4 Appendix

4.1 Additional related works

Machine unlearning and memorization. Many unlearning methods are proposed to effectively erase information of selected samples. Several basic but well-known methods such as random labeling of forget set (Graves et al., 2020) and explicit gradient ascent on the forget set (Warnecke et al., 2023) lay foundation for current unlearning methods. More recent works extend on those works to improve overall performance of unlearning. For example, SCRUB (Kurmanji et al., 2023) simultaneously perform gradient ascent on forget set and gradient descent in the retain set to better preserve performance of retain during unlearning. Influence based (Izzo et al., 2021) unlearning propose idea that takes into account of the Hessian information of datasets to perform update of the model weights. Saliency Unlearning (Fan et al., 2024) identity weights that react strongly to forget set through magnitude of gradient and perform unlearning only on those weight to achieve better performance. There are several theoretical studies about unlearning through the lens of the differential privacy and provide performance guarantee. For example, Langevin Unlearning Chien et al. (2025) study unlearning with privacy guarantee through projected noisy gradient descent. Sekhari et al. (2021) studies unlearning problem and provide performance guarantee and the corresponding sample complexity. There are also works discussing relationship between memorization and generalization. Attias et al. (2024) discuss the fundamental trade-off between generalization and memorization under information theory framework. Carlini et al. (2019) discuss different metrics for identifying sample of different type (memorized, prototypical and so on). Feldman (2021) provide theoretical and experimental analysis saying the memorization is necessary to achieve optimal performance. There are also several works studying memorization with different tasks and model architectures (Biderman et al. (2023); Li et al. (2025); Prashanth et al. (2025)).

4.2 Lemmas and proofs

Definition 6 (Full forget Hessian and retain Hessian)**.**

$$H_R = \frac{1}{n_r} \sum_{r \in D_r} H_r, \;\; H_F = \frac{1}{n_r} \sum_{f \in D_f} H_f, \tag{13}$$

Lemma 4.1. *l_1-l_2 norm inequality: For any $x \in \mathbb{R}$, $||x||_2 \leq ||x||_1 \leq \sqrt{d}||x||_2$*

Lemma 4.2. ***Binomial coefficient:*** *For all $n, k \in \mathbb{N}$ such that $k \leq n$, the binomial coefficients satisfy that*

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}. \tag{14}$$

Lemma 4.3. *For any matrix $M \in \mathbb{R}^{n \times n}$, $||M||_F \leq ||M||_{S_1} \leq \sqrt{n}||M||_F$, where $||M||_{S_p}$ is p norm of the spectrum of M, and the inequality is obtain through l_1-l_2 norm inequality.*

Lemma 4.4. *For matrices $M_1 ... M_k \in \mathbb{R}^{n \times n}$, $\mathrm{Tr}[M_1 ... M_k] \leq ||M_1 ... M_k||_{S_1}$ (see Bhatia (2013))*

Lemma 4.5 (Vershynin (2018))**.** *Consider n random gaussian vectors $x_1 ... x_n$ sampled i.i.d from $N(0, \sigma^2 I_d)$, there exist a constant C_1 such that with probability $1 - \delta$,*

$$\sum_{i=1}^{n} \|x_i\| \leq n\sqrt{2}\sigma \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} + \sqrt{\frac{n\sigma^2}{C_1} \log(\frac{2}{\delta})} \;\; \textit{for } n, d \textit{ large enough.} \tag{15}$$

Lemma 4.6 (Vershynin (2018))**.** *Consider n random gaussian vectors $x_1 ... x_n$ sampled i.i.d from $N(0, \sigma^2 I_d)$, there exist a constant C_1 such that with probability $1 - \delta$,*

$$\sum_{i=1}^{n} \|x_i\|^2 \leq n\sigma^2 d + \sqrt{\frac{n\sigma^4 d}{C_2} \log(\frac{2}{\delta})} \;\; \textit{for } n, d \textit{ large enough.} \tag{16}$$

4.3 Proof of lemma 2.1

Lemma 4.7 (Restated). *Consider the unlearning update operator J_k defined in* (3). *Define a sequence of PSD matrices $\{N_k\}_{k\geq 0}$ by $N_0 = I$ and for $k \geq 1$:*

$$N_k \;=\; C_f \sum_{i\in D_f} H_i\, N_{k-1}\, H_i \;+\; C_r \sum_{i\in D_r} H_i\, N_{k-1}\, H_i\,, \tag{17}$$

with C_r, C_f as given in Definition 2. Also let $M_k = J^{2k} + N_k$. Then:

1. ***(Lower bound)*** $\mathbb{E}\,\mathrm{Tr}\big(J_0^T \cdots J_{k-1}^T J_{k-1} \cdots J_0\big) \;\geq\; \mathrm{Tr}(M_k)$. *Moreover, if* $\mathrm{Tr}(N_k) \to \infty$ *as* $k \to \infty$, *then* $\mathbb{E}\|w_k\|^2 \to \infty$ *as well.*

2. ***(Upper bound)*** *If at each step J_k is spectrally bounded as $(1-\epsilon)I \succeq J \succeq -(1-\epsilon)I$ for some $\epsilon \in (0,1)$ (i.e. all eigenvalues of J lie in $[-(1-\epsilon),\, 1-\epsilon]$), then*

$$\mathbb{E}\,\mathrm{Tr}\big(J_0^T \cdots J_{k-1}^T J_{k-1} \cdots J_0\big) \;\leq\; \sum_{r=0}^{k-1} \binom{k}{r} (1-\epsilon)^{2(k-r)}\; \mathrm{Tr}(N_r)\,.$$

   If in addition $\mathrm{Tr}(N_r) \leq \epsilon$ for all r, then $\mathbb{E}\|w_k\|^2 \to 0$ as $k \to \infty$ (the unlearning update converges in mean square).

Proof. As we are taking the expectation value over the calculation, we can effectively transform the J_k into following with random variables involved:

$$J_k = (I - \eta(1-\alpha)\frac{1}{B}\sum_{i\in D_{r,k}} H_i + \eta\alpha\frac{1}{B}\sum_{i\in D_{f,k}} H_i) = (I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f), \tag{18}$$

where x_r, x_f are the corresponding Bernoulli random variables with probability $P(x_r = 1) = \frac{B}{n_r}$ and $P(x_f = 1) = \frac{B}{n_f}$ and 0 otherwise.

To initiate the first step in characterize the difference between the unlearning and usually learning process, we first calculate the $E[J_1^T J_1]$ as follows:

$$\begin{aligned} E[J_1^T J_1] &= E[(I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)^T (I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)] \\ &= E[(I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r)^T (I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r) \\ &+2(I - \eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r} x_r H_r)^T (\eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f) + (\eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)^T (\eta\alpha\frac{1}{B}\sum_{i\in D_f} x_f H_f)]. \end{aligned} \tag{19}$$

Here, we separate the above equation into three part and take the expectation accordingly:

$$
\begin{aligned}
&E[(I-\eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r}x_rH_r)^T(I-\eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r}x_rH_r)]\\
&=E[(I-2\eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r}x_rH_r+\eta^2(1-\alpha)^2\frac{1}{B}^2\sum_{r\in D_r}x_rH_r\sum_{r\in D_r}x_rH_r)],\\
&=I-2\eta(1-\alpha)\frac{1}{n_r}\sum_{r\in D_r}H_r+E[\eta^2(1-\alpha)^2(\frac{1}{B})^2\sum_{r\in D_r}x_rH_r\sum_{r\in D_r}x_rH_r],\\
&=I-2\eta(1-\alpha)\frac{1}{n_r}\sum_{r\in D_r}H_r+E[\eta^2(1-\alpha)^2(\frac{1}{B})^2\sum_{r'\in D_r}\sum_{r\in D_r}x_{r'}x_rH_{r'}H_r],\\
&=I-2\eta(1-\alpha)\frac{1}{n_r}\sum_{r\in D_r}H_r+\eta^2(1-\alpha)^2(\frac{1}{n_r})^2(\sum_{r\in D_r}H_r)^2+\eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}H_r^2,\\
&=I-2\eta(1-\alpha)H_R+\eta^2(1-\alpha)^2H_R^2+\eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}H_r^2,\\
&=(I-\eta(1-\alpha)H_R)^2+\eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}H_r^2.
\end{aligned}
\tag{20}
$$

The random variables x_r are independent to each other but not itself and therefore there exist one additional terms in the final line. Also, Compared to the original sgd there exists additional multiplication of the $(1-\alpha)^2$. Next, we move on to the interaction term:

$$
E[2(I-\eta(1-\alpha)\frac{1}{B}\sum_{r\in D_r}x_rH_r)^T(\eta\alpha\frac{1}{B}\sum_{i\in D_f}x_fH_f)]=2(I-\eta(1-\alpha)H_R)^T(\eta\alpha H_F). \tag{21}
$$

We can directly formulate this as above due to the fact that we assume the sampling process of retain set and forget set to be independent. Last, the term arising due to the forget set:

$$
E[(\eta\alpha\frac{1}{B}\sum_{f\in D_f}x_fH_f)^T(\eta\alpha\frac{1}{B}\sum_{f\in D_f}x_fH_f)]=\eta^2\alpha^2H_F^2+\eta^2\alpha^2\frac{1}{n_f}(\frac{1}{B}-\frac{1}{n_f})\sum_{f\in D_f}^{n_f}H_f^2. \tag{22}
$$

We then integrate the three part and reformulate the Jacobian:

$$
\begin{aligned}
E[J_1^TJ_1]&=(I-\eta(1-\alpha)H_R)(I-\eta(1-\alpha)H_R)+2(I-\eta(1-\alpha)H_R)^T(\eta\alpha H_F)+\eta^2\alpha^2H_F^2\\
&\quad+\eta^2\alpha^2\frac{1}{n_f}(\frac{1}{B}-\frac{1}{n_f})\sum_{f\in D_f}^{n_f}H_f^2+\eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}^{n_r}H_r^2,\\
&=(I-\eta(1-\alpha)H_R+\eta\alpha H_F)(I-\eta(1-\alpha)H_R+\eta\alpha H_F)\\
&\quad+\eta^2\alpha^2\frac{1}{n_f}(\frac{1}{B}-\frac{1}{n_f})\sum_{f\in D_f}^{n_f}H_f^2+\eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}^{n_r}H_r^2,\\
&=J^2+\eta^2\alpha^2\frac{1}{n_f}(\frac{1}{B}-\frac{1}{n_f})\sum_{f\in D_f}^{n_f}H_f^2+\eta^2(1-\alpha)^2\frac{1}{n_r}(\frac{1}{B}-\frac{1}{n_r})\sum_{r\in D_r}^{n_r}H_r^2,
\end{aligned}
\tag{23}
$$

where we define $J=I-\eta(1-\alpha)H_R+\eta\alpha H_F$. During the whole work, we will be analyzing on these terms to characterize the behavior of unlearning process.

We use inductive proof for both first and second part of the theory and we begin the proof as following:

First part:

Base case: k = 1

$$M_1 = J^2 + C_f \sum_{f \in D_f}^{n_f} H_f^2 + C_r \sum_{r \in D_r}^{n_r} H_r^2 = J^2 + N_1 \preceq E[J_1^T J_1], \tag{24}$$

where the left term match the equation 23 and therefore the basis case is set. Now we go further with inductive step.

Inductive step: k-1

$$\begin{aligned}
&E[J_k^T J_{k-1}^T ... J_1^T J_1 ... J_{k-1}] \succeq E[J_k^T M_{k-1} J_k], \\
&= E[(I - \eta(1-\alpha)\frac{1}{B}\sum_{i \in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{f \in D_f} x_f H_f) M_{k-1} (I - \eta(1-\alpha)\frac{1}{B}\sum_{i \in D_r} x_r H_r + \eta\alpha\frac{1}{B}\sum_{i \in D_f} x_f H_f)], \\
&= J M_{k-1} J + C_f \sum_{f \in D_f}^{n_f} H_f M_{k-1} H_f + C_r \sum_{r \in D_r}^{n_r} H_r M_{k-1} H_r, \\
&= J(J^{2(k-1)} + N_{k-1})J + C_f \sum_{f \in D_f}^{n_f} H_f (J^{2(k-1)} + N_{k-1}) H_f + C_r \sum_{r \in D_r}^{n_r} H_r (J^{2(k-1)} + N_{k-1}) H_r, \\
&= J^{2k} + C_f \sum_{f \in D_f}^{n_f} H_f N_{k-1} H_f + C_r \sum_{r \in D_r}^{n_r} H_r N_{k-1} H_r + J N_{k-1} J + C_f \sum_{f \in D_f}^{n_f} H_f J^{2(k-1)} H_f + C_r \sum_{r \in D_r}^{n_r} H_r J^{2(k-1)} H_r, \\
&= M_k + J N_{k-1} J + C_f \sum_{f \in D_f}^{n_f} H_f J^{2(k-1)} H_f + C_r \sum_{r \in D_r}^{n_r} H_r J^{2(k-1)} H_r, \\
&\succeq M_k.
\end{aligned} \tag{25}$$

The last equality is due to the later three terms are both PSD by assumption as they are symmetric in terms of left and right half of whole multiplication. As we can lower bound through M_k, diverging of N_k will lead to M_k and cause the whole product to diverge.

Second part:

Base step: k=1.

$$E[J_1^T J_1] = J^2 + N_1 \preceq (1-\epsilon)^2 I + N_1 = \sum_{r=0}^{1} \binom{1}{r} (1-\epsilon)^{2(1-r)} N_r. \tag{26}$$

The J^2 is bounded by $(1-\epsilon)^2 I$ due to our assumption. Now, we start with the inductive step

Inductive step: k-1.

$$
\begin{aligned}
&E[J_k^T J_{k-1}^T ... J_1^T J_1 ... J_{k-1} J_{k-1}] \preceq E[J_k^T (\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) J_k], \\
&= J(\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) J \\
&+ C_f \sum_{i \in D_f}^{n_f} H_i (\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) H_i + C_r \sum_{i \in D_r}^{n_r} H_i (\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) H_i, \\
&= J(\sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_r) J + \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} (C_f \sum_{i \in D_f}^{n_f} H_i N_r H_i + C_r \sum_{i \in D_r}^{n_r} H_i N_r H_i), \\
&\preceq \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-r)} N_r + \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_{r+1}, \\
&= \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-r)} N_r + \sum_{r=0}^{k-1} \binom{k-1}{r} (1-\epsilon)^{2(k-1-r)} N_{r+1}, \\
&= (1-\epsilon)^2 N_o + \sum_{r=1}^{k-1} (\binom{k-1}{r} + \binom{k-1}{r-1}) N_r + N_k, \\
&= (1-\epsilon)^2 N_o + \sum_{r=1}^{k-1} \binom{k}{r} N_r + N_k, \\
&= \sum_{r=0}^{k} \binom{k}{r} (1-\epsilon)^{2(k-r)} N_r.
\end{aligned}
\tag{27}
$$

□

The first and second inequality is due to the assumption in induction on previous step and we merge the coefficient in the last step through lemma 4.2. Finally, if we further have that $\mathrm{Tr}[N_r] \leq \epsilon \ \forall r$, then

$$
\begin{aligned}
&E[\mathrm{Tr}[J_k^T J_{k-1}^T ... J_1^T J_1 ... J_{k-1} J_{k-1}]], \\
&= \sum_{r=0}^{k} \binom{k}{r} (1-\epsilon)^{2k-r} \mathrm{Tr}[N_r], \\
&\leq \sum_{r=0}^{k} \binom{k}{r} (1-\epsilon)^{2(k-r)} \epsilon^r, \\
&\leq ((1-\epsilon)^2 + \epsilon)^k \leq (1-\epsilon)^k,
\end{aligned}
\tag{28}
$$

which will converge to zero when $k \to \infty$.

4.4 Proof of theorem 2.2

Theorem 4.8 (Restated). *Under the setup of Lemma 2.1, the unlearning process will diverge if the mix-Hessian eigenvalue exceeds a threshold determined by the coherence. In particular, if*

$$\lambda_{\max}(D) \;\geq\; \frac{\sqrt{2}\,\sigma}{\eta\Big((1-\alpha)\,n_f\,\sqrt{\frac{n_r}{B}-1}\;+\;\alpha\,n_r\,\sqrt{\frac{n_f}{B}-1}\Big)}\,, \tag{29}$$

then $\lim_{k\to\infty}\mathbb{E}\|w_k\|^2=\infty$. *Equivalently, condition* (7) *guarantees the unlearning algorithm will* escape *the original minima (diverge) due to the stochastic dynamics.*

Proof. To simplify the notation, we use the following

$$\begin{aligned} L_k &\in \{r,f\}^k \;\text{ (a string of length } k \text{ over the alphabet } \{r,f\}),\\ L_k[i] \;&\mapsto\; \text{the } i\text{-th symbol of } L_k,\;\; 1\leq i\leq k. \end{aligned} \tag{30}$$

We know that based on part one lemma 2.1, we can lower bound the N_k to lower bound the $E[\mathrm{Tr}[J_k^TJ_{k-1}^T...J_1^TJ_1...J_{k-1}J_k]]$. First, We can write the overall sum as follows:

$$\begin{aligned}
\mathrm{Tr}[N_k] &= \sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}\Big(\sum_{a_k\in D_{L_k[k]}}\sum_{a_{k-1}\in D_{L_k[k-1]}}\cdots\sum_{a_1\in D_{L_k[k]}}\Big)\,\mathrm{Tr}[H_{a_k}...H_{a_1}H_{a_1}...H_{a_k}],\\
&= \sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}\Big(\sum_{a_k\in D_{L_k[k]}}\sum_{a_{k-1}\in D_{L_k[k-1]}}\cdots\sum_{a_1\in D_{L_k[k]}}\Big)\|H_{a_k}...H_{a_1}\|_F^2,\\
&\geq \frac{1}{d}\sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}\Big(\sum_{a_k\in D_{L_k[k]}}\sum_{a_{k-1}\in D_{L_k[k-1]}}\cdots\sum_{a_1\in D_{L_k[k]}}\Big)\|H_{a_k}...H_{a_1}\|_{\mathcal{S}_1}^2,\\
&\geq \frac{1}{d}\sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}\Big(\sum_{a_k\in D_{L_k[k]}}\sum_{a_{k-1}\in D_{L_k[k-1]}}\cdots\sum_{a_1\in D_{L_k[k]}}\Big)\,\mathrm{Tr}[H_{a_k}...H_{a_1}]^2,\\
&\geq \frac{1}{n_rn_f}\frac{1}{d}\sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\,\mathrm{Tr}[H_{a_{L_k[k]}}...H_{a_{L_k[1]}}]^2,\\
&= \frac{1}{n_rn_f}\frac{1}{d}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\sum_{L_k\in\{r,f\}^k} C_r^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}\, C_f^{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}\,\mathrm{Tr}[H_{a_{L_k[k]}}...H_{a_{L_k[1]}}]^2,\\
&\geq \frac{1}{n_rn_f}\frac{1}{d}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\frac{1}{2^k}\Big(\sum_{L_k\in\{r,f\}^k}\mathrm{Tr}[C_r^{\frac{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}{2}}\, C_f^{\frac{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}{2}}H_{a_{L_k[k]}}...H_{a_{L_k[1]}}]\Big)^2,\\
&= \frac{1}{n_rn_f}\frac{1}{d}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\frac{1}{2^k}\Big(\mathrm{Tr}[\sum_{L_k\in\{r,f\}^k} C_r^{\frac{\sum_{i=1}^k \mathbf{1}\{L_k[i]=r\}}{2}}\, C_f^{\frac{\sum_{i=1}^k \mathbf{1}\{L_k[i]=f\}}{2}}H_{a_{L_k[k]}}...H_{a_{L_k[1]}}]\Big)^2,\\
&= \frac{1}{n_rn_f}\frac{1}{d}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\Big)\frac{1}{2^k}(C_r^{\frac12}+C_f^{\frac12})^{2k}\Big(\mathrm{Tr}[(\frac{C_r^{\frac12}}{C_r^{\frac12}+C_f^{\frac12}}H_{a_r}+\frac{C_f^{\frac12}}{C_r^{\frac12}+C_f^{\frac12}}H_{a_f})^k]\Big)^2,\\
&= \Big(\frac{1}{n_rn_f}\Big)^2\frac{1}{d}\frac{1}{2^k}(C_r^{\frac12}+C_f^{\frac12})^{2k}\Big(\sum_{a_r\in D_r}\sum_{a_f\in D_f}\mathrm{Tr}[(\frac{C_r^{\frac12}}{C_r^{\frac12}+C_f^{\frac12}}H_{a_r}+\frac{C_f^{\frac12}}{C_r^{\frac12}+C_f^{\frac12}}H_{a_f})^k]\Big)^2.
\end{aligned} \tag{31}$$

For the first and second inequality, we use lemma 4.3 and 4.4. For the third inequality, we reduce the summation to $\sum_{a_r\in D_r}\sum_{a_f\in D_f}$. As there are terms without D_r or D_f involved, we divided the whole equation by n_fn_r to ensure inequality. For the forth inequality, we use the lemma 4.1.

Before we try to connect the relationship between the quantity to the above, we first reindex the following:

$$\frac{1}{n_r n_f}\sum_{a_r\in D_r, a_f\in D_f}\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_r}+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_f}=\frac{1}{n_r n_f}\sum_{rf}D_{rf}=D, \tag{32}$$

where $D_{rf}=\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_r+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_f$ and the subscript indicates that summing over corresponding subset (retain and forget set). Now, we proceed to relate different quantities

$$\begin{aligned}
&\mathrm{Tr}[(\frac{1}{n_r n_f}\sum_{a_r\in D_r, a_f\in D_f}\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_r}+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_f})^k]=\mathrm{Tr}[(\frac{1}{n_r n_f})^k(\sum_{rf}D_{rf})^k],\\
&=\mathrm{Tr}[(\frac{1}{n_r n_f})^k\sum_{rf_1}\sum_{rf_2}...\sum_{rf_k}D_{rf_1}D_{rf_2}...D_{rf_{k-1}}D_{rf_k}],\\
&\le d(\frac{1}{n_r n_f})^k\sum_{rf_1}\sum_{rf_2}...\sum_{rf_k}\|D_{rf_k}^{\frac{1}{2}}D_{rf_1}^{\frac{1}{2}}\|_F\|D_{rf_1}^{\frac{1}{2}}D_{rf_2}^{\frac{1}{2}}\|_F...\|D_{rf_{k-1}}^{\frac{1}{2}}D_{rf_k}^{\frac{1}{2}}\|_F,\\
&=d(\frac{1}{n_r n_f})^k\sum_{rf_1}\sum_{rf_2}...\sum_{rf_k}S_{rf_k,rf_1}S_{rf_1,rf_2}...S_{rf_{k-1},rf_k},\\
&=d(\frac{1}{n_r n_f})^k\,\mathrm{Tr}(S^k),\\
&\le d^2(\frac{1}{n_r n_f})^k\lambda_1(S)^k.
\end{aligned} \tag{33}$$

Therefore, we say that

$$\mathrm{Tr}[D^k]\le d^2(\frac{1}{n_r n_f})^k\lambda_1(S)^k, \tag{34}$$

and we can have that

$$\begin{aligned}
\frac{(n_r n_f)^k\,\mathrm{Tr}[D^k]}{d^2\sigma^k}&\le\frac{(n_r n_f)^k\,\mathrm{Tr}[D^k]}{d^2\lambda_1(S)^k}\max_{i\in D_r j\in D_f}\lambda_1(\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_i+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_j)^k,\\
&\le\sum_{a_r\in D_r}\sum_{a_f\in D_f}\mathrm{Tr}[(\frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_r}+\frac{C_f^{\frac{1}{2}}}{C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}}}H_{a_f})^k].
\end{aligned} \tag{35}$$

Therefore, we can conclude that

$$\begin{aligned}
\mathrm{Tr}[N_k]&\ge\frac{1}{d}\frac{1}{n_f n_r}\frac{1}{2^k}(C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}})^{2k}(\frac{(n_r n_f)^k\,\mathrm{Tr}[D^k]}{d^2\sigma^k})^2,\\
&\ge\frac{1}{d^5}\frac{1}{n_f n_r}\frac{1}{2^k}(C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}})^{2k}\frac{(n_r n_f)^{2k}\lambda_1(D)^{2k}}{\sigma^{2k}},\\
&\ge\frac{1}{d^5}\frac{1}{n_f n_r}\frac{1}{2^k}(C_r^{\frac{1}{2}}+C_f^{\frac{1}{2}})^{2k}\frac{(n_r n_f)^{2k}\lambda_1(D)^{2k}}{\sigma^{2k}}.
\end{aligned} \tag{36}$$

Lastly, we can see that whether the trace diverge or not depend on those term with power of k. Therefore, by rearranging and plug in the definition of the coefficient into those terms, we can have that

$$\lambda_1(D)\ge\frac{\sqrt{2}\sigma}{\eta}\Big((1-\alpha)n_f(\frac{n_r}{B}-1)^{\frac{1}{2}}+\alpha n_r(\frac{n_f}{B}-1)^{\frac{1}{2}}\Big)^{-1}, \tag{37}$$

which is the condition for diverging behavior □

4.5 Proof of theorem 2.3

Theorem 4.9 ((Restate) Matching lower bound.)**.** *Suppose $\lambda_{\max}(D)$ and σ satisfy*

$$\lambda_{\max}(D) \;\le\; \frac{2\,\sigma}{\eta\, C_r' \left(\sigma + n_f\left(\frac{n_r}{B} - 1\right)\right)}, \tag{38}$$

where $C_r' = \sqrt{C_r}/(\sqrt{C_r} + \sqrt{C_f})$ (with C_r, C_f from Definition 2). Then there exists a choice of PSD Hessians $\{H_i\}$ for the retain and forget sets such that the unlearning update converges (i.e. $\lim_{k\to\infty} \mathbb{E}\|w_k\|^2 = 0$) under those Hessians.

Proof. We prove by construction in the following manner. We construct the retain set by setting $H_i = m \cdot e_1 e_1^T \;\; \forall\, i \;\in [\frac{\sigma}{n_f}]$. ($m = (C_r')^{-1}\frac{\lambda_1(D) n_r}{\frac{\sigma}{n_f}}$ and $C_r' = \frac{C_r^{\frac{1}{2}}}{C_r^{\frac{1}{2}} + C_f^{\frac{1}{2}}}$ and the definition of C_r and C_f are mentioned in definition 2.) Otherwise, we set the Hessian to be zero matrix. For the forget set, we set all matrix to be zero matrix.

We first verify that the eigenvalue of mix-Hessian is indeed the assigned value $\lambda_1(D)$.

$$D = \frac{1}{n_r n_f}\sum_{rf} C_r' H_r + C_f' H_f = \frac{1}{n_r n_f}\sum_{rf} C_r'(C_r')^{-1}\frac{\lambda_1(D) n_r}{\frac{\sigma}{n_f}} e_1 e_1^T = \lambda_1(D) e_1 e_1^T, \tag{39}$$

and we have that the construction indeed have the corresponding mix-Hessian eigenvalue.

We know verify that the coherence measure is of the assigned value σ. We first note that the element of the coherence matrix is:

$$S_{rf,r'f'} = \sqrt{\mathrm{Tr}[(C_r' H_r + C_f' H_f)(C_r' H_{r'} + C_f' H_{f'})]} = C_r' m = \frac{\lambda_1(D) n_r}{\frac{\sigma}{n_f}}, \;\; \forall r, r' \in [\frac{\sigma}{n_f}]. \tag{40}$$

else it is zero. We know that there is $n_f \cdot \frac{\sigma}{n_f} = \sigma$ nonzero elements for each row and column. We note that we will also need to divide the coherence matrix by $\max_{rf} D_{rf} = \max_{rf} C_r' H_r + C_f' H_f = \frac{\lambda_1(D) n_r}{\frac{\sigma}{n_f}}$. Finally, each element is 1 after this division, and we can get the eigenvalue of the matrix to be σ and verify that the construction is valid.

Now, we note that in our construction, we have each step J_i to commute to each other since every matrix involved is diagonal, so we can focus on one step to calculate the condition that lead to diverging or converging and since we only intentionally set our matrix to be one dimensional, we can study behavior on only one axis e_1 by plugging in the above as follows:

$$\begin{aligned} e_1 E[J_1 J_1] e_1 &= e_1[I - 2\eta(1-\alpha)H_R + \eta^2(1-\alpha)^2 H_r^2 + \eta^2(1-\alpha)^2 \sum_r H_r^2] e_1, \\ &= 1 - 2\eta(1-\alpha)(C_r')^{-1}\lambda_1(D) + \eta^2(1-\alpha)^2(C_r')^{-2}\lambda_1(D)^2 + \frac{(C_r')^{-2}}{\sigma}\lambda_1(D)^2\eta^2(1-\alpha)^2 n_f(\frac{n_r}{B} - 1). \end{aligned} \tag{41}$$

As we want to study the converging behavior, we want the above to be smaller than 1 to have repetitive multiplication lead to converging.

$$\begin{aligned} &1 - 2\eta(1-\alpha)(C_r')^{-1}\lambda_1(D) + \eta^2(1-\alpha)^2(C_r')^{-2}\lambda_1(D)^2 + \frac{(C_r')^{-2}}{\sigma}\lambda_1(D)^2\eta^2(1-\alpha)^2 n_f(\frac{n_r}{B} - 1) \le 1, \\ &\implies 2 \ge \eta(1-\alpha)(C_r')^{-1}\lambda_1(D)(1 + \frac{n_f}{\sigma}(\frac{n_r}{B} - 1)), \\ &\implies 2 \ge \frac{\eta}{\sigma}(1-\alpha)(C_r')^{-1}\lambda_1(D)(\sigma + n_f(\frac{n_r}{B} - 1)), \\ &\implies \lambda_1(D) \le \frac{2\sigma}{\eta} C_r'\Big((1-\alpha)(\sigma + n_f(\frac{n_r}{B} - 1))^{-1}\Big). \end{aligned} \tag{42}$$

□

4.6 Proof of theorem 4.6

Theorem 4.10 (Restate)**.** *Under the data model of Definition 5 and the two-layer ReLU CNN defined above, suppose the network is trained to near-zero training loss. Then with probability at least $1 - 8\delta$ (over the random draw of the dataset), the largest eigenvalue of the coherence matrix S for the retain/forget split satisfies*

$$\lambda_{\max}(S) \;\le\; \mathcal{O}\Big(n_r\, n_f\, d\sigma^2\Big[(\sqrt{C'_r}+\sqrt{C'_f})^2\,(\mathrm{SNR})^2 + (C'_r + C'_f)\Big]\Big)\,, \tag{43}$$

$$\max_{rf} \lambda_{\max}(D_{rf}) \le \mathcal{O}((C'_r + C'_f)(d\sigma^2(\mathrm{SNR})^2 + 1)), \tag{44}$$

where C'_r and C'_f are the normalized retain/forget weight fractions as defined in Theorem 2.3. Consider division of two quantities and we can find that for small SNR limit and large SNR limit:

$$\lim_{SNR\to 0} \frac{\lambda_{\max}(S)^{upper}}{\max_{rf} D_{rf}^{upper}} = \mathcal{O}(n_r n_f) \;\; , \lim_{SNR\to\infty} \frac{\lambda_{\max}(S)^{upper}}{\max_{rf} D_{rf}^{upper}} = \mathcal{O}(n_r n_f (1 + \frac{2\sqrt{C'_r C'_f}}{C'_r + C'_f})). \tag{45}$$

Proof. We first calculate the gradient of one sample respective to one of the $w_{j,r}$.

$$\frac{\partial \ell(y_i \cdot f(W, x_i))}{\partial w_{j,r}} = \ell_i' \cdot \frac{j}{m} \cdot (\mathbf{1}_{\{\langle w_{j,r}, y_i \cdot \boldsymbol{\mu}\rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j,r}, \xi_i\rangle > 0\}} \cdot y_i \cdot \xi_i). \tag{46}$$

There are several index in the above equation (i.e., j and r) which we use to take derivative with respect to a specific feature weight vector. We will continue to use this notation for future calculation. Now, we move to calculate the second derivative with respect to two different feature of weights for data i as follows:

$$\frac{\partial^2 \ell(y_i \cdot f(W, x_i))}{\partial w_{j,r} \partial w_{j',r'}} =$$
$$\ell_i'' \cdot \frac{jj'}{m^2} \cdot (\mathbf{1}_{\{\langle w_{j,r}, y_i \cdot \boldsymbol{\mu}\rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j,r}, \xi_i\rangle > 0\}} \cdot y_i \cdot \xi_i)(\mathbf{1}_{\{\langle w_{j',r'}, y_i \cdot \boldsymbol{\mu}\rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j',r'}, \xi_i\rangle > 0\}} \cdot y_i \cdot \xi_i)^T. \tag{47}$$

The above is one block of the Hessian. In the following, we will simplify the notation for indicator function (derivative of ReLU) to $\mathbf{1}_{j',r',y_i\cdot\boldsymbol{\mu}}$ and $\mathbf{1}_{j',r',\boldsymbol{\xi}}$ to ease the heavy notation. To calculate the coherence matrix, we need to calculate trace of Hessian product for different sample,

$$
\begin{aligned}
&\mathrm{Tr}[H_i H_k] = \sum_{j,j',r,r'} \mathrm{Tr}[\frac{\partial^2 \ell(y_i \cdot f(W,x_i))}{\partial w_{j,r} \partial w_{j',r'}} \frac{\partial^2 \ell(y_k \cdot f(W,x_k))}{\partial w_{j',r'} \partial w_{j,r}}], \\
&= \frac{\ell_i'' \ell_k''}{m^4} \sum_{j,j',r,r'} (\mathbf{1}_{j,r,y_k \cdot \boldsymbol{\mu}} \cdot \boldsymbol{\mu} + \mathbf{1}_{j,r,\xi_k} \cdot y_k \cdot \xi_k)^T (\mathbf{1}_{j,r,y_i \cdot \boldsymbol{\mu}} \cdot \boldsymbol{\mu} + \mathbf{1}_{j,r,\xi_i} \cdot y_i \cdot \xi_i) \\
&(\mathbf{1}_{j',r',y_i \cdot \boldsymbol{\mu}} \cdot \boldsymbol{\mu} + \mathbf{1}_{j',r',\xi_i} \cdot y_i \cdot \xi_i)^T (\mathbf{1}_{j',r',y_k \cdot \boldsymbol{\mu}} \cdot \boldsymbol{\mu} + \mathbf{1}_{j',r',\xi_k} \cdot y_k \cdot \xi_k), \\
&= \frac{\ell_i'' \ell_k''}{m^4} \Big((\sum_{j,r} \mathbf{1}_{j,r,y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j,r,y_i \cdot \boldsymbol{\mu}}) \|\boldsymbol{\mu}\|^2) + (\sum_{j,r} \mathbf{1}_{j,r,y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j,r,\xi_k}) \boldsymbol{\mu}^T \xi_k + \\
&(\sum_{j,r} \mathbf{1}_{j,r,y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j,r,\xi_i}) \boldsymbol{\mu}^T \xi_i + (\sum_{j,r} \mathbf{1}_{j,r,\xi_k} \mathbf{1}_{j,r,\xi_i}) \xi_k^T \xi_i \Big), \\
&\Big((\sum_{j,r} \mathbf{1}_{j',r',y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j',r',y_i \cdot \boldsymbol{\mu}}) \|\boldsymbol{\mu}\|^2) + (\sum_{j',r'} \mathbf{1}_{j',r',y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j',r',\xi_k}) \boldsymbol{\mu}^T \xi_k + \\
&(\sum_{j',r'} \mathbf{1}_{j',r',y_k \cdot \boldsymbol{\mu}} \mathbf{1}_{j',r',\xi_i}) \boldsymbol{\mu}^T \xi_i + (\sum_{j',r'} \mathbf{1}_{j',r',\xi_k} \mathbf{1}_{j',r',\xi_i}) \xi_k^T \xi_i \Big), \\
&\leq 4 \frac{\ell_i'' \ell_k''}{m^2} (\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_k| + |\boldsymbol{\mu}^T \xi_i| + |\xi_i^T \xi_k|)^2, \\
&\leq \frac{4}{m^2} (\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_k| + |\boldsymbol{\mu}^T \xi_i| + |\xi_i^T \xi_k|)^2.
\end{aligned} \tag{48}
$$

We now analyze each term in the coherence matrix.

$$
\begin{aligned}
S_{r_1 f_{1'}, r_2 f_{2'}} &= \sqrt{\mathrm{Tr}((C_r' H_{r_1} + C_f' H_{f_{1'}})(C_r' H_{r_2} + C_f' H_{f_{2'}}))}, \\
&= \sqrt{\mathrm{Tr}[C_r'^2 H_{r_1} H_{r_2}] + \mathrm{Tr}[C_r' C_f' H_{r_1} H_{f_{2'}}] + \mathrm{Tr}[C_r' C_f' H_{f_{1'}} H_{r_2}] + \mathrm{Tr}[C_f'^2 H_{f_{1'}} H_{r_{2'}}]}, \\
&\leq \sqrt{\mathrm{Tr}[C_r'^2 H_{r_1} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{r_1} H_{f_{2'}}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{f_{1'}} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_f'^2 H_{f_{1'}} H_{r_{2'}}]},
\end{aligned} \tag{49}
$$

where the C_r' and C_f' are respectively the normalized coefficient mentioned in the previous section.

As our goat is to estimate the largest eigenvalue of the coherence matrix and its relation between different variables in the design. To estimate the largest eigenvalue, we incur ϵ-net that is used random matrix theory

$$
\lambda_1 = \sup_{\|x\|=1} \langle x, Sx \rangle. \tag{50}
$$

For one vector x, we can write the expression as summation:

$$
\begin{aligned}
&\langle x, Sx \rangle = \sum_{r_1 f_{1'}, r_2 f_{2'}} S_{r_1 f_{1'}, r_2 f_{2'}} x_{r_1 f_{1'}} x_{r_2 f_{2'}}, \\
&\leq \sum_{r_1 f_{1'}, r_2 f_{2'}} (\sqrt{\mathrm{Tr}[C_r'^2 H_{r_1} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{r_1} H_{f_{2'}}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{f_{1'}} H_{r_2}]} + \sqrt{(\mathrm{Tr}[C_f'^2 H_{f_{1'}} H_{r_{2'}}])} x_{r_1 f_{1'}} x_{r_2 f_{2'}}.
\end{aligned} \tag{51}
$$

We can estimate the above through the random matrix theory and upper bound the largest eigenvalue through the elementwise calculation that we set up and use the tail bound for each random variable to provide relationship between controlled variable and the resulting largest eigenvalue. We first separate the discussion into several cases. First case, when we have four different samples r_1, r_2, f_1', f_2', we can have that

$$
\begin{aligned}
&(\sqrt{\mathrm{Tr}[C_r'^2 H_{r_1} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{r_1} H_{f_{2'}}]} + \sqrt{\mathrm{Tr}[C_r' C_f' H_{f_{1'}} H_{r_2}]} + \sqrt{\mathrm{Tr}[C_f'^2 H_{f_{1'}} H_{r_{2'}}]}) x_{r_1 f_{1'}} x_{r_2 f_{2'}}, \\
&\leq (C_r' \frac{2}{m}(\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_{r1}| + |\boldsymbol{\mu}^T \xi_{r2}| + |\xi_{r1}^T \xi_{r2}|) + \sqrt{C_r' C_f'} \frac{2}{m}(\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_{r1}| + |\boldsymbol{\mu}^T \xi_{f2'}| + |\xi_{r1}^T \xi_{f2'}|), \\
&+ \sqrt{C_r' C_f'} \frac{2}{m}(\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_{r2}| + |\boldsymbol{\mu}^T \xi_{f1'}| + |\xi_{f1'}^T \xi_{r2}|) + C_f' \frac{2}{m}(\|\boldsymbol{\mu}\|^2 + |\boldsymbol{\mu}^T \xi_{f2'}| + |\boldsymbol{\mu}^T \xi_{f1'}| + |\xi_{f1'}^T \xi_{f2'}|)) x_{r_1 f_{1'}} x_{r_2 f_{2'}}, \\
&\leq (\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|)(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r2}\| + \sqrt{C_f'}\|\xi_{f2'}\|) x_{r_1 f_{1'}} x_{r_2 f_{2'}}.
\end{aligned} \tag{52}
$$

Our aiming in the above is to establish relationship between different variables used in the CNN network. In the above, we can see that we can upper bound the eigenvalue by the cross product of the vector $v_{rf} = \sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|$ since the coherence matrix is upper bound elementwise by the vector. (i.e., $\lambda_1(S) \leq \lambda_1(vv^T) = \|v^T v\|^2$) and this turns the estimation of the eigenvalue into estimation of the magnitude of the vector.

Now, we analyze the $v^T v$,

$$
\begin{aligned}
v^T v &= \sum_{rf}(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|)(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\| + \sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|), \\
&= \sum_{rf}(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)^2 + 2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)(\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|) + (\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|)^2, \\
&= n_r n_f(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)^2 + 2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)\sum_{rf}(\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|) + \\
&\sum_{rf}(C_r'\|\xi_{r1}\|^2 + C_f'\|\xi_{f1'}\|^2 + \sqrt{C_r' C_f'}'\|\xi_{r1}\|\|\xi_{f1'}\|).
\end{aligned} \tag{53}
$$

We analyze different terms as follows:

$$
\begin{aligned}
&2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)\sum_{rf}(\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|) = \\
&\qquad 2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)(n_f \sum_r \sqrt{C_r'}\|\xi_{r1}\| + n_r \sum_f \sqrt{C_f'}\|\xi_{f1'}\|).
\end{aligned} \tag{54}
$$

We know that $\|\xi_{r1}\|, \|\xi_{f1'}\|$ are chi-distribution which is also sub-exponential distribution. We can utilize the tail bound for summation of the sub-exponential random variables to obtain high probability bound on the summation. We can have that with probability 2δ,

$$
\begin{aligned}
&2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)\sum_{rf}(\sqrt{C_r'}\|\xi_{r1}\| + \sqrt{C_f'}\|\xi_{f1'}\|), \\
&\leq 2(\sqrt{C_r'}\|\boldsymbol{\mu}\| + \sqrt{C_f'}\|\boldsymbol{\mu}\|)(n_r n_f \sqrt{C_r'}\sigma\sqrt{d} + n_f n_r \sqrt{C_f'}\sigma\sqrt{d} + n_f\sqrt{\frac{n_r\sigma^2}{C_1}\log(\frac{2}{\delta})} + n_r\sqrt{\frac{n_f\sigma^2}{C_1}\log(\frac{2}{\delta})}).
\end{aligned} \tag{55}
$$

Now, we move to the next chi-square distribution terms $C_r' \sum \|\xi_{r_1}\|^2$, $C_f' \sum \|\xi_{f_1'}\|^2$. By using the lemma 4.6, we can have that with probability $1-\delta$,

$$
C_r' \sum_{rf} \|\xi_{r_1}\|^2 \leq C_r'(n_f n_r \sigma^2 d + n_f \sqrt{\frac{n_r \sigma^4 d}{C_2}\log(\frac{2}{\delta})}). \tag{56}
$$

and so is the $C'_f \sum \|\xi_{f'_1}\|^2$,

$$C'_f \sum_{rf} \|\xi_{r_1}\|^2 \leq C'_f(n_f n_r \sigma^2 d + n_r \sqrt{\frac{n_f \sigma^4 d}{C_2} \log(\frac{2}{\delta})}). \tag{57}$$

The term $\sum_{rf} \sqrt{C'_r C'_f} \|\xi_{r1}\| \|\xi_{f1'}\|$ can also be dealt with in the same manner,

$$\begin{aligned}
&\sqrt{C'_r C'_f} \sum_{rf} \|\xi_{r1}\| \|\xi_{f1'}\| \leq \sqrt{C'_r C'_f}(\sum_r \|\xi_{r1}\|)(\sum_f \|\xi_{f1'}\|), \\
&\leq \sqrt{C'_r C'_f}(n_r \sqrt{2}\sigma \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} + \sqrt{\frac{n_r \sigma^2}{C_1} \log(\frac{2}{\delta})})(n_f \sqrt{2}\sigma \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} + \sqrt{\frac{n_f \sigma^2}{C_1} \log(\frac{2}{\delta})}).
\end{aligned} \tag{58}$$

To simplify the analysis, we only keep terms with magnitude at least $n_f n_r$. We will reach that with probability $1 - 6\delta$

$$\begin{aligned}
\lambda_1(S) \leq \mathcal{O}\Big(n_f n_r \big((\sqrt{C'_r} + \sqrt{C'_f})^2 \|\boldsymbol{\mu}\|^2 + 2\sqrt{2}(\sqrt{C'_r} + \sqrt{C'_f})^2 \|\boldsymbol{\mu}\| \sigma (\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}) \\
+ (C'_r + C'_f)\sigma^2 d + 2\sigma^2 \sqrt{C'_r C'_f} (\frac{\Gamma((d+1)/2)}{\Gamma(d/2)})^2)\Big).
\end{aligned} \tag{59}$$

To see how signal noise ratio (SNR $= \frac{\|\boldsymbol{\mu}\|}{\sigma \sqrt{d}}$) interact with the right hand side, we extract a factor $\sigma^2 d$ from all terms involved:

$$\begin{aligned}
\lambda_1(S) &\leq \mathcal{O}\Big(n_f n_r \sigma^2 d \big((\sqrt{C'_r} + \sqrt{C'_f})^2 (\text{SNR})^2 + \frac{2\sqrt{2}}{\sqrt{d}} (\sqrt{C'_r} + \sqrt{C'_f})^2 (\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}) (\text{SNR}) \\
&\quad + (C'_r + C'_f) + \frac{2}{d} \sqrt{C'_r C'_f} (\frac{\Gamma((d+1)/2)}{\Gamma(d/2)})^2)\Big), \\
&\leq \mathcal{O}\Big(n_f n_r \sigma^2 d \big((\sqrt{C'_r} + \sqrt{C'_f})^2 (\text{SNR})^2 + (C'_r + C'_f)\big)\Big).
\end{aligned} \tag{60}$$

where in the last equation, we omit terms with d in the denominator as it tends to be large when we consider larger network.

For the second part of the proof, we know that H_i have block structures as follows:

$$\begin{aligned}
&\frac{\partial^2 \ell(y_i \cdot f(W, x_i))}{\partial w_{j,r} \partial w_{j',r'}} = \\
&\ell''_i \cdot \frac{jj'}{m^2} \cdot (\mathbf{1}_{\{\langle w_{j,r}, y_i \cdot \boldsymbol{\mu} \rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j,r}, \xi_i \rangle > 0\}} \cdot y_i \cdot \xi_i)(\mathbf{1}_{\{\langle w_{j',r'}, y_i \cdot \boldsymbol{\mu} \rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j',r'}, \xi_i \rangle > 0\}} \cdot y_i \cdot \xi_i)^T.
\end{aligned} \tag{61}$$

We can see that the whole H_i matrix can be regarded as outer product of vector vv^T where we have v_{jr} being

$$v_{jr} = \frac{l''_i j}{m} (\mathbf{1}_{\{\langle w_{j,r}, y_i \cdot \boldsymbol{\mu} \rangle > 0\}} \cdot \boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j,r}, \xi_i \rangle > 0\}} \cdot y_i \cdot \xi_i). \tag{62}$$

We can immediately know that the eigenvalue of the H_i will be upper bounded by $v^T v$ as follows:

$$\begin{aligned}\lambda_{\max}(H_i) \leq v^T v &= \frac{l_i''}{m^2}\sum_{jr}(\mathbf{1}_{\{\langle w_{j,r}, y_i\cdot\boldsymbol{\mu}\rangle>0\}}\cdot\boldsymbol{\mu} + \mathbf{1}_{\{\langle w_{j,r},\xi_i\rangle>0\}}\cdot y_i\cdot\xi_i)^2,\\ &\leq 2\frac{l_i''}{m^2}\sum_{jr}\mathbf{1}_{\{\langle w_{j,r}, y_i\cdot\boldsymbol{\mu}\rangle>0\}}\|\boldsymbol{\mu}\|^2 + \mathbf{1}_{\{\langle w_{j,r},\xi_i\rangle>0\}}\|\xi_i\|^2,\\ &\leq 2\frac{l_i''}{m^2}\sum_{jr}\|\boldsymbol{\mu}\|^2 + \|\xi_i\|^2,\\ &\leq 2\frac{1}{m^2}\sum_{jr}\|\boldsymbol{\mu}\|^2 + \|\xi_i\|^2,\\ &= C(\|\boldsymbol{\mu}\|^2 + \|\xi_i\|^2),\end{aligned} \tag{63}$$

where we use C to encompass all constants.

To bound the $\max_{rf}\lambda_{\max}(D_{rf}) = \lambda_{\max}(C_r' H_r + C_f' H_f)$, we can use the following:

$$\lambda_{\max}(D_{rf}) = \lambda_{\max}(C_r' H_r + C_f' H_f) \leq C_r'\lambda_{\max}(H_r) + C_f'\lambda_{\max}(H_f). \tag{64}$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we can upper bound the the H_r with the following ($\|\xi_i\|$ is subexponential):

$$\begin{aligned}\max_{1\leq i\leq n_r} C\big(\|\boldsymbol{\mu}\|^2 + \|\xi_i\|^2\big) \;&\leq\; C\left(\|\boldsymbol{\mu}\|^2 + \sigma^2\Big[d \;+\; 2\sqrt{d\log\frac{n_r}{\delta}} \;+\; 2\log\frac{n_r}{\delta}\Big]\right),\\ &\leq \mathcal{O}(\|\boldsymbol{\mu}\|^2 + \sigma^2 d).\end{aligned} \tag{65}$$

Similarly, we can have the bound on H_f which is of same order and jointly we can have that with probability $1-8\delta$

$$\lambda_{\max}(D_{rf}) \leq \mathcal{O}((C_r' + C_f')(\|\boldsymbol{\mu}\|^2 + \sigma^2 d)) = \mathcal{O}((C_r' + C_f')\sigma^2 d(\text{SNR}^2 + 1)) \tag{66}$$

Last is the division and take the limit and we can have the following:

$$\lim_{\text{SNR}\to 0}\frac{\lambda_{\max}(S)^{\text{upper}}}{\max_{rf} D_{rf}^{\text{upper}}} = \mathcal{O}(n_r n_f) \;\;, \lim_{\text{SNR}\to\infty}\frac{\lambda_{\max}(S)^{\text{upper}}}{\max_{rf} D_{rf}^{\text{upper}}} = \mathcal{O}(n_r n_f(1 + \frac{2\sqrt{C_r' C_f'}}{C_r' + C_f'})). \tag{67}$$

□