MAML and ANIL Provably Learn Representations

Liam Collins [1] **Aryan Mokhtari** [1] **Sewoong Oh** [2] **Sanjay Shakkottai** [1]

Abstract

Recent empirical evidence has driven conventional wisdom to believe that gradient-based meta-learning (GBML) methods perform well at few-shot learning because they learn an expressive data representation that is shared across tasks. However, the mechanics of GBML have remained largely mysterious from a theoretical perspective. In this paper, we prove that two well-known GBML methods, MAML and ANIL, as well as their first-order approximations, are capable of learning common representation among a set of given tasks. Specifically, in the well-known multi-task linear representation learning setting, they are able to recover the ground-truth representation at an exponentially fast rate. Moreover, our analysis illuminates that the driving force causing MAML and ANIL to recover the underlying representation is that they adapt the final layer of their model, which harnesses the underlying task diversity to improve the representation in all directions of interest. To the best of our knowledge, these are the first results to show that MAML and/or ANIL learn expressive representations and to rigorously explain why they do so.

1. Introduction

A widely popular approach to achieve fast adaptation in multi-task learning settings is to learn a representation that extracts the important features shared across tasks (Maurer et al., 2016). However, our understanding of *how* multi-task representation learning should be done and *why* certain methods work well is still nascent.

Recently, a paradigm known as *meta-learning* has emerged as a powerful means of learning multi-task representations. This was sparked in large part by the introduction of Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), which achieved impressive results in few-shot image classification and reinforcement learning scenarios, and led to a series of related gradient-based meta-learning (GBML) methods (Raghu et al., 2020; Nichol & Schulman, 2018; Antoniou et al., 2019; Hospedales et al., 2021). Surprisingly, MAML does not explicitly try to learn a useful representation; instead, it aims to find a good initialization for a small number of task-specific gradient descent steps, agnostic of whether the learning model contains a representation. Nevertheless, Raghu et al. (2020) empirically argued that MAML's impressive performance on neural networks is likely due to its tendency to learn a shared representation across tasks. To make this argument, they noticed that MAML's representation does not change significantly when adapted to each task. Moreover, they showed that a modified version of MAML that freezes the representation during local adaptation, known as the Almost-No-Inner-Loop algorithm (ANIL), typically performs at least as well as MAML on few-shot image classification tasks. Yet it is still not well understood why these algorithms that search for a good initialization for gradient descent should find useful a global representation among tasks. Thus, in this paper, we aim to address the following questions:

> *Do MAML and ANIL provably learn high-quality representations? If so, why?*

To answer these questions we consider the multi-task linear representation learning setting (Maurer et al., 2016; Tripuraneni et al., 2021; Du et al., 2020) in which each task is a noisy linear regression problem in $\mathbb{R}^d$ with optimal solution lying in a shared k-dimensional subspace, where $k \ll d$. The learning model is a two-layer linear network consisting of a representation (the first layer of the model) and head (the last layer). The goal is to learn a representation that projects data onto the shared subspace so as to reduce the number of samples needed to find the optimal regressor for a new task from $\Omega(d)$ to $\Omega(k)$.

Main contributions. *We prove, for the first time, that both MAML and ANIL, as well their first-order approximations, are capable of representation learning and recover the ground-truth subspace in this setting.* Our analysis reveals that MAML and ANIL's distinctive adaptation updates

[1]Department of Electrical and Computer Engineering, The University of Texas at Austin [2]School of Computer Science and Engineering, University of Washington. Correspondence to: Liam Collins <liamc@utexas.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

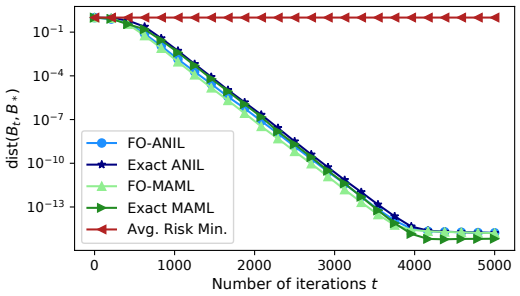

Figure 1. Distance of learned representation from the ground-truth for ANIL, MAML and average risk minimization run on task population losses in multi-task linear representation learning setting.

for the last layer of the learning model are critical to their recovery of the ground-truth representation. Figure 1 visualizes this observation: all meta-learning approaches (Exact ANIL, MAML, and their first-order (FO) versions that ignore second-order derivatives) approach the ground truth exponentially fast, while a non-meta learning baseline of average loss minimization empirically fails to recover the ground-truth. We show that the inner loop updates of the head exploit *task diversity* to make the outer loop updates bring the representation closer to the ground-truth. However, MAML's inner loop updates for the representation can inhibit this behavior, thus, our results for MAML require an initialization with error *related to task diversity*, whereas ANIL requires only *constant* error. We also show that ANIL learns the ground-truth representation with only $\tilde{O}(\frac{k^3 d}{n} + k^3) \ll d$ samples per task, demonstrating that ANIL's representation learning is *sample-efficient*.

Related work. Several works have studied why meta-learning algorithms are effective; please see Appendix A for a comprehensive discussion. Building off Raghu et al. (2020), most of these works have studied meta-learning from a representation learning perspective (Goldblum et al., 2020; Saunshi et al., 2021; Arnold et al., 2021; Wang et al., 2021a; Kao et al., 2022). Among these, Ni et al. (2021); Bouniot et al. (2020); Setlur et al. (2020) and Kumar et al. (2021) showed mixed empirical impacts of training task diversity on model performance. Most related to our work is (Saunshi et al., 2020), which proved that the continuous version of a first-order GBML method, Reptile (Nichol & Schulman, 2018), learns a one-dimensional linear representation in a two-task setting with a specific initialization, explicit regularization, and infinite samples per task. Other works studied multi-task representation learning in the linear setting we consider from a statistical perspective (Maurer et al., 2016; Du et al., 2020; Tripuraneni et al., 2021). Furthermore, Collins et al. (2021) and Thekumparampil et al. (2021) gave optimization results for gradient-based methods in this setting. However, the algorithms they studied are customized for the assumed low-dimensional linear representation model, which makes it relatively easy to learn the correct representation efficiently. A more challenging task is to understand how general purpose and model-agnostic meta-learning algorithms perform, such as the algorithms we study.

Notations. We use bold capital letters for matrices and bold lowercase letters for vectors. We use $\mathcal{O}^{d\times k}$ to denote the set of matrices in $\mathbb{R}^{d\times k}$ with orthonormal columns. A hat above a matrix, e.g. $\hat{\mathbf{B}}$, implies the matrix is a member of $\mathcal{O}^{d\times k}$. We let $\mathrm{col}(\mathbf{B})$ denote the column space of $\mathbf{B}$, and $\mathrm{col}^{\perp}(\mathbf{B})$ denote its orthogonal complement. $\mathcal{N}(0,\sigma^2)$ denotes the Gaussian distribution with mean 0 and variance σ^2. $O(\cdot)$ and $\Omega(\cdot)$ hide constant factors, and $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ hide constant and logarithmic factors.

2. Problem Formulation

We employ the multi-task linear representation learning framework (Maurer et al., 2016; Du et al., 2020; Tripuraneni et al., 2021) studied in prior works. Each task in this setting is a linear regression problem in $\mathbb{R}^d$. We index tasks by (t,i), corresponding to the i-th task sampled on iteration t. The inputs $\mathbf{x}_{t,i} \in \mathbb{R}^d$ and labels $y_{t,i} \in \mathbb{R}$ for the (t,i)-th task are sampled i.i.d. from a distribution $P_{t,i}$ over $\mathbb{R}^d \times \mathbb{R}$ such that:

$$\mathbf{x}_{t,i} \sim p, \quad z_{t,i} \sim \mathcal{N}(0,\sigma^2), \quad y_{t,i} = \langle \boldsymbol{\theta}_{*,t,i}, \mathbf{x}_{t,i} \rangle + z_{t,i}$$

where $\boldsymbol{\theta}_{*,t,i} \in \mathbb{R}^d$ is the ground-truth regressor for task (t,i), p is a distribution over $\mathbb{R}^d$ and $z_{t,i}$ is white Gaussian noise with variance σ^2. Each task has a set of m samples $\mathcal{D}_{t,i} := \{(\mathbf{x}_{t,i,j}, y_{t,i,j})\}_{j\in[m]}$ drawn i.i.d. from $P_{t,i}$ available for training.

To account for shared information across tasks, we suppose there exists a matrix $\mathbf{B}_* \in \mathcal{O}^{d\times k}$ such that the ground-truth regressors $\{\boldsymbol{\theta}_{*,t,i}\}_i$ for all tasks lie in $\mathrm{col}(\mathbf{B}_*)$, so they can be written as $\boldsymbol{\theta}_{*,t,i} = \mathbf{B}_* \mathbf{w}_{*,t,i}$ for all t,i, where $\mathbf{w}_{*,t,i} \in \mathbb{R}^k$. We refer to $\mathbf{B}_*$ as the ground-truth representation and $\mathbf{w}_{*,t,i}$ as the ground-truth head for task (t,i). The task environment consists of $\mathbf{B}_*$ and a distribution ν over ground-truth heads. With knowledge of $\mathrm{col}\,(\mathbf{B}_*)$, we can reduce the number of samples needed to solve a task from $\Omega(d)$ to $\Omega(k)$ by projecting the task data onto $\mathrm{col}\,(\mathbf{B}_*)$, then learning a head in $\mathbb{R}^k$. The question becomes how to learn the ground-truth subspace $\mathrm{col}\,(\mathbf{B}_*)$.

The learning model consists of a representation $\mathbf{B} \in \mathbb{R}^{d\times k}$ and a head $\mathbf{w} \in \mathbb{R}^k$. We would like the column space of $\mathbf{B}$ to be close to that of $\mathbf{B}_*$, measured as follows.

Definition 1 (Principle angle distance)**.** *Let $\hat{\mathbf{B}} \in \mathcal{O}^{d\times k}$ and $\hat{\mathbf{B}}_{*,\perp} \in \mathcal{O}^{d\times(d-k)}$ denote orthonormal matrices whose columns span $\mathrm{col}(\mathbf{B})$ and $\mathrm{col}^{\perp}(\mathbf{B}_*)$, respectively. Then the principle angle distance between $\mathbf{B}$ and $\mathbf{B}_*$ is*

$$\mathrm{dist}(\mathbf{B}, \mathbf{B}_*) := \|\hat{\mathbf{B}}_{*,\perp}^{\top} \hat{\mathbf{B}}\|_2. \tag{1}$$

For shorthand, we denote $\text{dist}_t := \text{dist}(\mathbf{B}_t, \mathbf{B}_*)$.

Notice that $\dim(\text{col}(\mathbf{B}_*)) = k$. Thus, the learned representation $\mathbf{B}$ must extract k orthogonal directions belonging to $\text{col}(\mathbf{B}_*)$. As we will show, MAML and ANIL's task-specific adaptation of the head critically leverages task diversity to learn k such directions.

3. Algorithms

Here we formally state the implementation of ANIL and MAML for the problem described above. First, letting $\boldsymbol{\theta} := [\mathbf{w}; \text{vec}(\mathbf{B})] \in \mathbb{R}^{(d+1)k}$ denote the vector of model parameters, we define the population loss for task (t, i):

$$\mathcal{L}_{t,i}(\boldsymbol{\theta}) := \tfrac{1}{2}\mathbb{E}_{(\mathbf{x}_{t,i}, y_{t,i})\sim P_{t,i}}\left[(\langle \mathbf{B}\mathbf{w}, \mathbf{x}_{t,i}\rangle - y_{t,i})^2\right].$$

Often we approximate this loss with the finite-sample loss for a dataset $\mathcal{D}'_{t,i} := \{(\mathbf{x}_{t,i,j}, y_{t,i,j})\}_{j\in[m']}$:

$$\hat{\mathcal{L}}_{t,i}(\boldsymbol{\theta}; \mathcal{D}'_{t,i}) := \tfrac{1}{2m'}\sum_{j=1}^{m'}(\langle \mathbf{B}\mathbf{w}, \mathbf{x}_{t,i,j}\rangle - y_{t,i,j})^2.$$

MAML. MAML minimizes the average loss across tasks *after* a small number of task-specific gradient updates. Here, we consider that the task-specific updates are one step of minibatch SGD with batch $\mathcal{D}^{in}_{t,i}$ consisting of m_{in} i.i.d. samples from $P_{t,i}$. Specifically, the loss function that MAML minimizes is

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{MAML}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{w}_{*,t,i}, \mathcal{D}^{in}_{t,i}}[\mathcal{L}_{t,i}(\boldsymbol{\theta} - \alpha\nabla_{\boldsymbol{\theta}}\hat{\mathcal{L}}_{t,i}(\boldsymbol{\theta}; \mathcal{D}^{in}_{t,i})))] \tag{2}$$

where for ease of notation we have written $\mathbb{E}_{\mathbf{w}_{*,t,i}, \mathcal{D}^{in}_{t,i}}$ as shorthand for $\mathbb{E}_{\mathbf{w}_{*,t,i}\sim\nu,\ \mathcal{D}^{in}_{t,i}\sim P^{m_{in}}_{t,i}}$. MAML essentially solves (2) with minibatch SGD. At iteration t, it draws n tasks with ground-truth heads $\{\mathbf{w}_{*,t,i}\}_{i\in[n]}$ drawn from ν, and for each drawn task, draws m samples contained in $\mathcal{D}_{t,i}$ i.i.d. from $P_{t,i}$. MAML then partitions $\mathcal{D}_{t,i}$ into $\mathcal{D}^{in}_{t,i}$ and $\mathcal{D}^{out}_{t,i}$ such that $|\mathcal{D}^{in}_{t,i}| = m_{in}$, $|\mathcal{D}^{out}_{t,i}| = m_{out}$, and $m_{in} + m_{out} = m$ (we assume $m_{in} < m$). For task (t, i), in what is known as the inner loop, MAML takes a task-specific stochastic gradient step from the initial model $(\mathbf{B}_t, \mathbf{w}_t)$ using the samples $\mathcal{D}^{in}_{t,i}$ and step size α to obtain the adapted parameters $\boldsymbol{\theta}_{t,i}$:

$$\boldsymbol{\theta}_{t,i} = \begin{bmatrix}\mathbf{w}_{t,i}\\ \text{vec}(\mathbf{B}_{t,i})\end{bmatrix} \leftarrow \begin{bmatrix}\mathbf{w}_t - \alpha\nabla_{\mathbf{w}}\hat{\mathcal{L}}_i(\mathbf{B}_t, \mathbf{w}_t; \mathcal{D}^{in}_{t,i})\\ \text{vec}(\mathbf{B}_t) - \alpha\nabla_{\text{vec}(\mathbf{B})}\hat{\mathcal{L}}_i(\mathbf{B}_t, \mathbf{w}_t; \mathcal{D}^{in}_{t,i})\end{bmatrix}$$

Then, in the so-called outer loop, MAML takes a minibatch SGD step with respect to the loss after task-specific adaptation using the samples $\{\mathcal{D}^{out}_{t,i}\}_{i\in[n]}$ and step size β:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \tfrac{\beta}{n}\sum_{i=1}^{n}(\mathbf{I} - \alpha\nabla^2_{\boldsymbol{\theta}}\hat{\mathcal{L}}_{t,i}(\boldsymbol{\theta}_t; \mathcal{D}^{out}_{t,i}))\nabla\hat{\mathcal{L}}_{t,i}(\boldsymbol{\theta}_{t,i}; \mathcal{D}^{out}_{t,i})$$

Note that the above Exact MAML update requires expensive second-order derivative computations. In practice, FO-MAML, which drops the Hessian, is often used, since it typically achieves similar performance (Finn et al., 2017).

ANIL. Surprisingly, Raghu et al. (2020) noticed that training neural nets with a modified version of MAML that lacks inner loop updates for the representation resulted in models that matched and sometimes even exceeded the performance of models trained by MAML on few-shot image classification tasks. This modified version is ANIL, and its inner loop updates in our linear case are given as follows:

$$\boldsymbol{\theta}_{t,i} = \begin{bmatrix}\mathbf{w}_{t,i}\\ \text{vec}(\mathbf{B}_{t,i})\end{bmatrix} = \begin{bmatrix}\mathbf{w}_t - \alpha\nabla_{\mathbf{w}}\hat{\mathcal{L}}_{t,i}(\mathbf{B}_t, \mathbf{w}_t; \mathcal{D}^{in}_{t,i})\\ \text{vec}(\mathbf{B}_t)\end{bmatrix}.$$

In the outer loop, ANIL again takes a minibatch SGD step with respect to the loss after the inner loop update. Then, the outer loop updates for Exact ANIL are given by:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \frac{\beta}{n}\sum_{i=1}^{n}\hat{\mathbf{H}}_{t,i}(\boldsymbol{\theta}_t; \mathcal{D}^{out}_{t,i})\nabla_{\boldsymbol{\theta}}\hat{\mathcal{L}}_{t,i}(\boldsymbol{\theta}_{t,i}, \mathcal{D}^{out}_{t,i})$$

where, for Exact ANIL,

$$\hat{\mathbf{H}}_{t,i}(\mathbf{B}_t, \mathbf{w}_t; \mathcal{D}^{out}_{t,i}) := \begin{bmatrix}\mathbf{I}_k - \alpha\nabla^2_{\mathbf{w}}\hat{\mathcal{L}}_{t,i}(\boldsymbol{\theta}_t; \mathcal{D}^{out}_{t,i}) & \mathbf{0}\\ -\alpha\frac{\partial^2}{\partial\,\text{vec}(\mathbf{B})\partial\mathbf{w}}\hat{\mathcal{L}}_{t,i}(\boldsymbol{\theta}_t; \mathcal{D}^{out}_{t,i}) & \mathbf{I}\end{bmatrix}$$

To avoid computing second order derivatives, we can instead treat $\hat{\mathbf{H}}_{t,i}$ as the identity operator, in which case we call the algorithm FO-ANIL.

3.1. Role of Adaptation

Now we present new intuition for MAML and ANIL's representation learning ability which motivates our proof structure. The key observation is that the outer loop gradients for the representation are *evaluated at the inner loop-adapted parameters*; this harnesses the power of task diversity to improve the representation in *all* k directions. This is easiest to observe in the FO-ANIL case with $m_{in} = m_{out} = \infty$. In this case, the update for the representation is given as:

$$\mathbf{B}_{t+1} = \mathbf{B}_t\Big(\underbrace{\mathbf{I}_k - \tfrac{\beta}{n}\sum_{i=1}^{n}\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top}_{\text{FO-ANIL prior weight}}\Big) + \mathbf{B}_*\underbrace{\tfrac{\beta}{n}\sum_{i=1}^{n}\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top}_{\text{FO-ANIL signal weight}} \tag{3}$$

If the 'prior weight' is small and the 'signal weight' is large, then the update replaces energy from $\text{col}(\mathbf{B}_t)$ with energy from $\text{col}(\mathbf{B}_*)$. Roughly, this is true as long as $\boldsymbol{\Psi}_t := \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top$ is well-conditioned, i.e. the $\mathbf{w}_{t,i}$'s are diverse. Assuming $\mathbf{w}_{t,i} \approx \mathbf{w}_{*,t,i}$ for each task, then $\boldsymbol{\Psi}_t$ is well-conditioned if and only if the tasks are diverse. This shows how task diversity causes the column space of the

representation learned by FO-ANIL to approach the ground-truth. For FO-MAML, we observe similar behavior, with a caveat. The representation update is:

$$\mathbf{B}_{t+1} \overset{(a)}{=} \mathbf{B}_t - \frac{\beta}{n}\sum_{i=1}^{n}\mathbf{B}_{t,i}\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top + \mathbf{B}_*\frac{\beta}{n}\sum_{i=1}^{n}\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top$$
$$\overset{(b)}{=} \mathbf{B}_t\underbrace{\Big(\mathbf{I}_k - (\mathbf{I}_k - \alpha\mathbf{w}_t\mathbf{w}_t^\top)\frac{\beta}{n}\sum_{i=1}^{n}\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\Big)}_{\text{FO-MAML prior weight}}$$
$$+\mathbf{B}_*\underbrace{\Big(\frac{\beta}{n}\sum_{i=1}^{n}(1-\alpha\langle\mathbf{w}_{t,i},\mathbf{w}_t\rangle)\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\Big)}_{\text{FO-MAML signal weight}}$$

Equation (a) is similar to (3) except that one $\mathbf{B}_t$ is replaced by the $\mathbf{B}_{t,i}$'s resulting from inner loop adaptation. Expanding $\mathbf{B}_{t,i}$ in (b), we notice that the prior weight is at least as large as in (3), since $\lambda_{\max}(\mathbf{I}_k - \alpha\mathbf{w}_t\mathbf{w}_t^\top) \le 1$, but it can still be small as long as the $\mathbf{w}_{t,i}$'s are diverse and $\|\mathbf{w}_t\|_2$ is small. Thus we conclude that FO-MAML also can learn the representation, yet its inner loop adaptation complicates its ability to do so.

Comparison with no inner-loop adaptation. Compare these updates to the case when there is no inner loop adaptation, i.e. we run SGD on the non-adaptive objective $\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{w}_{*,t,i}}[\mathcal{L}_{t,i}(\boldsymbol{\theta})]$ instead of (2). In this case, $\mathbf{B}_{t+1}$ is:

$$\mathbf{B}_{t+1} = \mathbf{B}_t\big(\mathbf{I}_k - \beta\mathbf{w}_t\mathbf{w}_t^\top\big) + \beta\mathbf{B}_*\bar{\mathbf{w}}_{*,t}\mathbf{w}_t^\top \tag{4}$$

where $\bar{\mathbf{w}}_{*,t} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{*,t,i}$. Observe that the coefficient of $\mathbf{B}_t$ in the update is rank $k-1$, while the coefficient of $\mathbf{B}_*$ is rank 1. Thus, $\mathrm{col}(\mathbf{B}_{t+1})$ can approach $\mathrm{col}(\mathbf{B}_*)$ in at most one direction on any iteration. Empirically, $\mathbf{w}_t$ points in roughly the same direction throughout training, preventing this approach from learning $\mathrm{col}(\mathbf{B}_*)$ (e.g. see Figure 1).

Technical challenges. The intuition on the role of adaptation, while appealing, makes strong assumptions; most notably that the $\mathbf{w}_{t,i}$'s are diverse enough to improve the representation and that the algorithm dynamics are stable. To show these points, we observe that $\mathbf{w}_{t,i}$ can be written as the linear combination of a vector distinct for each task and a vector that is shared across all tasks at time t. Showing that the shared vector is small implies the $\mathbf{w}_{t,i}$'s are diverse, and we can control the magnitude of the shared vector by controlling $\|\mathbf{w}_t\|_2$ and $\|\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t\|_2$. Showing that these quantities are small at all times also ensures the stability of the algorithms. Meanwhile, we must prove that $\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2$ and $\mathrm{dist}_t = \|\hat{\mathbf{B}}_{*,\perp}^\top\hat{\mathbf{B}}_t\|_2$ are contracting. It is not obvious that any of these conditions hold individually; in fact, they require a novel multi-way inductive argument to show that they hold simultaneously for each t (see Section 5).

4. Main Results

In this section we formalize our intuition discussed previously and prove that both MAML and ANIL and their first-order approximations are capable learning the column space of the ground-truth representation. To do so, we first make the following assumption concerning the diversity of the sampled ground-truth heads.

Assumption 1 (Task diversity)**.** *The eigenvalues of the symmetric matrix $\boldsymbol{\Psi}_{*,t} \coloneqq (\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top)$ are uniformly bounded below and above by μ_*^2 and L_*^2, respectively[1], i.e., $\mu_*^2\mathbf{I} \preceq \boldsymbol{\Psi}_{*,t} \preceq L_*^2\mathbf{I}$, for all $t \in [T]$.*

The lower bound on the eigenvalues of the matrix $\boldsymbol{\Psi}_{*,t}$ ensures that the $k \times k$ matrix $\boldsymbol{\Psi}_{*,t}$ is full rank and hence the vectors $\{\mathbf{w}_{*,t,i}\}_{i=1}^n$ span $\mathbb{R}^k$, therefore they are diverse. However, the diversity level of the tasks is defined by ratio of the eigenvalues of the matrix $\boldsymbol{\Psi}_{*,t}$, i.e., $\kappa_* \coloneqq \frac{L_*}{\mu_*}$. If this ratio is close to 1, then the ground-truth heads are very diverse and have equal energy in all directions. On the other hand, if κ_* is large, then the ground-truth heads are not very diverse as their energy is mostly focused in a specific direction. Hence, as the following results reveal, smaller κ_* leads to faster convergence for ANIL and MAML.

Now we are ready to state our main results for the ANIL and FO-ANIL algorithms in the infinite sample case.

Theorem 1. *Consider the infinite sample case for ANIL and FO-ANIL, where $m_{in} = m_{out} = \infty$. Further, suppose the conditions in Assumption 1 hold, the initial weights are selected as $\mathbf{w}_0 = \mathbf{0}_k$ and $\alpha\mathbf{B}_0^\top\mathbf{B}_0 = \mathbf{I}_k$. Let the step sizes are chosen as $\alpha = O(1/L_*)$ and $\beta = O(\alpha\kappa_*^{-4})$ for ANIL and $\beta = O(\alpha\kappa_*^{-4}\min(1, \mu_*^2/\eta_*^2))$ for FO-ANIL, where η_* satisfies $\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{*,t,i}\|_2 \le \eta_*$ for all times $t \in [T]$ almost surely. If the initial error satisfies the condition $\mathrm{dist}_0 \le \sqrt{0.9}$, then almost surely for both ANIL and FO-ANIL we have,*

$$\mathrm{dist}(\mathbf{B}_T, \mathbf{B}_*) \le \big(1 - 0.5\beta\alpha E_0\mu_*^2\big)^{T-1}, \tag{5}$$

where $E_0 \coloneqq 0.9 - \mathrm{dist}_0^2$.

Theorem 1 shows that both FO-ANIL and Exact ANIL learn a representation that approaches the ground-truth exponentially fast as long as the initial representation $\mathbf{B}_0$ is normalized and is a constant distance away from the ground-truth, the initial head $\mathbf{w}_0 = \mathbf{0}$, and the sampled tasks are diverse. Note that β is larger for ANIL and hence its convergence is faster, demonstrating the benefit of second-order updates.

[1] We could instead assume the ground-truth heads are sub-gaussian and use standard concentration results show that with $n = \Omega(k + \log(T))$, the set of ground-truth heads $\{\mathbf{w}_{*,t,i}\}_{i=1}^n$ sampled on iteration t are $(1 + O(1), 1 - O(1))$-diverse for all T iterations with high probability, but instead we assume generic bounds for simplicity.

Next, we state our results for FO-MAML and Exact MAML for the same infinite sample setting. Due to the adaptation of both the representation and head, the MAML and FO-MAML updates involve third- and fourth-order products of the ground-truth heads, unlike the ANIL and FO-ANIL updates which involve at most second-order products. To analyze the higher-order terms, we assume that the energy in each ground-truth head is balanced.

Assumption 2 (Task incoherence)**.** *For all times $t \in [T]$ and tasks $i \in [n]$, we almost surely have $\|\mathbf{w}_{*,t,i}\|_2 \leq c\sqrt{k}L_*$, where c is a constant.*

Next, as discussed in Section 3.1, MAML's adaptation of the representation complicates its ability to learn the ground-truth subspace. As a result, we require an additional condition to show that MAML learns the representation: the distance of the initialization to the ground-truth must small in the sense that it must scale with the task diversity and inversely with k. We formalize this in the following theorem.

Theorem 2. *Consider the infinite sample case for MAML, where $m_{in} = m_{out} = \infty$. Further, suppose the conditions in Assumptions 1 and 2 hold, the initial weights are selected as $\mathbf{w}_0 = \mathbf{0}_k$ and $\alpha\mathbf{B}_0^\top\mathbf{B}_0 = \mathbf{I}_k$, and the step sizes satisfy $\alpha = O(k^{-2/3}L_*^{-1}T^{-1/4})$ and $\beta = O(\alpha\kappa_*^{-4})$. If $\mathrm{dist}_0 = O(k^{-0.75}\kappa_*^{-1.5})$, then almost surely*

$$\mathrm{dist}(\mathbf{B}_T, \mathbf{B}_*) \leq \left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^{T-1},$$

where $E_0 \coloneqq 0.9 - \mathrm{dist}_0^2$.

Theorem 2 shows that the initial representation learning error for MAML must scale as $O(k^{-0.75}\kappa_*^{-1.5})$, which can be much smaller than the constant scaling that is sufficient for ANIL to learn the representation (see Theorem 1). Next we give the main result for FO-MAML, which requires an additional condition that the norm of the average of the ground-truth heads sampled on each iteration is small. This condition arises due to the fact that the FO-MAML updates are approximations of the exact MAML updates, and thus have a bias that depends on the average of the ground-truth heads. Without control of this bias, the iterates $\mathbf{B}_t$ and $\mathbf{w}_t$ may diverge.

Theorem 3. *Consider the infinite sample case for FO-MAML, where $m_{in} = m_{out} = \infty$. Further, suppose the conditions in Assumptions 1 and 2 hold, the initial weights are selected as $\mathbf{w}_0 = \mathbf{0}_k$ and $\alpha\mathbf{B}_0^\top\mathbf{B}_0 = \mathbf{I}_k$, and the step sizes satisfy $\alpha = O(\frac{1}{\sqrt{k}L_*})$ and $\beta = O(\alpha\kappa_*^{-4})$. If the initial error satisfies $\mathrm{dist}_0 = O(k^{-0.5}\kappa_*^{-1})$, and the average of the true heads almost surely satisfies $\|\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\|_2 = O(k^{-1.5}\kappa_*^{-3}\mu_*)$ for all times t, then almost surely*

$$\mathrm{dist}(\mathbf{B}_T, \mathbf{B}_*) \leq \left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^{T-1},$$

where $E_0 \coloneqq 0.9 - \mathrm{dist}_0^2$.

Theorem 3 shows that FO-MAML learns $\mathrm{col}(\mathbf{B}_*)$ as long as the initial principal angle is small and $\|\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\|_2 = O(k^{-1.5}\kappa_*^{-3}\mu_*)$ on all iterations, due to the biased updates. Note that the FO-ANIL updates are also biased, but this bias scales with $\|\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t\|_2$, which is eventually decreasing quickly enough to make the cumulative error induced by the bias negligible without any additional conditions. In contrast, $\|\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t\|_2$ is not guaranteed to decrease for FO-MAML due to the inner loop adaptation of the representation, so we need the additional condition.

To the best of our knowledge, the above theorems are the first results to show that ANIL, MAML, and their first-order approximations learn representations in any setting. Moreover, they are the first to show how *task diversity* plays a key role in representation learning from an optimization perspective, to the best of our knowledge. Due the the restrictions on β and α, Theorems 1 and 2 show that the rate of contraction of principal angle distance diminishes with less task diversity. Thus, the more diverse the tasks, i.e. the smaller κ_*, the faster that ANIL and MAML learn the representation. Additionally, the less diverse the tasks, the more accurate initialization that MAML requires, and the tighter that the true heads must be centered around zero to control the FO-MAML bias.

4.1. Finite-sample results

Thus far we have only considered the infinite sample case, i.e., $m_{in} = m_{out} = \infty$, to highlight the reasons that the adaptation updates in MAML and ANIL are essential for representation learning. Next, we study the finite sample setting. Indeed, establishing our results for the finite sample case is more challenging, but the mechanisms by which ANIL and MAML learn representations for finite m_{in} and m_{out} are very similar to the infinite-sample case, and the finite-sample problem reduces to showing concentration of the updates to the infinite-sample updates.

For MAML, this concentration requires assumptions on sixth and eighth-order products of the data which arise due to the inner-loop updates. In light of this, for the sake of readability we only give the finite-sample result for ANIL and FO-ANIL, whose analyses require only standard assumptions on the data, as we state below.

Assumption 3 (Sub-gaussian feature distribution)**.** *For $\mathbf{x} \sim p$, $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}_d$. Moreover, $\mathbf{x}$ is $\mathbf{I}_d$-sub-gaussian in the sense that $\mathbb{E}[\exp(\mathbf{v}^\top\mathbf{x})] \leq \exp(\frac{\|\mathbf{v}\|_2^2}{2})\ \forall\ \mathbf{v}$.*

Under this assumption, we can show the following.

Theorem 4 (ANIL Finite Samples)**.** *Consider the finite-sample case for ANIL and FO-ANIL. Suppose Assumptions 1, 2 and 3 hold, $\alpha = O((\sqrt{k}L_* + \sigma)^{-1})$ and β is chosen as in Theorem 1. For some $\delta > 0$ to be defined later, let $E_0 = 0.9 – \mathrm{dist}_0^2 - \delta$ and assume E_0 is lower bounded by a positive*

constant. Suppose the sample sizes satisfy $m_{in} = \tilde{\Omega}(M_{in})$ *and* $m_{out} = \tilde{\Omega}(M_{out})$ *for some expressions* M_{in}, M_{out} *to be defined later. Then both ANIL and FO-ANIL satisfy:*

$$\mathrm{dist}(\mathbf{B}_T, \mathbf{B}_*) \leq \left(1 - 0.5\beta\alpha\mu_*^2\right)^{T-1} + \tilde{O}(\delta)$$

where for ANIL,

$$M_{in} = k^3 + \tfrac{k^3 d}{n},\ M_{out} = k^2 + \tfrac{dk+k^3}{n},$$
$$\delta = (\sqrt{k}\kappa_*^2 + \tfrac{\kappa_*\sigma}{\mu_*} + \tfrac{\sigma^2}{\mu_*^2})(k + \tfrac{\sqrt{dk}}{\sqrt{n}})(\tfrac{1}{\sqrt{m_{in}}} + \tfrac{1}{\sqrt{m_{out}}})$$

and for FO-ANIL,

$$M_{in} = k^2,\ M_{out} = \tfrac{dk+k^3}{n}$$
$$\delta = (\sqrt{k}\kappa_*^2 + \tfrac{\kappa_*\sigma}{\mu_*} + \tfrac{\sigma^2}{\mu_*^2\sqrt{m_{in}}})\tfrac{\sqrt{dk}}{\sqrt{nm_{out}}}$$

with probability at least $1 - \frac{T}{\mathrm{poly}(n)} - \frac{T}{\mathrm{poly}(m_{in})} - O(Te^{-90k})$.

For ease of presentation, the $\tilde{\Omega}()$ notation excludes log factors and all parameters besides k, d and n; please refer to Theorem 8 in Appendix E for the full statement. We focus on dimension parameters and n here to highlight the sample complexity benefits conferred by ANIL and FO-ANIL compared to solving each task separately ($n = 1$). Theorem 4 shows that ANIL requires only $m_{in} + m_{out} = \tilde{\Omega}(k^3 + \frac{k^3 d}{n})$ samples per task to reach a neighborhood of the ground-truth solution. Since $k \ll d$ and n can be large, this sample complexity is far less than the $\tilde{\Omega}(d)$ required to solve each task individually (Hsu et al., 2012). Note that more samples are required for Exact ANIL because the second-order updates involve higher-order products of the data, which have heavier tails than the analogous terms for FO-ANIL.

5. Proof sketch

We now discuss how we prove the results in greater detail. We focus on the FO-ANIL case because the presentation is simplest yet still illuminates the key ideas used in all proofs.

5.1. Theorem 1 (FO-ANIL)

Intuition. Our goal is to show that the distance between the column spaces of $\mathbf{B}_t$ and $\mathbf{B}_*$, i.e. $\mathrm{dist}_t := \|\hat{\mathbf{B}}_{*,\perp}^\top \hat{\mathbf{B}}_t\|_2$ is converging to zero at a linear rate for all t. We will use an inductive argument in which we assume favorable conditions to hold up to time t, and will prove they continue to hold at time $t+1$. To show dist_{t+1} is linearly decaying, it is helpful to first consider the non-normalized energy in the subspace orthogonal to the ground-truth, namely $\|\hat{\mathbf{B}}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2$. We have observed in equation (3) that if the inner-loop adapted heads $\mathbf{w}_{t,i}$ at time t are diverse, then the FO-ANIL update of the representation subtracts energy from the previous representation and adds energy from the ground-truth representation. Examining (3) closer, we notice that the only energy in the column space of the new representation that can be orthogonal to the ground-truth subspace is contributed by the previous representation, and this energy is contracting at a rate proportional to the condition number of the matrix formed by the adapted heads. In particular, if we define the matrix $\mathbf{\Psi}_t := \frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top$, then we have

$$\begin{aligned}\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 &= \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t(\mathbf{I} - \beta\mathbf{\Psi}_t)\|_2 \\ &\leq (1 - \beta\lambda_{\min}(\mathbf{\Psi}_t))\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2, \end{aligned} \quad (6)$$

as long as $\beta \leq 1/\lambda_{\max}(\mathbf{\Psi}_t)$. Therefore, to show that the normalized energy $\|\hat{\mathbf{B}}_{*,\perp}^\top \hat{\mathbf{B}}_{t+1}\|_2$ approaches zero, we aim to show: (I) The condition number of $\mathbf{\Psi}_t$ continues to stay controlled and finite, which implies linear convergence of the non-normalized energy in $\mathrm{col}(\mathbf{B}_*)^\perp$ according to (6); and (II) The minimum singular value of the representation $\mathbf{B}_{t+1}$ is staying the same. Otherwise, the energy orthogonal to the ground-truth subspace could be decreasing, but the representation could be becoming singular, which would mean the distance to the ground-truth subspace is not decreasing.

To show (I), note that the adapted heads are given by:

$$\mathbf{w}_{t,i} = \underbrace{\mathbf{\Delta}_t \mathbf{w}_t}_{\text{non-unique}} + \alpha \underbrace{\mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i}}_{\text{unique}}, \quad (7)$$

where $\mathbf{\Delta}_t := \mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t$. The vector $\mathbf{\Delta}_t\mathbf{w}_t$ is present in every $\mathbf{w}_{t,i}$, so we refer to it as the non-unique part of $\mathbf{w}_{t,i}$. On the other hand, $\alpha\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}$ is the unique part of $\mathbf{w}_{t,i}$. Equation (7) shows that if the non-unique part of each $\mathbf{w}_{t,i}$ is relatively small compared to the unique part, then $\mathbf{\Psi}_t \approx \alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{\Psi}_{*,t}\mathbf{B}_*^\top\mathbf{B}_t$, meaning the $\mathbf{w}_{t,i}$'s are almost as diverse as the ground-truth heads. So we aim to show $\|\mathbf{\Delta}_t\|_2$ and $\|\mathbf{w}_t\|_2$ remain small for all t. We specifically need to show they are small compared to $\sigma_{\min}^2(\mathbf{B}_t^\top\mathbf{B}_*)$, since this quantity roughly lower bounds the energy in the diverse part of $\mathbf{w}_{t,i}$. One can show that $\sigma_{\min}^2(\hat{\mathbf{B}}_t^\top\mathbf{B}_*) = 1 - \mathrm{dist}_t^2$, so we need to use that dist_t is decreasing in order to lower bound the energy in the unique part of $\mathbf{w}_{t,i}$.

It is also convenient to track $\|\mathbf{\Delta}_t\|_2$ in order to show (II), since $\|\mathbf{\Delta}_{t+1}\|_2 \leq \varepsilon$ implies $\sigma_{\min}(\mathbf{B}_{t+1}) \geq \frac{\sqrt{1-\varepsilon}}{\sqrt{\alpha}}$. Note that for (II), we need control of $\|\mathbf{\Delta}_{t+1}\|_2$, whereas to show (I) we needed control of $\|\mathbf{\Delta}_t\|_2$. This difference in time indices is accounted for by the induction we will soon discuss.

It is now evident why it makes sense to initialize with $\|\mathbf{\Delta}_0\|_2 = 0$ and $\|\mathbf{w}_t\|_2 = 0$ (in fact, they do not have to be exactly zero; any initialization with $\|\mathbf{w}_0\|_2 = O(\sqrt{\alpha})$ and $\|\mathbf{\Delta}_t\|_2 = O(\alpha^2)$ would suffice). However, proving that $\|\mathbf{\Delta}_t\|_2$ and $\|\mathbf{w}_t\|_2$ remain small is difficult because the algorithm lacks explicit regularization or a normalization step after each round. Empirically, $\sigma_{\min}(\mathbf{B}_t)$ may decrease and $\|\mathbf{w}_t\|_2$ may increase on any particular round, so it is not clear why $\sigma_{\min}(\mathbf{B}_t)$ does not go to zero (i.e. $\|\mathbf{\Delta}_t\|_2$ does not go to 1) and $\|\mathbf{w}_t\|_2$ does not blow up. To address these issues, one could add an explicit regularization term

to the loss functions or an orthonormalization step to the algorithm, but doing so is empirically unnecessary and would not be consistent with the ANIL formulation or algorithm.

Inductive structure. We overcome the aforementioned challenges by executing a multi-way induction that involves the following six inductive hypotheses:

1. $A_1(t) \coloneqq \{\|\mathbf{w}_t\|_2 = O(\sqrt{\alpha}\min(1, \frac{\mu_*^2}{\eta_*^2})\eta_*)\}$,
2. $A_2(t) \coloneqq \big\{\|\mathbf{\Delta}_t\|_2 \le \rho\|\mathbf{\Delta}_{t-1}\|_2 + O(\beta^2\alpha^2 L_*^4 \operatorname{dist}_{t-1}^2)\big\}$,
3. $A_3(t) \coloneqq \big\{\|\mathbf{\Delta}_t\|_2 \le \frac{1}{10}\big\}$,
4. $A_4(t) \coloneqq \big\{0.9\alpha E_0\mu_*\mathbf{I}_k \preceq \mathbf{\Psi}_t \preceq 1.2\alpha L_*^2\mathbf{I}_k\big\}$,
5. $A_5(t) \coloneqq \big\{\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2 \le \rho\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t-1}\|_2\big\}$,
6. $A_6(t) \coloneqq \{\operatorname{dist}_t = \|\hat{\mathbf{B}}_{*,\perp}^\top\hat{\mathbf{B}}_t\|_2 \le \rho^{t-1}\}$,

where $\rho = 1 - 0.5\beta\alpha E_0\mu_*^2$. Our previous intuition motivates our choice of inductive hypotheses $A_1(t), \dots, A_5(t)$ as intermediaries to ultimately show that dist_t linearly converges to zero. More specifically, $A_1(t), A_2(t)$, and $A_3(t)$ bound $\|\mathbf{w}_t\|_2$ and $\|\mathbf{\Delta}_t\|_2$, $A_4(t)$ controls the diversity of the inner loop-adapted heads, and $A_5(t)$ and $A_6(t)$ confirm that the learned representation approaches the ground-truth. We employ two upper bounds on $\|\mathbf{\Delta}_t\|_2$ because we need to use that $\{\|\mathbf{\Delta}_t\|_2\}_t$ is both summable ($A_2(t)$) and uniformly small ($A_3(t)$) to complete different parts of the induction. In particular, if true for all t, $A_2(t)$ shows that $\|\mathbf{\Delta}_t\|_2$ may initially increase, but eventually linearly converges to zero due to the linear convergence of dist_t. The initialization implies each inductive hypothesis holds at time $t = 1$. We must show they hold at time $t + 1$ if they hold up to time t.

To do this, we employ the logic visualized in Figure 2. The top level events ($A_1(t+1), A_2(t+1), A_5(t+1)$) are most "immediate" in the sense that they follow directly from other events at all times up to and including t (via the dashed green arrows). The proofs of all other events at time $t+1$ require the occurrence of other events at time $t + 1$, with more logical steps needed as one moves down the graph, and solid red arrows denoting implications from and to time $t + 1$. In particular, $A_3(t + 1)$ requires the events up to and including time t *and* a top-level event at $t + 1$, namely $A_2(t + 1)$, so it is in the second level. Similarly, $A_6(t + 1)$ requires events up to and including time t and the second-level event at $t+1$, so it is in the third level, and so on.

Recall that our intuition is that diverse adapted heads lead to contraction of the non-normalized representation distance. This logic drives the implication $A_4(t) \implies A_5(t + 1)$. We then reasoned that contraction of the non-normalized distance leads to linear convergence of the distance as long as the minimum singular value of the representation is controlled from below. This intuition is captured in the implication $A_5(t + 1) \cap A_3(t + 1) \implies A_6(t + 1)$.

We also discussed that the diversity of the adapted heads depends on the global head being small, the representation being close to a scaled orthonormal matrix, and the representation distance being bounded away from 1 at the start of that iteration. This is ensured by the implication that the adapted heads are again diverse on iteration $t + 1$, in particular $A_1(t+1) \cap A_3(t+1) \cap A_6(t+1) \implies A_4(t+1)$. The other implications in the graph are technical and needed to control $\|\mathbf{w}_{t+1}\|_2$ and $\|\mathbf{\Delta}_{t+1}\|_2$.

Proving the implications. We now formally discuss each implication, starting with the top level. Full proofs are provided in Appendix C.

- $A_4(t) \implies A_5(t+1)$. This is true by equation (6).
- $A_1(t) \cap A_3(t) \cap A_6(t) \implies A_2(t+1)$. It can be shown that $\mathbf{\Delta}_{t+1}$ is of the form:

$$\mathbf{\Delta}_{t+1} = \mathbf{\Delta}_t(\mathbf{I}_k - \beta\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{\Psi}_{*,t}\mathbf{B}_*^\top\mathbf{B}_t) + \mathbf{N}_t \tag{8}$$

  for some matrix $\mathbf{N}_t$ whose norm is upper bounded by a linear combination of $\|\mathbf{\Delta}_t\|_2$ and dist_t. We next use

$$\begin{aligned}\lambda_{\min}(\mathbf{B}_t^\top\mathbf{B}_*\mathbf{\Psi}_{*,t}\mathbf{B}_*^\top\mathbf{B}_t) &\ge \mu_*^2\sigma_{\min}^2(\mathbf{B}_t^\top\mathbf{B}_*) \\ &\ge \tfrac{0.9}{\alpha}\mu_*^2(1-\operatorname{dist}_t^2)\end{aligned} \tag{9}$$

  where (9) follows by $\sigma_{\min}^2(\hat{\mathbf{B}}_t^\top\mathbf{B}_*) = 1-\operatorname{dist}_t^2$ and $A_3(t)$. The proof follows by applying $A_6(t)$ to control $1-\operatorname{dist}_t^2$.

- $(\cap_{s=1}^t A_2(s) \cap A_6(s)) \implies A_1(t+1)$. This is the most difficult induction to show. The FO-ANIL dynamics are such that $\|\mathbf{w}_t\|_2$ may increase on every iteration throughout the entire execution of the algorithm. However, we can exploit the fact that the amount that it increases is proportional to $\|\mathbf{\Delta}_t\|_2$, which we can show is summable due to the linear convergence of dist_t. First, we have

$$\mathbf{w}_{t+1} = (\mathbf{I}_k - \beta\mathbf{B}_t^\top\mathbf{B}_t\mathbf{\Delta}_t)\mathbf{w}_t + \frac{\beta}{n}\sum_{i=1}^n \mathbf{\Delta}_t\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}$$

  which implies $\|\mathbf{w}_t\|_2$ increases on each iteration by $O(\frac{\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2\eta_*)$. In particular,

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2 &\overset{(a)}{\le} (1 + \tfrac{2\beta}{\alpha}\|\mathbf{\Delta}_t\|_2)\|\mathbf{w}_t\|_2 + \tfrac{2\beta L_*}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2 \\ &\overset{(b)}{\le} \sum_{s=0}^{t-1}\tfrac{2\beta\eta_*}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2\prod_{r=s}^{t-1}(1 + \tfrac{2\beta}{\alpha}\|\mathbf{\Delta}_r\|_2) \\ &\overset{(c)}{\le} \sum_{s=0}^{t-1}\tfrac{2\beta\eta_*}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2\Big(1+\tfrac{1}{t-s}\sum_{r=s}^{t-1}\tfrac{2\beta}{\alpha}\|\mathbf{\Delta}_r\|_2\Big)^{t-s}\end{aligned}$$

  where (b) follows by recursively applying (a) for $t, t-1, \dots$ and (c) follows by the AM-GM inequality. Next, for any $s \in [t]$, recursively apply $A_2(s)$,$A_2(s-1)$, ... and

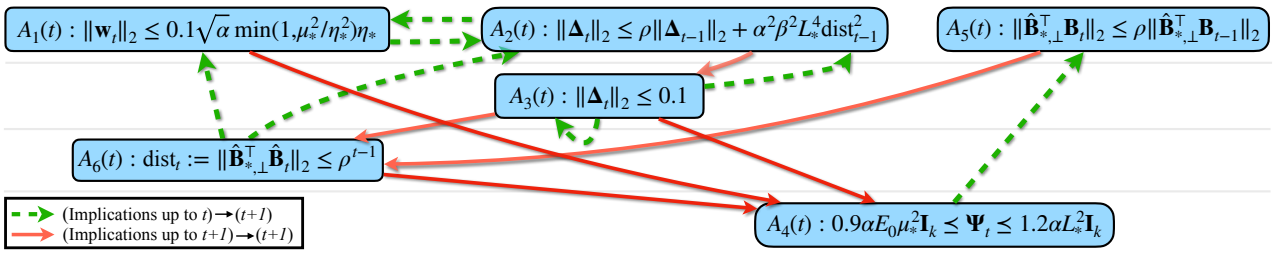

Figure 2. Logical flow of the proof. Note that there are no cycles among the implications from $t+1$ to $t+1$, so the logic is consistent.

use $A_6(r)\ \forall r \in [s]$ to obtain, for an absolute constant c,

$$\|\mathbf{\Delta}_s\|_2 \overset{(d)}{\le} c\sum_{r=0}^{s-1}\rho^{s-r}\beta^2\alpha^2 L_*^4\,\mathrm{dist}_r^2 \le c\rho^s\sum_{r=0}^{s-1}\rho^r\beta^2\alpha^2 L_*^4$$

Plugging (d) into (c), computing the sum of geometric series, and applying the choice of β completes the proof.

- $A_2(t+1)\cap A_3(t) \implies A_3(t+1)$. This follows straightforwardly since β is chosen sufficiently small.
- $A_3(t+1)\cap\left(\cap_{s=1}^{t+1}A_5(s)\right)\cap A_6(t) \implies A_6(t+1)$. Using the definition of the principal angle distance, the Cauchy-Schwarz inequality, and $\cap_{s=1}^{t+1}A_5(s)$, we can show

$$\mathrm{dist}_{t+1} \le \tfrac{1}{\sigma_{\min}(\mathbf{B}_{t+1})}\|\hat{\mathbf{B}}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 \le \tfrac{\sigma_{\max}(\mathbf{B}_0)}{\sigma_{\min}(\mathbf{B}_{t+1})}\rho^t\,\mathrm{dist}_0$$

from which the proof follows after applying $A_3(t+1)$ and the initial conditions. Note that here we have normalized the representation only once at time $t+1$ and used the contraction of the non-normalized energy to recurse from $t+1$ to 0, resulting in a $\frac{\sigma_{\max}(\mathbf{B}_0)}{\sigma_{\min}(\mathbf{B}_{t+1})}$ scaling error. If we instead tried to directly show the contraction of distance and thereby normalized analytically on every round, we would obtain $\mathrm{dist}_{t+1} \le \left(\prod_{s=0}^{t}\frac{\sigma_{\max}(\mathbf{B}_s)}{\sigma_{\min}(\mathbf{B}_{s+1})}\right)\rho^t\,\mathrm{dist}_0$, meaning a $\prod_{s=0}^{t}\frac{\sigma_{\max}(\mathbf{B}_s)}{\sigma_{\min}(\mathbf{B}_{s+1})}$ scaling error, which is too large because $\mathbf{B}_s$ is in fact not normalized on every round.

- $A_1(t+1)\cap A_3(t+1)\cap A_6(t+1) \implies A_4(t+1)$. This follows by expanding each $\mathbf{w}_{t,i}$ as in (7), and using similar logic as in (9).

5.2. Other results – ANIL, FO-MAML, and MAML

For ANIL, the inductive structure is nearly identical. The only meaningful change in the proof is that the second-order updates imply $\|\mathbf{w}_{t+1}\|_2 - \|\mathbf{w}_t\|_2 = O(\|\mathbf{\Delta}_t\|_2^2)$, which is smaller than the $O(\|\mathbf{\Delta}_t\|_2)$ for FO-ANIL, and thereby allows to control $\|\mathbf{w}_{t+1}\|_2$ with a potentially larger β.

For FO-MAML and MAML, recall that the inner loop update of the representation weakens the benefit of adapted head diversity (see Section 3.1). Thus, larger adapted head diversity is needed to learn $\mathrm{col}(\mathbf{B}_*)$. Specifically, we require a tighter bound of $\|\mathbf{\Delta}_t\|_2 = O(\alpha^2)$, compared to the $\|\mathbf{\Delta}_t\|_2 = O(1)$ bound in ANIL, and for FO-MAML, we also require a tighter bound on $\|\mathbf{w}_t\|_2$ (recall from Section 5 that smaller $\|\mathbf{\Delta}_t\|_2$ and $\|\mathbf{w}_t\|_2$ improves adapted head diversity). Moreover, to obtain tight bounds on $\|\mathbf{w}_{t+1}\|_2$ we can no longer use that $\|\mathbf{w}_{t+1}\|_2 - \|\mathbf{w}_t\|_2$ is controlled by $\|\mathbf{\Delta}_t\|_2$ due to to additional terms in the outer loop update. To overcome these issues, we must make stricter assumptions on the initial distance, and in the case of FO-MAML, on the average ground-truth head. Please see Appendix D for details.

Finally, the proof of Theorem 4 relies on showing concentration of the finite-sample gradients to the population gradients. The principal challenge is showing this concentration for fourth-order products of the data that arise in the ANIL updates, since we cannot apply standard methods to these higher-order products while maintaining $o(d)$ samples per task. Instead, we leverage the low-rankness of the products by applying a truncated version of the concentration result for low-rank random matrices from (Magen & Zouzias, 2011). We also use the L4-L2-hypercontractiveness of the data to control the bias in these higher-order products. Details are found in Appendix E.

6. Numerical simulations

In this section we run numerical simulations to verify our theoretical findings. First, we explore the effect of task diversity on MAML's rate of convergence to the ground-truth representation. In Figure 3, we execute MAML on the task population losses ($m_{in} = m_{out} = \infty$) in the multi-task linear representation learning setting. We set $d = 100$ and $k = n = 5$. On each round, the ground-truth heads are sampled i.i.d. from $\mathcal{N}(\mathbf{0}, \mathrm{diag}([1,\ldots,1,\mu^2]))$, where $\mu^2 < 1$. We randomly draw $\mathbf{B}_*$ and $\mathbf{B}_0$ at the start of algorithm execution. The parameter μ^2 controls task diversity, with larger μ^2 meaning the ground-truth heads are closer to isotropic and therefore more diverse. The results show that MAML's linear convergence rate improves with greater task diversity, consistent with Theorem 2.

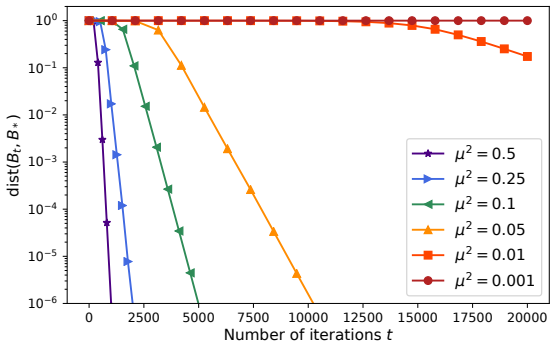

Figure 3. **Task diversity improves convergence rate.** Principal angle distance vs number of iterations for MAML with varying ground-truth head distributions. The larger value of μ^2, the more diverse ground-truth heads.

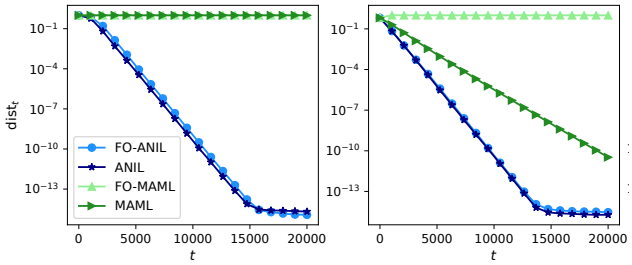

Figure 4. **MAML and FO-MAML initialization and ground-truth mean conditions are empirically necessary.** (Left) Random $\mathbf{B}_0$. (Right) Methodical $\mathbf{B}_0$. In both cases, the mean ground-truth head is far from zero.

Next, we show that the additional conditions relative to ANIL and FO-ANIL required for MAML and FO-MAML to learn the ground-truth representation are empirically necessary. That is, (i) MAML and FO-MAML require a good initialization relative to the underlying task diversity, and (ii) FO-MAML further requires the ground-truth heads to be concentrated around zero. To test these conditions, we set $d = 20$, $n = k = 3$, randomly draw $\mathbf{B}_*$, and use the task population losses. The ground-truth heads are drawn as $\mathbf{w}_{*,t,i} \sim \mathcal{N}(10\mathbf{1}_k, \mathbf{I}_k)$. Ground-truth task diversity is thus low, since most of the energy points in the direction $\mathbf{1}_k$. In Figure 4 (left), we use a random Gaussian initialization of $\mathbf{B}_0$, which has $\text{dist}_0 \approx 0.99$. In Figure 4 (right), we initialize with a noisy version of $\mathbf{B}_*$ satisfying $\text{dist}_0 \in [0.65, 0.7]$. The plots show that in this low-diversity setting, MAML requires good initialization to achieve linear convergence, whereas FO-MAML cannot obtain it even with good initialization, as $\|\mathbb{E}[\mathbf{w}_{*,t,i}]\| \gg 0$. Lastly, note that Figure 1 employs the same setting as Figure 4 (left), except that the mean of the ground-truth heads is zero in the former case, which leads to all four GBML approaches learning $\text{col}(\mathbf{B}_*)$.

7. Conclusion

Our analysis reveals that ANIL, MAML, and their first-order approximations exploit task diversity via inner adaptation steps of the head to recover the ground-truth representation in the multi-task linear representation learning setting. Further, task diversity helps these algorithms exhibit an implicit regularization that keeps the learned representation well-conditioned. However, the inner adaptation of the representation plays a restrictive role, inhibiting MAML and FO-MAML from achieving global convergence. To the best of our knowledge, these are the first results showing that GBML algorithms can learn a low-dimensional subspace.

Acknowledgements

This research is supported in part by NSF Grants 2127697, 2019844, 2107037, and 2112471, ARO Grant W911NF2110226, ONR Grant N00014-19-1-2566, the Machine Learning Lab (MLL) at UT Austin, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

References

Antoniou, A., Edwards, H., and Storkey, A. How to train your MAML. In *Seventh International Conference on Learning Representations*, 2019.

Arnold, S., Iqbal, S., and Sha, F. When MAML Can Adapt Fast and How to Assist When it Cannot. In *International Conference on Artificial Intelligence and Statistics*, pp. 244–252. PMLR, 2021.

Balcan, M.-F., Khodak, M., and Talwalkar, A. Provable Guarantees for Gradient-Based Meta-Learning. In *International Conference on Machine Learning*, pp. 424–433. PMLR, 2019.

Baxter, J. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Bernacchia, A. Meta-Learning with Negative Learning Rates. In *International Conference on Learning Representations*, 2020.

Bertinetto, L., Henriques, J. F., Torr, P., and Vedaldi, A. Meta-Learning with Differentiable Closed-form Solvers. In *International Conference on Learning Representations*, 2018.

Bouniot, Q., Redko, I., Audigier, R., Loesch, A., Zotkin, Y., and Habrard, A. Towards Better Understanding Meta-learning Methods through Multi-task Representation Learning Theory. *arXiv preprint arXiv:2010.01992*, 2020.

Bullins, B., Hazan, E., Kalai, A., and Livni, R. Generalize Across Tasks: Efficient Algorithms for Linear Representation Learning. In *Algorithmic Learning Theory*, pp. 235–246. PMLR, 2019.

Caruana, R. Multitask Learning. *Machine learning*, 28(1): 41–75, 1997.

Chen, J., Wu, X.-M., Li, Y., Li, Q., Zhan, L.-M., and Chung, F.-l. A Closer Look at the Training Strategy for Modern Meta-Learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Chua, K., Lei, Q., and Lee, J. D. How fine-tuning allows for effective Meta-Learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting Shared Representations for Personalized Federated Learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Collins, L., Mokhtari, A., and Shakkottai, S. How Does the Task Landscape Affect MAML performance? *arxiv preprint arXiv:2010.14672*, 2022.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. Incremental Learning-to-Learn with Statistical Guarantees. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, volume 34, pp. 457–466. AUAI, 2018.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-Shot Learning via Learning the Representation, Provably. In *International Conference on Learning Representations*, 2020.

Fallah, A., Mokhtari, A., and Ozdaglar, A. On the Convergence Theory of Gradient-Based Model-Agnostic Meta-Learning Algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1082–1092. PMLR, 2020a.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized Federated learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3557–3568, 2020b.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Generalization of Model-Agnostic Meta-Learning Algorithms: Recurring and Unseen Tasks. *Advances in Neural Information Processing Systems*, 34, 2021.

Finn, C., Abbeel, P., and Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

Finn, C., Xu, K., and Levine, S. Probabilistic Model-Agnostic Meta-Learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online Meta-Learning. In *International Conference on Machine Learning*, pp. 1920–1930. PMLR, 2019.

Goldblum, M., Reich, S., Fowl, L., Ni, R., Cherepanova, V., and Goldstein, T. Unraveling Meta-Learning: Understanding Feature Representations for Few-Shot Tasks. In *International Conference on Machine Learning*, pp. 3607–3616. PMLR, 2020.

Hospedales, T. M., Antoniou, A., Micaelli, P., and Storkey, A. J. Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Hsu, D., Kakade, S. M., and Zhang, T. Random Design Analysis of Ridge Regression. In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012.

Ji, K., Lee, J. D., Liang, Y., and Poor, H. V. Convergence of Meta-Learning with Task-Specific Adaptation over Partial Parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020a.

Ji, K., Yang, J., and Liang, Y. Multi-Step Model-Agnostic Meta-Learning: Convergence and Improved algorithms. *CoRR*, abs/2002.07836, 2020b.

Jiang, Y., Konečnỳ, J., Rush, K., and Kannan, S. Improving Federated Learning Personalization via Model Agnostic Meta Learning. *arXiv preprint arXiv:1909.12488*, 2019.

Kao, C. H., Chiu, W.-C., and Chen, P.-Y. MAML is a noisy contrastive learner in classification. In *International Conference on Learning Representations*, 2022.

Kumar, R., Deleu, T., and Bengio, Y. Effect of Diversity in Meta-Learning. In *Fifth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*, 2021.

Lee, K., Maji, S., Ravichandran, A., and Soatto, S. Meta-Learning with Differentiable Convex Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10657–10665, 2019.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-SGD: Learning to Learn Quickly for Few-Shot Learning. *arXiv preprint arXiv:1707.09835*, 2017.

Magen, A. and Zouzias, A. Low Rank Matrix-Valued Chernoff Bounds and Approximate Matrix Multiplication. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pp. 1422–1436. SIAM, 2011.

Maurer, A., Pontil, M., and Romera-Paredes, B. The Benefit of Multitask Representation Learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

McNamara, D. and Balcan, M.-F. Risk Bounds for Transferring Representations With and Without Fine-Tuning. In *International Conference on Machine Learning*, pp. 2373–2381. PMLR, 2017.

Ni, R., Goldblum, M., Sharaf, A., Kong, K., and Goldstein, T. Data augmentation for Meta-Learning. In *International Conference on Machine Learning*, pp. 8152–8161. PMLR, 2021.

Nichol, A. and Schulman, J. Reptile: A Scalable Meta-Learning Algorithm. *arXiv preprint arXiv:1803.02999*, 2:2, 2018.

Oh, J., Yoo, H., Kim, C., and Yun, S.-Y. BOIL: Towards Representation Change for Few-Shot Learning. In *International Conference on Learning Representations*, 2020.

Raghu, A., Raghu, M., Bengio, S., and Vinyals, O. Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML. In *International Conference on Learning Representations*, 2020.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-Learning with Implicit Gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

Ravi, S. and Larochelle, H. Optimization as a Model for Few-Shot Learning. 2016.

Saunshi, N., Zhang, Y., Khodak, M., and Arora, S. A Sample Complexity Separation Between Non-Convex and Convex Meta-Learning. In *International Conference on Machine Learning*, pp. 8512–8521. PMLR, 2020.

Saunshi, N., Gupta, A., and Hu, W. A Representation Learning Perspective on the Importance of Train-Validation Splitting in Meta-Learning. In *International Conference on Machine Learning*, pp. 9333–9343. PMLR, 2021.

Schmidhuber, J. *Evolutionary Principles in Self-Referential Learning, or on Learning how to Learn: the Meta-Meta-...* PhD thesis, Technische Universität München, 1987.

Setlur, A., Li, O., and Smith, V. Is Support Set Diversity Necessary for Meta-Learning? *arXiv preprint arXiv:2011.14048*, 2020.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Statistically and Computationally Efficient Linear Meta-Representation Learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Tripuraneni, N., Jordan, M., and Jin, C. On the Theory of Transfer Learning: The Importance of Task Diversity. *Advances in Neural Information Processing Systems*, 33: 7852–7862, 2020.

Tripuraneni, N., Jin, C., and Jordan, M. Provable Meta-Learning of Linear Representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching Networks for One Shot Learning. *Advances in Neural Information Processing Systems*, 29:3630–3638, 2016.

Wang, H., Zhao, H., and Li, B. Bridging Multi-task Learning and Meta-Learning: Towards Efficient Training and Effective Adaptation. In *International Conference on Machine Learning*. PMLR, 2021a.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. On the Global Optimality of Model-Agnostic Meta-Learning. In *International Conference on Machine Learning*, pp. 9837–9846. PMLR, 2020.

Wang, X., Yuan, S., Wu, C., and Ge, R. Guarantees for Tuning the Step Size Using a Learning-to-Learn Approach. In *International Conference on Machine Learning*, pp. 10981–10990. PMLR, 2021b.

Xu, Z. and Tewari, A. Representation Learning Beyond Linear Prediction Functions. *Advances in Neural Information Processing Systems*, 34, 2021.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. Bayesian Model-Agnostic Meta-Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7343–7353, 2018.

Zhou, P., Yuan, X., Xu, H., Yan, S., and Feng, J. Efficient Meta Learning via Minibatch Proximal Update. In *Advances in Neural Information Processing Systems*, pp. 1532–1542, 2019.

Zintgraf, L., Shiarli, K., Kurin, V., Hofmann, K., and Whiteson, S. Fast Context Adaptation via Meta-Learning. In *International Conference on Machine Learning*, pp. 7693–7702. PMLR, 2019.

A. Additional Related Work

Meta-learning background. Multi-task representation learning and meta-learning have been of theoretical interest for many years (Schmidhuber, 1987; Caruana, 1997; Baxter, 2000). Recently, meta-learning methods have garnered much attention due to successful implementations in few-shot learning scenarios with deep networks. These modern approaches are roughly grouped into three categories: model-based (Ravi & Larochelle, 2016), metric-based (Snell et al., 2017; Vinyals et al., 2016), and gradient-based (Finn et al., 2017). In this paper we focus on gradient-based methods.

Gradient-based meta-learning and MAML. The practicality and simplicity of model-agnostic meta-learning (MAML) (Finn et al., 2017) has led to many experimental and theoretical studies of gradient-based meta-learning in addition to those mentioned in Section 1. There have been numerous algorithms proposed as extensions of MAML (Li et al., 2017; Finn et al., 2018; Yoon et al., 2018; Antoniou et al., 2019; Nichol & Schulman, 2018; Rajeswaran et al., 2019; Zhou et al., 2019; Raghu et al., 2020; Zintgraf et al., 2019), and MAML has been applied to online (Finn et al., 2019) and federated (Fallah et al., 2020b; Jiang et al., 2019) learning settings. Theoretical analyses of MAML and related methods have included sample complexity guarantees in online settings (Balcan et al., 2019; Denevi et al., 2018), general convergence guarantees (Fallah et al., 2020a; Ji et al., 2020b;a), and landscape analysis (Wang et al., 2020; Collins et al., 2022). Other works have studied the choice of inner loop step size (Wang et al., 2021b; Bernacchia, 2020) and generalization (Chen et al., 2020; Fallah et al., 2021), all without splitting model parameters.

Gradient-based meta-learning and representation learning. A growing line of research has endeavored to develop and understand gradient-based meta-learning with a representation learning perspective. Besides ANIL, multiple other meta-learning methods fix the representation in the inner loop (Lee et al., 2019; Bertinetto et al., 2018). Goldblum et al. (2020) showed that these meta-learners learn representations that empirically exhibit the desirable behavior of clustering features by class. However, they also gave evidence suggesting this is not true for MAML since it adapts the feature extractor during the inner loop. Meanwhile, other works have argued for the benefits of adapting the representation in the inner loop both experimentally, when the head is fixed (Oh et al., 2020), and theoretically, when the task optimal solutions may not share a representation (Chua et al., 2021).

Two recent works have argued that ANIL behaves similarly to empirically successful approaches for representation learning. Wang et al. (2021a) showed that the models learned by ANIL and multi-task learning with a shared representation and unique heads are close in function space for sufficiently wide and deep ReLU networks, when the inner loop learning rate and number of inner adaptation steps for ANIL is small. Kao et al. (2022) noticed that ANIL with the global head initialized at zero at the start of each round is a "noisy contrastive learner" in the sense that outer loop update for the representation is a gradient step with respect to a contrastive loss, which suggests that ANIL should learn quality representations. Moreover, they showed that zeroing the global head at the start of each round empirically improves the performance of both ANIL and MAML. However, neither work shows that ANIL, let alone MAML, can in fact learn expressive representations. Additionally, our analysis rigorously explains the observation from Kao et al. (2022) that having small $\|\mathbf{w}_t\|_2$ aids representation learning.

Meta-learning and task diversity. Initial efforts to empirically understand the effects of meta-training task diversity on meta-learning performance with neural networks have shown a promising connection between the two, although the picture is not yet clear. Ni et al. (2021) and Bouniot et al. (2020) made modifications to the the meta-training task distribution and the meta-learning objective, respectively, to improve the effective task diversity, and both resulted in significant improvements in performance for a range of meta-learners. On the other hand, Setlur et al. (2020) and Kumar et al. (2021) empirically argued that reducing the overall diversity of the meta-training dataset does not restrict meta-learning. However, Setlur et al. (2020) only considered reducing intra-task data diversity, not the diversity of the tasks themselves (as no classes were dropped from the meta-training dataset), and the results due to Kumar et al. (2021) showed that reducing the overall number of tasks seen during meta-training hurts performance for most meta-learners, including MAML.

Multi-task linear representation learning. Several works have studied a similar multi-task linear representation learning setting as ours (Saunshi et al., 2021; Thekumparampil et al., 2021; Collins et al., 2021; Du et al., 2020; Tripuraneni et al., 2021; Bullins et al., 2019; Maurer et al., 2016; McNamara & Balcan, 2017), but did not analyze MAML or ANIL. Moreover, multiple works have shown that task diversity is necessary to learn generalizable representations from a statistical perspective (Du et al., 2020; Tripuraneni et al., 2020; Xu & Tewari, 2021; Tripuraneni et al., 2021). Our work complements these by showing the benefit of task diversity to gradient-based meta-learning methods from an optimization perspective.

B. General Lemmas

First we define the following notations used throughout the proofs.

Notation	Definition
$\mathbf{\Delta}_t$	$\mathbf{I}_k - \alpha \mathbf{B}_t^\top \mathbf{B}_t$
$\bar{\mathbf{\Delta}}_t$	$\mathbf{I}_d - \alpha \mathbf{B}_t \mathbf{B}_t^\top$
L_*	$\max_{t\in[T]} \sigma_{\max}^{0.5}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\right) \le L_*$
μ_*	$0 < \mu_* \le \min_{t\in[T]} \sigma_{\min}^{0.5}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\right)$
η_*	$\max_{t\in[T]} \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\right\|_2 \le \eta_* \le L_*$
$L_{\max}$	$\max_{t\in[T], i\in[n]} \|\mathbf{w}_{*,t,i}\|_2 \le L_{\max} \le c\sqrt{k}L_*$ for constant c
κ_*	L_*/μ_*
$\kappa_{*,\max}$	$L_{\max}/\mu_*$

Now we have the following general lemmas.

Lemma 1. *Suppose Assumption 1 holds and for some t, $\mathrm{dist}_t^2 \le \frac{1}{1-\tau}\mathrm{dist}_0^2$. Also, suppose $\|\mathbf{\Delta}_s\|_2 \le \tau$ for all $s \in [t]$. Then, for $E_0 \coloneqq 1 - \tau - \mathrm{dist}_0^2$,*

$$\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{B}_t\right) \ge \frac{E_0\mu_*^2}{\alpha} \tag{10}$$

Proof. First note that since $\sigma_{\min}(\mathbf{A}_1\mathbf{A}_2) \ge \sigma_{\min}(\mathbf{A}_1)\sigma_{\min}(\mathbf{A}_2)$ for any two square matrices $\mathbf{A}_1$, $\mathbf{A}_2$, we have

$$\begin{aligned}
\sigma_{\min}\left(\mathbf{B}_t^\top \mathbf{B}_* \tfrac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{B}_t\right) &\ge \sigma_{\min}^2(\mathbf{B}_*^\top \mathbf{B}_t)\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\right) \\
&\ge \sigma_{\min}^2(\mathbf{B}_*^\top \mathbf{B}_t)\mu_*^2 \\
&\ge \sigma_{\min}^2(\mathbf{B}_*^\top \hat{\mathbf{B}}_t)\sigma_{\min}^2(\mathbf{R}_t)\mu_*^2 \\
&\ge \sigma_{\min}^2(\mathbf{B}_*^\top \hat{\mathbf{B}}_t)\tfrac{1-\tau}{\alpha}\mu_*^2
\end{aligned} \tag{11}$$

where $\mathbf{B}_t = \hat{\mathbf{B}}_t\mathbf{R}_t$ is the QR-factorization of $\mathbf{B}_t$. Next, observe that

$$\begin{aligned}
\mathrm{dist}_t^2 \coloneqq \|\mathbf{B}_{*,\perp}^\top \hat{\mathbf{B}}_t\|_2^2 &= \|(\mathbf{I}_d - \mathbf{B}_*\mathbf{B}_*^\top)\hat{\mathbf{B}}_t\|_2^2 \\
&= \max_{\mathbf{u}\in\mathbb{R}^k:\|\mathbf{u}\|_2=1} \mathbf{u}^\top \hat{\mathbf{B}}_t^\top(\mathbf{I}_d - \mathbf{B}_*\mathbf{B}_*^\top)(\mathbf{I}_d - \mathbf{B}_*\mathbf{B}_*^\top)\hat{\mathbf{B}}_t\mathbf{u} \\
&= \max_{\mathbf{u}\in\mathbb{R}^k:\|\mathbf{u}\|_2=1} \mathbf{u}^\top \hat{\mathbf{B}}_t^\top(\mathbf{I}_d - \mathbf{B}_*\mathbf{B}_*^\top)\hat{\mathbf{B}}_t\mathbf{u} \\
&= \max_{\mathbf{u}\in\mathbb{R}^k:\|\mathbf{u}\|_2=1} \mathbf{u}^\top(\mathbf{I}_k - \hat{\mathbf{B}}_t^\top \mathbf{B}_*\mathbf{B}_*^\top \hat{\mathbf{B}}_t)\mathbf{u} \\
&= 1 - \sigma_{\min}^2(\mathbf{B}_*^\top \hat{\mathbf{B}}_t) \\
\implies \sigma_{\min}^2(\mathbf{B}_*^\top \hat{\mathbf{B}}_t) &= 1 - \mathrm{dist}_t^2 \\
&\ge 1 - \tfrac{1}{1-\tau}\mathrm{dist}_0^2\,.
\end{aligned} \tag{12}$$

which gives the result after combining with (11). □

Note that all four algorithms considered (FO-ANIL, Exact ANIL, FO-MAML, Exact MAML) execute the same inner loop update procedure for the head. The following lemma characterizes the diversity of the inner loop-updated heads for all four algorithms, under some assumptions on the behavior of dist_t and $\|\mathbf{\Delta}_t\|_2$ which we will show are indeed satisfied later.

Lemma 2. *Suppose Assumption 1 holds and that on some iteration t, FO-ANIL, Exact ANIL, FO-MAML, and Exact MAML*

satisfy $\text{dist}_t^2 \leq \frac{1}{1-\tau}\text{dist}_0^2$ *and* $\|\mathbf{\Delta}_s\|_2 \leq \tau$ *for all* $s \in [t]$. *Then the inner loop-updated heads on iteration* t *satisfy:*

$$\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \geq \alpha E_0\mu_*^2 - 2(1+\|\mathbf{\Delta}_t\|_2)\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* \tag{13}$$

$$\text{and } \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \leq (\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + \sqrt{\alpha(1+\|\mathbf{\Delta}_t\|_2)}L_*)^2. \tag{14}$$

Proof. We first lower bound the minimum singular value. Observe that $\frac{1-\|\mathbf{\Delta}_t\|_2}{\alpha} \leq \sigma_{\min}^2(\mathbf{B}_t) \leq \sigma_{\max}^2(\mathbf{B}_t) \leq \frac{1+\|\mathbf{\Delta}_t\|_2}{\alpha}$ by Weyl's inequality. Next, we have

$$\begin{aligned}
\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) &= \sigma_{\min}\Bigg(\frac{1}{n}\sum_{i=1}^n (\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t)\mathbf{w}_t\mathbf{w}_t^\top(\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t) \\
&\quad + \alpha(\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t)\mathbf{w}_t\mathbf{w}_{*,t,i}\mathbf{B}_*^\top\mathbf{B}_t + \alpha\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top(\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t) \\
&\quad + \alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}\mathbf{B}_*^\top\mathbf{B}_t\Bigg) \\
&\geq \sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}\mathbf{B}_*^\top\mathbf{B}_t\right) - 2\alpha\left\|\mathbf{\Delta}_t\mathbf{w}_t\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{B}_*^\top\mathbf{B}_t\right\|_2 \qquad (15) \\
&\geq \alpha E_0\mu_*^2 - 2\alpha\left\|\mathbf{\Delta}_t\mathbf{w}_t\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{B}_*^\top\mathbf{B}_t\right\|_2 \qquad (16) \\
&\geq \alpha E_0\mu_*^2 - 2(1+\|\mathbf{\Delta}_t\|_2)\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* \qquad (17)
\end{aligned}$$

where (15) follows by Weyl's inequality and the fact that $\mathbf{B}_t^\top\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_t \succeq \mathbf{0}$, (16) follows by Lemma 1, and (17) follows by the Cauchy-Schwarz inequality.

Now we upper bound the maximum singular value of $\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top$. We have

$$\begin{aligned}
\sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) &\leq \left\|(\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t)\mathbf{w}_t\mathbf{w}_t^\top(\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t)\right\|_2 \\
&\quad + 2\alpha\left\|(\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t)\mathbf{w}_t\right\|_2\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2 + \alpha^2\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2^2 \\
&\leq \|\mathbf{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^2 + 2\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\sqrt{\alpha(1+\|\mathbf{\Delta}_t\|_2)}\eta_* + \alpha(1+\|\mathbf{\Delta}_t\|_2)L_*^2 \\
&\leq (\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + \sqrt{\alpha(1+\|\mathbf{\Delta}_t\|_2)}L_*)^2. \qquad (18)
\end{aligned}$$

□

Lemma 3. *Suppose the sequence* $\{\mathbf{w}_s\}_{s=0}^{t+1}$ *satisfies:*

$$\begin{aligned}
\|\mathbf{w}_0\|_2 &= 0, \\
\|\mathbf{w}_{s+1}\|_2 &\leq (1+\xi_{1,s})\|\mathbf{w}_s\|_2 + \xi_{2,s}
\end{aligned} \tag{19}$$

where $\xi_{1,s} \geq 0$, $\xi_{2,s} \geq 0$ *for all* $s \in [t]$ *and* $\sum_{s=1}^t \xi_{1,s} \leq 1$. *Then:*

$$\|\mathbf{w}_{t+1}\|_2 \leq \sum_{s=1}^t \xi_{2,s}\left(1 + 2\sum_{r=s}^t \xi_{1,r}\right) \tag{20}$$

Proof. We have

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|_2 &\le (1+\xi_{1,t})\|\mathbf{w}_t\|_2 + \xi_{2,t} \\
&\le (1+\xi_{1,t})^2\|\mathbf{w}_{t-1}\|_2 + \xi_{2,t}(1-\xi_{1,t}) + \xi_{2,t} \\
&\vdots \\
&\le \|\mathbf{w}_0\|_2 \prod_{s=1}^{t}(1+\xi_{1,s}) + \sum_{s=1}^{t}\xi_{2,s}\prod_{r=s}^{t-1}(1+\xi_{1,r}) \\
&= \sum_{s=1}^{t}\xi_{2,s}\prod_{r=s}^{t-1}(1+\xi_{1,r}) \qquad (21) \\
&\le \sum_{s=1}^{t}\xi_{2,s}\left(1+\frac{1}{t-s}\sum_{r=s}^{t}\xi_{1,r}\right)^{t-s} \qquad (22)
\end{aligned}$$

where (21) is due to $\|\mathbf{w}_0\|_2 = 0$ and (22) follows from the AM-GM inequality. Next, note that $\left(1+\frac{1}{t-s}\sum_{r=s}^{t}\xi_{1,r}\right)^{t-s}$ is of the form $\left(1+\frac{a}{x}\right)^x$, where $x = t-s$ and $a = \sum_{r=s}^{t}\xi_{1,r}$. Since $\left(1+\frac{a}{x}\right)^x$ is an increasing function of x, we can upper bound it by its limit as $x \to \infty$, which is e^a. Thus we have

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|_2 &\le \sum_{s=1}^{t}\xi_{2,s}\exp\left(\sum_{r=s}^{t}\xi_{1,r}\right) \\
&\le \sum_{s=1}^{t}\xi_{2,s}\left(1+2\sum_{r=s}^{t}\xi_{1,r}\right) \qquad (23)
\end{aligned}$$

where (23) follows from the numerical inequality $\exp(x) \le 1+2x$ for all $x \in [0,1]$. □

Lemma 4. *Suppose that* $\mathbf{B}_{t+1} = \mathbf{B}_t - \beta\mathbf{G}_t$ *and*

$$\mathbf{G}_t = -\bar{\mathbf{\Delta}}_t\mathbf{S}_t\mathbf{B}_t - \chi\mathbf{S}_t\mathbf{B}_t\mathbf{\Delta}_t + \mathbf{N}_t \qquad (24)$$

for $\mathbf{N}_t \in \mathbb{R}^{d\times k}$ *and a positive semi-definite matrix* $\mathbf{S}_t \in \mathbb{R}^{k\times k}$. *Then*

$$\|\mathbf{\Delta}_{t+1}\|_2 \le \|\mathbf{\Delta}_t\|_2\left(1-(1+\chi)\beta\alpha\sigma_{\min}(\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t)\right) + 2\beta\alpha\|\mathbf{B}_t^\top\mathbf{N}_t\|_2 + \beta^2\alpha\|\mathbf{G}_t\|_2^2 \qquad (25)$$

Proof. By expanding $\mathbf{\Delta}_{t+1}$, $\mathbf{B}_{t+1}$, and $\mathbf{G}_t$, we obtain

$$\begin{aligned}
\mathbf{\Delta}_{t+1} &= \mathbf{I} - \alpha\mathbf{B}_{t+1}^\top\mathbf{B}_{t+1} \\
&= \mathbf{I} - \alpha\mathbf{B}_t^\top\mathbf{B}_t + \beta\alpha\mathbf{B}^\top\mathbf{G}_t + \beta\alpha\mathbf{G}_t^\top\mathbf{B}_t - \beta^2\alpha\mathbf{G}_t^\top\mathbf{G}_t \\
&= \mathbf{\Delta}_t - \beta\alpha\mathbf{\Delta}_t\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t - \chi\beta\alpha\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t\mathbf{\Delta}_t + \beta\alpha\mathbf{B}_t^\top\mathbf{N}_t \\
&\quad - \chi\beta\alpha\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t\mathbf{\Delta}_t - \beta\alpha\mathbf{\Delta}_t\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t + \beta\alpha\mathbf{N}_t^\top\mathbf{B}_t - \beta^2\alpha\mathbf{G}_t^\top\mathbf{G}_t \qquad (26) \\
&= \tfrac{1}{2}\mathbf{\Delta}_t\left(\mathbf{I}_k - (1+\chi)\beta\alpha\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t\right) \\
&\quad + \tfrac{1}{2}\left(\mathbf{I}_k - (1+\chi)\beta\alpha\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t\right)\mathbf{\Delta}_t + \beta\alpha(\mathbf{B}_t^\top\mathbf{N}_t + \mathbf{N}_t^\top\mathbf{B}_t) - \beta^2\alpha\mathbf{G}_t^\top\mathbf{G}_t \qquad (27)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathbf{\Delta}_{t+1}\|_2 &\le \|\mathbf{\Delta}_t\|_2\left\|\mathbf{I}_k - (1+\chi)\beta\alpha\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t\right\|_2 + 2\beta\alpha\|\mathbf{B}_t^\top\mathbf{N}_t\|_2 + \beta^2\alpha\|\mathbf{G}_t\|_2^2 \\
&\le \|\mathbf{\Delta}_t\|_2\left(1-(1+\chi)\beta\alpha\sigma_{\min}(\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t)\right) + 2\beta\alpha\|\mathbf{B}_t^\top\mathbf{N}_t\|_2 + \beta^2\alpha\|\mathbf{G}_t\|_2^2 \qquad (28)
\end{aligned}$$

where the last inequality follows by the triangle and Weyl inequalities. □

C. ANIL Infinite Samples

We start by considering the infinite sample case, wherein $m_{in} = m_{out} = \infty$. Let $E_0 := 0.9 - \text{dist}_0^2$. We restate Theorem 1 here with full constants.

Theorem 5 (ANIL Infinite Samples). *Let $m_{in} = m_{out} = \infty$ and define $E_0 := 0.9 - \text{dist}_0^2$. Suppose Assumption 1 holds and $\text{dist}_0 \leq \sqrt{0.9}$. Let $\alpha < \frac{1}{L_*}$, $\alpha \mathbf{B}_0^\top \mathbf{B}_0 = \mathbf{I}_k$ and $\mathbf{w}_t = \mathbf{0}$. Then FO-ANIL with $\beta \leq \frac{\alpha E_0^3 \mu_*}{180 \kappa_*^4} \min(1, \frac{\mu_*^2}{\eta_*^2})$ and Exact ANIL with $\beta \leq \frac{\alpha E_0^2}{40 \kappa_*^4}$ both satisfy that after T iterations,*

$$\text{dist}(\mathbf{B}_T, \mathbf{B}_*) \leq \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^{T-1} \tag{29}$$

Proof. The proof uses an inductive argument with the following six inductive hypotheses:

1. $A_1(t) := \{\|\mathbf{w}_t\|_2 \leq \frac{\sqrt{\alpha} E_0}{10} \min(1, \frac{\mu_*^2}{\eta_*^2}) \eta_*\}$
2. $A_2(t) := \{\|\mathbf{\Delta}_t\|_2 \leq (1 - 0.5\beta\alpha E_0 \mu_*^2) \|\mathbf{\Delta}_{t-1}\|_2 + \frac{5}{4}\alpha^2 \beta^2 L_*^4 \, \text{dist}_{t-1}^2\}$,
3. $A_3(t) := \{\|\mathbf{\Delta}_t\|_2 \leq \frac{1}{10}\}$,
4. $A_4(t) := \{0.9\alpha E_0 \mu_*^2 \mathbf{I}_k \preceq \frac{1}{n} \sum_{i=1}^n \mathbf{w}_{t,i} \mathbf{w}_{t,i}^\top \preceq 1.2\alpha L_*^2 \mathbf{I}_k\}$,
5. $A_5(t) := \{\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2 \leq \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right) \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t-1}\|_2\}$,
6. $A_6(t) := \{\text{dist}_t \leq \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t\}$.

These conditions hold for iteration $t = 0$ due to the choice of initialization $(\mathbf{B}_0, \mathbf{w}_0)$ satisfying $\mathbf{I}_k - \alpha \mathbf{B}_0^\top \mathbf{B}_0 = \mathbf{0}$ and $\mathbf{w}_0 = \mathbf{0}$. We will show that if they hold for all iterations up to and including iteration t for an arbitrary t, then they hold at iteration $t + 1$.

1. $\bigcap_{s=0}^t \{A_2(s) \cap A_6(s)\} \implies A_1(t+1)$. This is Lemma 5 for FO-ANIL and Lemma 9 for Exact ANIL.
2. $A_1(t) \cap A_3(t) \cap A_5(t) \implies A_2(t+1)$. This is Lemma 6 for FO-ANIL and Lemma 10 for Exact ANIL.
3. $A_2(t+1) \cap A_3(t) \implies A_3(t+1)$. This is Corollary 1 for FO-ANIL and Corollary 2 for Exact ANIL.
4. $A_1(t+1) \cap A_3(t+1) \cap A_6(t+1) \implies A_4(t+1)$. This is Lemma 7 for FO-ANIL and Lemma 12 for Exact ANIL.
5. FO-ANIL: $A_4(t) \implies A_5(t+1)$. This is Lemma 8.

   Exact ANIL: $A_1(t) \cap A_3(t) \cap A_4(t) \implies A_5(t+1)$. This is Lemma 11. The slight discrepancy here in the implications is due to the extra terms in the outer loop representation update for Exact ANIL.
6. $A_3(t+1) \cap \left\{\bigcap_{s=0}^{t+1} A_5(s)\right\} \implies A_6(t+1)$. Recall $\text{dist}_{t+1} = \|\mathbf{B}_{*,\perp}^\top \hat{\mathbf{B}}_{t+1}\|_2$ where $\hat{\mathbf{B}}_{t+1}$ is the orthogonal matrix resulting from the QR factorization of $\mathbf{B}_{t+1}$, i.e. $\mathbf{B}_{t+1} = \hat{\mathbf{B}}_{t+1} \mathbf{R}_{t+1}$ for an upper triangular matrix $\mathbf{R}_{t+1}$. By $A_3(t+1)$ and $\cap_{s=0}^{t+1} A_5(s)$ we have

$$\begin{aligned}
\frac{\sqrt{1 - \|\mathbf{\Delta}_{t+1}\|_2}}{\sqrt{\alpha}} \text{dist}_{t+1} &= \frac{\sqrt{1 - \|\mathbf{\Delta}_{t+1}\|_2}}{\sqrt{\alpha}} \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 \\
&\leq \sigma_{\min}(\mathbf{B}_{t+1}) \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 \\
&\leq \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 \\
&\leq \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_0\|_2 \\
&\leq \tfrac{1}{\sqrt{\alpha}} \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_0\|_2 \\
&= \tfrac{1}{\sqrt{\alpha}} \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t \text{dist}_0 \,.
\end{aligned}$$

Dividing both sides by $\frac{\sqrt{1-\|\mathbf{\Delta}_{t+1}\|_2}}{\sqrt{\alpha}}$ and using the facts that $\mathrm{dist}_0 \leq \frac{3}{\sqrt{10}}$ and $\|\mathbf{\Delta}_{t+1}\|_2 \leq \frac{1}{10}$ yields

$$\begin{aligned}\mathrm{dist}_{t+1} &\leq \tfrac{1}{\sqrt{1-\|\mathbf{\Delta}_{t+1}\|_2}} \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t \mathrm{dist}_0 \\ &\leq \tfrac{\sqrt{10}}{3} \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t \mathrm{dist}_0 \\ &\leq \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t, \end{aligned} \tag{30}$$

as desired.

□

C.1. FO-ANIL

First note that the inner loop updates for FO-ANIL can be written as:

$$\begin{aligned}\mathbf{w}_{t,i} &= \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} \mathcal{L}_{t,i}(\mathbf{B}_t, \mathbf{w}_t) \\ &= (\mathbf{I}_k - \alpha \mathbf{B}_t^\top \mathbf{B}_t)\mathbf{w}_t + \alpha \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i}, \end{aligned} \tag{31}$$

while the outer loop updates for the head and representation are:

$$\begin{aligned}\mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{\beta}{n} \sum_{i=1}^n \nabla_{\mathbf{w}} \mathcal{L}_{t,i}(\mathbf{B}_t, \mathbf{w}_{t,i}) \\ &= \mathbf{w}_t - \frac{\beta}{n} \sum_{i=1}^n \mathbf{B}_t^\top \mathbf{B}_t \mathbf{w}_{t,i} + \frac{\beta}{n} \sum_{i=1}^n \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} \end{aligned} \tag{32}$$

$$\begin{aligned}\mathbf{B}_{t+1} &= \mathbf{B}_t - \frac{\beta}{n} \sum_{i=1}^n \nabla_{\mathbf{B}} \mathcal{L}_{t,i}(\mathbf{B}_t, \mathbf{w}_{t,i}) \\ &= \mathbf{B}_t - \frac{\beta}{n} \sum_{i=1}^n \mathbf{B}_t \mathbf{w}_{t,i} \mathbf{w}_{t,i}^\top + \frac{\beta}{n} \sum_{i=1}^n \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_{t,i}^\top \end{aligned} \tag{33}$$

$$= \mathbf{B}_t - \beta(\mathbf{I}_d - \alpha \mathbf{B}_t \mathbf{B}_t^\top) \frac{1}{n} \sum_{i=1}^n (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_{t,i}^\top \tag{34}$$

Lemma 5 (FO-ANIL $A_1(t+1)$)**.** *Suppose we are in the setting of Theorem 1, and that the events $A_2(s)$ and $A_6(s)$ hold for all $s \in [t]$. Then*

$$\|\mathbf{w}_{t+1}\|_2 \leq \tfrac{1}{10}\sqrt{\alpha} E_0 \min(1, \tfrac{\mu_*^2}{\eta_*^2}) \eta_*. \tag{35}$$

Proof. For all $s = 1, \ldots, t$, the outer loop updates for FO-ANIL can be written as:

$$\mathbf{w}_{s+1} = \mathbf{w}_s - \frac{\beta}{n} \sum_{i=1}^n \nabla_{\mathbf{w}} \mathcal{L}_{s,i}(\mathbf{B}_s, \mathbf{w}_{s,i}) = \mathbf{w}_s - \frac{\beta}{n} \sum_{i=1}^n \mathbf{B}_s^\top \mathbf{B}_s \mathbf{w}_{s,i} + \frac{\beta}{n} \sum_{i=1}^n \mathbf{B}_s^\top \mathbf{B}_* \mathbf{w}_{*,s,i} \tag{36}$$

Substituting the definition of $\mathbf{w}_{s,i}$, we have

$$\begin{aligned}\mathbf{w}_{s+1} &= \mathbf{w}_s - \beta \mathbf{B}_s^\top \mathbf{B}_s (\mathbf{I} - \alpha \mathbf{B}_s^\top \mathbf{B}_s)\mathbf{w}_s - \frac{\alpha\beta}{n} \sum_{i=1}^n \mathbf{B}_s^\top \mathbf{B}_s \mathbf{B}_s^\top \mathbf{B}_* \mathbf{w}_{*,s,i} + \frac{\beta}{n} \sum_{i=1}^n \mathbf{B}_s^\top \mathbf{B}_* \mathbf{w}_{*,s,i} \\ &= (\mathbf{I}_k - \beta(\mathbf{I} - \alpha \mathbf{B}_s^\top \mathbf{B}_s)\mathbf{B}_s^\top \mathbf{B}_s)\mathbf{w}_s + \beta(\mathbf{I} - \alpha \mathbf{B}_s^\top \mathbf{B}_s)\mathbf{B}_s^\top \mathbf{B}_* \frac{1}{n} \sum_{i=1}^n \mathbf{w}_{*,s,i} \end{aligned} \tag{37}$$

Note that $\bigcup_{s=0}^t A_3(s)$ implies $\sigma_{\max}(\mathbf{B}_s^\top \mathbf{B}_s) \leq \frac{1+\|\mathbf{\Delta}_s\|_2}{\alpha} < \frac{1.1}{\alpha}$ for all $s \in \{0, \ldots, t+1\}$. Let $c := 1.1$. Using $\sigma_{\max}(\mathbf{B}_s^\top \mathbf{B}_s) \leq \frac{c}{\alpha}$ with (37), we obtain

$$\|\mathbf{w}_{s+1}\|_2 \leq (1 + \tfrac{c\beta}{\alpha}\|\mathbf{\Delta}_s\|_2)\|\mathbf{w}_s\|_2 + \tfrac{c\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2 \eta_* \tag{38}$$

for all $s \in \{0, \dots, t\}$. Therefore, by applying Lemma 3 with $\xi_{1,s} = \frac{c\beta}{\alpha}\|\mathbf{\Delta}_s\|_2$ and $\xi_{2,s} = \frac{c\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2\eta_*$, we have

$$\|\mathbf{w}_{t+1}\|_2 \leq \sum_{s=1}^{t} \tfrac{c\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2\eta_* \left(1 + 2c\sum_{r=s}^{t}\tfrac{\beta}{\alpha}\|\mathbf{\Delta}_r\|_2\right) \tag{39}$$

Next, let $\rho \coloneqq 1 - 0.5\beta\alpha E_0\mu_*^2$. By $\bigcup_{s=0}^{t} A_2(s)$, we have for any $s \in [t]$

$$\begin{aligned}
\|\mathbf{\Delta}_s\|_2 &\leq \rho\|\mathbf{\Delta}_{s-1}\|_2 + \tfrac{5}{4}\alpha^2\beta^2L_*^4\operatorname{dist}_{s-1}^2 \\
&\leq \rho^2\|\mathbf{\Delta}_{s-2}\|_2 + \tfrac{5}{4}\alpha^2\beta^2\rho L_*^4\operatorname{dist}_{s-2}^2 + \tfrac{5}{4}\alpha^2\beta^2L_*^4\operatorname{dist}_{s-1}^2 \\
&\vdots \\
&\leq \rho^s\|\mathbf{\Delta}_0\|_2 + \tfrac{5}{4}\alpha^2\beta^2L_*^4\sum_{r=0}^{s-1}\rho^{s-1-r}\operatorname{dist}_r^2 \\
&= \tfrac{5}{4}\alpha^2\beta^2L_*^4\sum_{r=0}^{s-1}\rho^{s-1-r}\operatorname{dist}_r^2
\end{aligned} \tag{40}$$

since $\|\mathbf{\Delta}_0\|_2 = 0$ by choice of initialization. Next, we have that $\operatorname{dist}_s \leq \rho^s$ for all $s \in \{0, ..., t\}$ by $\bigcup_{s=0}^{t} A_5(s)$. Thus, for any $s \in \{0, ..., t\}$, we can further bound $\|\mathbf{\Delta}_s\|_2$ as

$$\begin{aligned}
\|\mathbf{\Delta}_s\|_2 &\leq \tfrac{5}{4}\alpha^2\beta^2L_*^4\sum_{r=0}^{s-1}\rho^{s-1-r}\rho^{2r} \\
&= \tfrac{5}{4}\alpha^2\beta^2L_*^4\rho^{s-1}\sum_{r=0}^{s-1}\rho^r \\
&\leq \rho^{s-1}\tfrac{5\alpha^2\beta^2L_*^4}{4(1-\rho)} \\
&\leq \rho^{s-1}\tfrac{5\beta\alpha L_*^4}{2E_0\mu_*^2},
\end{aligned} \tag{41}$$

which means that

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|_2 &\leq \sum_{s=1}^{t}\tfrac{c\beta}{\sqrt{\alpha}}\rho^{s-1}\tfrac{5\beta\alpha L_*^4}{2E_0\mu_*^2}\eta_*\left(1 + 2c\sum_{r=s}^{t}\tfrac{\beta}{\alpha}\rho^{r-1}\tfrac{5\beta\alpha L_*^4}{2E_0\mu_*^2}\right) \\
&\leq 2.5c\beta^2\sqrt{\alpha}\frac{L_*^4\eta_*}{E_0\mu_*^2}\sum_{s=1}^{t}\rho^{s-1}\left(1 + 5c\beta^2\frac{L_*^4}{E_0\mu_*^2}\sum_{r=s}^{t}\rho^{r-1}\right) \\
&\leq 2.5c\beta^2\sqrt{\alpha}\frac{L_*^4\eta_*}{E_0\mu_*^2}\sum_{s=1}^{t}\rho^{s-1}\left(1 + 6\beta^2\frac{L_*^4\rho^s}{E_0\mu_*^2(1-\rho)}\right) \\
&\leq 3\beta^2\sqrt{\alpha}\frac{L_*^4\eta_*}{E_0\mu_*^2}\sum_{s=1}^{t}\rho^{s-1}\left(1 + 12\frac{\beta L_*^4}{\alpha E_0^2\mu_*^4}\right) \\
&\leq 6\beta^2\sqrt{\alpha}\frac{L_*^4\eta_*}{E_0\mu_*^2}\sum_{s=1}^{t}1.5\rho^{s-1} \qquad (42) \\
&\leq 18\frac{\beta\kappa_*^4\eta_*}{\sqrt{\alpha}E_0^2} \\
&\leq \tfrac{1}{10}\sqrt{\alpha}E_0\min(1, \tfrac{\mu_*^2}{\eta_*^2})\eta_* \qquad (43)
\end{aligned}$$

where (42) and (43) follow since $\beta \leq \frac{\alpha E_0^3}{180\kappa_*^4}\min(1, \frac{\mu_*^2}{\eta_*^2})\eta_*$. □

Remark 1. *As referred to in Section 5, it is not necessary to start with $\|\mathbf{w}_0\|_2$ and $\|\mathbf{\Delta}_0\|_2$ strictly equal to zero. Precisely, it can be shown that the above lemma still holds with $\|\mathbf{w}_0\|_2 \leq c\sqrt{\alpha}E_0\min(1, \frac{\mu_*^2}{\eta_*^2})\eta_*$ and $\|\mathbf{\Delta}_0\|_2 \leq c\frac{\beta\alpha L_*^4}{E_0\mu_*^2}$ for a sufficiently*

small absolute constant c. Inductive hypothesis $A_6(t+1)$ would also continue to hold under this initialization, with a slightly different constant. These are the only times we use $\|\mathbf{w}_0\|_2 = \|\mathbf{\Delta}_0\|_2 = 0$, so the rest of the proof would hold. Similar statements can be made regarding the rest of the algorithms.

Lemma 6 (FO-ANIL $A_2(t+1)$)**.** *Suppose we are in the setting of Theorem 1 and that $A_1(t), A_3(t), A_6(t)$ hold. Then $A_2(t+1)$ holds, i.e.*

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right) \|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2 L_*^4 \operatorname{dist}_t^2 . \tag{44}$$

Proof. Let $\mathbf{G}_t$ be the outer loop gradient for the representation, i.e. $\mathbf{G}_t = \frac{1}{\beta}(\mathbf{B}_t - \mathbf{B}_{t+1})$. We aim to apply Lemma 4, we write $\mathbf{G}_t$ as $-\bar{\mathbf{\Delta}}_t \mathbf{S}_t \mathbf{B}_t + \mathbf{N}_t$, for some positive definite matrix $\mathbf{S}_t$ and another matrix $\mathbf{N}_t$. We have

$$\begin{aligned}
\mathbf{G}_t &= \frac{1}{n}\sum_{i=1}^n (\mathbf{B}_t \mathbf{w}_{t,i} - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_{t,i}^\top \\
&= \frac{1}{n}\sum_{i=1}^n (\mathbf{B}_t \mathbf{\Delta}_t \mathbf{w}_t + \alpha \mathbf{B}_t \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_{t,i}^\top \\
&= \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Delta}}_t (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_{t,i}^\top \\
&= -\alpha \bar{\mathbf{\Delta}}_t \mathbf{B}_* \left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i} \mathbf{w}_{*,t,i}^\top \right) \mathbf{B}_*^\top \mathbf{B}_t + \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Delta}}_t \mathbf{B}_t \mathbf{w}_t \mathbf{w}_{t,i}^\top - \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Delta}}_t \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_t^\top \mathbf{\Delta}_t \qquad (45) \\
&= -\bar{\mathbf{\Delta}}_t \mathbf{S}_t \mathbf{B}_t + \mathbf{N}_t
\end{aligned}$$

where (45) follows by expanding $\mathbf{w}_{t,i}$, and $\mathbf{S}_t = \alpha \mathbf{B}_* \left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i} \mathbf{w}_{*,t,i}^\top \right) \mathbf{B}_*^\top$ and $\mathbf{N}_t = \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Delta}}_t \mathbf{B}_t \mathbf{w}_t \mathbf{w}_{t,i}^\top - \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Delta}}_t \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_t^\top \mathbf{\Delta}_t$. Since $\sigma_{\min}(\mathbf{B}_t^\top \mathbf{S}_t \mathbf{B}_t) \geq E_0 \mu_*^2$ (by Lemma 1), we have by Lemma 4

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq (1 - \beta\alpha E_0 \mu_*^2) \|\mathbf{\Delta}_t\|_2 + 2\beta\alpha \|\mathbf{B}_t^\top \mathbf{N}_t\|_2 + \beta^2 \alpha \|\mathbf{G}_t\|_2^2 \tag{46}$$

To bound $\|\mathbf{B}_t^\top \mathbf{N}_t\|_2$, we have

$$\begin{aligned}
\|\mathbf{B}_t^\top \mathbf{N}_t\|_2 &= \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Delta}_t \mathbf{B}_t^\top \mathbf{B}_t \mathbf{w}_t \mathbf{w}_{t,i}^\top - \frac{1}{n}\sum_{i=1}^n \mathbf{\Delta}_t \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_t^\top \mathbf{\Delta}_t \right\|_2 \\
&\leq \left\| \mathbf{\Delta}_t \mathbf{B}_t^\top \mathbf{B}_t \mathbf{w}_t \mathbf{w}_t^\top \mathbf{\Delta}_t \right\|_2 + \alpha \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Delta}_t \mathbf{B}_t^\top \mathbf{B}_t \mathbf{w}_t \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{B}_t \right\|_2 \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Delta}_t \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_t^\top \mathbf{\Delta}_t \right\|_2 \\
&\leq \tfrac{c}{\alpha} \|\mathbf{\Delta}_t\|_2^2 \|\mathbf{w}_t\|_2^2 + \tfrac{c}{\sqrt{\alpha}} (\|\mathbf{\Delta}_t\|_2 + \|\mathbf{\Delta}_t\|_2^2) \|\mathbf{w}_t\|_2 \eta_* \\
&\leq \frac{c E_0 \mu_*^2}{1000 \kappa_*^4} \|\mathbf{\Delta}_t\|_2 + \tfrac{11}{100} c E_0 \mu_*^2 \|\mathbf{\Delta}_t\|_2 \\
&\leq \tfrac{1}{8} \mu_*^2 \|\mathbf{\Delta}_t\|_2
\end{aligned} \tag{47}$$

where we have used $A_1(t)$ and $A_3(t)$ and the fact that $\min(1, \frac{\mu_*^2}{\eta_*^2})\eta_*^2 \leq \mu_*^2$. To bound $\|\mathbf{G}_t\|_2^2$ we have

$$\begin{aligned}
\|\mathbf{G}_t\|_2 &\leq \|\bar{\mathbf{\Delta}}_t \mathbf{S}_t \mathbf{B}_t\|_2 + \|\mathbf{N}_t\|_2 \\
&\leq c\sqrt{\alpha} L_*^2 (\|\mathbf{\Delta}_t\|_2 + \operatorname{dist}_t) + \left\| \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Delta}}_t \mathbf{B}_t \mathbf{w}_t \mathbf{w}_{t,i}^\top \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Delta}}_t \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_t^\top \mathbf{\Delta}_t \right\|_2 \\
&\leq c\sqrt{\alpha} L_*^2 (\|\mathbf{\Delta}_t\|_2 + \operatorname{dist}_t) + \tfrac{c}{\sqrt{\alpha}} \|\mathbf{\Delta}_t\|_2^2 \|\mathbf{w}_t\|_2^2 + 2c \|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2 \eta_* + \|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\| \eta_* \operatorname{dist}_t \\
&\leq c\sqrt{\alpha} L_*^2 \|\mathbf{\Delta}_t\|_2 + c\sqrt{\alpha} L_*^2 \operatorname{dist}_t + \tfrac{c}{1000}\sqrt{\alpha}\mu_*^2 \|\mathbf{\Delta}_t\|_2 + \tfrac{3c}{10}\sqrt{\alpha} L_* \eta_* \|\mathbf{\Delta}_t\|_2 \\
&\leq 1.5\sqrt{\alpha} L_*^2 \|\mathbf{\Delta}_t\|_2 + 1.1\sqrt{\alpha} L_*^2 \operatorname{dist}_t
\end{aligned} \tag{48}$$

where (48) follows since $\eta_* \leq L_*$. Therefore

$$\begin{aligned}\|\mathbf{G}_t\|_2^2 &\leq \alpha L_*^4(2.5\|\mathbf{\Delta}_t\|_2^2 + 3.3\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\operatorname{dist}_t^2)\\ &\leq 4\alpha L_*^4\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\alpha L_*^4\operatorname{dist}_t^2\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{\Delta}_{t+1}\|_2 &\leq \left(1 - \beta\alpha E_0\mu_*^2 + 0.25\beta\alpha E_0\mu_*^2 + 4\beta^2\alpha^2 L_*^4\right)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2 L_*^4\operatorname{dist}_t^2\\ &\leq \left(1 - 0.5\beta\alpha E_0\mu_*^2\right)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2 L_*^4\operatorname{dist}_t^2 \qquad (49)\end{aligned}$$

where in (49) we have used $\beta \leq \frac{\alpha E_0^3}{180\kappa_*^4}$, $\alpha \leq \frac{1}{L_*}$, and $E_0 \leq 1$. □

Corollary 1 (FO-ANIL $A_3(t+1)$)**.** *Suppose we are in the setting of Theorem 1. If inductive hypotheses $A_2(t+1)$ and $A_3(t)$ hold, then $A_3(t+1)$ holds, i.e.*

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq \tfrac{1}{10} \qquad (50)$$

Proof. Note that according to equation (49), we have

$$\begin{aligned}\|\mathbf{\Delta}_{t+1}\|_2 &\leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2 L_*^4\\ &\leq (1 - 0.5\beta\alpha E_0\mu_*^2)\tfrac{1}{10} + \tfrac{5}{720}E_0\beta\alpha\mu_*^2 \qquad (51)\\ &\leq \tfrac{1}{10}\end{aligned}$$

where equation (51) is satisfied by our choice of $\beta \leq \frac{\alpha E_0^3}{180\kappa_*^4}$ and $\alpha \leq \frac{1}{L_*}$ and inductive hypothesis $A_3(t)$. □

Lemma 7 (FO-ANIL $A_4(t+1)$)**.** *Suppose the conditions of Theorem 1 are satisfied and inductive hypotheses $A_1(t)$, $A_3(t)$ and $A_6(t)$ hold. Then $A_4(t+1)$ holds, i.e.*

$$\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t+1,i}\mathbf{w}_{t+1,i}^\top\right) \geq 0.9\alpha E_0\mu_*^2$$
$$\text{and } \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t+1,i}\mathbf{w}_{t+1,i}^\top\right) \leq 1.2\alpha L_*^2$$

Proof. By Lemma 2 and inductive hypotheses $A_1(t)$, $A_3(t)$ and $A_6(t)$, we have

$$\begin{aligned}\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) &\geq \alpha E_0\mu_*^2 - 0.022\alpha E_0\mu_*^2 \geq 0.9\alpha E_0\mu_*^2\\ \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) &\leq (\tfrac{1}{100}\sqrt{\alpha}E_0\kappa_*^{-1} + \sqrt{1.1\alpha}L_*)^2 \leq 1.2\alpha L_*^2 \qquad (52)\end{aligned}$$

where we have used the fact that $\min(1, \frac{\mu_*^2}{\eta_*^2})\eta_*^2 \leq \mu_*^2$ to lower bound the minimum singular value. □

Lemma 8 (FO-ANIL $A_5(t+1)$)**.** *Suppose the conditions of Theorem 1 are satisfied. If inductive hypotheses $A_4(t)$ holds, then $A_5(t+1)$ holds, i.e.*

$$\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \leq \left(1 - 0.5\beta\alpha E_0\mu_*^2\right)\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2. \qquad (53)$$

Proof. Note from $A_4(t+1)$ that $\left(\sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)\right)^{-1} \geq \frac{1}{\alpha L_*^2} \geq \frac{1}{L_*}$. Thus, since $\beta \leq \frac{\alpha E_0^3}{180\kappa_*^4} \leq \frac{1}{L_*} \leq \left(\sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)\right)^{-1}$, we have by Weyl's inequality that

$$\begin{aligned}\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 &\leq \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2\left\|\mathbf{I}_k - \frac{\beta}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right\|_2 \leq \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2\left(1 - \beta\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)\right)\\ &\leq \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2\left(1 - 0.5\beta\alpha E_0\mu_*^2\right).\end{aligned}$$

□

C.2. Exact ANIL

To study Exact ANIL, first note that the inner loop updates are identical to those for FO-MAML. However, the outer loop updates are different. Here, we have

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\beta}{n}\sum_{i=1}^{n} \nabla_{\mathbf{w}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t)$$
$$\mathbf{B}_{t+1} = \mathbf{B}_t - \frac{\beta}{n}\sum_{i=1}^{n} \nabla_{\mathbf{B}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t)$$

where for all t, i:

$$F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) \coloneqq \mathcal{L}_{t,i}(\mathbf{B}_t, \mathbf{w}_t - \alpha\nabla_{\mathbf{w}}\mathcal{L}_{t,i}(\mathbf{B}_t, \mathbf{w}_t)) \coloneqq \tfrac{1}{2}\|\mathbf{v}_{t,i}\|_2^2 \tag{54}$$

and

$$\mathbf{v}_{t,i} \coloneqq \mathbf{B}_t\boldsymbol{\Delta}_t\mathbf{w}_t + \alpha\mathbf{B}_t\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i} - \mathbf{B}_*\mathbf{w}_{*,t,i} = \bar{\boldsymbol{\Delta}}_t(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i}) \tag{55}$$

Therefore,

$$\nabla_{\mathbf{w}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) = \mathbf{B}_t^\top\bar{\boldsymbol{\Delta}}_t\mathbf{v}_{t,i} \tag{56}$$
$$\nabla_{\mathbf{B}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) = \mathbf{v}_{t,i}\mathbf{w}_t^\top\boldsymbol{\Delta}_t + \alpha\mathbf{v}_{t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t - \alpha\mathbf{B}_t\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t - \alpha\mathbf{B}_t\mathbf{B}_t^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top + \alpha\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{v}_{t,i}^\top\mathbf{B}_t \tag{57}$$

One can observe that for $\mathbf{w}$, the outer loop gradient is the same as in the FO-ANIL case but with an extra $\alpha\mathbf{B}_t^\top\bar{\boldsymbol{\Delta}}_t$ factor. Meanwhile, the first two terms in the outer loop gradient for $\mathbf{B}$ compose the outer loop gradient in the FO-ANIL case, while the other three terms are new. We deal with these differences in the following lemmas.

Lemma 9 (Exact ANIL $A_1(t+1)$)**.** *Suppose the conditions of Theorem 1 are satisfied and $A_2(s)$ and $A_6(s)$ hold for all $s \in [t]$, then $A_1(t+1)$ holds, i.e.*

$$\|\mathbf{w}_{t+1}\|_2 \le \tfrac{1}{10}\sqrt{\alpha}E_0\min(1, \tfrac{\mu_*^2}{\eta_*^2})\eta_*. \tag{58}$$

Proof. Similarly to the FO-ANIL case, we can show that for any $s \in [t]$,

$$\mathbf{w}_{s+1} = \mathbf{w}_s - \frac{\beta}{n}\sum_{i=1}^{n}\nabla_{\mathbf{w}}F_{s,i}(\mathbf{B}_s, \mathbf{w}_s) = (\mathbf{I}_k - \beta\boldsymbol{\Delta}_s\mathbf{B}_s^\top\mathbf{B}_s\boldsymbol{\Delta}_s)\mathbf{w}_s + \beta\boldsymbol{\Delta}_s^2\mathbf{B}_s^\top\mathbf{B}_*\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{*,s,i} \tag{59}$$

Note that $\bigcup_{s=0}^{t} A_3(s)$ implies $\sigma_{\max}(\mathbf{B}_s^\top\mathbf{B}_s) \le \frac{1+\|\boldsymbol{\Delta}_s\|_2}{\alpha} < \frac{1.1}{\alpha}$ for all $s \in \{0, \dots, t+1\}$. Let $c \coloneqq 1.1$.

Unlike in the first-order case, the coefficient of $\mathbf{w}_s$ in (59) is the identity matrix minus a positive semi-definite matrix, so this coefficient has spectral norm at most 1 (as β is sufficiently small). So,we can bound $\|\mathbf{w}_{s+1}\|_2$ as:

$$\|\mathbf{w}_{s+1}\|_2 \le \|\mathbf{w}_s\|_2 + \frac{c\beta}{\sqrt{\alpha}}\|\boldsymbol{\Delta}_s\|_2^2\eta_* \tag{60}$$

which allows us to apply Lemma 3 with $\xi_{1,s} = 0$ and $\xi_{2,s} = \frac{c\beta}{\sqrt{\alpha}}\|\boldsymbol{\Delta}_s\|_2^2\eta_*$ for all $s \in [t]$. This results in:

$$\|\mathbf{w}_{t+1}\|_2 \le \sum_{s=1}^{t}\frac{c\beta}{\sqrt{\alpha}}\|\boldsymbol{\Delta}_s\|_2^2\eta_*.$$

Next, note that

$$\begin{aligned}\|\mathbf{\Delta}_s\|_2 &\le (1-0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_{s-1}\|_2 + \tfrac{5}{4}\beta^2\alpha^2 L_*^4\,\mathrm{dist}_{s-1}^2\\ &\vdots\\ &\le \sum_{r=1}^{s-1}(1-0.5\beta\alpha E_0\mu_*^2)^{s-1-r}(\tfrac{5}{4}\beta^2\alpha^2 L_*^4\,\mathrm{dist}_r^2)\\ &\le \tfrac{5}{4}L_*^4\beta^2\alpha^2(1-0.5\beta\alpha E_0\mu_*^2)^{s-1}\sum_{r=1}^{s-1}(1-0.5\beta\alpha E_0\mu_*^2)^r\\ &\le \frac{5\beta\alpha L_*^4}{2E_0\mu_*^2}(1-0.5\beta\alpha E_0\mu_*^2)^{s-1} \end{aligned} \tag{61}$$

therefore

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2 &\le \sum_{s=1}^{t} c\frac{25\beta^3\alpha^{1.5}L_*^8}{4E_0^2\mu_*^4}(1-0.5\beta\alpha E_0\mu_*^2)^{2s-2}\eta_*\\ &\le \frac{14\beta^2\sqrt{\alpha}L_*^8}{E_0^3\mu_*^6}\eta_*\\ &\le \frac{\sqrt{\alpha}E_0}{10\kappa_*^2}\eta_* \qquad (62)\\ &\le \frac{\sqrt{\alpha}E_0}{10}\min(1,\tfrac{\mu_*^2}{\eta_*^2})\eta_*. \qquad (63)\end{aligned}$$

where (62) follows by choice of $\beta \le \frac{\alpha E_0^2}{40\kappa_*^4}$ and $\alpha \le 1/L_*$, and (63) follows since $\eta_* \le L_*$. □

Lemma 10 (Exact ANIL $A_2(t+1)$)**.** *Suppose the conditions of Theorem 1 are satisfied and $A_1(t)$, $A_3(t)$ and $A_5(t)$ hold, then $A_2(t+1)$ holds, i.e.*

$$\|\mathbf{\Delta}_{t+1}\|_2 \le (1-0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2 L_*^4\,\mathrm{dist}_t^2\,. \tag{64}$$

Proof. Let $\mathbf{G}_t \coloneqq \frac{1}{n}\sum_{i=1}^n \nabla_{\mathbf{B}} F_{t,i}(\mathbf{B}_t,\mathbf{w}_t) = \frac{1}{\beta}(\mathbf{B}_t - \mathbf{B}_{t+1})$ again be the outer loop gradient for the representation, where $\nabla_{\mathbf{B}} F_{t,i}(\mathbf{B}_t,\mathbf{w}_t)$ is written in (57). Note that $\mathbf{G}_t$ can be re-written as:

$$\mathbf{G}_t = -\bar{\mathbf{\Delta}}_t\mathbf{S}_t\mathbf{B}_t - \mathbf{S}_t\mathbf{B}_t\mathbf{\Delta}_t + \mathbf{N}_t \tag{65}$$

where $\mathbf{S}_t \coloneqq \alpha\mathbf{B}_*\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\right)\mathbf{B}_*^\top$ and

$$\begin{aligned}\mathbf{N}_t \coloneqq \tfrac{1}{n}\sum_{i=1}^n \Big(&\mathbf{v}_{t,i}\mathbf{w}_t^\top\mathbf{\Delta}_t + \alpha\bar{\mathbf{\Delta}}_t\mathbf{B}_t\mathbf{w}_t\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t - \alpha\mathbf{B}_t\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t - \alpha\mathbf{B}_t\mathbf{B}_t^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top\\ &+ \alpha\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_t\mathbf{\Delta}_t\Big)\end{aligned} \tag{66}$$

Since Lemma 1 shows that $\sigma_{\min}(\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t) \ge E_0\mu_*^2$, Lemma 4 (with $\chi = 1$) implies that

$$\|\mathbf{\Delta}_{t+1}\|_2 \le (1-2\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + 2\beta\alpha\|\mathbf{B}_t^\top\mathbf{N}_t\|_2 + \beta^2\alpha^2\|\mathbf{G}_t\|_2^2 \tag{67}$$

It remains to control $\|\mathbf{B}_t^\top \mathbf{N}_t\|_2$ and $\|\mathbf{G}_t\|_2$. Note that

$$\begin{aligned}
\|\mathbf{B}_t^\top \mathbf{N}_t\|_2 &\le \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top \mathbf{v}_{t,i}\mathbf{w}_t^\top \mathbf{\Delta}_t\right\|_2 + \alpha\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top \bar{\mathbf{\Delta}}_t \mathbf{B}_t \mathbf{w}_t \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{B}_t\right\|_2 \\
&\quad + \alpha\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top \mathbf{B}_t \mathbf{w}_t \mathbf{v}_{t,i}^\top \mathbf{B}_t\right\|_2 + \alpha\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top \mathbf{B}_t \mathbf{B}_t^\top \mathbf{v}_{t,i}\mathbf{w}_t^\top\right\|_2 \\
&\quad + \alpha\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i}\mathbf{w}_t^\top \mathbf{B}_t^\top \mathbf{B}_t \mathbf{\Delta}_t\right\|_2 \\
&\le \tfrac{c}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2^2\left(\tfrac{c}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2 + \eta_*\right) + 2\tfrac{c}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\eta_* \\
&\quad + 2\tfrac{c}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\left(\tfrac{c}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2 + \eta_*\right) \\
&\le \tfrac{5}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\eta_* + \tfrac{3}{\alpha}\|\mathbf{w}_t\|_2^2\|\mathbf{\Delta}_t\|_2 \\
&\le 0.6E_0\mu_*^2\|\mathbf{\Delta}_t\|_2
\end{aligned} \tag{68}$$

by inductive hypotheses $A_1(t)$ and $A_3(t)$ and the fact that $\min(1, \frac{\mu_*^2}{\eta_*^2})\eta_*^2 \le \mu_*^2$. Next,

$$\begin{aligned}
\|\mathbf{G}_t\|_2 &\le \|\bar{\mathbf{\Delta}}_t\mathbf{S}_t\mathbf{B}_t\|_2 + \|\mathbf{S}_t\mathbf{B}_t\mathbf{\Delta}_t\|_2 + \|\mathbf{N}_t\|_2 \\
&\le c\sqrt{\alpha}(2\|\mathbf{\Delta}_t\|_2 + \mathrm{dist}_t)L_*^2 + \|\mathbf{N}_t\|_2 \\
&\le c\sqrt{\alpha}(2\|\mathbf{\Delta}_t\|_2 + \mathrm{dist}_t)L_*^2 + \|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\left(\tfrac{c}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2 + (\|\mathbf{\Delta}_t\|_2 + \mathrm{dist}_t)\eta_*\right) \\
&\quad + 4c\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\eta_* + 2\tfrac{c^2}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2^2\|\mathbf{\Delta}_t\|_2 \\
&\le c\sqrt{\alpha}(2\|\mathbf{\Delta}_t\|_2 + \mathrm{dist}_t)L_*^2 + 6\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\eta_* + \tfrac{3}{\sqrt{\alpha}}\|\mathbf{w}_t\|_2^2\|\mathbf{\Delta}_t\|_2 \\
&\le 3\sqrt{\alpha}L_*^2\|\mathbf{\Delta}_t\|_2 + c\sqrt{\alpha}L_*^2\,\mathrm{dist}_t \\
\implies \|\mathbf{G}_t\|_2^2 &\le \alpha L_*^4(9\|\mathbf{\Delta}_t\|_2^2 + 7\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\,\mathrm{dist}_t^2) \\
&\le \alpha L_*^4(8\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\,\mathrm{dist}_t^2)
\end{aligned} \tag{69}$$

Combining (67), (68) and (69) yields

$$\begin{aligned}
\|\mathbf{\Delta}_{t+1}\|_2 &\le (1 - 2\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + 2\beta\alpha\|\mathbf{B}_t^\top\mathbf{N}_t\|_2 + \beta^2\alpha\|\mathbf{G}_t\|_2^2 \\
&\le (1 - 2\beta\alpha E_0\mu_*^2 + 1.2\beta\alpha E_0\mu_*^2 + 8\beta^2\alpha^2L_*^4)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^4\,\mathrm{dist}_t^2 \\
&\le (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^4\,\mathrm{dist}_t^2
\end{aligned} \tag{70}$$

where the last inequality follows since $\beta \le \frac{\alpha E_0^2}{40\kappa_*^4}$ and $\alpha \le 1/L_*$. □

Corollary 2 (Exact ANIL $A_3(t+1)$)**.** *Suppose the conditions of Theorem 1 are satisfied. If $A_2(t+1)$ and $A_3(t)$ hold. Then $A_3(t+1)$ holds, i.e.*

$$\|\mathbf{\Delta}_{t+1}\|_2 \le \tfrac{1}{10} \tag{71}$$

Proof. Note that according to equation (70), we have

$$\begin{aligned}
\|\mathbf{\Delta}_{t+1}\|_2 &\le (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2L_*^4 \\
&\le (1 - 0.5\beta\alpha E_0\mu_*^2)\tfrac{1}{10} + \tfrac{5}{4}E_0\beta\alpha\mu_*^2 \\
&\le \tfrac{1}{10}
\end{aligned} \tag{72}$$

where equation (72) is satisfied by the choice of $\beta \le \frac{\alpha E_0^2}{40\kappa_*^4}$. and inductive hypothesis $A_3(t)$. □

Lemma 11 (Exact-ANIL $A_4(t+1)$)**.** *Suppose the conditions of Theorem 1 are satisfied and that inductive hypotheses $A_1(t)$, $A_3(t)$ and $A_6(t)$ hold. Then $A_4(t+1)$ holds, i.e.*

$$\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{t+1,i}\mathbf{w}_{t+1,i}^\top\right) \geq 0.9\alpha E_0\mu_*^2$$
$$\text{and } \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{t+1,i}\mathbf{w}_{t+1,i}^\top\right) \leq 1.2\alpha L_*^2$$

Proof. The proof is identical to that of Lemma 7. □

Lemma 12 (Exact ANIL $A_5(t+1)$)**.** *Suppose the conditions of Theorem 1 are satisfied. If inductive hypothesis $A_4(t)$ holds, then $A_5(t+1)$ holds, that is*

$$\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \leq (1-0.5\beta\alpha E_0\mu_*^2)\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2$$

Proof. Note that from (57), the outer loop gradient for the (t,i)-th task can be re-written as:

$$\begin{aligned}\nabla_{\mathbf{B}}F_{t,i}(\mathbf{B}_t,\mathbf{w}_t) &= \mathbf{v}_{t,i}\mathbf{w}_t^\top\boldsymbol{\Delta}_t + \alpha\mathbf{v}_{t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t - \alpha\mathbf{B}_t\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t - \alpha\mathbf{B}_t\mathbf{B}_t^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top + \alpha\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{v}_{t,i}^\top\mathbf{B}_t\\ &= \mathbf{B}_t\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top - \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top - \alpha\mathbf{B}_t\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t - \alpha\mathbf{B}_t\mathbf{B}_t^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top + \alpha\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{v}_{t,i}^\top\mathbf{B}_t\end{aligned}$$

Therefore, noting $\mathbf{G}_t = \frac{1}{n}\sum_{i=1}^n\nabla_{\mathbf{B}}F_{t,i}(\mathbf{B}_t,\mathbf{w}_t)$, and using $\mathbf{B}_{*,\perp}^\top\mathbf{B}_* = \mathbf{0}$, we have

$$\begin{aligned}\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 &= \|\mathbf{B}_{*,\perp}^\top(\mathbf{B}_t-\beta\mathbf{G}_t)\|_2\\ &\leq \left\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\left(\mathbf{I}_k - \tfrac{\beta}{n}\sum_{i=1}^n\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top + \tfrac{\beta\alpha}{n}\sum_{i=1}^n(\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t + \mathbf{B}_t^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top)\right)\right\|_2\\ &\leq \left\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\left(\mathbf{I}_k - \tfrac{\beta}{n}\sum_{i=1}^n\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)\right\|_2 + 2\beta\alpha\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2\left\|\tfrac{1}{n}\sum_{i=1}^n\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t\right\|_2\\ &\leq \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2\left(1-0.9\beta\alpha E_0\mu_*^2 + \beta\alpha\tfrac{3E_0}{100}\min(1,\tfrac{\mu_*^2}{\eta_*^2})\eta_*^2\right) \qquad (73)\\ &\leq \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2(1-0.5\beta\alpha E_0\mu_*^2)\end{aligned}$$

where (73) follows by inductive hypotheses $A_1(t)$, $A_3(t)$, and $A_4(t)$, and $A_3(t)$ and the fact that $\min(1,\frac{\mu_*^2}{\eta_*^2})\eta_*^2 \leq \mu_*^2$. □

D. MAML Infnite Samples

D.1. FO-MAML

We consider FO-MAML when $m_{in} = m_{out} = \infty$. In this case, the inner loop updates are:

$$\begin{aligned}\mathbf{w}_{t,i} &= \mathbf{w}_t - \alpha\nabla_{\mathbf{w}}\mathcal{L}_{t,i}(\mathbf{B}_t,\mathbf{w}_t)\\ &= (\mathbf{I}_d - \alpha\mathbf{B}_t^\top\mathbf{B}_t)\mathbf{w}_t + \alpha\mathbf{B}_t^\top\hat{\mathbf{B}}_*\mathbf{w}_{*,t,i}\\ \mathbf{B}_{t,i} &= \mathbf{B}_t - \alpha\nabla_{\mathbf{B}}\mathcal{L}_{t,i}(\mathbf{B}_t,\mathbf{w}_t)\\ &= \mathbf{B}_t(\mathbf{I}_k - \alpha\mathbf{w}_t\mathbf{w}_t^\top) + \alpha\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top \qquad (74)\end{aligned}$$

The outer loop updates are:

$$\begin{aligned}\mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{\beta}{n}\sum_{i=1}^n\nabla_{\mathbf{w}}\mathcal{L}_{t,i}(\mathbf{B}_{t,i},\mathbf{w}_{t,i}) = \mathbf{w}_t - \frac{\beta}{n}\sum_{i=1}^n\mathbf{B}_{t,i}^\top(\mathbf{B}_{t,i}\mathbf{w}_{t,i} - \mathbf{B}_*\mathbf{w}_{*,t,i})\\ \mathbf{B}_{t+1} &= \mathbf{B}_t - \frac{\beta}{n}\sum_{i=1}^n\nabla_{\mathbf{B}}\mathcal{L}_{t,i}(\mathbf{B}_{t,i},\mathbf{w}_{t,i}) = \mathbf{B}_t - \frac{\beta}{n}\sum_{i=1}^n(\mathbf{B}_{t,i}\mathbf{w}_{t,i} - \mathbf{B}_*\mathbf{w}_{*,t,i})\mathbf{w}_{t,i}^\top \qquad (75)\end{aligned}$$

Now we state the main result for Exact MAML in the infinite sample case. Due to third and higher-order products of the ground-truth heads that arise in the FO-MAML and MAML updates, we require an upper bound on the maximum $\mathbf{w}_{*,t,i}$. We define the parameter $L_{\max}$ as follows.

Assumption 4. *There exists $L_{\max} < \infty$ such that almost surely for all $t \in [T]$, we have*

$$\max_{i\in[n]} \|\mathbf{w}_{*,t,i}\|_2 \leq L_{\max} \tag{76}$$

Note that if Assumption 2 holds, we have $L_{\max} = O(\sqrt{k}L_*)$. Here we prove a slightly more general version of Theorem 3 in which we allow for arbitrary finite $L_{\max}$. Note that Theorem 6 immediately implies Theorem 2 after applying Assumption 2. First we state the following assumption, then we prove the theorem.

Assumption 5 (Initialization and small average ground-truth heads)**.** *The following holds almost surely:*

$$\mathrm{dist}_0 \leq \frac{4\mu_*}{5L_{\max}} \quad \textit{and, for all } t \in [T], \quad \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_{*,t,i} \right\|_2 \leq \eta_* \leq \frac{2E_0^2\mu_*^4}{L_{\max}^3}. \tag{77}$$

Theorem 6 (FO-MAML Infinite Samples)**.** *Let $m_{in} = m_{out} = \infty$ and define $E_0 := 0.9 - \mathrm{dist}_0^2$. Suppose that $\alpha \leq \frac{1}{4L_{\max}}$, $\beta \leq \frac{\alpha E_0^2}{60\kappa_*^4}$, $\alpha\mathbf{B}_t^\top\mathbf{B}_t = \mathbf{I}_k$, $\mathbf{w}_0 = \mathbf{0}$ and Assumptions 1, 4 and 5 hold. Then FO-MAML satisfies that for all $T \in \mathbb{Z}_+$,*

$$\mathrm{dist}(\mathbf{B}_T, \mathbf{B}_*) \leq (1 - 0.5\beta\alpha E_0\mu_*^2)^{T-1}. \tag{78}$$

Proof. The proof follows by showing that the following inductive hypotheses hold for all $t \in [T]$:

1. $A_1(t) := \{\|\mathbf{w}_t\|_2 \leq \frac{E_0^2}{10}\sqrt{\alpha}\mu_*\kappa_{*,\max}^{-3}\}$
2. $A_2(t) := \{\|\mathbf{\Delta}_t\|_2 \leq \frac{E_0}{10}\alpha^2\mu_*^2\}$
3. $A_3(t) := \{\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2 \leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t-1}\|_2\}$
4. $A_4(t) := \{\mathrm{dist}_t \leq \frac{\sqrt{10}}{3}(1 - 0.5\beta\alpha E_0\mu_*^2)^{t-1}\,\mathrm{dist}_0\}$
5. $A_5(t) := \{\mathrm{dist}_t \leq (1 - 0.5\beta\alpha E_0\mu_*^2)^{t-1}\}$

These conditions hold for iteration $t = 0$ due to the choice of initialization. Now, assuming they hold for arbitrary t, we will show they hold at $t + 1$.

1. $A_1(t) \cap A_2(t) \cap A_4(t) \implies A_1(t+1)$. This is Lemma 13.
2. $A_1(t) \cap A_2(t) \cap A_4(t) \implies A_2(t+1)$. This is Lemma 14.
3. $A_1(t) \cap A_2(t) \cap A_4(t) \implies A_3(t+1)$. This is Lemma 15.
4. $A_2(t+1) \cap \bigcap_{s=1}^{t+1} A_3(s) \implies A_4(t+1) \cap A_5(t+1)$. Note that $A_2(t+1) \cap \bigcap_{s=1}^{t+1} A_3(s)$ implies

$$\begin{aligned}
\frac{\sqrt{1-\|\mathbf{\Delta}_{t+1}\|_2}}{\sqrt{\alpha}}\,\mathrm{dist}_{t+1} &= \frac{\sqrt{1-\|\mathbf{\Delta}_{t+1}\|_2}}{\sqrt{\alpha}}\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \\
&\leq \sigma_{\min}(\mathbf{B}_{t+1})\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \\
&\leq \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \\
&\leq \left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^t \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_0\|_2 \\
&\leq \tfrac{1}{\sqrt{\alpha}}\left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^t \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_0\|_2 \\
&= \tfrac{1}{\sqrt{\alpha}}\left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^t \mathrm{dist}_0\,.
\end{aligned}$$

Dividing both sides by $\frac{\sqrt{1-\|\mathbf{\Delta}_{t+1}\|_2}}{\sqrt{\alpha}}$ and using the facts that $\mathrm{dist}_0 \leq \frac{3}{\sqrt{10}}$ and $\|\mathbf{\Delta}_{t+1}\|_2 \leq \frac{1}{10}$ yields

$$\begin{aligned}\mathrm{dist}_{t+1} &\leq \tfrac{1}{\sqrt{1-\|\mathbf{\Delta}_{t+1}\|_2}} \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t \mathrm{dist}_0 \\ &\leq \tfrac{\sqrt{10}}{3} \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t \mathrm{dist}_0 \\ &\leq \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^t, \end{aligned} \tag{79}$$

as desired.

□

Lemma 13 (FO-MAML $A_1(t+1)$)**.** *Suppose the conditions of Theorem 6 are satisfied and $A_1(t), A_2(t)$ and $A_4(t)$ hold. Then $A_1(t+1)$ holds, i.e.*

$$\|\mathbf{w}_{t+1}\|_2 \leq \tfrac{E_0^2}{10}\sqrt{\alpha}\mu_* \kappa_{*,\max}^{-3}. \tag{80}$$

Proof. Let $\mathbf{G}_{t,i}$ be the inner loop gradient for the representation for the (t,i)-th task, in particular $\mathbf{G}_{t,i} = \mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top - \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top$. By expanding the outer loop update for the head, we obtain:

$$\begin{aligned}\mathbf{w}_{t+1} &= \frac{1}{n}\sum_{i=1}^n (\mathbf{I}_k - \beta\mathbf{B}_{t,i}^\top\mathbf{B}_{t,i}(\mathbf{I} - \alpha\mathbf{B}_t^\top\mathbf{B}_t))\mathbf{w}_t + \beta\frac{1}{n}\sum_{i=1}^n \mathbf{B}_{t,i}^\top(\mathbf{I} - \alpha\mathbf{B}_{t,i}\mathbf{B}_t^\top)\mathbf{B}_*\mathbf{w}_{*,t,i} \\ &= \frac{1}{n}\sum_{i=1}^n (\mathbf{I}_k - \beta\mathbf{B}_{t,i}^\top\mathbf{B}_{t,i}(\mathbf{I} - \alpha\mathbf{B}_t^\top\mathbf{B}_t))\mathbf{w}_t + \beta\alpha^2\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top\mathbf{G}_{t,i}\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i} \\ &\quad + \beta\frac{1}{n}\sum_{i=1}^n \mathbf{B}_{t,i}^\top(\mathbf{I} - \alpha\mathbf{B}_t\mathbf{B}_t^\top)\mathbf{B}_*\mathbf{w}_{*,t,i} - \beta\alpha^3\frac{1}{n}\sum_{i=1}^n \mathbf{G}_{t,i}^\top\mathbf{G}_{t,i}\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i} \\ &= \frac{1}{n}\sum_{i=1}^n (\mathbf{I}_k - \beta\mathbf{B}_{t,i}^\top\mathbf{B}_{t,i}(\mathbf{I} - \alpha\mathbf{B}_t^\top\mathbf{B}_t))\mathbf{w}_t \\ &\quad + \beta\alpha^2\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top(\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top - \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top)\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i} \\ &\quad + \beta\frac{1}{n}\sum_{i=1}^n \mathbf{B}_{t,i}^\top(\mathbf{I} - \alpha\mathbf{B}_t\mathbf{B}_t^\top)\mathbf{B}_*\mathbf{w}_{*,t,i} - \beta\alpha^3\frac{1}{n}\sum_{i=1}^n \mathbf{G}_{t,i}^\top\mathbf{G}_{t,i}\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i} \\ &= \frac{1}{n}\sum_{i=1}^n (\mathbf{I}_k - \beta\mathbf{B}_{t,i}^\top\mathbf{B}_{t,i}(\mathbf{I} - \alpha\mathbf{B}_t^\top\mathbf{B}_t))\mathbf{w}_t - \beta\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\right)\mathbf{B}_*^\top\mathbf{B}_t\mathbf{w}_t \\ &\quad + \beta\alpha^2\mathbf{B}_t^\top\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\right) \\ &\quad + \beta\frac{1}{n}\sum_{i=1}^n \mathbf{B}_{t,i}^\top(\mathbf{I} - \alpha\mathbf{B}_t\mathbf{B}_t^\top)\mathbf{B}_*\mathbf{w}_{*,t,i} - \beta\alpha^3\frac{1}{n}\sum_{i=1}^n \mathbf{G}_{t,i}^\top\mathbf{G}_{t,i}\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i} \\ &= \left(\mathbf{I}_k - \beta\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\right)\mathbf{B}_*^\top\mathbf{B}_t\right)\mathbf{w}_t + \mathbf{N}_t \end{aligned} \tag{81}$$

where $\mathbf{N}_t := -\beta\frac{1}{n}\sum_{i=1}^n \mathbf{B}_{t,i}^\top\mathbf{B}_{t,i}\mathbf{\Delta}_t\mathbf{w}_t + \beta\alpha^2\mathbf{B}_t^\top\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i} + \beta\frac{1}{n}\sum_{i=1}^n \mathbf{B}_{t,i}^\top\bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i} - \beta\alpha^3\frac{1}{n}\sum_{i=1}^n \mathbf{G}_{t,i}^\top\mathbf{G}_{t,i}\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}$. Since $\sigma_{\min}(\mathbf{B}_t^\top\mathbf{B}_*\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\right)\mathbf{B}_*^\top\mathbf{B}_t) \geq \frac{1}{\alpha}E_0\mu_*^2$ by Lemma 1, and $\beta \leq \frac{1}{2\alpha L_*^2}$, we have

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2 &\leq \left\|\mathbf{I}_k - \beta\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\right)\mathbf{B}_*^\top\mathbf{B}_t\right\|_2 \|\mathbf{w}_t\|_2 + \|\mathbf{N}_t\|_2 \\ &\leq (1 - \beta\alpha E_0\mu_*^2)\|\mathbf{w}_t\|_2 + \|\mathbf{N}_t\|_2 \end{aligned} \tag{82}$$

The remainder of the proof deals with bounding $\|\mathbf{N}_t\|_2$. First note that $\bigcup_{s=0}^{t} A_2(s)$ with $\alpha \leq 1/(4L_{\max})$ implies $\sigma_{\max}(\mathbf{B}_s^\top \mathbf{B}_s) \leq \frac{1+\|\mathbf{\Delta}_s\|_2}{\alpha} < \frac{1.1^{2/3}}{\alpha}$ for all $s \in \{0, \ldots, t+1\}$. In turn, this means that $\alpha^{1.5}\|\mathbf{B}_s\|_2^3 \leq 1.1$ Let $c \coloneqq 1.1$.

We consider each of the four terms in $\mathbf{N}_t$ separately. Using $\sqrt{\alpha}\|\mathbf{B}_t\|_2, \alpha^{1.5}\|\mathbf{B}_t\|_2^3 \leq c$ and the Cauchy-Schwarz and triangle inequalities, we have

$$\begin{aligned}
\beta \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{B}_{t,i}^\top \mathbf{B}_{t,i} \right) \mathbf{\Delta}_t \mathbf{w}_t \right\|_2 &\leq \beta \Bigg(\|\mathbf{B}_t\|_2^2 + 2\alpha \|\mathbf{B}_t\|_2 \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_{*,t,i} \right\|_2 \|\mathbf{w}_t\|_2 \\
&\qquad + \alpha^2 \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_{*,t,i} \mathbf{w}_{*,t,i}^\top \right\|_2 \|\mathbf{w}_t\|_2^2 \Bigg) \|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2 \\
&\leq \beta \left(\tfrac{c}{\alpha} + 2c\sqrt{\alpha} \|\mathbf{w}_t\|_2 \eta_* + \alpha^2 L_*^2 \|\mathbf{w}_t\|_2^2 \right) \|\mathbf{\Delta}_t\|_2 \\
\beta\alpha^2 \left\| \mathbf{B}_t^\top \mathbf{B}_t \mathbf{w}_t \mathbf{w}_t^\top \mathbf{B}_t^\top \mathbf{B}_* \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_{*,t,i} \right\|_2 &\leq c\beta\sqrt{\alpha}\eta_* \|\mathbf{w}_t\|_2^2 \\
\beta \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{B}_{t,i}^\top \bar{\mathbf{\Delta}}_t \mathbf{B}_* \mathbf{w}_{*,t,i} \right\|_2 &\leq \frac{c\beta}{\sqrt{\alpha}} \|\mathbf{\Delta}_t\|_2 \eta_* + \beta\alpha L_{\max}^2 \|\mathbf{w}_t\|_2 \|\mathbf{\Delta}_t\|_2 + \beta\alpha L_{\max}^2 \|\mathbf{w}_t\|_2 \operatorname{dist}_t^2 \qquad (83) \\
\beta\alpha^3 \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{G}_{t,i}^\top \mathbf{G}_{t,i} \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} \right\|_2 &\leq c\beta\alpha^{2.5} \left(\tfrac{c\|\mathbf{w}_t\|_2^2}{\alpha} \eta_* + \tfrac{2c\|\mathbf{w}_t\|_2 L_*^2}{\sqrt{\alpha}} + L_{\max}^3 \right) \|\mathbf{w}_t\|_2^2. \qquad (84)
\end{aligned}$$

Note that the dist_t^2 in (83) is due to the fact that $\|\mathbf{B}_*^\top (\mathbf{I}_k - \hat{\mathbf{B}}_t \hat{\mathbf{B}}_t^\top) \mathbf{B}_*\|_2 = \|\mathbf{B}_*^\top (\mathbf{I}_k - \hat{\mathbf{B}}_t \hat{\mathbf{B}}_t^\top)(\mathbf{I}_k - \hat{\mathbf{B}}_t \hat{\mathbf{B}}_t^\top) \mathbf{B}_*\|_2 \leq \operatorname{dist}_t^2$. Combining these bounds and applying inductive hypotheses $A_2(t)$ and $A_3(t)$ yields

$$\begin{aligned}
\|\mathbf{N}_t\|_2 &\leq \beta \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{B}_{t,i}^\top \mathbf{B}_{t,i} \right) \mathbf{\Delta}_t \mathbf{w}_t \right\|_2 + \beta\alpha^2 \left\| \mathbf{B}_t^\top \mathbf{B}_t \mathbf{w}_t \mathbf{w}_t^\top \mathbf{B}_t^\top \mathbf{B}_* \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_{*,t,i} \right\|_2 \\
&\quad + \beta \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{B}_{t,i}^\top \bar{\mathbf{\Delta}}_t \mathbf{B}_* \mathbf{w}_{*,t,i} \right\|_2 + \beta\alpha^3 \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{G}_{t,i}^\top \mathbf{G}_{t,i} \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} \right\|_2 \\
&\leq \beta \left(\tfrac{c}{\alpha} + 2c\sqrt{\alpha} \|\mathbf{w}_t\|_2 \eta_* + \alpha^2 L_*^2 \|\mathbf{w}_t\|_2^2 \right) \|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2 + c\beta\sqrt{\alpha}\eta_* \|\mathbf{w}_t\|_2^2 + \tfrac{c\beta}{\sqrt{\alpha}} \|\mathbf{\Delta}_t\|_2 \eta_* \\
&\quad + \beta\alpha L_{\max}^2 \|\mathbf{w}_t\|_2 \|\mathbf{\Delta}_t\|_2 + \beta\alpha L_{\max}^2 \|\mathbf{w}_t\|_2 \operatorname{dist}_t^2 + c\beta\alpha^{2.5} \left(\tfrac{c\|\mathbf{w}_t\|_2^2}{\alpha} \eta_* + \tfrac{2c\|\mathbf{w}_t\|_2 L_*^2}{\sqrt{\alpha}} + L_{\max}^3 \right) \|\mathbf{w}_t\|_2^2 \\
&\leq \tfrac{2c}{100} \beta\alpha^{1.5} \mu_*^3 \kappa_{*,\max}^{-3} E_0^3 + \tfrac{2c}{10} \beta\alpha^{1.5} \mu_*^2 \eta_* E_0 + \tfrac{1}{10} \beta\alpha^{1.5} \mu_*^3 \kappa_{*,\max} E_0^2 \operatorname{dist}_t^2
\end{aligned}$$

Thus we have

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|_2 &\leq \left(1 - \beta\alpha E_0 \mu_*^2\right) \|\mathbf{w}_t\|_2 + \tfrac{2c}{100} \beta\alpha^{1.5} \mu_*^3 \kappa_{*,\max}^{-3} E_0^3 + \tfrac{2c}{10} \beta\alpha^{1.5} \mu_*^2 \eta_* E_0 + \tfrac{1}{10} \beta\alpha^{1.5} \mu_*^3 \kappa_{*,\max} E_0^2 \operatorname{dist}_t^2 \\
&\leq \tfrac{1}{10} E_0^2 \sqrt{\alpha} \mu_* \kappa_{*,\max}^{-1} - \tfrac{1}{10} \beta\alpha^{1.5} E_0^3 \mu_*^3 \kappa_{*,\max}^{-1} + \tfrac{2c}{100} \beta\alpha^{1.5} \mu_*^3 \kappa_{*,\max}^{-3} E_0^3 + \tfrac{2c}{10} \beta\alpha^{1.5} \mu_*^2 \eta_* E_0 \\
&\quad + \tfrac{1}{10} \beta\alpha^{1.5} \mu_*^3 \kappa_{*,\max} E_0^2 \operatorname{dist}_0^2 \\
&\leq \tfrac{1}{10} E_0^2 \sqrt{\alpha} \mu_* \kappa_{*,\max}^{-1} \qquad (85)
\end{aligned}$$

where (85) follows by Assumption 5, namely:

$$\eta_* \leq \frac{2E_0^2 \mu_*^4}{L_{\max}^3} \quad \text{and} \quad \operatorname{dist}_0 \leq \frac{4\mu_*}{5L_{\max}}. \qquad (86)$$

□

Lemma 14 (FO-MAML $A_2(t+1)$)**.** *Suppose the conditions of Theorem 6 are satisfied and $A_1(t), A_2(t)$ and $A_4(t)$ hold. Then $A_2(t+1)$ holds almost surely, i.e.*

$$\|\mathbf{\Delta}_t\|_2 \leq \tfrac{E_0}{10} \alpha^2 \mu_*^2. \qquad (87)$$

Proof. We will employ Lemma (4), which requires writing the outer loop gradient for the representation, i.e. $\mathbf{G}_t \coloneqq \frac{1}{\beta}(\mathbf{B}_t - \mathbf{B}_{t+1})$, as $\mathbf{G}_t = -\bar{\mathbf{\Delta}}_t \mathbf{S}_t \mathbf{B}_t - \chi \mathbf{B}_t \mathbf{S}_t \mathbf{\Delta}_t + \mathbf{N}_t$, for some positive definite matrix $\mathbf{S}_t$ and a matrix $\mathbf{N}_t$ (note that this $\mathbf{N}_t$ is different from the $\mathbf{N}_t$ from that was used in the previous lemma). To this end, we expand the outer loop gradient:

$$
\begin{aligned}
\mathbf{G}_t &\coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbf{B}_{t,i}\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top - \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top \\
&= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{B}_{t,i}(\mathbf{I}_k - \alpha \mathbf{B}_t^\top \mathbf{B}_t)\mathbf{w}_t - (\mathbf{I}_d - \alpha \mathbf{B}_{t,i}\mathbf{B}_t^\top)\mathbf{B}_*\mathbf{w}_{*,t,i})\mathbf{w}_{t,i}^\top \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{B}_{t,i}\mathbf{\Delta}_t\mathbf{w}_t\mathbf{w}_{t,i}^\top - \bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top + \alpha^2 \mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top \\
&\qquad - \alpha^2 \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top \\
&= -\bar{\mathbf{\Delta}}_t\mathbf{B}_*\Big(\alpha\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\Big)\mathbf{B}_*^\top\mathbf{B}_t + \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{B}_{t,i}\mathbf{\Delta}_t\mathbf{w}_t\mathbf{w}_{t,i}^\top - \bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{\Delta}_t \\
&\quad + \alpha^2\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top - \alpha^2\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\Big) \\
&= -\bar{\mathbf{\Delta}}_t\mathbf{S}_t\mathbf{B}_t + \mathbf{N}_t
\end{aligned} \tag{88}
$$

where $\mathbf{S}_t \coloneqq \mathbf{B}_*\big(\alpha\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\big)\mathbf{B}_*^\top$,

$$
\begin{aligned}
\mathbf{N}_t \coloneqq \frac{1}{n}\sum_{i=1}^{n} &\big(\mathbf{B}_{t,i}\mathbf{\Delta}_t\mathbf{w}_t\mathbf{w}_{t,i}^\top - \bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{\Delta}_t + \alpha^2\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top \\
&- \alpha^2\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\big),
\end{aligned} \tag{89}
$$

and $\chi = 0$. Since $\sigma_{\min}(\mathbf{B}_t^\top\mathbf{S}_t\mathbf{B}_t) \geq E_0\mu_*^2$ (by Lemma 1), Lemma 4 shows

$$
\|\mathbf{\Delta}_{t+1}\|_2 \leq (1 - \beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + 2\beta\alpha\|\mathbf{B}_t^\top\mathbf{N}_t\|_2 + \beta^2\alpha\|\mathbf{G}_t\|_2^2 \tag{90}
$$

So, the remainder of the proof is to bound $\|\mathbf{B}_t^\top\mathbf{N}_t\|_2$ and $\|\mathbf{G}_t\|_2^2$. First we deal with $\|\mathbf{B}_t^\top\mathbf{N}_t\|_2$. We have

$$
\begin{aligned}
\|\mathbf{B}_t^\top\mathbf{N}_t\|_2 \leq &\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_t^\top\mathbf{B}_{t,i}\mathbf{\Delta}_t\mathbf{w}_t\mathbf{w}_{t,i}^\top\right\|_2 + \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_t^\top\bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{\Delta}_t\right\|_2 \\
&+ \left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t^\top\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\right\|_2 + \left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\right\|_2
\end{aligned} \tag{91}
$$

We consider each of the four terms in (91) separately.

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_t^\top\mathbf{B}_{t,i}\boldsymbol{\Delta}_t\mathbf{w}_t\mathbf{w}_{t,i}^\top\right\|_2 &\le \left\|\mathbf{B}_t^\top\mathbf{B}_t\boldsymbol{\Lambda}_t\boldsymbol{\Delta}_t\mathbf{w}_t\mathbf{w}_t\boldsymbol{\Delta}_t^\top\right\|_2 \\
&\quad+\alpha\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_t^\top\mathbf{B}_t\boldsymbol{\Lambda}_t\boldsymbol{\Delta}_t\mathbf{w}_t\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\right\|_2 \\
&\quad+\alpha\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\boldsymbol{\Delta}_t\mathbf{w}_t\mathbf{w}_t\boldsymbol{\Delta}_t^\top\right\|_2 \\
&\quad+\alpha^2\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\boldsymbol{\Delta}_t\mathbf{w}_t\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\right\|_2 \\
&\le \tfrac{c}{\alpha}\|\boldsymbol{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^2+\tfrac{c}{\sqrt{\alpha}}\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* \\
&\quad+c\sqrt{\alpha}\|\boldsymbol{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^3\eta_*+c\alpha\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^2L_*^2 \\
\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_t^\top\bar{\boldsymbol{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\boldsymbol{\Delta}_t\right\|_2 &\le \tfrac{c}{\sqrt{\alpha}}\|\boldsymbol{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2\eta_* \\
\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t^\top\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\right\|_2 &\le c\sqrt{\alpha}\|\mathbf{w}_t\|_2^2(\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_*+\sqrt{\alpha}L_*^2) \\
\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\right\|_2 &\le c\alpha\|\mathbf{w}_t\|_2(\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2L_*^2+\sqrt{\alpha}L_{\max}^3)
\end{aligned}$$

Therefore, after applying inductive hypotheses $A_1(t)$ and $A_2(t)$, we obtain

$$\|\mathbf{B}_t^\top\mathbf{N}_t\|_2 \le 2c\tfrac{E_0^2}{10}\alpha^2\mu_*^4.$$

Next we bound $\|\mathbf{G}_t\|_2^2$. Note that $\|\mathbf{G}_t\|_2 \le \|\bar{\boldsymbol{\Delta}}_t\mathbf{S}_t\mathbf{B}_t\|_2+\|\mathbf{N}_t\|_2$, and

$$\begin{aligned}
\|\bar{\boldsymbol{\Delta}}_t\mathbf{S}_t\mathbf{B}_t\|_2 &\le c\sqrt{\alpha}L_*^2(\|\boldsymbol{\Delta}_t\|_2+\mathrm{dist}_t) \\
&\le c\sqrt{\alpha}L_*^2(\alpha^2\mu_*^2E_0+\mathrm{dist}_t).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\|\mathbf{N}_t\|_2 &\le \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_{t,i}\boldsymbol{\Delta}_t\mathbf{w}_t\mathbf{w}_{t,i}^\top\right\|_2+\left\|\frac{1}{n}\sum_{i=1}^{n}\bar{\boldsymbol{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\boldsymbol{\Delta}_t\right\|_2 \\
&\quad+\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\right\|_2+\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top\right\|_2 \\
&\le \tfrac{3cE_0}{10}\alpha^{2.5}\mu_*^4
\end{aligned}$$

thus

$$\begin{aligned}
\|\mathbf{G}_t\|_2^2 &\le \left(c\sqrt{\alpha}L_*^2(\alpha^2\mu_*^2+\mathrm{dist}_t)+\tfrac{3cE_0}{10}\alpha^{2.5}\mu_*^4\right)^2 \\
&\le 3c^2\alpha^5L_*^4\mu_*^4+2c^2\alpha L_*^4\,\mathrm{dist}_t^2 \\
&\le 3\alpha L_*^4 \qquad (92)
\end{aligned}$$

which means that

$$\begin{aligned}
\|\boldsymbol{\Delta}_{t+1}\|_2 &\le (1-\beta\alpha E_0\mu_*^2)\|\boldsymbol{\Delta}_t\|_2+\tfrac{2cE_0^2}{10}\beta\alpha^3\mu_*^4+3\beta^2\alpha^2L_*^4 \\
&\le \tfrac{1}{10}\alpha^2E_0\mu_*^2-\tfrac{1}{10}\beta\alpha^3E_0^2\mu_*^4+\tfrac{2cE_0^2}{10}\beta\alpha^3\mu_*^4+3\beta^2\alpha^2L_*^4 \\
&\le \tfrac{1}{10}\alpha^2E_0\mu_*^2 \qquad (93)
\end{aligned}$$

where (93) follows by choice of $\beta \le \frac{\alpha E_0^2}{60\kappa_*^4}$. □

Lemma 15 (FO-MAML $A_3(t+1)$)**.** *Suppose the conditions of Theorem 6 are satisfied and $A_1(t)$, $A_2(t)$, and $A_4(t)$ hold. Then $A_3(t+1)$ holds almost surely, i.e.*

$$\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 \leq (1 - 0.5\beta\alpha E_0 \mu_*^2)\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2.$$

Proof. Recalling the definition of $\mathbf{B}_{t+1}$ from (75) and noting that $\mathbf{B}_{*,\perp}^\top \mathbf{B}_* = \mathbf{0}$, we obtain

$$\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1} = \mathbf{B}_{*,\perp}^\top \mathbf{B}_t \left(\mathbf{I}_k - \beta(\mathbf{I}_k - \alpha \mathbf{w}_t \mathbf{w}_t^\top) \frac{1}{n} \sum_{i=1}^n \mathbf{w}_{t,i} \mathbf{w}_{t,i}^\top \right) \tag{94}$$

Next, using the triangle and Cauchy-Schwarz inequalities, we obtain

$$\begin{aligned}
\left\| \mathbf{I}_k - \beta \left(\mathbf{I}_k - \alpha \mathbf{w}_t \mathbf{w}_t^\top \right) \frac{1}{n} \sum_{i=1}^n \mathbf{w}_{t,i} \mathbf{w}_{t,i}^\top \right\|_2 &\leq \left\| \mathbf{I}_k - \frac{\beta}{n} \sum_{i=1}^n \mathbf{w}_{t,i} \mathbf{w}_{t,i}^\top \right\| + \beta\alpha \left\| \mathbf{w}_t \mathbf{w}_t^\top \frac{1}{n} \sum_{i=1}^n \mathbf{w}_{t,i} \mathbf{w}_{t,i}^\top \right\|_2 \\
&\leq 1 - \beta \sigma_{\min} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{w}_{t,i} \mathbf{w}_{t,i}^\top \right) + \beta\alpha \|\mathbf{w}_t\|_2^2 \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{w}_{t,i} \mathbf{w}_{t,i}^\top \right\|_2 \\
&\leq 1 - \beta \left(\alpha E_0 \mu_*^2 - c\eta_* \sqrt{\alpha} \|\mathbf{w}_t\|_2 \|\mathbf{\Delta}_t\|_2 \right) \qquad (95) \\
&\quad + c\beta\alpha \|\mathbf{w}_t\|_2^2 \left(\|\mathbf{w}_t\|_2 \|\mathbf{\Delta}_t\|_2^2 + \eta_* \sqrt{\alpha} \|\mathbf{w}_t\|_2 \|\mathbf{\Delta}_t\|_2 + \alpha L_*^2 \right) \\
&\leq 1 - \beta\alpha E_0 \mu_*^2 + 2\tfrac{E_0^2}{100} \beta\alpha^3 \mu_*^3 \eta_* \kappa_{\max,*}^{-3} + c\tfrac{E_0^4}{100} \beta\alpha^3 \mu_*^2 L_*^2 \kappa_{*,\max}^{-6} \\
&\leq 1 - 0.5\beta\alpha E_0 \mu_*^2 \qquad (96)
\end{aligned}$$

where (95) follows by the diversity of the inner loop-updated heads (Lemma 2) and (96) follows from $\alpha \leq 1/(4L_{\max})$. □

D.2. Exact MAML

The first step in the analysis is to compute the second-order outer loop updates for Exact MAML. To do so, we must compute the loss on task i at iteration t after one step of gradient descent for both the representation and head. Let $\mathbf{\Lambda}_t := \mathbf{I}_k - \alpha \mathbf{w}_t \mathbf{w}_t^\top$, $\mathbf{\Delta}_t := \mathbf{I}_k - \alpha \mathbf{B}_t^\top \mathbf{B}_t$, and $\bar{\mathbf{\Delta}}_t := \mathbf{I}_d - \alpha \mathbf{B}_t \mathbf{B}_t^\top$. Note that

$$F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) := \mathcal{L}_{t,i}(\mathbf{B}_t - \alpha \nabla_{\mathbf{B}} \mathcal{L}_{t,i}(\mathbf{B}_t, \mathbf{w}_t), \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} \mathcal{L}_{t,i}(\mathbf{B}_t, \mathbf{w}_t)) = \tfrac{1}{2} \|\mathbf{v}_{t,i}\|_2^2 \tag{97}$$

where

$$\begin{aligned}
\mathbf{v}_{t,i} &= \mathbf{B}_t \mathbf{\Lambda}_t \mathbf{\Delta}_t \mathbf{w}_t + \alpha \mathbf{B}_t \mathbf{\Lambda}_t \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} + \alpha \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_t^\top \mathbf{\Delta}_t \mathbf{w}_t + \alpha^2 \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_t^\top \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} - \mathbf{B}_* \mathbf{w}_{*,t,i} \\
&= \bar{\mathbf{\Delta}}_t (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) - \alpha (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_t^\top \mathbf{\Delta}_t \mathbf{w}_t - \alpha^2 \mathbf{B}_t \mathbf{w}_t \mathbf{w}_t^\top \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} \\
&\quad + \alpha^2 \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{B}_t \mathbf{w}_t \\
&= (\bar{\mathbf{\Delta}}_t - (\alpha\omega_t + \alpha^2 a_{t,i}) \mathbf{I}_d)(\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) \qquad (98)
\end{aligned}$$

where $a_{t,i} := \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{B}_t \mathbf{w}_t \ \forall t, i$ and $\omega_t := \mathbf{w}_t^\top \mathbf{\Delta}_t \mathbf{w}_t \ \forall t$. The outer loop updates for Exact MAML are given by:

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{\beta}{n} \sum_{i=1}^n \nabla_{\mathbf{w}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) \\
\mathbf{B}_{t+1} &= \mathbf{B}_t - \frac{\beta}{n} \sum_{i=1}^n \nabla_{\mathbf{B}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t)
\end{aligned}$$

Again, we prove a more general version of Theorem 2 in which we allow for general $L_{\max}$. First we make the following assumption.

Assumption 6 (Exact MAML Initialization)**.** *The distance of the initial representation to the ground-truth representation satisfies:*

$$\text{dist}_0 \leq \tfrac{1}{17} \kappa_{*,\max}^{-1.5}. \tag{99}$$

Theorem 7 (Exact MAML Infinite Samples)**.** *Let $m_{in} = m_{out} = \infty$ and define $E_0 \coloneqq 0.9 - \mathrm{dist}_0^2$. Suppose that $\alpha \leq \frac{E_0^{1/4}\kappa_*^{3/4}(L_*/L_{\max})^{1/4}}{4L_{\max}T^{1/4}}$ and $\beta \leq \frac{E_0\alpha}{10\kappa_*^4}$, $\alpha\mathbf{B}_0^\top\mathbf{B}_0 = \mathbf{I}_k$, $\mathbf{w}_0 = \mathbf{0}$ and Assumptions 1, 6, and 4 hold. Then Exact MAML satisfies*

$$\mathrm{dist}_T \leq (1 - 0.5\beta\alpha E_0\mu_*^2)^{T-1} \tag{100}$$

Proof. The proof follows by showing that the following inductive hypotheses hold for all $t \in [T]$:

1. $A_1(t) \coloneqq \{\|\mathbf{w}_t\|_2 \leq \|\mathbf{w}_{t-1}\|_2 + 16\beta\alpha^{3.5}L_{\max}^5 t + 3\beta\alpha^{1.5}L_{\max}^3\,\mathrm{dist}_t^2\}$
2. $A_2(t) \coloneqq \{\|\mathbf{w}_t\|_2 \leq \frac{E_0}{20}\sqrt{\alpha}\mu_*\}$.
3. $A_3(t) \coloneqq \|\mathbf{\Delta}_t\|_2 \leq \alpha^2 L_{\max}^2$
4. $A_4(t) \coloneqq \{\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2 \leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t-1}\|_2\}$
5. $A_5(t) \coloneqq \{\mathrm{dist}_t \leq \frac{\sqrt{10}}{3}(1 - 0.5\beta\alpha E_0\mu_*^2)^{t-1}\,\mathrm{dist}_0\}$
6. $A_6(t) \coloneqq \{\mathrm{dist}_t \leq (1 - 0.5\beta\alpha E_0\mu_*^2)^{t-1}\}$

These conditions hold for iteration $t = 0$ due to the choice of initialization. Now, assuming they hold for arbitrary t, we will show they hold at $t + 1$.

1. $A_2(t) \cap A_3(t) \implies A_1(t+1)$. This is Lemma 16.
2. $\bigcap_{s=1}^t \{A_1(s) \cap A_5(s)\} \implies A_2(t+1)$. This is Lemma 17.
3. $A_2(t) \cap A_3(t) \cap A_5(t) \implies A_3(t+1)$. This is Lemma 18.
4. $A_2(t) \cap A_3(t) \cap A_5(t) \implies A_4(t+1)$. This is Lemma 19.
5. $A_3(t+1) \cap \bigcap_{s=1}^{t+1} A_4(s) \implies A_5(t+1) \cap A_6(t+1)$. Note that $A_3(t+1) \cap \bigcap_{s=1}^{t+1} A_4(s)$ and $\alpha \leq 1/(4L_{\max})$ implies

$$\begin{aligned}
\tfrac{\sqrt{1-0.1}}{\sqrt{\alpha}}\,\mathrm{dist}_{t+1} &= \tfrac{\sqrt{1-0.1}}{\sqrt{\alpha}}\|\mathbf{B}_{*,\perp}^\top\hat{\mathbf{B}}_{t+1}\|_2 \\
&\leq \sigma_{\min}(\mathbf{B}_{t+1})\|\mathbf{B}_{*,\perp}^\top\hat{\mathbf{B}}_{t+1}\|_2 \\
&\leq \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \\
&\leq \left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^t \|\mathbf{B}_{*,\perp}^\top\mathbf{B}_0\|_2 \\
&\leq \tfrac{1}{\sqrt{\alpha}}\left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^t \|\mathbf{B}_{*,\perp}^\top\hat{\mathbf{B}}_0\|_2 \qquad (101) \\
&= \tfrac{1}{\sqrt{\alpha}}\left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^t \mathrm{dist}_0,
\end{aligned}$$

   where (101) follows due to initialization $\|\mathbf{B}_0\|_2 = \frac{1}{\sqrt{\alpha}}$. This implies

$$\mathrm{dist}_{t+1} \leq \tfrac{\sqrt{10}}{3}\left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^t \mathrm{dist}_0 \leq \left(1 - 0.5\beta\alpha E_0\mu_*^2\right)^t \tag{102}$$

   since $\alpha \leq 1/(4L_{\max})$ and $\mathrm{dist}_0 \leq \frac{3}{\sqrt{10}}$ by Assumption 6.

□

Next, we complete the proof of Theorem 7 by proving the following lemmas.

Lemma 16 (Exact MAML $A_1(t)$)**.** *Suppose Assumptions 1 and 6 hold, and $A_2(t)$ and $A_3(t)$ hold. Then*

$$\|\mathbf{w}_{t+1}\|_2 \leq \|\mathbf{w}_t\|_2 + 16\alpha^{3.5}L_{\max}^5 + 3\alpha^{1.5}L_{\max}^3\,\mathrm{dist}_t^2 \tag{103}$$

Proof. Using (98) and the chain rule (while noting that $a_{t,i}$ is a function of $\mathbf{w}_t$), we find that for all $i \in [n]$, the gradient of $F_{t,i}(\mathbf{B}_t, \mathbf{w}_t)$ with respect to $\mathbf{w}_t$ is:

$$\begin{aligned}\nabla_{\mathbf{w}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) &= (\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)^\top(\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)\mathbf{w}_t \\ &\quad - \mathbf{B}_t^\top(\bar{\boldsymbol{\Delta}}_t - \alpha\omega_t\mathbf{I}_d - \alpha^2 a_{t,i}\mathbf{I}_d)^2\mathbf{B}_*\mathbf{w}_{*,t,i} \\ &\quad - 2\alpha\boldsymbol{\Delta}_t\mathbf{w}_t(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})^\top\mathbf{v}_{t,i} - \alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})^\top\mathbf{v}_{t,i} \\ &= (\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)^\top(\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)\mathbf{w}_t + \mathbf{N}_{t,i} \end{aligned} \tag{104}$$

where $\mathbf{N}_{t,i} \coloneqq -\mathbf{B}_t^\top(\bar{\boldsymbol{\Delta}}_t - \alpha\omega_t\mathbf{I}_d - \alpha^2 a_{t,i}\mathbf{I}_d)^2\mathbf{B}_*\mathbf{w}_{*,t,i} - 2\alpha\boldsymbol{\Delta}_t\mathbf{w}_t(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})^\top\mathbf{v}_{t,i}$
$- \alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})^\top\mathbf{v}_{t,i}$. Thus,

$$\begin{aligned}\mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{\beta}{n}\sum_{i=1}^n \nabla_{\mathbf{w}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) \\ &= \left(\mathbf{I}_k - \frac{\beta}{n}\sum_{i=1}^n (\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)^\top(\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)\right)\mathbf{w}_t - \frac{\beta}{n}\sum_{i=1}^n \mathbf{N}_{t,i} \end{aligned} \tag{105}$$

which implies that

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2 &\le \left\|\mathbf{I}_k - \frac{\beta}{n}\sum_{i=1}^n (\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)^\top(\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)\right\|_2 \|\mathbf{w}_t\|_2 \\ &\quad + \beta\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{N}_{t,i}\right\|_2 \\ &\le \|\mathbf{w}_t\|_2 + \beta\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{N}_{t,i}\right\|_2 \end{aligned} \tag{106}$$

where (106) follows since $\frac{1}{n}\sum_{i=1}^n (\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)^\top(\mathbf{B}_t\boldsymbol{\Delta}_t - \alpha\omega_t\mathbf{B}_t - \alpha^2 a_{t,i}\mathbf{B}_t)$ is PSD and β is sufficiently small. Next, we upper bound $\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{N}_{t,i}\right\|_2$, and to do so, we first use the triangle inequality to write

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{N}_{t,i}\right\|_2 &\le \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}_t^\top(\bar{\boldsymbol{\Delta}}_t - \alpha\omega_t\mathbf{I}_d - \alpha^2 a_{t,i}\mathbf{I}_d)^2\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2 \\ &\quad + 2\left\|\frac{1}{n}\sum_{i=1}^n \alpha\boldsymbol{\Delta}_t\mathbf{w}_t(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})^\top\mathbf{v}_{t,i}\right\|_2 \\ &\quad + \left\|\frac{1}{n}\sum_{i=1}^n \alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})^\top\mathbf{v}_{t,i}\right\|_2 \end{aligned} \tag{107}$$

We will bound each of the three terms above shortly. First, note that $A_4(t)$ and $\alpha \le 1/(4L_{\max})$ implies

$$\frac{15/16}{\alpha} \le \frac{1 - \alpha^2 L_{\max}^2}{\alpha} \le \sigma_{\min}^2(\mathbf{B}_t) \le \sigma_{\max}^2(\mathbf{B}_t) \le \frac{1 + \alpha^2 L_{\max}^2}{\alpha} \le \frac{17/16}{\alpha}. \tag{108}$$

In turn, this implies that $\|\mathbf{B}_t\|_2^3 \le 1.1$. Let $c \coloneqq 1.1$. Also, note that $\mathbf{B}_t^\top\bar{\boldsymbol{\Delta}}_t = \boldsymbol{\Delta}_t\mathbf{B}_t^\top$ and

$$\begin{aligned}\|\mathbf{B}_*^\top\bar{\boldsymbol{\Delta}}_t\mathbf{B}_*\|_2 &= \|\mathbf{B}_*^\top(\mathbf{I}_d - \alpha\mathbf{B}_t\mathbf{B}_t^\top)\mathbf{B}_*\|_2 \\ &\le \|\mathbf{B}_*^\top(\mathbf{I}_d - \hat{\mathbf{B}}_t\hat{\mathbf{B}}_t^\top)\mathbf{B}_*\|_2 + \|\mathbf{B}_*^\top(\hat{\mathbf{B}}_t\hat{\mathbf{B}}_t^\top - \alpha\mathbf{B}_t\mathbf{B}_t^\top)\mathbf{B}_*\|_2 \\ &\le \|\mathbf{B}_*^\top(\mathbf{I}_d - \hat{\mathbf{B}}_t\hat{\mathbf{B}}_t^\top)\|\|(\mathbf{I}_d - \hat{\mathbf{B}}_t\hat{\mathbf{B}}_t^\top)\mathbf{B}_*\|_2 + \|\hat{\mathbf{B}}_t\hat{\mathbf{B}}_t^\top - \alpha\mathbf{B}_t\mathbf{B}_t^\top\|_2 \\ &= \mathrm{dist}_t^2 + \|\hat{\mathbf{B}}_t(\mathbf{I}_k - \alpha\mathbf{R}_t\mathbf{R}_t^\top)\hat{\mathbf{B}}_t^\top\|_2 \\ &\le \mathrm{dist}_t^2 + \|\boldsymbol{\Delta}_t\|_2 \end{aligned} \tag{109}$$

where $\mathbf{R}_t \in \mathbb{R}^{k\times k}$ is the upper triangular matrix resulting from the QR decomposition of $\mathbf{B}_t$. We will use these observations along with inductive hypotheses $A_2(t)$ and $A_3(t)$ and the Cauchy-Schwarz and triangle inequalities to separately bound each of the terms from (107) as follows. Let $c_2 \coloneqq E_0/20$. Then we have:

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_t^\top(\bar{\mathbf{\Delta}}_t-(\alpha\omega_t+\alpha^2 a_{t,i})\mathbf{I}_d)^2\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2\\
&\quad\le\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Delta}_t^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2+2\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha\omega_t\mathbf{\Delta}_t\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2\\
&\qquad+2\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2 a_{t,i}\mathbf{\Delta}_t\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2+2\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^3\omega_t a_{t,i}\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2\\
&\qquad+\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\omega_t^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2+\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^4 a_{t,i}^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\right\|_2\\
&\quad\le c\alpha^{3.5}L_{\max}^4\eta_*+2c\alpha^{4.5}L_{\max}^4\eta_*\|\mathbf{w}_t\|_2^2+2c\alpha^3L_{\max}^2L_*^2\|\mathbf{w}_t\|_2\\
&\qquad+2c\alpha^4L_{\max}^2L_*^2\|\mathbf{w}_t\|_2^3+c\alpha^{3.5}L_{\max}^4\eta_*\|\mathbf{w}_t\|_2^4+c\alpha^{2.5}L_{\max}^3\|\mathbf{w}_t\|_2^2\\
&\quad\le c\alpha^{3.5}L_{\max}^4\eta_*+2cc_2^2\alpha^{5.5}L_{\max}^4\eta_*\mu_*^2\kappa_{*,\max}^{-2}+2cc_2\alpha^{3.5}L_{\max}^2L_*^2\mu_*\kappa_{*,\max}^{-1}\\
&\qquad+2cc_2^3\alpha^{5.5}L_{\max}^2L_*^2\mu_*^3\kappa_{*,\max}^{-3}+cc_2^4\alpha^{5.5}L_{\max}^4\eta_*\mu_*^4\kappa_{*,\max}^{-4}+cc_2^2\alpha^{3.5}L_{\max}^3\mu_*^2\kappa_{*,\max}^{-2}\\
&\quad\le 4c\alpha^{3.5}L_{\max}^4(\eta_*+\mu_*)
\end{aligned}
\tag{110}
$$

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha\mathbf{\Delta}_t\mathbf{w}_t(\mathbf{B}_t\mathbf{w}_t-\mathbf{B}_*\mathbf{w}_{*,t,i})^\top\mathbf{v}_{t,i}\right\|_2\\
&\quad\le\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha\mathbf{\Delta}_t\mathbf{w}_t\mathbf{w}_t^\top\mathbf{B}_t^\top(\bar{\mathbf{\Delta}}_t-\alpha\omega_t\mathbf{I}_d-\alpha^2a_{t,i}\mathbf{I}_d)(\mathbf{B}_t\mathbf{w}_t-\mathbf{B}_*\mathbf{w}_{*,t,i})\right\|_2\\
&\qquad+\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha\mathbf{\Delta}_t\mathbf{w}_t\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top(\bar{\mathbf{\Delta}}_t-\alpha\omega_t\mathbf{I}_d-\alpha^2a_{t,i}\mathbf{I}_d)(\mathbf{B}_t\mathbf{w}_t-\mathbf{B}_*\mathbf{w}_{*,t,i})\right\|_2\\
&\quad\le c\alpha^4L_{\max}^4\|\mathbf{w}_t\|_2^3+2c\alpha^{4.5}L_{\max}^4\eta_*\|\mathbf{w}_t\|_2^2+c\alpha^5L_{\max}^4\|\mathbf{w}_t\|_2^5+2c\alpha^{5.5}L_{\max}^4\eta_*\|\mathbf{w}_t\|_2^4\\
&\qquad+c\alpha^{3.5}L_{\max}^2\eta_*\|\mathbf{w}_t\|_2^4+2c\alpha^4L_{\max}^2L_*^2\|\mathbf{w}_t\|_2^3+c\alpha^5L_{\max}^6\|\mathbf{w}_t\|_2^3\\
&\qquad+c\alpha^{4.5}L_{\max}^5\|\mathbf{w}_t\|_2^2+c\alpha^5L_{\max}^6\|\mathbf{w}_t\|_2+c\alpha^3L_{\max}^4\operatorname{dist}_t^2\|\mathbf{w}_t\|_2\\
&\quad\le cc_2^3\alpha^{5.5}L_{\max}^4\mu_*^3\kappa_{*,\max}^{-3}+2cc_2^2\alpha^{5.5}L_{\max}^4\eta_*\mu_*^2\kappa_{*,\max}^{-2}+cc_2^5\alpha^{7.5}L_{\max}^4\mu_*^5\kappa_{*,\max}^{-5}\\
&\qquad+2cc_2^4\alpha^{7.5}L_{\max}^4\eta_*\mu_*^4\kappa_{*,\max}^{-4}+cc_2^4\alpha^{5.5}L_{\max}^2\eta_*\mu_*^4\kappa_{*,\max}^{-4}+2cc_2^3\alpha^{5.5}L_{\max}^2L_*^2\mu_*^3\kappa_{*,\max}^{-3}\\
&\qquad+cc_2^3\alpha^{7.5}L_{\max}^6\mu_*^3\kappa_{*,\max}^{-3}+cc_2^2\alpha^{5.5}L_{\max}^5\mu_*^2\kappa_{*,\max}^{-2}+cc_2\alpha^{5.5}L_{\max}^6\mu_*\kappa_{*,\max}^{-1}\\
&\qquad+cc_2\alpha^{3.5}L_{\max}^4\mu_*\kappa_{*,\max}^{-1}\operatorname{dist}_t^2\\
&\quad\le 9cc_2\alpha^{5.5}L_{\max}^5\mu_*^2+cc_2\alpha^{3.5}L_{\max}^3\mu_*^2\operatorname{dist}_t^2
\end{aligned}
\tag{111}
$$

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}(\mathbf{B}_t\mathbf{w}_t-\mathbf{B}_*\mathbf{w}_{*,t,i})^\top\mathbf{v}_{t,i}\right\|_2 \\
&\quad\le\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\mathbf{B}_t^\top(\bar{\mathbf{\Delta}}_t-\alpha\omega_t\mathbf{I}_d-\alpha^2a_{t,i}\mathbf{I}_d)(\mathbf{B}_t\mathbf{w}_t-\mathbf{B}_*\mathbf{w}_{*,t,i})\right\|_2 \\
&\qquad+\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top(\bar{\mathbf{\Delta}}_t-\alpha\omega_t\mathbf{I}_d-\alpha^2a_{t,i}\mathbf{I}_d)(\mathbf{B}_t\mathbf{w}_t-\mathbf{B}_*\mathbf{w}_{*,t,i})\right\|_2 \\
&\quad\le c\alpha^{2.5}L_{\max}^2\eta_*\|\mathbf{w}_t\|_2^2+c\alpha^3L_{\max}^2L_*^2\|\mathbf{w}_t\|_2+c\alpha^{3.5}L_{\max}^2\eta_*\|\mathbf{w}_t\|_2^4 \\
&\qquad+c\alpha^{2.5}L_{\max}^3\|\mathbf{w}_t\|_2^2+c\alpha^4L_{\max}^2L_*^2\|\mathbf{w}_t\|_2^3+c\alpha^2L_*^2\|\mathbf{w}_t\|_2^3 \\
&\qquad+c\alpha^3L_{\max}^2L_*^2\|\mathbf{w}_t\|_2+c\alpha^{2.5}L_{\max}^3\|\mathbf{w}_t\|_2^2 \\
&\qquad+c\alpha^{3.5}L_{\max}^2L_{\max}^3+c\alpha^{1.5}L_{\max}^3\operatorname{dist}_t^2+c\alpha^3L_{\max}^4\|\mathbf{w}_t\|_2 \\
&\qquad+c\alpha^{4.5}L_{\max}^5\|\mathbf{w}_t\|_2^2+c\alpha^4L_{\max}^2L_*^2\eta_*\|\mathbf{w}_t\|_2^3 \\
&\quad\le cc_2^2\alpha^{3.5}L_{\max}^2\eta_*\mu_*^2\kappa_{*,\max}^{-2}+cc_2\alpha^{3.5}L_{\max}^2L_*^2\mu_*\kappa_{*,\max}^{-1}+cc_2^4\alpha^{5.5}L_{\max}^2\eta_*\mu_*^4\kappa_{*,\max}^{-4} \\
&\qquad+cc_2^2\alpha^{3.5}L_{\max}^3\mu_*^2\kappa_{*,\max}^{-2}+cc_2^3\alpha^{5.5}L_{\max}^2L_*^2\mu_*^3\kappa_{*,\max}^{-3}+cc_2^3\alpha^{3.5}L_*^2\mu_*^3\kappa_{*,\max}^{-3} \\
&\qquad+cc_2\alpha^{3.5}L_{\max}^2L_*^2\mu_*\kappa_{*,\max}^{-1}+cc_2^2\alpha^{3.5}L_{\max}^3\mu_*^2\kappa_{*,\max}^{-2} \\
&\qquad+c\alpha^{3.5}L_{\max}^2L_{\max}^3+c\alpha^{1.5}L_{\max}^3\operatorname{dist}_t^2+cc_2\alpha^{3.5}L_{\max}^4\mu_*\kappa_{*,\max}^{-1} \\
&\qquad+cc_2^2\alpha^{5.5}L_{\max}^5\mu_*^2\kappa_{*,\max}^{-2}+c\alpha^{5.5}L_{\max}^2L_*^2\eta_*\mu_*^3\kappa_{*,\max}^{-3} \\
&\quad\le 9c\alpha^{3.5}L_{\max}^5+c\alpha^{1.5}L_{\max}^3\operatorname{dist}_t^2
\end{aligned}
\tag{112}
$$

where we have used $c_2 = E_0/20$ and $\alpha \le 1/(4L_{\max})$ to reduce terms. Next, combining the above bounds with (107) yields:

$$
\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{N}_{t,i}\right\|_2 \le 14c\alpha^{3.5}L_{\max}^5+2c\alpha^{1.5}L_{\max}^3\operatorname{dist}_t^2 \tag{113}
$$

Applying $c = 1.1$ yields the result. □

Lemma 17 (Exact MAML $A_2(t+1)$)**.** *Suppose the conditions of Theorem 7 are satisfied, and* $\bigcap_{s=1}^{t}\{A_1(s)\cap A_5(s)\}$. *Then*

$$
\|\mathbf{w}_{t+1}\|_2 \le \tfrac{E_0}{20}\sqrt{\alpha}\mu_*. \tag{114}
$$

Proof. By inductive hypotheses $A_1(1),\dots,A_1(t)$, we have $\|\mathbf{w}_{s+1}\|_2 \le \|\mathbf{w}_s\|_2 + 16\beta\alpha^{3.5}L_{\max}^5 + 3\beta\alpha^{1.5}L_{\max}^3\operatorname{dist}_s^2$ for all $s\in[t]$, so we can invoke Lemma 3 with $\xi_{1,s}=0\ \forall s\in[t]$ and $\xi_{2,s} = 16\beta\alpha^{3.5}L_{\max}^5+3\beta\alpha^{1.5}L_{\max}^3\operatorname{dist}_s^2$. This results in

$$
\|\mathbf{w}_{t+1}\|_2 \le \sum_{s=1}^{t}16\beta\alpha^{3.5}L_{\max}^5+3\beta\alpha^{1.5}L_{\max}^3\operatorname{dist}_s^2 \tag{115}
$$

Next, we invoke inductive hypotheses $A_6(1),\dots,A_6(t)$ to obtain $\operatorname{dist}_s^2 \le \frac{10}{9}(1-0.5\beta\alpha E_0\mu_*^2)^{2(s-1)}\operatorname{dist}_0^2$ for all $s\in[t]$. Therefore

$$
\begin{aligned}
\|\mathbf{w}_{t+1}\|_2 &\le \sum_{s=1}^{t}16\beta\alpha^{3.5}L_{\max}^5+3\beta\alpha^{1.5}L_{\max}^3\gamma^{2(s-1)} \\
&\le 16\beta\alpha^{3.5}L_{\max}^5t+3\beta\alpha^{1.5}L_{\max}^3\sum_{s=1}^{t}\tfrac{10}{9}(1-0.5\beta\alpha E_0\mu_*^2)^{2(s-1)}\operatorname{dist}_0^2 \\
&\le 16\beta\alpha^{3.5}L_{\max}^5t+\tfrac{10}{3}\beta\alpha^{1.5}L_{\max}^3\tfrac{\operatorname{dist}_0^2}{0.5\beta\alpha E_0\mu_*^2} \qquad (116)\\
&\le 16\beta\alpha^{3.5}L_{\max}^5t+\tfrac{20}{3}\sqrt{\alpha}L_{\max}^3\tfrac{\operatorname{dist}_0^2}{E_0\mu_*^2} \\
&\le \tfrac{E_0}{20}\sqrt{\alpha}\mu_* \qquad (117)
\end{aligned}
$$

where (116) is due to the sum of a geometric series and (117) follows since $\beta \leq \frac{\alpha}{10\kappa_*^4} \leq \frac{E_0\mu_*}{640\alpha^3 L_{\max}^5 T}$ (as α is sufficiently small) and the initial representation satisfies

$$\begin{aligned}\tfrac{20}{3}\sqrt{\alpha}L_{\max}^3\frac{\mathrm{dist}_0^2}{E_0\mu_*^2} &\leq \tfrac{E_0}{20}\sqrt{\alpha}\mu_*/2 \\ \iff 0 &\leq \mu_*^3 - (\tfrac{800}{3}L_{\max}^3 + 2\mu_*^3)\mathrm{dist}_0^2 + \mu_*^3\,\mathrm{dist}_0^4\end{aligned}$$

which is implied by

$$\mathrm{dist}_0 \leq \tfrac{1}{17}\kappa_{*,\max}^{-1.5} \tag{118}$$

□

Lemma 18 (Exact MAML $A_3(t+1)$)**.** *Suppose the conditions of Theorem 7 are satisfied and $A_2(t)$, $A_3(t)$ and $A_5(t)$ hold. Then $A_3(t+1)$ holds, namely*

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq \alpha^2 L_{\max}^2 \tag{119}$$

Proof. According to Lemma 4, we can control $\mathbf{\Delta}_{t+1}$ by controlling $\mathbf{G}_t$, recalling that $\mathbf{G}_t = \frac{1}{\beta}(\mathbf{B}_t - \mathbf{B}_{t+1}) \in \mathbb{R}^{d\times k}$ is the outer loop gradient with respect to the representation at time t. Before studying $\mathbf{G}_t$, we must compute the outer loop gradient with respect to the representation for task i. Again we use the fact that $F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) = \frac{1}{2}\|\mathbf{v}_{t,i}\|_2^2$ and apply the chain rule to obtain:

$$\begin{aligned}\nabla_{\mathbf{B}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t) &= \mathbf{v}_{t,i}\mathbf{w}_t^\top \mathbf{\Delta}_t\mathbf{\Lambda}_t - \alpha\mathbf{B}_t(\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t\mathbf{\Lambda}_t + \mathbf{\Lambda}_t\mathbf{B}_t^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top) \\ &\quad + \alpha(\mathbf{v}_{t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top + \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{v}_{t,i}^\top)\mathbf{B}_t\mathbf{\Lambda}_t \\ &\quad - 2\alpha^2(\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{v}_{t,i})\mathbf{B}_t\mathbf{w}_t\mathbf{w}_t^\top + \alpha^2\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{v}_{t,i}^\top\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_t^\top\end{aligned} \tag{120}$$

Note that $\mathbf{G}_t = \frac{1}{n}\sum_{i=1}^n \nabla_{\mathbf{B}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t)$. We aim to write $\mathbf{G}_t$ as $\mathbf{G}_t = -\bar{\mathbf{\Delta}}_t\mathbf{S}_t\mathbf{B}_t - \mathbf{S}_t\mathbf{B}_t\mathbf{\Delta}_t + \mathbf{N}_t$ for some positive definite $\mathbf{S}$ so that we can apply Lemma 4. It turns out that of the five terms in (120), the only one with 'sub'-terms that contribute to $\mathbf{S}_t$ is the third term. To see this, note that

$$\begin{aligned}\alpha(\mathbf{v}_{t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top + \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{v}_{t,i}^\top)\mathbf{B}_t\mathbf{\Lambda}_t &= -\alpha(\bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t + \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\mathbf{\Delta}_t) \\ &\quad + \alpha((\mathbf{v}_{t,i} + \bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i})\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top \\ &\quad + \mathbf{B}_*\mathbf{w}_{*,t,i}(\mathbf{v}_{t,i} + \mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\bar{\mathbf{\Delta}}_t))\mathbf{B}_t\mathbf{\Lambda}_t \\ &\quad + \alpha^2\bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t + \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\mathbf{\Delta}_t)\mathbf{w}_t\mathbf{w}_t^\top \\ &= -(\bar{\mathbf{\Delta}}_t\mathbf{S}_t\mathbf{B}_t + \mathbf{S}_t\mathbf{B}_t\mathbf{\Delta}_t) \\ &\quad + \alpha((\mathbf{v}_{t,i} + \bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i})\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top \\ &\quad + \mathbf{B}_*\mathbf{w}_{*,t,i}(\mathbf{v}_{t,i} + \mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\bar{\mathbf{\Delta}}_t))\mathbf{B}_t\mathbf{\Lambda}_t \\ &\quad + \alpha^2\bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t + \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\mathbf{\Delta}_t)\mathbf{w}_t\mathbf{w}_t^\top.\end{aligned}$$

where $\mathbf{S}_t \coloneqq \alpha\frac{1}{n}\sum_{i=1}^n \mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top$. Thus we can write

$$\mathbf{G}_t = -\bar{\mathbf{\Delta}}_t\mathbf{S}_t\mathbf{B}_t - \mathbf{S}_t\mathbf{B}_t\mathbf{\Delta}_t + \mathbf{N}_t \tag{121}$$

where

$$\begin{aligned}\mathbf{N}_t \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbf{v}_{t,i}\mathbf{w}_t^\top\mathbf{\Delta}_t\mathbf{\Lambda}_t - \frac{1}{n}\sum_{i=1}^n \alpha^2\mathbf{B}_t\mathbf{w}_t\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top + \frac{1}{n}\sum_{i=1}^n \alpha^2\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top \\ - \frac{1}{n}\sum_{i=1}^n \alpha\mathbf{B}_t(\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t\mathbf{\Lambda}_t + \mathbf{\Lambda}_t\mathbf{B}_t^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top) + \frac{1}{n}\sum_{i=1}^n \alpha(\mathbf{v}_{t,i} + \bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i})\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\mathbf{\Lambda}_t \\ + \frac{1}{n}\sum_{i=1}^n \alpha\bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\alpha\mathbf{w}_t\mathbf{w}_t^\top\end{aligned} \tag{122}$$

Note that $\|\mathbf{\Delta}_t\|_2 \leq \frac{1}{10}$ due to $A_3(t)$ and choice of $\alpha \leq \frac{1}{4L_{\max}}$. Therefore, Lemma 1 implies that $\sigma_{\min}(\mathbf{B}_t^\top \mathbf{S}_t \mathbf{B}_t) \geq E_0 \mu_*^2$ where $E_0 = 1 - \frac{1}{10} - \mathrm{dist}_0^2$. Thus by Lemma 4 with $\chi = 1$, we have

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq \|\mathbf{\Delta}_t\|_2(1 - 2\beta\alpha E_0 \mu_*^2) + 2\beta\alpha\|\mathbf{B}_t^\top \mathbf{N}_t\|_2 + \beta^2\alpha\|\mathbf{G}_t\|_2^2 \tag{123}$$

So it remains to control $\|\mathbf{B}_t^\top \mathbf{N}_t\|_2$ and $\|\mathbf{G}_t\|_2^2$. First we deal with $\|\mathbf{B}_t^\top \mathbf{N}_t\|_2$ by upper bounding the norm of $\mathbf{B}_t^\top$ times each of the six terms in (122). As before, we use $c = 1.1$ as an absolute constant that satisfies $\sigma_{\max}^3(\mathbf{B}_t) \leq c/\alpha^{1.5}$. We have

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{B}_t^\top \mathbf{v}_{t,i} \mathbf{w}_t^\top \mathbf{\Delta}_t \mathbf{\Lambda}_t \right\|_2 \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{B}_t^\top (\bar{\mathbf{\Delta}}_t - (\alpha\omega_t + \alpha^2 a_{t,i})\mathbf{I}_d)(\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_t^\top \mathbf{\Delta}_t \mathbf{\Lambda}_t \right\|_2 \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{\Delta}_t \mathbf{B}_t^\top (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_t^\top \mathbf{\Delta}_t \mathbf{\Lambda}_t \right\|_2 \\
&\quad + \left\| \frac{1}{n} \sum_{i=1}^n (\alpha\omega_t + \alpha^2 a_{t,i}) \mathbf{B}_t^\top (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_t^\top \mathbf{\Delta}_t \mathbf{\Lambda}_t \right\|_2 \\
&\leq \tfrac{c}{\alpha}\|\mathbf{\Delta}_t\|_2^2 \|\mathbf{w}_t\|_2^2 + \tfrac{c}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2^2 \|\mathbf{w}_t\|_2 \eta_* + c\|\mathbf{\Delta}_t\|_2^2 \|\mathbf{w}_t\|_2^4 \\
&\quad + c\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2^3 \eta_* + c\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2^2 \|\mathbf{w}_t\|_2^3 \eta_* + c\alpha\|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2^2 L_*^2
\end{aligned}$$

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \alpha^2 \mathbf{B}_t^\top \mathbf{B}_t \mathbf{w}_t \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{v}_{t,i} \mathbf{w}_t^\top \right\|_2 &\leq c \left\| \frac{1}{n} \sum_{i=1}^n \alpha \mathbf{w}_t \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top (\bar{\mathbf{\Delta}}_t - (\alpha\omega_t + \alpha^2 a_{t,i})\mathbf{I}_d) \right. \\
&\qquad \left. \times (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_t^\top \right\|_2 \\
&\leq c\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2^3 \eta_* + c\alpha(\|\mathbf{\Delta}_t\|_2 + \mathrm{dist}_t^2)\|\mathbf{w}_t\|_2^2 L_{\max}^2 \\
&\quad + c\alpha^{1.5}\|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2^5 \eta_* + c\alpha^2 \|\mathbf{\Delta}_t\|_2 \|\mathbf{w}\|_2^4 L_{\max}^2 \\
&\quad + c\alpha^2 \|\mathbf{w}_t\|_2^4 L_*^2 + c\alpha^{2.5}\|\mathbf{w}_t\|_2^3 L_{\max}^3
\end{aligned}$$

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \alpha^2 \mathbf{B}_t^\top \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{v}_{t,i} \mathbf{w}_t^\top \right\|_2 &\leq c\alpha\|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2^2 L_*^2 + c\alpha^{1.5}(\|\mathbf{\Delta}_t\|_2 + \mathrm{dist}_t^2)\|\mathbf{w}_t\|_2 L_{\max}^3 \\
&\quad + c\alpha^2 \|\mathbf{\Delta}_t\|_2 \|\mathbf{w}_t\|_2^4 L_*^2 + c\alpha^{2.5}\|\mathbf{\Delta}_t\|_2 \|\mathbf{w}\|_2^3 L_{\max}^3 \\
&\quad + c\alpha^{2.5}\|\mathbf{w}_t\|_2^3 L_{\max}^3 + c\alpha^3 \|\mathbf{w}_t\|_2^2 L_{\max}^4
\end{aligned}$$

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n \alpha \mathbf{B}_t^\top \mathbf{B}_t(\mathbf{w}_t\mathbf{v}_{t,i}^\top \mathbf{B}_t\mathbf{\Lambda}_t + \mathbf{\Lambda}_t\mathbf{B}_t^\top \mathbf{v}_{t,i}\mathbf{w}_t^\top)\right\|_2 &\le 2\left\|\frac{1}{n}\sum_{i=1}^n \alpha \mathbf{B}_t^\top \mathbf{B}_t\mathbf{w}_t\mathbf{v}_{t,i}^\top \mathbf{B}_t\mathbf{\Lambda}_t\right\|_2 \\
&\le 2c\Bigg\|\frac{1}{n}\sum_{i=1}^n \mathbf{w}_t(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})^\top \\
&\qquad \times (\bar{\mathbf{\Delta}}_t - (\alpha\omega_t + \alpha^2 a_{t,i})\mathbf{I}_d)\mathbf{B}_t\mathbf{\Lambda}_t\Bigg\|_2 \\
&\le \tfrac{2c}{\alpha}\|\mathbf{\Delta}_t\|\|\mathbf{w}_t\|_2^2 + \tfrac{2c}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* \\
&\quad + 2c\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2^4 + 2c\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2^3\eta_* \\
&\quad + 2c\sqrt{\alpha}\|\mathbf{w}_t\|_2^3\eta_* + 2c\alpha\|\mathbf{w}_t\|_2^2 L_*^2 \\
\left\|\frac{1}{n}\sum_{i=1}^n \alpha \mathbf{B}_t^\top(\mathbf{v}_{t,i} + \bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i})\mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{B}_t\mathbf{\Lambda}_t\right\|_2 &\le \tfrac{c}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* + c\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2^3\eta_* \\
&\quad + c\alpha\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2^2 L_*^2 + c\alpha\|\mathbf{w}_t\|_2^2 L_*^2 \\
&\quad + c\alpha^{3/2}\|\mathbf{w}_t\|_2 L_{\max}^3 \\
\left\|\frac{1}{n}\sum_{i=1}^n \alpha \mathbf{B}_t^\top \bar{\mathbf{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{B}_t\alpha\mathbf{w}_t\mathbf{w}_t^\top\right\|_2 &\le \alpha\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2^2 L_*^2
\end{aligned}$$

Let $c_2 \coloneqq E_0/20$. We can combine the above bounds and use inductive hypotheses $A_2(t)$ and $A_3(t)$ to obtain the following bound on $\|\mathbf{B}_t^\top \mathbf{N}_t\|_2$:

$$\begin{aligned}
\|\mathbf{B}_t^\top \mathbf{N}_t\|_2 &\le 2cc_2^2\alpha^4 L_*^4\mu_*^2\kappa_{*,\max}^{-2} + cc_2\alpha^4 L_*^4\eta_*\mu_*\kappa_{*,\max}^{-1} + 4cc_2^4\alpha^6 L^4\mu_*^4\kappa_{*,\max}^{-4} + 4cc_2^3\alpha^4 L_{\max}^2\eta_*\mu_*^3\kappa_{*,\max}^{-3} \\
&\quad + cc_2^3\alpha^6 L_*^4\eta_*\mu_*^3\kappa_{*,\max}^{-3} + 3cc_2^2\alpha^4 L_*^2 L_{\max}^3\mu_*\kappa_{*,\max}^{-1} + 3cc_2^2\alpha^2 L_{\max}^3\mu_*\kappa_{*,\max}^{-1}\operatorname{dist}_t^2 \\
&\quad + cc_2^5\alpha^6 L_*^2\eta_*\mu_*^5\kappa_{*,\max}^{-5} + 2cc_2^3\alpha^4\mu_*^3 L_{\max}^3\kappa_{*,\max}^{-3} + 2cc_2^3\alpha^6\mu_*^3 L_{\max}^3 L_*^2\kappa_{*,\max}^{-3} \\
&\quad + cc_2^2\alpha^4 L_{\max}^4\mu_*^2\kappa_{*,\max}^{-2} + 2cc_2^2\alpha^2 L_{\max}^2\mu_*^2\kappa_{*,\max}^{-2} \\
&\quad + 3cc_2\alpha^2 L_{\max}^2\eta_*\mu_*\kappa_{*,\max}^{-1} + 2cc_2^4\alpha^4 L_{\max}^2\mu_*^4\kappa_{*,\max}^{-4} + 2cc_2^3\alpha^2\eta_*\mu_*^3\kappa_{*,\max}^{-3} \\
&\quad + 4cc_2\alpha^2 L_{\max}^3\mu_*\kappa_{*,\max}^{-1} + 2cc_2\alpha^4 L_{\max}^2 L_*^2\mu_*^2\kappa_{*,\max}^{-2} \\
&\le 21cc_2\alpha^4(L_*^4\eta_* + L_{\max}^3 L_*^2)\mu_*\kappa_{*,\max}^{-1} + 3cc_2^2\alpha^2 L_{\max}^3\mu_*\kappa_{*,\max}^{-1}\operatorname{dist}_t^2 \\
&\quad + 4cc_2\alpha^2(L_{\max}^3 + L_{\max}^2(\eta_* + \mu_*))\mu_*\kappa_{*,\max}^{-1} \\
&\le 10cc_2\alpha^2 L_{\max}^2\mu_*^2 + 3cc_2^2\alpha^2 L_{\max}^2\mu_*^2\operatorname{dist}_t^2 \\
&\le 13cc_2\alpha^2 L_{\max}^2\mu_*^2 \\
&\le 15c_2\alpha^2 L_{\max}^2\mu_*^2
\end{aligned} \tag{124}$$

using that $\alpha \le 1/L_*$, $c = 1.1$ and combining like terms. We have not optimized constants. Next we bound $\|\mathbf{G}_t\|_2^2$. First, by (121) and the triangle and Cauchy-Schwarz inequalities,

$$\begin{aligned}
\|\mathbf{G}_t\|_2 &\le \|\bar{\mathbf{\Delta}}_t\mathbf{S}_t\|_2\|\mathbf{B}_t\|_2 + \|\mathbf{S}_t\|_2\|\mathbf{B}_t\|_2\|\mathbf{\Delta}_t\|_2 + \|\mathbf{N}_t\|_2 \\
&\le c\sqrt{\alpha}(\operatorname{dist}_t + 2\|\mathbf{\Delta}_t\|_2)L_*^2 + \|\mathbf{N}_t\|_2.
\end{aligned} \tag{125}$$

We have already bounded $\|\mathbf{B}_t^\top \mathbf{N}_t\|$ by separately bounding $\mathbf{B}_t^\top$ times each of the six terms in $\mathbf{N}_t$. We obtain a similar bound on $\|\mathbf{N}_t\|_2$ by separately considering each of the six terms in $\mathbf{N}_t$ (see equation (122)). Of these terms, all but the first and last can be easily bounded by multiplying our previous bounds by $\sqrt{\alpha}$ (to account for no $\mathbf{B}_t$). The other two terms are more complicated because we have previously made the reduction $\|\mathbf{B}_t^\top \bar{\mathbf{\Delta}}_t\|_2 = \|\mathbf{\Delta}_t\mathbf{B}_t^\top\|_2 \le \frac{c}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2$, but now that there is no $\mathbf{B}_t^\top$ to multiply with $\bar{\mathbf{\Delta}}_t$, we must control $\bar{\mathbf{\Delta}}_t$ via $\|\bar{\mathbf{\Delta}}_t\mathbf{B}_*\|_2 \le \|\mathbf{\Delta}_t\|_2 + \operatorname{dist}_t$. Specifically, for the easy four

terms we have

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_t\mathbf{w}_t\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top\right\|_2 &\le c\alpha\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^3\eta_* + c\alpha^{1.5}(\|\boldsymbol{\Delta}_t\|_2+\mathrm{dist}_t^2)\|\mathbf{w}_t\|_2^2L_{\max}^2\\
&\quad + c\alpha^2\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^5\eta_* + c\alpha^{2.5}\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}\|_2^4L_{\max}^2\\
&\quad + c\alpha^{2.5}\|\mathbf{w}_t\|_2^4L_*^2 + c\alpha^3\|\mathbf{w}_t\|_2^3L_{\max}^3\\
\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha^2\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top\right\|_2 &\le c\alpha^{1.5}\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^2L_*^2 + c\alpha^2(\|\boldsymbol{\Delta}_t\|_2+\mathrm{dist}_t^2)\|\mathbf{w}_t\|_2L_{\max}^3\\
&\quad + c\alpha^{2.5}\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^4L_*^2 + c\alpha^3\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}\|_2^3L_{\max}^3\\
&\quad + c\alpha^3\|\mathbf{w}_t\|_2^3L_{\max}^3 + c\alpha^{3.5}\|\mathbf{w}_t\|_2^2L_{\max}^4
\end{aligned}$$

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha\mathbf{B}_t(\mathbf{w}_t\mathbf{v}_{t,i}^\top\mathbf{B}_t\boldsymbol{\Lambda}_t + \boldsymbol{\Lambda}_t\mathbf{B}_t^\top\mathbf{v}_{t,i}\mathbf{w}_t^\top)\right\|_2 &\le \tfrac{2c}{\sqrt{\alpha}}\|\boldsymbol{\Delta}_t\|\|\mathbf{w}_t\|_2^2 + 2c\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_*\\
&\quad + 2c\sqrt{\alpha}\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^4 + 2c\alpha\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^3\eta_*\\
&\quad + 2c\alpha\|\mathbf{w}_t\|_2^3\eta_* + 2c\alpha^{1.5}\|\mathbf{w}_t\|_2^2L_*^2\\
\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha(\mathbf{v}_{t,i}+\bar{\boldsymbol{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i})\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\boldsymbol{\Lambda}_t\right\|_2 &\le c\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* + c\alpha\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^3\eta_*\\
&\quad + c\alpha^{1.5}\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^2L_*^2 + c\alpha^{1.5}\|\mathbf{w}_t\|_2^2L_*^2\\
&\quad + c\alpha^2\|\mathbf{w}_t\|_2L_{\max}^3
\end{aligned}$$

and for the first and last term from (122), we have

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_{t,i}\mathbf{w}_t^\top\boldsymbol{\Delta}_t\boldsymbol{\Lambda}_t\right\|_2 &\le \tfrac{c}{\sqrt{\alpha}}\|\boldsymbol{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^2 + c\|\boldsymbol{\Delta}_t\|_2(\|\boldsymbol{\Delta}_t\|_2+\mathrm{dist}_t)\|\mathbf{w}_t\|_2\eta_*\\
&\quad + c\sqrt{\alpha}\|\boldsymbol{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^4 + c\alpha\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^3\eta_*\\
&\quad + c\alpha\|\boldsymbol{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^3\eta_* + c\alpha^{1.5}\|\boldsymbol{\Delta}_t\|_2\|\mathbf{w}_t\|_2^2L_*^2\\
\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha\bar{\boldsymbol{\Delta}}_t\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{*,t,i}^\top\mathbf{B}_*^\top\mathbf{B}_t\alpha\mathbf{w}_t\mathbf{w}_t^\top\right\|_2 &\le \alpha^{1.5}(\|\boldsymbol{\Delta}_t\|_2+\mathrm{dist}_t)\|\mathbf{w}_t\|_2^2L_*^2
\end{aligned}$$

Combining these bounds and applying inductive hypotheses $A_2(t)$ and $A_3(t)$ yields

$$\begin{aligned}
\|\mathbf{N}_t\|_2 &\le cc_2^3\alpha^{4.5}L_{\max}^2\eta_*\mu_*^3\kappa_{*,\max}^{-3} + cc_2^2\alpha^{4.5}L_*^2L_{\max}^2\mu_*^2\kappa_{*,\max}^{-2} + cc_2^2\alpha^{2.5}L_{\max}^2\mu_*^2\kappa_{*,\max}^{-2}\operatorname{dist}_t^2 \\
&\quad + cc_2^5\alpha^{6.5}L_*^2\eta_*\mu_*^5\kappa_{*,\max}^{-5} + cc_2^4\alpha^{6.5}L_{\max}^2L_{\max}^2\mu_*^4\kappa_{*,\max}^{-4} + cc_2\alpha^{4.5}L_*^2\mu_*^4\kappa_{*,\max}^{-4} \\
&\quad + cc_2^3\alpha^{4.5}L_{\max}^3\mu_*^3\kappa_{*,\max}^{-3} \\
&\quad + cc_2^2\alpha^{4.5}L_*^2L_{\max}^2\mu_*^2\kappa_{*,\max}^{-2} + cc_2\alpha^{4.5}L_{\max}^3L_{\max}^2\mu_*\kappa_{*,\max}^{-1} + cc_2\alpha^{2.5}L_{\max}^3\mu_*\kappa_{*,\max}^{-1}\operatorname{dist}_t^2 \\
&\quad + cc_2^4\alpha^{6.5}L_*^2L_{\max}^2\mu_*^4\kappa_{*,\max}^{-4} + cc_2^3\alpha^{6.5}L_{\max}^3L_{\max}^2\mu_*^3\kappa_{*,\max}^{-3} + cc_2^3\alpha^{4.5}L_{\max}^3\mu_*^3\kappa_{*,\max}^{-3} \\
&\quad + cc_2^2\alpha^{4.5}L_{\max}^4\mu_*^2\kappa_{*,\max}^{-2} + 2cc_2^2\alpha^{2.5}L_*^2\mu_*^2\kappa_{*,\max}^{-2} \\
&\quad + 2cc_2\alpha^{2.5}L_*^2\eta_*\mu_*\kappa_{*,\max}^{-1} + 2cc_2^4\alpha^{4.5}L_{\max}^2\mu_*^4\kappa_{*,\max}^{-4} + 2cc_2^3\alpha^{4.5}L_{\max}^2\eta_*\mu_*^3\kappa_{*,\max}^{-3} \\
&\quad + 2cc_2^3\alpha^{2.5}\eta_*\mu_*^3\kappa_{*,\max}^{-3} + 2cc_2^2\alpha^{2.5}L_*^2\mu_*^2\kappa_{*,\max}^{-2} + cc_2\alpha^{2.5}L_{\max}^2\mu_*^2\kappa_{*,\max}^{-2} \\
&\quad + cc_2^3\alpha^{4.5}L_{\max}^2\eta_*\mu_*^3\kappa_{*,\max}^{-3} + cc_2^2\alpha^{4.5}L_{\max}^2L_*^2\mu_*^2\kappa_{*,\max}^{-2} \\
&\quad + cc_2^2\alpha^{2.5}L_*^2\mu_*^2\kappa_{*,\max}^{-2} + cc_2\alpha^{2.5}L_{\max}^3\mu_*\kappa_{*,\max}^{-1} + cc_2^2\alpha^{4.5}L^4\mu_*^2\kappa_{*,\max}^{-2} \\
&\quad + cc_2\alpha^{4.5}L^4\eta_*\mu_*\kappa_{*,\max}^{-1} + cc_2\alpha^{2.5}L_{\max}^2\eta_*\mu_*\kappa_{*,\max}^{-1}\operatorname{dist}_t^2 + cc_2^4\alpha^{4.5}L^4\mu_*^4\kappa_{*,\max}^{-4} \\
&\quad + cc_2^3\alpha^{4.5}L_{\max}^2\eta_*\mu_*^3\kappa_{*,\max}^{-3} + cc_2^3\alpha^{6.5}L^4\eta_*\mu_*^3\kappa_{*,\max}^{-3} + cc_2^2\alpha^{4.5}L_{\max}^2L_*^2\mu_*^2\kappa_{*,\max}^{-2} \\
&\quad + cc_2^2\alpha^{4.5}L_{\max}^2L_*^2\mu_*^2\kappa_{*,\max}^{-2} + cc_2^2\alpha^{2.5}L_*^2\mu_*^2\kappa_{*,\max}^{-2}\operatorname{dist}_t \\
&\le cc_2^2\alpha^{2.5}L_{\max}^2\mu_*^2\kappa_{*,\max}^{-2} \\
&\quad + 5cc_2^4\alpha^{6.5}L_{\max}^2\mu_*^6 \\
&\quad + 19cc_2\alpha^{4.5}L_{\max}^2L_{\max}^2\mu_*^2 + cc_2\alpha^{2.5}L_{\max}^3\mu_*\kappa_{*,\max}^{-1}\operatorname{dist}_t^2 \\
&\quad + 2cc_2^2\alpha^{2.5}L_*^2\mu_*^2\kappa_{*,\max}^{-2} + 2cc_2\alpha^{2.5}L_*^2\eta_*\mu_*\kappa_{*,\max}^{-1} \\
&\quad + 2cc_2^3\alpha^{2.5}\eta_*\mu_*^3\kappa_{*,\max}^{-3} + 2cc_2^2\alpha^{2.5}L_*^2\mu_*^2\kappa_{*,\max}^{-2} + cc_2\alpha^{2.5}L_{\max}^2\mu_*^2\kappa_{*,\max}^{-2} \\
&\quad + cc_2^2\alpha^{2.5}L_*^2\mu_*^2\kappa_{*,\max}^{-2} + cc_2\alpha^{2.5}L_{\max}^3\mu_*\kappa_{*,\max}^{-1} \\
&\quad + cc_2\alpha^{2.5}L_{\max}^2\eta_*\mu_*\kappa_{*,\max}^{-1}\operatorname{dist}_t^2 \\
&\quad + cc_2\alpha^{2.5}L_*^2\mu_*^2\kappa_{*,\max}^{-2}\operatorname{dist}_t \\
&\le 24cc_2\alpha^{4.5}L_{\max}^2L_{\max}^2\mu_*^2 + 12cc_2\alpha^{2.5}L_{\max}^2\mu_*^2 + 3cc_2\alpha^{2.5}L_{\max}^2\mu_*^2\operatorname{dist}_t \\
&\le 14cc_2\alpha^{2.5}L_{\max}^2\mu_*^2 + 3cc_2\alpha^{2.5}L_{\max}^2\mu_*^2\operatorname{dist}_t \\
&\le 17cc_2\alpha^{2.5}L_{\max}^2\mu_*^2 \\
&\le \tfrac{9}{8}cc_2\sqrt{\alpha}\mu_*^2
\end{aligned} \tag{126}$$

using $\alpha \le 1/(4L_{\max})$, $\operatorname{dist}_t \le 1$, and again, $c_2 = E_0/20$. Thus,

$$\begin{aligned}
\|\mathbf{G}_t\|_2^2 &\le \left(c\sqrt{\alpha}(\operatorname{dist}_t + 2\|\mathbf{\Delta}_t\|_2)L_*^2 + \tfrac{9}{8}cc_2\sqrt{\alpha}\mu_*^2\right)^2 \\
&\le \left(\tfrac{9}{8}c\sqrt{\alpha}L_*^2 + \tfrac{9}{8}cc_2\sqrt{\alpha}\mu_*^2\right)^2 \\
&\le \tfrac{3}{2}\alpha\left(L_*^2 + c_2\mu_*^2\right)^2 \\
&\le 3\alpha L_*^4
\end{aligned} \tag{127}$$

using $c = 1.1$ in (127). Returning to (123) and applying our bounds on $\|\mathbf{B}_t^\top\mathbf{N}_t\|_2$ and $\|\mathbf{G}_t\|_2^2$, along with inductive hypothesis $A_3(t)$, yields

$$\begin{aligned}
\|\mathbf{\Delta}_{t+1}\|_2 &= \|\mathbf{\Delta}_t\|_2(1 - 2\beta\alpha E_0\mu_*^2) + 30c_2\beta\alpha^3L_{\max}^2\mu_*^2 + 3\beta^2\alpha^2L_*^4 \\
&\le \alpha^2L_{\max}^2(1 - 2\beta\alpha E_0\mu_*^2) + 30c_2\beta\alpha^3L_{\max}^2\mu_*^2 + 3\beta^2\alpha^2L_*^4 \\
&\le \alpha^2L_{\max}^2 - 2\beta\alpha^3E_0L_{\max}^2\mu_*^2 + 30c_2\beta\alpha^3L_{\max}^2\mu_*^2 + 3\beta^2\alpha^2L_*^4 \\
&\le \alpha^2L_{\max}^2
\end{aligned} \tag{128}$$

where the last inequality follows due to $\beta \le \frac{E_0\alpha}{10\kappa_*^4} \le \frac{\alpha E_0L_{\max}^2\mu_*^2}{6L_*^4}$ and $c_2 = E_0/20$. □

Lemma 19 (Exact MAML $A_4(t+1)$)**.** *Suppose the conditions of Theorem 7 are satisfied and $A_2(t)$, $A_3(t)$ and $A_5(t)$ hold. Then $A_4(t+1)$ holds, i.e.*

$$\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 \le (1 - 0.5\beta\alpha E_0 \mu_*^2)\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2. \tag{129}$$

Proof. Recall from (121) that the outer loop gradient for the representation satisfies

$$\mathbf{G}_t = -\bar{\mathbf{\Delta}}_t \mathbf{S}_t \mathbf{B}_t - \mathbf{S}_t \mathbf{B}_t \mathbf{\Delta}_t + \mathbf{N}_t \tag{130}$$

where $\mathbf{S}_t \coloneqq \alpha \frac{1}{n}\sum_{i=1}^n \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top$ and $\|\mathbf{N}_t\|_2 \le \frac{9}{8} c c_2 \sqrt{\alpha}\mu_*^2$, where $c_2 \coloneqq E_0/20$. As a result,

$$\begin{aligned}
\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 &= \|\mathbf{B}_{*,\perp}^\top (\mathbf{B}_t - \beta(-\bar{\mathbf{\Delta}}_t \mathbf{S}_t \mathbf{B}_t - \mathbf{S}_t \mathbf{B}_t \mathbf{\Delta}_t + \mathbf{N}_t))\|_2 \\
&= \|\mathbf{B}_{*,\perp}^\top (\mathbf{B}_t + \beta \mathbf{S}_t \mathbf{B}_t - \beta\alpha \mathbf{B}_t \mathbf{B}_t^\top \mathbf{S}_t \mathbf{B}_t + \beta \mathbf{S}_t \mathbf{B}_t - \beta\alpha \mathbf{S}_t \mathbf{B}_t \mathbf{B}_t^\top \mathbf{B}_t - \beta \mathbf{N}_t\|_2 \\
&= \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t (\mathbf{I}_k - \beta\alpha \mathbf{B}_t^\top \mathbf{S}_t \mathbf{B}_t) - \beta \mathbf{B}_{*,\perp}^\top \mathbf{N}_t\|_2 \\
&\le \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2 \|\mathbf{I}_k - \beta\alpha \mathbf{B}_t^\top \mathbf{S}_t \mathbf{B}_t\|_2 + \beta \|\mathbf{B}_{*,\perp}^\top \mathbf{N}_t\|_2
\end{aligned} \tag{131}$$

where the last equality follows because $\mathbf{B}_{*,\perp}^\top \mathbf{S}_t = \alpha \frac{1}{n}\sum_{i=1}^n \mathbf{B}_{*,\perp}^\top \mathbf{B}_* \mathbf{w}_{*,t,i} \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top = \mathbf{0}$. Note that due to Lemma 1 and $\|\mathbf{\Delta}_t\|_2 \le \frac{1}{10}$, $\sigma_{\min}(\mathbf{B}_t^\top \mathbf{S}_t \mathbf{B}_t) \ge E_0 \mu_*^2$ where $E_0 = 1 - \frac{1}{10} - \mathrm{dist}_0^2$. Therefore, by Weyl's inequality,

$$\|\mathbf{I}_k - \beta\alpha \mathbf{B}_t^\top \mathbf{S}_t \mathbf{B}_t\|_2 \le 1 - \beta\alpha E_0 \mu_*^2. \tag{132}$$

Furthermore, from (126), we have

$$\begin{aligned}
&\|\mathbf{B}_{*,\perp}^\top \mathbf{N}_t\|_2 \le \|\mathbf{N}_t\|_2 \le \tfrac{9}{8} c c_2 \sqrt{\alpha}\mu_*^2 \le \tfrac{5}{4} c_2 \sqrt{\alpha}\mu_*^2 \\
\implies &\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 \le \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2 (1 - \beta\alpha E_0 \mu_*^2) + \tfrac{5}{4} c_2 \sqrt{\alpha}\mu_*^2
\end{aligned} \tag{133}$$

Next, recall that $\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2 \ge \sigma_{\min}(\mathbf{B}_{*,\perp}^\top \mathbf{B}_t) \ge \sqrt{\frac{9}{10}}\sigma_{\min}(\mathbf{B}_{*,\perp}^\top \hat{\mathbf{B}}_t)/\sqrt{\alpha} = \sqrt{\frac{9}{10}}\sqrt{1-\mathrm{dist}_t^2}/\sqrt{\alpha} \ge \sqrt{\frac{9}{10}}\sqrt{1 - \frac{10}{9}\mathrm{dist}_0^2}/\sqrt{\alpha} = \sqrt{E_0}/\sqrt{\alpha}$ due to inductive hypotheses $A_3(t)$ and $A_4(t)$ and $E_0 \coloneqq 0.9 - \mathrm{dist}_0^2$. Therefore, using $c_2 \le 2E_0^{3/2}/5$, we obtain

$$\begin{aligned}
&\tfrac{5}{4} c_2 \sqrt{\alpha}\mu_*^2 \le 0.5\sqrt{\alpha} E_0^{3/2} \mu_*^2 \le 0.5\alpha E_0 \mu_*^2 \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2 \\
\implies &\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 \le \|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2 (1 - 0.5\beta\alpha E_0 \mu_*^2)
\end{aligned} \tag{134}$$

□

E. ANIL Finite Samples

First we define the following notations for the finite-sample case.

The inner loop update for the head of the i-th task on iteration t is given by:

$$\begin{aligned}
\mathbf{w}_{t,i} &= \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} \hat{\mathcal{L}}_i(\mathbf{B}_t, \mathbf{w}_t, \mathcal{D}_i^{in}) \\
&= (\mathbf{I}_k - \alpha \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t)\mathbf{w}_t + \alpha \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_* \mathbf{w}_{*,t,i} + \tfrac{\alpha}{m_{in}} \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in}.
\end{aligned} \tag{135}$$

Notation	Explanation
$\mathbf{X}_{t,i}^{in} \coloneqq [\mathbf{x}_{t,i,1}^{in}, \ldots, \mathbf{x}_{t,n,m_{in}}^{in}]^\top$	Data for inner loop gradient
$\mathbf{\Sigma}_{t,i}^{in} \coloneqq \frac{1}{m_{in}} \sum_{j=1}^{m_{in}} \mathbf{x}_{t,i,j}^{in} (\mathbf{x}_{t,i,j}^{in})^\top$	Empirical covariance matrix for inner loop gradient
$\mathbf{z}_{t,i}^{in} \coloneqq [z_{t,i,1}, \ldots, z_{t,i,m_{in}}]$	Additive noise for samples for inner loop gradient
$\mathbf{\Delta}_{t,i}^{in} \coloneqq \mathbf{I}_k - \alpha \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t$	Finite-sample analogues of $\mathbf{\Delta}_t$
$\bar{\mathbf{\Delta}}_{t,i}^{in} \coloneqq \mathbf{I}_d - \alpha \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in}$	Finite-sample analogues of $\bar{\mathbf{\Delta}}_t$
$\mathbf{X}_{t,i}^{out} \coloneqq [\mathbf{x}_{t,i,1}^{out}, \ldots, \mathbf{x}_{t,i,m_{out}}^{out}]^\top$	Data for outer loop gradient
$\mathbf{\Sigma}_{t,i}^{out} \coloneqq \frac{1}{m_{out}} \sum_{j=1}^{m_{out}} \mathbf{x}_{t,i,j}^{out} (\mathbf{x}_{t,i,j}^{out})^\top$	Empirical covariance matrix for outer loop gradient
$\mathbf{z}_{t,i}^{out} \coloneqq [z_{t,i,1}, \ldots, z_{t,i,m_{out}}]$	Additive noise for samples for outer loop gradient
$\delta_{m,d_1} \coloneqq \frac{\sqrt{d_1} + 10\sqrt{\log(n)}}{\sqrt{m}}$	Local concentration parameter
$\bar{\delta}_{m,d_2} \coloneqq \frac{10\sqrt{d_2}}{\sqrt{nm}}$	Global concentration parameter

For Exact ANIL, the finite-sample loss after the inner loop update is given by:

$$
\begin{aligned}
&\hat{F}_{t,i}(\mathbf{B}_t, \mathbf{w}_t; \mathcal{D}_{t,i}^{in}, \mathcal{D}_{t,i}^{out}) \\
&\quad \coloneqq \hat{\mathcal{L}}_{t,i}(\mathbf{B}_t, \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} \hat{\mathcal{L}}_{t,i}(\mathbf{B}_t, \mathbf{w}_t; \mathcal{D}_{t,i}^{in}); \mathcal{D}_{t,i}^{out}) \\
&\quad = \tfrac{1}{2m_{out}} \sum_{j=1}^{m_{out}} (\mathbf{x}_{t,i,j}^\top \mathbf{B}_t (\mathbf{w}_t - \alpha \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \mathbf{w}_t + \alpha \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_* \mathbf{w}_{*,t,i} - \tfrac{\alpha}{m_{in}} \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in}) \\
&\qquad\qquad - \mathbf{x}_{t,i,j}^\top \mathbf{B}_* \mathbf{w}_{*,t,i}) - z_{t,i,j}^{out})^2 \\
&\quad = \tfrac{1}{2m_{out}} \sum_{j=1}^{m_{out}} (\mathbf{x}_{t,i,j}^\top \bar{\mathbf{\Delta}}_{t,i}^{in} (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) + \tfrac{\alpha}{m_{in}} \mathbf{x}_{t,i,j}^\top \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} - z_{t,i,j})^2 \\
&\quad = \tfrac{1}{2m_{out}} \|\hat{\mathbf{v}}_{t,i}\|_2^2 \\
&\hat{\mathbf{v}}_{t,i} \coloneqq \mathbf{X}_{t,i}^{out} \bar{\mathbf{\Delta}}_{t,i}^{in} (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) + \tfrac{\alpha}{m_{in}} \mathbf{X}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} - \mathbf{z}_{t,i}^{out}
\end{aligned}
$$

Therefore, using the chain rule, the exact outer loop gradients for the i-th task are:

$$
\begin{aligned}
\nabla_{\mathbf{B}} \hat{F}_{t,i}(\mathbf{B}_t, \mathbf{w}_t; \mathcal{D}_{t,i}^{in}, \mathcal{D}_{t,i}^{in}) &= (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \hat{\mathbf{v}}_{t,i} \mathbf{w}_t^\top - \alpha \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \hat{\mathbf{v}}_{t,i} \mathbf{w}_t^\top \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \\
&\quad + \alpha \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \hat{\mathbf{v}}_{t,i} \mathbf{w}_{*,t,i}^\top \mathbf{B}_*^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \\
&\quad - \alpha \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \mathbf{w}_t \hat{\mathbf{v}}_{t,i}^\top \tfrac{1}{m_{out}} \mathbf{X}_{t,i}^{out} \mathbf{B}_t + \alpha \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_* \mathbf{w}_{*,t,i} \hat{\mathbf{v}}_{t,i}^\top \tfrac{1}{m_{out}} \mathbf{X}_{t,i}^{out} \mathbf{B}_t \\
&\quad + \tfrac{\alpha^2}{m_{in} m_{out}} (\mathbf{X}_{t,i}^{out})^\top \hat{\mathbf{v}}_{t,i} (\mathbf{z}_{t,i}^{in})^\top \mathbf{X}_{t,i}^{in} \mathbf{B}_t + \tfrac{\alpha^2}{m_{in} m_{out}} (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} \hat{\mathbf{v}}_{t,i}^\top \mathbf{X}_{t,i}^{out} \mathbf{B}_t \\
\nabla_{\mathbf{w}} \hat{F}_{t,i}(\mathbf{B}_t, \mathbf{w}_t; \mathcal{D}_{t,i}^{in}, \mathcal{D}_{t,i}^{out}) &= \mathbf{B}_t^\top (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \hat{\mathbf{v}}_{t,i} - \tfrac{\alpha}{m_{out}} \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out}
\end{aligned}
$$

Meanwhile, the first-order outer loop gradients for the i-th task are

$$
\begin{aligned}
&\nabla_{\mathbf{B}} \hat{\mathcal{L}}_{t,i}(\mathbf{B}_t, \mathbf{w}_{t,i}; \mathcal{D}_{t,i}^{in}, \mathcal{D}_{t,i}^{out}) \\
&\quad = \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{w}_{t,i} - \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_* \mathbf{w}_{*,t,i} \\
&\quad = \mathbf{\Sigma}_{t,i}^{out} (\mathbf{B}_t \mathbf{w}_{t,i} - \mathbf{B}_* \mathbf{w}_{*,t,i}) \mathbf{w}_{t,i}^\top - \tfrac{\alpha}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \mathbf{w}_{t,i}^\top \\
&\quad = \mathbf{\Sigma}_{t,i}^{out} (\mathbf{B}_t (\mathbf{\Delta}_{t,i}^{in} \mathbf{w}_t + \alpha \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_* \mathbf{w}_{*,t,i}) - \mathbf{B}_* \mathbf{w}_{*,t,i}) (\mathbf{\Delta}_{t,i}^{in} \mathbf{w}_t + \alpha \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_* \mathbf{w}_{*,t,i})^\top \\
&\qquad - \tfrac{\alpha}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \mathbf{w}_{t,i}^\top \\
&\quad = \mathbf{\Sigma}_{t,i}^{out} \bar{\mathbf{\Delta}}_{t,i}^{in} (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) (\mathbf{\Delta}_{t,i}^{in} \mathbf{w}_t + \alpha \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_* \mathbf{w}_{*,t,i})^\top - \tfrac{\alpha}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \mathbf{w}_{t,i}^\top
\end{aligned}
$$

$$
\nabla_{\mathbf{w}} \hat{\mathcal{L}}_{t,i}(\mathbf{B}_t, \mathbf{w}_{t,i}; \mathcal{D}_{t,i}^{in}, \mathcal{D}_{t,i}^{out}) = \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \bar{\mathbf{\Delta}}_{t,i}^{in} (\mathbf{B}_t \mathbf{w}_t - \mathbf{B}_* \mathbf{w}_{*,t,i}) - \tfrac{\alpha}{m_{out}} \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out}
$$

Define

$$\hat{\mathbf{G}}_{\mathbf{B},t} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{B}} \hat{F}_{t,i}(\mathbf{B}_t, \mathbf{w}_t), \quad \mathbf{G}_{\mathbf{B},t} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{B}} F_{t,i}(\mathbf{B}_t, \mathbf{w}_t)$$
$$\hat{\mathbf{G}}_{\mathbf{w},t} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{w}} \hat{F}_{t,i}(\mathbf{B}_t, \mathbf{w}_t), \quad \mathbf{G}_{\mathbf{w},t} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{w}} \hat{F}_{t,i}(\mathbf{B}_t, \mathbf{w}_t)$$

Now we are ready to state the result.

Theorem 8 (ANIL Finite Samples)**.** *Suppose Assumptions 1, 2 and 3 hold. Let $E_0 \coloneqq 0.9 - \mathrm{dist}_0^2 - \delta$ for some $\delta \in (0,1)$ to be defined shortly and assume E_0 is a positive constant. Suppose the initialization further satisfies $\alpha \mathbf{B}_0^\top \mathbf{B}_0 = \mathbf{I}_k$ and $\mathbf{w}_0 = \mathbf{0}$, and let the step sizes be chosen as $\alpha \le \frac{c'}{\sqrt{k}L_* + \sigma}$, and $\beta \le \frac{c'\alpha E_0^2}{\kappa_*^4}$ for ANIL and $\beta \le \frac{c'\alpha E_0^3 \mu_*}{\kappa_*^4} \min\left(1, \frac{\mu_*^2}{\eta_*^2}\right)$ for FO-ANIL, for some absolute constant c'. Then there exists a constant $c > 0$ such that, for ANIL, if*

$$\begin{aligned} m_{out} &\ge cT^2 \tfrac{k^2(L_*+\sigma)^2}{n\eta_*^2\kappa_*^8} + cT\tfrac{k^3\kappa_*^2(\kappa_*^2+\sigma^2/\mu_*^2)}{n} + c\sqrt{T}(k + \tfrac{kd}{n} + \log(n))\kappa_*^{-2}(\tfrac{\sigma^2}{L_*^2} + k) + ck + c\log(n) \\ m_{in} &\ge cT^2(k^2 + k\log(n))\tfrac{(L_*+\sigma)^2}{\eta_*^2\kappa_*^8} + cT(k^3 + k\log(n))(\kappa_*^4 + \tfrac{\sigma^4}{\mu_*^4}) + c\sqrt{T}\tfrac{k^3 d\log(nm_{in})}{n}\kappa_*^{-2}(\tfrac{\sigma^2}{L_*^2} + 1) \end{aligned} \tag{136}$$

and for FO-ANIL, if

$$\begin{aligned} m_{out} &\ge c\tfrac{Tdk}{n\kappa_*^2} + c\tfrac{Tdk\sigma^2}{nL_*^2\kappa_*^2} + c\tfrac{T^2k^3\kappa_*^4}{n} + c\tfrac{T^2k^3\sigma^4}{n\mu_*^4} + c\tfrac{k\mu_*^2}{\eta_*^2\kappa_*^6} + c\tfrac{k\sigma^2}{\eta_*^2\kappa_*^8} \\ m_{in} &\ge cT(k + \log(n))(k\kappa_*^2 + \tfrac{\sigma^2}{\mu_*^2}) + c\tfrac{T^2k^3\kappa_*^4}{n} + c\tfrac{T^2k^2\kappa_*^2\sigma^2}{\mu_*^2 n} + c\tfrac{T^2k^2(L_*^2+\sigma^2)}{\eta_*^2\kappa_*^8 n} \end{aligned}$$

then both ANIL and FO-ANIL satisfy that after T iterations,

$$\mathrm{dist}(\mathbf{B}_T, \mathbf{B}_*) \le \left(1 - 0.5\beta\alpha E_0 \mu_*^2\right)^{T-1} + O(\delta) \tag{137}$$

with probability at least $1 - O(T\exp(-90k)) - \frac{T}{\mathrm{poly}(n)} - \frac{T}{\mathrm{poly}(m_{in})}$, where for ANIL,

$$\begin{aligned} \delta = {} & \tfrac{1}{m_{in}}\left(\sqrt{k} + \tfrac{\sigma}{L_*}\right) + \tfrac{1}{\sqrt{m_{in}}}\left((k\kappa_*^2 + \sqrt{k}\kappa_*\sigma/\mu_*)(\sqrt{k} + \sqrt{\log(n)})\right) \\ & + \tfrac{1}{\sqrt{m_{out}}}\left((k\kappa_*^2 + \sqrt{k}\kappa_*\sigma/\mu_*)(\sqrt{k} + \sqrt{\log(n)})\right) \\ & + \tfrac{1}{\sqrt{nm_{in}}}\Big((k\kappa_*^2 + \sqrt{k}\kappa_*\sigma/\mu_*)(k\sqrt{d\log(nm_{in})} + k\log(nm_{in}) + \sqrt{d}\log^{1.5}(nm_{in}) + \log^2(nm_{in})) \\ & + \tfrac{\sigma^2}{\mu_*^2}(\sqrt{kd} + \sqrt{d}\log(nm_{in}) + \log^{1.5}(nm_{in})) + (k\kappa_*^2 + \sqrt{k}\kappa_*\sigma/\mu_*)\sqrt{d}\Big) \\ & + \tfrac{1}{\sqrt{nm_{out}}}\left((k\kappa_*^2 + \sqrt{k}\kappa_*\sigma/\mu_*)\sqrt{d} + \tfrac{\sigma^2}{\mu_*^2}(\tfrac{\sqrt{d}}{\sqrt{m_{in}}} + \sqrt{k})\right) \end{aligned}$$

and for FO-ANIL,

$$\delta = (\sqrt{k}\kappa_*^2 + \tfrac{\kappa_*\sigma}{\mu_*} + \tfrac{\sigma^2}{\mu_*^2\sqrt{m_{in}}})\tfrac{\sqrt{dk}}{\sqrt{nm_{out}}} \tag{138}$$

Proof. The proof uses an inductive argument with the following five inductive hypotheses:

1. $A_1(t) \coloneqq \{\|\mathbf{w}_t\|_2 \le \frac{\sqrt{\alpha}E_0}{10}\min(1, \frac{\mu_*^2}{\eta_*^2})\eta_*\}$
2. $A_2(t) \coloneqq \{\|\mathbf{\Delta}_t\|_2 \le (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_{t-1}\|_2 + \frac{5}{4}\alpha^2\beta^2 L_*^4\,\mathrm{dist}_{t-1}^2 + \beta\alpha\zeta_2\}$,
3. $A_3(t) \coloneqq \{\|\mathbf{\Delta}_t\|_2 \le \frac{1}{10}\}$,
4. $A_4(t) \coloneqq \{\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2 \le \left(1 - 0.5\beta\alpha E_0\mu_*^2\right)\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t-1}\|_2 + \beta\sqrt{\alpha}\zeta_4\}$,

5. $A_5(t) \coloneqq \{\text{dist}_t \leq \frac{\sqrt{10}}{3}\left(1-0.5\beta\alpha E_0\mu_*^2\right)^t \text{dist}_0 + \delta\}$.

where ζ_2 is defined separately for ANIL and FO-ANIL in Lemmas 34 and 28, respectively, and ζ_4 is defined separately for ANIL and FO-ANIL in Lemmas 35 and 29, respectively. These conditions hold for iteration $t = 0$ due to the choice of initialization $(\mathbf{B}_0, \mathbf{w}_0)$. We will show that if they hold for all iterations up to and including iteration t for an arbitrary t, then they hold at iteration $t+1$ with probability at least $1 - \frac{1}{\text{poly}(n)} - \frac{1}{\text{poly}(m_{in})} - O(\exp(-90k))$.

1. $\bigcap_{s=0}^{t}\{A_2(s) \cap A_6(s)\} \implies A_1(t+1)$. This is Lemma 27 for FO-ANIL and Lemma 33 for Exact ANIL.
2. $A_1(t) \cap A_3(t) \cap A_5(t) \implies A_2(t+1)$. This is Lemma 28 for FO-ANIL and Lemma 34 for Exact ANIL.
3. $A_1(t) \cap A_2(t+1) \cap A_3(t) \cap A_5(t) \implies A_3(t+1)$. This is Corollary 3 for FO-ANIL and Corollary 4 for Exact ANIL.
4. $A_1(t) \cap A_3(t) \cap A_5(t) \implies A_4(t+1)$. This is Lemma 29 for FO-ANIL and Lemma 35 for Exact ANIL.
5. $A_3(t+1) \cap \left(\cap_{s=1}^{t+1} A_4(s)\right) \implies A_5(t+1)$. By $A_3(t+1)$ and $A_4(t+1)$ we have:

$$\begin{aligned}
\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t+1}\|_2 &\leq (1-0.5\beta\alpha E_0\mu_*^2)\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_t\|_2 + \beta\sqrt{\alpha}\zeta_4 \\
&\leq (1-0.5\beta\alpha E_0\mu_*^2)^2\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_{t-1}\|_2 + (1-0.5\beta\alpha E_0\mu_*^2)\beta\sqrt{\alpha}\zeta_4 + \beta\sqrt{\alpha}\zeta_4 \\
&\vdots \\
&\leq (1-0.5\beta\alpha E_0\mu_*^2)^t\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_0\|_2 + \beta\sqrt{\alpha}\zeta_4 \sum_{s=0}^{t}(1-0.5\beta\alpha E_0\mu_*^2)^s \\
&\leq (1-0.5\beta\alpha E_0\mu_*^2)^t\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_0\|_2 + \frac{\beta\sqrt{\alpha}\zeta_4}{1-(1-0.5\beta\alpha E_0\mu_*^2)} \\
&= (1-0.5\beta\alpha E_0\mu_*^2)^t\|\mathbf{B}_{*,\perp}^\top \mathbf{B}_0\|_2 + \frac{2\zeta_4}{\sqrt{\alpha}E_0\mu_*^2}. \qquad (139)
\end{aligned}$$

Now we orthogonalize $\mathbf{B}_t$ and $\mathbf{B}_0$ via the QR-factorization, writing $\mathbf{B}_t = \hat{\mathbf{B}}_t\mathbf{R}_t$ and $\mathbf{B}_0 = \hat{\mathbf{B}}_0\mathbf{R}_0$. By inductive hypothesis $A_3(t+1)$, we have $\sigma_{\min}(\mathbf{B}_{t+1}) \geq \frac{\sqrt{0.9}}{\sqrt{\alpha}}$, and by the initialization we have $\sigma_{\max}(\mathbf{B}_0) \leq \frac{1}{\sqrt{\alpha}}$. Thus, using (139) and the definition of the principal angle distance, we have

$$\begin{aligned}
\text{dist}(\mathbf{B}_{t+1}, \mathbf{B}_*) &\leq \left((1-0.5\beta\alpha E_0\mu_*^2)^t \text{dist}(\mathbf{B}_0, \mathbf{B}_*)\|\mathbf{R}_0\|_2 + \frac{2\zeta_4}{\sqrt{\alpha}E_0\mu_*^2}\right)\|\mathbf{R}_{t+1}^{-1}\|_2 \\
&\leq \frac{\sqrt{10}}{3}(1-0.5\beta\alpha E_0\mu_*^2)^t \text{dist}(\mathbf{B}_0, \mathbf{B}_*) + \frac{3\zeta_4}{E_0\mu_*^2} \qquad (140) \\
&\leq (1-0.5\beta\alpha E_0\mu_*^2)^t + \delta \qquad (141)
\end{aligned}$$

where $\varepsilon = O(\frac{\zeta_4}{\mu_*^2})$.

After T rounds, we have that the inductive hypotheses hold on every round with probability at least

$$(1 - \tfrac{1}{\text{poly}(n)} - \tfrac{1}{\text{poly}(m_{in})} - O(\exp(-90k)))^T \geq 1 - O(T\exp(-90k)) - \tfrac{T}{\text{poly}(n)} - \tfrac{T}{\text{poly}(m_{in})} \qquad (142)$$

where the inequality follows by the Weierstrass Inequality, completing the proof. □

Throughout the proof we will re-use c, c', c'', etc. to denote absolute constants.

E.1. General Concentration Lemmas

We start with generic concentration results for random matrices and vectors that will be used throughout the proof.

We use $\chi_{\mathcal{E}}$ to denote the indicator random variable for the event $\mathcal{E}$, i.e. $\chi_{\mathcal{E}} = 1$ if $\mathcal{E}$ holds and $\chi_{\mathcal{E}} = 0$ otherwise.

Lemma 20. *Let $\mathbf{X}_1 = [\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,m_1}]^\top \in \mathbb{R}^{m_1 \times d}$ have rows which are i.i.d. samples from a mean-zero, $\mathbf{I}_d$-sub-gaussian distribution, and let $\mathbf{X}_{1,1}, \ldots, \mathbf{X}_{1,n}$ be independent copies of $\mathbf{X}_1$. Likewise, let $\mathbf{X}_2 = [\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,m_2}]^\top \in \mathbb{R}^{m_2 \times d}$ have rows which are i.i.d. samples from a mean-zero, $\mathbf{I}_d$-sub-gaussian distribution, and let $\mathbf{X}_{2,1}, \ldots, \mathbf{X}_{2,n}$ be independent copies of $\mathbf{X}_2$ (and independent of $\mathbf{X}_{1,1}, \ldots, \mathbf{X}_{1,n}$). Define $\mathbf{\Sigma}_{1,i} \coloneqq \frac{1}{m_1}\mathbf{X}_{1,i}^\top \mathbf{X}_{1,i}$ and $\mathbf{\Sigma}_{2,i} \coloneqq \frac{1}{m_2}\mathbf{X}_{2,i}^\top \mathbf{X}_{2,i}$ for all $i \in [n]$. Let the elements of $\mathbf{z}_1 \in \mathbb{R}^{m_1}$ and $\mathbf{z}_2 \in \mathbb{R}^{m_2}$ be i.i.d. samples from $\mathcal{N}(0, \sigma^2)$. Further, let $\mathbf{C}_{\ell,i} \in \mathbb{R}^{d \times d_{\lfloor \ell/2 \rfloor}}$ for $\ell = 1, \ldots, 6$ be fixed matrices for $i \in [n]$, and let $c_\ell \coloneqq \max_{i \in [n]} \|\mathbf{C}_{\ell,i}\|_2$ for $\ell = 1, \ldots, 6$. Let $\delta_{m,d_l} \coloneqq c\frac{\sqrt{d_l}+10\sqrt{\log(n)}}{\sqrt{m}}$ and $\bar{\delta}_{m,d_l} \coloneqq c\frac{10\sqrt{d_l}}{\sqrt{nm}}$ for some absolute constant c. Assume that in all cases below, each δ and $\bar{\delta}$ is less than 1. Then the following hold:*

1. $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\right\|_2 \geq c_1c_2\bar{\delta}_{m_1,d_0+d_1}\right) \leq 2e^{-90(d_0+d_1)}$

2. $\mathbb{P}\Bigg(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\right\|_2$
$\geq \sigma c_1c_2c_3c_4\left((1+\delta_{m_2,d_1+d_2})\bar{\delta}_{m_1,d_0+d_2} + \bar{\delta}_{m_2,d_0+d_2}\right) \leq 2e^{-90(d_0+d_2)} + 2n^{-99}$

3. $\mathbb{P}\Bigg(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\right\|_2$
$\geq \sigma c_1c_2c_3c_4\left((1+\delta_{m_2,d_1+d_2})\bar{\delta}_{m_1,d_0+d_2} + \delta_{m_2,d_1+d_2}\right) \leq 2e^{-90(d_0+d_2)} + 2n^{-99}$

4. $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{X}_{1,i}^\top \mathbf{z}_{1,i}\right\|_2 \geq \sigma c_1\bar{\delta}_{m_1,d_0}\right) \leq 2e^{-90d_0}$

5. $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \frac{1}{m_2}\mathbf{X}_{2,i}^\top \mathbf{z}_{2,i}\right\|_2 \geq c_1c_2c_3(1+\bar{\delta}_{m_1,d_0})\delta_{m_2,d_1}\right) \leq 2e^{-90d_1} + 2n^{-99}$

6. $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \frac{1}{m_1}\mathbf{X}_{1,i}^\top \mathbf{z}_{1,i}\mathbf{c}_{2,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{3,i}\right\|_2 \geq \sigma c_1c_2c_3(1+\delta_{m_2,d_1})\bar{\delta}_{m_1,d_0}\right) \leq 2e^{-90d_0} + 2n^{-99}$

7. $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \frac{1}{m_1}\mathbf{X}_{1,i}^\top \mathbf{z}_{1,i}\frac{1}{m_2}\mathbf{z}_{2,i}^\top \mathbf{X}_{2,i}\mathbf{C}_{2,i}\right\|_2 \geq \sigma^2 c_1c_2\bar{\delta}_{m_1,d_0}\delta_{m_2,d_1}\right) \leq 2e^{-90d_0} + 2e^{-90d_1}$

8. $\mathbb{P}\Bigg(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{C}_{6,i}\right\|_2$
$\geq c_1c_2c_3c_4c_5c_6\left((1+\delta_{m_2,d_1+d_2})(1+\delta_{m_2,d_2+d_3})\bar{\delta}_{m_1,d_0+d_3} + \delta_{m_2,d_2+d_3} + (1+\delta_{m_2,d_2+d_3})\delta_{m_2,d_1+d_2}\right)\Bigg)$
$\leq 2e^{-90(d_0+d_3)} + 4n^{-99}$

9. $\mathbb{P}\Bigg(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{6,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{C}_{6,i}\right\|_2$
$\geq c_1c_2c_3c_4c_5c_6\left((1+\delta_{m_1,d_0+d_1})(1+\delta_{m_1,d_2+d_3})\bar{\delta}_{m_2,d_1+d_2} + \bar{\delta}_{m_1,d_0+d_3} + (1+\delta_{m_1,d_2+d_3})\delta_{m_1,d_0+d_1}\right)\Bigg)$
$\leq 2e^{-90(d_0+d_3)} + 4n^{-99}$

10. $\mathbb{P}\Bigg(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{X}_{2,i}^\top \mathbf{z}_{2,i}\mathbf{c}_{4,i}^\top \mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i}\right\|_2$
$\geq \sigma c_1c_2c_3c_4c_5c_6\left(1+\bar{\delta}_{m_1,d_0+d_1}\right)\left(1+\delta_{m_2,d_3}\right)\delta_{m_2,d_1}\Bigg) \leq 2e^{-90d_1} + 4n^{-99}$

11. $\mathbb{P}\Bigg(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{X}_{1,i}^\top \mathbf{z}_{1,i}\right\|_2$
$\geq \sigma c_1c_2c_3c_4c_5c_6\left(1+\delta_{m_1,d_0+d_1}\right)\left(1+\bar{\delta}_{m_2,d_1+d_2}\right)\delta_{m_1,d_2}\Bigg) \leq 2e^{-90d_2} + 4n^{-99}$

12. $\mathbb{P}\Bigg(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{X}_{1,i}^\top \mathbf{z}_{1,i}\mathbf{c}_{2,i}^\top \mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i}\right\|_2$
$\geq \sigma c_1c_2c_3c_4c_5c_6\left(1+\delta_{m_2,d_2+d_3}\right)\left(1+\delta_{m_2,d_1+d_2}\right)\delta_{m_1,d_0}\Bigg) \leq 2e^{-90d_0} + 6n^{-99}$

13. $\mathbb{P}\Bigg(\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{X}_{2,i}^\top \mathbf{z}_{2,i}\mathbf{z}_{2,i}^\top \mathbf{X}_{2,i}\mathbf{C}_{4,i} \right\|_2$
$\geq \sigma^2 c_1c_2c_3c_4c_5 \delta_{m_2,d_1}\delta_{m_2,d_2}\left(1+\bar{\delta}_{m_1,d_0+d_2}\right)\Bigg) \leq 2e^{-90d_1} + 2e^{-90d_2} + 4n^{-99}$

Proof. We give the proofs for (1), (2), and (8) since the rest of the proofs follow using analogous arguments. In all cases, the proofs are standard applications of Bernstein's inequality.

1. For any fixed unit vector $\mathbf{u} \in \mathbb{R}^{d_0}$, $r_{\mathbf{u},i,j} \coloneqq \mathbf{u}^\top \mathbf{C}_{1,i}^\top \mathbf{w}_{1,i,j}$ is sub-gaussian with sub-gaussian norm at most $c\|\mathbf{C}_{1,i}\|_2$. Likewise, for any fixed unit vector $\mathbf{v} \in \mathbb{R}^{d_1}$, $r_{\mathbf{v},i,j} \coloneqq \mathbf{v}^\top \mathbf{C}_{2,i}^\top \mathbf{x}_{2,i,j}$ is sub-gaussian with norm at most $c\|\mathbf{C}_{2,i}\|_2$ for an absolute constant c. Furthermore, $\mathbb{E}[r_{\mathbf{v},i,j}r_{\mathbf{u},i,j}] = \mathbf{u}^\top \mathbf{C}_{1,i}^\top \mathbf{x}_{1,i,j}\mathbf{x}_{1,i,j}^\top \mathbf{C}_{2,i}\mathbf{v} = \mathbf{u}^\top \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{v}$. Therefore,

$$\mathbf{v}^\top \left(\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i} \right) \mathbf{u} = \frac{1}{nm_1}\sum_{i=1}^n\sum_{j=1}^{m_1} \left(r_{\mathbf{v},i,j}r_{\mathbf{u},i,j} - \mathbb{E}[r_{\mathbf{v},i,j}r_{\mathbf{u},i,j}] \right) \tag{143}$$

is the sum of nm_1 independent, mean-zero, sub-exponential random variables with norm $O(\|\mathbf{C}_{1,i}\|_2\|\mathbf{C}_{2,i}\|_2)$. By Bernstein's inequality we have

$$\left| \frac{1}{nm_1}\sum_{i=1}^n\sum_{j=1}^{m_1} \left(\mathbb{E}[r_{\mathbf{v},i,j}r_{\mathbf{u},i,j}] - r_{\mathbf{v},i,j}r_{\mathbf{u},i,j} \right) \right| \leq c \max_{i\in[n]} \|\mathbf{C}_{1,i}\|_2\|\mathbf{C}_{2,i}\|_2 \max\left(\tfrac{\sqrt{d_0+d_1}+\lambda}{\sqrt{nm_1}}, \tfrac{(\sqrt{d_0+d_1}+\lambda)^2}{nm_1} \right)$$

for some absolute constant c and any $\lambda > 0$, with probability at least $1 - 2e^{-\lambda^2}$ over the outer loop samples. Let $\mathcal{S}^{d_0-1}$ and $\mathcal{S}^{d_1-1}$ denote the unit spheres in $\mathbb{R}^{d_0}$ and $\mathbb{R}^k$, respectively. From Corollary 4.2.13 in (Vershynin, 2018), we know that there exists $\frac{1}{4}$-nets $\mathcal{M}_1$ and $\mathcal{M}_2$ on $\mathcal{S}^{d_0-1}$ and $\mathcal{S}^{d_1-1}$ with cardinalities at most 9^{d_0} and 9^{d_1}, respectively. Thus, conditioning on using the variational definition of the spectral norm, and taking a union bound over the $\frac{1}{4}$-nets, we have

$$\begin{aligned}
\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i} \right\|_2 &= \max_{\mathbf{v}\in\mathcal{S}^{d_0-1},\mathbf{u}\in\mathcal{S}^{d_1-1}} \left| \frac{1}{nm_1}\sum_{i=1}^n\sum_{j=1}^{m_1} \left(\mathbb{E}[r_{\mathbf{v},i,j}r_{\mathbf{u},i,j}] - r_{\mathbf{v},i,j}r_{\mathbf{u},i,j} \right) \right| \\
&\leq 2 \max_{\mathbf{v}\in\mathcal{M}_1,\mathbf{u}\in\mathcal{M}_2} \left| \frac{1}{nm_{out}}\sum_{i=1}^n\sum_{j=1}^{m_{out}} \left(\mathbb{E}[r_{\mathbf{v},i,j}r_{\mathbf{u},i,j}] - r_{\mathbf{v},i,j}r_{\mathbf{u},i,j} \right) \right| \\
&\leq c' \max_{i\in[n]} \|\mathbf{C}_{1,i}\|_2\|\mathbf{C}_{2,i}\|_2 \max\left(\tfrac{\sqrt{d_0+d_1}+\lambda}{\sqrt{nm_1}}, \tfrac{(\sqrt{d_0+d_1}+\lambda)^2}{nm_1} \right)
\end{aligned}$$

for some absolute constant c, with probability at least $1 - 2 \times 9^{d_0+d_1}e^{-\lambda^2}$ over the outer loop samples. Choose $\lambda = 10\sqrt{d}$ and let $\sqrt{nm_1} \geq 11\sqrt{d_0+d_1}$ to obtain that,

$$\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i} \right\|_2 \leq c \max_{i\in[n]} \|\mathbf{C}_{1,i}\|_2\|\mathbf{C}_{2,i}\|_2 \bar{\delta}_{m_1,d_0+d_1} \leq cc_1c_2\bar{\delta}_{m_1,d_0+d_1}$$

with probability at least $1 - 2e^{-90(d_0+d_1)}$.

2. Let $\mathbf{E} \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}$. We have

$$\begin{aligned}
\|\mathbf{E}\|_2 &\leq \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i} \right\|_2 \\
&= \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top (\mathbf{\Sigma}_{1,i} - \mathbf{I}_d)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i} + \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i} \right\|_2 \\
&\leq \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top (\mathbf{\Sigma}_{1,i} - \mathbf{I}_d)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i} \right\|_2}_{:=\mathbf{E}_1} + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top (\mathbf{\Sigma}_{2,i} - \mathbf{I}_d)\mathbf{C}_{4,i} \right\|_2}_{:=\mathbf{E}_2}
\end{aligned} \tag{144}$$

We first consider $\|\mathbf{E}_1\|_2$. For any $i \in [n]$, we have by Theorem 4.6.1 in (Vershynin, 2018),

$$\left\|\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i} \mathbf{C}_{4,i} - \mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\right\|_2 \leq c \max_{i\in[n]} \|\mathbf{C}_{3,i}\|_2 \|\mathbf{C}_{4,i}\|_2 \delta_{m_2,d_1+d_2} = cc_3c_4\delta_{m_2,d_1+d_2} \tag{145}$$

with probability at least $1 - 2n^{-100}$. Union bounding over all $i \in [n]$ and using the triangle inequality gives

$$\mathbb{P}\left(\mathcal{A} := \left\{\{\mathbf{\Sigma}_{2,i}\}_{i\in[n]} : \left\|\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i} \mathbf{C}_{4,i} - \mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\right\|_2 \leq cc_3c_4(1+\delta_{m_2,d_1+d_2})\ \forall i \in [n]\right\}\right) \geq 1 - 2n^{-99}. \tag{146}$$

Next, for any fixed set $\{\mathbf{\Sigma}_{2,i}\}_{i\in[n]} \in \mathcal{A}$, the d_2-dimensional random vectors $\{\mathbf{x}_{1,i,j}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\}_{i\in[n],j\in[m]}$ are sub-gaussian with sub-gaussian norms at most $c'c_2c_3c_4(1+\delta_{m_2,d_1+d_2})$. Likewise, the d_0-dimensional random vectors $\{\mathbf{C}_{1,i}^\top \mathbf{x}_{i,j'}\}_{i\in[n],j'\in\{2,\ldots,m\}}$ are sub-gaussian with norms at most c. Thus using the same argument as in the proof of (1.), we have

$$\begin{aligned}
&\mathbb{P}\Bigg(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top(\mathbf{\Sigma}_{1,i} - \mathbf{I}_d)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\right\|_2 && (147)\\
&\quad < c''c_1c_2c_3c_4(1+\delta_{m_2,d_1+d_2})\bar{\delta}_{m_1,d_0+d_2} \Big| \{\mathbf{\Sigma}_{2,i}\}_{i\in[n]}, \{\mathbf{\Sigma}_{2,i}\}_{i\in[n]} \in \mathcal{A}\Bigg)\\
&\quad \geq 1 - 2e^{-90(d_0+d_2)}. && (148)
\end{aligned}$$

for an absolute constant c''. Integrating over all $\{\mathbf{\Sigma}_{2,i}\}_{i\in[n]} \in \mathcal{A}$ and using $\delta_{m_2,d_1+d_2} \leq 1$ yields

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top(\mathbf{\Sigma}_{1,i} - \mathbf{I}_d)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\right\|_2 < c''c_1c_2c_3c_4\bar{\delta}_{m_1,d_0+d_2} \Big| \mathcal{A}\right) \geq 1 - 2e^{-90(d_0+d_2)}. \tag{149}$$

Therefore, by the law of total probability and (146), we have

$$\mathbb{P}\left(\|\mathbf{E}_1\|_2 \leq c''c_1c_2c_3c_4\bar{\delta}_{m_1,d_0+d_2}\right) \leq 2e^{-90(d_0+d_2)} + \mathbb{P}(\mathcal{A}^c) \leq 2e^{-90(d_0+d_2)} + 2n^{-99}. \tag{150}$$

Next, we have from (1.) that $\|\mathbf{E}_2\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top(\mathbf{\Sigma}_{2,i} - \mathbf{I}_d)\mathbf{C}_{4,i}\right\|_2 \leq cc_1c_2c_3c_4\bar{\delta}_{m_2,d_0+d_2}$ with probability at least $1 - 2e^{-90(d_0+d_2)}$. Finally, combining our bounds on the two terms in (144) via a union bound yields

$$\mathbb{P}\left(\|\mathbf{E}\|_2 \leq c''c_1c_2c_3c_4(\bar{\delta}_{m_1,d_0+d_2} + (1+\delta_{m_2,d_1+d_2})\bar{\delta}_{m_2,d_0+d_2})\right) \tag{151}$$

as desired. Note that we could instead use (146) to bound $\|\mathbf{E}_2\|_2$, which would result in the bound (3.).

8. Let $\mathbf{E} := \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i} - \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{C}_{6,i}$. We make a similar argument as in the proof of (2.) We have

$$\begin{aligned}
\|\mathbf{E}\|_2 \leq &\underbrace{\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top(\mathbf{\Sigma}_{1,i} - \mathbf{I}_d)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i}\right\|_2}_{:=\mathbf{E}_1}\\
&+ \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top(\mathbf{\Sigma}_{2,i} - \mathbf{I}_d)\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i}\right\|_2}_{:=\mathbf{E}_2}\\
&+ \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{C}_{1,i}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top(\mathbf{\Sigma}_{2,i} - \mathbf{I}_d)\mathbf{C}_{6,i}\right\|_2}_{:=\mathbf{E}_3}
\end{aligned} \tag{152}$$

We know from Theorem 4.6.1 in (Vershynin, 2018) that $\mathbb{P}(\|\mathbf{C}_{3,i}^\top(\mathbf{\Sigma}_{2,i} - \mathbf{I}_d)\mathbf{C}_{4,i}\|_2 \leq cc_3c_4\delta_{m_2,d_1+d_2})$ $\geq 1 - 2n^{-100}$ and $\mathbb{P}(\|\mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i}\|_2 \leq cc_5c_6(1+\delta_{m_2,d_2+d_3})) \geq 1 - 2n^{-100}$. Union bounding these events over

$i \in [n]$ gives $\mathbb{P}(\|\mathbf{E}_2\|_2 \le cc_1c_2c_3c_4c_5c_6\delta_{m_2,d_1+d_2}(1+\delta_{m_2,d_2+d_3})) \ge 1-4e^{-99}$. Union bounding over the same events, we also have $\mathbb{P}(\|\mathbf{E}_3\|_2 \le cc_1c_2c_3c_4c_5c_6\delta_{m_2,d_2+d_3}) \ge 1-2n^{-99}$. Next, we make a similar argument as in (2.) to control $\|\mathbf{E}_1\|_2$, except that here $\mathcal{A}$ is defined as

$$\mathcal{A} := \Big\{ \{\mathbf{\Sigma}_{2,i}\}_{i\in[n]} : \big\|\mathbf{C}_{3,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\big\|_2 \le cc_3c_4(1+\delta_{m_2,d_1+d_2}), \\ \big\|\mathbf{C}_{5,i}^\top \mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i}\big\|_2 \le cc_5c_6(1+\delta_{m_2,d_2+d_3})\ \forall i\in[n]\Big\}, \tag{153}$$

which occurs with probability at least $1-4n^{-99}$ (which is implied by our discussion of bounding $\|\mathbf{E}_3\|_2$). Thus, following the logic in (2.), we obtain $\mathbb{P}(\|\mathbf{E}_1\|_2 \le cc_1c_2c_3c_4c_5c_6(1+\delta_{m_2,d_1+d_2})(1+\delta_{m_2,d_2+d_3})\bar{\delta}_{m_1,d_0+d_3}) \ge 1-4n^{-99}-2e^{-90(d_0+d_3)}$. Combining all bounds yields the desired result.

More generally, we add and subtract terms to show concentration through either a $\mathbf{\Sigma}-\mathbf{I}_d$ matrix, or an $\mathbf{Xz}$ matrix, with off terms bounded for each i by sub-gaussianity. □

Lemma 21. *Consider the setting described in Lemma 20. Further, suppose* $\min(d_1,d_2,d_3)=1$ *and* $\max(d_1,d_2,d_3)=k$. *Then the following events each hold with probability at most* $c'(e^{-100}+n^{-99}+{m_1}^{-99})$ *for absolute constants* c,c'*:*

- $\mathcal{U}_1 := \left\{ \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top\mathbf{\Sigma}_{1,i}\mathbf{C}_{6,i} - \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top\mathbf{C}_{6,i} \right\|_2 \ge cc_2c_3c_4c_5c_6(\tilde{\delta}+\frac{k}{m_1}) \right\}$
- $\mathcal{U}_2 := \left\{ \big\|\frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top\mathbf{X}_{1,i}^\top\mathbf{z}_{1,i}\big\|_2 \ge c\sigma c_2c_3c_4c_5\tilde{\delta} \right\}$
- $\mathcal{U}_3 := \left\{ \big\|\frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{1,i}\mathbf{c}_{2,i}\mathbf{z}_{1,i}^\top\mathbf{X}_{1,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i}\big\|_2 \ge c\sigma c_2c_3c_4c_5\tilde{\delta} \right\}$
- $\mathcal{U}_4 := \Big\{ \big\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_{1,i}^\top\mathbf{z}_{1,i}\mathbf{c}_{2,i}^\top\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{1,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{6,i}\big\|_2 \\ \ge c\sigma c_1c_2c_3c_4c_5c_6\frac{(\sqrt{kd}+\sqrt{d\log(nm_1)}+\log(nm_1))\log(nm_1)}{\sqrt{nm_1}} \Big\}$
- $\mathcal{U}_5 := \Big\{ \big\|\frac{1}{nm_1^2}\sum_{i=1}^n \mathbf{X}_{t,i}^\top\mathbf{z}_{t,i}\mathbf{z}_{t,i}^\top\mathbf{X}_{t,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\big\|_2 \\ \ge c\sigma^2c_2c_3c_4\left(\frac{(\sqrt{kd}+\sqrt{d\log(nm_1)}+\log(nm_1))\sqrt{\log(nm_1)}}{\sqrt{nm_1}}+\frac{1}{m_1}\right) \Big\}$

where

$$\tilde{\delta} := (k\sqrt{d\log(nm_1)}+k\log(nm_1)+\sqrt{d}\log^{1.5}(nm_1)+\log^2(nm_1))/\sqrt{nm_1}.$$

Proof. 1. Similarly to previous proofs involving sums of products of independent matrices, the idea is to first use that one set of matrices is small with high probability, then condition on these sets of matrices being small to isolate the randomness of the other matrices. Note that matrix $\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}$ has maximum dimension at most k, so by Lemma 20, for any $i\in[n]$, $\{\|\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\|_2 \ge cc_3c_4(1+\delta_{m_1,k})\}$ holds with probability at most n^{100}. Applying a union bound over $[n]$ gives that $\mathcal{A} := \cap_{i\in[n]}\{\|\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\|_2 \le cc_3c_4(1+\delta_{m_1,k})\}$ holds with probability at least $1-n^{-99}$. Conditioning on $\mathcal{A}$, and using $\delta_{m_1,k}\le 1$, we can apply Lemma 22 to obtain that

$$\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{1,i}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top\mathbf{\Sigma}_{1,i}\mathbf{C}_{6,i} - \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_{2,i}\mathbf{C}_{4,i}\mathbf{C}_{5,i}^\top\mathbf{C}_{6,i} \right\|_2 \ge cc_2c_3c_4c_5c_6(\tilde{\delta}+\tfrac{1+C^2}{m_1}) \tag{154}$$

occurs with probability at most . Since $\mathbb{P}(\mathcal{U}_1) \le \mathbb{P}(\mathcal{U}_1|\mathcal{A}) + \mathbb{P}(\mathcal{A}^c)$, we obtain the result.

2. We make the same argument as for (1) except that we apply Lemma 23 instead of Lemma 22.

3. Again, we use Lemma 23 as in (2).

4. Again, we use Lemma 23 as in (2).

5. Here we make the same argument as (1) except that we apply Lemma 24 instead of Lemma 22.

□

The following is a slightly generalized version of Theorem 1.1 in Magen & Zouzias (2011): here, the random matrices are not necessarily identically distributed, whereas they are identically distributed in Magen & Zouzias (2011). However, the proof from (Magen & Zouzias, 2011) does not rely on the matrices being identically distributed, so the same proof from Magen & Zouzias (2011) holds without modification for the below result.

Theorem 9 (Theorem 1.1 in (Magen & Zouzias, 2011))**.** *Let $0 < \epsilon < 1$ and $\mathbf{M}_1, \dots, \mathbf{M}_N$ be a sequence of independent symmetric random matrices that satisfy $\|\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\mathbf{M}_i]\|_2 \le 1$ and $\|\mathbf{M}_i\|_2 \le B$ and $\mathrm{rank}(\mathbf{M}_i) \le r$ almost surely for all $i \in [N]$. Set $N = \Omega(B\log(B/\epsilon^2)/\epsilon^2)$. If $r \le N$ almost surely, then*

$$\mathbb{P}\left(\left\|\frac{1}{N}\sum_{i=1}^N \mathbf{M}_i - \mathbb{E}[\mathbf{M}_i]\right)\right\|_2 \le \frac{1}{\mathrm{poly}(N)} \tag{155}$$

The following lemma again gives generic concentration results but for a more difficult set of matrices. The key technical contribution is a truncated version of Theorem 9.

Lemma 22. *Suppose that $\mathbf{x}$ is a random vector with $\mathbb{E}[\mathbf{x}] = \mathbf{0}_d$, and $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}_d$, and is $\mathbf{I}_d$-sub-gaussian. Let $\{\mathbf{x}_{i,j}\}_{i\in[n],j\in[m]}$ be nm independent copies of $\mathbf{x}$. Further, let $\mathbf{C}_{\ell,i} \in \mathbb{R}^{d\times d_{\lfloor \ell/2\rfloor}}$ for $\ell = 2,3,4$ be fixed matrices for $i \in [n]$, and let $c_\ell := \max_{i\in[n]} \|\mathbf{C}_{\ell,i}\|_2$ for $\ell = 2,3,4$. Denote $\mathbf{\Sigma}_i = \frac{1}{m}\sum_{j=1}^m \mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top$. Then, if $m \ge \max(1, C^2 c c_2 c_3 c_4 d_1)$,*

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_i\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_i\mathbf{C}_{4,i} - \frac{1}{n}\sum_{i=1}^n \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{C}_{4,i}\right\|_2 \\
&\le c(\sqrt{\log(nm)} + \sqrt{d_1})(\sqrt{\log(nm)} + \sqrt{d_2})\frac{\sqrt{d+d_2}+\sqrt{\log(m)}}{\sqrt{nm}}c_2c_3c_4 \\
&\quad + \frac{c}{\sqrt{nm}}(\sqrt{\log(nm)} + \sqrt{d})\left(\prod_{\ell=2}^4(\sqrt{\log(nm)} + \sqrt{d_{\lfloor \ell/2\rfloor}})\right)\max(c_2c_3c_4, 1) + c\tfrac{d_1}{m}c_2c_3c_4
\end{aligned}$$

for an absolute constant c, with probability at least $1 - 2m^{-99} - \frac{1}{\mathrm{poly}(nm)} - 2e^{-90d_1}$.

As in previous cases, in this lemma we would like to show concentration of fourth-order products of sub-gaussian random vectors with only $m = \mathrm{poly}(k)\tilde{O}(\frac{d}{n} + 1)$ samples per task. The issue here, unlike in the cases in Lemma 20, is that the leading $\mathbf{\Sigma}_i$ has no dimensionality reduction - there is no product matrix $\mathbf{C}_{1,i}$ to bring the d-dimensional random vectors that compose the leftmost $\mathbf{\Sigma}_i$ to a lower dimension. Thus, we would need $m = \Omega(d)$ samples per task to show concentration of each $\mathbf{\Sigma}_i$ (or $\mathbf{\Sigma}_i\mathbf{C}_{2,i}$). We must get around this by averaging over n. However, doing so requires dealing with fourth-order products of random vectors instead of bounding each of the two copies of $\mathbf{\Sigma}_i$ in the i-th term separately (perhaps along with their dimensionality-reducing products).

Due to the fourth-order products, we cannot apply standard concentrations based on sub-gaussian and sub-exponential tails. Instead, we leverage the low rank (at most k) of the matrices involved by applying a truncated version of the of the concentration result for bounded, low-rank random matrices in (Magen & Zouzias, 2011).

Proof. Throughout the proof we use c as a generic absolute constant. First note that by expanding $\mathbf{\Sigma}_i$ and the triangle

inequality,

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_i\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_i\mathbf{C}_{4,i}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{C}_{4,i}\right\|_2\\
&\le\left\|\frac{1}{nm^2}\sum_{i=1}^{n}\sum_{j,j'\neq j}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top\mathbf{C}_{4,i}-\frac{m(m-1)}{nm^2}\sum_{i=1}^{n}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{C}_{4,i}\right\|_2\\
&\quad+\left\|\frac{1}{nm^2}\sum_{i=1}^{n}\sum_{j}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}-\frac{m}{nm^2}\sum_{i=1}^{n}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{C}_{4,i}\right\|_2\\
&=\underbrace{\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}-\frac{(m-1)}{nm}\sum_{i=1}^{n}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{C}_{4,i}\right\|_2}_{=:\mathbf{E}'}\\
&\quad+\underbrace{\left\|\frac{1}{nm^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}-\frac{1}{nm}\sum_{i=1}^{n}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{C}_{4,i}\right\|_2}_{=:\mathbf{E}''}
\end{aligned}$$

Note that $\mathbf{E}'$ is unbiased while $\mathbf{E}''$ is biased due to the fourth-order product. We first bound $\|\mathbf{E}'\|_2$.

Step 1: Bound $\|\mathbf{E}'\|_2$. Add and subtract $\frac{(m-1)}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}$ to obtain

$$\begin{aligned}
\|\mathbf{E}'\|_2&\le\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}\right\|_2\\
&\quad+\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\left(\tfrac{1}{m}\sum_{j=1}^{m}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top-\mathbf{I}_d\right)\mathbf{C}_{4,i}\right\|_2\\
&\le\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}\right\|_2\\
&\quad+cc_2c_3c_4\bar{\delta}_{m,\max(d*,d_2)}
\end{aligned}\tag{156}$$

where (156) follows with probability at least $1-2e^{-90(d_1+d_2)}$ by Lemma 20, and $d*$ denotes d if the $\mathbf{C}_{2,i}$'s are distinct, and denotes d_1 otherwise (since if these matrices are equal, they can be factored out of the norm, in which case we show concentration of $d_1\times d_2$-dimensional random matrices). To deal with the first term in (156), note that as mentioned before, we need to show concentration over $i\in[n]$ to avoid requiring $m=\Omega(d)$. Ideally, we could also concentrate over $j\in[m]$, but we would lose independence of the summands. Thus, we reorder the sum and use the triangle inequality to write

$$\begin{aligned}
&\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}\right\|_2\\
&=\left\|\frac{1}{m}\sum_{j=1}^{m}\frac{m-1}{nm}\sum_{i=1}^{n}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}\right\|_2\\
&\le\frac{1}{m}\sum_{j=1}^{m}\underbrace{\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}\right\|_2}_{=:\mathbf{E}_j}
\end{aligned}$$

For each $j\in[m]$, define $\mathcal{E}_j$ as the event $\{\|\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\|_2\le(\gamma+\sqrt{d_1})c_3,\|\mathbf{C}_{4,i}^\top\mathbf{x}_{i,j}\|_2\le(\gamma+\sqrt{d_2})c_4\ \forall i\in[n]\}$ for some $\gamma>0$. Note that $\mathbf{x}_{i,j}$ and $\mathbf{C}_{2,i}^\top\mathbf{x}_{i,j}$ are d (resp. d_2)-dimensional sub-gaussian random vectors with sub-gaussian norm at most c (resp. cc_2). Thus $\mathcal{E}_j$ occurs with probability at least $1-2ne^{-c\gamma^2}$. Then using the law of total probability, for any $\epsilon>0$, we have

$$\mathbb{P}\left(\|\mathbf{E}_j\|_2\ge\epsilon\right)\le\mathbb{P}\left(\|\mathbf{E}_j\|_2\ge\epsilon|\mathcal{E}_j\right)+\mathbb{P}\left(\mathcal{E}_j^c\right)\le\mathbb{P}\left(\|\mathbf{E}_j\|_2\ge\epsilon|\mathcal{E}_j\right)+2ne^{-c\gamma^2}\tag{157}$$

Consider $\mathbf{E}_1$. For any fixed set $\{\mathbf{x}_{i,1}\}_{i\in[n]} \in \mathcal{E}_1$, the d_2-dimensional random vectors $\{\mathbf{x}_{i,j'}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{x}_{i,1}\mathbf{x}_{i,1}^\top \mathbf{C}_{4,i}\}_{i\in[n],j'\in\{2,\dots,m\}}$ are sub-gaussian with norms at most $c'(\gamma+\sqrt{d_1})(\gamma+\sqrt{d_2})c_2c_3c_4$. Likewise, the d-dimensional random vectors $\{\mathbf{x}_{i,j'}\}_{i\in[n],j'\in\{2,\dots,m\}}$ are sub-gaussian with norms at most c. Thus using Bernstein's inequality, we can bound

$$\mathbb{P}\Bigg(\|\mathbf{E}_1\|_2 < c''(\gamma+\sqrt{d_1})(\gamma+\sqrt{d_2})c_2c_3c_4 \max\left(\tfrac{\sqrt{d+d_2}+\lambda}{\sqrt{n(m-1)}}, \tfrac{d+d_2+\lambda^2}{n(m-1)}\right) \Big| \{\mathbf{x}_{i,1}\}_{i\in[n]}, \{\mathbf{x}_{i,1}\}_{i\in[n]} \in \mathcal{E}_1\Bigg) \geq 1-2e^{-\lambda^2}. \tag{158}$$

for $\lambda > 0$ and an absolute constant c''. Integrating over all $\{\mathbf{x}_{i,1}\}_{i\in[n]} \in \mathcal{E}_1$ yields

$$\mathbb{P}\left(\|\mathbf{E}_1\|_2 < c''(\gamma+\sqrt{d_1})(\gamma+\sqrt{d_2})c_2c_3c_4 \max\left(\tfrac{\sqrt{d+d_2}+\lambda}{\sqrt{n(m-1)}}, \tfrac{d+d_2+\lambda^2}{n(m-1)}\right) |\mathcal{E}_1\right) \geq 1-2e^{-\lambda^2}. \tag{159}$$

Therefore, using (157), we have

$$\mathbb{P}\left(\|\mathbf{E}_1\|_2 \geq c''(\gamma+\sqrt{d_1})(\gamma+\sqrt{d_2})c_2c_3c_4 \max\left(\tfrac{\sqrt{d+d_2}+\lambda}{\sqrt{n(m-1)}}, \tfrac{d+d_2+\lambda^2}{n(m-1)}\right)\right) \leq 2e^{-\lambda^2} + 2ne^{-c\gamma^2}. \tag{160}$$

Repeating the same argument for all $j \in [m]$ and applying a union bound gives

$$\mathbb{P}\left(\frac{1}{m}\sum_{j=1}^m \|\mathbf{E}_j\|_2 \geq c''(\gamma+\sqrt{d_1})(\gamma+\sqrt{d_2})c_2c_3c_4 \max\left(\tfrac{\sqrt{d+d_2}+\lambda}{\sqrt{n(m-1)}}, \tfrac{d+d_2+\lambda^2}{n(m-1)}\right)\right) \leq 2me^{-\lambda^2} + 2mne^{-c\gamma^2}. \tag{161}$$

Choose $\lambda = 10\sqrt{\log(m)}$ and $\gamma = 10\sqrt{\log(mn)}$, and use $\sqrt{n(m-1)} \geq \sqrt{d+d_2} + 10\sqrt{\log(m)}$ to obtain

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{m}\sum_{j=1}^m \|\mathbf{E}_j\|_2 \geq c'''c_2c_3c_4(\sqrt{\log(nm)}+\sqrt{d_1})(\sqrt{\log(nm)}+\sqrt{d_2})\tfrac{\sqrt{d+d_2}+\sqrt{\log(m)}}{\sqrt{nm}}\right) \\
&\qquad \leq 2m^{-99} + 2(mn)^{-99cd_1} + 2(mn)^{-99cd_2} \\
\implies &\mathbb{P}\Big(\|\mathbf{E}'\|_2 \geq c''c_2c_3c_4(\sqrt{\log(nm)}+\sqrt{d_1})(\sqrt{\log(nm)}+\sqrt{d_2})\tfrac{\sqrt{d+d_2}+\sqrt{\log(m)}}{\sqrt{nm}} \\
&\qquad + cc_2c_3c_4\bar{\delta}_{m,\max(d*,d_2)}\Big) \\
&\qquad \leq 2m^{-99} + 2(mn)^{-99cd_1} + 2(mn)^{-99cd_2} + 2e^{-90(d_1+d_2)} \qquad (162) \\
\implies &\mathbb{P}\left(\|\mathbf{E}'\|_2 \geq c'''c_2c_3c_4(\sqrt{\log(nm)}+\sqrt{d_1})(\sqrt{\log(nm)}+\sqrt{d_2})\tfrac{\sqrt{d+d_2}+\sqrt{\log(m)}}{\sqrt{nm}}\right) \\
&\qquad \leq 2m^{-99} + 2(mn)^{-99cd_1} + 2(mn)^{-99cd_2} + 2e^{-90(d_1+d_2)} \qquad (163)
\end{aligned}$$

where (162) follows from (156) and (163) follows by the fact that $\bar{\delta}_{m,\max(d*,d_2)}$ is dominated.

Step 2: Bound $\|\mathbf{E}''\|_2$. Bounding $\|\mathbf{E}''\|_2$ is challenging because we must deal with fourth-order products in $\mathbf{x}_{i,j}$, which may have heavy tails. However, we can leverage the independence and low-rank of the summands, combined with the sub-gaussian tails of each random vector. Second, we must control the bias in $\mathbf{E}'$, which we achieve by appealing to C-L4-L2 hypercontractivity. First note that by the triangle inequality

$$\|\mathbf{E}''\|_2 \leq \left\|\frac{1}{nm^2}\sum_{i=1}^n\sum_{j=1}^m = \mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top \mathbf{C}_{4,i}\right\|_2 + \left\|\frac{1}{nm}\sum_{i=1}^n \mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top \mathbf{C}_{4,i}\right\|_2 \tag{164}$$

It remains to control the first norm. To do so, we employ Theorem 9 (a.k.a. Theorem 1.1 from (Magen & Zouzias, 2011)) which characterizes the concentration of low-rank, bounded, symmetric random matrices with small expectation. Thus, in order to apply this theorem, we must truncate and symmetrize the random matrices, and control their expectation.

Define $\mathcal{E}_{\ell,i,j} \coloneqq \{\|\mathbf{C}_{\ell,i}\mathbf{x}_{i,j}\|_2 \le c(\rho + \sqrt{d_{\lfloor \ell/2 \rfloor}})c_\ell\}$ for some $\rho > 0$ and $\ell = 2,3,4$ and all i,j, and $\mathcal{E}_{1,i,j} \coloneqq \{\|\mathbf{x}_{i,j}\|_2 \le c(\rho + \sqrt{d})\}$ for some $\rho > 0$ and $\ell = 2,3,4$ and all i,j. Let $\chi_{\mathcal{E}_{\ell,i,j}}$ be the indicator random variable for the event $\mathcal{E}_{\ell,i,j}$. Define the truncated random variables $\bar{\mathbf{x}}_{\ell,i,j} \coloneqq \chi_{\mathcal{E}_{\ell,i,j}}\mathbf{C}_{\ell,i}\mathbf{x}_{i,j}$ for $\ell = 2,3,4$ and all i,j and $\bar{\mathbf{x}}_{1,i,j} \coloneqq \chi_{\mathcal{E}_{\ell,i,j}}\mathbf{x}_{i,j}$ for all i,j. Let $\mathbf{S}_{i,j} \coloneqq \mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{4,i}/m$ and $\bar{\mathbf{S}}_{i,j} \coloneqq \bar{\mathbf{x}}_{1,i,j}\bar{\mathbf{x}}_{2,i,j}^\top\bar{\mathbf{x}}_{3,i,j}\bar{\mathbf{x}}_{4,i,j}^\top/m$ for each i,j. Note that due to sub-gaussianity and earlier arguments, $\mathbb{P}(\cup_{i,j}\cup_{\ell=1}^4 \mathcal{E}_{\ell,i,j}) \le 2mn\sum_{\ell=1}^4 e^{-c\rho^2} = 8mne^{-c\rho^2}$. Thus, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbf{S}_{i,j}\right\|_2 \le \epsilon\right) \le \mathbb{P}\left(\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \bar{\mathbf{S}}_{i,j}\right\|_2 \le \epsilon\right) + 8nme^{-c\rho^2} \tag{165}$$

First, form the lifted, symmetric matrices

$$\tilde{\bar{\mathbf{S}}}_{i,j} \coloneqq \begin{bmatrix} \mathbf{0} & \bar{\mathbf{S}}_{i,j} \\ \bar{\mathbf{S}}_{i,j}^\top & \mathbf{0} \end{bmatrix} \tag{166}$$

for all i,j, and note that $\left\|\sum_{i=1}^n\sum_{j=1}^m \tilde{\bar{\mathbf{S}}}_{i,j}\right\|_2 = 2\left\|\sum_{i=1}^n\sum_{j=1}^m \bar{\mathbf{S}}_{i,j}\right\|_2$. Also note that by definition, $\|\tilde{\bar{\mathbf{S}}}_{i,j}\|_2 \le B \coloneqq 2(\rho+\sqrt{d})\prod_{\ell=2}^4(\rho + \sqrt{d_{\lfloor \ell/2 \rfloor}})\max(c_2c_3c_4, 1)$ for all i,j almost surely, and the $\tilde{\bar{\mathbf{S}}}_{i,j}$'s are independent.

We still must control $\|\mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\|_2$. We have that $\|\mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\|_2 = 2\|\mathbb{E}[\bar{\mathbf{S}}_{i,j}]\|_2$. Using Lemma 25 (with $\mathbf{C}_1 = \mathbf{I}_d$), we obtain $m\|\mathbb{E}[\bar{\mathbf{S}}_{i,j}]\|_2 \le mC^2\|\mathbf{C}_{2,i}\|_2\|\mathbf{C}_{3,i}\|_2\|\mathbf{C}_{4,i}\|_2 \le mC^2c_2c_3c_4d_1$ for all $i \in [n], j \in [m]$. Thus, $\|\mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\|_2 \le 1$ for all i,j as $m \ge 2C^2c_2c_3c_4d_1$.

Next, note that each $\bar{\mathbf{S}}_{i,j}$ is rank at most $\min(d,d_1,d_2)$, so $\tilde{\bar{\mathbf{S}}}_{i,j}$ is rank at most $2\min(d,d_1,d_2)$. Now we are ready to apply Theorem 9. Doing so, we obtain:

$$\mathbb{P}\left(\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \tilde{\bar{\mathbf{S}}}_{i,j} - \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\right\|_2 \ge \epsilon\right) \le \frac{1}{\mathrm{poly}(nm)} \tag{167}$$

as long as $nm \ge cB\log(B/\epsilon^2)/\epsilon^2$ and $nm \ge c\min(d,d_1,d_2)$. Setting $\epsilon = \frac{cB}{\sqrt{nm}}$ yields

$$\mathbb{P}\left(\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \tilde{\bar{\mathbf{S}}}_{i,j} - \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\right\|_2 \ge \frac{B'}{\sqrt{nm}}\right) \le \frac{1}{\mathrm{poly}(nm)} \tag{168}$$

as long as $nm \le Be^{c'B}$, which always holds since we will soon choose $\rho = \sqrt{\log(nm)}$ and we have chosen B appropriately. Therefore, with probability at least $\frac{1}{\mathrm{poly}(nm)}$, we have

$$\begin{aligned}\frac{1}{2}\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \tilde{\bar{\mathbf{S}}}_{i,j}\right\|_2 &\le \frac{1}{2}\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \tilde{\bar{\mathbf{S}}}_{i,j} - \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\right\|_2 + \frac{1}{2}\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\right\|_2 \\ &\le \frac{B}{2\sqrt{nm}} + \tfrac{C^2d_1}{nm}\sum_{i=1}^n \|\mathbf{C}_{2,i}\|_2\|\mathbf{C}_{3,i}\|_2\|\mathbf{C}_{4,i}\|_2\end{aligned}$$

which implies that

$$\|\mathbf{E}''\|_2 \le \frac{c}{\sqrt{nm}}(\rho+\sqrt{d})\left(\prod_{\ell=2}^4(\rho+\sqrt{d_{\lfloor \ell/2 \rfloor}})\right)\max(c_2c_3c_4,1) + \frac{(1+C^2)d_1}{m}c_2c_3c_4 \tag{169}$$

with probability at least $1 - \frac{1}{\mathrm{poly}(nm)} - 8nme^{-c\rho^2}$ by (164) and (165). Choose $\rho = 10\sqrt{\log(nm)}$ and recall that C is an absolute constant to obtain

$$\|\mathbf{E}''\|_2 \le \frac{c}{\sqrt{nm}}(\sqrt{\log(nm)}+\sqrt{d})\left(\prod_{\ell=2}^4(\sqrt{\log(nm)}+\sqrt{d_{\lfloor \ell/2 \rfloor}})\right)\max(c_2c_3c_4,1) + c\tfrac{d_1}{m}c_2c_3c_4 \tag{170}$$

with probability at least $1-\frac{1}{\mathrm{poly}(nm)}$. Combining Steps 1 and 2, we have

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_i\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{\Sigma}_i\mathbf{C}_{4,i}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{C}_{4,i}\right\|_2\\
&\le c(\sqrt{\log(nm)}+\sqrt{d_1})(\sqrt{\log(nm)}+\sqrt{d_2})\tfrac{\sqrt{d+d_2}+\sqrt{\log(m)}}{\sqrt{nm}}c_2c_3c_4\\
&\quad+\frac{c}{\sqrt{nm}}(\sqrt{\log(nm)}+\sqrt{d})\left(\prod_{\ell=2}^{4}(\sqrt{\log(nm)}+\sqrt{d_{\lfloor\ell/2\rfloor}})\right)\max(c_2c_3c_4,1)+\tfrac{d_1}{m}c_2c_3c_4
\end{aligned}$$

for an absolute constant c with probability at least $1-2m^{-99}-\frac{1}{\mathrm{poly}(nm)}-2e^{-90d_1}$. □

Lemma 23. *Suppose that* $\mathbf{x}$ *is a random vector with mean-zero,* $\mathbf{I}_d$*-sub-gaussian distribution over* $\mathbb{R}^d$*. Let* $\{\mathbf{x}_{i,j}\}_{i\in[n],j\in[m]}$ *be* nm *independent copies of* $\mathbf{x}$*. Denote* $\mathbf{\Sigma}_i=\frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top$ *and* $\mathbf{X}_i=[\mathbf{x}_{i,1},\ldots,\mathbf{x}_{i,m}]^\top$ *for all* $i\in[n]$*. Let* $\mathbf{z}=[z_1,\ldots,z_m]\in\mathbb{R}^m$ *be a vector whose elements are i.i.d. draws from* $\mathcal{N}(0,\sigma^2)$*, and let* $\{\mathbf{z}_i\}_{i\in[n]}$ *be* n *independent copies of* $\mathbf{z}$*. Further, let* $\mathbf{C}_{\ell,i}\in\mathbb{R}^{d\times d_{\lfloor\ell/2\rfloor}}$ *for* $\ell=2,3,5$ *be fixed matrices for* $i\in[n]$*, and let* $\mathbf{c}_{4,i}\in\mathbb{R}^d$*. Also define* $c_\ell\coloneqq\max_{i\in[n]}\|\mathbf{C}_{\ell,i}\|_2$ *for* $\ell=2,3,5$*,* $c_4\coloneqq\max_{i\in[n]}\|\mathbf{c}_{4,i}\|_2$*. Then,*

$$(i)\quad\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_i\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{X}_i\mathbf{z}_i\right\|_2\le\tfrac{c\sigma c_2c_3(\sqrt{d}+\sqrt{\log(nm)})(d_1+\log(nm))}{\sqrt{nm}}$$

$$(ii)\quad\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\top\mathbf{z}_i\mathbf{c}_{4,i}^\top\mathbf{\Sigma}_i\mathbf{C}_{5,i}^\top\right\|_2\le\tfrac{c\sigma c_4c_5(\sqrt{d}+\sqrt{\log(mn)})(\sqrt{d_1}+\sqrt{\log(nm)})\log(nm)}{\sqrt{nm}}$$

for an absolute constant c*, each with probability at least* $1-2m^{-99}-\frac{1}{\mathrm{poly}(nm)}$*.*

Proof. We only show the proof for (i) as the proof for (ii) follows by similar arguments. We argue similarly to the proof of Lemma 22. We have

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_i\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{X}_i\mathbf{z}_i\right\|_2&\le\underbrace{\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}\right\|_2}_{=:\mathbf{e}'}\\
&\quad+\underbrace{\left\|\frac{1}{nm^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}\right\|_2}_{=:\mathbf{e}''}
\end{aligned}$$

Step 1: $\|\mathbf{e}'\|_2$. Add and subtract $\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}$ to obtain

$$\begin{aligned}
\|\mathbf{e}'\|_2&\le\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}\right\|_2\\
&\quad+\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}\right\|_2\\
&\le c\sigma c_2c_3\bar{\delta}_{m,d_1}+\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}\right\|_2
\end{aligned}\tag{171}$$

where the second inequality follows with probability at least $1-e^{-90d_1}$ by Lemma 20. Next,

$$\begin{aligned}
&\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}\right\|_2\\
&\le\frac{1}{m}\sum_{j=1}^{m}\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}\right\|_2
\end{aligned}$$

By sub-gaussianity, we have that with probability at least $1-4(nm)^{-99}$, $\|\mathbf{C}_{3,i}\mathbf{x}_{i,j}\|_2 \le cc_3(\sqrt{d_1}+\sqrt{\log(nm)})$ and $\|z_{i,j}\|_2 \le c\sigma\sqrt{\log(nm)}$ for all $i\in[n], j\in[m]$. Thus, as in previous arguments, we have

$$\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top-\mathbf{I}_d\right)\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}\right\|_2 \le \tfrac{c\sigma c_2c_3(\sqrt{d}+\sqrt{\log(m)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}}{\sqrt{nm}}$$

for all $j\in[m]$ with probability at least $1-2m^{-99}-4(nm)^{-99}$, resulting in

$$\begin{aligned}\|\mathbf{e}'\|_2 &\le c\sigma c_2c_3\bar{\delta}_{m,d_1}+\frac{c\sigma c_2c_3(\sqrt{d}+\sqrt{\log(m)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}}{\sqrt{nm}}\\ &\le \frac{c'\sigma c_2c_3(\sqrt{d}+\sqrt{\log(m)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}}{\sqrt{nm}}\end{aligned} \tag{172}$$

with probability at least $1-2m^{-99}-4(nm)^{-99}$.

Step 2: $\|\mathbf{e}''\|_2$. For $\mathbf{e}''$, we again use Theorem 9. Define $\mathcal{E}_{\ell,i,j}$ and $\bar{\mathbf{x}}_{\ell,i,j}$ as in Lemma 22 for $\ell=1,2,3$ and $i\in[n]$ and $j\in[m]$. Define $\mathcal{E}_{4,i,j}=\{|z_{i,j}|\le c\sigma\sqrt{\log(nm)}\}$ and $\bar{z}_{i,j}=\chi_{\mathcal{E}_{4,i,j}}z_{i,j}$ for all $i\in[n]$ and $j\in[m]$. Define $\mathbf{s}_{i,j}=\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_{2,i}\mathbf{C}_{3,i}^\top\mathbf{x}_{i,j}z_{i,j}/m$ and $\bar{\mathbf{s}}_{i,j}=\bar{\mathbf{x}}_{1,i,j}\bar{\mathbf{x}}_{2,i,j}^\top\bar{\mathbf{x}}_{3,i,j}\bar{z}_{i,j}/m$, then we have $\mathbf{s}_{i,j}=\bar{\mathbf{s}}_{i,j}$ for all i,j with probability at least $1-\frac{1}{\mathrm{poly}(nm)}$. Also, $\|\bar{\mathbf{s}}_{i,j}\|\le B:=c\sigma c_2c_3(\sqrt{d}+\sqrt{\log(nm)})(d_1+\log(nm))\sqrt{\log(nm)}$. Next, by the symmetry of the Gaussian distribution, $\mathbb{E}[\bar{z}_{i,j}]=0$, thus $\|\mathbb{E}[\bar{\mathbf{s}}_{i,j}]\|_2=0$ by independence. Defining $\tilde{\bar{\mathbf{s}}}_{i,j}$ as in Lemma 22, we can now apply Theorem 9 as in Lemma 22 to obtain:

$$\mathbb{P}\left(\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{\bar{\mathbf{s}}}_{i,j}\right\|_2\ge\tfrac{c\sigma c_2c_3(\sqrt{d}+\sqrt{\log(nm)})(d_1+\log(nm))\sqrt{\log(nm)}}{\sqrt{nm}}\right)\le\tfrac{1}{\mathrm{poly}(nm)} \tag{173}$$

which, recalling $\|\mathbf{e}''\|_2=\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{s}_{i,j}\right\|_2$, implies

$$\mathbb{P}\big(\|\mathbf{e}''\|_2\ge\tfrac{c\sigma c_2c_3(\sqrt{d}+\sqrt{\log(nm)})(d_1+\log(nm))\sqrt{\log(nm)}}{\sqrt{nm}}\big)\le\tfrac{1}{\mathrm{poly}(nm)} \tag{174}$$

Combining (172) and (174) completes the proof. □

Lemma 24. *Suppose that $\mathbf{x}$ is a random vector with mean-zero, $\mathbf{I}_d$-sub-gaussian distribution over $\mathbb{R}^d$. Let $\{\mathbf{x}_{i,j}\}_{i\in[n],j\in[m]}$ be nm independent copies of $\mathbf{x}$. Denote $\mathbf{\Sigma}_i=\frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top$ and $\mathbf{X}_i=[\mathbf{x}_{i,1},\dots,\mathbf{x}_{i,m}]^\top$ for all $i\in[n]$. Let $\mathbf{z}=[z_1,\dots,z_m]\in\mathbb{R}^m$ be a vector whose elements are i.i.d. draws from $\mathcal{N}(0,\sigma^2)$, and let $\{\mathbf{z}_i\}_{i\in[n]}$ be n independent copies of $\mathbf{z}$. Further, let $\mathbf{C}_i\in\mathbb{R}^{d\times d_1}$ be fixed matrices for $i\in[n]$, and let $\bar{c}:=\max_{i\in[n]}\|\mathbf{C}_i\|_2$. Then,*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\top\mathbf{z}_i\mathbf{z}_i^\top\mathbf{X}_i\mathbf{C}_i\right\|_2\le\tfrac{c\sigma^2\bar{c}(\sqrt{d}+\sqrt{\log(nm)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}}{\sqrt{nm}}+\tfrac{\sigma^2\bar{c}}{m}$$

for an absolute constant c with probability at least $1-2m^{-99}-\frac{1}{\mathrm{poly}(nm)}$.

Proof. We have

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\top\mathbf{z}_i\mathbf{z}_i^\top\mathbf{X}_i\mathbf{C}_i\right\|_2 \le &\underbrace{\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\left(\tfrac{1}{m-1}\sum_{j'\neq j}\mathbf{x}_{i,j}z_{i,j}\right)z_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_i\right\|_2}_{=:\mathbf{E}'}\\ &+\underbrace{\left\|\frac{1}{nm^2}\sum_{i=1}^{n}\sum_{j=1}^{m}z_{i,j}^2\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{C}_i\right\|_2}_{=:\mathbf{E}''}\end{aligned}$$

Step 1: $\|\mathbf{E}'\|_2$**.** Note that

$$\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_{i,j}z_{i,j}\left(\tfrac{1}{m-1}\sum_{j'\neq j}z_{i,j'}\mathbf{x}_{i,j'}^\top\right)\mathbf{C}_i\right\|_2 \leq \frac{1}{m}\sum_{j=1}^{m}\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\mathbf{x}_{i,j}z_{i,j}\left(\tfrac{1}{m-1}\sum_{j'\neq j}z_{i,j'}\mathbf{x}_{i,j'}^\top\mathbf{C}_i\right)\right\|_2$$

Next, with probability at least $1-4(nm)^{-99}$, $\|\mathbf{C}_i\mathbf{x}_{i,j}\|_2 \leq c(\sqrt{d_1}+\sqrt{\log(nm)})$ and $\|z_{i,j}\|_2 \leq c\sigma\sqrt{\log(nm)}$ for all $i\in[n], j\in[m]$. Thus, by conditioning on this event as in previous arguments, we can show

$$\left\|\frac{m-1}{nm}\sum_{i=1}^{n}\mathbf{x}_{i,j}z_{i,j}\left(\tfrac{1}{m-1}\sum_{j'\neq j}z_{i,j'}\mathbf{x}_{i,j'}^\top\mathbf{C}_i\right)\right\|_2 \leq \frac{c\sigma^2\bar{c}(\sqrt{d}+\sqrt{\log(m)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}}{\sqrt{nm}}$$

for all $j\in[m]$ with probability at least $1-2m^{-99}-4(nm)^{-99}$, resulting in

$$\|\mathbf{E}'\|_2 \leq \frac{c\sigma^2\bar{c}(\sqrt{d}+\sqrt{\log(m)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}}{\sqrt{nm}} \tag{175}$$

with probability at least $1-2m^{-99}-4(nm)^{-99}$.

Step 2: $\|\mathbf{E}''\|_2$**.** Define $\mathcal{E}_{1,i,j}=\{\|\mathbf{x}_{i,j}\|\leq c(\sqrt{d}+\sqrt{\log(nm)})\}$, $\mathcal{E}_{2,i,j}=\{\|\mathbf{C}_i\mathbf{x}_{i,j}\|_2\leq c(\sqrt{d_1}+\sqrt{\log(nm)})\}$ and $\mathcal{E}_{3,i,j}=\{|z_{i,j}|\leq c\sigma\sqrt{\log(nm)}\}$ for all $i\in[n], j\in[m]$. Define $\bar{\mathbf{x}}_{1,i,j}=\chi_{\mathcal{E}_{1,i,j}}\mathbf{x}_{i,j}$, $\bar{\mathbf{x}}_{2,i,j}=\chi_{\mathcal{E}_{2,i,j}}\mathbf{C}_i^\top\mathbf{x}_{i,j}$, and $\bar{z}_{i,j}=\chi_{\mathcal{E}_{3,i,j}}z_{i,j}$ for all $i\in[n]$ and $j\in[m]$. Define $\mathbf{S}_{i,j}=z_{i,j}^2\mathbf{x}_{1,i,j}\mathbf{x}_{2,i,j}^\top\mathbf{C}_i/m$ and $\bar{\mathbf{S}}_{i,j}=\bar{z}_{i,j}^2\bar{\mathbf{x}}_{1,i,j}\bar{\mathbf{x}}_{2,i,j}^\top/m$, then we have $\mathbf{S}_{i,j}=\bar{\mathbf{S}}_{i,j}$ for all i,j with probability at least $1-\frac{1}{\mathrm{poly}(nm)}$. Also, $\|\bar{\mathbf{S}}_{i,j}\|\leq B\coloneqq c\sigma^2\bar{c}(\sqrt{d}+\sqrt{\log(nm)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}$. Note that by the law of total expectation,

$$\begin{aligned}
\|\mathbb{E}[\bar{\mathbf{S}}_{i,j}]\|_2 &= \|\mathbb{E}[\mathbf{S}_{i,j}\big|\mathcal{E}_{1,i,j}\cap\mathcal{E}_{2,i,j}\cap\mathcal{E}_{3,i,j}]\|_2\mathbb{P}(\mathcal{E}_{1,i,j},\mathcal{E}_{2,i,j},\mathcal{E}_{3,i,j})\\
&\leq \|\mathbb{E}[\mathbf{S}_{i,j}\big|\mathcal{E}_{1,i,j}\cap\mathcal{E}_{2,i,j}\cap\mathcal{E}_{3,i,j}]\|_2\mathbb{P}(\mathcal{E}_{1,i,j}\cap\mathcal{E}_{2,i,j}\cap\mathcal{E}_{3,i,j})\\
&\quad+\|\mathbb{E}[\mathbf{S}_{i,j}\big|\mathcal{E}_{1,i,j}^c\cup\mathcal{E}_{2,i,j}^c\cup\mathcal{E}_{3,i,j}^c]\|_2\mathbb{P}(\mathcal{E}_{1,i,j}^c\cup\mathcal{E}_{2,i,j}^c\cup\mathcal{E}_{3,i,j})\\
&= \|\mathbb{E}[\mathbf{S}_{i,j}]\|_2\\
&= \tfrac{\sigma^2}{m}\|\mathbf{C}_i\|_2
\end{aligned}$$

Now, defining $\tilde{\bar{\mathbf{S}}}_{i,j}$ as in Lemma 22, we can now apply Theorem 9 as in Lemma 22 to obtain for $m\geq\sigma^2\bar{c}$:

$$\mathbb{P}\left(\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{\bar{\mathbf{S}}}_{i,j}-\mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\right\|_2\geq\frac{c\sigma^2\bar{c}(\sqrt{d}+\sqrt{\log(nm)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}}{\sqrt{nm}}\right)\leq\tfrac{1}{\mathrm{poly}(nm)} \tag{176}$$

Now, note that

$$\begin{aligned}
\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\bar{\mathbf{S}}_{i,j}\right\|_2 &\leq \left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\bar{\mathbf{S}}_{i,j}-\mathbb{E}[\bar{\mathbf{S}}_{i,j}]\right\|_2+\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\bar{\mathbf{S}}_{i,j}]\right\|_2\\
&\leq \left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{\bar{\mathbf{S}}}_{i,j}-\mathbb{E}[\tilde{\bar{\mathbf{S}}}_{i,j}]\right\|_2+\tfrac{\sigma^2\bar{c}}{m}
\end{aligned} \tag{177}$$

Thus, recalling $\|\mathbf{E}''\|_2=\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{S}_{i,j}\right\|_2$, we have

$$\mathbb{P}\left(\|\mathbf{E}''\|_2\leq\frac{c\sigma^2\bar{c}(\sqrt{d}+\sqrt{\log(nm)})(\sqrt{d_1}+\sqrt{\log(nm)})\sqrt{\log(nm)}}{\sqrt{nm}}+\tfrac{\sigma^2\bar{c}}{m}\right)\leq 1-\frac{1}{\mathrm{poly}(nm)} \tag{178}$$

Combining (175) and (178) completes the proof. □

Fact 1. *Suppose* $\mathbf{x}\sim p$ *satisfies* $\mathbb{E}[\mathbf{x}]=\mathbf{0}$, $\mathrm{Cov}(\mathbf{x})=\mathbf{I}_d$ *and* $\mathbf{x}$ *is* $\mathbf{I}_d$*-sub-gaussian, as in Assumption 3. Then* $\mathbf{x}$ *is* C*-L4-L2 hypercontractive for an absolute constant* C*, that is for any* $\mathbf{u}\in\mathbb{R}^d:\|\mathbf{u}\|_2=1$*,*

$$\mathbb{E}[\langle\mathbf{u},\mathbf{x}_{i,j}\rangle^4]\leq C^2(\mathbb{E}[\langle\mathbf{u},\mathbf{x}_{i,j}\rangle^2])^2 \tag{179}$$

Lemma 25 (L4-L2 hypercontractive implication). *Suppose $\mathbf{x} \in \mathbb{R}^d$ is C-L4-L2 hypercontractive, $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, and $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}_d$. Further, let $\mathbf{C}_\ell \in \mathbb{R}^{d \times d_{\lfloor \ell/2 \rfloor}}$ for $\ell = 1,2,3,4$ be fixed matrices for $i \in [n]$, and let $c_\ell \coloneqq \max_{i \in [n]} \|\mathbf{C}_{\ell,i}\|_2$ for $\ell = 1,2,3,4$. Given scalar thresholds a_ℓ for $\ell = 1,\dots,4$, form the truncated random vectors $\bar{\mathbf{x}}_\ell \coloneqq \chi_{\|\mathbf{C}_\ell^\top \mathbf{x}\|_2 \le a_\ell} \mathbf{C}_\ell \mathbf{x}$. Then,*

$$\|\mathbb{E}[\bar{\mathbf{x}}_1 \bar{\mathbf{x}}_2^\top \bar{\mathbf{x}}_3 \bar{\mathbf{x}}_4^\top]\|_2 \le C^2 \|\mathbf{C}_1\|_2 \|\mathbf{C}_2\|_2 \|\mathbf{C}_3\|_2 \|\mathbf{C}_4\|_2 d_1. \tag{180}$$

Proof. First we note that if a random vector $\mathbf{x}$ is C-L4-L2 hypercontractive, then for any fixed matrix $\mathbf{C} \in \mathbb{R}^{d \times d_1}$, then the random vector $\mathbf{C}^\top \mathbf{x} \in \mathbb{R}^{d_1}$ is also C-L4-L2 hypercontractive, since for any unit vector $\mathbf{u}$,

$$\begin{aligned}
&\tfrac{1}{\|\mathbf{C}\mathbf{u}\|_2^4} \mathbb{E}[\langle \mathbf{u}, \mathbf{C}^\top \mathbf{x} \rangle^4] = \mathbb{E}[\langle \tfrac{\mathbf{C}\mathbf{u}}{\|\mathbf{C}\mathbf{u}\|_2}, \mathbf{x} \rangle^4] \le C^2 (\mathbb{E}[\langle \tfrac{\mathbf{C}\mathbf{u}}{\|\mathbf{C}\mathbf{u}\|_2}, \mathbf{x} \rangle^2])^2 = \tfrac{1}{\|\mathbf{C}\mathbf{u}\|_2^4} C^2 (\mathbb{E}[\langle \mathbf{u}, \mathbf{C}^\top \mathbf{x} \rangle^2])^2 \\
&\implies \mathbb{E}[\langle \mathbf{u}, \mathbf{C}^\top \mathbf{x} \rangle^4] \le C^2 (\mathbb{E}[\langle \mathbf{u}, \mathbf{C}^\top \mathbf{x} \rangle^2])^2
\end{aligned}$$

Also, if the random vector $\mathbf{x}$ is C-L4-L2 hypercontractive then the truncated random vector $\bar{\mathbf{x}} \coloneqq \chi_{\|\mathbf{x}\|_2 \le a} \|\mathbf{x}\|_2$ is also C-L4-L2 hypercontractive. To see this, observe that by the law of total expectation,

$$\mathbb{E}\big[\langle \mathbf{u}, \bar{\mathbf{x}} \rangle^4\big] = \mathbb{E}\big[\langle \mathbf{u}, \chi_{\|\mathbf{x}\|_2 \le a} \mathbf{x} \rangle^4\big] = \mathbb{E}\big[\langle \mathbf{u}, \mathbf{x} \rangle^4 \big| \|\mathbf{x}\|_2 \le a\big] \mathbb{P}(\|\mathbf{x}\|_2 \le a) \le \mathbb{E}[\langle \mathbf{u}, \mathbf{x} \rangle^4] \le C^2 (\mathbb{E}[\langle \mathbf{u}, \mathbf{x} \rangle^2])^2 \tag{181}$$

So we have that the truncated random vectors $\{\bar{\mathbf{x}}_h\}_{h=1}^4$ are C-L4-L2 hypercontractive. Next, pick some $\mathbf{u} \in \mathbb{R}^{d_1} : \|\mathbf{u}\|_2 \le 1$ and $\mathbf{v} \in \mathbb{R}^{d_4} : \|\mathbf{v}\| \le 1$. By the Cauchy-Schwarz inequality and C-L4-L2 hypercontractivity, we have

$$\begin{aligned}
&\mathbb{E}[\mathbf{u}^\top \bar{\mathbf{x}}_1 \bar{\mathbf{x}}_2^\top \bar{\mathbf{x}}_3 \bar{\mathbf{x}}_4^\top \mathbf{v}] \\
&\quad \le (\mathbb{E}[(\mathbf{u}^\top \bar{\mathbf{x}}_1 \bar{\mathbf{x}}_4^\top \mathbf{v})^2] \mathbb{E}[(\bar{\mathbf{x}}_2^\top \bar{\mathbf{x}}_3)^2])^{1/2} \\
&\quad \le (\mathbb{E}[(\mathbf{u}^\top \bar{\mathbf{x}}_1)^4] \mathbb{E}[(\bar{\mathbf{x}}_4^\top \mathbf{v})^4])^{1/4} (\mathbb{E}[\mathrm{Tr}(\bar{\mathbf{x}}_2 \bar{\mathbf{x}}_3^\top)^2])^{1/2} \\
&\quad \le C (\mathbb{E}[(\mathbf{u}^\top \bar{\mathbf{x}}_1)^2] \mathbb{E}[(\bar{\mathbf{x}}_4^\top \mathbf{v})^2])^{1/2} \left(\mathbb{E}\left[\left(\sum_{\ell=1}^d \mathbf{e}_\ell^\top \bar{\mathbf{x}}_2 \bar{\mathbf{x}}_3^\top \mathbf{e}_\ell \right)^2 \right] \right)^{1/2} \\
&\quad \le C (\mathbb{E}[(\mathbf{u}^\top \bar{\mathbf{x}}_1)^2] \mathbb{E}[(\bar{\mathbf{x}}_4^\top \mathbf{v})^2])^{1/2} \left(\sum_{\ell,\ell'} \mathbb{E}\left[\mathbf{e}_\ell^\top \bar{\mathbf{x}}_2 \bar{\mathbf{x}}_3^\top \mathbf{e}_\ell \mathbf{e}_{\ell'}^\top \bar{\mathbf{x}}_2 \bar{\mathbf{x}}_3^\top \mathbf{e}_{\ell'} \right] \right)^{1/2} \\
&\quad \le C (\mathbb{E}[(\mathbf{u}^\top \bar{\mathbf{x}}_1)^2] \mathbb{E}[(\bar{\mathbf{x}}_4^\top \mathbf{v})^2])^{1/2} \left(\sum_{\ell,\ell'} \left(\mathbb{E}\left[(\mathbf{e}_\ell^\top \bar{\mathbf{x}}_2)^4 \right] \mathbb{E}[(\bar{\mathbf{x}}_3^\top \mathbf{e}_\ell)^4] \mathbb{E}[(\mathbf{e}_{\ell'}^\top \bar{\mathbf{x}}_2)^4] \mathbb{E}[(\bar{\mathbf{x}}_3^\top \mathbf{e}_{\ell'})^4] \right)^{1/4} \right)^{1/2} \\
&\quad \le C^2 (\mathbb{E}[(\mathbf{u}^\top \bar{\mathbf{x}}_1)^2] \mathbb{E}[(\bar{\mathbf{x}}_4^\top \mathbf{v})^2])^{1/2} \\
&\qquad \times \left(\sum_{\ell,\ell'} \left((\mathbb{E}[(\mathbf{e}_\ell^\top \bar{\mathbf{x}}_2)^2])^2 (\mathbb{E}[(\bar{\mathbf{x}}_3^\top \mathbf{e}_\ell)^2])^2 (\mathbb{E}[(\mathbf{e}_{\ell'}^\top \bar{\mathbf{x}}_2)^2])^2 (\mathbb{E}[(\bar{\mathbf{x}}_3^\top \mathbf{e}_{\ell'})^2])^2 \right)^{1/4} \right)^{1/2}
\end{aligned} \tag{182}$$

where $\mathbf{e}_\ell$ is the ℓ-th standard basis vector in $\mathbb{R}^{d_1}$. Note that by the law of total expectation and the nonnegativity of $\mathbf{U}^\top \mathbf{C}_1^\top \mathbf{x} \mathbf{x}^\top \mathbf{C}_1 \mathbf{U}$,

$$\begin{aligned}
\mathbb{E}[(\bar{\mathbf{x}}_1^\top \mathbf{u})^2] &= \mathbb{E}\left[\mathbf{u}^\top \mathbf{C}_1^\top \mathbf{x} \mathbf{x}^\top \mathbf{C}_1 \mathbf{u} \big| \|\mathbf{C}_1^\top \mathbf{x}\|_2 \le a \right] \mathbb{P}(\|\mathbf{C}_1^\top \mathbf{x}\|_2 \le a) \\
&\le \mathbb{E}[\mathbf{u}^\top \mathbf{C}_1^\top \mathbf{x} \mathbf{x}^\top \mathbf{C}_1 \mathbf{u}] = \mathbf{u}^\top \mathbf{C}_1^\top \mathbf{C}_1 \mathbf{u} \le \|\mathbf{C}_1\|_2^2
\end{aligned}$$

Therefore, applying the same logic for $\mathbb{E}[(\mathbf{e}_\ell^\top \bar{\mathbf{x}}_2)^2]$, $\mathbb{E}[(\mathbf{e}_\ell^\top \bar{\mathbf{x}}_3)^2]$, and $\mathbb{E}[(\mathbf{e}_\ell^\top \bar{\mathbf{x}}_4)^2]$, and using (182), we obtain

$$\mathbb{E}[\mathbf{u}^\top \bar{\mathbf{x}}_1 \bar{\mathbf{x}}_2^\top \bar{\mathbf{x}}_3 \bar{\mathbf{x}}_4^\top \mathbf{v}] \le C^2 \|\mathbf{C}_1\|_2 \|\mathbf{C}_4\|_2 \left(\sum_{\ell,\ell'} \|\mathbf{C}_2\|_2^2 \|\mathbf{C}_3\|_2^2 \right)^{1/2} = C^2 \|\mathbf{C}_1\|_2 \|\mathbf{C}_2\|_2 \|\mathbf{C}_3\|_2 \|\mathbf{C}_4\|_2 d_1$$

Repeating this argument over all unit vectors $\mathbf{u}, \mathbf{v}$ completes the proof. □

Next, we characterize the diversity of the inner loop-updated heads for both ANIL and FO-ANIL. Note that now we are analyzing ANIL and FO-ANIL specifically rather than studying generic matrix concentration.

Lemma 26. *Let $\mathbf{w}_{t,i}$ be the inner loop-updated head for the i-th task at iteration t for ANIL and FO-ANIL for all $i \in [n]$. Define $\mu^2 \coloneqq \sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)$ and $L^2 \coloneqq \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)$. Assume $\|\mathbf{\Delta}_t\|_2 \leq \frac{1}{10}$ and Assumption 1, 2, and 3 hold. Then*

$$\sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \leq L^2 \coloneqq 2\left(\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_* + \delta_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma)\right)^2 \tag{183}$$

$$\begin{aligned}\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \geq \mu^2 \coloneqq\ & 0.9\alpha E_0\mu_*^2 - 2.2\sqrt{\alpha}\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\eta_* \\ & - 2\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\bar{\delta}_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma) \\ & - 2.2\sqrt{\alpha}\bar{\delta}_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_* + \sqrt{\alpha}\sigma)L_{\max}\end{aligned} \tag{184}$$

with probability at least $1 - 4n^{-99} - 6e^{-90k}$.

Proof. Note that $\mathbf{w}_{t,i}$ can be written as:

$$\mathbf{w}_{t,i} = \mathbf{w}_t - \alpha\mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t\mathbf{w}_t + \alpha\mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_*\mathbf{w}_{*,t,i} = \mathbf{r} + \mathbf{s}_i + \mathbf{p}_{1,i} + \mathbf{p}_{2,i} + \mathbf{p}_{3,i} \tag{185}$$

where $\mathbf{r} = \mathbf{\Delta}_t\mathbf{w}_t$, $\mathbf{s}_i = \alpha\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i}$, $\mathbf{p}_{1,i} \coloneqq \alpha(\mathbf{B}_t^\top\mathbf{B}_t - \mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t)\mathbf{w}_t$, $\mathbf{p}_{2,i} \coloneqq -\alpha(\mathbf{B}_t^\top\mathbf{B}_* - \mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_*)\mathbf{w}_{*,t,i}$, and $\mathbf{p}_{3,i} \coloneqq \frac{\alpha}{m_{in}}\mathbf{B}_t^\top(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}$ for all $i \in [n]$ (for ease of notation we drop the iteration index t). Note that since $\|\mathbf{\Delta}_t\|_2 \leq \frac{1}{10}$, $\|\mathbf{B}_t\|_2 \leq \frac{\sqrt{11/10}}{\sqrt{\alpha}}$. As a result, for any $i \in [n]$, from Lemma 20 we have

$$\|\mathbf{p}_{1,i}\|_2 \leq 1.1\|\mathbf{w}_t\|_2\delta_{m_{in},k}, \quad \|\mathbf{p}_{2,i}\|_2 \leq \sqrt{1.1\alpha}L_{\max}\delta_{m_{in},k}, \quad \|\mathbf{p}_{3,i}\|_2 \leq \sqrt{1.1\alpha}\sigma\delta_{m_{in},k} \tag{186}$$

each with probability at least $1 - 2n^{-100}$. Thus, all of these events happen simultaneously with probability at least $1 - 6n^{-100}$ via a union bound. Further, a union bound over all $i \in [n]$ shows that $A \coloneqq \left\{\cap_{i\in[n]}\left\{\|\mathbf{p}_{1,i}\|_2 \leq 1.1\|\mathbf{w}_t\|_2\delta_{m_{in},k} \cap \|\mathbf{p}_{2,i}\|_2 \leq \sqrt{1.1\alpha}L_{\max}\delta_{m_{in},k} \cap \|\mathbf{p}_{3,i}\|_2 \leq \sqrt{1.1\alpha}\sigma\delta_{m_{in},k}\right\}\right\}$ occurs with probability at least $1 - 6n^{-99}$. Thus by the triangle inequality, a

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{p}_{1,i} + \mathbf{p}_{2,i} + \mathbf{p}_{3,i}\right\|_2 \leq 1.1\delta_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma) \tag{187}$$

with probability at least $1 - 6n^{-99}$, and

$$\left\|\frac{1}{n}\sum_{i=1}^n (\mathbf{r} + \mathbf{s}_i)(\mathbf{r} + \mathbf{s}_i)^\top\right\|_2 \leq \|\mathbf{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^2 + 2.2\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* + 1.1\alpha L_*^2 \tag{188}$$

So,

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right\|_2 \\
&\leq \left\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}+\mathbf{s}_i)(\mathbf{r}+\mathbf{s}_i)^\top\right\|_2 + 2\left\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}+\mathbf{s}_i)(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})^\top\right\|_2 \\
&\qquad + \left\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})^\top\right\|_2 \\
&\leq \|\mathbf{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^2 + 2.2\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* + 1.1\alpha L_*^2 + 2\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{r}(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})^\top\right\|_2 \\
&\quad + 2\alpha\|\mathbf{B}_t^\top\mathbf{B}_*\|_2\left\|\frac{1}{n}\mathbf{W}_{*,t}^\top\begin{bmatrix}\vdots\\(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})^\top\\\vdots\end{bmatrix}\right\|_2 + \max_{i\in[n]}\|\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i}\|_2^2 \qquad (189)\\
&\leq \|\mathbf{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2^2 + 2.2\sqrt{\alpha}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\eta_* + 1.1\alpha L_*^2 \\
&\quad + 2.2\|\mathbf{\Delta}_t\|_2^2\|\mathbf{w}_t\|_2\delta_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma) \\
&\quad + 2.2\sqrt{\alpha}L_*\delta_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma) + 1.1^2(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma)^2\delta_{m_{in},k}^2 \\
&\leq 2\left(\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_* + \delta_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma)\right)^2 \qquad (190)
\end{aligned}$$

where $\mathbf{W}_{*,t} = [\mathbf{w}_{*,t,1}, \ldots, \mathbf{w}_{*,t,n}]^\top$, (189) follows from the triangle inequality, and (190) follows with probability at least $1 - 6n^{-99}$ from the discussion above.

We make an analogous argument to lower bound $\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)$. This time, we only need to bound first-order products of the $\mathbf{p}$ matrices, which concentrate around zero as n becomes large. So now we are able to obtain finite-sample dependence on $\bar{\delta}_{m_{in},k}$ (which decays with $\frac{1}{\sqrt{n}}$) instead of $\delta_{m_{in},k}$ (which does not), as follows.

$$\begin{aligned}
\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) &= \sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}+\mathbf{s}_i)(\mathbf{r}+\mathbf{s}_i)^\top + (\mathbf{r}+\mathbf{s}_i)(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})^\top\right. \\
&\qquad + (\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})(\mathbf{r}+\mathbf{s}_i)^\top \\
&\qquad \left. + (\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})^\top\right) \\
&\geq \sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}+\mathbf{s}_i)(\mathbf{r}+\mathbf{s}_i)^\top\right) - 2\left\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}+\mathbf{s}_i)(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})^\top\right\|_2 \\
&\geq \sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}_i\mathbf{s}_i^\top\right) - 2\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{r}\mathbf{s}_i^\top\right\|_2 - 2\left\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}+\mathbf{s}_i)(\mathbf{p}_{1,i}+\mathbf{p}_{2,i}+\mathbf{p}_{3,i})^\top\right\|_2 \\
&\geq 0.9\alpha E_0\mu_*^2 - 2.2\sqrt{\alpha}\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\eta_* \\
&\quad - 2\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\bar{\delta}_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma) \\
&\quad - 2.2\sqrt{\alpha}\bar{\delta}_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_* + \sqrt{\alpha}\sigma)L_{\max}
\end{aligned}$$

where the last inequality follows with probability at least $1 - 6e^{-90k}$. □

E.2. FO-ANIL

For FO-ANIL, inner loop update for the head of the i-th task on iteration t is given by:

$$\begin{aligned}
\mathbf{w}_{t,i} &= \mathbf{w}_t - \alpha\nabla_{\mathbf{w}}\hat{\mathcal{L}}_i(\mathbf{B}_t, \mathbf{w}_t, \mathcal{D}_i^{in}) \\
&= (\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t)\mathbf{w}_t + \alpha\mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_*\mathbf{w}_{*,t,i} + \tfrac{\alpha}{m_{in}}\mathbf{B}_t^\top(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}. \qquad (191)
\end{aligned}$$

The outer loop updates for the head and representation are:

$$\begin{aligned}\mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{\beta}{n}\sum_{i=1}^{n}\nabla_{\mathbf{w}}\hat{\mathcal{L}}_i(\mathbf{B}_t, \mathbf{w}_{t,i}, \mathcal{D}_{t,i}^{out}) \\ &= \mathbf{w}_t - \frac{\beta}{n}\sum_{i=1}^{n}\left(\mathbf{B}_t^\top\boldsymbol{\Sigma}_{t,i}^{out}\mathbf{B}_t\mathbf{w}_{t,i} - \mathbf{B}_t^\top\boldsymbol{\Sigma}_{t,i}^{out}\mathbf{B}_*\mathbf{w}_{*,i} - \tfrac{2}{m_{out}}\mathbf{B}_t^\top(\mathbf{X}_{t,i}^{out,g})^\top\mathbf{z}_{t,i}^{out,g}\right) \qquad (192)\\ \mathbf{B}_{t+1} &= \mathbf{B}_t - \frac{\beta}{n}\sum_{i=1}^{n}\nabla_{\mathbf{B}}\hat{\mathcal{L}}_i(\mathbf{B}_t, \mathbf{w}_{t,i}, \mathcal{D}_{t,i}^{out}) \\ &= \mathbf{B}_t - \frac{\beta}{n}\sum_{i=1}^{n}\left(\boldsymbol{\Sigma}_{t,i}^{out}\mathbf{B}_t\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top - \boldsymbol{\Sigma}_{t,i}^{out}\mathbf{B}_*\mathbf{w}_{*,t,i}\mathbf{w}_{t,i}^\top - \tfrac{2}{m_{out}}(\mathbf{X}_{t,i}^{out})^\top\mathbf{z}_{t,i}^{out}\mathbf{w}_{t,i}^\top\right) \qquad (193)\end{aligned}$$

Lemma 27 (FO-ANIL, Finite samples $A_1(t+1)$)**.** *For any t, suppose that $A_2(s), A_3(s)$ and $A_4(s)$ occur for all $s \in [t]$. Then*

$$\|\mathbf{w}_{t+1}\|_2 \le \tfrac{1}{10}\sqrt{\alpha}E_0\min(1, \tfrac{\mu_*^2}{\eta_*^2})\eta_* \qquad (194)$$

with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$.

Proof. The proof follows similar structure as in the analogous proof for the infinite-sample case. Recall the outer loop updates for ANIL (here we replace t with s):

$$\begin{aligned}\mathbf{w}_{s+1} = {}& (\mathbf{I}_k - \beta\mathbf{B}_s^\top\mathbf{B}_s(\mathbf{I}_k - \alpha\mathbf{B}_s^\top\mathbf{B}_s))\mathbf{w}_s + \beta(\mathbf{I}_k - \alpha\mathbf{B}_s^\top\mathbf{B}_s)\mathbf{B}_s^\top\mathbf{B}_*\frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{*,s,i} \\ &+ \alpha\beta\mathbf{B}_s^\top\mathbf{B}_s\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{B}_s^\top\mathbf{B}_s - \mathbf{B}_s^\top\boldsymbol{\Sigma}_{s,i}^{in}\mathbf{B}_s\right)\mathbf{w}_s - \alpha\beta\mathbf{B}_s^\top\mathbf{B}_s\frac{1}{n}\sum_{i=1}^{n}(\mathbf{B}_s^\top\mathbf{B}_* - \mathbf{B}_s^\top\boldsymbol{\Sigma}_{s,i}^{in}\mathbf{B}_*)\mathbf{w}_{*,s,i} \\ &+ \alpha\beta\mathbf{B}_s^\top\mathbf{B}_s\frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_s^\top(\mathbf{X}_{s,i}^{in})^\top\mathbf{z}_{s,i}^{in} + \frac{\beta}{n}\sum_{i=1}^{n}(\mathbf{B}_s^\top\mathbf{B}_s - \mathbf{B}_s^\top\boldsymbol{\Sigma}_{s,i}^{out}\mathbf{B}_s)\mathbf{w}_{s,i} \\ &- \frac{\beta}{n}\sum_{i=1}^{n}(\mathbf{B}_s^\top\mathbf{B}_* - \mathbf{B}_s^\top\boldsymbol{\Sigma}_{s,i}^{out}\mathbf{B}_*)\mathbf{w}_{*,s,i} + \frac{2\beta}{nm_{out}}\sum_{i=1}^{n}\mathbf{B}_s^\top(\mathbf{X}_{s,i}^{out})^\top\mathbf{z}_{s,i}^{out} \qquad (195)\end{aligned}$$

Note that $\bigcup_{s=0}^{t}A_3(s)$ implies $\sigma_{\max}(\mathbf{B}_s^\top\mathbf{B}_s) \le \frac{1+\|\boldsymbol{\Delta}_s\|_2}{\alpha} < \frac{1.1}{\alpha}$ for all $s \in \{0, \dots, t+1\}$. Also, we can straightforwardly use Lemma 20 with the Cauchy-Schwartz inequality to obtain, for some absolute constant c,

$$\begin{aligned}\left\|\alpha\beta\mathbf{B}_s^\top\mathbf{B}_s\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{B}_s^\top\mathbf{B}_s - \mathbf{B}_s^\top\boldsymbol{\Sigma}_{s,i}^{in}\mathbf{B}_s\right)\mathbf{w}_t\right\|_2 &\le c\tfrac{\beta}{\alpha}\|\mathbf{w}_s\|_2\bar{\delta}_{m_{in},k} \\ \left\|\alpha\beta\mathbf{B}_s^\top\mathbf{B}_s\frac{1}{n}\sum_{i=1}^{n}(\mathbf{B}_s^\top\mathbf{B}_* - \mathbf{B}_t^\top\boldsymbol{\Sigma}_{s,i}^{in}\mathbf{B}_s)\mathbf{w}_{*,s,i}\right\|_2 &\le c\tfrac{\beta}{\sqrt{\alpha}}L_{\max}\bar{\delta}_{m_{in},k} \qquad (196)\end{aligned}$$

$$\begin{aligned}\left\|\tfrac{\beta}{n}\sum_{i=1}^{n}(\mathbf{B}_s^\top\mathbf{B}_s - \mathbf{B}_s^\top\boldsymbol{\Sigma}_{s,i}^{out}\mathbf{B}_s)\mathbf{w}_{s,i}\right\|_2 &\le c\tfrac{\beta}{\alpha}\bar{\delta}_{m_{out},k}\max_{i\in[n]}\|\mathbf{w}_{s,i}\|_2 \\ &\le c\tfrac{\beta}{\alpha}\bar{\delta}_{m_{out},k}(\|\boldsymbol{\Delta}_s\|_2\|\mathbf{w}_s\|_2 + \sqrt{\alpha}L_{\max} + \delta_{m_{in},k}\|\mathbf{w}_s\|_2 \\ &\quad + \sqrt{\alpha}\delta_{m_{in},k}L_{\max} + \sqrt{\alpha}\sigma\delta_{m_{in},k}) \\ &\le c\tfrac{\beta}{\alpha}\bar{\delta}_{m_{out},k}(\|\boldsymbol{\Delta}_s\|_2\|\mathbf{w}_s\|_2 + c'\sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma\delta_{m_{in},k}) \qquad (197)\\ \left\|\frac{\beta}{n}\sum_{i=1}^{n}(\mathbf{B}_s^\top\mathbf{B}_* - \mathbf{B}_s^\top\boldsymbol{\Sigma}_{s,i}^{out}\mathbf{B}_*)\mathbf{w}_{*,s,i}\right\|_2 &\le c\tfrac{\beta}{\sqrt{\alpha}}L_{\max}\bar{\delta}_{m_{out},k} \\ \left\|\tfrac{2\beta}{nm_{out}}\sum_{i=1}^{n}\mathbf{B}_s^\top(\mathbf{X}_{s,i}^{out})^\top\mathbf{z}_{s,i}^{out}\right\|_2 &\le c\tfrac{\beta}{\sqrt{\alpha}}\sigma\bar{\delta}_{m_{out},k}\end{aligned}$$

using that $\delta_{m_{in},k} < 1$ and $\|\mathbf{w}_s\|_2 \leq c\sqrt{\alpha}\eta_*$ in (197). Thus using (195) and the Cauchy-Schwarz and triangle inequalities, we have for an absolute constant c:

$$\begin{aligned}\|\mathbf{w}_{s+1}\|_2 &\leq (1 + c\tfrac{\beta}{\alpha}\|\mathbf{\Delta}_s\|_2)\|\mathbf{w}_s\|_2 + c\tfrac{\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2\eta_* + c\tfrac{\beta}{\alpha}\|\mathbf{w}_s\|_2\bar{\delta}_{m_{in},k} \\ &\quad + c\tfrac{\beta}{\sqrt{\alpha}}\bar{\delta}_{m_{in},k}(L_{\max} + \sigma) + c\tfrac{\beta}{\alpha}\bar{\delta}_{m_{out},k}\sqrt{\alpha}L_{\max} + c\tfrac{\beta}{\sqrt{\alpha}}L_{\max}\bar{\delta}_{m_{out},k} + c\tfrac{\beta}{\sqrt{\alpha}}\sigma\bar{\delta}_{m_{out},k} \\ &\leq (1 + c\tfrac{\beta}{\alpha}\|\mathbf{\Delta}_s\|_2)\|\mathbf{w}_s\|_2 + t\tfrac{\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2\eta_* + t\tfrac{\beta}{\sqrt{\alpha}}\left(L_{\max} + \sigma\right)\left(\bar{\delta}_{m_{in},k} + \bar{\delta}_{m_{out},k}\right) \\ &= (1 + c\tfrac{\beta}{\alpha}\|\mathbf{\Delta}_s\|_2)\|\mathbf{w}_s\|_2 + c\tfrac{\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2\eta_* + \tfrac{\beta}{\sqrt{\alpha}}\zeta_1 \end{aligned} \tag{198}$$

using $\|\mathbf{w}_{s,i}\|_2 \leq \sqrt{\alpha}L_{\max}$, where $\zeta_1 = c(L_{\max} + \sigma)(\bar{\delta}_{m_{in},k} + \bar{\delta}_{m_{out},k})$. Thus, by Lemma 3, we have

$$\|\mathbf{w}_{t+1}\|_2 \leq c\tfrac{\beta}{\sqrt{\alpha}}\sum_{s=1}^{t}(\|\mathbf{\Delta}_s\|_2\eta_* + \zeta_1)\left(1 + 2\sum_{r=s}^{t}\frac{\beta}{\alpha}\|\mathbf{\Delta}_r\|_2\right) \tag{199}$$

Next, let $\rho := 1 - 0.5\beta\alpha E_0\mu_*^2$ and $\zeta_2 = O\big((L_{\max} + \sigma)^2\bar{\delta}_{m_{out},k} + (L_{\max}^2 + L_{\max}\sigma)(\bar{\delta}_{m_{in},k} + \delta_{m_{in},k}^2) + \sigma^2\delta_{m_{in},k}^2 + \beta\alpha(L_{\max} + \sigma)^4\bar{\delta}_{m_{out},d}^2\big)$ as defined in (219). By $\bigcup_{r=0}^{s} A_2(r)$, we have

$$\begin{aligned}\|\mathbf{\Delta}_{s+1}\|_2 &\leq \rho\|\mathbf{\Delta}_s\|_2 + c\alpha^2\beta^2L_*^4\,\mathrm{dist}_s^2 + \beta\alpha\zeta_2 \\ &\leq \rho^2\|\mathbf{\Delta}_{s-1}\|_2 + \rho(c\alpha^2\beta^2\,\mathrm{dist}_{s-1}^2 + \beta\alpha\zeta_2) + c\alpha^2\beta^2\,\mathrm{dist}_s^2 + \beta\alpha\zeta_2 \\ &\vdots \\ &\leq \rho^{s+1}\|\mathbf{\Delta}_0\|_2 + \sum_{r=0}^{s}\rho^{s-r}(c\alpha^2\beta^2L_*^4\,\mathrm{dist}_r^2 + \beta\alpha\zeta_2) \\ &= \sum_{r=0}^{s}\rho^{s-r}(c\alpha^2\beta^2L_*^4\,\mathrm{dist}_r^2 + \beta\alpha\zeta_2) \end{aligned} \tag{200}$$

since $\|\mathbf{I} - \alpha\mathbf{B}_0^\top\mathbf{B}_0\|_2 = 0$ by choice of initialization. Now, we have that $\mathrm{dist}_s \leq \rho^s + \varepsilon$ for all $s \in \{0, ..., t\}$ by $\bigcup_{s=0}^{t} A_5(s)$. Thus, for any $s \in \{0, ..., t\}$, we have

$$\begin{aligned}\|\mathbf{\Delta}_{s+1}\|_2 &\leq \sum_{r=0}^{s}\rho^{s-r}(c\alpha^2\beta^2L_*^4\,\mathrm{dist}_r^2 + \beta\alpha\zeta_2) \\ &\leq \sum_{r=0}^{s}\rho^{s-r}(c\alpha^2\beta^2L_*^4(2\rho^{2r} + 2\varepsilon^2) + \beta\alpha\zeta_2) \\ &\leq 2c\alpha^2\beta^2L_*^4\sum_{r=0}^{s}\rho^{s-r}\rho^{2r} + 2c\alpha^2\beta^2L_*^4\sum_{r=0}^{s}\rho^{s-r}\varepsilon^2 + \beta\alpha\sum_{r=0}^{s}\rho^{s-r}\zeta_2 \\ &\leq 2c\rho^s\tfrac{\alpha^2\beta^2L_*^4}{1-\rho} + (2c\alpha^2\beta^2L_*^4\varepsilon^2 + \beta\alpha\zeta_2)\tfrac{1}{1-\rho} \\ &\leq 2c\rho^s\beta\alpha L_*^2\kappa_*^2/E_0 + 2c\varepsilon^2\beta\alpha L_*^2\kappa_*^2/E_0 + \zeta_2/(E_0\mu_*^2) \\ &=: \epsilon_s \end{aligned} \tag{201}$$

Now, applying equation (199) yields

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2 &\leq \frac{\beta}{\sqrt{\alpha}}\sum_{s=1}^{t}(\epsilon_s\eta_* + \zeta_1)\left(1 + 2\sum_{r=s}^{t}\frac{\beta}{\alpha}\epsilon_r\right) \\ &\leq \frac{\beta}{\sqrt{\alpha}}\sum_{s=1}^{t}(\epsilon_s\eta_* + \zeta_1)\left(1 + 4c\beta\rho^s\kappa_*^4/(\alpha E_0^2) + 4c(t-s)\beta^2\varepsilon^2\kappa_*^2L_*^2/E_0 + 2(t-s)\beta\zeta_2/(\alpha E_0\mu_*^2)\right) \end{aligned} \tag{202}$$

$$\leq \frac{\beta}{\sqrt{\alpha}}\sum_{s=1}^{t}(\epsilon_s\eta_* + \zeta_1)\left(1 + 4c\beta\kappa_*^4/(\alpha E_0^2) + 4cT\beta^2\varepsilon^2\kappa_*^2L_*^2/E_0 + 2T\beta\zeta_2/(\alpha E_0\mu_*^2)\right) \tag{203}$$

where (202) follows by plugging in the definition of ϵ_r and using the sum of a geometric series. In order for the RHS of (203) to be at most $\frac{1}{10}\sqrt{\alpha}E_0\min(1,\frac{\mu_*^2}{\eta_*^2})\eta_*$ as desired, we can ensure that $\left(4c\beta\kappa_*^4/(\alpha E_0^2)+4cT\beta^2\varepsilon^2\kappa_*^2L_*^2/E_0+2T\beta\zeta_2/(\alpha E_0\mu_*^2)\right) \leq 1$ for all s and $\frac{\beta}{\sqrt{\alpha}}\sum_{s=1}^t(\epsilon_s\eta_*+\zeta_1) \leq \frac{1}{10}\sqrt{\alpha}E_0\min(1,\frac{\mu_*^2}{\eta_*^2})\eta_*$. To satisfy the first condition, it is sufficient to have

$$\begin{aligned}\beta &\leq c'\tfrac{\alpha E_0^2}{\kappa_*^4}\\ \varepsilon^2 &\leq c'\tfrac{E_0}{T\beta^2\kappa_*^2L_*^2}\\ \zeta_2 &\leq c'\tfrac{\alpha E_0\mu_*^2}{T\beta}\end{aligned}$$

For the second condition, it is sufficient to have

$$\begin{aligned}\beta &\leq c\alpha E_0^3\kappa_*^{-4}\min(1,\tfrac{\mu_*^2}{\eta_*^2})\\ \zeta_1 &\leq c'\tfrac{\kappa_*^4\eta_*}{TE_0^2}\\ \sum_{s=1}^t\epsilon_s &\leq c'\tfrac{\kappa_*^4}{E_0^2}\\ \implies \varepsilon^2 &\leq c''\tfrac{1}{T\beta\alpha\mu_*^2E_0}\\ \zeta_2 &\leq c''\tfrac{L_*^2\kappa_*^2}{TE_0}\end{aligned} \tag{204}$$

However, for Corollary 3, will need a tighter bound on ζ_2, namely $\zeta_2 \leq \frac{c'E_0\mu_*^2}{T}$. In summary, the tightest bounds are:

$$\beta \leq c\alpha E_0^3\kappa_*^{-4}\min(1,\tfrac{\mu_*^2}{\eta_*^2}) \tag{205}$$
$$\varepsilon^2 \leq c\tfrac{1}{T\beta\alpha\mu_*^2E_0} \tag{206}$$
$$\zeta_1 \leq c\tfrac{\kappa_*^4\eta_*}{TE_0^2} \tag{207}$$
$$\zeta_2 \leq c\tfrac{E_0\mu_*^2}{T} \tag{208}$$

To determine when these conditions hold, we must recall the scaling of $\varepsilon,\zeta_1,\zeta_2$.

$$\begin{aligned}\varepsilon &= O(\tfrac{L_{\max}(L_{\max}+\sigma)}{\mu_*^2}\bar{\delta}_{m_{out},d})\\ \zeta_1 &= O((L_{\max}+\sigma)(\bar{\delta}_{m_{in},k}+\bar{\delta}_{m_{out},k}))\\ \zeta_2 &= O\big(\left((L_{\max}^2+L_{\max}\sigma)(\bar{\delta}_{m_{in},k}+\bar{\delta}_{m_{out},k}+\delta_{m_{in},k}^2)+\sigma^2\delta_{m_{in},k}^2\right)\\ &\quad+\beta\alpha\bar{\delta}_{m_{out},d}^2(L_{\max}+\sigma)^2(L_{\max}+\sigma\delta_{m_{in},k})^2\big)\end{aligned}$$

Thus, in order to satisfy (205)-(208), we can choose:

$$m_{out} \geq c'\left(\beta\alpha\tfrac{dTE_0(L_{\max}+\sigma)^4}{n\mu_*^2}+\tfrac{T^2kE_0^2L_{\max}^2(L_{\max}+\sigma)^2}{n\mu_*^4}+\tfrac{T^2E_0^4k(L_{\max}+\sigma)^2}{n\eta_*^2\kappa_*^8}+\beta\alpha\tfrac{TL_{\max}^2(L_{\max}+\sigma)^2E_0d}{n\mu_*^2}\right).$$

Recalling that $L_{\max}\leq c\sqrt{k}L_*$, $\beta\leq\alpha\kappa_*^{-4}$, and $\alpha\leq\frac{1}{L_{\max}+\sigma}$, we see that our choice of m_{out} as

$$m_{out} \geq c\left(\frac{Tdk}{n\kappa_*^2}+\frac{Tdk\sigma^2}{nL_*^2\kappa_*^2}+\frac{T^2k^3\kappa_*^4}{n}+\frac{T^2k^3\sigma^4}{n\mu_*^4}+\frac{k\mu_*^2}{\eta_*^2\kappa_*^6}+\frac{k\sigma^2}{\eta_*^2\kappa_*^8}\right)$$

is sufficient, where we have treated E_0 as a constant. For m_{in}, we can choose:

$$m_{in} \geq c'\frac{T(k+\log(n))E_0(L_{\max}+\sigma)^2}{\mu_*^2}+c'\frac{T^2kE_0^2L_{\max}^2(L_{\max}+\sigma)^2}{n\mu_*^4}+c'\frac{T^2kE_0^4(L_{\max}+\sigma)^2}{n\eta_*^2\kappa_*^8}$$

which is satisfied by

$$m_{in} \geq cT(k+\log(n))(k\kappa_*^2 + \tfrac{\sigma^2}{\mu_*^2}) + c\tfrac{T^2k^3\kappa_*^4}{n} + c\tfrac{T^2k^2\kappa_*^2\sigma^2}{\mu_*^2 n} + c\tfrac{T^2k^2(L_*^2+\sigma^2)}{\eta_*^2\kappa_*^8 n}$$

Since m_{in} and m_{out} satisfy these conditions, we have completed the proof.

□

Lemma 28 (FO-ANIL, Finite samples, $A_2(t+1)$)**.** *Suppose the conditions of Theorem 8 are satisfied and inductive hypotheses $A_1(t)$, $A_3(t)$ and $A_5(t)$ hold. Then $A_2(t+1)$ holds with high probability, i.e.*

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + c\beta^2\alpha^2 L_*^4 \operatorname{dist}_t^2 + \beta\alpha\zeta_2 \tag{209}$$

for an absolute constant c and $\zeta_2 = O\big((L_{\max}+\sigma)^2\bar{\delta}_{m_{out},k} + (L_{\max}^2 + L_{\max}\sigma)(\bar{\delta}_{m_{in},k} + \delta^2_{m_{in},k}) + \sigma^2\delta^2_{m_{in},k} + \beta\alpha(L_{\max}+\sigma)^4\bar{\delta}^2_{m_{out},d}\big)$*, with probability at least* $1 - \frac{1}{\operatorname{poly}(n)}$.

Proof. Note that we can write:

$$\begin{aligned}
\mathbf{B}_{t+1} &= \mathbf{B}_t - \beta\bar{\mathbf{\Delta}}_t \tfrac{1}{n}\sum_{i=1}^n (\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})((\mathbf{I}_k - \alpha\mathbf{B}_t^\top\mathbf{B}_t)\mathbf{w}_t + \alpha\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i})^\top \\
&\quad + \beta\bar{\mathbf{\Delta}}_t \tfrac{1}{n}\sum_{i=1}^n (\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})(\alpha(\mathbf{B}_t^\top\mathbf{B}_t - \mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t)\mathbf{w}_t + \alpha(\mathbf{B}_t^\top\mathbf{B}_* - \mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_*)\mathbf{w}_{*,t,i})^\top && (210)\\
&\quad + \beta\bar{\mathbf{\Delta}}_t \tfrac{1}{n}\sum_{i=1}^n (\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})(\alpha\mathbf{B}_t^\top \tfrac{1}{m_{in}}(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i})^\top && (211)\\
&\quad + \tfrac{\beta}{n}\sum_{i=1}^n \left(\mathbf{I}_d - \mathbf{\Sigma}_{t,i}^{out}\right)(\mathbf{B}_t\mathbf{w}_{t,i} - \mathbf{B}_*\mathbf{w}_{*,t,i})\,\mathbf{w}_{t,i}^\top + \tfrac{\beta}{nm_{out}}\sum_{i=1}^n (\mathbf{X}_{t,i}^{out})^\top\mathbf{z}_{t,i}^{out}\mathbf{w}_{t,i}^\top && (212)\\
&\quad + \tfrac{\beta\alpha}{n}\sum_{i=1}^n \mathbf{B}_t\mathbf{B}_t^\top\left(\mathbf{I}_d - \mathbf{\Sigma}_{t,i}^{in}\right)(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})\,\mathbf{w}_{t,i}^\top + \tfrac{\beta\alpha}{nm_{in}}\sum_{i=1}^n \mathbf{B}_t\mathbf{B}_t^\top(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}\mathbf{w}_{t,i}^\top && (213)\\
&= \mathbf{B}_{t+1}^{pop} + \beta(\mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \mathbf{E}_4) && (214)
\end{aligned}$$

where $\mathbf{B}_{t,pop} \coloneqq \mathbf{B}_t - \beta\bar{\mathbf{\Delta}}_t\frac{1}{n}\sum_{i=1}^n(\mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i})(\mathbf{\Delta}_t\mathbf{w}_t + \alpha\mathbf{B}_t^\top\mathbf{B}_*\mathbf{w}_{*,t,i})^\top$ denotes the update of the representation in the infinite sample case, and $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ and $\mathbf{E}_4$ are the finite-sample error terms in lines (210), (211), (212) and (213), respectively. From (214) and the triangle inequality, we can compute the final bound.

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq \|\mathbf{I}_k - \alpha\mathbf{B}_{t,pop}^\top\mathbf{B}_{t,pop}\|_2 + 2\beta\alpha\|\mathbf{B}_{t,pop}^\top(\mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \mathbf{E}_4)\|_2 + \beta^2\alpha\|\mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \mathbf{E}_4\|_2^2 \tag{215}$$

Note that from Corollary 1 and the fact that $\|\mathbf{B}_t\|_2 \leq 1.1/\sqrt{\alpha}$ by $A_3(t)$, and β, α are sufficiently small, we have that $\|\mathbf{B}_{t+1}^{pop}\|_2 \leq \frac{1.1}{\sqrt{\alpha}}$. Also, clearly $\mathbf{B}_{t+1}^{pop} \in \mathbb{R}^{d\times k}$. Therefore by the concentration results in Lemma 20 and the triangle and

Cauchy-Schwarz inequalities, we have, for an absolute constant c,

$$\begin{aligned}
\max_{i\in[n]}\|\mathbf{w}_{t,i}\|_2 &\le \|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + c\sqrt{\alpha}L_{\max} + \delta_{m_{in},k}\|\mathbf{w}_t\|_2 + c\sqrt{\alpha}\sigma\delta_{m_{in},k}\\
\|\mathbf{B}_{t,pop}^\top\mathbf{E}_1\|_2 &\le \tfrac{c}{\alpha}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2^2\bar{\delta}_{m_{in},k} + \tfrac{c}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2 L_{\max}\|\mathbf{w}_t\|_2\bar{\delta}_{m_{in},k} + \tfrac{c}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 L_{\max}\bar{\delta}_{m_{in},k}\\
&\quad + \|\mathbf{\Delta}_t\|_2 L_{\max}^2\bar{\delta}_{m_{in},k}\\
&\le cL_{\max}^2\bar{\delta}_{m_{in},k}\\
\|\mathbf{B}_{t,pop}^\top\mathbf{E}_2\|_2 &\le \tfrac{c}{\sqrt{\alpha}}\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\sigma\bar{\delta}_{m_{in},k} + c\|\mathbf{\Delta}_t\|_2 L_{\max}\sigma\bar{\delta}_{m_{in},k} \le cL_{\max}\sigma\bar{\delta}_{m_{in},k}\\
\|\mathbf{B}_{t,pop}^\top\mathbf{E}_3\|_2 &\le \tfrac{c}{\alpha}\bar{\delta}_{m_{out},k}\left(\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \delta_{m_{in},k}\|\mathbf{w}_t\|_2 + \sqrt{\alpha}\sigma\delta_{m_{in},k}\right)^2\\
&\quad + \tfrac{c}{\sqrt{\alpha}}\bar{\delta}_{m_{out},k}L_{\max}\left(\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \delta_{m_{in},k}\|\mathbf{w}_t\|_2 + \sqrt{\alpha}\sigma\delta_{m_{in},k}\right)\\
&\quad + \tfrac{c}{\sqrt{\alpha}}\sigma\bar{\delta}_{m_{out},k}\left(\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \delta_{m_{in},k}\|\mathbf{w}_t\|_2 + \sqrt{\alpha}\sigma\delta_{m_{in},k}\right)\\
&\le c\bar{\delta}_{m_{out},k}(L_{\max}+\sigma)\left(L_{\max}+\sigma\delta_{m_{in},k}\right)\\
\|\mathbf{B}_{t,pop}^\top\mathbf{E}_4\|_2 &\le c\delta_{m_{in},k}\left(\tfrac{\|\mathbf{w}_t\|_2}{\alpha} + \tfrac{L_{\max}}{\sqrt{\alpha}}\right)\left(\tfrac{\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2}{\sqrt{n}} + L_{\max}\tfrac{\sqrt{\alpha}}{\sqrt{n}} + \delta_{m_{in},k}\|\mathbf{w}_t\|_2 + \sqrt{\alpha}\sigma\delta_{m_{in},k}\right)\\
&\quad + \tfrac{c}{\sqrt{\alpha}}(\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\sigma\bar{\delta}_{m_{in},k} + L_{\max}\sigma\bar{\delta}_{m_{in},k} + \sigma\delta^2(m_{in},k)\tfrac{\|\mathbf{w}_t\|_2}{\sqrt{\alpha}} + \sigma^2\delta^2(m_{in},k))\\
&\le c\delta_{m_{in},k}L_{\max}(\tfrac{L_{\max}}{\sqrt{n}} + L_{\max}\delta_{m_{in},k} + \sigma\delta_{m_{in},k})\\
&\quad + c(L_{\max}\sigma\bar{\delta}_{m_{in},k} + L_{\max}\sigma\delta^2(m_{in},k) + \sigma^2\delta^2(m_{in},k))\\
&\le c\bar{\delta}_{m_{out},k}(L_{\max}+\sigma)\left(L_{\max}+\sigma\delta_{m_{in},k}\right) + L_{\max}^2(\delta^2(m_{in},k) + \bar{\delta}_{m_{in},k})\\
&\quad + cL_{\max}\sigma(\delta_{m_{in},k}^2 + \bar{\delta}_{m_{in},k}) + c\sigma^2\delta_{m_{in},k}^2
\end{aligned} \tag{216}$$

with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$. Thus

$$\begin{aligned}
\|\mathbf{B}_{t,pop}^\top(\mathbf{E}_1+\mathbf{E}_2+\mathbf{E}_3+\mathbf{E}_4)\|_2 \le{}& \frac{c\bar{\delta}_{m_{out},k}}{\sqrt{n}}(L_{\max}+\sigma)\left(L_{\max}+\sigma\delta_{m_{in},k}\right) + c(L_{\max}^2 + L_{\max}\sigma)\frac{\delta_{m_{in},k}}{\sqrt{n}}\\
&+ L_{\max}^2\delta_{m_{in},k}^2 + L_{\max}\sigma\delta_{m_{in},k}^2 + \sigma^2\delta_{m_{in},k}^2
\end{aligned} \tag{217}$$

Similarly,

$$\begin{aligned}
\|\mathbf{E}_1\|_2 &\le (\tfrac{\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2}{\sqrt{\alpha}} + \mathrm{dist}_t\, L_{\max} + \|\mathbf{\Delta}_t\|_2 L_{\max})(\delta_{m_{in},k}\|\mathbf{w}_t\|_2 + \sqrt{\alpha}\delta_{m_{in},k}L_{\max})\tfrac{1}{\sqrt{n}}\\
&\le \sqrt{\alpha}L_{\max}^2\frac{\delta_{m_{in},k}}{\sqrt{n}}\\
\|\mathbf{E}_2\|_2 &\le (\tfrac{\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2}{\sqrt{\alpha}} + \mathrm{dist}_t\, L_{\max} + \|\mathbf{\Delta}_t\|_2 L_{\max})\sqrt{\alpha}\sigma\tfrac{\delta_{m_{in},k}}{\sqrt{n}}\\
&\le \sqrt{\alpha}L_{\max}\sigma\frac{\delta_{m_{in},k}}{\sqrt{n}}\\
\|\mathbf{E}_3\|_2 &\le \bar{\delta}_{m_{out},d}\sqrt{\alpha}(L_{\max}+\sigma)(L_{\max}+\sigma\delta_{m_{in},k})\\
\|\mathbf{E}_4\|_2 &\le \sqrt{\alpha}\Bigg(\delta_{m_{in},k}\left(\tfrac{\|\mathbf{w}_t\|_2}{\alpha} + \tfrac{L_{\max}}{\sqrt{\alpha}}\right)\left(\tfrac{\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2}{\sqrt{n}} + \tfrac{\sqrt{\alpha}L_{\max}}{\sqrt{n}} + \delta_{m_{in},k}\|\mathbf{w}_t\|_2 + \sqrt{\alpha}\sigma\delta_{m_{in},k}\right)\\
&\quad + \left(\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\sigma\tfrac{\bar{\delta}_{m_{in},k}}{\sqrt{\alpha}} + L_{\max}\sigma\bar{\delta}_{m_{in},k} + \sigma\delta_{m_{in},k}^2\tfrac{\|\mathbf{w}_t\|_2}{\sqrt{\alpha}} + \sigma^2\delta_{m_{in},k}^2\right)\Bigg)\\
&\le \sqrt{\alpha}\delta_{m_{in},k}L_{\max}\big(\tfrac{L_{\max}}{\sqrt{n}} + L_{\max}\delta_{m_{in},k} + \sigma\delta_{m_{in},k}\big) + \sqrt{\alpha}\big(L_{\max}\sigma\frac{\delta_{m_{in},k}}{\sqrt{n}} + L_{\max}\sigma\delta_{m_{in},k}^2 + \sigma^2\delta_{m_{in},k}^2\big)\\
&\le c\sqrt{\alpha}\bar{\delta}_{m_{out},k}(L_{\max}+\sigma)\left(L_{\max}+\sigma\delta_{m_{in},k}\right) + \sqrt{\alpha}L_{\max}^2(\delta_{m_{in},k}^2 + \bar{\delta}_{m_{in},k})\\
&\quad + \sqrt{\alpha}L_{\max}\sigma(\delta_{m_{in},k}^2 + \bar{\delta}_{m_{in},k}) + \sqrt{\alpha}\sigma^2\delta_{m_{in},k}^2
\end{aligned}$$

thus

$$\|\mathbf{E}_1+\mathbf{E}_2+\mathbf{E}_3+\mathbf{E}_4\|_2 \le \sqrt{\alpha}(L_{\max}^2+L_{\max}\sigma)(\bar{\delta}_{m_{in},k}+\delta_{m_{in},k}^2) + \bar{\delta}_{m_{out},d}\sqrt{\alpha}(L_{\max}+\sigma)(L_{\max}+\sigma\delta_{m_{in},k}) + \sqrt{\alpha}\sigma^2\delta_{m_{in},k}^2$$

Now, from (215) and the triangle inequality, we can compute the final bound.

$$\begin{aligned}
&\|\mathbf{\Delta}_{t+1}\|_2 \\
&\le \|\mathbf{I}_k-\alpha\mathbf{B}_{t,pop}^\top\mathbf{B}_{t,pop}\|_2 \\
&\quad + c\beta\alpha\left(\bar{\delta}_{m_{out},k}(L_{\max}+\sigma)\left(L_{\max}+\sigma\delta_{m_{in},k}\right)+(L_{\max}^2+L_{\max}\sigma)(\bar{\delta}_{m_{in},k}+\delta_{m_{in},k)}^2+\sigma^2\delta_{m_{in},k}^2\right) \\
&\quad + c\beta^2\alpha^2\left((L_{\max}^2+L_{\max}\sigma)(\bar{\delta}_{m_{in},k}+\delta_{m_{in},k}^2)+\bar{\delta}_{m_{out},d}(L_{\max}+\sigma)(L_{\max}+\sigma\delta_{m_{in},k})+\sigma^2\delta_{m_{in},k}^2\right)^2 \\
&\le \|\mathbf{I}_k-\alpha\mathbf{B}_{t,pop}^\top\mathbf{B}_{t,pop}\|_2 \\
&\quad + c\beta\alpha\left(\bar{\delta}_{m_{out},k}(L_{\max}+\sigma)\left(L_{\max}+\sigma\delta_{m_{in},k}\right)+(L_{\max}^2+L_{\max}\sigma)(\bar{\delta}_{m_{in},k}+\delta_{m_{in},k}^2)+\sigma^2\delta_{m_{in},k}^2\right) \\
&\quad + c\beta^2\alpha^2\bar{\delta}_{m_{out},d}^2(L_{\max}+\sigma)^2(L_{\max}+\sigma\delta_{m_{in},k})^2 \\
&= \|\mathbf{I}_k-\alpha\mathbf{B}_{t,pop}^\top\mathbf{B}_{t,pop}\|_2+\beta\alpha\zeta_2 \\
&\le (1-0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2+c\beta^2\alpha^2L_*^4\,\mathrm{dist}_t^2+\beta\alpha\zeta_2 \qquad (218)
\end{aligned}$$

where the last line follows from Lemma 6 (note that all conditions for that lemma are satisfied by $\|\mathbf{I}_k-\alpha\mathbf{B}_{t,pop}^\top\mathbf{B}_{t,pop}\|_2$), and

$$\zeta_2 = O\big((L_{\max}+\sigma)^2\bar{\delta}_{m_{out},k}+(L_{\max}^2+L_{\max}\sigma)(\bar{\delta}_{m_{in},k}+\delta_{m_{in},k}^2)+\sigma^2\delta_{m_{in},k}^2+\beta\alpha(L_{\max}+\sigma)^4\bar{\delta}_{m_{out},d}^2\big) \qquad (219)$$

□

Corollary 3 (FO-ANIL, Finite samples $A_3(t+1)$)**.** *Suppose that $A_2(t+1)$ and $A_3(t)$ hold. Then*

$$\|\mathbf{\Delta}_{t+1}\|_2 \le \tfrac{1}{10} \qquad (220)$$

Proof. From $A_2(t+1)$ we have

$$\begin{aligned}
\|\mathbf{\Delta}_{t+1}\|_2 &\le (1-0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2+c\beta^2\alpha^2L_*^4\,\mathrm{dist}_t^2+\beta\alpha\zeta_2 \\
&\le (1-0.5\beta\alpha E_0\mu_*^2)\tfrac{1}{10}+c\beta^2\alpha^2L_*^4+\beta\alpha\zeta_2 \\
&\le \tfrac{1}{10}-0.25\beta\alpha E_0\mu_*^2+c\beta^2\alpha^2L_*^4 \qquad (221) \\
&\le \tfrac{1}{10} \qquad (222)
\end{aligned}$$

where (221) follows as long as $\zeta_2 \le 0.25E_0\mu_*^2$, and (222) follows since $\beta \le c'\alpha E_0^3\kappa^{-4}$. □

Lemma 29 (FO-ANIL, Finite samples, $A_4(t+1)$)**.** *Suppose $A_1(t)$, $A_3(t)$ and $A_5(t)$ hold. Then $A_4(t+1)$ holds, i.e.*

$$\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \le (1-0.5\beta\alpha E_0\mu^2)\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2+\beta\sqrt{\alpha}\zeta_4$$

where $\zeta_4 = O((L_{\max}+\sigma)(L_{\max}+\sigma\delta_{m_{in},k})\bar{\delta}_{m_{out},d})$ with probability at least $1-\frac{1}{\mathrm{poly}(n)}$.

Proof. Using (213), we have

$$\begin{aligned}
\hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_{t+1} &= \hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_t\left(\mathbf{I}_k-\frac{\beta}{n}\sum_{i=1}^n\mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \\
&\quad+\beta\underbrace{\frac{1}{n}\sum_{i=1}^n\mathbf{B}_{*,\perp}^\top\left(\mathbf{I}_d-\mathbf{\Sigma}_{t,i}^{out}\right)(\mathbf{B}_t\mathbf{w}_{t,i}-\mathbf{B}_*\mathbf{w}_{*,t,i})\mathbf{w}_{t,i}^\top}_{=:\mathbf{E}_1}+\beta\underbrace{\frac{1}{nm_{out}}\sum_{i=1}^n\mathbf{B}_{*,\perp}^\top(\mathbf{X}_{t,i}^{out})^\top\mathbf{z}_{t,i}^{out}\mathbf{w}_{t,i}^\top}_{=:\mathbf{E}_2} \qquad (223)
\end{aligned}$$

Next, we can use the concentration results in Lemma 20 to show that all of the following inequalities hold with probability at least $1 - \frac{1}{\text{poly}(n)}$

$$\max_{i\in[n]} \|\mathbf{w}_{t,i}\|_2 \leq \|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + c\sqrt{\alpha}L_{\max} + \delta_{m_{in},k}\|\mathbf{w}_t\|_2 + c\sqrt{\alpha}\sigma\delta_{m_{in},k} \leq c'\sqrt{\alpha}L_{\max} + c\sqrt{\alpha}\sigma\delta_{m_{in},k}$$

$$\|\mathbf{E}_1\|_2 \leq cL_{\max}\max_{i\in[n]}\|\mathbf{w}_{t,i}\|_2\bar{\delta}_{m_{out},d} \tag{224}$$

$$\|\mathbf{E}_2\|_2 \leq c\sigma\max_{i\in[n]}\|\mathbf{w}_{t,i}\|_2\bar{\delta}_{m_{out},d} \tag{225}$$

Thus we have

$$\|\hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \leq \left\|\hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_t(\mathbf{I}_k - \frac{\beta}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top)\right\|_2 + \beta\sqrt{\alpha}\zeta_4$$

where $\zeta_4 = O((L_{\max}+\sigma)(L_{\max}+\sigma\delta_{m_{in},k})\bar{\delta}_{m_{out},d})$ with probability at least $1 - \frac{1}{\text{poly}(n)}$. Next, recall from Lemma 29 that

$$\sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \leq L^2 \coloneqq 2\left(\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_* + \delta_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma)\right)^2$$

$$\begin{aligned}\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \geq \mu^2 \coloneqq\ & 0.9\alpha E_0\mu_*^2 - 2.2\sqrt{\alpha}\|\mathbf{w}_t\|_2\|\mathbf{\Delta}_t\|_2\eta_* \\ & - 2\|\mathbf{\Delta}_t\|_2\|\mathbf{w}_t\|_2\bar{\delta}_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_{\max} + \sqrt{\alpha}\sigma) \\ & - 2.2\sqrt{\alpha}\bar{\delta}_{m_{in},k}(\|\mathbf{w}_t\|_2 + \sqrt{\alpha}L_* + \sqrt{\alpha}\sigma)L_{\max}\end{aligned}$$

with probability at least $1 - \frac{1}{\text{poly}(n)}$. Apply inductive hypotheses $A_1(t)$ and $A_3(t)$ to obtain

$$\sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \leq 4\alpha L_*^2 + 4\alpha(L_{\max}+\sigma)^2\delta_{m_{in},k}^2 \leq 12\alpha L_*^2$$

by choice of $m_{in} = \Omega((k+\log(n))(L_{\max}+\sigma)^2)$. This means that we have $\beta \leq \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)^{-1}$ since we have chosen $\beta = O(\alpha\kappa_*^{-4})$. Also, we have

$$\sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right) \geq 0.8\alpha E_0\mu_*^2 - 2.3\alpha L_{\max}(L_{\max}+\sigma)\bar{\delta}_{m_{in},k} \geq 0.5\alpha E_0\mu_*^2 \tag{226}$$

where the last inequality follows since $m_{in} = \Omega\left(\frac{k^2}{n}(k\kappa_*^4 + \kappa_*^2\sigma^2\mu_*^{-2})\right)$, recalling that $L_{\max} \leq c\sqrt{k}L_*$. Thus, using the above and Weyl's inequality with $\beta \leq \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}\mathbf{w}_{t,i}^\top\right)^{-1}$, we obtain:

$$\|\hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \leq \left\|\hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_t\right\|_2(1 - 0.5\beta\alpha E_0\alpha^2) + \beta\sqrt{\alpha}\zeta_4 \tag{227}$$

□

E.3. Exact ANIL

Lemma 30 (Exact ANIL FS representation concentration I)**.** *For Exact ANIL, consider any $t \in [T]$. With probability at least $1 - \frac{1}{\text{poly}(n)} - \frac{1}{\text{poly}(m_{in})} - ce^{-90k}$,*

$$\|\hat{\mathbf{G}}_{\mathbf{B},t} - \mathbf{G}_{\mathbf{B},t}\|_2 = \sqrt{\alpha}\zeta_{2,a}, \tag{228}$$

where

$$\zeta_{2,a} = O\Bigg(\Bigg(\tfrac{1}{m_{in}}\Bigg(\tfrac{(L_{\max}+\sigma)L_*}{\kappa_*^2}\Bigg) + \tfrac{1}{\sqrt{m_{in}}}\Bigg(L_{\max}(L_{\max}+\sigma)(\sqrt{k}+\sqrt{\log(n)})\Bigg) \\ + \tfrac{1}{\sqrt{m_{out}}}\Bigg(L_{\max}(L_{\max}+\sigma)(\sqrt{k}+\sqrt{\log(n)})\Bigg) \\ + \tfrac{1}{\sqrt{nm_{in}}}\Bigg(L_{\max}(L_{\max}+\sigma)(k\sqrt{d\log(nm_{in})}+k\log(nm_{in})+\sqrt{d}\log^{1.5}(nm_{in})+\log^2(nm_{in})) \\ + \sigma^2(\sqrt{kd}+\sqrt{d}\log(nm_{in})+\log^{1.5}(nm_{in}))+L_{\max}(L_{\max}+\sigma)\sqrt{d}\Bigg) \\ + \tfrac{1}{\sqrt{nm_{out}}}\Bigg(L_{\max}(L_{\max}+\sigma)\sqrt{d}+\sigma^2(\tfrac{\sqrt{d}}{\sqrt{m_{in}}}+\sqrt{k})\Bigg)\Bigg)\Bigg)$$

Proof. Let $\mathbf{q}_{t,i} \coloneqq \mathbf{B}_t\mathbf{w}_t - \mathbf{B}_*\mathbf{w}_{*,t,i}$. First recall that $\hat{\mathbf{G}}_{\mathbf{B},t} = \frac{1}{n}\sum_{i=1}^n \nabla_{\mathbf{B}}\hat{F}_{t,i}(\mathbf{B}_t,\mathbf{w}_t)$, where

$$\nabla_{\mathbf{B}}\hat{F}_{t,i}(\mathbf{B}_t,\mathbf{w}_t) = (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top\tfrac{1}{m_{out}}(\mathbf{X}_{t,i}^{out})^\top\hat{\mathbf{v}}_{t,i}\mathbf{w}_t^\top - \alpha\tfrac{1}{m_{out}}(\mathbf{X}_{t,i}^{out})^\top\hat{\mathbf{v}}_{t,i}\mathbf{q}_{t,i}^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t \\ - \alpha\mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i}\hat{\mathbf{v}}_{t,i}^\top\tfrac{1}{m_{out}}\mathbf{X}_{t,i}^{out}\mathbf{B}_t + \tfrac{\alpha^2}{m_{in}m_{out}}(\mathbf{X}_{t,i}^{out})^\top\hat{\mathbf{v}}_{t,i}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t \\ + \tfrac{\alpha^2}{m_{in}m_{out}}(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}\hat{\mathbf{v}}_{t,i}^\top\mathbf{X}_{t,i}^{out}\mathbf{B}_t$$

where $\hat{\mathbf{v}}_{t,i} = \mathbf{X}_{t,i}^{out}\bar{\mathbf{\Delta}}_{t,i}^{in}\mathbf{q}_{t,i} + \frac{\alpha}{m_{in}}\mathbf{X}_{t,i}^{out}\mathbf{B}_t\mathbf{B}_t^\top\mathbf{X}_{t,i}^{in}\mathbf{z}_{t,i}^{in} - \mathbf{z}_{t,i}^{out}$. Also, $\mathbf{G}_{\mathbf{B},t} = \frac{1}{n}\sum_{i=1}^n \nabla_{\mathbf{B}}F_{t,i}(\mathbf{B}_t,\mathbf{w}_t)$, where

$$\nabla_{\mathbf{B}}F_{t,i}(\mathbf{B}_t,\mathbf{w}_t) = \bar{\mathbf{\Delta}}_t\mathbf{v}_{t,i}\mathbf{w}_t^\top - \alpha\mathbf{v}_{t,i}\mathbf{q}_{t,i}^\top\mathbf{B}_t - \alpha\mathbf{q}_{t,i}\mathbf{v}_{t,i}^\top\mathbf{B}_t$$

and $\mathbf{v}_{t,i} = \bar{\mathbf{\Delta}}_t\mathbf{q}_{t,i}$. Thus,

$$\begin{aligned}
&\|\hat{\mathbf{G}}_{\mathbf{B},t} - \mathbf{G}_{\mathbf{B},t}\|_2 \\
&\le \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n(\bar{\mathbf{\Delta}}_{t,i}^{in})^\top\tfrac{1}{m_{out}}\mathbf{X}_{t,i}^\top\hat{\mathbf{v}}_{t,i}\mathbf{w}_t^\top - \bar{\mathbf{\Delta}}_t\mathbf{v}_{t,i}\mathbf{w}_t^\top\right\|_2}_{=:\mathbf{E}_1} \\
&\quad + \alpha\underbrace{\left\|\frac{1}{n}\sum_{i=1}^n\tfrac{1}{m_{out}}(\mathbf{X}_{t,i}^{out})^\top\hat{\mathbf{v}}_{t,i}\mathbf{q}_{t,i}^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t - \mathbf{v}_{t,i}\mathbf{q}_{t,i}^\top\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_2} \\
&\quad + \alpha\underbrace{\left\|\frac{1}{n}\sum_{i=1}^n\mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i}\hat{\mathbf{v}}_{t,i}^\top\tfrac{1}{m_{out}}\mathbf{X}_{t,i}^{out}\mathbf{B}_t - \mathbf{q}_{t,i}\mathbf{v}_{t,i}^\top\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_3} + \alpha\underbrace{\left\|\frac{1}{nm_{in}m_{out}}\sum_{i=1}^n(\mathbf{X}_{t,i}^{out})^\top\hat{\mathbf{v}}_{t,i}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_4} \\
&\quad + \alpha\underbrace{\left\|\frac{1}{nm_{in}m_{out}}\sum_{i=1}^n(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}\hat{\mathbf{v}}_{t,i}^\top\mathbf{X}_{t,i}^{out}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_5}
\end{aligned} \tag{229}$$

We will further decompose each of the above terms into terms for which we can apply concentration results from Lemmas

20 and 21. First we bound $\|\mathbf{E}_1\|_2$. We have

$$
\begin{aligned}
\|\mathbf{E}_1\|_2 &= \left\| \frac{1}{n}\sum_{i=1}^{n} (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \tfrac{1}{m_{out}} \mathbf{X}_{t,i}^\top \hat{\mathbf{v}}_{t,i} \mathbf{w}_t^\top - \bar{\mathbf{\Delta}}_t \mathbf{v}_{t,i} \mathbf{w}_t^\top \right\|_2 \\
&= \left\| \frac{1}{n}\sum_{i=1}^{n} (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \mathbf{\Sigma}_{t,i}^{out} \bar{\mathbf{\Delta}}_{t,i}^{in} \mathbf{q}_{t,i} \mathbf{w}_t^\top - \bar{\mathbf{\Delta}}_t \bar{\mathbf{\Delta}}_t \mathbf{q}_{t,i} \mathbf{w}_t^\top \right\|_2 \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha}{m_{in}} (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} \mathbf{w}_t^\top \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^{n} (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \mathbf{w}_t^\top \right\|_2 \\
&\le \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Sigma}_{t,i}^{out} \mathbf{q}_{t,i} \mathbf{w}_t^\top - \mathbf{q}_{t,i} \mathbf{w}_t^\top \right\|_2}_{=:\mathbf{E}_{1,1}} + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \alpha \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{q}_{t,i} \mathbf{w}_t^\top - \alpha \mathbf{B}_t \mathbf{B}_t^\top \mathbf{q}_{t,i} \mathbf{w}_t^\top \right\|_2}_{=:\mathbf{E}_{1,2}} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \alpha \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{q}_{t,i} \mathbf{w}_t^\top - \alpha \mathbf{B}_t \mathbf{B}_t^\top \mathbf{q}_{t,i} \mathbf{w}_t^\top \right\|_2}_{=:\mathbf{E}_{1,3}} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \alpha^2 \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{q}_{t,i} \mathbf{w}_t^\top - \alpha^2 \mathbf{B}_t \mathbf{B}_t^\top \mathbf{B}_t \mathbf{B}_t^\top \mathbf{q}_{t,i} \mathbf{w}_t^\top \right\|_2}_{=:\mathbf{E}_{1,4}} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha}{m_{in}} \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} \mathbf{w}_t^\top \right\|_2}_{=:\mathbf{E}_{1,5}} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha^2}{m_{in}} \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} \mathbf{w}_t^\top \right\|_2}_{=:\mathbf{E}_{1,6}} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \mathbf{w}_t^\top \right\|_2}_{=:\mathbf{E}_{1,7}} + \alpha \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \mathbf{B}_t^\top \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \mathbf{w}_t^\top \right\|_2}_{=:\mathbf{E}_{1,8}}
\end{aligned}
$$

Note that after factoring out trailing $\mathbf{w}_t$'s where necessary, each of the above matrices is in the form that is bounded in Lemma 20 or Lemma 21. We apply the bounds from those lemmas and use $\alpha\|\mathbf{B}_t\|_2^2 = O(1)$, $\|\mathbf{w}_t\|_2 = O(\sqrt{\alpha}\min(1, \eta_*^2/\mu_*^2)\eta_*)$, and $\max_{i\in[n]} \|\mathbf{q}_{t,i}\|_2 = O(L_{\max})$ to obtain that each of the following bounds hold with probability at least $1 - \frac{1}{\mathrm{poly}(n)} - \frac{1}{\mathrm{poly}(m_{in})} - c'e^{-90k}$, for some absolute constants c, c'.

$$
\begin{aligned}
\|\mathbf{E}_{1,1}\|_2 &\le c\sqrt{\alpha}\tfrac{L_{\max}L_*}{\kappa_*^2}\bar{\delta}_{m_{out},d} \\
\|\mathbf{E}_{1,2}\|_2 + \|\mathbf{E}_{1,3}\|_2 &\le c\sqrt{\alpha}\tfrac{L_{\max}L_*}{\kappa_*^2}(\bar{\delta}_{m_{out},d} + \bar{\delta}_{m_{in},d}) \\
\|\mathbf{E}_{1,4}\|_2 &\le c\sqrt{\alpha}\tfrac{L_{\max}L_*}{\kappa_*^2}\left(\tfrac{k\sqrt{d\log(nm_{in})}+\sqrt{d}\log^{1.5}(nm_{in})+\log^2(nm_{in})}{\sqrt{nm_{in}}} + \tfrac{1+C^2k}{m_{in}} + \bar{\delta}_{m_{out},k}\right) \\
\|\mathbf{E}_{1,5}\|_2 &\le c\sqrt{\alpha}\tfrac{\sigma L_*}{\kappa_*^2}\delta_{m_{in},k} \\
\|\mathbf{E}_{1,6}\|_2 &\le c\sqrt{\alpha}\tfrac{\sigma L_*}{\kappa_*^2}\left(\tfrac{k\sqrt{d\log(nm_{in})}+\sqrt{d}\log^{1.5}(nm_{in})+\log^2(nm_{in})}{\sqrt{nm_{in}}} + \bar{\delta}_{m_{out},k}\right) \\
\|\mathbf{E}_{1,7}\|_2 &\le c\sqrt{\alpha}\tfrac{\sigma L_*}{\kappa_*^2}\bar{\delta}_{m_{out},d} \\
\|\mathbf{E}_{1,8}\|_2 &\le c\sqrt{\alpha}\tfrac{\sigma L_*}{\kappa_*^2}\delta_{m_{out},k}
\end{aligned} \tag{230}
$$

For $\|\mathbf{E}_2\|_2$, we have

$$\begin{aligned}
\|\mathbf{E}_2\|_2 &\le \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{out} \bar{\mathbf{\Delta}}_{t,i}^{in} \mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t - \bar{\mathbf{\Delta}}_t \mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{B}_t \right\|_2 \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{out} \tfrac{\alpha}{m_{in}} \mathbf{B}_t\mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^n \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t \right\|_2 \\
&\le \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{out}\mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t - \mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{B}_t \right\|_2}_{=:\mathbf{E}_{2,1}} \\
&\quad + \alpha \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t\mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t - \mathbf{B}_t\mathbf{B}_t^\top \mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{B}_t \right\|_2}_{=:\mathbf{E}_{2,2}} \\
&\quad + \alpha \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t\mathbf{B}_t^\top \tfrac{1}{m_{in}} (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t \right\|_2}_{=:\mathbf{E}_{2,3}} + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t \right\|_2}_{=:\mathbf{E}_{2,4}}
\end{aligned}$$

As before, we apply the bounds from Lemmas 20 and 21 and use $\alpha\|\mathbf{B}_t\|_2^2 = O(1)$, $\|\mathbf{w}_t\|_2 = O(\sqrt{\alpha}\eta_*/\kappa_*^2)$, and $\max_{i\in[n]} \|\mathbf{q}_{t,i}\|_2 = O(L_{\max})$ to obtain that each of the following bounds hold with probability at least $1 - \frac{1}{\mathrm{poly}(n)} - \frac{1}{\mathrm{poly}(m_{in})} - c'e^{-90k}$, for some absolute constants c, c'.

$$\begin{aligned}
\|\mathbf{E}_{2,1}\|_2 &\le \tfrac{cL_{\max}^2}{\sqrt{\alpha}}(\bar{\delta}_{m_{out},d} + \delta_{m_{in},k}) \\
\|\mathbf{E}_{2,2}\|_2 &\le \tfrac{cL_{\max}^2}{\sqrt{\alpha}}(\bar{\delta}_{m_{out},d} + \delta_{m_{in},k}) \\
\|\mathbf{E}_{2,3}\|_2 &\le \tfrac{cL_{\max}\sigma}{\sqrt{\alpha}}\delta_{m_{in},k} \\
\|\mathbf{E}_{2,4}\|_2 &\le \tfrac{cL_{\max}\sigma}{\sqrt{\alpha}}\bar{\delta}_{m_{out},d}
\end{aligned}$$

For $\|\mathbf{E}_3\|_2$, we have

$$\begin{aligned}
\|\mathbf{E}_3\|_2 &\le \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t - \mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \bar{\mathbf{\Delta}}_t \mathbf{B}_t \right\|_2 \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i} \tfrac{\alpha}{m_{in}} (\mathbf{z}_{t,i}^{in})^\top (\mathbf{X}_{t,i}^{in}) \mathbf{B}_t\mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t \right\|_2 \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i} \tfrac{1}{m_{out}} (\mathbf{z}_{t,i}^{out})^\top \mathbf{X}_{t,i}^{out}\mathbf{B}_t \right\|_2 \\
&\le \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t - \mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{B}_t \right\|_2}_{=:\mathbf{E}_{3,1}} \\
&\quad + \underbrace{\left\| \frac{\alpha}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t\mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t - \mathbf{q}_{t,i}\mathbf{q}_{t,i}^\top \bar{\mathbf{\Delta}}_t \mathbf{B}_t \right\|_2}_{=:\mathbf{E}_{3,2}} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i} \tfrac{\alpha}{m_{in}} (\mathbf{z}_{t,i}^{in})^\top (\mathbf{X}_{t,i}^{in}) \mathbf{B}_t\mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t \right\|_2}_{=:\mathbf{E}_{3,3}} + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i} \tfrac{1}{m_{out}} (\mathbf{z}_{t,i}^{out})^\top \mathbf{X}_{t,i}^{out}\mathbf{B}_t \right\|_2}_{=:\mathbf{E}_{3,4}}
\end{aligned}$$

Each term is bounded as follows with probability at least $1-\frac{1}{\text{poly}(n)}-\frac{1}{\text{poly}(m_{in})}-c'e^{-90k}$, for some absolute constants c, c'.

$$\begin{aligned}
\|\mathbf{E}_{3,1}\|_2 &\le \tfrac{cL_{\max}^2}{\sqrt{\alpha}}(\bar{\delta}_{m_{in},d}+\delta_{m_{out},k})\\
\|\mathbf{E}_{3,2}\|_2 &\le \tfrac{cL_{\max}^2}{\sqrt{\alpha}}\left(\tfrac{\sqrt{kd\log(nm_{in})}+\sqrt{k}\log(nm_{in})+\sqrt{d}\log^{1.5}(nm_{in})+\log^2(nm_{in})}{\sqrt{nm_{in}}}+\tfrac{\sqrt{k}}{\sqrt{nm_{out}}}\right)\\
\|\mathbf{E}_{3,3}\|_2 &\le \tfrac{cL_{\max}\sigma}{\sqrt{\alpha}}\left(\tfrac{k\sqrt{d\log(nm_{in})}+k\log(nm_{in})+\sqrt{d}\log^{1.5}(nm_{in})+\log^2(nm_{in})}{\sqrt{nm_{in}}}+\tfrac{\sqrt{k}}{\sqrt{nm_{out}}}\right)\\
\|\mathbf{E}_{3,4}\|_2 &\le \tfrac{cL_{\max}\sigma}{\sqrt{\alpha}}\delta_{m_{out},k}
\end{aligned}$$

For $\|\mathbf{E}_4\|_2$, we have

$$\begin{aligned}
\|\mathbf{E}_4\|_2 \le{}& \left\|\frac{1}{nm_{in}}\sum_{i=1}^n \mathbf{\Sigma}_{t,i}^{out}\bar{\mathbf{\Delta}}_{t,i}^{in}\mathbf{q}_{t,i}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\right\|_2+\alpha\left\|\frac{1}{n}\sum_{i=1}^n\mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t\mathbf{B}_t^\top\tfrac{1}{m_{in}^2}(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\right\|_2\\
&+\left\|\frac{1}{nm_{in}m_{out}}\sum_{i=1}^n(\mathbf{X}_{t,i}^{out})^\top\mathbf{z}_{t,i}^{out}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\right\|_2\\
\le{}&\underbrace{\left\|\frac{1}{nm_{in}}\sum_{i=1}^n\mathbf{\Sigma}_{t,i}^{out}\mathbf{q}_{t,i}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_{4,1}}+\alpha\underbrace{\left\|\frac{1}{nm_{in}}\sum_{i=1}^n\mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t\mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{q}_{t,i}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_{4,2}}\\
&+\alpha\underbrace{\left\|\frac{1}{n}\sum_{i=1}^n\mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t\mathbf{B}_t^\top\tfrac{1}{m_{in}^2}(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_{4,3}}\\
&+\underbrace{\left\|\frac{1}{nm_{in}m_{out}}\sum_{i=1}^n(\mathbf{X}_{t,i}^{out})^\top\mathbf{z}_{t,i}^{out}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_{4,4}}
\end{aligned}$$

Each term is bounded as follows with probability at least $1-\frac{1}{\text{poly}(n)}-\frac{1}{\text{poly}(m_{in})}-c'e^{-90k}$, for some absolute constants c, c'.

$$\begin{aligned}
\|\mathbf{E}_{4,1}\|_2 &\le \tfrac{cL_{\max}\sigma}{\sqrt{\alpha}}\delta_{m_{in},k}\\
\|\mathbf{E}_{4,2}\|_2 &\le \tfrac{cL_{\max}\sigma}{\sqrt{\alpha}}\delta_{m_{in},k}\\
\|\mathbf{E}_{4,3}\|_2 &\le \tfrac{c\sigma^2}{\sqrt{\alpha}}\delta_{m_{in},k}^2\\
\|\mathbf{E}_{4,4}\|_2 &\le \tfrac{c\sigma^2}{\sqrt{\alpha}}\bar{\delta}_{m_{out},d}\delta_{m_{in},k}
\end{aligned}$$

For $\|\mathbf{E}_5\|_2$, we have

$$\begin{aligned}
\|\mathbf{E}_5\|_2 \le{}&\underbrace{\left\|\frac{1}{nm_{in}}\sum_{i=1}^n(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}\mathbf{q}_{t,i}^\top\mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_{5,1}}+\underbrace{\left\|\frac{\alpha}{nm_{in}}\sum_{i=1}^n(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}\mathbf{q}_{t,i}^\top\mathbf{\Sigma}_{t,i}^{in}\mathbf{B}_t\mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_{5,2}}\\
&+\underbrace{\left\|\frac{\alpha}{nm_{in}^2}\sum_{i=1}^n(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}(\mathbf{z}_{t,i}^{in})^\top\mathbf{X}_{t,i}^{in}\mathbf{B}_t\mathbf{B}_t^\top\mathbf{\Sigma}_{t,i}^{out}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_{5,3}}+\underbrace{\left\|\frac{1}{nm_{in}m_{out}}\sum_{i=1}^n(\mathbf{X}_{t,i}^{in})^\top\mathbf{z}_{t,i}^{in}(\mathbf{z}_{t,i}^{out})^\top\mathbf{X}_{t,i}^{out}\mathbf{B}_t\right\|_2}_{=:\mathbf{E}_{5,4}}
\end{aligned}$$

Each term is bounded as follows with probability at least $1-\frac{1}{\text{poly}(n)}-\frac{1}{\text{poly}(m_{in})}-c'e^{-90k}$, for some absolute constants

c, c'.

$$\|\mathbf{E}_{5,1}\|_2 \leq \tfrac{cL_{\max}\sigma}{\sqrt{\alpha}}\bar{\delta}_{m_{in},d}$$
$$\|\mathbf{E}_{5,2}\|_2 \leq \tfrac{cL_{\max}\sigma}{\sqrt{\alpha}}\left(\tfrac{\sqrt{kd}\log(nm_{in})+\sqrt{d}\log^{1.5}(nm_{in})+\log^2(nm_{in})}{\sqrt{nm_{in}}} + \tfrac{\sqrt{k}}{\sqrt{nm_{out}}}\right)$$
$$\|\mathbf{E}_{5,3}\|_2 \leq \tfrac{c\sigma^2}{\sqrt{\alpha}}\left(\tfrac{\sqrt{kd}\sqrt{\log(nm_{in})}+\sqrt{d}\log(nm_{in})+\log^{1.5}(nm_{in})}{\sqrt{nm_{in}}} + \tfrac{\sqrt{k}}{\sqrt{nm_{out}}}\right)$$
$$\|\mathbf{E}_{5,4}\|_2 \leq \tfrac{c\sigma^2}{\sqrt{\alpha}}\bar{\delta}_{m_{in},d}\delta_{m_{out},k}$$

Applying a union bound over these events yields that

$$\begin{aligned}
&\|\hat{\mathbf{G}}_{\mathbf{B},t} - \mathbf{G}_{\mathbf{B},t}\|_2 \\
&\leq c\sqrt{\alpha}\Bigg(\tfrac{1}{m_{in}}\left(\tfrac{kL_{\max}L_*}{\kappa_*^2}\right) + \tfrac{1}{\sqrt{m_{in}}}\left(L_{\max}(L_{\max}+\sigma)(\sqrt{k}+\sqrt{\log(n)})\right) \\
&\quad + \tfrac{1}{\sqrt{m_{out}}}\left(L_{\max}(L_{\max}+\sigma)(\sqrt{k}+\sqrt{\log(n)})\right) \\
&\quad + \tfrac{1}{\sqrt{nm_{in}}}\Bigg(L_{\max}(L_{\max}+\sigma)(k\sqrt{d\log(nm_{in})} + k\log(nm_{in}) + \sqrt{d}\log^{1.5}(nm_{in}) + \log^2(nm_{in})) \\
&\quad + \sigma^2(\sqrt{kd} + \sqrt{d}\log(nm_{in}) + \log^{1.5}(nm_{in}))\Bigg) \\
&\quad + \tfrac{1}{\sqrt{nm_{out}}}\left(L_{\max}(L_{\max}+\sigma)\sqrt{d} + \sigma^2(\tfrac{\sqrt{d}}{\sqrt{m_{in}}} + \sqrt{k})\right)\Bigg) \\
&:= \sqrt{\alpha}\zeta_{2,a}
\end{aligned}$$

with probability at least $1 - \frac{1}{\text{poly}(n)} - \frac{1}{\text{poly}(m_{in})} - c'e^{-90k}$ for absolute constants c, c'. □

Lemma 31 (Exact ANIL FS representation concentration II)**.** *For Exact ANIL, consider any $t \in [T]$. With probability at least $1 - ce^{-100k} - \frac{1}{\text{poly}(n)}$ for an absolute constant c:*

$$\|\mathbf{B}_t^\top \hat{\mathbf{G}}_{\mathbf{B},t} - \mathbf{B}_t^\top \mathbf{G}_{\mathbf{B},t}\|_2 \leq \zeta_{2,b}$$

where

$$\zeta_{2,b} := O\left(\tfrac{\sqrt{k}+\sqrt{\log(n)}}{\sqrt{m_{in}}}\big(L_{\max}(L_{\max}+\sigma) + \sigma^2(\tfrac{\sqrt{k}+\sqrt{\log(n)}}{\sqrt{m_{in}}} + \tfrac{\sqrt{k}}{\sqrt{nm_{out}}})\big) + \tfrac{\sqrt{k}}{\sqrt{nm_{out}}}\big(L_{\max}(L_{\max}+\sigma)\big)\right). \tag{231}$$

Proof. We adapt the proof of Lemma 30. Multiplying $\hat{\mathbf{G}}_{\mathbf{B},t} - \mathbf{G}_{\mathbf{B},t}$ on the left by $\mathbf{B}_t^\top$ serves to reduce the dimensionality of $\hat{\mathbf{G}}_{\mathbf{B},t} - \mathbf{G}_{\mathbf{B},t}$ from $\mathbb{R}^{d\times k}$ to $\mathbb{R}^{k\times k}$. This means that all of the d dependence in the previous concentration result for $\|\hat{\mathbf{G}}_{\mathbf{B},t} - \mathbf{G}_{\mathbf{B},t}\|_2$ is reduced to k. Moreover, we no longer need to apply the complicated bounds on sums of fourth-order products (Lemma 21) to show concentration at a rate of $\frac{\sqrt{d}}{\sqrt{nm_{in}}}$, since we can afford to show concentration of each second order product at a rate $\frac{\sqrt{k}+\sqrt{\log(n)}}{\sqrt{m_{in}}}$ (see Lemma 20). Finally, we must divide the remaining bound from Lemma 30 by $\sqrt{\alpha}$ since $\|\mathbf{B}_t\|_2 = \Theta(\frac{1}{\sqrt{\alpha}})$. Making these changes yields the result.

□

Lemma 32 (Exact ANIL FS head concentration)**.** *For Exact ANIL, consider any $t \in [T]$. With probability at least $1 - ce^{-100k} - \frac{1}{\text{poly}(n)}$ for an absolute constant c, we have*

$$\|\hat{\mathbf{G}}_{\mathbf{w},t} - \mathbf{G}_{\mathbf{w},t}\|_2 \leq \tfrac{1}{\sqrt{\alpha}}\zeta_1, \tag{232}$$

where $\zeta_1 = O(\frac{(L_{\max}+\sigma)(\sqrt{k}+\sqrt{\log(n)})}{\sqrt{m_{in}}} + \frac{(L_{\max}+\sigma)\sqrt{k}}{\sqrt{nm_{out}}})$.

Proof. We have:

$$
\begin{aligned}
\|\hat{\mathbf{G}}_{\mathbf{w},t} - \mathbf{G}_{\mathbf{w},t}\|_2 &= \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{B}_t^\top (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \hat{\mathbf{v}}_{t,i} - \tfrac{1}{m_{out}} \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} - \mathbf{B}_t^\top \bar{\mathbf{\Delta}}_t \bar{\mathbf{\Delta}}_t \mathbf{q}_{t,i} \right\|_2 \\
&\le \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{B}_t^\top (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \mathbf{\Sigma}_{t,i}^{out} \bar{\mathbf{\Delta}}_{t,i}^{in} \mathbf{q}_{t,i} - \mathbf{\Delta}_t \mathbf{\Delta}_t \mathbf{B}_t^\top \mathbf{q}_{t,i} \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{m_{out}} \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \right\|_2 \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^{n} \tfrac{\alpha}{m_{in}} \mathbf{B}_t^\top (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} \right\|_2 \\
&\quad + \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{B}_t^\top (\bar{\mathbf{\Delta}}_{t,i}^{in})^\top \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \right\|_2 \\
&\le \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Delta}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \bar{\mathbf{\Delta}}_t \mathbf{q}_{t,i} - \mathbf{\Delta}_t \mathbf{B}_t^\top \bar{\mathbf{\Delta}}_t \mathbf{q}_{t,i} \right\|_2}_{=:\mathbf{E}_1} + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{m_{out}} \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \right\|_2}_{=:\mathbf{E}_2} \\
&\quad + \underbrace{\left\| \frac{\alpha^2}{n}\sum_{i=1}^{n} \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{q}_{t,i} - \frac{\alpha^2}{n}\sum_{i=1}^{n} \mathbf{B}_t^\top \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{q}_{t,i} \right\|_2}_{=:\mathbf{E}_3} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \tfrac{\alpha}{m_{in}} \mathbf{B}_t^\top \bar{\mathbf{\Delta}}_t \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} \right\|_2}_{=:\mathbf{E}_4} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \tfrac{\alpha^2}{m_{in}} \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{in} \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} \right\|_2}_{=:\mathbf{E}_5} \\
&\quad + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \tfrac{\alpha^2}{m_{in}} \mathbf{B}_t^\top \mathbf{B}_t \mathbf{B}_t^\top \mathbf{\Sigma}_{t,i}^{out} \mathbf{B}_t \mathbf{B}_t^\top (\mathbf{X}_{t,i}^{in})^\top \mathbf{z}_{t,i}^{in} \right\|_2}_{=:\mathbf{E}_6} + \underbrace{\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{B}_t^\top \bar{\mathbf{\Delta}} \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \right\|_2}_{=:\mathbf{E}_7} \\
&\quad + \underbrace{\left\| \frac{\alpha}{n}\sum_{i=1}^{n} \mathbf{B}_t^\top (\mathbf{\Sigma}_{t,i}^{in} - \mathbf{I}_d) \mathbf{B}_t \mathbf{B}_t^\top \tfrac{1}{m_{out}} (\mathbf{X}_{t,i}^{out})^\top \mathbf{z}_{t,i}^{out} \right\|_2}_{=:\mathbf{E}_8}
\end{aligned}
$$

By Lemma 20 and the facts that $\|\mathbf{B}_t\|_2 = \frac{1}{\sqrt{\alpha}}$, $\|\mathbf{\Delta}_t\|_2 = O(1)$, and $\max_i \|\mathbf{q}_{t,i}\|_2 = L_{\max}$ we have

$$
\begin{aligned}
\mathbb{P}(\|\mathbf{E}_1\|_2 \ge \|\mathbf{\Delta}_t\|_2 \|\mathbf{B}_t\|_2 \max_i \|\bar{\mathbf{\Delta}}_t \mathbf{q}_{t,i}\|_2 \bar{\delta}_{m_{out},k}) &\le 2e^{-90k} \\
\mathbb{P}(\|\mathbf{E}_2\|_2 \ge \sigma \|\mathbf{B}_t\|_2 \bar{\delta}_{m_{out},k}) &\le 2e^{-90k} \\
\mathbb{P}(\|\mathbf{E}_3\|_2 \ge \alpha^2 \|\mathbf{B}_t\|_2^5 \max_i \|\mathbf{q}_{t,i}\|_2 \delta_{m_{in},k} &\le 8n^{-99} \\
\mathbb{P}(\|\mathbf{E}_4\|_2 \ge \alpha \|\mathbf{B}_t\|_2^3 \sigma \|\mathbf{\Delta}_t\|_2 \bar{\delta}_{m_{in},k}) &\le 4n^{-99} \\
\mathbb{P}(\|\mathbf{E}_5\|_2 \ge \alpha^2 \|\mathbf{B}_t\|_2^5 \sigma \delta_{m_{in},k} &\le 6n^{-99} \\
\mathbb{P}(\|\mathbf{E}_6\|_2 \ge \alpha^2 \|\mathbf{B}_t\|_2^5 \sigma \bar{\delta}_{m_{in},k}) &\le 4n^{-99} \\
\mathbb{P}(\|\mathbf{E}_7\|_2 \ge \|\mathbf{B}_t\|_2 \sigma \|\mathbf{\Delta}_t\|_2 \bar{\delta}_{m_{out},k}) &\le 2e^{-90k} \\
\mathbb{P}(\|\mathbf{E}_8\|_2 \ge \alpha \|\mathbf{B}_t\|_2^3 \sigma \delta_{m_{in},k} \bar{\delta}_{m_{out},k}) &\le 2e^{-90k} + 2n^{-99}
\end{aligned}
$$

Combining these bounds with a union bound yields:

$$
\|\hat{\mathbf{G}}_{\mathbf{w},t} - \mathbf{G}_{\mathbf{w},t}\|_2 \le \tfrac{c}{\sqrt{\alpha}} \tfrac{(L_{\max}+\sigma)(\sqrt{k}+\sqrt{\log(n)})}{\sqrt{m_{in}}} + \tfrac{c}{\sqrt{\alpha}} \tfrac{(L_{\max}+\sigma)\sqrt{k}}{\sqrt{n m_{out}}} \tag{233}
$$

with probability at least $1 - c'e^{-90k} - \frac{1}{\text{poly}(n)}$ for absolute constants c, c'. □

Lemma 33 (Exact ANIL, Finite samples, $A_1(t+1)$)**.** *For Exact ANIL, suppose $A_2(s)$ and $A_5(s)$ hold for all $s \in [t]$. Then*

$$\|\mathbf{w}_{t+1}\|_2 \leq \tfrac{1}{10}\sqrt{\alpha}E_0 \min\left(1, \tfrac{\mu_*^2}{\eta_*^2}\right)\eta_* \tag{234}$$

with probability at least $1 - ce^{-100k} - \frac{1}{\text{poly}(n)}$ for an absolute constant c.

Proof. For any $s \in [t]$, we have

$$\begin{aligned}
\|\mathbf{w}_{s+1}\|_2 &= \|\mathbf{w}_s - \beta\mathbf{G}_{\mathbf{w},s} + \beta(\mathbf{G}_{\mathbf{w},s} - \hat{\mathbf{G}}_{\mathbf{w},s})\|_2 \\
&\leq \|\mathbf{w}_s - \beta\mathbf{G}_{\mathbf{w},s}\|_2 + \beta\|\mathbf{G}_{\mathbf{w},s} - \hat{\mathbf{G}}_{\mathbf{w},s}\|_2 \\
&\leq \|\mathbf{w}_s\|_2 + c\tfrac{\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2^2\eta_* + \beta\|\mathbf{G}_{\mathbf{w},s} - \hat{\mathbf{G}}_{\mathbf{w},s}\|_2 \qquad (235) \\
&\leq \|\mathbf{w}_s\|_2 + c\tfrac{\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2^2\eta_* + \tfrac{\beta}{\sqrt{\alpha}}\zeta_1 \qquad (236)
\end{aligned}$$

where ζ_1 is defined as in Lemma 32, (235) follows from equation (60) and (236) follows from Lemma 32. This will allow us to apply Lemma 3 with $\xi_{1,s} = 0$ and $\xi_{2,s} = \frac{c\beta}{\sqrt{\alpha}}(\|\mathbf{\Delta}_s\|_2^2\eta_* + \zeta_1)$.

Before doing so, let ζ_2 be defined as in Lemma 34 and $\zeta_4 \coloneqq \zeta_{2,a}$, corresponding to Lemma 35. Observe that for any $s \in [t]$, we can recursively apply $A_2(s), A_2(s-1), \ldots$ to obtain

$$\begin{aligned}
\|\mathbf{\Delta}_s\|_2 &\leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_{s-1}\|_2 + \beta^2\alpha^2 L_*^4 \operatorname{dist}_{s-1}^2 + \beta\alpha\zeta_2 \\
&\vdots \\
&\leq \sum_{r=1}^{s-1}(1 - 0.5\beta\alpha E_0\mu_*^2)^{s-1-r}(c\beta^2\alpha^2 L_*^4 \operatorname{dist}_r^2 + \beta\alpha\zeta_2) \\
&\leq \sum_{r=1}^{s-1}(1 - 0.5\beta\alpha E_0\mu_*^2)^{s-1-r}(c\beta^2\alpha^2 L_*^4(2\rho^{2r} + \beta^2\alpha\zeta_4^2) + \beta\alpha\zeta_2) \\
&\leq c'\alpha^2\beta^2 L_*^4 \sum_{r=1}^{s}(1 - 0.5\beta\alpha E_0\mu_*^2)^{s-1-r}\rho^{2r} + c'\alpha^3\beta^4 L_*^4 \sum_{r=1}^{s-1}(1 - 0.5\beta\alpha E_0\mu_*^2)^{s-1-r}\zeta_4^2 \\
&\quad + \beta\alpha\sum_{r=1}^{s-1}(1 - 0.5\beta\alpha E_0\mu_*^2)^{s-1-r}\zeta_2 \\
&\leq \frac{c'\alpha^2\beta^2 L_*^4(1 - 0.5\beta\alpha E_0\mu_*^2)^{s-1}}{0.5\beta\alpha\mu_*^2} + \frac{c'\alpha^3\beta^4 L_*^4\zeta_4^2}{0.5\beta\alpha\mu_*^2} + \frac{\beta\alpha\zeta_2}{0.5\beta\alpha\mu_*^2} \\
&\leq c''\alpha\beta\kappa_*^2 L_*^2(1 - 0.5\beta\alpha E_0\mu_*^2)^{s-1} + c''\alpha^2\beta^3\kappa_*^2 L_*^2\zeta_4^2 + \frac{\zeta_2}{\mu_*^2} \qquad (237)
\end{aligned}$$

Therefore, via Lemma 3,

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|_2 &\leq \sum_{s=1}^{t} c\tfrac{\beta}{\sqrt{\alpha}}\|\mathbf{\Delta}_s\|_2^2\eta_* + \tfrac{c\beta}{\sqrt{\alpha}}\zeta_1 \\
&\leq c'\sum_{s=1}^{t}\tfrac{\beta^3\alpha^{1.5}L_*^8}{E_0^2\mu_*^4}(1 - 0.5\beta\alpha E_0\mu_*^2)^{2s-2}\eta_* + t\beta^7\alpha^{3.5}\kappa_*^4 L_*^4\eta_*\zeta_4^4 + t\tfrac{\beta}{\sqrt{\alpha}}\tfrac{\eta_*}{\mu_*^4}\zeta_2^2 + t\tfrac{\beta}{\sqrt{\alpha}}\zeta_1 \\
&\leq c''\tfrac{\beta^2\sqrt{\alpha}L_*^8}{E_0^3\mu_*^6}\eta_* + c''t\beta^7\alpha^{3.5}\kappa_*^4 L_*^4\eta_*\zeta_4^4 + c''t\tfrac{\beta}{\sqrt{\alpha}}\tfrac{\eta_*}{\mu_*^4}\zeta_2^2 + c''t\tfrac{\beta}{\sqrt{\alpha}}\zeta_1 \qquad (238) \\
&\leq c'''\tfrac{\sqrt{\alpha}E_0}{\kappa_*^2}\min(1, \tfrac{\mu_*^2}{\eta_*^2})\eta_* + t\beta^7\alpha^{3.5}\kappa_*^4 L_*^4\eta_*\zeta_4^4 + t\tfrac{\sqrt{\alpha}E_0}{L_*^4}\min(1, \tfrac{\mu_*^2}{\eta_*^2})\eta_*\zeta_2^2 + t\tfrac{\sqrt{\alpha}E_0}{\kappa_*^4}\min(1, \tfrac{\mu_*^2}{\eta_*^2})\zeta_1 \qquad (239) \\
&\leq c''''\big(\tfrac{\sqrt{\alpha}E_0}{\kappa_*^2}\min(1, \tfrac{\mu_*^2}{\eta_*^2})\eta_* + t\tfrac{\sqrt{\alpha}E_0}{L_*^4}\min(1, \tfrac{\mu_*^2}{\eta_*^2})\eta_*\zeta_{2,b}^2 + t\beta^2\alpha^{2.5}\tfrac{E_0}{L_*^4}\min(1, \tfrac{\mu_*^2}{\eta_*^2})\eta_*\zeta_{2,a}^4 \\
&\quad + t\tfrac{\sqrt{\alpha}E_0}{\kappa_*^4}\min(1, \tfrac{\mu_*^2}{\eta_*^2})\zeta_1\big) \qquad (240)
\end{aligned}$$

where (238) follows by the sum of a geometric series and (239) follows by choice of $\beta \leq c\frac{\alpha E_0^2}{\kappa_*^4}$ for a sufficiently small constant c, (240) follows by using the definitions of ζ_4 and ζ_2, the numerical inequality $(a+b)^2 \leq 2a^2+2b^2$, and subsuming the dominated term.

In order for the RHS (240) to be at most $\frac{\sqrt{\alpha}E_0}{10}\min(1, \frac{\mu_*^2}{\eta_*^2})\eta_*$, we require the following:

$$\zeta_1 \leq \tfrac{c}{T}, \quad \zeta_{2,b} \leq \tfrac{L_*^2}{\sqrt{T}}, \quad \zeta_{2,a} \leq \tfrac{L_*}{\sqrt{\beta\alpha}T^{0.25}} \tag{241}$$

However, from Corollary 4 we require tighter bounds on $\zeta_{2,b}$ and $\zeta_{2,a}$ when T is small. Accounting for these, it is sufficient to choose

$$\begin{aligned}\zeta_1 &\leq \tfrac{c\kappa_*^4\eta_*}{T}\\ \zeta_{2,b} &\leq c\tfrac{E_0\mu_*^2}{\sqrt{T}},\\ \zeta_{2,a} &\leq c\tfrac{\sqrt{E_0}\mu_*}{\sqrt{\beta\alpha}T^{0.25}}\end{aligned} \tag{242}$$

We also require $m_{out} \geq ck + c\log(n)$ so that the concentration results hold. This implies that we need

$$\begin{aligned}m_{out} \geq\ & cT^2\tfrac{k(L_{\max}+\sigma)^2}{n\eta_*^2\kappa_*^8} + cT\tfrac{kL_{\max}^2(L_{\max}+\sigma)^2}{nE_0^2\mu_*^4} + c\sqrt{T}\tfrac{\beta\alpha}{E_0\mu_*^2}(L_{\max}^2(L_{\max}+\sigma)^2(k+\log(n)) + (L_{\max}+\sigma)^4\tfrac{d}{n})\\ &+ ck + c\log(n)\end{aligned}$$

For m_{in}, we need

$$\begin{aligned}m_{in} \geq\ & cT^2\tfrac{(L_{\max}+\sigma)^2(k+\log(n))}{\eta_*^2\kappa_*^8} + cT\tfrac{(L_{\max}+\sigma)^4(k+\log(n))}{E_0^2\mu_*^4} + cT^{0.25}\tfrac{\sqrt{\beta\alpha}L_{\max}k}{\sqrt{E_0}\kappa_*}\\ &+ c\sqrt{T}\tfrac{\beta\alpha L_{\max}^2(L_{\max}+\sigma)^2(k+\log(n))}{E_0\mu_*^2} + c\sqrt{T}\tfrac{\beta\alpha(L_{\max}+\sigma)^4k^2d\log(nm_{in})}{nE_0\mu_*^2}\end{aligned} \tag{243}$$

under the natural assumption that $k = \Omega(\log(nm_{in}))$. Note that if m_{in} satisfies the above lower bound, this implies $m_{in} >> k + \log(n)$, as needed. Using our upper bounds on β and α, replacing $L_{\max}$ with $\sqrt{k}L_*$, and treating E_0 as a constant gives the final results:

$$\begin{aligned}m_{out} &\geq cT^2\tfrac{k^2(L_*+\sigma)^2}{n\eta_*^2\kappa_*^8} + cT\tfrac{k^3\kappa_*^2(\kappa_*^2+\sigma^2/\mu_*^2)}{n} + c\sqrt{T}(k + \tfrac{kd}{n} + \log(n))\kappa_*^{-2}(\tfrac{\sigma^2}{L_*^2}+k) + ck + c\log(n)\\ m_{in} &\geq cT^2(k^2+k\log(n))\tfrac{(L_*+\sigma)^2}{\eta_*^2\kappa_*^8} + cT(k^3+k\log(n))(\kappa_*^4+\tfrac{\sigma^4}{\mu_*^4}) + c\sqrt{T}\tfrac{k^3d\log(nm_{in})}{n}\kappa_*^{-2}(\tfrac{\sigma^2}{L_*^2}+1)\end{aligned}$$

□

Lemma 34 (Exact ANIL Finite samples $A_2(t+1)$)**.** *Suppose the conditions of Theorem 8 are satisfied and $A_1(t), A_3(t)$ and $A_5(t)$ hold. Then with probability at least $1 - ce^{-90k} - \frac{1}{\text{poly}(n)} - \frac{1}{\text{poly}(m_{in})}$ for an absolute constant c,*

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2L_*^4\,\text{dist}_t^2 + \beta\alpha\zeta_2, \tag{244}$$

$\zeta_2 \coloneqq 2\zeta_{2,b} + \beta\alpha\zeta_{2,a}^2$, and $\zeta_{2,a}$ and $\zeta_{2,b}$ are defined in Lemmas 30 and 31, respectively.

Proof. As in Lemma 28, let $\mathbf{B}_{t+1} = \mathbf{B}_{t+1}^{pop} + \beta(\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t})$, and let $\mathbf{\Delta}_{t+1}^{pop} = \mathbf{I}_k - \alpha(\mathbf{B}_{t+1}^{pop})^\top\mathbf{B}_{t+1}^{pop}$. Note that the bound from Lemma 10 applies to $\|\mathbf{\Delta}_{t+1}^{pop}\|_2$ This results in

$$\begin{aligned}\|\mathbf{\Delta}_{t+1}\|_2 &= \|\mathbf{\Delta}_{t+1}^{pop} - \beta\alpha\mathbf{B}_t^\top(\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t}) - \beta\alpha(\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t})^\top\mathbf{B}_t\\ &\quad + \beta^2\alpha(\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t})^\top(\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t})\|_2\\ &\leq \|\mathbf{\Delta}_{t+1}^{pop}\|_2 + 2\beta\alpha\|\mathbf{B}_t^\top(\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t})\|_2 + \beta^2\alpha\|\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t})\|_2^2\\ &\leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2L_*^4\,\text{dist}_t^2 + 2\beta\alpha\|\mathbf{B}_t^\top(\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t})\|_2\\ &\quad + \beta^2\alpha\|\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t}\|_2^2 \qquad (245)\\ &\leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2L_*^4\,\text{dist}_t^2 + 2\beta\alpha\zeta_{2,b} + \beta^2\alpha^2\zeta_{2,a}^2 \qquad (246)\end{aligned}$$

where (245) follows from Lemma 10 and (246) $\zeta_{2,a}$ and $\zeta_{2,b}$ are defined in Lemmas 30 and 31, respectively. Define $\zeta_2 \coloneqq 2\zeta_{2,b} + \beta\alpha\zeta_{2,a}^2$ to complete the proof.

□

Corollary 4 (Exact ANIL, Finite samples, $A_3(t+1)$)**.** *Suppose the conditions of Theorem 8 are satisfied and $A_2(t+1)$ and $A_3(t)$ hold. Then*

$$\|\mathbf{\Delta}_{t+1}\|_2 \leq \tfrac{1}{10} \tag{247}$$

Proof. By $A_2(t+1)$ and $A_3(t)$, we have

$$\begin{aligned}
\|\mathbf{\Delta}_{t+1}\|_2 &\leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{\Delta}_t\|_2 + \tfrac{5}{4}\beta^2\alpha^2 L_*^2\,\mathrm{dist}_t^2 + 2\beta\alpha\zeta_{2,b} + \beta^2\alpha^2\zeta_{2,a}^2 \\
&\leq \tfrac{1}{10} - 0.05\beta\alpha E_0\mu_*^2 + \tfrac{5}{4}\beta^2\alpha^2 L_*^4 + 2\beta\alpha\zeta_{2,b} + \beta^2\alpha^2\zeta_{2,a}^2 \\
&\leq \tfrac{1}{10} - 0.04\beta\alpha E_0\mu_*^2 + 2\beta\alpha\zeta_{2,b} + \beta^2\alpha^2\zeta_{2,a}^2 \qquad (248) \\
&\leq \tfrac{1}{10} \qquad (249)
\end{aligned}$$

where (248) follows by choice of $\beta = c\frac{\alpha E_0}{\kappa_*^4}$ and (249) follows by $\zeta_{2,b} \leq cE_0\mu_*^2$ and $\zeta_{2,a}^2 \leq c\frac{E_0\mu_*^2}{\beta\alpha}$ for a sufficiently small constant c. □

Lemma 35 (Exact-ANIL, Finite samples, $A_4(t+1)$)**.** *Suppose the conditions of Theorem 8 are satisfied and $A_1(t)$, $A_3(t)$ and $A_5(t)$ hold. Then $A_4(t+1)$ holds with high probability , i.e.*

$$\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 \leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\mathbf{B}_{*,\perp}^\top\mathbf{B}_t\|_2 + \beta\sqrt{\alpha}\zeta_4 \tag{250}$$

where $\zeta_4 = \zeta_{2,a}$ where $\zeta_{2,a}$ is defined in Lemma 30, with probability at least $1 - ce^{-90k} - \frac{1}{\mathrm{poly}(n)} - \frac{1}{\mathrm{poly}(m_{in})}$ for an absolute constant c.

Proof. We have

$$\begin{aligned}
\|\hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_{t+1}\|_2 &= \|\hat{\mathbf{B}}_{*,\perp}^\top(\mathbf{B}_t - \beta\mathbf{G}_{\mathbf{B},t}) + \beta\hat{\mathbf{B}}_{*,\perp}^\top(\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t})\|_2 \\
&\leq \|\hat{\mathbf{B}}_{*,\perp}^\top(\mathbf{B}_t - \beta\mathbf{G}_{\mathbf{B},t})\|_2 + \beta\|\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t}\|_2 \\
&\leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_t\|_2 + \beta\|\mathbf{G}_{\mathbf{B},t} - \hat{\mathbf{G}}_{\mathbf{B},t}\|_2 \qquad (251) \\
&\leq (1 - 0.5\beta\alpha E_0\mu_*^2)\|\hat{\mathbf{B}}_{*,\perp}^\top\mathbf{B}_t\|_2 + \beta\sqrt{\alpha}\zeta_{2,a} \qquad (252)
\end{aligned}$$

where (251) follows by Lemma 12 (note that all the required conditions are satisfied) and (252) holds with probability at least $1 - ce^{-90k} - \frac{1}{\mathrm{poly}(n)} - \frac{1}{\mathrm{poly}(m_{in})}$ for an absolute constant c according to Lemma 30, where $\zeta_{2,a}$ is defined therein.

□

F. Additional simulation and details

In all experiments, we generated $\mathbf{B}_*$ by sampling a matrix in $\mathbb{R}^{d\times k}$ with i.i.d. standard normal elements, then orthogonalizing this matrix by computing its QR-factorization. The same procedure was used to generate $\mathbf{B}_0$ in cases with random initialization, except that the result of the QR-factorization was scaled by $\frac{1}{\sqrt{\alpha}}$ such that $\mathbf{\Delta}_0 = \mathbf{0}$, and for the case of methodical initialization (Figure 4 (right)), we initialized with an orthogonalized and scaled linear combination of Gaussian noise and $\mathbf{B}_*$ such that $\mathrm{dist}_0 \in [0.65, 0.7]$ and $\|\mathbf{\Delta}_0\| = 0$. Meanwhile, we set $\mathbf{w}_0 = \mathbf{0}$. We used step sizes $\beta = \alpha = 0.05$ in all cases for Figure 4, which were tuned optimally. Figure 1 uses the same setting of $d = 20$, $n = k = 3$, and Gaussian ground-truth heads as in Figure 4, except that the mean of the ground-truth heads is shifted to zero. We are therefore able to use the larger step sizes of $\alpha = \beta = 0.1$ and observe faster convergence in this case, as task diversity is larger since the ground-truth heads are isotropic, and L_* and $L_{\max}$ are smaller. Additionally, in Figure 1, Avg. Risk Min. is the algorithm that tries to minimize $\mathbb{E}_{\mathbf{w}_{*,t,i}}[\mathcal{L}_{t,i}(\mathbf{B}, \mathbf{w})]$ via standard mini-batch SGD. It is equivalent to ANIL and MAML with no inner loop ($\alpha = 0$). In Figure 3, we use $d{=}100$, $k{=}n{=}5$ and $\alpha = \beta = 0.1$. All results are averaged over 5 random trials.