m\times D} = - \underbrace{\hat{\Theta}_\theta^{-1}}_{\in\mathbb{R}^{m\times m}} \cdot \underbrace{(\nabla_\theta \hat{\Theta}_\theta)}_{\in\mathbb{R}^{m\times m\times D}} \cdot \underbrace{\hat{\Theta}_\theta^{-1}}_{\mathbb{R}^{m\times m}}$, where we apply a dot product in the first two dimensions of the 3-tensor and matrices.

  Note that $\nabla_\theta \hat{\Theta}_\theta = \sqrt{1}l \nabla_\theta^2 f_\theta(X') \cdot \nabla_\theta f_\theta(X')^\top + \sqrt{1}l \nabla_\theta f_\theta(X') \cdot \nabla_\theta^2 f_\theta(X')^\top$. Obviously, by (39) and (48), we can easily prove that

$$\|\hat{\Theta}_\theta(x, X') \left(\nabla_\theta \hat{\Theta}_\theta^{-1}\right) (I - e^{-\lambda \hat{\Theta}_\theta \tau})(f_\theta(X') - Y')\|_F \to 0 \tag{52}$$

- Bound (47): $\hat{\Theta}_\theta(x, X') \hat{\Theta}_\theta^{-1} \left(\nabla_\theta (I - e^{-\lambda \hat{\Theta}_\theta \tau})\right) (f_\theta(X') - Y')$

  Since $\underbrace{\nabla_\theta (I - e^{-\lambda \hat{\Theta}_\theta \tau})}_{\in\mathbb{R}^{m\times m\times D}} = \lambda \tau \cdot \underbrace{e^{-\lambda \hat{\Theta}_\theta \tau}}_{\in\mathbb{R}^{m\times m}} \cdot \underbrace{\nabla_\theta \hat{\Theta}_\theta}_{\in\mathbb{R}^{m\times m\times D}}$, we can easily obtain the following result by (39) and (48),

$$\|\hat{\Theta}_\theta(x, X') \hat{\Theta}_\theta^{-1} \left(\nabla_\theta (I - e^{-\lambda \hat{\Theta}_\theta \tau})\right) (f_\theta(X') - Y')\|_F \to 0 \tag{53}$$

Step IV. Final result: prove (44) and (43).

Combining (51), (52) and (53), we can prove (44)

$$\begin{aligned}
&\|\nabla_\theta \left(\hat{\Theta}_\theta(x, X') \hat{\Theta}_\theta^{-1} (I - e^{-\lambda \hat{\Theta}_\theta \tau})\right) (f_\theta(X') - Y')|_F && (54)\\
&\leq \|\nabla_\theta \left(\hat{\Theta}_\theta(x, X')\right) \hat{\Theta}_\theta^{-1} (I - e^{-\lambda \hat{\Theta}_\theta \tau})(f_\theta(X') - Y')\|_F \\
&+ \|\hat{\Theta}_\theta(x, X') \left(\nabla_\theta \hat{\Theta}_\theta^{-1}\right) (I - e^{-\lambda \hat{\Theta}_\theta \tau})(f_\theta(X') - Y')\|_F \\
&+ \|\hat{\Theta}_\theta(x, X') \hat{\Theta}_\theta^{-1} \left(\nabla_\theta (I - e^{-\lambda \hat{\Theta}_\theta \tau})\right) (f_\theta(X') - Y')\|_F \\
&\to 0 && (55)
\end{aligned}$$

Then, since $\hat{\Theta}_\theta(X, X')$ can be seen as a stack of n copies of $\hat{\Theta}_\theta(x, X')$, where n is a finite constant, we can easily prove (43) by

$$\begin{aligned}
&\|\nabla_\theta \left(\hat{\Theta}_\theta(X, X') \hat{\Theta}_\theta^{-1} (I - e^{-\lambda \hat{\Theta}_\theta \tau})\right) (f_\theta(X') - Y')\|_F && (56)\\
&\leq \sum_{i\in[n]} \|\nabla_\theta \left(\hat{\Theta}_\theta(x_i, X') \hat{\Theta}_\theta^{-1} (I - e^{-\lambda \hat{\Theta}_\theta \tau})\right) (f_\theta(X') - Y')\|_F \\
&\to 0 && (57)
\end{aligned}$$

where we denote $X = (x_i)_{i=1}^n$. □

B.2. Proof of Lemma 1

Proof of Lemma 1. Consider an arbitrary task $\mathcal{T} = (X, Y, X', Y')$. Given sufficiently large width l, for any parameters in the neighborhood of the initialization, i.e., $\theta \in B(\theta_0, Cl^{-1/2})$, based on [37], we know the meta-output can be decomposed into a terms of f_θ,

$$F_\theta(X, X', Y') = f_\theta(X) - \hat{\Theta}_\theta(X, X') \hat{\Theta}_\theta^{-1} (I - e^{-\lambda \hat{\Theta}_\theta \tau})(f_\theta(X') - Y'), \tag{58}$$

where $\hat{\Theta}_\theta(X, X') = \frac{1}{l} \nabla_\theta f_\theta(X) \nabla_\theta f_\theta(X')^\top$, and $\hat{\Theta}_\theta \equiv \hat{\Theta}_\theta(X', X')$ for convenience.

Then, we consider $\nabla_\theta F_\theta(X, X', Y')$, the gradient of $F_\theta(X, X', Y')$ in (58),

$$\begin{aligned}\nabla_\theta F_\theta(X, X', Y') = \nabla_\theta f_\theta(X) - \hat{\Theta}_\theta(X, X')\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})\nabla_\theta f_\theta(X') \\ - \nabla_\theta \left(\hat{\Theta}_\theta(X, X')\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})\right)(f_\theta(X') - Y')\end{aligned} \tag{59}$$

By Lemma 5, we know the last term of (59) has a vanishing Frobenius norm as the width increases to infinity. Thus, for any $\varepsilon > 0$ and $0 < \delta < 1$, there exists $l^* > 0$ s.t. for width $l > l^*$, with probability at least $1 - \delta$, the last term of (59) is of $\mathcal{O}(\varepsilon)$, i.e.,

$$\nabla_\theta F_\theta(X, X', Y') = \nabla_\theta f_\theta(X) - \hat{\Theta}_\theta(X, X')\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})\nabla_\theta f_\theta(X') + \mathcal{O}(\varepsilon) \tag{60}$$

Since $\mathcal{O}(\varepsilon)$ is of a negligible order, we do not carry it in the remaining proof steps for simplicity, and it does not affect the correctness of the derivations (since the bounds of this Lemma are probabilistic).

Now, let us consider the SVD decomposition on $\frac{1}{\sqrt{l}}\nabla_\theta f_\theta(X') \in \mathbb{R}^{km\times D}$, where $X' \in \mathbb{R}^{k\times m}$ and $\theta \in \mathbb{R}^D$. such that $\frac{1}{\sqrt{l}}\nabla_\theta f_\theta(X') = U\Sigma V^\top$, where $U \in \mathbb{R}^{km\times km}, V \in \mathbb{R}^{D\times km}$ are orthonormal matrices while $\Sigma \in \mathbb{R}^{km\times km}$ is a diagonal matrix. Note that we take $km \leq D$ here since the width is sufficiently wide.

Then, since $\hat{\Theta}_\theta = \frac{1}{l}\nabla_\theta f_\theta(X')\nabla_\theta f_\theta(X')^\top = U\Sigma V^\top V\Sigma U^\top = U\Sigma^2 U^\top$, we have $\hat{\Theta}_\theta^{-1} = U\Sigma^{-2}U^\top$. Also, by Taylor expansion, we have

$$I - e^{-\lambda\hat{\Theta}_\theta\tau} = I - \sum_{i=0}^{\infty}\frac{(-\lambda\tau)^i}{i!}(\hat{\Theta}_\theta)^i = U\left(I - \sum_{i=0}^{\infty}\frac{(-\lambda\tau)^i}{i!}(\Sigma)^i\right)U^\top = U\left(I - e^{-\lambda\Sigma\tau}\right)U^\top. \tag{61}$$

With these results of SVD, (60) becomes

$$\begin{aligned}&\nabla_\theta F((X, X', Y'), \theta) \\ &= \nabla_\theta f_\theta(X) - \frac{1}{l}\nabla_\theta f_\theta(X)\nabla_\theta f_\theta(X')^\top\hat{\Theta}_\theta^{-1}(I - e^{-\lambda\hat{\Theta}_\theta\tau})\nabla_\theta f_\theta(X') \\ &= \nabla_\theta f_\theta(X) - \frac{1}{l}\nabla_\theta f_\theta(X)(\sqrt{l}V\Sigma U^\top)(U\Sigma^{-2}U^\top)[U\left(I - e^{-\lambda\Sigma\tau}\right)U^\top](\sqrt{l}U\Sigma V^\top) \\ &= \nabla_\theta f_\theta(X) - \nabla_\theta f_\theta(X)V\Sigma^{-1}\left(I - e^{-\lambda\Sigma\tau}\right)\Sigma V^\top \\ &= \nabla_\theta f_\theta(X) - \nabla_\theta f_\theta(X)V\left(I - e^{-\lambda\Sigma\tau}\right)V^\top \\ &= \nabla_\theta f_\theta(X) - \nabla_\theta f_\theta(X)(I - e^{-\lambda H_\theta\tau}) \\ &= \nabla_\theta f_\theta(X)e^{-\lambda H_\theta\tau}\end{aligned} \tag{62}$$

where $H_\theta \equiv H_\theta(X', X') = \frac{1}{l}\nabla_\theta f_\theta(X')^\top\nabla_\theta f_\theta(X') \in \mathbb{R}^{D\times D}$, and the step (62) can be easily obtained by a Taylor expansion similar to (61).

Note that H_θ is a product of $\nabla_\theta f_\theta(X')^\top$ and its transpose, hence it is positive semi-definite, and so does $e^{-\lambda H\tau}$. By eigen-decomposition on H, we can easily see that the eigenvalues of $e^{-\lambda H\tau}$ are all in the range $[0, 1)$ for arbitrary $\tau > 0$. Therefore, it is easy to get that for arbitrary $\tau > 0$,

$$\|\nabla_\theta F((X, X', Y'), \theta)\|_F = \|\nabla_\theta f_\theta(X)e^{-\lambda H_\theta\tau}\|_F \leq \|\nabla_\theta f_\theta(X)\|_F \tag{63}$$

By Lemma 1 of [37], we know that there exists a $K_0 > 0$ such that for any X and θ,

$$\|\frac{1}{\sqrt{l}}\nabla f_\theta(X)\|_F \leq K_0. \tag{64}$$

Combining (63) and (64), we have

$$\|\frac{1}{\sqrt{l}}\nabla_\theta F((X, X', Y'), \theta)\|_F \leq \|\frac{1}{\sqrt{l}}\nabla_\theta f_\theta(X)\|_F \leq K_0, \tag{65}$$

which is equivalent to

$$\frac{1}{\sqrt{l}}\|J(\theta)\|_F \leq K_0 \tag{66}$$

Now, let us study the other term of interest, $\|J(\theta)-J(\bar{\theta})\|_F = \|\frac{1}{\sqrt{l}}\nabla_\theta F((X,X',Y'),\theta) - \frac{1}{\sqrt{l}}\nabla_\theta F((X,X',Y'),\bar{\theta})\|_F$, where $\theta, \bar{\theta} \in B(\theta_0, Cl^{-1/2})$.

To bound $\|J(\theta)-J(\bar{\theta})\|_F$, let us consider

$$\begin{aligned}
&\|\nabla_\theta F((X,X',Y'),\theta) - \nabla_\theta F((X,X',Y'),\bar{\theta})\|_{op} \qquad (67)\\
&= \|\nabla_\theta f_\theta(X) e^{-\lambda H_\theta \tau} - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X) e^{-\lambda H_{\bar{\theta}} \tau}\|_{op}\\
&= \frac{1}{2}\|\left(\nabla_\theta f_\theta(X) - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\right)\left(e^{-\lambda H_\theta \tau} + e^{-\lambda H_{\bar{\theta}} \tau}\right) \qquad (68)\\
&\quad + \left(\nabla_\theta f_\theta(X) + \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\right)\left(e^{-\lambda H_\theta \tau} - e^{-\lambda H_{\bar{\theta}} \tau}\right)\|_{op}\\
&\leq \frac{1}{2}\|\nabla_\theta f_\theta(X) - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op}\|e^{-\lambda H_\theta \tau} + e^{-\lambda H_{\bar{\theta}} \tau}\|_{op} \qquad (69)\\
&\quad + \frac{1}{2}\|\nabla_\theta f_\theta(X) + \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op}\|e^{-\lambda H_\theta \tau} - e^{-\lambda H_{\bar{\theta}} \tau}\|_{op}\\
&\leq \frac{1}{2}\|\nabla_\theta f_\theta(X) - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op}\left(\|e^{-\lambda H_\theta \tau}\|_{op} + \|e^{-\lambda H_{\bar{\theta}} \tau}\|_{op}\right) \qquad (70)\\
&\quad + \frac{1}{2}\left(\|\nabla_\theta f_\theta(X)\|_{op} + \|\nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op}\right)\|e^{-\lambda H_\theta \tau} - e^{-\lambda H_{\bar{\theta}} \tau}\|_{op} \qquad (71)
\end{aligned}$$

It is obvious that $\|e^{-\lambda H_\theta \tau}\|_{op}, \|e^{-\lambda H_{\bar{\theta}} \tau}\|_{op} \leq 1$. Also, by the relation between the operator norm and the Frobenius norm, we have

$$\|\nabla_\theta f_\theta(X) - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op} \leq \|\nabla_\theta f_\theta(X) - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_F \tag{72}$$

Besides, Lemma 1 of [37] indicates that there exists a $K_1 > 0$ such that for any X and $\theta, \bar{\theta} \in B(\theta_0, Cl^{-1/2})$,

$$\|\frac{1}{\sqrt{l}}\nabla_\theta f_\theta(X) - \frac{1}{\sqrt{l}}\nabla_\theta f_{\bar{\theta}}(X)\|_F \leq K_1\|\theta - \bar{\theta}\|_2 \tag{73}$$

Therefore, (72) gives

$$\|\nabla_\theta f_\theta(X) - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op} \leq K_1\sqrt{l}\|\theta - \bar{\theta}\|_2 \tag{74}$$

and then (70) is bounded as

$$\frac{1}{2}\|\nabla_\theta f_\theta(X) - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op}\left(\|e^{-\lambda H_\theta \tau}\|_{op} + \|e^{-\lambda H_{\bar{\theta}} \tau}\|_{op}\right) \leq K_1\sqrt{l}\|\theta - \bar{\theta}\|_2. \tag{75}$$

As for (71), notice that $\|\cdot\|_{op} \leq \|\cdot\|_F$ and (64) give us

$$\frac{1}{2}\left(\|\nabla_\theta f_\theta(X)\|_{op} + \|\nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op}\right) \leq \sqrt{l}K_0. \tag{76}$$

Then, to bound $\|e^{-\lambda H_\theta \tau} - e^{-\lambda H_{\bar{\theta}} \tau}\|_{op}$ in (71), let us bound the following first

$$\begin{aligned}
\|H_\theta - H_{\bar{\theta}}\|_F &= \|\frac{1}{l}\nabla_\theta f_\theta(X')^\top \nabla_\theta f_\theta(X') - \frac{1}{l}\nabla_{\bar{\theta}} f_{\bar{\theta}}(X')^\top \nabla_{\bar{\theta}} f_{\bar{\theta}}(X')\|_F\\
&= \frac{1}{l}\|\frac{1}{2}(\nabla_\theta f_\theta(X')^\top + \nabla_{\bar{\theta}} f_{\bar{\theta}}(X')^\top)(\nabla_\theta f_\theta(X') - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X'))\\
&\quad + \frac{1}{2}(\nabla_\theta f_\theta(X')^\top - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X')^\top)(\nabla_\theta f_\theta(X') + \nabla_{\bar{\theta}} f_{\bar{\theta}}(X'))\|_F\\
&\leq \frac{1}{l}\|\nabla_\theta f_\theta(X') + \nabla_{\bar{\theta}} f_{\bar{\theta}}(X')\|_F\|\nabla_\theta f_\theta(X') - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X')\|_F\\
&\leq \frac{1}{l}\left(\|\nabla_\theta f_\theta(X')\|_F + \|\nabla_{\bar{\theta}} f_{\bar{\theta}}(X')_F\|\right)\|\nabla_\theta f_\theta(X') - \nabla_{\bar{\theta}} f_{\bar{\theta}}(X')\|_F\\
&\leq 2K_0K_1\|\theta - \bar{\theta}\|_2 \qquad (77)
\end{aligned}$$

Then, with the results above and a perturbation bound[13] on matrix exponentials from [29], we have

$$\begin{aligned}
\|e^{-\lambda H_\theta \tau} - e^{-\lambda H_{\bar{\theta}} \tau}\|_{op} &\leq \|H_\theta - H_{\bar{\theta}}\|_{op} \cdot \left(\lambda \tau e^{-\lambda\tau\cdot(\|H_\theta\|_{op} - \|H_\theta - H_{\bar{\theta}}\|_{op})}\right) \\
&\leq \frac{\|H_\theta - H_{\bar{\theta}}\|_{op}}{\|H_\theta\|_{op} - \|H_\theta - H_{\bar{\theta}}\|_{op}} \\
&\leq \mathcal{O}(\|H_\theta - H_{\bar{\theta}}\|_{op}) \\
&\leq 2K_0 K_1 K_2 \|\theta - \bar{\theta}\|_2
\end{aligned} \tag{78}$$

where we used the facts $\|H_\theta\|_{op} = \|\hat{\Theta}_\theta\|_{op} \geq \mathcal{O}(1)$ [10, 71] and $\|H_\theta - H_{\bar{\theta}}\|_{op} \leq \mathcal{O}(\|\theta - \bar{\theta}\|_2) \leq \mathcal{O}(\frac{1}{\sqrt{l}})$.

Hence, by (76) and (78), we can bound (71) as

$$\frac{1}{2}\left(\|\nabla_\theta f_\theta(X)\|_{op} + \|\nabla_{\bar{\theta}} f_{\bar{\theta}}(X)\|_{op}\right) \|e^{-\lambda H_\theta \tau} - e^{-\lambda H_{\bar{\theta}} \tau}\|_{op} \leq 2\sqrt{l} K_0^2 K_1 K_2 \|\theta - \bar{\theta}\|_2 \tag{79}$$

Finally, with (75) and (79), we can bound (67) as

$$\|\nabla_\theta F((X, X', Y'), \theta) - \nabla_\theta F((X, X', Y'), \bar{\theta})\|_{op} \leq (K_1 + 2K_0^2 K_1 K_2)\sqrt{l}\|\theta - \bar{\theta}\|_2$$

Finally, combining these bounds on (70) and (71), we know that

$$\begin{aligned}
\|J(\theta) - J(\bar{\theta})\|_F &= \|\frac{1}{\sqrt{l}} \nabla_\theta F((X, X', Y'), \theta) - \frac{1}{\sqrt{l}} \nabla_\theta F((X, X', Y'), \bar{\theta})\|_F \\
&\leq \frac{\sqrt{kn}}{\sqrt{l}} \|\nabla_\theta F((X, X', Y'), \theta) - \nabla_\theta F((X, X', Y'), \bar{\theta})\|_{op} \\
&\leq \sqrt{kn}(K_1 + 2K_0^2 K_1 K_2)\|\theta - \bar{\theta}\|_2
\end{aligned} \tag{80}$$

Define $K_3 = \sqrt{kn}(K_1 + 2K_0^2 K_1 K_2)$, we have

$$\|J(\theta) - J(\bar{\theta})\|_F \leq K_3 \|\theta - \bar{\theta}\|_2 \tag{81}$$

Taking $K = \max\{K_0, K_3\}$ completes the proof. □

B.3. Proof of Lemma 2

Proof of Lemma 2. It is known that $f_{\theta_0}(\cdot)$ converges in distribution to a mean zero Gaussian with the covariance $\mathcal{K}$ determined by the parameter initialization [37]. As a result, for arbitrary $\delta_1 \in (0, 1)$ there exist constants $l_1 > 0$ and $R_1 > 0$, such that: $\forall\ l \geq l_1$, over random initialization, the following inequality holds true with probability at least $(1 - \delta_1)$,

$$\|f_{\theta_0}(X) - Y\|_2, \|f_{\theta_0}(X') - Y'\|_2 \leq R_1 \tag{82}$$

We know that $\forall \mathcal{T} = (X, Y, X', Y') \in D$,

$$F_{\theta_0}(X, X', Y') = f_{\theta_0'}(X)$$

where θ_0' is the parameters after τ-step update on θ_0 over the meta-test task (X', Y'):

$$\begin{aligned}
&\theta_\tau = \theta', \ \ \theta_0 = \theta, \\
&\theta_{i+1} = \theta_i - \lambda \nabla_{\theta_i} \ell(f_{\theta_i}(X'), Y') \ \ \forall i = 0, ..., \tau - 1,
\end{aligned} \tag{83}$$

Suppose the learning rate λ is sufficiently small, then similar to (58), we have

$$F_{\theta_0}(X, X', Y') = f_{\theta_0}(X) + \hat{\Theta}_0(X, X')\hat{\Theta}_0^{-1}(I - e^{-\lambda \hat{\Theta}_0 \tau})(f_{\theta_0}(X') - Y'). \tag{84}$$

where $\hat{\Theta}_0(\cdot, \star) = \nabla_{\theta_0} f_{\theta_0}(\cdot) \nabla_{\theta_0} f_{\theta_0}(\star)^\top$ and we use a shorthand $\hat{\Theta}_0 \equiv \hat{\Theta}_0(X', X')$.

[13] This bound is also derived in [61].

[26] proves that for sufficiently large width, $\hat{\Theta}_0$ is positive definite and converges to Θ, the Neural Tangent Kernel, a full-rank kernel matrix with bounded positive eigenvalues. Let $\sigma_{\min}(\Theta)$ and $\sigma_{\max}(\Theta)$ denote the least and largest eigenvalue of Θ, respectively. Then, it is obvious that for a sufficiently over-parameterized neural network, the operator norm of $\hat{\Theta}(X, X')\hat{\Theta}^{-1}(I - e^{-\lambda\hat{\Theta}\tau})$ can be bounded based on $\sigma_{\min}(\Theta)$ and $\sigma_{\max}(\Theta)$. Besides, [4, 37] demonstrate that the neural net output at initialization, $f_{\theta_0}(\cdot)$, is a zero-mean Gaussian with small-scale covaraince. Combining these results and (82), we know there exists $R(R_1, N, \sigma_{\min}(\Theta), \sigma_{\max}(\Theta))$ such that

$$\|F_{\theta_0}(X, X', Y') - Y\|_2 \leq R(R_1, N, \sigma_{\min}(\Theta), \sigma_{\max}(\Theta)) \tag{85}$$

By taking an supremum over $R(R_1, N, \sigma_{\min}, \sigma_{\max})$ for each training task in $\{\mathcal{T}_i = (X_i, Y_i, X'_i, Y'_i)\}_{i\in[N]}$, we can get R_2 such that $\forall i \in [N]$

$$\|F_{\theta_0}(X_i, X'_i, Y'_i) - Y_i\|_2 \leq R_2 \tag{86}$$

and for $R_0 = \sqrt{N}R_2$, define δ_0 as some appropriate scaling of δ_1, then the following holds true with probability $(1 - \delta_0)$ over random initialization,

$$\|g(\theta_0)\|_2 = \sqrt{\sum_{X,Y,X',Y'\in D} \|F((X, X', Y'), \theta_0) - y\|_2^2} \leq R_0 \tag{87}$$

□

B.4. Proof of Lemma 3

Proof of Lemma 3. The learning rate for meta-adaption, λ, is sufficiently small, so the inner-loop optimization becomes *continuous-time* gradient descent. Based on [37], for any task $\mathcal{T} = (X, Y, X', Y')$,

$$F_0(X, X', Y') = f_0(X) + \hat{\Theta}_0(X, X')\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X', \tau)\left(Y' - f_0(X')\right), \tag{88}$$

where $\hat{\Theta}_0(\cdot, \star) = \frac{1}{l}\nabla_{\theta_0} f_0(\cdot)\nabla_{\theta_0} f_0(\star)^\top$, and $\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(\cdot, \tau) := \hat{\Theta}_0(\cdot, \cdot)^{-1}(I - e^{-\lambda\hat{\Theta}_0(\cdot,\cdot)\tau})$.

Then, we consider $\nabla_{\theta_0} F_0(X, X', Y')$, the gradient of $F_0(X, X', Y')$ in (88). By Lemma 5, we know that for sufficiently wide networks, the gradient of $F_0(X, X', Y')$ becomes

$$\nabla_{\theta_0} F_0(X, X', Y') = \nabla_{\theta_0} f_0(X) - \hat{\Theta}_0(X, X')T^{\lambda}_{\hat{\Theta}_0}(X', \tau)\nabla_{\theta_0} f_0(X') \tag{89}$$

Since $\hat{\Phi}_0 \equiv \hat{\Phi}_0((\mathcal{X}, \mathcal{X}', \mathcal{Y}'), (\mathcal{X}, \mathcal{X}', \mathcal{Y}')) = \frac{1}{l}\nabla_{\theta_0} F_0(\mathcal{X}, \mathcal{X}', \mathcal{Y}')\nabla_{\theta_0} F_0(\mathcal{X}, \mathcal{X}', \mathcal{Y}')^\top$ and $F_0(\mathcal{X}, \mathcal{X}', \mathcal{Y}') = (F_0(X_i, X'_i, Y'_i))_{i=1}^N \in \mathbb{R}^{knN}$, we know $\hat{\Phi}_0$ is a block matrix with $N \times N$ blocks of size $kn \times kn$. For $i, j \in [N]$, the (i, j)-th block can be denoted as $[\hat{\Phi}_0]_{ij}$ such that

$$\begin{aligned}
[\hat{\Phi}_0]_{ij} &= \frac{1}{l}\nabla_{\theta_0} F_0(X_i, X'_i, Y'_i)\nabla_{\theta_0} F_0(X_j, X'_j, Y'_j)^\top \\
&= \quad \frac{1}{l}\nabla_{\theta_0} f_0(X_i)\nabla_{\theta_0} f_0(X_j)^\top \\
&\quad + \frac{1}{l}\hat{\Theta}_0(X_i, X'_i)\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X'_i, \tau)\nabla_{\theta_0} f_0(X'_i)\nabla_{\theta_0} f_0(X'_j)^\top\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X'_j, \tau)^\top\hat{\Theta}_0(X'_j, X_j) \\
&\quad - \frac{1}{l}\nabla_{\theta_0} f_0(X_i)\nabla_{\theta_0} f_0(X'_j)^\top\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X'_j, \tau)^\top\hat{\Theta}_0(X'_j, X_j) \\
&\quad - \frac{1}{l}\hat{\Theta}_0(X_i, X'_i)\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X'_i, \tau)\nabla_{\theta_0} f_0(X'_i)\nabla_{\theta_0} f_0(X_j)^\top \\
&= \quad \hat{\Theta}_0(X_i, X_j) \\
&\quad + \hat{\Theta}_0(X_i, X'_i)\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X'_i, \tau)\hat{\Theta}_0(X'_i, X'_j)\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X'_j, \tau)^\top\hat{\Theta}_0(X'_j, X_j) \\
&\quad - \hat{\Theta}_0(X_i, X'_j)\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X'_j, \tau)^\top\hat{\Theta}_0(X'_j, X_j) \\
&\quad - \hat{\Theta}_0(X_i, X'_i)\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X'_i, \tau)\hat{\Theta}_0(X'_i, X_j)
\end{aligned} \tag{90}$$

where we used the equivalences $\hat{\Theta}_0(\cdot,\star) = \hat{\Theta}_0(\star,\cdot)^\top$ and $\frac{1}{l}\nabla_{\theta_0} f_0(\cdot)\nabla_{\theta_0} f_0(\star) = \hat{\Theta}_0(\cdot,\star)$.

By Algebraic Limit Theorem for Functional Limits, we have

$$\begin{aligned}
&\lim_{l\to\infty}[\hat{\Phi}_0]_{ij} \\
=& \lim_{l\to\infty}\hat{\Theta}_0(X_i, X_j) \\
&+ \lim_{l\to\infty}\hat{\Theta}_0(X_i, X_i')T^{\lambda}_{\lim_{l\to\infty}\hat{\Theta}_0}(X_i',\tau)\lim_{l\to\infty}\hat{\Theta}_0(X_i', X_j')T^{\lambda}_{\lim_{l\to\infty}\hat{\Theta}_0}(X_j',\tau)^\top\lim_{l\to\infty}\hat{\Theta}_0(X_j', X_j) \\
&- \lim_{l\to\infty}\hat{\Theta}_0(X_i, X_j')T^{\lambda}_{\lim_{l\to\infty}\hat{\Theta}_0}(X_j',\tau)^\top\lim_{l\to\infty}\hat{\Theta}_0(X_j', X_j) \\
&- \lim_{l\to\infty}\hat{\Theta}_0(X_i, X_i')T^{\lambda}_{\lim_{l\to\infty}\hat{\Theta}_0}(X_i',\tau)\hat{\Theta}_0(X_i', X_j) \\
=& \quad \Theta(X_i, X_j) \\
&+ \Theta(X_i, X_i')\widetilde{T}^{\lambda}_{\Theta}(X_i',\tau)\Theta(X_i', X_j')\widetilde{T}^{\lambda}_{\Theta}(X_j',\tau)^\top\Theta(X_j', X_j) \\
&- \Theta(X_i, X_j')\widetilde{T}^{\lambda}_{\Theta}(X_j',\tau)^\top\Theta(X_j', X_j) \\
&- \Theta(X_i, X_i')\widetilde{T}^{\lambda}_{\Theta}(X_i',\tau)\Theta(X_i', X_j)
\end{aligned} \tag{91}$$

where $\Theta(\cdot,\star) = \lim_{l\to\infty}\hat{\Theta}_0(\cdot,\star)$ is a deterministic kernel function, the Neural Tangent Kernel function (NTK) from the literature on supervised learning [4,26,37]. Specifically, $\hat{\Theta}_0(\cdot,\star)$ converges to $\Theta(\cdot,\star)$ in probability as the width l approaches infinity.

Hence, for any $i, j \in [N]$, as the width l approaches infinity, $[\hat{\Phi}_0]_{ij}$ converges in probability to a deterministic matrix $\lim_{l\to\infty}[\hat{\Phi}_0]_{ij}$, as shown by (91). Thus, the whole block matrix $\hat{\Phi}_0$ converges in probability to a deterministic matrix in the infinite width limit. Denote $\Phi = \lim_{l\to\infty}\hat{\Phi}_0$, then we know Φ is a deterministic matrix with each block expressed as (91).

Since $\hat{\Phi}_0 \equiv \hat{\Phi}_0((\mathcal{X},\mathcal{X}',\mathcal{Y}'),(\mathcal{X},\mathcal{X}',\mathcal{Y}')) = \frac{1}{l}\nabla_{\theta_0}F_0(\mathcal{X},\mathcal{X}',\mathcal{Y}')\nabla_{\theta_0}F_0(\mathcal{X},\mathcal{X}',\mathcal{Y}')^\top$, it is a symmetric square matrix. Hence all eigenvalues of $\hat{\Phi}_0$ are greater or equal to 0, which also holds true for Φ. In addition, because of Assumption 4, Φ is positive definite, indicating $\sigma_{\text{min}}(\Phi) > 0$. On the other hand, from [4], we know diagonal entries and eigenvalues of $\Theta(\cdot,\star)$ are positive real numbers upper bounded by $\mathcal{O}(L)$, as a direct result, it is easy to verify that the diagonal entries of the matrix Φ are also upper bounded, indicating $\sigma_{\text{max}}(\Phi) < \infty$. Hence, we have $0 < \sigma_{\text{min}}(\Phi) < \sigma_{\text{max}}(\Phi) < \infty$.

Extension. It is easy to extend (91), the expression for $\Phi \equiv \lim_{l\to\infty}\hat{\Phi}_0((\mathcal{X},\mathcal{X}',\mathcal{Y}'),(\mathcal{X},\mathcal{X}',\mathcal{Y}'))$, to more general cases. Specifically, we can express $\Phi(\cdot,\star)$ analytically for arbitrary inputs. To achieve this, let us define a kernel function, $\phi : (\mathbb{R}^{n\times k}\times\mathbb{R}^{m\times k})\times(\mathbb{R}^{n\times k}\times\mathbb{R}^{m\times k}) \mapsto \mathbb{R}^{nk\times nk}$ such that

$$\begin{aligned}
\phi((\cdot,*),(\bullet,\star)) = {}& \Theta(\cdot,\bullet) + \Theta(\cdot,*)\widetilde{T}^{\lambda}_{\Theta}(*,\tau)\Theta(*,\star)\widetilde{T}^{\lambda}_{\Theta}(\star,\tau)^\top\Theta(\star,\bullet) \\
& - \Theta(\cdot,*)\widetilde{T}^{\lambda}_{\Theta}(*,\tau)\Theta(*,\bullet) - \Theta(\cdot,\star)\widetilde{T}^{\lambda}_{\Theta}(\star,\tau)^\top\Theta(\star,\bullet).
\end{aligned} \tag{92}$$

Then, it is obvious that for $i, j \in [N]$, the (i,j)-th block of Φ can be expressed as $[\Phi]_{ij} = \phi((X_i, X_i'),(X_j, X_j'))$.

For cases such as $\Phi((X,X'),(\mathcal{X},\mathcal{X}')) \in \mathbb{R}^{kn\times knN}$, it is also obvious that $\Phi((X,X'),(\mathcal{X},\mathcal{X}'))$ is a block matrix that consists of $1\times N$ blocks of size $kn\times kn$, with the $(1,j)$-th block as follows for $j \in [N]$,

$$[\Phi((X,X'),(\mathcal{X},\mathcal{X}'))]_{1,j} = \phi((X,X'),(X_j,X_j')).$$

□

B.5. Proof of Theorem 3

Proof of Theorem 3. Based on these lemmas presented above, we can prove Theorem 3.

Lemma 2 indicates that there exist R_0 and l^* such that for any width $l \geq l^*$, the following holds true over random initialization with probability at least $(1-\delta_0/10)$,

$$\|g(\theta_0)\|_2 \leq R_0\,. \tag{93}$$

Consider $C = \frac{3KR_0}{\sigma}$ in Lemma 1.

First, we start with proving (29) and (32) by induction. Select $\widetilde{l} > l^*$ such that (93) and (25) hold with probability at least $1 - \frac{\delta_0}{5}$ over random initialization for every $l \geq \widetilde{l}$. As $t = 0$, by (28) and (25), we can easily verify that (29) and (32) hold true

$$\begin{cases} \|\theta_1 - \theta_0\|_2 & = \| - \eta J(\theta_0)^\top g(\theta_0)\|_2 \leq \eta \|J(\theta_0)\|_{op} \|g(\theta_0)\|_2 \leq \frac{\eta_0}{l} \|J(\theta_0)\|_F \|g(\theta_0)\|_2 \leq \frac{K\eta_0}{\sqrt{l}} R_0 \,. \\ \|g(\theta_0)\|_2 & \leq R_0 \end{cases}$$

Assume (29) and (32) hold true for any number of training step j such that $j < t$. Then, by (25) and (32), we have

$$\|\theta_{t+1} - \theta_t\|_2 \leq \eta \|J(\theta_t)\|_{op} \|g(\theta_t)\|_2 \leq \frac{K\eta_0}{\sqrt{l}} \left(1 - \frac{\eta_0 \sigma_{\min}}{3}\right)^t R_0 \,.$$

Beside, with the mean value theorem and (28), we have the following

$$\begin{aligned} \|g(\theta_{t+1})\|_2 &= \|g(\theta_{t+1} - g(\theta_t) + g(\theta_t))\|_2 \\ &= \|J(\theta_t^\mu)(\theta_{t+1} - \theta_t) + g(\theta_t)\|_2 \\ &= \|(I - \eta J(\theta_t^\mu) J(\theta_t)^\top) g(\theta_t)\|_2 \\ &\leq \|I - \eta J(\theta_t^\mu) J(\theta_t)^\top\|_{op} \|g(\theta_t)\|_2 \\ &\leq \|I - \eta J(\theta_t^\mu) J(\theta_t)^\top\|_{op} \left(1 - \frac{\eta_0 \sigma_{\min}}{3}\right)^t R_0 \end{aligned}$$

where we define θ_t^μ as a linear interpolation between θ_t and θ_{t+1} such that $\theta_t^\mu := \mu\theta_t + (1-\mu)\theta_{t+1}$ for some $0 < \mu < 1$.

Now, we will show that with probability $1 - \frac{\delta_0}{2}$,

$$\|I - \eta J(\theta_t^\mu) J(\theta_t)^\top\|_{op} \leq 1 - \frac{\eta_0 \sigma_{\min}}{3}.$$

Recall that $\hat{\Phi}_0 \to \Phi$ in probability, proved by Lemma 3. Then, there exists $\hat{l}$ such that the following holds with probability at least $1 - \frac{\delta_0}{5}$ for any width $l > \hat{l}$,

$$\|\Phi - \hat{\Phi}_0\|_F \leq \frac{\eta_0 \sigma_{\min}}{3}.$$

Our assumption $\eta_0 < \frac{2}{\sigma_{\max} + \sigma_{\min}}$ makes sure that

$$\|I - \eta_0 \Phi\|_{op} \leq 1 - \eta_0 \sigma_{\min} \,.$$

Therefore, as $l \geq (\frac{18K^3 R_0}{\sigma_{\min}^2})^2$, with probability at least $1 - \frac{\delta_0}{2}$ the following holds,

$$\begin{aligned} &\|I - \eta J(\theta_t^\mu) J(\theta_t)^\top\|_{op} \\ &= \|I - \eta_0 \Phi + \eta_0 \Phi - \hat{\Phi}_0 + \eta \left(J(\theta_0) J(\theta_0)^\top - J(\theta_t^\mu) J(\theta_t)^\top\right) \|_{op} \\ &\leq \|I - \eta_0 \Phi\|_{op} + \eta_0 \|\Phi - \hat{\Phi}_0\|_{op} + \eta \|J(\theta_0) J(\theta_0)^\top - J(\theta_t^\mu) J(\theta_t)^\top\|_{op} \\ &\leq 1 - \eta_0 \sigma_{\min} + \frac{\eta_0 \sigma_{\min}}{3} + \eta_0 K^2 (\|\theta_t - \theta_0\|_2 + \|\theta_t^\mu - \theta_0\|_2) \\ &\leq 1 - \eta_0 \sigma_{\min} + \frac{\eta_0 \sigma_{\min}}{3} + \frac{6\eta_0 K^3 R_0}{\sigma_{\min} \sqrt{l}} \\ &\leq 1 - \frac{\eta_0 \sigma_{\min}}{3} \end{aligned}$$

where we used the equality $\frac{1}{l} J(\theta_0) J(\theta_0)^\top = \hat{\Phi}_0$.

Hence, as we choose $\Lambda = \max\{l^*, \widetilde{l}, \hat{l}, \frac{18K^3 R_0}{\sigma_{\min}^2})^2\}$, the following holds for any width $l > \Lambda$ with probability at least $1 - \delta_0$ over random initialization

$$\|g(\theta_{t+1}\|_2 \leq \|I - \eta J(\theta_t^\mu) J(\theta_t)^\top\|_{op} \left(1 - \frac{\eta_0 \sigma_{\min}}{3}\right)^t R_0 \leq \left(1 - \frac{\eta_0 \sigma_{\min}}{3}\right)^{t+1} R_0, \tag{94}$$

which finishes the proof (32).

Finally, we prove (30) by

$$\begin{aligned}\|\hat{\Phi}_0 - \hat{\Phi}_t\|_F &= \frac{1}{l}\|J(\theta_0)J(\theta_0)^\top - J(\theta_t)J(\theta_t)^\top\|_F \\ &\leq \frac{1}{l}\|J(\theta_0)\|_{op}\|J(\theta_0)^\top - J(\theta_t)^\top\|_F + \frac{1}{l}\|J(\theta_t) - J(\theta_0)\|_{op}\|J(\theta_t)^\top\|_F \\ &\leq 2K^2\|\theta_0 - \theta_t\|_2 \\ &\leq \frac{6K^3R_0}{\sigma_{\min}\sqrt{l}},\end{aligned}$$

where we used (29) and Lemma 1. □

C. Analytical Expression of MAML Output

In this section, we will present Corollary 3.1. Briefly speaking, with the help of Theorem 3, we first show the training dynamics of MAML with over-parameterized DNNs can be described by a differential equation, which is analytically solvable. By solving this differential equation, we obtain the expression for MAML output on any training or test task.

Remarks. This corollary implies for a sufficiently over-parameterized neural network, the training of MAML is *determined* by the *parameter initialization*, θ_0. Given access to θ_0, we can compute the functions $\hat{\Phi}_0$ and F_0, and then the trained MAML output can be obtained by simple calculations, without the need for running gradient descent on θ_0. This nice property enables us to perform a deeper analysis on MAML with DNNs.

Corollary 3.1 (MAML Output (Corollary 3.1 Restated))**.** *In the setting of Theorem 1, the training dynamics of the MAML can be described by a differential equation*

$$\frac{dF_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')}{dt} = -\eta\,\hat{\Phi}_0(F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}') - \mathcal{Y})$$

where we denote $F_t \equiv F_{\theta_t}$ and $\hat{\Phi}_0 \equiv \hat{\Phi}_{\theta_0}((\mathcal{X},\mathcal{X}',\mathcal{Y}'),(\mathcal{X},\mathcal{X}',\mathcal{Y}'))$ for convenience.

Solving this differential equation, we obtain the meta-output of MAML on training tasks at any training time as

$$F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}') = (I - e^{-\eta\hat{\Phi}_0 t})\mathcal{Y} + e^{-\eta\hat{\Phi}_0 t}F_0(\mathcal{X},\mathcal{X}',\mathcal{Y}')\,. \tag{95}$$

Similarly, on arbitrary test task $\mathcal{T} = (X, Y, X', Y')$, the meta-output of MAML is

$$F_t(X,X',Y') = F_0(X,X',Y') + \hat{\Phi}_0(X,X',Y')T^{\eta}_{\hat{\Phi}_0}(t)\,(\mathcal{Y} - F_0(\mathcal{X},\mathcal{X}',\mathcal{Y}')) \tag{96}$$

where $\hat{\Phi}_0(\cdot) \equiv \hat{\Phi}_{\theta_0}(\cdot,(\mathcal{X},\mathcal{X}',\mathcal{Y}'))$ and $T^{\eta}_{\hat{\Phi}_0}(t) = \hat{\Phi}_0^{-1}\left(I - e^{-\eta\hat{\Phi}_0 t}\right)$ are shorthand notations.

Proof. For the optimization of MAML, the gradient descent on θ_t with learning rate η can be expressed as

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta\nabla_{\theta_t}\mathcal{L}(\theta_t) \\ &= \theta_t - \frac{1}{2}\eta\nabla_{\theta_t}\|F_{\theta_t}(\mathcal{X},\mathcal{X}',\mathcal{Y}') - \mathcal{Y}\|_2^2 \\ &= \theta_t - \eta\nabla_{\theta_t}F_{\theta_t}(\mathcal{X},\mathcal{X}',\mathcal{Y}')^\top\,(F_{\theta_t}(\mathcal{X},\mathcal{X}',\mathcal{Y}') - \mathcal{Y})\end{aligned} \tag{97}$$

Since the learning rate η is sufficiently small, the *discrete-time* gradient descent above can be re-written in the form of *continuous-time* gradient descent (i.e., gradient flow),

$$\frac{d\theta_t}{dt} = -\eta\nabla_{\theta_t}F_{\theta_t}(\mathcal{X},\mathcal{X}',\mathcal{Y}')^\top\,(F_{\theta_t}(\mathcal{X},\mathcal{X}',\mathcal{Y}') - \mathcal{Y})$$

Then, the training dynamics of the meta-output $F_t(\cdot) \equiv F_{\theta_t}(\cdot)$ can be described by the following differential equation,

$$\begin{aligned}\frac{dF_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')}{dt} &= \nabla_{\theta_t}F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')\frac{d\theta_t}{dt} \\ &= -\eta\nabla_{\theta_t}F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')\nabla_{\theta_t}F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')^\top\,(F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}') - \mathcal{Y}) \\ &= -\eta\hat{\Phi}_t\,(F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}') - \mathcal{Y})\end{aligned} \tag{98}$$

where $\hat{\Phi}_t = \hat{\Phi}_t((\mathcal{X},\mathcal{X}',\mathcal{Y}'),(\mathcal{X},\mathcal{X}',\mathcal{Y}')) = \nabla_{\theta_t} F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')\nabla_{\theta_t} F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')^\top$.

On the other hand, Theorem 3 gives the following bound in (30),

$$\sup_t \|\hat{\Phi}_0 - \hat{\Phi}_t\|_F \leq \frac{6K^3 R_0}{\sigma_{\min}} l^{-\frac{1}{2}}, \tag{99}$$

indicating $\hat{\Phi}_t$ stays almost constant during training for sufficiently over-parameterized neural networks (i.e., large enough width l). Therefore, similar to [37], we can replace $\hat{\Phi}_t$ by $\hat{\Phi}_0$ in (98), and get

$$\frac{dF_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')}{dt} = -\eta\hat{\Phi}_0\left(F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}') - \mathcal{Y}\right), \tag{100}$$

which is an ordinary differential equation (ODE) for the meta-output $F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}')$ w.r.t. the training time t.

This ODE is analytically solvable with a unique solution. Solving it, we obtain the meta-output on training tasks at any training time t as,

$$F_t(\mathcal{X},\mathcal{X}',\mathcal{Y}') = (I - e^{-\eta\hat{\Phi}_0 t})\mathcal{Y} + e^{-\eta\hat{\Phi}_0 t}F_0(\mathcal{X},\mathcal{X}',\mathcal{Y}'). \tag{101}$$

The solution can be easily extended to any test task $\mathcal{T} = (X, Y, X', Y')$, and the meta-output on the test task at any training time is

$$F_t(X,X',Y') = F_0(X,X',Y') + \hat{\Phi}_0(X,X',Y')T^{\eta}_{\hat{\Phi}_0}(t)\left(\mathcal{Y} - F_0(\mathcal{X},\mathcal{X}',\mathcal{Y}')\right), \tag{102}$$

where $\hat{\Phi}_0(\cdot) \equiv \hat{\Phi}_{\theta_0}(\cdot,(\mathcal{X},\mathcal{X}',\mathcal{Y}'))$ and $T^{\eta}_{\hat{\Phi}_0}(t) = \hat{\Phi}_0^{-1}\left(I - e^{-\eta\hat{\Phi}_0 t}\right)$ are shorthand notations. □

D. Gradient-Based Meta-Learning as Kernel Regression

In this section, we first make an assumption on the scale of parameter initialization, then we restate Theorem 2. After that, we provide the proof for Theorem 2.

[37] shows the output of a neural network randomly initialized following (16) is a zero-mean Gaussian with covariance determined by σ_w and σ_b, the variances corresponding to the initialization of weights and biases. Hence, small values of σ_w and σ_b can make the outputs of randomly initialized neural networks approximately zero. We adopt the following assumption from [4] to simplify the expression of the kernel regression in Theorem 2.

Assumption 5 (Small Scale of Parameter Initialization)**.** *The scale of parameter initialization is sufficiently small, i.e., σ_w, σ_b in (16) are small enough, so that $f_0(\cdot) \simeq 0$.*

Note the goal of this assumption is to make the output of the randomly initialized neural network negligible. The assumption is quite common and mild, since, in general, the outputs of randomly initialized neural networks are of small scare compared with the outputs of trained networks [37].

Theorem 4 (MAML as Kernel Regression (Theorem 2 Restated))**.** *Suppose learning rates η and λ are infinitesimal. As the network width l approaches infinity, with high probability over random initialization of the neural net, the MAML output, (8), converges to a special kernel regression,*

$$F_t(X,X',Y') = G^{\tau}_{\Theta}(X,X',Y') + \Phi((X,X'),(\mathcal{X},\mathcal{X}'))T^{\eta}_{\Phi}(t)\left(\mathcal{Y} - G^{\tau}_{\Theta}(\mathcal{X},\mathcal{X}',\mathcal{Y}')\right) \tag{103}$$

where G is a function defined below, Θ is the neural tangent kernel (NTK) function from [26] that can be analytically calculated without constructing any neural net, and Φ is a new kernel, which name as Meta Neural Kernel (MNK). The expression for G is

$$G^{\tau}_{\Theta}(X,X',Y') = \Theta(X,X')\widetilde{T}^{\lambda}_{\Theta}(X',\tau)Y'. \tag{104}$$

where $\widetilde{T}^{\lambda}_{\Theta}(\cdot,\tau) := \Theta(\cdot,\cdot)^{-1}(I - e^{-\lambda\Theta(\cdot,\cdot)\tau})$. Besides, $G^{\tau}_{\Theta}(\mathcal{X},\mathcal{X}',\mathcal{Y}') = (G^{\tau}_{\Theta}(X_i,X'_i,Y'_i))_{i=1}^N$.

The MNK is defined as $\Phi \equiv \Phi((\mathcal{X},\mathcal{X}'),(\mathcal{X},\mathcal{X}')) \in \mathbb{R}^{knN\times knN}$, which is a block matrix that consists of $N\times N$ blocks of size $kn\times kn$. For $i,j\in[N]$, the (i,j)-th block of Φ is

$$[\Phi]_{ij} = \phi((X_i,X'_i),(X_j,X'_j)) \in \mathbb{R}^{kn\times kn}, \tag{105}$$

where $\phi : (\mathbb{R}^{n\times k} \times \mathbb{R}^{m\times k}) \times (\mathbb{R}^{n\times k} \times \mathbb{R}^{m\times k}) \to \mathbb{R}^{nk\times nk}$ *is a kernel function defined as*

$$\begin{aligned}\phi((\cdot,*),(\bullet,\star)) = \Theta(\cdot,\bullet) &+ \Theta(\cdot,*)\widetilde{T}^{\lambda}_{\Theta}(*,\tau)\Theta(*,\star)\widetilde{T}^{\lambda}_{\Theta}(\star,\tau)^{\top}\Theta(\star,\bullet) \\ &- \Theta(\cdot,*)\widetilde{T}^{\lambda}_{\Theta}(*,\tau)\Theta(*,\bullet) - \Theta(\cdot,\star)\widetilde{T}^{\lambda}_{\Theta}(\star,\tau)^{\top}\Theta(\star,\bullet). \qquad (106)\end{aligned}$$

Here $\Phi((X,X'),(\mathcal{X},\mathcal{X}')) \in \mathbb{R}^{kn\times knN}$ *in (12) is also a block matrix, which consists of* $1\times N$ *blocks of size* $kn\times kn$*, with the* $(1,j)$*-th block as follows for* $j\in[N]$*,*

$$[\Phi((X,X'),(\mathcal{X},\mathcal{X}'))]_{1,j} = \phi((X,X'),(X_j,X'_j)). \qquad (107)$$

Proof. First, (8) shows that the output of MAML on any test task $\mathcal{T} = (X,Y,X',Y')$ can be expressed as

$$F_t(X,X',Y') = F_0(X,X',Y') + \hat{\Phi}_0(X,X',Y')T^{\eta}_{\hat{\Phi}_0}(t)\left(\mathcal{Y} - F_0(\mathcal{X},\mathcal{X}',\mathcal{Y}')\right) \qquad (108)$$

Note (88) in Appendix B.4 shows that

$$F_0(X,X',Y') = f_0(X) + \hat{\Theta}_0(X,X')\widetilde{T}^{\lambda}_{\hat{\Theta}_0}(X',\tau)\left(Y' - f_0(X')\right), \qquad (109)$$

With Assumption 5, we can drop the terms $f_0(X)$ and $f_0(X')$ in (109). Besides, from [4,26,37], we know $\lim_{l\to\infty}\hat{\Theta}_0(\cdot,\star) = \Theta(\cdot,\star)$, the Neural Tangent Kernel (NTK) function, a determinisitc kernel function. Therefore, $F_0(X,X',Y')$ can be described by the following function as the width appraoches infinity,

$$\lim_{l\to\infty} F_0(X,X',Y') = G^{\tau}_{\Theta}(X,X',Y') = \Theta(X,X')\widetilde{T}^{\lambda}_{\Theta}(X',\tau)Y'. \qquad (110)$$

where $\widetilde{T}^{\lambda}_{\Theta}(\cdot,\tau) := \Theta(\cdot,\cdot)^{-1}(I - e^{-\lambda\Theta(\cdot,\cdot)\tau})$. Besides, $G^{\tau}_{\Theta}(\mathcal{X},\mathcal{X}',\mathcal{Y}') = (G^{\tau}_{\Theta}(X_i,X'_i,Y'_i))_{i=1}^{N}$.

In addition, from Lemma 3, we know $\lim_{l\to\infty}\hat{\Phi}_0(\cdot,\star) = \Phi(\cdot,\star)$. Combined this with (110), we can express (108) in the infinite width limit as

$$F_t(X,X',Y') = G^{\tau}_{\Theta}(X,X',Y') + \Phi((X,X'),(\mathcal{X},\mathcal{X}'))T^{\eta}_{\Phi}(t)\left(\mathcal{Y} - G^{\tau}_{\Theta}(\mathcal{X},\mathcal{X}',\mathcal{Y}')\right) \qquad (111)$$

where $\Phi(\cdot,\star)$ is a kernel function that we name as Meta Neural Kernel function. The derivation of its expression shown in (105)-(107) can be found in Appendix B.4. □

ANIL Kernel The above theorem derives the analytical expression of the kernel induced by MAML. Certainly that variants algorithms of MAML will induce kernels with (slightly) different expressions. A recent impactful variant of MAML is Almost-No-Inner-Loop (ANIL) [50]. ANIL is a simplification of MAML that retains the performance of MAML while enjoying a significant training speedup. The key idea of ANIL is to remove the inner-loop updates on the hidden layers; in other words, ANIL only update the last linear layer in the inner loop, resulting in a much smaller computation and memory cost compared with MAML. Following procedures in Appendix C and D, one can straightforwardly derive the expression of the kernel induced by ANIL, which just replaces Eq. (106) (kernel function induced by MAML) by

$$\begin{aligned}\phi((\cdot,*),(\bullet,\star)) = \Theta(\cdot,\bullet) &+ \mathcal{K}(\cdot,*)\widetilde{T}^{\lambda}_{\mathcal{K}}(*,\tau)\Theta(*,\star)\widetilde{T}^{\lambda}_{\mathcal{K}}(\star,\tau)^{\top}\mathcal{K}(\star,\bullet) \\ &- \mathcal{K}(\cdot,*)\widetilde{T}^{\lambda}_{\mathcal{K}}(*,\tau)\Theta(*,\bullet) - \Theta(\cdot,\star)\widetilde{T}^{\lambda}_{\mathcal{K}}(\star,\tau)^{\top}\mathcal{K}(\star,\bullet). \qquad (112)\end{aligned}$$

where $\mathcal{K}$ is the neural tangent kernel function corresponds to neural networks with frozen hidden layers (i.e., only the last linear layer is optimized by gradient descent). The appearance of $\mathcal{K}$ directly results from the special inner-loop update rule of ANIL (i.e., only updates the last linear layer in the inner loop).

E. More Details on Experiments

Training Data Augmentation Following previous few-shot learning works [38, 60], in the training stage, we adopt data augmentation operations, including random cropping, color jittering, and random horizontal flip.

Training Batch Size For all 5-cells experiments, a batch size of 64 is used. For 8-cells experiments, we set the batch size to 40 for miniImageNet and 56 for tieredImageNet to accommodate the GPU memory of a single GPU card.

Dropout Rate We use dropout rate of 0.2 and 0.1 for miniImageNet and tieredImageNet, respectively. Following DARTS [42], we gradually increase the dropout rate during the training.

Normalization Layers To enable efficient computation of per-sample-gradients with Opacus [76] (it does not support BatchNorm so far), we first convert all the BatchNorm [25] layers to GroupNorm [70] layers with 16 number of groups in the search stage. After obtaining the cells, we train and evaluate the selected architectures with BatchNorm layers.

Hyper-parameters for Computing MetaNTK MAML kernels (defined in Definition 2)) and ANIL kernels (defined in Eq. (112)) are used for 5-cells and 8-cells experiments, respectively. To write more concisely, We denote the product of inner loop learning rate and training time as $\lambda\tau$. An $\lambda\tau = \infty$ and a regularization coefficient of 0.001 are used for all 5-cells experiments. For 8-cells experiments, an $\lambda\tau = 1$ and a kernel regularization coefficient of 10^{-5} are used for miniImageNet experiments while an $\lambda\tau = \infty$ and a kernel regularization coefficient of 0.001 are used for tieredImageNet experiments.

Hyper-parameters for Evaluation In the evaluation stage, we fine-tune the last layer of the learned neural net on the labelled support samples of each test task, and then evaluate its prediction accuracy on the query samples. Following the evaluation strategies of RFS [60], (*i*) we normalize the last hidden layer output of each sample to unit norm before passing to the last layer during the evaluation; (*ii*) we enlarge the set of support samples by applying data augmentation (used in the training stage) to create 5x augmented support samples for fine-tuning. We use cross-entropy loss and hinge loss for the fine-tuning, both with ℓ_2 regularization. For cross-entropy fine-tuning, we use the Logistic Regression (LR) solver provided in scikit-learn [49]; for the hinge loss fine-tuning, we adopt the C-Support Vector Classification (SVC) with linear kernel provided in scikit-learn [49]. Notice that these the ℓ_2 regularization in scikit-learn solvers is controlled by a regularization parameter $C = \frac{1}{\ell_2 \text{ penalty}}$ On mini-ImageNet: (*i*) in the 5-cells case, we use SVC with $C = 0.2$ for 1-shot and LR with $C = 0.6$ for the 5-shot experiments; (*ii*) in the 8-cells case, we use SVC with $C = 0.35$ for 1-shot and LR with $C = 0.4$ for the 5-shot experiments. On tiered-ImageNet: (i) in the 5-cells case, we use SVC with $C = 0.75$ for 1-shot and LR with $C = 0.8$ for the 5-shot experiments; (*ii*) in the 8-cells case, we use LR with $C = 0.95$ for 1-shot and LR with $C = 0.5$ for the 5-shot experiments.