UNDERSTANDING AND MITIGATING THE OVERCORRECTION IN MACHINE UNLEARNING VIA FEATURE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to remove the influence of particular data points from a pre-trained model in response to growing privacy and ethical concerns. While gradient-ascent (GA)-based unlearning provides a simple and widely used strategy, its optimization dynamics remain poorly understood. In this work, we identify a striking non-monotonic behavior in GA-based unlearning, which we call the overcorrection phenomenon: the forgetting quality evaluated only on the forgetting data first improves and then deteriorates as unlearning proceeds. Specifically, the Kolmogorov-Smirnov (KS) distance between the truth-ratio distributions of the unlearned model and the retrained model first decreases and then increases, indicating that the unlearned model initially approaches the retrained reference but later deviates from it. To explain this phenomenon, we provide a three-stage theoretical characterization under a two-layer convolutional neural network model. Our analysis specifies when the forgetting quality enters a favorable regime, how long it remains there, and how it eventually deteriorates. The theoretical results further reveal the key role of the learning rate, motivating a simple learning-rate adjustment strategy that mitigates overcorrection. Experiments on both synthetic and real-world datasets support our theoretical predictions and demonstrate the effectiveness of the proposed strategy.

1 INTRODUCTION

150unlearning (MU) aims to remove the influence of designated data on an already trained model (Shaik et al., 2024). Originally motivated by data governance requirements such as the right to be forgotten (Hoofnagle et al., 2019), MU has evolved into an important paradigm for enhancing the trustworthiness of modern machine learning systems. In computer vision, MU has been studied for eliminating the effects of backdoor poisoning attacks (Liu et al., 2022a), mitigating undesirable biases (Oesterling et al., 2024), and refining pre-training pipelines to improve representation transferability (Jain et al., 2023; Jia et al., 2023). Its importance is further amplified in generative models and large language models, where memorization and reproduction of training content (Carlini et al., 2021) raise serious concerns about harmful content generation (Gandikota et al., 2023), privacy leakage (Huang et al., 2022), and copyright infringement (Eldan & Russinovich, 2023).

A direct way to perform machine unlearning is to retrain the model from scratch (Guo et al., 2020; Chen et al., 2022), which is commonly regarded as the gold standard (Thudi et al., 2022a;b). However, this strategy is often computationally prohibitive for modern deep networks, especially when models are large or deletion requests arrive repeatedly. To reduce this cost, approximate unlearning methods aim to efficiently edit a trained model so that its behavior approaches that of the retrained reference model. Representative approaches include influence unlearning (Izzo et al., 2021; Warnecke et al., 2021), Fisher forgetting (Golatkar et al., 2020), and finetuning-based methods (Jang et al., 2023; Wang et al., 2023). Among them, finetuning-based unlearning (Chen & Yang, 2023; Yao & Xu, 2024; Liu et al., 2024; Li et al., 2024b) has become particularly attractive due to its simplicity and broad applicability. A common strategy is gradient ascent (Golatkar et al., 2020; Jang et al., 2023; Yao & Xu, 2024; Chen & Yang, 2023; Maini et al., 2024), which increases the prediction loss on the forgetting set and is motivated by the intuition of approximately reversing the original gradient descent update.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we identify a striking and previously under-explored phenomenon in GA-based machine unlearning. Specifically, the forgetting quality (Maini et al., 2024; Zhang et al., 2024) on the forgetting data, measured by the Kolmogorov-Smirnov (KS) distance between the truth-ratio distributions of the unlearned model and the retrained model on the forgetting set, first improves and then deteriorates during unlearning: the KS distance first decreases and then increases. In other words, the unlearned model initially becomes closer to the retrained model on the forgetting data, but later deviates from it as unlearning continues. We refer to this non-monotonic behavior as the **overcorrection phenomenon**. Notably, related notions such as over-forgetting have also been discussed in Hatua et al. (2024); He et al. (2025), but mainly in terms of their adverse effects on the retained data. In contrast, our overcorrection phenomenon highlights a distinct failure mode on the forgetting data itself: excessive unlearning can degrade the forgetting quality. This naturally raises a question:

How does the overcorrection phenomenon occur and how can it be mitigated?

In this work, we seek to answer the above question by providing a three-stage theoretical analysis of the overcorrection phenomenon.

1.1 Main contributions

We identify and analyze the overcorrection phenomenon in GA-based machine unlearning, where the forgetting quality on the forgetting data first improves and then deteriorates, and develop both a theoretical explanation and a practical mitigation strategy.

Empirically, we first present findings that demonstrate the overcorrection phenomenon. We train a ResNet-18 model on Tiny ImageNet and perform unlearning on 10% of the classes. Following Zhang et al. (2024), we evaluate forgetting quality by the KS distance between the truth-ratio distributions of the unlearned model f^u and the retrained model f^r on the forgetting data. A smaller KS distance indicates better forgetting quality, whereas a larger value indicates worse forgetting quality. As shown in Figure 1, the KS distance first decreases and then increases during unlearning, suggesting that the forgetting quality first improves and then deteriorates. The goal of this work is to rigorously characterize the overcorrection phenomenon *in an idealized theoretical model*.

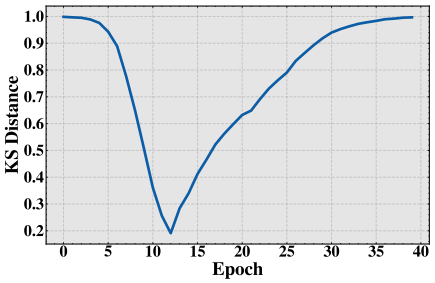


Figure 1: **Non-monotonic forgetting quality on Tiny ImageNet**. The KS distance between the truth-ratio distributions of the unlearned model and the retrained model first decreases and then increases.

Theoretically, we provide a three-stage characterization of overcorrection in a setting with two binary classification tasks, where the goal is to unlearn one of them. We study the dynamics of a two-layer convolutional neural network and characterize when the forgetting quality enters a favorable regime, how long it remains there, and how it eventually leaves this regime as unlearning continues. These theoretical results are summarized in the following theorem.

Theorem 1.1 (Informal)**.** *Under suitable conditions on the network width m, the forgetting-set size n_f, the learning rate η, the signal strength $\|\boldsymbol{\mu}_f\|_2$, and the noise scale $\sigma_p\sqrt{d}$, suppose that the initial model before unlearning satisfies loss $L_S(\mathbf{W}^b) \leq \epsilon$. Then, with high probability, the overcorrection dynamics can be characterized by the following three stages: (i)* ***First stage.*** *The forgetting quality first enters a favorable regime after $t_1 = \Theta(\frac{n_f m}{\eta \sigma_p^2 d \epsilon})$. (ii)* ***Second stage.*** *The forgetting quality remains in the favorable regime for $t_2 - t_1 = \Theta(\frac{m}{C\eta\|\boldsymbol{\mu}_f\|_2^2})$. (iii)* ***Third stage.*** *As unlearning continues beyond t_2, the forgetting quality gradually deteriorates.*

We make the following remarks on our main results.

- **Three Stages of the Overcorrection Phenomenon.** The theorem characterizes the three-stage dynamics of overcorrection. First, after $t_1 = \Theta(\frac{n_f m}{\eta \sigma_p^2 d \epsilon})$ iterations, the unlearned model becomes sufficiently close to the retrained model, and the forgetting quality enters a favorable regime. Second, this favorable regime lasts for $t_2 - t_1 = \Theta(\frac{m}{C\eta\|\boldsymbol{\mu}_f\|_2^2})$ iterations. Finally, the unlearned model gradually deviates from the retrained model, causing the forgetting quality to deteriorate.

- **A Learning-rate Adjustment Strategy.** The theorem highlights the key role of the learning rate in overcorrection: a relatively large learning rate can accelerate entry into the favorable regime, while reducing the learning rate in the second stage can help keep the model near the retrained reference and mitigate overcorrection. Motivated by this insight, we propose a simple learning-rate adjustment strategy and empirically evaluate its effectiveness in Section 5.

2 Problem Setup

We primarily focus on a setting involving two binary classification tasks, where the objective is to unlearn one of them. In this section, we introduce the notation, the data model, the neural network, GA-based MU, and KL-divergence metric used in our analysis.

Notation. We use lowercase letters for scalars, lowercase boldface letters for vectors, and uppercase boldface letters for matrices. The ℓ_2 norm of a vector $\mathbf{v}$ is denoted by $\|\mathbf{v}\|_2$. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_2$ to denote its spectral norm. For two sequences $\{a_k\}$ and $\{b_k\}$, we write $a_k = O(b_k)$ if there exists an absolute constant $C > 0$ such that $|a_k| \leq C|b_k|$. We write $a_k = \Omega(b_k)$ if $b_k = O(a_k)$, and $a_k = \Theta(b_k)$ if both $a_k = O(b_k)$ and $a_k = \Omega(b_k)$ hold. We also use $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$, and $\tilde{\Theta}(\cdot)$ to hide logarithmic factors. Moreover, we write $x_k = \mathrm{poly}(y_k)$ if $x_k = O(y_k^D)$ for some positive constant D. Finally, for any positive integer n, we denote $[n] = \{1, 2, \ldots, n\}$.

Data model. We extend the signal-noise data model in Cao et al. (2022); Kou et al. (2023b); Wan et al. (2025) to the following setting tailored to machine unlearning.

Definition 2.1. *Let $\boldsymbol{\mu}_f, \boldsymbol{\mu}_r \in \mathbb{R}^d$ be fixed vectors representing the signal components contained in the data points of the forgetting dataset S_f and the retain dataset S_r respectively, and let the two signal vectors be orthogonal. Each data point in the forgetting dataset S_f is denoted by $(\mathbf{x}_f, y_f)$, where $\mathbf{x}_f = [\mathbf{x}_f^{(1)\top}, \mathbf{x}_f^{(2)\top}]^\top \in \mathbb{R}^{2d}$ and $y_f \in \{-1, 1\}$. Similarly, each data point in the retain dataset S_r is denoted by $(\mathbf{x}_r, y_r)$, where $\mathbf{x}_r = [\mathbf{x}_r^{(1)\top}, \mathbf{x}_r^{(2)\top}]^\top \in \mathbb{R}^{2d}$ and $y_r \in \{-1, 1\}$. Both datasets are generated according to the following distributional assumptions:*

1. *Label y_f, y_r are generated as Rademacher random variables, i.e. $\mathbb{P}[y_f = 1] = \mathbb{P}[y_f = -1] = \mathbb{P}[y_r = 1] = \mathbb{P}[y_r = -1] = 1/2$.*
2. *Noise vectors $\boldsymbol{\xi}_f, \boldsymbol{\xi}_r$ are generated from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$.*
3. *One of $\mathbf{x}_f^{(1)}, \mathbf{x}_f^{(2)}$ is assigned the signal component $y_f \boldsymbol{\mu}_f$, while the other is assigned the noise vector $\boldsymbol{\xi}_f$. Similarly, for the retained data, one of $\mathbf{x}_r^{(1)}, \mathbf{x}_r^{(2)}$ is given by the signal $y_r \boldsymbol{\mu}_r$, and the other is given by the noise $\boldsymbol{\xi}_r$.*

The data distribution consists of label-dependent signal components and label-independent noise components. It is motivated by image data, where only a subset of patches contains class-relevant information, while the remaining patches are irrelevant to the label. The orthogonality of the two signal vectors is mainly for analytical convenience.

Neural network. We consider a two-layer CNN model with the ReLU activation function $\sigma(z) = \max\{0, z\}$. Given an input $\mathbf{x}$, the network output is expressed as $f(\mathbf{W}, \mathbf{x}) = F_{+1}(\mathbf{W}_{+1}, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$, where the second-layer weights are fixed as either $+1/m$ or $-1/m$. Specifically, the terms $F_{+1}(\mathbf{W}_{+1}, \mathbf{x})$ and $F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$ are defined as:

$$
\begin{aligned}
F_j(\mathbf{W}_j, \mathbf{x}) &= \frac{1}{m} \sum_{s=1}^{m} \big[\sigma(\langle \mathbf{w}_{j,s}, \mathbf{x}^{(1)} \rangle) + \sigma(\langle \mathbf{w}_{j,s}, \mathbf{x}^{(2)} \rangle)\big] \\
&= \frac{1}{m} \sum_{s=1}^{m} \big[\sigma(\langle \mathbf{w}_{j,s}, y\boldsymbol{\mu} \rangle) + \sigma(\langle \mathbf{w}_{j,s}, \boldsymbol{\xi} \rangle)\big],
\end{aligned}
$$

where m is the network width, and $\mathbf{w}_{j,s}$ represents the s-th first-layer neuron associated with the sign j. We use $\mathbf{W}$ to denote the collection of all trainable parameters.

GA-based MU. We first train an initial model f^b on the full dataset $S = \{(\mathbf{x}_{f,i_f}, y_{f,i_f})\}_{i_f=1}^{n_f} \cup \{(\mathbf{x}_{r,i_r}, y_{r,i_r})\}_{i_r=1}^{n_r}$, which contains both the forgetting and retained examples. Specifically, f^b is obtained by minimizing the logistic loss.

$$L_S(\mathbf{W}^b) = \frac{\sum_{i_f=1}^{n_f} \ell[y_{f,i_f} f(\mathbf{W}^b, \mathbf{x}_{f,i_f})] + \sum_{i_r=1}^{n_r} \ell[y_{r,i_r} f(\mathbf{W}^b, \mathbf{x}_{r,i_r})]}{n_f + n_r}, \tag{2.1}$$

where $\ell(z) = \log(1 + \exp(-z))$, and $f(\mathbf{W}^b, \mathbf{x})$ is the two-layer ReLU CNN defined above. The model is trained by gradient descent from Gaussian initialization, where each entry of $\mathbf{W}^{b,0}$ is independently sampled from $\mathcal{N}(0, \sigma_0^2)$.

In the machine unlearning stage, the unlearned model f^u is naturally initialized from the final weights of the initial model, i.e., $\mathbf{w}_{j,s}^{u,0} = \mathbf{w}_{j,s}^{b,t_{end}}$. We then construct the unlearned model f^u by performing gradient ascent to maximize the following objective

$$L_{S_f}(\mathbf{W}^u) = \frac{1}{n_f} \sum_{i_f=1}^{n_f} \ell[y_{f,i_f} f(\mathbf{W}^u, \mathbf{x}_{f,i_f})], \tag{2.2}$$

thereby degrading the model's classification performance on the forgetting data. The gradient ascent update of the filters in the CNN can be written as

$$\begin{aligned} \mathbf{w}_{j,s}^{u,t+1} &= \mathbf{w}_{j,s}^{u,t} + \eta \cdot \nabla_{\mathbf{w}_{j,s}} L_{S_f}(\mathbf{W}^{u,t}) \\ &= \mathbf{w}_{j,s}^{u,t} + \frac{\eta}{n_f m} \sum_{i_f=1}^{n_f} \ell_{f,i_f}'^{u,t} \big[\sigma'(\langle \mathbf{w}_{j,s}^{u,t}, y_{f,i_f} \boldsymbol{\mu}_f \rangle) j \boldsymbol{\mu}_f + \sigma'(\langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle) j y_{f,i_f} \boldsymbol{\xi}_{f,i_f}\big] \end{aligned} \tag{2.3}$$

for $j \in \{\pm 1\}$ and $s \in [m]$, where we denote the derivative of the loss by $\ell_{f,i_f}'^{u,t} = \ell'[y_{f,i_f} \cdot f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})]$ and explicitly set $\sigma'(0) = 1$ for the ReLU activation. The goal of machine unlearning is for the unlearned model f^u to approximate the behavior of the retrained model f^r, which is obtained by training from scratch solely on the retain set S_r using gradient descent.

KL divergence. To analyze the overcorrection phenomenon during unlearning, we follow the spirit of Zhang et al. (2024) and use KL divergence as a tractable surrogate for the KS distance between the truth-ratio distributions of f^u and f^r on the forgetting set:

$$\mathrm{D}_{\mathrm{KL}}(f^u \| f^r) = \frac{1}{n_f} \sum_{i_f=1}^{n_f} \Big[p_{i_f}^u \log \frac{p_{i_f}^u}{p_{i_f}^r} + (1 - p_{i_f}^u) \log \frac{1 - p_{i_f}^u}{1 - p_{i_f}^r} \Big], \tag{2.4}$$

where $p_{i_f}^u = q(y_{f,i_f} f^u(\mathbf{x}_{f,i_f}))$, $p_{i_f}^r = q(y_{f,i_f} f^r(\mathbf{x}_{f,i_f}))$, and $q(z) = 1/[1 + \exp(-z)]$.

3 Main Results

In this section, we provide a theoretical characterization of the overcorrection phenomenon, showing when the KL divergence enters a small regime, and how long it remains there. These results are established under the following conditions.

Condition 3.1. *Suppose there exists a sufficiently large constant C, such that the following hold:*

1. *Dimension d is sufficiently large: $d \geq C \max\big\{n_f \sigma_p^{-2} \|\boldsymbol{\mu}_f\|_2^2 \log(T^*), n_r \sigma_p^{-2} \|\boldsymbol{\mu}_r\|_2^2 \log(T^*), n_f^2 \log(n_f m/\delta)(\log(T^*))^2, n_r^2 \log(n_r m/\delta)(\log(T^*))^2\big\}$.*
2. *Training sample size $n_f \geq C \log(m/\delta), n_r \geq C \log(m/\delta)$; neural network width $m \geq C \max\big\{\log(n_f/\delta), \log(n_r/\delta)\big\}$.*
3. *The norm $\|\boldsymbol{\mu}_f\|_2^2 \geq C \cdot \sigma_p^2 \log(\max\{n_f, n_r\}/\delta), \|\boldsymbol{\mu}_r\|_2^2 \geq C \cdot \sigma_p^2 \log(\max\{n_f, n_r\}/\delta)$.*
4. *The learning rate $\eta \leq \big(C \max\big\{\sigma_p^2 d^{3/2}/(n_f^2 m \sqrt{\log(n_f/\delta)}), \sigma_p^2 d^{3/2}/(n_r^2 m \sqrt{\log(n_r/\delta)}), \sigma_p^2 d/n_f, \sigma_p^2 d/n_r\big\}\big)^{-1}$.*

5. *The standard deviation of Gaussian initialization σ_0 is appropriately chosen such that $\sigma_0 \leq \big(C\max\{\sigma_p d/\sqrt{n_f}, \sigma_p d/\sqrt{n_r}, \sqrt{\log(m/\delta)}\cdot\|\boldsymbol{\mu}_f\|_2, \sqrt{\log(m/\delta)}\cdot\|\boldsymbol{\mu}_r\|_2\}\big)^{-1}$.*

Here, $T^* = \eta^{-1}\mathrm{poly}(\epsilon^{-1}, d^{-1}, n, m)$ with $n = \max\{n_f, n_r\}$.

The conditions on d, m, n_f and n_r ensure a sufficiently over-parameterized regime and provide certain statistical properties needed for the training data and random initialization. The conditions on η and σ_0 guarantee that gradient descent effectively minimize the training loss. Similar conditions can be found in Chatterji & Long (2021); Frei et al. (2022); Cao et al. (2022); Kou et al. (2023b).

Before presenting our main theoretical results, we emphasize that they build on the properties of the initial model f^b established in Kou et al. (2023b). Specifically, for any $\epsilon > 0$, under Condition 3.1, if $n_f\|\boldsymbol{\mu}_f\|_2^4 \geq C_1\sigma_p^4 d$ and $n_r\|\boldsymbol{\mu}_r\|_2^4 \geq C_1\sigma_p^4 d$, then there exists $t_{end} = \tilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$ such that training loss $L_S(\mathbf{W}^{b,t_{end}}) \leq \epsilon$ and test error $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{b,t_{end}}) \leq \exp(-n\|\boldsymbol{\mu}\|_2^4/(C_2\sigma_p^4 d))$. Here, C_1 is the same constant as in the result of Kou et al. (2023b). Now we present the main results in the following theorem.

Theorem 3.2. *Under Condition 3.1, if $n_f\|\boldsymbol{\mu}_f\|_2^4 \geq C_1\sigma_p^4 d$ and $n_r\|\boldsymbol{\mu}_r\|_2^4 \geq C_1\sigma_p^4 d$, then with probability at least $1-\delta$ there exists $t_1 = \Theta(\frac{n_f m}{\eta\sigma_p^2 d\epsilon})$ and $t_2 = t_1 + \Theta(\frac{m}{C\eta\|\boldsymbol{\mu}_f\|_2^2})$ such that*

1. *When $t_1 \leq t \leq t_2$, the KL divergence $\mathrm{D}_{\mathrm{KL}}(f^{u,t}\|f^r) \leq 2^{-1}C^{-2}$.*

2. *When $t > t_2$, the KL divergence $\mathrm{D}_{\mathrm{KL}}(f^{u,t}\|f^r) \to \frac{1}{n_f}\sum_{i_f=1}^{n_f}\log\Big(1+\exp\big(y_{f,i_f}f^r(\mathbf{x}_{f,i_f})\big)\Big)$.*

Here, $\frac{1}{n_f}\sum_{i_f=1}^{n_f}\log(1+\exp(y_{f,i_f}f^r(\mathbf{x}_{f,i_f})))$ is a positive constant, and C is a sufficiently large constant as in Condition 3.1.

Theorem 3.2 shows that the KL divergence between the unlearned model f^u and the retrained model f^r first enters a small regime, where $\mathrm{D}_{\mathrm{KL}}(f^{u,t}\|f^r) \leq 2^{-1}C^{-2}$, after $t_1 = \Theta(\frac{n_f m}{\eta\sigma_p^2 d\epsilon})$, indicating good forgetting quality. It then remains in this small regime for $t_2 - t_1 = \Theta(\frac{m}{C\eta\|\boldsymbol{\mu}_f\|_2^2})$, before eventually leaving the small regime and converging to a positive constant as unlearning continues, which indicates degraded forgetting quality.

4 Proof Roadmap

This section outlines the main technical ideas to prove Theorem 3.2. The complete proofs can be found in Appendix E, F, G.

4.1 Signal-noise decomposition

Based on the gradient ascent update rule in (2.3), the weight vector $\mathbf{w}_{j,s}^{u,t}$ can be expressed as a linear combination of its initialization $\mathbf{w}_{j,s}^{u,0}$, the signal vector $\boldsymbol{\mu}_f$, and the noise components $\{\boldsymbol{\xi}_{f,i_f}\}_{i_f=1}^{n_f}$. Therefore, for $s \in [m]$, the weight vector can be decomposed as such:

$$\mathbf{w}_{j,s}^{u,t} = \mathbf{w}_{j,s}^{u,0} + j\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)\|\boldsymbol{\mu}_f\|_2^{-2}\boldsymbol{\mu}_f + \sum_{i_f=1}^{n_f}\rho_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})\|\boldsymbol{\xi}_{f,i_f}\|_2^{-2}\boldsymbol{\xi}_{f,i_f}. \tag{4.1}$$

Here, $\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)$ and $\rho_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$ represent the relevant coefficients. We refer to (4.1) as the signal-noise decomposition of $\mathbf{w}_{j,s}^{u,t}$, where the factors $\|\boldsymbol{\mu}_f\|_2^{-2}$ and $\|\boldsymbol{\xi}_{f,i_f}\|_2^{-2}$ are introduced for normalization. In particular, $\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)$ captures the signal component, while $\rho_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$ characterizes the noise memorization during unlearning. The decomposition reduces the analysis of CNN unlearning dynamics to tracking the evolution of these coefficients, forming the foundation of our results. Further denote $\overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) := \rho_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})\mathbb{1}(\rho_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) \geq 0)$ and $\underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) := \rho_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})\mathbb{1}(\rho_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) \leq 0)$. Combining the decomposition in (4.1) with the update rule in (2.3), the iteration of coefficients are given:

Lemma 4.1. *The coefficients* $\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f), \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}), \underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$ *in decomposition* (4.1) *satisfy the following equations:*

$$\gamma_{j,s}^{u,t+1}(\boldsymbol{\mu}_f) = \gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \cdot \sum_{i_f=1}^{n_f} \ell'^{u,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,t}, y_{f,i_f} \cdot \boldsymbol{\mu}_f \rangle) \cdot \|\boldsymbol{\mu}_f\|_2^2, \tag{4.2}$$

$$\overline{\rho}_{j,s,i_f}^{u,t+1}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{n_f m} \cdot \ell'^{u,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = j), \tag{4.3}$$

$$\underline{\rho}_{j,s,i_f}^{u,t+1}(\boldsymbol{\xi}_{f,i_f}) = \underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta}{n_f m} \cdot \ell'^{u,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = -j) \tag{4.4}$$

for all $s \in [m]$, $j \in \{\pm 1\}$ *and* $i_f \in [n_f]$.

Lemma 4.1 characterizes the coefficient dynamics induced by the gradient ascent update, thereby providing a basis for tracking the optimization trajectory of GA-based machine unlearning. The complete proof can be found in Appendix E.2.

4.2 Decoupling with a Three-stage Analysis

First stage. In this stage, we characterize how long it takes for the KL divergence to enter a small regime, i.e., $\mathrm{D}_{\mathrm{KL}}(f^u \| f^r) \leq 2^{-1}C^{-2}$. We first identify a sufficient condition that certifies this small-KL regime through the following lemma.

Lemma 4.2. *Suppose that* $|y_{f,i_f} f^u(\mathbf{x}_{f,i_f}) - y_{f,i_f} f^r(\mathbf{x}_{f,i_f})| \leq 2C^{-1}$ *for all* $i_f \in [n_f]$, *then we have*

$$\mathrm{D}_{\mathrm{KL}}(f^u || f^r) = \frac{1}{n_f} \sum_{i_f=1}^{n_f} \Big[p_{i_f}^u \log \frac{p_{i_f}^u}{p_{i_f}^r} + (1 - p_{i_f}^u) \log \frac{1 - p_{i_f}^u}{1 - p_{i_f}^r} \Big] \leq \frac{1}{2C^2},$$

where $p_{i_f}^u = q(y_{f,i_f} f^u(\mathbf{x}_{f,i_f}))$, $p_{i_f}^r = q(y_{f,i_f} f^r(\mathbf{x}_{f,i_f}))$, $f^u(\mathbf{x}_{f,i_f}) = f(\mathbf{W}^u, \mathbf{x}_{f,i_f})$, $f^r(\mathbf{x}_{f,i_f}) = f(\mathbf{W}^r, \mathbf{x}_{f,i_f})$ *and* $q(z) = 1/[1 + \exp(-z)]$.

Lemma 4.2 shows that the small-KL regime can be certified once $|y_{f,i_f} f^u(\mathbf{x}_{f,i_f}) - y_{f,i_f} f^r(\mathbf{x}_{f,i_f})|$ is sufficiently small. Following the single-task binary classification result in Kou et al. (2023b), $|y_{f,i_f} f^r(\mathbf{x}_{f,i_f})|$ remains fixed and is bounded by C^{-1}. Thus, it remains to characterize when $y_{f,i_f} f^u(\mathbf{x}_{f,i_f}) \leq C^{-1}$. At the start of unlearning, $y_{f,i_f} f^u(\mathbf{x}_{f,i_f})$ is governed by the average coefficient $\bar{\overline{\rho}}_t = \frac{1}{|S_{f,i_f}^{b,0}|} \sum_{s \in S_{f,i_f}^{b,0}} \overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$, where $S_{f,i_f}^{b,0} = \{s \in [m] : \langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, \boldsymbol{\xi}_{f,i_f} \rangle > 0\}$ is the set of neurons activated by the learned feature in the initial model. We define t_1 as the first iteration such that $y_{f,i_f} f^u(\mathbf{x}_{f,i_f}) \leq C^{-1}$. To estimate t_1, we first establish two key properties: balanced coefficient updates and invariant signal-to-noise coefficient ratio. The balanced update property is given below.

Lemma 4.3. *Under Condition 3.1, for* $0 \leq t \leq t_1$, *the following property hold :*

$$\ell'^{u,t}_{f,i_f} / \ell'^{u,t}_{f,i'_f} \leq C_3$$

for all $i_f, i'_f \in [n_f]$, *where* C_3 *is a positive constant.*

Lemma 4.3 shows that the loss gradients of all forgetting samples are balanced. The proof relies on the the property of the logistic loss gradient $\ell'(z) = -1/(1 + \exp(z))$, which implies that $\ell'(z_1)/\ell'(z_2)$ is controlled by $\exp(z_2 - z_1)$. Thus, it suffices to bound the difference between $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ and $y_{f,i'_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i'_f})$. Then $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ can be approximated by the accumulated coefficient $\sum_{s=1}^m \overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$. Together with the iteration analysis of (4.3), we obtain $\sum_{s=1}^m [\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t}(\boldsymbol{\xi}_{f,i'_f})] \leq \kappa$, where κ is a positive constant. This proves the balanced-updates property. Based on the balanced coefficient updates and Lemma 4.1, we further obtain the following invariant signal-to-noise coefficient ratio.

Lemma 4.4. *Under the same conditions as Theorem 3.2, we have*

$$\gamma_{j',s'}^{u,t}(\boldsymbol{\mu}_f) / \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) = \Theta(n_f \mathrm{SNR}_{\mathrm{f}}^2)$$

for $j = y_{f,i_f}, j' \in \{\pm 1\}$, $s \in S_{f,i_f}^{b,0}$, $s' \in [m]$ *and* $0 \leq t \leq t_1$, *where* $\mathrm{SNR}_{\mathrm{f}} = \|\boldsymbol{\mu}_f\|_2 / (\sigma_p \sqrt{d})$.

Lemma 4.4 shows that the ratio $\gamma_{j',s'}^{u,t}(\boldsymbol{\mu}_f)/\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$ is time-invariant. Therefore, as in the initial stage of unlearning, $y_{f,i_f}f^{u,t}(\mathbf{x}_{f,i_f})$ is governed by the averaged coefficient $\bar{\overline{\rho}}_t = \frac{1}{|S_{f,i_f}^{b,0}|}\sum_{s\in S_{f,i_f}^{b,0}}\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$, where $S_{f,i_f}^{b,0} = \{s\in[m] : \langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$. Combining the discrete update in (4.3) with the ODE approximation $\frac{dz(\tau)}{d\tau} = -\frac{\eta\sigma_p^2 d}{n_f m}\cdot\frac{1}{e^{\kappa z(\tau)}+1}$, we can estimate the first hitting time t_1 for $y_{f,i_f}f^{u,t}(\mathbf{x}_{f,i_f}) \le C^{-1}$. The formal statement is given in the following lemma.

Lemma 4.5. *Define $\bar{\overline{\rho}}_t \triangleq \frac{1}{|S_{f,i_f}^{b,0}|}\sum_{s\in S_{f,i_f}^{b,0}}\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$. Under the same conditions as Theorem 3.2, there exists*

$$t_1 = \frac{n_f m}{\eta\sigma_p^2 d}\left[\exp\left(\Theta(\bar{\overline{\rho}}_0)\right) - \exp\left(C^{-1}\right) + (\bar{\overline{\rho}}_0 - C^{-1})\right] = \Theta\left(\frac{n_f m}{\eta\sigma_p^2 d\epsilon}\right)$$

such that the KL divergence $\mathrm{D}_{\mathrm{KL}}(f^{u,t_1}||f^r) \le 2^{-1}C^{-2}$.

Lemma 4.5 provides an estimate of the first hitting time t_1 when $y_{f,i_f}f^{u,t}(\mathbf{x}_{f,i_f}) \le C^{-1}$. Together with Lemma 4.2, this implies $\mathrm{D}_{\mathrm{KL}}(f^{u,t_1}\|f^r) \le 2^{-1}C^{-2}$. See Appendix G.1 for the full proof.

Second stage. In this stage, we study how long the small-KL regime $\mathrm{D}_{\mathrm{KL}}(f^{u,t}\|f^r) \le 2^{-1}C^{-2}$ can be maintained after t_1. Equivalently, we analyze how long it takes for $y_{f,i_f}f^u(\mathbf{x}_{f,i_f})$ to decrease from C^{-1} to $-C^{-1}$. We define t_2 as the last iteration before $y_{f,i_f}f^u(\mathbf{x}_{f,i_f})$ drops below $-C^{-1}$. Unlike the first stage, where t_1 can be estimated through $\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$, the dynamics after t_1 are mainly driven by $\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)$. By Lemma 4.1, once $\langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f}\rangle < 0$, the coefficient $\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$ remains unchanged, whereas $\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)$ can continue to evolve. Specifically, if $\langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\mu}_f\rangle \ge 0$, then $\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)$ is updated by samples with $y_{f,i_f} = 1$; if $\langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\mu}_f\rangle < 0$, then it is updated by samples with $y_{f,i_f} = -1$. Therefore, when $\langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f}\rangle < 0$, the $y_{f,i_f}f^u(\mathbf{x}_{f,i_f})$ is governed by the averaged coefficient $\bar{\gamma}_{j,t} = \frac{1}{m}\sum_{s=1}^m \gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)$. It remains to characterize how long $\bar{\gamma}_{j,t}$ stays within the small-margin regime. The formal estimate is given in the following lemma.

Lemma 4.6. *Define $\bar{\gamma}_{j,t} \triangleq \frac{1}{m}\sum_{s=1}^m \gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)$. Under the same conditions as Theorem 3.2, there exists*

$$t_2 = t_1 + \Theta\left(\frac{m}{C\eta\|\boldsymbol{\mu}_f\|_2^2}\right)$$

such that $-C^{-1} \le \bar{\gamma}_{j,t} \le n_f \mathrm{SNR}_{\mathrm{f}}^2 C^{-1}$ for $t_1 \le t \le t_2$.

The estimate of t_2 in Lemma 4.6 is based on the observation that $\bar{\gamma}_{j,t}$ decreases at a linear rate. This follows from the uniform bound on the loss gradients, $C_4 \le -\ell'^{u,t}_{f,i_f} \le C_5$. Combining Lemma 4.6 with Lemma 4.2, we further obtain the following lemma.

Lemma 4.7. *Under the same conditions as Theorem 3.2, we have*

$$\mathrm{D}_{\mathrm{KL}}(f^{u,t}||f^r) \le 2^{-1}C^{-2},$$

for $t_1 \le t \le t_2$, where t_2 is defined as in Lemma 4.6.

Lemma 4.6 and Lemma 4.7 characterize how long the KL divergence can remain in the small regime. Together with Lemma 4.5, which estimates the first iteration at which the KL divergence enters this regime, they establish the first part of Theorem 3.2. The complete proof is provided in Appendix G.2. While the KL divergence remains in the small regime, we further have the following property.

Lemma 4.8. *Under the same conditions as Theorem 3.2, we have*

$$\left|L_{S_f}(\mathbf{W}^{u,t}) - L_{S_f}(\mathbf{W}^r)\right| \le 2C^{-1},$$

for $t_1 \le t \le t_2$, where C is a large constant and t_1, t_2 are defined as in Lemma 4.6.

Lemma 4.8 shows that, during the second stage where the KL divergence remains in the small regime, the gap between the forgetting losses of the unlearned and retrained models remains small. This observation motivates our learning-rate adjustment strategy, which uses the forgetting loss as a proxy to adjust the learning rate and mitigate overcorrection. More discussion is provided in Section 5.

Third stage. In this stage, we characterize the evolution of $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ and the corresponding KL divergence. In particular, we study the rate at which $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ continues to decrease. Similar to the second-stage analysis, $y_{f,i_f} f^u(\mathbf{x}_{f,i_f})$ is governed by the averaged coefficient $\bar{\gamma}_{j,t} = \frac{1}{m}\sum_{s=1}^{m} \gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f)$. Therefore, it suffices to analyze the decay of $\bar{\gamma}_{j,t}$. The formal characterization is given in the following lemma.

Lemma 4.9. *Under the same conditions as Theorem 3.2, when $t \geq t_2$, we have that*

1. *$\bar{\gamma}_{j,t}$ and $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ diverge to negative infinity at a linear rate.*
2. $\mathrm{D}_{\mathrm{KL}}(f^{u,t} \| f^r) \to \frac{1}{n_f}\sum_{i_f=1}^{n_f} \log\Big(1 + \exp\big(y_{f,i_f} f^r(\mathbf{x}_{f,i_f})\big)\Big)$.

Here, $\bar{\gamma}_{j,t}$ is defined as in Lemma 4.6.

Lemma 4.9 shows that $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ diverges to negative infinity at a linear rate, and consequently the KL divergence converges to a positive constant. This establishes the second part of Theorem 3.2. The proof relies on the uniform gradient bound $C_4 \leq -\ell'^{u,t}_{f,i_f} \leq 1$, analogous to the second-stage analysis. The complete proof is provided in Appendix G.3.

5 Experiments

In this section, we provide additional experiments to further investigate the overcorrection phenomenon and validate our theoretical results.

Overcorrection in GA-based unlearning. We empirically validate the overcorrection phenomenon on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and Tiny ImageNet (Le et al., 2015). For each dataset, we first train a ResNet-18 (He et al., 2016) for 182 epochs with an initial learning rate of 0.1 to obtain the initial model, following the setup of Fan et al. (2024). We then conduct class-wise unlearning by removing 10% of the classes from each dataset. The unlearning learning rates are set to 5×10^{-5}, 10^{-4} and 10^{-4} for CIFAR-10, CIFAR-100 and Tiny ImageNet respectively. To construct the retrained reference model, we use the same training configuration as the initial model, but train the model from scratch only on the retained data.

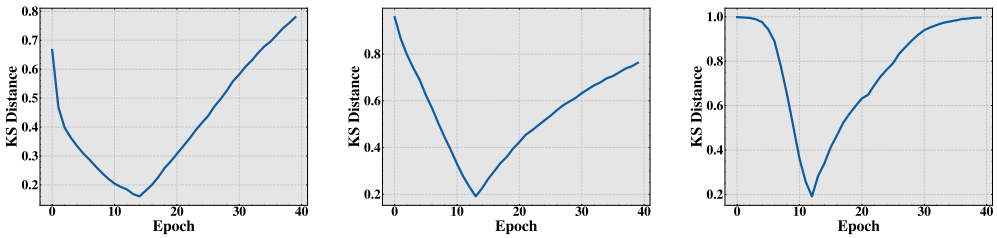


Figure 2: Overcorrection phenomenon on CIFAR-10, CIFAR-100, and Tiny ImageNet. KS distance between the unlearned model and the retrained model first decreases and then increases.

Figure 2 presents the evolution of the KS distance between the truth-ratio distributions of the unlearned model and the retrained model on the forgetting data during unlearning. We observe a clear non-monotonic pattern: the KS distance first decreases and then increases. This suggests that the forgetting quality initially improves, but subsequently deteriorates, which characterizes the overcorrection phenomenon.

Validation of theoretical results. We further conduct experiments on synthetic data to validate our theoretical findings. The samples are generated according to Definition 2.1. We set the forgetting set size and the retain set size to $n_f = 200$ and $n_r = 1000$ respectively. The input dimension is set to $d = 100$. The signal strengths are set to $\|\boldsymbol{\mu}_f\|_2 = \|\boldsymbol{\mu}_r\|_2 = 5$, and the noise scale is set to $\sigma_p = 1$. We train a two-layer ReLU network with width $m = 10$ for 20000 steps using a learning rate of 10^{-3} to obtain the initial model. Starting from this initial model, we then perform gradient ascent on the forgetting data for 5000 steps with an unlearning learning rate of 6×10^{-4}. The retrained model is

obtained using the same training configuration as the initial model, except that it is trained only on the retained data.

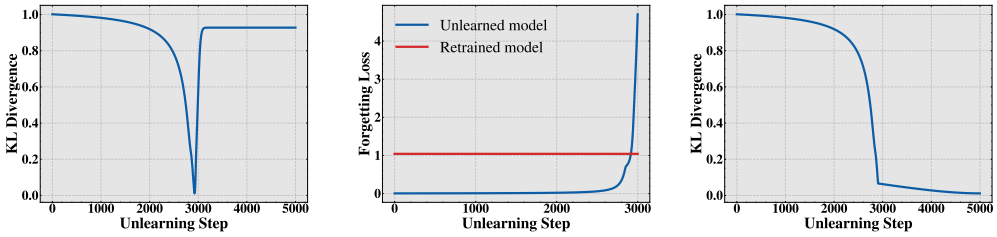


Figure 3: Validation of Theorem 3.2, Lemma 4.8, and the effectiveness of the learning-rate adjustment strategy on synthetic data. The KL divergence remains small almost simultaneously with the alignment of the forgetting losses of the unlearned model and the retrained model.

Figure 3 validates our theoretical characterization on synthetic data. The first plot shows that the KL divergence between the unlearned model and the retrained model on the forgetting data first decreases and then increases, consistent with Theorem 3.2. Combining the first and second plots, we observe that the small-KL regime nearly coincides with the regime where the loss gap between the unlearned and retrained models on the forgetting data remains small, supporting Lemma 4.8. This observation suggests a practical proxy for monitoring forgetting quality: once the forgetting loss indicates that the model has entered a favorable forgetting regime, we can reduce the unlearning learning rate to mitigate overcorrection. Specifically, we set a loss threshold and reduce the learning rate once the forgetting loss exceeds this threshold. On synthetic data, we set the threshold to 0.92 and reduce the learning rate to 0.01 of its original value after crossing this threshold. As shown in the third plot, this loss-threshold-based learning-rate adjustment keeps the KL divergence within a small range for longer, thereby mitigating overcorrection.

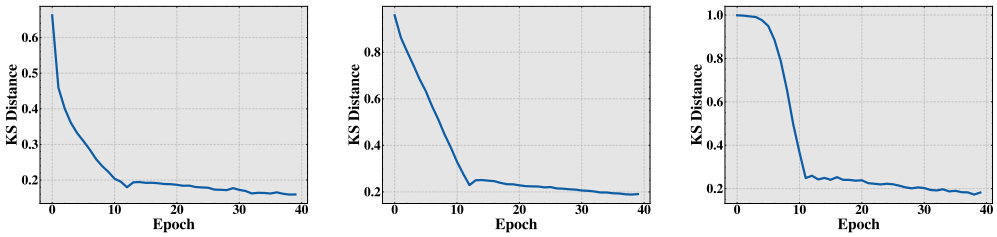


Figure 4: Learning-rate adjustment mitigates the overcorrection phenomenon on CIFAR-10, CIFAR-100, and Tiny ImageNet. As a result, the KS distance remains within a small range.

Mitigating overcorrection via learning-rate adjustment. We further evaluate the effectiveness of the learning-rate adjustment strategy on CIFAR-10, CIFAR-100 and Tiny ImageNet. Taking CIFAR-10 as an example, after obtaining the initial model, we perform gradient-ascent-based unlearning with an initial learning rate of 5×10^{-4}. We set the loss threshold to 8.0 and reduce the learning rate to 0.1 of its original value once the forgetting loss exceeds this threshold. The detailed settings for CIFAR-100 and Tiny ImageNet are provided in Appendix C.1. As shown in Figure 4, compared with Figure 2, the KS distance remains within a smaller range during unlearning, demonstrating the effectiveness of the proposed learning-rate adjustment strategy.

6 Conclusion

We identify and study the overcorrection phenomenon in GA-based machine unlearning, where forgetting quality first improves and then deteriorates during unlearning. We characterize this non-monotonic behavior through a three-stage theoretical analysis under a two-layer convolutional neural network model, explaining when the unlearned model approaches the retrained reference, how long it remains in this favorable regime, and why it eventually deviates. Our theory further motivates a simple learning-rate adjustment strategy to mitigate overcorrection. Experiments on synthetic and real-world datasets validate our theoretical results and the effectiveness of the proposed strategy.

AI USE STATEMENT

In this work, we use generative AI tools solely for translation and language polishing. Their use is limited to improving the linguistic presentation of the manuscript and does not extend to developing theoretical models, deriving mathematical results, designing or implementing methods, conducting experiments, or interpreting results. The authors are responsible for ensuring that AI-assisted revisions preserve the intended scientific meaning and take full responsibility for the accuracy, originality, and integrity of the final manuscript.

ETHICS STATEMENT

Our work focuses on understanding and mitigating the overcorrection phenomenon in gradient-ascent-based machine unlearning. Our experiments use synthetic data and publicly available image-classification benchmarks, including CIFAR-10, CIFAR-100, and Tiny ImageNet, without involving human-subject studies or collecting new personal or sensitive data. By improving the reliability of machine unlearning, this research can support privacy protection and responsible data governance. We do not identify any direct negative societal impacts specific to the proposed learning-rate adjustment strategy, and we remain committed to academic integrity and transparent reporting of our methods and findings.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the data model and learning setup in Section 2 and explicitly state the assumptions underlying our theoretical results in Section 3. Supporting lemmas, derivations, and detailed proofs are presented in Appendices D–G. Our empirical evaluations use synthetic data generated according to Definition 2.1 and the publicly available CIFAR-10, CIFAR-100, and Tiny ImageNet datasets. Section 5 and Appendix C describe the model architectures, training and unlearning settings, learning-rate adjustment parameters, and computational resources. Appendix C also reports additional experiments with different random seeds and forgetting-set proportions. An anonymous implementation repository is linked in Appendix C to support replication of the reported experiments.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd annual symposium on foundations of computer science (FOCS)*, pp. 977–988. IEEE, 2022.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Alexander Becker and Thomas Liebig. Evaluating machine unlearning via epistemic uncertainty. *arXiv preprint arXiv:2208.10836*, 2022.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12041–12052, 2023.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pp. 499–513, 2022.

Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. Why does sharpness-aware minimization generalize better than sgd? *Advances in neural information processing systems*, 36:72325–72376, 2023.

Jiali Cheng, George Dasoulas, Huan He, Chirag Agarwal, and Marinka Zitnik. Gnndelete: A general strategy for unlearning in graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Eli Chien, Chao Pan, and Olgica Milenkovic. Certified graph unlearning. *arXiv preprint arXiv:2206.09140*, 2022.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. *Advances in neural information processing systems*, 37: 79666–79703, 2024.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9304–9312, 2020.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11516–11524, 2021.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3832–3842, 2020.

Amartya Hatua, Trung T Nguyen, Filip Cano, and Andrew H Sung. Machine unlearning using forgetting neural networks. *arXiv preprint arXiv:2410.22374*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhengbao He, Tao Li, Xinwen Cheng, Zhehao Huang, and Xiaolin Huang. Towards natural machine unlearning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & communications technology law*, 28(1):65–98, 2019.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2038–2047, 2022.

Wei Huang, Ye Shi, Zhongyi Cai, and Taiji Suzuki. Understanding convergence and generalization in federated learning through feature learning theory. In *International Conference on Learning Representations*, volume 2024, pp. 25655–25686, 2024.

Wei Huang, Yuan Cao, Haonan Wang, Xin Cao, and Taiji Suzuki. Quantifying the optimization and generalization advantages of graph neural networks over multilayer perceptrons. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, pp. 2008–2016. PMLR, 2021.

Saachi Jain, Hadi Salman, Alaa Khaddaj, Eric Wong, Sung Min Park, and Aleksander Mądry. A data-based perspective on transfer learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3613–3622, 2023.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, 2023.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pp. 9965–10040. PMLR, 2022.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.

Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization. *Advances in Neural Information Processing Systems*, 37:135464–135625, 2024.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. In *International Conference on Machine Learning*, pp. 31275–31298. PMLR, 2025.

Yiwen Kou, Zixiang Chen, Yuan Cao, and Quanquan Gu. How does semi-supervised learning with pseudo-labelers work? a case study. In *The Eleventh International Conference on Learning Representations*, 2023a.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning*, pp. 17615–17659. PMLR, 2023b.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations (ICLR 2023)*, 2023.

Hongkang Li, Meng Wang, Tengfei Ma, Sijia Liu, Zaixi Zhang, and Pin-Yu Chen. What improves the generalization of graph transformers? a theoretical dive into the self-attention and positional encoding. In *International Conference on Machine Learning*, pp. 28784–28829. PMLR, 2024a.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *Proceedings of Machine Learning Research*, 235: 28525–28550, 2024b.

Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pp. 280–289. IEEE, 2022a.

Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pp. 1749–1758. IEEE, 2022b.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1817–1829, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. In *First Conference on Language Modeling*, 2024.

Odelia Melamed, Gilad Yehudai, and Gal Vardi. Provable unlearning with gradient ascent on two-layer relu neural networks. *arXiv preprint arXiv:2510.14844*, 2025.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Alex Oesterling, Jiaqi Ma, Flavio Calmon, and Himabindu Lakkaraju. Fair machine unlearning: Data removal while mitigating disparities. In *International Conference on Artificial Intelligence and Statistics*, pp. 3736–3744. PMLR, 2024.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):11676–11696, 2024.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022a.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX security symposium (USENIX Security 22)*, pp. 4007–4022, 2022b.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pp. 4126–4142. PMLR, 2021.

Hongyi Wan, Shiyuan Ren, Wei Huang, Miao Zhang, Xiang Deng, Yixin Bao, and Liqiang Nie. Understanding the forgetting of (replay-based) continual learning via feature learning: Angle matters. In *International Conference on Machine Learning*, pp. 61956–62019. PMLR, 2025.

Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM web conference 2022*, pp. 622–632, 2022.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. Kga: A general machine unlearning framework based on knowledge gap alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13264–13276, 2023.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

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

Leijie Wu, Song Guo, Junxiao Wang, Zicong Hong, Jie Zhang, and Yaohong Ding. Federated unlearning: Guarantee the right of clients to forget. *IEEE Network*, 36(5):129–135, 2022.

Yuanshun Yao and Xiaojun Xu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. In *International Conference on Learning Representations (ICLR)*, 2023.

Appendix

CONTENTS

A Limitations and Broader Impacts **16**

B Additional Related Works **16**

B.1 Machine Unlearning 16

B.2 Feature Learning Theory 16

C Additional Experiments **17**

C.1 Detailed Settings for the Experiments in Figure 4 17

C.2 More Empirical Results on Overcorrection 18

D Preliminary Lemmas **18**

E Signal-noise Decomposition Analysis **20**

E.1 Initial Model 20

E.2 Unlearned Model 22

E.3 Retrained Model 22

F Fundamental Properties of the Initial Model **23**

G Decoupling with a Three-stage Analysis **26**

G.1 First Stage 26

G.2 Second Stage 40

G.3 Third Stage 43

A Limitations and Broader Impacts

Our current theoretical analysis is developed under a specific network architecture and unlearning task setup. In particular, we focus on a two-layer convolutional neural network and analyze unlearning in a structured binary-classification setting. Extending the analysis to more general architectures, such as deeper convolutional networks, would be an important direction for future work. In addition, studying more complex unlearning scenarios, including multi-class unlearning and heterogeneous forget sets, may further improve the theoretical understanding of MU in practical settings.

This work aims to provide a theoretical explanation of the overcorrection phenomenon in machine unlearning. We are not aware of any immediate negative societal impacts resulting from this research.

B Additional Related Works

B.1 Machine Unlearning

Machine unlearning (MU) studies how to remove the effect of specified training samples from a trained model, and was originally motivated by privacy-related requirements after model deployment (Ginart et al., 2019; Neel et al., 2021; Ullah et al., 2021; Sekhari et al., 2021). A conceptually clean solution is exact unlearning, which retrains the model from scratch after excluding the data to be forgotten. However, this strategy is often computationally prohibitive for modern models and large-scale datasets. To reduce the cost of retraining, early studies investigated probabilistic relaxations, including approaches based on differential privacy (DP) (Ginart et al., 2019; Guo et al., 2020; Neel et al., 2021; Ullah et al., 2021; Sekhari et al., 2021). Despite their theoretical appeal, such methods may suffer from limited practical effectiveness, particularly when evaluated against membership inference attacks (Dwork et al., 2006; Graves et al., 2021). This has motivated a growing line of work on more practical and efficient unlearning algorithms (Golatkar et al., 2020; Becker & Liebig, 2022; Thudi et al., 2022b; Jia et al., 2023; Chen & Yang, 2023; Warnecke et al., 2021). Beyond standard supervised learning, MU has also been studied in broader settings, including federated learning (Wang et al., 2022; Liu et al., 2022b; Wu et al., 2022), graph neural networks (Chen et al., 2022; Chien et al., 2022; Cheng et al., 2023), and large language models (Jang et al., 2023; Wang et al., 2023; Chen & Yang, 2023; Yao & Xu, 2024; Eldan & Russinovich, 2023; Liu et al., 2024; Li et al., 2024b). Among these approaches, several representative unlearning methods perform GA on the forgetting loss to suppress the influence of the target data (Jang et al., 2023; Yao & Xu, 2024; Chen & Yang, 2023; Maini et al., 2024).

On the theoretical side, Guo et al. (2020) proposed certified data removal, which makes the unlearned model statistically close to retraining without the removed data, and developed an efficient Newton-style update for regularized linear models. Chien et al. (2024) proposed Langevin unlearning, a noisy-gradient framework for approximate certified unlearning that supports non-convex objectives and sequential removal requests. Koloskova et al. (2025) studied certified unlearning for neural networks and proposed noisy retain-set fine-tuning to provide provable removal guarantees with an explicit efficiency-accuracy trade-off. Melamed et al. (2025) theoretically analyzed GA-based unlearning in two-layer ReLU networks, showing that gradient ascent can remove the influence of a single data point while approximately preserving the KKT conditions on the retained data. In contrast, our work identifies and studies the overcorrection phenomenon in unlearning for the first time, where the forgetting quality evaluated solely on the forgetting data first improves and then deteriorates.

B.2 Feature Learning Theory

Feature learning theory has become an important framework for understanding the training dynamics and generalization behavior of modern neural networks (Cao et al., 2022; Kou et al., 2023b). It has been applied to a broad range of settings, including neural architectures such as convolutional neural networks (Cao et al., 2022; Kou et al., 2023b), vision transformers (Jelassi et al., 2022; Li et al., 2023; Jiang et al., 2024) and graph neural networks (Huang et al., 2025; Li et al., 2024a); learning paradigms such as knowledge distillation (Allen-Zhu & Li, 2023), semi-supervised and self-supervised learning (Kou et al., 2023a; Wen & Li, 2021), adversarial training (Allen-Zhu & Li, 2022), continual learning (Wan et al., 2025) and federated learning (Huang et al., 2024); and optimization methods such as Adam (Zou et al., 2023), momentum (Jelassi & Li, 2022), sharpness-aware minimization (Chen et al.,

2023). In contrast to these prior studies, our work leverages feature learning theory to understand the overcorrection phenomenon in machine unlearning, where the forgetting quality on the forgetting data first improves and then deteriorates.

C Additional Experiments

All experiments were implemented in PyTorch and conducted on an NVIDIA RTX 3050 GPU. The synthetic experiments require less than one hour on a single NVIDIA RTX 3050 GPU, while the real-world experiments take approximately 2–3 days on the same GPU, with the main computational cost coming from training the initial and retrained models. We provide the code at `https://anonymous.4open.science/r/anonymous-mu-overcorrection-8CFA`.

C.1 Detailed Settings for the Experiments in Figure 4

For the CIFAR-100 experiment with 10% class unlearning, we set the initial learning rate of gradient-ascent unlearning to 10^{-4} and the forgetting-loss threshold to 8.5. Once the forgetting loss exceeds this threshold, the learning rate is reduced to 0.06 times its initial value. For the Tiny ImageNet experiment with 10% class unlearning, we use the same initial learning rate of 10^{-4}, set the forgetting-loss threshold to 9.0, and reduce the learning rate to 0.03 times its initial value after the threshold is exceeded. The results are shown in the second and third plots of Figure 4, where the learning-rate adjustment strategy effectively alleviates the overcorrection phenomenon. In addition, we validate both the overcorrection phenomenon and the effectiveness of the learning-rate adjustment strategy under different random seeds. The results are reported in Figures 5, 6, 7, and 8.

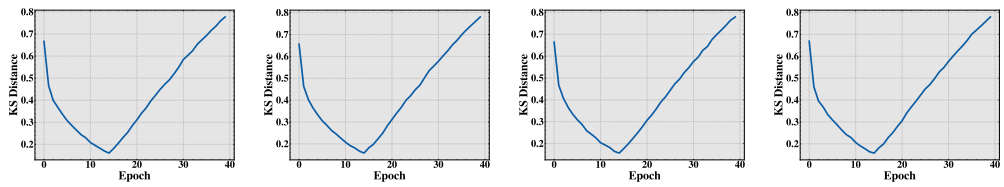


Figure 5: Overcorrection phenomenon on CIFAR-10 under different random seeds. KS distance between the truth-ratio distributions of the unlearned model and the retrained model first decreases and then increases.

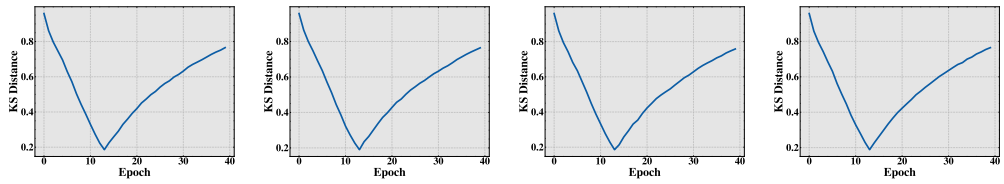


Figure 6: Overcorrection phenomenon on CIFAR-100 under different random seeds. KS distance between the truth-ratio distributions of the unlearned model and the retrained model first decreases and then increases.

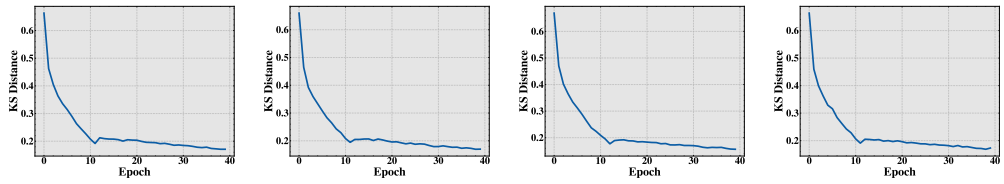


Figure 7: Learning-rate adjustment mitigates the overcorrection phenomenon on CIFAR-10 under different random seeds. As a result, the KS distance remains within a small range.

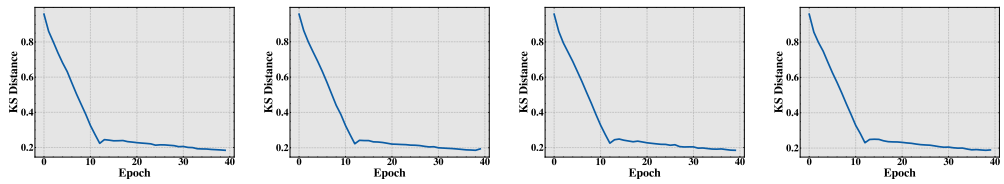


Figure 8: Learning-rate adjustment mitigates the overcorrection phenomenon on CIFAR-100 under different random seeds. As a result, the KS distance remains within a small range.

C.2 More Empirical Results on Overcorrection

To provide further empirical evidence for overcorrection, we conduct additional experiments on CIFAR-100 and Tiny ImageNet by unlearning 1% and 5% of the classes. For CIFAR-100, we set the learning rates to 5×10^{-4} and 2×10^{-4} for the 1% and 5% class-unlearning settings respectively. For Tiny ImageNet, we use a learning rate of 10^{-4} for both settings. Figure 9 reports the results under these four settings. In all cases, the KS distance first decreases and then increases as unlearning proceeds, confirming the presence of the overcorrection phenomenon.

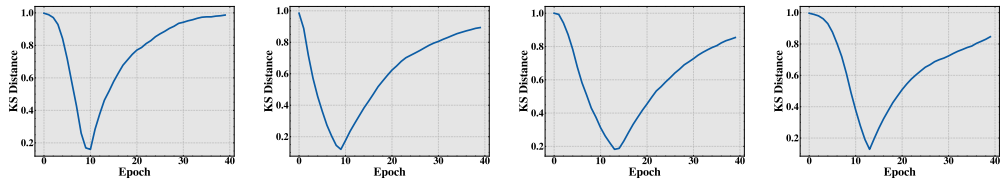


Figure 9: Overcorrection phenomenon on CIFAR-100 (unlearning 1% and 5%), and Tiny ImageNet (unlearning 1% and 5%). KS distance between the truth-ratio distributions of the unlearned model and the retrained model first decreases and then increases.

We further evaluate the effectiveness of the learning-rate adjustment strategy. The initial learning rates for unlearning are kept the same as those used in the experiments for validating the overcorrection phenomenon. For CIFAR-100 with 1% and 5% class unlearning, the forgetting-loss thresholds are set to 8.0 and 9.0 respectively. Once the forgetting loss exceeds the threshold, the learning rate is reduced to 0.01 and 0.05 times its initial value respectively. For Tiny ImageNet with 1% and 5% class unlearning, the forgetting-loss thresholds are set to 8.3 and 9.0 respectively, and the learning rate is reduced to 0.08 and 0.04 times its initial value after the threshold is exceeded. As shown in Figure 10, the learning-rate adjustment strategy effectively alleviates the overcorrection phenomenon.

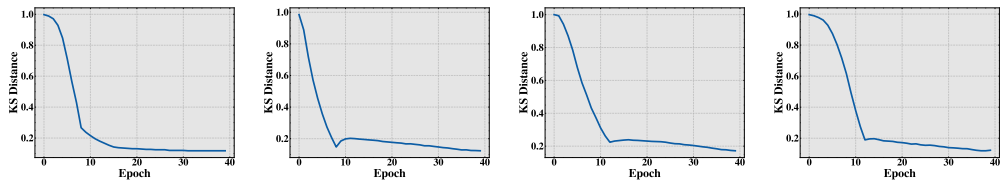


Figure 10: Learning-rate adjustment mitigates the overcorrection phenomenon on CIFAR-100 (unlearning 1% and 5%), and Tiny ImageNet (unlearning 1% and 5%). As a result, the KS distance remains within a small range.

D Preliminary Lemmas

In this section, we present several key lemmas that characterize important properties of the data and network parameters at random initialization.

Let $S_{f,1} = \{i_f | y_{f,i_f} = 1\}$, $S_{f,-1} = \{i_f | y_{f,i_f} = -1\}$, $S_{r,1} = \{i_r | y_{r,i_r} = 1\}$ and $S_{r,-1} = \{i_r | y_{r,i_r} = -1\}$. The following lemma characterizes the sizes of these sets.

Lemma D.1 (Kou et al. (2023b)). *Suppose that $\delta > 0$ and $n_f, n_r \geq 8\log(4/\delta)$. Then with probability at least $1-\delta$,*

$$|S_{f,1}|, |S_{f,-1}| \in [n_f/4, 3n_f/4],$$
$$|S_{r,1}|, |S_{r,-1}| \in [n_r/4, 3n_r/4].$$

The following lemma bounds the norms of the noise vectors, and provides upper bounds on their pairwise inner products as well as their inner products with the signal vectors.

Lemma D.2 (Kou et al. (2023b)). *Suppose that $\delta > 0$ and $d = \Omega(\log(6n/\delta))$, where $n = \max\{n_f, n_r\}$. Then with probability at least $1-\delta$,*

$$\sigma_p^2 d/2 \leq \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \leq 3\sigma_p^2 d/2,$$
$$\sigma_p^2 d/2 \leq \|\boldsymbol{\xi}_{r,i_r}\|_2^2 \leq 3\sigma_p^2 d/2,$$
$$|\langle \boldsymbol{\xi}_{f,i_f}, \boldsymbol{\xi}_{f,i'_f}\rangle| \leq 2\sigma_p^2 \cdot \sqrt{d\log(6n_f^2/\delta)},$$
$$|\langle \boldsymbol{\xi}_{r,i_r}, \boldsymbol{\xi}_{r,i'_r}\rangle| \leq 2\sigma_p^2 \cdot \sqrt{d\log(6n_r^2/\delta)},$$
$$|\langle \boldsymbol{\xi}_{f,i_f}, \boldsymbol{\xi}_{r,i_r}\rangle| \leq 2\sigma_p^2 \cdot \sqrt{d\log(6n^2/\delta)},$$
$$|\langle \boldsymbol{\xi}_{f,i_f}, \boldsymbol{\mu}_f\rangle| \leq \|\boldsymbol{\mu}_f\|_2\sigma_p \cdot \sqrt{2\log(6n_f/\delta)},$$
$$|\langle \boldsymbol{\xi}_{f,i_f}, \boldsymbol{\mu}_r\rangle| \leq \|\boldsymbol{\mu}_r\|_2\sigma_p \cdot \sqrt{2\log(6n_f/\delta)},$$
$$|\langle \boldsymbol{\xi}_{r,i_r}, \boldsymbol{\mu}_f\rangle| \leq \|\boldsymbol{\mu}_f\|_2\sigma_p \cdot \sqrt{2\log(6n_r/\delta)},$$
$$|\langle \boldsymbol{\xi}_{r,i_r}, \boldsymbol{\mu}_r\rangle| \leq \|\boldsymbol{\mu}_r\|_2\sigma_p \cdot \sqrt{2\log(6n_r/\delta)}$$

for all $i_f, i'_f \in [n_f]$ and $i_r, i'_r \in [n_r]$.

The following lemma studies the inner products between randomly initialized convolutional filters $\mathbf{w}_{j,s}^{b,0}$ and the signal and noise vectors, revealing how the network at initialization captures both components.

Lemma D.3 (Kou et al. (2023b)). *Suppose that $d = \Omega(\log(mn/\delta))$ and $m = \Omega(\log(1/\delta))$, where $n = \max\{n_f, n_r\}$. Then with probability at least $1-\delta$,*

$$\sigma_0^2 d/2 \leq \|\mathbf{w}_{j,s}^{b,0}\|_2^2 \leq 3\sigma_0^2 d/2,$$
$$|\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_f\rangle| \leq \sqrt{2\log(12m/\delta)} \cdot \sigma_0\|\boldsymbol{\mu}_f\|_2,$$
$$|\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_r\rangle| \leq \sqrt{2\log(12m/\delta)} \cdot \sigma_0\|\boldsymbol{\mu}_r\|_2,$$
$$|\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle| \leq 2\sqrt{\log(12mn_f/\delta)} \cdot \sigma_0\sigma_p\sqrt{d},$$
$$|\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{r,i_r}\rangle| \leq 2\sqrt{\log(12mn_r/\delta)} \cdot \sigma_0\sigma_p\sqrt{d}$$

for all $s \in [m]$, $j \in \{\pm 1\}$, $i_f \in [n_f]$ and $i_r \in [n_r]$. Moreover,

$$\sigma_0\|\boldsymbol{\mu}_f\|_2/2 \leq \max_{s\in[m]} j \cdot \langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_f\rangle \leq \sqrt{2\log(12m/\delta)} \cdot \sigma_0\|\boldsymbol{\mu}_f\|_2,$$
$$\sigma_0\|\boldsymbol{\mu}_r\|_2/2 \leq \max_{s\in[m]} j \cdot \langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_r\rangle \leq \sqrt{2\log(12m/\delta)} \cdot \sigma_0\|\boldsymbol{\mu}_r\|_2,$$
$$\sigma_0\sigma_p\sqrt{d}/4 \leq \max_{r\in[m]} j \cdot \langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle \leq 2\sqrt{\log(12mn_f/\delta)} \cdot \sigma_0\sigma_p\sqrt{d},$$
$$\sigma_0\sigma_p\sqrt{d}/4 \leq \max_{r\in[m]} j \cdot \langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{r,i_r}\rangle \leq 2\sqrt{\log(12mn_r/\delta)} \cdot \sigma_0\sigma_p\sqrt{d}$$

for all $j \in \{\pm 1\}$, $i_f \in [n_f]$ and $i_r \in [n_r]$.

Next, we define $S_{f,i_f}^{b,t} = \{s \in [m] : \langle \mathbf{w}_{y_{f,i_f},s}^{b,t}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$, $S_{f,i_f}^{u,t} = \{s \in [m] : \langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$ and $S_{f,j,s}^{b,t} = \{i_f \in [n_f] : y_{f,i_f} = j, \langle \mathbf{w}_{j,s}^{b,t}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$, $S_{f,j,s}^{u,t} = \{i_f \in [n_f] : y_{f,i_f} = j, \langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$ for $j \in \{\pm 1\}$ and $s \in [m]$. The following two lemmas provide lower bounds on $|S_{f,i_f}^{b,0}|$ and $|S_{f,j,s}^{b,0}|$.

Lemma D.4 (Kou et al. (2023b)). *Suppose that $\delta > 0$ and $m \geq 50\log(2n_f/\delta)$. Then with probability at least $1-\delta$,*

$$|S_{f,i_f}^{b,0}| \geq 0.4m,\ \forall i_f \in [n_f].$$

Lemma D.5 (Kou et al. (2023b)). *Suppose that $\delta > 0$ and $n_f \geq 32\log(4m/\delta)$. Then with probability at least $1-\delta$,*

$$|S_{f,j,r}^{b,0}| \geq n_f/8,\ \forall j \in \{\pm 1\}, s \in [m].$$

E Signal-noise Decomposition Analysis

In this section, we present a series of signal-noise decomposition results for both the initial model and the unlearned model. These results build upon the conclusions in Appendix D, which hold with high probability.

E.1 Initial Model

We obtain the initial model f^b by minimizing (2.1) via gradient descent. The corresponding update rule for the weights can be written as:

$$\begin{aligned}
\mathbf{w}_{j,s}^{b,t+1} &= \mathbf{w}_{j,s}^{b,t} - \eta \cdot \nabla_{\mathbf{w}_{j,s}} L_S(\mathbf{W}^{b,t}) \\
&= \mathbf{w}_{j,s}^{b,t} - \frac{\eta}{m(n_f+n_r)} \sum_{i_f=1}^{n_f} \ell_{f,i_f}'^{b,t} \big[\sigma'(\langle \mathbf{w}_{j,s}^{b,t}, y_{f,i_f}\boldsymbol{\mu}_f\rangle) j\boldsymbol{\mu}_f + \sigma'(\langle \mathbf{w}_{j,s}^{b,t}, \boldsymbol{\xi}_{f,i_f}\rangle) j y_{f,i_f}\boldsymbol{\xi}_{f,i_f}\big] \\
&\quad - \frac{\eta}{m(n_f+n_r)} \sum_{i_r=1}^{n_r} \ell_{r,i_r}'^{b,t} \big[\sigma'(\langle \mathbf{w}_{j,s}^{b,t}, y_{r,i_r}\boldsymbol{\mu}_r\rangle) j\boldsymbol{\mu}_r + \sigma'(\langle \mathbf{w}_{j,s}^{b,t}, \boldsymbol{\xi}_{r,i_r}\rangle) j y_{r,i_r}\boldsymbol{\xi}_{r,i_r}\big]
\end{aligned} \tag{E.1}$$

for $j \in \{\pm 1\}$ and $s \in [m]$, where $\ell_{f,i_f}'^{b,t} = \ell'[y_{f,i_f} \cdot f(\mathbf{W}^{b,t}, \mathbf{x}_{f,i_f})]$ and $\ell_{r,i_r}'^{b,t} = \ell'[y_{r,i_r} \cdot f(\mathbf{W}^{b,t}, \mathbf{x}_{r,i_r})]$. Based on the gradient descent update in (E.1), we obtain the following signal-noise decomposition:

$$\begin{aligned}
\mathbf{w}_{j,s}^{b,t} = \mathbf{w}_{j,s}^{b,0} &+ j\gamma_{j,s}^{b,t}(\boldsymbol{\mu}_f)\frac{\boldsymbol{\mu}_f}{\|\boldsymbol{\mu}_f\|_2^2} + \sum_{i_f=1}^{n_f} \rho_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f})\frac{\boldsymbol{\xi}_{f,i_f}}{\|\boldsymbol{\xi}_{f,i_f}\|_2^2} \\
&+ j\gamma_{j,s}^{b,t}(\boldsymbol{\mu}_r)\frac{\boldsymbol{\mu}_r}{\|\boldsymbol{\mu}_r\|_2^2} + \sum_{i_r=1}^{n_r} \rho_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r})\frac{\boldsymbol{\xi}_{r,i_r}}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2}.
\end{aligned} \tag{E.2}$$

Further denote $\overline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) := \rho_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f})\,\mathbb{1}(\rho_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) \geq 0)$, $\underline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) := \rho_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f})\,\mathbb{1}(\rho_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) \leq 0)$, $\overline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) := \rho_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r})\,\mathbb{1}(\rho_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) \geq 0)$, $\underline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) := \rho_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r})\,\mathbb{1}(\rho_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) \leq 0)$, then we have the following lemma.

Lemma E.1. *The coefficients $\gamma_{j,s}^{b,t}(\boldsymbol{\mu}_f), \gamma_{j,s}^{b,t}(\boldsymbol{\mu}_r), \overline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}), \underline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}), \overline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r})$ and $\underline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r})$ in decomposition* (E.2) *satisfy the following iterative equations:*

$$\begin{aligned}
&\gamma_{j,s}^{b,0}(\boldsymbol{\mu}_f), \gamma_{j,s}^{b,0}(\boldsymbol{\mu}_r), \overline{\rho}_{j,s,i_f}^{b,0}(\boldsymbol{\xi}_{f,i_f}), \underline{\rho}_{j,s,i_f}^{b,0}(\boldsymbol{\xi}_{f,i_f}), \overline{\rho}_{j,s,i_r}^{b,0}(\boldsymbol{\xi}_{r,i_r}), \underline{\rho}_{j,s,i_r}^{b,0}(\boldsymbol{\xi}_{r,i_r}) = 0, \\
&\gamma_{j,s}^{b,t+1}(\boldsymbol{\mu}_f) = \gamma_{j,s}^{b,t}(\boldsymbol{\mu}_f) - \frac{\eta}{m(n_f+n_r)} \cdot \sum_{i_f=1}^{n_f} \ell_{f,i_f}'^{b,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{b,t}, y_{f,i_f}\cdot\boldsymbol{\mu}_f\rangle) \cdot \|\boldsymbol{\mu}_f\|_2^2, \\
&\gamma_{j,s}^{b,t+1}(\boldsymbol{\mu}_r) = \gamma_{j,s}^{b,t}(\boldsymbol{\mu}_r) - \frac{\eta}{m(n_f+n_r)} \cdot \sum_{i_r=1}^{n_r} \ell_{r,i_r}'^{b,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{b,t}, y_{r,i_r}\cdot\boldsymbol{\mu}_r\rangle) \cdot \|\boldsymbol{\mu}_r\|_2^2, \\
&\overline{\rho}_{j,s,i_f}^{b,t+1}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta}{m(n_f+n_r)} \cdot \ell_{f,i_f}'^{b,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{b,t}, \boldsymbol{\xi}_{f,i_f}\rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = j),
\end{aligned}$$

$$\underline{\rho}_{j,s,i_f}^{b,t+1}(\boldsymbol{\xi}_{f,i_f}) = \underline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{m(n_f+n_r)} \cdot \ell_{f,i_f}'^{b,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{b,t}, \boldsymbol{\xi}_{f,i_f}\rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = -j)$$

$$\overline{\rho}_{j,s,i_r}^{b,t+1}(\boldsymbol{\xi}_{r,i_r}) = \overline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) - \frac{\eta}{m(n_f+n_r)} \cdot \ell_{r,i_r}'^{b,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{b,t}, \boldsymbol{\xi}_{r,i_r}\rangle) \cdot \|\boldsymbol{\xi}_{r,i_r}\|_2^2 \cdot \mathbb{1}(y_{r,i_r} = j),$$

$$\underline{\rho}_{j,s,i_r}^{b,t+1}(\boldsymbol{\xi}_{r,i_r}) = \underline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) + \frac{\eta}{m(n_f+n_r)} \cdot \ell_{r,i_r}'^{b,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{b,t}, \boldsymbol{\xi}_{r,i_r}\rangle) \cdot \|\boldsymbol{\xi}_{r,i_r}\|_2^2 \cdot \mathbb{1}(y_{r,i_r} = -j)$$

for all $s \in [m]$, $j \in \{\pm 1\}$, $i_f \in [n_f]$ *and* $i_r \in [n_r]$.

Proof of Lemma E.1. First, by iterating the gradient descent update rule (E.1) for t steps, we obtain

$$\mathbf{w}_{j,s}^{b,t} = \mathbf{w}_{j,s}^{b,0} - \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \sum_{i_f=1}^{n_f} \ell_{f,i_f}'^{b,\tau} \big[\sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, y_{f,i_f}\boldsymbol{\mu}_f\rangle) j\boldsymbol{\mu}_f + \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, \boldsymbol{\xi}_{f,i_f}\rangle) j y_{f,i_f}\boldsymbol{\xi}_{f,i_f}\big]$$
$$- \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \sum_{i_r=1}^{n_r} \ell_{r,i_r}'^{b,\tau} \big[\sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, y_{r,i_r}\boldsymbol{\mu}_r\rangle) j\boldsymbol{\mu}_r + \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, \boldsymbol{\xi}_{r,i_r}\rangle) j y_{r,i_r}\boldsymbol{\xi}_{r,i_r}\big]$$

According to the decomposition (E.2),

$$\mathbf{w}_{j,s}^{b,t} = \mathbf{w}_{j,s}^{b,0} + j\gamma_{j,s}^{b,t}(\boldsymbol{\mu}_f)\frac{\boldsymbol{\mu}_f}{\|\boldsymbol{\mu}_f\|_2^2} + \sum_{i_f=1}^{n_f} \rho_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f})\frac{\boldsymbol{\xi}_{f,i_f}}{\|\boldsymbol{\xi}_{f,i_f}\|_2^2}$$
$$+ j\gamma_{j,s}^{b,t}(\boldsymbol{\mu}_r)\frac{\boldsymbol{\mu}_r}{\|\boldsymbol{\mu}_r\|_2^2} + \sum_{i_r=1}^{n_r} \rho_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r})\frac{\boldsymbol{\xi}_{r,i_r}}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2}.$$

Since these vectors are linearly independent with probability 1, we have the following unique representation

$$\gamma_{j,s}^{b,t}(\boldsymbol{\mu}_f) = -\frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \sum_{i_f=1}^{n_f} \ell_{f,i_f}'^{b,\tau} \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, y_{f,i_f}\boldsymbol{\mu}_f\rangle)\|\boldsymbol{\mu}_f\|_2^2, \tag{E.3}$$

$$\gamma_{j,s}^{b,t}(\boldsymbol{\mu}_r) = -\frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \sum_{i_r=1}^{n_r} \ell_{r,i_r}'^{b,\tau} \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, y_{r,i_r}\boldsymbol{\mu}_r\rangle)\|\boldsymbol{\mu}_r\|_2^2, \tag{E.4}$$

$$\rho_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) = -\frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell_{f,i_f}'^{b,\tau} \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, \boldsymbol{\xi}_{f,i_f}\rangle)\|\boldsymbol{\xi}_{f,i_f}\|_2^2 j y_{f,i_f},$$

$$\rho_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) = -\frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell_{r,i_r}'^{b,\tau} \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, \boldsymbol{\xi}_{r,i_r}\rangle)\|\boldsymbol{\xi}_{r,i_r}\|_2^2 j y_{r,i_r}.$$

Based on the notation of $\overline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f})$, $\underline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f})$, $\overline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r})$, $\underline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r})$ and $\ell_{f,i_f}'^{b,\tau} < 0$, $\ell_{r,i_r}'^{b,\tau} < 0$, we get

$$\overline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) = -\frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell_{f,i_f}'^{b,\tau} \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, \boldsymbol{\xi}_{f,i_f}\rangle)\|\boldsymbol{\xi}_{f,i_f}\|_2^2 \mathbb{1}(y_{f,i_f} = j), \tag{E.5}$$

$$\underline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) = \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell_{f,i_f}'^{b,\tau} \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, \boldsymbol{\xi}_{f,i_f}\rangle)\|\boldsymbol{\xi}_{f,i_f}\|_2^2 \mathbb{1}(y_{f,i_f} = -j), \tag{E.6}$$

$$\overline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) = -\frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell_{r,i_r}'^{b,\tau} \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, \boldsymbol{\xi}_{r,i_r}\rangle)\|\boldsymbol{\xi}_{r,i_r}\|_2^2 \mathbb{1}(y_{r,i_r} = j), \tag{E.7}$$

$$\underline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) = \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell_{r,i_r}'^{b,\tau} \sigma'(\langle \mathbf{w}_{j,s}^{b,\tau}, \boldsymbol{\xi}_{r,i_r}\rangle)\|\boldsymbol{\xi}_{r,i_r}\|_2^2 \mathbb{1}(y_{r,i_r} = -j). \tag{E.8}$$

Writing out the iterative versions of (E.3), (E.4), (E.5), (E.6), (E.7) and (E.8) completes the proof. □

E.2 UNLEARNED MODEL

Lemma E.2 (Restatement of Lemma 4.1)**.** *The coefficients* $\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f), \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}), \underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$ *in decomposition* (4.1) *satisfy the following equations:*

$$\gamma_{j,s}^{u,0}(\boldsymbol{\mu}_f) = \gamma_{j,s}^{b,t_{end}}(\boldsymbol{\mu}_f), \overline{\rho}_{j,s,i_f}^{u,0}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}_{j,s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}), \underline{\rho}_{j,s,i_f}^{u,0}(\boldsymbol{\xi}_{f,i_f}) = \underline{\rho}_{j,s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}),$$

$$\gamma_{j,s}^{u,t+1}(\boldsymbol{\mu}_f) = \gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \cdot \sum_{i_f=1}^{n_f} \ell'^{u,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,t}, y_{f,i_f} \cdot \boldsymbol{\mu}_f \rangle) \cdot \|\boldsymbol{\mu}_f\|_2^2,$$

$$\overline{\rho}_{j,s,i_f}^{u,t+1}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{n_f m} \cdot \ell'^{u,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = j),$$

$$\underline{\rho}_{j,s,i_f}^{u,t+1}(\boldsymbol{\xi}_{f,i_f}) = \underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta}{n_f m} \cdot \ell'^{u,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = -j)$$

for all $s \in [m]$, $j \in \{\pm 1\}$ *and* $i_f \in [n_f]$.

Proof of Lemma E.2. First, iterating (2.3) for t steps yields

$$\mathbf{w}_{j,s}^{u,t} = \mathbf{w}_{j,s}^{u,0} + \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \sum_{i_f=1}^{n_f} \ell'^{u,\tau}_{f,i_f} \big[\sigma'(\langle \mathbf{w}_{j,s}^{u,\tau}, y_{f,i_f} \boldsymbol{\mu}_f \rangle) j \boldsymbol{\mu}_f + \sigma'(\langle \mathbf{w}_{j,s}^{u,\tau}, \boldsymbol{\xi}_{f,i_f} \rangle) j y_{f,i_f} \boldsymbol{\xi}_{f,i_f} \big]$$

Using the same proof technique as in Lemma E.1, we obtain

$$\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f) = \gamma_{j,s}^{u,0}(\boldsymbol{\mu}_f) + \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \sum_{i_f=1}^{n_f} \ell'^{u,\tau}_{f,i_f} \sigma'(\langle \mathbf{w}_{j,s}^{u,\tau}, y_{f,i_f} \boldsymbol{\mu}_f \rangle) \|\boldsymbol{\mu}_f\|_2^2, \tag{E.9}$$

$$\rho_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) = \rho_{j,s,i_f}^{u,0}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell'^{u,\tau}_{f,i_f} \sigma'(\langle \mathbf{w}_{j,s}^{u,\tau}, \boldsymbol{\xi}_{f,i_f} \rangle) \|\boldsymbol{\xi}_{f,i_f}\|_2^2 j y_{f,i_f}.$$

Using the notation of $\overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$ and $\underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$, together with $\ell'^{u,\tau}_{f,i_f} < 0$, we obtain

$$\overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}_{j,s,i_f}^{u,0}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell'^{u,\tau}_{f,i_f} \sigma'(\langle \mathbf{w}_{j,s}^{u,\tau}, \boldsymbol{\xi}_{f,i_f} \rangle) \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \mathbb{1}(y_{f,i_f} = j), \tag{E.10}$$

$$\underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) = \underline{\rho}_{j,s,i_f}^{u,0}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta}{m(n_f+n_r)} \sum_{\tau=0}^{t-1} \ell'^{u,\tau}_{f,i_f} \sigma'(\langle \mathbf{w}_{j,s}^{u,\tau}, \boldsymbol{\xi}_{f,i_f} \rangle) \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \mathbb{1}(y_{f,i_f} = -j). \tag{E.11}$$

Writing out the iterative versions of (E.9), (E.10) and (E.11) completes the proof. □

E.3 RETRAINED MODEL

Similar to the initial model f^b, the retrained model f^r admits the following signal–noise decomposition:

$$\mathbf{w}_{j,s}^{r,t} = \mathbf{w}_{j,s}^{r,0} + j\gamma_{j,s}^{r,t}(\boldsymbol{\mu}_r) \frac{\boldsymbol{\mu}_r}{\|\boldsymbol{\mu}_r\|_2^2} + \sum_{i_r=1}^{n_r} \overline{\rho}_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) \frac{\boldsymbol{\xi}_{r,i_r}}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2} + \sum_{i_r=1}^{n_r} \underline{\rho}_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) \frac{\boldsymbol{\xi}_{r,i_r}}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2}, \tag{E.12}$$

where $\overline{\rho}_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) = \rho_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) \mathbb{1}(\rho_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) \geq 0)$, $\underline{\rho}_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) = \rho_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) \mathbb{1}(\rho_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) \leq 0)$, and $\mathbf{w}_{j,s}^{r,0} = \mathbf{w}_{j,s}^{b,0}$ in the retrained model. Using a similar argument to that in the proof of Lemma E.1, we obtain the following lemma.

Lemma E.3. *The coefficients* $\gamma_{j,s}^{r,t}(\boldsymbol{\mu}_r), \overline{\rho}_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r})$ *and* $\underline{\rho}_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r})$ *in decomposition* (E.12) *satisfy the following iterative equations:*

$$\gamma_{j,s}^{r,0}(\boldsymbol{\mu}_r), \overline{\rho}_{j,s,i_r}^{b,0}(\boldsymbol{\xi}_{r,i_r}), \underline{\rho}_{j,s,i_r}^{b,0}(\boldsymbol{\xi}_{r,i_r}) = 0,$$

$$\gamma_{j,s}^{r,t+1}(\boldsymbol{\mu}_r) = \gamma_{j,s}^{r,t}(\boldsymbol{\mu}_r) - \frac{\eta}{mn_r} \cdot \sum_{i_r=1}^{n_r} \ell_{r,i_r}'^{r,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{r,t}, y_{r,i_r} \cdot \boldsymbol{\mu}_r \rangle) \cdot \|\boldsymbol{\mu}_r\|_2^2,$$

$$\overline{\rho}_{j,s,i_r}^{r,t+1}(\boldsymbol{\xi}_{r,i_r}) = \overline{\rho}_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) - \frac{\eta}{mn_r} \cdot \ell_{r,i_r}'^{r,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{r,t}, \boldsymbol{\xi}_{r,i_r} \rangle) \cdot \|\boldsymbol{\xi}_{r,i_r}\|_2^2 \cdot \mathbb{1}(y_{r,i_r} = j),$$

$$\underline{\rho}_{j,s,i_r}^{r,t+1}(\boldsymbol{\xi}_{r,i_r}) = \underline{\rho}_{j,s,i_r}^{r,t}(\boldsymbol{\xi}_{r,i_r}) + \frac{\eta}{mn_r} \cdot \ell_{r,i_r}'^{r,t} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{r,t}, \boldsymbol{\xi}_{r,i_r} \rangle) \cdot \|\boldsymbol{\xi}_{r,i_r}\|_2^2 \cdot \mathbb{1}(y_{r,i_r} = -j)$$

for all $s \in [m]$, $j \in \{\pm 1\}$, *and* $i_r \in [n_r]$.

F Fundamental Properties of the Initial Model

We primarily focus on the regime where $n_f\|\boldsymbol{\mu}_f\|_2^4 \geq C_1\sigma_p^4 d$ and $n_r\|\boldsymbol{\mu}_r\|_2^4 \geq C_1\sigma_p^4 d$, under which the initial model f^b exhibits strong generalization performance (as characterized in Lemma F.1). Moreover, all results in this section rely on the conclusions established in Appendix D, which hold with high probability.

Lemma F.1 (Kou et al. (2023b)). *For any $\epsilon > 0$, under Condition 3.1, suppose that $n_f\|\boldsymbol{\mu}_f\|_2^4 \geq C_1\sigma_p^4 d$ and $n_r\|\boldsymbol{\mu}_r\|_2^4 \geq C_1\sigma_p^4 d$. Then with probability at least $1-\delta$, there exists $t_{end} = \tilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$ such that:*

1. *The training loss converges to ϵ, i.e., $L_S(\mathbf{W}^{b,t_{end}}) \leq \epsilon$.*
2. *The CNN achieves a small classification error: $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{b,t_{end}}) \leq \exp\Big(-n\|\boldsymbol{\mu}\|_2^4/(C_2\sigma_p^4 d)\Big)$.*

Here C_1, C_2 are some absolute constants, $\|\boldsymbol{\mu}\|_2 = \max\{\|\boldsymbol{\mu}_1\|_2, \|\boldsymbol{\mu}_2\|_2\}$, and $n = \max\{n_f, n_r\}$.

Moreover, Kou et al. (2023b) further established that the initial model f^b satisfies the following properties:

Proposition F.2. *Let $\beta = 2\max_{j,s,i_f,i_r}\{|\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_f\rangle|, |\mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_r\rangle|, |\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle|, |\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{r,i_r}\rangle|\}$ and $\alpha = 4\log(T^*)$, where $T^* = \eta^{-1}\,\mathrm{poly}(n, m, d^{-1}, \epsilon^{-1})$ and $n = \max\{n_f, n_r\}$. Given Condition 3.1 and $\epsilon > 0$, then for all $0 \leq t \leq T^*$, $i_f \in [n_f]$, $i_r \in [n_r]$, $s \in [m]$, and $j \in \{\pm 1\}$, the following bounds hold:*

$$0 \leq \gamma_{j,s}^{b,t}(\boldsymbol{\mu}_f) \leq C' n_f \mathrm{SNR}_{\mathrm{f}}^2 \alpha, \quad 0 \leq \overline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) \leq \alpha, \tag{F.1}$$

$$0 \leq \gamma_{j,s}^{b,t}(\boldsymbol{\mu}_r) \leq C' n_r \mathrm{SNR}_{\mathrm{r}}^2 \alpha, \quad 0 \leq \overline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) \leq \alpha, \tag{F.2}$$

$$0 \geq \underline{\rho}_{j,s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) \geq -\beta - 20\sqrt{\frac{\log(6n^2/\delta)}{d}} n\alpha \geq -\alpha. \tag{F.3}$$

$$0 \geq \underline{\rho}_{j,s,i_r}^{b,t}(\boldsymbol{\xi}_{r,i_r}) \geq -\beta - 20\sqrt{\frac{\log(6n^2/\delta)}{d}} n\alpha \geq -\alpha, \tag{F.4}$$

where C' is a positive constant, $\mathrm{SNR}_{\mathrm{f}} = \|\boldsymbol{\mu}_f\|_2/(\sigma_p\sqrt{d})$ and $\mathrm{SNR}_{\mathrm{r}} = \|\boldsymbol{\mu}_r\|_2/(\sigma_p\sqrt{d})$.

Lemma F.3 (Kou et al. (2023b)). *Under the same conditions as Lemma F.1, the following properties hold for any iteration $t \leq t_{end}$:*

1. *$\sum_{s=1}^m \big[\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i_f'},s,i_f'}^{b,t}(\boldsymbol{\xi}_{f,i_f'})\big] \leq \kappa$ for all $i_f, i_f' \in [n_f]$.*
2. *$y_{f,i_f} \cdot f(\mathbf{W}^{b,t}, \mathbf{x}_{f,i_f}) - y_{f,i_f'} \cdot f(\mathbf{W}^{b,t}, \mathbf{x}_{f,i_f'}) \leq C_6$ for all $i_f, i_f' \in [n_f]$,*
3. *$\ell_{f,i_f}'^{b,t}/\ell_{f,i_f'}'^{b,t} \leq C_3 = \exp(C_6)$ for all $i_f, i_f' \in [n]$.*
4. *$S_{f,i_f}^{b,0} \subseteq S_{f,i_f}^{b,t}$, where $S_{f,i_f}^{b,t} := \{s \in [m] : \langle \mathbf{w}_{y_{f,i_f},s}^{b,t}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$, and $|S_{f,i_f}^{b,t}| \geq |S_{f,i_f}^{b,0}| \geq 0.4m$ for all $i_f \in [n_f]$.*
5. *$S_{f,j,s}^{b,0} \subseteq S_{f,j,s}^{b,t}$, where $S_{f,j,s}^{b,t} := \{i_f \in [n_f] : \langle \mathbf{w}_{y_{f,i_f},s}^{b,t}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$, and hence $|S_{f,j,s}^{b,t}| \geq n_f/8$ for all $j \in \{\pm 1\}, s \in [m]$.*

Here, κ and C_6 are positive constants.

Building on Lemma F.3, we provide a more refined characterization as follows.

Lemma F.4 (Kou et al. (2023b)). *Let $g(z) = \ell'(z) = -1/(1+\exp(z))$, then for all $z_2 - c = z_1 \geq -1$ where $c \geq 0$, we have that*

$$\frac{\exp(c)}{4} \leq \frac{\exp(1)+\exp(c)}{\exp(1)+1} \leq \frac{g(z_1)}{g(z_2)} \leq \exp(c).$$

Lemma F.5. *Under the same conditions as Lemma F.1, for any iteration $0 \leq t \leq t_{end}$, we have that*

$$\overline{\rho}^{b,t}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,t}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) \leq \kappa,$$

for any $i_f, i'_f \in [n_f]$ and $s, s' \in S^{b,0}_{f,i_f}$, where κ is a positive constant.

Proof of Lemma F.5. We prove the lemma by induction. For the base case $t = 0$, since $\overline{\rho}^{b,0}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) = 0$ for any $i_f \in [n_f]$ and $s \in S^{b,0}_{f,i_f}$, it follows that $\overline{\rho}^{b,t}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,t}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) = 0$, and thus the claim holds at $t = 0$. Now suppose that there exists $\tilde{t} \leq t$ such that the property holds for all $0 \leq t \leq \tilde{t}-1$. We aim to show that it also holds at $t = \tilde{t}$.

We begin by deriving the iterative update for $\overline{\rho}^{b,t}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,t}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f})$. From Lemma E.1, recall that

$$\overline{\rho}^{b,t+1}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}^{b,t}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta}{m(n_f+n_r)} \cdot \ell'^{b,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}^{b,t}_{j,s}, \boldsymbol{\xi}_{f,i_f} \rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = j),$$

for all $j \in \{\pm 1\}$, $s \in [m]$, and $i_f \in [n_f]$. By the definition of $S^{b,0}_{f,i_f}$, we obtain

$$\begin{aligned}
\overline{\rho}^{b,t+1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,t+1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) &= \overline{\rho}^{b,t}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,t}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) \\
&\quad - \frac{\eta}{m(n_f+n_r)}\Big(\ell'^{b,t}_{f,i_f}\|\boldsymbol{\xi}_{f,i_f}\|_2^2 - \ell'^{b,t}_{f,i'_f}\|\boldsymbol{\xi}_{f,i'_f}\|_2^2\Big),
\end{aligned}$$

for any $i_f, i'_f \in [n_f]$. We next distinguish two cases depending on the value of $\overline{\rho}^{b,t}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,t}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f})$:

When $\overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) \leq 0.9\kappa$,

$$\begin{aligned}
\overline{\rho}^{b,\tilde{t}}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) &\leq \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) - \frac{\eta\|\boldsymbol{\xi}_{f,i_f}\|_2^2 \ell'^{b,\tilde{t}-1}_{f,i_f}}{m(n_f+n_r)} \\
&\leq \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) + \frac{\eta\|\boldsymbol{\xi}_{f,i_f}\|_2^2}{m(n_f+n_r)} \\
&\leq 0.1\kappa + 0.9\kappa \\
&\leq \kappa,
\end{aligned}$$

where the first inequality is by $\ell'^{b,\tilde{t}-1}_{f,i_f} < 0$; the second inequality is by $-\ell'^{b,\tilde{t}-1}_{f,i_f} < 1$; the third inequality is by $\overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) \leq 0.9\kappa$ and $\eta \leq C^{-1}m(n_f+n_r)\sigma_p^{-2}d^{-1}$.

When $\overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) \geq 0.9\kappa$, by adopting a proof technique similar to that used in Lemma C.7 of Kou et al. (2023b), and setting $\kappa = 3.25$, we obtain

$$\begin{aligned}
&y_{f,i_f} f(\mathbf{W}^{b,\tilde{t}-1}, \mathbf{x}_{f,i_f}) - y_{f,i'_f} f(\mathbf{W}^{b,\tilde{t}-1}, \mathbf{x}_{f,i'_f}) \\
&\geq \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f})\big] - 1.75
\end{aligned}$$

$$\geq \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) - 0.54\kappa$$
$$\geq 0.36\kappa,$$

where the second inequality follows from the fact that $\overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s',i_f}(\boldsymbol{\xi}_{f,i_f})$ for $s, s' \in S^{b,0}_{f,i_f}$, which is implied by Lemma E.1 together with the definition of $S^{b,0}_{f,i_f}$. Then by Lemma F.4, we have that

$$\frac{\ell'^{b,\tilde{t}-1}_{f,i_f}}{\ell'^{b,\tilde{t}-1}_{f,i'_f}} \leq \exp\big(y_{f,i_f} f(\mathbf{W}^{b,\tilde{t}-1}, \mathbf{x}_{f,i_f}) - y_{f,i'_f} f(\mathbf{W}^{b,\tilde{t}-1}, \mathbf{x}_{f,i'_f})\big) \leq \exp(-0.36\kappa) < 0.2.$$

According to Lemma D.2, we have that

$$-\ell'^{b,\tilde{t}-1}_{f,i_f}\|\boldsymbol{\xi}_{f,i_f}\|_2^2 \leq -0.2\ell'^{b,\tilde{t}-1}_{f,i'_f} \cdot 3\sigma_p^2 d/2 \leq \ell'^{b,\tilde{t}-1}_{f,i'_f} \cdot \sigma_p^2 d/2 \leq -\ell'^{b,\tilde{t}-1}_{f,i'_f}\|\boldsymbol{\xi}_{f,i'_f}\|_2^2.$$

Then it follows that

$$\overline{\rho}^{b,\tilde{t}}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) \leq \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{b,\tilde{t}-1}_{y_{f,i'_f},s',i'_f}(\boldsymbol{\xi}_{f,i'_f}) \leq \kappa,$$

which completes the proof. □

Regarding the comparison between signal learning and noise memorization, Kou et al. (2023b) established the following lemma:

Lemma F.6 (Kou et al. (2023b))**.** *Under the same conditions as Lemma F.1, we have*

$$\sum_{i_f=1}^{n_f} \overline{\rho}^{b,t}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f}) / \gamma^{b,t}_{j',s'}(\boldsymbol{\mu}_f) = \Theta(\mathrm{SNR}_{\mathrm{f}}^{-2}) \tag{F.5}$$

for all $j, j' \in \{\pm 1\}$, $s, s' \in [m]$ and $t_1 \leq t \leq t_{end}$, where $\mathrm{SNR}_{\mathrm{f}} = \|\boldsymbol{\mu}_f\|_2/(\sigma_p\sqrt{d})$ and $t_1 = C_3\eta^{-1}nm\sigma_p^{-2}d^{-1}$.

Based on Lemma F.6, we provide a more refined version:

Lemma F.7. *Under the same conditions as Lemma F.1, we have*

$$\overline{\rho}^{b,t}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f}) / \gamma^{b,t}_{j',s'}(\boldsymbol{\mu}_f) = \Theta(n_f^{-1}\mathrm{SNR}_{\mathrm{f}}^{-2}) \tag{F.6}$$

for $j = y_{f,i_f}, j' \in \{\pm 1\}$, $s \in S^{b,0}_{f,i_f}, s' \in [m]$ and $t_1 \leq t \leq t_{end}$, where $\mathrm{SNR}_{\mathrm{f}} = \|\boldsymbol{\mu}_f\|_2/(\sigma_p\sqrt{d})$ and $t_1 = C_3\eta^{-1}nm\sigma_p^{-2}d^{-1}$.

Proof of Lemma F.7. We prove the claim by induction. As shown in Lemma D.1 of Kou et al. (2023b), we have $\overline{\rho}^{b,t}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f}) = \Theta(1)$ and $\gamma^{b,t}_{j',s'}(\boldsymbol{\mu}_f) = \Theta(n_f\mathrm{SNR}_{\mathrm{f}}^2)$ at t_1 for $s \in S^{b,0}_{f,i_f}, s' \in [m]$, and thus (F.6) holds. Now suppose that there exists $\tilde{t} \in [t_1, t_{end}]$ such that (F.6) holds for all $t \in [t_1, \tilde{t}-1]$. Recalling Lemma E.1, we have that

$$\gamma^{b,t+1}_{j',s'}(\boldsymbol{\mu}_f) = \gamma^{b,t}_{j',s'}(\boldsymbol{\mu}_f) - \frac{\eta}{m(n_f+n_r)} \cdot \sum_{i_f=1}^{n_f} \ell'^{b,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}^{b,t}_{j,s}, y_{f,i_f} \cdot \boldsymbol{\mu}_f\rangle) \cdot \|\boldsymbol{\mu}_f\|_2^2,$$

$$\overline{\rho}^{b,t+1}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}^{b,t}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta}{m(n_f+n_r)} \cdot \ell'^{b,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}^{b,t}_{j,s}, \boldsymbol{\xi}_{f,i_f}\rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = j).$$

Then we have that

$$\begin{aligned}
\gamma^{b,\tilde{t}}_{j',s'}(\boldsymbol{\mu}_f) &= \gamma^{b,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f) - \frac{\eta}{m(n_f+n_r)} \sum_{i_f \in S_{f,1}} \ell'^{b,\tilde{t}-1}_{f,i_f} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\geq \gamma^{b,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f) + \frac{\eta n_f \|\boldsymbol{\mu}_f\|_2^2}{4m(n_f+n_r)} \min_{i_f \in S_{f,1}} |\ell'^{b,\tilde{t}-1}_{f,i_f}|, \text{ if } \langle \mathbf{w}^{b,\tilde{t}-1}_{j',s'}, \boldsymbol{\mu}_f\rangle \geq 0, \\
\gamma^{b,\tilde{t}}_{j',s'}(\boldsymbol{\mu}_f) &= \gamma^{b,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f) - \frac{\eta}{m(n_f+n_r)} \sum_{i_f \in S_{f,-1}} \ell'^{b,\tilde{t}-1}_{f,i_f} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\geq \gamma^{b,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f) + \frac{\eta n_f \|\boldsymbol{\mu}_f\|_2^2}{4m(n_f+n_r)} \min_{i_f \in S_{f,-1}} |\ell'^{b,\tilde{t}-1}_{f,i_f}|, \text{ if } \langle \mathbf{w}^{b,\tilde{t}-1}_{j',s'}, \boldsymbol{\mu}_f\rangle < 0,
\end{aligned} \tag{F.7}$$

where the inequality is by Lemma D.1. And we have that

$$\overline{\rho}_{j,s,i_f}^{b,\tilde{t}}(\boldsymbol{\xi}_{f,i_f}) \leq \overline{\rho}_{j,s,i_f}^{b,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) + \frac{3\eta\sigma_p^2 d}{2m(n_f+n_r)} \cdot |\ell_{f,i_f}'^{b,\tilde{t}-1}|, \tag{F.8}$$

where the inequality is by Lemma D.2. Then by (F.7), (F.8) and Lemma F.3, we have

$$\frac{\overline{\rho}_{j,s,i_f}^{b,\tilde{t}}(\boldsymbol{\xi}_{f,i_f})}{\gamma_{j',s'}^{b,\tilde{t}}(\boldsymbol{\mu}_f)} \leq \max\left\{\frac{\overline{\rho}_{j,s,i_f}^{b,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f})}{\gamma_{j',s'}^{b,\tilde{t}-1}(\boldsymbol{\mu}_f)}, \frac{6C_4\sigma_p^2 d}{n_f\|\boldsymbol{\mu}_f\|_2^2}\right\} = \Theta(n_f^{-1}\mathrm{SNR}_f^{-2}). \tag{F.9}$$

On the other hand,

$$\begin{aligned}
\gamma_{j',s'}^{b,\tilde{t}}(\boldsymbol{\mu}_f) &= \gamma_{j',s'}^{b,\tilde{t}-1}(\boldsymbol{\mu}_f) - \frac{\eta}{m(n_f+n_r)} \sum_{i_f\in S_{f,1}} \ell_{f,i_f}'^{b,\tilde{t}-1} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\leq \gamma_{j',s'}^{b,\tilde{t}-1}(\boldsymbol{\mu}_f) + \frac{3\eta n_f\|\boldsymbol{\mu}_f\|_2^2}{4m(n_f+n_r)} \max_{i_f\in S_{f,1}} |\ell_{f,i_f}'^{b,\tilde{t}-1}|, \text{ if } \langle \mathbf{w}_{j',s'}^{b,\tilde{t}-1}, \boldsymbol{\mu}_f\rangle \geq 0, \\
\gamma_{j',s'}^{b,\tilde{t}}(\boldsymbol{\mu}_f) &= \gamma_{j',s'}^{b,\tilde{t}-1}(\boldsymbol{\mu}_f) - \frac{\eta}{m(n_f+n_r)} \sum_{i_f\in S_{f,-1}} \ell_{f,i_f}'^{b,\tilde{t}-1} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\leq \gamma_{j',s'}^{b,\tilde{t}-1}(\boldsymbol{\mu}_f) + \frac{3\eta n_f\|\boldsymbol{\mu}_f\|_2^2}{4m(n_f+n_r)} \max_{i_f\in S_{f,-1}} |\ell_{f,i_f}'^{b,\tilde{t}-1}|, \text{ if } \langle \mathbf{w}_{j',s'}^{b,\tilde{t}-1}, \boldsymbol{\mu}_f\rangle < 0,
\end{aligned} \tag{F.10}$$

where the inequality is by Lemma D.1. We also have that

$$\begin{aligned}
\overline{\rho}_{j,s,i_f}^{b,\tilde{t}}(\boldsymbol{\xi}_{f,i_f}) &= \overline{\rho}_{j,s,i_f}^{b,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta}{m(n_f+n_r)} \cdot \ell_{f,i_f}'^{b,\tilde{t}-1} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{b,\tilde{t}-1}, \boldsymbol{\xi}_{f,i_f}\rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \\
&= \overline{\rho}_{j,s,i_f}^{b,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta}{m(n_f+n_r)} \cdot \ell_{f,i_f}'^{b,\tilde{t}-1} \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \\
&\geq \overline{\rho}_{j,s,i_f}^{b,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta\sigma_p^2 d}{2m(n_f+n_r)} \cdot |\ell_{f,i_f}'^{b,\tilde{t}-1}|,
\end{aligned} \tag{F.11}$$

where the inequality is by Lemma D.2. Then by (F.10), (F.11) and Lemma F.3, we have

$$\frac{\overline{\rho}_{j,s,i_f}^{b,\tilde{t}}(\boldsymbol{\xi}_{f,i_f})}{\gamma_{j',s'}^{b,\tilde{t}}(\boldsymbol{\mu}_f)} \geq \min\left\{\frac{\overline{\rho}_{j,s,i_f}^{b,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f})}{\gamma_{j',s'}^{b,\tilde{t}-1}(\boldsymbol{\mu}_f)}, \frac{2\sigma_p^2 d}{3C_4 n_f\|\boldsymbol{\mu}_f\|_2^2}\right\} = \Theta(n_f^{-1}\mathrm{SNR}_f^{-2}). \tag{F.12}$$

By combining (F.9) and (F.12), we have that

$$\frac{\overline{\rho}_{j,s,i_f}^{b,\tilde{t}}(\boldsymbol{\xi}_{f,i_f})}{\gamma_{j',s'}^{b,\tilde{t}}(\boldsymbol{\mu}_f)} = \Theta(n_f^{-1}\mathrm{SNR}_f^{-2}),$$

which completes the induction. □

G Decoupling with a Three-stage Analysis

In this section, we provide a three-stage theoretical analysis of the overcorrection phenomenon in machine unlearning, characterizing when the KL divergence enters a small regime and when it leaves this regime.

G.1 First Stage

We begin by studying the first stage, aiming to characterize how long it takes for the KL divergence to satisfy $\mathrm{D}_{\mathrm{KL}}(f^u||f^r) \leq 2^{-1}C^{-2}$. To this end, we first present the following lemma, which provides a condition for $\mathrm{D}_{\mathrm{KL}}(f^u||f^r) \leq 2^{-1}C^{-2}$.

Lemma G.1. *Suppose that $|y_{f,i_f}f^u(\mathbf{x}_{f,i_f}) - y_{f,i_f}f^r(\mathbf{x}_{f,i_f})| \leq 2C^{-1}$ for all $i_f \in [n_f]$, then we have*

$$\mathrm{D}_{\mathrm{KL}}(f^u||f^r) = \frac{1}{n_f}\sum_{i_f=1}^{n_f}\Big[p_{i_f}^u \log\frac{p_{i_f}^u}{p_{i_f}^r} + (1-p_{i_f}^u)\log\frac{1-p_{i_f}^u}{1-p_{i_f}^r}\Big] \leq \frac{1}{2C^2},$$

where $p_{i_f}^u = q(y_{f,i_f}f^u(\mathbf{x}_{f,i_f}))$, $p_{i_f}^r = q(y_{f,i_f}f^r(\mathbf{x}_{f,i_f}))$, $f^u(\mathbf{x}_{f,i_f}) = f(\mathbf{W}^u, \mathbf{x}_{f,i_f})$, $f^r(\mathbf{x}_{f,i_f}) = f(\mathbf{W}^r, \mathbf{x}_{f,i_f})$ and $q(z) = 1/[1+\exp(-z)]$.

Proof of Lemma G.1. For convenience, define $z^u_{i_f} \triangleq y_{f,i_f} f^u(\mathbf{x}_{f,i_f})$ and $z^r_{i_f} \triangleq y_{f,i_f} f^r(\mathbf{x}_{f,i_f})$, then $p^u_{i_f} = q(z^u_{i_f})$ and $p^r_{i_f} = q(z^r_{i_f})$. We first rewrite the KL divergence term for each sample. Let

$$\Phi(z) \triangleq \log(1 + \exp(z)),$$

then

$$\Phi'(z) = \frac{\exp(z)}{1 + \exp(z)} = \frac{1}{1 + \exp(-z)} = q(z),$$

and

$$\Phi''(z) = q(z)\big(1 - q(z)\big).$$

Since $q(z) \in (0, 1)$ for all $z \in \mathbb{R}$, we have

$$q(z)\big(1 - q(z)\big) \le \frac{1}{4},$$

and therefore

$$\Phi''(z) \le \frac{1}{4}, \quad \forall z \in \mathbb{R}.$$

Next, we show that the Bernoulli KL divergence with logistic parameters can be written as a Bregman divergence induced by Φ. For each i_f, define

$$\mathrm{KL}_{i_f} \triangleq p^u_{i_f} \log \frac{p^u_{i_f}}{p^r_{i_f}} + (1 - p^u_{i_f}) \log \frac{1 - p^u_{i_f}}{1 - p^r_{i_f}}.$$

Substituting $p^u_{i_f} = q(z^u_{i_f})$ and $p^r_{i_f} = q(z^r_{i_f})$, we obtain

$$\mathrm{KL}_{i_f} = q(z^u_{i_f}) \log \frac{q(z^u_{i_f})}{q(z^r_{i_f})} + \big(1 - q(z^u_{i_f})\big) \log \frac{1 - q(z^u_{i_f})}{1 - q(z^r_{i_f})}.$$

Using

$$q(z) = \frac{\exp(z)}{1 + \exp(z)}, \quad 1 - q(z) = \frac{1}{1 + \exp(z)},$$

we have

$$\begin{aligned}
\log \frac{q(z^u_{i_f})}{q(z^r_{i_f})} &= \log \frac{\exp(z^u_{i_f})}{1 + \exp(z^u_{i_f})} - \log \frac{\exp(z^r_{i_f})}{1 + \exp(z^r_{i_f})} \\
&= z^u_{i_f} - z^r_{i_f} - \log(1 + \exp(z^u_{i_f})) + \log(1 + \exp(z^r_{i_f})) \\
&= z^u_{i_f} - z^r_{i_f} - \Phi(z^u_{i_f}) + \Phi(z^r_{i_f}),
\end{aligned}$$

and similarly,

$$\begin{aligned}
\log \frac{1 - q(z^u_{i_f})}{1 - q(z^r_{i_f})} &= \log \frac{1}{1 + \exp(z^u_{i_f})} - \log \frac{1}{1 + \exp(z^r_{i_f})} \\
&= -\Phi(z^u_{i_f}) + \Phi(z^r_{i_f}).
\end{aligned}$$

Hence,

$$\mathrm{KL}_{i_f} = q(z^u_{i_f})\Big[z^u_{i_f} - z^r_{i_f} - \Phi(z^u_{i_f}) + \Phi(z^r_{i_f})\Big] + \big(1 - q(z^u_{i_f})\big)\Big[-\Phi(z^u_{i_f}) + \Phi(z^r_{i_f})\Big],$$

Expanding the right-hand side gives

$$\mathrm{KL}_{i_f} = q(z^u_{i_f})(z^u_{i_f} - z^r_{i_f}) - \Phi(z^u_{i_f}) + \Phi(z^r_{i_f}).$$

Recalling that $\Phi'(z) = q(z)$, we further obtain

$$\mathrm{KL}_{i_f} = \Phi(z^r_{i_f}) - \Phi(z^u_{i_f}) - \Phi'(z^u_{i_f})(z^r_{i_f} - z^u_{i_f}).$$

Therefore, KL_{i_f} is exactly the Bregman divergence generated by Φ between $z^r_{i_f}$ and $z^u_{i_f}$. Now we apply Taylor's theorem to Φ at the point $z^u_{i_f}$. There exists some $\widetilde{z}_{i_f}$ lying between $z^u_{i_f}$ and $z^r_{i_f}$ such that

$$\Phi(z^r_{i_f}) = \Phi(z^u_{i_f}) + \Phi'(z^u_{i_f})(z^r_{i_f} - z^u_{i_f}) + \frac{1}{2}\Phi''(\widetilde{z}_{i_f})(z^r_{i_f} - z^u_{i_f})^2.$$

Substituting this into the above expression for KL_{i_f} yields

$$\mathrm{KL}_{i_f} = \frac{1}{2}\Phi''(\widetilde{z}_{i_f})(z^r_{i_f} - z^u_{i_f})^2.$$

Using the bound $\Phi''(\widetilde{z}_{i_f}) \le \frac{1}{4}$, we obtain

$$\mathrm{KL}_{i_f} \le \frac{1}{2}\cdot\frac{1}{4}(z^r_{i_f} - z^u_{i_f})^2 = \frac{1}{8}(z^r_{i_f} - z^u_{i_f})^2.$$

By the assumption of the lemma,

$$\left|y_{f,i_f} f^u(\mathbf{x}_{f,i_f}) - y_{f,i_f} f^r(\mathbf{x}_{f,i_f})\right| \le 2C^{-1},$$

that is,

$$|z^u_{i_f} - z^r_{i_f}| \le 2C^{-1}.$$

Thus,

$$(z^r_{i_f} - z^u_{i_f})^2 \le 4C^{-2},$$

and hence

$$\mathrm{KL}_{i_f} \le \frac{1}{8}\cdot 4C^{-2} = \frac{1}{2C^2}.$$

Finally, averaging over all $i_f \in [n_f]$, we have

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(f^u \| f^r) &= \frac{1}{n_f}\sum_{i_f=1}^{n_f}\Big[p^u_{i_f}\log\frac{p^u_{i_f}}{p^r_{i_f}} + (1-p^u_{i_f})\log\frac{1-p^u_{i_f}}{1-p^r_{i_f}}\Big] \\ &= \frac{1}{n_f}\sum_{i_f=1}^{n_f}\mathrm{KL}_{i_f} \le \frac{1}{n_f}\sum_{i_f=1}^{n_f}\frac{1}{2C^2} = \frac{1}{2C^2}.\end{aligned}$$

This completes the proof. □

Lemma G.1 indicates that the KL divergence lies in a small regime is closely related to the quantity $|y_{f,i_f} f^u(\mathbf{x}_{f,i_f}) - y_{f,i_f} f^r(\mathbf{x}_{f,i_f})|$. Since $y_{f,i_f} f^r(\mathbf{x}_{f,i_f})$ remains unchanged during the unlearning process, we first focus on analyzing retrained model f^r and $y_{f,i_f} f^r(\mathbf{x}_{f,i_f})$.

Lemma G.2 (Kou et al. (2023b)). *For any $\epsilon > 0$, under Condition 3.1, suppose that $n_r\|\boldsymbol{\mu}_r\|_2^4 \ge C_1\sigma_p^4 d$. Then with probability at least $1-\delta$, there exists $t_{end} = \tilde{O}(\eta^{-1}\epsilon^{-1}mn_r d^{-1}\sigma_p^{-2})$ such that: $L_{S_r}(\mathbf{W}^{r,t_{end}}) \le \epsilon$ and $L^{0-1}_{\mathcal{D}_r}(\mathbf{W}^{r,t_{end}}) \le \exp\Big(-n\|\boldsymbol{\mu}_r\|_2^4/(C_2\sigma_p^4 d)\Big)$, where C_1, C_2 are some absolute constants.*

Proposition G.3 (Kou et al. (2023b)). *Let $\beta = 2\max_{j,s,i_r}\{|\mathbf{w}^{r,0}_{j,s}, \boldsymbol{\mu}_r\rangle|, |\langle \mathbf{w}^{r,0}_{j,s}, \boldsymbol{\xi}_{r,i_r}\rangle|\}$ and $\alpha = 4\log(T^*)$, where $T^* = \eta^{-1}\,\mathrm{poly}(n_r, m, d^{-1}, \epsilon^{-1})$. Given Condition 3.1 and $\epsilon > 0$, then for all $0 \le t \le T^*$, $i_r \in [n_r]$, $s \in [m]$, and $j \in \{\pm 1\}$, the following bounds hold:*

$$0 \le \gamma^{r,t}_{j,s}(\boldsymbol{\mu}_r) \le C' n_r \mathrm{SNR}_{\mathrm{r}}^2\alpha, \quad 0 \le \overline{\rho}^{r,t}_{j,s,i_r}(\boldsymbol{\xi}_{r,i_r}) \le \alpha, \tag{G.1}$$

$$0 \ge \underline{\rho}^{r,t}_{j,s,i_r}(\boldsymbol{\xi}_{r,i_r}) \ge -\beta - 10\sqrt{\frac{\log(6n_r^2/\delta)}{d}}n_r\alpha \ge -\alpha, \tag{G.2}$$

where C' is a positive constant, and $\mathrm{SNR}_{\mathrm{r}} = \|\boldsymbol{\mu}_r\|_2/(\sigma_p\sqrt{d})$.

Lemma G.4. *Under the same conditions as Lemma G.2, we have that*

$$y_{f,i_f} f^{r,t_{end}}(\mathbf{x}_{f,i_f}) = y_{f,i_f} f(\mathbf{W}^{r,t_{end}}, \mathbf{x}_{f,i_f}) \le C^{-1},$$

where C is a sufficiently large constant, as specified in Condition 3.1.

Proof of Lemma G.4. Recalling (E.12), we have

$$\mathbf{w}^{r,t}_{j,s} = \mathbf{w}^{r,0}_{j,s} + j\gamma^{r,t}_{j,s}(\boldsymbol{\mu}_r)\frac{\boldsymbol{\mu}_r}{\|\boldsymbol{\mu}_r\|_2^2} + \sum_{i_r=1}^{n_r}\overline{\rho}^{r,t}_{j,s,i_r}(\boldsymbol{\xi}_{r,i_r})\frac{\boldsymbol{\xi}_{r,i_r}}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2} + \sum_{i_r=1}^{n_r}\underline{\rho}^{r,t}_{j,s,i_r}(\boldsymbol{\xi}_{r,i_r})\frac{\boldsymbol{\xi}_{r,i_r}}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2}.$$

Then we have that

$$
\begin{aligned}
|\langle \mathbf{w}_{y_{f,i_f},s}^{r,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle| &= |\langle \mathbf{w}_{y_{f,i_f},s}^{r,0}, y_{f,i_f}\boldsymbol{\mu}_f\rangle + \sum_{i_r=1}^{n_r} \rho_{y_{f,i_f},s,i_r}^{r,t_{end}}(\boldsymbol{\xi}_{r,i_r})\frac{\langle \boldsymbol{\xi}_{r,i_r}, y_{f,i_f}\boldsymbol{\mu}_f\rangle}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2}| \\
&\le |\langle \mathbf{w}_{y_{f,i_f},s}^{r,0}, y_{f,i_f}\boldsymbol{\mu}_f\rangle| + \sum_{i_r=1}^{n_r} |\rho_{y_{f,i_f},s,i_r}^{r,t_{end}}(\boldsymbol{\xi}_{r,i_r})|\frac{|\langle \boldsymbol{\xi}_{r,i_r}, y_{f,i_f}\boldsymbol{\mu}_f\rangle|}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2} \\
&\le \sqrt{2\log(12m/\delta)}\sigma_0\|\boldsymbol{\mu}_f\|_2 + \frac{2\|\boldsymbol{\mu}_f\|_2\sigma_p\sqrt{2\log(6n_r/\delta)}n_r\alpha}{\sigma_p^2 d} \\
&\le C^{-1},
\end{aligned}
$$

where the second inequality is by Lemma D.2 and Lemma D.3, the third inequality is by Condition 3.1. Then we also have that

$$
\begin{aligned}
|\langle \mathbf{w}_{y_{f,i_f},s}^{r,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle| &\le |\langle \mathbf{w}_{y_{f,i_f},s}^{r,0}, \boldsymbol{\xi}_{f,i_f}\rangle| + |\gamma_{y_{f,i_f},s}^{r,t_{end}}(\boldsymbol{\mu}_r)|\frac{|\langle \boldsymbol{\mu}_r, \boldsymbol{\xi}_{f,i_f}\rangle|}{\|\boldsymbol{\mu}_r\|_2^2} \\
&\quad + \sum_{i_r=1}^{n_r} |\rho_{y_{f,i_f},s,i_r}^{r,t_{end}}(\boldsymbol{\xi}_{r,i_r})|\frac{|\langle \boldsymbol{\xi}_{r,i_r}, \boldsymbol{\xi}_{f,i_f}\rangle|}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2} \\
&\le 2\sqrt{\log(12mn_f/\delta)}\sigma_0\sigma_p\sqrt{d} + \frac{n_r\|\boldsymbol{\mu}_r\|_2^2\alpha}{\sigma_p^2 d}\frac{\sigma_p\sqrt{2\log(6n_f/\delta)}}{\|\boldsymbol{\mu}_r\|_2} \\
&\quad + \frac{4\sigma_p^2\sqrt{d\log(6n^2/\delta)}n_r\alpha}{\sigma_p^2 d} \\
&\le C^{-1}
\end{aligned}
$$

where the second inequality is by Proposition G.3 Lemma D.2 and Lemma D.3, the third inequality is by Condition 3.1. Similarly,

$$
|\langle \mathbf{w}_{-y_{f,i_f},s}^{r,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle| \le C^{-1}, \quad |\langle \mathbf{w}_{-y_{f,i_f},s}^{r,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle| \le C^{-1}
$$

Then it follows from the definition of the two-layer CNN model that

$$
\begin{aligned}
y_{f,i_f} f^{r,t_{end}}(\mathbf{x}_{f,i_f}) &= \frac{1}{m}\sum_{s=1}^{m}\big[\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{r,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle) + \sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{r,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle)\big] \\
&\quad - \frac{1}{m}\sum_{s=1}^{m}\big[\sigma(\langle \mathbf{w}_{-y_{f,i_f},s}^{r,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle) + \sigma(\langle \mathbf{w}_{-y_{f,i_f},s}^{r,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle)\big] \\
&\le C^{-1},
\end{aligned}
$$

which completes the proof. □

We next characterize the initialization of $y_{f,i_f} f^u(\mathbf{x}_{f,i_f})$. Prior to this, we introduce several key lemmas.

Lemma G.5. *Under the same conditions as Lemma F.1, we have that* $\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle = \Theta(\gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f))$, *for* $s \in [m]$.

Proof of Lemma G.5. By Condition 3.1, Lemma D.2 and Lemma D.3, we have

$$
\begin{aligned}
&|\langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, y_{f,i_f}\boldsymbol{\mu}_f\rangle| + |\sum_{i_f=1}^{n_f}\rho_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})\frac{\langle \boldsymbol{\xi}_{f,i_f}, y_{f,i_f}\boldsymbol{\mu}_f\rangle}{\|\boldsymbol{\xi}_{f,i_f}\|_2^2}| \\
&\quad + |\sum_{i_r=1}^{n_r}\rho_{y_{f,i_f},s,i_r}^{b,t_{end}}(\boldsymbol{\xi}_{r,i_r})\frac{\langle \boldsymbol{\xi}_{r,i_r}, y_{f,i_f}\boldsymbol{\mu}_f\rangle}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2}| \\
&\le \sqrt{2\log(12m/\delta)}\sigma_0\|\boldsymbol{\mu}_f\|_2 + \frac{2\|\boldsymbol{\mu}_f\|_2\sigma_p\sqrt{2\log(6n_r/\delta)}n_f\alpha}{\sigma_p^2 d} + \frac{2\|\boldsymbol{\mu}_f\|_2\sigma_p\sqrt{2\log(6n_r/\delta)}n_r\alpha}{\sigma_p^2 d}
\end{aligned}
$$

$\leq C^{-1}$

Then by (E.2), we have

$$\begin{aligned}\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle &= \langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, y_{f,i_f}\boldsymbol{\mu}_f\rangle + \gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f) + \sum_{i_f=1}^{n_f} \rho_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})\frac{\langle \boldsymbol{\xi}_{f,i_f}, y_{f,i_f}\boldsymbol{\mu}_f\rangle}{\|\boldsymbol{\xi}_{f,i_f}\|_2^2} \\ &\quad + \sum_{i_r=1}^{n_r} \rho_{y_{f,i_f},s,i_r}^{b,t_{end}}(\boldsymbol{\xi}_{r,i_r})\frac{\langle \boldsymbol{\xi}_{r,i_r}, y_{f,i_f}\boldsymbol{\mu}_f\rangle}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2} \\ &= \Theta(\gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f)),\end{aligned}$$

where the equality is by $\gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f) = \Theta(n_f \mathrm{SNR}_{\mathrm{f}}^2)$ from Kou et al. (2023b), which completes the proof. □

Using a proof technique similar to that in the proof of Lemma G.5, we derive the following lemma.

Lemma G.6. *Under the same conditions as Lemma F.1, we have that* $\langle \mathbf{w}_{-y_{f,i_f},s}^{b,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle = \Theta(-\gamma_{-y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f))$*, for* $s \in [m]$.

Lemma G.7. *Under the same conditions as Lemma F.1, we have that* $\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle = \Theta(\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}))$*, for* $s \in S_{f,i_f}^{b,0}$.

Proof of Lemma G.7. By Condition 3.1, Lemma D.2 and Lemma D.3, we have

$$\begin{aligned}&|\langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle| + |\gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f)\frac{\langle \boldsymbol{\mu}_f, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\mu}_f\|_2^2}| + |\gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_r)\frac{\langle \boldsymbol{\mu}_r, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\mu}_r\|_2^2}| \\ &\quad + |\sum_{i'_f \neq i_f} \overline{\rho}_{y_{f,i'_f},s,i'_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i'_f})\frac{\langle \boldsymbol{\xi}_{f,i'_f}, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\xi}_{f,i'_f}\|_2^2}| + |\sum_{i_r=1}^{n_r} \overline{\rho}_{y_{f,i_f},s,i_r}^{b,t_{end}}(\boldsymbol{\xi}_{r,i_r})\frac{\langle \boldsymbol{\xi}_{r,i_r}, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2}| \\ &\leq 2\sqrt{\log(12mn_f/\delta)}\sigma_0\sigma_p\sqrt{d} + \frac{n_f\|\boldsymbol{\mu}_f\|_2^2\alpha}{\sigma_p^2 d}\frac{\sigma_p\sqrt{2\log(6n_f/\delta)}}{\|\boldsymbol{\mu}_f\|_2} + \frac{n_r\|\boldsymbol{\mu}_r\|_2^2\alpha}{\sigma_p^2 d}\frac{\sigma_p\sqrt{2\log(6n_f/\delta)}}{\|\boldsymbol{\mu}_r\|_2} \\ &\quad + \frac{4\sigma_p^2\sqrt{d\log(6n_f^2/\delta)}n_f\alpha}{\sigma_p^2 d} + \frac{4\sigma_p^2\sqrt{d\log(6n^2/\delta)}n_r\alpha}{\sigma_p^2 d} \leq C^{-1}\end{aligned}$$

Then by (E.2), we have

$$\begin{aligned}\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle &= \langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle + \gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f)\frac{\langle \boldsymbol{\mu}_f, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\mu}_f\|_2^2} + \gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_r)\frac{\langle \boldsymbol{\mu}_r, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\mu}_r\|_2^2} \\ &\quad + \overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}) + \sum_{i'_f \neq i_f} \overline{\rho}_{y_{f,i'_f},s,i'_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i'_f})\frac{\langle \boldsymbol{\xi}_{f,i'_f}, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\xi}_{f,i'_f}\|_2^2} \\ &\quad + \sum_{i_r=1}^{n_r} \overline{\rho}_{y_{f,i_f},s,i_r}^{b,t_{end}}(\boldsymbol{\xi}_{r,i_r})\frac{\langle \boldsymbol{\xi}_{r,i_r}, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2} \\ &= \Theta(\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}))\end{aligned}$$

where the equality is by $\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}) = \Theta(1)$ for $s \in S_{f,i_f}^{b,0}$ from Kou et al. (2023b), which completes the proof. □

Lemma G.8. *Under the same conditions as Lemma F.1, we have* $|\langle \mathbf{w}_{-y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle| \leq C^{-1}$ *for* $s \in [m]$*, where* C *is a sufficiently large constant, as specified in Condition 3.1.*

Proof of Lemma G.8. It has been shown in Lemma D.6 of Kou et al. (2023b) that $\max_{j,s,i_f} |\underline{\rho}_{j,s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})| = \max\big\{O\big(\sqrt{\log(mn_f/\delta)}\sigma_0\sigma_p\sqrt{d}\big), O\big(n_f\sqrt{\log(n_f/\delta)}\log(T^*)/\sqrt{d}\big)\big\} = O(C^{-1})$. Then by (E.2), Condition 3.1, Lemma D.2 and Lemma D.3, we have

$$
\begin{aligned}
|\langle \mathbf{w}_{-y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle| &\le |\langle \mathbf{w}_{-y_{f,i_f},s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle| + |\gamma_{-y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f)\frac{\langle \boldsymbol{\mu}_f, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\mu}_f\|_2^2}| \\
&\quad + |\gamma_{-y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_r)\frac{\langle \boldsymbol{\mu}_r, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\mu}_r\|_2^2}| + |\sum_{i_r=1}^{n_r}\underline{\rho}_{-y_{f,i_f},s,i_r}^{b,t_{end}}(\boldsymbol{\xi}_{r,i_r})\frac{\langle \boldsymbol{\xi}_{r,i_r}, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\xi}_{r,i_r}\|_2^2}| \\
&\quad + |\underline{\rho}_{-y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})| + |\sum_{i'_f\neq i_f}\underline{\rho}_{-y_{f,i'_f},s,i'_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i'_f})\frac{\langle \boldsymbol{\xi}_{f,i'_f}, \boldsymbol{\xi}_{f,i_f}\rangle}{\|\boldsymbol{\xi}_{f,i'_f}\|_2^2}| \\
&\le 2\sqrt{\log(12mn_f/\delta)}\sigma_0\sigma_p\sqrt{d} + \frac{n_f\|\boldsymbol{\mu}_f\|_2^2\alpha}{\sigma_p^2 d}\frac{\sigma_p\sqrt{2\log(6n_f/\delta)}}{\|\boldsymbol{\mu}_f\|_2} \\
&\quad + \frac{n_r\|\boldsymbol{\mu}_r\|_2^2\alpha}{\sigma_p^2 d}\frac{\sigma_p\sqrt{2\log(6n_f/\delta)}}{\|\boldsymbol{\mu}_r\|_2} + |\underline{\rho}_{-y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})| \\
&\quad + \frac{4\sigma_p^2\sqrt{d\log(6n_f^2/\delta)}n_f\alpha}{\sigma_p^2 d} + \frac{4\sigma_p^2\sqrt{d\log(6n^2/\delta)}n_r\alpha}{\sigma_p^2 d} \le C^{-1},
\end{aligned}
$$

which completes the proof. □

Lemma G.9. *Under the same conditions as Lemma F.1, we have that*

$$
\frac{1}{m}\sum_{s=1}^{m}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle) = \Theta\Big(\sum_{s\in S_{f,i_f}^{b,0}}\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}|\Big),
$$

where $S_{f,i_f}^{b,0} = \{s\in[m] : \langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$.

Proof of Lemma G.9. On the one hand,

$$
\frac{1}{m}\sum_{s=1}^{m}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle) \le \frac{\sum_{s\in S_{f,i_f}^{b,0}}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle)}{|S_{f,i_f}^{b,0}|}.
$$

Then by Lemma F.5 and Lemma G.7, we have

$$
\frac{1}{m}\sum_{s=1}^{m}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle) = O\Big(\sum_{s\in S_{f,i_f}^{b,0}}\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}|\Big).
$$

On the other hand,

$$
\frac{1}{m}\sum_{s=1}^{m}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle) \ge \frac{1}{m}\sum_{s\in S_{f,i_f}^{b,0}}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle) \ge 0.4\frac{\sum_{s\in S_{f,i_f}^{b,0}}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle)}{|S_{f,i_f}^{b,0}|}
$$

Then by Lemma F.5 and Lemma G.7, we have

$$
\frac{1}{m}\sum_{s=1}^{m}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{b,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle) = \Omega\Big(\sum_{s\in S_{f,i_f}^{b,0}}\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}|\Big),
$$

which completes the proof. □

Now we present the following lemma to characterize the initialization of $y_{f,i_f}f^u(\mathbf{x}_{f,i_f})$.

Lemma G.10. *Under the same conditions as Lemma F.1, we have that*

$$y_{f,i_f} f^{u,0}(\mathbf{x}_{f,i_f}) = y_{f,i_f} f(\mathbf{W}^{b,t_{end}}, \mathbf{x}_{f,i_f}) = \Theta\Big(\sum_{s\in S_{f,i_f}^{b,0}} \overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}|\Big),$$

where $S_{f,i_f}^{b,0} = \{s \in [m] : \langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$.

Proof of Lemma G.10. Recalling Lemma F.7, we have

$$\gamma_{y_{f,i_f},s}^{b,t_{end}}(\boldsymbol{\mu}_f) = \frac{n_f\|\boldsymbol{\mu}_f\|_2^2}{\sigma_p^2 d}\overline{\rho}_{y_{f,i_f},s',i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}) \le \frac{1}{C\log(T^*)}\overline{\rho}_{y_{f,i_f},s',i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}),$$

where the inequality is by Condition 3.1. Then we have that

$$\begin{aligned}\frac{1}{m}\sum_{s=1}^{m}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{r,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle) &\le \max_{s\in[m]}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{r,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle) = \gamma_{y_{f,i_f},s*}^{b,t_{end}}(\boldsymbol{\mu}_f)\\ &\le \frac{1}{C\log(T^*)}\overline{\rho}_{y_{f,i_f},s',i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}),\end{aligned}$$

where the equality is by Lemma G.5. Then by the definition of the two-layer CNN, we have

$$\begin{aligned}y_{f,i_f} f^{r,t_{end}}(\mathbf{x}_{f,i_f}) &= \frac{1}{m}\sum_{s=1}^{m}\big[\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{r,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle) + \sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{r,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle)\big]\\ &\quad - \frac{1}{m}\sum_{s=1}^{m}\big[\sigma(\langle \mathbf{w}_{-y_{f,i_f},s}^{r,t_{end}}, y_{f,i_f}\boldsymbol{\mu}_f\rangle) + \sigma(\langle \mathbf{w}_{-y_{f,i_f},s}^{r,t_{end}}, \boldsymbol{\xi}_{f,i_f}\rangle)\big]\\ &= \Theta\Big(\sum_{s\in S_{f,i_f}^{b,0}}\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}|\Big),\end{aligned}$$

where the second equality is by Lemma G.6, G.8 and G.9, which completes the proof. □

By Lemma G.10, it suffices to characterize the time required for $\sum_{s\in S_{f,i_f}^{b,0}}\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}|$ to reach $\sum_{s\in S_{f,i_f}^{b,0}}\overline{\rho}_{y_{f,i_f},s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}| \le C^{-1}$. Before proceeding, we present several key lemmas.

Proposition G.11. *Let* $\beta = 2\max_{j,s,i_f,i_r}\{|\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_f\rangle|, |\mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_r\rangle|, |\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle|, |\langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{r,i_r}\rangle|\}$ *and* $\alpha = 4\log(T^*)$*, where* $T^* = \eta^{-1}\,\mathrm{poly}(n, m, d^{-1}, \epsilon^{-1})$ *and* $n = \max\{n_f, n_r\}$*. Given Condition 3.1 and* $\epsilon > 0$*, then for the first stage and* $i_f \in [n_f]$*,* $s \in [m]$*, and* $j \in \{\pm 1\}$*, the following bounds hold:*

$$0 \le \gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f) \le C' n_f \mathrm{SNR}_{\mathrm{f}}^2\alpha, \quad 0 \le \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) \le \alpha, \tag{G.3}$$

$$0 \ge \underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) \ge -\beta - 20\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \ge -\alpha. \tag{G.4}$$

where C' *is a positive constant, and* $\mathrm{SNR}_{\mathrm{f}} = \|\boldsymbol{\mu}_f\|_2/(\sigma_p\sqrt{d})$.

Before proving Proposition G.11, we introduce several technical lemmas.

Lemma G.12 (Kou et al. (2023b))**.** *Under Condition 3.1, suppose* (G.3) *and* (G.4) *hold at iteration* t*. Then, for all* $s \in [m]$*,* $j \in \{\pm 1\}$ *and* $i_f \in [n_f]$*,*

$$\big|\langle \mathbf{w}_{j,s}^{u,t} - \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\mu}_f\rangle - j\cdot\gamma_{j,s}^{u,t}\big| \le 2\mathrm{SNR}_{\mathrm{f}}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, \tag{G.5}$$

$$\big|\langle \mathbf{w}_{j,s}^{u,t} - \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle - \underline{\rho}_{j,s,i_f}^{u,t}\big| \le 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,\ j \ne y_{f,i_f}, \tag{G.6}$$

$$\big|\langle \mathbf{w}_{j,s}^{u,t} - \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle - \overline{\rho}_{j,s,i_f}^{u,t}\big| \le 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,\ j = y_{f,i_f}, \tag{G.7}$$

where $n = \max\{n_f, n_r\}$.

Then by (G.5), (G.6) and (G.7), we have the following lemma:

Lemma G.13 (Kou et al. (2023b))**.** *Under Condition 3.1, suppose* (G.3) *and* (G.4) *hold at iteration t. Then, for all $j \neq y_i$, $j \in \{\pm 1\}$ and $i_f \in [n_f]$, $F_j(\mathbf{W}_j^{u,t}, \mathbf{x}_{f,i_f}) \leq 0.5$.*

Lemma G.14 (Kou et al. (2023b))**.** *Under Condition 3.1, suppose* (G.3) *and* (G.4) *hold at iteration t. Then we have*

$$\langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle \geq -0.25,$$
$$\langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle \leq \sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle) \leq \langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle + 0.25,$$

for any $i_f \in [n_f]$.

Lemma G.15. *Under Condition 3.1, suppose* (G.3) *and* (G.4) *hold for any iteration $t' \leq t$, the following properties hold for any iteration $t' \leq t$:*

1. $\sum_{s=1}^{m} \big[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t'}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t'}(\boldsymbol{\xi}_{f,i'_f})\big] \leq \kappa$ *for all* $i_f, i'_f \in [n_f]$.
2. $y_{f,i_f} \cdot f(\mathbf{W}^{u,t'}, \mathbf{x}_{f,i_f}) - y_{f,i'_f} \cdot f(\mathbf{W}^{u,t'}, \mathbf{x}_{f,i'_f}) \leq C_6$ *for all* $i_f, i'_f \in [n_f]$,
3. $\ell'^{u,t'}_{f,i_f} / \ell'^{u,t'}_{f,i'_f} \leq C_3 = \exp(C_6)$ *for all* $i_f, i'_f \in [n]$.

Here, κ and C_6 are positive constants.

Proof of Lemma G.15. We prove the lemma by induction. For the base case $t = 0$, we have $\mathbf{w}_{j,s}^{u,0} = \mathbf{w}_{j,s}^{b,t_{end}}$; thus, by Lemma F.3, all three properties hold. Now suppose there exists $\tilde{t} \leq t'$ such that these properties hold for all $0 \leq t' \leq \tilde{t} - 1$. We aim to show that they also hold at $t' = \tilde{t}$.

We begin by showing $y_{f,i_f} f(\mathbf{W}^{u,t'}, \mathbf{x}_{f,i_f}) - y_{f,i'_f} f(\mathbf{W}^{u,t'}, \mathbf{x}_{f,i'_f})$ can be approximated by $\sum_{s=1}^{m} \big[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t'}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t'}(\boldsymbol{\xi}_{f,i'_f})\big]$

$$\begin{aligned}
&y_{f,i_f} f(\mathbf{W}^{u,t'}, \mathbf{x}_{f,i_f}) - y_{f,i'_f} f(\mathbf{W}^{u,t'}, \mathbf{x}_{f,i'_f}) \\
&= F_{-y_{f,i'_f}}(\mathbf{W}_{-y_{f,i'_f}}^{u,t'}, \mathbf{x}_{f,i'_f}) - F_{-y_{f,i_f}}(\mathbf{W}_{-y_{f,i_f}}^{u,t'}, \mathbf{x}_{f,i_f}) \\
&\quad + \frac{1}{m}\sum_{s=1}^{m}[\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, y_{f,i_f}\boldsymbol{\mu}_f \rangle) + \sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i_f} \rangle)] \\
&\quad - \frac{1}{m}\sum_{s=1}^{m}[\sigma(\langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, y_{f,i'_f}\boldsymbol{\mu}_f \rangle) + \sigma(\langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i'_f} \rangle)] \\
&= \underbrace{F_{-y_{f,i'_f}}(\mathbf{W}_{-y_{f,i'_f}}^{u,t'}, \mathbf{x}_{f,i'_f}) - F_{-y_{f,i_f}}(\mathbf{W}_{-y_{f,i_f}}^{u,t'}, \mathbf{x}_{f,i_f})}_{I_1} \\
&\quad \underbrace{+ \frac{1}{m}\sum_{s=1}^{m}[\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, y_{f,i_f}\boldsymbol{\mu}_f \rangle) - \sigma(\langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, y_{f,i'_f}\boldsymbol{\mu}_f \rangle)]}_{I_2} \\
&\quad \underbrace{+ \frac{1}{m}\sum_{s=1}^{m}[\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i_f} \rangle) - \sigma(\langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i'_f} \rangle)]}_{I_3},
\end{aligned} \tag{G.8}$$

For $|I_1|$,

$$\begin{aligned}
|I_1| &\leq |F_{-y_{f,i'_f}}(\mathbf{W}_{-y_{f,i'_f}}^{u,t'}, \mathbf{x}_{f,i'_f})| + |F_{-y_{f,i_f}}(\mathbf{W}_{-y_{f,i_f}}^{u,t'}, \mathbf{x}_{f,i_f})| \\
&= F_{-y_{f,i'_f}}(\mathbf{W}_{-y_{f,i'_f}}^{u,t'}, \mathbf{x}_{f,i'_f}) - F_{-y_{f,i_f}}(\mathbf{W}_{-y_{f,i_f}}^{u,t'}, \mathbf{x}_{f,i_f}) \leq 1,
\end{aligned} \tag{G.9}$$

where the second inequality is by Lemma G.13. For $|I_2|$,

$$\begin{aligned}|I_2| &\le \max\left\{\frac{1}{m}\sum_{s=1}^{m}\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, y_{f,i_f}\boldsymbol{\mu}_f\rangle), \frac{1}{m}\sum_{s=1}^{m}\sigma(\langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, y_{f,i'_f}\boldsymbol{\mu}_f\rangle)\right\}\\ &\le 3\max\left\{\beta, C'n_f\mathrm{SNR}_{\mathrm{f}}^2\alpha, 2\mathrm{SNR}_{\mathrm{f}}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha\right\}\\ &\le 0.25,\end{aligned} \tag{G.10}$$

where the second inequality is by (G.3) and (G.5); the third inequality is by Condition 3.1. For I_3, on the one hand,

$$\begin{aligned}I_3 &= \frac{1}{m}\sum_{s=1}^{m}\left[\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i_f}\rangle) - \sigma(\langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i'_f}\rangle)\right]\\ &\le \frac{1}{m}\sum_{s=1}^{m}\left[\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i_f}\rangle - \langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i'_f}\rangle\right] + 0.25\\ &\le \frac{1}{m}\sum_{s=1}^{m}\left[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t'}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t'}(\boldsymbol{\xi}_{f,i'_f}) + 20\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha + 0.25\right]\\ &\le \frac{1}{m}\sum_{s=1}^{m}\left[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t'}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t'}(\boldsymbol{\xi}_{f,i'_f})\right] + 0.5,\end{aligned} \tag{G.11}$$

where the first inequality is by Lemma G.14; the second inequality is due to (G.7) in Lemma G.12; the last inequality is by Condition 3.1. On the other hand,

$$\begin{aligned}I_3 &= \frac{1}{m}\sum_{s=1}^{m}\left[\sigma(\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i_f}\rangle) - \sigma(\langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i'_f}\rangle)\right]\\ &\ge \frac{1}{m}\sum_{s=1}^{m}\left[\langle \mathbf{w}_{y_{f,i_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i_f}\rangle - \langle \mathbf{w}_{y_{f,i'_f},s}^{u,t'}, \boldsymbol{\xi}_{f,i'_f}\rangle\right] - 0.25\\ &\ge \frac{1}{m}\sum_{s=1}^{m}\left[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t'}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t'}(\boldsymbol{\xi}_{f,i'_f}) - 20\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha - 0.25\right]\\ &\ge \frac{1}{m}\sum_{s=1}^{m}\left[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t'}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t'}(\boldsymbol{\xi}_{f,i'_f})\right] - 0.5,\end{aligned} \tag{G.12}$$

where the first inequality is by Lemma G.14; the second inequality is due to (G.7) in Lemma G.12; the last inequality is by Condition 3.1. Plugging (G.9)-(G.11) into (G.8), we get

$$\begin{aligned}&\left|y_{f,i_f}f(\mathbf{W}^{u,t'}, \mathbf{x}_{f,i_f}) - y_{f,i'_f}f(\mathbf{W}^{u,t'}, \mathbf{x}_{f,i'_f}) - \frac{1}{m}\sum_{s=1}^{m}\left[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t'}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t'}(\boldsymbol{\xi}_{f,i'_f})\right]\right|\\ &\le 1.75.\end{aligned} \tag{G.13}$$

By Lemma E.2,

$$\overline{\rho}_{j,s,i_f}^{u,t+1}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{n_f m}\cdot \ell'^{u,t}_{f,i_f}\cdot\sigma'(\langle \mathbf{w}_{j,s}^{u,t}, \boldsymbol{\xi}_{f,i_f}\rangle)\cdot\|\boldsymbol{\xi}_{f,i_f}\|_2^2\cdot\mathbb{1}(y_{f,i_f}=j),$$

for all $j\in\{\pm 1\}$, $s\in[m]$, and $i_f\in[n_f]$. By the definition of $S_{f,i_f}^{u,t}$, we obtain

$$\begin{aligned}&\sum_{s=1}^{m}\left[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t+1}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t+1}(\boldsymbol{\xi}_{f,i'_f})\right]\\ &= \sum_{s=1}^{m}\left[\overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}_{y_{f,i'_f},s,i'_f}^{u,t}(\boldsymbol{\xi}_{f,i'_f})\right] + \frac{\eta\left(|S_{f,i_f}^{u,t}|\ell'^{u,t}_{f,i_f}\|\boldsymbol{\xi}_{f,i_f}\|_2^2 - |S_{f,i'_f}^{u,t}|\ell'^{u,t}_{f,i'_f}\|\boldsymbol{\xi}_{f,i'_f}\|_2^2\right)}{mn_f},\end{aligned}$$

for any $i_f, i'_f \in [n_f]$. We next distinguish two separate cases depending on the value of $\sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}-1}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})]$:

When $\sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}-1}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})] \leq 0.9\kappa$,

$$
\begin{aligned}
&\sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})] \\
&\leq \sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}-1}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})] - \frac{\eta |S^{u,\tilde{t}-1}_{f,i'_f}| \ell'^{u,\tilde{t}-1}_{f,i'_f} \|\boldsymbol{\xi}_{f,i'_f}\|_2^2}{mn_f} \\
&\leq \sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}-1}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})] + \frac{\eta \|\boldsymbol{\xi}_{f,i'_f}\|_2^2}{n_f} \\
&\leq 0.1\kappa + 0.9\kappa \\
&= \kappa,
\end{aligned}
$$

where the first inequality is by $\ell'^{u,\tilde{t}-1}_{f,i_f} < 0$; the second inequality is by $|S^{u,\tilde{t}-1}_{f,i'_f}| \leq m$, $-\ell'^{u,\tilde{t}-1}_{f,i_f} < 1$; the third inequality is by $\eta \leq C^{-1} n_f \sigma_p^{-2} d^{-1}$.

When $\sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}-1}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})] \geq 0.9\kappa$, setting $\kappa = 3.25$ as in Kou et al. (2023b), we obtain from the (G.13) that

$$
\begin{aligned}
&y_{f,i'_f} f(\mathbf{W}^{u,\tilde{t}-1}, \mathbf{x}_{f,i'_f}) - y_{f,i_f} f(\mathbf{W}^{u,\tilde{t}-1}, \mathbf{x}_{f,i_f}) \\
&\leq \frac{1}{m} \sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}-1}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f}) - \overline{\rho}^{u,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f})] + 1.75 \\
&\leq -0.9\kappa + 0.54\kappa \\
&= -0.36\kappa.
\end{aligned}
\tag{G.14}
$$

Then by Lemma F.4, we have

$$
\frac{\ell'^{u,\tilde{t}-1}_{f,i'_f}}{\ell'^{u,\tilde{t}-1}_{f,i_f}} \leq \exp\left(y_{f,i'_f} f(\mathbf{W}^{u,\tilde{t}-1}, \mathbf{x}_{f,i'_f}) - y_{f,i_f} f(\mathbf{W}^{u,\tilde{t}-1}, \mathbf{x}_{f,i_f})\right) \leq \exp(-0.36\kappa).
$$

Then by Lemma D.4,

$$
\frac{|S^{u,\tilde{t}-1}_{f,i'_f}| \ell'^{u,\tilde{t}-1}_{f,i'_f}}{|S^{u,\tilde{t}-1}_{f,i_f}| \ell'^{u,\tilde{t}-1}_{f,i_f}} \leq 2.5 \exp(-0.36\kappa) \leq 0.8.
$$

Then by Lemma D.3, we have

$$
-\ell'^{u,\tilde{t}-1}_{f,i'_f} |S^{u,\tilde{t}-1}_{f,i'_f}| \|\boldsymbol{\xi}_{f,i'_f}\|_2^2 \leq -\ell'^{u,\tilde{t}-1}_{f,i_f} |S^{u,\tilde{t}-1}_{f,i_f}| \|\boldsymbol{\xi}_{f,i_f}\|_2^2.
$$

Then it follows that

$$
\sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})] \leq \sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}-1}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}-1}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})] \leq \kappa,
$$

which completes the proof of the first property. Then by (G.13), we have

$$
\begin{aligned}
&y_{f,i_f} f(\mathbf{W}^{u,\tilde{t}}, \mathbf{x}_{f,i_f}) - y_{f,i'_f} f(\mathbf{W}^{u,\tilde{t}}, \mathbf{x}_{f,i'_f}) \\
&\leq \frac{1}{m} \sum_{s=1}^m [\overline{\rho}^{u,\tilde{t}}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f}) - \overline{\rho}^{u,\tilde{t}}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})] + 1.75 \\
&\leq \kappa + 1.75 \leq C_6.
\end{aligned}
$$

Then by Lemma F.4, we have

$$\frac{\ell'^{u,\tilde{t}}_{f,i_f}}{\ell'^{u,\tilde{t}}_{f,i'_f}} \leq \exp\big(y_{f,i_f} f(\mathbf{W}^{u,\tilde{t}}, \mathbf{x}_{f,i_f}) - y_{f,i'_f} f(\mathbf{W}^{u,\tilde{t}}, \mathbf{x}_{f,i'_f})\big) \leq \exp(C_6) = C_3,$$

which completes the induction. □

Now we prove Proposition G.11.

Proof of Proposition G.11. By Proposition F.2 and $\mathbf{w}_{j,s}^{u,0} = \mathbf{w}_{j,s}^{b,t_{end}}$, (G.3) and (G.4) hold trivially at $t = 0$. By Lemma G.10, we know that the end of the first stage, denoted by t_1, is the first time such that $\sum_{s\in S_{f,i_f}^{b,0}} \overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}| \leq C^{-1}$. Moreover, since $\overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$ is monotonically decreasing, it follows that $0 \leq \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) \leq \alpha$ for $0 \leq t \leq t_1$. Also, Kou et al. (2023b) has established that $\underline{\rho}_{j,s,i_f}^{u,0}(\boldsymbol{\xi}_{f,i_f}) = \underline{\rho}_{j,s,i_f}^{b,t_{end}}(\boldsymbol{\xi}_{f,i_f}) \leq -0.5\beta - 10\sqrt{\log(6n^2/\delta)/d}n\alpha$. Consequently, by (G.6), we have $\langle \mathbf{w}_{j,s}^{u,0}, \boldsymbol{\xi}_{f,i_f}\rangle \leq \underline{\rho}_{j,s,i_f}^{u,0} + \langle \mathbf{w}_{j,s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle + 10\sqrt{\log(6n^2/\delta)/d}n\alpha \leq 0$, which further implies that $0 \geq \underline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) \geq -\beta - 20\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \geq -\alpha$.

Now suppose there exists $\tilde{t} \leq t_1$ such that $0 \leq \gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f) \leq C' n_f \mathrm{SNR}_{\mathrm{f}}^2 \alpha$ holds for all $0 \leq t \leq \tilde{t} - 1$. We aim to show that it also holds at $t = \tilde{t}$. By Lemma E.2, we have

$$\gamma_{j,s}^{u,\tilde{t}}(\boldsymbol{\mu}_f) = \gamma_{j,s}^{u,\tilde{t}-1}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \cdot \sum_{i_f=1}^{n_f} \ell'^{u,\tilde{t}-1}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,\tilde{t}-1}, y_{f,i_f} \cdot \boldsymbol{\mu}_f\rangle) \cdot \|\boldsymbol{\mu}_f\|_2^2,$$

$$\overline{\rho}_{j,s,i_f}^{u,\tilde{t}}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}_{j,s,i_f}^{u,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{n_f m} \cdot \ell'^{u,\tilde{t}-1}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,\tilde{t}-1}, \boldsymbol{\xi}_{f,i_f}\rangle) \cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \cdot \mathbb{1}(y_{f,i_f} = j).$$

For $\gamma_{j,s}^{u,\tilde{t}}(\boldsymbol{\mu}_f)$,

$$\begin{aligned}
\gamma_{j,s}^{u,\tilde{t}}(\boldsymbol{\mu}_f) &= \gamma_{j,s}^{u,\tilde{t}-1}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \sum_{i_f \in S_{f,1}} \ell'^{u,\tilde{t}-1}_{f,i_f} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\leq \gamma_{j,s}^{u,\tilde{t}-1}(\boldsymbol{\mu}_f) - \frac{\eta\|\boldsymbol{\mu}_f\|_2^2}{4m} \min_{i_f \in S_{f,1}} |\ell'^{u,\tilde{t}-1}_{f,i_f}|, \text{ if } \langle \mathbf{w}_{j,s}^{u,\tilde{t}-1}, \boldsymbol{\mu}_f\rangle \geq 0, \\
\gamma_{j,s}^{u,\tilde{t}}(\boldsymbol{\mu}_f) &= \gamma_{j,s}^{u,\tilde{t}-1}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \sum_{i_f \in S_{f,-1}} \ell'^{u,\tilde{t}-1}_{f,i_f} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\leq \gamma_{j,s}^{u,\tilde{t}-1}(\boldsymbol{\mu}_f) - \frac{\eta\|\boldsymbol{\mu}_f\|_2^2}{4m} \min_{i_f \in S_{f,-1}} |\ell'^{u,\tilde{t}-1}_{f,i_f}|, \text{ if } \langle \mathbf{w}_{j,s}^{u,\tilde{t}-1}, \boldsymbol{\mu}_f\rangle < 0,
\end{aligned} \tag{G.15}$$

where the inequality is by Lemma D.1. For $\overline{\rho}_{j,s,i_f}^{u,\tilde{t}}(\boldsymbol{\xi}_{f,i_f})$ with $s \in S_{f,i_f}^{b,0}$,

$$\overline{\rho}_{j,s,i_f}^{u,\tilde{t}}(\boldsymbol{\xi}_{f,i_f}) \geq \overline{\rho}_{j,s,i_f}^{u,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) - \frac{3\eta\sigma_p^2 d}{2mn_f} \cdot |\ell'^{u,\tilde{t}-1}_{f,i_f}|, \tag{G.16}$$

where the inequality is by Lemma D.2. Then by (G.15), (G.16) and Lemma G.15, we have

$$\frac{\gamma_{j,s}^{u,\tilde{t}}(\boldsymbol{\mu}_f)}{\overline{\rho}_{j,s,i_f}^{u,\tilde{t}}(\boldsymbol{\xi}_{f,i_f})} \leq \max\left\{\frac{\gamma_{j,s}^{u,\tilde{t}-1}(\boldsymbol{\mu}_f)}{\overline{\rho}_{j,s,i_f}^{u,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f})}, \frac{C_4 n_f\|\boldsymbol{\mu}_f\|_2^2}{6\sigma_p^2 d}\right\} \leq C' n_f \mathrm{SNR}_f^2,$$

which completes the induction. □

Now we have already proved Proposition G.11. Then by Lemma G.15, the following proposition directly follows.

Proposition G.16. *Under Condition 3.1, for $0 \leq t \leq t_1$, the following properties hold :*

1. $\sum_{s=1}^{m}\big[\overline{\rho}^{u,t}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f})-\overline{\rho}^{u,t}_{y_{f,i'_f},s,i'_f}(\boldsymbol{\xi}_{f,i'_f})\big]\le\kappa$ *for all* $i_f,i'_f\in[n_f]$.

2. $y_{f,i_f}\cdot f(\mathbf{W}^{u,t},\mathbf{x}_{f,i_f})-y_{f,i'_f}\cdot f(\mathbf{W}^{u,t},\mathbf{x}_{f,i'_f})\le C_6$ *for all* $i_f,i'_f\in[n_f]$,

3. $\ell'^{u,t}_{f,i_f}/\ell'^{u,t}_{f,i'_f}\le C_3=\exp(C_6)$ *for all* $i_f,i'_f\in[n]$.

Here, κ *and* C_6 *are positive constants.*

Lemma G.17. *Under the same conditions as Theorem 3.2, we have*

$$\overline{\rho}^{u,t}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f})/\gamma^{u,t}_{j',s'}(\boldsymbol{\mu}_f)=\Theta(n_f^{-1}\mathrm{SNR}_{\mathrm{f}}^{-2}) \tag{G.17}$$

for $j=y_{f,i_f}, j'\in\{\pm1\}$, $s\in S^{b,0}_{f,i_f}$, $s'\in[m]$ *and* $0\le t\le t_1$, *where* $\mathrm{SNR}_{\mathrm{f}}=\|\boldsymbol{\mu}_f\|_2/(\sigma_p\sqrt{d})$.

Proof of Lemma G.17. We prove the claim by induction. As shown in Lemma F.7, we have $\overline{\rho}^{u,0}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f})/\gamma^{u,0}_{j',s'}(\boldsymbol{\mu}_f)=\Theta(n_f^{-1}\mathrm{SNR}_{\mathrm{f}}^{-2})$ for $s\in S^{b,0}_{f,i_f}, s'\in[m]$, and thus (G.17) holds at $t=0$. Now suppose that there exists $\tilde{t}\in[0,t_1]$ such that (G.17) holds for all $t\in[0,\tilde{t}-1]$. By Lemma E.2, we have that

$$\begin{aligned}
\gamma^{u,\tilde{t}}_{j',s'}(\boldsymbol{\mu}_f)&=\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)+\frac{\eta}{n_fm}\cdot\sum_{i_f=1}^{n_f}\ell'^{u,\tilde{t}-1}_{f,i_f}\cdot\sigma'(\langle\mathbf{w}^{u,\tilde{t}-1}_{j',s'},y_{f,i_f}\cdot\boldsymbol{\mu}_f\rangle)\cdot\|\boldsymbol{\mu}_f\|_2^2,\\
\overline{\rho}^{u,\tilde{t}}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f})&=\overline{\rho}^{u,\tilde{t}-1}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f})+\frac{\eta}{n_fm}\cdot\ell'^{u,\tilde{t}-1}_{f,i_f}\cdot\sigma'(\langle\mathbf{w}^{u,\tilde{t}-1}_{j,s},\boldsymbol{\xi}_{f,i_f}\rangle)\cdot\|\boldsymbol{\xi}_{f,i_f}\|_2^2\cdot\mathbb{1}(y_{f,i_f}=j).
\end{aligned}$$

Then we have that

$$\begin{aligned}
\gamma^{u,\tilde{t}}_{j',s'}(\boldsymbol{\mu}_f)&=\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)+\frac{\eta}{n_fm}\sum_{i_f\in S_{f,1}}\ell'^{u,\tilde{t}-1}_{f,i_f}\cdot\|\boldsymbol{\mu}_f\|_2^2\\
&\le\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)-\frac{\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}\min_{i_f\in S_{f,1}}|\ell'^{u,\tilde{t}-1}_{f,i_f}|,\ \text{if }\langle\mathbf{w}^{u,\tilde{t}-1}_{j',s'},\boldsymbol{\mu}_f\rangle\ge0,\\
\gamma^{u,\tilde{t}}_{j',s'}(\boldsymbol{\mu}_f)&=\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)+\frac{\eta}{n_fm}\sum_{i_f\in S_{f,-1}}\ell'^{u,\tilde{t}-1}_{f,i_f}\cdot\|\boldsymbol{\mu}_f\|_2^2\\
&\le\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)-\frac{\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}\min_{i_f\in S_{f,-1}}|\ell'^{u,\tilde{t}-1}_{f,i_f}|,\ \text{if }\langle\mathbf{w}^{u,\tilde{t}-1}_{j',s'},\boldsymbol{\mu}_f\rangle<0,
\end{aligned} \tag{G.18}$$

where the inequality is by Lemma D.1. And we have that

$$\overline{\rho}^{u,\tilde{t}}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f})\ge\overline{\rho}^{u,\tilde{t}-1}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f})-\frac{3\eta\sigma_p^2d}{2mn_f}\cdot|\ell'^{u,\tilde{t}-1}_{f,i_f}|, \tag{G.19}$$

where the inequality is by Lemma D.2. Then by (G.18), (G.19) and Lemma G.15, we have

$$\frac{\overline{\rho}^{u,\tilde{t}}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f})}{\gamma^{u,\tilde{t}}_{j',s'}(\boldsymbol{\mu}_f)}\ge\min\left\{\frac{\overline{\rho}^{u,\tilde{t}-1}_{j,s,i_f}(\boldsymbol{\xi}_{f,i_f})}{\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)},\frac{6\sigma_p^2d}{C_4n_f\|\boldsymbol{\mu}_f\|_2^2}\right\}=\Theta(n_f^{-1}\mathrm{SNR}_f^{-2}). \tag{G.20}$$

On the other hand,

$$\begin{aligned}
\gamma^{u,\tilde{t}}_{j',s'}(\boldsymbol{\mu}_f)&=\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)+\frac{\eta}{n_fm}\sum_{i_f\in S_{f,1}}\ell'^{u,\tilde{t}-1}_{f,i_f}\cdot\|\boldsymbol{\mu}_f\|_2^2\\
&\ge\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)-\frac{3\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}\max_{i_f\in S_{f,1}}|\ell'^{u,\tilde{t}-1}_{f,i_f}|,\ \text{if }\langle\mathbf{w}^{u,\tilde{t}-1}_{j',s'},\boldsymbol{\mu}_f\rangle\ge0,\\
\gamma^{u,\tilde{t}}_{j',s'}(\boldsymbol{\mu}_f)&=\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)+\frac{\eta}{n_fm}\sum_{i_f\in S_{f,-1}}\ell'^{u,\tilde{t}-1}_{f,i_f}\cdot\|\boldsymbol{\mu}_f\|_2^2\\
&\ge\gamma^{u,\tilde{t}-1}_{j',s'}(\boldsymbol{\mu}_f)-\frac{3\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}\max_{i_f\in S_{f,-1}}|\ell'^{u,\tilde{t}-1}_{f,i_f}|,\ \text{if }\langle\mathbf{w}^{u,\tilde{t}-1}_{j',s'},\boldsymbol{\mu}_f\rangle<0,
\end{aligned} \tag{G.21}$$

where the inequality is by Lemma D.1. We also have that

$$\begin{aligned}\overline{\rho}_{j,s,i_f}^{u,\tilde{t}}(\boldsymbol{\xi}_{f,i_f}) &= \overline{\rho}_{j,s,i_f}^{u,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{n_f m}\cdot \ell_{f,i_f}'^{u,\tilde{t}-1}\cdot \sigma'(\langle \mathbf{w}_{j,s}^{u,\tilde{t}-1}, \boldsymbol{\xi}_{f,i_f}\rangle)\cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \\ &= \overline{\rho}_{j,s,i_f}^{u,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{n_f m}\cdot \ell_{f,i_f}'^{u,\tilde{t}-1}\cdot \|\boldsymbol{\xi}_{f,i_f}\|_2^2 \\ &\le \overline{\rho}_{j,s,i_f}^{u,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f}) - \frac{\eta\sigma_p^2 d}{2mn_f}\cdot |\ell_{f,i_f}'^{u,\tilde{t}-1}|,\end{aligned} \tag{G.22}$$

where the inequality is by Lemma D.2. Then by (G.21), (G.22) and Lemma F.3, we have

$$\frac{\overline{\rho}_{j,s,i_f}^{u,\tilde{t}}(\boldsymbol{\xi}_{f,i_f})}{\gamma_{j',s'}^{u,\tilde{t}}(\boldsymbol{\mu}_f)} \le \max\left\{\frac{\overline{\rho}_{j,s,i_f}^{u,\tilde{t}-1}(\boldsymbol{\xi}_{f,i_f})}{\gamma_{j',s'}^{u,\tilde{t}-1}(\boldsymbol{\mu}_f)}, \frac{2C_4\sigma_p^2 d}{3n_f\|\boldsymbol{\mu}_f\|_2^2}\right\} = \Theta(n_f^{-1}\mathrm{SNR}_f^{-2}). \tag{G.23}$$

By combining (G.20) and (G.23), we have that

$$\frac{\overline{\rho}_{j,s,i_f}^{u,\tilde{t}}(\boldsymbol{\xi}_{f,i_f})}{\gamma_{j',s'}^{u,\tilde{t}}(\boldsymbol{\mu}_f)} = \Theta(n_f^{-1}\mathrm{SNR}_f^{-2}),$$

which completes the induction. □

By an argument similar to that in the proof of Lemma G.10, together with Lemma G.17, we obtain the following lemma:

Lemma G.18. *Under the same conditions as Theorem 3.2, we have that*

$$y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f}) = \Theta\Big(\sum_{s\in S_{f,i_f}^{b,0}} \overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}|\Big),$$

for $0 \le t \le t_1$, *where* $S_{f,i_f}^{b,0} = \{s \in [m] : \langle \mathbf{w}_{y_{f,i_f},s}^{b,0}, \boldsymbol{\xi}_{f,i_f}\rangle > 0\}$.

It remains to characterize how long it takes for $\sum_{s\in S_{f,i_f}^{b,0}} \overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})/|S_{f,i_f}^{b,0}| \le C^{-1}$ to hold.

Lemma G.19. *Define* $\bar{\overline{\rho}}_t \triangleq \frac{1}{|S_{f,i_f}^{b,0}|}\sum_{s\in S_{f,i_f}^{b,0}} \overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f})$. *Under the same conditions as Theorem 3.2, there exists*

$$t_1 = \frac{n_f m}{\eta\sigma_p^2 d}\left[\exp\left(\Theta(\bar{\overline{\rho}}_0)\right) - \exp\left(C^{-1}\right) + (\bar{\overline{\rho}}_0 - C^{-1})\right] = \Theta\left(\frac{n_f m}{\eta\sigma_p^2 d\epsilon}\right)$$

such that the KL divergence $\mathrm{D}_{\mathrm{KL}}(f^{u,t_1}||f^r) \le 2^{-1}C^{-2}$.

Proof of Lemma G.19. For simplicity, define

$$\bar{\overline{\rho}}_t \triangleq \frac{1}{|S_{f,i_f}^{b,0}|}\sum_{s\in S_{f,i_f}^{b,0}} \overline{\rho}_{y_{f,i_f},s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}).$$

By Lemma E.2, for $j = y_{f,i_f}$ and any $s \in S_{f,i_f}^{b,0}$,

$$\overline{\rho}_{j,s,i_f}^{u,t+1}(\boldsymbol{\xi}_{f,i_f}) = \overline{\rho}_{j,s,i_f}^{u,t}(\boldsymbol{\xi}_{f,i_f}) + \frac{\eta}{n_f m}\ell_{f,i_f}'^{u,t}\|\boldsymbol{\xi}_{f,i_f}\|_2^2,$$

where

$$\ell_{f,i_f}'^{u,t} = \ell'\big(y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})\big) = -\frac{1}{\exp\big(y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})\big) + 1}.$$

Averaging over $s \in S_{f,i_f}^{b,0}$ gives

$$\bar{\overline{\rho}}_{t+1} = \bar{\overline{\rho}}_t + \frac{\eta}{n_f m}\ell_{f,i_f}'^{u,t}\|\boldsymbol{\xi}_{f,i_f}\|_2^2. \tag{G.24}$$

By Lemma D.2, we have

$$\|\boldsymbol{\xi}_{f,i_f}\|_2^2 = \Theta(\sigma_p^2 d).$$

Moreover, by Lemma G.18, throughout the first stage,

$$y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f}) = \Theta(\bar{\bar{\rho}}_t).$$

Hence, there exists an absolute constant $\kappa > 0$ such that

$$y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f}) = \kappa \bar{\bar{\rho}}_t \tag{G.25}$$

up to absolute-constant rescaling, and therefore (G.24) can be written as

$$\bar{\bar{\rho}}_{t+1} = \bar{\bar{\rho}}_t - \frac{\eta \sigma_p^2 d}{n_f m} \cdot \frac{1}{e^{\kappa \bar{\bar{\rho}}_t} + 1}, \tag{G.26}$$

where we have absorbed all absolute constants into κ.

To characterize when $\bar{\bar{\rho}}_t$ decreases to the level C^{-1}, consider the continuous-time dynamics associated with (G.26):

$$\frac{dz(\tau)}{d\tau} = -\frac{\eta \sigma_p^2 d}{n_f m} \cdot \frac{1}{e^{\kappa z(\tau)} + 1}, \quad z(0) = \bar{\bar{\rho}}_0. \tag{G.27}$$

Separating variables yields

$$\left(e^{\kappa z(\tau)} + 1\right) dz = -\frac{\eta \sigma_p^2 d}{n_f m} \, d\tau.$$

Integrating both sides from 0 to τ, we obtain

$$\int_{z(\tau)}^{\bar{\bar{\rho}}_0} \left(e^{\kappa u} + 1\right) du = \frac{\eta \sigma_p^2 d}{n_f m} \tau.$$

Evaluating the integral gives

$$\frac{1}{\kappa} e^{\kappa z(\tau)} + z(\tau) = \frac{1}{\kappa} e^{\kappa \bar{\bar{\rho}}_0} + \bar{\bar{\rho}}_0 - \frac{\eta \sigma_p^2 d}{n_f m} \tau. \tag{G.28}$$

Now let τ_1 be the first time such that $z(\tau_1) = C^{-1}$. Substituting $z(\tau_1) = C^{-1}$ into (G.28) yields

$$\tau_1 = \frac{n_f m}{\eta \sigma_p^2 d} \left[\frac{1}{\kappa} \left(e^{\kappa \bar{\bar{\rho}}_0} - e^{\kappa C^{-1}} \right) + (\bar{\bar{\rho}}_0 - C^{-1}) \right]. \tag{G.29}$$

Therefore, the original discrete recursion (G.26) reaches the level C^{-1} within

$$\begin{aligned} t_1 &= \frac{n_f m}{\eta \sigma_p^2 d} \left[\frac{1}{\kappa} \left(e^{\kappa \bar{\bar{\rho}}_0} - e^{\kappa C^{-1}} \right) + (\bar{\bar{\rho}}_0 - C^{-1}) \right] \\ &= \frac{n_f m}{\eta \sigma_p^2 d} \left[\exp\left(\Theta(\bar{\bar{\rho}}_0)\right) - \exp\left(C^{-1}\right) + (\bar{\bar{\rho}}_0 - C^{-1}) \right] \end{aligned}$$

iterations. By adopting a proof technique similar to that used in Lemma D.17 of Wan et al. (2025), we obtain

$$\bar{\bar{\rho}}_0 = \Theta\left(\log(\epsilon^{-1})\right)$$

Then we have that

$$t_1 = \Theta\left(\frac{n_f m}{\eta \sigma_p^2 d \epsilon}\right)$$

Hence, at iteration t_1, we have

$$\bar{\bar{\rho}}_{t_1} \le C^{-1}.$$

By (G.25), this implies

$$\left| y_{f,i_f} f(\mathbf{W}^{u,t_1}, \mathbf{x}_{f,i_f}) \right| \le C^{-1}.$$

By Lemma G.4, we further have

$$\left| y_{f,i_f} f(\mathbf{W}^{u,t_1}, \mathbf{x}_{f,i_f}) - y_{f,i_f} f^r(\mathbf{x}_{f,i_f}) \right| \le 2C^{-1}.$$

Finally, applying Lemma G.1, we conclude that

$$\mathrm{D}_{\mathrm{KL}}(f^{u,t_1} \| f^r) \le \frac{1}{2} C^{-2}.$$

This completes the proof. □

G.2 SECOND STAGE

In this section, we study how long the KL divergence $\mathrm{D}_{\mathrm{KL}}(f^{u,t_1}||f^r) \leq 2^{-1}C^{-2}$ can be maintained. Equivalently, we analyze how long it takes for $y_{f,i_f} f^u(\mathbf{x}_{f,i_f})$ to decay from C^{-1} to $-C^{-1}$. We begin by estimating $\ell'^{u,t}_{f,i_f}$ over this time interval.

Lemma G.20. *Under the same conditions as Theorem 3.2, we have*

$$C_4 \leq -\ell'^{u,t}_{f,i_f} \leq C_5,$$

for the second stage, where $C_4 < C_5 < 1$ are positive constants and $i_f \in [n_f]$.

Proof of Lemma G.20. For the second stage, we know

$$-C^{-1} \leq y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f}) \leq C^{-1}.$$

Then by

$$\ell'^{u,t}_{f,i_f} = \ell'\big(y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})\big) = -\frac{1}{\exp\big(y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})\big) + 1},$$

we have

$$\frac{1}{\exp(C^{-1}) + 1} \leq -\ell'^{u,t}_{f,i_f} \leq \frac{1}{\exp(-C^{-1}) + 1}.$$

Let $C_4 = \frac{1}{\exp(C^{-1})+1}$ and $C_5 = \frac{1}{\exp(-C^{-1})+1}$, then we completes the proof. □

By Lemma E.2, once $\langle \mathbf{w}^{u,t}_{y_{f,i_f},s}, \boldsymbol{\xi}_{f,i_f} \rangle \leq 0$, $\bar{\rho}^{u,t}_{y_{f,i_f},s,i_f}(\boldsymbol{\xi}_{f,i_f})$ remains unchanged, whereas $\gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f)$ is unaffected and can continue to evolve. We next characterize how long it takes for $\gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f)$ to decay from $n_f \mathrm{SNR}_{\mathrm{f}}^2 C^{-1}$ at t_1 to $-C^{-1}$.

Lemma G.21. *Define $\bar{\gamma}_{j,t} \triangleq \frac{1}{m}\sum_{s=1}^m \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f)$. Under the same conditions as Theorem 3.2, there exists*

$$t_2 = t_1 + \Theta\left(\frac{m}{C\eta\|\boldsymbol{\mu}_f\|_2^2}\right)$$

such that $-C^{-1} \leq \bar{\gamma}_{j,t} \leq n_f \mathrm{SNR}_{\mathrm{f}}^2 C^{-1}$ for $t_1 \leq t \leq t_2$.

Proof of Lemma G.21. By Lemma E.2, we have that

$$\gamma^{u,t+1}_{j,s}(\boldsymbol{\mu}_f) = \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \cdot \sum_{i_f=1}^{n_f} \ell'^{u,t}_{f,i_f} \cdot \sigma'(\langle \mathbf{w}^{u,t}_{j,s}, y_{f,i_f} \cdot \boldsymbol{\mu}_f \rangle) \cdot \|\boldsymbol{\mu}_f\|_2^2.$$

Then we have that

$$\begin{aligned}
\gamma^{u,t+1}_{j,s}(\boldsymbol{\mu}_f) &= \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \sum_{i_f \in S_{f,1}} \ell'^{u,t}_{f,i_f} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\leq \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) - \frac{C_5 \eta \|\boldsymbol{\mu}_f\|_2^2}{4m}, \text{ if } \langle \mathbf{w}^{u,t}_{j,s}, \boldsymbol{\mu}_f \rangle \geq 0, \\
\gamma^{u,t+1}_{j,s}(\boldsymbol{\mu}_f) &= \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \sum_{i_f \in S_{f,-1}} \ell'^{u,t}_{f,i_f} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\leq \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) - \frac{C_5 \eta \|\boldsymbol{\mu}_f\|_2^2}{4m}, \text{ if } \langle \mathbf{w}^{u,t}_{j,s}, \boldsymbol{\mu}_f \rangle \leq 0,
\end{aligned} \tag{G.30}$$

where the inequality is by Lemma D.1 and Lemma G.20. On the other hand,

$$\begin{aligned}
\gamma^{u,t+1}_{j,s}(\boldsymbol{\mu}_f) &= \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \sum_{i_f \in S_{f,1}} \ell'^{u,t}_{f,i_f} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\geq \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) - \frac{3C_6 \eta \|\boldsymbol{\mu}_f\|_2^2}{4m}, \text{ if } \langle \mathbf{w}^{u,t}_{j,s}, \boldsymbol{\mu}_f \rangle \geq 0, \\
\gamma^{u,t+1}_{j,s}(\boldsymbol{\mu}_f) &= \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) + \frac{\eta}{n_f m} \sum_{i_f \in S_{f,-1}} \ell'^{u,t}_{f,i_f} \cdot \|\boldsymbol{\mu}_f\|_2^2 \\
&\geq \gamma^{u,t}_{j,s}(\boldsymbol{\mu}_f) - \frac{3C_6 \eta \|\boldsymbol{\mu}_f\|_2^2}{4m}, \text{ if } \langle \mathbf{w}^{u,t}_{j,s}, \boldsymbol{\mu}_f \rangle \leq 0,
\end{aligned} \tag{G.31}$$

where the inequality is by Lemma D.1 and Lemma G.20. Then combining (G.30) and (G.31), we have

$$\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f) - \frac{3C_5\eta\|\boldsymbol{\mu}_f\|_2^2}{4m} \le \gamma_{j,s}^{u,t+1}(\boldsymbol{\mu}_f) \le \gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f) - \frac{C_4\eta\|\boldsymbol{\mu}_f\|_2^2}{4m} \tag{G.32}$$

Averaging (G.32) over $s \in [m]$, and recalling that

$$\bar{\gamma}_{j,t} \triangleq \frac{1}{m}\sum_{s=1}^{m}\gamma_{j,s}^{u,t}(\boldsymbol{\mu}_f),$$

we obtain

$$\bar{\gamma}_{j,t} - \frac{3C_5\eta\|\boldsymbol{\mu}_f\|_2^2}{4m} \le \bar{\gamma}_{j,t+1} \le \bar{\gamma}_{j,t} - \frac{C_4\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}. \tag{G.33}$$

Hence $\bar{\gamma}_{j,t}$ decreases at a constant rate throughout the second stage. Iterating (G.33), for any integer $k \ge 0$, yields

$$\bar{\gamma}_{j,t_1} - \frac{3C_5\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}k \le \bar{\gamma}_{j,t_1+k} \le \bar{\gamma}_{j,t_1} - \frac{C_4\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}k. \tag{G.34}$$

By the definition of t_1 from the first stage and Lemma G.17, we have

$$\bar{\gamma}_{j,t_1} = \Theta\left(n_f\mathrm{SNR}_f^2 C^{-1}\right).$$

Moreover, since $\bar{\gamma}_{j,t}$ is decreasing, it follows that for all $t \ge t_1$,

$$\bar{\gamma}_{j,t} \le n_f\mathrm{SNR}_f^2 C^{-1}.$$

Now define

$$t_2 \triangleq \max\left\{t \ge t_1 : \bar{\gamma}_{j,t} \ge -C^{-1}\right\}.$$

Then, by definition, for every $t_1 \le t \le t_2$, we have

$$-C^{-1} \le \bar{\gamma}_{j,t} \le n_f\mathrm{SNR}_f^2 C^{-1}.$$

It remains to estimate $t_2 - t_1$. Let $k \triangleq t_2 - t_1$. Since $\bar{\gamma}_{j,t_2} \ge -C^{-1}$, applying the upper bound in (G.34) at time $t_2 = t_1 + k$ gives

$$-C^{-1} \le \bar{\gamma}_{j,t_2} \le \bar{\gamma}_{j,t_1} - \frac{C_4\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}k,$$

which implies

$$k \le \frac{4m\left(\bar{\gamma}_{j,t_1} + C^{-1}\right)}{C_4\eta\|\boldsymbol{\mu}_f\|_2^2}. \tag{G.35}$$

On the other hand, by the maximality of t_2, we have

$$\bar{\gamma}_{j,t_2+1} < -C^{-1}.$$

Applying the lower bound in (G.34) at time $t_2 + 1 = t_1 + k + 1$, we obtain

$$\bar{\gamma}_{j,t_1} - \frac{3C_5\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}(k+1) \le \bar{\gamma}_{j,t_2+1} < -C^{-1},$$

and therefore

$$k + 1 > \frac{4m\left(\bar{\gamma}_{j,t_1} + C^{-1}\right)}{3C_5\eta\|\boldsymbol{\mu}_f\|_2^2}. \tag{G.36}$$

Combining (G.35) and (G.36), we conclude that

$$\frac{4m\left(\bar{\gamma}_{j,t_1} + C^{-1}\right)}{3C_5\eta\|\boldsymbol{\mu}_f\|_2^2} - 1 < t_2 - t_1 \le \frac{4m\left(\bar{\gamma}_{j,t_1} + C^{-1}\right)}{C_4\eta\|\boldsymbol{\mu}_f\|_2^2}.$$

Since $\bar{\gamma}_{j,t_1} = n_f\mathrm{SNR}_f^2 C^{-1}$ at the beginning of the second stage, we obtain

$$t_2 - t_1 = \Theta\left(\frac{m\left(n_f\mathrm{SNR}_f^2 + 1\right)}{C\eta\|\boldsymbol{\mu}_f\|_2^2}\right).$$

Then by $n_f \mathrm{SNR}_f^2 \leq C^{-1}$ from Condition 3.1, we have

$$t_2 - t_1 = \Theta\bigg(\frac{m}{C\eta\|\boldsymbol{\mu}_f\|_2^2}\bigg).$$

Therefore,

$$t_2 = t_1 + \Theta\bigg(\frac{m}{C\eta\|\boldsymbol{\mu}_f\|_2^2}\bigg),$$

which completes the proof. □

We next characterize $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ over the interval $t_1 \leq t \leq t_2$.

Lemma G.22. *Under the same conditions as Theorem 3.2, we have*

$$|y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})| \leq C^{-1},$$

for $t_1 \leq t \leq t_2$, where C is a large constant and t_2 is defined as in Lemma G.21.

Proof of Lemma G.22. When $\langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle > 0$, Lemma G.18 implies that $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ is determined by $\bar{\bar{\rho}}_t$. Then by the proof of Lemma G.19, we have

$$\bar{\bar{\rho}}_t \leq \bar{\bar{\rho}}_{t_1} \leq C^{-1}.$$

Subsequently, by adopting an argument similar to that used in the proof of Lemma G.18, we obtain

$$|y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})| \leq C^{-1}.$$

When $\langle \mathbf{w}_{y_{f,i_f},s}^{u,t}, \boldsymbol{\xi}_{f,i_f} \rangle \leq 0$, by the proof of Lemma G.18, $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ is governed by $\bar{\gamma}_{j,t}$. Then by $-C^{-1} \leq \bar{\gamma}_{j,t} \leq n_f \mathrm{SNR}_{\mathrm{f}}^2 C^{-1}$ from Lemma G.21, we have

$$|y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})| \leq C^{-1},$$

which completes the proof. □

Next we characterize $\mathrm{D}_{\mathrm{KL}}(f^{u,t} || f^r)$ over the interval $t_1 \leq t \leq t_2$.

Lemma G.23. *Under the same conditions as Theorem 3.2, we have*

$$\mathrm{D}_{\mathrm{KL}}(f^{u,t} || f^r) \leq 2^{-1} C^{-2},$$

for $t_1 \leq t \leq t_2$, where t_2 is defined as in Lemma G.21.

Proof of Lemma G.23. By Lemma G.22,

$$|y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})| \leq C^{-1}.$$

By Lemma G.4, we further have

$$\big|y_{f,i_f} f(\mathbf{W}^{u,t_1}, \mathbf{x}_{f,i_f}) - y_{f,i_f} f^r(\mathbf{x}_{f,i_f})\big| \leq 2C^{-1}.$$

Finally, applying Lemma G.1, we conclude that

$$\mathrm{D}_{\mathrm{KL}}(f^{u,t} \| f^r) \leq \frac{1}{2} C^{-2},$$

which completes the proof. □

Moreover, in the second stage, we have the following property.

Lemma G.24. *Under the same conditions as Theorem 3.2, we have*

$$\big|L_{S_f}(\mathbf{W}^{u,t}) - L_{S_f}(\mathbf{W}^r)\big| \leq 2C^{-1},$$

for $t_1 \leq t \leq t_2$, where C is a large constant and t_2 is defined as in Lemma G.21.

Proof of Lemma G.24. For simplicity, define

$$a_{i_f} := y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f}), \quad b_{i_f} := y_{f,i_f} f(\mathbf{W}^{r}, \mathbf{x}_{f,i_f}).$$

Then by (2.2),

$$L_{S_f}(\mathbf{W}^{u,t}) - L_{S_f}(\mathbf{W}^{r}) = \frac{1}{n_f} \sum_{i_f=1}^{n_f} \big(\ell(a_{i_f}) - \ell(b_{i_f})\big).$$

where $\ell(z) = \log(1 + \exp(-z))$. We first note that

$$\ell'(z) = -\frac{1}{1 + \exp(z)},$$

and hence

$$|\ell'(z)| \le 1, \quad \forall z \in \mathbb{R}.$$

Therefore, ℓ is 1-Lipschitz on $\mathbb{R}$, i.e.,

$$|\ell(a) - \ell(b)| \le |a - b|, \quad \forall a, b \in \mathbb{R}.$$

Applying this inequality to each pair (a_{i_f}, b_{i_f}) yields

$$|\ell(a_{i_f}) - \ell(b_{i_f})| \le |a_{i_f} - b_{i_f}| \le 2C^{-1},$$

where the inequality is by the proof of Lemma G.23. Thus,

$$\big|L_{S_f}(\mathbf{W}^{u,t}) - L_{S_f}(\mathbf{W}^{r})\big| \le \frac{1}{n_f} \sum_{i_f=1}^{n_f} |\ell(a_{i_f}) - \ell(b_{i_f})| \le 2C^{-1},$$

which completes the proof. □

G.3 Third Stage

In this section, we characterize that for $t > t_2$, $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ diverges to negative infinity at a linear rate, while the KL divergence approaches a positive constant C_7.

Lemma G.25. *Under the same conditions as Theorem 3.2, when $t \ge t_2$, we have that*

1. *$\bar{\gamma}_{j,t}$ and $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ diverge to negative infinity at a linear rate.*
2. $\mathrm{D}_{\mathrm{KL}}(f^{u,t} \| f^{r}) \to \frac{1}{n_f} \sum_{i_f=1}^{n_f} \log\Big(1 + \exp\big(y_{f,i_f} f^{r}(\mathbf{x}_{f,i_f})\big)\Big)$.

Here, $\bar{\gamma}_{j,t}$ is defined as in Lemma G.21.

Proof of Lemma G.25. By the proof of Lemma G.21, we have

$$\bar{\gamma}_{j,t} - \frac{3\eta\|\boldsymbol{\mu}_f\|_2^2}{4m} \le \bar{\gamma}_{j,t+1} \le \bar{\gamma}_{j,t} - \frac{C_4\eta\|\boldsymbol{\mu}_f\|_2^2}{4m}, \tag{G.37}$$

where $0 < C_4 < 1$. Hence, for all $t \ge t_2$, $\bar{\gamma}_{j,t}$ decreases by an absolute-constant amount at each iteration, and therefore diverges to negative infinity at a linear rate.

Moreover, by the proof of Lemma G.22, when $t \ge t_2$, the quantity $y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f})$ is dominated by the term $\bar{\gamma}_{j,t}$. Therefore,

$$y_{f,i_f} f(\mathbf{W}^{u,t}, \mathbf{x}_{f,i_f}) \to -\infty$$

at a linear rate as well. This proves part 1.

We next prove part 2. For each $i_f \in [n_f]$, define

$$p_{i_f}^{u,t} \triangleq q\big(y_{f,i_f} f^{u,t}(\mathbf{x}_{f,i_f})\big), \qquad p_{i_f}^{r} \triangleq q\big(y_{f,i_f} f^{r}(\mathbf{x}_{f,i_f})\big),$$

where

$$q(z) \triangleq \frac{1}{1 + \exp(-z)}.$$

Then

$$\mathrm{D}_{\mathrm{KL}}(f^{u,t}\|f^r) = \frac{1}{n_f}\sum_{i_f=1}^{n_f}\left[p_{i_f}^{u,t}\log\frac{p_{i_f}^{u,t}}{p_{i_f}^r} + (1-p_{i_f}^{u,t})\log\frac{1-p_{i_f}^{u,t}}{1-p_{i_f}^r}\right].$$

By part 1,

$$y_{f,i_f} f^{u,t}(\mathbf{x}_{f,i_f}) \to -\infty.$$

Since $q(z) = 1/(1+\exp(-z))$, we obtain

$$p_{i_f}^{u,t} \to 0.$$

Now fix any $i_f \in [n_f]$. We analyze the two terms in the i_f-th summand separately. First,

$$p_{i_f}^{u,t}\log\frac{p_{i_f}^{u,t}}{p_{i_f}^r} = p_{i_f}^{u,t}\log p_{i_f}^{u,t} - p_{i_f}^{u,t}\log p_{i_f}^r.$$

Since $p_{i_f}^{u,t} \to 0$, we have

$$p_{i_f}^{u,t}\log p_{i_f}^{u,t} \to 0.$$

Also, $p_{i_f}^r$ is fixed with respect to t, so

$$p_{i_f}^{u,t}\log p_{i_f}^r \to 0.$$

Therefore,

$$p_{i_f}^{u,t}\log\frac{p_{i_f}^{u,t}}{p_{i_f}^r} \to 0.$$

For the second term, since $p_{i_f}^{u,t} \to 0$, we have $1 - p_{i_f}^{u,t} \to 1$, and hence

$$(1-p_{i_f}^{u,t})\log\frac{1-p_{i_f}^{u,t}}{1-p_{i_f}^r} \to \log\frac{1}{1-p_{i_f}^r}.$$

Combining the above two displays yields

$$p_{i_f}^{u,t}\log\frac{p_{i_f}^{u,t}}{p_{i_f}^r} + (1-p_{i_f}^{u,t})\log\frac{1-p_{i_f}^{u,t}}{1-p_{i_f}^r} \to \log\frac{1}{1-p_{i_f}^r}.$$

Finally, using the identity

$$1 - q(z) = \frac{1}{1+\exp(z)},$$

we have

$$\log\frac{1}{1-p_{i_f}^r} = \log\frac{1}{1-q\big(y_{f,i_f}f^r(\mathbf{x}_{f,i_f})\big)} = \log\Big(1+\exp\big(y_{f,i_f}f^r(\mathbf{x}_{f,i_f})\big)\Big).$$

Therefore, for each $i_f \in [n_f]$,

$$p_{i_f}^{u,t}\log\frac{p_{i_f}^{u,t}}{p_{i_f}^r} + (1-p_{i_f}^{u,t})\log\frac{1-p_{i_f}^{u,t}}{1-p_{i_f}^r} \to \log\Big(1+\exp\big(y_{f,i_f}f^r(\mathbf{x}_{f,i_f})\big)\Big).$$

Since the sum is finite, we may pass the limit through the summation, which gives

$$\mathrm{D}_{\mathrm{KL}}(f^{u,t}\|f^r) \to \frac{1}{n_f}\sum_{i_f=1}^{n_f}\log\Big(1+\exp\big(y_{f,i_f}f^r(\mathbf{x}_{f,i_f})\big)\Big).$$

By Lemma G.4, we know that $\frac{1}{n_f}\sum_{i_f=1}^{n_f}\log\Big(1+\exp\big(y_{f,i_f}f^r(\mathbf{x}_{f,i_f})\big)\Big)$ is a constant, which we denote by C_7. This completes the proof. □