A Proof of (1)

Lemma 2 (Optimization comparison lemma [35]). Suppose
\[ x^* \in \arg\min_x \varphi_1(x) + \varphi_0(x) \quad \text{and} \quad y^* \in \arg\min_x \varphi_2(x) + \varphi_0(x). \] (10)

for \( \varphi_1 \) and \( \varphi_2 \) differentiable and \( \varphi_0 \) convex.

Proof. The (sub)differentiability assumptions and the optimality of \( x_{\varphi_1} \) and \( x_{\varphi_2} \) imply that \( 0 \in \partial \varphi_2 \) and \( u = 0 + \nabla (\varphi_1 - \varphi_2)(x_{\varphi_1}) \) for some \( u \in \partial \varphi_2 \). The gradient growth condition implies
\[ \nu_{\varphi_2}(\| x_{\varphi_1} - x_{\varphi_2} \|_2) \leq \langle x_{\varphi_1} - x_{\varphi_2}, u - 0 \rangle = \langle x_{\varphi_1} - x_{\varphi_2}, \nabla (\varphi_2 - \varphi_1)(x_{\varphi_1}) \rangle. \] (11)

Lemma 3 (Learning guarantee for \( \hat{\theta}_n(\lambda) \)). Given, \( F_n \) satisfies Assumption 1 or 2 and any distribution \( \mathcal{D}, \) let \( S = \{ z_i \}_{i=1}^n \) where \( z_i \sim \mathcal{D}^n \). Then the empirical minimizer \( \theta_n(\lambda) \) of \( F_n(\theta, \lambda, z) \) satisfies
\[ E[ F(\hat{\theta}_n(\lambda)) - F(\theta^*(\lambda)) ] \leq \frac{4L^2}{\mu n}. \]

Proof. Given \( F_n \) is \( \mu \)-strongly convex this follows from Claim 6.2 in [29].

A.1 Proof of (6b): Closeness of \( \hat{\theta}_n(\lambda) \) and \( \hat{\theta}_{n,U}(\lambda) \)

Suppose we have deleted \( m \) users in a set \( U \). Define \( 
\hat{F}_{n,U} = \frac{n-m}{n} F_{n,U} \) where \( F_{n,U} = \frac{1}{n-m} \sum_{i \in U} f(z_i, \theta, \lambda) \) and note that \( \hat{F}_{n,U} \) and \( F_{n,U} \) have the same minimizers. We will work with \( 
\hat{F}_{n,U} \). By the optimizer comparison lemma 2 and strong convexity of \( F_n \)
\[ \mu \| \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2^2 \leq \langle \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda), \nabla F_n(z, \hat{\theta}_n(\lambda), \lambda) - \nabla \hat{F}_{n,U}(z, \hat{\theta}_n(\lambda), \lambda) \rangle 
\leq \frac{1}{n} \sum_{i \in U} \| \nabla \ell(z_i, \hat{\theta}_n(\lambda)) \|_2 \| \nabla \ell(z_i, \hat{\theta}_n(\lambda)) \|_2 
\leq \frac{1}{n} \| \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2 \cdot mL \]

Dividing both sides by \( \| \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2 \) and rearranging gives the desired bound of
\[ \| \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2 \leq \frac{mL}{\mu n}. \]

A.1.1 Proof of (6b): Closeness of \( \hat{\theta}_{n,U}(\lambda) \) and \( \hat{\theta}_{n,U}(\lambda) \)

We define:
• \( \psi_1 = \hat{F}_{n,U}(z, \theta, \lambda) \)
• \( \psi_2 = \langle \nabla \hat{F}_{n,U}(\hat{\theta}_n(\lambda)), \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda) \rangle + \langle \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda), \nabla^2 \hat{F}_{n,U}(\hat{\theta}_n(\lambda)) [\hat{\theta}_n(\lambda) - \theta] \rangle 
• \( \psi_3 = \langle \nabla \hat{F}_{n,U}(\hat{\theta}_n(\lambda)), \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda) \rangle + \langle \hat{\theta}_n(\lambda) - \hat{\theta}_{n,U}(\lambda), \nabla^2 \hat{F}_{n,U}(\hat{\theta}_n(\lambda)) [\hat{\theta}_n(\lambda) - \theta] \rangle 
• \( \hat{\theta}_{n,U}(\lambda) = \arg\min \psi_1(\theta), \)
• \( \hat{\theta}_{n,U}(\lambda) = \arg\min \psi_3(\theta) \)

The optimizer comparison theorem and strong convexity of \( F_n \) implies the following upper bound:
\[ \frac{\mu}{2} \| \hat{\theta}_{n,U}(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2^2 \leq \langle \hat{\theta}_{n,U}(\lambda) - \hat{\theta}_{n,U}(\lambda), \nabla (\psi_3 - \psi_1) (\hat{\theta}_{n,U}(\lambda)) \rangle 
\leq \| \hat{\theta}_{n,U}(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2 \| \nabla (\psi_3 - \psi_1) (\hat{\theta}_{n,U}(\lambda)) \|_2 \]
We show that the batch and streaming version of the algorithms are equivalent. This inductive argument shows both batch and streaming algorithms are the same.

\[
\frac{2}{\mu} \| \hat{\theta}_{n,U}(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2 \leq \frac{2m^2CL}{\mu^2n^2} + \frac{Mn^2L^2}{\mu^4n^2}
\]

Inequality \((\text{i})\) follows from smoothness of the objective function. Dividing both sides by \(\frac{2}{\mu}\), gives the desired bound of

\[
\| \hat{\theta}_{n,U}(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2 \leq \frac{2m^2CL}{\mu^2n^2} + \frac{Mn^2L^2}{\mu^4n^2}
\]

For the non-smooth version of our algorithm, the same proof holds where we define

- \(\psi_1 = \ell_{n,U}(z, \theta, \lambda)\)
- \(\psi_2 = \langle \nabla \ell_{n,U}(\hat{\theta}_n(\lambda)), \hat{\theta}_n(\lambda) - \theta \rangle + \langle \hat{\theta}_n(\lambda) - \theta, \nabla^2 \ell_{n,U}(\hat{\theta}_n(\lambda)) [\hat{\theta}_n(\lambda) - \theta] \rangle + \pi(\theta)\)
- \(\psi_3 = \langle \nabla \ell_{n,U}(\hat{\theta}_n(\lambda)), \hat{\theta}_n(\lambda) - \theta \rangle + \langle \hat{\theta}_n(\lambda) - \theta, \nabla^2 \ell_{n,U}(\hat{\theta}_n(\lambda)) [\hat{\theta}_n(\lambda) - \theta] \rangle + \pi(\theta)\)
- \(\hat{\theta}_{n,U}(\lambda) = \text{argmin} \psi_1(\theta)\)
- \(\hat{\theta}_{n,U}(\lambda) = \text{argmin} \psi_3(\theta)\)

\[
\frac{2}{\mu} \| \hat{\theta}_{n,U}(\lambda) - \hat{\theta}_{n,U}(\lambda) \|_2 \leq \frac{2m^2CL}{\mu^2n^2} + \frac{Mn^2L^2}{\mu^4n^2}
\]

A.2 Comparisons between batch and streaming algorithm

We show that the batch and streaming version of the algorithms are equivalent.

Case 1: \(\pi\) is smooth. The bounds we have proved are for the minimizer of \(\varphi_3\), namely

\[
\hat{\theta}_{n,U}(\lambda) = \hat{\theta}_n(\lambda) - \nabla^2 \ell_{n,U}(\hat{\theta}_n(\lambda))^{-1} \nabla \ell_{n,U}(\hat{\theta}_n(\lambda))
\]

Now suppose 1 datapoint (user \(j\)) requests to be deleted. Then the streaming and batch algorithms agree, as the update becomes

\[
\hat{\theta}_{n,-i}(\lambda) = \hat{\theta}_n(\lambda) + \frac{1}{n} \left( \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(z_i, \hat{\theta}_n(\lambda), \lambda) \right) - \frac{1}{n} \sum_{i \in U} \nabla \ell(z_i, \hat{\theta}_n(\lambda))
\]

Now suppose the algorithms are consistent for all deletion requests in the set \(U\). When an additional user \(j\) requests to delete their data the streaming algorithm returns

\[
\hat{\theta}_{n,(-U\cup\{j\})}(\lambda) = \hat{\theta}_{n,U}(\lambda) + \frac{1}{n} \left( \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(z_i, \hat{\theta}_n(\lambda), \lambda) \right) - \frac{1}{n} \sum_{i \in U} \nabla \ell(z_i, \hat{\theta}_n(\lambda))
\]

which matches the batch version of the deletion algorithm. This inductive argument shows both batch and streaming algorithms are the same.
Case 2: $\pi$ is not smooth. When $\pi$ is not smooth, the minimizer of $\varphi_3$ satisfies
$$\tilde{\theta}_{n,-(U \cup \{j\})}(\lambda) = \tilde{\theta}_{n,-U}(\lambda) + \frac{1}{n} \left( \sum_{i=1}^{n} \nabla^2 F(z_i, \tilde{\theta}_n(\lambda), \lambda) \right)^{-1} \nabla \ell(z_j, \tilde{\theta}_n(\lambda)) + \lambda \nabla \pi(\tilde{\theta}_{n,-(U \cup \{j\})}(\lambda))$$

When 1 datapoint (user $j$) requests to be deleted, the streaming and batch algorithms agree given $U = \emptyset$. Now suppose the algorithms are consistent for all deletion requests in the set $U$. When an additional user $j$ requests to delete their data the streaming algorithm returns an estimator that satisfies
$$\tilde{\theta}_{n,-(U \cup \{j\})}(\lambda) = \tilde{\theta}_{n,-U}(\lambda) + \frac{1}{n} H^{-1}_\ell \nabla \ell(z_j, \tilde{\theta}_n(\lambda)) + \lambda H^{-1}_\ell \nabla \pi(\tilde{\theta}_{n,-(U \cup \{j\})}(\lambda))$$

which matches the batch version of the deletion algorithm. This inductive arguments show both batch and streaming algorithms are the same.

A.3 Proof of excess empirical risk

Second, we prove the excess empirical risk of our unlearning algorithm (1).

Proof.

$$\mathbb{E}[F_n(\tilde{\theta}_{n,-U}(\lambda)) - F_n(\theta^*(\lambda))] = \mathbb{E}[F_n(\tilde{\theta}_{n,-U}(\lambda)) - F_n(\hat{\theta}_n(\lambda)) + F_n(\hat{\theta}_n(\lambda)) - F_n(\theta^*(\lambda))]$$

$$\leq \mathbb{E}[F_n(\tilde{\theta}_{n,-U}(\lambda)) - F_n(\hat{\theta}_n(\lambda))] + \mathbb{E}[F_n(\hat{\theta}_n(\lambda)) - F_n(\theta^*(\lambda))]$$

where $\(1\)$ comes from Lemma 3 given that $F_n$ satisfies Assumption 1 or 2.

Next we upper bound $\mathbb{E}[\|\tilde{\theta}_{n,-U}(\lambda) - \hat{\theta}_n(\lambda)\|]$:

$$\mathbb{E}[\|\tilde{\theta}_{n,-U}(\lambda) - \hat{\theta}_n(\lambda)\|] = \mathbb{E}[\|\tilde{\theta}_{n,-U}(\lambda) - \hat{\theta}_{n,-U}(\lambda) + \hat{\theta}_{n,-U}(\lambda) - \hat{\theta}_n(\lambda)\|]$$

$$\leq \mathbb{E}[\|\tilde{\theta}_{n,-U}(\lambda) - \hat{\theta}_{n,-U}(\lambda)\|] + \mathbb{E}[\|\hat{\theta}_{n,-U}(\lambda) - \hat{\theta}_n(\lambda)\|]$$

$$\leq \mathbb{E}[\|\tilde{\theta}_{n,-U}(\lambda) - \hat{\theta}_{n,-U}(\lambda)\|] + \frac{\mu L}{\mu n}$$

$$\leq \mathbb{E}[\|\tilde{\theta}_{n,-U}(\lambda) - \hat{\theta}_{n,-U}(\lambda)\|] + \sigma ||| + \frac{\mu L}{\mu n}$$

where $\(2\)$ comes from Lemma 1 and $\(3\)$ comes from Jensen’s inequality and Lemma 1 (Equation 6b).

Now we substitute this back into our earlier bound:

$$\mathbb{E}[F_n(\tilde{\theta}_{n,-U}(\lambda)) - F_n(\theta^*(\lambda))] \leq L(\frac{2m^2 CL}{\mu^2 n^2} + \frac{m^2 L^2}{\mu^2 n^2} + \sqrt{d} \sqrt{\frac{2m(1.25/\delta)}{e}} (\frac{2m^2 CL}{\mu^2 n^2} + \frac{m^2 L^2}{\mu^2 n^2}) + \frac{\mu L}{\mu n})$$

$$\leq \frac{2m^2 CL}{\mu^2 n^2} + \frac{m^2 L^2}{\mu^2 n^2} + \sqrt{d} \sqrt{\frac{2m(1.25/\delta)}{e}} (\frac{2m^2 CL}{\mu^2 n^2} + \frac{m^2 L^2}{\mu^2 n^2}) + \frac{\mu L}{\mu n}$$

$$\leq (1 + \sqrt{d} \sqrt{\frac{2m(1.25/\delta)}{e}}) (\frac{2m^2 CL}{\mu^2 n^2} + \frac{m^2 L^2}{\mu^2 n^2}) + 4mL^2 \frac{\mu}{\mu n}$$

$$\leq (1 + \sqrt{d} \sqrt{\frac{2m(1.25/\delta)}{e}}) (\frac{2m^2 CL}{\mu^2 n^2} + \frac{m^2 L^2}{\mu^2 n^2}) + 4mL^2 \frac{\mu}{\mu n}$$

$\square$
Finally, we prove that our unlearning algorithm (1) results in $(\epsilon, \delta)$-certifiable removal of datapoint \( z \in U \subseteq S \).

**Proof.** We use a similar technique to the proof of the differential privacy guarantee for the Gaussian mechanism ([9]).

Let \( \hat{\theta}_n(\lambda) \) be the output of learning algorithm \( A \) trained on dataset \( S \) and \( \tilde{\theta}_{n,-U}(\lambda) \) be the output of unlearning algorithm \( M \) run on the sequence of delete requests \( U, \theta_n(\lambda) \), and the data statistics \( T(S) \). We also note the output of \( M \) before adding noise as \( \hat{\theta}_{n,-U}(\lambda) \). Finally, we denote \( \tilde{\theta}_{n,-U}(\lambda) \) as the output of \( A \) trained on the dataset \( S \setminus U \).

We note that in Algorithm 1 that \( \tilde{\theta}_{n,-U}(\lambda) \) is simply \( \tilde{\theta}_{n,-U}(\lambda) = \tilde{\theta}_{n,-U}(\lambda) + \sigma \). The noise \( \sigma \) is sampled from \( \mathcal{N}(0, c^2 I) \) with \( c = \| \tilde{\theta}_{n,-U}(\lambda) - \tilde{\theta}_{n,-U}(\lambda) \|_2 \cdot \frac{\sqrt{2m(\log(1/\delta))}}{\epsilon} \). Where \( \| \tilde{\theta}_{n,-U}(\lambda) - \tilde{\theta}_{n,-U}(\lambda) \|_2 \leq 2m^2\mu L^2 + \frac{m^2\mu L^2}{n^2} \) (6b). Following the same proof for the DP guarantee of the Gaussian mechanism as Dwork et al. [9] (Theorem A.1) given the noise is sampled from the previously described Gaussian distribution we get for any \( \Theta \):

\[
P(\hat{\theta}_{n,-U} \in \Theta) \leq e^\epsilon P(\tilde{\theta}_{n,-U} \in \Theta) + \delta, \quad \text{and} \quad P(\tilde{\theta}_{n,-U} \in \Theta) \leq e^\epsilon P(\hat{\theta}_{n,-U} \in \Theta) + \delta
\]

resulting in \((\epsilon, \delta)\)-unlearning.

\[\square\]

**B Proof of Algorithm 1 Deletion Capacity**

The upper bound on the excess risk (Theorem 1) implies that we can delete at least:

\[
m_{e,\delta,\gamma}(d, n) \geq c \cdot \frac{n\sqrt{\tau}}{(\log(1/\delta))^{\frac{3}{2}}}
\]

where \( c \) depends on the properties of function \( F(z, \theta, \lambda) \). We specifically derive the value of \( c \) below by substituting our deletion capacity bound as \( m \) into the empirical excess risk upper bound:

\[
\mathbb{E}[F(\hat{\theta}_{n,-U}(\lambda)) - F(\theta^*(\lambda))] = O \left( \frac{(2C\mu + M)L^2m^2 \sqrt{\tau(\log(1/\delta))}}{\mu^2n} + \frac{4mL^2}{\mu n} \right)
\]

Plugging in the deletion capacity bound \( m = c \cdot \frac{n\sqrt{\tau}}{(\log(1/\delta))^{\frac{3}{2}}} \) into the excess risk bound (12) then

\[
\frac{(2C\mu + M)L^2m^2 \sqrt{\tau(\log(1/\delta))}}{\mu^2n} + \frac{4mL^2}{\mu n} = \frac{c^2(2C\mu + M)L^2}{\mu^2} + \frac{4L^2}{\mu} \leq c \left( \frac{c(2C\mu + M)L^2}{\mu^2} + \frac{4L^2}{\mu} \right)
\]

Therefore,

\[
c \leq \gamma \left( \frac{\mu^2}{(2C\mu + M)L^2} + \frac{\mu}{4L^2} \right) \implies \mathbb{E}[F(\hat{\theta}_{n,-U}(\lambda)) - F(\theta^*(\lambda))] \leq \gamma
\]

given \( c \leq 1 \). Note that the third line follows from the fact that \( \frac{\sqrt{\tau}}{(\log(1/\delta))^{\frac{3}{2}}} \leq 1 \) given \( \epsilon \leq 1 \) and \( \delta \leq 0.005 \).

**C Extension of non-smooth regularizer to [28]**

Given a function \( F(z, \theta, \lambda) \) with a non-smooth regularizer \( \pi(\theta) \) which satisfies Assumption 2, the algorithm from Sekhri et al. [28] can use non-smooth regularizers with the same deletion capacity,
generalization, and unlearning guarantees as Algorithm 1. This follows from the fact that the removal mechanism introduced by Sekhari et al. [28] minimizes $\psi_2$ in Appendix A.1. Therefore the optimizer comparison theorem can be applied and the distance between the estimator and the leave-U-out estimator can be upper bounded by the same terms (more precisely, we can upper bound this distance by $\frac{n^2 M^2}{M n^2}$).

D Dataset Details

**MNIST** We consider digit classification from the MNIST dataset which contains 60000 images of digits from 1-9. We select only digits 3 and 8 to simplify the task to binary classification. We flatten the original images which are $28 \times 28$ into a vector of 784 pixels. Additionally, we allow for either random sampling or adaptive sampling where the probability of sampling a 3 is set to 10% and the probability of sampling an 8 is set to 90%.

**SVHN** We consider digit recognition from street signs from the SVHN dataset which contains 60000 images of street sign images that contain digits from 1-9. We select only digits 3 and 8 to simplify the task to binary classification. We flatten the original images which are $28 \times 28$ into a vector of 784 pixels. Additionally, we allow for either random sampling or adaptive sampling where the probability of sampling a 3 is set to 10% and the probability of sampling an 8 is set to 90%.

**Warfarin Dosing** Warfarin is a prescription drug used to treat symptoms stemming from blood clots (e.g. deep vein thrombosis) and to help reduce the incidence of stroke and heart attack in at-risk patients. It is an anticoagulant which inhibits blood clotting but overdosing leads to excessive bleeding. The appropriate dosage for a patient dependent on demographic and physiologic factors resulting in high variance between patients. We focus on predicting small or large dosages for patients (defined as > 30mg/week) from a dataset released by the International Warfarin Pharmacogenetics Consortium [8] which contains both demographic and physiological measurements for patients. The dataset contains 5528 examples each with 62 features.

E Additional Experiments

**Logistic Regression with Smooth Regularizers** We present the test accuracy results for the remaining values of $\lambda = \{10^{-4}, 10^{-5}, 10^{-6}\}$.

![Figure 4: IJ vs. RT and TA for smooth regularizers. Comparing both the test accuracy of the unlearned models in our $\ell_2$ logistic regression setup for $\lambda = 10^{-4}$ for random vs adaptive sampling.](image-url)
Figure 5: IJ vs. RT and TA for smooth regularizers. Comparing both the test accuracy of the unlearned models in our $\ell_2$ logistic regression setup for $\lambda = 10^{-5}$ for random vs adaptive sampling.

Figure 6: IJ vs. RT and TA for smooth regularizers. Comparing both the test accuracy of the unlearned models in our $\ell_2$ logistic regression setup for $\lambda = 10^{-6}$ for random vs adaptive sampling.

Logistic Regression with Non-Smooth Regularizers

We present the test accuracy results for the remaining values of $\lambda = \{10^{-4}, 10^{-5}, 10^{-6}\}$.

Figure 7: IJ vs. RT for non-smooth regularizers. Comparing the test accuracy of the unlearned models in our $\ell_1$ logistic regression setup for $\lambda \in \{10^{-4}, 10^{-5}, 10^{-6}\}$.

Non-Conxex: Logistic Regression with Differentially Private Feature Extractor

We present the test accuracy results for the remaining values of $\lambda = \{10^{-4}, 10^{-5}, 10^{-6}\}$.

Figure 8: IJ vs. TA and RT for non-convex training. Comparing both the test accuracy of the unlearned models in our DP feature extractor + $\ell_2$ setup for $\lambda = 10^{-4}$.
Figure 9: IJ vs. TA and RT for non-convex training. Comparing both the test accuracy of the unlearned models in our DP feature extractor + $\ell_2$ setup for $\lambda = 10^{-5}$.

Figure 10: IJ vs. TA and RT for non-convex training. Comparing both the test accuracy of the unlearned models in our DP feature extractor + $\ell_2$ setup for $\lambda = 10^{-6}$.

E.1 Runtimes

Figure 11: IJ vs. RT vs. TA for smooth regularizers on MNIST. Demonstrating runtime improvements across different hyperparameter settings of $10^{-4}$, $10^{-5}$, $10^{-6}$.

Figure 12: IJ vs. RT vs. TA for non-convex settings on SVHN. Demonstrating runtime improvements across different hyperparameter settings of $10^{-4}$, $10^{-5}$, $10^{-6}$.
Figure 13: IJ vs. RT for non-smooth settings on Warfarin. Demonstrating runtime improvements across different hyperparameter settings of $10^{-4}$, $10^{-5}$, $10^{-6}$. 