The Mechanism of Prediction Head in Non-contrastive Self-supervised Learning

Zixin Wen
Machine Learning Department
Carnegie Mellon University
zixinw@andrew.cmu.edu

Yuanzhi Li
Machine Learning Department
Carnegie Mellon University
yuanzhil@andrew.cmu.edu

Abstract

The surprising discovery of the BYOL method shows the negative samples can be replaced by adding a prediction head to the neural network. It is mysterious why even when there exist trivial collapsed global optimal solutions, neural networks trained by (stochastic) gradient descent can still learn competitive representations. In this work, we present our empirical and theoretical discoveries on non-contrastive self-supervised learning. Empirically, we find that when the prediction head is initialized as an identity matrix with only its off-diagonal entries being trainable, the network can learn competitive representations even though the trivial optima still exist in the training objective. Theoretically, we characterized the substitution effect and acceleration effect of the trainable, but identity-initialized prediction head. The substitution effect happens when learning the stronger features in some neurons can substitute for learning these features in other neurons through updating the prediction head. And the acceleration effect happens when the substituted features can accelerate the learning of other weaker features to prevent them from being ignored. These two effects enable the neural networks to learn diversified features rather than focus only on learning the strongest features, which is likely the cause of the dimensional collapse phenomenon. To the best of our knowledge, this is also the first end-to-end optimization guarantee for non-contrastive methods using nonlinear neural networks with a trainable prediction head and normalization.

1 Introduction

Self-supervised learning is about learning representations of real-world vision or language data without human supervision, and contrastive learning [62, 43, 41, 24, 20, 34] is one of the most successful self-supervised learning approaches. It has been known that the behavior of contrastive learning depends critically on the minimization of the *negative term*, which corresponds to contrasting the representations of *negative pairs*, i.e., pairs of different data points. However, the surprising finding of the *Bootstrap Your Own Latent* (`BYOL`) method by Grill et al. [37] initiated the research of *non-contrastive self-supervised learning*, which refers to contrastive learning methods without using the negative pairs. `BYOL` achieved state-of-the-art results in various computer vision benchmarks and there are plenty of follow-up works [39, 26, 21, 17, 33, 87, 44, 61] in this direction.

On a high level, in non-contrastive self-supervised learning, one wishes to learn a network ϕ such that $\phi(x)$ aligns in direction with $\phi(x')$, where x and x' are called the *positive pair*, generated by random augmentations from the same sample. Without the negative samples, collapsed global optima exist in the training objectives. The *complete collapse* is when $\phi(\cdot)$ is a constant vector whose variance is zero. Another trivial solutions called **dimensional collapse** by [44] is when all the coordinates $\phi_i(\cdot)$ are exactly aligned. Nevertheless, adding a trainable prediction head on top of (one branch of) $\phi(x)$ magically avoids learning such solutions, **even though the prediction head *can* possibly learn the identity mapping and render itself useless**. A more formal introduction will be given in Section 2.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Since the proposition of BYOL, there have been lots of empirical studies trying to understand non-contrastive learning. The SimSiam method by Chen and He [26] shows the exponential moving average (EMA) is not necessary for avoiding collapsed solutions while **stop-gradient** is necessary. [68] empirically disproved using batch normalization (BN) is the reason why BYOL can avoid collapse. [21, 88] further explored other similar approaches. If one wishes to work without both asymmetry and the negative pairs, one must add extra diversity-enforcing structures as in *Barlow Twins* [87] or [33, 44, 17]. Although these previous papers provided some empirical insights, in theory, the question of how the prediction head helps in learning those diverse features is still unanswered.

Despite the great empirical progress, there is very little theoretical progress towards explaining them. Most of existing theories focus on contrastive learning, especially from the statistical learning perspective [79, 81, 14, 80, 40, 82, 13, 15, 47, 45, 59]. However, due to the existence of trivial collapsed *global optimal* solutions (even with the prediction head) of the non-contrastive methods, to the best of our knowledge, *there is no well-established statistical framework for those methods yet*. To explain the non-contrastive learning, it is inevitable to study how the solutions are chosen during the optimization. Therefore, our research questions are:

Our theoretical questions: the role of prediction head

Why do most non-contrastive self-supervised methods learn collapsed solutions when the so-called prediction head is absent in the network architecture? How does the *trainable* prediction head help **optimizing** the neural network to learn more diversified representations in non-contrastive self-supervised learning?

Due to the existence of trivial collapsed optimal solutions of the non-contrastive learning objective, we need to understand the **implicit bias in optimization** posed by the prediction head. However, to the best of our knowledge, all of the previous implicit biases theories focus only on the supervised learning tasks, and thus cannot be applied to our question. On a high level, the results in this paper are summarized as follows:

Our empirical contributions. In non-contrastive self-supervised learning, we obtain the following experimental results:

- We discover empirically that even when the prediction head is **linear** and initialized as an identity matrix with only off-diagonal entries being trainable, the performance of learned representation is comparable to using the usual non-linear two-layer MLP or randomly initialized (trainable) linear prediction head. See Figure 1.
- We empirically verified that even when the prediction head is an identity-initialized matrix (with fixed diagonal entries), its off-diagonal entries display a rise-and-fall pattern, and it does not always converge to a symmetric matrix. See Figure 3.

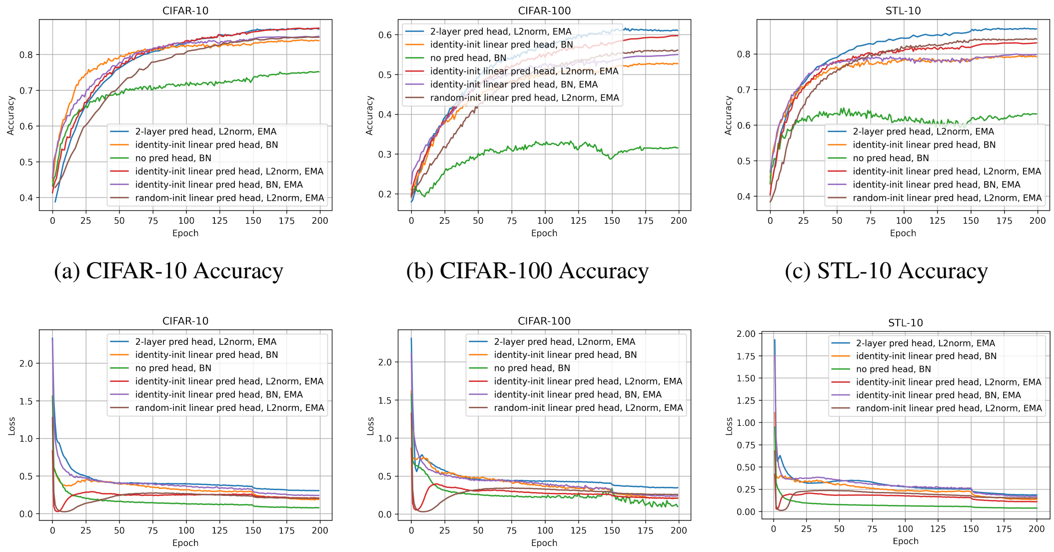
Our theoretical contributions. We based our theory on a very simple setting, where the data consist of two features: the strong feature and the weak feature. Intuitively, the strong features in a dataset are the ones that show up more frequently or with large magnitude, and weak features as those that show up rarely or with small magnitude. We consider learning with a **two-layer non-linear neural network with output normalization** using (stochastic) gradient descent. Under this setting:

- We prove that without a prediction head, even with BN on the output to avoid complete collapse, the networks will still converge to dimensional collapsed solutions, which provides a theoretical explanation to the dimensional collapse phenomenon observed in [44].
- We prove that the trainable prediction head, combined with suitable output normalization and stop-gradient operation, can learn diversified features to avoid the dimensional collapse problem. We characterize two effects leveraged by the prediction head: the **substitution effect** and the **acceleration effect**, as intuitively described below:

The effects of the trainable prediction head

In our setting, we prove that the trainable prediction head can help to learn diversed features by leveraging two effects: the **substitution effect** and the **acceleration effect**. The substitution effect happens when by learning the prediction head, the learned stronger features in some neurons can substitute for learning the same features in other neurons. The acceleration effect happens when the substituted features from the prediction head further accelerate learning the weaker features in those substituted neurons.

Besides the above effects, we also explain in our setting, how the two common components in non-contrastive learning: *stop-gradient* operation and *output normalization*, can assist the prediction



(a) CIFAR-10 Accuracy (b) CIFAR-100 Accuracy (c) STL-10 Accuracy

(d) CIFAR-10 Loss (e) CIFAR-100 Loss (f) STL-10 Loss

Figure 1: Performances of using different prediction heads. Here in CIFAR-10, CIFAR-100 and STL-10, identity-initialized linear prediction head can achieve good accuracies comparable to commonly used two-layer non-linear MLP or randomly-initialized linear head. All the prediction heads are trainable, while for identity-init prediction head only the off-diagonals are trainable. Here BN or L2norm represents the output normalization, and EMA represents using exponential moving average to update the target network as in `BYOL` [39].

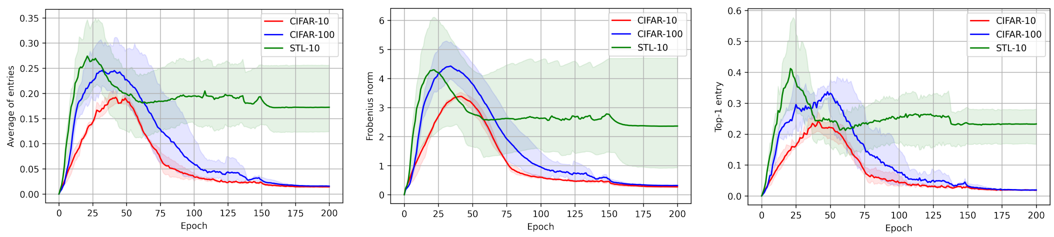
head in creating those effects during training, which will be further discussed in Section 5.3. There are already some theoretical papers [78, 83, 63] that try to address similar questions. Our results provide a completely different perspective compared to them: **We explain why *training* the prediction head can encourage the network to learn diversified features and avoid dimensional collapses**, even when the trivial collapsed optima still exist in the training objective, which is not covered by the prior works, as shall be discussed below.

1.1 Comparison to Similar Studies

In this section, we will clarify the differences between our results and some similar studies. We point out that all the claims below are derived **only in our theoretical setting** and are partially verified in experiments over datasets such as CIFAR-10, CIFAR-100, and STL-10.

Can eigenspace alignment explain the effects of training the prediction head? The paper [78] presented a theoretical statement that (symmetric) linear prediction head will *converge* to a matrix that commutes with the covariance matrix of linear representations *at the end of training*, and they provided experiments to support their theory. However, our theory suggests that *the intermediate stage of training the prediction head matters more to the feature learning of the encoder network than the convergence*. Indeed, as shown in Figure 3, in many cases, the trainable projection head will **converge back to identity** after training, which commutes with any covariance matrix. Moreover, the experiments in Figure 3a shows *the training trajectory of the prediction head displays a clear two-stage separation*, which demonstrates that the convergence result (e.g., the eigenspace alignment result in [78]) is not sufficient to understand the trainable prediction head.

Can the symmetric prediction head explain the trainable prediction head? In the paper [78], experiments over the STL-10 dataset showed that the linear prediction head converges to a symmetric matrix during training. And the follow-up paper [83] established a theory under the symmetric prediction head (which is not trained but manually set at each iteration). Specifically, under their linear network setting, where W is the weight matrix of the base encoder, they manually set the prediction head W_p at iteration t to be $W_p^{(t)} \leftarrow W^{(t)}\mathbb{E}_{x_1} x_1 x_1^\top (W^{(t)})^\top$ and the outputs of both online and target network are not normalized. Under this manual update rule of the prediction head, they proved a subspace learning result over spherical gaussian data. Nevertheless, our experiments in



(a) Average of off-diag entries (b) F-norm of off-diag matrix (c) Maximum of off-diag entries

Figure 2: Trajectories of the identity-initialized prediction head with a $(\min, \max)$ confidence band, average over 3 runs. In all three datasets, we observe a consistent rise and fall trajectory pattern.

Figure 1 and Figure 3b show that even if we initialize the prediction head using a symmetric matrix (identity), **the trainable prediction head can be very asymmetric at the early training stage when the encoder network learn most of its features**. Actually, in the presence of feature imbalance (e.g., $\mathbb{E}_{x_1} x_1 x_1^\top$ has huge eigen-gap), the symmetric prediction head is also likely to learn a rank-one matrix where W focus on learning the largest eigenvector of $\mathbb{E}_{x_1} x_1 x_1^\top$.

The role of stop-gradient and output-normalization. It is discussed in the theory of Tian et al. [78] that without the stop-gradient, the linear network will learn the zero (constant) solution. [83] also incorporated the stop-gradient into their theory, but did not explain why it is necessary for their setting. As a comparison, we proved in our setting that the stop-gradient and output-normalization together can turn the features substituted via the prediction head into a factor in the gradient of the slower learning neurons, thereby creating the acceleration effect. In contrast, analyses in [78, 83] did not incorporate the output normalization, even though their experiments have used certain forms of normalizations. To the best of our knowledge, our paper is the first to explain the effects of output-normalization in optimizing nonlinear neural networks in self-supervised learning.

2 Preliminaries on Non-contrastive Learning

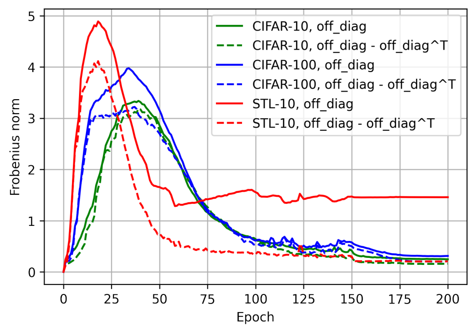
In this section, we formally define what is non-contrastive self-supervised learning. To do this, we first introduce contrastive learning following [24, 85] as background. We use $[N]$ as a shorthand for the index set $\{1, \dots, N\}$.

Background on contrastive learning. Letting $\phi_W(\cdot)$ be the neural networks, contrastive learning aims to learn good representations ϕ_W via contrasting representations of similar data samples to those of dissimilar ones. Usually we are given a batch of data points $\{X_i\}_{i\in[N]}$, and we construct for each $i \in [N]$ a positive pair $(X_i^{(1)}, X_i^{(2)})$ by applying random data augmentations to X_i, and collect negative pairs $(X_i^{(1)}, X_j^{(2)})$ for $i \neq j \in [N]$. Now given $z_i = \phi_W(X_i^{(1)}),\ z_i' = \phi_W(X_i^{(2)}),\ i \in [N]$, we train the network ϕ_W to minimize the contrastive loss:

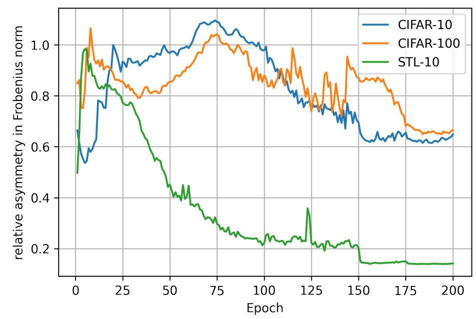
$$L_{\text{contrastive}}(\phi_W) := \frac{1}{N} \sum_{i\in[N]} \underbrace{-\mathbf{sim}(z_i, z_i')/\tau}_{\text{positive term}} + \underbrace{\log \left[\sum_{j\in[N]} \exp\left(\mathbf{sim}(z_i, z_j')/\tau\right) \right]}_{\text{negative term}} \tag{2.1}$$

where $\mathbf{sim}(\cdot, \cdot)$ is the similarity metric, often defined as the cosine similarity, and τ is the so-called temperature hyper-parameter. Intuitively, minimizing the contrastive loss can be roughly viewed as trying to classify the representation z_i as z_i' instead of $z_j', j \neq i$. It is a common belief that in order for the network ϕ_W to be able to "distinguish" data points X_i from $X_j, j \neq i$, merely minimizing the positive term of contrastive loss is not sufficient.

Non-contrastive self-supervised learning. We choose the `SimSiam` method [26] as our primary framework, whose differerence with `BYOL` is a EMA component that is proven inessential in [26]. Following the same notations as above, except that $z_i' = \mathsf{StopGrad}[\phi_W(X_i^{(2)})]$ is detached from



(a) $\|\text{off-diag}(E^{(t)})\|_F$ and $\|E^{(t)} - (E^{(t)})^\top\|_F$



(b) $\|E^{(t)} - (E^{(t)})^\top\|_F / \|\text{off-diag}(E^{(t)})\|_F$

Figure 3: Trajectories of the identity-initialized prediction head. off-diag(E) is obtained by setting the diagonal of E to be zero. In (a), we discover that the Frobenius norm of our identity-initialized prediction head's off-diagonal matrix clearly display a two stage separation, more precisely, a rise and fall pattern; In (b), The off-diagonal matrix of the prediction head is not symmetric in CIFAR-10 and CIFAR-100.

gradient computation, the loss objective become: (the symmetric network version)

$$L'_{\texttt{SimSiam}}(\phi_W) = \tfrac{1}{N}\textstyle\sum_{i\in[N]} -\mathbf{sim}(z_i, z'_i) \tag{2.2}$$

which is just the positive term in contrastive loss (2.1) (not divided by τ). Clearly there exist plenty trivial **global optimal** solutions for this objective. For example, the *complete collapse* refers to when $\phi_W(\cdot)$ learns some constant vector. Another solution called **dimensional collapse** [44] is when all the coordinates $[\phi_W(\cdot)]_i$ has correlation ± 1. The dimensional collapsed solution can minimize the objective (2.2) even when $\phi_W(\cdot)$ is BN-normalized to avoid learning a constant vector [44, 88].

However, by adding a *trainable prediction head* on top of z_i, the training miraculously succeeds and outputs a state-of-the-art feature extractor. Let $g(\cdot)$ be a shallow feed-forward network (often one or two-layer, or even simply linear), we train g and ϕ_W simultaneously on the following objective:

$$L_{\texttt{SimSiam}}(\phi_W, g) = \tfrac{1}{N}\textstyle\sum_{i\in[N]} -\mathbf{sim}(g(z_i), z'_i) \tag{2.3}$$

where z'_i is still detached from gradient computation. The $g(z_i) = g \circ \phi_W(X_i^{(1)})$ and the detached part $z'_i = \mathsf{StopGrad}[\phi_W(X_i^{(2)})]$ are often called the *online* network and the *target* network respectively following [39], known as two branches of non-contrastive learning. Note that the trainable prediction head can represent identity function, so the objective (2.3) still has the collapsed optima.

3 Problem Setup

In this section, we present the setting of our theoretical results. We first define the data distribution.

Notations. We use O, Ω, Θ notations to hide universal constants with respect to d and $\widetilde{O}, \widetilde{\Omega}, \widetilde{\Theta}$ notations to hide polynomial factors of $\log d$. We denote $a = o(1)$ if $a \to 0$ when $d \to \infty$. We use the notations $\mathsf{poly}(d)$, $\mathsf{polylog}(d)$ to represent large constant degree polynomials of d or $\log d$. We use $\mathcal{N}(\mu, \Sigma)$ to denote standard normal distribution in with mean μ and covariance matrix Σ. We use the bracket $\langle \cdot, \cdot \rangle$ to denote the inner product and $\|\cdot\|_2$ the ℓ_2-norm in Euclidean space. And for a subspace $V \subset \mathbb{R}^d$, we denote $V^\perp$ as its orthogonal complement. We use $\mathbb{1}_B$ to denote the indicator function of event B. We use I_m to denote the $m \times m$ identity matrix.

Following the standard structure of image datasets, we consider data divided into patches, where each patch can contain either features or noises.

Definition 3.1 (data distribution and features)**.** Let $X \sim \mathcal{D}$ be $X = (X_1, \dots, X_P) \in \mathbb{R}^{d\times P}$ where each $X_i \in \mathbb{R}^d$ is a patch. We assume that there are two feature vectors v_1, v_2 such that $\|v_\ell\|_2 = 1, \ell = 1, 2$ and are orthogonal to each other. To generate a sample X, we uniformly sampled $\ell \in [2]$ and generate for each $p \in [P]$:

$$X_p = z_p(X) v_\ell + \xi_p \mathbb{1}_{z_p = 0}, \quad \mathbb{E}_{X\sim\mathcal{D}}[z_p(X)] = 0, \quad \forall p \in [P]$$

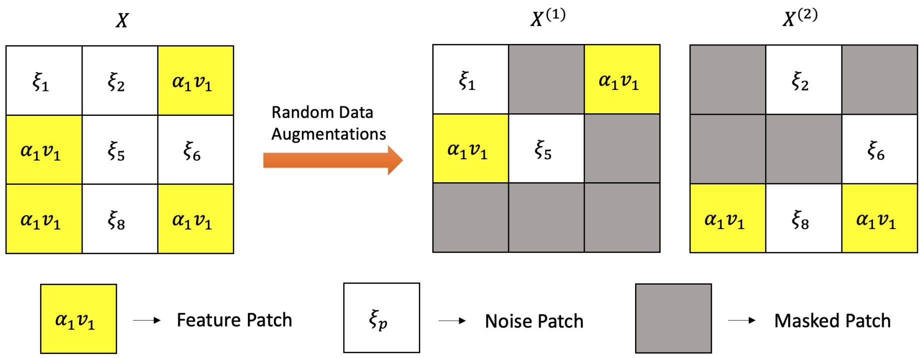


Figure 4: Illustration of the data distribution and data augmentations. Each data is equipped with a feature, either v_1 or v_2, and contains a lot of noise patches. After the data augmentations, the positive pair $(X^{(1)}, X^{(2)})$ is constructed by randomly masking out half of non-overlapping patches for each positive sample. The reason for constructing positive pair with non-overlapping patches is because of the strong noise assumption we made in Assumption 3.2 and the *feature decoupling* principle in [85].

where $z_p(X)$ is the latent vector of X, ξ_p is the noise vector of patch $p \in [P]$ whose assumption will be given in Assumption 3.2. We denote $\mathcal{S}(X) = \{p : z_p(X) \neq 0\} \subseteq [P]$ as the set of feature patches, where $z_p(X) = z_{p'}(X) \in \{0, \pm\alpha_\ell\}, \forall p, p' \in [P]$, where α_ℓ will be picked afterwards. We assume $P = \mathsf{polylog}(d)$, $\mathcal{S}(X) \equiv P_0 = \Theta(\log d)$ for every X. A figurative illustration is given in Figure 4.

Strong and weak features. We pick $\alpha_1 = 2^{\mathsf{polyloglog}(d)}$ and $\alpha_2 = \alpha_1/\mathsf{polylog}(d)$. Hence v_1 is the *strong feature* and v_2 is the *weak feature*, and we want the learner network to learn both v_1, v_2 (but by different neurons) as their learning goal. This is a simplification of the real scenario. Intuitively, we can think of the strong features in a dataset are the ones that show up more frequently or with larger magnitude, and weak features as those that show up rarely or with smaller magnitude.

Assumption 3.2 (noise)**.** Denoting $V = \mathrm{span}(v_1, v_2)$, we assume $\xi_p \in V^\perp$ is independent for each $p \in [P] \setminus S(X)$, where $X = (X_p)_{p \in [P]} \sim \mathcal{D}$, and:

(a) For any unit vector $u \in V^\perp$, $\mathbb{E}[\langle \xi_p, u \rangle] = 0$, and $\mathbb{E}[\langle \xi_p, u \rangle^6] = \sigma^6$ for some $\sigma = \Theta(1)$;

(b) It holds for some $\varrho \in [0, \frac{1}{d^{\Omega(1)}}]$ it holds $|\mathbb{E}[\langle u_1, \xi_p \rangle^3 \langle u_2, \xi_p \rangle^3]| \leq \varrho$ and $|\mathbb{E}[\langle u_1, \xi_p \rangle^5 \langle u_2, \xi_p \rangle]| \leq \varrho$ for any two vectors $u_1, u_2 \in V^\perp$ that are orthogonal to each other.

Remark 3.3. A simple example of our noise ξ_p is the spherical Gaussian noise in $V^\perp$. Assumption 3.2b ensures that the prediction head cannot be used to cancel the noise correlation between different neurons. We point out that the features in our data can be learned via clustering, but we emphasize that we do not intend to compare our algorithm with any clustering method in this setting since our goal is to study how the prediction head helps in learning the features.

3.1 Learner Network

Following the `SimSiam` framework, the online and target network share the same encoder network in our setting, as explained in Section 2. We consider the base encoder network f as a simple convolutional neural network: Let $W = (w_1, \ldots, w_m) \in \mathbb{R}^{d \times m}$ be the weight matrix, where $w_i \in \mathbb{R}^d$, the **encoder network** f is defined by

$$f_j(X) := \textstyle\sum_{p \in [P]} \sigma(\langle w_j, X_p \rangle), \qquad \forall j \in [m]$$

Here we use the cubic activation function $\sigma(z) = z^3$, as polynomial activations are standard in literatures of deep learning theory [9, 35, 50, 2, 52, 23] and also has comparable performance in practice [2]. The (identity initialized) prediction head is defined as a matrix $E = [E_{i,j}]_{(i,j) \in [m]^2}$ with $E_{i,i} \equiv 1, i \in [m]$, where only the the off-diagonals $E_{i,j}, i \neq j$ are trainable parameters. The **online**

network $\widetilde{F}$ is defined by: given $j \in [m]$, we let $F_j(X) := f_j(X) + \sum_{r \neq j} E_{j,r} f_r(X)$, and

$$\widetilde{F}_j(X) := \mathsf{BN}\left(F_j(X)\right) = \mathsf{BN}\Big[\sum_{p \in [P]} \Big(\sigma(\langle w_j, X_p \rangle) + \sum_{r \neq j} E_{j,r} \sigma(\langle w_r, X_p \rangle) \Big) \Big]$$

where the batch normalization BN[1] here is defined as follows: Given a batch of inputs $\{z_i\}_{i \in [N]}$,

$$\mathsf{BN}(z_i) := \frac{z_i - \frac{1}{N} \sum_{i \in [N]} z_i}{\sqrt{\frac{1}{N} \sum_{i \in [N]} z_i^2 - \left(\frac{1}{N} \sum_{i \in [N]} z_i \right)^2}} \tag{3.1}$$

And the we define the **target network** G as $\widetilde{G}_j(X) := \mathsf{BN}\left(G_j(X)\right) = \mathsf{BN}\left(f_j(X)\right), \forall j \in [m]$.

3.2 Training Algorithm

Data augmentation. We use a very simple data augmentation: for each data $X = (X_p)_{p \in [P]}$, we randomly and uniformly sample half of the patches $\mathcal{P} \subseteq [P]$ to generate the *positive pair*:

$$X^{(1)} = (X_p \mathbb{1}_{p \in \mathcal{P}})_{p \in [P]}, \quad X^{(2)} = (X_p \mathbb{1}_{p \notin \mathcal{P}})_{p \in [P]} \tag{3.2}$$

Our data augmentation is similar to the common random cropping used in contrastive learning [22, 76]. It is also analogous to the data augmentations studied in theoretical literatures [85, 47, 58].

Non-contrastive loss function. Now we define the loss function as follows: we sample N data points $\{X_i\}_{i \in [N]}, X_i \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ and apply our data augmentation (3.2) to obtain $\mathcal{S} = \{X^{(i,1)}, X^{(i,2)}\}_{i \in [N]}$. Now we define

$$L_{\mathcal{S}}(W, E) := \frac{1}{N} \sum_{i \in [N]} \left\| \widetilde{F}(X^{(i,1)}) - \mathsf{StopGrad}[\widetilde{G}(X^{(i,2)})] \right\|_2^2 \tag{3.3}$$

where the StopGrad operator detach gradient computation of the target network $\widetilde{G}(\cdot)$. This form of objective (3.3) is first defined in [37] and is equivalent to (2.3) in Chen and He [26] when $\widetilde{F}$ and $\widetilde{G}$ share the same encoder network $f(\cdot)$ and their outputs are normalized.

Intuitions of the data augmentation and collapse. In Definition 3.1, the features v_1, v_2 appear in multiple patches, but the noises are independent across different patches (see Figure 4). As our data augmentation produces positive pairs with non-overlapping patches, learning to emphasize noises cannot align the representations of the positive pair, but learning **either one of** the features $\phi(X) = \sum_p \sigma(\langle v_1, X_p \rangle)$ or $\phi(X) = \sum_p \sigma(\langle v_2, X_p \rangle)$ is sufficient. **We consider learning the same feature v_i in *all the neurons f_j in the encoder network f* as the dimensional collapsed solution.**

Initialization and hyper-parameters. At $t = 0$, we initialize W and E as $W_{i,j}^{(0)} \sim \mathcal{N}(0, \frac{1}{d})$ and $E^{(0)} = I_m$ and we only train the off-diagonal entries of $E^{(t)}$. For the simplicity of analysis, **we let $m = 2$, which suffices to illustrate our main message.** For the learning rates, we let $\eta \in (0, \frac{1}{\mathsf{poly}(d)}]$ be sufficiently small and $\eta_E \in [\eta / \alpha_1^{O(1)}, \eta / \mathsf{polylog}(d)]$, which is smaller than η[2].

Optimization algorithm Given the data augmentation and the loss function, we perform (stochastic) gradient descent on the training objective (3.3) as follows: at each iteration $t = 0, \ldots, T-1$, we sample a new batch of augmented data $\mathcal{S}_t = \{X^{(t,i,1)}, X^{(t,i,2)}\}_{i \in [N]}$ and update

$$W^{(t+1)} = W^{(t)} - \eta \nabla_W L_{\mathcal{S}_t}(W^{(t)}, E^{(t)}), \quad E_{i,j}^{(t+1)} = E_{i,j}^{(t)} - \eta_E \nabla_{E_{i,j}} L_{\mathcal{S}_t}(W^{(t)}, E^{(t)}), \ \forall i \neq j.$$

If we do not train the prediction head, we just simply keep $E^{(t)} \equiv I_m$.

---

[1] We use batch normalization as a output-normalization method, rather than for the supposed implicit negative term effects as disproved in Richemond et al. [68].

[2] We conjecture that by modifying certain assumptions for the noise (especially by allowing the noise to span the feature subspace V), one can prove a similar result for the case $\eta_E = \eta$.

4 Statements of Main Results

In this section, we shall present our main theoretical results on the mechanism of learning the prediction head in non-contrastive learning. To measure the correlation between neurons, we introduce the following notion: letting $\mathbf{Var}(\psi(X)) := \mathbb{E}_{X\sim\mathcal{D}}[(\psi(X) - \mathbb{E}[\psi(X)])^2]$ be the variance of any function ψ of $X \sim \mathcal{D}$, we denote the correlation $\mathbf{Corr}(\psi(X), \psi'(X))$ of any two function ψ, ψ' over $\mathcal{D}$ as

$$\mathbf{Corr}(\psi(X), \psi'(X)) := \frac{\mathbb{E}[(\psi(X) - \mathbb{E}[\psi(X)])(\psi'(X) - \mathbb{E}[\psi'(X)])]}{\sqrt{\mathbf{Var}(\psi(X))}\sqrt{\mathbf{Var}(\psi'(X))}}$$

Now we present the main theorem of training with a prediction head, and set $m = 2$.

Theorem 4.1 (learning with prediction head and BN, see Theorem F.2)**.** *For every $d > 2$, let $N \geq \mathsf{poly}(d)$, $\eta \in (0, \frac{1}{\mathsf{poly}(d)}]$ be sufficiently small, and $\eta_E \in [\frac{\eta}{\alpha_1^{O(1)}}, \frac{\eta}{\mathsf{polylog}(d)}]$. Then with probability $1 - o(1)$, after training for $T = \mathsf{poly}(d)/\eta$ many iterations, we shall have for some $\ell \in [2]$:*

$$w_1^{(T)} = \beta_1 v_\ell + \varepsilon_1, \quad w_2^{(T)} = \beta_2 v_{3-\ell} + \varepsilon_2 \qquad \textit{with} \quad |\beta_1|, |\beta_2| = \Theta(1),\ \|\varepsilon_1\|_2, \|\varepsilon_2\|_2 \leq \widetilde{O}(\frac{1}{\sqrt{d}})$$

Furthermore, the objective converges: $\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^N}[L_{\mathcal{S}}(W^{(T)}, E^{(T)})] \leq \mathsf{OPT} + \frac{1}{\mathsf{poly}(d)} \leq O(\frac{1}{\log d})$. Where OPT stands for the global optimum[3].

Theorem 4.1 clearly shows the network learn all the desired features, even under huge imbalance between v_1 and v_2. This leads to the following corollary.

Corollary 4.2. *Under the same hyper-parameter in Theorem 4.1, with probability $1 - o(1)$, after training for $T = \mathsf{poly}(d)/\eta$ many iterations, then the encoder f* ***avoids dimensional collapse:***

$$|\mathbf{Corr}(f_1(X), f_2(X))| \leq O(\frac{1}{\sqrt{d}}).$$

In contrast, learning without the prediction head will create strong correlations between any two neurons. To emphasize that this problem cannot be alleviated by having more neurons, we let the number of neurons m be any positive integer in the following theorem.

Theorem 4.3 (learning without prediction head but with BN, see Theorem G.1)**.** *Let $N \geq \mathsf{poly}(d)$, $\eta = o(1)$ and the number of neurons $m \leq o(\alpha_1/\alpha_2)$. Suppose we freeze $E^{(t)} \equiv I_m$, then with probability $1 - o(1)$, after training for $T = \mathsf{poly}(d)/\eta$ many iterations, we shall have:*

$$w_j^{(T)} = \beta_j v_1 + \varepsilon_j \qquad \textit{with} \quad |\beta_j| = \Theta(1),\ \|\varepsilon_j\|_2 \leq \widetilde{O}(\frac{1}{\sqrt{d}}) \qquad \text{for all } j \in [m]$$

Furthermore, the objective converges: $\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^N}[L_{\mathcal{S}}(W^{(T)}, E^{(T)})] \leq \mathsf{OPT} + \frac{1}{\mathsf{poly}(d)} \leq O(\frac{1}{\log d})$. This means the collapsed solution also reaches the global minimum of the objective. Again OPT stands for the global optimum.

Note that since we have used BN as our output normalization, the learner is immune to complete collapse and must have a certain variance in the outputs. Immediately, we have a corollary.

Corollary 4.4. *Under the same hyper-parameter in Theorem 4.3, with probability $1 - o(1)$, after training with $E^{(t)} \equiv I_m$ for $T = \mathsf{poly}(d)/\eta$ many iterations, we shall have* ***dimensional collapse****:*

$$|\mathbf{Corr}(f_i(X), f_j(X))| \geq 1 - O(\frac{1}{\sqrt{d}}), \qquad \textit{for all } i, j \in [m].$$

In the following section, we shall give some intuitions by digging through the training process and separately discuss the four phases of the training process.

[3]Under our setting described in Section 2, the *global minimum* of our objective (3.3) in population is

$$\mathsf{OPT} = 2 - 2\frac{\mathbb{E}[|\mathcal{S}(X) \cap \mathcal{P}| \cdot |\mathcal{S}(X) \setminus \mathcal{P}|]}{\mathbb{E}[|\mathcal{S}(X) \cap \mathcal{P}|^2]} = \Theta(\frac{1}{\log d})$$

5 The Four Phases of the Learning Process

We divide the complete training process into four phases: phase I for learning the stronger feature, phase II for the substitution effect, phase III for the acceleration effect, and the end phase for convergence. The first three phases explain how the prediction head can help learn the base encoder, and the last phase explains why the off-diagonal entries often shrink later in training.

5.1 Phase I: Learning the Stronger Feature

At the beginning of training, the stronger feature v_1 enjoys a much larger gradient as opposed to the weaker feature v_2, so naturally, v_1 will be learned first. Without loss of generality, let us assume at initialization, the neuron f_1 has larger v_1 between $f_j, j \in [2]$, then we can show:

Lemma 5.1 (learning the stronger feature, see Lemma C.13)**.** *After some $t \geq T_1 = d^{2+o(1)}/\eta$, the feature v_1 in neuron f_1 will be learn to $\langle w_1^{(t)}, v_1\rangle = \Omega(1)$, while all other features $\langle w_j^{(t)}, v_\ell\rangle = o(1)$ for $(j, \ell) \neq (1, 1)$. And the prediction head $\|E^{(t)} - I_2\|_2 \leq d^{-\Omega(1)}$ is still close to the initialization.*

In this phase, the prediction head has not come into play. The substitution effect can only happen after the feature v_1 in neuron f_1 is learned to a certain degree, and neuron f_2 remains largely unlearned.

5.2 Phase II: The Substitution Effect

To illustrate the substitution effect, let us keep assuming that neuron f_1 has already learned some significant amount of the strong feature v_1, say $w_1 = \beta_1 v_1 + \textit{residual}$ with $|\beta_1| = \Omega(\|\textit{residual}\|)$, then we have: (recall $f_j(\cdot), j \in [2]$ are the neurons of the encoder network)

Lemma 5.2 (substitution effect, formal statement see Lemma D.8)**.** *After $|\langle w_1^{(t)}, v_1\rangle| = \Omega(1)$, we shall have $|E_{2,1}^{(t)}|$ increasing until $|E_{2,1}^{(t)} f_1(X^{(1)})| \gg |f_2(X^{(1)})|$ when X is equipped with the strong feature v_1, for $T_2 - T_1 = o(T_1)$ iterations.*

Intuition of the substitution effect. After the stronger feature is learned in neuron f_1, the optimal way to align two positive representations $F_2(X^{(1)})$ and $G_2(X^{(2)})$ is not learning features in weight w_2, but use the prediction head to "substitute" the features in f_1 into F_2. This is how the substitution effect happens when trained with a prediction head.

5.3 Phase III: The Acceleration Effect

After the substitution of v_1 in F_2, our concern is, $w_2^{(t)}$ will learn v_2 and only v_2 eventually, according to the acceleration effect in the following lemma.

Lemma 5.3 (acceleration effect, formal statement see Lemma E.8)**.** *After $E_{2,1}^{(t)}$ is learned in Lemma 5.2, learning v_2 in $w_2^{(t)}$ will be much faster than v_1, until $\|w_2^{(t)} - \beta_2 v_2\| = o(1)$ for some $\beta_2 = \Theta(1)$.*

The acceleration effect is caused by the interactions between the prediction head, the stop gradient operation, and the normalization method (which in this case is the batch normalization).

What is the role of the stop-gradient? Thanks to the stop-gradient operation, when we compute the gradient $-\nabla_{w_2} F_2(X^{(1)}) \cdot \mathsf{StopGrad}[G_2(X^{(2)})]$ to learn f_2, this negative gradient will only try to maximize $f_2(X^{(1)}) \cdot f_2(X^{(2)})$, rather than to maximize $F_2(X^{(1)}) \cdot f_2(X^{(2)})$. This is because the stop-gradient is on G not on F: while F_2 has a large component of v_1 borrowed from f_1 using E, G_2 does not have this component. So the gradient of F_2 is to align with the features in G_2 that does not contain many v_1, while the gradient of G_2 is to aligned with F_2 that contains a lot of v_1.

What is the role of the output normalization? Again due to the $\mathsf{StopGrad}$ operation, the gradient of $\widetilde{F}_2$ is taken with respect to the ratio $f_2(X^{(1)})/\sqrt{\mathbf{Var}[F_2(X^{(1)})]}$. As gradient descent tries to maximize this ratio, a direct computation gives

$$\nabla_{w_2} \frac{f_2(X^{(1)})}{\sqrt{\mathbf{Var}(F_2(X^{(1)}))}} \propto \textstyle\sum_{\ell \in [2]} \left([E_{2,1}^{(t)} \langle w_1^{(t)}, v_{3-\ell}\rangle^3]^2 + \mathbf{Var}[f_2(X^{(1)})] \right) \langle \nabla_{w_2} f_2(X^{(1)}), v_\ell\rangle v_\ell$$

which borrow the *substituted feature* v_1 from $f_1(\cdot)$ to adjust the gradient of v_2 in $f_2(\cdot)$, via the prediction head $E_{2,1}^{(t)}$. Without the output normalization, the learning of v_1 will dominate that of v_2 even when we train the prediction head.

5.4 The End Phase: Convergence

As the weak features are learned, we have already obtained a good encoder network $f(\cdot)$ as shown in Theorem 4.1. The rest of Theorem F.2 also contains the following result:

Proposition 5.4 (convergence of the prediction head, see Theorem F.2c)**.** *After some* $t \geq T = \mathsf{poly}(d)/\eta$ *iterations, we shall have* $\|E^{(t)} - I_2\|_F \leq \frac{1}{\mathsf{poly}(d)}$.

While we admit that only some of our real-world experiments in Figure 3 show the convergence to zero for the off-diagonal entries of the prediction head, most of the experiments do display a rise and fall trajectory pattern of off-diagonal entries consistently, which supports our theory to some degree.

6 Additional Related Work

Self-supervised learning Self-supervised learning has created huge success in natural language processing [30, 86, 18] and established the pretrain-finetune paradigm for deep learning. In vision, contrastive learning [75, 41, 24, 20, 27, 28, 34, 64, 33] became dominant in many downstream tasks recently. Another approach is the generative learning [65, 16, 42], which also gives promising results. Applications such as [64, 66] also illustrate the power of contrastive learning in multiple domains.

Theory of self-supervised learning The theoretical side of self-supervised learning developed quickly due to the success of contrastive learning. Since [12], plenty of papers have studied the contrastive learning. [25, 69] discussed many interesting phenomena associated with the negative term. Saunshi et al. [71] provided evidence that function class agnostic analyses is vacuous. [85] took a feature learning view, and inspired our analysis in the non-contrastive setting. For generative learning, [51, 74] provides downstream performance guarantees. [70, 84] studied the natural language tasks. [58] gave a recovery guarantee for tensors under hidden Markov models. [4] provided an optimization guarantee for GANs trained by stochastic gradient descent ascent.

Feature learning theory of deep learning Our theoretical results are also inspired by the recent progress of the feature learning theory of neural networks [55, 56, 5, 3, 49, 90, 46]. [55] initiate the study of the speed difference in learning different types of features. [1] developed theory for learning two-layer neural networks beyond the *neural tangent kernel* (NTK) [7, 8, 6, 32, 11]. [5, 3, 2] further studied how features are learned in different deep learning tasks. Before this recent progress, [77, 89, 19, 72, 31, 53, 54] also studied how shallow neural networks can learn on certain simple data distributions, but all of them focus on the supervised learning. There are also plenty of studies [73, 38, 10, 60, 48, 67, 29] on the implicit bias of optimization in deep learning, but none of their techniques are designed for analyzing self-supervised learning.

7 Conclusion

In this paper, we showed how the prediction head can ensure the neural network learns all the features in non-contrastive learning through theoretical investigation. Our key contribution is that we proved the prediction head can leverage two effects called substitution effect and acceleration effect during the training process. We also gave an explanation for the dimensional collapse phenomenon. We believe our theory, although based on a very simple setup, can provide some insights into the inner workings of non-contrastive self-supervised learning.

Acknowledgments and Disclosure of Funding

Funding in direct support of this work includes NSF Award 2145703.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[4] Zeyuan Allen-Zhu and Yuanzhi Li. Forward super-resolution: How can gans learn hierarchical generative models for real-world distributions. *arXiv preprint arXiv:2106.02619*, 2021.

[5] Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 977–988. IEEE, 2021. URL `https://doi.org/10.1109/FOCS52979.2021.00098`.

[6] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pages 6158–6169, 2019.

[7] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *ICML 2019 : Thirty-sixth International Conference on Machine Learning*, pages 242–252, 2019.

[8] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 6676–6688, 2019.

[9] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International conference on machine learning*, pages 1908–1916. PMLR, 2014.

[10] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[12] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[13] Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. *arXiv preprint arXiv:2106.09943*, 2021.

[14] Yamini Bansal, Gal Kaplun, and Boaz Barak. For self-supervised learning, rationality implies generalization, provably. *arXiv preprint arXiv:2010.08508*, 2020.

[15] Han Bao, Yoshihiro Nagano, and Kento Nozawa. Sharp learning bounds for contrastive unsupervised representation learning. *arXiv preprint arXiv:2110.02501*, 2021.

[16] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[17] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[18] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[19] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34^{th} International Conference on Machine Learning-Volume 70*, pages 605–614. JMLR. org, 2017.

[20] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[22] Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

[23] Sitan Chen, Jerry Li, Yuanzhi Li, and Anru R Zhang. Learning polynomial transformations. *arXiv preprint arXiv:2204.04209*, 2022.

[24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 1597–1607, 2020.

[25] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34, 2021.

[26] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[27] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[28] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[29] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[31] Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? In *International Conference on Learning Representations*, 2018.

[32] Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *36^{th} International Conference on Machine Learning, ICML 2019*, pages 1675–1685, 2019.

[33] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR, 2021.

[34] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[35] Antoine Gautier, Quynh N Nguyen, and Matthias Hein. Globally optimal training of generalized polynomial neural networks with nonlinear spectral methods. *Advances in Neural Information Processing Systems*, 29, 2016.

[36] Sven Gowal, Po-Sen Huang, Aaron van den Oord, Timothy Mann, and Pushmeet Kohli. Self-supervised adversarial robustness for the low-label, high-data regime. In *International Conference on Learning Representations*, 2020.

[37] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[38] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[39] Zhaohan Daniel Guo, Bernardo Avila Pires, Bilal Piot, Jean-Bastien Grill, Florent Altché, Rémi Munos, and Mohammad Gheshlaghi Azar. Bootstrap latent-predictive representations for multitask reinforcement learning. In *International Conference on Machine Learning*, pages 3875–3886. PMLR, 2020.

[40] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *arXiv preprint arXiv:2106.04156*, 2021.

[41] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[42] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[43] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

[44] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9598–9608, 2021.

[45] Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.

[46] Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning, 2022. URL `https://openreview.net/forum?id=lf0W6tcWmh-`.

[47] Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*, 2021.

[48] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

[49] Stefani Karp, Ezra Winston, Yuanzhi Li, and Aarti Singh. Local signal adaptivity: Provable feature learning in neural networks beyond kernels. *Advances in Neural Information Processing Systems*, 34, 2021.

[50] Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. *Advances in neural information processing systems*, 32, 2019.

[51] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[52] Yuanzhi Li and Zehao Dou. Making method of moments great again?–how can gans learn distributions. *arXiv preprint arXiv:2003.04033*, 2020.

[53] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in neural information processing systems*, pages 597–607, 2017.

[54] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *COLT 2018: 31st Annual Conference on Learning Theory*, pages 2–47, 2018.

[55] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pages 11674–11685, 2019.

[56] Yuanzhi Li, Tengyu Ma, and Hongyang R. Zhang. Learning over-parametrized two-layer relu neural networks beyond ntk. In *COLT*, pages 2613–2682, 2020.

[57] Bingbin Liu, Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Analyzing and improving the optimization landscape of noise-contrastive estimation. *arXiv preprint arXiv:2110.11271*, 2021.

[58] Bingbin Liu, Daniel Hsu, Pradeep Ravikumar, and Andrej Risteski. Masked prediction tasks: a parameter identifiability view. *arXiv preprint arXiv:2202.09305*, 2022.

[59] Zeping Luo, Cindy Weng, Shiyou Wu, Mo Zhou, and Rong Ge. One objective for all models– self-supervised learning for topic models. *arXiv preprint arXiv:2203.03539*, 2022.

[60] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[61] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[62] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[63] Ashwini Pokle, Jinjin Tian, Yuchen Li, and Andrej Risteski. Contrasting the landscape of contrastive and non-contrastive learning. *arXiv preprint arXiv:2203.15702*, 2022.

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[65] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[66] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *OpenAI blog*, 2022. URL `https://cdn.openai.com/papers/dall-e-2.pdf`.

[67] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

[68] Pierre H. Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, and Michal Valko. BYOL works even without batch statistics. *arXiv:2010.10241 [cs, stat]*, October 2020.

[69] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in Neural Information Processing Systems*, 34, 2021.

[70] Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A mathematical exploration of why language models help solve downstream tasks. *arXiv preprint arXiv:2010.03648*, 2020.

[71] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037*, 2022.

[72] Mahdi Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, pages 2007–2017, 2017.

[73] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[74] Jiaye Teng, Weiran Huang, and Haowei He. Can pretext-based self-supervised learning be boosted by downstream data? a theoretical analysis. *arXiv preprint arXiv:2103.03568*, 2021.

[75] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[76] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[77] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3404–3413. JMLR. org, 2017.

[78] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

[79] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *arXiv preprint arXiv:2003.02234*, 2020.

[80] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

[81] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

[82] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34, 2021.

[83] Xiang Wang, Xinlei Chen, Simon S Du, and Yuandong Tian. Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint arXiv:2110.04947*, 2021.

[84] Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *Advances in Neural Information Processing Systems*, 34, 2021.

[85] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.

[86] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

[87] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

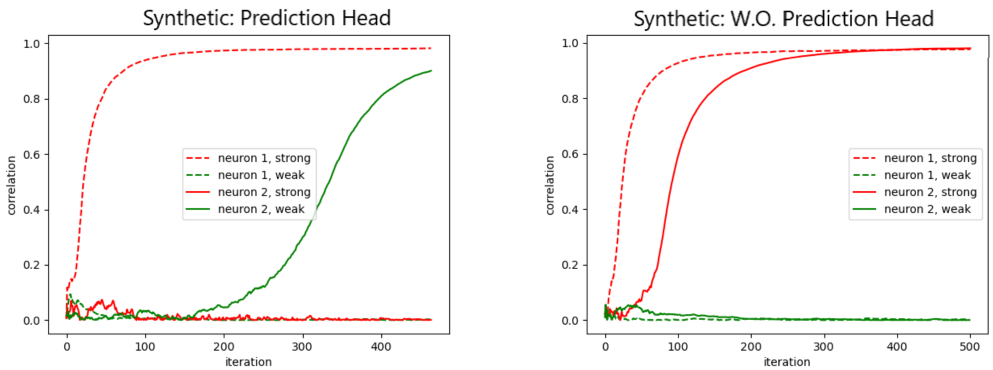
[88] Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X Pham, Chang D Yoo, and In So Kweon. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. *arXiv preprint arXiv:2203.16262*, 2022.

[89] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149. JMLR. org, 2017.

[90] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [No]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]



(a) Identity-initialized (trainable) prediction head

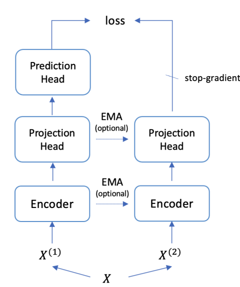
(b) Learning without prediction head

Figure 6: The feature learning process over synthetic data. When trained with the prediction head, after the strong feature is learned in the faster learning neuron, the weak feature can be learned in the slower learning neuron. When trained without the prediction head, both neurons will learn the strong feature and ignore the weak feature.

A Experiment Details

The framework we use in our experiments is shown in Figure 5. We use a modified version of the codebase shared by the authors of [33], and we use the same data augmentation in their implementation. All our experiments (except for Figure 7) use the following architecture and hyper-parameters: we choose standard ResNet-18 as base encoder architecture, 0.003 as the learning rate for Adam optimizer, a two-layer MLP with ReLU activation and 512 hidden neurons as the projection head, an identity-initialized but diagonally froze linear matrix (with shape (64x64)) as the prediction head and a non-tracking-stats, non-affine, non-momentum BN layer as the output normalization. Our experiments in Figure 3 use the same architecture and hyper-parameters, but some runs are trained with EMA with momentum 0.99, with output BN replaced by ℓ_2-norm or using different prediction heads (such as a two-layer MLP or a linear head, with Pytorch default initialization). Evaluation in Figure 1 is by training a linear classifier on top of frozen encoder with no data augmentation.

Figure 5: Framework.



B Notations and Gradients

In this section, we will give some useful notations and warm-up computations for the technical proofs in subsequent sections. We summarize here the notations that will also be defined in later sections:

Notations. We denote $\mathcal{E}_j = \mathbb{E}[\langle w_j, \xi_p\rangle^6]$, $\mathcal{E}_{j,3-j} = \mathbb{E}\left[(\langle w_j, \xi_p\rangle^3 + E_{j,3-j}\langle w_{3-j}, \xi_p\rangle^3)^2\right]$, and

$$C_0 = \frac{\mathbb{E}[|\mathcal{S}(X)\cap\mathcal{P}|\cdot|\mathcal{S}(X)\setminus\mathcal{P}|]}{2}, \quad C_1 = \frac{\mathbb{E}\left[|\mathcal{S}(X)\cap\mathcal{P}|^2\right]}{2}, \quad C_2 = P - |\mathcal{S}(X)|,$$

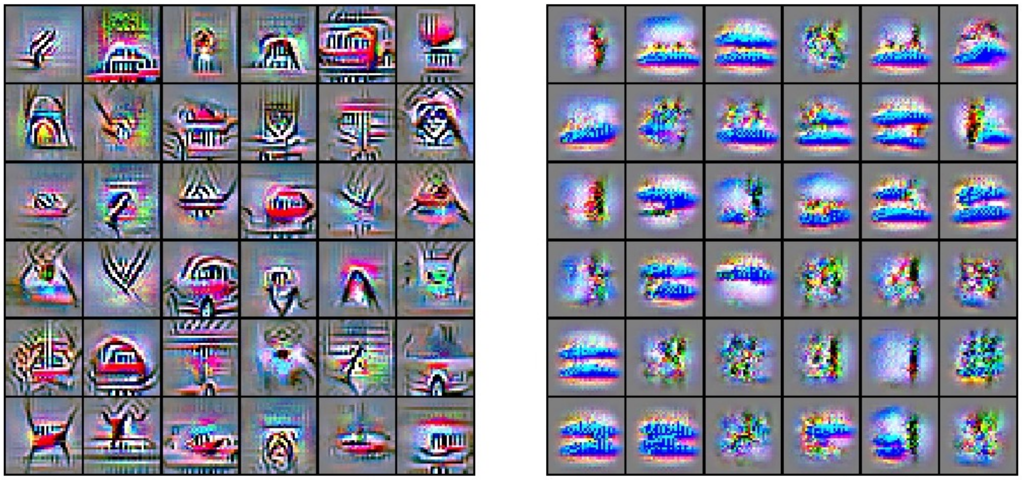
$$\bar{B}^3_{j,\ell} = \mathsf{StopGrad}[\langle w_j, v_\ell\rangle^3], \qquad B_{j,\ell} = \langle w_j, v_\ell\rangle, \qquad Q_j = (\mathbb{E}[\mathsf{StopGrad}[G_j^2(X^{(2)})]])^{-1/2}.$$

and

$$\begin{aligned} U_j &:= \mathbb{E}[F_j^2(X^{(1)})] = \textstyle\sum_{\ell\in[2]} C_1\alpha_\ell^6(B^3_{j,\ell} + E_{j,3-j}B^3_{3-j,\ell})^2 + C_2\mathcal{E}_{j,3-j} \\ H_{j,\ell} &:= C_1\alpha_\ell^6(B^3_{j,\ell} + E_{j,3-j}B^3_{3-j,\ell})^2 + C_2\mathcal{E}_{j,3-j}, \\ K_{j,\ell} &:= C_1\alpha_\ell^6(B^3_{j,\ell} + E_{j,3-j}B^3_{3-j,\ell})(B^3_{j,3-\ell} + E_{j,3-j}B^3_{3-j,3-\ell}) \end{aligned}$$

Moreover, we denote $\Phi_j := Q_j/U_j^{3/2}$, and (recall $V := \mathrm{span}(v_1, v_2)$)

$$R_j := \langle \Pi_{V^\perp} w_j, w_j\rangle \qquad R_{1,2} := \langle \Pi_{V^\perp} w_1, w_2\rangle \qquad \overline{R}_{1,2} := \frac{\langle \Pi_{V^\perp} w_1, w_2\rangle}{\|\Pi_{V^\perp} w_1\|_2\|\Pi_{V^\perp} w_2\|_2}$$



(a) Features learned with prediction head (b) Features learned without prediction head

Figure 7: Feature visualization of deep neural network. We visualized the features of an Wide-ResNet-16x5 following the `BYORL` method by Gowal et al. [36], a adversarial robust version of `BYOL`. Features learned with prediction head obviously have more variety than features learned without the prediction head. Our feature visualization technique follows from [5].

For any $j \in [2]$, the gradient $-\nabla_{w_j} L(W, E)$ can be decomposed as

$$\begin{aligned}
-\nabla_{w_j} L(W, E) &= \sum_{\ell \in [2]} (\Lambda_{j,\ell} + \Gamma_{j,\ell} - \Upsilon_{j,\ell}) v_\ell - \sum_{(j',\ell) \in [2] \times [2]} \Sigma_{j',\ell} \nabla_{w_j} \mathcal{E}_{j',3-j'} \\
\Lambda_{j,\ell} &:= C_0 \Phi_j \alpha_\ell^6 B_{j,\ell}^5 H_{j,3-\ell} \\
\Gamma_{j,\ell} &:= C_0 \Phi_{3-j} E_{3-j,j} \alpha_\ell^6 B_{3-j,\ell}^3 B_{j,\ell}^2 H_{3-j,3-\ell} \\
\Upsilon_{j,\ell} &:= C_0 \alpha_{3-\ell}^6 \left(\Phi_j B_{j,3-\ell}^3 B_{j,\ell}^2 K_{j,\ell} + \Phi_{3-j} E_{3-j,j} B_{3-j,3-\ell}^3 B_{j,\ell}^2 K_{3-j,\ell} \right) \\
\Sigma_{j,\ell} &:= C_0 C_2 \Phi_j \alpha_\ell^6 B_{j,\ell}^3 (B_{j,\ell}^3 + E_{j,3-j} B_{3-j,\ell}^3)
\end{aligned}$$

Sometimes we need to decompose $\Upsilon_{j,\ell} = \Upsilon_{j,\ell,1} + \Upsilon_{j,\ell,2}$ which is straightforward from its expression. In Section E, we further define

$$\begin{aligned}
\Xi_j^{(t)} &= C_0 C_1 \alpha_1^6 \alpha_2^6 \Phi_j^{(t)} \left((B_{1,1}^{(t)})^6 (B_{2,2}^{(t)})^6 + (B_{2,1}^{(t)})^6 (B_{1,2}^{(t)})^6 \right) \\
\Delta_{j,\ell}^{(t)} &= C_0 \Phi_j^{(t)} \alpha_\ell^6 (B_{j,\ell}^{(t)})^3 (B_{3-j,\ell}^{(t)})^3 C_2 \mathcal{E}_{j,3-j}^{(t)}
\end{aligned}$$

for the gradients of the prediction head.

B.1 Gradient Computation

Let us $L(W, E)$ to be the population version of the objective. Because $\mathbb{E}[F_j(X^{(1)})]$ and $\mathbb{E}[G_j(X^{(2)})]$ are both zero (which can be verified easily from the zero-mean assumptions of $z_p(X)$ and ξ_p), a direct computation gives:

$$L(W, E) = 2 - \sum_{j \in [2]} \frac{\mathbb{E}[F_j(X^{(1)}) \cdot \mathsf{StopGrad}[G_j(X^{(2)})]]}{\sqrt{\mathbb{E}[F_j^2(X^{(1)})]} \sqrt{\mathbb{E}[\mathsf{StopGrad}[G_j^2(X^{(2)})]]}}$$

We first calculate the normalizing quantity $\mathbb{E}[F_j^2(X^{(1)})]$:

$$\mathbb{E}[F_j^2(X^{(1)})] = \mathbb{E}\left[\left(\sum_{p\in[P]}\sigma(\langle w_j, X_p^{(1)}\rangle) + E_{j,3-j}\sigma(\langle w_{3-j}, X_p^{(1)}\rangle)\right)^2\right]$$

$$= \frac{1}{2}\sum_{\ell\in[2]}\mathbb{E}\left[|\mathcal{S}(X)\cap\mathcal{P}|^2\alpha_\ell^6(\langle w_j, v_\ell\rangle^3 + E_{j,3-j}\langle w_{3-j}, v_\ell\rangle^3)^2\right]$$

(Because all signal patches has the same sign within the same data)

$$+ \mathbb{E}\left[|\mathcal{P}\setminus|\mathcal{S}(X)|(\langle w_j, \xi_p\rangle^3 + E_{j,3-j}\langle w_{3-j}, \xi_p\rangle^3)^2\right]$$

(Because noise patches are independent and have mean zero)

$$= \sum_{\ell\in[2]}\alpha_\ell^6(\langle w_j, v_\ell\rangle^3 + E_{j,3-j}\langle w_{3-j}, v_\ell\rangle^3)^2\frac{\mathbb{E}\left[|\mathcal{S}(X)\cap\mathcal{P}|^2\right]}{2} + (P - |\mathcal{S}(X)|)\mathcal{E}_{j,3-j}$$

where we let

$$\mathcal{E}_{j,3-j} \stackrel{\text{def}}{=} \mathbb{E}\left[(\langle w_j, \xi_p\rangle^3 + E_{j,3-j}\langle w_{3-j}, \xi_p\rangle^3)^2\right]$$
$$= \mathbb{E}\left[\langle w_j, \xi_p\rangle^6 + 2E_{j,3-j}\langle w_j, \xi_p\rangle^3\langle w_{3-j}, \xi_p\rangle^3 + E_{j,3-j}^2\langle w_{3-j}, \xi_p\rangle^6\right]$$

On the other hand, we have

$$\mathbb{E}[F_j(X^{(1)})\cdot\mathsf{StopGrad}[G_j(X^{(2)})]]$$
$$= \mathbb{E}\left[\left(\sum_{p\in[P]}\sigma(\langle w_j, X_p^{(1)}\rangle) + E_{j,3-j}\sigma(\langle w_{3-j}, X_p^{(1)}\rangle)\right)\times\left(\sum_{p\in[P]}\sigma(\langle w_j, X_p^{(2)}\rangle)\right)\right]$$
$$= \frac{1}{2}\sum_{\ell\in[2]}\mathbb{E}\left[\sum_{p\in|\mathcal{S}(X)\cap\mathcal{P}}\alpha_\ell^3(\langle w_j, v_\ell\rangle^3 + E_{j,3-j}\langle w_{3-j}, v_\ell\rangle^3)\times\sum_{p\in|\mathcal{S}(X)\setminus\mathcal{P}}\alpha_\ell^3\mathsf{StopGrad}[\langle w_j, v_\ell\rangle^3]\right]$$
$$= \sum_{\ell\in[2]}\alpha_\ell^6(\langle w_j, v_\ell\rangle^3 + E_{j,3-j}\langle w_{3-j}, v_\ell\rangle^3)\cdot\mathsf{StopGrad}[\langle w_j, v_\ell\rangle^3]\cdot\frac{\mathbb{E}[|\mathcal{S}(X)\cap\mathcal{P}|\cdot|\mathcal{S}(X)\setminus\mathcal{P}|]}{2}$$

Now, by denoting

$$C_0 = \frac{\mathbb{E}[|\mathcal{S}(X)\cap\mathcal{P}|\cdot|\mathcal{S}(X)\setminus\mathcal{P}|]}{2}, \quad C_1 = \frac{\mathbb{E}\left[|\mathcal{S}(X)\cap\mathcal{P}|^2\right]}{2}, \quad C_2 = P - |\mathcal{S}(X)|,$$
$$\bar{B}_{j,\ell}^3 = \mathsf{StopGrad}[\langle w_j, v_\ell\rangle^3], \quad B_{j,\ell} = \langle w_j, v_\ell\rangle, \quad Q_j = (\mathbb{E}[\mathsf{StopGrad}[G_j^2(X^{(2)})]])^{-1/2}.$$

we denote $U_j := \mathbb{E}[F_j^2(X^{(1)})]$, where the expanded expression is

$$U_j = \mathbb{E}[F_j^2(X^{(1)})] = \sum_{\ell\in[2]}C_1\alpha_\ell^6(B_{j,\ell}^3 + E_{j,3-j}B_{3-j,\ell}^3)^2 + C_2\mathcal{E}_{j,3-j}$$

and we can rewrite the objective as follows

$$L(W,E) = 2 - \sum_{j\in[2]}\sum_{\ell\in[2]}\frac{Q_jC_0\alpha_\ell^6\bar{B}_{j,\ell}^3(B_{j,\ell}^3 + E_{j,3-j}B_{3-j,\ell}^3)}{U_j^{1/2}} \tag{B.1}$$

Now denote

$$H_{j,\ell} = C_1\alpha_\ell^6(B_{j,\ell}^3 + E_{j,3-j}B_{3-j,\ell}^3)^2 + C_2\mathcal{E}_{j,3-j},$$
$$K_{j,\ell} = C_1\alpha_\ell^6(B_{j,\ell}^3 + E_{j,3-j}B_{3-j,\ell}^3)(B_{j,3-\ell}^3 + E_{j,3-j}B_{3-j,3-\ell}^3)$$

It is easy to calculate

$$\begin{aligned}
Q_j^{-2} &= \mathbb{E}[\mathsf{StopGrad}[G_j^2(X^{(2)})]] \\
&= \mathbb{E}\left[\left(\sum_{p\in[P]} \sigma(\langle w_j, X_p^{(2)}\rangle)\right)^2\right] \\
&= \frac{1}{2}\sum_{\ell\in[2]} \alpha_\ell^6 \langle w_j, v_\ell\rangle^6 \mathbb{E}\left[|\mathcal{S}(X)\cap\mathcal{P}|^2\right] + \mathbb{E}\left[|\mathcal{P}\setminus|\mathcal{S}(X)|\langle w_j, \xi_p\rangle^6\right] \\
&= \sum_{\ell\in[2]} C_1\alpha_\ell^6 B_{j,\ell}^6 + C_2\mathcal{E}_j
\end{aligned}$$

where $\mathcal{E}_j = \mathbb{E}[\langle w_j, \xi_p\rangle^6]$. And thus the gradient can be computed as (notice $\bar{B}_{j,\ell}^3 = B_{j,\ell}^3$)

$$\begin{aligned}
-\nabla_{w_j} L(W,E) &= \sum_{\ell\in[2]}\left(\frac{C_0 Q_j \alpha_\ell^6 H_{j,3-\ell} B_{j,\ell}^5}{U_j^{3/2}}\right) v_\ell + \sum_{\ell\in[2]}\left(\frac{C_0 Q_{3-j} E_{3-j,j}\alpha_\ell^6 B_{3-j,\ell}^3 B_{j,\ell}^2 H_{3-j,3-\ell}}{U_{3-j}^{3/2}}\right) v_\ell \\
&\quad - \sum_{\ell\in[2]}\left(\frac{C_0 Q_j \alpha_{3-\ell}^6 B_{j,3-\ell}^3 B_{j,\ell}^2 K_{j,\ell}}{U_j^{3/2}} + \frac{C_0 Q_{3-j} E_{3-j,j}\alpha_{3-\ell}^6 B_{3-j,3-\ell}^3 B_{j,\ell}^2 K_{3-j,\ell}}{U_{3-j}^{3/2}}\right) v_\ell \\
&\quad - \sum_{j'\in[2]}\sum_{\ell\in[2]} \frac{C_0 C_2 Q_{j'}\alpha_\ell^6 B_{j',\ell}^3 (B_{j',\ell}^3 + E_{j',3-j'} B_{3-j',\ell}^3)}{U_{j'}^{3/2}} \nabla_{w_j}\mathcal{E}_{j',3-j'} \\
&= \sum_{\ell\in[2]} (\Lambda_{j,\ell} + \Gamma_{j,\ell} - \Upsilon_{j,\ell}) v_\ell - \sum_{(j',\ell)\in[2]\times[2]} \Sigma_{j',\ell}\nabla_{w_j}\mathcal{E}_{j',3-j'} \qquad \text{(B.2)}
\end{aligned}$$

where

$$\begin{aligned}
\nabla_{w_j}\mathcal{E}_{j,3-j} &= 6\mathbb{E}[\langle w_j, \xi_p\rangle^5 \xi_p + E_{j,3-j}\langle w_j, \xi_p\rangle^2 \langle w_{3-j}, \xi_p\rangle^3 \xi_p] \\
\nabla_{w_j}\mathcal{E}_{3-j,j} &= 6\mathbb{E}[E_{3-j,j}^2 \langle w_j, \xi_p\rangle^5 \xi_p + E_{3-j,j}\langle w_{3-j}, \xi_p\rangle^3 \langle w_j, \xi_p\rangle^2 \xi_p]
\end{aligned}$$

As for the gradient of the prediction head, we can calculate

$$\begin{aligned}
-\nabla_{E_{j,3-j}} L(W,E) &= \sum_{\ell\in[2]} \frac{C_0 Q_j \alpha_\ell^6 B_{j,\ell}^3 B_{3-j,\ell}^3 U_j}{U_j^{3/2}} \\
&\quad - \sum_{\ell\in[2]} \frac{C_0 Q_j \alpha_\ell^6 B_{j,\ell}^3 (B_{j,\ell}^3 + E_{j,3-j} B_{3-j,\ell}^3)\sum_{\ell'\in[2]} C_1\alpha_{\ell'}^6 (B_{j,\ell'}^3 + E_{j,3-j} B_{3-j,\ell'}^3) B_{3-j,\ell'}^3}{U_j^{3/2}} \\
&\quad - \sum_{\ell\in[2]} \frac{C_0 C_2 Q_j \alpha_\ell^6 B_{j,\ell}^3 (B_{j,\ell}^3 + E_{j,3-j} B_{3-j,\ell}^3)}{U_j^{3/2}} \nabla_{E_{j,3-j}}\mathcal{E}_{j,3-j} \\
&= \sum_{\ell\in[2]} \frac{C_0 Q_j \alpha_\ell^6 B_{j,\ell}^3 (B_{3-j,\ell}^3 H_{j,3-\ell} - B_{3-j,3-\ell}^3 K_{j,3-\ell})}{U_j^{3/2}} \\
&\quad - \sum_{\ell\in[2]} \Sigma_{j,\ell}\mathbb{E}\left[2\langle w_j, \xi_p\rangle^3 \langle w_{3-j}, \xi_p\rangle^3 + 2E_{j,3-j}\langle w_{3-j}, \xi_p\rangle^6\right]
\end{aligned}$$

where $\Sigma_{j,\ell}$ is defined in (B.2). In fact, all the above gradient expressions can be simplified by letting $\Phi_j := Q_j/U_j^{3/2}$ for $j \in [2]$, which is what we shall do in later sections.

Summarizing the notations. We shall define some useful notations to simplify the proof. We define $V = \text{span}(v_1, v_2)$. Let Π_A be the projection operator to subspace $A \subset \mathbb{R}^d$, then

$$R_j := \langle \Pi_{V^\perp} w_j, w_j\rangle \qquad R_{1,2} := \langle \Pi_{V^\perp} w_1, w_2\rangle \qquad \overline{R}_{1,2} := \frac{\langle \Pi_{V^\perp} w_1, w_2\rangle}{\|\Pi_{V^\perp} w_1\|_2 \|\Pi_{V^\perp} w_2\|_2}$$

B.2 Some Useful Bounds for Gradients

In this section we use the superscript $^{(t)}$ to denote the iteration t during training. Below we present a claim which comes from direct calculations of $\Sigma_{j,\ell}^{(t)}$ and $\nabla_{w_j}\mathcal{E}_{j',3-j'}^{(t)}$, which is very useful in the following sections.

Claim B.1 (on $\Sigma_{j,\ell}^{(t)}$ and $\nabla_{w_j}\mathcal{E}_{j',3-j'}^{(t)}$)**.** *Let $R_j, R_{1,2}^{(t)}$ be defined as above, then we have*

(a) $\Sigma_{j,\ell}^{(t)} = O(\Sigma_{1,1}^{(t)})\frac{(B_{j,\ell}^{(t)})^6 + E_{j,3-j}^{(t)}(B_{3-j,\ell}^{(t)})^3(B_{j,\ell}^{(t)})^3}{(B_{1,1}^{(t)})^6}\frac{\Phi_j^{(t)}}{\Phi_1^{(t)}}$;

(b) $\langle \nabla_{w_j}\mathcal{E}_{j,3-j}^{(t)}, \Pi_{V^\top} w_j^{(t)}\rangle = \Theta([R_j^{(t)}]^3) \pm \Theta(E_{j,3-j}^{(t)})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$;

(c) $\langle \nabla_{w_j}\mathcal{E}_{3-j,j}^{(t)}, w_j^{(t)}\rangle = \Theta((E_{3-j,j}^{(t)})^2)[R_j^{(t)}]^3 \pm O(E_{3-j,j}^{(t)})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$

(d) $\langle \nabla_{w_j}\mathcal{E}_{j,3-j}^{(t)}, w_{3-j}^{(t)}\rangle = (\Theta(\overline{R}_{1,2}^{(t)}) \pm \varrho)[R_j^{(t)}]^{5/2}[R_{3-j}^{(t)}]^{1/2} + O(E_{j,3-j}^{(t)})R_j^{(t)}[R_{3-j}^{(t)}]^2$;

(e) $\langle \nabla_{w_j}\mathcal{E}_{3-j,j}^{(t)}, w_{3-j}^{(t)}\rangle = ((E_{3-j,j}^{(t)})^2(\Theta(\overline{R}_{1,2}^{(t)}) \pm \varrho)[R_j^{(t)}]^{5/2}[R_{3-j}^{(t)}]^{1/2} + O(E_{3-j,j}^{(t)})R_j^{(t)}[R_{3-j}^{(t)}]^2)$

Proof. The part on $\Sigma_{j,\ell}^{(t)}$ is trivial from its expression, we shall focus on proving (b) – (d).
On $\langle \nabla_{w_j}\mathcal{E}_{j',3-j'}^{(t)}, w_j^{(t)}\rangle$: If $j = j'$, then

$$\begin{aligned}\langle \nabla_{w_j}\mathcal{E}_{j,3-j}^{(t)}, w_j^{(t)}\rangle &= \Theta(1)\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^6 + E_{j,3-j}^{(t)}\langle w_j^{(t)}, \xi_p\rangle^3\langle w_{3-j}^{(t)}, \xi_p\rangle^3] \\ &= \Theta(1)\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^6] + O(E_{j,3-j}^{(t)})\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^3(\langle w_{3-j}^{(t)}, \xi_p\rangle^3 \\ &\quad - \langle (I - \bar{w}_{j,t}\bar{w}_{j,t}^\top)w_{3-j}^{(t)}, \xi_p\rangle^3)] \\ &\quad + O(E_{j,3-j}^{(t)})\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^3\langle (I - \bar{w}_{j,t}\bar{w}_{j,t}^\top)w_{3-j}^{(t)}, \xi_p\rangle^3]\end{aligned}$$

Write $\bar{w}_{j,t} = \frac{\Pi_{V^\perp} w_j^{(t)}}{\|\Pi_{V^\perp} w_j^{(t)}\|_2}$, we can derive

$$\begin{aligned}&\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^3(\langle w_{3-j}^{(t)}, \xi_p\rangle^3 - \langle (I - \bar{w}_{j,t}\bar{w}_{j,t}^\top)w_{3-j}^{(t)}, \xi_p\rangle^3)] \\ &= \mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^3\langle \bar{w}_{j,t}\bar{w}_{j,t}^\top w_{3-j}^{(t)}, \xi_p\rangle O(\langle w_{3-j}^{(t)}, \xi_p\rangle^2)] \\ &= O(\frac{R_{1,2}^{(t)}}{\|\Pi_{V^\perp} w_j^{(t)}\|_2^2})\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^4\langle w_{3-j}^{(t)}, \xi_p\rangle^2] \\ &\le O(\frac{R_{1,2}^{(t)}}{\|\Pi_{V^\perp} w_j^{(t)}\|_2^2})\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^6]^{\frac{2}{3}}\mathbb{E}[\langle w_{3-j}^{(t)}, \xi_p\rangle^6]^{\frac{1}{3}} \qquad \text{(by Hölder's inequality)} \\ &\le O(\overline{R}_{1,2}^{(t)})\|\Pi_{V^\perp} w_j^{(t)}\|_2^3\|\Pi_{V^\perp} w_{3-j}^{(t)}\|_2^3\end{aligned}$$

and by our assumption on noise ξ_p, we also have

$$\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^3\langle (I - \bar{w}_{j,t}\bar{w}_{j,t}^\top)w_{3-j}^{(t)}, \xi_p\rangle^3] \le O(\varrho)\|\Pi_{V^\perp} w_j^{(t)}\|_2^3\|\Pi_{V^\perp} w_{3-j}^{(t)}\|_2^3$$

Combined with the fact that $\mathbb{E}[\langle w_j^{(t)}, \xi_p\rangle^6] = O(\|\Pi_{V^\perp} w_j^{(t)}\|_2^3)$, we can get

$$\langle \nabla_{w_j}\mathcal{E}_{j,3-j}^{(t)}, w_j^{(t)}\rangle = O(\|\Pi_{V^\perp} w_j^{(t)}\|_2^6) \pm O(E_{j,3-j}^{(t)})(R_{1,2}^{(t)} + \varrho)\|\Pi_{V^\perp} w_j^{(t)}\|_2^3\|\Pi_{V^\perp} w_{3-j}^{(t)}\|_2^3$$

when $j' = 3 - j$, we also have

$$\begin{aligned}\langle \nabla_{w_j}\mathcal{E}_{3-j,j}^{(t)}, w_j^{(t)}\rangle &= \Theta(1)\mathbb{E}[(E_{3-j,j}^{(t)})^2\langle w_j^{(t)}, \xi_p\rangle^6 + E_{3-j,j}^{(t)}\langle w_j^{(t)}, \xi_p\rangle^3\langle w_{3-j}^{(t)}, \xi_p\rangle^3] \\ &= O((E_{3-j,j}^{(t)})^2)\|\Pi_{V^\perp} w_j^{(t)}\|_2^6 \pm O(E_{3-j,j}^{(t)})(R_{1,2}^{(t)} + \varrho)\|\Pi_{V^\perp} w_j^{(t)}\|_2^3\|\Pi_{V^\perp} w_{3-j}^{(t)}\|_2^3\end{aligned}$$

On $\langle \nabla_{w_j} \mathcal{E}^{(t)}_{j',3-j'}, w^{(t)}_{3-j}\rangle$: when $j' = j$, we have

$$\begin{aligned}
\langle \nabla_{w_j} \mathcal{E}^{(t)}_{j,3-j}, w^{(t)}_{3-j}\rangle &= O(1)\mathbb{E}[\langle w^{(t)}_j, \xi_p\rangle^5 \langle w^{(t)}_{3-j}, \xi_p\rangle + E^{(t)}_{j,3-j}\langle w^{(t)}_j, \xi_p\rangle^2 \langle w^{(t)}_{3-j}, \xi_p\rangle^4] \\
&= O(1)\mathbb{E}[\langle w^{(t)}_j, \xi_p\rangle^5 \langle (I - \bar{w}_{j,t}\bar{w}_{j,t}^\top + \bar{w}_{j,t}\bar{w}_{j,t}^\top) w^{(t)}_{3-j}, \xi_p\rangle] \\
&\quad + O(1)\mathbb{E}[E^{(t)}_{j,3-j}\langle w^{(t)}_j, \xi_p\rangle^2 \langle w^{(t)}_{3-j}, \xi_p\rangle^4]
\end{aligned} \tag{B.3}$$

Using Hölder's inequality and our assumpsion on ξ_p, we have

$$\mathbb{E}[\langle w^{(t)}_j, \xi_p\rangle^5 \langle (I - \bar{w}_{j,t}\bar{w}_{j,t}^\top) w^{(t)}_{3-j}, \xi_p\rangle] \lesssim \varrho \|\Pi_{V^\perp} w^{(t)}_j\|_2^5 \|\Pi_{V^\perp} w^{(t)}_{3-j}\|_2$$

In the meantime, we also have

$$\mathbb{E}[\langle w^{(t)}_j, \xi_p\rangle^5 \langle \bar{w}_{j,t}\bar{w}_{j,t}^\top w^{(t)}_{3-j}, \xi_p\rangle] = \Theta(\overline{R}^{(t)}_{1,2})\mathbb{E}[\langle w^{(t)}_j, \xi_p\rangle^6][R^{(t)}_j]^{-1/2}[R^{(t)}_{3-j}]^{1/2} = \Theta(\overline{R}^{(t)}_{1,2})[R^{(t)}_j]^{5/2}[R^{(t)}_{3-j}]^{1/2}$$

for the last term in (B.3), we can also use Hölder's inequality to get

$$E^{(t)}_{j,3-j}\mathbb{E}[\langle w^{(t)}_j, \xi_p\rangle^2 \langle w^{(t)}_{3-j}, \xi_p\rangle^4] \lesssim E^{(t)}_{j,3-j}\mathbb{E}[\langle w^{(t)}_j, \xi_p\rangle^6]^{1/3}\mathbb{E}[\langle w^{(t)}_{3-j}, \xi_p\rangle^6]^{2/3} \lesssim E^{(t)}_{j,3-j} R^{(t)}_j [R^{(t)}_{3-j}]^2$$

Therefore, we can combine above analysis to get

$$\langle \nabla_{w_j} \mathcal{E}^{(t)}_{j,3-j}, w^{(t)}_{3-j}\rangle = (\Theta(\overline{R}^{(t)}_{1,2}) \pm \varrho)[R^{(t)}_j]^{5/2}[R^{(t)}_{3-j}]^{1/2} + O(E^{(t)}_{j,3-j}) R^{(t)}_j [R^{(t)}_{3-j}]^2$$

When $j' = 3-j$, we also have

$$\begin{aligned}
\langle \nabla_{w_j} \mathcal{E}^{(t)}_{3-j,j}, w^{(t)}_{3-j}\rangle &= 6\mathbb{E}[(E^{(t)}_{3-j,j})^2\langle w^{(t)}_j, \xi_p\rangle^5 \langle w^{(t)}_{3-j}, \xi_p\rangle + E^{(t)}_{3-j,j}\langle w^{(t)}_j, \xi_p\rangle^2 \langle w^{(t)}_{3-j}, \xi_p\rangle^4] \\
&= 6(E^{(t)}_{3-j,j})^2(\Theta(\overline{R}^{(t)}_{1,2}) \pm \varrho)[R^{(t)}_j]^{5/2}[R^{(t)}_{3-j}]^{1/2} + E^{(t)}_{3-j,j} R^{(t)}_j [R^{(t)}_{3-j}]^2
\end{aligned}$$

which proves the claim. □

C Phase I: Learning the Stronger Feature

In this section, we shall discuss the initial phase of learning the stronger feature. Firstly, we establish some properties at the initialization for our induction afterwards.

Initialization properties. We prove the following properties for our network at initialization. Recall our initialization is $w^{(0)}_j \sim \mathcal{N}(0, I_d/d), \forall j \in [2]$ and $E^{(0)} = I_2$.

Lemma C.1 (properties at initialization)**.** *Recall that without loss of generality we let $|B^{(0)}_{1,1}| = \max_{j\in[2]} |B^{(0)}_{j,1}|$. With probability $1 - o(1)$, the following holds:*

(a) $\|w^{(0)}_j\|_2^2 = 1 \pm \widetilde{O}(\frac{1}{\sqrt{d}})$ *for all* $j \in [2]$, *and* $|\langle w^{(0)}_1, w^{(0)}_2\rangle| \le \widetilde{O}(\frac{1}{\sqrt{d}})$;

(b) $\max_{j,\ell} |B^{(0)}_{j,\ell}| \le O(\sqrt{\log d/d})$ *and* $\min_{j,\ell} |B^{(0)}_{j,\ell}| \ge \Omega(\frac{1}{\log d}) \max_{j,\ell} |B^{(0)}_{j,\ell}|$;

(c) $|B^{(0)}_{1,1}| \ge |B^{(0)}_{2,1}|(1 + \frac{1}{\log d})$;

(d) $\mathcal{E}^{(0)}_j = (1 - O(\frac{1}{d^3}))\sigma^6 \|w^{(0)}_j\|_2^6 = \Theta(1)$ *for all* $j \in [2]$;

(e) $H^{(0)}_{j,\ell} = C_2 \mathcal{E}^{(0)}_j (1 + \widetilde{O}(\frac{1}{\sqrt{d}}))$ *for all* $(j,\ell) \in [2] \times [2]$;

(f) $U^{(0)}_j = C_2 \mathcal{E}^{(0)}_j (1 + \widetilde{O}(\frac{\alpha_1^6}{\sqrt{d}}))$ *for all* $j \in [2]$;

(g) $(Q^{(0)}_j)^{-2} = C_2 \mathcal{E}^{(0)}_j (1 + \widetilde{O}(\frac{\alpha_1^6}{\sqrt{d}}))$ *for all* $j \in [2]$;

(h) $K^{(0)}_{j,\ell} \le \widetilde{O}(\alpha_\ell^6/d^3)$ *for all* $(j,\ell) \in [2] \times [2]$.

Let us first introduce a fact about Gaussian ratio distribution without proof.

Fact C.2 (Gaussian ratio distribution)**.** If X and Y are two independent standard Gaussian variables, then the probability density of $Z = X/Y$ is $p(z) = \frac{1}{\pi(1+z^2)}, z \in (-\infty, \infty)$.

Proof of Lemma C.1. a. Norm bound comes from simple χ^2 concentration inequality and our initialization $w_j^{(0)} \sim \mathcal{N}(0, \frac{I_d}{d})$. The inner product bound comes from Gaussian concentration.

b. It is from a direct calculation under our initialization, and some application of Gaussian c.d.f. and a union bound.

c. It is from a probability distribution of Gaussian ratio distribution from Fact C.2 to bound the probability of $|B_{1,1}^{(0)}|/|B_{2,1}^{(0)}| \leq (1 + \frac{1}{\log d})$ (WLOG we let $|B_{1,1}^{(0)}| = \max_{j\in[2]} |B_{j,1}^{(0)}|$).

d. It can be directly proven from our assumption on noise ξ_p in the subspace $V^\perp$ and (a).

e. Since at the initialization we have $B_{j,\ell}^{(0)} = \widetilde{O}(\frac{1}{\sqrt{d}}), j, \ell \in [2]$ and $E_{j,3-j}^{(0)} = 0$, it is easy to directly upper bound the errors.

f. Again from $B_{j,\ell}^{(0)} = \widetilde{O}(\frac{1}{\sqrt{d}}), \forall j, \ell \in [2]$ at initialization and a direct upper bound.

g. Proof is similar to (e).

h. Directly from a naive upper bound using (b).

□

C.1 Induction in Phase I

We define phase I as all iterations $t \leq T_1$, where $T_1 := \min\{t : B_{1,1}^{(t)} \geq 0.01\}$, we will prove the existence of T_1 at the end of this section. We state the following induction hypotheses, which will hold throughout the phase I:

Inductions C.3. *For each $t \leq T_1$, all of the followings hold:*

(a). $\|w_j^{(t)}\|_2 = \|w_j^{(0)}\|_2 \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$ *for each* $j \in [2]$*;*

(b). $|B_{1,2}^{(t)}|, |B_{2,1}^{(t)}|, |B_{2,2}^{(t)}| = \widetilde{\Theta}(\frac{1}{\sqrt{d}})$*;*

(c). $|B_{1,1}^{(t)}| \geq \Omega(\frac{1}{\log d}) \max(|B_{1,2}^{(t)}|, |B_{2,2}^{(t)}|, |B_{2,1}^{(t)}|)$*;*

(d). $|E_{1,2}^{(t)}| \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})\frac{\eta_E}{\eta}|B_{1,1}^{(t)}|$ *and* $|E_{2,1}^{(t)}| \leq \widetilde{O}(\frac{1}{d})$*;*

(e). $R_1^{(t)}, R_2^{(t)} = \Theta(1)$, $|R_{1,2}^{(t)}| \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$

Remark C.4. Since we have chosen $\eta_E \leq \eta$ and $\varrho \leq \frac{1}{d^{\Omega(1)}}$, Induction C.3d implies $|E_{j,3-j}^{(t)}| = o(1)$ throughout $t \leq T_1$.

We shall prove the above induction holds in later sections, but first we need some useful claims assuming our induction holds in this phase.

C.2 Computing Variables at Phase I

Firstly we establish a claim controlling the noise terms $\mathcal{E}_j, \mathcal{E}_{j,3-j}$ during this phase.

Claim C.5. *At each iteration $t \leq T_1$, if Induction C.3 holds, then*

(a) $\mathcal{E}_1^{(t)} = \mathcal{E}_2^{(t)} \pm O(\sum_{\ell\in[2]} |B_{j,\ell}^{(t)}| + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}))$

(b) $\mathcal{E}_j^{(t)} = \mathcal{E}_j^{(0)} \pm O(\sum_{\ell\in[2]} |B_{j,\ell}^{(t)}| + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}))$

(c) $\mathcal{E}_{j,3-j}^{(t)} = \mathcal{E}_j^{(t)} \pm \widetilde{O}(E_{j,3-j}^{(t)}(\varrho + \frac{1}{\sqrt{d}}) + (E_{j,3-j}^{(t)})^2)$;

Proof. For (a), we can simply write down

$$\mathcal{E}_j^{(t)} = \mathbb{E}[\langle w_j, \xi_p\rangle^6] = \sigma^6 \|\Pi_{V^\perp} w_j^{(t)}\|_2^6$$

Note that by Induction C.3a we always have $\|w_j^{(t)}\|_2 = \|w_j^{(0)}\|_2 \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$, and by Lemma C.1a we also have $\|w_j^{(0)}\|_2 = (1 \pm \widetilde{O}(\frac{1}{\sqrt{d}}))\|w_j^{(0)}\|_2$, which implies

$$\begin{aligned}
\|\Pi_{V^\perp} w_j^{(t)}\|_2 - \|\Pi_{V^\perp} w_{3-j}^{(t)}\|_2 &= \|w_j^{(t)}\|_2 - \|w_{3-j}^{(t)}\|_2 \pm O(\sum_{j,\ell\in[2]^2} B_{j,\ell}^{(t)}) \\
&= \|w_j^{(0)}\|_2 - \|w_{3-j}^{(0)}\|_2 \pm O(\sum_{j,\ell\in[2]^2} B_{j,\ell}^{(t)}) \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \\
&= \widetilde{O}(\frac{1}{\sqrt{d}}) \pm O(\sum_{j,\ell\in[2]^2} B_{j,\ell}^{(t)}) \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})
\end{aligned}$$

By the elementary equality $x^n - y^n = (x-y)\sum_{0\le i\le n-1} x^i y^{n-1-i}$, we can obtain (a). The proof of (b) is almost the same as (a), and the proof of (c) is just direct calculation. □

Equipped with Claim C.5, we can establish the following lemma, which will be frequently applied to bound the gradient in our induction argument.

Lemma C.6 (variables control in phase I)**.** *Suppose Induction C.3 holds at some iteration $t \le T_1$, then we have:*

(a) *if* $\forall \ell \in [2], \alpha_\ell |B_{j,\ell}^{(t)}| \le O(1)$, *then* $\Phi_j^{(t)} = (C_2 \mathcal{E}_j^{(t)})^{-2}(1 \pm \frac{1}{\mathsf{polylog}(d)})$;

(b) *if* $\exists \ell \in [2], |B_{j,\ell}^{(t)}| \ge \Omega(\frac{1}{\alpha_\ell})$, *then* $\Phi_j^{(t)} = O((C_2\mathcal{E}_j^{(t)} + \sum_{\ell\in[2]} C_1\alpha_\ell^6 (B_{j,\ell}^{(t)})^6)^{-2})$;

(c) *if* $\alpha_\ell |B_{j,\ell}^{(t)}| \le O(1)$, $H_{j,\ell}^{(t)} = C_2\mathcal{E}_j^{(t)}(1 + \frac{1}{\mathsf{polylog}(d)}) = \Theta(C_2)$, *otherwise* $H_{j,\ell}^{(t)} \in [\Omega(C_2), \widetilde{O}(\alpha_\ell^6)]$

(d) $|K_{j,\ell}^{(t)}| \le \widetilde{O}(\alpha_\ell^6/d^{3/2})$

Proof. (a) From our assumptions that $|B_{1,2}^{(t)}|, |B_{2,1}^{(t)}|, |B_{2,2}^{(t)}| \le \widetilde{O}(\frac{1}{\sqrt{d}})$ and $\alpha_1 B_{1,1}^{(t)} \le O(1)$, and also the fact that $\mathcal{E}_j^{(t)} = \Omega(\sigma^6) = \Omega(1)$, $C_2 = \Theta(\mathsf{polylog}(d)) \gg C_1$, we can calculate

$$\begin{aligned}
U_j^{(t)} &= \sum_{\ell\in[2]} C_1\alpha_\ell^6((B_{j,\ell}^{(t)})^3 + E_{j,3-j}^{(t)}(B_{3-j,\ell}^{(t)})^3)^2 + C_2\mathcal{E}_{j,3-j}^{(t)} \\
&= O(C_1) + C_2\mathcal{E}_j^{(t)} + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \\
&= C_2\mathcal{E}_j^{(t)}(1 \pm \frac{1}{\mathsf{polylog}(d)})
\end{aligned}$$

Meanwhile, we can also compute similarly

$$Q_j^{(t)} = \sum_{\ell\in[2]} C_1\alpha_\ell^6(B_{j,\ell}^{(t)})^6 + C_2\mathcal{E}_j = C_2\mathcal{E}_j^{(t)}(1 \pm \frac{1}{\mathsf{polylog}(d)})$$

Therefore $\Phi_j^{(t)} = Q_j^{(t)}/(U_j^{(t)})^{3/2} = (C_2\mathcal{E}_j^{(t)}(1 \pm \frac{1}{\mathsf{polylog}(d)}))^{-2}$ as desired.

(b) The proof is similar to that of (a).

(c) when $\alpha_1 B_{1,1}^{(t)} \le O(1)$, the proof is similar to (a). When $\alpha_1 B_{1,1}^{(t)} \ge O(1)$, we have from Induction C.3a and $H_{j,\ell}^{(t)}$'s expression that

$$H_{j,\ell}^{(t)} = C_1\alpha_\ell^6((B_{j,\ell}^{(t)})^3 + E_{j,3-j}^{(t)}(B_{3-j,\ell}^{(t)})^3)^2 + C_2\mathcal{E}_{j,3-j}^{(t)} \le \widetilde{O}(\alpha_\ell^6)$$

And since $T_1 := \min\{t : B_{1,1}^{(t)} \geq 0.01\}$, so for $t \leq T_1$, we have

$$H_{j,\ell}^{(t)} \geq C_2 \mathcal{E}_{j,3-j}^{(t)} \overset{①}{\geq} C_2 \mathcal{E}_j^{(t)} - |E_{j,3-j}^{(t)}| \overset{②}{\geq} \Omega(C_2)$$

where ① is from Claim C.5b and ② is from Induction C.3d.

(d) Since we have assumed $|B_{1,2}^{(t)}|, |B_{2,1}^{(t)}|, |B_{2,2}^{(t)}| \leq \widetilde{O}(\frac{1}{\sqrt{d}})$, it is direct to bound $|K_{j,\ell}^{(t)}| \leq \widetilde{O}(\alpha_\ell^6 / d^{1.5})$.

□

Claim C.7 (about $\Sigma_{j,\ell}^{(t)}$ and $\nabla_{w_j} \mathcal{E}_{j',3-j'}^{(t)}$). *If Induction C.3 holds at iteration $t \leq T_1$, then*

(a) $\Sigma_{j,\ell}^{(t)} = O(\Lambda_{1,1}^{(t)} B_{1,1}^{(t)}) \frac{(B_{j,\ell}^{(t)})^6 + E_{j,3-j}^{(t)} (B_{3-j,\ell}^{(t)})^3 (B_{j,\ell}^{(t)})^3}{(B_{1,1}^{(t)})^6} \frac{\Phi_j^{(t)}}{\Phi_1^{(t)}}$;

(b) $\langle \nabla_{w_j} \mathcal{E}_{j,3-j}^{(t)}, w_j^{(t)} \rangle = O(1) \pm O(E_{j,3-j}^{(t)})(R_{1,2}^{(t)} + \varrho)$;

(c) $\langle \nabla_{w_j} \mathcal{E}_{3-j,j}^{(t)}, w_j^{(t)} \rangle = O((E_{3-j,j}^{(t)})^2) \pm O(E_{3-j,j}^{(t)})(R_{1,2}^{(t)} + \varrho)$

(d) $|\langle \nabla_{w_j} \mathcal{E}_{j,3-j}^{(t)}, w_{3-j}^{(t)} \rangle| = O(R_{1,2}^{(t)} + \varrho) + O(E_{j,3-j}^{(t)})$;

(e) $|\langle \nabla_{w_j} \mathcal{E}_{3-j,j}^{(t)}, w_{3-j}^{(t)} \rangle| = O(R_{1,2}^{(t)} + \varrho)(E_{3-j,j}^{(t)})^2 + O(E_{3-j,j}^{(t)})$

Proof. Notice that $\|\Pi_{V^\perp} w_j^{(t)}\|_2 = \Theta(1), \forall j \in [2]$ for $t \leq T_1$, which is because of $\|w_j^{(t)}\|_2 = \sqrt{2} \pm o(1)$ from Induction C.3a and $\max_{j,\ell} |B_{j,\ell}^{(t)}| < 0.02$[4]. Now we can apply Claim B.1 to obtain the bounds. □

C.3 Gradient Lemmas for Phase I

We first present an interesting lemma regarding the effects of Batch-Normalization on the gradients of weights. The following lemma allow us maintain the norm of weights to above a constant throughout phase I.

Lemma C.8 (effects of BN on gradients). *For any $W = (w_1, w_2)$ and E, it holds*

(a) $\sum_{j\in[2]} \langle \nabla_{w_j} L(W, E), w_j \rangle = 0$;

Further, if Induction C.3 holds for each $t \leq T_1$, we have

(b) $|\langle \nabla_{w_j} L(W^{(t)}, E^{(t)}), w_j^{(t)} \rangle| \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) |\Lambda_{1,1}| \sum_{j\in[2]} |E_{j,3-j}^{(t)}|$ *for each $j \in [2]$.*

Proof. **Proof of (a):** We first calculate the gradient term as follows:

$$\begin{aligned}
\nabla_W L(W, E) &= \nabla_W \sum_{j\in[2]} \frac{\mathbb{E}[F_j(X^{(1)}) \cdot \mathsf{StopGrad}[G_j(X^{(2)})]]}{\sqrt{\mathbb{E}[F_j^2(X^{(1)})]} \sqrt{\mathbb{E}[\mathsf{StopGrad}[G_j^2(X^{(2)})]]}} \\
&= \sum_{j\in[2]} \frac{\mathbb{E}[(\nabla_W F_j(X^{(1)})) \cdot [G(X^{(2)})]_j] \cdot \mathbb{E}[F_j^2(X^{(1)})]}{(\mathbb{E}[F_j^2(X^{(1)})])^{3/2} \sqrt{\mathbb{E}[G_j^2(X^{(2)})]}} \\
&\quad - \sum_{j\in[2]} \frac{\mathbb{E}[(\nabla_W F_j(X^{(1)})) \cdot F_j(X^{(1)})] \cdot \mathbb{E}[[F_j(X^{(1)}) \cdot [G(X^{(2)})]_j]}{(\mathbb{E}[F_j^2(X^{(1)})])^{3/2} \sqrt{\mathbb{E}[G_j^2(X^{(2)})]}}
\end{aligned}$$

Since by our definition $\langle \nabla_W F_j(X^{(1)}), W \rangle = \sum_{i\in[2]} \langle \nabla_{w_i} [F_j(X^{(1)}), w_i \rangle = 3[F_j(X^{(1)})$, we immediately have $\sum_{j\in[2]} \langle \nabla_{w_j} L(W, E), w_j \rangle = 0$.

[4]due to our choice of $\eta = \frac{1}{\mathsf{poly}(d)}$ is small, we can make sure when $T_1 = \min\{t : B_{1,1}^{(t)} \geq 0.01\}$, $B_{1,1}^{(T_1)} < 0.02$.

Proof of (b): Firstly we define a new notion

$$\nabla_{i,j} = \nabla_{w_i} \frac{\mathbb{E}[F_j(X^{(1)}) \cdot \mathsf{StopGrad}[G_j(X^{(2)})]]}{\sqrt{\mathbb{E}[F_j^2(X^{(1)})]}\sqrt{\mathbb{E}[\mathsf{StopGrad}[G_j^2(X^{(2)})]]}}$$

Then it is straghtforward to verify that $\sum_{i\in[2]}\langle \nabla_{i,j}, w_i\rangle = 0$ for any $j \in [2]$, which implies that $|\langle \nabla_{j',j}, w_{j'}\rangle| = |\langle \nabla_{3-j',j}, w_{3-j'}\rangle|$. So in order to obtain an upper bound for $|\langle \nabla_{w_j} L(W, E), w_j\rangle| = |\sum_{j'\in[2]}\langle \nabla_{j,j'}, w_j\rangle|$, we only need to upper bound $|\langle \nabla_{j,j'}, w_{3-j'}\rangle|$, each of which can be calculated as (ignoring all time superscript $^{(t)}$)

$$\begin{aligned}|\langle \nabla_{3-j,j}, w_{3-j}\rangle| = &\frac{\mathbb{E}\left[\sum_{p\in[P]\cap\mathcal{P}} E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle)\cdot[G(X^{(2)})]_j\right]\cdot \mathbb{E}[F_j^2(X^{(1)})]}{(\mathbb{E}[F_j^2(X^{(1)})])^{3/2}\sqrt{\mathbb{E}[G_j^2(X^{(2)})]}} \\ &- \frac{\mathbb{E}\left[\sum_{p\in[P]\cap\mathcal{P}} E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle)\cdot F_j(X^{(1)})\right]\cdot \mathbb{E}[[F_j(X^{(1)})\cdot[G(X^{(2)})]_j]}{(\mathbb{E}[F_j^2(X^{(1)})])^{3/2}\sqrt{\mathbb{E}[G_j^2(X^{(2)})]}}\end{aligned}$$

Now we compute

$$\begin{aligned}\mathbb{E}\left[\sum_{p\in[P]\cap\mathcal{P}} E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle)[G(X^{(2)})]_j\right] &= \mathbb{E}\left[\sum_{p\in[P]\cap\mathcal{P}} E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle)\sum_{p\in[P]\setminus\mathcal{P}}\sigma(\langle w_j, X_p\rangle)\right] \\ &= \sum_{\ell\in[2]} E_{j,3-j}C_0\alpha_\ell^6 B_{3-j,\ell}^3 B_{j,\ell}^3\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\left[\sum_{p\in[P]\cap\mathcal{P}} E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle)\cdot F_j(X^{(1)})\right] \\ &= \mathbb{E}\left[\sum_{p\in[P]\cap\mathcal{P}} E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle)\cdot\sum_{p\in[P]\cap\mathcal{P}}(\sigma(\langle w_j, X_p\rangle) + E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle))\right] \\ &= \sum_{\ell\in[2]} E_{j,3-j}C_1\alpha_\ell^6 B_{3-j,\ell}^3(B_{j,\ell}^3 + E_{j,3-j}B_{3-j,\ell}^3) + C_2E_{j,3-j}\mathbb{E}[\langle w_j, \xi_p\rangle^3\langle w_{3-j}, \xi_p\rangle^3 + E_{j,3-j}\langle w_{3-j}, \xi_p\rangle^6]\end{aligned}$$

So we can further obtain the nominator in the expression of $|\langle\nabla_{3-j,j}, w_{3-j}\rangle|$ as

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{p\in[P]\cap\mathcal{P}} E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle)\cdot[G(X^{(2)})]_j\right]\cdot\mathbb{E}[F_j^2(X^{(1)})]\\
&-\mathbb{E}\left[\sum_{p\in[P]\cap\mathcal{P}} E_{j,3-j}\sigma(\langle w_{3-j}, X_p\rangle)\cdot F_j(X^{(1)})\right]\cdot\mathbb{E}[[F_j(X^{(1)})\cdot[G(X^{(2)})]_j]\\
=&\left(\sum_{\ell\in[2]} E_{j,3-j}C_0\alpha_\ell^6 B_{3-j,\ell}^3 B_{j,\ell}^3\right)\cdot\left(\sum_{\ell\in[2]} C_1\alpha_\ell^6(B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3)^2+C_2\mathcal{E}_{j,3-j}\right)\\
&-\left(\sum_{\ell\in[2]} E_{j,3-j}C_1\alpha_\ell^6 B_{3-j,\ell}^3(B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3)\right)\cdot\left(\sum_{\ell\in[2]} C_0\alpha_\ell^6 B_{j,\ell}^3(B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3)\right)\\
&-C_2E_{j,3-j}\mathbb{E}[\langle w_j,\xi_p\rangle^3\langle w_{3-j},\xi_p\rangle^3+E_{j,3-j}\langle w_{3-j},\xi_p\rangle^6]\cdot\left(\sum_{\ell\in[2]} C_0\alpha_\ell^6 B_{j,\ell}^3(B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3)\right)\\
=&E_{j,3-j}\sum_{\ell\in[2]} C_0\alpha_\ell^6 B_{3-j,\ell}^3(B_{j,\ell}^3H_{j,3-\ell}-B_{j,3-\ell}^3K_{j,3-\ell})\\
&-C_2E_{j,3-j}\mathbb{E}[\langle w_j,\xi_p\rangle^3\langle w_{3-j},\xi_p\rangle^3+E_{j,3-j}\langle w_{3-j},\xi_p\rangle^6]\cdot\left(\sum_{\ell\in[2]} C_0\alpha_\ell^6 B_{j,\ell}^3(B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3)\right)
\end{aligned}
$$

Now can sum over $j'\in[2]$ to get

$$
\begin{aligned}
&|\langle\nabla_{w_j}L(W,E), w_j\rangle|\\
\le&\sum_{j\in[2]}\sum_{\ell\in[2]} C_0E_{j,3-j}\left|\Phi_j\alpha_\ell^6 B_{3-j,\ell}^3 B_{j,\ell}^3 H_{j,3-\ell}\right|+\sum_{j\in[2]}\sum_{\ell\in[2]}\left|C_0E_{j,3-j}\Phi_j\alpha_\ell^3 B_{3-j,\ell}^3 B_{j,3-\ell}^3 K_{j,3-\ell}\right|\\
&+\sum_{j\in[2]}\sum_{\ell\in[2]}\left|C_2E_{j,3-j}\Phi_j\mathbb{E}[\langle w_j,\xi_p\rangle^3\langle w_{3-j},\xi_p\rangle^3+E_{j,3-j}\langle w_{3-j},\xi_p\rangle^6]C_0\alpha_\ell^6 B_{j,\ell}^3(B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3)\right|
\end{aligned}
$$

Next we are going to bound each term, for the first term of LHS we have

$$
\begin{aligned}
\sum_{j\in[2]}\sum_{\ell\in[2]}\left|C_0E_{j,3-j}\Phi_j\alpha_\ell^6 B_{3-j,\ell}^3 B_{j,\ell}^3 H_{j,3-\ell}\right| &\le \sum_{j\in[2]}\sum_{\ell\in[2]}|E_{j,3-j}||\Lambda_{j,\ell}|\left|\frac{B_{3-j,\ell}^3}{B_{j,\ell}^2}\right|\\
&\le|\Lambda_{1,1}|\sum_{j\in[2]}|E_{j,3-j}|\left|\frac{B_{3-j,\ell}^3 B_{j,\ell}^3\Phi_j}{B_{1,1}^5\Phi_1}\right|\\
&\le\widetilde{O}(\frac{d^{o(1)}}{\sqrt{d}})|\Lambda_{1,1}|\sum_{j\in[2]}|E_{j,3-j}|
\end{aligned}
$$

where the last inequality is because

- By Lemma C.6a,b, we have $\Phi_j^{(t)}/\Phi_1^{(t)}\le O(\alpha_1^O(1))\le d^{o(1)}$ during $t\le T_1$.
- $(B_{3-j,\ell}^{(t)})^3(B_{j,\ell}^{(t)})^3\le\widetilde{O}(\frac{1}{\sqrt{d}})(B_{1,1}^{(t)})^5$ from Induction C.3b,c.

Similarly, we can also compute

$$
\begin{aligned}
\sum_{j\in[2]}\sum_{\ell\in[2]}\left|C_0E_{j,3-j}\Phi_j\alpha_\ell^3 B_{3-j,\ell}^3 B_{j,3-\ell}^3 K_{j,3-\ell}\right| &\le\sum_{j\in[2]}\sum_{\ell\in[2]} E_{j,3-j}|\Lambda_{1,1}|\left|\frac{B_{3-j,\ell}^3 B_{j,3-\ell}^3 K_{j,3-\ell}}{B_{1,1}^5 H_{j,3-\ell}}\right|\\
&\le\widetilde{O}(\frac{d^{o(1)}}{d^2})|\Lambda_{1,1}|\sum_{j\in[2]}|E_{j,3-j}|
\end{aligned}
$$

and

$$\sum_{j\in[2]}\sum_{\ell\in[2]}\left|C_2 E_{j,3-j}\Phi_j\mathbb{E}[\langle w_j,\xi_p\rangle^3\langle w_{3-j},\xi_p\rangle^3+E_{j,3-j}\langle w_{3-j},\xi_p\rangle^6]C_0\alpha_\ell^6 B_{j,\ell}^3(B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3)\right|$$
$$\overset{①}{\leq}\sum_{j\in[2]}\sum_{\ell\in[2]}|E_{j,3-j}\Lambda_{j,\ell}|\left|\frac{B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3}{B_{j,\ell}^2}\right|\left|\mathbb{E}[\langle w_j,\xi_p\rangle^3\langle w_{3-j},\xi_p\rangle^3+E_{j,3-j}\langle w_{3-j},\xi_p\rangle^6]\right|$$
$$\overset{②}{\leq}\sum_{j\in[2]}\sum_{\ell\in[2]}|E_{j,3-j}\Lambda_{j,\ell}|\left|\frac{B_{j,\ell}^3+E_{j,3-j}B_{3-j,\ell}^3}{B_{j,\ell}^2}\right|(O(R_{1,2}+\varrho)+O(E_{j,3-j}))$$
$$\leq\widetilde{O}(R_{1,2}+\varrho)|\Lambda_{1,1}|\sum_{j\in[2]}|E_{j,3-j}|$$

where ① is due to Lemma C.6c, ② is from the same calculation in Claim C.7 for $\mathbb{E}[\langle w_j,\xi_p\rangle^3\langle w_{3-j},\xi_p\rangle^3]$ and Induction C.3a. Now combining the above and Induction C.3e together we have

$$|\langle\nabla_{w_j}L(W,E),w_j\rangle|\leq\widetilde{O}(\varrho+\frac{1}{\sqrt{d}})|\Lambda_{1,1}|\sum_{j\in[2]}|E_{j,3-j}|$$

which gives the desired bound. □

Next we give a lemma characterizing the gradient of feature v_1 in this phase.

Lemma C.9 (learning feature v_1 in phase I)**.** *For each $t\leq T_1$, if Induction C.3 holds at iteration t, then using notations of* (B.2)*, we have:*

(a) $\langle-\nabla_{w_1}L(W^{(t)},E^{(t)}),v_1\rangle=(1\pm\widetilde{O}(\frac{1}{d}))\Lambda_{1,1}^{(t)}$

(b) $\langle-\nabla_{w_2}L(W^{(t)},E^{(t)}),v_1\rangle=(1\pm O(\frac{1}{\sqrt{d}}))\Lambda_{2,1}^{(t)}+\Gamma_{2,1}^{(t)}\leq(1\pm O(\frac{1}{\sqrt{d}}))\Lambda_{2,1}^{(t)}\pm\frac{(B_{2,1}^{(t)})^2}{(B_{1,1}^{(t)})^2}E_{1,2}^{(t)}\Lambda_{1,1}^{(t)}$

Proof. From (B.2), we write down the gradient formula for $B_{j,1}^{(t)}$ as follows:

$$\langle-\nabla_{w_j}L_{\mathcal{D}}(W^{(t)},E^{(t)}),v_1\rangle=\Lambda_{j,1}^{(t)}+\Gamma_{j,1}^{(t)}-\Upsilon_{j,1}^{(t)}$$

where (ignoring the superscript $^{(t)}$ for the RHS)

$$\Lambda_{j,1}^{(t)}=C_0\Phi_j\alpha_1^6B_{j,1}^5H_{j,2}$$
$$\Gamma_{j,1}^{(t)}=C_0\Phi_{3-j}E_{3-j,j}\alpha_1^6B_{3-j,1}^3B_{j,1}^2H_{3-j,2}$$
$$\Upsilon_{j,1}^{(t)}=C_0\alpha_2^6\left(\Phi_jB_{j,2}^3B_{j,1}^2K_{j,1}+\Phi_{3-j}E_{3-j,j}B_{3-j,2}^3B_{j,1}^2K_{3-j,1}\right)$$

We first prove (a), and we deal with each term individually:

Comparing $\Lambda_{1,1}^{(t)}$ and $\Gamma_{1,1}^{(t)}$: When $t\leq T_{1,1}$, we have from Lemma C.6a that

$$\Phi_1^{(t)}H_{1,2}^{(t)}=\frac{1}{C_2\mathcal{E}_1^{(t)}}(1\pm\frac{1}{\mathsf{polylog}(d)})=\frac{1}{C_2\mathcal{E}_2^{(t)}}(1\pm\frac{1}{\mathsf{polylog}(d)})=\Phi_2^{(t)}H_{2,2}^{(t)}(1\pm\frac{1}{\mathsf{polylog}(d)})$$

Further, by Induction C.3b,c,d and our definition of stage 1, we know $E_{1,2}^{(t)}\leq\widetilde{O}(\frac{1}{d})$. Now from Induction C.3b that $B_{2,1}^{(t)}\leq\widetilde{O}(\frac{1}{\sqrt{d}})$, together we have

$$\Gamma_{1,1}^{(t)}=C_0\alpha_1^6E_{2,1}^{(t)}\Phi_2^{(t)}H_{2,2}^{(t)}(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^2\leq\widetilde{O}(\frac{1}{d})C_0\alpha_1^6\Phi_1^{(t)}H_{1,2}^{(t)}(B_{1,1}^{(t)})^5=\widetilde{O}(\frac{\Lambda_{1,1}^{(t)}}{d})$$

When $t\in[T_{1,1},T_1]$, by Lemma C.6b we have

$$\Phi_1^{(t)}H_{1,2}^{(t)}\geq\Omega(\frac{C_2}{(C_1\alpha_1^6(B_{1,1}^{(t)})^6+O(C_2))^2})\geq\omega(\frac{1}{d^{0.1}}),\quad\text{and}\quad E_{2,1}^{(t)}\Phi_2^{(t)}H_{2,2}^{(t)}\leq\widetilde{O}(\frac{1}{d})$$

Now from our definition of stage 2, it holds that $B_{1,1}^{(t)} \geq \Omega(\frac{1}{\alpha_1})$ while $B_{2,1}^{(t)} \leq \widetilde{O}(\frac{1}{\sqrt{d}})$ by Induction C.3b, which gives

$$\Gamma_{1,1}^{(t)} = C_0 \alpha_1^6 E_{2,1}^{(t)} \Phi_2^{(t)} H_{2,2}^{(t)} (B_{2,1}^{(t)})^3 (B_{1,1}^{(t)})^2 \leq \widetilde{O}(\frac{1}{d}) C_0 \alpha_1^6 \Phi_1^{(t)} H_{1,2}^{(t)} (B_{1,1}^{(t)})^5 = \widetilde{O}(\frac{\Lambda_{1,1}^{(t)}}{d})$$

Comparing $\Lambda_{1,1}^{(t)}$ and $\Upsilon_{1,1}^{(t)}$: Now consider $\Upsilon_{1,1}^{(t)}$, by Lemma C.6, we can follow the same analysis as above to get

$$\Phi_j^{(t)} K_{j,\ell}^{(t)} \leq \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/2}}) \Phi_1^{(t)} H_{1,2}^{(t)} \qquad \text{for any } (j,\ell) \in [2] \times [2]$$

Combined with $E_{2,1}^{(t)} \leq o(1)$, we can derive

$$\begin{aligned}
\Upsilon_{1,1}^{(t)} &= C_0 \alpha_2^6 \left(\Phi_1^{(t)} K_{1,1}^{(t)} (B_{1,2}^{(t)})^3 (B_{1,1}^{(t)})^2 + E_{1,2}^{(t)} \Phi_2^{(t)} K_{2,1}^{(t)} (B_{2,2}^{(t)})^3 (B_{1,1}^{(t)})^2 \right) \\
&\leq \widetilde{O}(\frac{\alpha_1^{O(1)} \alpha_2^6}{d^{3/2}}) C_0 \alpha_1^6 \Phi_1^{(t)} H_{1,2}^{(t)} (B_{1,1}^{(t)})^5 \\
&= \widetilde{O}(\frac{\Lambda_{1,1}^{(t)}}{d^{3/2 - o(1)}}) \qquad \text{(since } C_1 = \widetilde{O}(1) \text{ and } \alpha_1, \alpha_2 = d^{o(1)})
\end{aligned}$$

Comparing $\Lambda_{2,1}^{(t)}$ and $\Upsilon_{2,1}^{(t)}$: Till now (a) is proved, we can deal with (b) by only comparing $\Lambda_{2,1}^{(t)}$ with $\Upsilon_{2,1}^{(t)}$. Similar to the above arguments, we have by Induction C.3b we know $K_{j,1}^{(t)} = \widetilde{O}(\frac{C_1 \alpha_1^6}{d^{3/2}}), \forall j \in [2]$, and thus

$$\Phi_j^{(t)} K_{j,\ell}^{(t)} \leq \widetilde{O}(\frac{\alpha_1^6}{d^{3/2}}) \Phi_2^{(t)} H_{2,2}^{(t)} \qquad \text{for any } (j,\ell) \in [2] \times [2]$$

By Induction C.3e we know $E_{1,2}^{(t)} \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$. Also, note that from Induction C.3b we have $\widetilde{O}((B_{1,2}^{(t)})^3/d) \leq \widetilde{O}((B_{2,1}^{(t)})^5)$, and thus

$$E_{1,2}^{(t)} \Phi_1^{(t)} K_{1,1}^{(t)} (B_{1,2}^{(t)})^3 (B_{2,1}^{(t)})^2 \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \widetilde{O}(\frac{\alpha_1^6}{d^{5/2}}) \Phi_2^{(t)} H_{2,2}^{(t)} \widetilde{O}(B_{1,2}^{(t)})^3 \leq O(\frac{1}{d^{3/2}}) \Phi_2^{(t)} H_{2,2}^{(t)} (B_{2,1}^{(t)})^5$$

So together we have

$$\begin{aligned}
|\Upsilon_{2,1}^{(t)}| &= |C_0 \alpha_2^6 \left(\Phi_2^{(t)} K_{2,1}^{(t)} (B_{2,2}^{(t)})^3 (B_{2,1}^{(t)})^2 + E_{2,1}^{(t)} \Phi_1^{(t)} K_{1,1}^{(t)} (B_{1,2}^{(t)})^3 (B_{2,1}^{(t)})^2 \right)| \\
&\leq O(\frac{1}{d^{3/2}}) C_0 \alpha_1^6 \Phi_2^{(t)} H_{2,2}^{(t)} |(B_{2,1}^{(t)})^5| \\
&= O(\frac{1}{d^{3/2}}) |\Lambda_{2,1}^{(t)}|
\end{aligned}$$

Comparing $\Gamma_{2,1}^{(t)}$ with $\Lambda_{1,1}^{(t)}$: It suffices to notice that

$$|\Gamma_{2,1}^{(t)}| \leq |E_{1,2}^{(t)}| C_0 \alpha_1^6 \Phi_1^{(t)} H_{1,2}^{(t)} |B_{1,1}^{(t)}|^3 (B_{2,1}^{(t)})^2 = \frac{(B_{2,1}^{(t)})^2}{(B_{1,1}^{(t)})^2} |E_{1,2}^{(t)}| |\Lambda_{1,1}^{(t)}|$$

Combining the bounds for $\Lambda_{2,1}^{(t)}$ and $\Gamma_{2,1}^{(t)}$, we obtain the proof of (b). □

Then we can also calculate the gradients of feature v_2 in this phase.

Lemma C.10 (learning feature v_2 in phase I)**.** *For each $t \leq T_1$, if Induction C.3 holds at iteration t, then using notations of* (B.2)*, we have for each $j \in [2]$:*

$$\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_2 \rangle = \left(1 \pm \widetilde{O}(\alpha_1^6)(E_{3-j,j}^{(t)} + (B_{j,1}^{(t)})^3) \right) \Lambda_{j,2}^{(t)} \tag{C.1}$$

Proof. Again as in the proof of Lemma C.9, we expand the notations: (ignoring the superscript $^{(t)}$ for the RHS)

$$\Lambda_{j,2}^{(t)} = C_0\alpha_2^6\Phi_j H_{j,1}B_{j,2}^5$$
$$\Gamma_{j,2}^{(t)} = C_0\alpha_2^6\Phi_j E_{3-j,j}B_{3-j,2}^3 B_{j,2}^2 H_{3-j,1}$$
$$\Upsilon_{j,2}^{(t)} = C_0\alpha_1^6\left(\Phi_j B_{j,1}^3 B_{j,2}^2 K_{j,2} + \Phi_{3-j}E_{3-j,j}B_{3-j,1}^3 B_{j,2}^2 K_{3-j,2}\right)$$

We first compare $\Lambda_{j,2}^{(t)}$ and $\Gamma_{j,2}^{(t)}$ as follows: Lemma C.6 we have

- $B_{3-j,2}^{(t)} \le \widetilde{O}(B_{j,2}^{(t)})$ by Induction C.3b;
- From Lemma C.6a,b we can have $\Phi_{3-j}^{(t)} \le \widetilde{O}(\alpha_1^{O(1)})\Phi_j^{(t)}, \forall j \in [2]$.

Together they imply:

$$\begin{aligned} C_0\alpha_2^6 E_{3-j,j}^{(t)}(B_{3-j,2}^{(t)})^3(B_{j,2}^{(t)})^2\Phi_{3-j}^{(t)}H_{3-j,1}^{(t)} &\le \widetilde{O}(\alpha_1^{O(1)}E_{3-j,j}^{(t)})C_0\alpha_2^6\Phi_j^{(t)}H_{j,2}^{(t)}(B_{j,2}^{(t)})^5 \\ &= \widetilde{O}(\alpha_1^{O(1)}E_{j,3-j}^{(t)})\Lambda_{j,2}^{(t)} \end{aligned} \tag{C.2}$$

Now we turn to compare $\Lambda_{j,2}^{(t)}$ with $\Upsilon_{j,2}^{(t)}$. We split $\Upsilon_{j,2}^{(t)}$ into two terms $\Upsilon_{j,2,1}^{(t)}, \Upsilon_{j,2,2}^{(t)}$

$$\Upsilon_{j,2,1}^{(t)} = C_0\alpha_1^6\Phi_j^{(t)}(B_{j,1}^{(t)})^3(B_{j,2}^{(t)})^2K_{j,2}^{(t)}, \quad \Upsilon_{j,2,2}^{(t)} = C_0\alpha_1^6\Phi_{3-j}^{(t)}E_{3-j,j}^{(t)}(B_{3-j,1}^{(t)})^3(B_{j,2}^{(t)})^2K_{3-j,2}^{(t)}$$

For $\Upsilon_{j,2,1}^{(t)}$, we can calculate

$$\begin{aligned} \Upsilon_{j,2,1}^{(t)} &= C_0\alpha_1^6\Phi_j^{(t)}(B_{j,1}^{(t)})^3(B_{j,2}^{(t)})^2K_{j,2}^{(t)} \\ &\le \widetilde{O}(\frac{C_1\alpha_2^6}{d^{3/2}})(B_{j,1}^{(t)})^3\cdot C_0\alpha_1^6\Phi_j^{(t)}H_{j,1}^{(t)}(B_{j,2}^{(t)})^2 && (K_{j,\ell}^{(t)} \le \widetilde{O}(\tfrac{C_1\alpha_\ell^6}{d^{3/2}}) \text{ from Lemma C.6d}) \\ &\le \widetilde{O}(\alpha_1^6(B_{j,1}^{(t)})^3)C_0\alpha_2^6\Phi_j^{(t)}H_{j,1}^{(t)}(B_{j,2}^{(t)})^5 && (\widetilde{O}(\tfrac{C_1}{d^{3/2}}) \le \widetilde{O}((B_{j,2}^{(t)})^3) \text{ from Induction C.3b}) \\ &= \widetilde{O}(\alpha_1^6(B_{j,1}^{(t)})^3)\Lambda_{j,2}^{(t)} \end{aligned} \tag{C.3}$$

And for $\Upsilon_{j,2,1}^{(t)}$, we use Induction C.3b and Lemma C.6d again to get

$$(B_{3-j,1}^{(t)})^3(B_{3-j,2}^{(t)})^2K_{3-j,2}^{(t)} \le \widetilde{O}(C_1\alpha_2^6(B_{j,2}^{(t)})^5)$$

and thus combined with $\Phi_{3-j}^{(t)} \le \widetilde{O}(\alpha_1^6)\Phi_j^{(t)}, \forall j \in [2]$ from Lemma C.6a,b, we can derive

$$\begin{aligned} \Upsilon_{j,2,2}^{(t)} &= C_0\alpha_1^6\Phi_{3-j}^{(t)}E_{j,3-j}^{(t)}(B_{3-j,1}^{(t)})^3(B_{j,2}^{(t)})^2K_{3-j,2}^{(t)} \\ &\le \widetilde{O}(\alpha_1^6E_{3-j,j}^{(t)})C_0\alpha_2^6\Phi_j^{(t)}H_{j,1}^{(t)}(B_{j,2}^{(t)})^5 \\ &= \widetilde{O}(\alpha_1^6E_{3-j,j}^{(t)})\Lambda_{j,2}^{(t)} \end{aligned} \tag{C.4}$$

Now combine the results of (C.2), (C.3) and (C.4) finishes the proof of (C.1). □

Lemma C.11 (learning prediction head $E_{1,2}, E_{2,1}$ in phase I)**.** *If Induction C.3 holds at iteration $t \le T_1$, then we have*

(a) $-\nabla_{E_{1,2}}L(W^{(t)},E^{(t)}) = O(\Lambda_{1,1}^{(t)}B_{1,1}^{(t)})\left(-O(E_{1,2}^{(t)}) + \widetilde{O}(\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^3}) + O(R_{1,2}^{(t)})\right)$*;*

(b) $-\nabla_{E_{2,1}}L(W^{(t)},E^{(t)}) = \widetilde{O}(\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^2})\Lambda_{1,1}^{(t)} + \sum_{\ell\in[2]}C_2\Lambda_{2,\ell}^{(t)}B_{2,\ell}^{(t)}\left(-O(E_{2,1}^{(t)}) + O(R_{1,2}^{(t)})\right)$

Proof. We first write down the gradient for $E_{j,3-j}^{(t)}$: (ignoring the time superscript $^{(t)}$)

$$-\nabla_{E_{j,3-j}}L(W,E) = \sum_{\ell\in[2]}C_0\Phi_j\alpha_\ell^6B_{j,\ell}^3(B_{3-j,\ell}^3H_{j,3-\ell} - B_{3-j,3-\ell}^3K_{j,3-\ell}) - \sum_{\ell\in[2]}\Sigma_{j,\ell}\nabla_{E_{j,3-j}}\mathcal{E}_{j,3-j}$$

where $\nabla_{E_{j,3-j}}\mathcal{E}_{j,3-j} = \mathbb{E}\left[2\langle w_j, \xi_p\rangle^3\langle w_{3-j}, \xi_p\rangle^3 + 2E_{j,3-j}\langle w_{3-j}, \xi_p\rangle^6\right]$. Thus we have

$$\nabla_{E_{j,3-j}}\mathcal{E}^{(t)}_{j,3-j} = O(1)E^{(t)}_{j,3-j} + O(R^{(t)}_{1,2})$$

and by Claim C.5 and Lemma C.6a,b

$$\Sigma^{(t)}_{j,\ell} = O(\Lambda^{(t)}_{1,1}B^{(t)}_{1,1})\frac{(B^{(t)}_{j,\ell})^6 + E^{(t)}_{j,3-j}(B^{(t)}_{3-j,\ell})^3(B^{(t)}_{j,\ell})^3}{(B^{(t)}_{1,1})^6}\frac{\Phi^{(t)}_j}{\Phi^{(t)}_1} \le O(\Lambda^{(t)}_{1,1}B^{(t)}_{1,1})$$

Now let us look at $\nabla_{E_{1,2}}L(W^{(t)}, E^{(t)})$, first we consider the term

$$\sum_{\ell\in[2]} C_0\Phi^{(t)}_1\alpha^6_\ell(B^{(t)}_{1,\ell})^3((B^{(t)}_{2,\ell})^3H^{(t)}_{1,3-\ell} - (B^{(t)}_{2,3-\ell})^3K^{(t)}_{1,3-\ell})$$

Using Lemma C.6 and Induction C.3b,c, we know

- $H^{(t)}_{1,1} \le \widetilde{O}(H^{(t)}_{1,2})$ at $t \le T_{1,1}$ and $H^{(t)}_{1,1} \le \widetilde{O}(\alpha^6_1 H^{(t)}_{1,2})$ for $t \in [T_{1,1}, T_1]$;
- $B^{(t)}_{2,1}, B^{(t)}_{1,2}, B^{(t)}_{2,2} \le \widetilde{O}(B^{(t)}_{2,1}) \le \widetilde{O}(B^{(t)}_{1,1})$;
- $K^{(t)}_{1,3-\ell} \le \widetilde{O}(\alpha^6_1/d^{3/2})$.

It can be computed that

$$C_0\Phi^{(t)}_1\alpha^6_2(B^{(t)}_{1,2})^3(B^{(t)}_{2,2})^3H^{(t)}_{1,1} \le \widetilde{O}(1)\left(\frac{B^{(t)}_{2,1}}{B^{(t)}_{1,1}}\right)^3 C_0\Phi^{(t)}_1\alpha^3_1(B^{(t)}_{1,1})^6H^{(t)}_{1,2}$$

$$\sum_{\ell\in[2]}\left|C_0\Phi^{(t)}_1\alpha^6_\ell(B^{(t)}_{1,\ell})^3(B^{(t)}_{2,\ell})^3K^{(t)}_{1,3-\ell}\right| \le \widetilde{O}(\frac{\alpha^6_1}{d^{3/2}})\frac{(B^{(t)}_{2,1})^3}{(B^{(t)}_{1,1})^3}C_0\Phi^{(t)}_1\alpha^6_1(B^{(t)}_{1,1})^6H^{(t)}_{1,2}$$

Now we turn to $\nabla_{E_{2,1}}L(W^{(t)}, E^{(t)})$, similarly we have

$$C_0\Phi^{(t)}_2\alpha^6_1(B^{(t)}_{2,1})^3(B^{(t)}_{1,1})^3H^{(t)}_{2,2} \le \widetilde{O}(1)\left(\frac{B^{(t)}_{2,1}}{B^{(t)}_{1,1}}\right)^3 C_0\Phi^{(t)}_1\alpha^6_1(B^{(t)}_{1,1})^6H^{(t)}_{1,2}$$

and since $H^{(t)}_{2,1} \le O(C_2) = O(H^{(t)}_{1,2})$ by Lemma C.6c, we can go through the same arguments again to obtain

$$\left|C_0\Phi^{(t)}_2\alpha^6_2(B^{(t)}_{1,2})^3(B^{(t)}_{2,2})^3H^{(t)}_{2,1}\right| \le \widetilde{O}(1)\left(\frac{B^{(t)}_{1,2}}{B^{(t)}_{1,1}}\right)^3 C_0\Phi^{(t)}_1\alpha^6_1(B^{(t)}_{1,1})^6H^{(t)}_{1,2}$$

$$\left|C_0\Phi^{(t)}_2\alpha^6_2(B^{(t)}_{1,2})^3(B^{(t)}_{2,1})^3K^{(t)}_{2,1}\right| \le \widetilde{O}(\frac{\alpha^6_1}{d^{3/2}})\left(\frac{B^{(t)}_{1,2}}{B^{(t)}_{1,1}}\right)^3 C_0\Phi^{(t)}_1\alpha^6_1(B^{(t)}_{1,1})^6H^{(t)}_{1,2}$$

Now the proof is complete. □

Also, we will need the following lemma controlling gradient bounds for the noise term.

Lemma C.12 (update of $R^{(t)}_{1,2}$ in phase I)**.** *Suppose Induction C.3 holds at iteration $t \le T_1$, then we have*

(a) $|\langle -\nabla_{w_1}L(W^{(t)}, E^{(t)}), \Pi_{V^\perp}w^{(t)}_2\rangle| \le \widetilde{O}(\frac{1}{\sqrt{d}} + \varrho)\Lambda^{(t)}_{1,1}B^{(t)}_{1,1}$

(b) $|\langle -\nabla_{w_2}L(W^{(t)}, E^{(t)}), \Pi_{V^\perp}w^{(t)}_1\rangle| \le \widetilde{O}(\frac{1}{\sqrt{d}} + \varrho)\Lambda^{(t)}_{1,1}B^{(t)}_{1,1}$

Proof. **Proof of (a):** Firstly, by Claim C.7a, we can directly write

$$\langle \nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} \rangle = -\sum_{j,\ell} \Sigma_{j,\ell}^{(t)} \langle \nabla_{w_1} \mathcal{E}_{j,3-j}^{(t)}, w_2^{(t)} \rangle$$

$$= -\Lambda_{1,1}^{(t)} B_{1,1}^{(t)} \sum_{(j,\ell)\in[2]^2} \frac{(B_{j,\ell}^{(t)})^6 + E_{j,3-j}^{(t)} (B_{3-j,\ell}^{(t)})^3 (B_{j,\ell}^{(t)})^3}{(B_{1,1}^{(t)})^6} \frac{\Phi_j^{(t)}}{\Phi_1^{(t)}} \langle \nabla_{w_1} \mathcal{E}_{j,3-j}^{(t)}, w_1^{(t)} \rangle \quad \text{(C.5)}$$

Now we discuss each summand respectively: for $(j,\ell) = (1,1)$, we have

$$\frac{(B_{j,\ell}^{(t)})^6 + E_{j,3-j}^{(t)} (B_{3-j,\ell}^{(t)})^3 (B_{j,\ell}^{(t)})^3}{(B_{1,1}^{(t)})^6} = 1 + E_{1,2}^{(t)} \frac{(B_{2,1}^{(t)})^3}{(B_{1,1}^{(t)})^3} = 1 + o(\frac{1}{d^{3/2}(B_{1,1}^{(t)})^3}) \quad \text{(C.6)}$$

where the last one is due to Induction C.3d. And for $\ell = 2$, we can see from Induction C.3b and d, that $\max_{(j,\ell)\neq(1,1)} |B_{j,\ell}^{(t)}| = \widetilde{O}(\frac{1}{\sqrt{d}})$ and $E_{j,3-j}^{(t)} \leq o(1)$ to give

$$\frac{(B_{j,2}^{(t)})^6 + E_{j,3-j}^{(t)} (B_{3-j,2}^{(t)})^3 (B_{j,2}^{(t)})^3}{(B_{1,1}^{(t)})^6} \frac{\Phi_j^{(t)}}{\Phi_1^{(t)}} \leq \widetilde{O}(\frac{1}{d^3}) \frac{1}{(B_{1,1}^{(t)})^6} \frac{\Phi_j^{(t)}}{\Phi_1^{(t)}}$$

On one hand, when $t \leq T_{1,1}$, we have $\alpha_\ell B_{j,\ell}^{(t)} \leq O(1)$ for all $(j,\ell) \in [2]^2$, so Lemma C.6a applies for both $\Phi_j^{(t)}$ and results in $\Phi_2^{(t)}/\Phi_1^{(t)} \leq O(1)$. We can also apply Induction C.3c to have $B_{j,2}^{(t)}/B_{1,1}^{(t)} \leq \widetilde{O}(1)$. On the other hand, when $t \in [T_{1,1}, T_1]$, we have by Induction C.3b and Lemma C.6a,b that $\Phi_2^{(t)}/\Phi_1^{(t)} \leq \widetilde{O}(\alpha_1^{O(1)}) = d^{o(1)}$, but now $B_{1,1}^{(t)} = d^{-o(1)} \gg \widetilde{O}(d^{-1/2})$, therefore

$$\widetilde{O}(\frac{1}{d^3}) \frac{1}{(B_{1,1}^{(t)})^6} \frac{\Phi_2^{(t)}}{\Phi_1^{(t)}} \leq \widetilde{O}(\frac{1}{d^{3/2}}) \frac{1}{(B_{1,1}^{(t)})^3}$$

So together, they imply

$$\frac{(B_{j,2}^{(t)})^6 + E_{j,3-j}^{(t)} (B_{3-j,2}^{(t)})^3 (B_{j,2}^{(t)})^3}{(B_{1,1}^{(t)})^6} \frac{\Phi_j^{(t)}}{\Phi_1^{(t)}} \leq \widetilde{O}(\frac{1}{d^{3/2}(B_{1,1}^{(t)})^3}) \quad \text{(C.7)}$$

and similarly, we have

$$\frac{(B_{2,1}^{(t)})^6 + E_{2,1}^{(t)} (B_{1,1}^{(t)})^3 (B_{2,1}^{(t)})^3}{(B_{1,1}^{(t)})^6} \frac{\Phi_2^{(t)}}{\Phi_1^{(t)}} \leq \widetilde{O}(\frac{1}{d^{3/2}(B_{1,1}^{(t)})^3}) \quad \text{(C.8)}$$

Next we turn to $\langle \nabla_{w_1} \mathcal{E}_{j,3-j}^{(t)}, w_2^{(t)} \rangle$. When $j = 1$, we can apply Claim C.7d to get

$$\langle \nabla_{w_1} \mathcal{E}_{1,2}^{(t)}, w_2^{(t)} \rangle = O(R_{1,2}^{(t)} + \varrho) + O(E_{1,2}^{(t)}) = O(\varrho + \frac{1}{\sqrt{d}}) + O(E_{1,2}^{(t)}) \leq O(\varrho + \frac{1}{\sqrt{d}}) \quad \text{(C.9)}$$

and when $j = 2$, we can apply Claim C.7e to get

$$\langle \nabla_{w_1} \mathcal{E}_{2,1}^{(t)}, w_2^{(t)} \rangle = -(E_{2,1}^{(t)})^2 O(R_{1,2}^{(t)} + \varrho) + O(E_{2,1}) = \widetilde{O}(\frac{1}{d^2})(\varrho + \frac{1}{\sqrt{d}}) + O(\frac{1}{d}) \quad \text{(C.10)}$$

Combining (C.5), (C.6), (C.7), (C.8), (C.9), and (C.10) completes the proof of (a).

Proof of (b): The $\Sigma_{j,\ell}^{(t)}$ part is the same as in the proof of (a), so we only deal with $\langle \nabla_{w_2} \mathcal{E}_{1,2}^{(t)}, w_1^{(t)} \rangle$ and $\langle \nabla_{w_2} \mathcal{E}_{2,1}^{(t)}, w_1^{(t)} \rangle$ here. For $\langle \nabla_{w_2} \mathcal{E}_{2,1}^{(t)}, w_1^{(t)} \rangle$, we apply Claim C.7d to get

$$\langle \nabla_{w_2} \mathcal{E}_{2,1}^{(t)}, w_1^{(t)} \rangle = O(R_{1,2}^{(t)} + \varrho) + O(1) E_{1,2}^{(t)} \quad \text{(C.11)}$$

and for $\langle \nabla_{w_2} \mathcal{E}_{1,2}^{(t)}, w_1^{(t)} \rangle$, we have

$$\langle \nabla_{w_2} \mathcal{E}_{1,2}^{(t)}, w_1^{(t)} \rangle = O(R_{1,2}^{(t)} + \varrho)(E_{2,1}^{(t)})^2 + O(1) E_{2,1}^{(t)} \quad \text{(C.12)}$$

Inserting (C.6), (C.7), (C.8) and (C.11), (C.12) into the expression of $\langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)} \rangle$ finishes the proof of (b). □

C.4 At the End of Phase I

Lemma C.13 (Phase I). *Suppose $\eta \le \frac{1}{\mathsf{poly}(d)}$ is sufficiently small, then Induction C.3 holds for at least all $t \le T_1 = O(\frac{d^2}{\eta})$, and at iteration $t = T_1$, we have*

(a) $B_{1,1}^{(T_1)} = \Omega(1)$*;*

(b) $\|w_j^{(T_1)}\|_2 = 1 \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$*;*

(c) $B_{2,1}^{(T_1)} = \widetilde{\Theta}(\frac{1}{\sqrt{d}})$ *and* $B_{j,2}^{(T_1)} = B_{j,2}^{(0)}(1 \pm o(1))$ *for* $j \in [2]$*;*

(d) $E_{2,1}^{(T_1)} = \widetilde{O}(\frac{\eta_E/\eta}{d})$ *and* $E_{1,2}^{(T_1)} \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$*;*

(e) $R_{1,1}^{(T_1)}, R_2^{(T_1)} = \Theta(1)$ *and* $R_{1,2}^{(T_1)} = \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$.

Proof. We begin by first prove the existence of $T_1 := \min\{t : B_{1,1}^{(t)} \ge 0.01\} = O(\frac{d^2}{\eta})$ if Induction C.3 holds whenever $B_{1,1}^{(t)} \le 0.01$, then we will turn back to prove Induction C.3 holds throughout $t \le T_1$. We split the analysis into two stages:

Proof of $T_1 \le O(\frac{d^2}{\eta})$: By Lemma C.9a we can write down the update of $B_{1,1}^{(t)}$ as

$$B_{1,1}^{(t+1)} = B_{1,1}^{(t)} + \eta(1 \pm \widetilde{O}(\frac{1}{d}))\Lambda_{1,1}^{(t)} = B_{1,1}^{(t)} + \eta(1 \pm \widetilde{O}(\frac{1}{d}))\Phi_1^{(t)} C_0 \alpha_1^6 H_{1,2}^{(t)} (B_{1,1}^{(t)})^5 \tag{C.13}$$

When $\alpha_1 B_{1,1}^{(t)} \le O(1)$, by Lemma C.6a,c we have $\Phi_1^{(t)} = \Theta(\frac{1}{C_2^2})$ and $H_{1,2}^{(t)} = \Omega(C_2)$, this means we can lower bound the update as

$$B_{1,1}^{(t+1)} \ge B_{1,1}^{(t)} + \Omega(\frac{\eta C_0 \alpha_1^6}{C_2})(B_{1,1}^{(t)})^5$$

since $\frac{C_0\alpha_1^6}{C_2}$ is a constant, we know there exist some $t' \ge 0$ such that $B_{1,1}^{(t')} \ge \Omega(\frac{1}{\alpha_1})$. Also recall that $T_{1,1} := \min\{t : B_{1,1}^{(t)} \ge \Omega(\frac{1}{\alpha_1})\}$. So by Lemma H.1, where $\eta = \frac{1}{\mathsf{poly}(d)}, C_t = \Omega(\frac{C_0\alpha_1^6}{C_2})$ $\delta = \frac{1}{\mathsf{polylog}(d)}$ and $A = \Omega(\frac{1}{\alpha_1}), \log(A/B_{1,1}^{(0)}) = \widetilde{O}(1)$, we have

$$T_{1,1} = O(\frac{C_2}{\eta C_0 \alpha_1^6}) \sum_{x_t \le O(\frac{1}{\alpha_1})} \eta C_t \le O(\frac{C_2}{\eta C_0 \alpha_1^6}) \left(O(1) + \frac{\widetilde{O}(\eta)}{B_{1,1}^{(0)}} \right) \frac{1}{(B_{1,1}^{(0)})^4} \le \widetilde{O}(\frac{1}{\eta \alpha_1^6 (B_{1,1}^{(0)})^4})$$

Since $(B_{1,1}^{(0)})^4 \ge \widetilde{\Omega}(\frac{1}{d^2})$ from our initialization, we have $T_{1,1} \le O(\frac{d^2}{\eta})$ and thus $T_{1,1}$ exists. Now we consider when $B_{1,1}^{(t)} \ge \Omega(\frac{1}{\alpha_1})$. Now by Lemma C.6b,c, we have $\Phi_1^{(t)} \ge \Omega((C_2 + \alpha_1^6)^{-2})$, which gives an update:

$$B_{1,1}^{(t+1)} \ge B_{1,1}^{(t)} + \Omega(\frac{\eta C_0 \alpha_1^6}{(C_2 + \alpha_1^6)^2})(B_{1,1}^{(t)})^5$$

so again by Lemma H.1, choosing $C_t = \Omega(\frac{C_0\alpha_1^6}{(C_2+\alpha_1^6)^2})$,

$$T_1 = \frac{O((C_2 + \alpha_1^6)^2)}{\eta C_0 \alpha_1^6} \sum_{x_t \in [\Omega(\frac{1}{\alpha_1}), 0.01]} \eta C_t \le \left(O(1) + \frac{\widetilde{O}(\eta)}{B_{1,1}^{(T_{1,1})}} \right) \frac{\widetilde{O}(\alpha_1^{12})}{(B_{1,1}^{(T_{1,1})})^4} \le \widetilde{O}(\frac{\alpha_1^6}{\eta (B_{1,1}^{(T_{1,1})})^4}) \le O(\frac{\alpha_1^6}{\eta})$$

where $O(\frac{\alpha_1^6}{\eta}) \ll O(\frac{d^2}{\eta})$, so we have proved that T_1 exist. Now we begin to prove that Induction C.3 holds for all $t \le T_1$.

Proof of Induction C.3: We first prove (b)–(d), and then come back to prove (a) and (d). At $t = 0$, we know all induction holds from Properties C.1. Now we suppose Induction C.3 holds for all iterations $\leq t-1$ and prove it holds at t.

The growth of $B_{2,1}^{(t)}$: Applying Lemma C.9, we have for $t \leq T_{1,1}$

$$B_{1,1}^{(t+1)} \geq B_{1,1}^{(t)} + \eta(1-\widetilde{O}(\frac{1}{d}))\Lambda_{1,1}^{(t)}$$

$$B_{2,1}^{(t+1)} \leq B_{2,1}^{(t)} + \eta(1+O(\frac{1}{\sqrt{d}}))\Lambda_{2,1}^{(t)} + \eta\frac{(B_{2,1}^{(t)})^2}{(B_{1,1}^{(t)})^2}E_{1,2}^{(t)}\Lambda_{1,1}^{(t)}$$

For some $t_1' := \min\{t : B_{1,1}^{(t)} \geq \frac{\Omega(1)}{d^{0.49}}\}$, we have $E_{1,2}^{(t)} \leq \widetilde{O}(B_{1,1}^{(t)}\varrho) \lesssim \frac{1}{d^{0.49}}$ during $t \leq t_1'$, and

$$\frac{(B_{2,1}^{(t)})^2}{(B_{1,1}^{(t)})^2}E_{1,2}^{(t)}\Lambda_{1,1}^{(t)} \lesssim \frac{(B_{2,1}^{(t)})^2}{d^{0.49}(B_{1,1}^{(t)})^2}\Lambda_{1,1}^{(t)} \leq \widetilde{O}(\frac{1}{d^{0.49}})\Lambda_{2,1}^{(t)}$$

which allow us to give an upper bound to $B_{2,1}^{(t+1)}$ as

$$\begin{aligned} B_{2,1}^{(t+1)} &\leq (1+O(\frac{1}{\sqrt{d}}))\Lambda_{2,1}^{(t)} + \widetilde{O}(\frac{1}{d^{0.49}})\Lambda_{2,1}^{(t)} \\ &\leq (1+\widetilde{O}(\frac{1}{d^{0.49}}))\Phi_2^{(t)}C_0\alpha_1^6C_2\mathcal{E}_2^{(t)}(1+\frac{1}{\mathsf{polylog}(d)})(B_{2,1}^{(t)})^5 \qquad \text{(when } t \leq t_1'\text{)} \end{aligned}$$

Since we also have

$$B_{1,1}^{(t+1)} \geq (1-\widetilde{O}(\frac{1}{d}))\Lambda_{1,1}^{(t)} \geq (1-\widetilde{O}(\frac{1}{d}))\Phi_1^{(t)}C_0\alpha_1^6\mathcal{E}_1^{(t)}(1-\frac{1}{\mathsf{polylog}(d)})(B_{1,1}^{(t)})^5$$

Since $B_{1,1}^{(0)} \geq B_{2,1}^{(0)}(1+\Omega(\frac{1}{\log d}))$, we can now apply Corollary H.2 to the two sequence $B_{1,1}^{(t+1)}$ and $B_{2,1}^{(t+1)}$, where $S_t = \frac{\Phi_1^{(t)}\mathcal{E}_1^{(t)}}{\Phi_2^{(t)}\mathcal{E}_2^{(t)}}(1+\frac{1}{\mathsf{polylog}(d)})$ to get

$$B_{1,1}^{(t_1')} \geq \frac{1}{d^{0.499}} \quad \text{while} \quad B_{2,1}^{(t_1')} \leq \widetilde{O}(\frac{1}{\sqrt{d}})$$

Note that here the update of $B_{2,1}^{(t)}$ at every step satisfies $\operatorname{sign}(B_{2,1}^{(t+1)} - B_{2,1}^{(t)}) = \operatorname{sign}(B_{2,1}^{(t)})$ which implies $B_{2,1}^{(t_1')} = \widetilde{\Theta}(\frac{1}{\sqrt{d}})$. Now for every $T \in [t_1', T_1]$, we can apply Lemma H.3 to get that

$$\sum_{t\in[t_1',T]}\eta\frac{(B_{2,1}^{(t)})^2}{(B_{1,1}^{(t)})^2}E_{1,2}^{(t)}\Lambda_{1,1}^{(t)} \leq \widetilde{O}(\varrho+\frac{1}{\sqrt{d}})O(\frac{1}{B_{1,1}^{(t_1')}})\max_{t\leq T}\{(B_{2,1}^{(t)})^2\} \leq O(\frac{1}{d^{0.5+\Omega(1)}})$$

Suppose we have proved that $B_{2,1}^{(t)} \leq \widetilde{O}(\frac{1}{\sqrt{d}})$ for each $t \leq T$, we define a new sequence

$$\widetilde{B}_{2,1}^{(t+1)} = \widetilde{B}_{2,1}^{(t)} + \eta(1+\widetilde{O}(\frac{1}{d^{0.49}}))\Phi_2^{(t)}C_0\alpha_1^6C_2\mathcal{E}_2^{(t)}(1+\frac{1}{\mathsf{polylog}(d)})(\widetilde{B}_{2,1}^{(t)})^5,$$

$$\text{where } \widetilde{B}_{2,1}^{(t_1')} = B_{2,1}^{(t_1')} + \sum_{t\in[t_1',T]}\eta\frac{(B_{2,1}^{(t)})^2}{(B_{1,1}^{(t)})^2}E_{1,2}^{(t)}\Lambda_{1,1}^{(t)} = (1\pm o(1))\widetilde{B}_{2,1}^{(t_1')}$$

It can be directly seen that $|\widetilde{B}_{2,1}^{(t)} - \widetilde{B}_{2,1}^{(0)}| \geq |B_{2,1}^{(t)} - B_{2,1}^{(0)}|$ for all $t \in [t_1', T]$. Notice that now $\widetilde{B}_{2,1}^{(t_1')} \leq d^{\Omega(1)}B_{1,1}^{(t_1')}$, we can now apply Corollary H.2 again to get

$$|B_{2,1}^{(T)} - B_{2,1}^{(0)}| \leq |\widetilde{B}_{2,1}^{(T)} - \widetilde{B}_{2,1}^{(0)}| \leq \frac{1}{\sqrt{d}\mathsf{polylog}(d)} \qquad \text{(for every } T \leq T_{1,1}\text{)}$$

Now we deal with $t \in [T_{1,1}, T_1]$. During this stage, we can directly apply Corollary H.2 to $\widetilde{B}_{2,1}^{(t)}$ and $B_{1,1}^{(t)}$, where $S_t = \frac{\Phi_1^{(t)}H_{1,2}^{(t)}}{\Phi_2^{(t)}H_{2,2}^{(t)}} \leq O(\alpha_1^{O(1)})$, to get that

$$|B_{2,1}^{(T)} - B_{2,1}^{(0)}| \leq |\widetilde{B}_{2,1}^{(T)} - \widetilde{B}_{2,1}^{(0)}| \leq \frac{1}{\sqrt{d}\mathsf{polylog}(d)} \qquad \text{(for every } T \leq T_1\text{)}$$

And thus by Lemma C.1, we have $B_{2,1}^{(T)} = B_{2,1}^{(0)}(1 \pm o(1))$.

The growth of $B_{1,2}^{(t)}$ and $B_{2,2}^{(t)}$: By Lemma C.10, we can write down the update as

$$B_{j,2}^{(t+1)} = B_{j,2}^{(t)} + \eta\left(1 \pm \widetilde{O}(\alpha_1^6)(E_{3-j,j}^{(t)} + (B_{j,1}^{(t)})^3)\right)\Lambda_{j,2}^{(t)}$$

Since $B_{2,1}^{(t)} \leq \widetilde{O}(\frac{1}{\sqrt{d}})$ and $E_{1,2}^{(t)} \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})B_{1,1}^{(t)}, E_{2,1}^{(t)} \leq \widetilde{O}(\frac{1}{d})$ because we chose $\eta_E \leq \eta$, we only need to care about $(B_{1,1}^{(t)})^3$ in the update expression. Now define $t_2' := \min\{t : B_{1,1}^{(t)} \geq \Omega(\frac{1}{\alpha_1^2})\}$, we have

- For $t \leq t_2'$, by Corollary H.2 and setting $x_t = B_{1,1}^{(t)}$, $C_t = (1 - \widetilde{O}(\frac{1}{d}))\Phi_1^{(t)} C_0 \alpha_1^6 H_{1,2}^{(t)}$, $S_t = O(\frac{\alpha_2^6 \Phi_j^{(t)} H_{j,1}^{(t)}}{\alpha_1^6 \Phi_1^{(t)} H_{1,2}^{(t)}}) \leq \widetilde{O}(\frac{\alpha_2^6}{\alpha_1^6}) \ll \frac{1}{\mathsf{polylog}(d)}$ (by Lemma C.6a,c), we have $|B_{j,2}^{(t)} - B_{j,2}^{(0)}| \leq O(\frac{\alpha_2^6}{\alpha_1^6}\frac{1}{\sqrt{d}}) \lesssim \frac{1}{\sqrt{d}\mathsf{polylog}(d)}$ for all $t \leq t_2'$, which implies $B_{j,2}^{(t_2')} = B_{j,2}^{(0)} \pm \frac{1}{\sqrt{d}\mathsf{polylog}(d)} \in [\Omega(\frac{1}{\sqrt{d}\log d}), O(\frac{\sqrt{\log d}}{\sqrt{d}})]$ by Lemma C.1.
- For $t \in [t_2', T_1]$, we can use Corollary H.2 again and let $x_t = B_{1,1}^{(t)}$, we know $B_{1,1}^{(t_2')} \geq d^{\Omega(1)} B_{2,1}^{(t_2')}$. Setting $C_t = (1 - \widetilde{O}(\frac{1}{d}))\Phi_1^{(t)} C_0 \alpha_1^6 H_{1,2}^{(t)}$, $S_t = O((1 + \alpha_1^6)\frac{\alpha_2^6 \Phi_j^{(t)} H_{j,1}^{(t)}}{\alpha_1^6 \Phi_1^{(t)} H_{1,2}^{(t)}}) \leq O(\alpha^{O(1)})$, we can have $|B_{j,2}^{(t)} - B_{j,2}^{(t_2')}| \lesssim \frac{1}{\sqrt{d}\mathsf{polylog}(d)}$, which implies $B_{j,2}^{(t)} \in [\Omega(\frac{1}{\sqrt{d}\log d}), O(\frac{\sqrt{\log d}}{\sqrt{d}})]$ for all $t \in [t_2', T_1]$.

This proves Induction C.3b. Indeed, simple calculations also proves Induction C.3c, since the update of $B_{1,1}^{(t)}$ is always larger than others' during $t \leq T_1$.

For Induction C.3d: From Lemma C.11, we can write down the update

$$-\nabla_{E_{1,2}} L(W^{(t)}, E^{(t)}) = O(\Lambda_{1,1}^{(t)} B_{1,1}^{(t)})\left(-C_1 E_{1,2}^{(t)} + \widetilde{O}(\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^3}) + C_2(R_{1,2}^{(t)} + \varrho)\right)$$

for some constants $C_1, C_2 = \Theta(1)$. Applying Lemma H.3 to $O(\Lambda_{1,1}^{(t)} B_{1,1}^{(t)})\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^3}$, we can obtain

$$\sum_{t \leq T} O(\eta_E \Lambda_{1,1}^{(t)} B_{1,1}^{(t)})\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^3} = \frac{\eta_E}{\eta}\sum_{t \leq T} O(\eta \Lambda_{1,1}^{(t)})\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^2} \leq \widetilde{O}(\frac{\eta_E/\eta}{d^{3/2}})\frac{1}{B_{1,1}^{(0)}} \leq \widetilde{O}(\frac{\eta_E/\eta}{d})$$

So here it suffices to notice that whenever $|E_{1,2}^{(t)}| < 2\frac{C_2}{C_1}(R_{1,2}^{(t)} + \varrho)$ (which is obviously satisified at $t = 0$), we would have

$$O(\Lambda_{1,1}^{(t)} B_{1,1}^{(t)})\left(-O(E_{1,2}^{(t)}) + C_2(R_{1,2}^{(t)} + \varrho)\right) = -O(\Lambda_{1,1}^{(t)} B_{1,1}^{(t)})\widetilde{O}(R_{1,2}^{(t)} + \varrho) \leq O(\Lambda_{1,1}^{(t)} B_{1,1}^{(t)})\widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$$

In that case, we will always have (since $E_{1,2}^{(0)} = 0$)

$$E_{1,2}^{(t+1)} \leq \left|\sum_{t \leq T} \widetilde{O}(\eta_E \Lambda_{1,1}^{(t)} B_{1,1}^{(t)})\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^3}\right| + \sum_{s \leq t} O(\eta_E \Lambda_{1,1}^{(s)} B_{1,1}^{(s)})(R_{1,2}^{(s)} + \varrho) \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})\frac{\eta_E}{\eta} B_{1,1}^{(t+1)}$$

Similarly for $\nabla_{E_{2,1}} L(W^{(t)}, E^{(t)})$, we can write down

$$-\nabla_{E_{2,1}} L(W^{(t)}, E^{(t)}) = \widetilde{O}(\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^2})\Lambda_{1,1}^{(t)} + \sum_{\ell \in [2]} C_2 \Lambda_{2,\ell}^{(t)} B_{2,\ell}^{(t)}\left(-O(E_{2,1}^{(t)} R_2^{(t)}) + O(R_{1,2}^{(t)})\right)$$

by Lemma H.3, we have

$$\sum_{t \le T_1} \eta_E \widetilde{O}(\frac{(B_{1,2}^{(t)})^3}{(B_{1,1}^{(t)})^2}) \Lambda_{1,1}^{(t)} \le \widetilde{O}(\frac{\eta_E/\eta}{d})$$

and since from previous comparison results we know that

$$\sum_{t \le T_1} \sum_{\ell \in [2]} \eta_E C_2 \Lambda_{2,\ell}^{(t)} B_{2,\ell}^{(t)} = \frac{\eta_E}{\eta} \sum_{t \le T_1} \sum_{\ell \in [2]} \eta C_2 \Lambda_{2,\ell}^{(t)} B_{2,\ell}^{(t)} \le \widetilde{O}(\frac{\eta_E/\eta}{d})$$

we can then prove the claim.

For Induction C.3a: We can write down the update of $\|w_j^{(t)}\|_2^2$ as follows:

$$\begin{aligned}
\|w_j^{(t+1)}\|_2^2 &= \|w_j^{(t)} - \eta \nabla_{w_j} L(W^{(t)}, E^{(t)})\|_2^2 \\
&= \|w_j^{(t)}\|_2^2 - \eta \langle \nabla_{w_j} L(W^{(t)}, E^{(t)}), w_j^{(t)} \rangle + \eta^2 \|\nabla_{w_j} L(W^{(t)}, E^{(t)})\|_2^2
\end{aligned}$$

from (B.2) and Induction C.3a,b,c at iteration t and our assumption on ξ_p, we know

$$\|\nabla_{w_j} L(W^{(t)}, E^{(t)})\|_2^2 \le \widetilde{O}(d)$$

which allow us to choose $\eta \le \frac{1}{\mathsf{poly}(d)}$ to be small enough so that $\eta d T_1 \le \frac{1}{\eta \mathsf{poly}(d)}$. Then by Lemma C.8b, we have

$$\begin{aligned}
\|w_j^{(t+1)}\|_2^2 &= \|w_j^{(0)}\|_2^2 \pm \eta \sum_{s \le t} |\langle \nabla_{w_j} L(W^{(s)}, E^{(s)}), w_j^{(s)} \rangle| \pm \frac{1}{\mathsf{poly}(d)} \\
&\le \|w_j^{(0)}\|_2^2 \pm \eta \sum_{s \le t} \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) |\Lambda_{1,1}^{(s)}| \sum_{j \in [2]} |E_{j,3-j}^{(s)}| \pm \frac{1}{\mathsf{poly}(d)}
\end{aligned}$$

Since from the above analysis of the update of $B_{1,1}^{(t)}$, we know $\sum_{t \le T_1} \Lambda_{1,1}^{(t)} \le O(1)$. Moreover, we also know that $|B_{1,1}^{(t)}|$ is increasing and $\mathrm{sign}(\Lambda_{1,1}^{(t)}) = \mathrm{sign}(\Lambda_{1,1}^{(s)})$ for any $s, t \le T_1$. Thus they imply $\sum_{s \le t} |\Lambda_{1,1}^{(s)}| = |\sum_{s \le t} \Lambda_{1,1}^{(s)}| = O(1)$, which can be combine with Induction C.3d to prove the claim.

Proof of Induction C.3e: We can write down the update of $R_{1,2}^{(t)} = \langle \Pi_{V^\perp} w_1^{(t)}, w_2^{(t)} \rangle$ as follows

$$\begin{aligned}
\langle \Pi_{V^\perp} w_1^{(t+1)}, w_2^{(t+1)} \rangle &= \langle \Pi_{V^\perp} w_1^{(t)} - \Pi_{V^\perp} \eta \nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} - \Pi_{V^\perp} \eta \nabla_{w_2} L(W^{(t)}, E^{(t)}) \rangle \\
&= R_{1,2}^{(t)} - \eta \langle \nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} \rangle - \eta \langle \nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)} \rangle \\
&\quad + \eta^2 \langle \Pi_{V^\perp} \nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} \nabla_{w_2} L(W^{(t)}, E^{(t)}) \rangle
\end{aligned}$$

By Cauchy-Schwarz inequality and the same analysis above we have

$$\begin{aligned}
|\langle \Pi_{V^\perp} \nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} \nabla_{w_2} L(W^{(t)}, E^{(t)}) \rangle| &\le \|\nabla_{w_1} L(W^{(t)}, E^{(t)})\|_2 \|\nabla_{w_2} L(W^{(t)}, E^{(t)})\|_2 \\
&\le \widetilde{O}(d)
\end{aligned}$$

so by our choice of η

$$\sum_{t \le T_1} \eta^2 |\langle \Pi_{V^\perp} \nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} \nabla_{w_2} L(W^{(t)}, E^{(t)}) \rangle| \le \frac{1}{\mathsf{poly}(d)}$$

and by Lemma C.12 we have

$$\left| -\eta \langle \nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} \rangle - \eta \langle \nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)} \rangle \right| \le \eta \widetilde{O}(\Lambda_{1,1}^{(t)} B_{1,1}^{(t)})(\varrho + \frac{1}{\sqrt{d}})$$

which implies

$$|\langle \Pi_{V^\perp} w_1^{(t+1)}, w_2^{(t+1)} \rangle| \le |\langle \Pi_{V^\perp} w_1^{(0)}, w_2^{(0)} \rangle| + \sum_{s\le t}\sum_{j\in[2]} \eta |\langle \nabla_{w_j} L(W^{(s)}, E^{(s)}), \Pi_{V^\perp} w_{3-j}^{(s)} \rangle| + \frac{1}{\mathsf{poly}(d)}$$
$$\le \widetilde{O}(\frac{1}{\sqrt{d}}) + \sum_{s\le t} \eta \widetilde{O}(\Lambda_{1,1}^{(s)} B_{1,1}^{(s)}) + \frac{1}{\mathsf{poly}(d)}$$
$$\le \widetilde{O}(\frac{1}{\sqrt{d}}) + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) B_{1,1}^{(t+1)}$$
$$\le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$$

which completes the proof of Induction C.3. As for (a) – (e) of Lemma C.13, they are just direct corrolary of our induction at $t = T_1$. □

D Phase II: The Substitution Effect of Prediction Head

In this phase, As $B_{1,1}^{(t)}$ is learned to become very large ($B_{1,1}^{(t)} \gtrsim \|w_1^{(t)}\|_2$). The focus now shift to grow $E_{2,1}^{(t)}$, because we want $C_1\alpha_1^6((B_{2,1}^{(t)})^3 + E_{2,1}^{(t)}(B_{1,1}^{(t)})^3)^2$ in $H_{2,1}^{(t)}$ to dominate $\mathcal{E}_{2,1}^{(t)}$. We can write down the gradient of $E_{2,1}^{(t)}$ as

$$-\nabla_{E_{2,1}} L(W^{(t)}, E^{(t)}) = \sum_{\ell\in[2]} C_0 \Phi_2^{(t)} \alpha_\ell^6 (B_{2,\ell}^{(t)})^3((B_{1,\ell}^{(t)})^3 H_{2,3-\ell}^{(t)} - (B_{2,3-\ell}^{(t)})^3 K_{2,3-\ell}^{(t)}) - \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} \nabla_{E_{2,1}} \mathcal{E}_{2,1}^{(t)}$$

Now let us define

$$T_2 := \min\{t : R_2^{(t)} < \frac{1}{\log d} |E_{1,2}^{(t)}|\} \tag{D.1}$$

We will prove that $E_{2,1}^{(T_2)}$ reaches at most $O(\sqrt{\eta_E/\eta})$ and the following induction hypothesis holds throughout $t \in [T_1, T_2]$. In this phase, the learning of $E_{2,1}^{(t)}$ is much faster than the growth of the first feature v_1 such that $T_2 - T_1 = o(T_1/\sqrt{d})$, which is due to the acceleration effects brought by $B_{1,1}^{(t)} = \Omega(1)$ during this phase.

D.1 Induction in Phase II

We will be based on the following induction hypothesis during phase II.

Inductions D.1 (Phase II)**.** *When $t \in [T_1, T_2]$, we hypothesize the followings would hold*

(a) $B_{1,1}^{(t)} = \Theta(1)$, $B_{j,\ell}^{(t)} = B_{j,\ell}^{(T_1)}(1 \pm o(1)) = \widetilde{\Theta}(\frac{1}{\sqrt{d}})$ *for* $(j,\ell) \ne (1,1)$ *and* $\mathrm{sign}(B_{j,\ell}^{(t)}) = \mathrm{sign}(B_{j,\ell}^{(T_1)})$*;*

(b) $|R_{1,2}^{(t)}| = \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \alpha_1^{O(1)} [R_1^{(t)}]^{1/2} [R_2^{(t)}]^{1/2}$*;*

(c) $R_1^{(t)} \in [\Omega(\frac{1}{d^{3/4}\alpha_1^2}), O(1)]$*,* $R_2^{(t)} \in [\Omega(\frac{1}{\log d}\sqrt{\eta_E/\eta}), O(1)]$*;*

(d) $E_{1,2}^{(t)} \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}$ *and* $E_{2,1}^{(t)} \le O(\sqrt{\eta_E/\eta})$*.*

Under Induction D.1, we have some results as direct corollary.

Claim D.2. *At each iteration $t \in [T_1, T_2]$, if Induction C.3 holds, then*

(a) $\mathcal{E}_j^{(t)} = \Theta(C_2[R_j^{(t)}]^3)$*;*

(b) $\mathcal{E}_{j,3-j}^{(t)} = \mathcal{E}_j^{(t)} \pm \widetilde{O}(E_{j,3-j}^{(t)}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}) + O((E_{j,3-j}^{(t)})^2[R_{3-j}^{(t)}]^3)$ *for each* $j \in [2]$*;*

Proof. It is trivial to derive (a) from the expression of $\mathcal{E}_j^{(t)}$ and our assumption of ξ_p. For (b) it suffices to directly calculate the expression of $\mathcal{E}_{j,3-j}^{(t)}$ along with Induction D.1b. $\square$

Lemma D.3 (variables control in phase II)**.** *In Phase II ($t \in [T_1, T_2]$), if Induction D.1 holds, then*

(a) $\Phi_1^{(t)} = \widetilde{\Theta}(\frac{1}{\alpha_1^{12}})$, $\Phi_2^{(t)} = \Theta((C_2[R_2^{(t)}]^3 + C_1\alpha_1^6(E_{2,1}^{(t)})^2)^{-2})$;

(b) $K_{1,\ell}^{(t)} = \widetilde{O}(\alpha_\ell^6/d^{3/2})$, $K_{2,\ell}^{(t)} = \widetilde{O}(E_{2,1}^{(t)}\alpha_\ell^6/d^{3/2} + \alpha_\ell^6/d^3)$

(c) $H_{1,1}^{(t)} = \Theta(C_1\alpha_1^6)$, $H_{1,2}^{(t)} = \widetilde{O}([R_1^{(t)}]^3)$, $H_{2,2}^{(t)} = \Theta(C_2[R_2^{(t)}]^3)$, $H_{2,1}^{(t)} = \Theta(C_2[R_2^{(t)}]^3 + C_1\alpha_1^6(E_{2,1}^{(t)})^2)$.

Proof. The proof of (a) directly follows from Induction D.1a,c and Claim D.2. The proof of (b) follows directly from the expression of $K_{j,\ell}$ and Induction D.1a,d. The proof of (c) is also similar. $\square$

D.2 Gradient Lemmas for Phase II

Lemma D.4 (learning prediction head $E_{1,2}, E_{2,1}$ in phase II)**.** *If Induction D.1 holds at iteration $t \in [T_1, T_2]$, then we have*

$$(a)\quad -\nabla_{E_{1,2}} L(W^{(t)}, E^{(t)}) = (1 \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/2}}))\Sigma_{1,1}^{(t)}(-2E_{1,2}^{(t)}[R_2^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2})$$
$$\pm \Sigma_{1,1}^{(t)}\widetilde{O}(\frac{\eta_E/\eta}{\sqrt{d}})\max\{[R_1^{(t)}]^3, \frac{\alpha_1^{O(1)}}{d^{5/2}}\},$$
$$(b)\quad -\nabla_{E_{2,1}} L(W^{(t)}, E^{(t)}) = (1 \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/2}}))C_0\Phi_2^{(t)}\alpha_1^6(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^3H_{2,2}^{(t)}$$
$$\pm O(\Sigma_{2,1}^{(t)})(|E_{2,1}^{(t)}|[R_1^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2})$$

Proof. We first write down the gradient for $E_{j,3-j}^{(t)}$: (ignoring the time superscript $^{(t)}$)

$$-\nabla_{E_{j,3-j}} L(W, E) = \sum_{\ell\in[2]} C_0\Phi_j\alpha_\ell^6 B_{j,\ell}^3(B_{3-j,\ell}^3 H_{j,3-\ell} - B_{3-j,3-\ell}^3 K_{j,3-\ell}) - \sum_{\ell\in[2]}\Sigma_{j,\ell}\nabla_{E_{j,3-j}}\mathcal{E}_{j,3-j}$$

where $\nabla_{E_{j,3-j}}\mathcal{E}_{j,3-j} = \mathbb{E}\left[2\langle w_j, \xi_p\rangle^3\langle w_{3-j}, \xi_p\rangle^3 + 2E_{j,3-j}\langle w_{3-j}, \xi_p\rangle^6\right]$. Thus we have

$$\nabla_{E_{j,3-j}}\mathcal{E}_{j,3-j}^{(t)} = 2E_{j,3-j}^{(t)}[R_{3-j}^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$$

and by Claim B.1 and Induction D.1a, if $(j,\ell) \neq (1,1)$

$$\Sigma_{j,\ell}^{(t)} = O(\Sigma_{1,1}^{(t)})\frac{(B_{j,\ell}^{(t)})^6 + E_{j,3-j}^{(t)}(B_{3-j,\ell}^{(t)})^3(B_{j,\ell}^{(t)})^3}{(B_{1,1}^{(t)})^6}\frac{\Phi_j^{(t)}}{\Phi_1^{(t)}} \le o(\frac{1}{d^{3/2}})\Sigma_{1,1}^{(t)}\frac{\Phi_j^{(t)}}{\Phi_1^{(t)}}$$

Therefore for $j = 1$:

$$\sum_{\ell\in[2]}\Sigma_{1,\ell}^{(t)}\nabla_{E_{1,2}}\mathcal{E}_{1,2}^{(t)} = (1 \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/2}}))\Sigma_{1,1}^{(t)}\nabla_{E_{1,2}}\mathcal{E}_{1,2}^{(t)}$$

Now by Induction D.1a,c and Lemma D.3b,c we have $(B_{1,\ell}^{(t)})^3 H_{1,3-\ell}^{(t)} \le \max\{\Theta(C_2[R_1^{(t)}]^3), \widetilde{O}(\frac{\alpha_1^6}{d^{3/2}})\}$, which leads to the bounds

$$|(B_{1,\ell}^{(t)})^3(B_{2,\ell}^{(t)})^3 H_{1,3-\ell}^{(t)}| \le \widetilde{O}(\frac{1}{d^{3/2}})\max\{[R_1^{(t)}]^3, \frac{\alpha_1^6}{d^3}\}, \qquad |(B_{1,\ell}^{(t)})^3(B_{2,3-\ell}^{(t)})^3 K_{1,3-\ell}^{(t)}| \le \widetilde{O}(\frac{1}{d^3})$$

which implies

$$\left|\sum_{\ell\in[2]} C_0\Phi_1^{(t)}\alpha_\ell^6(B_{1,\ell}^{(t)})^3((B_{2,\ell}^{(t)})^3 H_{1,3-\ell}^{(t)} - (B_{2,3-\ell}^{(t)})^3 K_{1,3-\ell}^{(t)})\right| \lesssim \widetilde{O}(\frac{\eta_E/\eta}{\sqrt{d}})\Sigma_{1,1}^{(t)}\max\{[R_1^{(t)}]^3, \frac{\alpha_1^{O(1)}}{d^{5/2}}\}$$

Combining above together, we have

$$
\begin{aligned}
&-\nabla_{E_{1,2}} L(W^{(t)}, E^{(t)}) \\
&= (1+o(\frac{1}{d^{3/2}}))\Sigma_{1,1}^{(t)}(-2E_{1,2}^{(t)}[R_2^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \pm \widetilde{O}(\frac{\eta_E/\eta}{\sqrt{d}})\max\{[R_1^{(t)}]^3, \frac{\alpha_1^{O(1)}}{d^{5/2}}\})
\end{aligned}
$$

For $-\nabla_{E_{2,1}} L(W^{(t)}, E^{(t)})$, the expression is slightly different, we first observe that by Induction D.1a

$$\Delta_{2,2}^{(t)} \le \widetilde{O}(\frac{1}{d^{3/2}})\Delta_{2,1}^{(t)}$$

Meanwhile, by Induction D.1a and Lemma D.3b,c , we have

$$\Xi_2^{(t)} \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^3})C_0C_2\Phi_2^{(t)}[R_2^{(t)}]^3,$$

Moreover, we can also calculate $\Sigma_{2,1}^{(t)} = C_0C_2\alpha_1^6E_{2,1}^{(t)}\Phi_2^{(t)}(B_{2,1}^{(t)})^3) = \widetilde{O}(\frac{\alpha_1^6}{d^{3/2}})\Phi_2^{(t)}$, $\Sigma_{2,2}^{(t)} = \widetilde{O}(\frac{\alpha_2^6}{d^3})\Phi_2^{(t)}$, which gives

$$\sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\nabla_{E_{2,1}}\mathcal{E}_{2,1}^{(t)} = \Sigma_{2,1}^{(t)}(-\Theta(E_{2,1}^{(t)})[R_1^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2})$$

Now we combine the above results and get

$$
\begin{aligned}
-\nabla_{E_{2,1}} L(W^{(t)}, E^{(t)}) &= (1\pm\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/2}}))C_0\Phi_2^{(t)}\alpha_1^6(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^3\mathcal{E}_{2,1}^{(t)} \\
&\quad \pm O(\Sigma_{2,1}^{(t)})(|E_{2,1}^{(t)}|[R_1^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2})
\end{aligned}
$$

□

Lemma D.5 (reducing noise in phase II)**.** *Suppose Induction D.1 holds at $t\in[T_1,T_2]$, then*

(a) $\langle -\nabla_{w_1}L(W^{(t)},E^{(t)}), \Pi_{V^\perp}w_1^{(t)}\rangle = \Sigma_{1,1}^{(t)}\Theta(-[R_1^{(t)}]^3 \pm \widetilde{O}(|E_{1,2}^{(t)}| + \frac{|E_{2,1}^{(t)}|^2}{d^{3/2}})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2})$;

(b) $\langle -\nabla_{w_1}L(W^{(t)},E^{(t)}), \Pi_{V^\perp}w_2^{(t)}\rangle = \Sigma_{1,1}^{(t)}((-\Theta(\overline{R}_{1,2}^{(t)}) + O(\varrho))[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} + \widetilde{O}(|E_{1,2}^{(t)}| + \frac{|E_{2,1}^{(t)}|^2}{d^{3/2}})R_1^{(t)}[R_2^{(t)}]^2)$

And furthermore

$$
\begin{aligned}
(c)\quad \langle -\nabla_{w_2}L(W^{(t)},E^{(t)}), \Pi_{V^\perp}w_2^{(t)}\rangle &= -\Theta([R_2^{(t)}]^3)\Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big) \\
&\quad \pm O\Big(\sum_{j,\ell}\Sigma_{j,\ell}^{(t)}E_{j,3-j}^{(t)}(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big); \\
(d)\quad \langle -\nabla_{w_2}L(W^{(t)},E^{(t)}), \Pi_{V^\perp}w_1^{(t)}\rangle &= \Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big)(-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} \\
&\quad + O\Big(\sum_{j,\ell}\Sigma_{j,\ell}^{(t)}E_{j,3-j}^{(t)}R_2^{(t)}[R_1^{(t)}]^2\Big)
\end{aligned}
$$

Proof. The proof can be obtained directly from some calculation using Claim B.1 as follows:
Proof of (a): From (B.2), we can obtain that

$$\langle -\nabla_{w_1}L(W^{(t)},E^{(t)}), \Pi_{V^\perp}w_1^{(t)}\rangle = -\sum_{j,\ell}\Sigma_{j,\ell}^{(t)}\langle\nabla_{w_1}\mathcal{E}_{j,3-j}^{(t)}, w_1^{(t)}\rangle$$

Now from Claim B.1a and Induction D.1a, we know $(B_{j,\ell}^{(t)})^3 \le \widetilde{O}(\frac{1}{d^{3/2}})$ and the following

$$\Sigma_{j,\ell}^{(t)} = O(\Sigma_{1,1}^{(t)})\frac{(B_{j,\ell}^{(t)})^6 + E_{j,3-j}^{(t)}(B_{3-j,\ell}^{(t)})^3(B_{j,\ell}^{(t)})^3}{(B_{1,1}^{(t)})^6}\frac{\Phi_j^{(t)}}{\Phi_1^{(t)}} \le \widetilde{O}(\frac{E_{j,3-j}^{(t)}}{d^{3/2}})\Sigma_{1,1}^{(t)}\frac{\Phi_j^{(t)}}{\Phi_1^{(t)}} \quad \text{for any } (j,\ell)\ne(1,1)$$

From Induction D.1a,c, we know $((B_{2,\ell}^{(t)})^3 + E_{2,1}^{(t)}(B_{1,\ell}^{(t)})^3)^2 \le \widetilde{O}(\frac{1}{d^{3/2}})E_{2,1}^{(t)}$ and $R_2^{(t)} = \Theta(1)$, which by Claim D.2a,b and Lemma D.3a gives $\Phi_2^{(t)}/\Phi_1^{(t)} \le \widetilde{O}(\alpha_1^{O(1)})$. Combine the bounds above, we can obtain $\Sigma_{j,\ell}^{(t)} = \widetilde{O}(E_{j,3-j}^{(t)}/d^{3/2})\Sigma_{1,1}^{(t)}$. We can then directly apply Claim B.1 to prove Lemma D.5a as follows

$$\begin{aligned}
&\langle -\nabla_{w_1} L(W^{(t)},E^{(t)}), \Pi_{V^\perp} w_1^{(t)}\rangle \\
&= (1\pm\widetilde{O}(E_{1,2}^{(t)}))\Sigma_{1,1}^{(t)}\Big(-\Theta([R_1^{(t)}]^3) \pm O(E_{1,2}^{(t)})(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) \\
&\quad + \widetilde{O}(E_{2,1}^{(t)}/d^{3/2})\Sigma_{1,1}^{(t)}\Big(-\Theta((E_{2,1}^{(t)})^2)[R_1^{(t)}]^3 \pm O(E_{2,1}^{(t)})(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) \\
&= \Theta(\Sigma_{1,1}^{(t)})\Big(-[R_1^{(t)}]^3 \pm \widetilde{O}(|E_{1,2}^{(t)}| + \frac{|E_{2,1}^{(t)}|^2}{d^{3/2}})(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) \\
&\qquad \text{(Since } |E_{1,2}^{(t)}| \le d^{-\Omega(1)} \text{ by Induction D.1c,d)}
\end{aligned}$$

Proof of (b): For Lemma D.5b, we can use the same analysis for $\Sigma_{1,1}^{(t)}$ above and Claim B.1(d,e) to get (again we have used $\Sigma_{j,\ell}^{(t)} = \widetilde{O}(E_{j,3-j}^{(t)})\Sigma_{1,1}^{(t)} = o(\Sigma_{1,1}^{(t)})$)

$$\begin{aligned}
&\langle -\nabla_{w_1} L(W^{(t)},E^{(t)}), \Pi_{V^\perp} w_2^{(t)}\rangle \\
&= (1\pm\widetilde{O}(E_{1,2}^{(t)}))\Sigma_{1,1}^{(t)}\Big((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} + E_{1,2}^{(t)}R_1^{(t)}[R_2^{(t)}]^2\Big) \\
&\quad + \widetilde{O}(E_{2,1}^{(t)}/d^{3/2})\Sigma_{1,1}^{(t)}\Big((-\Theta(\overline{R}_{1,2}^{(t)}) + O(\varrho))(E_{2,1}^{(t)})^2[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} + E_{2,1}^{(t)}R_1^{(t)}[R_2^{(t)}]^2\Big) \\
&= \Sigma_{1,1}^{(t)}((-\Theta(\overline{R}_{1,2}^{(t)}) + O(\varrho))[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} + \widetilde{O}(|E_{1,2}^{(t)}| + \frac{|E_{2,1}^{(t)}|^2}{d^{3/2}})R_1^{(t)}[R_2^{(t)}]^2)
\end{aligned}$$

Proof of (c): Similarly to the proof of (a), we can also expand as follows

$$\begin{aligned}
&\langle -\nabla_{w_2} L(W^{(t)},E^{(t)}), \Pi_{V^\perp} w_2^{(t)}\rangle \\
&= (1\pm O(E_{1,2}^{(t)}))\Sigma_{1,1}^{(t)}\Big(-[R_2^{(t)}]^3\Theta((E_{1,2}^{(t)})^2) \pm O(E_{1,2}^{(t)})(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) \\
&\quad - \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big([R_2^{(t)}]^3 \pm O(E_{2,1}^{(t)})(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) \\
&= -[R_2^{(t)}]^3\Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big) \pm O\Big(\sum_{j,\ell}\Sigma_{j,\ell}^{(t)}E_{j,3-j}^{(t)}(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big)
\end{aligned}$$

Proof of (d): Similarly, we can calculate (again by $\Sigma_{j,\ell}^{(t)} = \widetilde{O}(E_{j,3-j}^{(t)})\Sigma_{1,1}^{(t)} = o(\Sigma_{1,1}^{(t)})$)

$$\begin{aligned}
&\langle -\nabla_{w_2} L(W^{(t)},E^{(t)}), \Pi_{V^\perp} w_1^{(t)}\rangle \\
&= \sum_{\ell\in[2]}\Sigma_{1,\ell}^{(t)}\Big((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))(E_{1,2}^{(t)})^2[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} + E_{1,2}^{(t)}R_2^{(t)}[R_1^{(t)}]^2\Big) \\
&\quad + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} + E_{2,1}^{(t)}R_2^{(t)}[R_1^{(t)}]^2\Big) \\
&= (1\pm\widetilde{O}(E_{1,2}^{(t)}))\Sigma_{1,1}^{(t)}\Big((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))(E_{1,2}^{(t)})^2[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} + E_{1,2}^{(t)}R_2^{(t)}[R_1^{(t)}]^2\Big) \\
&\quad + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} + E_{1,2}^{(t)}R_2^{(t)}[R_1^{(t)}]^2\Big) \\
&= \Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big)(-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} + O\Big(\sum_{j,\ell}\Sigma_{j,\ell}^{(t)}E_{j,3-j}^{(t)}R_2^{(t)}[R_1^{(t)}]^2\Big)
\end{aligned}$$

which completes the proof. $\square$

Lemma D.6 (learning feature v_2 in phase II)**.** *For each $t \in [T_1, T_2]$, if Induction D.1 holds at iteration t, then we have for each $j \in [2]$:*

$$|\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_2\rangle| \le \widetilde{O}(\frac{\alpha_2^6 \alpha_1^6}{d^{5/2}})\Big(\Phi_j^{(t)}(|E_{j,3-j}^{(t)}| + [R_j^{(t)}]^3) + \Phi_{3-j}^{(t)}(|E_{3-j,j}^{(t)}|[R_{3-j}^{(t)}]^3 + \frac{|E_{3-j,j}^{(t)}|^2}{d^{3/2}})\Big)$$

Proof. Again as in the proof of Lemma C.9, we expand the notations: (ignoring the superscript $^{(t)}$ for the RHS)

$$\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_2\rangle = \Lambda_{j,2}^{(t)} + \Gamma_{j,2}^{(t)} - \Upsilon_{j,2}^{(t)} \tag{D.2}$$

where

$$\begin{aligned}
\Lambda_{j,2}^{(t)} &= C_0 \alpha_2^6 \Phi_j^{(t)} H_{j,1}^{(t)} (B_{j,2}^{(t)})^5 \\
\Gamma_{j,2}^{(t)} &= C_0 \alpha_2^6 \Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,2}^{(t)})^3 (B_{j,2}^{(t)})^2 H_{3-j,1}^{(t)} \\
\Upsilon_{j,2}^{(t)} &= C_0 \alpha_1^6 \left(\Phi_j^{(t)} (B_{j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{j,2}^{(t)} + \Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{3-j,2}^{(t)}\right)
\end{aligned}$$

Now we further write $\Upsilon_{j,2}^{(t)} = \Upsilon_{j,2,1}^{(t)} + \Upsilon_{j,2,2}^{(t)}$, where

$$\Upsilon_{j,2,1}^{(t)} = C_0 \alpha_1^6 \Phi_j^{(t)} (B_{j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{j,2}^{(t)}, \qquad \Upsilon_{j,2,2}^{(t)} = \Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{3-j,2}^{(t)})^2 K_{3-j,2}^{(t)}$$

According to (D.2), we can first compute

$$\begin{aligned}
\Lambda_{j,2}^{(t)} - \Upsilon_{j,2,1}^{(t)} &= C_0 \alpha_2^6 \Phi_j^{(t)} (B_{j,2}^{(t)})^5 H_{j,1}^{(t)} - C_0 \alpha_1^6 \Phi_j^{(t)} (B_{j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{j,2}^{(t)} \\
&= C_0 \alpha_2^6 \Phi_j^{(t)} (B_{j,2}^{(t)})^5 \left(C_1 \alpha_1^6 ((B_{j,1}^{(t)})^3 + E_{j,3-j}^{(t)} (B_{3-j,1}^{(t)})^3)^2 + C_2 \mathcal{E}_{j,3-j}^{(t)}\right) \\
&\quad - C_0 \alpha_1^6 \Phi_j^{(t)} (B_{j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 C_1 \alpha_2^6 ((B_{j,2}^{(t)})^3 + E_{j,3-j}^{(t)} (B_{3-j,2}^{(t)})^3)((B_{j,1}^{(t)})^3 + E_{j,3-j}^{(t)} (B_{3-j,1}^{(t)})^3) \\
&= C_0 \alpha_2^6 C_1 \alpha_1^6 \Phi_j^{(t)} (B_{j,2}^{(t)})^5 \left(E_{j,3-j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{j,1}^{(t)})^3 + (E_{j,3-j}^{(t)})^2 (B_{3-j,1}^{(t)})^6\right) \\
&\quad - C_0 \alpha_2^6 C_1 \alpha_1^6 \Phi_j^{(t)} (B_{j,2}^{(t)})^2 (B_{3-j,2}^{(t)})^3 E_{j,3-j}^{(t)} \left((B_{j,1}^{(t)})^6 + E_{j,3-j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{j,1}^{(t)})^3\right) \\
&\quad + C_0 \alpha_2^6 \Phi_j^{(t)} (B_{j,2}^{(t)})^5 C_2 \mathcal{E}_{j,3-j}^{(t)}
\end{aligned}$$

Then we can apply Induction D.1a,c,d, Claim D.2a,b and Lemma D.3a,c to get

$$|\Lambda_{j,2}^{(t)} - \Gamma_{j,2,1}^{(t)}| \le \widetilde{O}(\frac{\alpha_2^6}{\alpha_1^6 d^{5/2}}) \Phi_j^{(t)} (|E_{j,3-j}^{(t)}| + [R_j^{(t)}]^3)$$

where the last inequality is due to Lemma D.3a,c. Similarly, we can also compute for $\Gamma_{j,2}^{(t)} - \Upsilon_{j,2,2}^{(t)}$:

$$\begin{aligned}
|\Gamma_{j,2}^{(t)} - \Upsilon_{j,2,2}^{(t)}| &\le \left|C_0 \alpha_2^6 \Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,2}^{(t)})^3 (B_{j,2}^{(t)})^2 H_{3-j,1}^{(t)}\right| \\
&\quad + \left|C_0 \alpha_1^6 \Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{3-j,2}^{(t)}\right| \\
&\le \widetilde{O}(\frac{\alpha_1^6 \alpha_2^6}{d^{5/2}}) \Phi_{3-j}^{(t)} |E_{3-j,j}^{(t)}| ([R_{3-j}^{(t)}]^3 + \frac{|E_{3-j,j}^{(t)}|}{d^{3/2}})
\end{aligned}$$

This completes the proof $\square$

Lemma D.7 (learning feature v_1 in Phase II)**.** *For each $t \in [T_1, T_2]$, if Induction D.1 holds at iteration t, then we have:*

(a) $\langle -\nabla_{w_1} L(W^{(t)}, E^{(t)}), v_1\rangle = \Theta(\Sigma_{1,1}^{(t)})[R_1^{(t)}]^3 + \Gamma_{1,1}^{(t)} \pm \widetilde{O}(\alpha_1^{O(1)}/d^{5/2})$;

(b) $\langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), v_1\rangle = \widetilde{O}(\alpha_1^{O(1)}/d^{5/2}) + \widetilde{O}(\frac{\alpha_1^6}{d}) E_{1,2}^{(t)} \Phi_1^{(t)} [R_1^{(t)}]^3$

Proof. As in the proof of Lemma D.6, we expand the gradient terms:

$$\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_1\rangle = \Lambda_{j,2}^{(t)} + \Gamma_{j,2}^{(t)} - \Upsilon_{j,2}^{(t)} \tag{D.3}$$

where

$$\Lambda_{j,1}^{(t)} = C_0\alpha_1^6\Phi_j^{(t)}H_{j,2}^{(t)}(B_{j,1}^{(t)})^5$$
$$\Gamma_{j,1}^{(t)} = C_0\alpha_1^6\Phi_{3-j}^{(t)}E_{3-j,j}^{(t)}(B_{3-j,1}^{(t)})^3(B_{j,1}^{(t)})^2H_{3-j,2}^{(t)}$$
$$\Upsilon_{j,1}^{(t)} = C_0\alpha_1^6\left(\Phi_j^{(t)}(B_{j,2}^{(t)})^3(B_{j,1}^{(t)})^2K_{j,1}^{(t)} + \Phi_{3-j}^{(t)}E_{3-j,j}^{(t)}(B_{3-j,2}^{(t)})^3(B_{j,1}^{(t)})^2K_{3-j,1}^{(t)}\right)$$

Indeed, when $j=1$, by Induction D.1a and Lemma D.3a,c, we can compute

$$\Lambda_{1,1}^{(t)} = C_0\alpha_1^6\Phi_1^{(t)}(B_{1,1}^{(t)})^5H_{1,2}^{(t)} = \Theta(\Sigma_{1,1}^{(t)})[R_1^{(t)}]^3$$

and with additionally Lemma D.3b, we also have

$$|\Upsilon_{1,1}^{(t)}| = \left|C_0\alpha_1^6\left(\Phi_1^{(t)}(B_{1,2}^{(t)})^3(B_{1,1}^{(t)})^2K_{1,1}^{(t)} + \Phi_2^{(t)}E_{2,j}^{(t)}(B_{2,2}^{(t)})^3(B_{1,1}^{(t)})^2K_{2,1}^{(t)}\right)\right| \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}})$$

which gives the proof of (a). For (b), we can also apply Induction D.1a and Lemma D.3a,c to get

$$\Lambda_{2,1}^{(t)} = C_0\alpha_1^6\Phi_2^{(t)}H_{2,2}^{(t)}(B_{2,1}^{(t)})^5 \le \widetilde{O}(\alpha_1^{O(1)}/d^{5/2})$$
$$\Gamma_{2,1}^{(t)} = C_0\alpha_1^6\Phi_1^{(t)}E_{1,2}^{(t)}(B_{1,1}^{(t)})^3(B_{2,1}^{(t)})^2H_{1,2}^{(t)} \le \widetilde{O}(\frac{1}{d})E_{1,2}^{(t)}\Phi_1^{(t)}\frac{[R_1^{(t)}]^3}{\alpha_1^6}$$
$$\Upsilon_{2,1}^{(t)} = C_0\alpha_1^6\left(\Phi_2^{(t)}(B_{2,2}^{(t)})^3(B_{2,1}^{(t)})^2K_{2,1}^{(t)} + \Phi_1^{(t)}E_{1,2}^{(t)}(B_{1,2}^{(t)})^3(B_{2,1}^{(t)})^2K_{1,1}^{(t)}\right) \le \widetilde{O}(\frac{\alpha_1^6}{d^4})$$

this finishes the proof. □

D.3 At the End of Phase II

Now we shall present the main theorem of this section, which gives the result of prediction head $E_{2,1}^{(t)}$ growth after the feature v_1 is learned in the first stage.

Lemma D.8 (Phase II). *Suppose $\eta = \frac{1}{\mathsf{poly}(d)}$ is sufficiently small, then Induction D.1 holds for all iteration $t \in [T_1, T_2]$, and at iteration $t = T_2$, the followings holds:*

(a) $B_{1,1}^{(T_2)} = \Theta(1)$, $B_{j,\ell}^{(T_2)} = B_{j,\ell}^{(T_1)}(1 \pm o(1)) = \widetilde{\Theta}(\frac{1}{\sqrt{d}})$ for $(j,\ell) \neq (1,1)$

(b) $R_1^{(T_2)} \le \widetilde{O}(\frac{1}{d^{3/4}})$, $R_2^{(T_2)} = \Theta(\sqrt{\eta_E/\eta})$, and $\overline{R}_{1,2}^{(T_2)} \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$;

(c) $|E_{1,2}^{(T_2)}| = \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$ and $|E_{2,1}^{(T_2)}| = \Theta(\sqrt{\eta_E/\eta})$

Where the part of learning $E_{2,1}^{(t)}$ is what we called substitution effect. One can easily verify that $|E_{2,1}^{(t)}f_1(X^{(1)})| \gg |f_2(X^{(1)})|$ when X is equipped with feature v_1, as stated in Lemma 5.2.

Proof. We first will prove Induction D.1 holds for all iteration $t \in [T_1, T_2]$. We shall first prove that if Induction D.1 continues to hold when $R_2^{(t)} \ge |E_{2,1}^{(t)}|$, we shall have $[R_1^{(t)}]$ decreasing at an exponential rate.

Proof of the decrease of $R_1^{(t)}$: Firstly, we write down the update of $R_1^{(t)}$ using Lemma D.5a:

$$R_1^{(t+1)} = R_1^{(t)} + \eta\Sigma_{1,1}^{(t)}\Theta(-[R_1^{(t)}]^3 \pm O(|E_{1,2}^{(t)}| + \frac{|E_{2,1}^{(t)}|^2}{d^{3/2}})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2})$$

from the expression of $\Sigma_{1,1}^{(t)}$ in (B.2), and by Induction D.1a and Lemma D.3a,c, we can compute

$$\Sigma_{1,1}^{(t)} = \Theta(C_0C_2\Phi_1^{(t)}) = \Theta(\frac{C_0C_2}{\alpha_1^{12}})$$

Moreover, from Induction D.1c we know that

$$(|E_{1,2}^{(t)}| + \frac{|E_{2,1}^{(t)}|^2}{d^{3/2}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \le (\widetilde{\Theta}(\frac{1}{d^{3/2}}) + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$$
$$\le (\widetilde{\Theta}(\frac{1}{d^{3/2}}) + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2})[R_1^{(t)}]^{3/2}$$

Therefore whenever $R_1^{(t)} \ge \frac{\alpha_1^{18}}{d^{3/4}}$ (which $t \le T_2$ suffices), we shall have always have

$$(\overline{R}_{1,2}^{(t)} + \varrho)(\widetilde{\Theta}(\frac{1}{d^{3/2}}) + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2})[R_1^{(t)}]^{3/2} \le o([R_1^{(t)}]^3)$$

which implies, if we set $T_2' := \min\{t : R_1^{(t)} \ge \frac{1}{d^{3/4}\alpha_1^2}\}$, then for all $t \in [T_1, T_2']$, we will have

$$\begin{aligned} R_1^{(t+1)} &= R_1^{(t)} + \eta\Sigma_{1,1}^{(t)}\Theta(-[R_1^{(t)}]^3 \pm O(|E_{1,2}^{(t)}| + \frac{|E_{2,1}^{(t)}|^2}{d^{3/2}})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}) \\ &= R_1^{(t)} - \Theta(\eta\Sigma_{1,1}^{(t)})[R_1^{(t)}]^3 && \text{(D.4)} \\ &\le R_1^{(t)}(1 - \Theta(\frac{\eta C_0 C_2}{\alpha_1^{12}})\frac{1}{d^{3/2}\alpha_1^2}) && (\text{since } R_1^{(t)} \ge \tfrac{1}{d^{3/4}}) \end{aligned}$$

From the last inequality we know that after $T_2 = T_1 + \widetilde{\Theta}(\frac{d^{1.5}}{\eta\alpha_1^{\Omega(1)}})$, we shall have $R_1^{(t)} \le O(\frac{\alpha_1^{O(1)}}{d^{3/4}})$. Moreover, suppose $T_2' < T_2$, (which just mean $R_1^{(s)} \le O(\frac{1}{d^{3/4}\alpha_1^2})$ for some iteration $s \in [T_1, T_2]$) we also have

$$\begin{aligned} R_1^{(t+1)} &= R_1^{(t)} - \Theta(\eta\Sigma_{1,1}^{(t)})[R_1^{(t)}]^3 \\ &\ge R_1^{(t)}(1 - \Theta(\frac{\eta C_0 C_2}{\alpha_1^{14}})\frac{1}{d^{3/2}}) \end{aligned}$$

So when $T_2 \le T_1 + \widetilde{O}(\frac{d^{1.5}\alpha_1^{12}}{\eta})$ iterations, we will have $R_1^{(t)} \ge R_1^{(s)}(1 - \Theta(\frac{\eta C_0 C_2}{d^{3/2}\alpha_1^{14}}))^{T_2 - T_1} \ge \Omega(R_1^{(t)})$ for all $t \in [s, T_2]$, which means we have a lower bound $R_1^{(t)} \ge \frac{1}{d^{3/4}\alpha_1^2}$ throughout $t \in [T_1, T_2]$. This proves Lemma D.8a and also our induction on $R_1^{(t)}$.

Proof of induction for $E_{1,2}^{(t)}$: By Lemma D.4a, we can write

$$\begin{aligned} -\nabla_{E_{1,2}} L(W^{(t)}, E^{(t)}) &= (1 + \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/2}}))\Sigma_{1,1}^{(t)}(-2E_{1,2}^{(t)}[R_2^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}) \\ &\quad \pm \Sigma_{1,1}^{(t)}\widetilde{O}(\frac{\eta_E/\eta}{\sqrt{d}})\max\{[R_1^{(t)}]^3, \frac{\alpha_1^{O(1)}}{d^{5/2}}\} \\ &= -\Theta(\Sigma_{1,1}^{(t)}[R_2^{(t)}]^3)E_{1,2}^{(t)} \pm O(\Sigma_{1,1}^{(t)})\Big((\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} + \widetilde{O}(\frac{\eta_E/\eta}{\sqrt{d}})[R_1^{(t)}]^3\Big) \end{aligned}$$

Since again from Induction D.1b,c that $\overline{R}_{1,2}^{(t)} \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}), R_1^{(t)} = O(1), R_2^{(t)} \in [\sqrt{\eta_E/\eta}, O(1)]$, we can obtain the update of $E_{1,2}^{(t)}$ as

$$\begin{aligned} E_{1,2}^{(t+1)} &= E_{1,2}^{(t)}(1 - \Theta(\eta_E\Sigma_{1,1}^{(t)}[R_2^{(t)}]^3)) \pm \widetilde{O}(\eta_E\Sigma_{1,1}^{(t)})\Big((\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} + \widetilde{O}(\frac{\eta_E/\eta}{\sqrt{d}})[R_1^{(t)}]^3\Big) \\ &= E_{1,2}^{(t)}(1 - \Theta(\eta_E\Sigma_{1,1}^{(t)}[R_2^{(t)}]^3)) \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})\eta_E\Sigma_{1,1}^{(t)}[R_1^{(t)}]^{3/2} \\ &= E_{1,2}^{(t)}(1 - \Theta(\eta_E\Sigma_{1,1}^{(t)}[R_2^{(t)}]^3)) \pm \eta_E\Sigma_{1,1}^{(t)}J_{1,2}^{(t)} \end{aligned}$$

where $J_{1,2}^{(t)} = \widetilde{C}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2} > 0$ and $\widetilde{C} = \widetilde{\Theta}(1)$ is larger than the hidden constant (including the polylog(d) factors) of $E_{2,1}^{(T_1)} \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$ in Lemma C.13d. And then we can compute

$$\begin{aligned} J_{1,2}^{(t+1)} &= \widetilde{C}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t+1)}]^{3/2} \\ &= \widetilde{C}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}(1 - \Theta(\eta\Sigma_{1,1}^{(t)})[R_1^{(t)}]^2)^{3/2} && \text{(due to calculations in (D.4))} \\ &= J_{1,2}^{(t)}(1 - \Theta(\eta^{3/2}(\Sigma_{1,1}^{(t)})^{3/2})[R_1^{(t)}]^3) && \text{(because } \eta\Sigma_{1,1}^{(t)} = \tfrac{\alpha_1^{O(1)}}{\mathsf{poly}(d)} \text{ is very small)} \end{aligned}$$

Now by Lemma C.13d, we know $|E_{1,2}^{(T_1)}| \leq J_{1,2}^{(T_1)}$; then we begin our induction that $|E_{1,2}^{(t)}| < (\log\log d)J_{1,2}^{(t)}$ at for all iterations $t \in [T_1, T_2]$. Now assume we have $|E_{1,2}^{(t)}| = \frac{1}{2}(\log\log d)J_{1,2}^{(t)}$[5], from above calculations it holds that $|E_{1,2}^{(t+1)}| = |E_{1,2}^{(t)}|(1 - \Theta(\eta\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3))$. Then we would have

$$\frac{J_{1,2}^{(t+1)}}{J_{1,2}^{(t)}} \geq (1 - \Theta(\eta^{3/2}(\Sigma_{1,1}^{(t)})^{3/2})[R_1^{(t)}]^3) \geq (1 - \Theta(\eta_E\Sigma_{1,1}^{(t)}[R_2^{(t)}]^3)) \geq \frac{|E_{1,2}^{(t+1)}|}{|E_{1,2}^{(t)}|}$$
$$\text{(because of the range of } R_1^{(t)} \text{ and } R_2^{(t)})$$

This proved that $|E_{1,2}^{(t+1)}| \lesssim \log\log d \cdot J_{1,2}^{(t+1)} \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t+1)}]^{3/2}$ and also the induction can go on until $t = T_2$.

Proof of the growth of $E_{2,1}^{(t)}$ and $T_2 \leq T_1 + O(\frac{d^{1.5}}{\eta\alpha_1^4})$: According to Lemma D.4b, we can write down the update of $E_{2,1}^{(t)}$ as

$$\begin{aligned} -\nabla_{E_{2,1}} L(W^{(t)}, E^{(t)}) = (1 \pm O(\frac{\alpha_1^{O(1)}}{d^{3/2}}))\Delta_{2,1}^{(t)} \\ \pm O(\Sigma_{2,1}^{(t)})(|E_{2,1}^{(t)}|[R_1^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}) \end{aligned}$$

Then, from Lemma D.3a,c and Induction D.1, we have

$$O(\Sigma_{2,1}^{(t)})(|E_{2,1}^{(t)}|[R_1^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}) \leq O(\frac{\mathsf{polylog}(d)}{d^{3/2}\alpha_1^2})\Phi_2^{(t)} \leq O(\frac{1}{d^{3/2}\alpha_1})\Phi_2^{(t)}$$

and also

$$\left|(1 \pm \widetilde{O}(\frac{\alpha_1^6}{d^{0.3}}))C_0\Phi_2^{(t)}\alpha_1^6(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^3H_{2,2}^{(t)}\right| \geq \widetilde{\Theta}(\frac{\alpha_1^6}{d^{3/2}})\Phi_2^{(t)}$$

Now by Lemma D.3a and Induction D.1a, it allow us to simplify the update to

$$\begin{aligned} E_{2,1}^{(t+1)} &= E_{2,1}^{(t)} - \eta_E\nabla_{E_{2,1}} L(W^{(t)}, E^{(t)}) \\ &= E_{2,1}^{(t)} + (1 \pm \frac{1}{\alpha_1^{\Omega(1)}})\eta_E C_0 C_2 \alpha_1^6 \Phi_2^{(t)}(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^3\mathcal{E}_{2,1}^{(t)} \\ &\geq E_{2,1}^{(t)} + \eta_E\widetilde{\Theta}(\frac{1}{d^{3/2}\alpha_1^6})\mathrm{sign}(B_{1,1}^{(t)})\mathrm{sign}(B_{2,1}^{(t)}) && \text{(by Induction D.1 and Claim D.2)} \end{aligned}$$

[5]If we want $|E_{1,2}^{(t)}| > (\log\log d)J_{1,2}^{(t)}$, then as long as $\eta = \frac{1}{\mathsf{poly}(d)}$ is small enough, we can always assume to have found some iteration $t' \in (T_1, t]$ such that $|E_{1,2}^{(t')}| = \frac{1}{2}(\log\log d)J_{1,2}^{(t)}$, and we set $t = t'$ and start our argument from that iteration.

Now since $\mathrm{sign}(B_{j,1}^{(t)}) = \mathrm{sign}(B_{j,1}^{(T_1)})$, we know there is an iteration $T'_{2,1} \le T_1 + O(\frac{d^{1/2}\alpha_1^{O(1)}}{\eta})$ such that for all $t \in [T'_{2,1}, T_2]$, it holds

$$
\begin{aligned}
|E_{2,1}^{(t)}| &= \left| E_{2,1}^{(T_1)} + \sum_{t\in[T_1,T'_{2,1}]} \Theta(\eta_E C_0 C_2 \alpha_1^6)\Phi_2^{(t)}(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^3[R_2^{(t)}]^3 \right| \\
&= \left| |E_{2,1}^{(T_1)}| \pm \sum_{s\in[T_1,T'_{2,1}]} \eta_E \widetilde{\Theta}(\frac{1}{d^{3/2}\alpha_1^{O(1)}}) \right| \\
&\in \left[2|E_{2,1}^{(T_1)}|, \widetilde{O}(\frac{\alpha_1^{O(1)}}{d}) \right]
\end{aligned}
$$

and thus $\mathrm{sign}(E_{2,1}^{(t)}) = \prod_{j\in[2]} \mathrm{sign}(B_{j,1}^{(t)})$ and $|E_{2,1}^{(t)}|$ will be increasing during $t \in [T'_{2,1}, T_2]$. Thus as long as $R_2^{(t)} \ge |E_{2,1}^{(t)}|$ continues to hold, after at most $\widetilde{\Theta}(\frac{d^{1.5}}{\eta\alpha_1^6})$ iterations starting from T_1, we shall have $|E_{2,1}^{(t)}| \ge \Omega(\sqrt{\eta_E/\eta})$.

However, in order to actually prove $|E_{2,1}^{(T_2)}| = \Theta(\sqrt{\eta_E/\eta})$, we will need to ensure that (1) there exist some constant $C = \Omega(\sqrt{\eta_E/\eta})$ such that $|E_{2,1}^{(t)}| > C$ while $R_2^{(s)} \ge \frac{1}{\log d}|E_{2,1}^{(t)}|$ for all $s \in [T_1, t]$; (2) we shall have a upper bound $|E_{2,1}^{(t)}| < O(\sqrt{\eta_E/\eta})$. They will be done below.

Proof of $E_{2,1}^{(T_2)} = \Theta(\sqrt{\eta_E/\eta})$ and $T_2 = T_1 + \widetilde{O}(\frac{d^{3/2}\alpha_1^{O(1)}}{\eta})$: In fact, Induction D.1c are already proved since we have already calculated the dynamics of $R_1^{(t)}$ and its upper bound and lower bound. In this part we are going to prove $T_2 = T_1 + \widetilde{\Theta}(\frac{d^{1.5}\alpha_1^{12}}{\eta})$ (which means that $R_2^{(t)} \le |E_{2,1}|$ can be achieved in $\widetilde{O}(\frac{d^{3/2}\alpha_1^{12}}{\eta})$ many iterations). From Lemma D.5c, we can write down the update for $R_2^{(t)}$ as

$$
\begin{aligned}
R_2^{(t+1)} &= R_2^{(t)} - 2\eta\langle \nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} \rangle + \eta^2 \|\Pi_{V^\perp} \nabla_{w_2} L(W^{(t)}, E^{(t)})\|_2^2 \\
&= R_2^{(t)} - \eta\Theta([R_2^{(t)}]^3)\Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}\Big) \\
&\quad \pm \eta O\Big(\sum_{j,\ell} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} (\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) + \frac{\eta}{\mathsf{poly}(d)}
\end{aligned}
$$

where we have used the fact that $\|\Pi_{V^\perp}\nabla_{w_2} L(W^{(t)}, E^{(t)})\|_2^2 \le \widetilde{O}(d^2)$ from our assumption on the noise ξ_p and a simple bound for $\Sigma_{j,\ell}^{(t)}$ as we have done before. Next we can resort to Induction D.1d that $|E_{1,2}^{(t)}| \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}$ to derive

$$
\begin{aligned}
\sum_{s\in[T_1,t]} \eta\Sigma_{1,1}^{(s)}\Theta((E_{1,2}^{(s)})^2) &\le \sum_{s\in[T_1,t]} \widetilde{O}(\varrho^2 + \frac{1}{d})\eta\Sigma_{1,1}^{(s)}[R_1^{(s)}]^3 \\
&\le \widetilde{O}(\varrho^2 + \frac{1}{d})
\end{aligned}
$$

which is because $\sum_{t\in[T_1,T_2]} \Theta(\eta\Sigma_{1,1}^{(t)})[R_1^{(t)}]^3 \le O(1)$ and $\Sigma_{1,1}^{(t)} > 0$ as we have calculated in the proof of Induction D.1a above. Similarly, we can also bound

$$
\sum_{s\in[T_1,t]} \Sigma_{1,\ell}^{(s)}|E_{1,2}^{(s)}|(|\overline{R}_{1,2}^{(s)}| + \varrho)[R_1^{(s)}]^{3/2}[R_2^{(s)}]^{3/2} \le \sum_{s\in[T_1,t]} \widetilde{O}(\varrho^2 + \frac{1}{d})\eta\Sigma_{1,\ell}^{(s)}[R_1^{(s)}]^3 \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})
$$

Moreover, because $T_2 \le T_1 + \widetilde{O}(\frac{d^{3/2}\alpha_1^{12}}{\eta})$ and $|E_{2,1}^{(t)}| \le O(1)$, $\Phi_2^{(t)} \le \alpha_1^{O(1)}$ from Induction D.1, we have for each $t \le T_2$:

$$\begin{aligned}\sum_{s\in[T_1,t]} \eta\Sigma_{2,\ell}^{(s)}|E_{2,1}^{(s)}|(|\overline{R}_{1,2}^{(s)}| + \varrho)[R_1^{(s)}]^{3/2}[R_2^{(s)}]^{3/2} &\le \widetilde{O}(\frac{|E_{2,1}^{(s)}|^2}{d^{3/2}}) \sum_{s\in[T_1,t]} \eta\Phi_2^{(s)}\widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \\ &\le \widetilde{O}(\frac{\eta}{d^{3/2}}) \cdot \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \cdot \widetilde{O}(\frac{d^{3/2}\alpha_1^{12}}{\eta}) \\ &\le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})\alpha_1^{O(1)} = o(\frac{1}{\log d})\end{aligned}$$

Thus combining all the bounds above, we have proved that for each $t \in [T_1, T_2]$, it holds

$$\begin{aligned}R_2^{(t)} &= R_2^{(T_1)} - \sum_{s\in[T_1,t]} \Theta(\eta\Sigma_{2,1}^{(t)})[R_2^{(t)}]^3 \pm o(1) \\ &= R_2^{(T_1)} - \sum_{s\in[T_1,t]} \Theta(\eta C_0 C_2)E_{2,1}^{(t)}\alpha_1^6\Phi_2^{(t)}(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^3[R_2^{(t)}]^3 \pm o(\frac{1}{\log d}) \qquad \text{(D.5)} \\ &= R_2^{(T_1)} - \sum_{s\in[T_1,t]} \eta E_{2,1}^{(t)}\widetilde{\Theta}(\frac{1}{d^{3/2}})\Phi_2^{(t)}[R_2^{(t)}]^3 \cdot \mathrm{sign}(E_{2,1}^{(t)}) \cdot \mathrm{sign}(B_{2,1}^{(T_1)}) \cdot \mathrm{sign}(B_{1,1}^{(T_1)}) \pm o(\frac{1}{\log d}) \qquad \text{(D.6)}\end{aligned}$$

where the last equality is because $\mathrm{sign}(B_{j,\ell}^{(t)}) \equiv \mathrm{sign}(B_{j,\ell}^{(T_1)})$ by Induction D.1a. Now from what we have proved above on the growth of $E_{2,1}^{(t)}$ that $\mathrm{sign}(E_{2,1}^{(t)}) = \mathrm{sign}(B_{1,1}^{(t)}B_{2,1}^{(t)}) \equiv \mathrm{sign}(B_{1,1}^{(T_1)}B_{2,1}^{(T_1)})$ throughout the rest of phase II (which is just $t \in [T'_{2,1}, T_2]$). Recall that

$$R_2^{(T'_{2,1})} = R_2^{(T_1)} \pm o(1), \quad \text{and} \quad E_{2,1}^{(t)} - E_{2,1}^{(T'_{2,1})} = \sum_{s\in[T'_{2,1},t]} \Theta(\eta_E C_0 C_2)\Phi_2^{(s)}(B_{2,1}^{(s)})^3(B_{1,1}^{(s)})^3$$

The above arguments imply for $t \in [T'_{2,1}, T_2]$:

$$\begin{aligned}R_2^{(t+1)} &= R_2^{(T_1)} - \sum_{s\in[T'_{2,1},t]} \Theta(\eta C_0 C_2)E_{2,1}^{(s)}\Phi_2^{(s)}(B_{2,1}^{(s)})^3(B_{1,1}^{(s)})^3[R_2^{(t)}]^3 \pm o(\frac{1}{\log d}) \\ &= R_2^{(T_1)} - \Theta(\frac{\eta}{\eta_E}|E_{2,1}^{(t)}|^2) - o(\frac{1}{\log d})\end{aligned}$$

Now we can confirm

(1) there exist a constant $C = \Theta(\sqrt{\eta_E/\eta})$ such that $E_{2,1}^{(t)} = C$ if $R_2^{(t)}$ falls below $\frac{1}{\log d}|E_{2,1}^{(t)}|$;

(2) $T_2 = T_1 + \widetilde{\Theta}(\frac{d^{3/2}\alpha_1^{12}}{\eta})$ due to the growth $|E_{2,1}^{(t+1)}| = |E_{2,1}^{(t)}| + \eta_E\widetilde{\Theta}(\frac{1}{d^{3/2}\alpha_1^{12}\sqrt{\eta_E/\eta}})$ for $t \in [T'_{2,1}, T_2]$.

which are the desired results.

Proof of Induction D.1a: We first obtain from Lemma D.7a that the update of $B_{1,1}^{(t)}$ can be written as

$$B_{1,1}^{(t+1)} = B_{1,1}^{(t)} + \eta\left(\Theta(\Sigma_{1,1}^{(t)})\mathrm{sign}(B_{1,1}^{(t)})[R_1^{(t)}]^3 + \Gamma_{1,1}^{(t)} \pm \widetilde{O}(\alpha_1^{O(1)}/d^{5/2})\right)$$

Now by what we have calculated above in (D.4), the total decrease of $R_1^{(t)}$ is (since $R_1^{(t)}$ is monotone in this phase)

$$\sum_{t\in[T_1,T_2]} \Theta(\eta\Sigma_{1,1}^{(t)})[R_1^{(t)}]^3 \le O(R_1^{(T_1)} - R_1^{(T_2)}) \le O(1)$$

And also since $T_2 \le T_1 + \widetilde{\Theta}(\frac{d^{3/2}\alpha_1^{12}}{\eta})$, we can bound

$$\sum_{t\in[T_1,T_2]} \widetilde{O}(\alpha_1^6/d^{5/2}) \le \widetilde{O}(\alpha_1^{O(1)}/d^{5/2}) \cdot \widetilde{O}(\frac{d^{3/2}}{\eta\alpha_1^6}) \le \widetilde{O}(\alpha_1^{O(1)}/d)$$

Now we consider how the $\Gamma_{1,1}^{(t)}$ term accumulates

$$\begin{aligned}
\sum_{t\in[T_1,T_2]} \eta\Gamma_{1,1}^{(t)} &= \left(\sum_{t\in[T_1,T'_{2,1}]} + \sum_{t\in[T'_{2,1},T_2]}\right)\eta C_0\alpha_1^6 E_{2,1}^{(t)}\Phi_2^{(t)}(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^2 H_{2,2}^{(t)} \\
&\overset{①}{=} \widetilde{O}(\frac{\alpha_1^{12}}{d}) + \sum_{t\in[T'_{2,1},T_2]} O\left(\eta C_0\alpha_1^6\Phi_2^{(t)}|B_{2,1}^{(t)}|^3|B_{1,1}^{(t)}|^3 H_{2,2}^{(t)}\right)\mathrm{sign}(B_{1,1}^{(t)}) \\
&= \pm o(1) + O(1)\mathrm{sign}(B_{1,1}^{(t)})
\end{aligned}$$

where in ① we have used $|E_{2,1}^{(t)}| \le O(1) \le O(B_{1,1}^{(t)})$ and $\mathrm{sign}(E_{2,1}^{(t)}) = \prod_{j\in[2]}\mathrm{sign}(B_{j,1}^{(t)})$ when $t \in [T'_{2,1}, T_2]$. These calculations tell us $B_{1,1}^{(t)} = B_{1,1}^{(T_1)} + O(1)\mathrm{sign}(B_{1,1}^{(T_1)}) \pm O(\frac{1}{\alpha_1}) = \Theta(1)$ for all iterations $t \in [T_1, T_2]$. Similarly from Lemma D.7b, for $B_{2,1}^{(t)}$ we can also write

$$B_{2,1}^{(T+1)} = B_{2,1}^{(t)} + \eta\widetilde{O}(\alpha_1^{O(1)}/d^{5/2}) + \widetilde{O}(\frac{\alpha_1^6}{d})E_{2,1}^{(t)}\Phi_1^{(t)}[R_1^{(t)}]^3$$

From similar calculations, it holds $B_{2,1}^{(t)} = B_{2,1}^{(T_1)} \pm \widetilde{O}(\alpha_1^{O(1)}/d)$, which proves that $B_{2,1}^{(t)} = B_{2,1}^{(T_1)}(1 \pm o(1))$ when $t \in [T_1, T_2]$. Now we turn to feature v_2. By Lemma D.6 we have for $j \in [2]$:

$$\begin{aligned}
|\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_2\rangle| &\le \widetilde{O}(\frac{\alpha_2^6\alpha_1^6}{d^{5/2}})\left(\Phi_j^{(t)}(|E_{j,3-j}^{(t)}| + [R_j^{(t)}]^3) + \Phi_{3-j}^{(t)}(|E_{3-j,j}^{(t)}|[R_{3-j}^{(t)}]^3 + \frac{|E_{3-j,j}^{(t)}|^2}{d^{3/2}})\right) \\
&\le \widetilde{O}(\frac{\alpha_2^6\alpha_1^6}{d^{5/2}})
\end{aligned}$$

where the last inequality is from Lemma D.3a and Induction D.1c,d. Thus when $t \le T_2 = T_1 + \widetilde{O}(\frac{d^{3/2}\alpha_1^{12}}{\eta})$ we would have

$$B_{j,2}^{(t)} = B_{j,2}^{(T_1)} \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d}) = B_{j,2}^{(T_1)}(1 \pm o(1)) \qquad \text{since } B_{j,2}^{(T_1)} = \widetilde{\Theta}(\tfrac{1}{\sqrt{d}}) \text{ by Lemma C.13c}$$

Together they proved Induction D.1a and Lemma D.8a. Moreover, we have also

Proof of Induction D.1b: Firstly, we write down the update of $R_{1,2}^{(t)}$ using Lemma D.5b,d as follows:

$$\begin{aligned}
R_{1,2}^{(t+1)} &= R_{1,2}^{(t)} - \eta\langle\nabla_{w_1}L(W^{(t)}, E^{(t)}), \Pi_{V^\perp}w_2^{(t)}\rangle - \eta\langle\nabla_{w_2}L(W^{(t)}, E^{(t)}), \Pi_{V^\perp}w_1^{(t)}\rangle \\
&\quad + \eta^2\langle\Pi_{V^\perp}\nabla_{w_1}L(W^{(t)}, E^{(t)}), \Pi_{V^\perp}\nabla_{w_2}L(W^{(t)}, E^{(t)})\rangle \\
&= R_{1,2}^{(t)} + \eta\Sigma_{1,1}^{(t)}((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} + \widetilde{O}(|E_{1,2}^{(t)}| + \frac{|E_{2,1}^{(t)}|^2}{d^{3/2}})R_1^{(t)}[R_2^{(t)}]^2) \\
&\quad + \eta\Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big)(-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} \\
&\quad + O\Big(\sum_{j,\ell}\eta\Sigma_{j,\ell}^{(t)}E_{j,3-j}^{(t)}R_2^{(t)}[R_1^{(t)}]^2\Big) + \frac{\eta}{\mathsf{poly}(d)}
\end{aligned}$$

where in the last inequality we have used

$$\begin{aligned}
&|\langle\Pi_{V^\perp}\nabla_{w_1}L(W^{(t)}, E^{(t)}), \Pi_{V^\perp}\nabla_{w_2}L(W^{(t)}, E^{(t)})\rangle| \\
&\le \|\Pi_{V^\perp}\nabla_{w_1}L(W^{(t)}, E^{(t)})\|_2\|\Pi_{V^\perp}\nabla_{w_2}L(W^{(t)}, E^{(t)})\|_2 \le \widetilde{O}(d)
\end{aligned}$$

Now from Induction D.1c,d that $R_2^{(t)} = \Theta(1)$ and $|E_{1,2}^{(t)}| \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}$, $|E_{2,1}^{(t)}| \le O(\sqrt{\eta_E/\eta})$, we can further obtain $|\Sigma_{2,2}^{(t)}| = \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/2}})|\Sigma_{2,1}^{(t)}|$, and the bound

$$R_{1,2}^{(t+1)} = R_{1,2}^{(t)}\Big(1 - \Theta(\eta\Sigma_{1,1}^{(t)})[R_1^{(t)}]^2 - \Theta(\eta(\Sigma_{1,1}^{(t)}(E_{1,2}^{(t)})^2 + \Sigma_{2,1}^{(t)}))[R_2^{(t)}]^2\Big)$$
$$\pm \eta O(\varrho)[R_2^{(t)}]^{1/2}[R_1^{(t)}]^{1/2}\left(O(\Sigma_{1,1}^{(t)})[R_1^{(t)}]^2 + \left(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \Sigma_{2,1}^{(t)}\right)[R_2^{(t)}]^2\right)$$

Notice here that there exist a constant $C = \Theta(1)$, whenever $|R_{1,2}^{(t)}| \ge C(\varrho + \frac{1}{\sqrt{d}})[R_2^{(t)}]^{1/2}[R_1^{(t)}]^{1/2}$, it will holds

$$\begin{aligned} R_{1,2}^{(t+1)} &= R_{1,2}^{(t)}\Big(1 - \Theta(\eta\Sigma_{1,1}^{(t)}[R_1^{(t)}]^2) - \Theta(\eta(\Sigma_{1,1}^{(t)}(E_{1,2}^{(t)})^2 + \Sigma_{2,1}^{(t)}))[R_2^{(t)}]^2\Big) \\ &= R_{1,2}^{(t)}\Big(1 - \Theta(\eta\Sigma_{1,1}^{(t)}[R_1^{(t)}]^2) - \Theta(\eta(\Sigma_{1,1}^{(t)}(E_{1,2}^{(t)})^2 + \frac{\alpha_1^6}{d^{3/2}}\Sigma_{2,1}^{(t)}))[R_2^{(t)}]^2\Big) \end{aligned}$$

Thus we can go through the same analysis as in the proof of induction for $E_{1,2}^{(t)}$ to derive that

$$|R_{1,2}^{(t)}| \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_2^{(t)}]^{1/2}[R_1^{(t)}]^{1/2}$$

which is the desired result. Note that at the end of phase II

Induction D.1a	$\Longrightarrow$	Lemma D.8a
Induction D.1b,c	$\Longrightarrow$	Lemma D.8b
Induction D.1d	$\Longrightarrow$	Lemma D.8c

We now complete the proof of Lemma D.8. □

E Phase III: The Acceleration Effect of Prediction Head

We shall prove in this section that the growth of $E_{2,1}^{(t)}$ in the previous phase creates an acceleration effect to the growth of $B_{2,2}^{(t)}$, which will finally outrun the growth of $B_{2,1}^{(t)}$ to win the lottery. We define

$$T_3 := \min\left\{t : |B_{2,2}^{(t)}| \ge \frac{1}{2}\min\{|B_{1,1}^{(t)}|, \sqrt{\frac{\eta}{\eta_E}}|E_{2,1}^{(t)}|\}\right\} \tag{E.1}$$

and we call iterations $t \in [T_2, T_3]$ as the phase III of training and $t \ge T_3$ as the end phase of training.

E.1 Induction in Phase III

Inductions E.1 (Phase III). *During $t \in [T_2, T_3]$, we hypothesize the following conditions holds.*

(a) $|B_{1,1}^{(t)}| = \Theta(1)$, $B_{2,1}^{(t)} = B_{2,1}^{(T_2)}(1 \pm o(1))$, $B_{1,2}^{(t)} = B_{1,2}^{(T_2)}(1 \pm o(1))$, $|B_{2,2}^{(t)}| \in [|B_{2,2}^{(T_2)}|, O(1)]$*;*

(b) $|E_{2,1}^{(t)}| = \Theta(\sqrt{\eta_E/\eta})$, $\mathrm{sign}(E_{2,1}^{(t)}) = \mathrm{sign}(E_{2,1}^{(T_2)})$ *and* $|E_{1,2}^{(t)}| \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$*;*

(c) $R_1^{(t)} \in [\Omega(\frac{1}{d}), O(\frac{d^{o(1)}}{d^{3/4}})]$, $[R_2^{(t)}] \in [\frac{1}{\sqrt{d}}, O(\frac{1}{\log d}\sqrt{\eta_E/\eta})]$.

As usual, before we prove the induction, we need to derive some useful claims. But firstly we shall give a much cleaner form of $\nabla_{E_{j,3-j}}L(W^{(t)}, E^{(t)})$ to help us understand the learning process of phase III and the end phase.

Fact E.2. Let us write

$$\begin{aligned} \Xi_j^{(t)} &= C_0C_1\alpha_1^6\alpha_2^6\Phi_j^{(t)}\Big((B_{1,1}^{(t)})^6(B_{2,2}^{(t)})^6 + (B_{2,1}^{(t)})^6(B_{1,2}^{(t)})^6\Big) \\ \Delta_{j,\ell}^{(t)} &= C_0\Phi_j^{(t)}\alpha_\ell^6(B_{j,\ell}^{(t)})^3(B_{3-j,\ell}^{(t)})^3C_2\mathcal{E}_{j,3-j}^{(t)} \end{aligned}$$

Then the gradient of $E_{j,3-j}^{(t)}$ can be written as

$$-\nabla_{E_{j,3-j}}L(W^{(t)}, E^{(t)}) = -\Xi_j^{(t)}E_{j,3-j}^{(t)} + \sum_{\ell\in[2]}\Delta_{j,\ell}^{(t)} - \sum_{\ell\in[2]}\Sigma_{j,\ell}^{(t)}\nabla_{E_{j,3-j}}\mathcal{E}_{j,3-j}^{(t)}$$

Proof. By expanding the gradients of $E_{j,3-j}^{(t)}$, we can verify by checking each monomial of polynomials of $B_{j,\ell}$ to obtain the first term, and leave the $\mathcal{E}_{j,3-j}^{(t)}$ part for the second term. $\square$

Lemma E.3 (variables control at phase III). *For $t \in [T_2, T_3]$, if Induction E.1 holds at iteration t, then we have*

(a) $\Phi_1^{(t)} = \widetilde{\Theta}(\frac{1}{\alpha_1^{12}})$, $[Q_2^{(t)}]^{-2} = \Theta(C_2[R_2^{(t)}]^3 + C_1\alpha_2^6(B_{2,2}^{(t)})^6)$, $U_2^{(t)} = \Theta(C_1(\alpha_1^6(E_{2,1}^{(t)})^2 + \alpha_2^6(B_{2,2}^{(t)})^6))$;

(b) $H_{1,1}^{(t)} = \Theta(C_1\alpha_1^6)$, $H_{1,2}^{(t)} \leq O(C_2[R_1^{(t)}]^3) + \widetilde{O}(\frac{\alpha_2^6}{d^3})$;

(c) $H_{2,1}^{(t)} = \Theta(C_1\alpha_1^6(E_{2,1}^{(t)})^2)$, $H_{2,2}^{(t)} = \Theta(C_2[R_2^{(t)}]^3)$;

(d) $\Sigma_{1,2}^{(t)} \leq \widetilde{O}(\frac{|E_{1,2}^{(t)}|}{d^{3/2}})\Sigma_{1,1}^{(t)}$;

(e) $\mathcal{E}_{j,3-j}^{(t)} = (1 \pm o(1))\mathcal{E}_j^{(t)} = O(C_2[R_j^{(t)}]^3)$

Proof. Assuming Induction E.1 holds at $t \in [T_2, T_3]$, we can recall the expression of these variables and prove their bounds directly. The bounds for Φ_1 and $H_{1,1}$ comes from $|B_{1,1}^{(t)}| = \Theta(1)$ and $|B_{1,2}^{(t)}|, |E_{1,2}^{(t)}| = o(1)$. The bounds for Q_2, U_2 comes from our definition of T_3 in (E.1). The rest of the claims can be derived by similar arguments using Induction E.1. $\square$

E.2 Gradient Lemmas for Phase III

In this subsection, we would give some gradient lemmas concerning the dynamics of our network in Phase III.

Lemma E.4 (learning feature v_2 in phase III). *For each $t \in [T_2, T_3]$, if Induction E.1 holds at iteration t, then we have:*

(a) $\langle -\nabla_{w_1} L(W^{(t)}, E^{(t)}), v_2 \rangle = \Theta(\frac{(B_{1,2}^{(t)})^2}{(B_{2,2}^{(t)})^2}) E_{2,1}^{(t)} \Lambda_{2,2}^{(t)} \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})|E_{2,1}^{(t)}|^2 \Phi_2^{(t)} \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}})$;

(b) $\langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), v_2 \rangle = (1 \pm \widetilde{O}(\frac{1}{d}))\Lambda_{2,2}^{(t)}$

Proof. Since $\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_2 \rangle = \Lambda_{j,2}^{(t)} + \Gamma_{j,2}^{(t)} - \Upsilon_{j,2}^{(t)}$, let us write down the definition of $\Lambda_{j,2}^{(t)}, \Gamma_{j,2}^{(t)}, \Upsilon_{j,2}^{(t)}$ respectively:

$$
\begin{aligned}
\Lambda_{j,2}^{(t)} &= C_0\alpha_2^6\Phi_j^{(t)} H_{j,1}^{(t)} (B_{j,2}^{(t)})^5 \\
\Gamma_{j,2}^{(t)} &= C_0\alpha_2^6\Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,2}^{(t)})^3 (B_{j,2}^{(t)})^2 H_{3-j,1}^{(t)} \\
\Upsilon_{j,2}^{(t)} &= C_0\alpha_1^6 \left(\Phi_j^{(t)} (B_{j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{j,2}^{(t)} + \Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{3-j,2}^{(t)} \right)
\end{aligned}
$$

Again we decompose $\Upsilon_{j,2}^{(t)} = \Upsilon_{j,2,1}^{(t)} + \Upsilon_{j,2,2}^{(t)}$ as in the proof of Lemma D.6, where

$$
\Upsilon_{j,2,1}^{(t)} = C_0\alpha_1^6\Phi_j^{(t)} (B_{j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{j,2}^{(t)}, \qquad \Upsilon_{j,2,2}^{(t)} = \Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{3-j,2}^{(t)})^2 K_{3-j,2}^{(t)}
$$

This gives

$$
\begin{aligned}
\Lambda_{j,2}^{(t)} - \Upsilon_{j,2,1}^{(t)} &= C_0\alpha_2^6\Phi_j^{(t)} (B_{j,2}^{(t)})^5 H_{j,1}^{(t)} - C_0\alpha_1^6\Phi_j^{(t)} (B_{j,1}^{(t)})^3 (B_{j,2}^{(t)})^2 K_{j,2}^{(t)} \\
&= C_0\alpha_2^6 C_1\alpha_1^6\Phi_j^{(t)} (B_{j,2}^{(t)})^5 \left(E_{j,3-j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{j,1}^{(t)})^3 + (E_{j,3-j}^{(t)})^2 (B_{3-j,1}^{(t)})^6 \right) \\
&\quad - C_0\alpha_2^6 C_1\alpha_1^6\Phi_j^{(t)} (B_{j,2}^{(t)})^2 (B_{3-j,2}^{(t)})^3 E_{j,3-j}^{(t)} \left((B_{j,1}^{(t)})^6 + E_{j,3-j}^{(t)} (B_{3-j,1}^{(t)})^3 (B_{j,1}^{(t)})^3 \right) \\
&\quad + C_0\alpha_2^6\Phi_j^{(t)} (B_{j,2}^{(t)})^5 C_2\mathcal{E}_{j,3-j}^{(t)}
\end{aligned}
$$

When $j=1$, from Induction E.1 and Lemma E.3a (which gives $\Phi_1^{(t)} \le \alpha_1^{O(1)}\Phi_2^{(t)}$), we can crudely obtain

$$\left|C_0\alpha_2^6 C_1\alpha_1^6\Phi_1^{(t)}(B_{1,2}^{(t)})^5\left(E_{1,2}^{(t)}(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^3+(E_{1,2}^{(t)})^2(B_{2,1}^{(t)})^6\right)\right| \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})\Phi_1^{(t)}|E_{1,2}^{(t)}|$$

$$\left|C_0\alpha_2^6 C_1\alpha_1^6\Phi_1^{(t)}(B_{1,2}^{(t)})^2(B_{2,2}^{(t)})^3E_{1,2}^{(t)}\left((B_{1,1}^{(t)})^6+E_{1,2}^{(t)}(B_{2,1}^{(t)})^3(B_{1,1}^{(t)})^3\right)\right| \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d})\Lambda_{2,2}^{(t)}|E_{1,2}^{(t)}|$$

$$\left|C_0\alpha_2^6\Phi_1^{(t)}(B_{1,2}^{(t)})^5C_2\mathcal{E}_{1,2}^{(t)}\right| = \widetilde{O}(\frac{\alpha_1^6}{d^{5/2}})\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3$$

So we have

$$\Lambda_{1,2}^{(t)}-\Upsilon_{1,2,1}^{(t)} = \widetilde{O}(\frac{\alpha_1^6}{d^{5/2}})\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3 \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d})\Lambda_{2,2}^{(t)}|E_{1,2}^{(t)}|$$

When $j=2$, we can also derive using Lemma E.3 about $H_{2,1}^{(t)}$ and Induction E.1 about $B_{2,1}^{(t)}$ and some rearrangement to obtain

$$C_0\alpha_2^6\Phi_2^{(t)}(B_{2,2}^{(t)})^5\left[C_1\alpha_1^6\left(E_{2,1}^{(t)}(B_{1,1}^{(t)})^3(B_{2,1}^{(t)})^3+(E_{2,1}^{(t)})^2(B_{1,1}^{(t)})^6\right)+C_2\mathcal{E}_{2,1}^{(t)}\right] = (1\pm\widetilde{O}(\frac{1}{d}))\Lambda_{2,2}^{(t)}$$

$$\left|C_0\alpha_2^6C_1\alpha_1^6\Phi_2^{(t)}(B_{2,2}^{(t)})^2(B_{1,2}^{(t)})^3E_{2,1}^{(t)}\left((B_{2,1}^{(t)})^6+E_{2,1}^{(t)}(B_{1,1}^{(t)})^3(B_{2,1}^{(t)})^3\right)\right| \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^3})|E_{2,1}^{(t)}|\Phi_2^{(t)}$$

which leads to the approximation

$$\Lambda_{2,2}^{(t)}-\Upsilon_{1,2,2}^{(t)} = (1\pm\widetilde{O}(\frac{1}{d}))\Lambda_{2,2}^{(t)} \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^3})|E_{2,1}^{(t)}|\Phi_2^{(t)}$$

Similarly, we can also calculate

$$\begin{aligned}\Gamma_{j,2}^{(t)}-\Upsilon_{j,2,2}^{(t)} &= C_0\alpha_2^6\Phi_{3-j}^{(t)}E_{3-j,j}^{(t)}(B_{3-j,2}^{(t)})^3(B_{j,2}^{(t)})^2H_{3-j,1}^{(t)} - C_0\alpha_1^6\Phi_{3-j}^{(t)}E_{3-j,j}^{(t)}(B_{3-j,1}^{(t)})^3(B_{j,2}^{(t)})^2K_{3-j,2}^{(t)}\\
&= C_0\alpha_2^6C_1\alpha_1^6\Phi_{3-j}^{(t)}(B_{3-j,2}^{(t)})^3(B_{j,2}^{(t)})^2E_{3-j,j}^{(t)}\left(E_{3-j,j}^{(t)}(B_{j,1}^{(t)})^3(B_{3-j,1}^{(t)})^3+(E_{3-j,j}^{(t)})^2(B_{j,1}^{(t)})^6\right)\\
&\quad - C_0\alpha_2^6C_1\alpha_1^6\Phi_{3-j}^{(t)}(B_{j,2}^{(t)})^5(E_{3-j,j}^{(t)})^2\left((B_{3-j,1}^{(t)})^6+E_{3-j,j}^{(t)}(B_{j,1}^{(t)})^3(B_{3-j,1}^{(t)})^3\right)\\
&\quad + C_0\alpha_2^6\Phi_{3-j}^{(t)}E_{3-j,j}^{(t)}(B_{3-j,2}^{(t)})^3(B_{j,2}^{(t)})^2C_2\mathcal{E}_{3-j,j}^{(t)}\end{aligned}$$

When $j=1$, following similar procedure as above, we can apply Induction E.1 and Lemma E.3 to give

$$\Gamma_{1,2}^{(t)}-\Upsilon_{1,2,2}^{(t)} = \Theta(\frac{(B_{1,2}^{(t)})^2}{(B_{2,2}^{(t)})^2})E_{2,1}^{(t)}\Lambda_{2,2}^{(t)} \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})|E_{2,1}^{(t)}|^2\Phi_2^{(t)}$$

Note that the first term on the RHS dominates the term $\pm\widetilde{O}(\frac{\alpha_1^{O(1)}}{d})\Lambda_{2,2}^{(t)}|E_{1,2}^{(t)}|$ in the approximation for $\Lambda_{1,2}^{(t)}-\Upsilon_{1,2,1}^{(t)}$ due to Induction E.1a,b. When $j=2$, since $\Phi_1^{(t)} \le \widetilde{\Theta}(\frac{1}{\alpha_1^{12}}) \le \alpha_1^{O(1)}\Phi_2^{(t)}H_{2,1}^{(t)}$ in this phase and $|B_{1,1}^{(t)}| = O(1)$, we can derive

$$|\Gamma_{2,2}^{(t)}-\Upsilon_{2,2,2}^{(t)}| \le \widetilde{\Theta}(\frac{\alpha_1^{O(1)}}{d^3})(E_{1,2}^{(t)})^2\Phi_1^{(t)} + \alpha_1^{O(1)}(E_{1,2}^{(t)})^2\Lambda_{2,2}^{(t)}$$

It can be seen that $(E_{1,2}^{(t)})^2\Phi_1^{(t)} \le (E_{2,1}^{(t)})^2\Phi_2^{(t)}$ by Induction E.1 and Lemma E.3. And by similar arguments we can have $(1\pm\widetilde{O}(\frac{1}{d}))\Lambda_{2,2}^{(t)} \ge \frac{1}{d^{\Omega(1)}}\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^3})|E_{2,1}^{(t)}|\Phi_2^{(t)}$. Combining all the results above, we can finish the proof. □

Lemma E.5 (learning feature v_1 in Phase III)**.** *For each $t\in[T_2,T_3]$, if Induction E.1 holds at iteration t, then we have: (recall that Δ-notation is from Fact E.2)*

(a) $\langle -\nabla_{w_1}L(W^{(t)},E^{(t)}), v_1\rangle = \Theta(\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3) \pm O(\frac{(B_{1,2}^{(t)})^3}{(B_{2,2}^{(t)})^3}+\frac{1}{\sqrt{d}})\alpha_1^{O(1)}\Lambda_{2,2}^{(t)} + \frac{E_{2,1}^{(t)}}{B_{1,1}^{(t)}}\Delta_{2,1}^{(t)} - \frac{B_{2,2}^{(t)}}{B_{1,1}^{(t)}}\Lambda_{2,2}^{(t)}$;

(b) $\langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), v_1\rangle = \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}})\Phi_2^{(t)}[R_2^{(t)}]^3 \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d})\Lambda_{2,2}^{(t)} \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^3})$

Proof. Recall that $\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_1\rangle = \Lambda_{j,1}^{(t)} + \Gamma_{j,1}^{(t)} - \Upsilon_{j,1}^{(t)}$. Similar to the proof of Lemma E.4, we can decompose $\Upsilon_{j,1}^{(t)} = \Upsilon_{j,1,1}^{(t)} + \Upsilon_{j,1,2}^{(t)}$ and do similar calculations:

$$\begin{aligned}\Lambda_{j,1}^{(t)} - \Upsilon_{j,1,1}^{(t)} &= C_0C_1\alpha_1^6\alpha_2^6\Phi_j^{(t)}(B_{j,1}^{(t)})^5\left(E_{j,3-j}^{(t)}(B_{3-j,2}^{(t)})^3(B_{j,2}^{(t)})^3 + (E_{j,3-j}^{(t)})^2(B_{3-j,2}^{(t)})^6\right)\\ &\quad - C_0C_1\alpha_1^6\alpha_2^6\Phi_j^{(t)}(B_{j,1}^{(t)})^2(B_{3-j,1}^{(t)})^3E_{j,3-j}^{(t)}\left((B_{j,2}^{(t)})^6 + E_{j,3-j}^{(t)}(B_{3-j,2}^{(t)})^3(B_{j,2}^{(t)})^3\right)\\ &\quad + C_0\alpha_1^6\Phi_j^{(t)}(B_{j,1}^{(t)})^5C_2\mathcal{E}_{j,3-j}^{(t)}\end{aligned}$$

When $j = 1$, from Induction E.1 and Lemma E.3a we know $\Phi_1^{(t)} \le \alpha_1^{(O(1))}$ during $t \in [T_2, T_3]$, which allow us to derive

$$\begin{aligned}&C_0C_1\alpha_1^6\alpha_2^6\Phi_1^{(t)}(B_{1,1}^{(t)})^5\left(E_{1,2}^{(t)}(B_{2,2}^{(t)})^3(B_{1,2}^{(t)})^3 + (E_{1,2}^{(t)})^2(B_{2,2}^{(t)})^6\right)\\ &\le \widetilde{O}(\Sigma_{1,1}^{(t)}(E_{1,2}^{(t)})^2) + C_0C_1\alpha_1^6\alpha_2^6(B_{1,1}^{(t)})^5E_{1,2}^{(t)}(B_{2,2}^{(t)})^3(B_{1,2}^{(t)})^3\\ &\le O(\frac{(B_{1,2}^{(t)})^3}{(B_{2,2}^{(t)})^3})\alpha_1^{O(1)}\Lambda_{2,2}^{(t)}|E_{1,2}^{(t)}| + \Theta(\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3)\end{aligned}$$

And

$$\left|C_0C_1\alpha_1^6\alpha_2^6\Phi_1^{(t)}(B_{1,1}^{(t)})^2(B_{2,1}^{(t)})^3E_{1,2}^{(t)}\left((B_{1,2}^{(t)})^6 + E_{1,2}^{(t)}(B_{2,2}^{(t)})^3(B_{1,2}^{(t)})^3\right)\right| \le \widetilde{O}(\frac{1}{d^{3/2}})|E_{1,2}^{(t)}|\Lambda_{2,2}^{(t)}$$

which can be summarized as

$$\Lambda_{1,1}^{(t)} - \Upsilon_{1,1,1}^{(t)} = \Theta(\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3) \pm O(\frac{(B_{1,2}^{(t)})^3}{(B_{2,2}^{(t)})^3} + (B_{2,1}^{(t)})^3 + \frac{1}{\sqrt{d}})|E_{1,2}^{(t)}|\alpha_1^{O(1)}\Lambda_{2,2}^{(t)}$$

A similar calculation also gives

$$\Lambda_{2,1}^{(t)} - \Upsilon_{2,1,1}^{(t)} = \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}})\Phi_2^{(t)}[R_2^{(t)}]^3 \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})\Phi_2^{(t)}|E_{2,1}^{(t)}| \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d})\Lambda_{2,2}^{(t)}B_{2,2}^{(t)}$$

Now we turn to the other terms in the gradient, from similar calculations in the proof of Lemma D.6, we have

$$\begin{aligned}\Gamma_{j,1}^{(t)} - \Upsilon_{j,1,2}^{(t)} &= C_0\alpha_2^6C_1\alpha_1^6\Phi_{3-j}^{(t)}(B_{3-j,1}^{(t)})^3(B_{j,1}^{(t)})^2E_{3-j,j}^{(t)}\left(E_{3-j,j}^{(t)}(B_{j,2}^{(t)})^3(B_{3-j,2}^{(t)})^3 + (E_{3-j,j}^{(t)})^2(B_{j,2}^{(t)})^6\right)\\ &\quad - C_0\alpha_2^6C_1\alpha_1^6\Phi_{3-j}^{(t)}(B_{j,1}^{(t)})^5(E_{3-j,j}^{(t)})^2\left((B_{3-j,2}^{(t)})^6 + E_{3-j,j}^{(t)}(B_{j,2}^{(t)})^3(B_{3-j,2}^{(t)})^3\right)\\ &\quad + C_0\alpha_2^6\Phi_{3-j}^{(t)}E_{3-j,j}^{(t)}(B_{3-j,1}^{(t)})^3(B_{j,1}^{(t)})^2C_2\mathcal{E}_{3-j,j}^{(t)}\end{aligned}$$

which also similarly gives

$$\Gamma_{1,1}^{(t)} - \Upsilon_{1,1,2}^{(t)} = \frac{E_{2,1}^{(t)}}{B_{1,1}^{(t)}}\Delta_{2,1}^{(t)} - \frac{B_{2,2}^{(t)}}{B_{1,1}^{(t)}}\Lambda_{2,2}^{(t)} \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/2}})\Lambda_{2,2}^{(t)}$$

and

$$|\Gamma_{2,1}^{(t)} - \Upsilon_{2,1,2}^{(t)}| \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d})\Phi_1^{(t)}((E_{1,2}^{(t)})^2 + |E_{1,2}^{(t)}|[R_1^{(t)}]^3) \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^3})$$

which finishes the proof. $\square$

Lemma E.6 (reducing noise in phase III). *Suppose Induction E.1 holds at $t \in [T_2, T_3]$, then we have*

$$
\begin{aligned}
(a)\quad \langle -\nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)} \rangle &= -\Theta([R_1^{(t)}]^3)\Big(\Sigma_{1,1}^{(t)} + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} (E_{2,1}^{(t)})^2\Big) \\
&\quad \pm O\Big(\sum_{j,\ell} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} (\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big); \\
(b)\quad \langle -\nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} \rangle &= \Big(\Sigma_{1,1}^{(t)} + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} (E_{2,1}^{(t)})^2\Big)(-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} \\
&\quad + O\Big(\sum_{(j,\ell)\neq(1,2)} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} R_1^{(t)} [R_2^{(t)}]^2\Big) \\
(c)\quad \langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} \rangle &= -\Theta([R_2^{(t)}]^3)\Big(\sum_{\ell\in[2]} \Sigma_{1,1}^{(t)} \Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}\Big) \\
&\quad \pm O\Big(\sum_{j,\ell} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} (\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big); \\
(d)\quad \langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)} \rangle &= \Big(\Sigma_{1,1}^{(t)} \Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}\Big)(-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} \\
&\quad + O\Big(\sum_{(j,\ell)\neq(1,2)} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} R_2^{(t)} [R_1^{(t)}]^2\Big)
\end{aligned}
$$

Proof. The proof of Lemma E.6 is very similar to Lemma D.5, but we write it down to stress some minor differences. As in (B.2), we first write down

$$\langle -\nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)} \rangle = -\sum_{j,\ell} \Sigma_{j,\ell}^{(t)} \langle \nabla_{w_1} \mathcal{E}_{j,3-j}^{(t)}, w_1^{(t)} \rangle$$

Proof of (a): Combine the bounds above, we can obtain for each $j \in [2]$: $\Sigma_{1,2}^{(t)} = \widetilde{O}(E_{1,2}^{(t)}/d^{3/2})\Sigma_{1,1}^{(t)}$. We can then directly apply Claim B.1 to prove Lemma E.6a as follows

$$
\begin{aligned}
&\langle -\nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)} \rangle \\
&= (1 \pm \widetilde{O}(E_{1,2}^{(t)}/d^{3/2}))\Sigma_{1,1}^{(t)}\Big(-\Theta([R_1^{(t)}]^3) \pm O(E_{1,2}^{(t)})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) \\
&\quad + (\Sigma_{2,1}^{(t)} + \Sigma_{2,2}^{(t)})\Big(-\Theta((E_{2,1}^{(t)})^2)[R_1^{(t)}]^3 \pm O(E_{2,1}^{(t)})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) \\
&= -\Theta(\Sigma_{1,1}^{(t)} + \Sigma_{2,1}^{(t)} + \Sigma_{2,2}^{(t)})[R_1^{(t)}]^3 \pm O(\sum_{j,\ell} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} (\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big)
\end{aligned}
$$

(Since $|E_{1,2}^{(t)}| \leq d^{-\Omega(1)}$ by Induction E.1)

Proof of (b): For Lemma D.5b, we can use the same analysis for $\Sigma_{1,1}^{(t)}$ above and Claim B.1d,e to get (again we have used $\Sigma_{1,2}^{(t)} = \widetilde{O}(E_{1,2}^{(t)}/d^{3/2})\Sigma_{1,1}^{(t)}$)

$$
\begin{aligned}
&\langle -\nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} \rangle \\
&= (1 \pm \widetilde{O}(E_{1,2}^{(t)}/d^{3/2}))\Sigma_{1,1}^{(t)}\Big((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} + E_{1,2}^{(t)} R_1^{(t)} [R_2^{(t)}]^2\Big) \\
&\quad + \Theta(\Sigma_{2,1}^{(t)} + \Sigma_{2,2}^{(t)})\Big((-\Theta(\overline{R}_{1,2}^{(t)}) + O(\varrho))(E_{2,1}^{(t)})^2[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} + E_{2,1}^{(t)} R_1^{(t)} [R_2^{(t)}]^2\Big) \\
&= \Big(\Sigma_{1,1}^{(t)} + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} (E_{2,1}^{(t)})^2\Big)((-\Theta(\overline{R}_{1,2}^{(t)}) + O(\varrho))[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2}) \\
&\quad + O\Big(\sum_{(j,\ell)\neq(2,1)} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} R_1^{(t)} [R_2^{(t)}]^2\Big)
\end{aligned}
$$

Proof of (c): Similarly to the proof of (a), we can also expand as follows

$$\begin{aligned}
&\langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)} \rangle \\
&= (1 \pm \widetilde{O}(E_{1,2}^{(t)}/d^{3/2}))\Sigma_{1,1}^{(t)} \Big(-[R_2^{(t)}]^3 \Theta((E_{1,2}^{(t)})^2) \pm O(E_{1,2}^{(t)})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \Big) \\
&\quad - \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} \Big([R_2^{(t)}]^3 \pm O(E_{2,1}^{(t)})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \Big) \\
&= -\Theta([R_2^{(t)}]^3)\Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} \Big) \pm O\Big(\sum_{j,\ell} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} (\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \Big)
\end{aligned}$$

Proof of (d): Similarly, we can calculate

$$\begin{aligned}
&\langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)} \rangle \\
&= (1 \pm \widetilde{O}(E_{1,2}^{(t)}/d^{3/2}))\Sigma_{1,1}^{(t)} \Big((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))(E_{1,2}^{(t)})^2 [R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} + E_{1,2}^{(t)} R_2^{(t)} [R_1^{(t)}]^2 \Big) \\
&\quad + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} \Big((-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} + E_{1,2}^{(t)} R_2^{(t)} [R_1^{(t)}]^2 \Big) \\
&= \Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} \Big)(-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} \\
&\quad + O\Big(\sum_{(j,\ell)\neq(2,1)} \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} R_2^{(t)} [R_1^{(t)}]^2 \Big)
\end{aligned}$$

which completes the proof. □

Lemma E.7 (learning the prediction head in phase III)**.** *If Induction E.1 holds at iteration $t \in [T_2, T_3]$, then using the notations from Fact E.2, we have*

$$\begin{aligned}
-\nabla_{E_{j,3-j}} L(W^{(t)}, E^{(t)}) = \Theta\Big(\sum_{\ell\in[2]} \Sigma_{j,\ell}^{(t)} \Big)(-E_{j,3-j}^{(t)}[R_{3-j}^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}) \\
- \Xi_j^{(t)} E_{j,3-j}^{(t)} + \sum_{\ell\in[2]} \Delta_{j,\ell}^{(t)}
\end{aligned}$$

Proof. By Fact E.2, we only need to bound the last term $\sum_{\ell\in[2]} \Sigma_{j,\ell}^{(t)} \nabla_{E_{1,2}} \mathcal{E}_{j,3-j}^{(t)}$, which can be directly obtained from applying Claim B.1. □

E.3 At the End of Phase III

In order to argue that $B_{2,2}^{(T_2)} = \Omega(1)$ at the end of phase III, we need to define some auxiliary notions. Recall that T_3 is defined in (E.1), and now we further define

$$T_{3,1} := \min\{t : C_1 \alpha_2^6 (B_{2,2}^{(t)})^6 \geq C_2 [R_2^{(t)}]^3\}, \qquad T_{3,2}^{(t)} = \min\Big\{t : |B_{2,2}^{(t)}| \geq \frac{1}{3}\min\{|E_{2,1}^{(t)}|, |B_{1,1}^{(t)}|\}\Big\} \tag{E.2}$$

It can be observed that if Induction E.1 holds for $t \in [T_2, T_3]$ and our learning rate η is small enough, we shall have $T_2 < T_{3,1} \leq T_{3,2} < T_3$. Now we are ready to present the main lemma we want to prove in this phase.

Lemma E.8 (Phase III)**.** *Let T_3 be defined as in* (E.1)*. Suppose $\eta = \frac{1}{\mathsf{poly}(d)}$ is sufficiently small, then Induction E.1 holds for all iteration $t \in [T_2, T_3]$, and at iteration $t = T_3$, the followings holds:*

(a) $|B_{1,1}^{(T_3)}| = \Theta(1)$, $|B_{2,2}^{(T_3)}| = \Theta(1)$, $B_{j,\ell}^{(T_3)} = B_{j,\ell}^{(T_2)}(1 \pm o(1))$ for $j \neq \ell$;

(b) $R_1^{(T_3)} = \widetilde{O}(\frac{1}{d^{3/4}})$, $R_2^{(T_3)} \in [\widetilde{O}(\frac{1}{d^{1/2}}), \widetilde{O}(\frac{1}{d^{1/4}})]$, and $\overline{R}_{1,2}^{(T_3)} \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$;

(c) $|E_{2,1}^{(T_2)}| = \Theta(\sqrt{\eta_E/\eta})$ and $|E_{1,2}^{(T_2)}| = \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} = \widetilde{O}(\frac{1}{d})$.

Moreover, $|B_{2,2}^{(t)}|$ is increasing and $R_2^{(t)}$ is decreasing. The part of learning $|B_{2,2}^{(t)}|$ till $\Omega(1)$ and keeping $B_{2,1}^{(t)}$ close to its initialization is what's been accelerated by the prediction head $E_{2,1}^{(t)}$.

The proof of Lemma E.8 will be proven after we have proven Induction E.1, which will again be proven after some intermediate results are proven.

Lemma E.9 (The growth of $B_{2,2}^{(t)}$ before $T_{3,1}$)**.** *Let $T_{3,1}$ be defined as in* (E.2)*. If Induction E.1 holds for $t \in [T_2, T_{3,1}]$, then we have $R_2^{(T_{3,1})} \le \frac{\alpha_1^{12}}{d^{1/4}}$ and $B_{2,2}^{(T_{3,1})} \in [\frac{1}{d^{1/4}}, O(\frac{\alpha_1^{O(1)}}{d^{1/4}})]$ and $T_{3,1} \le T_2 + \widetilde{O}(\frac{d^{1.625}\alpha_1^{O(1)}}{\eta})$.*

Proof. Firstly by Lemma E.6b , we can write down the update of $R_2^{(t)}$: (as in Lemma D.8)

$$R_2^{(t+1)} = R_2^{(t)} - \eta\Theta([R_2^{(t)}]^3)\Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big)$$
$$\pm O\Big(\sum_{j,\ell}\eta\Sigma_{j,\ell}^{(t)}E_{j,3-j}^{(t)}(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\Big) \pm \frac{\eta}{\mathsf{poly}(d)}$$

Next, by Claim B.1 and Lemma E.3a combined with Induction E.1a,b, we have $\widetilde{O}(\frac{|E_{2,1}^{(t)}|}{d^{3/2}})\Sigma_{1,1}^{(t)}\frac{\Phi_1^{(t)}}{\Phi_2^{(t)}} \le \widetilde{O}(\Sigma_{2,1}^{(t)})$, which leads to the bound

$$\eta\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) \le \widetilde{O}(\varrho^2 + \frac{1}{d})\alpha_1^{O(1)}\eta\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3[R_2^{(t)}]^3 \le O(\frac{1}{d^{9/4}})\eta\Sigma_{1,1}^{(t)}[R_2^{(t)}]^3 \le O(\frac{\alpha_1^{O(1)}}{d^{3/4}})\eta\Sigma_{2,1}^{(t)}[R_2^{(t)}]^3$$

Similarly, we can bound the following term

$$\begin{aligned}\sum_{\ell\in[2]}\eta\Sigma_{1,\ell}^{(t)}|E_{1,2}^{(t)}|(|\overline{R}_{1,2}^{(t)}| + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} &\le \widetilde{O}(\varrho^2 + \frac{1}{d})\alpha_1^{O(1)}\sum_{\ell\in[2]}\eta\Sigma_{1,\ell}^{(t)}[R_1^{(t)}]^3[R_2^{(t)}]^3\\ &\le \widetilde{O}(\varrho^2 + \frac{1}{d})\alpha_1^{O(1)}\frac{1}{d^{9/4}}\sum_{\ell\in[2]}\eta\Sigma_{1,\ell}^{(t)}[R_2^{(t)}]^3\\ &\le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{3/4}})\eta\Sigma_{2,1}^{(t)}[R_2^{(t)}]^3\end{aligned}$$

Moreover, from Induction E.1c that $R_2^{(t)} \ge R_1^{(t)}$, we can also calculate for each $t \in [T_2, T_{3,1}]$:

$$\eta\Sigma_{2,\ell}^{(s)}|E_{2,1}^{(t)}|(|\overline{R}_{1,2}^{(t)}| + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})\alpha_1^{O(1)}\eta\Sigma_{2,\ell}^{(t)}[R_2^{(t)}]^3$$

Thus by combining the results above, we have the update of $R_2^{(t)}$ at $t \in [T_2, T_3]$ as follows:

$$\begin{aligned}R_2^{(t+1)} &= R_2^{(t)} - \eta\Theta([R_2^{(t)}]^3)\Big(\Sigma_{1,1}^{(t)}\Theta((E_{1,2}^{(t)})^2) + \sum_{\ell\in[2]}\Sigma_{2,\ell}^{(t)}\Big)\\ &= R_2^{(t)} - \eta(\Sigma_{2,1}^{(t)} + \Sigma_{2,2}^{(t)})[R_2^{(t)}]^3 \qquad \text{(E.3)}\end{aligned}$$

which implies that $R_2^{(t)}$ is decreasing throughout phase III. From Lemma E.3a and Induction E.1b, we know that for $t \in [T_2, T_{3,1}]$:

$$\Phi_2^{(t)} = Q_2^{(t)}/[U_2^{(t)}]^{3/2} = \Theta(\frac{1}{\sqrt{C_2[R_2^{(t)}]^3}(C_1\alpha_1^6(E_{2,1}^{(t)})^2)^{3/2}})$$

which implies (also using a bit of Claim B.1 and Induction E.1a)

$$\begin{aligned}
\Sigma_{2,1}^{(t)}[R_2^{(t)}]^3 &= (1 \pm \widetilde{O}(\frac{1}{d^{3/2}}))E_{2,1}^{(t)}\Delta_{2,1}^{(t)} \\
&= (1 \pm \widetilde{O}(\frac{1}{d^{3/2}}))(1 \pm \widetilde{O}(\frac{1}{d^{3/2}}))C_0 C_2 \alpha_1^6 \Phi_2^{(t)} E_{2,1}^{(t)}(B_{1,1}^{(t)})^3(B_{2,1}^{(t)})^3[R_2^{(t)}]^3 \\
&= \Theta(\frac{C_2^{1/2}[R_2^{(t)}]^{3/2}}{(U_2^{(t)})^{3/2}})C_0\alpha_1^6 E_{2,1}^{(t)}(B_{1,1}^{(t)})^3(B_{2,1}^{(t)})^3 \\
&= \Theta(\frac{C_0 C_2^{1/2}|B_{2,1}^{(T_2)}|^3}{C_1^{3/2}\alpha_1^3|E_{2,1}^{(T_2)}|})[R_2^{(t)}]^{3/2}
\end{aligned}$$

(because $B_{2,1}^{(t)} = B_{2,1}^{(T_2)}(1 \pm o(1))$, $B_{1,1}^{(t)} = \Theta(B_{1,1}^{(T_2)})$ and $E_{2,1}^{(t)} = \Theta(E_{2,1}^{(T_2)})\text{sign}(B_{1,1}^{(T_2)}B_{2,1}^{(T_2)})$)

And for $\Sigma_{2,2}^{(t)}$, from some simple calcualtions (using Claim B.1), we have

- when $|B_{2,2}^{(t)}| \leq \frac{\alpha_1}{\alpha_2}\sqrt{|B_{2,1}^{(T_2)}|}$, we would have $\Sigma_{2,2}^{(t)} \leq O(\Sigma_{2,1}^{(t)})$;
- otherwise, we have $\Sigma_{2,1}^{(t)} + \Sigma_{2,2}^{(t)} = \Theta(\Sigma_{2,2}^{(t)})$.

So by (E.3), we know R_2 is decreasing for $t \in [T_2, T_{3,1}]$ by at least

$$R_2^{(t+1)} \leq R_2^{(t)} - \eta\Theta(\frac{C_0 C_2^{1/2}|B_{2,1}^{(T_2)}|^3}{C_1^{3/2}\alpha_1^3|E_{2,1}^{(T_2)}|})[R_2^{(t)}]^{3/2} \leq R_2^{(t)}(1 - \eta\zeta[R_2^{(t)}]^{1/2}) \tag{E.4}$$

where $\zeta := \Theta(\frac{C_0 C_2^{1/2}|B_{2,1}^{(T_2)}|^3}{C_1^{3/2}\alpha_1^3|E_{2,1}^{(T_2)}|}) = \widetilde{\Theta}(\frac{\sqrt{\eta/\eta_E}}{d^{3/2}\alpha_1^3})$. By this update, we can prove $T_{3,1} \leq T_2 + O(\frac{d^{3/2+1/8}\alpha_1^{O(1)}}{\eta})$. In order to do that, we can first see that for some $t'_{3,1} \in [T_2 + \widetilde{\Theta}(\frac{d^{3/2}\alpha_1^2\sqrt{\eta_E/\eta}}{\eta}), T_2 + \widetilde{\Theta}(\frac{d^{3/2}\alpha_1^4\sqrt{\eta_E/\eta}}{\eta})]$, we shall have $R_2^{(t'_{3,1})} \leq d^{-1/4}$. Indeed, suppose otherwise $R_2^{(t'_{3,1}-1)} \geq d^{-1/4}$, then (E.4) implies

$$\begin{aligned}
R_2^{(t'_{3,1})} &\leq R_2^{(t'_{3,1}-1)}(1 - \eta\zeta[R_2^{(t'_{3,1}-1)}]^{1/2}) \leq R_2^{(t'_{3,1}-1)}(1 - \eta\zeta\frac{1}{d^{1/8}}) \\
&\leq R_2^{(T_2)}\left(1 - \Theta(\frac{C_0 C_2^{1/2}\sqrt{\eta/\eta_E}}{C_1^{3/2}d^{3/2}\alpha_1^3})\frac{\eta}{d^{1/8}}\right)^{t'_{3,1}-T_2-1} \\
&\leq O(\sqrt{\eta_E/\eta})\left(1 - \Theta(\frac{C_0 C_2^{1/2}\sqrt{\eta/\eta_E}}{C_1^{3/2}d^{3/2}\alpha_1^3})\frac{\eta}{d^{1/8}}\right)^{t'_{3,1}-T_2-1}
\end{aligned}$$

which means there must exist an iteration $t'_{3,1} \in [T_2 + \widetilde{\Theta}(\frac{d^{3/2}\alpha_1^2\sqrt{\eta_E/\eta}}{\eta}), T_2 + \widetilde{\Theta}(\frac{d^{3/2}\alpha_1^4\sqrt{\eta_E/\eta}}{\eta})]$ such that $R_2^{(t'_{3,1}-1)} \geq d^{-1/4}$ (so the above update bound is still valid when the RHS is for $t \leq t'_{3,1} - 1$) and $R_2^{(t'_{3,1})} < d^{-1/4}$. Next we need to prove that at $t = t'_{3,1}$, it holds $C_1\alpha_2^6(B_{2,2}^{(t)})^6 \geq C_2[R_2^{(t)}]^3$. Let us discuss several possible cases:

1. Suppose $|B_{2,2}^{(t'_{3,1})}| \geq \frac{\alpha_1}{\alpha_2}|B_{2,1}^{(T_1)}|^{1/2} \geq \Theta(\frac{1}{d^{1/4}})$ (by Induction E.1a and Lemma E.8), then we already have $C_1\alpha_2^6(B_{2,2}^{(t'_{3,1})})^6 \geq C_2[R_2^{(t'_{3,1})}]^3$ and $T_{3,1} \leq t'_{3,1}$;
2. Suppose otherwise $|B_{2,2}^{(t'_{3,1})}| \leq \frac{\alpha_1}{\alpha_2}|B_{2,1}^{(T_1)}|^{1/2}$, then we shall have $\Sigma_{2,2}^{(t)} \leq O(\Sigma_{2,1}^{(t)})$. So the update of $R_2^{(t)}$ during $t \in [T_2, T_{3,1}]$ can be written as

$$R_2^{(t+1)} = R_2^{(t)} - \Theta(\eta\Sigma_{2,1}^{(t)})[R_2^{(t)}]^3 = R_2^{(t)}(1 - \Theta(\eta\zeta)[R_2^{(t)}]^{1/2})$$

Let $t'_{3,2} = \min\{t : R_2^{(t)} \leq 2d^{-1/4}\}$ be an iteration between T_2 and $t'_{3,1}$, we shall have

$$\sum_{t\in[t'_{3,2},t'_{3,1}]} \eta\zeta[R_2^{(t)}]^{3/2} = \Theta(R_2^{(t'_{3,2})} - R_2^{(t'_{3,1})}) = \Theta(\frac{1}{d^{1/4}}) \quad \text{and} \quad R_2^{(t)} \in [0.99\frac{1}{d^{1/4}}, 2.01\frac{1}{d^{1/4}}]$$

which also implies $t'_{3,1} - t'_{3,2} = \Theta(\frac{d^{1/8}}{\eta\zeta}) = \widetilde{\Theta}(\frac{d^{3/2+1/8}\alpha_1^3\sqrt{\eta_E/\eta}}{\eta})$. In this case, let us look at the update of $B_{2,2}^{(t)}$ at $t \in [T_2, T_3]$. By Lemma E.42, we have

$$B_{2,2}^{(t+1)} = B_{2,2}^{(t)} + \eta(1 \pm \widetilde{O}(\frac{1}{d}))\Lambda_{2,2}^{(t)}$$

It is not hard to see $|B_{2,2}^{(t)}|$ is monotonically increasing. Also by Induction E.1a and Lemma E.3a, if we sum together the update between $t'_{3,2}$ and $t'_{3,1}$ as follows: (suppose the sign of $B_{2,2}^{(t'_{3,2})}$ is positive for now, the negative case can be similarly dealt with)

$$\begin{aligned}
B_{2,2}^{(t'_{3,2})} + \sum_{t\in[t'_{3,2},t'_{3,1}]} \eta(1 \pm \widetilde{O}(\frac{1}{d}))\Lambda_{2,2}^{(t)} &= \sum_{t\in[t'_{3,2},t'_{3,1}]} \Theta(\frac{\eta C_0 C_1 \alpha_1^6 \alpha_2^6 (E_{2,1}^{(T_2)})^2}{\sqrt{C_2[R_2^{(t)}]^3 (C_1\alpha_1^6(E_{2,1}^{(T_2)})^2)^{3/2}}})(B_{2,2}^{(t)})^5 \\
&\geq B_{2,2}^{(t'_{3,2})} + (B_{2,2}^{(T_2)})^4 \sum_{t\in[t'_{3,2},t'_{3,1}]} \Theta(\frac{\eta C_0 \alpha_1^3 \alpha_2^6 B_{2,2}^{(t)}}{C_1^{1/2} C_2^{1/2} [R_2^{(t)}]^{3/2} |E_{2,1}^{(T_2)}|}) \\
&\geq B_{2,2}^{(t'_{3,2})} \prod_{t=t'_{3,2}}^{t'_{3,1}} \Big(1 + \eta\widetilde{\Theta}(\frac{\alpha_1^3\alpha_2^6}{d^{3/2+1/8}\sqrt{\eta_E/\eta}})\Big) \\
&\geq \widetilde{\Theta}(\frac{1}{\sqrt{d}})\Big(1 + \eta\widetilde{\Theta}(\frac{\alpha_1^3\alpha_2^6}{d^{3/2+1/8}\sqrt{\eta_E/\eta}})\Big)^{\widetilde{\Theta}(d^{3/2}\alpha_1^3\sqrt{\eta_E/\eta}/\eta)} \\
&\geq \Omega(e^{\alpha_1})
\end{aligned}$$

which is a contradiction to our assumption $|B_{2,2}^{(t'_{3,1})}| \leq \frac{\alpha_1}{\alpha_2}|B_{2,1}^{(T_1)}|^{1/2}$. Since $|B_{2,2}^{(t)}|$ is monotonically increasing, we know there must exist some iteration $t \leq t'_{3,1}$ such that $|B_{2,2}^{(t)}| \geq \frac{\alpha_1}{\alpha_2}|B_{2,1}^{(T_1)}|^{1/2}$, which means $T_{3,1} \leq t'_{3,1}$.

Thus we proved the bound of $T_{3,1} \leq T_2 + \widetilde{\Theta}(\frac{d^{3/2}\alpha_1^{O(1)}}{\eta})$.

Using similar arguments, we can prove that $R_2^{(T_{3,1})} \leq \frac{\alpha_1^{O(1)}}{d^{1/4}}$. Indeed, we can set $T_{3,3} := \min\{t : |B_{2,2}^{(t'_{3,1})}| \geq \frac{\alpha_1}{\alpha_2}|B_{2,1}^{(T_1)}|^{1/2}\}$. From our arguments in this proof, we know $\Sigma_{2,2}^{(t)} \leq O(\Sigma_{2,2}^{(t)})$ for $t \leq T_{3,3}$. Now we can further choose $t'_{3,3} = \min\{t : R_2^{(t)} \leq a\}$ for some $a = \frac{\alpha_1^{12}}{d^{1/4}}$ to be some iteration with $R_2^{(t)} \geq a$ for $t \in [T_2, t'_{3,3}]$ and $t'_{3,3} - T_2 = \Theta(\frac{\sqrt{a}\log d}{\eta\zeta})$. Now we can work out the update of $B_{2,2}^{(t)}$ during $t \in [T_2, t'_{3,3}]$ again to see that $|B_{2,2}^{(t'_{3,3})}| \leq B_{2,2}^{(T_2)}\Big(1 + \eta\widetilde{\Theta}(\frac{\alpha_1^3\alpha_2^6}{d^2 a^{3/2}\sqrt{\eta_E/\eta}})\Big)^{\frac{\sqrt{a}}{\eta\zeta}} \leq \widetilde{O}(\frac{1}{\sqrt{d}})$. This would prove that $t'_{3,3} \leq T_{3,3}$ and $R_2^{(T_{3,3})} \leq \frac{\alpha_1^{O(1)}}{d^{1/4}}$. So we also have $|B_{2,2}^{(T_{3,3})}| \leq \frac{\alpha_1^{O(1)}}{d^{1/4}}$ because of the definition of $T_{3,1}$. But since $T_{3,3} \geq T_{3,1}$ by our arguments above and the fact that $|B_{2,2}^{(t)}|$ is increasing, we shall have $|B_{2,2}^{(T_{3,1})}| \in [\frac{1}{d^{1/4}}, \frac{\alpha_1^{O(1)}}{d^{1/4}}]$. □

Now we proceed to characterize the learning of $B_{2,2}^{(t)}$ during $t \in [T_{3,1}, T_{3,2}]$.

Lemma E.10 (The growth of $B_{2,2}^{(t)}$ until T_3)**.** *Let $T_{3,1}, T_{3,2}$ be defined as in* (E.2)*. If Induction E.1 holds true for all $t \in [T_2, T_3]$, then we have $T_{3,2} = T_{3,1} + \widetilde{O}(\frac{d^{1/4}\alpha_1^{O(1)}}{\eta})$ and $T_3 \leq T_{3,2} + \widetilde{O}(\frac{\alpha_1^{O(1)}}{\eta})$.*

Proof. We first calculate the bound for $T_{3,2}$. After $T_{3,1}$, since $|B_{2,2}^{(t)}|$ is increasing while $R_2^{(t)}$ is decreasing by Induction E.1. So by Lemma E.3a, we have

$$[Q_2^{(t)}]^{-2} = \Theta(C_1\alpha_2^6(B_{2,2}^{(t)})^6), \quad \Phi_2^{(t)} = Q_2^{(t)}/[U_2^{(t)}]^{3/2} = \Theta((C_1^{3/2}\alpha_2^3\alpha_1^9|B_{2,2}^{(t)})|^3|E_{2,1}^{(t)}|^3)^{-1})$$

So according to Lemma E.4, we would have for all $t \in [T_{3,1}, T_{3,2})$:

$$\langle -\nabla_{w_2} L(W^{(t)}, E^{(t)}), v_2\rangle = (1 \pm o(1))\Lambda_{2,2}^{(t)} = \Theta(\frac{1}{C_1^{3/2}\alpha_1^9|E_{2,1}^{(T_2)}|^3})(B_{2,2}^{(t)})^2\mathrm{sign}(B_{2,2}^{(t)})$$

where we have used $(E_{2,1}^{(t)})^3 = \Theta((E_{2,1}^{(T_2)})^3)$ from Induction E.1a. So when $t \in [T_{3,1}, T_{3,2}]$, we can write down the explicit form of $\Lambda_{2,2}^{(t)}$ and use Lemma E.3d to derive

$$\begin{aligned}|B_{2,2}^{(t+1)}| &= |B_{2,2}^{(t)}| + \eta\Theta(\frac{C_1\alpha_1^6|E_{2,1}^{(T_2)}|^2}{C_1^{3/2}\alpha_1^9|E_{2,1}^{(T_2)}|^3})(B_{2,2}^{(t)})^2 \\ &\geq |B_{2,2}^{(t)}|\left(1 + \Theta(\frac{1}{C_1\alpha_1^{O(1)}})|B_{2,2}^{(T_{3,1})}|\right) \\ &\geq |B_{2,2}^{(t)}|\left(1 + \Theta(\frac{1}{C_1\alpha_1^{O(1)}})\frac{1}{d^{1/4}}\right)\end{aligned}$$

Thus after $\widetilde{O}(\frac{d^{1/4}\alpha^{O(1)}}{\eta})$ many iterations, we would have $|B_{2,2}^{(t)}| \geq \frac{1}{3}\min\{|E_{2,1}^{(t)}|, |B_{1,1}^{(t)}|\}$. Now let us deal with the growth of $|B_{2,2}^{(t)}|$ at $t \in [T_{3,2}, T_{3,3}]$. During this stage, since $B_{2,2}^{(t)}$ is still increasing and $|E_{2,1}^{(t)}| = |E_{2,1}^{(T_2)}|$ by Induction E.1, we have from Lemma E.3a that

$$\Phi_2^{(t)} = Q_2^{(t)}/[U_2^{(t)}]^{3/2} = \Theta(\frac{1}{C_1^2\alpha_2^{12}(B_{2,2}^{(t)})^{12}}) \geq \Theta(\frac{1}{C_1^2\alpha_1^{O(1)}})$$

And we can redo the calcualtions as above to get $T_3 \leq T_{3,2} + \widetilde{O}(\frac{\alpha_1^{O(1)}}{\eta})$ since $\sqrt{\eta/\eta_E}|E_{2,1}^{(t)}|$ and $|B_{1,1}^{(t)}|$ are both $\Theta(1)$ according to Induction E.1a,b ☐

Proving The Main Lemma. Now we finally begin to prove Lemma E.8.

Proof of Lemma E.8. We start with proving Induction E.1.

Proof of Induction E.1a: From Lemma E.5, we know the update of $B_{1,1}^{(t)}$ can be written as

$$B_{1,1}^{(t+1)} = B_{1,1}^{(t)} + \Theta(\eta\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3) \pm \eta O(\frac{(B_{1,2}^{(t)})^3}{(B_{2,2}^{(t)})^3} + \frac{1}{\sqrt{d}})\alpha_1^{O(1)}\Lambda_{2,2}^{(t)} + \frac{E_{2,1}^{(t)}}{B_{1,1}^{(t)}}\eta\Delta_{2,1}^{(t)} - \frac{B_{2,2}^{(t)}}{B_{1,1}^{(t)}}\eta\Lambda_{2,2}^{(t)}$$

Since from Lemma E.9 and Lemma E.10, we know $T_3 \leq \widetilde{O}(\frac{d^{1.625}\alpha_1^{O(1)}}{\eta})$ and from Claim B.1 and Induction E.1a,c we have $\Sigma_{1,1}^{(t)}[R_1^{(t)}]^3 \leq \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{2.25}})$, we shall have

$$\sum_{s\in[T_2,t)} \Theta(\eta\Sigma_{1,1}^{(s)}[R_1^{(s)}]^3) \leq \widetilde{O}(\frac{d^{1.625}\alpha_1^{O(1)}}{\eta})\widetilde{O}(\frac{\eta\alpha_1^{O(1)}}{d^{2.25}}) \leq \frac{1}{\sqrt{d}} = o(1)$$

Further more, by applying Lemma H.3 to $x_t = B_{2,2}^{(t)}$ with $q' = q - 2$, and notice that $\mathrm{sign}(B_{j,2}^{(t)}) = \mathrm{sign}(B_{j,2}^{(T_2)})$ for all $t \in [T_2, T_3]$, we also have

$$\left|\sum_{s\in[T_2,t)} O(\frac{(B_{1,2}^{(s)})^3}{(B_{2,2}^{(s)})^3})\alpha_1^{O(1)}\eta\Lambda_{2,2}^{(s)}\right| \leq \widetilde{O}(\frac{\alpha_1^{O(1)}}{\sqrt{d}})$$

Now we turn to the last two terms. We first see that from the expression (E.3) of $R_2^{(t)}$'s update, we have that (note that $\mathrm{sign}(E_{2,1}^{(t)}\Delta_{2,1}^{(t)}) = 1$)

$$\sum_{s\in[T_2,t)} \frac{E_{2,1}^{(s)}}{|B_{1,1}^{(s)}|}\eta\Delta_{2,1}^{(s)} = \sum_{s\in[T_2,t)} \frac{1}{|B_{1,1}^{(s)}|}\Theta(\eta\Sigma_{2,1}^{(s)}[R_2^{(s)}]^3) = \Theta(\frac{\sqrt{\eta_E/\eta}}{|B_{1,1}^{(T_2)}|}) = \Theta(\sqrt{\eta_E/\eta})$$

where we have used the fact that $\Sigma_{2,1}^{(t)}[R_2^{(t)}]^3 = (1 \pm O(\frac{1}{d}))E_{2,1}^{(t)}\Delta_{2,1}^{(t)}$ and $\sum_{s\in[T_2,t)}\eta\Sigma_{2,1}^{(s)}[R_2^{(s)}]^3 \lesssim R_2^{(T_2)}$ from (E.3) (which holds for all $t\in[T_2,T_3]$). And also, the analysis above shows that

$$|B_{1,1}^{(t)}| = |B_{1,1}^{(T_2)}| + O(\sqrt{\eta_E/\eta}) - \sum_{s\in[T_2,t]} \frac{B_{2,2}^{(s)}}{B_{1,1}^{(s)}}\eta\Lambda_{2,2}^{(s)}$$

for all $t\in[T_2,T_3]$, which means that either $\sum_{s\in[T_2,t]}\frac{B_{2,2}^{(s)}}{|B_{1,1}^{(s)}|}\eta\Lambda_{2,2}^{(s)} \le \sum_{s\in[T_2,t)}\frac{E_{2,1}^{(s)}}{|B_{1,1}^{(s)}|}\eta\Delta_{2,1}^{(s)}$ and we have $|B_{1,1}^{(t)}| \ge |B_{1,1}^{(T_2)}|$ holds throughout $t\in[T_2,T_3]$, or that $\sum_{s\in[T_2,t]}\frac{B_{2,2}^{(s)}}{|B_{1,1}^{(s)}|}\eta\Lambda_{2,2}^{(s)} \ge \Omega(\sqrt{\eta_E/\eta})$, in which case we would have $|B_{1,1}^{(t)}|$ to be actually decreasing (as $B_{2,2}^{(t)}$ is increasing). Now that since $B_{1,1}^{(T_2)} = \Theta(1)$, we can easily see by our definition of T_3 and the monotonicity of $B_{1,1}^{(t)}$ after going below $B_{1,1}^{(T_2)} - \Omega(\sqrt{\eta_E/\eta})$ that $B_{1,1}^{(t)} \ge 0.49B_{1,1}^{(T_2)} = \Omega(1)$ for all $t\in[T_2,T_3]$.

Next let us look at the change of $B_{2,1}^{(t)}$. From Lemma E.5, we can write down the update of $B_{2,1}^{(t)}$:

$$B_{2,1}^{(t+1)} = B_{2,1}^{(t)} + \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}})\eta\Phi_2^{(t)}[R_2^{(t)}]^3 \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{d})\eta\Lambda_{2,2}^{(t)} \pm \widetilde{O}(\frac{\eta\alpha_1^{O(1)}}{d^3})$$

For the first term, according to Lemma E.9 and Lemma E.10 and $R_2^{(t)} \le O(\sqrt{\eta_E/\eta}) = o(1)$ for all $t\in[T_2,T_3]$ by Induction E.1c, we have $\Phi_2^{(t)}[R_2^{(t)}]^3 \le \alpha_1^{O(1)}$ for all $t\in[T_2,T_3]$ and

$$\sum_{s\in[T_2,t]} \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}})\eta\Phi_2^{(s)}[R_2^{(s)}]^3 \le \widetilde{O}(\frac{d^{1.625}\alpha_1^{O(1)}}{\eta})\eta\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}}) \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{7/8}})$$

And similarly as in the proof of induction for $B_{1,1}^{(t)}$, we have

$$\sum_{s\in[T_2,t]} \widetilde{O}(\frac{\alpha_1^{O(1)}}{d})\eta\Lambda_{2,2}^{(s)} \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d}), \quad \sum_{s\in[T_2,t]} \widetilde{O}(\frac{\eta\alpha_1^{O(1)}}{d^3}) \le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d})$$

which proved the induction for $B_{2,1}^{(t)}$ since $|B_{2,1}^{(T_2)}| = \widetilde{\Theta}(\frac{1}{\sqrt{d}})$.
Next we go on for the induction of $B_{1,2}^{(t)}$, we write down its update:

$$B_{1,2}^{(t+1)} = B_{1,2}^{(t)} + \Theta(\frac{(B_{1,2}^{(t)})^2}{(B_{2,2}^{(t)})^2})E_{2,1}^{(t)}\eta\Lambda_{2,2}^{(t)} \pm \eta\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})|E_{2,1}^{(t)}|^2\Phi_2^{(t)} \pm \eta\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}})$$

By Lemma E.9 and Lemma E.10, we have for any $t\in[T_2,T_3]$

$$\sum_{s\in[T_2,t]} \eta\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^{5/2}}) \le \frac{1}{\sqrt{d}\,\mathsf{polylog}(d)}$$

and also

$$\begin{aligned}\sum_{s\in[T_2,t]} \eta\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})|E_{2,1}^{(t)}|^2\Phi_2^{(t)} &\le \left(\sum_{s\in[T_2,T_{3,1}]} + \sum_{s\in[T_{3,1},T_3]}\right)\eta\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})|E_{2,1}^{(t)}|^2\Phi_2^{(t)} \\ &\le \eta\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})\cdot(T_{3,1}-T_2)\cdot O(\alpha_1^{O(1)}d^{3/8}) + \eta\widetilde{O}(\frac{\alpha_1^{O(1)}}{d^4})(T_3 - T_{3,1}) \\ &\le \widetilde{O}(\frac{\alpha_1^{O(1)}}{d^2})\end{aligned}$$

Now we consider the term $\Theta(\frac{(B_{1,2}^{(t)})^2}{(B_{2,2}^{(t)})^2})E_{2,1}^{(t)}\eta\Lambda_{2,2}^{(t)}$, we have by Induction E.1a that

$$\left|\sum_{s\in[T_2,t]}\Theta(\frac{(B_{1,2}^{(t)})^2}{(B_{2,2}^{(t)})^2})E_{2,1}^{(t)}\eta\Lambda_{2,2}^{(t)}\right| \le O(\sqrt{\eta_E/\eta}(B_{1,2}^{(T_2)})^2)\sum_{s\in[T_2,t]}\eta\frac{|\Lambda_{2,2}^{(t)}|}{(B_{2,2}^{(t)})^2}$$

where we have used our induction hypothesis that $B_{1,2}^{(t)} = B_{1,2}^{(T_2)}(1\pm o(1))$. Using Lemma H.3 by setting $x_t = B_{2,2}^{(t)}$, $q' = 3$, and $A = \Theta(1) \ge d^{\Omega(1)}B_{2,2}^{(T_2)}$, it holds that

$$\left|\sum_{s\in[T_2,t]}\Theta(\frac{(B_{1,2}^{(t)})^2}{(B_{2,2}^{(t)})^2})E_{2,1}^{(t)}\eta\Lambda_{2,2}^{(t)}\right| \le O(\sqrt{\eta_E/\eta})\frac{(B_{1,2}^{(T_2)})^2}{|B_{2,2}^{(T_2)}|} \le O(\sqrt{\eta_E/\eta})\frac{(B_{1,2}^{(0)})^2}{|B_{2,2}^{(0)}|} \le \frac{1}{\sqrt{d}\mathsf{polylog}(d)}$$

where in the second inequality we have used Lemma C.13c, Lemma D.8a and Lemma C.1, and in the last our choice of $\eta_E/\eta \le \frac{1}{\mathsf{polylog}(d)}$. This ensures the induction can go on until $t = T_3$. And we finished our proof of Induction E.1a.

Proof of Induction E.1b: Let us write down the update of $E_{1,2}^{(t)}$ using Lemma E.7:

$$\begin{aligned}
E_{1,2}^{(t+1)} &= E_{1,2}^{(t)}(1-\eta_E\Xi_1^{(t)}) + \sum_{\ell\in[2]}\Theta(\eta_E\Sigma_{1,\ell}^{(t)})(-E_{1,2}^{(t)}[R_2^{(t)}]^3 \pm O(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}) + \sum_{\ell\in[2]}\eta_E\Delta_{1,\ell}^{(t)} \\
&= E_{1,2}^{(t)}(1-\eta_E\Xi_1^{(t)} - \sum_{\ell\in[2]}\Theta(\eta_E\Sigma_{1,\ell}^{(t)})[R_2^{(t)}]^3) + \widetilde{O}(\frac{\eta_E}{d^{3/2}})\Phi_1^{(t)}[R_1^{(t)}]^3 \\
&\quad \pm \widetilde{O}(\varrho+\frac{1}{\sqrt{d}})\sum_{\ell\in[2]}\eta_E\Sigma_{1,\ell}^{(t)}[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \\
&= E_{1,2}^{(t)}(1-\eta_E\Xi_1^{(t)} - \Theta(\eta_E\Sigma_{1,1}^{(t)})[R_2^{(t)}]^3) \pm \widetilde{O}(\varrho+\frac{1}{\sqrt{d}})\eta_E\Sigma_{1,1}^{(t)}[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}
\end{aligned}$$

where in the last inequality we have used $R_2^{(t)} \ge R_1^{(t)}$ from Induction E.1c and $\Sigma_{1,1}^{(t)} \ge \Omega(\Phi_1^{(t)})$, $\Sigma_{2,1}^{(t)} \le \widetilde{O}(\frac{1}{d^{3/2}})\Sigma_{1,1}^{(t)}$ from Claim B.1 and Induction E.1a. Now we can use the same analysis in the proof of Lemma D.8 on $E_{1,2}^{(t)}$ to prove the desired claim, which we do not repeat here.

As for $E_{2,1}^{(t)}$, we can obtain similar expressions:

$$\begin{aligned}
E_{2,1}^{(t+1)} &= E_{2,1}^{(t)}(1-\eta_E\Xi_2^{(t)} - \sum_{\ell\in[2]}\Theta(\eta_E\Sigma_{2,\ell}^{(t)})[R_1^{(t)}]^3) \\
&\quad \pm \widetilde{O}(\varrho+\frac{1}{\sqrt{d}})\sum_{\ell\in[2]}\Theta(\eta_E\Sigma_{2,\ell}^{(t)})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} + \sum_{\ell\in[2]}\eta_E\Delta_{2,\ell}^{(t)}
\end{aligned}$$

Now we can obtain bounds for each terms as

$$\sum_{s\in[T_2,t]}\sum_{\ell\in[2]}\Theta(\eta_E\Sigma_{2,\ell}^{(s)})[R_1^{(s)}]^3 \le \widetilde{O}(\frac{\eta_E\alpha_1^{O(1)}}{d^2})\cdot\widetilde{O}(\frac{d^{1.625}\alpha_1^{O(1)}}{\eta}) \le \frac{1}{d^{3/4}}$$

and by (E.3) in Lemma E.9, we also have for any $t\in[T_2,T_3]$

$$\begin{aligned}
\sum_{s\in[T_2,t]}\widetilde{O}(\varrho+\frac{1}{\sqrt{d}})\sum_{\ell\in[2]}\Theta(\eta_E\Sigma_{2,\ell}^{(s)})[R_1^{(s)}]^{3/2}[R_2^{(s)}]^{3/2} &\le \widetilde{O}(\varrho+\frac{1}{\sqrt{d}})\sum_{s\in[T_2,t]}\sum_{\ell\in[2]}\Theta(\eta_E\Sigma_{2,\ell}^{(s)})[R_2^{(s)}]^3 \\
&\le \widetilde{O}(\varrho+\frac{1}{\sqrt{d}})R_2^{(T_2)} \\
&\le \widetilde{O}(\varrho+\frac{1}{\sqrt{d}})
\end{aligned}$$

And also by using our induction and by (E.3) in Lemma E.9:

$$\sum_{s\in[T_2,t]}\sum_{\ell\in[2]}\eta_E\Delta_{2,\ell}^{(s)} \le \sum_{s\in[T_2,t]}\frac{\eta_E/\eta}{|E_{2,1}^{(t)}|}\Theta(\eta\Sigma_{2,1}^{(s)}+\eta\Sigma_{2,2}^{(s)})[R_2^{(s)}]^3 \le \frac{\eta_E/\eta}{|E_{2,1}^{(T_2)}|}R_2^{(T_2)} \le O(\frac{\eta_E/\eta}{\log d}) = o(\sqrt{\eta_E/\eta})$$

Finally, we can calculate

$$\sum_{s\in[T_2,t]} \eta_E \Xi_2^{(t)} E_{2,1}^{(t)} = \sum_{s\in[T_2,t]} \frac{\eta_E}{\eta} \frac{B_{2,2}^{(t)}}{E_{2,1}^{(t)}} \eta \Lambda_{2,2}^{(t)}$$

By resorting to the defintion of T_3 and go through similar analysis as for the induction of $B_{1,1}^{(t)}$, we can obtain that $|E_{2,1}^{(t)}|$ is either above $|E_{2,1}^{(T_2)}|(1+o(1))$ or is decreasing and always above $\frac{1}{2}|E_{2,1}^{(T_2)}|$. This proves Induction E.1b.

Proof of Induction E.1c: The proof of induction of $R_2^{(t)}$ is half done in Lemma E.9, we only need to complete the part when $t \in [T_{3,1}, T_3]$, since by (E.3), we always have $R_2^{(t)}$ to be decreasing by

$$R_2^{(t+1)} = R_2^{(t)} (1 - \sum_{\ell\in[2]} \Theta(\eta \Sigma_{2,\ell}^{(s)}) [R_2^{(t)}]^2)$$

And when $t \in [T_{3,1}, T_3]$, we have

$$\sum_{\ell\in[2]} \Theta(\eta \Sigma_{2,\ell}^{(s)} \leq \widetilde{O}(\eta d^{3/8+o(1)})$$

So if we suppose $R_2^{(T_3)} \leq \frac{1}{\sqrt{d}}$, we shall have for $T_3 - T_{3,1} = O(d^{1/4+o(1)}/\eta)$ many iterations that

$$R_2^{(t+1)} \geq R_2^{(T_{3,1})} (1 - \frac{\eta}{d^{5/8}})^{T_3 - T_{3,1}} \geq \Omega(R_2^{(T_{3,1})}) \geq \frac{1}{d^{1/4}} \qquad \text{(by Lemma E.9)}$$

So it negates our supposition, which completes the proof of the induction for $R_2^{(t)}$ in $t \in [T_2, T_3]$.

Now we turn to the proof of induction for $R_1^{(t)}$, we write down its update: (as in Lemma D.8)

$$\begin{aligned} R_1^{(t+1)} &= R_1^{(t)} - \Theta(\eta [R_1^{(t)}]^3) \Big(\Sigma_{1,1}^{(t)} + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)} (E_{2,1}^{(t)})^2 \Big) \\ &\quad \pm O\Big(\sum_{j,\ell} \eta \Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)} (\overline{R}_{1,2}^{(t)} + \varrho) [R_1^{(t)}]^{3/2} [R_2^{(t)}]^{3/2} \Big) \pm \frac{\eta}{\mathsf{poly}(d)} \end{aligned}$$

It is straightforward to derive

$$\sum_{\ell\in[2]} \Sigma_{1,\ell}^{(t)} |E_{1,2}^{(t)}| (\overline{R}_{1,2}^{(t)} + \varrho) [R_1^{(t)}]^{3/2} [R_2^{(t)}]^{3/2} \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})^2 \sum_{\ell\in[2]} \Sigma_{1,\ell}^{(t)} [R_1^{(t)}]^3 [R_2^{(t)}]^3$$

and when $t \in [T_2, T_{3,1}]$:

$$\begin{aligned} \sum_{s\in[T_2,t]} \sum_{\ell\in[2]} \eta \Sigma_{2,\ell}^{(s)} |E_{2,1}^{(s)}| (\overline{R}_{1,2}^{(s)} + \varrho) [R_1^{(s)}]^{3/2} [R_2^{(s)}]^{3/2} &\leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \frac{d^{o(1)} d^{3/8}}{d^{9/8}} \sum_{s\in[T_2,t]} \sum_{\ell\in[2]} \eta \Sigma_{2,\ell}^{(s)} [R_2^{(s)}]^3 \\ &\leq o(\frac{d^{o(1)}}{d^{3/4}}) \end{aligned}$$

and when $t \in [T_{3,1}, T_3]$:

$$\sum_{s\in[T_2,t]} \sum_{\ell\in[2]} \eta \Sigma_{2,\ell}^{(s)} |E_{2,1}^{(s)}| (\overline{R}_{1,2}^{(s)} + \varrho) [R_1^{(s)}]^{3/2} [R_2^{(s)}]^{3/2} \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \frac{d^{o(1)} d^{3/8}}{d^{9/8}} \eta \widetilde{O}(\frac{d^{1/4+o(1)}}{\eta}) \leq O(\frac{1}{d})$$

So these combined with Lemma D.8 proved that $R_1^{(t)} \leq O(\frac{d^{o(1)}}{d^{3/4}})$ for all $t \in [T_2, T_3]$. We can go through some similar analysis about $R_2^{(t)}$ to get that $R_1^{(t)} \geq \frac{1}{d}$ for all $t \in [T_2, T_3]$.

Finally we begin to prove the induction of $\overline{R}_{1,2}^{(t)}$. Similarly as in the proof of Lemma D.8, we first write down

$$\begin{aligned}
R_{1,2}^{(t+1)} &= R_{1,2}^{(t)} - \eta\langle\nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_2^{(t)}\rangle - \eta\langle\nabla_{w_2} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp} w_1^{(t)}\rangle \\
&\quad + \eta^2\langle\Pi_{V^\perp}\nabla_{w_1} L(W^{(t)}, E^{(t)}), \Pi_{V^\perp}\nabla_{w_2} L(W^{(t)}, E^{(t)})\rangle \\
&= R_{1,2}^{(t)} + \eta\Big(\Sigma_{1,1}^{(t)} + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}(E_{2,1}^{(t)})^2\Big)(-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_1^{(t)}]^{5/2}[R_2^{(t)}]^{1/2} \\
&\quad + \eta\Big(\Sigma_{1,1}^{(t)}\Theta(E_{1,2}^{(t)})^2 + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}\Big)(-\Theta(\overline{R}_{1,2}^{(t)}) \pm O(\varrho))[R_2^{(t)}]^{5/2}[R_1^{(t)}]^{1/2} \\
&\quad + O\Big(\sum_{(j,\ell)\neq(1,2)} \eta\Sigma_{j,\ell}^{(t)} E_{j,3-j}^{(t)}(R_1^{(t)}[R_2^{(t)}]^2 + R_2^{(t)}[R_1^{(t)}]^2)\Big) \pm \frac{\eta}{\mathsf{poly}(d)}
\end{aligned}$$

Note that since $|E_{1,2}^{(t)}| \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_2^{(t)}]^{3/2}[R_1^{(t)}]^{3/2}$ and $R_1^{(t)} \le O(\frac{1}{d^{3/4}})$, it holds

$$\begin{aligned}
\sum_{(j,\ell)\neq(1,2)} \eta\Sigma_{j,\ell}^{(t)}|E_{j,3-j}^{(t)}|R_2^{(t)}[R_1^{(t)}]^2 &\le \sum_{(j,\ell)\neq(1,2)} \eta\Sigma_{j,\ell}^{(t)}|E_{j,3-j}^{(t)}|R_1^{(t)}[R_2^{(t)}]^2 \\
&\le o\left(\Sigma_{1,1}^{(t)}[R_1^{(t)}]^2 + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}[R_2^{(t)}]^2\right)\widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_2^{(t)}]^{1/2}[R_1^{(t)}]^{1/2}
\end{aligned}$$

so the update becomes

$$\begin{aligned}
R_{1,2}^{(t+1)} = R_{1,2}^{(t)}&\left(1 - \eta\Theta\Big(\Sigma_{1,1}^{(t)} + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}(E_{2,1}^{(t)})^2\Big)[R_1^{(t)}]^2 - \eta\Theta\Big(\Sigma_{1,1}^{(t)}(E_{1,2}^{(t)})^2 + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}\Big)[R_2^{(t)}]^2\right) \\
&\pm \eta\widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{1/2}[R_2^{(t)}]^{1/2}\Theta\Big(\Sigma_{1,1}^{(t)} + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}(E_{2,1}^{(t)})^2\Big)[R_1^{(t)}]^2 \\
&\pm \eta\widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{1/2}[R_2^{(t)}]^{1/2}\Theta\Big(\Sigma_{1,1}^{(t)}(E_{1,2}^{(t)})^2 + \sum_{\ell\in[2]} \Sigma_{2,\ell}^{(t)}\Big)[R_2^{(t)}]^2
\end{aligned}$$

Now we can use the same arguments as in the proof of $\overline{R}_{1,2}^{(t)}$ in Lemma D.8 to conclude.

Proof of Lemma E.8a,b,c: Indeed, at the end of phase III:

$$\begin{aligned}
\text{Induction E.1a} &\implies \text{Lemma E.8a} \\
\text{Induction E.1b} &\implies \text{Lemma E.8c} \\
\text{Induction E.1c} &\implies \text{Lemma E.8b}
\end{aligned}$$

Now we have completed the whole proof. □

F The End Phase: Convergence

When we arrive at $t = T_3$, we have already obtained the representation we want for the encoder network $f(X)$, where v_1 and v_2 are satisfactorily learned by different neurons. In the last phase, we prove that such features are the solutions that the algorithm are converging to, which gives a stronger guarantee than just accidentally finding the solution at some intermediate steps.

To prove the convergence, we need to ensure all the good properties that we got through the training still holds. Fortunately, mosts of Induction E.1 still hold, as we summarized below:

Inductions F.1. *At the end phase, i.e. when $t \in [T_3, T]$, Induction E.1a continues to hold except that $|B_{2,2}^{(t)}| = \Theta(1)$, Induction E.1b will hold except that for $|E_{2,1}^{(t)}|$ only the upper bound still holds, and the upper bounds in Induction E.1c still hold while the lower bounds for $R_1^{(t)}, R_2^{(t)}$ is $1/\mathsf{poly}(d)$. Moreover, there is a constant $C = O(1)$ such that when $t \ge T_3 + \frac{\alpha_1^C}{\eta}$, we would have $|E_{2,1}^{(t)}| \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$.*

Now we present the main theorem of the paper, which we shall prove in this section.

Theorem F.2 (End phase: convergence)**.** *For some $T_4 = T_3 + \frac{d^{2+o(1)}}{\eta}$ and $T = \mathsf{poly}(d)/\eta$, we have for all $t \in [T_4, T]$ that Induction F.1 holds true and:*

(a) Successful learning of both v_1, v_2: $|B_{1,1}^{(t)}|, |B_{2,2}^{(t)}| = \Theta(1)$ while $|B_{2,1}^{(t)}|, |B_{1,2}^{(t)}| = \widetilde{O}(\frac{1}{\sqrt{d}})$.

(b) Successful denoising at the end: $R_j^{(t)} \le R_j^{(T_3)}(1 - \widetilde{\Theta}(\frac{1}{\alpha_j^6})[R_j^{(t)}]^2)$ for all $j \in [2]$.

(c) Prediction head is close to identity: $|E_{j,3-j}^{(t)}| \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_1^{(t)}]^{3/2}$ for all $j \in [2]$;

In fact, (b) and (c) also imply for some sufficiently large $t = \mathsf{poly}(d)/\eta$, it holds $R_j^{(t)} \le \frac{1}{\mathsf{poly}(d)}$ and $|E_{j,3-j}^{(t)}| \le \frac{1}{\mathsf{poly}(d)}$ for all $j \in [2]$.

And we have a simple corollary for the objective convergence.

Corollary F.3 (objective convergence, with prediction head)**.** *Let OPT denote the global minimum of the population objective* (B.1)*. It is easy to derive that $\mathsf{OPT} = 2 - 2\frac{C_0}{C_1} = \Theta(\frac{1}{\log d})$. We have for some sufficiently large $t \ge \mathsf{poly}(d)/\eta$:*

$$L(W^{(t)}, E^{(t)}) \le \mathsf{OPT} + \frac{1}{\mathsf{poly}(d)}$$

Now we need to establish some auxiliary lemmas:

Lemma F.4. *For some $t \in [T_3, \mathsf{poly}(d)/\eta]$, if Induction F.1 holds from T_3 to t, we have Lemma E.6 holds at t.*

Proof. Simple from similar calculations in the proof of Lemma E.6 . □

Lemma F.5. *For some $t \in [T_3, \mathsf{poly}(d)/\eta]$, if Induction F.1 holds from T_3 to t, we have for each $j \in [2]$ that*

$$\sum_{s\in[T_3,t]}\sum_{\ell\in[2]} \eta\Sigma_{j,\ell}^{(s)}[R_j^{(s)}]^3 \le O(R_j^{(T_3)}), \quad \forall j \in [2]$$

Proof. Notice that when Induction F.1 holds, we always have

$$\sum_{\ell\in[2]}(\Sigma_{j,\ell}^{(t)} + \Sigma_{3-j,\ell}^{(t)}(E_{3-j,j}^{(t)})^2) = (1 \pm o(1))\sum_{\ell\in[2]}\Sigma_{j,\ell}^{(t)}$$

we can use Lemma F.4 to obtain the update of $R_2^{(t)}$ as in the calculations when we obtained (E.3):

$$R_2^{(t)} = R_2^{(T_3)} - \sum_{s\in[T_3,t)}\sum_{\ell\in[2]}\Theta(\eta\Sigma_{2,\ell}^{(s)})[R_2^{(s)}]^3$$

which means that $R_2^{(t)}$ is decreasing from T_3 to t. Summing up the update, the part of $R_2^{(t)}$ is solved. For the part of $R_1^{(t)}$, we separately discuss when $|E_{2,1}^{(t)}|$ is larger than or smaller than $\widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$. When the former happens, which we know from Induction F.1 that it cannot last until some $t_4' = T_3 + \frac{\alpha_1^{O(1)}}{\eta}$ many iterations, we have for $t \in [T_3, t_4']$

$$\sum_{s\in[T_3,t)}\sum_{(j,\ell)\in[2]^2}\eta\Sigma_{j,\ell}^{(s)}|E_{j,3-j}^{(s)}|(\overline{R}_{1,2}^{(s)} + \varrho)[R_1^{(s)}]^{3/2}[R_2^{(s)}]^{3/2} \le \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})\frac{\alpha_1^{O(1)}}{d}R_1^{(T_3)} \le \frac{1}{d}R_1^{(T_3)}$$

Now for $t \geq t'_4$ we can simply go through similar calculations as in the proof of Induction E.1c to obtain

$$\begin{aligned}\sum_{s\in[t'_4,t)}\sum_{(j,\ell)\in[2]^2}\eta\Sigma_{j,\ell}^{(s)}|E_{j,3-j}^{(s)}|(\overline{R}_{1,2}^{(s)}+\varrho)[R_1^{(s)}]^{3/2}[R_2^{(s)}]^{3/2} &\leq \sum_{s\in[t'_4,t)}\widetilde{O}(\varrho+\frac{1}{\sqrt{d}})^2\sum_{(j,\ell)\in[2]^2}\eta\Sigma_{j,\ell}^{(s)}[R_1^{(s)}]^3[R_2^{(s)}]^3\\ &\leq \widetilde{O}(\varrho+\frac{1}{\sqrt{d}})^2R_2^{(T_3)}\max_{s\in[t'_4,t)}[R_1^{(s)}]^3\\ &\leq \frac{1}{d}R_1^{(T_3)}\end{aligned}$$

So by applying Lemma F.4a and Lemma E.6, we have

$$R_1^{(t)}=(1\pm o(1))R_1^{(T_3)}-\sum_{s\in[T_3,t)}\sum_{\ell\in[2]}\Theta(\eta\Sigma_{j,\ell}^{(s)})[R_1^{(s)}]^3$$

which proves the claim. □

Lemma F.6. *For some $t\in[T_3,\mathsf{poly}(d)/\eta]$, if Induction F.1 holds from T_3 to t. Then we have $|E_{j,3-j}^{(t)}|$ is decreasing until $|E_{j,3-j}^{(t)}|\leq O(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}+\widetilde{O}(\frac{1}{d^{3/2}})[R_j^{(t)}]^3$. Moreover, we have for each $t\in[T_3,T]$ that*

$$\left|\sum_{s\in[T_3,t]}\eta_E\Xi_j^{(t)}E_{j,3-j}^{(s)}\right|\leq|E_{j,3-j}^{(T_3)}|+\widetilde{O}(\varrho+\frac{1}{\sqrt{d}})\leq O(\sqrt{\eta_E/\eta})$$

Proof. We can go through the same calculations in the proof of Induction E.1b (using Fact E.2) to obtain

$$\begin{aligned}E_{j,3-j}^{(t+1)}&=E_{j,3-j}^{(t)}(1-\eta_E\Xi_j^{(t)})+\sum_{\ell\in[2]}\eta_E\Delta_{j,\ell}^{(t)}\\ &\quad+\sum_{\ell\in[2]}\Theta(\eta_E\Sigma_{j,\ell}^{(t)})(-E_{j,3-j}^{(t)}[R_{3-j}^{(t)}]^3\pm O(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2})\\ &=E_{j,3-j}^{(t)}(1-\eta_E\Xi_j^{(t)}-\eta_E\Theta(\Sigma_{j,j}^{(t)}[R_{3-j}^{(t)}]^3))+\widetilde{O}(\frac{1}{d^{3/2}})\sum_{\ell\in[2]}\eta_E\Sigma_{j,\ell}^{(t)}[R_j^{(t)}]^3\\ &\quad\pm O(\eta_E\Sigma_{j,j}^{(t)})(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}\end{aligned}$$

where we have used in the second equality that $\sum_{\ell\in[2]}\Delta_{j,\ell}^{(t)}\leq\widetilde{O}(\frac{1}{d^{3/2}})\sum_{\ell\in[2]}\Sigma_{j,\ell}^{(t)}[R_j^{(t)}]^3$ and also $\Sigma_{j,3-j}^{(t)}\leq O(\frac{1}{d^{3/2}})\Sigma_{j,j}^{(t)}$ for both $j\in[2]$ when Induction F.1 holds. Note that from above calculations, there exist a constant C such that if $|E_{j,3-j}^{(t)}|\geq C(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}+\sum_{\ell\in[2]}\eta_E\Delta_{j,\ell}^{(t)}$, we have $|E_{2,1}^{(t)}|$ to be decreasing. Now it suffices to observe that:

$$\begin{aligned}\sum_{s\in[T_3,t]}O(\eta_E\Sigma_{j,j}^{(t)})(\overline{R}_{1,2}^{(t)}+\varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}&\leq\sum_{s\in[T_3,t]}O(\eta_E\Sigma_{1,1}^{(t)}+\eta_E\Sigma_{2,2}^{(t)})(\overline{R}_{1,2}^{(t)}+\varrho)([R_1^{(t)}]^3+[R_2^{(t)}]^3)\\ &\leq\widetilde{O}(\varrho+\frac{1}{\sqrt{d}})\end{aligned}$$

which is from Induction F.1, Induction E.1c and Lemma F.4. Also note that $\Sigma_{j,j}^{(t)}[R_{3-j}^{(t)}]^3\leq O(\frac{d^{o(1)}}{d^{3/4}})\Xi_j^{(t)}$ at this stage, we have

$$E_{3-j,j}^{(t)}=E_{j,3-j}^{(T_3)}-\sum_{s\in[T_3,t)}\Xi_j^{(s)}E_{j,3-j}^{(s)}+\widetilde{O}(\varrho+\frac{1}{\sqrt{d}})$$

Recalling the expression of $\Xi_j^{(t)}$ finishes the proof. □

Lemma F.7. *Recall T_2 defined in* (D.1) *and T_3 defined in* (E.1)*, we have*

$$\sqrt{\eta/\eta_E}\max_{t\le T_3}|E_{2,1}^{(t)}| \le \sum_{t\le T_2}\frac{\eta\Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|}\mathcal{E}_{1,2}^{(t)} + \frac{1}{\alpha_1^{\Omega(1)}}$$

To prove this lemma, we need a simple claim.

Claim F.8. *If $\{x_t\}_{t<T}, x_t \ge 0$ is an increasing sequence and $C = \Theta(1)$ is a constant such that $x_{t+1} - x_t \le O(\eta)$ and $\sum_{t<T} x_t(x_{t+1} - x_t) = C$, then for each $\delta \in (\frac{1}{d}, 1)$ it holds $|x_T - \sqrt{C}| \le O(\delta^2 + x_0^2 + O(\frac{\log d}{d}))$.*

Proof. Indeed, for every $g \in 0, 1, \ldots$, we define $\mathcal{T}_g := \min\{t : x_t \ge (1+\delta)^g x_0\}$. and define $b := \min\{g : ((1+\delta)^g x_0)^2 \ge C - \delta^2\}$. Now for any $g < b$, we have

$$\sum_{t\in[\mathcal{T}_g,\mathcal{T}_{g+1}]} x_t(x_{t+1}-x_t) \ge x_{\mathcal{T}_g}(x_{\mathcal{T}_{g+1}} - x_{\mathcal{T}_g}) \ge (1+\delta)^g\delta(1+\delta)^{g-1}x_0^2 - \frac{1}{d} = \delta(1+\delta)^{2g-1}x_0^2 - \frac{1}{d}$$

By our definition of $\mathcal{T}_g$, we can further get

$$C = \sum_{t<T} x_t(x_{t+1}-x_t) = \sum_{g=1}^{b}\sum_{t\in[\mathcal{T}_g,\mathcal{T}_{g+1}]} x_t(x_{t+1}-x_t) \ge (1+\delta)^{2b}x_0^2 - x_0^2 - \frac{b}{d} \ge C - \delta^2 - x_0^2 - \frac{b}{d}$$

And also we have $C \le (\max_{t\le T} x_t)\sum_{t<T}(x_{t+1}-x_t) = x_T^2$, so we have $|x_T^2 - C| \le \delta^2 + x_0^2 + \frac{b}{d}$, where $b = O(\log(C)/\log(1+\delta)) \le O(\log d)$, which proves the claim. ☐

Proof of Lemma F.7. From the proof of Lemma D.8 and Lemma E.8 we know that

$$\max_{t\le T_3}|E_{2,1}^{(t)}| \le \sum_{t\le T_3}(1 \pm \frac{1}{\alpha_1^{\Omega(1)}})\eta_E|\Delta_{2,1}^{(t)}| + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$$

And since from the proof of Lemma D.8 we know that

$$\begin{aligned}
R_2^{(T_3)} &= R_2^{(0)} - \sum_{t\le T_3}(1 \pm \widetilde{O}(\frac{1}{d^{3/2}}))\eta\Sigma_{2,1}^{(t)}\mathcal{E}_{2,1}^{(t)} \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \\
&= (1 \pm \widetilde{O}(\frac{1}{d^{3/2}}))\sum_{t\le T_3} E_{2,1}^{(t)}\Delta_{2,1}^{(t)} \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})
\end{aligned}$$

We can define some alternative variables $\widetilde{E}_{2,1}^{(t)}$ updated as $\widetilde{E}_{2,1}^{(t+1)} = \widetilde{E}_{2,1}^{(t)} + \eta_E\Delta_{2,1}^{(t)}$ and $\widetilde{R}_2^{(t+1)} = \widetilde{R}_2^{(t)} - \widetilde{E}_{2,1}^{(t)}\Delta_{2,1}^{(t)}$. It is easy to see that $|E_{2,1}^{(t)} - \widetilde{E}_{2,1}^{(t)}| \le \frac{1}{\alpha_1^{\Omega(1)}}\max_{t\le T_3}|E_{2,1}^{(t)}|$. From above calculations, we know $\frac{\eta}{\eta_E}\sum_{t\in[T_1,T_3]}\widetilde{E}_{2,1}^{(t)}(\widetilde{E}_{2,1}^{(t+1)} - \widetilde{E}_{2,1}^{(t)}) = \widetilde{R}_2^{(T_1)} \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) + O(\frac{1}{d^{1/4}})$, which by Claim F.8 implies that

$$\sqrt{\eta/\eta_E}|\widetilde{E}_{2,1}^{(T_3)}| = \sqrt{\widetilde{R}_2^{(T_1)}} \pm O(\frac{1}{d^{1/4}}) = \sqrt{2} \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \pm O(\frac{1}{d^{1/4}})$$

And when we turn back, we shall have $\sqrt{\eta/\eta_E}\max_{t\le T_3}|E_{2,1}^{(t)}| \le \sqrt{2} + \frac{1}{\alpha_1^{\Omega(1)}}$. Now we can use similar techniques on $B_{1,1}^{(t)}$ and $R_1^{(t)}$. Indeed, from (D.4) and similar arguments in phase I, we know for all $t \in [T_1, T_2]$

$$R_1^{(t+1)} = R_1^{(0)} - \sum_{s\le t}(1 \pm \widetilde{O}(\frac{1}{d^{3/2}}))\eta\Sigma_{1,1}^{(s)}\mathcal{E}_{1,2}^{(s)} \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) \tag{F.1}$$

$$R_1^{(t+1)} \le R_1^{(t)}(1 - \widetilde{O}(\frac{\eta}{\alpha_1^6})[R_1^{(t)}]^2)$$

So one can obtain that at some iteration $t' = T_1 + O(\frac{d\alpha_1^{O(1)}}{\eta})$, we shall have $R_1^{(t)} \leq O(\frac{1}{\sqrt{d}})$ for all $t \geq t'$. Now let us consider the growth of $B_{1,1}^{(t)}$ before t', which clearly constitutes of

$$\begin{aligned}
B_{1,1}^{(t')} &= B_{1,1}^{(T_1)} + \sum_{t\in[T_1,t')} (\Lambda_{1,1}^{(t)} + \Gamma_{1,1}^{(t)} - \Upsilon_{1,1}^{(t)}) \\
&= B_{1,1}^{(T_1)} + \sum_{t\in[T_1,t')} \left(\frac{\eta\Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{1,2}^{(t)} \mathrm{sign}(B_{1,1}^{(t)}) + \eta\Gamma_{1,1}^{(t)} - \eta\Upsilon_{1,1}^{(t)} \right) \\
&= B_{1,1}^{(0)} + \sum_{t<t'} \frac{\eta\Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{1,2}^{(t)} \mathrm{sign}(B_{1,1}^{(t)}) + \sum_{t\in[T_1,t')} \eta \left(\Gamma_{1,1}^{(t)} - \Upsilon_{1,1}^{(t)} \right) + \widetilde{O}(\frac{1}{\sqrt{d}})
\end{aligned}$$

where the last one comes from the proof of Lemma C.13. Moreover by using the same arguments in the proof of Lemma D.8 we can easily prove that

$$\left| \sum_{t\in[T_1,t')} (\Gamma_{1,1}^{(t)} - \Upsilon_{1,1}^{(t)}) \right| \leq \widetilde{O}(\frac{1}{\sqrt{d}}) \quad \Longrightarrow \quad \sum_{t<t'} \frac{\eta\Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{1,2}^{(t)} \geq |B_{1,1}^{(t')}| - |B_{1,1}^{(0)}| - \widetilde{O}(\frac{1}{\sqrt{d}})$$

And for $t \in [t', T_2]$, we also have by (F.1) that

$$\sum_{t\in[t',T_2]} \frac{\eta\Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{1,2}^{(t)} \leq \sum_{t\in[t',T_2)} \eta\Sigma_{1,1}^{(t)} \mathcal{E}_{1,2}^{(t)} \leq O(\frac{1}{\sqrt{d}})$$

Recall $R_1^{(0)} = \sum_{t\in[0,t')} (1 \pm \widetilde{O}(\frac{1}{d^{3/2}}))\eta\Sigma_{1,1}^{(t)}\mathcal{E}_{1,2}^{(t)} \pm \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$ by (F.1) and $R_1^{(t)} \leq O(\frac{1}{\sqrt{d}})$ for $t \geq t'$. Now we can finally go through the same analysis using Claim F.8 on $B_{1,1}^{(t)}$ and $R_1^{(t)}$ during $t \in [0, t']$ as above to obtain that

$$\sum_{t\leq T_2} \frac{\eta\Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{1,2}^{(t)} \geq (1 - \widetilde{O}(\frac{1}{d^{3/2}}))\sqrt{R_1^{(0)}} - \widetilde{O}(\frac{1}{\sqrt{d}}) = 1 - \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$$

Combining the results, we finishes the proof. □

Now we are prepared to prove Theorem F.2.

F.1 Proof of Convergence

Proof of Theorem F.2. First we start with the $B_{j,\ell}^{(t)}$s. Indeed, we can go through similar calculations to see that all gradients $\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_\ell \rangle$ can be decomposed into

$$\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_\ell \rangle = (\Lambda_{j,\ell}^{(t)} - \Upsilon_{j,\ell,1}^{(t)}) + (\Gamma_{j,\ell}^{(t)} - \Upsilon_{j,\ell,2}^{(t)})$$

where $\Lambda_{j,\ell}^{(t)} - \Upsilon_{j,\ell,1}^{(t)}$ and $\Gamma_{j,\ell}^{(t)} - \Upsilon_{j,\ell,2}^{(t)}$ can be expressed as

$$\begin{aligned}
\Lambda_{j,\ell}^{(t)} - \Upsilon_{j,\ell,1}^{(t)} &= C_0\alpha_2^6 C_1\alpha_1^6 \Phi_j^{(t)} (B_{j,\ell}^{(t)})^5 \left(E_{j,3-j}^{(t)} (B_{3-j,3-\ell}^{(t)})^3 (B_{j,3-\ell}^{(t)})^3 + (E_{j,3-j}^{(t)})^2 (B_{3-j,3-\ell}^{(t)})^6 \right) \\
&\quad - C_0\alpha_2^6 C_1\alpha_1^6 \Phi_j^{(t)} (B_{j,\ell}^{(t)})^2 (B_{3-j,\ell}^{(t)})^3 E_{j,3-j}^{(t)} \left((B_{j,3-\ell}^{(t)})^6 + E_{j,3-j}^{(t)} (B_{3-j,3-\ell}^{(t)})^3 (B_{j,3-\ell}^{(t)})^3 \right) \\
&\quad + C_0\alpha_2^6 \Phi_j^{(t)} (B_{j,\ell}^{(t)})^5 C_2 \mathcal{E}_{j,3-j}^{(t)} \\
\Gamma_{j,\ell}^{(t)} - \Upsilon_{j,\ell,2}^{(t)} &= C_0\alpha_2^6 C_1\alpha_1^6 \Phi_{3-j}^{(t)} (B_{3-j,\ell}^{(t)})^3 (B_{j,\ell}^{(t)})^2 E_{3-j,j}^{(t)} \left(E_{3-j,j}^{(t)} (B_{j,3-\ell}^{(t)})^3 (B_{3-j,3-\ell}^{(t)})^3 + (E_{3-j,j}^{(t)})^2 (B_{j,3-\ell}^{(t)})^6 \right) \\
&\quad - C_0\alpha_2^6 C_1\alpha_1^6 \Phi_{3-j}^{(t)} (B_{j,\ell}^{(t)})^5 (E_{3-j,j}^{(t)})^2 \left((B_{3-j,3-\ell}^{(t)})^6 + E_{3-j,j}^{(t)} (B_{j,3-\ell}^{(t)})^3 (B_{3-j,3-\ell}^{(t)})^3 \right) \\
&\quad + C_0\alpha_2^6 \Phi_{3-j}^{(t)} E_{3-j,j}^{(t)} (B_{3-j,\ell}^{(t)})^3 (B_{j,\ell}^{(t)})^2 C_2 \mathcal{E}_{3-j,j}^{(t)}
\end{aligned}$$

Firstly, for all the terms that contain factors of $(B_{j,\ell}^{(t)})^2 (B_{3-j,\ell}^{(t)})^2$ (or $(B_{j,\ell}^{(t)})^2 (B_{j,3-\ell}^{(t)})^2$), we can apply Lemma F.6, our Induction F.1 assumption and $|E_{j,3-j}^{(t)}| \leq O(1), \forall t \in [T_3, T]$ to obtain that their (multiplicated by η) summation over $t \in [T_3, T]$ is absolutely bounded by $\widetilde{O}(\frac{1}{d})$. So we can move on to deal with all other terms. When $j = \ell$, Using Lemma F.6, we have

$$\sum_{t\in[T_3,T]} \eta C_0 \alpha_2^6 C_1 \alpha_1^6 \Phi_j^{(t)} |B_{j,\ell}^{(t)}|^5 (E_{j,3-j}^{(t)})^2 (B_{3-j,3-\ell}^{(t)})^6 = \sum_{t\in[T_3,T]} \frac{\eta \Xi_j^{(t)}}{|B_{j,\ell}^{(t)}|} (E_{j,3-j}^{(t)})^2$$
$$\leq \sqrt{\frac{\eta}{\eta_E}} |E_{j,3-j}^{(T_3)}| + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) = O(1)$$

And the sign of LHS is $\mathrm{sign}(B_{j,\ell}^{(t)})$. Moreover, for $j = \ell = 1$, from Lemma F.7 and Lemma F.6 we also have

$$\sum_{t\in[T_3,T]} \eta C_0 \alpha_2^6 C_1 \alpha_1^6 \Phi_2^{(t)} |B_{1,1}^{(t)}|^5 (E_{2,1}^{(t)})^2 (B_{2,2}^{(t)})^6 \leq \sqrt{\frac{\eta}{\eta_E}} \left| \sum_{t\in[T_3,T]} \eta_E \Xi_j^{(t)} E_{j,3-j}^{(t)} \right|$$
$$\leq \sqrt{\frac{\eta}{\eta_E}} |E_{2,1}^{(T_3)}| + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})$$
$$\leq \sum_{t\leq T_2} \frac{\eta \Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{1,2}^{(t)} + \frac{1}{\alpha_1^{\Omega(1)}}$$

Since we have

$$B_{1,1}^{(T_2)} = \sum_{s\leq T_2} \frac{\eta \Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{1,2}^{(t)} + \sum_{s\leq T_2} \frac{\eta \Sigma_{2,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{2,1}^{(t)} - \sum_{t\in[T_3,T]} \frac{\eta \Xi_j^{(t)}}{|B_{j,\ell}^{(t)}|} (E_{j,3-j}^{(t)})^2$$

And since by Induction D.1 we have $|B_{1,1}^{(t)}| = \Theta(1)$ during $t \in [T_1, T_2]$ and $\sum_{t\in[T_1,T_2]} \eta \Sigma_{2,1}^{(t)} \geq R^{(T_1)} - o(1) = \sqrt{2} - o(1)$. For all the other terms in the gradient , we can apply Lemma F.6, our Induction F.1 assumption and $|E_{j,3-j}^{(t)}| \leq O(1)$ so we have for $t \in [T_3, T]$

$$|B_{1,1}^{(t)}| = \sum_{s\leq T_2} \frac{\eta \Sigma_{1,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{1,2}^{(t)} + \sum_{s\leq T_2} \frac{\eta \Sigma_{2,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{2,1}^{(t)} - \sum_{t\in[T_3,T]} \frac{\eta \Xi_j^{(t)}}{|B_{j,\ell}^{(t)}|} (E_{j,3-j}^{(t)})^2 - o(1)$$
$$\geq \sqrt{\eta/\eta_E} \max_{t\leq T_3} |E_{2,1}^{(t)}| + \sum_{s\leq T_2} \frac{\eta \Sigma_{2,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{2,1}^{(t)} - \sqrt{\frac{\eta}{\eta_E}} |E_{j,3-j}^{(T_3)}| + \widetilde{O}(\varrho + \frac{1}{\sqrt{d}}) - o(1)$$
$$\geq \sum_{s\leq T_2} \frac{\eta \Sigma_{2,1}^{(t)}}{|B_{1,1}^{(t)}|} \mathcal{E}_{2,1}^{(t)} - o(1) \geq \Omega(1)$$

which also proved $|B_{1,1}^{(t)}| = O(1)$ since all the terms on the RHS are absolutely $O(1)$ bounded. Since one can see from Lemma F.6 that $|E_{2,1}^{(t)}|$ is decreasing before it reaches $\frac{1}{d}$). Moreover this proves $\sqrt{\eta/\eta_E}|E_{2,1}^{(t)}| \leq B_{1,1}^{(t)}$ for all $t \in [T_3, T]$, and also the fact that

$$B_{1,1}^{(t)} \geq \Omega(1), \quad \forall t \in [T_3, T]$$

The case of $B_{2,2}^{(t)}$ is much more simple as $E_{1,2}^{(t)} \leq \widetilde{O}(\frac{1}{d})$ throughout $t \in [T_3, T]$ by Lemma F.6 and Lemma E.8c, Now we can go through the similar calculations again to obtain that $B_{2,2}^{(t)} = \Theta(1)$ for all $t \in [T_3, T]$. When $j \neq \ell$, all the terms calculated in the expansion of $\Lambda_{j,\ell}^{(t)} - \Upsilon_{j,\ell,1}^{(t)}$ and $\Gamma_{j,\ell}^{(t)} - \Upsilon_{j,\ell,2}^{(t)}$ contain factors of $(B_{2,1}^{(t)})^2 = \widetilde{O}(\frac{1}{d})$ or $(B_{1,2}^{(t)})^2 = \widetilde{O}(\frac{1}{d})$. So we can similarly use Lemma F.6 as before to derive that $B_{j,3-j}^{(t)} = B_{j,3-j}^{(T_3)}(1 \pm \widetilde{O}(\frac{\alpha_1^{O(1)}}{\sqrt{d}}))$ for all $t \in [T_3, T]$ and $j \in [2]$.

As for the prediction head, the induction of $E_{1,2}^{(t)}$ follows from exactly the same proof in Lemma E.8. The part of $E_{2,1}^{(t)}$ is half done in Lemma F.6. It suffices to notice that $\Xi_2^{(t)} = \widetilde{\Theta}(\frac{\alpha_1^6}{\alpha_2^6})$ and if

$|E_{2,1}^{(t)}| \geq C(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$ for some $C = O(1)$, then

$$\begin{aligned} E_{2,1}^{(t+1)} &= E_{2,1}^{(t)}(1 - \eta_E \Xi_2^{(t)} - \eta_E \Theta(\Sigma_{2,2}^{(t)}[R_1^{(t)}]^3)) + \widetilde{O}(\frac{1}{d^{3/2}}) \sum_{\ell \in [2]} \eta_E \Sigma_{2,\ell}^{(t)}[R_2^{(t)}]^3 \\ &\quad \pm O(\eta_E \Sigma_{2,2}^{(t)})(\overline{R}_{1,2}^{(t)} + \varrho)[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \\ &\leq E_{2,1}^{(t)}(1 - \widetilde{\Theta}(\frac{\eta \alpha_1^6}{\alpha_2^6})) \end{aligned}$$

So after $\frac{\alpha_1^{O(1)}}{\eta}$ many epochs will we have

$$|E_{2,1}^{(t)}| \leq (\log d)|\overline{R}_{1,2}^{(t)} + \varrho|[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2} \leq \widetilde{O}(\varrho + \frac{1}{\sqrt{d}})[R_1^{(t)}]^{3/2}[R_2^{(t)}]^{3/2}$$

as desired. And the rest of the induction of $E_{2,1}^{(t)}$ is the same as in the induction arguments of $E_{1,2}^{(t)}$ in Lemma E.8.

The induction of $R_1^{(t)}, R_2^{(t)}$ and $R_{1,2}^{(t)}$ is exactly the same as those in the proof of Lemma E.8 except here we only need $R_1^{(t)}/R_2^{(t)} \in [\frac{1}{\alpha_1^{O(1)}}, \alpha_1^{O(1)}]$ after T_4. Indeed, from the update of $R_j^{(t)}$ (which can be easily worked out), we have

$$R_j^{(t+1)} = R_j^{(t)}(1 - \Theta(\eta \Sigma_{j,j}^{(t)})[R_j^{(t)}]^2) = R_j^{(t)}(1 - \widetilde{\Theta}(\frac{\eta}{\alpha_j^6})[R_j^{(t)}]^2)$$

Now after $\frac{d^2 \alpha_1^{O(1)}}{\eta}$ many epochs, we can obtain from similar arguments in Lemma E.8 that $R_1^{(t)}/R_2^{(t)} \in [\frac{1}{\alpha_1^{O(1)}}, \alpha_1^{O(1)}]$ and $R_j^{(t)} \leq \frac{1}{d}$. The induction can go on untill $t = \mathsf{poly}(d)/\eta$.

For the convergence of $B_{1,1}^{(t)}$ and $B_{2,2}^{(t)}$ after $t = T_4$, notice that their change depends on $\sum_{t \geq T_4} \frac{E_{j,3-j}^{(t)}}{B_{j,j}^{(t)}} \Xi_j^{(t)}$, which stays very small after T_4, we have that $|B_{j,j}^{(t)} - B_{j,j}^{(T_4)}| \leq o(1)$ for all $j \in [2]$. This finishes the whole proof. □

G Learning Without Prediction Head

When we do not use prediction head in the network architecture, the analysis is much simpler. We can reuse most of the gradient calculations in previous sections as long as we set $E^{(t)}$ to the identity. Note that here we allow $m \geq 1$ to be any positive integer.

Theorem G.1 (learning without the prediction head)**.** *Let m be any positive integer. If we keep $E^{(t)} \equiv I_m$ during the whole training process, then for all $t \in [\widetilde{\Omega}(\frac{d^2}{\eta}), \mathsf{poly}(d)/\eta]$, we shall have $|B_{j,1}^{(t)}| = \Theta(1)$, $|B_{j,2}^{(t)}| = \widetilde{O}(\frac{1}{\sqrt{d}})$ and $R_j^{(t)} = O(\frac{1}{d^{1-o(1)}})$ for all $j \in [m]$ with probability $1 - o(1)$. Moreover, for a longer training time $t = \mathsf{poly}(d)/\eta$, we would have $R_j^{(t)} \leq \frac{1}{\mathsf{poly}(d)}$ for all $j \in [m]$.*

Moreover, it is direct to obtain a objective convergence result similar to Corollary F.3.

Corollary G.2 (objective convergence, without prediction head)**.** *Let OPT denote the global minimum of the population objective* (B.1)*. When trained with $E^{(t)} \equiv I_m$, we have for some sufficiently large $t \geq \mathsf{poly}(d)/\eta$:*

$$L(W^{(t)}, I_m) \leq \mathsf{OPT} + \frac{1}{\mathsf{poly}(d)}$$

Proof of Theorem G.1. The proof is easy to obtain since it is very similar to some proofs in previous sections, and we only sketch it here. Indeed, using the calculations in Lemma E.5 and Lemma E.4 and set $E_{i,j}^{(t)}, i \neq j \in [m]$ to zero. We shall have (note that here $\mathcal{E}_{j,r}^{(t)} \equiv \mathcal{E}_j^{(t)}$ for any $r \neq j$)

$$\langle -\nabla_{w_j} L(W^{(t)}, E^{(t)}), v_\ell \rangle = C_0 C_2 \alpha_\ell^6 (B_{j,\ell}^{(t)})^5 \Phi_j^{(t)} \mathcal{E}_j^{(t)} = \Theta(C_0 C_2 \alpha_\ell^6 \Phi_j^{(t)} (B_{j,\ell}^{(t)})^5 [R_j^{(t)}]^3)$$

Now we can go through the similar induction arguments as in the proof of Lemma C.13 (with TPM lemma to distinguish the learning speed) to obtain that for each $j \in [m]$:

$$|B_{j,1}^{(t)}| = \Theta(1), \quad |B_{j,2}^{(t)}| = |B_{j,2}^{(0)}|(1 \pm o(1)), \quad \forall j \in [m] \qquad (\text{when } t \geq \frac{d^2}{\eta})$$

When this is proven, we can also reuse the calculations as in the proof of Lemma D.5 to obtain that

$$R_j^{(t+1)} = R_j^{(t)}(1 - \Theta(\eta \Sigma_{j,1}^{(t)})[R_j^{(t)}]^2) = R_j^{(t)}(1 - \Theta(\eta C_0 C_2 \alpha_1^6 \Phi_j^{(t)} (B_{j,1}^{(t)})^6 [R_j^{(t)}]^2), \quad \forall j \in [m]$$

So again after some $t = \widetilde{O}(\frac{d^2}{\eta})$, we shall have $R_j^{(t)} \leq O(\frac{d^{o(1)}}{d})$. While the decrease of $R_j^{(t)}$ is happening, we can make induction that $|B_{j,2}^{(t)}| = |B_{j,2}^{(0)}|(1 \pm o(1))$, since if it holds for all previous iterations before t, then

$$\begin{aligned}\sum_{s \leq t-1} \eta |\langle -\nabla_{w_j} L(W^{(s)}, E^{(s)}), v_2 \rangle| &= \sum_{s \leq t-1} \eta C_0 \alpha_2^6 \Phi_j^{(s)} |B_{j,2}^{(s)}|^5 C_2 \mathcal{E}_j^{(s)} \\ &\overset{①}{\leq} \frac{1}{\mathsf{polylog}(d)} |B_{j,2}^{(0)}|\end{aligned}$$

where ① is due to Corollary H.2, where $x_t = |B_{j,1}^{(t)}|$ and $y_t = |B_{j,2}^{(t)}|$ and $S_t \leq \frac{1}{\mathsf{polylog}(d)}$, $y_0 \leq O(\log d) x_0$. which finishes the proof. $\square$

H Tensor Power Method Bounds

In this section, we give two lemmas related to the tensor power method that can help us in previous sections' proofs.

Lemma H.1 (TPM, adapted from [3]). *Consider an increasing sequence $x_t \geq 0$ defined by $x_{t+1} = x_t + \eta C_t x_t^q$ for some integer $q \geq 3$ and $C_t > 0$, and suippose for some $A > 0$ there exist $t' \geq 0$ such that $x_{t'} \geq A$. Then for every $\delta > 0$, and every $\eta \in (0, 1)$:*

$$\sum_{t \geq 0, x_t \leq A} \eta C_t \geq \left(\frac{\delta(1+\delta)^{-1}}{(1+\delta)^{q-1} - 1} \left(1 - \left(\frac{(1+\delta)x_0}{A} \right)^{q-1} \right) - \frac{O(\eta A^q)}{x_0} \frac{\log(A/x_0)}{\log(1+\delta)} \right) \cdot \frac{1}{x_0^{q-1}}$$

$$\sum_{t \geq 0, x_t \leq A} \eta C_t \leq \left(\frac{(1+\delta)^{q-1}}{q-1} + \frac{O(\eta A^q)}{x_0} \frac{\log(A/x_0)}{\log(1+\delta)} \right) \cdot \frac{1}{x_0^{q-1}}$$

This lemma has a corollary:

Corollary H.2 (TPM, from [3]). *Let $q \geq 3$ be a constant and $x_0, y_0 = o(1)$ and $A = O(1)$. Let $\{x_t, y_t\}_{t \geq 0}$ be two positive sequences updated as*

- $x_{t+1} = x_t + \eta C_t x_t^q$ *for some $C_t > 0$;*
- $y_{t+1} = y_t + \eta S_t C_t y_t^q$ *for some $S_t > 0$.*

Suppose $x_0 \geq y_0 (\max_{t: x_t \leq A} S_t)^{\frac{1}{q-1}} (1 + \frac{1}{\mathsf{polylog}(d)})$, then $y_t \leq \widetilde{O}(y_0)$ for all t such that $x_t \leq A$. Moreover, if $x_0 \geq y_0 (\max_{t: x_t \leq A} S_t)^{\frac{1}{q-1}} \log(d)$, we would have $|y_t - y_0| \lesssim \frac{|y_0|}{\mathsf{polylog}(d)}$.

Moreover, we prove the following lemma for comparing the updates of different variables.

Lemma H.3 (TPM of different degrees). *Consider an increasing sequences $x_t \geq 0$ defined by $x_{t+1} = x_t + \eta C_t x_t^q$, for some integer $q > q' \geq 3$ and $q' \leq q - 2$, and $C_t > 0$, and further suppose given $A = O(1)$, there exists $t' \geq 0, x_{t'} \geq A$. Then for every $\delta > 0$ and every $\eta \in (0, 1)$:*

$$\sum_{t \geq 0, x_t \leq A} \eta C_t x_t^{q'} \leq (1+\delta)^{q'} \left(O(1) + \eta b A^q \right) \frac{1}{x_0^{q-q'-1}}$$

$$\sum_{t \geq 0, x_t \leq A} \eta C_t x_t^{q'} \geq (1+\delta)^{-q'} \left(\delta(1+\delta)^{-1} \frac{1 - (1+\delta)^{-b(q-q'-1)}}{1 - (1+\delta)^{-(q-q'-1)}} - \eta b A^q \right) \frac{1}{x_0^{q-q'-1}}$$

where $b = \Theta(\log(A/x_0)/\log(1+\delta))$. *When* $A = x_0 d^{\Theta(1)}$, $\eta = o(\frac{1}{A^q \delta})$ *and* $q = O(1)$*, then*

$$\sum_{t \geq 0, x_t \leq A} \eta C_t x_t^{q'} = \Theta(\frac{1}{x_0^{q-q'-1}})$$

Proof. For every $g \in 0, 1, \ldots$, we define $\mathcal{T}_g := \min\{t : x_t \geq (1+\delta)^g x_0\}$. and define $b := \min\{g : (1+\delta)^g \geq A\}$, we can write down the following two inequalities according to the update of x_t:

$$\sum_{t \in [\mathcal{T}_g, \mathcal{T}_{g+1}]} \eta C_t [(1+\delta)^g x_0]^q \leq (1+\delta) x_{\mathcal{T}_g} - x_{\mathcal{T}_g} + \eta A^q \leq \delta (1+\delta)^g x_0 + \eta A^q$$

$$\sum_{t \in [\mathcal{T}_g, \mathcal{T}_{g+1}]} \eta C_t [(1+\delta)^{g+1} x_0]^q \geq (1+\delta) x_{\mathcal{T}_g} - x_{\mathcal{T}_g} - \eta A^q \geq \delta (1+\delta)^g x_0 - \eta A^q$$

where $g+1 \leq b$. Dividing both sides by $[(1+\delta)^g x_0]^{q-q'}$ in the first inequality and $[(1+\delta)^{g+1} x_0]^{q-q'}$ in the second, we have

$$\sum_{t \in [\mathcal{T}_g, \mathcal{T}_{g+1}]} \eta C_t [(1+\delta)^g x_0]^{q'} \leq \frac{\delta}{(1+\delta)^{g(q-q'-1)}} \frac{1}{x_0^{q-q'-1}} + \frac{\eta A^q}{x_0^{q-q'-1}}$$

$$\sum_{t \in [\mathcal{T}_g, \mathcal{T}_{g+1}]} \eta C_t [(1+\delta)^{g+1} x_0]^{q'} \geq \frac{\delta (1+\delta)^{-1}}{(1+\delta)^{(g+1)(q-q'-1)}} \frac{1}{x_0^{q-q'-1}} - \frac{\eta A^q}{x_0^{q-q'-1}}$$

Therefore if we sum over $g = 0, \ldots, b$, then

$$\begin{aligned}
\sum_{t \geq 0, x_t \leq A} \eta C_t x_t^{q'} &\leq \sum_{t \geq 0, x_t \leq A} \eta C_t [(1+\delta)^{g+1} x_0]^{q'} \\
&= (1+\delta)^{q'} \sum_{t \geq 0, x_t \leq A} \eta C_t [(1+\delta)^g x_0]^{q'} \\
&\leq (1+\delta)^{q'} \sum_{0 \leq g \leq b} \left(\frac{\delta}{(1+\delta)^{g(q-q'-1)}} \frac{1}{x_0^{q-q'-1}} + \frac{\eta A^q}{x_0^{q-q'-1}} \right) \\
&= (1+\delta)^{q'} O\left(\frac{\delta}{(1+\delta)^{q-q'-1} - 1} + \eta b A^q \right) \frac{1}{x_0^{q-q'-1}} \\
&\leq (1+\delta)^{q'} O\left(\frac{1}{q-q'-1} + \eta b A^q \right) \frac{1}{x_0^{q-q'-1}}
\end{aligned}$$

For the lower bound, we also have

$$\begin{aligned}
\sum_{t \geq 0, x_t \leq A} \eta C_t x_t^{q'} &\geq (1+\delta)^{-q'} \sum_{t \geq 0, x_t \leq A} \eta C_t [(1+\delta)^{g+1} x_0]^{q'} \\
&\geq (1+\delta)^{-q'} \sum_{0 \leq g \leq b} \left(\frac{\delta (1+\delta)^{-1}}{(1+\delta)^{(g+1)(q-q'-1)}} - \eta A^q \right) \frac{1}{x_0^{q-q'-1}} \\
&= (1+\delta)^{-q'} \left(\delta (1+\delta)^{-1} \frac{1 - (1+\delta)^{-b(q-q'-1)}}{1 - (1+\delta)^{-(q-q'-1)}} - \eta b A^q \right) \frac{1}{x_0^{q-q'-1}} \\
&= (1+\delta)^{-q'} \left(\delta (1+\delta)^{-1} \frac{1 - (1+\delta)^{-b(q-q'-1)}}{1 - (1+\delta)^{-(q-q'-1)}} - \eta b A^q \right) \frac{1}{x_0^{q-q'-1}}
\end{aligned}$$

Inserting $b = \Theta(\log(A/x_0)/\log(1+\delta))$ proves the lower bound. For the last one we can choose $\delta = \frac{1}{\sqrt{\log d}}$ to get:

$$b = \Theta(\mathsf{polylog}(d)), \quad \frac{\delta(1 - (1+\delta)^{-b(q-q'-1)})}{1 - (1+\delta)^{-(q-q'-1)}} = \Omega(1), \quad (1+\delta)^{-q'} = \Omega(1),$$

which proves the claim. $\square$