Appendix

In this appendix we provide additional experiments (Appendix A), experimental details (Appendix B) and theoretical results (Appendix C).

A. Additional Experiments

A.1. Forgetting an entire class

In the main paper we considered forgetting a random subset of 10% of the training data. Here we consider instead the problem of completely forgetting all samples of a given class in a single forgetting request. In Figures 6 and 7, we observe that also in this setting our proposed method outperforms other methods and is robust to different readout functions. Note that for the case of removing an entire class the target forget error (i.e. the error on the class to forget) is 100%.

Figure 6. Readout function plot similar to Figure 3 for Caltech-256 dataset, where we forget an entire class rather a sequence of randomly sampled data subsets.

Figure 7. Readout function plot similar to Figure 3 for FGVC-Aicrafts dataset, where we forget an entire class rather a sequence of randomly sampled data subsets.

A.2. Role of L_2-Regularization

We plot the amount of remaining information and the test error as a function of the L_2 regularization coefficient. Note that instead of incorporating weight decay directly in the optimization step, as it is often done, we explicitly add the L_2 regularization to the loss function. As expected theoretically (Theorem 3), increasing the regularization coefficient makes the training optimization problem more strongly convex, which in turn makes forgetting easy. However, increasing weight decay too much also hurts the accuracy of the model. Hence there is a trade-off between the amount of remaining information and the amount of regularization with respect to the regularization. We plot the trade-off in Figure 8.

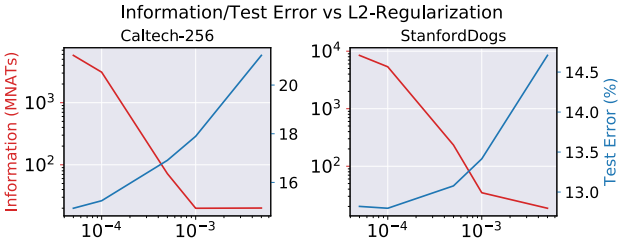

Figure 8. Plot of the amount of remaining information and test error vs the L_2 regularization coefficient. We forget 10% of the training data sequentiall through 10 forgetting request.

A.3. More experiments using SGD for forgetting

We repeat the same experiments as in Figure 3 on the following datasets: *Stanford Dogs, MIT-67, CIFAR-10, CUB-200, FGVC Aircrafts*. Overall, we observe consistent results over all datasets.

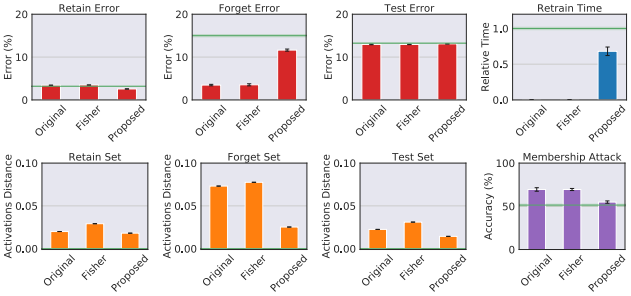

Figure 9. Same experiments as Figure 3 for StanfordDogs.

Figure 10. Same experiments as Figure 3 for MIT-67.

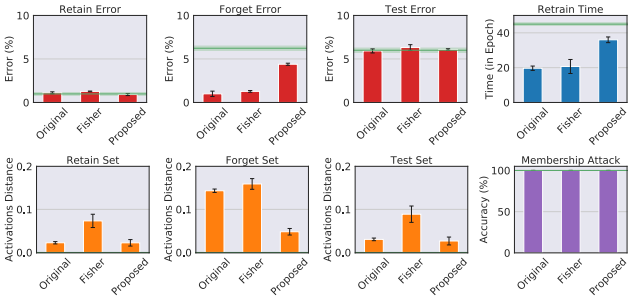

Figure 11. Same experiments as Figure 3 for CIFAR-10.

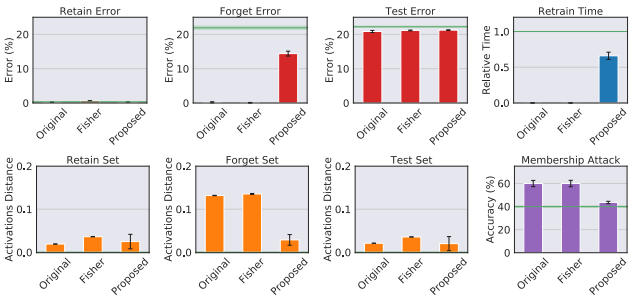

Figure 12. Same experiments as Figure 3 for CUB-200.

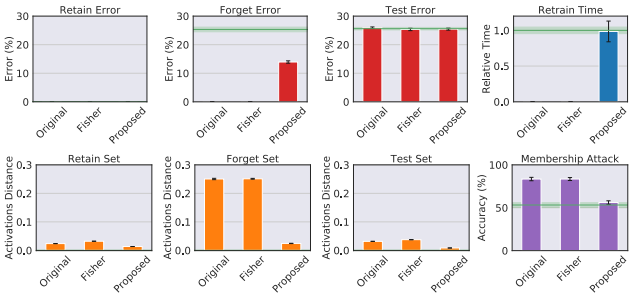

Figure 13. Same experiments as Figure 3 for FGVC-Aircrafts.

A.4. Information vs Noise/Epochs

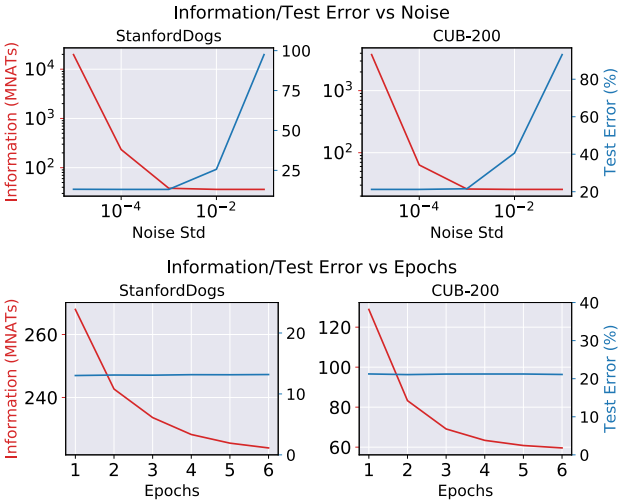

Figure 14. Same experiment as Figure 2 for Stanforddogs and CUB-200 datasets.

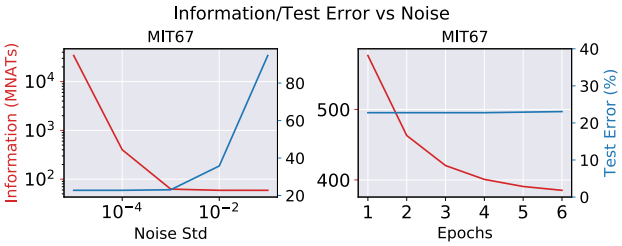

Figure 15. Same experiment as Figure 2 for MIT67.

B. Experimental Details

We use a ResNet-50 pre-trained on ImageNet. For the plots in Figure 1, we train ML-Forgetting model using SGD for 50 epochs with batch size 64, learning rate lr=0.05, momentum=0.9, weight decay=0.00001 where the learning rate is annealed by 0.1 at 25 and 40 epochs. We explicitly add the L_2 regularization to the loss function instead of incorporating it in the SGD update equation. We only linearize the final layers of ResNet-50 and scale the one-hot vectors by 5 while using the MSE loss. For fine-grained datasets, FGVC-Aircrafts and CUB-200, in addition to the ImageNet pre-training, we also pre-train them using randomly sampled 30% of the training data (which we assume is part of the core set).

For the training the ML-Forgetting model in the readout functions and information plots using SGD, we use the same experimental setting as above with a increased weight decay=0.0005 for Caltech-256,StanfordDogs and CIFAR-10 and 0.001 for MIT-67,CUB-200 and FGVC-Aircrafts. We use a higher value of weight decay to increase the strong convexity constant of the training loss function, which facilitates forgetting (see Lemma 4).

For forgetting using ML-Forgetting model in the readout function/information plots using SGD (ML-Forgetting to minimize eq. (11)), we use momentum=0.999 and decrease the learning rate by 0.5 per epoch. We run SGD for 3 epochs with an initial lr=0.01 for Caltech-256, StanfordDogs and CIFAR-10 and run it for 4 epochs with initial lr=0.025 for MIT-67, CUB-200 and FGVC-Aircrafts.

C. Theoretical Results

Lemma 1. *Let* $\mathbf{x}, \mathbf{y}$ *be two random vectors such that* $\mathbb{E}\|\mathbf{x}\|^2, \mathbb{E}\|\mathbf{y}\|^2 > 0$. *Then we have the following, for any* $\alpha > 0$*:*

$$\mathbb{E}\|\mathbf{x}+\mathbf{y}\|^2 \leq (1+\alpha)\mathbb{E}\|\mathbf{x}\|^2 + (1+\frac{1}{\alpha})\mathbb{E}\|\mathbf{y}\|^2$$

Proof.

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}+\mathbf{y}\|^2 &= \mathbb{E}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle) \\
&\leq \mathbb{E}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2|\langle \mathbf{x}, \mathbf{y}\rangle|) \\
&\overset{(a)}{\leq} \mathbb{E}\Big(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\sqrt{\|\mathbf{x}\|^2}\sqrt{\|\mathbf{y}\|^2}\Big) \\
&= \mathbb{E}\Big(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\sqrt{\|\mathbf{x}\|^2\alpha}\sqrt{\frac{\|\mathbf{y}\|^2}{\alpha}}\Big) \\
&\overset{(b)}{\leq} \mathbb{E}\Big(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{x}\|^2\alpha + \|\mathbf{y}\|^2\frac{1}{\alpha}\Big) \\
&= (1+\alpha)\mathbb{E}\|\mathbf{x}\|^2 + (1+\frac{1}{\alpha})\mathbb{E}\|\mathbf{y}\|^2
\end{aligned} \tag{16}$$

for any $\alpha > 0$, where (a) follows from the Cauchy-Schwarz inequality and (b) follows from the AM-GM inequality. □

Lemma 2. *Let* $C = \{\mathbf{w} | \mathbf{w} \in \mathbb{R}^d$ *and* $\|\mathbf{w}\| \leq R < \infty\}$ *and* $\ell(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^+$ *be a strongly convex function with* $max_{\mathbf{w}\in C}\ell(\mathbf{w}) < \infty$, $G \triangleq max_{\mathbf{w}\in C}\|\nabla\ell(\mathbf{w})\| < \infty$ *and* $\mathbf{w}^* = argmin_{\mathbf{w}\in\mathbb{R}^d}\ell(\mathbf{w})$ *such that* $\|\mathbf{w}^*\| < \infty$. *Then* $\forall \mathbf{w}_1, \mathbf{w}_2 \in C$, *we have that* $|\ell(\mathbf{w}_1) - \ell(\mathbf{w}_2)| \leq G\|\mathbf{w}_1 - \mathbf{w}_2\|$. *When* $\ell(\mathbf{w})$ *is also quadratic with* $\beta = \lambda_{Max}(\nabla^2\ell(\mathbf{w}))$, *the maximum eigen value of the Hessian, we have that* $G = \beta\big(R + \|\mathbf{w}^*\|\big)$.

Proof. Let $g(t) = \ell(\mathbf{w}_1 t + \mathbf{w}_2(1-t))$, where $t \in [0,1]$ then from Mean Value Theorem (MVT) we have that $g(1) - g(0) = g^{'}(c)$ for some c in between 0 and 1. This implies that $g(1) = \ell(\mathbf{w}_1)$, $g(0) = \ell(\mathbf{w}_2)$ and $g^{'}(c) = \big\langle \nabla\ell(\mathbf{w}_1 t + \mathbf{w}_2(1-t)), \mathbf{w}_1 - \mathbf{w}_2\big\rangle$. Thus from MVT we get:

$$\begin{aligned}
|\ell(\mathbf{w}_1) - \ell(\mathbf{w}_2)| &= |\big\langle \nabla\ell(\mathbf{w}_1 t + \mathbf{w}_2(1-t)), \mathbf{w}_1 - \mathbf{w}_2\big\rangle| \\
&\overset{(a)}{\leq} \|\nabla\ell(\mathbf{w}_1 t + \mathbf{w}_2(1-t))\|\|\mathbf{w}_1 - \mathbf{w}_2\| \\
&\overset{(b)}{\leq} (\max_{\mathbf{w}\in C}\|\nabla\ell(\mathbf{w})\|)\|\mathbf{w}_1 - \mathbf{w}_2\|
\end{aligned} \tag{17}$$

where (a) follows from the Cauchy-Schwarz inequality and (b) follows from the fact that $\mathbf{w}_1 t + \mathbf{w}_2(1-t) \in C$ and $\forall \mathbf{w} \in C$, $\|\nabla\ell(\mathbf{w})\| \leq \max_{\mathbf{w}\in C}\|\nabla\ell(\mathbf{w})\|$.

When $\ell(\mathbf{w})$ is quadratic, then we can always write $\ell(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T Q(\mathbf{w} - \mathbf{w}^*) + c_0$, where $Q = \nabla^2\ell(\mathbf{w})$ is a constant symmetric matrix and $c_0 = \ell(\mathbf{w}^*)$. From our definition of $\ell(\mathbf{w})$ we can write:

$$\begin{aligned}
\max_{\mathbf{w}\in C}\|\nabla\ell(\mathbf{w})\| &= \max_{\mathbf{w}\in C}\|Q(\mathbf{w} - \mathbf{w}^*)\| \\
&\overset{(a)}{\leq} \max_{\mathbf{w}\in C}\beta\|\mathbf{w} - \mathbf{w}^*\| \\
&\overset{(b)}{\leq} \beta(\max_{\mathbf{w}\in C}\|\mathbf{w}\| + \|\mathbf{w}^*\|) \\
&\leq \beta(R + \|\mathbf{w}^*\|)
\end{aligned}$$

where (a) follows from the definition of β and (b) follows from the triangle inequality. Substituting this result in Equation (17) we get:

$$|\ell(\mathbf{w}_1) - \ell(\mathbf{w}_2)| \leq \beta\big(R + \|\mathbf{w}^*\|\big)\|\mathbf{w}_1 - \mathbf{w}_2\|$$

□

Lemma 3. *Consider a function* $L_{\mathcal{D}}(\mathbf{w}) = \frac{1}{n}\sum_{i\in\mathcal{D}}\ell_i(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w}\|^2$, *where* $\ell_i(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^+$, $\ell_i(0) \leq M$ *and* $\mathcal{D}$ *is a dataset of size* n. *Let* $\mathbf{w}^*_{\mathcal{D}} = argmin_{\mathbf{w}\in\mathbb{R}^d} L_{\mathcal{D}}(\mathbf{w})$, *then* $\|\mathbf{w}^*_{\mathcal{D}}\| \leq \sqrt{\frac{2M}{\mu}}$.

Proof.

$$\frac{\mu}{2}\|\mathbf{w}^*_{\mathcal{D}}\|^2 \overset{(a)}{\leq} L_{\mathcal{D}}(\mathbf{w}^*_{\mathcal{D}}) \overset{(b)}{\leq} L_{\mathcal{D}}(0) \overset{(c)}{\leq} M$$

Thus, $\|\mathbf{w}^*_{\mathcal{D}}\| \leq \sqrt{\frac{2M}{\mu}}$, where (a) follows from the assumption that $\ell_i(\mathbf{w})$ is non-negative, (b) follows from the fact that $\mathbf{w}^*_{\mathcal{D}}$ is the minimizer of $L_{\mathcal{D}}(\mathbf{w})$ and (c) follows from the assumption that $\ell_i(0) \leq M$. Note that the result is independent of n, thus, the empirical risk minimizers of the datasets obtained by removing a subset of samples will also lie within a $d-$dimensional sphere of radius $\sqrt{\frac{2M}{\mu}}$.

□

Lemma 4. *Consider a function* $L_{\mathcal{D}}(\mathbf{w}) = \frac{1}{n}\sum_{i\in\mathcal{D}}\widehat{\ell}_i(\mathbf{w})$, *where* $\widehat{\ell}_i(\mathbf{w}) = \ell_i(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w}\|^2$, $\ell_i(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^+$ *is strongly convex function with* $\ell_i(0) \leq M < \infty$. *Let* $\mathcal{D}_r$ *(retain set) be the dataset remaining after removing* b *samples* $\mathcal{D}_f$ *(forget set) from* $\mathcal{D}$ *i.e.* $\mathcal{D}_r = \mathcal{D} - \mathcal{D}_f$. *Let* $\mathbf{w}^*_{\mathcal{D}} \triangleq argmin_{\mathbf{w}\in\mathbb{R}^d} L_{\mathcal{D}}(\mathbf{w})$ *and* $\mathbf{w}^*_{\mathcal{D}_r} \triangleq argmin_{\mathbf{w}\in\mathbb{R}^d} L_{\mathcal{D}_r}(\mathbf{w})$. *Let* $C = \{\mathbf{w}|\mathbf{w} \in \mathbb{R}^d$ *and* $\|\mathbf{w}\| \leq \sqrt{2M/\mu}\}$ *and* $G \triangleq max_{\mathbf{w}\in C}\|\nabla\ell(\mathbf{w})\| < \infty$. *Then we have that:*

$$\|\mathbf{w}^*_{\mathcal{D}} - \mathbf{w}^*_{\mathcal{D}_r}\| \leq \frac{2bG}{n\mu}$$

When $\ell_i(\mathbf{w})$ *is also quadratic with* $\beta = max_{i\in\mathcal{D}}\lambda_M(\nabla^2\hat{\ell}_i(\mathbf{w}))$, *the smoothness constant of* $L_{\mathcal{D}}(\mathbf{w})$, *we have that* $G = 2\beta\sqrt{2M/\mu}$.

Proof. We use the same technique as proposed in [35].

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{w}^*_{\mathcal{D}_r}) &= \frac{n-b}{n}L_{\mathcal{D}_r}(\mathbf{w}^*_{\mathcal{D}_r}) + \frac{b}{n}L_{\mathcal{D}_f}(\mathbf{w}^*_{\mathcal{D}_r}) \\ &\overset{(a)}{\leq} \frac{n-b}{n}L_{\mathcal{D}_r}(\mathbf{w}^*_{\mathcal{D}}) + \frac{b}{n}L_{\mathcal{D}_f}(\mathbf{w}^*_{\mathcal{D}_r}) \\ &= L_{\mathcal{D}}(\mathbf{w}^*_{\mathcal{D}}) + \frac{b}{n}L_{\mathcal{D}_f}(\mathbf{w}^*_{\mathcal{D}_r}) - \frac{b}{n}L_{\mathcal{D}_f}(\mathbf{w}^*_{\mathcal{D}}) \end{aligned} \tag{18}$$

where, (a) first inequality follows from the fact that $\mathbf{w}^*_{\mathcal{D}_r}$ is the minimizer of $L_{\mathcal{D}_r}(\mathbf{w})$.

From Lemma 3 we know that $\|\mathbf{w}^*_{\mathcal{D}}\|, \|\mathbf{w}^*_{\mathcal{D}_r}\|, \|\mathbf{w}^*_{\mathcal{D}_f}\| \leq \sqrt{\frac{2M}{\mu}}$. Also from the definition of β we have that $\lambda_M(\nabla^2 L_{\mathcal{D}_f}(\mathbf{w})) \leq \beta$. Then applying Lemma 2 with $R = \sqrt{\frac{2M}{\mu}}$ for $L_{\mathcal{D}_f}(\mathbf{w})$ we get that

$$|L_{\mathcal{D}_f}(\mathbf{w}^*_{\mathcal{D}}) - L_{\mathcal{D}_f}(\mathbf{w}^*_{\mathcal{D}_r})| \leq G\|\mathbf{w}^*_{\mathcal{D}} - \mathbf{w}^*_{\mathcal{D}_r}\| \tag{19}$$

From the the definition of $L_{\mathcal{D}_f}(\mathbf{w})$ we know that it is a $\mu-$strongly convex function. So we have the following property:

$$L_{\mathcal{D}_f}(\mathbf{w}^*_{\mathcal{D}_r}) \geq L_{\mathcal{D}_f}(\mathbf{w}^*_{\mathcal{D}}) + \frac{\mu}{2}\|\mathbf{w}^*_{\mathcal{D}} - \mathbf{w}^*_{\mathcal{D}_r}\|^2 \tag{20}$$

Substituting Equation (19) and Equation (20) in Equation (18) we get:

$$\begin{aligned} \frac{\mu}{2}\|\mathbf{w}^*_{\mathcal{D}} - \mathbf{w}^*_{\mathcal{D}_r}\|^2 &\leq bG\|\mathbf{w}^*_{\mathcal{D}} - \mathbf{w}^*_{\mathcal{D}_r}\| \\ \|\mathbf{w}^*_{\mathcal{D}} - \mathbf{w}^*_{\mathcal{D}_r}\| &\leq \frac{2bG}{n\mu} \end{aligned}$$

When $\ell_i(\mathbf{w})$ is also quadratic with $\beta = \max_{i\in\mathcal{D}}\lambda_{\text{Max}}(\nabla^2\hat{\ell}_i(\mathbf{w}))$, the smoothness constant, then from Lemma 2 we have $G = \beta\big(R + \|\mathbf{w}^*_{\mathcal{D}}\|\big) \leq \beta(2\sqrt{2M/\mu})$.

$$\|\mathbf{w}^*_{\mathcal{D}} - \mathbf{w}^*_{\mathcal{D}_r}\| \leq \frac{4b\beta\sqrt{2M}}{n\mu^{3/2}}$$

□

Lemma 5. *Let $\ell(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^+$ be a convex and $\beta-$ smooth with minimizer, $\mathbf{w}^* = \text{argmin}_{\mathbf{w}\in\mathbb{R}^d}\ell(\mathbf{w})$. Then we have that:*

1. $\ell(\mathbf{w}) - \ell(\mathbf{w}^*) \leq \frac{\beta}{2}\|\mathbf{w} - \mathbf{w}^*\|^2$
2. $\|\nabla\ell(\mathbf{w})\|^2 \leq 2\beta(\ell(\mathbf{w}) - \ell(\mathbf{w}^*))$

Proof. From the definition of $\beta-$smoothness we have that:

$$\ell(\mathbf{w}_1) \leq \ell(\mathbf{w}_2) + \big\langle\nabla\ell(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\big\rangle + \frac{\beta}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|^2 \tag{21}$$

Setting $\mathbf{w}_1 = \mathbf{w}$, $\mathbf{w}_2 = \mathbf{w}^*$ and using the fact that $\nabla\ell(\mathbf{w}^*) = 0$, we get that:

$$\ell(\mathbf{w}) - \ell(\mathbf{w}^*) \leq \frac{\beta}{2}\|\mathbf{w} - \mathbf{w}^*\|^2$$

For (2) we minimize Equation (21) with respect to $\mathbf{w}_1$:

$$\begin{aligned}
\min_{\mathbf{w}_1\in\mathbb{R}^d}\ell(\mathbf{w}_1) &\leq \min_{\mathbf{w}_1\in\mathbb{R}^d}\Big(\ell(\mathbf{w}_2) + \big\langle\nabla\ell(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\big\rangle + \frac{\beta}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|^2\Big) \\
&= \ell(\mathbf{w}_2) + \min_{\mathbf{w}_1\in\mathbb{R}^d}\Big(\big\langle\nabla\ell(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\big\rangle + \frac{\beta}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|^2\Big) \\
&\overset{(a)}{=} \ell(\mathbf{w}_2) - \frac{\|\nabla\ell(\mathbf{w}_2)\|^2}{2\beta} \qquad (22) \\
& \qquad (23)
\end{aligned}$$

where (a) follows from the result that $\mathbf{w}_1 = \mathbf{w}_2 - \dfrac{\nabla\ell(\mathbf{w}_2)}{\beta} = \text{argmin}_{\mathbf{w}_1\in\mathbb{R}^d}\big(\big\langle\nabla\ell(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\big\rangle + \frac{\beta}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|^2\big)$. Setting $\mathbf{w}_2 = \mathbf{w}$, $\min_{\mathbf{w}_1\in\mathbb{R}^d} = \ell(\mathbf{w}^*)$ in Equation (23) and re-arranging the terms we get (2). □

Theorem 2. *(SGD) Consider $L_{\mathcal{D}}(\mathbf{w}) = \dfrac{1}{n}\sum_{i\in\mathcal{D}}\ell_i(\mathbf{w})$, where $\ell_i(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^+$ is $\beta-$smooth and $\mu-$ strongly convex. Let $\mathbf{w}^* = \text{argmin}_{\mathbf{w}\in\mathbb{R}^d}L_{\mathcal{D}}(\mathbf{w})$ and $\mathbf{w}^*_i = \text{argmin}_{\mathbf{w}\in\mathbb{R}^d}\ell_i(\mathbf{w})$. Then we have the following result after t steps of SGD with batch-size B and constant learning rate $\eta = \mu/\beta^2$:*

$$\mathbb{E}\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq \Big(1 - \frac{\mu^2}{\beta^2}\Big)^t\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{2\sigma_\ell}{B\beta}$$

where $\sigma_\ell = \dfrac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*) - \dfrac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*_i)$

Proof. We do not use the bounded gradient assumption for the convergence of SGD and instead use the smoothness of our loss function [4]. This is because the training loss in our case is a quadratic function whose gradient is linear and not bounded.

Consider a mini-batch SGD update,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)$$

where the examples $\{i_{t+1}^{(1)}, i_{t+1}^{(2)}, \cdots i_{t+1}^{(B)}\}$ are sampled uniformly at random with replacement for all the iterations t.

Then by expanding $\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2$,

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 &= \left\|\mathbf{w}_t - \eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t) - \mathbf{w}^*\right\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}^*\|^2 - 2\left\langle \mathbf{w}_t - \mathbf{w}^*, \eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)\right\rangle + \left\|\eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)\right\|^2
\end{aligned}$$

Now taking expectation over the randomness of sampling we get that,

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \leq \mathbb{E}\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \underbrace{2\,\mathbb{E}\left\langle \mathbf{w}_t - \mathbf{w}^*, \eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)\right\rangle}_{(T_1)} \\
\underbrace{+\,\mathbb{E}\left\|\eta \cdot \frac{1}{m} \sum_{j=1}^{B} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)\right\|^2}_{(T_2)}
\end{aligned} \tag{24}$$

We will lower bound (T_1) and upper bound (T_2). Let $\xi_t = \{i_t^{(1)}, i_t^{(2)}, \cdots i_t^{(B)}\}$. Then for (T_1) we have,

$$\begin{aligned}
\mathbb{E}\left[\left\langle \mathbf{w}_t - \mathbf{w}^*, \eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)\right\rangle\right] &= \mathbb{E}_{\xi_1 \cdots \xi_t}\left[\mathbb{E}_{\xi_{t+1}}\left[\left\langle \mathbf{w}_t - \mathbf{w}^*, \eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)\right\rangle \middle| \xi_1 \cdots \xi_t\right]\right] \\
&= \mathbb{E}_{\xi_1 \cdots \xi_t}\left[\left\langle \mathbf{w}_t - \mathbf{w}^*, \eta \cdot \frac{1}{B} \sum_{j=1}^{B} \mathbb{E}_{\xi_{t+1}} \nabla \ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{w}_t - \mathbf{w}^*, \eta \cdot \frac{1}{B} \sum_{j=1}^{B} \nabla L_{\mathcal{D}}(\mathbf{w}_t)\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{w}_t - \mathbf{w}^*, \eta \nabla L_{\mathcal{D}}(\mathbf{w}_t)\right\rangle\right] \\
&\overset{(a)}{\geq} \mu\eta \cdot \mathbb{E}\|\mathbf{w}_t - \mathbf{w}^*\|^2
\end{aligned} \tag{25}$$

where (a) follows from the strong convexity of $L_{\mathcal{D}}$. Note that if $\ell_i(\mathbf{w})$ is $\mu-$strongly convex then even $L_{\mathcal{D}}$ is $\mu-$strongly convex.

Using the same conditioning argument as before and the fact that each sample in a batch is sampled i.i.d. for (T_2) we get that:

$$
\begin{aligned}
\mathbb{E}\Big\|\eta\cdot\frac{1}{B}\sum_{j=1}^{B}\nabla\ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)\Big\|^2 &= \eta^2\Big(\frac{1}{B}\mathbb{E}\big[\|\nabla\ell_{i_{t+1}^{(1)}}(\mathbf{w}_t)\|^2\big]+\frac{B-1}{B}\mathbb{E}\|\nabla L_{\mathcal{D}}(\mathbf{w}_t)\|^2\Big)\\
&\overset{(a)}{\leq}\eta^2\left(\frac{2\beta}{B}\Big(\mathbb{E}L(\mathbf{w}_t)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}_i^*)\Big)+\frac{2\beta(B-1)}{B}\Big(\mathbb{E}L_{\mathcal{D}}(\mathbf{w}_t)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*)\Big)\right)\\
&=\eta^2\left(\frac{2\beta}{B}\Big(\mathbb{E}L_{\mathcal{D}}(\mathbf{w}_t)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*)+\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}_i^*)\Big)\right.\\
&\left.+\frac{2\beta(B-1)}{B}\Big(\mathbb{E}L_{\mathcal{D}}(\mathbf{w}_t)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*)\Big)\right)\\
&=\eta^2\left(\frac{2\beta}{B}\Big(\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}_i^*)\Big)\right.\\
&\left.+2\beta\Big(\mathbb{E}L_{\mathcal{D}}(\mathbf{w}_t)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*)\Big)\right)\\
&\overset{(b)}{\leq}\eta^2\left(\frac{2\beta}{B}\Big(\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}_i^*)\Big)+\beta^2\mathbb{E}\|\mathbf{w}_t-\mathbf{w}^*\|^2\right)\\
&\leq\eta^2\cdot\beta^2\cdot\mathbb{E}\|\mathbf{w}_t-\mathbf{w}^*\|^2+\frac{2\beta\eta^2}{B}\underbrace{\Big(\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}^*)-\frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}_i^*)\Big)}_{\sigma_\ell}\\
&=\eta^2\cdot\beta^2\cdot\mathbb{E}\|\mathbf{w}_t-\mathbf{w}^*\|^2+\frac{2\beta\eta^2\sigma_\ell}{B}
\end{aligned}
\tag{26}
$$

where (a) follows from applying Lemma 5 (1) to $\ell_{i_{t+1}^{(j)}}(\mathbf{w}_t)$ and $L_{\mathcal{D}}(\mathbf{w}_t)$ and (b) follows from applying Lemma 5 (2).

Now substituting Equation (25) and Equation (26) in Equation (24), we get that,

$$
\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\leq(1-2\eta\mu+\eta^2\beta^2)\mathbb{E}\|\mathbf{w}_t-\mathbf{w}^*\|^2+\frac{2\beta\eta^2\sigma_\ell}{B}
\tag{27}
$$

Minimizing the coefficient with respect to η we get $\eta^*=\frac{\mu}{\beta^2}$, which gives the following update equation:

$$
\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\leq\Big(1-\frac{\mu^2}{\beta^2}\Big)\mathbb{E}\|\mathbf{w}_t-\mathbf{w}^*\|^2+\frac{2\mu^2\sigma_\ell}{B\beta^3}
\tag{28}
$$

Now applying this update recursively we get:

$$
\mathbb{E}\|\mathbf{w}_t-\mathbf{w}^*\|^2\leq\Big(1-\frac{\mu^2}{\beta^2}\Big)^t\mathbb{E}\|\mathbf{w}_0-\mathbf{w}^*\|^2+\frac{2\sigma_\ell}{B\beta}
$$

□

Definition 1. (Forgetting: Single Request) Consider an ERM problem with $L_{\mathcal{D}}(\mathbf{w})=\frac{1}{n}\sum_{i\in\mathcal{D}}\hat{\ell}_i(\mathbf{w})$, where $\hat{\ell}_i=\ell_i(\mathbf{w})+\frac{\mu}{2}\|\mathbf{w}\|^2$, $\ell_i(\mathbf{w}):\mathbb{R}^d\to\mathbb{R}^+$ is a quadratic convex function and $\mathcal{D}$ is a dataset of size n. Let $\mathbf{w}_{\mathcal{D}}=\mathcal{A}_{\mathcal{T}}(L_{\mathcal{D}}(\mathbf{w}))$ be the weights obtained after $\mathcal{T}$ steps of algorithm $\mathcal{A}$ (SGD in our case) on $L_{\mathcal{D}}(\mathbf{w})$. Then given a request to forget a set $\mathcal{D}_f\subset\mathcal{D}$ we apply the following scrubbing procedure:

$$S(\mathbf{w}_{\mathcal{D}}, \mathcal{D}, \mathcal{D}_f) \triangleq \mathbf{w}_{\mathcal{D}} - \Delta \mathbf{w}_{\mathcal{D},\mathcal{D}_f} + z \tag{29}$$

where $\Delta \mathbf{w}_{\mathcal{D},\mathcal{D}_f} = \mathcal{A}_\tau(\tilde{L}_{\mathcal{D}-\mathcal{D}_f}(\mathbf{w}))$, τ is the number of steps of $\mathcal{A}$ to minimize $\tilde{L}_{\mathcal{D}_r}(\mathbf{w})$ (where $\mathcal{D}_r = \mathcal{D} - \mathcal{D}_f$). Here

$$\tilde{L}_{\mathcal{D}_r}(\mathbf{w}) = 0.5 \cdot \mathbf{w}^T H_{\mathcal{D}_r} \mathbf{w} - \mathbf{w}^T g_{\mathcal{D}_r}(\mathbf{w}_{\mathcal{D}}) \tag{30}$$

$z \sim \mathcal{N}(0, \sigma I)$ and $H_{\mathcal{D}_r}$ is the hessian on the remaining data ($\mathcal{D}_r$). We compute the *residual gradient*, $g_{\mathcal{D}_r}(\mathbf{w}_{\mathcal{D}}) = \nabla L_{\mathcal{D}_r}(\mathbf{w}_{\mathcal{D}})$ once over entire $\mathcal{D}_r$, while $\mathbf{w}^T H_{\mathcal{D}_r}(\mathbf{w}_{\mathcal{D}})\mathbf{w}$ is compute stochastically in $\mathcal{A}$.

Definition 2. (Forgetting: Multiple Requests) Consider that we are provided with a sequence of forgetting request $(\mathcal{D}_f^j)$. Let $\mathcal{D}_j \subset \mathcal{D}$ be the dataset remaining, $\mathbf{w}_{\mathcal{D}_j}$ (or simply $\mathbf{w}_j$) be the weights obtained after j forgetting requests. Then given the $j+1^{\text{th}}$ request to forget $\mathcal{D}_f^{j+1} \subset \mathcal{D}_j$, from Equation (29) in Definition 1 we have that:

$$\mathbf{w}_{j+1} \triangleq S(\mathbf{w}_j, \mathcal{D}_j, \mathcal{D}_f^{j+1}) = \mathbf{w}_j - \Delta \mathbf{w}_{\mathcal{D}_j,\mathcal{D}_f^{j+1}} + z \tag{31}$$

where $z \sim \mathcal{N}(0, \sigma I)$ and $\mathbf{w}_0 = \mathbf{w}_{\mathcal{D}}$ are the weights obtained after training on the entire data $\mathcal{D}$.

Theorem 3. *(Formal) Consider an empirical risk, $L_{\mathcal{D}}(\mathbf{w}) = \frac{1}{n}\sum_{i\in\mathcal{D}} \widehat{\ell}_i(\mathbf{w})$, where $\widehat{\ell}_i(\mathbf{w}) = \ell_i(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w}\|^2$, $\ell_i(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^+$ is quadratic convex function with symmetric $\nabla^2\ell(\mathbf{w})$, $\beta = max_{i\in\mathcal{D}}\lambda_M(\nabla^2\hat{\ell}_i(\mathbf{w}))$ is the smoothness constant of $L_{\mathcal{D}}(\mathbf{w})$, $\ell_i(0) \le M < \infty$ and $\mathcal{D}$ is a dataset of size n. Let $\mathcal{A}$ be SGD with mini-batch size B, $\sigma_\ell > 0$ be some constant associated with SGD, $\gamma = 1 - \mu^2/\beta^2$, τ be the number of steps of $\mathcal{A}$ performed while forgetting and $a \in (0, 1/\gamma^\tau - 1)$. From Definition 2, let $\mathbf{w}_{j-1}, \mathcal{D}_{j-1}$ be the scrubbed weights and the dataset remaining after $j-1$ removal requests, then given a forgetting request to remove $\mathcal{D}_f^j$ ($|\mathcal{D}_f^j| = b$) we obtain the following bound on the amount of information remaining in the weights after using the scrubbing procedure in Definition 2:*

$$I(\cup_{k=1}^j \mathcal{D}_f^k, S(\mathbf{w}_{j-1}, \mathcal{D}_{j-1}, \mathcal{D}_f^j)) \le \frac{2\gamma^\tau(2 + 1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}\sqrt{\sigma}}\right)^2 + d\right] + \frac{8\sigma_\ell}{B\beta\sigma}}{1 - (1+\alpha)\gamma^\tau}$$

Proof. We follow similar proof technique to [35]. Consider a dataset $\mathcal{D}$ of size n and a forgetting sequence $\mathcal{D}_{Forget} = (\mathcal{D}_f^j)_j$ (batches of data of size b that we want to forget). Let $\mathbf{w}_j, \mathbf{w}_j^{'}, \hat{\mathbf{w}}_j, \mathcal{D}_j$ be the scrubbed weights with noise, scrubbed weights without noise, weights obtained by re-training from scratch using SGD and the remaining dataset after j requests of forgetting. Let n_j be the size of $\mathcal{D}_j$ and $\mathbf{w}_j^* = \text{argmin}_{\mathbf{w}\in\mathbb{R}^d} L_{\mathcal{D}_j}(\mathbf{w})$. Then from Definition 2 we have that

$$\mathbf{w}_j = \mathbf{w}_j^{'} + z \tag{32}$$

$$\mathbf{w}_j^{'} = \mathbf{w}_{j-1} - \Delta \mathbf{w}_{\mathcal{D}_{j-1},\mathcal{D}_f^j} \tag{33}$$

where $z \sim \mathcal{N}(0, \sigma^2 I)$, $\Delta \mathbf{w}_{\mathcal{D}_{j-1},\mathcal{D}_f^j} = \mathcal{A}_\tau(\tilde{L}_{\mathcal{D}_{j-1}-\mathcal{D}_f^j}(\mathbf{w}))$, $\mathcal{A}_\tau$ is τ steps of SGD and $\tilde{L}_{\mathcal{D}_{j-1}-\mathcal{D}_f^j}(\mathbf{w}) = \frac{1}{2|\mathcal{D}_{j-1} - \mathcal{D}_f^j|}\sum_{i\in\mathcal{D}_{j-1}-\mathcal{D}_f^j} \mathbf{w}^T\nabla^2\ell_i(\mathbf{w}_{\mathcal{D}})\mathbf{w} - \langle \nabla L_{\mathcal{D}_{j-1}-\mathcal{D}_f^j}(\mathbf{w}_{\mathcal{D}}), \mathbf{w}\rangle$. Note that $\hat{\mathbf{w}}_j$ are weights obtained at the end of training with SGD while $\mathbf{w}_j^*$ is the true empirical risk minimizer of $L_{\mathcal{D}_j}(\mathbf{w})$.

After re-training from scratch on $\mathcal{D}_j$ for $\mathcal{T}_j$ steps using SGD with mini-batch size of B, we have the following relation for $j \ge 0$, using Theorem 2:

$$\mathbb{E}\|\hat{\mathbf{w}}_j - \mathbf{w}_j^*\|^2 \le \gamma^{\mathcal{T}_j}\|\hat{\mathbf{w}}_{\text{init}} - \mathbf{w}_j^*\|^2 + \frac{2\sigma_\ell}{B\beta} \tag{34}$$

where $\gamma = 1 - \mu^2/\beta^2$ and $\hat{\mathbf{w}}_{\text{init}} = 0$ is the training initialization, $\hat{\mathbf{w}}_0$ are the weights obtained by training on $\mathcal{D}_0 = \mathcal{D}$, which is the complete dataset before receiving any forgetting request. When training the linearized model the user weights are initialized to 0 since they correspond to the first order perturbation of the non-linear weights.

Let us select $\mathcal{T}_j \geq \tau + \frac{2\log(n_j\mu/4b\beta)}{\log 1/\gamma}$, where $n_j \geq \frac{n}{2}$. Here τ is the number of steps of SGD used to remove one batch (b samples) of data during forgetting. Substituting τ_j in Equation (34) we get:

$$\mathbb{E}\|\hat{\mathbf{w}}_j - \mathbf{w}_j^*\|^2 \leq \gamma^\tau \left(\frac{4b\beta}{n_j\mu}\right)^2 \|\hat{\mathbf{w}}_{\text{init}} - \mathbf{w}_j^*\|^2 + \frac{2\sigma_\ell}{B\beta} \overset{(a)}{\leq} \gamma^\tau \left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + \frac{2\sigma_\ell}{B\beta} \tag{35}$$

where (a) follows from $\|\hat{\mathbf{w}}_{\text{init}} - \mathbf{w}_j^*\| \leq \sqrt{\frac{2M}{\mu}}$ and $1/n_j \leq 2/n$.

Now we will compute a similar bound for the weights obtained by applying the forgetting procedure. For any $j \geq 1$, using the scrubbing procedure in Definition 2 we have the following relation:

$$\mathbb{E}\|\mathbf{w}_j^{'} - \mathbf{w}_j^*\|^2 \leq \frac{\gamma^\tau(1+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{2\sigma_\ell}{B\beta}}{1-(1+\alpha)\gamma^\tau} \tag{36}$$

for $0 < \alpha < 1/\gamma^\tau - 1$. Note that $\mathbf{w}_j$ are the weights after applying the newton update but before adding the scrubbing noise.

From the scrubbing procedure described in eq. (32) and eq. (33) to solve for $\Delta\mathbf{w}_{\mathcal{D}_{j-1},\mathcal{D}_f^j}$ we minimize $\hat{L}_{\mathcal{D}_{j-1}-\mathcal{D}_f^j}$ using SGD. While the optimal value of $\Delta\mathbf{w}^*_{\mathcal{D}_{j-1},\mathcal{D}_f^j} = \mathbf{w}_{j-1} - \mathbf{w}_j^*$, thus, $\Delta\mathbf{w}_{\mathcal{D}_{j-1},\mathcal{D}_f^j} - \Delta\mathbf{w}^*_{\mathcal{D}_{j-1},\mathcal{D}_f^j} = \Delta\mathbf{w}_{\mathcal{D}_{j-1},\mathcal{D}_f^j} - \mathbf{w}_{j-1} + \mathbf{w}_j^* = \mathbf{w}_j^* - \mathbf{w}_j^{'}$. We can bound $\mathbb{E}\|\Delta\mathbf{w}_{\mathcal{D}_{j-1},\mathcal{D}_f^j} - \Delta\mathbf{w}^*_{\mathcal{D}_{j-1},\mathcal{D}_f^j}\|^2$ using Theorem 2 and thus also bound $\mathbb{E}\|\mathbf{w}_j^* - \mathbf{w}_j^{'}\|$.

More precisely we have:

$$\begin{aligned}
\mathbb{E}\|\Delta\mathbf{w}_{\mathcal{D}_{j-1},\mathcal{D}_f^j} - \Delta\mathbf{w}^*_{\mathcal{D}_{j-1},\mathcal{D}_f^j}\|^2 &\overset{(a)}{=} \mathbb{E}\|\mathbf{w}_j^{'} - \mathbf{w}_j^*\|^2 \\
\mathbb{E}\|\mathbf{w}_j^{'} - \mathbf{w}_j^*\|^2 &\overset{(b)}{\leq} \gamma^\tau \mathbb{E}\|\Delta\mathbf{w}^{(0)}_{\mathcal{D}_{j-1},\mathcal{D}_f^j} - \Delta\mathbf{w}^*_{\mathcal{D}_{j-1},\mathcal{D}_f^j}\|^2 + \frac{2\sigma_\ell}{B\beta} \\
\mathbb{E}\|\mathbf{w}_j^{'} - \mathbf{w}_j^*\|^2 &\overset{(c)}{=} \gamma^\tau \mathbb{E}\|\mathbf{w}_{j-1} - \mathbf{w}_j^*\| + \frac{2\sigma_\ell}{B\beta}
\end{aligned} \tag{37}$$

where the (a) follows from the definition of the scrubbing update as shown above, (b) follows from Theorem 2 and (c) follows from $\Delta\mathbf{w}^{(0)}_{\mathcal{D}_{j-1},\mathcal{D}_{f_j}} = 0$. Note that both while training ($\mathcal{T}_j$ iterations) and forgetting (τ iterations) we use SGD with a constant step-size. We will use eq. (37) along with induction to prove eq. (39). For $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 I)$, we have:

$$\mathbb{E}\|\mathbf{z}\|^2 = d\sigma^2 \tag{38}$$

To prove eq. (39) we use induction. Lets consider the base case $j = 1$.

$$
\begin{aligned}
\mathbb{E}\|\mathbf{w}_1^{'} - \mathbf{w}_1^*\|^2 &\overset{(1)}{\leq} \gamma^\tau \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}_1^*\|^2 + \frac{2\sigma_\ell^2}{B\beta} \\
&\overset{(2)}{=} \gamma^\tau \mathbb{E}\|\hat{\mathbf{w}}_0 - \mathbf{w}_0^* + \mathbf{w}_0^* - \mathbf{w}_1^*\|^2 + \frac{2\sigma_\ell}{B\beta} \\
&\overset{(3)}{\leq} \gamma^\tau (1+\alpha)\mathbb{E}\|\hat{\mathbf{w}}_0 - \mathbf{w}_0^*\|^2 + \gamma^\tau(1+1/\alpha)\mathbb{E}\|\mathbf{w}_0^* - \mathbf{w}_1^*\|^2 + \frac{2\sigma_\ell}{B\beta} \\
&\overset{(4)}{\leq} \gamma^\tau (1+\alpha)\mathbb{E}\|\hat{\mathbf{w}}_0 - \mathbf{w}_0^*\|^2 + \underbrace{\gamma^\tau(1+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{2\sigma_\ell}{B\beta}}_{(A)} \\
&\overset{(5)}{\leq} \gamma^\tau(1+\alpha)\underbrace{\left[\gamma^\tau\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + \frac{2\sigma_\ell}{B\beta}\right]}_{\leq (A)} + (A) \\
&\leq \gamma^\tau(1+\alpha)(A) + (A) \\
&\leq \frac{\gamma^\tau(1+\alpha)}{1-(1+\alpha)\gamma^\tau}(A) + (A) \\
&= \frac{\gamma^\tau(1+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{2\sigma_\ell}{B\beta}}{1-(1+\alpha)\gamma^\tau}
\end{aligned}
$$

where (1) follows from eq. (37), (2) follows from the Definition 2, (3) follows from Lemma 1, (4) follows from Lemma 4 and eq. (38), (5) follows from eq. (35). Note that in the base case $\mathbf{w}_0 = \hat{\mathbf{w}}_0$ which are the weights obtained by training on the complete data. For Lemma 4, $\mathcal{D}_j$ and $\mathcal{D}_{j-1}$ differ in b samples and $n_j \geq n/2$. Also note that the expectation above is with respect to the randomness of SGD.

Now that we have the base case, for any general $j > 1$ we get:

$$
\begin{aligned}
\mathbb{E}\|\mathbf{w}_j' - \mathbf{w}_j^*\|^2 &\overset{(1)}{\leq} \gamma^\tau \mathbb{E}\|\mathbf{w}_{j-1} - \mathbf{w}_j^*\|^2 + \frac{2\sigma_\ell}{B\beta} \\
&\overset{(2)}{=} \gamma^\tau \mathbb{E}\|\mathbf{w}_{j-1}^{'} - \mathbf{w}_j^* + \mathbf{z}\|^2 + \frac{2\sigma_\ell}{B\beta} \\
&= \gamma^\tau \mathbb{E}\|\mathbf{w}_{j-1}^{'} - \mathbf{w}_{j-1}^* + \mathbf{w}_{j-1}^* - \mathbf{w}_j^* + \mathbf{z}\| + \frac{2\sigma_\ell}{B\beta} \\
&\overset{(3)}{\leq} \gamma^\tau (1+\alpha)\mathbb{E}\|\mathbf{w}_{j-1}^{'} - \mathbf{w}_{j-1}^*\|^2 + \gamma^\tau(1+1/\alpha)\mathbb{E}\|\mathbf{w}_{j-1}^* - \mathbf{w}_j^* + \mathbf{z}\|^2 + \frac{2\sigma_\ell}{B\beta} \\
&\overset{(4)}{\leq} \gamma^\tau(1+\alpha)\underbrace{\mathbb{E}\|\mathbf{w}_{j-1}^{'} - \mathbf{w}_{j-1}^*\|^2}_{\leq \frac{(A)}{1-\gamma^\tau(1+\alpha)}} + \underbrace{\gamma^\tau(1+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{2\sigma_\ell}{B\beta}}_{(A)} \\
&\leq \frac{\gamma^\tau(1+\alpha)}{1-\gamma^\tau(1+\alpha)}(A) + (A) \\
&= \frac{\gamma^\tau(1+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{2\sigma_\ell}{B\beta}}{1-(1+\alpha)\gamma^\tau}
\end{aligned}
$$

where (1) follows from eq. (37), (2) follows from the eq. (32), (3) follows from Lemma 1, (4) follows from induction update, Lemma 4 and eq. (38). For Lemma 4, $\mathcal{D}_i$ and $\mathcal{D}_{i-1}$ differ in b samples. The expectation above is with respect to the randomness of SGD and the scrubbing noise $\mathbf{z}$.

Thus we have:

$$\mathbb{E}\|\mathbf{w}'_j - \mathbf{w}^*_j\|^2 \leq \frac{\gamma^\tau(1+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{2\sigma_\ell}{B\beta}}{1-(1+\alpha)\gamma^\tau} \tag{39}$$

Combining eq. (35) and eq. (39) using Lemma 1 we get:

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}_j^{'} - \hat{\mathbf{w}}_j\|^2 &= \mathbb{E}\|\mathbf{w}_j^{'} - \mathbf{w}^*_j + \mathbf{w}^*_j - \hat{\mathbf{w}}_j\|_2^2 \\
&\overset{(1)}{\leq} 2\mathbb{E}\|\mathbf{w}_j^{'} - \mathbf{w}^*_j\|^2 + 2\mathbb{E}\|\mathbf{w}^*_j - \hat{\mathbf{w}}_j\|^2 \\
&\leq \frac{2\gamma^\tau(1+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{4\sigma_\ell}{B\beta}}{1-(1+\alpha)\gamma^\tau} + 2\gamma^T\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + \frac{4\sigma_\ell}{m\beta} \\
&\leq \frac{2\gamma^\tau(2+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{8\sigma_\ell}{B\beta}}{1-(1+\alpha)\gamma^\tau}
\end{aligned}$$

where (1) follows from Lemma 1.

Thus, we get:

$$\mathbb{E}\|\mathbf{w}_j^{'} - \hat{\mathbf{w}}_j\|^2 \leq \frac{2\gamma^\tau(2+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}}\right)^2 + d\sigma^2\right] + \frac{8\sigma_\ell}{B\beta}}{1-(1+\alpha)\gamma^\tau} \tag{40}$$

Since the problem is quadratic, the hessian will be same at all points. From Proposition 1 in [16], we have:

$$\begin{aligned}
I(\cup_{k=1}^j \mathcal{D}_f^k, S(\mathbf{w}_{j-1}, \mathcal{D}_{j-1}, \mathcal{D}_f^j)) &\leq \mathbb{E}\big[(\mathbf{w}'_j - \hat{\mathbf{w}}_j)^T(\sigma^2 I)^{-1}(\mathbf{w}'_j - \hat{\mathbf{w}}_j)\big] \\
&\leq \frac{\mathbb{E}\|\mathbf{w}'_i - \hat{\mathbf{w}}_i\|_2^2}{\sigma^2}
\end{aligned}$$

Note that the expectation in the previous expression is with respect to the randomness in the training algorithm and the forgetting algorithm, plus the set $\mathcal{D}_f^k$. Then using eq. (40) with the previous equation we obtain that:

$$I(\cup_{k=1}^j \mathcal{D}_f^k, S(\mathbf{w}_{j-1}, \mathcal{D}_{j-1}, \mathcal{D}_f^j)) \leq \frac{2\gamma^\tau(2+1/\alpha)\left[\left(\frac{8b\beta\sqrt{2M}}{n\mu^{3/2}\sigma}\right)^2 + d\right] + \frac{8\sigma_\ell}{B\beta\sigma^2}}{1-(1+\alpha)\gamma^\tau} \tag{41}$$

□