Appendix

Overview of the Appendix

The appendix mainly consists of three parts. In Section A we provide more detailed introduction to the set up of meta-learning as well as neural tangent kernels that are missing from the main text due to page limit. In Section B we provide all the missing proofs of the lemmas and theorems presented in the main paper. In Section C we discuss in depth about the experiments in the paper. For the convenience of readers, we also provide a copy of the reference at the end of this appendix.

A. More on Meta-Learning and Neural Net Setup

In this section, we will provide more information on

- Appendix A.1: Query-support split of meta-learning.
- Appendix A.2: Unified framework for gradient-based meta-learning that optimizes all layers in the inner loop.
- Appendix A.3: NTK parameterization.

A.1. Query-Support Split

Sec. 3.2 introduces meta-training in the setting without query-support split. In this section, we adopt the notation of Sec. 3.2, and describe meta-training in the setting with query-support split below.

The n labelled samples in each training task is divided into two sets, n_q query samples and n_s support samples, i.e., for $i \in [N]$, the i-th task consists of

$$\begin{aligned} &n_q \text{ Query Samples \& Labels:} && X_i^q \in \mathbb{R}^{n_q \times d},\ Y_i^q \in \mathbb{R}^{n_q k} \\ &n_s \text{ Support Samples \& Labels:} && X_i^s \in \mathbb{R}^{n_s \times d},\ Y_i^s \in \mathbb{R}^{n_q k} \end{aligned}$$

The optimization objective of ANIL on the training data $\{X_i^q, Y_i^q, X_i^s, Y_i^s\}_{i=1}^N$ is

$$\min_{\theta} \mathcal{L}_{ANIL}(\theta) \coloneqq \sum_{i \in [N]} \ell(\phi_{\theta^{<L}}^{\top}(X_i^q)\, w_i',\ Y_i^q) \tag{19}$$

$$\text{s.t. } w_i' = \texttt{InnerLoop}(w, \phi_{\theta^{<L}}^{\top}(X_i^s), Y_i^s, \tau, \lambda) \tag{20}$$

It is clear that the `InnerLoop` operation is performed on *support* data (X_i^s, Y_i^s), while the loss evaluation is on the *query* data (X_i^q, Y_i^q).

A.2. Unified Framework for Gradient-Based Meta-Learning that Optimizes All Layers in the Inner Loop

For GBML algorithms that optimize all layers in the inner loop, their objectives can be summarized into the following unified framework. In contrast to (13), we have

$$\min_{\theta} \overbrace{\Big[\min_{\{\theta_i\}_{i=1}^N} \sum_{i \in [N]} \ell\Big(f_{\theta_i}(X_i), Y_i\Big) + R(\theta_i) \Big]}^{\mathcal{L}_{GBML}(\theta)}. \tag{21}$$

Note that similar to (13), the parameters $\{\theta_i\}_{i=1}^N$ in (21) are transient, in the sense that GBML algorithms do not explicitly save them during training. In contrast, θ contains all parameters to optimize in (21), and θ is optimized over $\ell_{GBML}(\theta)$, which is obtained by plugging in the minimizer of $\{\theta_i\}_{i=1}^N$ on the regularized loss. In other words, (21) is a bi-level optimization problem, with outer-loop optimization on network parameters θ and inner-loop optimization on the transient parameters $\{\theta_i\}_{i=1}^N$.

A.3. NTK Parameterization

NTK parameterization is a neural net parameterization that can be used to provide theoretical analyses of neural net optimization and convergence (Lee et al., 2019a; Xiao et al., 2020). The training dynamics and predictions of NTK-parameterized neural nets are the same as those of standard neural nets (Lee et al., 2019a), up to a width-dependent factor in the learning rate. In what follows, we take a single-head neural net as an example to describe the NTK parameterization. Notice that multi-head networks share the same parameterization with single-head networks, and the only difference is that N-head networks have N copies of the output heads (parameterized in the same way as the output heads of single-head networks).

In this paper, we consider a fully-connected feed-forward network with L layers. Each hidden layer has width l_i, for $i = 1, ..., L-1$. The readout layer (i.e., output layer) has width $l_L = k$. 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

$$\begin{cases} h^{i+1} &= z^i W^{i+1} + b^{i+1} \\ z^{i+1} &= \sigma\left(h^{i+1}\right) \end{cases} \text{ and } \begin{cases} W^i_{\mu,\nu} &= \omega^i_{\mu\nu} \sim \mathcal{N}(0, \frac{\sigma_\omega}{\sqrt{l_i}}) \\ b^i_\nu &= \beta^i_\nu \sim \mathcal{N}(0, \sigma_b) \end{cases}, \tag{22}$$

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

We present all the missing proofs from the main paper, summarized as follows:

- Appendix B.1: Proves the **global convergence** of MTL and ANIL, and demonstrates that neural net output and meta-output functions are linearized under over-parameterization.
- Appendix B.2: Studies the **training dynamics** of MTL and ANIL, and derives analytic expressions for their predictors.
- Appendix B.3: Derives the expression of **kernels** for MTL and ANIL, and proves **Lemma 1**.
- Appendix B.4: Characterizes the **structures and spectra** of ANIL and MTL kernels for deep **ReLU** nets.
- Appendix B.5: Proves our main theorem, i.e., **Theorem 1.**
- Appendix B.6: Extends Theorem 1 to **residual** ReLU networks.

Shorthand. As described in Sec. 3.4, for both MTL and ANIL, we randomly initialize a test head w_{test} for fine-tuning in the test phase. Now, we define the following shorthand for convenience.

- $\theta^{test} = \{\theta^{<L}, w_{test}\}$: a parameter set including first $L-1$ layers' parameters of θ and the test head w_{test}.
- $\hat{\theta}^{test} = \{\hat{\theta}^{<L}, w_{test}\}$: a parameter set including first $L-1$ layers' parameters of θ and the test head w_{test}.

B.1. Global Convergence of ANIL and MTL with Over-parameterized Deep Neural Nets

Throughout the paper, we use the squared loss as the objective function of training neural nets: $\ell(\hat{y}, y) := \frac{1}{2}\|\hat{y} - y\|_2^2$. To ease the presentation, we define the following meta-output functions.

Definition 2 (Meta-Output Functions)**.** *On any task $\mathcal{T} = (X, Y, X', Y')$, for the given adaptation steps τ, we define the meta-output function as*

$$F^\tau_\theta(X, X', Y') = f_{\theta^{test}}(X) \in \mathbb{R}^{nk} \tag{23}$$

where the adapted parameters θ^{test} is obtained as follows: use θ as the initial parameter and update it by τ steps of gradient descent on support samples and labels (X', Y'), with learning rate λ and loss function ℓ. Mathematically, $\forall j = 0, ..., \tau - 1$, we have

$$\theta = \theta_0, \;\; \theta^{test} = \theta_\tau, \text{ and } \theta_{j+1} = \theta_j - \lambda \nabla_{\theta_j} \ell(f_{\theta_j}(X'), Y') \tag{24}$$

Shorthand To make the notation uncluttered, we define some shorthand for the meta-output function,

- $F^\tau_\theta(\mathcal{X}, \mathcal{X}, \mathcal{Y}) \triangleq (F^\tau_\theta(X_i, X_i, Y_i))_{i=1}^N$: the concatenation of meta-outputs on *all* training tasks.
- $F^\tau_t \triangleq F^\tau_{\theta_t}$: shorthand for the meta-output function with parameters θ_t at training time t.

ANIL Loss With the squared loss function, the training objective of ANIL is expressed as

$$\mathcal{L}_{ANIL}(\theta) = \sum_{i=1}^{N} \ell(F_\theta^\tau(X_i, X_i, Y_i), Y_i) = \frac{1}{2}\sum_{i=1}^{N} \|F_\theta^\tau(X_i, X_i, Y_i) - Y_i\|_2^2 = \frac{1}{2}\|F_\theta^\tau(\mathcal{X}, \mathcal{X}, \mathcal{Y}) - \mathcal{Y}\|_2^2 \tag{25}$$

MTL Loss With the squared loss function, the objective of MTL is

$$\mathcal{L}_{MTL}(\hat{\theta}) = \sum_{i=1}^{N} \ell(\hat{f}_{\hat{\theta}}(X_i, i), Y_i) = \frac{1}{2}\sum_{i=1}^{N} \|\hat{f}_{\hat{\theta}}(X_i, i) - Y_i\|_2^2 = \frac{1}{2}\|\hat{f}_{\hat{\theta}}(\mathcal{X}) - \mathcal{Y}\|_2^2 \tag{26}$$

where we define the notation $\hat{f}_{\hat{\theta}}(\mathcal{X})$ to be $\mathrm{vec}(\{\hat{f}_{\hat{\theta}}(X_i, i)\}_{i=1}^N)$.

Tangent Kernels Now, we define tangent kernels for MTL and ANIL, following Wang et al. (2020). Denote h as the minimum width across hidden layers, i.e., $h = \min_{l\in[L-1]} h_l$. Then, the tangent kernels of MTL and ANIL are defined as

$$\Phi_{MTL} = \lim_{h\to\infty} \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X}) \cdot \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top \tag{27}$$

$$\Phi_{ANIL} = \lim_{h\to\infty} \nabla_{\theta_0} F_{\theta_0}^\tau(\mathcal{X}, \mathcal{X}, \mathcal{Y}) \cdot \nabla_{\theta_0} F_{\theta_0}^\tau(\mathcal{X}, \mathcal{X}, \mathcal{Y})^\top \tag{28}$$

Notice that by Wang et al. (2020), we know both kernels are deterministic positive-definite matrices, independent of the initializations θ_0 and $\hat{\theta}_0$.

Next, we present the following theorem that characterizes the global convergence of the above two algorithms on over-parametrized neural networks.

Theorem 3 (Global Convergence of ANIL and MTL with Over-parameterized Deep Neural Nets)**.** *Define*

$$\eta_0 = \min\left\{ \frac{2}{\sigma_{\min}(\Phi_{MTL}) + \sigma_{\max}(\Phi_{ANIL})}, \frac{2}{\sigma_{\min}(\Phi_{MTL}) + \sigma_{\max}(\Phi_{ANIL})} \right\}.$$

For arbitrarily small $\delta > 0$, there exists constants $R, \lambda_0, h^* > 0$ such that for networks with width greater than h^*, running gradient descent on $\mathcal{L}_{MTL}$ and $\mathcal{L}_{ANIL}$ with learning rate $\eta > \eta_0$ and inner-loop learning rate $\lambda < \lambda_0$, the following bounds on training losses hold true with probability at least $1 - \delta$ over random initialization,

$$\mathcal{L}_{ANIL}(\theta_t) \le \left(1 - \frac{1}{3}\eta_0 \cdot \sigma_{\min}(\Phi_{ANIL})\right)^{2t} R \tag{29}$$

$$\mathcal{L}_{MTL}(\hat{\theta}_t) \le \left(1 - \frac{1}{3}\eta_0 \cdot \sigma_{\min}(\Phi_{MTL})\right)^{2t} R \tag{30}$$

where $t \in \mathbb{N}$ is the number of training steps. Furthermore, the displacement of the parameters during the training process can be bounded by

$$\sup_{t\ge 0} \frac{1}{\sqrt{h}}\|\theta_t - \theta_0\|_2 = \mathcal{O}(h^{-\frac{1}{2}}), \quad \sup_{t\ge 0} \frac{1}{\sqrt{h}}\|\hat{\theta}_t - \hat{\theta}_0\|_2 = \mathcal{O}(h^{-\frac{1}{2}}) \tag{31}$$

Remarks. Notice the bounds in (31) are derived in the setting of NTK parameterization (see Appendix A.3). When switching to the standard parameterization, as shown by Theorem G.2 of Lee et al. (2019a), (31) is transformed to

$$\sup_{t\ge 0} \|\theta_t - \theta_0\|_2 = \mathcal{O}(h^{-\frac{1}{2}}), \quad \sup_{t\ge 0} \|\hat{\theta}_t - \hat{\theta}_0\|_2 = \mathcal{O}(h^{-\frac{1}{2}}), \tag{32}$$

indicating a closeness between the initial and trained parameters as the network width h is large.

Proof. For ANIL, the global convergence can be straightforwardly obtained by following the same steps of Theorem 4 of Wang et al. (2020), which proves the global convergence for MAML in the same setting[10].

[10]Notice the only difference between ANIL and MAML is the layers to optimize in the inner loop, where ANIL optimizes less layers than MAML. Hence, bounds on the inner loop optimization in Theorem 4 of Wang et al. (2020) cover that of ANIL, and the proof steps of that theorem applies to the case of ANIL.

For MTL, it can be viewed as a variant of MAML with multi-head neural nets and inner-loop learning rate $\tau = 0$, since it only has the outer-loop optimization. Then, the global convergence of MTL can also be straightforwardly obtained by following the proof steps of Theorem 4 from Wang et al. (2020). ■

Linearization at Large Width. The following corollary provides us a useful toolkit to analyze the training dynamics of both ANIL and MTL in the over-parametrization regime, which is adopted and rephrased from Wang et al. (2020) and Lee et al. (2019a).

Corollary 3.1 (Linearized (Meta) Output Functions)**.** *For arbitrarily small $\delta > 0$, there exists $h^* > 0$ s.t. as long as the network width h is greater than h^*, during the training of ANIL and MTL, with probability at least $1 - \delta$ over random initialization, the network parameters stay in the neighbourhood of the initialization s.t. $\theta_t \in \{\theta : \|\theta - \theta_0\|_2 \leq \mathcal{O}(1/h^2)\}$ or $\hat{\theta}_t \in \{\hat{\theta} : \|\hat{\theta} - \hat{\theta}_0\|_2 \leq \mathcal{O}(1/h^2)\}$, where $\theta_0 = \{\theta_0^{\leq L}, w_0\}$ and $\hat{\theta}_0 = \{\theta_0^{\leq L}\} \cup \{\hat{w}_0^{(i)}\}_{i\in[N]}$ are the initial parameters of networks trained by ANIL and MTL, respectively. Then, for any network trained by ANIL, its output on any $x \in \mathbb{R}^d$ is effectively linearized, i.e.,*

$$f_\theta(x) = f_{\theta_0}(x) + \nabla_{\theta_0} f_{\theta_0}(x)(\theta - \theta_0) + \mathcal{O}(\frac{1}{\sqrt{h}}) \tag{33}$$

Similarly, for any network trained by MTL, the output of the multi-head neural net on x with head index $i \in [N]$ is characterized by

$$\hat{f}_{\hat{\theta}}(x, i) = \hat{f}_{\hat{\theta}_0}(x, i) + \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(x, i)(\hat{\theta} - \hat{\theta}_0) + \mathcal{O}(\frac{1}{\sqrt{h}}) \tag{34}$$

Besides, the meta-output function is also effectively linearized, i.e., for any task $\mathcal{T} = (X, Y, X', Y')$,

$$F_\theta^\tau(X, X', Y') = F_{\theta_0}^\tau(X, X', Y') + \nabla_{\theta_0} F_{\theta_0}^\tau(X, X', Y')(\theta - \theta_0) + \mathcal{O}(\frac{1}{\sqrt{h}}), \tag{35}$$

where $F_{\theta_0}^\tau(X, X', Y')$ can be expressed as

$$F_{\theta_0}^\tau(X, X', Y') = f_{\theta_0}(X) + \hat{\mathcal{K}}_{w_0}(X, X')\hat{\mathcal{K}}_{w_0}^{-1}(X', X') \left(I - e^{-\lambda \hat{\mathcal{K}}_{w_0}(X', X')\tau}\right) [Y' - f_{\theta_0}(X')] + \mathcal{O}(\frac{1}{\sqrt{h}}), \tag{36}$$

and the gradient $\nabla_{\theta_0} F_{\theta_0}^\tau(X, X', Y')$ as[11]

$$\nabla_{\theta_0} F_{\theta_0}^\tau(X, X', Y') = \nabla_{\theta_0} f_{\theta_0}(X) - \hat{\mathcal{K}}_{w_0}(X, X')\hat{\mathcal{K}}_{w_0}^{-1}(X', X') \left(I - e^{-\lambda \hat{\mathcal{K}}_{w_0}(X', X')\tau}\right) \nabla_{\theta_0} f_{\theta_0}(X') + \mathcal{O}(\frac{1}{\sqrt{h}}), \tag{37}$$

with $\hat{\mathcal{K}}_{w_0}$ defined as

$$\hat{\mathcal{K}}_{w_0}(\cdot, *) = \nabla_w f_{\theta_0}(\cdot) \cdot \nabla_w f_{\theta_0}(*)^\top$$

Remarks. One can replace θ_0 in (35) with $\{\theta_0^{\leq L}, w_{test}\}$ or $\{\hat{\theta}_0^{\leq L}, w_{test}\}$, and similar results apply.

Proof. Notice that the proof of Theorem 3 above is based on Theorem 4 of Wang et al. (2020), which also proves that the trained parameters stay in the neighborhood of the initialization with radius of $\mathcal{O}(\frac{1}{\sqrt{h}})$. Hence, following the proof steps of Theorem 4 of Wang et al. (2020), one can also straightforwardly prove the same result for ANIL and MTL.

With the global convergence and the neighborhood results above, we can directly invoke Theorem H.1 of Lee et al. (2019a), and obtain (33), (34) and (35). Notice, the expressions in (36) and (37) are derived in Sec. 2.3.1 of Lee et al. (2019a). ■

[11] The proof of the gradient expression can be straightforwardly obtained by Lemma 6 of (Wang et al., 2020).

B.2. Training Dynamics of MTL and ANIL

Definition 4 (Empirical Tangent Kernels of ANIL and MTL)**.** *We define the following empirical tangent kernels of ANIL and MTL, in a similar way to (Wang et al., 2020; Lee et al., 2019a):*

$$\hat{\Phi}_{ANIL}(\mathcal{X},\mathcal{X}) = \nabla_{\theta_0} F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y}) \cdot \nabla_{\theta_0} F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})^{\top} \in \mathbb{R}^{Nn\times Nn} \tag{38}$$

$$\hat{\Phi}_{MTL}(\mathcal{X},\mathcal{X}) = \nabla_{\theta_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X}) \cdot \nabla_{\theta_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^{\top} \in \mathbb{R}^{Nn\times Nn} \tag{39}$$

Shorthand. To simplify expressions, we define the following shorthand. For any kernel function $\hat{\Phi}$, learning rate η and optimization steps t, we have

$$T^{\eta,t}_{\hat{\Phi}}(\cdot) = \hat{\Phi}^{-1}(\cdot,\cdot)\left(I - e^{-\eta\hat{\Phi}(\cdot,\cdot)t}\right) \tag{40}$$

Lemma 2 (ANIL and MTL in the Linearization Regime)**.** *With linearized output functions shown in Corollary 3.1, the training dynamics of ANIL and MTL under gradient descent on squared losses can be characterized by analytically solvable ODEs, giving rise to the solutions:*

- *ANIL.*
  - ☐ *Trained parameters at time t:*

$$\theta_t = \theta_0 + \nabla_{\theta_0} F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})^{\top} \hat{\Phi}^{-1}_{ANIL}(\mathcal{X},\mathcal{X})\left(I - e^{-\eta\hat{\Phi}_{ANIL}(\mathcal{X},\mathcal{X})t}\right)\left[\mathcal{Y} - F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})\right] + \mathcal{O}(\frac{1}{\sqrt{h}}) \tag{41}$$

  - ☐ *Prediction on any test task $\mathcal{T} = (X, Y, X', Y')$ with adaptation steps $\hat{\tau}$ (i.e., we take the hidden layers of the trained network $\theta^{<L}$ and append a randomly initialized head w_{test} to fine-tune):*

$$\begin{aligned} &F^{\hat{\tau}}_{\theta^{test}_t}(X,X',Y') \\ &= F^{\hat{\tau}}_{\theta^{test}_0}(X,X',Y') + \nabla_{\theta^{<L}_0} F^{\hat{\tau}}_{\theta^{test}_0}(X,X',Y')\nabla_{\theta^{<L}_0} F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})^{\top} T^{\eta,t}_{\hat{\Phi}_{ANIL}}(\mathcal{X})\left[\mathcal{Y} - F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})\right] + \mathcal{O}(\frac{1}{\sqrt{h}}) \end{aligned} \tag{42}$$

    where $\hat{\theta}^{test}_t = \{\hat{\theta}^{<L}_t, w_{test}\}$ and $\hat{\theta}^{test}_0 = \{\hat{\theta}^{<L}_0, w_{test}\}$.
- *MTL.*
  - ☐ *Trained parameters:*

$$\hat{\theta}_t = \hat{\theta}_0 + \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^{\top} T^{\eta,t}_{\hat{\Phi}_{MTL}}(\mathcal{X})\left[\mathcal{Y} - \hat{f}_{\hat{\theta}_0}(\mathcal{X})\right] + \mathcal{O}(\frac{1}{\sqrt{h}}) \tag{43}$$

  - ☐ *Prediction on test task $\mathcal{T} = (X, Y, X', Y')$ with adaptation steps $\hat{\tau}$ (i.e., we take the hidden layers of the trained network $\hat{\theta}^{<L}$ and append a randomly initialized head w_{test} to fine-tune):*

$$\begin{aligned} &F^{\hat{\tau}}_{\hat{\theta}^{test}_t}(X,X',Y') \\ &= F^{\hat{\tau}}_{\hat{\theta}^{test}_0}(X,X',Y') + \nabla_{\hat{\theta}^{<L}_0} F^{\hat{\tau}}_{\hat{\theta}^{test}_0}(X,X',Y')\nabla_{\hat{\theta}^{<L}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^{\top} T^{\eta,t}_{\hat{\Phi}_{MTL}}(\mathcal{X})\left[\mathcal{Y} - \hat{f}_{\hat{\theta}_0}(\mathcal{X})\right] + \mathcal{O}(\frac{1}{\sqrt{h}}) \end{aligned} \tag{44}$$

    where $\hat{\theta}^{test}_t = \{\hat{\theta}^{<L}_t, w_{test}\}$, $\hat{\theta}^{test}_0 = \{\hat{\theta}^{<L}_0, w_{test}\}$.

Proof. Similar to Sec. 2.2 of Lee et al. (2019a), with linearized functions (34) and (35), the training dynamics of MTL and ANIL under gradient flow with squared losses are governed by the ODEs,

- Training dynamics of ANIL.

$$\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = -\eta\nabla_{\theta_0} F_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})^{\top}\left(F_{\theta_t}(\mathcal{X},\mathcal{X},\mathcal{Y}) - \mathcal{Y}\right) \tag{45}$$

$$\frac{\mathrm{d}F_{\theta_t}(\mathcal{X},\mathcal{X},\mathcal{Y})}{\mathrm{d}t} = -\eta\hat{\Phi}_{ANIL}(\mathcal{X},\mathcal{X})\left(F_{\theta_t}(\mathcal{X},\mathcal{X},\mathcal{Y}) - \mathcal{Y}\right) \tag{46}$$

  Solving the set of ODEs, we obtain the solution to θ_t as

$$\theta_t = \theta_0 - \nabla_{\theta_0} F_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})^{\top}\hat{\Phi}_{ANIL}(\mathcal{X},\mathcal{X})^{-1}\left(I - e^{-\eta\hat{\Phi}_{ANIL}(\mathcal{X},\mathcal{X})t}\right)\left(F_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y}) - \mathcal{Y}\right) \tag{47}$$

  up to an error of $\mathcal{O}(\frac{1}{\sqrt{h}})$. See Theorem H.1 of Lee et al. (2019a) for the bound on the error across training.

- Training dynamics of MTL.

$$\frac{\mathrm{d}\hat{\theta}_t}{\mathrm{d}t} = -\eta \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top \left(\hat{f}_{\theta_t}(\mathcal{X}) - \mathcal{Y}\right) \tag{48}$$

$$\frac{\mathrm{d}\hat{f}_{\theta_t}(\mathcal{X})}{\mathrm{d}t} = -\eta \hat{\Phi}_{MTL}(\mathcal{X}, \mathcal{X}) \left(\hat{f}_{\theta_0}(\mathcal{X}) - \mathcal{Y}\right) \tag{49}$$

Solving the set of ODEs, we obtain the solution to $\hat{\theta}_t$ as

$$\hat{\theta}_t = \hat{\theta}_0 - \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top \hat{\Phi}_{MTL}(\mathcal{X}, \mathcal{X})^{-1} \left(I - e^{-\eta \hat{\Phi}_{MTL}(\mathcal{X}, \mathcal{X})t}\right) \left(\hat{f}_{\theta_t}(\mathcal{X}) - \mathcal{Y}\right) \tag{50}$$

up to an error of $\mathcal{O}(\frac{1}{\sqrt{h}})$. See Theorem H.1 of Lee et al. (2019a) for the bound on the error across training.

Now, with the derived expressions of trained parameters, we can certainly plug them in the linearized functions (34) and (35) to obtain the outputs of trained ANIL and MTL models. Notice that during test, the predictions of ANIL and MTL are obtained from a fine-tuned *test head* that are randomly initialized (see Sec. 3.4 for details). Thus, we need to take care of the test heads when plugging trained parameters into the linearized functions. Specifically, for an arbitrary test task $\mathcal{T} = (X, Y, X', Y')$, the test predictions of ANIL and MTL are derived below.

- Test predictions of ANIL. For notational simplicity, we define

$$\hat{\mathcal{K}}_t(\cdot, *) = \nabla_{w_{test}} f_{\theta_t^{test}}(\cdot) \nabla_{w_{test}} f_{\theta_t^{test}}(*)^\top$$

Then, since the fine-tuning is on the test head w_{test}, following the Sec. 2.3.1. of Lee et al. (2019a), we know

$$F^{\hat{\tau}}_{\theta_t^{test}}(X, X', Y') = f_{\theta_t^{test}}(X) + \hat{\mathcal{K}}_t(X, X') T^{\lambda, \hat{\tau}}_{\hat{\mathcal{K}}_t}(X') \left(Y' - f_{\theta_t^{test}}(X')\right) + \mathcal{O}(\frac{1}{\sqrt{h}}) \tag{51}$$

where

$$\begin{aligned}
& f_{\theta_t^{test}}(X) \qquad (52)\\
&= f_{\theta_0^{test}}(X) + \nabla_{\theta_0^{test}} f_{\theta_0^{test}}(X)(\theta_t^{test} - \theta_0^{test}) + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
&= f_{\theta_0^{test}}(X) + \nabla_{\theta_0^{<L}} f_{\theta_0^{test}}(X)(\theta_t^{<L} - \theta_0^{<L}) + \nabla_{w_{test}} f_{\theta_0^{test}}(X)(w_{test} - w_{test}) + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
&= f_{\theta_0^{test}}(X) + \nabla_{\theta_0^{<L}} f_{\theta_0^{test}}(X)(\theta_t^{<L} - \theta_0^{<L}) + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
&= f_{\theta_0^{test}}(X) + \nabla_{\theta_0^{<L}} f_{\theta_0^{test}}(X) \nabla_{\theta_0^{<L}} F_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})^\top \hat{\Phi}_{ANIL}(\mathcal{X}, \mathcal{X})^{-1} \left(I - e^{-\eta \hat{\Phi}_{ANIL}(\mathcal{X}, \mathcal{X})t}\right) (\mathcal{Y} - F_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})) \\
&\quad + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
&= f_{\theta_0^{test}}(X) + \nabla_{\theta_0^{<L}} f_{\theta_0^{test}}(X) \nabla_{\theta_0^{<L}} F_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})^\top T^{\eta, t}_{\hat{\Phi}_{ANIL}}(\mathcal{X}) (\mathcal{Y} - F_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})) + \mathcal{O}(\frac{1}{\sqrt{h}})
\end{aligned}$$

and

$$\nabla_{w_{test}} f_{\theta_t^{test}}(X) = \nabla_{w_{test}} f_{\theta_0^{test}}(X) + \mathcal{O}(\frac{1}{\sqrt{h}}) \tag{53}$$

Pluging in everything, we have

$$
\begin{aligned}
& F^{\hat{\tau}}_{\theta^{test}_t}(X, X', Y') \qquad (54)\\
&= f_{\theta^{test}_0}(X) + \nabla_{\theta^{<L}_0} f_{\theta^{test}_0}(X) \nabla_{\theta^{<L}_0} F_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})^\top T^{\eta,t}_{\hat{\Phi}_{ANIL}}(\mathcal{X}) \left(\mathcal{Y} - F_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})\right)\\
&\quad + \hat{\mathcal{K}}_0(X, X') T^{\lambda,\hat{\tau}}_{\hat{\mathcal{K}}_0}(X') \left(Y' - f_{\theta^{test}_t}(X')\right) + \mathcal{O}(\frac{1}{\sqrt{h}})\\
&= f_{\theta^{test}_0}(X) + \hat{\mathcal{K}}_0(X, X') T^{\lambda,\hat{\tau}}_{\hat{\mathcal{K}}_0}(X')(Y' - f_{\theta^{test}_0}(X'))\\
&\quad + \left(\nabla_{\theta^{<L}_0} f_{\theta^{test}_0}(X) - \hat{\mathcal{K}}_0(X, X') T^{\lambda,\hat{\tau}}_{\hat{\mathcal{K}}_0}(X') \nabla_{\theta^{<L}_0} f_{\theta^{test}_0}(X')\right) \nabla_{\theta^{<L}_0} F_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})^\top T^{\eta,t}_{\hat{\Phi}_{ANIL}}(\mathcal{X}) \left(\mathcal{Y} - F_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})\right)\\
&\quad + \mathcal{O}(\frac{1}{\sqrt{h}})\\
&= F^{\hat{\tau}}_{\theta^{test}_0}(X, X', Y') + \nabla_{\theta^{<L}_0} F^{\hat{\tau}}_{\theta^{test}_0}(X, X', Y') \nabla_{\theta^{<L}_0} F^{\tau}_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})^\top T^{\eta,t}_{\hat{\Phi}_{ANIL}}(\mathcal{X}) \left[\mathcal{Y} - F^{\tau}_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})\right] + \mathcal{O}(\frac{1}{\sqrt{h}})
\end{aligned}
$$

- Test prediction of MTL. Following the derivation for the test prediction of ANIL above, one can straightforwardly derive that

$$
F^{\hat{\tau}}_{\hat{\theta}^{test}_t}(X, X', Y') = F^{\hat{\tau}}_{\hat{\theta}^{test}_0}(X, X', Y') + \nabla_{\hat{\theta}^{<L}_0} F^{\hat{\tau}}_{\hat{\theta}^{test}_0}(X, X', Y') \nabla_{\hat{\theta}^{<L}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top T^{\eta,t}_{\hat{\Phi}_{MTL}}(\mathcal{X}) \left[\mathcal{Y} - \hat{f}_{\hat{\theta}_0}(\mathcal{X})\right] + \mathcal{O}(\frac{1}{\sqrt{h}})
$$

■

B.3. Derivation of Kernels and Outputs for ANIL and MTL.

Notation 1 (NTK and NNGP)**.** *We denote*

- $\Theta(\cdot, *)$*: kernel function of Neural Tangent Kernel (NTK).*
- $\mathcal{K}(\cdot, *)$*: kernel function of Neural Network Gaussian Process (NNGP).*

Equivalence to Kernels Lee et al. (2019a) shows that as the network width h approaches infinity, for parameter initialization $\theta_0 = \{\theta_0^{<L}, w_0\}$, we have the following equivalence relations,

$$
\nabla_{\theta_0} f_{\theta_0}(\cdot) \nabla_{\theta_0} f_{\theta_0}(*)^\top = \Theta(\cdot, *) \qquad (55)
$$

$$
\nabla_{w} f_{\theta_0}(\cdot) \nabla_{\theta_0} f_{w}(*)^\top = \mathcal{K}(\cdot, *) \qquad (56)
$$

Lemma 3 (ANIL and MTL Kernels)**.** *As the width of neural nets increases to infinity, i.e., $h \to \infty$, we define the following kernels for ANIL and MTL, and they converge to corresponding analytical expressions shown below.*

- ***ANIL kernels.***
  - ☐ $\Phi_{ANIL}(\mathcal{X}, \mathcal{X}) = \nabla_{\theta_0} F^{\tau}_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y}) \cdot \nabla_{\theta_0} F^{\tau}_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})^\top$ *is a block matrix of $N \times N$ blocks. $\forall i, j \in [N]$, its (i, j)-th block is*

$$
[\Phi_{\text{ANIL}}(\mathcal{X}, \mathcal{X})]_{ij} = e^{-\lambda \mathcal{K}(X_i, X_i)\tau} \Theta(X_i, X_j) e^{-\lambda \mathcal{K}(X_j, X_j)\tau},
$$

  - ☐ $\Phi'_{ANIL}((X, X', \hat{\tau}), \mathcal{X}) = \nabla_{\theta^{<L}_0} F^{\hat{\tau}}_{\theta^{test}_0}(X, X', Y') \nabla_{\theta^{<L}_0} F^{\tau}_{\theta_0}(\mathcal{X}, \mathcal{X}, \mathcal{Y})^\top$ *is a block matrix of $1 \times N$ blocks, with the $(1, j)$-th block as*

$$
[\Phi'_{ANIL}((X, X', \hat{\tau}), \mathcal{X})]_{1j} = \left[\Theta(X, X_j) - \mathcal{K}(X, X') T^{\hat{\tau}}_{\mathcal{K}}(X') \Theta(X', X_j)\right] e^{-\lambda \mathcal{K}(X_j, X_j)\tau}
$$

- ***MTL Kernels.***
  - ☐ $\Phi_{MTL}(\mathcal{X}, \mathcal{X}) = \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X}) \cdot \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top$ *is also a block matrix of $N \times N$ blocks. $\forall i, j \in [N]$, its (i, j)-th block is*

$$
[\Phi_{MTL}(\mathcal{X}, \mathcal{X})]_{ij} = \Theta(X_i, X_j) - \mathbb{1}[i \neq j] \mathcal{K}(X_i, X_j),
$$

□ $\Phi'_{MTL}((X,X',\hat{\tau}),\mathcal{X}) = \nabla_{\hat{\theta}_0^{<L}} F^{\hat{\tau}}_{\hat{\theta}_0^{test}}(X,X',Y')\nabla_{\hat{\theta}_0^{<L}}\hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top$ *is a block matrix of* $1\times N$ *blocks, with the* $(1,j)$*-th block as*

$$\begin{aligned}[\Phi'_{MTL}((X,X',\hat{\tau}),\mathcal{X})]_{1j} &= \Theta(X,X_j) - \mathcal{K}(X,X_j) - \mathcal{K}(X,X')T^{\hat{\tau}}_{\mathcal{K}}(X')\Big[\Theta(X',X_j) - \mathcal{K}(X',X_j)\Big] \\ &= \Theta(X,X_j) - \mathcal{K}(X,X')T^{\hat{\tau}}_{\mathcal{K}}(X')\Theta(X',X_j) - \mathcal{K}(X,X')e^{-\lambda\mathcal{K}(X',X')\hat{\tau}}\mathcal{K}(X',X_j)\end{aligned} \tag{57}$$

Proof. The proof is presented in the same structure as the lemma statement above.

- **ANIL Kernels**

□ $\Phi_{ANIL}(\mathcal{X},\mathcal{X})$. With (37), we know

$$\begin{aligned}\nabla_{\theta_0}F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y}) &= \left(\nabla_{\theta_0}F^{\tau}_{\theta_0}(X_i,X_i,Y_i)\right)_{i=1}^N \\ &= \left(\nabla_{\theta_0}f_{\theta_0}(X_i) - \mathcal{K}(X_i,X_i)\mathcal{K}^{-1}(X_i,X_i)\left(I - e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\right)\nabla_{\theta_0}f_{\theta_0}(X_i)\right)_{i=1}^N \\ &= \left(e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\nabla_{\theta_0}f_{\theta_0}(X_i)\right)_{i=1}^N\end{aligned} \tag{58}$$

Thus, the (i,j)-th block of $\Phi_{ANIL}(\mathcal{X},\mathcal{X}) = \nabla_{\theta_0}F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})\cdot\nabla_{\theta_0}F^{\tau}_{\theta_0}(\mathcal{X},\mathcal{X},\mathcal{Y})^\top$ is

$$\begin{aligned}[\Phi_{ANIL}(\mathcal{X},\mathcal{X})]_{ij} &= \nabla_{\theta_0}F^{\tau}_{\theta_0}(X_i,X_i,Y_i)\nabla_{\theta_0}F^{\tau}_{\theta_0}(X_j,X_j,Y_j)^\top \\ &= e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\nabla_{\theta_0}f_{\theta_0}(X_i)f_{\theta_0}(X_j)^\top e^{-\lambda\mathcal{K}(X_j,X_j)\tau} \\ &= e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\Theta(X_i,X_j)e^{-\lambda\mathcal{K}(X_j,X_j)\tau}\end{aligned} \tag{59}$$

Then, the whole matrix can be expressed as

$$\Phi_{ANIL}(\mathcal{X},\mathcal{X}) = \mathrm{diag}\left(\{e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\}_{i=1}^N\right)\cdot\Theta(\mathcal{X},\mathcal{X})\cdot\mathrm{diag}\left(\{e^{-\lambda\mathcal{K}(X_j,X_j)\tau}\}_{j=1}^N\right) \tag{60}$$

where $\mathrm{diag}\left(\{e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\}_{i=1}^N\right)$ is a diagonal block matrix with the i-th block as $e^{-\lambda\mathcal{K}(X_i,X_i)\tau}$.

□ $\Phi'_{ANIL}((X,X',\hat{\tau}),\mathcal{X})$. With (37), we can derive that

$$\begin{aligned}&[\Phi'_{ANIL}((X,X',\hat{\tau}),\mathcal{X})]_{1j} \\ &= \nabla_{\theta_0^{<L}}F^{\hat{\tau}}_{\theta_0^{test}}(X,X',Y')\nabla_{\theta_0^{<L}}F^{\tau}_{\theta_0}(X_j,X_j,Y_j)^\top \\ &= \left(\nabla_{\theta_0^{<L}}f_{\theta_0^{test}}(X) - \mathcal{K}(X,X')\mathcal{K}^{-1}(X',X')\left(I - e^{-\lambda\mathcal{K}(X',X')\hat{\tau}}\right)\nabla_{\theta_0^{<L}}f_{\theta_0^{<L}}(X')\right) \\ &\quad\cdot\left(\nabla_{\theta_0^{<L}}f_{\theta_0}(X_j) - \mathcal{K}(X_j,X_j)\mathcal{K}^{-1}(X_j,X_j)\left(I - e^{-\lambda\mathcal{K}(X_j,X_j)\tau}\right)\nabla_{\theta_0^{<L}}f_{\theta_0}(X_j)\right)^\top \\ &= \left(\nabla_{\theta_0^{<L}}f_{\theta_0^{test}}(X) - \mathcal{K}(X,X')\mathcal{K}^{-1}(X',X')\left(I - e^{-\lambda\mathcal{K}(X',X')\hat{\tau}}\right)\nabla_{\theta_0^{<L}}f_{\theta_0}(X')\right)\cdot\nabla_{\theta_0^{<L}}f_{\theta_0}(X_j)^\top e^{-\lambda\mathcal{K}(X_j,X_j)\tau} \\ &= \left[(\Theta(X,X_j) - \mathcal{K}(X,X_j)) - \mathcal{K}(X,X')\mathcal{K}^{-1}(X',X')\left(I - e^{-\lambda\mathcal{K}(X',X')\hat{\tau}}\right)(\Theta(X',X_j) - \mathcal{K}(X',X_j))\right]e^{-\lambda\mathcal{K}(X_j,X_j)\tau} \\ &= \left[(\Theta(X,X_j) - \mathcal{K}(X,X_j)) - \mathcal{K}(X,X')T^{\lambda,\hat{\tau}}_{\mathcal{K}}(X')(\Theta(X',X_j) - \mathcal{K}(X',X_j))\right]e^{-\lambda\mathcal{K}(X_j,X_j)\tau}\end{aligned} \tag{61}$$

where we used the equivalence

$$\nabla_{\theta_0^{<L}}f_{\theta_0^{test}}(\cdot)\cdot\nabla_{\theta_0^{<L}}f_{\theta_0}(*)^\top = \Theta(\cdot,*) - \mathcal{K}(\cdot,*) \tag{62}$$

in the infinite width limit at initialization.

- **MTL**

□ $\Phi_{MTL}(\mathcal{X},\mathcal{X}) = \nabla_{\hat{\theta}_0}\hat{f}_{\hat{\theta}_0}(\mathcal{X})\cdot\nabla_{\hat{\theta}_0}\hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top$. Notice that for any input with head index i, we have

$$\begin{aligned}\nabla_{\hat{\theta}_0}\hat{f}_{\hat{\theta}_0}(\cdot,i) &= \nabla_{\hat{\theta}_0^{<L}}\hat{f}_{\hat{\theta}_0}(\cdot,i) + \sum_{j=1}^{N+1}\nabla_{\hat{w}^{(j)}}\hat{f}_{\hat{\theta}_0}(\cdot,i) \\ &= \nabla_{\hat{\theta}_0^{<L}}\hat{f}_{\hat{\theta}_0}(\cdot,i) + \nabla_{\hat{w}^{(i)}}\hat{f}_{\hat{\theta}_0}(\cdot,i)\end{aligned} \tag{63}$$

since for $j \neq i$, we have $\nabla_{\hat{w}^{(j)}} \hat{f}_{\hat{\theta}_0}(x, i) = 0$ based on the multi-head structure.
Thus, we can write down the (i, j)-th block of $\Phi_{MTL}(\mathcal{X}, \mathcal{X})$ as

$$\begin{aligned}[\Phi_{MTL}(\mathcal{X}, \mathcal{X})]_{ij} &= \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(X_i, i) \nabla_{\hat{\theta}_0} \hat{f}_{\hat{\theta}_0}(X_j, j)^\top \\ &= \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_i, i) \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_j, j)^\top + \nabla_{\hat{w}^{(i)}} \hat{f}_{\hat{\theta}_0}(X_i, i) \nabla_{\hat{w}^{(j)}} \hat{f}_{\hat{\theta}_0}(X_j, j)^\top\end{aligned}$$

Note that for $i \neq j$, we have $\nabla_{\hat{w}^{(i)}} \hat{f}_{\hat{\theta}_0}(X_i, i) \nabla_{\hat{w}^{(j)}} \hat{f}_{\hat{\theta}_0}(X_j, j)^\top = 0$, since $\hat{w}^{(i)}$ and $\hat{w}^{(j)}$ are in different dimensions of $\hat{\theta}$. Thus,

* as $i \neq j$, we have[12]

$$[\Phi_{MTL}(\mathcal{X}, \mathcal{X})]_{ij} = \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_i, i) \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_j, j)^\top = \Theta(X_i, X_j) - \mathcal{K}(X_i, X_j)$$

* as $i = j$, we have

$$[\Phi_{MTL}(\mathcal{X}, \mathcal{X})]_{ii} = \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_i, i) \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_i, i)^\top + \nabla_{\hat{w}^{(i)}} \hat{f}_{\hat{\theta}_0}(X_i, i) \nabla_{\hat{w}^{(i)}} \hat{f}_{\hat{\theta}_0}(X_i, i)^\top = \Theta(X_i, X_i)$$

In conclusion, for $i, j \in [N]$, we have

$$[\Phi_{MTL}(\mathcal{X}, \mathcal{X})]_{ij} = \Theta(X_i, X_j) - \mathbb{1}[i \neq j] \mathcal{K}(X_i, X_j)$$

Thus,

$$\Phi_{MTL}(\mathcal{X}, \mathcal{X}) = \Theta(\mathcal{X}, \mathcal{X}) - \mathcal{K}(\mathcal{X}, \mathcal{X}) + \operatorname{diag}\left(\{\mathcal{K}(X_i, X_i)\}_{i=1}^N\right) \tag{64}$$

- $\Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X}) = \nabla_{\hat{\theta}_0^{<L}} \hat{F}^{\hat{\tau}}_{\hat{\theta}_0}(X, X', Y') \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top$.
  Based on (63), following (61), we can express the $(1, j)$-th block of $\Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X})$ as

$$\begin{aligned}&[\Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X})]_{1j} \\ &= \nabla_{\hat{\theta}_0^{<L}} F^{\hat{\tau}}_{\hat{\theta}_0^{test}}(X, X', Y') \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(\mathcal{X})^\top \\ &= \left(\nabla_{\hat{\theta}_0^{<L}} f_{\hat{\theta}^{test}}(X) - \mathcal{K}(X, X') \mathcal{K}^{-1}(X', X') \left(I - e^{-\lambda \mathcal{K}(X', X')\hat{\tau}}\right) \nabla_{\hat{\theta}_0^{<L}} f_{\hat{\theta}^{test}}(X')\right) \cdot \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_j, j)^\top \\ &= \nabla_{\hat{\theta}_0^{<L}} f_{\hat{\theta}^{test}}(X) \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_j, j)^\top - \mathcal{K}(X, X') \mathcal{K}^{-1}(X', X') \left(I - e^{-\lambda \mathcal{K}(X', X')\hat{\tau}}\right) \nabla_{\hat{\theta}_0^{<L}} f_{\hat{\theta}^{test}}(X') \nabla_{\hat{\theta}_0^{<L}} \hat{f}_{\hat{\theta}_0}(X_j, j)^\top \\ &= \left[\Theta(X, X_j) - \mathcal{K}(X, X_j)\right] - \mathcal{K}(X, X') \mathcal{K}^{-1}(X', X') \left(I - e^{-\lambda \mathcal{K}(X', X')\hat{\tau}}\right) \left[\Theta(X', X_j) - \mathcal{K}(X', X_j)\right] \\ &= \Theta(X, X_j) - \mathcal{K}(X, X') T^{\hat{\tau}}_{\mathcal{K}}(X') \Theta(X', X_j) - \mathcal{K}(X, X') \mathcal{K}^{-1}(X', X') e^{-\lambda \mathcal{K}(X', X')\hat{\tau}} \mathcal{K}(X', X_j)\end{aligned} \tag{65}$$

■

Remarks. Notice that (61) and (65) indicate the following relation:

$$[\Phi'_{ANIL}((X, X', \hat{\tau}), \mathcal{X})]_{1j} = [\Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X})]_{1j} \, e^{-\lambda \mathcal{K}(X_j, X_j)\tau}$$

Furthermore, it is straightforward to show that

$$\Phi'_{ANIL}((X, X', \hat{\tau}), \mathcal{X}) = \Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X}) \cdot \operatorname{diag}\left(\{e^{-\lambda \mathcal{K}(X_j, X_j)\tau}\}_{j=1}^N\right) \tag{66}$$

where $\operatorname{diag}\left(\{e^{-\lambda \mathcal{K}(X_j, X_j)\tau}\}_{j=1}^N\right)$ is a diagonal block matrix with the j-th block as $e^{-\lambda \mathcal{K}(X_j, X_j)\tau}$.

[12] The following equivalence can be straightforwardly dervied based on Appendix D and E of (Lee et al., 2019a).

B.3.1. Proof of Lemma 1

Now, we can prove Lemma 1 shown in Sec. 4.3, by leveraging Lemma 2 and Lemma 3 that we just proved. In particular, without loss of generality, following Arora et al. (2019), we assume the outputs of randomly initialized networks have a much smaller magnitude compared with the magnitude of training labels such that $\|f_{\theta_0}(x)\|_2 \leq \|y\|_2 \leq \mathcal{O}(h^{-\frac{1}{2}})$. Notice this can be always achieved by choosing smaller initialization scale or scaling down the neural net output (Arora et al., 2019), without any effect on the training dynamics and the predictions, up to a width-dependent factor on the learning rate. Below, we present the steps of the proof in detail.

Proof of Lemma 1. Plugging the kernels expressions derived by Lemma 3 into (42) and (44), and combining with the fact that $\lim_{h\to\infty} \hat{\mathcal{K}}_{w_0} \to \mathcal{K}$ (proved by Corollary 1 of Lee et al. (2019a)), we obtain the expressions of (16) and (15) in Lemma 1 in the infinite width limit. Notice that we consider sufficiently large width h, then the discrepancy between the infinite-width kernels and their finite-width counter-parts (i.e., the finite-width correction) is bounded by $\mathcal{O}(\frac{1}{\sqrt{h}})$ with arbitrarily large probability, indicated by Theorem 1 of Hanin & Nica (2020). Thus, the finite-width correction terms are absorbed into the $\mathcal{O}(\frac{1}{\sqrt{h}})$ error terms in (42) and (44).

■

B.3.2. Discrepancy between Predictions of ANIL and MTL

Based on (60), (64), and (66), for small $\lambda\tau$, the discrepancy between ANIL and MTL predictions can be written as (Note: we consider neural nets trained under ANIL and MTL for infinite time $t=\infty$, then take their parameters θ_∞ and $\hat{\theta}_\infty$ for test on any task $\mathcal{T}=(X,Y,X',Y')$),

$$
\begin{aligned}
&F_{ANIL}(X,X',Y') - F_{MTL}(X,X',Y') \\
&= F_{\theta_\infty^{test}}(X,X',Y') - F_{\hat{\theta}_\infty^{test}}(X,X',Y') \\
&= \left[\Phi'_{ANIL}((X,X',\hat{\tau}),\mathcal{X})\Phi^{-1}_{ANIL}(\mathcal{X},\mathcal{X}) - \Phi'_{MTL}((X,X',\hat{\tau}),\mathcal{X})\Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\right]\mathcal{Y} \\
&\quad - \Phi'_{ANIL}((X,X',\hat{\tau}),\mathcal{X})\Phi^{-1}_{ANIL}(\mathcal{X},\mathcal{X})\overbrace{G_\tau(\mathcal{X},\mathcal{X}',\mathcal{Y}')}^{=\mathcal{O}(\lambda\tau\sigma_{\max}(\mathcal{K}))} + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
&= \Big[\Phi'_{MTL}((X,X',\hat{\tau}),\mathcal{X})\cdot \mathrm{diag}\left(\{e^{-\lambda\mathcal{K}(X_j,X_j)\tau}\}_{j=1}^N\right)\mathrm{diag}\left(\{e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\}_{i=1}^N\right)^{-1}\Theta(\mathcal{X},\mathcal{X})^{-1}\mathrm{diag}\left(\{e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\}_{i=1}^N\right)^{-1} \\
&\quad - \Phi'_{MTL}((X,X',\hat{\tau}),\mathcal{X})\Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\Big]\mathcal{Y} + \mathcal{O}(\lambda\tau) + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
&= \Phi'_{MTL}((X,X',\hat{\tau}),\mathcal{X})\left[\Theta(\mathcal{X},\mathcal{X})^{-1}\mathrm{diag}\left(\{e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\}_{i=1}^N\right)^{-1} - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\right]\mathcal{Y} + \mathcal{O}(\lambda\tau) + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
&= \Phi'_{MTL}((X,X',\hat{\tau}),\mathcal{X})\Big[\Theta(\mathcal{X},\mathcal{X})^{-1}\underbrace{\mathrm{diag}\left(\{e^{-\lambda\mathcal{K}(X_i,X_i)\tau}\}_{i=1}^N\right)^{-1}}_{=I+\mathcal{O}(\lambda\tau\sigma_{\max}(\mathcal{K}))} - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\Big]\mathcal{Y} + \mathcal{O}(\lambda\tau\sigma_{\max}(\mathcal{K})) + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
&= \Phi'_{MTL}((X,X',\hat{\tau}),\mathcal{X})\Big[\Theta(\mathcal{X},\mathcal{X})^{-1} - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\Big]\mathcal{Y} + \mathcal{O}(\lambda\tau\sigma_{\max}(\mathcal{K})) + \mathcal{O}(\frac{1}{\sqrt{h}})
\end{aligned} \tag{67}
$$

where $\sigma_{\max}(\mathcal{K}) \triangleq \max\left(\{\sigma_{\max}(\mathcal{K}(X_i,X_i))\}_{i=1}^N\right)$.

Remarks. (67) indicates that for small $\lambda\tau$, the discrepancy between ANIL's and MTL's test predictions is determined by

$$
\Theta(\mathcal{X},\mathcal{X})^{-1} - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X}). \tag{68}
$$

Thus, if this difference vanishes in some limit, ANIL and MTL will output almost the same predictions on any test task.

B.4. Kernel Structures for Deep ReLU Nets

Setup. As described by Sec. 4.3, we focus on networks that adopt ReLU activation and He's initialization, and we consider the inputs are normalized to have unit variance, without loss of generality. Besides, we also assume any pair of samples in the training set are distinct.

NTK and NNGP Kernel Structures. Xiao et al. (2020) shows that for ReLU networks with He's initialization and unit-variance inputs, the corresponding NTK and NNGP kernels have some special structures. Specifically, at large depth, the spectra of these kernels can be characterized explicitly, as shown by Lemma 4 below, which is adopted and rephrased from the Appendix C.1 of Xiao et al. (2020).

Lemma 4 (Kernel Structures of NTK and NNGP)**.** *For sufficiently large depth L, NTK and NNGP kernels have the following expressions*[13] *(Note: we use the superscript $^{(L)}$ to mark the kernels' dependence on the depth L)*

$$\Theta^{(L)}(\mathcal{X},\mathcal{X}) = L\left(\frac{1}{4}\mathbf{1}_{Nn}\mathbf{1}_{Nn}^\top + \frac{3}{4}I\right) + \mathbf{A}^{(L)}_{\mathcal{X},\mathcal{X}} \tag{69}$$

$$\mathcal{K}^{(L)}(\mathcal{X},\mathcal{X}) = \mathbf{1}_{Nn}\mathbf{1}_{Nn}^\top + \frac{1}{L^2}\mathbf{B}^{(L)}_{\mathcal{X},\mathcal{X}} \tag{70}$$

where $\mathbf{A}^{(L)}_{\mathcal{X},\mathcal{X}}, \mathbf{B}^{(L)}_{\mathcal{X},\mathcal{X}} \in \mathbb{R}^{Nn\times Nn}$ is a symmetric matrix with elements of $\mathcal{O}(1)$.

The eigenvalues of $\Theta^{(L)}(\mathcal{X},\mathcal{X})$ and $\mathcal{K}^{(L)}(\mathcal{X},\mathcal{X})$ are all positive since Θ and $\mathcal{K}$ are guaranteed to be positive definite, and these eigenvalues can be characterized as

$$\begin{cases}\sigma_{\max}(\Theta(\mathcal{X},\mathcal{X})) = \frac{Nn+3}{4}L + \mathcal{O}(1)\\ \sigma_{\text{bulk}}(\Theta(\mathcal{X},\mathcal{X})) = \frac{3}{4}L + \mathcal{O}(1)\end{cases} \qquad \begin{cases}\sigma_{\max}(\mathcal{K}(\mathcal{X},\mathcal{X})) = Nn + \mathcal{O}(\frac{1}{L}^2)\\ \sigma_{\text{bulk}}(\mathcal{K}(\mathcal{X},\mathcal{X})) = \mathcal{O}(\frac{1}{L^2})\end{cases} \tag{71}$$

where $\sigma_{\text{bulk}}(\cdot)$ denotes the eigenvalues besides the largest eigenvalue.

Discrepancy between Kernel Inverses. As shown by Appendix B.3.2, the discrepancy between the predictions of ANIL and MTL is controlled by (68), i.e., $\Theta^{-1}(\mathcal{X},\mathcal{X}) - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})$. In the lemma below, we study (68) in the setting of ReLU nets with He's initialization, and prove a bound over the operator norm of (68).

Lemma 5 (Discrepancy between Kernel Inverses)**.** *There exists $L^* \in \mathbb{N}^+$ s.t. for $L \geq L^*$,*

$$\begin{cases}\sigma_{\max}\left(\Theta^{(L)}(\mathcal{X},\mathcal{X})\right) & \simeq \mathcal{O}(NnL) \gg \sigma_2\left(\Theta^{(L)}(\mathcal{X},\mathcal{X})\right)\\ \frac{1}{Nn}\mathbf{1}_{Nn}^\top\Theta^{(L)}(\mathcal{X},\mathcal{X})\mathbf{1}_{Nn} & \simeq \mathcal{O}(NnL) \gg \sigma_2\left(\Theta^{(L)}(\mathcal{X},\mathcal{X})\right)\\ \sigma_{\max}\left(\Theta^{(L)}(\mathcal{X},\mathcal{X})\right) & \geq \mathcal{O}(L)\cdot\sigma_{\max}\left(\mathcal{K}^{(L)}(\mathcal{X},\mathcal{X})\right)\end{cases} \tag{72}$$

where $\sigma_2(\cdot)$ denotes the second largest eigenvalue. Then, we have

$$\|\Theta(\mathcal{X},\mathcal{X})^{-1} - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\|_{\text{op}} \leq \mathcal{O}(\frac{1}{L^2}) \tag{73}$$

Proof. From (64), we know (Note: we omit the superscript $^{(L)}$ for simplicity in this proof)

$$\begin{aligned}\Phi_{MTL}(\mathcal{X},\mathcal{X}) &= \Theta(\mathcal{X},\mathcal{X}) - \mathcal{K}(\mathcal{X},\mathcal{X}) + \text{diag}\left(\{\mathcal{K}(X_i,X_i)\}_{i=1}^N\right)\\ &= \Theta(\mathcal{X},\mathcal{X}) - \widetilde{\mathcal{K}}(\mathcal{X},\mathcal{X})\end{aligned}$$

where we denote $\widetilde{\mathcal{K}}(\mathcal{X},\mathcal{X}) = \mathcal{K}(\mathcal{X},\mathcal{X}) + \text{diag}\left(\{\mathcal{K}(X_i,X_i)\}_{i=1}^N\right)$ for simplicity.

Case I: $n = 1$.

In this case, obviously, for each $i \in [N]$, we have $\mathcal{K}(X_i,X_i) = 1 + \mathcal{O}(\frac{1}{L}^2) \in \mathbb{R}$. We can define a perturbed NNGP matrix as

$$\widetilde{\mathcal{K}}(\mathcal{X},\mathcal{X}) = \mathcal{K}(\mathcal{X},\mathcal{X}) - \text{diag}\left(\{\mathcal{K}(X_i,X_i)\}_{i=1}^N\right) \tag{74}$$

$$= \mathbf{1}_N\mathbf{1}_N^\top - I + \frac{1}{L^2}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}} \tag{75}$$

where we define $\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}} = \mathbf{B}^{(L)}_{\mathcal{X},\mathcal{X}} - \left(\text{diag}\left(\{\mathcal{K}(X_i,X_i)\}_{i=1}^N\right) - I\right)$, i.e., $\mathbf{B}^{(L)}_{\mathcal{X},\mathcal{X}}$ with the $\mathcal{O}(\frac{1}{L}^2)$ terms from $\text{diag}\left(\{\mathcal{K}(X_i,X_i)\}_{i=1}^N\right)$.

[13] Notice that we use the little-o notation here: $f(x) = o(g(x))$ indicates that $g(x)$ grows much faster than $f(x)$. Thus the $o(\cdot)$ terms are negligible here.

For convenience, let us define a perturbed NTK matrix as

$$\begin{aligned}\widetilde{\Theta}(\mathcal{X},\mathcal{X}) &= \Theta(\mathcal{X},\mathcal{X}) - \left(\widetilde{\mathcal{K}}(\mathcal{X},\mathcal{X}) - \mathbf{1}_N\mathbf{1}_N^\top\right) \\ &= \Theta(\mathcal{X},\mathcal{X}) + I - \frac{1}{L^2}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}}. \end{aligned} \tag{76}$$

Obviously, we have

$$\begin{aligned}\|\Theta(\mathcal{X},\mathcal{X})^{-1} - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\|_{\text{op}} &= \|\Theta(\mathcal{X},\mathcal{X})^{-1} - \widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X}) + \widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X}) - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\|_{\text{op}} \\ &\le \|\Theta(\mathcal{X},\mathcal{X})^{-1} - \widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X})\|_{\text{op}} + \|\widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X}) - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\|_{\text{op}} \end{aligned} \tag{77}$$

Thus, we can prove (73) by providing bounds for $\|\Theta(\mathcal{X},\mathcal{X})^{-1} - \widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X})\|_{\text{op}}$ and $\|\widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X}) - \Theta^{-1}(\mathcal{X},\mathcal{X})\|_{\text{op}}$ separately.

- **Bound** $\|\Phi^{-1}_{MTL} - \widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X})\|_{\text{op}}$**.**
  By the Woodbury identity, we have

$$\begin{aligned}\Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X}) &= \left(\Theta(\mathcal{X},\mathcal{X}) - \widetilde{\mathcal{K}}(\mathcal{X},\mathcal{X})\right)^{-1} \\ &= \Big(\overbrace{\big[\Theta(\mathcal{X},\mathcal{X}) + I - \frac{1}{L^2}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}} - o(\frac{1}{L^2})\big]}^{\widetilde{\Theta}(\mathcal{X},\mathcal{X})\triangleq} - \mathbf{1}_N\mathbf{1}_N^\top\Big)^{-1} \\ &= \left(\widetilde{\Theta}(\mathcal{X},\mathcal{X}) - \mathbf{1}_N\mathbf{1}_N^\top\right)^{-1} \\ &= \widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1} - \rho\cdot\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathbf{1}_N\mathbf{1}_N^\top\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1} \end{aligned}$$

  where

$$\rho = \frac{1}{1 - \mathbf{1}_N^\top\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathbf{1}_N}$$

  By (72) and some eigendecomposition analysis, we can easily derive that

$$\begin{aligned}\rho = \frac{1}{1 - \mathbf{1}_N^\top\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathbf{1}_N} &\simeq \frac{1}{1 - \mathcal{O}(\frac{1}{L})} \\ \widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathbf{1}_N\mathbf{1}_N^\top\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1} &\simeq \mathcal{O}\left(\frac{1}{N^2L^2}\right)\mathbf{1}_N\mathbf{1}_N^\top \end{aligned}$$

  Thus

$$\Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X}) = \widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1} - \mathcal{O}\left(\frac{1}{N^2L^2(1-\mathcal{O}(\frac{1}{L}))}\right)\mathbf{1}_N\mathbf{1}_N^\top \tag{78}$$

  where the last term is negligible since its *maximum* eigenvalue is $\mathcal{O}(\frac{1}{NL^2(1-\mathcal{O}(\frac{1}{L}))})$, while the *minimum* eigenvalue for the first term is $\mathcal{O}(\frac{1}{NL})$.
  Thus, we can write

$$\|\Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X}) - \widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\|_{\text{op}} = \|\mathcal{O}\left(\frac{1}{N^2L^2(1-\mathcal{O}(\frac{1}{L}))}\right)\mathbf{1}_N\mathbf{1}_N^\top\|_{\text{op}} \le \mathcal{O}(\frac{1}{NL^2}) \tag{79}$$

- Bound $\|\widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X}) - \Theta^{-1}(\mathcal{X},\mathcal{X})\|_{\text{op}}$
  By (69), (76), we know

$$\begin{aligned}\widetilde{\Theta}(\mathcal{X},\mathcal{X}) &= \Big(\overbrace{\frac{L}{4}\mathbf{1}_N\mathbf{1}_{Nn}^\top + \frac{3L}{4}I + \mathbf{A}^{(L)}_{\mathcal{X},\mathcal{X}}}^{\Theta(\mathcal{X},\mathcal{X})}\Big) - \Big(\overbrace{\mathbf{1}_N\mathbf{1}_N^\top - I + \frac{1}{L^2}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}}}^{\widetilde{\mathcal{K}}(\mathcal{X},\mathcal{X}) - \mathbf{1}_N\mathbf{1}_N^\top}\Big) \\ &= \left(\frac{L}{4} - 1\right)\mathbf{1}_N\mathbf{1}_N^\top + \left(\frac{3L}{4} + 1\right)I + \left(\mathbf{A}^{(L)}_{\mathcal{X},\mathcal{X}} - \frac{1}{L^2}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}}\right) \end{aligned}$$

By observation, it is obvious that for relatively large L, the perturbation $\mathbf{1}_N\mathbf{1}_N^\top - I + \frac{1}{L^2}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}}$ has minimal effect, e.g., the spectrum of $\widetilde{\Theta}(\mathcal{X},\mathcal{X})$ is almost identical to $\Theta(\mathcal{X},\mathcal{X})$.
Now, let us bound the inverse of the perturbed matrix $\widetilde{\Theta}(\mathcal{X},\mathcal{X})$ formally.
Leveraging the identity $(A+B)^{-1} = A^{-1}A^{-1}B(A+B)^{-1}$ from (Henderson & Searle, 1981). Defining

$$\hat{\Delta} = \widetilde{\Theta}(\mathcal{X},\mathcal{X}) - \Theta(\mathcal{X},\mathcal{X}) = \mathbf{1}_N\mathbf{1}_N^\top - I + \frac{1}{L^2}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}}$$

then we have

$$\begin{aligned}
&\left\|\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1} - \Theta(\mathcal{X},\mathcal{X})^{-1}\right\|_{op} \\
&= \left\|\left(\Theta(\mathcal{X},\mathcal{X}) + \hat{\Delta}\right)^{-1}\right\|_{op} \\
&= \left\|\Theta(\mathcal{X},\mathcal{X})^{-1} + \Theta(\mathcal{X},\mathcal{X})^{-1}\hat{\Delta}\left(\Theta(\mathcal{X},\mathcal{X}) + \Delta\right)^{-1} - \Theta(\mathcal{X},\mathcal{X})^{-1}\right\|_{op} \\
&= \left\|\Theta(\mathcal{X},\mathcal{X})^{-1}\hat{\Delta}\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{op} \\
&= \left\|\Theta(\mathcal{X},\mathcal{X})^{-1}\left(\mathbf{1}_N\mathbf{1}_N^\top - I + \frac{1}{L^2}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}}\right)\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{op} \\
&\le \left\|\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\mathbf{1}_N\mathbf{1}_N^\top\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{\text{op}} + \left\|\Theta(\mathcal{X},\mathcal{X})^{-1}I\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{\text{op}} + \frac{1}{L}^2\left\|\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\widetilde{\mathbf{B}}^{(L)}_{\mathcal{X},\mathcal{X}}\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{\text{op}} \qquad (80) \\
&\le \mathcal{O}\left(\frac{1}{NL^2}\right) + \mathcal{O}\left(\frac{1}{L^2}\right) + \mathcal{O}\left(\frac{1}{L^4}\right) \\
&\le \mathcal{O}\left(\frac{1}{L^2}\right) \qquad (81)
\end{aligned}$$

Finally, combining (77), (79) and (81), we have

$$\begin{aligned}
\|\Theta(\mathcal{X},\mathcal{X})^{-1} - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\|_{\text{op}} &\le \|\Theta(\mathcal{X},\mathcal{X})^{-1} - \widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X})\|_{\text{op}} + \|\widetilde{\Theta}^{-1}(\mathcal{X},\mathcal{X}) - \Phi^{-1}_{MTL}(\mathcal{X},\mathcal{X})\|_{\text{op}} \\
&\le \mathcal{O}(\frac{1}{NL^2}) + \mathcal{O}(\frac{1}{L^2}) = \mathcal{O}(\frac{1}{L^2}) \qquad (82)
\end{aligned}$$

Case II: $n > 1$.

Compared to the case of $n = 1$, the only difference with (82) is caused by the term $\left\|\Theta(\mathcal{X},\mathcal{X})^{-1}I\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{\text{op}}$ in (80) is converted to

$$\left\|\Theta(\mathcal{X},\mathcal{X})^{-1}\text{diag}\left(\{\mathbf{1}_n\mathbf{1}_n^\top\}_{i=1}^N\right)\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{\text{op}}$$

Since

$$\|\text{diag}\left(\{\mathbf{1}_n\mathbf{1}_n^\top\}_{i=1}^N\right)\|_{op} = \|\mathbf{1}_n\mathbf{1}_n^\top\|_{op} = n = \mathcal{O}(1)\,,$$

we have

$$\begin{aligned}
\left\|\Theta(\mathcal{X},\mathcal{X})^{-1}\text{diag}\left(\{\mathbf{1}_n\mathbf{1}_n^\top\}_{i=1}^N\right)\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{\text{op}} &\le \left\|\widetilde{\Theta}(\mathcal{X},\mathcal{X})^{-1}\right\|_{\text{op}}^2\left\|\text{diag}\left(\{\mathbf{1}_n\mathbf{1}_n^\top\}_{i=1}^N\right)\right\|_{\text{op}} \\
&\le \mathcal{O}(\frac{1}{L^2})
\end{aligned}$$

■

B.5. Proof of Theorem 1

The proof of Theorem 1 can be straightforwardly derived based on Lemma 5.

Proof. By (67), (73), we have

$$\begin{aligned}
&\|F_{ANIL}(X, X', Y') - F_{MTL}(X, X', Y')\|_2 \\
\leq\ &\|\Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X})\|_{\text{op}} \|\Theta(\mathcal{X}, \mathcal{X})^{-1} - \Phi^{-1}_{MTL}(\mathcal{X}, \mathcal{X})\|_{\text{op}} \|\mathcal{Y}\|_2 + \mathcal{O}(\lambda\tau\sigma_{\max}(\mathcal{K})) + \mathcal{O}(\frac{1}{\sqrt{h}}) \\
\leq\ &\mathcal{O}(\frac{1}{L}) + \mathcal{O}(\lambda\tau) + \mathcal{O}(\frac{1}{\sqrt{h}})
\end{aligned}$$

where we used the facts that $\|\Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X})\|_{\text{op}} = \mathcal{O}(L)$, which can be straightforwardly derived from Lemma 3 and 4. ■

B.6. Extension to Residual ReLU Networks

Corollary 1.1 states that the theoretical results of Theorem 1 apply to residual ReLU networks and residual ReLU networks with LayerNorm. The proof of this corollary is simply derived from Appendix C.2 and C.4 of Xiao et al. (2020).

Proof. For residual ReLU networks, the corresponding NTK and NNGP have a factor of e^L compared (69) and (70), which has no effect on the predictors F_{ANIL} and F_{MTL}, since the factors from the kernel and kernel inverse cancel out (e.g., $e^L\Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X}) \cdot (e^L\Phi_{MTL}(\mathcal{X}, \mathcal{X}))^{-1} = \Phi'_{MTL}((X, X', \hat{\tau}), \mathcal{X})\Phi_{MTL}(\mathcal{X}, \mathcal{X})^{-1}$). Thus, Theorem 1 applies to this class of networks.

For residual ReLU networks with LayerNorm, Appendix C.3 of Xiao et al. (2020) shows the kernel structures of NTK and NNGP is the same as ReLU networks without residual connections. Thus, Theorem 1 directly applies to this class of networks. ■

C. Details of Experiments

In this section, we will provide more details about the experiment in Sec. 5. Specifically,

- Appendix C.1: presents more experimental details about Sec. 5.1, the empirical validation of Theorem 1.
- Appendix C.2: presents more experimental details about Sec. 5.2, the empirical study on few-shot image classification benchmarks.

C.1. Empirical Validation of Theorem 1

Implementation. We implement MTL and ANIL kernels with Neural Tangents (Novak et al., 2020), a codebase built on JAX (Bradbury et al., 2018), which is a package designed for high-performance machine learning research in Python. Since MTL and ANIL kernel functions are composite kernel functions built upon NTK and NNGP functions, we directly construct NTKs and NNGPs using Neural Tangents and then compose them into MTL and ANIL kernels.

About Figure 1. Note that the value at $L = 10$ in the first image is a little smaller than the value at $\lambda\tau = 0$ in the second image. That is because the random seeds using in the two images are different. Even though we take an average over 5 random seeds when plotting each image, there still exists some non-negligible variance.

C.2. Experiments on Few-Shot Image Classification Benchmarks

Fine-Tuning in Validation and Test. In the meta-validation and meta-testing stages, following Sec. 3.4, we fine-tune a linear classifier on the features (i.e., outputs of the last hidden layer) with the cross-entropy loss and a ℓ_2 regularization. Specifically, similar to Tian et al. (2020), we use the logistic regression classifier from sklearn for the fine-tuning (Pedregosa et al., 2011), and we set the ℓ_2 regularization strength to be 0.33 based on the following ablation study on ℓ_2 penalty (i.e., Table 5.

ℓ_2 Penalty	0.0001	0.001	0.01	0.1	0.33	1	3
Test Accuracy(%)	76.86	77.02	77.28	77.61	**77.72**	77.55	76.82

Table 5. Ablation study of the ℓ_2 penalty on the fine-tuned linear layer. Evaluated on mini-ImageNet (5-way 5-shot classification).