REPASTA-CU: CERTIFIED UNLEARNING BEYOND GLOBAL SMOOTHNESS AND UNIFORM GRADIENT BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Certified machine unlearning seeks to remove the influence of deleted data without retraining from scratch. Existing first-order guarantees for nonconvex models typically impose global smoothness or uniformly bounded gradients, and some methods use a comparison target different from retraining on the retained data. We propose RePASTA-CU, a checkpoint-based method that keeps non-private training and uses the model obtained by retraining on the retained dataset as its comparison target. It stores a checkpoint after $T - K$ clipped SGD updates, applies K PASTA updates using only retained data after deletion, and adds Gaussian noise at release. Public clipping and projection bound the pre-release distance between the unlearned model and the retained-data retraining model; calibrating the release noise to this bound yields bidirectional (ε, δ)-certified unlearning without smoothness, Lipschitz continuity, weak convexity, or uniformly bounded gradients. Under weak convexity on a bounded convex domain and a quadratic-growth bound on the raw second moment of stochastic subgradients, we prove an $O(K^{-1/6})$ root-mean-square Moreau-stationarity rate before release noise, with a separate noise-dependent term for the released model. Experiments on synthetic, image, text, federated, and convex tasks compare RePASTA-CU with R2D and other certified-unlearning methods across utility, certificate scale, stability, and deletion cost.

1 INTRODUCTION

Machine unlearning aligns to legislation after model training (Cao & Yang, 2015; Ginart et al., 2019; Bourtoule et al., 2021). A deletion request should remove a record's influence from model parameters, while retraining from scratch is excessively computationally expensive. For a sample of size n, certified machine unlearning seeks a computationally efficient update whose output from the unlearning algorithm U is approximately indistinguishable from the output of the learning algorithm A trained on retained data, with certificate guaranteed by (ε, δ)-indistinguishability.

Existing certified unlearning requires assumptions on convexity, smoothness, curvature, or gradient bounds. Guo et al. (Guo et al., 2019) study regularized linear and logistic models under convex smooth losses, bounded features or gradients, and Lipschitz Hessians, and bound the Newton residual. Descent-to-Delete (Neel et al., 2021) uses projected gradient descent under convex or strongly convex smooth losses, with optimization and population risk bounds. Sekhari et al. (Sekhari et al., 2021) assume Lipschitz losses and smooth Hessians, with bounds for test risk, generalization, and supported deletions. Zhang et al. (Zhang et al., 2024) allow weak convexity for neural networks but retain smooth gradients and Hessians, bounded parameters, and local regularization, with a local Newton error bound. Chien et al. (Chien et al., 2024) use noisy dynamics but retain smooth losses together with bounded domains, clipping, or mixing conditions, with noisy update error bounds. R2D (Mu & Klabjan, 2026) is the weakest strict first-order black-box method with the learning algorithm A as fixed retraining reference. It removes convexity but retains global smoothness and uniformly bounded gradients. Its utility controls retained-data gradient norm, with stronger risk and generalization bounds under PL condition. Koloskova et al. (Koloskova et al., 2025) remove loss assumptions from clipping and noisy fine-tuning, but compare with a separate retained-data reference

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.

based mainly on the unlearning algorithm U, which yields a different certification guarantee. Their utility analysis describes the tradeoff among noise, iterations, and accuracy.

Open gap of the certified unlearning

Can certified unlearning remove the regularity, i.e., smoothness or gradient bounds?

PASTA (Fazla et al., 2026) provides the optimization tools for this gap. Its oracle analysis motivates a quadratic-growth second-moment condition, so the stochastic gradient need not be uniformly bounded. We use the condition in the form $\mathbb{E}\|g_R(x;\zeta)\|^2 \le B_v^2\|x-a\|^2 + b_v^2$ and call it RG (raw growth). PASTA uses two mechanisms: **Halpern anchoring** keeps each update close to a reference point, and **Tikhonov regularization** adds local curvature. We propose **Re**wind with **PASTA** for **C**ertified **U**nlearning (**RePASTA-CU**), a dual-phase learning and unlearning algorithm pair. In the learning phase, the learning algorithm A runs $T-K$ clipped SGD updates on the full data and stores the resulting checkpoint. It then runs a K-step PASTA to finish training, and injects Gaussian noise for publishing. In the unlearning phase, the unlearning algorithm U receives the checkpoint and a deletion request, uses only the retained data, and runs the same K-step PASTA and the noise injection. The unlearning algorithm U does not repeat the earlier $T-K$ updates.

The (ε,δ) certification follows from a bound on the distance between the two model-weight vectors before output perturbation. Public clipping and projection provide a finite worst-case bound, although this bound need not decrease as K grows. We choose the Gaussian noise standard deviation σ from this bound and the target (ε,δ), and add independent noise with this standard deviation to both vectors. The resulting released distributions satisfy both directions of the certified-unlearning inequality. The retained-data retraining run is used only for this theoretical comparison and is not executed by U. This certification proof does not require smoothness, Lipschitz continuity, weak convexity, or uniformly bounded gradients of the loss.

We further guarantee the utility of RePASTA-CU, which is usually defined as the stationarity in nonconvex learning. The utility analysis assumes a weakly convex loss, with the quadratic stochastic-subgradient growth condition, and a closed parameter domain. We analyze the K PASTA updates of U starting from its stored checkpoint. Stationarity is measured by the gradient norm of the Moreau envelope, which is a smooth proxy for the nonsmooth retained-data objective. Before adding noise, the expected squared stationarity decreases with root-mean-square rate $K^{-1/6}$. Adding Gaussian noise to the released model weights contributes a separate term controlled by the noise scale σ. Under the minimal calibration in (7), the released-model bound vanishes only when the distance bound also vanishes. The utility result thus covers weakly convex, nonsmooth losses without global smoothness or uniformly bounded gradients.

Our contributions are as follows.

- We develop RePASTA-CU, a first-order learning–unlearning pair that combines nonprivate training, checkpoint reuse, and PASTA updates.
- We prove that RePASTA-CU is an (ε,δ) certified unlearning without smoothness, Lipschitz continuity, weak convexity, or uniformly bounded gradients.
- Under weak convexity and quadratic growth of the second moments of stochastic subgradients on a bounded convex domain, we derive a Moreau-stationarity utility bound. Its root-mean-square rate before publishing-noise injection is $K^{-1/6}$, and publishing-noise injection introduces an additional σ-dependent utility error term.

2 Preliminaries

2.1 Notation and Certified Unlearning

Let $\mathcal{Z}$ be the instance space and let $\mathcal{D}=(z_1,\dots,z_n)\in\mathcal{Z}^n$ be the training set. Let $\Theta\subseteq\mathbb{R}^d$ be the parameter domain. For a dataset E, write $f_E(\theta)=|E|^{-1}\sum_{z\in E} f_z(\theta)$. A deletion request specifies $Z\subseteq\mathcal{D}$, where $m=|Z|<n$, and leaves $\mathcal{D}'=\mathcal{D}\setminus Z$. The randomized learning algorithm is A and the unlearning algorithm is U. We use T for the total number of learning updates, K for the updates after the stored checkpoint, and σ for the standard deviation of the final Gaussian perturbation.

Definition 1 ((ε, δ)-certified unlearning). *For $\varepsilon \geq 0$ and $0 \leq \delta < 1$, the pair (A, U) satisfies (ε, δ)-certified unlearning if, for every dataset $\mathcal{D}$, deletion set Z, and measurable $S \subseteq \mathbb{R}^d$,*

$$\Pr[U(A(\mathcal{D}), \mathcal{D}, Z) \in S] \leq e^{\varepsilon} \Pr[A(\mathcal{D}') \in S] + \delta,$$
$$\Pr[A(\mathcal{D}') \in S] \leq e^{\varepsilon} \Pr[U(A(\mathcal{D}), \mathcal{D}, Z) \in S] + \delta.$$

Here $A(\mathcal{D}')$ is the model obtained by retraining A from initialization on retained data. It is used only for the certificate. In the first argument of U, $A(\mathcal{D})$ denotes the training state that contains the stored checkpoint; U uses this checkpoint and $\mathcal{D}'$. Both output distributions include the final perturbation $\mathcal{N}(0, \sigma^2 I_d)$. Let θ_{T-K} be the checkpoint and let θ_K^U be the model after the K retained-data updates. For a nonsmooth objective, we use the Moreau envelope $f_\lambda(\theta) = \inf_u \{f(u) + (\lambda/2)\|u - \theta\|_2^2\}$. Weak convexity means that $f + (\rho/2)\|\cdot\|_2^2$ is convex for some $\rho \geq 0$ (Rockafellar & Wets, 1998; Bauschke & Combettes, 2020). PASTA controls the expected squared stochastic-subgradient error by a quadratic function of the distance from its anchor.

2.2 Related Work

Guo et al. (Guo et al., 2019) introduced certified removal for regularized linear models. A Newton correction has residual $O(1/n)$ under differentiable convex losses, bounded per-example gradients and features, and a Lipschitz second derivative; perturbing this residual gives an (ε, δ) certificate. Descent-to-Delete (Neel et al., 2021) extends the setting to addition and deletion sequences. Its projected-gradient bounds require convexity or strong convexity, Lipschitz losses, smooth gradients, and a bounded parameter domain; strong convexity gives geometric accuracy in the deletion-update budget. Sekhari et al. (Sekhari et al., 2021) use auxiliary statistics and a Hessian correction. Under strong convexity, Lipschitz losses, and a Hessian-Lipschitz condition, the correction error is $O(m^2/n^2)$, yielding population-risk bounds and deletion capacity $\widetilde{O}(n/d^{1/4})$. These methods all retain substantial convexity or smoothness assumptions. Approximate deletion for linear and logistic models uses low-dimensional corrections (Izzo et al., 2021), while SISA training limits the influence of a point through data partitioning (Bourtoule et al., 2021). Selective forgetting modifies trained weights without retraining from scratch (Golatkar et al., 2020).

For nonconvex models, Zhang et al. (Zhang et al., 2024) use a regularized inverse-Hessian correction on a bounded parameter ball. Lipschitz gradients and Hessians give an error bound Δ to the retrained model, and Gaussian noise of order $\Delta\sqrt{\log(1/\delta)}/\varepsilon$ certifies the output. Chien et al. (Chien et al., 2024) use projected noisy SGD on a bounded convex set. With smooth, bounded-gradient, strongly convex losses, their Rényi unlearning loss decreases as $O(Z_{\mathcal{B}}^2 c^{2Kn/b}/(\eta\sigma^2))$, where $c = 1 - \eta\mu$; their full-batch excess-risk bound is $O(Mc^K Z_{\mathcal{B}} + Ld\sigma^2/\mu)$. Their method adds noise throughout training and unlearning. Recent certified schemes study online corrections, source-data-free updates, and utility after deletion (Qiao et al., 2025; Basaran et al., 2025; Allouah et al., 2025).

R2D (Mu & Klabjan, 2026) is the closest first-order baseline. It stores the model after $T - K$ gradient-descent updates and performs K updates on the retained-data objective. Its final noise is calibrated as

$$\sigma = \frac{2mGh(K)}{Ln\varepsilon}\sqrt{2\log\left(\frac{1.25}{\delta}\right)}, \qquad h(K) = \left(\left(1 + \frac{\eta Ln}{n-m}\right)^{T-K} - 1\right)(1+\eta L)^K.$$

Theorem 3.1 assumes global L-smoothness and a uniform gradient bound G; its average squared-gradient bound is of order $O(\sigma^2 + n/[T(n-m)] + (T-K-1)m/[T(n-m)])$. Koloskova et al. (Koloskova et al., 2025) remove loss assumptions from certification with noisy fine-tuning and clipping, but use an existential comparison algorithm and inject noise during fine-tuning; they do not give a nonsmooth stationarity theorem.

The gap is a first-order checkpoint method that keeps non-private training and the fixed retraining target $A(\mathcal{D}')$, while removing global smoothness and uniform gradient bounds from certification. RePASTA-CU fills this gap with public clipping and projection for certification, and PASTA updates for utility under weak convexity and quadratic stochastic-subgradient growth. The surrounding literature also includes privacy leakage after unlearning and approximate removal limits (Suriyakumar & Wilson, 2022; Chourasia & Shah, 2023).

Algorithm 1 Learning algorithm A

Require: Full data $\mathcal{D}$, initialization $\theta_0 \in \Theta$, total steps T, checkpoint length $K = SN$, and public schedules

for $t = 0, \dots, T-K-1$ **do**
  Draw a stochastic subgradient $g_t(\mathcal{D}, \theta_t)$
  Update $\theta_{t+1} \leftarrow \Pi_\Theta(\theta_t - \eta_t \bar{g}_t(\mathcal{D}, \theta_t))$
end for
Store $a_0 \leftarrow \theta_{T-K}$ for future deletion requests
for $s = 0, \dots, S-1$ **do**
  Set $u_{s,0} \leftarrow a_s$
  for $\ell = 0, \dots, N-1$ **do**
    Draw $g_{s,\ell}(\mathcal{D}, u_{s,\ell})$
    Update $u_{s,\ell+1} \leftarrow \Pi_\Theta((1-\beta)u_{s,\ell} + \beta a_s - \eta \bar{g}_{s,\ell}(\mathcal{D}, u_{s,\ell}))$
  end for
  Set $a_{s+1} \leftarrow N^{-1} \sum_{\ell=0}^{N-1} u_{s,\ell}$
end for
Draw $J \sim \mathrm{Unif}\{0, \dots, S-1\}$ and $\xi_A \sim \mathcal{N}(0, \sigma^2 I_d)$
return $\widetilde{\theta}_A \leftarrow a_J + \xi_A$

3 REPASTA-CU ALGORITHM

Algorithm overview. RePASTA-CU has two phases. In the learning phase, A initializes $\theta_0 \in \Theta$, performs $T - K$ clipped-SGD updates on the full data $\mathcal{D}$, and stores the checkpoint $a = \theta_{T-K}$ before the final release noise is added. Starting from a, it performs K PASTA updates on $\mathcal{D}$, adds Gaussian noise, and releases the model. After a deletion request, U loads a, uses the retained data $\mathcal{D}'$, performs the same K PASTA updates, and releases the result after adding Gaussian noise. Thus, U does not repeat the first $T - K$ updates. The model obtained by rerunning A on $\mathcal{D}'$ is used only in the certification comparison and is not computed by U.

Procedure. The learning run is $\mathcal{D}$ prefix training $\to$ hidden checkpoint $\to$ K-call PASTA suffix training $\to$ Gaussian release. A deletion run loads the checkpoint, replaces $\mathcal{D}$ by $\mathcal{D}'$, executes the same suffix training, and releases its sampled epoch point. Retraining on $\mathcal{D}'$ is used only as the certification target.

For a dataset E and an anchor a, let $g_t(E, \theta_t)$ be a stochastic subgradient. We use the public clipping rule

$$\bar{g}_t(E, \theta_t) = \mathrm{clip}_{C_t}(g_t(E, \theta_t)), \qquad \mathrm{clip}_{C_t}(v) = v \min\left\{1, \frac{C_t}{\|v\|_2}\right\}, \tag{1}$$

with $\mathrm{clip}_{C_t}(0) = 0$. Here C_t is fixed by the public algorithm specification. The clipped-SGD update used in the first phase is

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta_t \bar{g}_t(E, \theta_t)), \tag{2}$$

where η_t is the step size and Π_Θ is Euclidean projection onto the parameter domain. In the second phase, RePASTA uses the PASTA update of Arda et al. (Fazla et al., 2026),

$$\theta_{t+1} = \Pi_\Theta((1-\beta_t)\theta_t + \beta_t a - \eta_t \bar{g}_t(E, \theta_t)), \qquad 0 \le \beta_t \le 1. \tag{3}$$

The term with β_t is the Halpern anchor. The schedule in Section 4 couples β_t and η_t as the Tikhonov regularization required by PASTA.

Learning algorithm. Algorithm 1 adds no Gaussian or other privacy noise to its training updates. The checkpoint is stored before the final publishing noise.

Unlearning algorithm. Algorithm 2 uses the checkpoint as its initial state and performs only the $K = SN$ PASTA updates on $\mathcal{D}'$. It does not reconstruct the first $T - K$ updates. The retrained model $A(\mathcal{D}')$ is computed only in the certification argument.

Algorithm 2 Unlearning algorithm U

Require: Stored checkpoint a_0, retained data $\mathcal{D}'$, and the public PASTA and publishing schedules

Set $a_0^U \leftarrow a_0$
for $s = 0, \dots, S-1$ **do**
  Set $u_{s,0}^U \leftarrow a_s^U$
  for $\ell = 0, \dots, N-1$ **do**
    Draw $g_{s,\ell}(\mathcal{D}', u_{s,\ell}^U)$
    Update $u_{s,\ell+1}^U \leftarrow \Pi_\Theta\Big((1-\beta)u_{s,\ell}^U + \beta a_s^U - \eta \bar{g}_{s,\ell}(\mathcal{D}', u_{s,\ell}^U)\Big)$
  end for
  Set $a_{s+1}^U \leftarrow N^{-1}\sum_{\ell=0}^{N-1} u_{s,\ell}^U$
end for
Draw $J \sim \mathrm{Unif}\{0, \dots, S-1\}$ and independent $\xi_U \sim \mathcal{N}(0, \sigma^2 I_d)$
return $\widetilde{\theta}_U \leftarrow a_J^U + \xi_U$

Implementation details. The domain Θ, clipping radii C_t, step sizes η_t, anchor weights β_t, and the publishing-noise rule are part of the public algorithm specification. The learning and unlearning phases use the same update map, and their only difference is whether the update reads $\mathcal{D}$ or $\mathcal{D}'$. The checkpoint a is stored once during learning. The learning cost is T stochastic first-order updates, while each deletion request costs K updates after the checkpoint and stores one d-dimensional model state. Section 4 chooses σ and analyzes the certification and utility of the two released models.

4 Analyses

4.1 Certified Unlearning Guarantee

We first bound the distance used for the final Gaussian perturbation. Let $a_{\mathcal{D}}$ be the checkpoint produced by the first $T-K$ clipped-SGD updates on $\mathcal{D}$. Let $a_{\mathcal{D}'}^{\mathrm{ret}}$ be the checkpoint obtained by rerunning those same updates from the same initialization on $\mathcal{D}'$. Let a_s^U and a_s^{ret} denote the epoch points produced by the $K = SN$ retained-data PASTA updates of U and of the retraining run, respectively. For the proof, couple the two runs using identically distributed random seeds and the same epoch index J. This preserves each output distribution; actual executions use independent randomness.

The following distance bounds are derived for the RePASTA-CU update map. Because projection is nonexpansive and every clipped vector has norm at most its public radius, the prefix training checkpoints obey

$$B_{\mathrm{pre}} := 2\sum_{t=0}^{T-K-1} \eta_t C_t, \qquad \|a_{\mathcal{D}} - a_{\mathcal{D}'}^{\mathrm{ret}}\|_2 \le B_{\mathrm{pre}}. \tag{4}$$

For the s-th PASTA epoch, let $d_s = \|a_s^U - a_s^{\mathrm{ret}}\|_2$. The two inner updates start at points whose distance is d_s, use anchors whose distance is also d_s, and use clipped vectors with norms at most $C_{s,\ell}$. Nonexpansiveness of projection therefore gives

$$\|u_{s,\ell+1}^U - u_{s,\ell+1}^{\mathrm{ret}}\|_2 \le d_s + 2\sum_{j=0}^{\ell} \eta_{s,j} C_{s,j}. \tag{5}$$

Averaging the inner iterates and iterating over the epochs yields

$$d_J \le \Delta_{T,K} := B_{\mathrm{pre}} + B_{\mathrm{post}}, \qquad B_{\mathrm{post}} := 2\sum_{s=0}^{S-1}\sum_{\ell=0}^{N-1} \eta_{s,\ell} C_{s,\ell}. \tag{6}$$

The same bound holds for every possible value of J. It is finite and public. Decay of $\Delta_{T,K}$ with K requires contraction or stability of the PASTA update map. If Θ is bounded, its diameter can be used as an additional upper bound.

Let $x_U = a_J^U$ and $x_{\mathrm{ret}} = a_J^{\mathrm{ret}}$ be the two model weights before the final publishing perturbation. Define the released models by $\widetilde{x}_U = x_U + \xi_U$ and $\widetilde{x}_{\mathrm{ret}} = x_{\mathrm{ret}} + \xi_{\mathrm{ret}}$, where the two noises are

independent and both follow $\mathcal{N}(0, \sigma^2 I_d)$. For $0 < \varepsilon \leq 1$ and $0 < \delta < 1$, choose

$$\sigma \geq \frac{\Delta_{T,K}}{\varepsilon}\sqrt{2\log\left(\frac{1.25}{\delta}\right)}. \tag{7}$$

The calibration in (7) is the standard Gaussian-shift calibration (Dwork & Roth, 2014; Mironov, 2017; Abadi et al., 2016; Balle & Wang, 2018; Wang et al., 2019, Theorem A.1). The implementation uses the minimal value $\sigma_{\min} = (\Delta_{T,K}/\varepsilon)\sqrt{2\log(1.25/\delta)}$; the theorem also permits any larger public scale.

Theorem 1 (Certified unlearning guarantee)**.** *Assume a common initialization, a nonempty closed convex projection set, finite public clipping radii and nonnegative step sizes, and anchor weights in* $[0, 1]$*. Fix the same schedules,* T*,* K*, and* σ *for learning, unlearning, and retraining. If* σ *satisfies (7), then for every fixed deletion set and every measurable* $\mathcal{B} \subseteq \mathbb{R}^d$*,*

$$\Pr[\widetilde{x}_U \in \mathcal{B}] \leq e^{\varepsilon}\Pr[\widetilde{x}_{\mathrm{ret}} \in \mathcal{B}] + \delta,$$
$$\Pr[\widetilde{x}_{\mathrm{ret}} \in \mathcal{B}] \leq e^{\varepsilon}\Pr[\widetilde{x}_U \in \mathcal{B}] + \delta.$$

Consequently, (A, U) *satisfies Definition 1. The theorem does not assume differentiability, smoothness, Lipschitz continuity, weak convexity, or a uniform bound on the un-clipped gradients of the loss.*

The complete proof is given in Appendix A; the main text records the distance recursion and the Gaussian calibration used by the theorem.

The proof uses bounded clipped updates and nonexpansive projection. The distance bound is determined by the public update budgets and has no deletion-ratio factor m/n.

4.2 Utility Guarantee

The finite-sample utility bound below is derived for the clipped epoch schedule, in the spirit of stochastic first-order analyses for nonconvex objectives (Ghadimi & Lan, 2013). Write $R = \mathcal{D}'$ and condition on the stored checkpoint $a = a_0$. Group the $K = SN$ calls into S epochs, set $\eta_{s,\ell} = \eta$, $\beta_{s,\ell} = \lambda\eta$, and use radius H_S. Assume bounded closed convex Θ with $\Theta \subseteq \operatorname{dom} f_R$, proper lsc f_R bounded below and ρ-weakly convex on Θ, $\lambda > \rho$, and $a \in \operatorname{dom} f_R$.

At every iterate, conditional on the entire history before its fresh oracle call, the oracle satisfies:

$$\mathbb{E}[g_R(x;\zeta) \mid x] \in \partial f_R(x), \qquad \mathbb{E}[\|g_R(x;\zeta)\|_2^2 \mid x] \leq B_v^2\|x - a\|_2^2 + b_v^2. \tag{8}$$

Condition (8) bounds the raw second moment and allows individual stochastic subgradients to be unbounded. We call (8) RG (raw growth). It enters the utility analysis, while certification uses public clipping and projection. The proof below uses the raw second-moment condition in (8), not a centered BG-0 variance bound. Condition (8) supplies the raw second moment used in the clipping-bias and descent bounds below. Define the derived quantities

$$R_\Theta := \sup_{x\in\Theta}\|x - a\|_2, \qquad V_\Theta := B_v^2 R_\Theta^2 + b_v^2, \qquad \mu := \lambda - \rho. \tag{9}$$

By the domain bound, the checkpoint definition, and (8), the quantities in (9) are finite and bound the terms used below. On Θ, clipping has bias at most

$$\left\|\mathbb{E}[\mathrm{clip}_{H_S}(g_R) - g_R \mid x]\right\|_2 \leq \frac{V_\Theta}{H_S}, \tag{10}$$

because $\|v - \mathrm{clip}_{H_S}(v)\|_2 \leq \|v\|_2^2/H_S$.

Define the retained-data Moreau envelope with the PASTA coefficient by

$$M_R(x) := \min_{u\in\Theta}\left\{f_R(u) + \frac{\lambda}{2}\|u - x\|_2^2\right\}. \tag{11}$$

Equation (11) is the Moreau envelope of the constrained objective $f_R + \iota_\Theta$ with parameter $1/\lambda$, where ι_Θ is zero on Θ and $+\infty$ elsewhere. It is defined for all $x \in \mathbb{R}^d$. The constrained Moreau construction in (11) is standard for weakly convex objectives (Davis & Drusvyatskiy, 2019). For $\lambda > \rho$, the minimizer in (11) is unique and the envelope is differentiable (Davis & Drusvyatskiy, 2019; Davis & Grimmer, 2019). Let $\Gamma_R = f_R(a) - \inf_{u\in\Theta} f_R(u)$.

Theorem 2 (Utility bound)**.** *Assume that Θ is bounded, closed, and convex with $\Theta \subseteq \operatorname{dom} f_R$; that f_R is proper, lower semicontinuous, bounded below, and ρ-weakly convex on Θ; that $\lambda > \rho$; that $a \in \operatorname{dom} f_R$; and that the stochastic oracle satisfies (8) at every iterate conditional on the history before its fresh call. If $0 < \eta \le 1/\lambda$ and $N \ge 4/(\mu\eta)$, let J be uniform on $\{0, \dots, S-1\}$ and let a_J^U be the epoch point produced by U before the publishing perturbation. Then*

$$\mathbb{E}\|\nabla M_R(a_J^U)\|_2^2 \le B_K, \tag{12}$$

where

$$B_K := \frac{8\lambda^2\Gamma_R}{3\mu S} + \frac{4\lambda^2\eta(H_S + 2\lambda R_\Theta)^2}{3\mu} + \frac{8\lambda^2 V_\Theta^2}{3\mu^2 H_S^2}. \tag{13}$$

The expectation is conditional on a and includes the oracle calls and J. Averaging both sides over the training randomness gives the unconditional bound, with $\mathbb{E}[B_K]$ on the right.

The complete descent argument is given in Appendix A. Choose

$$\eta_S = \Theta(S^{-2}), \qquad H_S = \Theta(S^{1/2}), \qquad N_S = \left\lceil \frac{4}{\mu\eta_S} \right\rceil. \tag{17}$$

For bounded $\Gamma_R, R_\Theta, V_\Theta$, this gives $K = \Theta(S^3)$, $B_K = O(K^{-1/3})$, and

$$\left(\mathbb{E}\|\nabla M_R(a_J^U)\|_2^2\right)^{1/2} = O(K^{-1/6}). \tag{18}$$

For the released model $\widetilde{a}_J^U = a_J^U + \xi_U$, Lipschitzness of the Moreau gradient and $\mathbb{E}\|\xi_U\|^2 = d\sigma^2$ give

$$\mathbb{E}\|\nabla M_R(\widetilde{a}_J^U)\|_2^2 \le 2B_K + 2\left(\frac{\lambda^2}{\mu}\right)^2 d\sigma^2. \tag{19}$$

Under this schedule, the optimization term in (19) decays with K. The Gaussian term is determined by the certified distance in (7), while the suffix training budget in (6) is generally $\Theta(S^{3/2})$. The moving anchor removes the fixed-anchor bias from the epoch objective.

5 Experiments

Tasks and methods. We evaluate AG News, CIFAR-100, Tiny-ImageNet, and FEMNIST with DeBERTa-v3-small (He et al., 2020), ViT-B/16 (Dosovitskiy, 2020), ConvNeXt-Tiny (Liu et al., 2022), and the LEAF writer split (Caldas et al., 2018). We compare RePASTA-CU with retained-data retraining, R2D (Mu & Klabjan, 2026), and SCRUB (Kurmanji et al., 2023); clipped-SGD and PASTA without release noise are ablations. The Llama-3.1-8B performance test uses LoRA (Hu et al., 2021) and is reported in Appendix B. The logistic control uses Guo–Newton (Guo et al., 2019).

Experimental setting. Each deep model receives the same public AdamW initialization (Loshchilov et al., 2017). RePASTA-CU and retraining then start from that checkpoint and use $T = 30$, $\eta = 2 \times 10^{-5}$, and $C = 0.5$. The main table uses random deletion with $m/n \approx 0.10$ and $K/T = 0.10$; loss-based deletion is reported separately. We report released test accuracy and deletion time. Deep-network runs use one seed, while vector and logistic controls use five seeds.

Table 1: Released test accuracy (%) and deletion time (s) at $m/n \approx 0.10$ and $K/T = 0.10$. All methods start from the same public checkpoint. Bold underlining marks the highest released accuracy in each task; it is a within-table comparison.

Task	Retrain	RePASTA-CU	R2D	SCRUB	RePASTA time	Retrain time
AG News	90.05	**90.16**	89.97	89.93	0.612	2.548
CIFAR-100	77.72	78.15	78.41	**78.45**	0.535	3.016
Tiny-ImageNet	60.86	60.62	62.56	**62.67**	0.696	6.178
FEMNIST	**77.95**	77.65	77.83	77.82	0.052	0.255

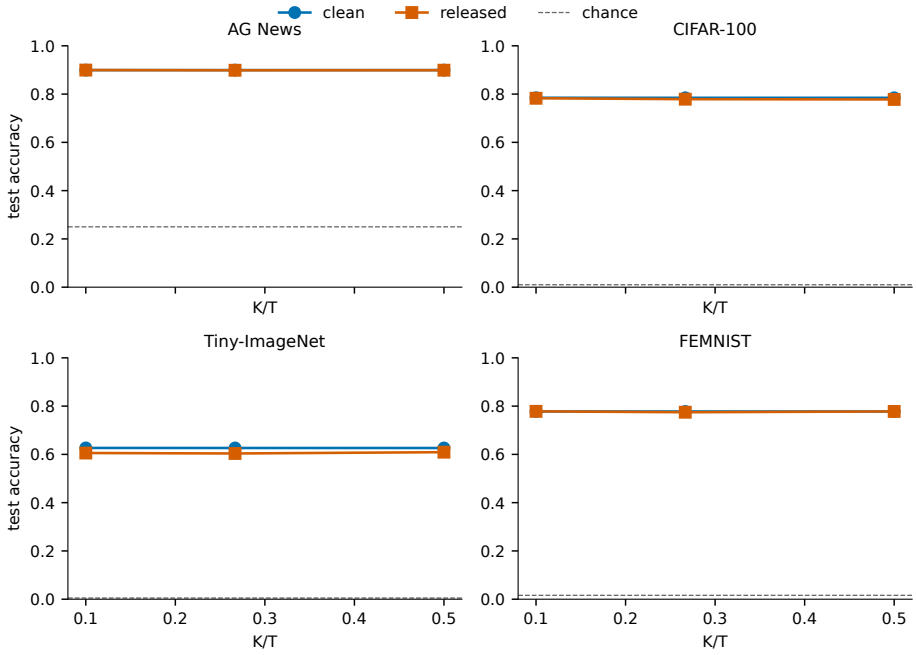


Figure 1: Test accuracy as the retained-data suffix training budget changes. Points show the clean and released models; dashed lines mark the task chance level.

Classification utility. At $m/n \approx 0.10$ and $K/T = 0.10$, RePASTA-CU is within 0.43 percentage points of retained-data retraining on all four tasks (Table 1). The deletion continuation is 4.16–8.87 times faster than retraining. It is the best released method on AG News; the gaps to the best baseline on the other tasks are 0.30, 2.05, and 0.30 percentage points. Figure 1 shows the accuracy trend as the suffix training budget changes. Under loss-based deletion, released accuracies are 90.01%, 77.86%, 59.36%, and 77.53% for AG News, CIFAR-100, Tiny-ImageNet, and FEMNIST, respectively; the corresponding retraining values are 89.93%, 78.29%, 60.71%, and 77.51% in Table 2. Vector controls use their natural metric and are summarized in Table 3; the logistic control and the Llama-3.1-8B performance test are reported in the appendix.

The loss-based comparison uses the same retained set for both methods. It tests whether the checkpoint continuation remains close to retraining when the deleted examples are selected by their training loss.

Table 2: Released test accuracy (%) under loss-based deletion at $m/n \approx 0.10$ and $K/T = 0.10$. RePASTA-CU and retained-data retraining use the same retained sets. Bold underlining marks the higher accuracy in each task.

Task	RePASTA-CU	Retraining	Deletion time (s)
AG News	**90.01**	89.93	0.592
CIFAR-100	77.86	**78.29**	0.540
Tiny-ImageNet	59.36	**60.71**	0.703
FEMNIST	**77.53**	77.51	0.052

Table 3: Moreau-proxy values for the vector controls at $m/n \approx 0.10$ and $K/T = 0.10$. The released value is evaluated after the Gaussian release.

Task	Clean Moreau^2	Released Moreau^2
Synthetic 1D	1.32×10^{-4}	2.2292
Synthetic 2D	0.0400	1.3456
Smooth 1D	9.14×10^{-5}	1.7786

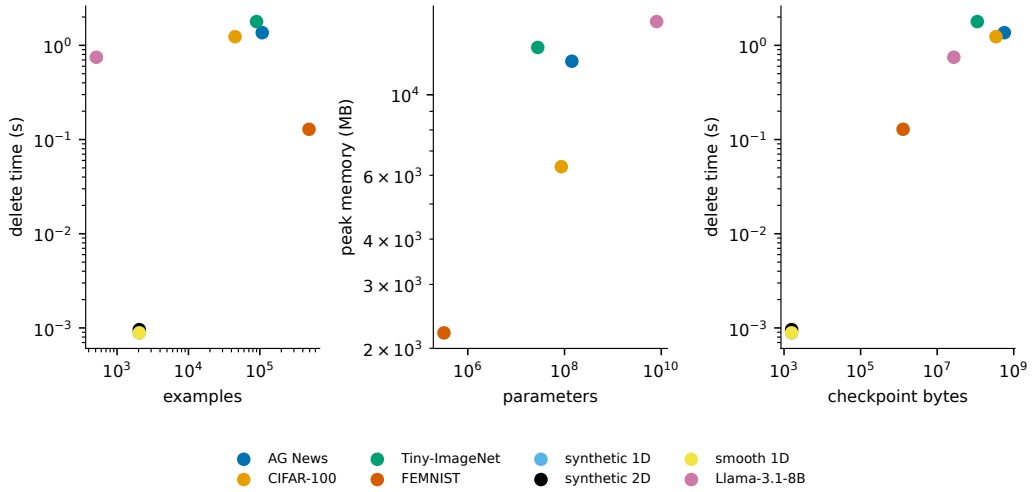


Figure 2: Deletion time and peak memory across task scale. Each point is one task; the horizontal axes show the number of examples, parameter count, and checkpoint size.

Figure 2 reports deletion time and peak memory across the evaluated tasks. On the four classification tasks, deletion takes 0.052–0.696 seconds, compared with 0.255–6.178 seconds for retained-data retraining. Runtime follows suffix training computation, while peak memory tracks the stored parameter and checkpoint footprint; together they expose the compute and storage costs of checkpoint reuse.

6 Conclusion

Certified unlearning must remove a data record's influence without replaying the full training run. RePASTA-CU addresses this requirement with a checkpoint-based learning–unlearning pair. The learning algorithm stores the state after the first $T - K$ clipped updates; after deletion, the unlearning algorithm reuses this state and runs only the retained-data PASTA suffix training. Public clipping and projection bound the distance to the fixed retained-data retraining reference. Calibrating one final Gaussian perturbation to this distance gives bidirectional (ε, δ) certified unlearning. This certification argument does not require differentiability, smoothness, Lipschitz continuity, weak convexity, or a uniform bound on the raw gradients.

The utility guarantee assumes weak convexity on a bounded convex domain and the RG raw second-moment condition in (8). Moving-anchor PASTA yields a Moreau-stationarity bound for the retained-data suffix training. With the schedule in Section 4, the pre-release RMS stationarity rate is $O(K^{-1/6})$. The released-model bound adds the explicit Gaussian term in (19), whose size is determined by the certificate scale. Experiments on text, image, and federated classification show test accuracy close to retained-data retraining with a shorter deletion continuation. The vector controls quantify the increase in the Moreau proxy after Gaussian release (Table 3). The guarantee is stated for one fixed deletion request; repeated or adaptive requests require composition for the stored state and released outputs.

At the reported deletion fraction, released accuracy differs from retained-data retraining by at most 0.43 percentage points, while the deletion continuation is 4.16–8.87 times faster. The vector controls also show that the release term can dominate the Moreau proxy when the certified perturbation is large.

The certificate and utility analyses use different parts of the construction. Clipping and projection set a finite release scale for the certified comparison, whereas weak convexity and RG control the retained-data suffix training utility. This separation keeps the certification statement applicable to nonsmooth objectives while the $O(K^{-1/6})$ rate is stated under the assumptions required by PASTA.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing. All AI-assisted polishing were reviewed against the original text and revised by the authors to ensure accuracy. We take full responsibility for the final content of this work, including any text produced with the assistance of generative AI.

ETHICS STATEMENT

This work focuses on the theory and does not involve human subjects, private or personally identifiable information, or interactions with production systems. All experiments are conducted in controlled settings using publicly available models and datasets. The proposed watermarking method is intended to support content provenance and the identification of machine-generated text. Although watermarking research may have dual-use implications, this study does not collect or release sensitive data and explicitly presents its assumptions and limitations. We identify no additional ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

Codes, all experimental results, and all hyperparameter settings are submitted to the Supplementary Material on OpenReview.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Youssef Allouah, Joshua Kazdan, Rachid Guerraoui, and Sanmi Koyejo. The utility and complexity of in-and out-of-distribution machine unlearning. In *International Conference on Learning Representations*, volume 2025, pp. 2713–2738, 2025.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International conference on machine learning*, pp. 394–403. PMLR, 2018.

Umit Yigit Basaran, Sk Miraj Ahmed, Amit Roy-Chowdhury, and Basak Guler. A certified unlearning approach without access to source data. *arXiv preprint arXiv:2506.06486*, 2025.

Heinz H Bauschke and Patrick L Combettes. Correction to: convex analysis and monotone operator theory in hilbert spaces. In *Convex analysis and monotone operator theory in Hilbert spaces*, pp. C1–C4. Springer, 2020.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Certified machine unlearning via noisy stochastic gradient descent. *Advances in Neural Information Processing Systems*, 37:38852–38887, 2024.

Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In *International conference on machine learning*, pp. 6028–6073. PMLR, 2023.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Arda Fazla, Ege C Kaya, Antesh Upadhyay, and Abolfazl Hashemi. Lower bounds and proximally anchored sgd for non-convex minimization under unbounded variance. *arXiv preprint arXiv:2604.16620*, 2026.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9301–9309. IEEE, 2020.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, pp. 2008–2016. PMLR, 2021.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. *arXiv preprint arXiv:2506.06985*, 2025.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5, 2017.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Siqiao Mu and Diego Klabjan. Rewind-to-delete: Certified machine unlearning for nonconvex functions. *Advances in Neural Information Processing Systems*, 38:114911–114947, 2026.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic learning theory*, pp. 931–962. PMLR, 2021.

Xinbao Qiao, Meng Zhang, Ming Tang, and Ermin Wei. Hessian-free online certified unlearning. In *International Conference on Learning Representations*, volume 2025, pp. 32675–32711, 2025.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

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.

Vinith Suriyakumar and Ashia C Wilson. Algorithms that approximate data removal: New results and limitations. *Advances in Neural Information Processing Systems*, 35:18892–18903, 2022.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1226–1235. PMLR, 2019.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep neural networks. *arXiv preprint arXiv:2408.00920*, 2024.

A COMPLETE PROOFS

Lemma 1 (prefix training checkpoint distance)**.** *For two prefix training runs with the same initialization, common coupled randomness, public radii C_t, and projection onto a closed convex set,*

$$\|a_{\mathcal{D}} - a^{\mathrm{ret}}_{\mathcal{D}'}\|_2 \le 2 \sum_{t=0}^{T-K-1} \eta_t C_t = B_{\mathrm{pre}}.$$

Proof. Write x_t and y_t for the two prefix training iterates, and let $\bar{g}^x_t$ and $\bar{g}^y_t$ be the clipped oracle vectors generated by the coupled calls at step t. The public clipping rule gives

$$\|\bar{g}^x_t\|_2 \le C_t, \qquad \|\bar{g}^y_t\|_2 \le C_t.$$

Before projection, the two next points therefore satisfy

$$\begin{aligned}&\|x_t - \eta_t \bar{g}^x_t - (y_t - \eta_t \bar{g}^y_t)\|_2 \\ &\quad \le \|x_t - y_t\|_2 + \eta_t \|\bar{g}^x_t\|_2 + \eta_t \|\bar{g}^y_t\|_2 \le \|x_t - y_t\|_2 + 2\eta_t C_t.\end{aligned}$$

For a closed convex set, Euclidean projection is nonexpansive, namely $\|\Pi_\Theta(v) - \Pi_\Theta(w)\|_2 \le \|v - w\|_2$. Applying this inequality at every step gives, with $\Delta_t = \|x_t - y_t\|_2$,

$$\Delta_{t+1} \le \Delta_t + 2\eta_t C_t.$$

The common initialization gives $\Delta_0 = 0$. Induction on t yields

$$\Delta_t \le 2 \sum_{j=0}^{t-1} \eta_j C_j.$$

At $t = T - K$, the left side is the distance between the two checkpoints, which is the claimed bound. □

Lemma 2 (Coarse suffix training distance)**.** *For the PASTA suffix training in Algorithms 1–2, under the same coupling,*

$$\|a^U_J - a^{\mathrm{ret}}_J\|_2 \le B_{\mathrm{pre}} + B_{\mathrm{post}} \quad \textit{for every } J,$$

where

$$B_{\mathrm{post}} = 2 \sum_{s=0}^{S-1} \sum_{\ell=0}^{N-1} \eta_{s,\ell} C_{s,\ell}.$$

Proof. Let

$$\Delta_s = \|a^U_s - a^{\mathrm{ret}}_s\|_2, \qquad q_{s,\ell} = \|u^U_{s,\ell} - u^{\mathrm{ret}}_{s,\ell}\|_2.$$

At the beginning of epoch s, both inner iterates equal their epoch anchors, so $q_{s,0} = \Delta_s$. Define the unprojected next points by

$$z^U_{s,\ell} = (1 - \beta_{s,\ell}) u^U_{s,\ell} + \beta_{s,\ell} a^U_s - \eta_{s,\ell} \bar{g}^U_{s,\ell},$$

and define $z^{\mathrm{ret}}_{s,\ell}$ analogously. Since both anchor weights are in $[0, 1]$,

$$\begin{aligned}\|z^U_{s,\ell} - z^{\mathrm{ret}}_{s,\ell}\|_2 &\le (1 - \beta_{s,\ell}) q_{s,\ell} + \beta_{s,\ell} \Delta_s \\ &\quad + \eta_{s,\ell} \|\bar{g}^U_{s,\ell} - \bar{g}^{\mathrm{ret}}_{s,\ell}\|_2.\end{aligned}$$

Each clipped vector has norm at most $C_{s,\ell}$, so the last term is at most $2\eta_{s,\ell} C_{s,\ell}$. Projection is nonexpansive, hence

$$q_{s,\ell+1} \le (1 - \beta_{s,\ell}) q_{s,\ell} + \beta_{s,\ell} \Delta_s + 2\eta_{s,\ell} C_{s,\ell}.$$

We claim by induction on ℓ that

$$q_{s,\ell+1} \le \Delta_s + 2 \sum_{j=0}^{\ell} \eta_{s,j} C_{s,j}.$$

For $\ell = 0$, this follows from $q_{s,0} = \Delta_s$. If the claim holds at ℓ, then

$$q_{s,\ell+1} \le (1-\beta_{s,\ell}) \left(\Delta_s + 2\sum_{j=0}^{\ell-1} \eta_{s,j} C_{s,j} \right) + \beta_{s,\ell}\Delta_s + 2\eta_{s,\ell} C_{s,\ell}$$
$$\le \Delta_s + 2\sum_{j=0}^{\ell} \eta_{s,j} C_{s,j},$$

where the empty sum is zero. Thus every inner iterate in epoch s is at distance at most $\Delta_s + 2\sum_{j=0}^{N-1} \eta_{s,j} C_{s,j}$ from its counterpart. The next epoch anchor is the average of the N inner iterates. Convexity of the norm gives

$$\Delta_{s+1} \le \frac{1}{N} \sum_{\ell=0}^{N-1} q_{s,\ell} \le \Delta_s + 2\sum_{\ell=0}^{N-1} \eta_{s,\ell} C_{s,\ell}.$$

Iterating this recurrence and using Lemma 1, $\Delta_0 \le B_{\mathrm{pre}}$, gives

$$\Delta_J \le B_{\mathrm{pre}} + 2\sum_{s=0}^{S-1}\sum_{\ell=0}^{N-1} \eta_{s,\ell} C_{s,\ell} = B_{\mathrm{pre}} + B_{\mathrm{post}}.$$

The argument does not depend on J, so the bound holds for every epoch index. □

Lemma 3 (Gaussian shift)**.** *If $\|x-y\|_2 \le \Delta$ and $0 < \varepsilon \le 1$, then the two Gaussian laws $\mathcal{N}(x, \sigma^2 I_d)$ and $\mathcal{N}(y, \sigma^2 I_d)$ satisfy both (ε, δ) inequalities whenever*

$$\sigma \ge \frac{\Delta}{\varepsilon}\sqrt{2\log(1.25/\delta)}.$$

Proof. Let $P = \mathcal{N}(x, \sigma^2 I_d)$, $Q = \mathcal{N}(y, \sigma^2 I_d)$, and $d = x - y$. If $d = 0$, the two laws coincide. Otherwise their density ratio has privacy loss

$$L(z) = \log \frac{p(z)}{q(z)} = \frac{\langle d, z - (x+y)/2 \rangle}{\sigma^2}.$$

If $Z \sim P$, then

$$L(Z) = \frac{\|d\|_2^2}{2\sigma^2} + \frac{\|d\|_2}{\sigma} G, \qquad G \sim \mathcal{N}(0,1).$$

Set $h = \|d\|_2/\sigma$ and $L_\delta = \log(1.25/\delta)$. The assumed calibration implies $h \le \varepsilon/\sqrt{2L_\delta}$. Hence

$$\Pr_P[L > \varepsilon] = \Pr\left[G > \frac{\varepsilon}{h} - \frac{h}{2} \right].$$

The threshold is decreasing in h, so it is at least

$$t_0 = \sqrt{2L_\delta} - \frac{\varepsilon}{2\sqrt{2L_\delta}} \ge \sqrt{2L_\delta} - \frac{1}{2\sqrt{2L_\delta}}.$$

If $t_0 \le 0$, then $L_\delta \le 1/4$ and consequently $\delta \ge 1.25e^{-1/4} > 1/2$, so the tail probability is at most $1/2 < \delta$. If $t_0 > 0$, use $\Pr[G \ge t] \le \frac{1}{2}e^{-t^2/2}$ to obtain

$$\Pr_P[L > \varepsilon] \le \frac{1}{2}\exp\left[-\frac{1}{2}\left(\sqrt{2L_\delta} - \frac{1}{2\sqrt{2L_\delta}} \right)^2 \right]$$
$$= \frac{1}{2}\exp\left[-L_\delta + \frac{1}{2} - \frac{1}{16L_\delta} \right] \le 1.25e^{-L_\delta} = \delta.$$

For any measurable set $\mathcal{B}$,

$$P(\mathcal{B}) = \int_{\mathcal{B}\cap\{L\le\varepsilon\}} e^{L(z)}\, dQ(z) + P(\mathcal{B}\cap\{L>\varepsilon\})$$
$$\le e^{\varepsilon} Q(\mathcal{B}) + \Pr_P[L > \varepsilon]$$
$$\le e^{\varepsilon} Q(\mathcal{B}) + \delta.$$

Exchanging x and y gives the reverse inequality, because the same distance and calibration bound apply in the other direction. □

Theorem 3 (Certified unlearning guarantee)**.** *Under the assumptions of Theorem 1, the released unlearning and retained-data retraining distributions satisfy both inequalities in Definition 1 when σ obeys (7).*

Proof. Couple the internal random choices of the unlearning and retained-data retraining runs so that each marginal run has its prescribed law. Under a fixed sampling rule, common uniform seeds provide one such coupling; the coupling is used only for the proof. Include the common epoch index J in this internal randomness. For every coupled realization ω, let $x_U(\omega) = a_J^U(\omega)$ and $x_{\mathrm{ret}}(\omega) = a_J^{\mathrm{ret}}(\omega)$. Lemmas 1 and 2 give

$$\|x_U(\omega) - x_{\mathrm{ret}}(\omega)\|_2 \le \Delta_{T,K}.$$

Condition on ω. The only remaining randomness is the independent Gaussian publishing noise. Lemma 3 therefore gives, for every measurable $\mathcal{B}$,

$$\Pr[\widetilde{x}_U \in \mathcal{B} \mid \omega] \le e^{\varepsilon} \Pr[\widetilde{x}_{\mathrm{ret}} \in \mathcal{B} \mid \omega] + \delta.$$

Integrating this inequality with respect to the law of ω preserves the marginal release distributions and gives

$$\Pr[\widetilde{x}_U \in \mathcal{B}] \le e^{\varepsilon} \Pr[\widetilde{x}_{\mathrm{ret}} \in \mathcal{B}] + \delta.$$

Applying the same argument after swapping the two runs gives the reverse inequality. These are exactly the two inequalities in Definition 1. □

Lemma 4 (Clipping bias under RG)**.** *Under (8), for every $x \in \Theta$,*

$$\left\|\mathbb{E}[\mathrm{clip}_{H_S}(g_R) - g_R \mid x]\right\|_2 \le \frac{V_\Theta}{H_S}.$$

Proof. For a vector v, if $\|v\|_2 \le H_S$ then clipping does nothing. If $\|v\|_2 > H_S$, then

$$\|v - \mathrm{clip}_{H_S}(v)\|_2 = \|v\|_2 - H_S \le \frac{\|v\|_2^2}{H_S}.$$

Thus the same inequality holds in both cases. Because $\|x - a\|_2 \le R_\Theta$ on Θ, (8) implies

$$\mathbb{E}[\|g_R(x;\zeta)\|_2^2 \mid x] \le B_v^2 R_\Theta^2 + b_v^2 = V_\Theta.$$

Taking conditional expectations, using the triangle inequality for the norm of an expectation, and applying the pointwise clipping inequality gives

$$\begin{aligned}\left\|\mathbb{E}[\mathrm{clip}_{H_S}(g_R) - g_R \mid x]\right\|_2 &\le \mathbb{E}[\|\mathrm{clip}_{H_S}(g_R) - g_R\|_2 \mid x] \\ &\le \frac{1}{H_S}\mathbb{E}[\|g_R\|_2^2 \mid x] \le \frac{V_\Theta}{H_S}.\end{aligned}$$

□

Lemma 5 (One-epoch descent)**.** *Let $Q_s(u) = f_R(u) + \lambda\|u - a_s^U\|_2^2/2$ and let p_s be its minimizer. Under the assumptions of Theorem 2,*

$$\mathbb{E}[Q_s(a_{s+1}^U) - Q_s(p_s) \mid a_s^U] \le \frac{r_s^2}{2\eta N} + \frac{\eta(H_S + 2\lambda R_\Theta)^2}{2} + \frac{V_\Theta^2}{\mu H_S^2},$$

where $r_s = \|a_s^U - p_s\|_2$.

Proof. Fix an epoch and condition on the history up to its anchor a_s^U. For readability write $a = a_s^U$, $u_\ell = u_{s,\ell}^U$, $p = p_s$, $r_\ell = u_\ell - p$, and $r_s = \|a - p\|_2$. Since Θ is convex, every projected inner iterate and every averaged epoch anchor belongs to Θ. Moreover, both u_ℓ and a lie in Θ, so

$$\|u_\ell - a\|_2 \le \|u_\ell - a_0\|_2 + \|a - a_0\|_2 \le 2R_\Theta.$$

The update with $\beta = \lambda\eta$ can be written as

$$u_{\ell+1} = \Pi_\Theta(u_\ell - \eta d_\ell), \qquad d_\ell = \mathrm{clip}_{H_S}(g_R(u_\ell;\zeta_\ell)) + \lambda(u_\ell - a).$$

Therefore

$$\|d_\ell\|_2 \le H_S + 2\lambda R_\Theta =: G_H.$$

Let $\mathcal{F}_\ell$ be the history before the fresh oracle call, and define

$$v_\ell = \mathbb{E}[g_R(u_\ell;\zeta_\ell) \mid \mathcal{F}_\ell], \qquad e_\ell = \mathbb{E}[\mathrm{clip}_{H_S}(g_R) - g_R \mid \mathcal{F}_\ell].$$

The conditional-mean condition in (8) gives $v_\ell \in \partial f_R(u_\ell)$, and Lemma 4 gives $\|e_\ell\|_2 \le b_H := V_\Theta/H_S$. The vector

$$q_\ell = v_\ell + \lambda(u_\ell - a)$$

is a subgradient of Q_s at u_ℓ. Since Q_s is μ-strongly convex,

$$Q_s(u_\ell) - Q_s(p) \le \langle q_\ell, r_\ell\rangle - \frac{\mu}{2}\|r_\ell\|_2^2. \tag{A.1}$$

Nonexpansiveness of projection and $p \in \Theta$ imply

$$\begin{aligned}\mathbb{E}_\ell\|r_{\ell+1}\|_2^2 &\le \mathbb{E}_\ell\|u_\ell - \eta d_\ell - p\|_2^2 \\ &= \|r_\ell\|_2^2 - 2\eta\langle q_\ell + e_\ell, r_\ell\rangle + \eta^2\mathbb{E}_\ell\|d_\ell\|_2^2 \\ &\le \|r_\ell\|_2^2 - 2\eta\langle q_\ell + e_\ell, r_\ell\rangle + \eta^2 G_H^2.\end{aligned}$$

Rearranging and using $-\langle e_\ell, r_\ell\rangle \le \|e_\ell\|_2\|r_\ell\|_2$ gives

$$\langle q_\ell, r_\ell\rangle \le \frac{\|r_\ell\|_2^2 - \mathbb{E}_\ell\|r_{\ell+1}\|_2^2}{2\eta} + \frac{\eta G_H^2}{2} + b_H\|r_\ell\|_2. \tag{A.2}$$

Substitute (A.2) into (A.1) and apply Young's inequality $b_H\|r_\ell\|_2 \le \mu\|r_\ell\|_2^2/4 + b_H^2/\mu$:

$$Q_s(u_\ell) - Q_s(p) \le \frac{\|r_\ell\|_2^2 - \mathbb{E}_\ell\|r_{\ell+1}\|_2^2}{2\eta} + \frac{\eta G_H^2}{2} + \frac{b_H^2}{\mu} - \frac{\mu}{4}\|r_\ell\|_2^2. \tag{A.3}$$

Sum (A.3) over $\ell = 0, \dots, N-1$, take the conditional expectation, and drop the final negative term. The distance terms telescope, and $r_0 = a - p$, so

$$\frac{1}{N}\sum_{\ell=0}^{N-1}\mathbb{E}[Q_s(u_\ell) - Q_s(p) \mid a_s^U] \le \frac{r_s^2}{2\eta N} + \frac{\eta G_H^2}{2} + \frac{b_H^2}{\mu}.$$

The function Q_s is convex on Θ, and $a_{s+1}^U = N^{-1}\sum_{\ell=0}^{N-1} u_\ell$. Jensen's inequality gives $Q_s(a_{s+1}^U) \le N^{-1}\sum_\ell Q_s(u_\ell)$. Substituting $G_H = H_S + 2\lambda R_\Theta$ and $b_H = V_\Theta/H_S$ proves the lemma. □

Theorem 4 (Utility bound)**.** *Under the assumptions of Theorem 2, the bound (12)–(13) holds for a uniformly selected epoch point.*

Proof. Let p_s and r_s be as in Lemma 5, and set

$$c = \frac{\eta(H_S + 2\lambda R_\Theta)^2}{2} + \frac{V_\Theta^2}{\mu H_S^2}.$$

Because $N \ge 4/(\mu\eta)$, $1/(2\eta N) \le \mu/8$. Lemma 5 therefore gives

$$\mathbb{E}[Q_s(a_{s+1}^U) - Q_s(p_s) \mid a_s^U] \le \frac{\mu}{8}r_s^2 + c. \tag{A.4}$$

The function Q_s is μ-strongly convex on Θ and p_s is its minimizer. The strong-convexity inequality applied along the segment joining p_s to a_s^U gives

$$Q_s(a_s^U) - Q_s(p_s) \ge \frac{\mu}{2}\|a_s^U - p_s\|_2^2 = \frac{\mu}{2}r_s^2. \tag{A.5}$$

The anchor penalty vanishes at a_s^U, so $Q_s(a_s^U) = f_R(a_s^U)$, while the penalty is nonnegative and hence $f_R(a_{s+1}^U) \le Q_s(a_{s+1}^U)$. Combining (A.4)–(A.5) yields

$$\mathbb{E}[f_R(a_s^U) - f_R(a_{s+1}^U) \mid a_s^U] \ge \frac{3\mu}{8}r_s^2 - c. \tag{A.6}$$

Take total expectations and sum (A.6) over $s = 0, \dots, S-1$. The left side telescopes because the epoch points form one trajectory, and $f_R(a_S^U) \geq \inf_{u\in\Theta} f_R(u)$. With $\Gamma_R = f_R(a) - \inf_{u\in\Theta} f_R(u)$, we obtain

$$\frac{3\mu}{8}\sum_{s=0}^{S-1}\mathbb{E}r_s^2 \leq \Gamma_R + Sc,$$

which is equivalent to

$$\frac{1}{S}\sum_{s=0}^{S-1}\mathbb{E}r_s^2 \leq \frac{8\Gamma_R}{3\mu S} + \frac{8c}{3\mu}. \tag{A.7}$$

For $\lambda > \rho$, the proximal objective in (11) is μ-strongly convex, so its minimizer at a_s^U is unique and equals p_s. The envelope gradient formula proved in Lemma 6 gives $\nabla M_R(a_s^U) = \lambda(a_s^U - p_s)$ and hence $\|\nabla M_R(a_s^U)\|_2^2 = \lambda^2 r_s^2$. Since J is uniform and independent of the fresh suffix training randomness conditional on the trajectory,

$$\mathbb{E}\|\nabla M_R(a_J^U)\|_2^2 = \frac{\lambda^2}{S}\sum_{s=0}^{S-1}\mathbb{E}r_s^2.$$

Substitute (A.7) and the definition of c to obtain exactly (12)–(13). □

Lemma 6 (Lipschitzness of the Moreau gradient)**.** *For $\lambda > \rho$, the constrained Moreau envelope in (11) is differentiable and*

$$\|\nabla M_R(x) - \nabla M_R(y)\|_2 \leq \frac{\lambda^2}{\mu}\|x - y\|_2.$$

Proof. Let $F = f_R + \iota_\Theta$ and let

$$p_x = \arg\min_u \left\{ F(u) + \frac{\lambda}{2}\|u - x\|_2^2 \right\}.$$

Since F is ρ-weakly convex, the objective is μ-strongly convex, so p_x is unique. First-order optimality gives

$$g_x := \lambda(x - p_x) \in \partial F(p_x).$$

For x, y, write $\Delta x = x - y$, $\Delta p = p_x - p_y$ and $\Delta g = g_x - g_y$. Hypomonotonicity of the subdifferential of a ρ-weakly convex function gives

$$\langle \Delta g, \Delta p\rangle \geq -\rho\|\Delta p\|_2^2.$$

Using $\Delta g = \lambda(\Delta x - \Delta p)$ and $\mu = \lambda - \rho$ yields

$$\lambda\langle \Delta x, \Delta p\rangle \geq \mu\|\Delta p\|_2^2. \tag{A.8}$$

If $\Delta x = 0$, (A.8) implies $\Delta p = 0$. Otherwise, if $\Delta p \neq 0$, Cauchy–Schwarz in (A.8) gives

$$\frac{\|\Delta p\|_2}{\|\Delta x\|_2} \leq \frac{\lambda}{\mu}.$$

Set $t = \|\Delta p\|_2/\|\Delta x\|_2$ and $c_0 = \langle \Delta x, \Delta p\rangle/(\|\Delta x\|_2\|\Delta p\|_2)$. Then $0 \leq t \leq \lambda/\mu$ and (A.8) implies $c_0 \geq (\mu/\lambda)t$. Consequently,

$$\begin{aligned}\frac{\|\Delta x - \Delta p\|_2^2}{\|\Delta x\|_2^2} &= 1 + t^2 - 2c_0 t \\ &\leq 1 + \left(1 - \frac{2\mu}{\lambda}\right)t^2.\end{aligned}$$

If $\lambda \leq 2\mu$, the last expression is at most 1. If $\lambda > 2\mu$, it is maximized over $0 \leq t \leq \lambda/\mu$ at the endpoint, where it equals $(\lambda/\mu - 1)^2$, which is also at most $(\lambda/\mu)^2$. Thus in all cases

$$\|\Delta x - \Delta p\|_2 \leq \frac{\lambda}{\mu}\|\Delta x\|_2, \qquad \|\Delta g\|_2 \leq \frac{\lambda^2}{\mu}\|\Delta x\|_2. \tag{A.9}$$

It remains to justify the envelope gradient formula. For any h, using p_x as a feasible point at $x + h$ gives

$$M_R(x + h) - M_R(x) \leq \langle g_x, h \rangle + \frac{\lambda}{2}\|h\|_2^2.$$

Using p_{x+h} as a feasible point at x gives the reverse inequality

$$M_R(x + h) - M_R(x) \geq \langle g_{x+h}, h \rangle - \frac{\lambda}{2}\|h\|_2^2.$$

Inequality (A.9) shows that $g_{x+h} \to g_x$ as $h \to 0$. Dividing the two inequalities by $\|h\|_2$ and letting $h \to 0$ proves that M_R is differentiable with $\nabla M_R(x) = g_x$. □

Corollary 1 (Released-model utility)**.** *Under the assumptions of Theorem 2, if* $\widetilde{a}_J^U = a_J^U + \xi_U$ *with* $\xi_U \sim \mathcal{N}(0, \sigma^2 I_d)$*, then*

$$\mathbb{E}\|\nabla M_R(\widetilde{a}_J^U)\|_2^2 \leq 2B_K + 2\left(\frac{\lambda^2}{\mu}\right)^2 d\sigma^2.$$

Proof. Let $L_M = \lambda^2/\mu$. By Lemma 6,

$$\|\nabla M_R(a_J^U + \xi_U)\|_2 \leq \|\nabla M_R(a_J^U)\|_2 + L_M\|\xi_U\|_2.$$

The inequality $(r + s)^2 \leq 2r^2 + 2s^2$ gives

$$\|\nabla M_R(a_J^U + \xi_U)\|_2^2 \leq 2\|\nabla M_R(a_J^U)\|_2^2 + 2L_M^2\|\xi_U\|_2^2.$$

Taking expectations, using Theorem 2, and noting that $\mathbb{E}\|\xi_U\|_2^2 = \mathrm{tr}(\sigma^2 I_d) = d\sigma^2$, proves the claim. □

Table 4: Guo–Newton (Guo et al., 2019) on breast-cancer logistic regression for deletion fractions near 0.10.

m/n range	Test loss	Test accuracy (%)	Deletion time (s)
0.0989–0.1033	0.6930	33.33	0.005

Table 5: Test loss and deletion time for the Llama-3.1-8B LoRA performance test. Lower test loss is better; bold underlining marks the lower loss at each suffix training budget. The matched retraining run uses the same public checkpoint.

Method	K/T	Test loss	Deletion time (s)
RePASTA-CU	0.25	1.2701	61.90
Retraining	0.25	**1.2697**	682.26
RePASTA-CU	0.50	**1.2687**	703.70
Retraining	0.50	1.2720	682.26

B ADDITIONAL EXPERIMENTAL RESULTS

The appendix reports logistic regression and the Llama-3.1-8B performance test. Each table keeps the metric used by its task.