NEAR-OPTIMAL MACHINE UNLEARNING UTILITY FOR SMOOTH STRONGLY CONVEX LOSSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning is motivated by legal and user-facing requirements to remove the influence of individuals' data from trained models, such as the right to be forgotten. Prior work has developed algorithms and error bounds for unlearning in smooth strongly convex stochastic optimization but the fundamental statistical cost of unlearning has remained unclear. We nearly resolve this problem by proving upper and lower bounds on the excess population risk of approximate (ε, δ)-unlearning; our bounds are tight up to a condition-number factor. For mean estimation over the unit ball, our upper and lower bounds match up to constants. In fact, our algorithm achieves ε-unlearning, which implies a notable separation between differential privacy (DP) and unlearning: (ε, δ)-unlearning has no statistical advantage over pure ε-unlearning in stochastic convex optimization.

The optimal rate is the usual sampling error plus an unlearning penalty that interpolates between the retraining-from-scratch rate and an exponentially smaller term as ε/d grows, where d is the dimension of the model. In particular, retraining-from-scratch is statistically optimal when $\varepsilon \lesssim d$. On the other hand, for $\varepsilon \gg d$ and large unlearning requests, our ε-unlearning algorithm offers an exponential accuracy improvement over retraining the model from scratch and over DP baselines.

1 INTRODUCTION

Machine unlearning (Cao & Yang, 2015) is motivated by legal, institutional, and user-facing requirements to remove the influence of individuals' data from trained models, such as the right to be forgotten (European Union, 2016; Guo et al., 2020). Given a model trained on a dataset, an unlearning procedure receives a request to "delete" or "unlearn" a subset of the training samples and must update the model so that the result is statistically close to what would have been produced had those samples never been used.

We consider unlearning in stochastic convex optimization (SCO). Given i.i.d. samples $Z = (z_1, \ldots, z_n) \sim P^n$, the goal is to approximately minimize the population loss function

$$\min_{w \in \mathcal{W}} \left\{ F_P(w) := \mathop{\mathbb{E}}_{z \sim P}[f(w, z)] \right\}, \tag{1}$$

where $\mathcal{W} \subseteq \mathbb{R}^d$ is a convex parameter domain and $f(\cdot, z)$ is a loss function. The quality of a solution w is measured by its *excess population risk* $\Delta F_P(w) := F_P(w) - F_P^*$, where $F_P^* := \min_{w' \in \mathcal{W}} F_P(w')$. In the unlearning setting, a learning algorithm $\mathcal{A}$ first receives the full dataset Z and outputs a model, possibly together with side information. Later, an unlearning algorithm $\tilde{\mathcal{A}}$ receives an unlearning request $U \subseteq Z$, with $|U| \leq m$, and must output an updated model. Informally, approximate unlearning requires that this updated model be statistically difficult to distinguish (in the max-divergence sense, à la differential privacy) from the output of the learning algorithm run directly on the retained dataset $Z \setminus U$.

There are two naïve baseline methods for unlearning. First, *differential privacy* (DP) (Dwork et al., 2006) automatically gives unlearning: if the original training algorithm is private enough to hide the contribution of any possible unlearning set, then no special update is needed at unlearning time. However, this can be overly conservative, since a DP algorithm must hide all possible changes in advance, before the unlearning set is known; it does not leverage knowledge of U. Second, discarding

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

the trained model and *retraining the model from scratch* on $Z \setminus U$ gives exact 0-unlearning; but it discards part of the dataset and intuitively seems inefficient from a utility perspective. Can we statistically improve over these two baseline approaches?

In this paper, we investigate the *smallest excess population risk that is achievable* in smooth strongly convex SCO subject to this approximate unlearning constraint. While some use the term "unlearning" exclusively to evoke the associated computational concerns ("How does one unlearn faster than retraining from scratch?"), the operation of removing points from a trained model brings with it a wealth of other considerations – for example, what information must be retained to support data deletion? What statistical phenomena arise when part of our training data is removed? The latter is our focus: indeed, adversarially removing training data from a trained model fundamentally breaks the i.i.d. assumption, and how this affects the downstream utility of the model is not at all obvious. Therefore, we consider it essential to understand these statistical properties before turning to additional challenges such as computation. This mirrors a celebrated line of research on the adjacent problem of private SCO. Bassily et al. (2019) first established the optimal statistical rate for private SCO but with suboptimal time complexity. Subsequent work built upon this foundation to achieve the optimal rate by faster algorithms and in increasing generality (Feldman et al., 2020; Bassily et al., 2020; Asi et al., 2021a;b).

A series of works including Neel et al. (2021); Ullah et al. (2021) consider unlearning under excess *empirical* risk relative to retraining from scratch (RFS), rather than *population* risk. These works necessarily consider only computational tradeoffs, whereas the statistical goal of machine learning is performance on unseen test data (i.e., small excess population risk). This consideration motivated a recent line of work (Sekhari et al., 2021; Allouah et al., 2025; Van Waerebeke et al., 2025; Zou et al., 2025; Qiao et al., 2025) to develop algorithms and excess population risk bounds for unlearning in smooth strongly convex stochastic optimization. Despite progress showing that unlearning can improve over differentially private baselines in certain regimes, the following fundamental question has remained open:

> **Question.** What is the minimax optimal excess population risk for (ε, δ)-unlearning up to m samples in smooth strongly convex stochastic optimization?

Our contributions. We resolve the question above up to a condition-number factor. Our upper and lower bounds characterize the minimax optimal excess-risk for (ε, δ)-unlearning of smooth strongly convex SCO as

$$\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}. \tag{2}$$

The rate (2) holds for all $\varepsilon > 0, \delta \in [0, 1/5]$ and has a natural interpretation that yields structural insights that may be relevant to the broader unlearning program. The $1/n$ term is the usual sampling error present even without unlearning requirements (Nemirovski & Yudin, 1983), while the second term is the unlearning penalty. We show that RFS achieves unlearning penalty $(m/n)^2$, which improves over the standard DP baseline. It is clear that this is statistically optimal for exact 0-unlearning; our novel lower bound shows that this is also optimal for $\varepsilon \lesssim d$, which is typical in high-dimensional settings. On the other hand, for a weaker unlearning constraint $\varepsilon \gg d$, our algorithm can leverage information from the unlearned data to reduce the unlearning penalty by the exponential factor $\exp(-2\varepsilon/(d+2))$. Thus, for large unlearning requests $m \gg \sqrt{n}$ and a weak unlearning constraint $\varepsilon \gg d$, an optimal unlearner can be exponentially more accurate than RFS, DP-based unlearning, and prior unlearning algorithms. Comparisons are summarized in Table 1, which reports the closest apples-to-apples baselines for our rate comparison; broader related work is discussed in Appendix A.

Our main contributions are:

1. **Nearly tight minimax bounds for smooth strongly convex SCO.** We give an $(\varepsilon, 0)$-unlearning algorithm for smooth strongly convex stochastic optimization whose excess risk matches (2) up to a condition-number factor. We complement this with a lower bound showing that no unlearning algorithm can improve the dependence on $n, m, \varepsilon, \delta, d$.
2. **A sharp characterization for mean estimation.** For mean estimation over the unit ball in $\mathbb{R}^d$, our upper and lower bounds match up to constants.

Table 1: Excess-risk rates for ε-unlearning up to m points. We suppress logarithmic, Lipschitz, strong-convexity, smoothness, and condition-number factors. For prior algorithms, we report the sharpened or independently derived rates proved in the Appendix.

Method / result	Excess-risk rate	Reference / comments
DP baseline	$\frac{1}{n} + \frac{d^2m^2}{\varepsilon^2 n^2}$	Group DP and Asi et al. (2021b)
Retrain from scratch (RFS)	$\frac{1}{n} + \left(\frac{m}{n}\right)^2$	Exact 0-unlearning; Theorem 5
ERM / warm-start unlearning	$\frac{1}{n} + \left(\frac{m}{n}\right)^2$	(Allouah et al., 2025; 2026); see Theorem 7 for an independent analysis; approximate unlearning; logarithmic speedup over RFS
Newton-step unlearning	$\frac{1}{n} + \left(\frac{m}{n}\right)^2 + \frac{d^2m^4}{\varepsilon^2 n^4}$	(Sekhari et al., 2021); sharpened in Theorem 8; requires Lipschitz Hessian
Our upper/lower bound	$\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}$	Optimal up to condition number; tight for mean estimation

3. **Sharper analyses of baseline and prior algorithms.** We give a sharp analysis of the exact 0-unlearning baseline retrain-from-scratch, showing that it achieves $1/n + (m/n)^2$ excess risk and is optimal when $\varepsilon \lesssim d$. In the appendix, we sharpen the analysis of the Newton-step algorithm of Sekhari et al. (2021) to get a quadratic improvement in the unlearning error term. The original analysis of Allouah et al. (2025) contained a gap in its population-risk argument, which we identified. The authors subsequently corrected this in an updated preprint (Allouah et al., 2026). We provide an independent analysis of their warm-start ERM procedure, recovering the retraining-from-scratch rate.

Our lower bounds apply to all (ε, δ)-unlearning algorithms, unlike the DP-based lower bounds of Huang & Canonne (2025). In particular, allowing $\delta > 0$ offers no meaningful utility improvements for unlearning. This is in stark contrast to differential privacy, where many statistical tasks, including SCO, exhibit a significant separation between the $\delta = 0$ and $\delta > 0$ settings (Steinke & Ullman, 2016).

Techniques. Our key algorithmic idea is a novel way to exploit information available at unlearning time. The empty-deletion run on a retained dataset $S = Z \setminus U$ returns a point near $\hat{w}_S$ with high probability, but also places a small amount of probability mass over a region covering all possible full-data ERM solutions. When an unlearning request U arrives, the unlearner forms $S = Z \setminus U$, and reconstructs this noise distribution by retraining from scratch. It then moves the high-probability mass near the reduced-data ERM solution $\hat{w}_{Z\setminus U}$ to the full-data ERM solution $\hat{w}_Z$. We calibrate the optimal geometry and probabilities needed to satisfy the ε-likelihood-ratio constraint while achieving an advantage over retraining from scratch by a factor of $e^{-\Theta(\varepsilon/d)}$.

Our lower bound uses packing techniques, but departs from standard DP packing arguments. In unlearning, the algorithm may see the unlearning set and arbitrary side information, so DP-style neighboring-dataset indistinguishability alone is not enough. Instead, we construct many separated mean-estimation instances whose nonzero signal samples can be deleted with high probability. After deletion, the retained datasets collapse to a common reference dataset, so ε-unlearning forces the output distributions for all packed instances to be close to one common reference distribution. A packing argument then yields the matching $e^{-\Theta(\varepsilon/d)}$ unlearning penalty.

1.1 Preliminaries

Let $\|\cdot\|$ denote the Euclidean norm. Let $\mathbb{B}_d$ denote the d-dimensional Euclidean unit ball and let $\mathcal{W} \subseteq \mathbb{R}^d$ be a domain. For a differentiable function $h : \mathcal{W} \to \mathbb{R}$, we say h is *L-Lipschitz* if

$|h(w) - h(u)| \leq L\|w - u\|$ for all $w, u \in \mathcal{W}$; *μ-strongly convex* if $h(u) \geq h(w) + \langle \nabla h(w), u - w\rangle + \frac{\mu}{2}\|u - w\|^2$ for all $w, u \in \mathcal{W}$; and *β-smooth* if $h(u) \leq h(w) + \langle \nabla h(w), u - w\rangle + \frac{\beta}{2}\|u - w\|^2$ for all $w, u \in \mathcal{W}$. For smooth strongly convex losses, $\kappa := \beta/\mu$ denotes the *condition number*.

We make the following assumption throughout, which is standard in the theoretical study of unlearning (Sekhari et al., 2021; Allouah et al., 2025). Unlike Sekhari et al. (2021), we do not require that the Hessian of the loss is Lipschitz.

Assumption 1 (Smooth strongly convex SCO)**.** *The following conditions hold:*

1. *The parameter domain $\mathcal{W} \subseteq \mathbb{R}^d$ is closed, convex, and has finite diameter $D = \sup_{w,w' \in \mathcal{W}} \|w - w'\|$.*
2. *For every $z \in \mathcal{Z}$, the loss $f(\cdot, z) : \mathcal{W} \to \mathbb{R}$ is differentiable and L-Lipschitz on $\mathcal{W}$.*
3. *For every $z \in \mathcal{Z}$, the loss $f(\cdot, z)$ is μ-strongly convex on $\mathcal{W}$.*
4. *For every $z \in \mathcal{Z}$, the loss $f(\cdot, z)$ is β-smooth on $\mathcal{W}$.*
5. *The population minimizer $w_P^* := \text{argmin}_{w \in \mathcal{W}} F_P(w)$ satisfies $\nabla F_P(w_P^*) = 0$.*

Moreover, we will fix throughout a dataset size $n \geq 1$ as well as unlearning capacity $1 \leq m < n$. As in Sekhari et al. (2021), non-trivial unlearning guarantees can only be provided in the regime where $m \leq cn$ for a constant $c < 1$. In particular, we assume throughout that $n - m = \Omega(n)$. Now, for a dataset $Z = (z_1, \ldots, z_n)$, let $\widehat{F}_Z(w) := \frac{1}{n}\sum_{i=1}^n f(w, z_i)$ and, for an unlearning set $U \subseteq Z$, let $Z \setminus U$ denote the retained dataset.

Approximate Unlearning. For random variables X, Y on the same measurable space, write $X \approx_{\varepsilon,\delta} Y$ if for every measurable event E, both $\Pr(X \in E) \leq e^\varepsilon \Pr(Y \in E) + \delta$ and $\Pr(Y \in E) \leq e^\varepsilon \Pr(X \in E) + \delta$. We follow the definition of approximate unlearning used in Sekhari et al. (2021) but do not impose any storage constraints on the unlearner. On input data Z, any learning algorithm $\mathcal{A}(Z)$ is applied. Afterward, the unlearner $\tilde{\mathcal{A}}$ takes as inputs the unlearning set U and the dataset Z and must return a model that is $\approx_{\varepsilon,\delta}$-close to the model that the unlearner would return given a learner input $Z \setminus U$ in the absence of any unlearning request. In general, the unlearner would also take the output of the learner $\mathcal{A}(Z)$ as well as some stored information $T(Z)$ instead of necessarily the whole dataset Z. In our case, we allow the unlearner full storage, meaning $T(Z) = Z$. Thus, in our setting, the unlearner has access to the full dataset Z and the unlearning set U, as well as the learned model $\mathcal{A}(Z)$.

Definition 1 (Approximate ε-unlearning)**.** *An algorithm $\tilde{\mathcal{A}}$ satisfies (ε, δ)-unlearning for up to m deletions if, for every dataset $Z \in \mathcal{Z}^n$ and every deletion set $U \subseteq Z$ with $|U| \leq m$,*

$$\tilde{\mathcal{A}}(U, Z) \approx_{\varepsilon,\delta} \tilde{\mathcal{A}}(\emptyset, Z \setminus U).$$

If $\delta = 0$, we say that $\tilde{\mathcal{A}}$ satisfies ε-unlearning.

We measure utility by *worst-case expected excess population risk after up to m deletions*.

Definition 2. *We say that $\tilde{\mathcal{A}}$ achieves expected excess unlearning risk α if for every distribution $P \in \Delta(\mathcal{Z})$ and $n - m \leq N \leq n$ and any adversarial unlearning request $U(Z) \subseteq Z$ such that $|Z \setminus U(Z)| \geq n - m$, we have*

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}, Z \sim P^N} \left[F_P(\tilde{\mathcal{A}}(U(Z), Z)) - F_P^* \right] \leq \alpha.$$

In the weaker setting of data-independent random deletion requests, our lower bound does not apply and one can show by standard techniques that RFS is already statistically optimal (Shalev-Shwartz et al., 2009). We note that our definition is slightly different than that of prior work including Sekhari et al. (2021), which only requires good performance on $\tilde{\mathcal{A}}(U, Z), Z \sim P^n$. Our stronger definition also requires good performance on $\tilde{\mathcal{A}}(\emptyset, S), S \sim P^{n-m}$, the model that would have been trained without the unlearned data. We present in Section C analogous matching bounds for the weaker definition from prior work. We argue that the optimal unlearner in that setting behaves contrary to the spirit of machine unlearning and we thus advocate the adoption of our stronger notion of utility.

Retraining from scratch (RFS) is defined by $\tilde{\mathcal{A}}(U, Z) := \operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}_{Z \setminus U}(w)$ in combination with empty unlearning request run $\tilde{\mathcal{A}}(\emptyset, Z \setminus U) := \operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}_{Z \setminus U}(w)$; this satisfies exact 0-unlearning by definition.

Differential privacy gives another generic route: if $\mathcal{A}$ is ε-DP for groups of size m, then $\mathcal{A}(Z)$ is already ε-close in distribution to $\mathcal{A}(Z \setminus U)$, so no unlearning-time update is needed (i.e., $\tilde{\mathcal{A}}(U, Z) = \mathcal{A}(Z)$). In particular, by group privacy, an (ε/m)-DP algorithm at the individual-sample level yields ε-unlearning for unlearning sets of size at most m.

2 Upper Bounds for ε-Unlearning

In this section, we provide a novel algorithm for ε-unlearning that achieves excess population risk

$$O\left(\kappa \frac{L^2}{\mu}\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}\right)\right),$$

for SCO and $O(1/n + (m/n)^2 e^{-2\varepsilon/(d+2)})$ mean squared error for mean estimation on a unit ball. Thus, our algorithm gives an exponential improvement over the retrain-from-scratch penalty of $\Theta((m/n)^2)$ in the regime $\varepsilon \gg d$. These upper bounds are tight up to $O(\kappa)$ and $O(1)$ respectively.

Our algorithm has a simple implementation with equivalent runtime to retraining from scratch. Nonetheless, rather than computation, our focus is to demonstrate key statistical phenomena of SCO unlearning. First, RFS is only optimal in the regime $\varepsilon \lesssim d$, outside of which exponential improvement is possible by partially retaining unlearned data. Secondly, the optimal rate is achieved by an ε-unlearning algorithm, whereas our lower bound holds for (ε, δ)-unlearning. This shows a novel separation from DP, where commonly $\delta > 0$ leads to significantly smaller minimax rates.

For any dataset S, write $\hat{w}_S \in \operatorname{argmin}_{w \in \mathcal{W}} \widehat{F}_S(w)$ where $\widehat{F}_S(w) = \frac{1}{|S|} \sum_{z \in S} f(w, z)$.

Our algorithm. Consider the following two extreme algorithms for unlearning. To achieve high accuracy at the expense of providing no unlearning guarantees, an ERM unlearner given a dataset Z and an unlearning request U should simply ignore the request and return an approximate full-data ERM solution $\tilde{w}_Z$. On the other hand, to achieve a perfect unlearning guarantee at the expense of accuracy, the unlearner should retrain from scratch and return an approximate ERM solution on the reduced dataset $\tilde{w}_{Z \setminus U}$.

At a high-level, our ERM unlearner works by optimally interpolating between these two algorithms to provide the desired level of unlearning while maximizing the accuracy. More precisely, given a dataset Z and a (possibly empty) unlearning request U, our unlearner returns with high probability a solution sampled uniformly from a small ball centered around the full-data ERM solution $\tilde{w}_Z$. However, when the learner is run "dry" on a reduced dataset S and an empty unlearning request, the unlearner must provide plausible deniability to any true runs of the unlearner provided with a dataset Z and an unlearning request U that resolves to $S = Z \setminus U$. To achieve this, the unlearner also returns, with low probability, a sample from a wider ball centered around the reduced-data ERM $\tilde{w}_S$ that is just large enough to capture any full-data ERM solutions $\tilde{w}_Z$ that may reduce to $S = Z \setminus U$ by a legal unlearning request. We note that, critically, the approximate ERM solution $\tilde{w}_S$ is deterministic in the sense that it only depends on Z and U through $Z \setminus U$. Figure 1 visualizes the sampling distributions of the unlearning process when run dry compared to a true run of the unlearner. Algorithm 1 provides the pseudocode, which is written for a generic input dataset T. In the true unlearning call, $T = Z$ and the deletion set is U. In the reference empty-deletion dry run, $T = Z \setminus U$ and the deletion set is $\emptyset$.

Guarantees of Algorithm 1. We now formally record our main upper bound guarantees. Note that by gradient query we mean a single evaluation of $\nabla_w f(w, z)$ and by projection query we mean a Euclidean projection onto the convex set $\mathcal{W}$. In particular, under the assumption that gradient and projection queries can be evaluated in $O(d)$ time, the overall runtime of our algorithm is $\tilde{O}(\kappa \cdot nd)$.

Theorem 1 (Main Upper Bound)**.** *Grant Assumption 1. Then the unlearner Algorithm 1 satisfies ε-unlearning, requires $O(d)$ time for randomization as well as $\tilde{O}(\kappa \cdot n)$ gradient and projection queries. Moreover, Algorithm 1 achieves expected excess unlearning risk*

$$\alpha = O\left(\kappa \frac{L^2}{\mu}\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}\right)\right).$$

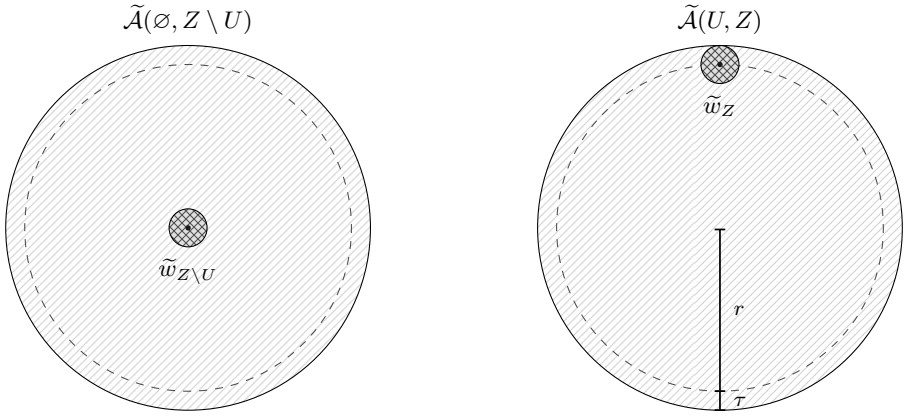

Figure 1: We compare the distribution of $\tilde{\mathcal{A}}(\emptyset, Z \setminus U)$ to $\tilde{\mathcal{A}}(U, Z)$. Lightly shaded regions are sampled uniformly with low probability and heavily shaded regions are sampled uniformly with high probability. With high probability, $\tilde{\mathcal{A}}(\emptyset, Z \setminus U)$ returns a solution very close to the RFS ERM solution $\tilde{w}_{Z\setminus U}$ and, with small probability, a sample from a ball surrounding it. $\tilde{\mathcal{A}}(U, Z)$ has the same distribution except that ball centered around the inaccurate RFS solution $\tilde{w}_{Z\setminus U}$ is shifted to capture the full ERM solution $\tilde{w}_Z$.

Algorithm 1 Core-swap ε-unlearning for ERM

Require: Dataset T; deletion set $U \subseteq T$; parameters $\mathcal{W}, L, \mu, n, m, d, \varepsilon$

1: Set $S \leftarrow T \backslash U$, $\rho \leftarrow \frac{L}{\mu\sqrt{n}}$, $r \leftarrow \frac{2Lm}{\mu(n-m)} + 2\rho$, $\tau \leftarrow r \min\{1, 2e^{-\frac{\varepsilon}{d+2}}\}$, and $\eta \leftarrow \frac{1}{(e^{\varepsilon}-1)\left(\frac{\tau}{\tau+r}\right)^d + 1}$

2: Compute deterministic $\tilde{w}_T$ and $\tilde{w}_S$ such that $\|\tilde{w}_T - \hat{w}_T\|, \|\tilde{w}_S - \hat{w}_S\| \leq \rho$

3: Sample $w \sim \begin{cases} \mathrm{Unif}((\tau + r)\mathbb{B}_d + \tilde{w}_S) & \text{w.p. } \eta \\ \mathrm{Unif}(\tau\mathbb{B}_d + \tilde{w}_T) & \text{w.p. } 1 - \eta \end{cases}$

4: **return** $\mathrm{argmin}_{w' \in \mathcal{W}} \|w' - w\|$

In the special case of mean estimation over the unit ball, i.e. $\mathcal{W} = \mathcal{Z} = \mathbb{B}_d$ and $f(w, z) = \frac{1}{2}\|w - z\|^2$, Algorithm 1 achieves expected excess unlearning risk

$$\alpha = O\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}\right).$$

The key proof idea is that the empty-deletion distribution for $S = Z \setminus U$ places small probability mass near every possible full-data ERM $\tilde{w}_Z$. The unlearning step moves the high-probability mass from a region near $\tilde{w}_S$ onto a region near $\tilde{w}_Z$. Verifying the unlearning guarantee thus involves bounding just one likelihood ratio, while utility follows from carefully controlling the slack radius τ, the probability η of selecting a noisy solution, as well as the ERM approximation error. Our utility analysis separates the statistical error of $\tilde{w}_Z$ from the randomization cost of unlearning.

To formally prove Theorem 1, we proceed in three steps. First, Proposition 1 shows that Algorithm 1 satisfies ε-unlearning. Second, Proposition 2 shows that its output is close to the approximate full-data ERM $\tilde{w}_Z$. Finally, we combine this ERM-distance guarantee with the standard stability-induced distance bound between $\hat{w}_Z$ and the population minimizer w_P^* to obtain our excess population risk guarantee.

Proposition 1. *The unlearner Algorithm 1, satisfies ε-unlearning.*

A proof is provided in Section B.

Proposition 2 (Distance to the Full-Data ERM)**.** *For every $n - m \leq N \leq n$, dataset $Z \in \mathcal{Z}^N$, and deletion request $U \subseteq Z$ with $|Z \setminus U| \geq n - m$, the output $\tilde{\mathcal{A}}(U, Z)$ of Algorithm 1 satisfies*

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}}\left[\left\|\tilde{\mathcal{A}}(U, Z) - \tilde{w}_Z\right\|^2\right] = O\left(\frac{L^2}{\mu^2}\left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)} + \frac{L^2}{\mu^2 n}\right).$$

A proof is provided in Section B.

Lemma 1 (ERM Distance to the Population Minimizer)**.** *For any $N \geq n - m$ and distribution P over $\mathcal{Z}$, we have*

$$\mathop{\mathbb{E}}_{Z\sim P^N}\left[\|\hat{w}_Z - w_P^*\|^2\right] = O\left(\frac{L^2}{\mu^2 n}\right),$$

where $w_P^* = \mathit{argmin}_{w\in\mathcal{W}} F_P(w)$.

Proof. This follows from the standard expected excess-risk bound for ERM with L-Lipschitz, μ-strongly convex losses,

$$\mathop{\mathbb{E}}_{Z\sim P^N}\left[F_P(\hat{w}_Z) - F_P^*\right] = O\left(\frac{L^2}{\mu N}\right) = O\left(\frac{L^2}{\mu n}\right),$$

see e.g. Shalev-Shwartz et al. (2009), together with μ-strong convexity of F_P and $N = \Omega(n)$. □

Lemma 2 (ERM Approximation)**.** *For any error tolerance $\rho > 0$, there exists an algorithm that takes a dataset $Z \in \mathcal{Z}^N$, makes $O(\kappa \cdot N \cdot \log(D/\rho))$ gradient and projection queries, and deterministically computes $\tilde{w}_Z$ such that $\|\tilde{w}_Z - \hat{w}_Z\| \leq \rho$.*

An algorithm achieving the claimed gradient query complexity[1] and error rate is projected gradient descent, which is analyzed in Theorem 3.10 of Bubeck (2015).

We now combine the above results to prove our main upper bound theorem. Details are provided in Section B.

3 Lower Bounds for (ε, δ)-Unlearning

In this section we prove a lower bound nearly matching the rate of our ERM unlearner. Surprisingly, we show that the optimal statistical rate cannot be improved by taking $\delta > 0$. This is in sharp contrast to differential privacy. We begin with a mean estimation lower bound.

Theorem 2. *Let $\tilde{\mathcal{A}}(U, Z)$ be an (ε, δ)-unlearning algorithm with $\delta \leq 1/5$ and suppose that for all distributions P on $\mathbb{B}_d$ with mean μ_P, all unlearning requests $U(Z) \subseteq Z$ with $|Z \setminus U(Z)| \geq n - m$, and any $n - m \leq N \leq n$, we have the mean squared error guarantee*

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}, Z\sim P^N}\left[\|\tilde{\mathcal{A}}(U(Z), Z) - \mu_P\|^2\right] \leq \alpha.$$

Then, it must be the case that

$$\alpha = \Omega\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}\right).$$

The first term $\Omega(1/n)$ is the mean squared error lower bound even without unlearning requirements (see, e.g. Duchi (2021)). The second term is what we will prove in this section. As a consequence of our mean estimation lower bound, we obtain our SCO lower bound:

Corollary 1. *Suppose the loss satisfies Assumption 1 and that there is an (ε, δ)-unlearning algorithm $\tilde{\mathcal{A}}$ with $\delta \leq 1/5$ and excess unlearning risk α. Then we must have*

$$\alpha \geq \Omega\left(\frac{L^2}{\mu}\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}\right)\right).$$

The first term $\Omega(L^2/(\mu n))$ holds for SCO without unlearning constraints (Nemirovski & Yudin, 1983). The second follows from Theorem 2 and a standard reduction from SCO to mean estimation (see e.g., Lowy & Razaviyayn (2025, Proof of Theorem 8)).

[1] In fact, accelerated projected gradient descent improves the κ dependence to $\sqrt{\kappa}$ (Beck, 2017, Theorem 10.42). We do not pursue such improvements as runtime and condition number dependence are not our focus.

Lower-bound intuition. We construct distributions whose means point in many separated directions, but whose nonzero signal samples can all be removed with high probability by a valid unlearning request. After this deletion, all hard instances induce the same retained dataset, so the unlearning guarantee forces their output distributions to be close to a common reference distribution. The packing size of the possible mean directions then limits how accurately all means can be recovered.

Proof of Theorem 2. We now develop the tools that will be needed to prove Theorem 2. Just as our algorithm exploits covering geometry, our lower bound will exploit the packing geometry of the ball $\mathbb{B}_d$.

Lemma 3. *For any $0 < \tau < 1$, there exists $\mathcal{V} \subseteq \mathbb{B}_d \setminus \frac{1}{2}\mathbb{B}_d$ of size $|\mathcal{V}| \geq \frac{1}{2}\left(\frac{1}{\tau}\right)^d$ for which $\|v - v'\| > \tau$ for any $v \neq v' \in \mathcal{V}$.*

Indeed, by a standard volumetric packing argument (see e.g. Section 4.2 of Vershynin (2018)), we can find a τ-separated subset $\mathcal{V} \subseteq \mathbb{B}_d \setminus \frac{1}{2}\mathbb{B}_d$ of size $|\mathcal{V}| \geq \frac{\mathrm{Vol}(\mathbb{B}_d \setminus \frac{1}{2}\mathbb{B}_d)}{\mathrm{Vol}(\tau\mathbb{B}_d)} = \frac{\mathrm{Vol}(\mathbb{B}_d) - 2^{-d}\mathrm{Vol}(\mathbb{B}_d)}{\tau^d \mathrm{Vol}(\mathbb{B}_d)} \geq \frac{1}{2}\left(\frac{1}{\tau}\right)^d$.

Next, we show that, for a well-structured class of contaminated mixture distributions, deleting on the order of m samples leaves a common distribution. Consequently, any algorithm that handles unlearning requests of size up to m cannot effectively distinguish these distributions.

Lemma 4. *Let $n \geq 1$, $v \in \mathbb{B}_d$, and $\eta \in [0, 1/5]$ be such that $m := 5\eta n$ is an integer. There exists a distribution $P_{v,\eta}$ on $\mathbb{B}_d$ with mean ηv such that, given $Z \in \mathbb{B}_d^n$, we can construct an unlearning request $U(Z) \subseteq Z$ of size m for which*

$$\mathop{\mathbb{P}}_{Z \sim P_{v,\eta}^n}(Z \setminus U(Z) \neq \boldsymbol{0}_{n-m}) \leq 2^{-m},$$

where $\boldsymbol{0}_{n-m} := (0, \ldots, 0)$ denotes the dataset consisting of $n - m$ zeroes.

Proof. Consider the contaminated mixture

$$P_{v,\eta} := (1 - \eta)1_0 + \eta 1_v.$$

Sample $Z \sim P_{v,\eta}^n$ and let $K \sim \mathrm{Bin}(n, \eta)$ be the number of non-zero entries in Z. Now, consider the unlearning request $U(Z)$ such that, when $K > m$, $U(Z)$ removes any m entries from Z and, when $K \leq m$, $U(Z)$ removes all of the non-zero entries as well as other entries arbitrarily so that exactly m entries are removed in total.

Clearly, $Z \setminus U(Z) = \mathbf{0}_{n-m}$ as long as $K \leq m$, so by a multiplicative Chernoff bound we get that

$$\mathop{\mathbb{P}}_{Z \sim P_{v,\eta}^n}(Z \setminus U(Z) \neq \mathbf{0}_{n-m}) \leq \mathop{\mathbb{P}}_{K \sim \mathrm{Bin}(n,\eta)}(K > 5\eta n) \leq \left(\frac{e^4}{5^5}\right)^{\eta n} \leq (2^{-5})^{\eta n} = 2^{-m}. \qquad \square$$

Finally, we require a slight variant of the standard technique of packing lower bounds from the differential privacy literature.

Lemma 5. *Let $P^*, P_1, \ldots, P_k$ be $\mathbb{R}^d$-valued distributions such that we can find disjoint events $E_1, \ldots, E_k$ for which $P_i(E_i) \geq p$ as well as $P_i \approx_{\varepsilon,\delta} P^*$ for each $i \leq k$. Assume additionally that we can find $\mu \in \mathbb{R}^d$ so that $\mathbb{E}_{W \sim P^*}[\|W - \mu\|^2] \leq \alpha$ and $\|w - \mu\| \geq \eta$ for each $w \in E_i$ and $i \leq k$. Then*

$$e^\varepsilon \geq \frac{k\eta^2}{\alpha}(p - \delta).$$

Proof. By disjointness of the E_i, we have

$$\alpha \geq \mathop{\mathbb{E}}_{W \sim P^*}[\|W - \mu\|^2] \geq \sum_{i=1}^k \int_{E_i} \|w - \mu\|^2 \, dP^*(w) \geq \eta^2 \sum_{i=1}^k P^*(E_i) \geq k\eta^2 e^{-\varepsilon}(p - \delta),$$

which immediately yields the desired bound. $\square$

We now have assembled all of the tools needed to show Theorem 2. For convenience, we will write $g(X)|_{X \sim P}$ to denote the distribution of $g(X), X \sim P$.

Proof of Theorem 2. Set $\eta := \frac{m}{5n}$. If $\sqrt{\alpha} \geq \frac{1}{8}\eta$, we are done, so assume that $\sqrt{\alpha} < \frac{1}{8}\eta$ and set $\tau := \frac{4\sqrt{\alpha}}{\eta} < 1$. By Lemma 3, we can find a τ-separated $\mathcal{V} \subseteq \mathbb{B}_d \setminus \frac{1}{2}\mathbb{B}_d$ of size $|\mathcal{V}| \geq \frac{1}{2}\left(\frac{\eta}{4\sqrt{\alpha}}\right)^d$.

Now, for each $v \in \mathcal{V}$, recall the hard distribution $P_{v,\eta}$ and corresponding delete request $U_v(Z)$ as in Lemma 4. We claim that $\tilde{\mathcal{A}}(U_v(Z), Z)|_{Z\sim P_{v,\eta}^n} \approx_{\varepsilon,\delta+2^{-m}} \tilde{\mathcal{A}}(\emptyset, \mathbf{0}_{n-m})$. Indeed, for any event E,

$$\begin{aligned}
\mathop{\mathbb{P}}_{\tilde{\mathcal{A}},Z\sim P_{v,\eta}^n}(\tilde{\mathcal{A}}(U_v(Z), Z) \in E) &\leq \mathop{\mathbb{P}}_{\tilde{\mathcal{A}},Z\sim P_{v,\eta}^n}(Z \setminus U_v(Z) = \mathbf{0}_{n-m} \text{ and } \tilde{\mathcal{A}}(U_v(Z), Z) \in E) + 2^{-m} \\
&\leq \mathop{\mathbb{E}}_{Z\sim P_{v,\eta}^n}\left[\mathop{\mathbb{P}}_{\tilde{\mathcal{A}}}(\tilde{\mathcal{A}}(U_v(Z), Z) \in E) \,\middle|\, Z \setminus U_v(Z) = \mathbf{0}_{n-m}\right] + 2^{-m} \\
&\leq \mathop{\mathbb{E}}_{Z\sim P_{v,\eta}^n}\left[e^{\varepsilon}\mathop{\mathbb{P}}_{\tilde{\mathcal{A}}}(\tilde{\mathcal{A}}(\emptyset, Z \setminus U_v(Z)) \in E) + \delta \,\middle|\, Z \setminus U_v(Z) = \mathbf{0}_{n-m}\right] + 2^{-m} \\
&= e^{\varepsilon}\mathop{\mathbb{P}}_{\tilde{\mathcal{A}}}(\tilde{\mathcal{A}}(\emptyset, \mathbf{0}_{n-m}) \in E) + \delta + 2^{-m}
\end{aligned}$$

and, analogously, $\mathbb{P}_{\tilde{\mathcal{A}}}(\tilde{\mathcal{A}}(\emptyset, \mathbf{0}_{n-m}) \in E) \leq e^{\varepsilon}\,\mathbb{P}_{\tilde{\mathcal{A}},Z\sim P_{v,\eta}^n}(\tilde{\mathcal{A}}(U_v(Z), Z) \in E) + \delta + 2^{-m}$ as well, though this is not required to invoke Lemma 5.

On the other hand, by Markov's inequality and our assumption on the mean squared error, for every $v \in \mathcal{V}$ we have $\mathbb{P}_{\tilde{\mathcal{A}},Z\sim P_{v,\eta}^n}(\tilde{\mathcal{A}}(U_v(Z), Z) \in E_v) \geq 3/4$ where $E_v := \{w : \|w - \eta v\| \leq 2\sqrt{\alpha}\}$. Since the packing $\mathcal{V}$ has separation $\tau = \frac{4\sqrt{\alpha}}{\eta}$, the E_v are disjoint. In addition, we have

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}}\left[\|\tilde{\mathcal{A}}(\emptyset, \mathbf{0}_{n-m})\|^2\right] = \mathop{\mathbb{E}}_{\tilde{\mathcal{A}},Z\sim 1_0^{n-m}}\left[\|\tilde{\mathcal{A}}(\emptyset, Z)\|^2\right] \leq \alpha$$

by the mean squared error assumption. Finally, for all $v \in \mathcal{V}$ and $w \in E_v$, since $\|v\| > 1/2$, we must have

$$\|w\| \geq \|\eta v\| - \|w - \eta v\| \geq \eta/2 - 2\sqrt{\alpha} \geq \eta/4.$$

Altogether, Lemma 5 yields

$$e^{\varepsilon} \geq \frac{|\mathcal{V}|(\eta/4)^2}{\alpha}\left(\frac{3}{4} - \delta - 2^{-m}\right) \geq \frac{\eta^2}{640\alpha}\left(\frac{\eta}{4\sqrt{\alpha}}\right)^d \geq \left(\frac{\eta}{640\sqrt{\alpha}}\right)^{d+2} = \left(\frac{m}{3200n\sqrt{\alpha}}\right)^{d+2}.$$

In particular, we have $\alpha \gtrsim \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}$. □

4 Conclusion

We determined the minimax optimal rate for (ε, δ)-approximate machine unlearning in smooth strongly convex stochastic optimization up to a condition-number factor. Our results show that the optimal excess population risk consists of the usual statistical error plus an unlearning penalty that depends exponentially on the ratio ε/d. In particular, retraining from scratch is statistically optimal when $\varepsilon \lesssim d$, whereas in the regime $\varepsilon \gg d$, our novel algorithm improves the deletion cost exponentially over retraining from scratch. For mean estimation over the unit ball, our upper and lower bounds match, giving an essentially sharp characterization of the statistical price of unlearning in this canonical setting. Finally, we show that the lower bound is statistically achievable by an ε-unlearning algorithm. This demonstrates a notable difference between the unlearning and DP settings, where typically $\delta > 0$ enables information-theoretically stronger algorithms.

Several questions remain open for future work. First, our upper and lower bounds for general smooth strongly convex SCO differ by a condition-number factor; closing this gap would give a fully sharp minimax characterization beyond mean estimation and would help pave the way for nonsmooth and non-strongly convex SCO algorithms. Extending the minimax theory of unlearning to nonconvex optimization is an important goal to aim for. Finally, we leave open the optimal rates for unlearning under explicit storage and/or computational constraints. In particular, it is an important question whether the statistically optimal rate can be achieved by a sublinear time algorithm that does not require a full pass over the dataset.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance tasks. For example, we use LLMs to look for typos, double check proofs, improve narrative flow, and to search over related works to ensure relevant works are mentioned. We have not used AI generated text or research ideas and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. Any LLM generated writing feedback is manually assessed and iterated upon. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Youssef Allouah, Joshua Kazdan, Rachid Guerraoui, and Sanmi Koyejo. The utility and complexity of in-and out-of-distribution machine unlearning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Youssef Allouah, Joshua Kazdan, Rachid Guerraoui, and Sanmi Koyejo. The utility and complexity of in-and out-of-distribution machine unlearning, 2026. URL https://arxiv.org/abs/2412.09119v3.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *ICML*, 2021a.

Hilal Asi, Daniel Levy, and John Duchi. Adapting to function difficulty and growth conditions in private optimization. volume 34, pp. 19069–19081, 2021b.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33: 4381–4391, 2020.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *Proceedings of the 42nd IEEE Symposium on Security and Privacy*, pp. 141–159. IEEE, 2021. URL https://arxiv.org/abs/1912.03817.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 987–1004. IEEE, 2015.

Eli Chien, Haoyu Peter Wang, Ziang Chen, and Pan Li. Certified machine unlearning via noisy stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/forum?id=h3k2NXu5bJ.

John Duchi. Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford.edu/class/stats311/Lectures/full_notes.pdf*, 2021.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88, 2016. URL https://gdpr-info.eu/.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM on the Theory of Computing*, pp. 439–449, 2020.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842. PMLR, 2020. URL https://arxiv.org/abs/1911.03030.

Yiyang Huang and Clément L. Canonne. Tight bounds for machine unlearning via differential privacy. *Journal of Privacy and Confidentiality*, 15(2), 2025. doi: 10.29012/jpc.924. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/924.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2008–2016. PMLR, 2021. URL https://proceedings.mlr.press/v130/izzo21a.html.

Jiaqi Liu, Jian Lou, Zhan Qin, and Kui Ren. Certified minimax unlearning with generalization rates and deletion capacity. *Advances in Neural Information Processing Systems*, 36:62821–62852, 2023.

Andrew Lowy and Meisam Razaviyayn. Private stochastic optimization with large worst-case lipschitz parameter. *Journal of Privacy and Confidentiality*, 15:1, 2025.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 931–962. PMLR, 2021. URL https://proceedings.mlr.press/v132/neel21a.html.

A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization.* Chichester, 1983.

Xinbao Qiao, Meng Zhang, Ming Tang, and Ermin Wei. Hessian-free online certified unlearning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=C3TrHWanh5.

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

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2, pp. 5, 2009.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2016.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A. Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4126–4142. PMLR, 2021. URL https://arxiv.org/abs/2102.13179.

Martin Van Waerebeke, Marco Lorenzi, Giovanni Neglia, and Kevin Scaman. When to forget? complexity trade-offs in machine unlearning. *arXiv preprint arXiv:2502.17323*, 2025.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Haolin Zou, Arnab Auddy, Yongchan Kwon, Kamiar Rahnama Rad, and Arian Maleki. Certified machine unlearning under high dimensional regime. *Journal of Machine Learning Research*, 26 (308):1–58, 2025. URL https://jmlr.org/papers/v26/25-1751.html.

APPENDIX

A ADDITIONAL RELATED WORK

We discuss additional related work on machine unlearning, certified removal, and differentially private baselines. See the introduction for the works most directly comparable to our results.

Certified removal and approximate data unlearning. Guo et al. (2020) introduced certified data removal, requiring that a model after data removal be statistically indistinguishable from one trained without the removed data, and developed certified-removal mechanisms for linear classifiers. Izzo et al. (2021) proposed approximate unlearning methods for linear and logistic models whose unlearning-time cost is linear in the feature dimension and independent of the number of training samples. These works emphasize efficient approximate unlearning for specific model classes, while our work characterizes minimax population-risk rates for ε-unlearning in smooth strongly convex stochastic optimization.

Practical machine unlearning and exact unlearning frameworks. Bourtoule et al. (2021) introduced SISA training, a practical framework for accelerating unlearning by sharding, isolating, slicing, and aggregating the training procedure. Their work helped popularize machine unlearning as a practical data-governance problem and focused primarily on reducing unlearning-time computation relative to retraining. This line of work is complementary to ours: we focus on the information-theoretic statistical cost of approximate unlearning in stochastic optimization.

Gradient-based and stability-based unlearning. Neel et al. (2021) study data unlearning for convex models and introduce gradient-based unlearning algorithms that can handle long sequences of adversarial updates with per-unlearning runtime and steady-state error not growing with the sequence length. Ullah et al. (2021) connect unlearning to total-variation stability and design noisy-SGD-based algorithms with efficient unlearning procedures. Chien et al. (2024) later study certified machine unlearning via projected noisy stochastic gradient descent and establish approximate unlearning guarantees under convexity assumptions. These works are algorithmic and computationally motivated; whereas our focus is the minimax statistical rate.

Certified unlearning for convex and strongly convex learning. Sekhari et al. (2021) initiated a population-risk study of certified unlearning for convex learning and gave algorithms with unlearning-capacity guarantees, including a Newton-step method for smooth strongly convex losses. Their method improves over DP baselines in some regimes, but requires stronger smoothness assumptions such as Lipschitz Hessians. In Appendix F, we sharpen the analysis of this algorithm. Allouah et al. (2025) study utility guarantees for in-distribution unlearning and propose a warm-start ERM procedure. Their original published population-risk analysis contained a gap, which was subsequently corrected in an updated preprint (Allouah et al., 2026). In Appendix E, we give an independent analysis recovering the same retraining-level rate. Van Waerebeke et al. (2025) study computational aspects of unlearning for strongly convex losses. These works are the closest algorithmic precursors to ours.

DP-based and restricted unlearning. Differential privacy gives a generic route to unlearning: by group privacy, an algorithm private enough for changes of size m automatically satisfies an unlearning guarantee for unlearning sets of size m. Huang & Canonne (2025) give tight bounds for machine unlearning via differential privacy and for restricted unlearning models, including settings with a DP-based algorithm that does not use side information or the unlearning set. Their results show that DP-based unlearning is optimal in these restricted models. Our results show that unrestricted unlearning is more powerful than DP-based unlearning: by using the actual unlearning set and side information, retrain-from-scratch achieves superior excess risk to DP when $\varepsilon \leq d$ and our approximate unlearning algorithm can achieve an exponentially smaller unlearning penalty when $\varepsilon \gg d$.

Other optimization settings. Liu et al. (2023) study certified unlearning for minimax models and derive generalization rates and unlearning-capacity bounds for convex-concave and strongly convex-strongly concave settings. Zou et al. (2025) study certified machine unlearning in proportional high-dimensional regimes, analyzing Newton-style procedures when the model dimension and sample size grow together. These works address different optimization or asymptotic settings and are complementary to our minimax analysis for smooth strongly convex stochastic optimization.

Broader unlearning literature. There is also a large empirical and systems-oriented literature on machine unlearning for neural networks, graphs, federated learning, recommendation systems, and foundation models. These works address important practical settings and evaluation questions,

but typically do not provide excess risk bounds under certified ε-unlearning. We therefore focus our theoretical comparisons on certified unlearning methods and DP-based baselines with formal guarantees.

B Proofs for Section 2

In this section, we provide proofs for results in Section 2 not included in the main text.

Proof of Proposition 1. Unlearning is preserved by postprocessing, so we may exclude the final projection onto $\mathcal{W}$ from our analysis. Now, fix $U \subseteq Z \in \mathcal{Z}^N$ with $n - m \leq N \leq n$ and $|Z \setminus U| \geq n - m$. The distribution of $\tilde{\mathcal{A}}(\emptyset, Z \setminus U)$ before projection is

$$W_{\emptyset, Z\setminus U} = (1-\eta)\mathrm{Unif}(\underbrace{\tau\mathbb{B}_d + \tilde{w}_{Z\setminus U}}_{=:G}) + \eta\mathrm{Unif}(\underbrace{(\tau+r)\mathbb{B}_d + \tilde{w}_{Z\setminus U}}_{=:B})$$

whereas the distribution of running $\tilde{\mathcal{A}}(U, Z)$ before projection is

$$W_{U,Z} = (1-\eta)\mathrm{Unif}(\underbrace{\tau\mathbb{B}_d + \tilde{w}_{Z}}_{=:G'}) + \eta\mathrm{Unif}(\underbrace{(\tau+r)\mathbb{B}_d + \tilde{w}_{Z\setminus U}}_{=B}).$$

We just need to argue that the likelihood ratio between these distributions lies in $[e^{-\varepsilon}, e^{\varepsilon}]$. Indeed, by the standard stability bound for strongly convex ERM (c.f. Sekhari et al. (2021, Lemma 6)),

$$\|\tilde{w}_Z - \tilde{w}_{Z\setminus U}\| \leq \|\hat{w}_Z - \hat{w}_{Z\setminus U}\| + 2\rho \leq \frac{2L|U|}{\mu|Z\setminus U|} + 2\rho \leq \frac{2Lm}{\mu(n-m)} + 2\rho = r.$$

In particular, G' is contained in B and hence the likelihood ratio between these distributions differs only for w contained in either G or G', exclusively. In the first case, the ratio of probability density functions is

$$\frac{dW_{\emptyset,Z\setminus U}}{dW_{U,Z}}(w) = \frac{(1-\eta)/\mathrm{Vol}(G) + \eta/\mathrm{Vol}(B)}{\eta/\mathrm{Vol}(B)} = 1 + \left(\frac{1}{\eta} - 1\right)\frac{\mathrm{Vol}(B)}{\mathrm{Vol}(G)} = 1 + \left(\frac{1}{\eta} - 1\right)\left(\frac{\tau+r}{\tau}\right)^d = e^{\varepsilon}$$

by choice of η and, analogously, we have $\frac{dW_{\emptyset,Z\setminus U}}{dW_{U,Z}}(w) = e^{-\varepsilon}$ for $w \in G' \setminus G$. In particular, $W_{\emptyset,Z\setminus U}$ and $W_{U,Z}$ are $(\varepsilon, 0)$-indistinguishable, as desired. □

Proof of Proposition 2. Recall as in the proof of Proposition 1 that $\tilde{\mathcal{A}}(U, Z)$ is exactly the projection of

$$w \sim W = (1-\eta)\mathrm{Unif}(\tau\mathbb{B}_d + \tilde{w}_Z) + \eta\mathrm{Unif}((\tau+r)\mathbb{B}_d + \tilde{w}_{Z\setminus U})$$

onto the convex domain $\mathcal{W}$. But $\tilde{w}_Z \in \mathcal{W}$, so

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}}[\|\tilde{\mathcal{A}}(U,Z) - \tilde{w}_Z\|^2] \leq \mathop{\mathbb{E}}_{w\sim W}[\|w - \tilde{w}_Z\|^2] \leq (1-\eta)\tau^2 + \eta \max_{w\in(\tau+r)\mathbb{B}_d + \tilde{w}_{Z\setminus U}} \|w - \tilde{w}_Z\|^2.$$

Recalling that $n - m = \Omega(n)$, the first term is bounded by

$$\tau^2 = \left(r\min\left\{1, 2e^{-\varepsilon/(d+2)}\right\}\right)^2 \lesssim \frac{L^2}{\mu^2}\left(\left(\frac{m}{n-m}\right)^2 + \frac{1}{n}\right)e^{-2\varepsilon/(d+2)} \lesssim \frac{L^2}{\mu^2}\left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)} + \frac{L^2}{\mu^2 n}.$$

As for the second term, recalling that $\|\tilde{w}_{Z\setminus U} - \tilde{w}_Z\| \leq r$ as well as our choice of $\tau \leq r$, we have

$$\|w - \tilde{w}_Z\|^2 \lesssim \|w - \tilde{w}_{Z\setminus U}\|^2 + \|\tilde{w}_{Z\setminus U} - \tilde{w}_Z\|^2 \leq (\tau+r)^2 + 2r^2 \lesssim r^2 \lesssim \frac{L^2}{\mu^2}\left(\frac{m}{n}\right)^2 + \frac{L^2}{\mu^2 n}$$

for any $w \in (\tau+r)\mathbb{B}_d + \tilde{w}_{Z\setminus U}$. Moreover, if $2e^{-\varepsilon/(d+2)} > 1$, then $\eta \leq 1 = 1^2 \lesssim e^{-2\varepsilon/(d+2)}$. More importantly, if $2e^{-\varepsilon/(d+2)} \leq 1$, then clearly $1 \leq \frac{1}{2}e^{\varepsilon}$ and $\tau = 2re^{-\varepsilon/(d+2)}$ by choice of τ, so it follows that

$$\eta = \frac{1}{(e^{\varepsilon}-1)\left(\frac{\tau}{\tau+r}\right)^d + 1} \lesssim e^{-\varepsilon}\left(\frac{2r}{\tau}\right)^d = e^{-\varepsilon}\left(e^{\varepsilon/(d+2)}\right)^d = e^{-\varepsilon\left(1-\frac{d}{d+2}\right)} = e^{-2\varepsilon/(d+2)}$$

in this case as well. The result now follows by combining these bounds. □

Proof of Theorem 1. **Unlearning.** The unlearning claim is exactly Proposition 1.

Runtime. Uniform sampling from d-dimensional balls of the form $a\mathbb{B}_d + b$ can be implemented in $O(d)$ time by noticing that

$$G \sim N(0,1)^d, T \sim \mathrm{Unif}([0,1]) \implies T^{1/d} \cdot \frac{G}{\|G\|} \sim \mathrm{Unif}(\mathbb{B}_d).$$

Furthermore, by choice of $\rho = \frac{L}{\mu\sqrt{n}}$ Lemma 2 ensures we may compute $\tilde{w}_T$ and $\tilde{w}_S$ with the desired precision using $O(\kappa \cdot n \cdot \log(D/\rho)) = \tilde{O}(\kappa \cdot n)$ gradient and projection queries.

Excess risk. Fix any unlearning request strategy $U(Z) \subseteq Z$ with $|Z \setminus U(Z)| \geq n - m$ as well as $n - m \leq N \leq n$ and consider smooth strongly convex SCO. By smoothness of F_P and the assumption $\nabla F_P(w_P^*) = 0$,

$$F_P(w) - F_P^* \leq \frac{\beta}{2}\|w - w_P^*\|^2.$$

Applying $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ with $w = \tilde{\mathcal{A}}(U(Z), Z)$, we get

$$F_P(\tilde{\mathcal{A}}(U(Z),Z)) - F_P^* \lesssim \beta\|\tilde{\mathcal{A}}(U(Z),Z) - \tilde{w}_Z\|^2 + \beta\|\hat{w}_Z - w_P^*\|^2 + \beta\|\tilde{w}_Z - \hat{w}_Z\|^2.$$

In particular, Proposition 2 and Lemma 1 together yield

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}, Z\sim P^N}\left[F_P(\tilde{\mathcal{A}}(U(Z),Z)) - F_P^*\right] = O\left(\kappa\frac{L^2}{\mu}\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}\right)\right),$$

where we used $\beta/\mu = \kappa$. This is the claimed SCO bound.

Similarly, for mean estimation over $\mathbb{B}_d$, $w_P^* = \mu_P := \mathbb{E}_P[z]$ and $F_P(w) - F_P^* = \frac{1}{2}\|w - \mu_P\|^2$, so it follows that

$$F_P(\tilde{\mathcal{A}}(U(Z),Z)) - F_P^* \lesssim \|\tilde{\mathcal{A}}(U(Z),Z) - \tilde{w}_Z\|^2 + \|\hat{w}_Z - \mu_P\|^2 + \|\tilde{w}_Z - \hat{w}_Z\|^2$$

Moreover, $\hat{w}_Z = \frac{1}{N}\sum_{i=1}^N z_i$, so $\mathbb{E}[\|\hat{w}_Z - \mu_P\|^2] \leq \frac{1}{N} \lesssim \frac{1}{n}$ and, combining with Proposition 2, we get the desired excess unlearning risk

$$O\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/(d+2)}\right). \qquad \square$$

C Matching Bounds for Unlearning Under the Weaker Utility Assumption

In this section, we consider the effect of replacing the expected excess unlearning risk measure in Definition 2 with the weaker definition seen in prior work, which we call expected excess *post-unlearning* risk.

Definition 3. *We say that $\tilde{\mathcal{A}}$ achieves expected excess post-unlearning risk α if for every distribution $P \in \Delta(\mathcal{Z})$ and any adversarial unlearning request $U(Z) \subseteq Z$ of size $|U(Z)| \leq m$, we have*

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}, Z\sim P^n}\left[F_P(\tilde{\mathcal{A}}(U(Z),Z)) - F_P^*\right] \leq \alpha.$$

C.1 Upper Bound for Post-Unlearning

Theorem 3 (Post-Unlearning Upper Bound)**.** *Grant Assumption 1 and assume $d \geq 2$. Then the unlearner Algorithm 2 satisfies ε-unlearning and requires $O(d)$ time for randomization as well as $\tilde{O}(\kappa \cdot n)$ gradient and projection queries. Moreover, Algorithm 2 achieves expected excess post-unlearning risk*

$$\alpha = O\left(\kappa\frac{L^2}{\mu}\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/d}\right)\right).$$

In the special case of mean estimation over the unit ball, i.e. $\mathcal{W} = \mathcal{Z} = \mathbb{B}_d$ and $f(w,z) = \frac{1}{2}\|w-z\|^2$, Algorithm 2 achieves expected excess post-unlearning risk

$$\alpha = O\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/d}\right).$$

Algorithm 2 Core-diffusion ε-unlearning for ERM

Require: Dataset T; deletion set $U \subseteq T$; parameters $\mathcal{W}, L, \mu, n, m, d, \varepsilon$

1: Set $S \leftarrow T \setminus U$, $\rho \leftarrow \frac{L}{\mu\sqrt{n}}$, $r \leftarrow \frac{2Lm}{\mu n} + 2\rho$, $\tau \leftarrow r \min\{1, 2e^{-\varepsilon/d}\}$, $K \leftarrow \left(\frac{\tau+r}{\tau}\right)^d$, and $\eta \leftarrow \max\{e^{-\varepsilon}, \frac{K-e^{\varepsilon}}{K-1}\}$

2: Compute deterministic $\tilde{w}_T$ and $\tilde{w}_S$ such that $\|\tilde{w}_T - \hat{w}_T\|, \|\tilde{w}_S - \hat{w}_S\| \leq \rho$

3: **if** $U = \emptyset$ and $|T| < n$ **then**

4: Sample $w \sim \mathrm{Unif}((\tau + r)\mathbb{B}_d + \tilde{w}_S)$

5: **else**

6: Sample $w \sim \begin{cases} \mathrm{Unif}((\tau + r)\mathbb{B}_d + \tilde{w}_S) & \text{w.p. } \eta \\ \mathrm{Unif}(\tau\mathbb{B}_d + \tilde{w}_T) & \text{w.p. } 1 - \eta \end{cases}$

7: **return** $\mathrm{argmin}_{w' \in \mathcal{W}} \|w' - w\|$

Critically, we note that Algorithm 2 inspects the sizes of the dataset and the unlearning request to determine whether it is performing a dry or a true unlearning run. Although this algorithm achieves the optimal rate for the post-unlearning utility model, we suggest that this algorithm's behaviour goes against the spirit of machine unlearning. Therefore we advocate replacing the post-unlearning utility model Definition 3 with the stronger utility model Definition 2.

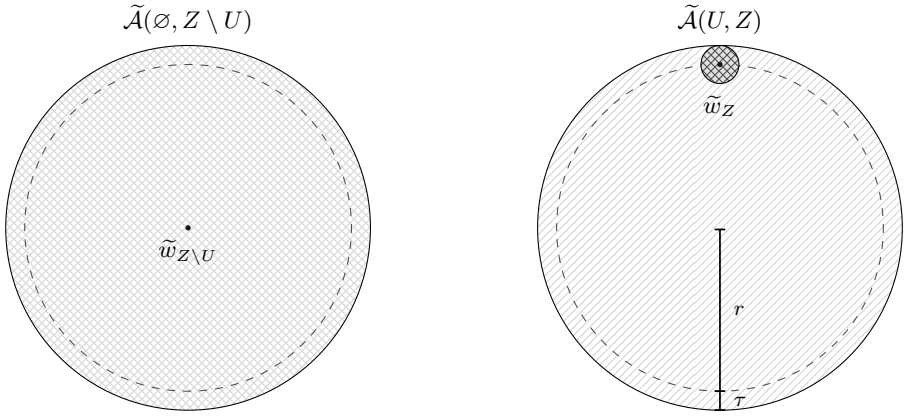

Figure 2: We compare the distribution of $\tilde{\mathcal{A}}(\emptyset, Z \setminus U)$ to $\tilde{\mathcal{A}}(U, Z)$ for the unlearner Algorithm 2. Lightly shaded regions are sampled uniformly with low probability and heavily shaded regions are sampled uniformly with high probability. In this case, $\tilde{\mathcal{A}}(\emptyset, Z \setminus U)$ returns a noisy solution centered around the RFS ERM solution $\hat{w}_{Z\setminus U}$. $\tilde{\mathcal{A}}(U, Z)$ has a similar distribution except some probability mass is transported into a small ball around the full ERM solution $\hat{w}_Z$.

Proposition 3. *The unlearner Algorithm 2 satisfies ε-unlearning.*

Proof. Unlearning is preserved by postprocessing, so we may exclude the final projection onto $\mathcal{W}$ from our analysis. Now, fix $U \subseteq Z \in \mathcal{Z}^n$, $|U| \leq m$. If $U = \emptyset$, then clearly $\tilde{\mathcal{A}}(U, Z) \equiv \tilde{\mathcal{A}}(\emptyset, Z \setminus U)$, so assume $U \neq \emptyset$. In this case, the distribution of $\tilde{\mathcal{A}}(\emptyset, Z \setminus U)$ before projection is

$$W_{\emptyset, Z\setminus U} = \mathrm{Unif}(\underbrace{(\tau + r)\mathbb{B}_d + \tilde{w}_{Z\setminus U}}_{=:B})$$

whereas the distribution of running $\tilde{\mathcal{A}}(U, Z)$ before projection is

$$W_{U,Z} = (1 - \eta)\mathrm{Unif}(\underbrace{\tau\mathbb{B}_d + \tilde{w}_Z}_{=:G'}) + \eta\mathrm{Unif}(\underbrace{(\tau + r)\mathbb{B}_d + \tilde{w}_{Z\setminus U}}_{=B}).$$

We just need to argue that the likelihood ratio between these distributions lies in $[e^{-\varepsilon}, e^{\varepsilon}]$. Indeed, by the standard stability bound for strongly convex ERM (c.f. Sekhari et al. (2021, Lemma 6)),

$$\|\tilde{w}_Z - \tilde{w}_{Z\setminus U}\| \leq \|\hat{w}_Z - \hat{w}_{Z\setminus U}\| + 2\rho \leq \frac{2Lm}{\mu n} + 2\rho = r.$$

In particular, G' is contained in B and hence the likelihood ratio between these distributions differs only for w contained in either G' or B, exclusively. In the first case, if $w \in G'$, then

$$\frac{dW_{\emptyset,Z\setminus U}}{dW_{U,Z}}(w) = \frac{1/\mathrm{Vol}(B)}{(1-\eta)/\mathrm{Vol}(G) + \eta/\mathrm{Vol}(B)} = \frac{1}{(1-\eta)K + \eta} = \frac{1}{K - \eta(K-1)} \in [e^{-\varepsilon}, 1]$$

since $\frac{K-e^{\varepsilon}}{K-1} \le \eta \le 1$. In the latter case, if $w \in B \setminus G'$, we have

$$\frac{dW_{\emptyset,Z\setminus U}}{dW_{U,Z}}(w) = \frac{1/\mathrm{Vol}(B)}{\eta/\mathrm{Vol}(B)} = \frac{1}{\eta} \in [1, e^{\varepsilon}]$$

since $e^{-\varepsilon} \le \eta \le 1$. In particular, $W_{\emptyset,Z\setminus U}$ and $W_{U,Z}$ are $(\varepsilon, 0)$-indistinguishable, as desired. □

Proposition 4. *Assume $d \ge 2$. For every dataset $Z \in \mathcal{Z}^n$ and every deletion set $U \subseteq Z$ with $|U| \le m$, the output $\tilde{\mathcal{A}}(U, Z)$ of Algorithm 2 satisfies*

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}}\left[\left\|\tilde{\mathcal{A}}(U,Z) - \tilde{w}_Z\right\|^2\right] = O\left(\frac{L^2}{\mu^2}\left(\frac{m}{n}\right)^2 e^{-2\varepsilon/d} + \frac{L^2}{\mu^2 n}\right).$$

Proof. Recall as in the proof of Proposition 3 that $\tilde{\mathcal{A}}(U, Z)$ is exactly the projection of

$$w \sim W = (1-\eta)\mathrm{Unif}(\tau\mathbb{B}_d + \tilde{w}_Z) + \eta\mathrm{Unif}((\tau + r)\mathbb{B}_d + \tilde{w}_{Z\setminus U})$$

onto the convex domain $\mathcal{W}$. But $\tilde{w}_Z \in \mathcal{W}$, so

$$\mathop{\mathbb{E}}_{\tilde{\mathcal{A}}}[\|\tilde{\mathcal{A}}(U,Z) - \tilde{w}_Z\|^2] \le \mathop{\mathbb{E}}_{w\sim W}[\|w - \tilde{w}_Z\|^2] \le (1-\eta)\tau^2 + \eta \max_{w \in (\tau+r)\mathbb{B}_d + \tilde{w}_{Z\setminus U}} \|w - \tilde{w}_Z\|^2.$$

The first term is bounded by

$$\tau^2 = \left(r\min\left\{1, 2e^{-\varepsilon/d}\right\}\right)^2 \lesssim \frac{L^2}{\mu^2}\left(\frac{m}{n}\right)^2 e^{-2\varepsilon/d} + \frac{L^2}{\mu^2 n}.$$

As for the second term, recalling that $\|\tilde{w}_{Z\setminus U} - \tilde{w}_Z\| \le r$ as well as our choice of $\tau \le r$, we have

$$\|w - \tilde{w}_Z\|^2 \lesssim \|w - \tilde{w}_{Z\setminus U}\|^2 + \|\tilde{w}_{Z\setminus U} - \tilde{w}_Z\|^2 \le (\tau + r)^2 + r^2 \lesssim r^2 \lesssim \frac{L^2}{\mu^2}\left(\frac{m}{n}\right)^2 + \frac{L^2}{\mu^2 n}$$

for any $w \in (\tau + r)\mathbb{B}_d + \hat{w}_{Z\setminus U}$. Moreover, if $2e^{-\varepsilon/d} > 1$, then $\eta \le 1 = 1^2 \lesssim e^{-2\varepsilon/d}$. More importantly, if $2e^{-\varepsilon/d} \le 1$, then $\tau = 2re^{-\varepsilon/d}$ by choice of τ, in which case

$$K = \left(\frac{\tau + r}{\tau}\right)^d \le \left(\frac{2r}{\tau}\right)^d = \left(\frac{2r}{2re^{-\varepsilon/d}}\right)^d = e^{\varepsilon},$$

and hence

$$\eta = \max\left\{e^{-\varepsilon}, \frac{K - e^{\varepsilon}}{K-1}\right\} = e^{-\varepsilon} \le e^{-2\varepsilon/d}$$

in this case as well. The result now follows by combining these bounds. □

As in the proof of Theorem 1, Theorem 3 now follows by combining Propositions 3 and 4 with Lemmas 1 and 2.

C.2 Lower Bounds for (ε, δ)-Post-Unlearning

In this section we prove a lower bound matching the rate of Algorithm 2. At its core, this reduces to showing a mean estimation lower bound.

Theorem 4. *Let $\tilde{\mathcal{A}}(U, Z)$ be an (ε, δ)-unlearning algorithm with $\delta \leq 1/7$ and suppose that, for all distributions P on $\mathbb{B}_d$ with mean μ_P and unlearning requests $U(Z) \subseteq Z$ with $|U(Z)| \leq m$, we have the mean squared error guarantee*

$$\mathop{\mathbb{P}}_{\tilde{\mathcal{A}}, Z\sim P^n}\left(\|\tilde{\mathcal{A}}(U(Z), Z) - \mu_P\|^2 \leq \alpha\right) \geq \frac{2}{3}.$$

Then, it must be the case that

$$\alpha = \Omega\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/d}\right).$$

By Markov's inequality, the same lower bound applies up to constants for any algorithm with *expected* mean squared error $O(\alpha)$. Moreover, as in Section 3 the sampling error lower bound $\Omega(1/n)$ holds even without unlearning requirements, so it just remains to derive the second term. As in Section 3, this implies the following SCO lower bound:

Corollary 2. *Suppose the loss satisfies Assumption 1 and that there is an (ε, δ)-unlearning algorithm $\tilde{\mathcal{A}}$ with $\delta \leq 1/7$ and excess post-unlearning risk α. Then we must have*

$$\alpha \geq \Omega\left(\frac{L^2}{\mu}\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/d}\right)\right)$$

The argument for Theorem 4 is very similar to that of Section 3, except that we rely on a weaker form of the packing technique.

Lemma 6. *Let $P^*, P_1, \ldots, P_k$ be distributions over a common space such that we can find disjoint events $E_1, \ldots, E_k$ for which $P_i(E_i) \geq p$ as well as $P_i \approx_{\varepsilon,\delta} P^*$ for each $i \in \{1, \ldots, k\}$. Then*

$$e^\varepsilon \geq k(p - \delta).$$

Proof. Indeed, by disjointness, we have

$$1 \geq \sum_{i=1}^k P^*(E_i) \geq \sum_{i=1}^k e^{-\varepsilon}(P_i(E_i) - \delta) \geq \sum_{i=1}^k e^{-\varepsilon}(p - \delta) = ke^{-\varepsilon}(p - \delta). \qquad \square$$

Proof of Theorem 4. Set $\eta := \frac{m}{5n}$. If $\sqrt{\alpha} \geq \frac{1}{2}\eta$, we are done, so assume that $\sqrt{\alpha} < \frac{1}{2}\eta$ and set $\tau := \frac{2\sqrt{\alpha}}{\eta} < 1$. By Lemma 3, we can find a τ-separated $\mathcal{V} \subseteq \mathbb{B}_d$ of size

$$|\mathcal{V}| \geq \frac{1}{2}\left(\frac{\eta}{2\sqrt{\alpha}}\right)^d.$$

Now, for each $v \in \mathcal{V}$, recall the hard distribution $P_{v,\eta}$ and corresponding delete request $U_v(Z)$ as in Lemma 4. As in the proof of Theorem 2, we have that $\tilde{\mathcal{A}}(U_v(Z), Z)|_{Z\sim P_{v,\eta}^n} \approx_{\varepsilon,\delta+2^{-m}} \tilde{\mathcal{A}}(\emptyset, \mathbf{0}_{n-m})$ because $\tilde{\mathcal{A}}$ is (ε, δ)-unlearning.

On the other hand, by our assumption on the mean estimation error, for every $v \in \mathcal{V}$ we have

$$\frac{2}{3} \leq \mathop{\mathbb{P}}_{\tilde{\mathcal{A}}, Z\sim P_{v,\eta}^n}\left(\|\tilde{\mathcal{A}}(U_v(Z), Z) - \eta v\| \leq \sqrt{\alpha}\right).$$

As the $v \in \mathcal{V}$ are $\frac{2\sqrt{\alpha}}{\eta}$-separated, the $\{w : \|w - \eta v\| \leq \sqrt{\alpha}\}$ are disjoint and hence Lemma 6 yields

$$e^\varepsilon \geq |\mathcal{V}|\left(\frac{2}{3} - \delta - 2^{-m}\right) \geq \frac{1}{42}\left(\frac{\eta}{2\sqrt{\alpha}}\right)^d \geq \left(\frac{\eta}{84\sqrt{\alpha}}\right)^d = \left(\frac{m}{420n\sqrt{\alpha}}\right)^d.$$

In particular, we have $\alpha \gtrsim \left(\frac{m}{n}\right)^2 e^{-2\varepsilon/d}$. $\square$

D RETRAIN-FROM-SCRATCH UPPER BOUND

We first record the performance of the most basic exact-unlearning algorithm: retraining from scratch. Given an unlearning request $U \subseteq Z$, define

$$\tilde{\mathcal{A}}_{\mathrm{RFS}}(U, Z) := \hat{w}_{Z\setminus U} \in \mathrm{argmin}_{w\in\mathcal{W}} \widehat{F}_{Z\setminus U}(w),$$

where

$$\widehat{F}_{Z\setminus U}(w) := \frac{1}{|Z \setminus U|} \sum_{z_i \in Z\setminus U} f(w, z_i).$$

This algorithm ignores the original model and recomputes an empirical risk minimizer on the retained data. Its main drawback is that it requires storing the full dataset.

Theorem 5 (Retrain-from-scratch upper bound)**.** *Let f satisfy Assumption 1 and let $\kappa = \beta/\mu$. Then retraining from scratch is exact 0-unlearning. Moreover,*

$$\mathop{\mathbb{E}}_{Z\sim P^n}\left[\max_{U\subseteq Z:\,|U|\le m}\left(F_P(\hat{w}_{Z\setminus U}) - F_P^*\right)\right] \lesssim \kappa\frac{L^2}{\mu}\left(\frac{1}{n} + \left(\frac{m}{n}\right)^2\right).$$

Proof. The exact-unlearning guarantee is immediate. After receiving U, the algorithm outputs $\hat{w}_{Z\setminus U}$, which is exactly the output obtained by running the learning algorithm directly on the retained dataset $Z \setminus U$. Hence the two output distributions are identical.

It remains to prove the excess-risk bound. Let

$$\hat{w}_Z \in \mathrm{argmin}_{w\in\mathcal{W}} \widehat{F}_Z(w), \qquad w^* \in \mathrm{argmin}_{w\in\mathcal{W}} F_P(w).$$

By β-smoothness and the stationarity assumption, we have

$$F_P(w) - F_P^* \le \frac{\beta}{2}\|w - w^*\|^2 \qquad \forall w \in \mathcal{W}.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\max_{U\subseteq Z:\,|U|\le m}\left(F_P(\hat{w}_{Z\setminus U}) - F_P^*\right)\right] \\
&\quad\le \frac{\beta}{2}\,\mathbb{E}\left[\max_{U\subseteq Z:\,|U|\le m}\|\hat{w}_{Z\setminus U} - w^*\|^2\right] \\
&\quad\le \beta\,\mathbb{E}\,\|\hat{w}_Z - w^*\|^2 + \beta\,\mathbb{E}\left[\max_{U\subseteq Z:\,|U|\le m}\|\hat{w}_{Z\setminus U} - \hat{w}_Z\|^2\right], \qquad (3)
\end{aligned}$$

where the last step uses $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$.

We bound the two terms in (3). First, standard stability/generalization bounds for L-Lipschitz, μ-strongly convex ERM imply

$$\mathbb{E}[F_P(\hat{w}_Z) - F_P^*] \lesssim \frac{L^2}{\mu n}.$$

By μ-strong convexity of F,

$$\frac{\mu}{2}\,\mathbb{E}\,\|\hat{w}_Z - w^*\|^2 \le \mathbb{E}[F_P(\hat{w}_Z) - F_P^*],$$

and hence

$$\mathbb{E}\,\|\hat{w}_Z - w^*\|^2 \lesssim \frac{L^2}{\mu^2 n}. \qquad (4)$$

Second, the unlearning stability bound for strongly convex ERM gives, uniformly over all $U \subseteq Z$ with $|U| \le m$,

$$\|\hat{w}_{Z\setminus U} - \hat{w}_Z\| \lesssim \frac{Lm}{\mu n}.$$

For example, this is precisely the unlearning stability estimate used in Sekhari et al. (2021, Lemma 6). Therefore,

$$\mathbb{E}\left[\max_{U\subseteq Z:\,|U|\le m}\|\hat{w}_{Z\setminus U} - \hat{w}_Z\|^2\right] \lesssim \frac{L^2m^2}{\mu^2n^2}. \qquad (5)$$

Combining (3), (4), and (5) yields

$$\mathbb{E}\left[\max_{U\subseteq Z:\,|U|\le m}\left(F_P(\hat{w}_{Z\setminus U})-F_P^*\right)\right]\lesssim \beta\frac{L^2}{\mu^2}\left(\frac{1}{n}+\left(\frac{m}{n}\right)^2\right).$$

Since $\beta/\mu=\kappa$, this is

$$\lesssim \kappa\frac{L^2}{\mu}\left(\frac{1}{n}+\left(\frac{m}{n}\right)^2\right),$$

as claimed. □

Remark 6. *The proof uses only Lipschitzness, strong convexity, smoothness on the constrained domain, as well as stationarity of the minimizer. Under these assumptions, the condition-number factor multiplies both the usual statistical term and the unlearning term. Removing the factor κ from the $1/n$ term would require an additional condition such as a smooth self-bounding inequality.*

E Independent analysis of the ERM warm-start algorithm of Allouah et al. (2025)

In this section we revisit the ERM-based warm-start unlearning algorithm of Allouah et al. (2025). At a high level, their algorithm is very close to retraining from scratch: after receiving an unlearning request, it approximately minimizes the empirical risk on the retained dataset, using the original trained model as a warm start, and then adds noise to certify unlearning. The original published population-risk analysis contained a gap, which the authors subsequently corrected in an updated preprint (Allouah et al., 2026). We give an independent utility analysis showing that this approach achieves a population-risk bound matching retraining from scratch, up to the optimization error and the noise variance.

We use our notation rather than the notation of Allouah et al. (2025). Let $Z=(z_1,\dots,z_n)\sim P^n$, and for an unlearning set $U\subseteq Z$, let

$$\widehat{F}_{Z\setminus U}(w):=\frac{1}{|Z\setminus U|}\sum_{z_i\in Z\setminus U} f(w,z_i),\qquad \hat{w}_{Z\setminus U}\in \operatorname{argmin}_{w\in\mathcal{W}}\widehat{F}_{Z\setminus U}(w).$$

The algorithm. The learner first computes an empirical risk minimizer $\hat{w}_Z$ on the full dataset and stores the data Z. Upon receiving an unlearning request U, the unlearning algorithm runs an optimization method, initialized at $\hat{w}_Z$, on the retained empirical objective $\widehat{F}_{Z\setminus U}$. Let w_U denote the resulting approximate retained-data ERM. Finally, the algorithm outputs

$$\tilde{w}_U:=\Pi_{\mathcal{W}}(w_U+v), \tag{6}$$

where v is the noise used to certify unlearning. When $w_U=\hat{w}_{Z\setminus U}$ and $v=0$, this is exactly retraining from scratch.

The theorem below isolates the utility guarantee of this template. It applies to the warm-start algorithm of Allouah et al. (2025) once their optimization accuracy and noise calibration are substituted.

Theorem 7 (Utility bound for ERM warm-start unlearning)**.** *Let f satisfy Assumption 1 and let $\kappa=\beta/\mu$. Suppose that for every unlearning set $U\subseteq Z$ with $|U|\le m$, the unlearning-time optimizer returns $w_U\in\mathcal{W}$ satisfying*

$$\widehat{F}_{Z\setminus U}(w_U)-\widehat{F}_{Z\setminus U}(\hat{w}_{Z\setminus U})\le\gamma. \tag{7}$$

Assume also that the noise v in (6) *satisfies*

$$\mathbb{E}\,\|v\|^2\le\sigma^2.$$

Then

$$\mathop{\mathbb{E}}_{Z\sim P^n}\left[\sup_{U\subseteq Z:\,|U|\le m}\mathbb{E}[F_P(\tilde{w}_U)-F_P^*\mid Z,U]\right]\lesssim\kappa\frac{L^2}{\mu}\left(\frac{1}{n}+\left(\frac{m}{n}\right)^2\right)+\frac{\beta}{\mu}\gamma+\beta\sigma^2.$$

In particular, if

$$\gamma\lesssim\frac{L^2}{\mu}\left(\frac{1}{n}+\left(\frac{m}{n}\right)^2\right)\qquad\text{and}\qquad\sigma^2\lesssim\frac{L^2}{\mu^2}\left(\frac{1}{n}+\left(\frac{m}{n}\right)^2\right),$$

then the algorithm achieves the retraining from scratch rate

$$\mathop{\mathbb{E}}_{Z\sim P^n}\left[\sup_{U\subseteq Z:\,|U|\le m}\mathbb{E}[F_P(\tilde{w}_U)-F_P^*\mid Z,U]\right]\lesssim \kappa\frac{L^2}{\mu}\left(\frac{1}{n}+\left(\frac{m}{n}\right)^2\right).$$

Proof. Fix Z and $U\subseteq Z$ with $|U|\le m$. Let

$$w^*\in \operatorname{argmin}_{w\in\mathcal{W}}F_P(w).$$

By β-smoothness and the stationarity assumption,

$$F_P(w)-F_P^*\le\frac{\beta}{2}\|w-w^*\|^2\qquad\forall w\in\mathcal{W}. \tag{8}$$

Using nonexpansiveness of projection and the inequality $\|a_1+\cdots+a_4\|^2\le 4\sum_{j=1}^4\|a_j\|^2$, we obtain

$$\begin{aligned}\mathbb{E}[F_P(\tilde{w}_U)-F_P^*\mid Z,U]&\le\frac{\beta}{2}\,\mathbb{E}\big[\|\tilde{w}_U-w^*\|^2\mid Z,U\big]\\&\lesssim\beta\|\hat{w}_Z-w^*\|^2+\beta\|\hat{w}_{Z\setminus U}-\hat{w}_Z\|^2\\&\quad+\beta\|w_U-\hat{w}_{Z\setminus U}\|^2+\beta\,\mathbb{E}\,\|v\|^2.\end{aligned}\tag{9}$$

We now bound each term. First, as in the proof of Theorem 5, standard stability/generalization bounds for L-Lipschitz, μ-strongly convex ERM imply

$$\mathop{\mathbb{E}}_{Z}\|\hat{w}_Z-w^*\|^2\lesssim\frac{L^2}{\mu^2 n}.$$

Second, the unlearning stability bound for strongly convex ERM gives, uniformly over all $U\subseteq Z$ with $|U|\le m$,

$$\|\hat{w}_{Z\setminus U}-\hat{w}_Z\|\lesssim\frac{Lm}{\mu n}.$$

Third, by μ-strong convexity of $\widehat{F}_{Z\setminus U}$ and the empirical accuracy condition (7),

$$\frac{\mu}{2}\|w_U-\hat{w}_{Z\setminus U}\|^2\le\widehat{F}_{Z\setminus U}(w_U)-\widehat{F}_{Z\setminus U}(\hat{w}_{Z\setminus U})\le\gamma,$$

and hence

$$\|w_U-\hat{w}_{Z\setminus U}\|^2\le\frac{2\gamma}{\mu}.$$

Finally, $\mathbb{E}\,\|v\|^2\le\sigma^2$ by assumption.

Substituting these bounds into (9), taking the supremum over U, and then taking expectation over Z, gives

$$\mathop{\mathbb{E}}_{Z}\left[\sup_{U\subseteq Z:\,|U|\le m}\mathbb{E}[F_P(\tilde{w}_U)-F_P^*\mid Z,U]\right]\lesssim\beta\frac{L^2}{\mu^2}\left(\frac{1}{n}+\left(\frac{m}{n}\right)^2\right)+\frac{\beta}{\mu}\gamma+\beta\sigma^2.$$

Since $\beta/\mu=\kappa$, the claimed bound follows. □

Gap in the original ICLR 2025 analysis. The proof of Allouah et al. (2025, Proposition 1) attempts to convert an empirical-risk guarantee on the retained sample into a population-risk guarantee. The key error occurs in the line after the authors invoke smoothness of the loss. In our notation, smoothness can only give

$$F_P(w)-F_P(\theta_S^*)\le\langle\nabla F_P(\theta_S^*),w-\theta_S^*\rangle+\frac{\beta}{2}\|w-\theta_S^*\|^2,$$

where θ_S^* is the empirical minimizer used in their argument. The proof then effectively treats the linear term

$$\langle\nabla F_P(\theta_S^*),w-\theta_S^*\rangle$$

as zero. This is not justified: θ_S^* minimizes the empirical risk, not the population risk, so in general

$$\nabla F_P(\theta_S^*) \neq 0.$$

Moreover, in constrained optimization, even the population minimizer need not have zero gradient. Thus the omitted first-order term can dominate the claimed bound, and the original proof does not establish the stated population-risk guarantee. The authors subsequently corrected this argument in Allouah et al. (2026).

Our independent analysis above avoids this step. Instead of expanding the population risk around an empirical minimizer and dropping the first-order term, we compare the warm-start output to the exact retained-data ERM, use strong convexity to convert empirical optimization error into parameter error, and then apply the same stability argument as retraining from scratch. This yields a retraining-level rate, plus the explicit contributions of optimization error and unlearning noise.

F A Sharper Analysis of the Newton-step Algorithm of Sekhari et al. (2021)

We revisit the Newton-step unlearning algorithm of Sekhari et al. (2021). Their algorithm was introduced as a way to improve over generic differentially private baselines for smooth strongly convex losses. In this section, we show that the same algorithm admits a sharper population-risk analysis than the one originally given.

Throughout this section, assume that for every $z \in \mathcal{Z}$, the loss $f(\cdot, z)$ is L-Lipschitz, μ-strongly convex, and β-smooth over $\mathcal{W}$. We write $\kappa = \beta/\mu$. We additionally assume, as in Sekhari et al. (2021), that the Hessian is M-Lipschitz:

$$\|\nabla^2 f(w, z) - \nabla^2 f(w', z)\| \leq M\|w - w'\| \qquad \forall w, w' \in \mathcal{W},\ z \in \mathcal{Z}. \tag{10}$$

For a dataset $Z = (z_1, \ldots, z_n)$, write

$$\widehat{F}_Z(w) := \frac{1}{n}\sum_{i=1}^{n} f(w, z_i), \qquad \hat{w}_Z \in \operatorname{argmin}_{w\in\mathcal{W}} \widehat{F}_Z(w).$$

The algorithm of Sekhari et al. (2021) initializes at the full-data ERM $\hat{w}_Z$. After receiving an unlearning set $U \subseteq Z$, it takes one Newton step with respect to the retained empirical objective. Equivalently, the deterministic part of the update can be written as

$$\bar{w}_U := \hat{w}_Z - \nabla^2 \widehat{F}_{Z\setminus U}(\hat{w}_Z)^{-1} \nabla \widehat{F}_{Z\setminus U}(\hat{w}_Z), \tag{11}$$

where

$$\widehat{F}_{Z\setminus U}(w) := \frac{1}{|Z \setminus U|} \sum_{z_i \in Z\setminus U} f(w, z_i).$$

The unlearning algorithm outputs a noisy version of this update,

$$\tilde{w}_U := \Pi_{\mathcal{W}}(\bar{w}_U + v), \tag{12}$$

where v is calibrated Gaussian noise for (ε, δ)-unlearning, or calibrated multivariate Laplace noise for pure ε-unlearning. The projection is post-processing and can only improve the distance-to-w^* bounds used below.

The original analysis of Sekhari et al. (2021) gives, up to constants,

$$\frac{L^2}{\mu}\frac{m}{n} + \frac{L^3 M}{\mu^3}\frac{m^2\sqrt{d\log(1/\delta)}}{\varepsilon n^2}$$

for the Gaussian-noise (ε, δ)-unlearning version when $m \leq n/2$. When $\delta = 0$, substituting Laplace noise for Gaussian in their algorithm yields the same bound but with $\sqrt{d\log(1/\delta)}$ replaced by d. We show that a more direct population-risk argument improves the generalization term from m/n to $1/n + (m/(n-m))^2$, which is $1/n + (m/n)^2$ when $m \leq n/2$, and squares the privacy contribution when converting parameter error to excess risk.

Theorem 8 (Improved analysis of the Newton-step algorithm)**.** *Let f satisfy Assumption 1 and additionally assume that $f(\cdot, z)$ has M-Lipschitz Hessian for every $z \in \mathcal{Z}$. Let $\tilde{w}_U$ be the Newton-step unlearning algorithm of Sekhari et al. (2021) for unlearning sets of size at most $m < n$.*

For the Gaussian-noise version calibrated to satisfy (ε, δ)-unlearning,

$$\sup_{U \subseteq Z:\, |U| \le m} \mathbb{E}[F_P(\tilde{w}_U) - F_P^*] \lesssim \kappa \frac{L^2}{\mu} \left(\frac{1}{n} + \left(\frac{m}{n-m} \right)^2 \right) + \beta \left(\frac{ML^2}{\mu^3} \right)^2 \frac{d\, m^4 \log(1/\delta)}{\varepsilon^2 n^4}.$$

For the pure ε-unlearning version obtained by replacing Gaussian noise with multivariate Laplace noise,

$$\sup_{U \subseteq Z:\, |U| \le m} \mathbb{E}[F_P(\tilde{w}_U) - F_P^*] \lesssim \kappa \frac{L^2}{\mu} \left(\frac{1}{n} + \left(\frac{m}{n-m} \right)^2 \right) + \beta \left(\frac{ML^2}{\mu^3} \right)^2 \frac{d^2 m^4}{\varepsilon^2 n^4}.$$

In particular, if $m \le n/2$, then the deterministic part of both bounds simplifies to

$$\kappa \frac{L^2}{\mu} \left(\frac{1}{n} + \left(\frac{m}{n} \right)^2 \right).$$

Proof. Fix any unlearning set $U \subseteq Z$ with $|U| \le m$. The bounds below are uniform over such U, so taking the supremum gives the theorem.

Let $w^* \in \operatorname{argmin}_{w \in \mathcal{W}} F_P(w)$. By β-smoothness and the stationarity assumption,

$$F_P(w) - F_P^* \le \frac{\beta}{2} \|w - w^*\|^2 \qquad \forall w \in \mathcal{W}. \tag{13}$$

Using nonexpansiveness of projection and $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\begin{aligned} \mathbb{E}[F_P(\tilde{w}_U) - F_P^*] &\le \frac{\beta}{2}\, \mathbb{E}\, \|\tilde{w}_U - w^*\|^2 \\ &\le \beta\, \mathbb{E}\, \|\bar{w}_U - \hat{w}_Z\|^2 + \beta\, \mathbb{E}\, \|\hat{w}_Z - w^*\|^2 + \beta\, \mathbb{E}\, \|v\|^2. \end{aligned} \tag{14}$$

We now bound the three terms in (14). First, by the definition of the Newton step (11),

$$\bar{w}_U - \hat{w}_Z = -\nabla^2 \widehat{F}_{Z \setminus U}(\hat{w}_Z)^{-1} \nabla \widehat{F}_{Z \setminus U}(\hat{w}_Z).$$

The proof of Sekhari et al. (2021, Lemma 9) gives, uniformly over all $U \subseteq Z$ with $|U| \le m$,

$$\|\bar{w}_U - \hat{w}_Z\| \lesssim \frac{Lm}{\mu(n-m)}. \tag{15}$$

Consequently,

$$\mathbb{E}\, \|\bar{w}_U - \hat{w}_Z\|^2 \lesssim \frac{L^2 m^2}{\mu^2 (n-m)^2}. \tag{16}$$

Second, standard stability/generalization bounds for L-Lipschitz, μ-strongly convex ERM imply

$$\mathbb{E}[F_P(\hat{w}_Z) - F_P^*] \lesssim \frac{L^2}{\mu n}. \tag{17}$$

By μ-strong convexity of F,

$$\frac{\mu}{2}\, \mathbb{E}\, \|\hat{w}_Z - w^*\|^2 \le \mathbb{E}[F_P(\hat{w}_Z) - F_P^*]\,.$$

Combining this with (17) gives

$$\mathbb{E}\, \|\hat{w}_Z - w^*\|^2 \lesssim \frac{L^2}{\mu^2 n}. \tag{18}$$

Third, Sekhari et al. (2021) show that the deterministic Newton-step map has unlearning sensitivity

$$\Delta \lesssim \frac{ML^2}{\mu^3} \frac{m^2}{n^2}. \tag{19}$$

Thus, for the Gaussian mechanism calibrated to (ε, δ)-unlearning,

$$\mathbb{E}\,\|v\|^2 \lesssim \frac{d\,\Delta^2 \log(1/\delta)}{\varepsilon^2} \lesssim \left(\frac{ML^2}{\mu^3}\right)^2 \frac{d\,m^4 \log(1/\delta)}{\varepsilon^2 n^4}. \tag{20}$$

For the pure ε-unlearning version based on multivariate Laplace noise,

$$\mathbb{E}\,\|v\|^2 \lesssim \frac{d^2\Delta^2}{\varepsilon^2} \lesssim \left(\frac{ML^2}{\mu^3}\right)^2 \frac{d^2 m^4}{\varepsilon^2 n^4}. \tag{21}$$

Plugging (16), (18), and either (20) or (21) into (14) gives

$$\mathbb{E}[F_P(\tilde{w}_U) - F_P^*] \lesssim \beta \frac{L^2}{\mu^2}\left(\frac{1}{n} + \left(\frac{m}{n-m}\right)^2\right) + \beta\,\mathbb{E}\,\|v\|^2.$$

Since $\beta L^2/\mu^2 = \kappa L^2/\mu$, the claimed bounds follow. If $m \le n/2$, then $(n-m)^{-1} \le 2n^{-1}$, giving the simplified deterministic term. □

Remark 9. *The improvement comes from analyzing population risk through the squared distance to the full-data ERM $\hat{w}_Z$. The original analysis effectively pays a first-order generalization term of order m/n. By instead using smoothness of F, the deterministic displacement of the Newton update contributes quadratically, giving $(m/(n-m))^2$, while the full-data ERM contributes the usual $1/n$ statistical term, up to the condition-number factor.*