Tight Bounds for Machine Unlearning via Differential Privacy (Supplementary)

Anonymous Author(s)
Affiliation
Address
email

1 Proof of Theorem 3.1 (Lower Bound)

Theorem 1.1 (Deletion capacity from unlearning via DP, Lower Bound (Theorem 3.1 in Submission))**.** *Suppose $\mathcal{W} \subseteq \mathbb{R}^d$, and fix any Lipschitz convex loss function. Then there exists a lazy (ε, δ)-unlearning algorithm $(\bar{A}, A)$, where $\bar{A}$ has the form $\bar{A}(U, A(S), T(S)) := A(S)$ (and thus, in particular, takes no side information) with deletion capacity*

$$m^{A,\bar{A}}_{\varepsilon,\delta}(\alpha) \geq \Omega\left(\frac{\varepsilon n \alpha}{\sqrt{d \log(1/\delta)}}\right)$$

where the constant in the $\Omega(\cdot)$ only depends on the properties of the loss function.

We first restate some useful results before diving into the proof, starting with some results on Concentrated DP (zCDP).

Proposition 1.2 (k-distance group privacy of ρ-zCDP [Bun and Steinke, 2016, Proposition 1.9])**.** *Let $M : \mathcal{X}^n \to \mathcal{Y}$ satisfy ρ-zCDP. Then, M is $(k^2\rho)$-zCDP for every $X, X' \in \mathcal{X}^n$ that differs in at most k entries.*

Lemma 1.3 (zCDP mini-batch noisy SGD Feldman et al. [2020])**.** *Fix any L-Lipschitz convex loss function over a convex subset $\mathcal{B}$ of $\mathbb{R}^d$ of diameter D. Then there exists an algorithm A which satisfies $(\rho^2/2)$-zCDP with excess population loss:*

$$\mathbb{E}\left[F(\theta) - \min_{\theta \in \mathcal{B}} F(\theta)\right] \leq O\left(DL \cdot \left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\rho n}\right)\right)$$

where the expectation is taken over the randomness of A.

Proof of Theorem 1.1. The proof follows the same setting as in Sekhari et al. [2021]. The main change is that we apply group privacy bounds in terms of zCDP instead of the standard DP guarantee provided by [Bassily et al., 2019, Theorem 3.2].

We first establish a tighter bound for algorithm that achieves m-entries group privacy via Lemma 1.3. Feldman et al. [2020] provides a zCDP version of [Bassily et al., 2019, Theorem 3.2] with $\rho^2/2$-zCDP, hence by group privacy, we yield $\frac{m^2\rho^2}{2}$-zCDP by Proposition 1.2 for neighboring datasets differing in m entries. Then, translating $\frac{m^2\rho^2}{2}$-zCDP to (ε, δ)-DP yields $\varepsilon = O\left(m\rho\sqrt{\log(1/\delta)}\right)$.

By the above discussion, using this zCDP-private learning algorithm with $\rho = \Theta\left(\frac{\varepsilon}{m\sqrt{\ln(1/\delta)}}\right)$, we get an excess population loss bounded by

$$O\left(DL\left(\frac{1}{\sqrt{n}} + \frac{m\sqrt{d\ln(1/\delta)}}{\varepsilon n}\right)\right) \tag{1}$$

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

It only remains to show how the claimed deletion capacity bound follows from this excess population risk guarantee. Construct, as discussed earlier, an unlearning algorithm $\bar{A}$ that returns the input without making any changes (and in particular does not require any additional statistics $T(S)$, and satisfies the laziness assumption). Since A is (ε, δ)-DP, for any set $U \subseteq S, |U| = m$, and $W \subseteq \mathcal{W}$,

$$\Pr[\, A(S) \in W\,] \leq e^{\varepsilon} \Pr[\, A(S') \in W\,] + \delta$$
$$\Pr[\, A(S') \in W\,] \leq e^{\varepsilon} \Pr[\, A(S) \in W\,] + \delta$$

. But since $\bar{A}(U, A(S)) = A(S)$, this readily yields, letting $S' := S \setminus U$:

$$\Pr\left[\, \bar{A}(U, A(S)) \in W\,\right] \leq e^{\varepsilon} \Pr\left[\, \bar{A}(\emptyset, A(S')) \in W\,\right] + \delta$$
$$\Pr\left[\, \bar{A}(\emptyset, A(S')) \in W\,\right] \leq e^{\varepsilon} \Pr\left[\, \bar{A}(U, A(S)) \in W\,\right] + \delta$$

which implies that $(A, \bar{A})$ is indeed (ε, δ)-unlearning for U of size (up to) m.

Recalling the definition of deletion capacity, we finally deduce from (1) the deletion capacity with excess population risk less than α:

$$m^{A,\bar{A}}_{\varepsilon,\delta}(\alpha) \geq m = \Omega\left(\frac{\varepsilon n \alpha}{\sqrt{d \ln (1/\delta)}} \right)$$

where the $O(\cdot)$ hides constant factors depending only on the loss function (namely, the Lipschitz function L, and the diameter D). □

2 Proof of Theorem 3.3 (Upper Bound)

Theorem 2.1 (Deletion capacity from unlearning via DP, Upper Bound (Theorem 3.3 in Submission))**.** *There exists a Lipschitz convex loss function (indeed,* linear*) for which any $\varepsilon, \delta)$-unlearning algorithm $(\bar{A}, A)$ which takes no side information must have deletion capacity*

$$m^{A,\bar{A}}_{\varepsilon,\delta}(\alpha) \leq O\left(\frac{\varepsilon n \alpha}{\sqrt{d \log (1/\delta)}} \right).$$

Proof of Theorem 2.1. We will consider the following linear (and therefore convex and Lipschitz) loss function $\mathcal{L}(\theta, S)$:

$$\mathcal{L}(\theta, S) := -\langle \theta, \sum_{i=1}^{n} x_i \rangle \tag{2}$$

for dataset S of n points $x_1, \dots, x_n \in \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$. We also define the 1-way marginal query, i.e. average, as:

$$q(S) := \frac{1}{n} \sum_{i=1}^{n} x_i\,. \tag{3}$$

To establish our deletion capacity lower bound with respect to this loss function, we will proceed in three stages: the first, relatively standard, is to relate population loss (what we are interested in) to *empirical* loss – which allows us to focus on the existence of a "hard dataset." The second step is then to establish a sample complexity lower bound on the empirical risk (for this loss function) of any (ε, δ)-DP algorithm, via a reduction to differentially private computing of 1-marginals. This step is similar to the one underlying the (weaker) lower bound of Sekhari et al. [2021] (itself relying on an argument of [Bassily et al., 2019]), although a more careful choice of building blocks for the reduction already allows us to obtain an improvement by logarithmic factors.

Third, lift this DP lower bound to a stronger lower bound for DP with respect to edit distance m. This step is quite novel, as it morally corresponds to establishing the converse of the grouposition property of differential privacy (for our specific setting), a converse which does *not* hold in general. Our argument, relatively simple, will crucially rely on the linearity of our loss function.

We omit the details of the first step (reduction from population to empirical loss) in this detailed outline, as it is quite standard. For the second step, our starting point is the following lower bound of Steinke and Ullman:

Theorem 2.2 (Lower bound for one-way marginals [Steinke and Ullman, 2016, Main Theorem]). *For every $\varepsilon \in (0,1)$, every function $\delta = \delta(n)$ such that $\delta \geq 2^{-o(n)}$ and $\delta \leq 1/n^{1+\Omega(1)}$, and for every $\alpha \leq 1/10$, if $A : \{\pm 1\}^{n \times d} \to [\pm 1]^d$ is (ε, δ)-differentially private and $\mathbb{E}[\|\mathcal{A}(S) - q(S)\|_1] \leq \alpha d$ (i.e., with average-case accuracy α) for all $S \in \{\pm 1\}^{n \times d}$, then we must have*

$$n \geq \Omega\left(\frac{\sqrt{d \ln(1/\delta)}}{\varepsilon\alpha}\right).$$

Using this lower bound as a blackbox, we then can adapt the argument of [Bassily et al., 2014, Lemma 5.1, Part 2] to obtain the following stronger result:

Lemma 2.3 (Lower bound for Privately Computing 1-way Marginals). *Let $n, d \in \mathbb{N}, \varepsilon > 0, 2^{-on} \leq \delta(n) \leq 1/n^{1+\Omega(1)}$. For all $\alpha \leq 1/10$, if $\mathcal{A}$ is (ε, δ)-differentially private. Then, for $S \subseteq \{\pm \frac{1}{\sqrt{d}}\}^{n \times d}$, one must have*

$$\mathbb{E}[\|\mathcal{A}(S) - q(S)\|_2] = \min\left(\alpha, \Omega\left(\frac{\sqrt{d \ln(1/\delta)}}{n\varepsilon}\right)\right),$$

where $q(S) = \frac{1}{n}\sum_{i=1}^{n} x_i$ as before. Moreover, this still holds under the assumption that $\|q(S)\|_2 \in [\frac{M-1}{n}, \frac{M+1}{n}]$, where $M = \Omega(\min(n\alpha, \frac{\sqrt{d \ln(1/\delta)}}{\varepsilon}))$.

Proof of Lemma 2.3. Our proof follows the same outline as in Bassily et al. [2014], but using the result of Theorem 2.2 as a black-box instead of the packing argument of Bassily et al. [2014]. Before doing so, we have to translate the result from Theorem 2.2 into our setting, and handle the slightly different choice of parameterization ($\{\pm 1\}^d$ instead of $\{\pm 1/\sqrt{d}\}^d$).

Let $n_\alpha \coloneqq C \cdot \frac{\sqrt{d \ln(1/\delta)}}{\varepsilon\alpha}$, where $C > 0$ is (strictly smaller than) the constant hidden in the $\Omega(\cdot)$ of Theorem 2.2. By contradiction, suppose that, for some $n \leq n_\alpha$, we have an (ε, δ)-differentially private algorithm $\mathcal{A}$ that takes in a dataset $S \subseteq \{\pm \frac{1}{\sqrt{d}}\}^{n \times d}$ and outputs an estimate $\mathcal{A}(S)$ of $q(S)$ with expected L_2 error α. Rescaling, we get that the algorithm $\mathcal{A}'$ which, on input $S' \subseteq \{\pm 1\}^{n \times d}$, computes $S \coloneqq S'/\sqrt{d} \subseteq \{\pm \frac{1}{\sqrt{d}}\}^{n \times d}$ and outputs $\sqrt{d} \cdot \mathcal{A}(S)$ is (1) (ε, δ)-DP by post-processing, and (2) since q is linear, has error related to that of $\mathcal{A}$ by

$$\mathbb{E}[\|\mathcal{A}'(S') - q(S')\|_2] = \sqrt{d} \cdot \mathbb{E}[\|\mathcal{A}(S) - q(S)\|_2] \leq \sqrt{d} \cdot \alpha \tag{4}$$

However, by Theorem 2.2, $\mathcal{A}'$ must have expected L_1 error at least αd since $n \leq n_\alpha$. By Cauchy–Schwarz,

$$\alpha d < \mathbb{E}[\|\mathcal{A}'(S') - q(S')\|_1] \overset{\text{CS}}{\leq} \sqrt{d} \cdot \mathbb{E}[\|\mathcal{A}'(S') - q(S')\|_2] \overset{(4)}{\leq} \sqrt{d} \cdot (\alpha\sqrt{d}) = \alpha d$$

leading to a contradiction. So for $n \leq n_\alpha$, any (ε, δ)-DP algorithm to estimate q must have expected L_2 error at least α, i.e., $\mathbb{E}[\|\mathcal{A}(S) - q(S)\|_2] \geq \alpha$. Further, one can see by inspection of the proof of Theorem 2.2 that $\|q(S)\|_2$ satisfies the assumption in the "Moreover."

Now, for $n \geq n_\alpha$ (assume, for simplicity and without loss of generality, that $n - n_\alpha$ is even), we use a padding argument to establish the other part of the bound. Let $\mathcal{A}$ be any (ε, δ)-differentially private algorithm for answering q on datasets of size n. Suppose for the sake of contradiction, that $\mathcal{A}$ satisfies

$$\mathbb{E}[\|\mathcal{A}(S) - q(S)\|_2] < \frac{n_\alpha}{n} \cdot \alpha \tag{5}$$

for every dataset S of size n.

Fix an arbitrary point $\mathbf{c} \in \{\pm 1/\sqrt{d}\}^d$. Given any dataset $S = (x^{(1)}, \ldots, x^{(n_\alpha)}) \in \{\pm 1\}^{d \times n_\alpha}$ of size n_α, we construct $\hat{S}$ of size n as follows. Its first n_α entries are $x^{(1)}, \ldots, x^{(n_\alpha)}$; then for the remaining $n - n_\alpha$, we have (1) the first $\lceil \frac{n-n_\alpha}{2} \rceil$ (i.e. the first half) of those entries are all copies of $\mathbf{c}$, and (2) the remaining $\lfloor \frac{n-n_\alpha}{2} \rfloor$ are copies of $-\mathbf{c}$.

Note that we have

$$q(\hat{S}) = \frac{n_\alpha}{n} q(S)$$

for every S, and in particular $\|q(\hat{S})\|_2$ satisfies the assumption in the "Moreover."

Now, we define an algorithm $\hat{\mathcal{A}}$ for approximating q on datasets of size n_α as follows. On input $S \in \{\pm 1\}^{d \times n_\alpha}$, $\hat{\mathcal{A}}$:

1. Computes $\hat{S} \in \{\pm 1\}^{d \times n}$ as above
2. Outputs $\frac{n}{n_\alpha}\mathcal{A}(\hat{S})$

Since $\mathcal{A}$ is already differentially private, $\hat{\mathcal{A}}$ is also (ε, δ)-DP due to the post-processing property of differential privacy. Moreover,

$$\mathbb{E}\Big[\|\hat{\mathcal{A}}(S) - q(S)\|_2\Big] = \mathbb{E}\left[\left\|\frac{n}{n_\alpha}\mathcal{A}(\hat{S}) - \frac{n}{n_\alpha}q(\hat{S})\right\|_2\right] = \frac{n}{n_\alpha}\mathbb{E}\left[\left\|\mathcal{A}(\hat{S}) - q(\hat{S})\right\|_2\right] \overset{(5)}{<} \frac{n}{n_\alpha}\cdot\frac{n_\alpha}{n}\alpha = \alpha$$

and so $\hat{\mathcal{A}}$ achieves expected error strictly smaller than α on datasets of size n_α; which contradicts the first part of the lower bound we already established. So for $n > n_\alpha$, any (ε, δ)-DP algorithm to estimate q must have expected L_2 error at least $\frac{n_\alpha}{n} \cdot \alpha = C \cdot \frac{\sqrt{d \ln(1/\delta)}}{n\varepsilon}$.

Finally, we we have shown that for every n and every $\varepsilon > 0$, there is a constant $C > 0$ such that every (ε, δ)-differentially private algorithm $\mathcal{A}$ answering the linear query q must have, on some dataset S of size n, expected L_2 error at least

$$\mathbb{E}[\|\mathcal{A}(S) - q(S)\|_2] = \min\left(\alpha, C \cdot \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right).$$

proving the lemma. □

Combining the above with the argument strategy of [Bassily et al., 2014, Theorem 5.3] finally yields the main lemma for the second step of our proof for Theorem 1.1:

Lemma 2.4 (Lower bound on empirical loss of (ε, δ)-DP algorithms)**.** *Let $n, d \in \mathbb{N}, \varepsilon > 0$, and $\delta = o(1/n)$. For every (ε, δ)-differentially private algorithm with output θ^{priv}, there is a dataset $S = \{x_1, \ldots, x_n\} \subseteq \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$ such that*

$$\mathbb{E}\big[\mathcal{L}(\theta^{priv}, S) - \mathcal{L}(\theta^*, S)\big] = \min\left(\alpha^2, \Omega\left(\frac{d\log(1/\delta)}{n^2\varepsilon^2}\right)\right)$$

where $\theta^* := \frac{\sum_{i=1}^n x_i}{\|\sum_{i=1}^n x_i\|_2}$ is the minimizer of $\mathcal{L}(\theta, S) := -\langle\theta, \frac{1}{n}\sum_{i=1}^n x_i\rangle$ (which is linear and, as such, Lipschitz and convex).

Proof of Lemma 2.4. This proof follows the same structure as that of [Bassily et al., 2014, Theorem 5.3] but adapt the bound in terms of expectation.

First, observe that for any $\theta \in \mathbb{B}$ and dataset S we have:

$$\mathcal{L}(\theta, S) - \mathcal{L}(\theta^*, S) = \frac{1}{2}\|q(S)\|_2\|\theta - \theta^*\|_2^2,$$

since $\|\theta - \theta^*\|_2^2 = \|\theta^*\|_2^2 + \|\theta\|_2^2 - 2\langle\theta, \theta^*\rangle = 2(1 - \langle\theta, \theta^*\rangle)$ using the fact that $\theta^*, \theta \in \mathbb{B}$ have $\|\theta\|_2, \|\theta^*\|_2 = 1$.

Suppose that there is an (ε, δ)-differentially private algorithm $\mathcal{A}$ that outputs θ^{priv} such that, for every dataset $S \subseteq \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$, we have:

$$\mathbb{E}\big[\mathcal{L}(\theta^{priv}, S) - \mathcal{L}(\theta^*, S)\big] \leq \Delta$$

for a sufficiently small constant $C > 0$, and some $\Delta \geq 0$. We will prove a lower bound on Δ. To do so, recall $q(S) = \theta^* \cdot \|q(S)\|_2$; and that the lower bound from Lemma 2.3 still holds when the dataset S is promised to be such that $q(S) \in [(M \pm 1)/n]$, for $M = \Theta(\min(n\alpha, \sqrt{d\log(1/\delta)}/\varepsilon)$.

Consider the algorithm (private by post-processing) $\mathcal{A}$ which outputs $\mathcal{A}(S) = \frac{M}{n}\theta^{priv}$. Then, for any dataset S such that $\|\sum_{i=1}^{n} x_i\|_2 \in [M-1, M+1]$,

$$\mathbb{E}[\|\mathcal{A}(S) - q(S)\|_2] \le \mathbb{E}\big[\|\mathcal{A}(S) - q(S)\|_2^2\big]^{1/2} = \mathbb{E}\left[\|\frac{M}{n}\theta^{priv} - q(S)\|_2^2\right]^{1/2}.$$

On the other hand,

$$\begin{aligned}\mathbb{E}\left[\|\frac{M}{n}\theta^{priv} - q(S)\|_2^2\right] &\le 2\left(\mathbb{E}\big[\|q(S)\|_2^2\|\theta^{priv} - \theta^*\|_2^2\big] + \mathbb{E}\left[\|\frac{M}{n}\theta^{priv} - \|q(S)\|_2\theta^{priv}\|_2^2\right]\right) \\ &= 4\|q(S)\|_2\mathbb{E}\big[\mathcal{L}(\theta^{priv}, S) - \mathcal{L}(\theta^*, S)\big] + 2\left(\frac{M}{n} - \|q(S)\|_2\right)^2 \\ &\le \frac{4(M+1)}{n}\mathbb{E}\big[\mathcal{L}(\theta^{priv}, S) - \mathcal{L}(\theta^*, S)\big] + \frac{2}{n^2} \\ &\qquad \text{(as } n\|q(S)\|_2 \in [M-1, M+1]\text{)} \\ &\le \frac{4(M+1)\Delta}{n} + \frac{2}{n^2}\end{aligned}$$

By Lemma 2.3, we know that $\mathbb{E}[\|\mathcal{A}(S) - q(S)\|_2] = \min\left(\alpha, C \cdot \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right)$, for some absolute constant $C > 0$, in the worst case. Hence, we must have

$$\frac{\Delta \cdot M}{n} \ge \min\left(\alpha^2, \frac{d\ln(1/\delta)}{n^2\varepsilon^2}\right);$$

recalling the setting of M, we get $\mathbb{E}\big[\mathcal{L}(\theta^{priv}, S) - \mathcal{L}(\theta^*, S)\big] = \min\left(\alpha, \Omega\left(\sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}}\right)\right)$. ☐

The above lemma establishes a lower bound on the empirical loss of any (ε, δ)-differentially private algorithm. To derive from this our claimed lower bound on unlearning algorithms, we need to introduce a dependence on m, the deletion capacity (i.e., number of points to unlearn). This is done in the last (third) step of our argument, via a reduction which establishes a (restricted) converse to the grouposition property of DP.

Recall that an algorithm $M\colon \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-DP for edit distance m if for every pair of neighboring datasets X, X' *that differ in up to m entries*, and every $S \subseteq \mathcal{Y}$:

$$\Pr[\,M(X) \in S\,] \le e^{\varepsilon} \Pr[\,M(X') \in S\,] + \delta.$$

We apply this m-edit distance (ε, δ)-DP on Lemma 2.4 by a reduction that shows: for any differentially private algorithm with respect to edit distance at most m must incur an empirical loss given by Lemma 2.4.

Lemma 2.5. *Suppose there exists an m-edit distance (ε, δ)-DP algorithm $\mathcal{M}$ that takes in a dataset S of size n to approximate $q(S)$ (as defined in* (3)*), with empirical loss γ. Then, we can construct a* 1*-edit distance (i.e., standard) (ε, δ)-DP algorithm $\mathcal{M}'$ that, on input a dataset S' of $N = n/m$ data points, approximates $q(S')$ to error γ.*

Proof of Lemma 2.5. The reduction is quite simple: given $\mathcal{M}$, construct $\mathcal{M}'$ as follows for $N = \frac{n}{m}$ inputs:

$$\mathcal{M}'(x_1, \dots, x_N) = \mathcal{M}(\underbrace{x_1, \dots, x_1}_{m}, \underbrace{x_2, \dots, x_2}_{m}, \dots, \underbrace{x_N, \dots, x_N}_{m}).$$

We immediately have that $\mathcal{M}'$ is (ε, δ)-DP for the usual 1-edit distance between datasets, since $\mathcal{M}$ is DP with respect to edit distance m. The sample complexity and error bound then follows direction from $n = N \times m$, where $n \ge N, N \in \mathbb{N}, m \ge 1$, and the fact that $q(x_1, \dots, x_N) = q(x_1, \dots, x_1, x_2, \dots, x_2, \dots, x_N, \dots, x_N)$ due to the definition of q. ☐

Combining Lemma 2.5 with Lemma 2.4, we get that any m-edit distance (ε, δ)-DP algorithm $\mathcal{M}$ approximating q on datasets of size $n = N \times m$ must have error γ at least

$$\min\left(\alpha, \Omega\left(\frac{\sqrt{d\log(1/\delta)}}{N\varepsilon}\right)\right) = \min\left(\alpha, \Omega\left(\frac{m\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)\right)$$

which, reorganising the terms and recalling the definition of deletion capacity, yields the claimed bound on $m^{A,\bar{A}}_{\varepsilon,\delta}$, and hence completes the proof for Theorem 2.1. □

The proof of Theorem 1.2 (the strongly convex case), restated below, is analogous to those of Theorems 1.1 and 2.1, but using [Feldman et al., 2020, Theorem 5.1] for the upper bound (instead of Lemma 1.3) and [Steinke and Ullman, 2016, Theorem 5.2] for the lower bound (instead of Theorem 2.2).

Theorem 2.6 (Unlearning For Strongly Convex Loss Functions (Theorem 1.2, restated))**.** *Let $f\colon \mathcal{W}\times\mathcal{X}\to\mathbb{R}$ be a 1-Lipschitz* strongly *convex loss function. There exists an (ε,δ)-machine unlearning algorithm which, trained on a dataset $S\subseteq\mathcal{X}^n$, does not store* any *side information about the training set besides the learned model, and can unlearn up to*

$$m = O\left(\frac{n\varepsilon\sqrt{\alpha}}{\sqrt{d\log(1/\delta)}}\right)$$

datapoints without incurring excess population risk greater than α. Moreover, this is tight.

3 Proof of (ε,δ)-unlearning properties

The laziness assumption defined below is essential for the proof, and a natural requirement for practical applications.

Assumption 3.1 (Unlearning Laziness (Assumption 1.3 in Submission))**.** *An unlearning algorithm $(\bar{A},A)$ is said to be* lazy *if, when provided with an* empty *set of deletion requests, the unlearning algorithm $\bar{A}$ does not update the model. That is, $\bar{A}(\emptyset, A(X), T(X)) = A(X)$ for all datasets X.*

Theorem 3.2 (Post-processing of unlearning (Theorem 1.4 in Submission))**.** *Let $(\bar{A},A)$ be an (ε,δ)-unlearning algorithm taking no side information. Let $f\colon\mathcal{W}\to\mathcal{W}$ be an arbitrary (possibly randomized) function. Then $(f\circ\bar{A},A)$ is also an (ε,δ)-unlearning algorithm.*

Proof. The proof follows exactly same as post-processing property of differential privacy. We consider the case that f is a deterministic function here without loss of generality.

Let $T=\{r\in\mathbb{R}^d \mid f(r)\in\mathcal{Y}\}$ and $\mathcal{Y}\subseteq\mathbb{R}^d$. Consider for any $\mathcal{Y}\subseteq\mathbb{R}^d$:

$$\begin{aligned}\Pr\big[\, f(\bar{A}(A(S),U))\in\mathcal{Y}\,\big] &= \Pr\big[\,\bar{A}(A(S),U)\in T\,\big]\\ &\le e^{\varepsilon}\Pr\big[\,\bar{A}(A(S),U)\in T\,\big]+\delta\\ &= e^{\varepsilon}\Pr\big[\, f(\bar{A}(A(S),U))\in\mathcal{Y}\,\big]+\delta\end{aligned}$$

□

Under our laziness assumption, we can establish bounds on applying unlearning algorithm repeatedly when the overall deletion requests is within the deletion capacity:

Theorem 3.3 (Chaining of unlearning (Theorem 1.5 in Submission))**.** *Let $(\bar{A},A)$ be a* lazy *(ε,δ)-unlearning algorithm taking no side information, and able to handle up to m deletion requests. Then, the algorithm which uses $(\bar{A},A)$ to sequentially unlearn k disjoint deletion requests $U_1,\dots,U_k\subseteq X$ such that $|\cup_i U_i|\le m$, outputting*

$$\bar{A}(U_k,\dots,\bar{A}(U_1,A(X))\dots)$$

is an (ε',δ')-unlearning algorithm, with $\varepsilon'=k\varepsilon$ and $\delta'=\delta\cdot\frac{e^{k\varepsilon}-1}{e^{\varepsilon}-1}=O(k\delta)$ (for $k=O(1/\varepsilon)$).

Proof. We proceed by induction on $n\ge 1$. Given a pair of (ε,δ)-unlearning algorithm $(\bar{A},A)$ and deletion requests $D_1,\dots,D_n\subseteq S\in\mathbb{R}^{n\times d}$ such that $|\cup_i D_i|\le m^{\bar{A},A}_{\varepsilon,\delta}$ with $D_i\cap D_j=,\forall i\ne j$ for $i,j\in[n]$.

(1) For $n=1$:

$$\Pr\big[\,\bar{A}(A(S),D_1)\in T\,\big]\le e^{n\varepsilon}\Pr\big[\,\bar{A}(A(S\setminus D_1),\emptyset)\,\big]+\delta$$

by the definition of (ε,δ)-unlearning. Hence the case $n=1$ holds.

(2) Assume $n = k$ is true:

$$\Pr\big[\,\bar{A}(\ldots\bar{A}(A(S), D_1), \ldots, D_k) \in T\,\big] \le e^{k\varepsilon}\Pr\big[\,\bar{A}(A(S\setminus\bar{D}_k), \emptyset)\,\big] + \sum_{i=0}^{k-1} e^{i\varepsilon}\cdot\delta \tag{6}$$

(3) Then for $n = k+1$:

$$\begin{aligned}\Pr\big[\,\bar{A}(\ldots\bar{A}(A(S), D_1), \ldots, D_{k+1}) \in T\,\big] &\overset{(6)}{\le} e^{k\varepsilon}\Pr\big[\,\bar{A}(\bar{A}(A(S\setminus\bar{D}_k), \emptyset), D_{k+1})\,\big] + \sum_{i=0}^{k-1} e^{i\varepsilon}\cdot\delta \\ &= e^{k\varepsilon}\Pr\big[\,\bar{A}(A(S\setminus\bar{D}_k), D_{k+1})\,\big] + \sum_{i=0}^{k-1} e^{i\varepsilon}\cdot\delta \\ &\le e^{(k+1)\varepsilon}\Pr\big[\,\bar{A}(A(S\setminus\bar{D}_{k+1}), \emptyset) \in T\,\big] + \sum_{i=0}^{(k+1)-1} e^{i\varepsilon}\cdot\delta\end{aligned}$$

where the first and third inequality result from the definition of (ε, δ)-unlearning and the second equality is due to Laziness Assumption 3.1.

Hence, by induction, the claim holds for all $n \in \mathbb{N}$. $\square$

Theorem 3.4 (Advanced composition of unlearning (Theorem 1.6 in Submission))**.** *Let $(\bar{A}_1, A), \ldots, (\bar{A}_k, A)$ be lazy (ε, δ)-unlearning (with common learning algorithm A) taking no side information, and define the composition of those unlearning algorithms, $\tilde{A}$ as*

$$\tilde{A}(U, A(X)) = f\big(\bar{A}_1(U, A(X)), \ldots, \bar{A}_k(U, A(X))\big)\,.$$

where $f\colon \mathcal{W}^k \to \mathcal{W}$ is any (possibly randomized) function. Then, for every $\delta' > 0$, $(\tilde{A}, A)$ is an (ε', δ')-unlearning taking no side information, where $\varepsilon' = \frac{k}{2}\varepsilon^2 + \varepsilon\sqrt{2k\ln(1/\delta')}$.

Proof. The proof follows the same argument as in [Vadhan, 2017, Lemma 2.4]. We consider the case of $\delta > 0$ only as the $\delta = 0$ is same with the pure DP proof.

Fix two datasets, S (original dataset) and $S' := S \setminus U$ ("forgotten dataset") where U is the set of delete requests with $|U| \le m^{\bar{A},A}_{\varepsilon,\delta}$. Note that S, S' differs in m entries.

For an output $y = (y_1, \ldots, y_k) \in \mathcal{Y}$, define "memory" loss (which is just privacy loss in differential privacy) to be:

$$\mathcal{L}^{S\to S'}_{\mathcal{A}}(y) = \ln\frac{\Pr[\,\mathcal{A}(A(S), U) = y\,]}{\Pr[\,\mathcal{A}(A(S'), \emptyset) = y\,]}$$

where $|\mathcal{L}^{S\to S'}_{\mathcal{A}}(y)| \le \varepsilon$.

Then, by [Vadhan, 2017, Lemma 1.5] we know that $\bar{A}_i(A(S), U), \bar{A}_i(A(S'), \emptyset)$ are (ε, δ)-indistinguishable, hence there are events $E = E_1 \wedge \ldots \wedge E_k, E' = E'_1 \wedge \ldots \wedge E'_k$ such that w.p. at least $1 - k\delta$ by, for all $y_i, i \in [k]$,

$$\begin{aligned}\mathbb{E}\Big[\mathcal{L}^{S\to S'}_{\mathcal{A}}(y)\Big] &= \mathbb{E}\left[\ln\frac{\Pr[\,\mathcal{A}(A(S), U) = y \mid E\,]}{\Pr[\,\mathcal{A}(A(S'), \emptyset) = y \mid E'\,]}\right] \\ &= \sum_{i=1}^{k}\mathbb{E}\left[\ln\left(\frac{\Pr\big[\,\bar{A}_i(A(S), U) = y \mid E_i\,\big]}{\Pr\big[\,\bar{A}_i(A(S'), \emptyset) = y \mid E'_i\,\big]}\right)\right] \\ &= \sum_{i=1}^{k}\mathbb{E}\Big[\mathcal{L}^{S\to S'}_{\bar{A}_i}(y)\Big]\end{aligned}$$

where we observe that the expectation of the loss is just KL-divergence between the distributions of $\bar{A}_i(A(S), U)$ and $\bar{A}_i(A(S'), \emptyset)$ conditioned on E and E'. Hence:

$$\mathbb{E}\Big[\mathcal{L}^{S\to S'}_{\mathcal{A}}(y)\Big] = \sum_{i=1}^{k} \mathrm{D}_{\mathrm{KL}}(\bar{A}_i(A(S), U)\|\bar{A}_i(A(S'), \emptyset)) \le \frac{k}{2}\varepsilon^2$$

where the inequality is a result from [Bun and Steinke, 2016, Proposition 3.3] when $\alpha = 1$. This proposition is applicable because the conditional distribution of $\bar{A}_i$ is (ε, δ)-indistinguishable, which shares the max-divergence definition.

Then by Hoeffding's inequality where the loss is bounded by $[-\varepsilon, \varepsilon]$, with probability at least $1 - \delta'$, we have:

$$\begin{aligned} \exp\left(-\frac{t^2}{2k\varepsilon^2}\right) &\geq \Pr\Big[\, \mathcal{L}_{\mathcal{A}}^{S \to S'}(y) > \mathbb{E}\Big[\mathcal{L}_{\mathcal{A}}^{S \to S'}(y)\Big] + t \Big] \\ &\geq \Pr\Big[\, \mathcal{L}_{\mathcal{A}}^{S \to S'}(y) > \frac{k}{2}\varepsilon^2 + t \Big] \\ &= \Pr\Big[\, \mathcal{L}_{\mathcal{A}}^{S \to S'}(y) > \varepsilon' \,\Big] \end{aligned}$$

Now for $\delta' := \exp\left(-\frac{t^2}{2k\varepsilon^2}\right)$, we have $t = \varepsilon\sqrt{2k \ln(1/\delta')}$ and $\varepsilon' := \frac{k}{2}\varepsilon^2 + \varepsilon\sqrt{2k \ln(1/\delta')}$.

Hence, for any set $T \in \mathcal{Y}$:

$$\begin{aligned} \Pr[\, \mathcal{A}(A(S), U) \in T \,] &\leq \Pr\Big[\, \mathcal{L}_{\mathcal{A}}^{S \to S'}(y) > \varepsilon' \,\Big] + \sum_{y \in T : \mathcal{L}_{\mathcal{A}}^{S \to S'}(y) \leq \varepsilon'} \Pr[\, \mathcal{A}(A(S), U) = y \,] \\ &\leq \delta' + \sum_{y \in T : \mathcal{L}_{\mathcal{A}}^{S \to S'}(y) \leq \varepsilon'} e^{\varepsilon'} \Pr[\, \mathcal{A}(A(S'), \emptyset) = y \,] \\ &\leq \delta' + e^{\varepsilon'} \Pr[\, \mathcal{A}(A(S'), \emptyset) \in T \,] \end{aligned}$$

where the second inequality is from the definition of unlearning. Thus, along with an application of [Vadhan, 2017, Lemma 1.5], this proves that $\mathcal{A} = (\bar{A}_1, \ldots, \bar{A}_k)$ is indeed $(\varepsilon', \delta' + k\delta)$-unlearning w.r.t. learning algorithm A. $\square$

References

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 464–473. IEEE Computer Society, 2014. doi: 10.1109/FOCS.2014.56.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, pages 11279–11288, 2019.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I*, volume 9985 of *Lecture Notes in Computer Science*, pages 635–658, 2016. doi: 10.1007/978-3-662-53641-4_24.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 439–449, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384335.

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

Thomas Steinke and Jonathan R. Ullman. Between pure and approximate differential privacy. *J. Priv. Confidentiality*, 7(2), 2016. doi: 10.29012/jpc.v7i2.648.

Salil P. Vadhan. The complexity of differential privacy. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, 2017. doi: 10.1007/978-3-319-57048-8_7. URL `https://doi.org/10.1007/978-3-319-57048-8_7`.