CERTIFIED UNLEARNING AS A STABILITY TOOL: RATES, LIMITS, AND IMPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The relationship between differential privacy and the generalization of a learning algorithm has long been an open problem. We prove that the existence of a stable "privacy wrapper" implies generalization, a sufficient condition for generalization weaker than differential privacy. Specifically, we view certified unlearning (CU), which guarantees indistinguishability between unlearned and retrained models, as a privacy wrapper for its learning counterpart. We show that an (ε, δ)-unlearning certificate, combined with unlearning stability, controls the learner's algorithmic stability and generalization without requiring differential privacy during learning. Our conversion theorem transfers (ε, δ)-unlearning certification to algorithmic stability by replacing full-trajectory stability with that accumulated over the shorter unlearning process. The resulting stability bound depends on the certified unlearning dynamic. For a contractive unlearner with an anchored checkpoint, the learner has $O(n^{-1})$ stability, uniformly in the training horizon T. For a non-expansive (but non-contractive) unlearner, the stability bound is $O(n^{-1/3})$. In the expansive regime, the bound admits no vanishing rate in n.

Another surprising result is *"ununlearnablility"*. Reversing the conversion theorem yields a contrapositive impossibility result: for nonsmooth convex, expansive smooth nonconvex, and ill-conditioned strongly convex losses, no stable, shallow, and cheaply certifiable unlearner can satisfy the required guarantees. Experiments with linear heads on frozen CIFAR-10 and SST-2 features confirm the predicted scaling laws and phase transitions (Appendix B).

1 INTRODUCTION

The relationship between differential privacy and the generalization of learning algorithms has long been an open problem in learning theory. Algorithmic stability (Defs. 6, 7) provides a useful tool for studying this relationship because it measures how much a learner's output changes when one training example is replaced. The intuition is closely related to differential privacy: if an algorithm limits the information carried by any one record, that record should have little effect on the output. Several works formalize this connection by showing that differential privacy implies algorithmic stability (Defs. 6, 7) and hence generalization (Wang et al., 2016; He et al., 2021; Kang et al., 2026). Other work makes the privacy–utility tradeoff explicit by relating privacy to excess risk and optimization error (Bassily et al., 2014; 2019; Wang et al., 2024).

However, differential privacy limits the influence of every individual record throughout the learning process, which is a rather strong requirement. We therefore ask whether a learner must itself be differentially private to generalize. Perhaps it is enough that the influence of any single record can be *uniformly removed* after training, producing an output law that is indistinguishable from the output law of the learner retrained without that record. We focus on certified unlearning (CU), which is widely investigated as such a post-training deletion procedure (Guo et al., 2019; Ullah et al., 2021; Chien et al., 2024a; Mu & Klabjan, 2026). This intuition leads to our central question:

Question. Does the existence of a CU algorithm imply algorithmic stability and generalization of the underlying learning algorithm?

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.

To answer this question, we establish the Conversion Theorem 1. For deletion indices $i \neq j$, the theorem bounds the distance between the retrained output laws $P_{\mathcal{A}(\mathcal{D}_N^{-i})}$ and $P_{\mathcal{A}(\mathcal{D}_N^{-j})}$ by the distance between $P_{u^{(i)}}$ and $P_{u^{(j)}}$, plus the distance from $P_{u^{(i)}}$ to $P_{\mathcal{A}(\mathcal{D}_N^{-i})}$ and the distance from $P_{u^{(j)}}$ to $P_{\mathcal{A}(\mathcal{D}_N^{-j})}$. The two distances from unlearned output laws to their corresponding retrained output laws are bounded together by $2r_{T_u,p}$ when the worst-case distance to retraining is available, or by the CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$ when only the CU inequalities and the parameter-domain diameter are available. The Conversion Theorem therefore studies two T_u-step deletion trajectories instead of accumulating the one-record perturbation over the full T-step training trajectory; the tightness theorems show that both upper bounds can be attained.

We next study anchored certified unlearners (Def. 2), in which every deletion run starts from the same checkpoint. The contraction or expansion factor of the deletion update determines the learner-stability bound (Def. 7). Under contraction (Def. 26), the distance between the two deletion output laws and each distance from a deletion output law to its corresponding retrained output law are $O(n^{-1})$, giving learner stability $O(n^{-1})$ uniformly in T. Under non-expansive but non-contractive dynamics, the distance between the two deletion output laws grows linearly with T_u, while the CU-implied distance decreases as $T_u^{-1/2}$; balancing these contributions gives $O(n^{-1/3})$. Under expansion by $E > 1$, the distance between the two deletion output laws grows as E^{T_u}, and each distance from an unlearned output law to its corresponding retrained output law contains a factor of order E^T. Under conditions (C1)–(C4), the expansive construction violates the utility condition once $T \geq T^\dagger$ (Appendix P); the ill-conditioned quadratic construction either has a learner-stability floor when $\beta/\gamma = \Omega(n)$ or forces $T + T_u = \Omega(n^a \log n)$ when $\beta/\gamma = \Theta(n^a)$. These three update regimes show how the deletion procedure's update geometry determines the generalization bound obtained for learner $\mathcal{A}$.

Our contributions are as follows.

1. We prove the Conversion Theorem 1. It converts an (ε, δ) CU comparison and unlearning stability (Def. 10) into learner stability (Def. 7); for a neighboring pair indexed by i and j, the proof uses the comparison of $u^{(i)}$ with $\mathcal{A}(\mathcal{D}_N^{-i})$ and the comparison of $u^{(j)}$ with $\mathcal{A}(\mathcal{D}_N^{-j})$.
2. We derive algorithmic stability bounds for anchored certified unlearners across the three update regimes. Under contraction, directly comparing each unlearned output with its retrained counterpart gives $O(1/n)$ learner stability (Def. 7) uniformly in T; under non-expansive but non-contractive dynamics, the optimized rate is $O(n^{-1/3})$; expansive dynamics retain a factor of order E^T.
3. Rearranging the Conversion Theorem gives conditional impossibility and depth results under (C1)–(C4): the nonsmooth convex family violates the vanishing-unlearning-stability condition, the expansive smooth family violates the utility condition once $T \geq T^\dagger$, and the ill-conditioned quadratic family has either a stability floor or the depth lower bound $T + T_u = \Omega(n^a \log n)$.
4. We test the $O(1/n)$ contractive scaling, the $O(n^{-1/3})$ non-contractive scaling, and the expansive phase boundary on CIFAR-10 and SST-2 (Appendix B).

2 PRELIMINARIES

Notations and definitions used below are recorded here. Related work, samplers, loss assumptions, contractive transition rules, and admissible impossibility profiles appear in Appendices E.1, D, G, L, and P.

2.1 NOTATIONS

Basic notations. Let $\Theta \subseteq \mathbf{R}^d$ be a nonempty closed convex set of diameter $D_\Theta := \mathrm{diam}(\Theta)$. For compactness, the main text writes $D := D_\Theta$ whenever a diameter appears in a bound. Let Π_Θ be Euclidean projection onto Θ. $(\mathcal{Z}, \mathfrak{Z})$ is a standard Borel sample space carrying an unknown distribution $\mathcal{P}$. We write $\mathcal{M}(\Theta)$ for the set of Borel probability measures on Θ. For $p \geq 1$, let $\mathcal{W}_p$ denote the p-Wasserstein distance. Let $\mathcal{L} : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ be jointly Borel measurable, and $\mathcal{L}(\cdot, z)$ is

the loss function. The population risk is $R(\theta) := \mathbb{E}_{z\sim\mathcal{P}}[\mathcal{L}(\theta, z)]$, the empirical risk of a dataset S is $R_S(\theta) := |S|^{-1}\sum_{z\in S}\mathcal{L}(\theta, z)$, and $R^\star := \inf_\Theta R$.

Datasets. Fix $N = n + 1$ and draw $z_1, \ldots, z_N$ i.i.d. from $\mathcal{P}$. The set $\mathcal{D}_N^{-i}$ has size n, while $\mathcal{D}_N^{(i)}$ replaces z_i by an independent copy. The averaged notions use an independent sample $\mathcal{D}'_n$. We assume permutation-invariant learners, so leave-one-out pairs are replacement pairs (Lem. I.1).

Algorithms. We use a family of learners $\{\mathcal{A}_m : \mathcal{Z}^m \mapsto \mathcal{M}(\Theta)\}_{m\geq 1}$. The member $\mathcal{A}_m$ is realized by T steps with step sizes $\eta_t > 0$ (written η when constant). An unlearner is a Markov transition rule $\mathcal{U} : \mathcal{M}(\Theta) \times \mathcal{Z}^m \mapsto \mathcal{M}(\Theta)$ taking a trained model and a forget set to a published law. We suppress oracle access to the full dataset in the notation. Write $u^{(i)} := \mathcal{U}(\mathcal{A}_N(\mathcal{D}_N), \{z_i\})$ and use $\mathcal{A}_n(\mathcal{D}_N^{-i})$ as the retraining reference. When the sample size is clear, $\mathcal{A}$ denotes the relevant member $\mathcal{A}_m$. A point $\theta \in \Theta$ is identified with its Dirac law. We call T_u the unlearning depth. We write E for the worst-case distance-expansion factor of one deletion update; the formal one-step inequality defining E appears in equation 21.

For each deletion index i, let $r_{T_u,p}$ be the worst-case distance to retraining: the largest $\mathcal{W}_p$ distance between the unlearned output law and the law produced by retraining without z_i. For two deletion indices, the direct endpoint term is $2r_{T_u,p}$. When only the CU parameters ϵ, δ and the parameter-space diameter D_Θ are known, the CU inequalities and Lemmas I.2(i) and I.2(iii) give the CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$:

$$r_{T_u,p} := \sup_{\mathcal{D}_N, i} \mathcal{W}_p\Big(P_{u^{(i)}}, P_{\mathcal{A}(\mathcal{D}_N^{-i})}\Big),$$
$$C_{\epsilon,\delta} := \min\{1, \delta + (1-\delta)\tanh(\epsilon/2)\}, \quad P^{\mathrm{TV}}(\epsilon, \delta) := 2D_\Theta C_{\epsilon,\delta}^{1/p}.$$

The CU guarantee bounds the total variation distance between each unlearned output law and its retrained reference. Together with D_Θ, this defines the CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$, which bounds the sum of the two endpoint distances.

2.2 Definitions

The conversion uses three ingredients: the CU certificate, anchored unlearning dynamics, and learner/unlearner stability (Defs. 6, 7, 9, 10). We give the short definitions needed for the main statements here. Appendix D gives the remaining CU, stability, and sampler definitions; Appendix G gives the loss assumptions; Appendix L gives the Wasserstein and KL contraction assumptions; and Appendix P gives the admissible stability profiles used in the impossibility results.

Definition 1 ((ϵ, δ)-certified unlearning)**.** *Let $\epsilon \geq 0$, $\delta \in [0, 1/2]$. $\mathcal{U}$ is (ϵ, δ)-*CU *for $\mathcal{A}$ if for every $\mathcal{D}_N$, every forget set $\mathcal{D}_f \subseteq \mathcal{D}_N$ and every Borel $O \subseteq \Theta$,*

$$\Pr\big[\mathcal{U}(\mathcal{A}(\mathcal{D}_N), \mathcal{D}_f) \in O\big] \leq e^\epsilon \Pr\big[\mathcal{A}(\mathcal{D}_N \backslash \mathcal{D}_f) \in O\big] + \delta,$$

and the reverse inequality with the two probability laws interchanged is also required.

The CU definition uses the retrained output law as the reference law, whereas differential privacy uses the output law of a neighbouring learning run. Bounding the log-density ratio by ϵ pointwise instead gives ϵ-PCU (Def. 5); every CU mechanism with $\delta = 0$ satisfies the pointwise density-ratio condition in Def. 5, which we also use below.

Definition 2 ($(\theta^{\mathrm{a}}, T_u)$-anchored unlearner)**.** *Let $\mathcal{A}_N$ train on $\mathcal{D}_N$ for T updates and let θ_t denote its full-data training states. An anchor is a random state $\theta^{\mathrm{a}} \in \Theta$ that is determined by this training run before the deletion set is known. For a single deletion of z_i, an unlearner is anchored with depth T_u if it starts from the same anchor, $\theta_0^{(i)} = \theta^{\mathrm{a}}$, and performs T_u updates using the retained dataset $\mathcal{D}_N^{-i}$:*

$$\theta_{k+1}^{(i)} = \Pi_\Theta\big(\theta_k^{(i)} - \eta_k g_k^{(i)} + \sqrt{2\eta_k}\,\sigma_{\mathrm{u}}\xi_k\big), \quad \xi_k \overset{iid}{\sim} \mathcal{N}(0, I_d), \tag{1}$$

where $g_k^{(i)}$ is a full-batch or sampled gradient of $R_{\mathcal{D}_N^{-i}}$ at $\theta_k^{(i)}$ and $\sigma_{\mathrm{u}} \geq 0$ is the update-noise scale. The unlearned output is $u^{(i)} := \theta_{T_u}^{(i)}$, possibly followed by the publication mechanism used in Definition 1.

In the rewind-to-delete construction analyzed in this paper, $\theta^{\mathrm{a}} = \theta_{T-T_u}$. The retraining reference starts from the original initialization and runs the learner for all T updates on $\mathcal{D}_N^{-i}$; the unlearner runs only the final T_u retained-data updates from the stored checkpoint. Thus the two procedures use the same retained-data transition only during those final T_u updates. Shared initialization and shared random draws are coupling assumptions used later to derive distance bounds; they are not part of the definition of CU.

Stability parameters (Defs. 6, 7, 8, 9, 10, 12). For $p \geq 1$ set

$$\begin{aligned} \tau &:= \sup_{\mathcal{D}_N, i} \mathcal{W}_p\big(\mathcal{A}(\mathcal{D}_N), \mathcal{A}(\mathcal{D}_N^{(i)})\big), \\ \Delta &:= \sup_{\mathcal{D}_N, i \neq j} \mathcal{W}_p\big(u^{(i)}, u^{(j)}\big), \end{aligned} \tag{2}$$

These are the $\mathcal{W}_p$-uniform stability parameters of the learner (Zhu et al., 2023) (Def. 7) and its unlearning analogue (Def. 10); the unlearning parameter measures how much the identity of the deleted point changes the unlearned output. We also use two analogous versions of the conversion. Averaged stability (Kuzborskij & Lampert, 2018) replaces the Wasserstein suprema by an average over deletion indices and expected loss differences (Defs. 6, 9); KL stability (Dadi & Cevher, 2025) replaces Wasserstein distance by relative entropy (Defs. 8, 12). The symbols τ and Δ denote the learner and unlearner stability parameters for whichever of these three routes is being discussed.

Whenever a bound names L or β, the loss is L-Lipschitz and β-smooth in θ uniformly in z (Defs. 16, 21). Each theorem states any additional function-class assumption that the theorem needs. Main-text bounds are asymptotic in n but retain all parameters carrying qualitative information, and the appendix restates each result with every constant explicit.

3 Converting Unlearning Stability (Defs. 9, 10) to Learning Stability (Defs. 6, 7)

3.1 Certificates Transfer Unlearning Stability (Defs. 9, 10) to Learning Stability (Defs. 6, 7)

We ask whether two properties of a deletion procedure can bound the algorithmic stability of learner $\mathcal{A}$: an (ϵ, δ)-CU comparison between an unlearned output law and the law of the model retrained without the deleted record, and either expected-loss unlearning stability (Def. 9) or Wasserstein unlearning stability (Def. 10). Let $\mathcal{A}$ train on $\mathcal{D}_N$, let $\mathcal{U}$ delete z_i from the trained output, and let $\mathcal{A}(\mathcal{D}_N^{-i})$ be the output of retraining without z_i. For two deletion indices $i \neq j$, the Conversion Theorem 1 bounds the distance between the retrained output laws $P_{\mathcal{A}(\mathcal{D}_N^{-i})}$ and $P_{\mathcal{A}(\mathcal{D}_N^{-j})}$ by the distance between the unlearned output laws $P_{u^{(i)}}$ and $P_{u^{(j)}}$, plus the two distances from each unlearned output law to its corresponding retrained output law. The theorem therefore replaces a learner-stability bound accumulated over learner $\mathcal{A}$'s full T-step training trajectory (Def. 7) with a stability bound for the two T_u-step deletion trajectories (Def. 10).

For the expected-loss route, R^{un} and R^{fg} denote the average loss on the deleted record under the unlearned and corresponding retrained output laws.

Theorem 1 (Conversion Theorem)**.** *Let $\mathcal{U}$ be (ϵ, δ)-CU for $\mathcal{A}$ on single-point deletions. Let $r_{T_u,p}$ and $P^{\mathrm{TV}}(\epsilon, \delta)$ be the quantities defined in the Notations subsection of Section 2. Then:* ***(i)*** *if $|\mathcal{L}| \leq \rho$ and Δ is expected-loss unlearning stability (Def. 9), then $\tau \leq \Delta + 4\rho C_{\epsilon,\delta}$; under ϵ-PCU and $\mathcal{L} \geq 0$, the CU-implied distance is $\tau \leq e^{\epsilon}\Delta + 2\sinh(\epsilon)\,\mathbb{E}[R^{\mathrm{un}} \vee R^{\mathrm{fg}}]$;* ***(ii)*** *if Δ is $\mathcal{W}_p$-unlearning stability (Def. 10), then*

$$\tau \leq \Delta + 2r_{T_u,p} \leq \Delta + P^{\mathrm{TV}}(\epsilon, \delta); \tag{3}$$

(iii) *if Δ is KL-unlearning stability (Def. 12) and the certificate is ϵ-PCU, then $\tau \leq e^{\epsilon}\Delta + \epsilon(1 + \tanh(\epsilon/2))$ and the learner is KL-stable (Def. 8).*

All three conclusions hold uniformly over the datasets and randomness covered by their assumptions. Exact constants and proofs appear in Appendix I.

3.2 Bounding the Distance Between Unlearned and Retrained Models

For two deletion indices $i \neq j$, the Conversion Theorem contains one endpoint distance for each unlearned output:

$$d_i := \mathcal{W}_p\Big(P_{u^{(i)}}, P_{\mathcal{A}(\mathcal{D}_N^{-i})}\Big), \qquad d_j := \mathcal{W}_p\Big(P_{u^{(j)}}, P_{\mathcal{A}(\mathcal{D}_N^{-j})}\Big).$$

There are two ways to bound their sum. If the worst-case distance to retraining can be estimated directly, then $d_i + d_j \leq 2r_{T_u,p}$. If the analysis knows only the (ϵ, δ)-CU inequalities and the diameter D_Θ, Lemmas I.2(i) and I.2(iii) give $d_i + d_j \leq P^{\mathrm{TV}}(\epsilon, \delta) = 2D_\Theta C_{\epsilon,\delta}^{1/p}$. We use the first bound whenever the unlearning and retraining trajectories can be compared; the second is the CU-implied distance justified by the CU guarantee and the diameter.

A concrete case in which we can use the worst-case distance to retraining is an anchored unlearner whose retained-data update is contractive. Full-data training runs for T updates and stores the checkpoint $heta_{T-T_u}$. After z_i is deleted, the unlearning run starts from this checkpoint and applies the retained-data update for T_u steps. To analyze its distance to retraining, consider a run that starts from the original initialization and trains on $\mathcal{D}_N^{-i}$ for all T steps. Its last T_u updates use the same retained-data update. Let b_0 be the distance between the unlearning checkpoint and the retraining state immediately before these last T_u updates. If one retained-data update contracts distances by $q < 1$ under shared random draws, then after T_u updates the two states are at distance at most $q^{T_u} b_0$. A common nonexpansive publication rule does not increase this distance, so

$$r_{T_u,p} \leq q^{T_u} b_0.$$

Thus, the worst-case distance to retraining decreases geometrically with the unlearning depth. In the full-batch setting of Lemma L.2(ii), $b_0 \leq 2L/(\kappa n)$, giving $r_{T_u,p} = O(q^{T_u}/n)$. This direct bound requires the stated comparison between the unlearning and retraining trajectories; the CU guarantee alone does not provide it.

When this comparison is unavailable, the CU guarantee and the diameter may be the only information about the endpoint distances. The next theorem shows that the resulting bound need not decrease with n.

Theorem 2 (CU-implied distance is tight)**.** *For every $D > 0$, $p \geq 1$, and (ϵ, δ) with $C_{\epsilon,\delta} \leq 1$, there are a learner and an (ϵ, δ)-CU unlearner with $\Delta = 0$ and*

$$\tau = P^{\mathrm{TV}}(\epsilon, \delta)/2 = DC_{\epsilon,\delta}^{1/p}$$

for every $n \geq 2$.

The factor two in the direct sum is also necessary, as shown by the following secondary sharpness result.

Proposition 1 (The factor two in the direct bound is necessary)**.** *For every $p \geq 1$, $s > 0$, $\epsilon \in (0, 1]$, and $\delta \in (0, 1/2)$, if a common Gaussian publication mechanism has scale $\sigma \geq s\sqrt{2\ln(1.25/\delta)}/\epsilon$, there are a learner and an anchored CU unlearner with $T_u = 0$, $\Delta = 0$, $r_{T_u,p} = s$, and*

$$\tau = 2s = \Delta + 2r_{T_u,p}. \tag{4}$$

The directly analyzable case is summarized by the following main result.

Theorem 3 (Direct bound for an anchored contractive unlearner)**.** *In the anchored construction above, assume that the final T_u retained-data transitions use shared random draws and contract pointwise by $q := 1 - \kappa\eta \in [0, 1)$, that the state distance before this segment is at most b_0, and that the publication rule is common and nonexpansive. If*

$$b_0 \leq \frac{2L}{\kappa N}(1 - q^{T-T_u}) \leq \frac{2L}{\kappa n},$$

then, for every $p \geq 1$,

$$r_{T_u,p} \leq q^{T_u} b_0, \qquad 2r_{T_u,p} \leq \frac{4L}{\kappa n} q^{T_u}.$$

Finite optimization error in the retraining reference is added to b_0. The next subsection analyzes the other term in the Conversion Theorem, the unlearning stability Δ.

Case	Assumptions on the loss and update	Quantity controlling one deletion update	Bound on unlearning stability Δ	Term added by the Conversion Theorem	Resulting learner-stability bound τ	Example method or lower-bound construction
Metric-contractive updates ($E < 1$; Def. 26): the direct distance to retraining decreases geometrically.						
A	strongly convex loss	$E = 1 - \eta\gamma$	$O(L/\gamma n)$	$O(L(1-\gamma\eta)^{T_u}/\gamma n)$	$O(L/\gamma n)$	PNSGD (Chien et al., 2024a)
C	dissipative, smooth loss; verified contraction	$E = 1 - \eta\kappa$	$O(L/\kappa n)$	$O(L(1-\kappa\eta)^{T_u}/\kappa n)$	$O(L/\kappa n)$	R2D (Mu & Klabjan, 2026)
KL-contractive update: a log-Sobolev condition controls the separate KL comparison.						
D	log-Sobolev condition; bounded-domain rule	KL contraction	$O(1/n)$	$O(\epsilon)$ (no diameter factor)	$O(1/n) + O(\epsilon)$	Langevin (Chien et al., 2024b)
PL and quadratic-growth conditions: quadratic growth converts objective error into distance to the empirical minimizer; PL alone does not imply contraction of the update.						
E	PL + unique minimizer + quadratic growth	Objective-gap control μ_{PL}	Bound through distance to the empirical minimizer	Depends on the worst-case distance to retraining	Conditional on a separate bound for $r_{T_u,p}$	D2D (Neel et al., 2021)
Non-expansive but non-contractive updates ($E = 1$): under the certificate relation used in rows B and F, and with the additive $2D_\Theta\delta$ term controlled, Δ grows as T_u while $P^{\mathrm{TV}}(\epsilon,\delta)$ decreases as $T_u^{-1/2}$; balancing the two terms yields $n^{-1/3}$.						
B	convex, smooth loss	$E = 1$	$O(\eta L T_u/n)$	$P^{\mathrm{TV}}(\epsilon,\delta) = O(\frac{D_\Theta^2}{\sigma\sqrt{T_u}})$	$O(n^{-1/3})$	TV-stable (Def. 13) (Ullah et al., 2021)
F	smooth loss without a global Lipschitz bound	$E = 1$	$O(\eta T_u\sqrt{\beta\Lambda}/n)$	$P^{\mathrm{TV}}(\epsilon,\delta) = O(\frac{D_\Theta^2}{\sigma\sqrt{T_u}})$	$O(n^{-1/3})$ at the interior optimum	Noisy GD
Expansive updates or a stability floor ($E > 1$ or a non-vanishing lower bound): for the families in Appendix P, no depth satisfying (C1)–(C4) makes τ vanish.						
C†	smooth nonconvex	$E = 1 + \eta\beta$	$O(\frac{L}{\beta n}E^{T_u})$	$\Omega(\frac{\eta L}{n}E^{T-1})$	vacuous	None (Thm. 6(ii))
G	convex, nonsmooth loss	$E = 1$; stability floor	$\Omega(\eta L\sqrt{T_u})$	No choice of T_u removes the floor	No useful vanishing bound	None (Thm. 6(i))

Table 1: Learner-stability bounds and obstructions from the Conversion Theorem for anchored certified unlearning. Columns list the assumptions, the quantity controlling one deletion update, the unlearning-stability term Δ, the additional term from the two comparisons with retraining, and the resulting bound on τ. The additional term is $2r_{T_u,p}$ when the two distances to retraining are directly bounded, and is $P^{\mathrm{TV}}(\epsilon,\delta) = 2D_\Theta C_{\epsilon,\delta}^{1/p}$ when only the CU inequalities and D_Θ are available. Metric rows use the one-step factor E in equation 21; row D uses KL contraction and row E uses objective-gap control. Rates suppress fixed problem parameters.

3.3 Bounding Unlearning Stability for Anchored CU

After the endpoint distances are bounded, the remaining term in the Conversion Theorem is the distance between the outputs of two deletion runs. Anchoring makes their initial distance zero. If one deletion update expands a parameter distance by at most E and replacing the deleted record changes one update by at most $2\eta L/n$, shared random draws give

$$d_0 = 0, \qquad d_{k+1} \le E d_k + 2\eta L/n.$$

Proposition N.1 therefore gives

$$\Delta = O\left(\frac{\eta L}{n}\sum_{k<T_u} E^k\right),$$

uniformly in the full training length T. The factor E classifies the three cases.

If $E < 1$, the geometric sum is bounded and $\Delta = O(1/n)$. Together with the direct estimate $r_{T_u,p} = O(q^{T_u}/n)$, this gives an $O(1/n)$ learner stability bound uniformly in T. If $E = 1$, then $\Delta = O(\eta L T_u/n)$. Without a direct distance-to-retraining estimate, Lemma L.3 gives the CU-implied distance of order $T_u^{-1/2}$; balancing the two terms yields $T_u = \Theta(n^{2/3})$ and learner stability $O(n^{-1/3})$. If $E > 1$, the unlearning-stability term grows geometrically with T_u, while the expansive families in Appendix P have a distance-to-retraining term containing a factor E^{T-1}. The impossibility result shows why a sublinear deletion depth cannot make both contributions vanish for those families.

The complete learner-stability bound is obtained by adding the two terms: use $\tau \le \Delta + 2r_{T_u,p}$ when the worst-case distance to retraining is directly controlled, and use $\tau \le \Delta + P^{\mathrm{TV}}(\epsilon,\delta)$ when only the CU inequalities and D_Θ are available. Table 1 records these combinations for reference. The next section adds the PL–QG case, which is not a direct contraction result, and then derives its excess-risk bound.

4 From PL and Quadratic Growth to Excess Risk

Sections 3.2–3.3 already analyze the cases in which the retained-data update itself contracts, does not contract, or expands. This section treats a different situation. A retained empirical risk may satisfy the Polyak–Lojasiewicz (PL) condition and a quadratic-growth condition even when its update rule does not contract parameter distances. The PL condition controls how quickly the empirical objective decreases, while quadratic growth turns that objective decrease into a distance to an empirical minimizer. We first use these two facts to bound learner stability (Def. 7) through the Conversion Theorem, and then convert that stability bound into an excess-risk bound.

4.1 PL AND QUADRATIC GROWTH GIVE LEARNER STABILITY

The following result compares two deletion runs. The first term in the bound comes from the change of the empirical minimizer when one record is replaced. The second term is the error left after the two runs have taken T_u updates from their common anchor.

Theorem 4 (Learner stability under PL and quadratic growth (Def. 7))**.** *Let $\mathcal{U}$ be an anchored (ϵ, δ)-CU unlearner for an L-Lipschitz, β-smooth loss. Suppose every retained empirical risk satisfies the μ_{PL}-PL condition and the matching quadratic-growth condition, has a unique interior minimizer, and satisfies the minimizer-stability condition in Lem. J.8. For full-batch updates with a shared anchor and $\eta \le 1/\beta$, define*

$$\Psi_0 := \max_i \{R_{\mathcal{D}_N^{-i}}(\theta^{\mathrm{a}}) - \min R_{\mathcal{D}_N^{-i}}\}.$$

Then

$$\Delta \le \frac{4L}{\mu_{\mathrm{PL}} n} + 2\sqrt{\frac{2\Psi_0}{\mu_{\mathrm{PL}}}}(1 - \eta\mu_{\mathrm{PL}})^{T_u/2}.$$

If

$$T_u \ge \frac{1}{\eta\mu_{\mathrm{PL}}} \ln \frac{2\Psi_0 \mu_{\mathrm{PL}} n^2}{4L^2},$$

then the Conversion Theorem gives

$$\tau \le \frac{8L}{\mu_{\mathrm{PL}} n} + 2r_{T_u,p}.$$

The PL and quadratic-growth assumptions control the unlearning stability term Δ (Def. 10). They do not by themselves control the distance between an unlearned model and the model retrained without the deleted record. That distance is the separate quantity $r_{T_u,p}$ in the last display. If a direct bound on $r_{T_u,p}$ is unavailable, the Conversion Theorem instead uses the CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$; the corresponding risk statement is obtained by replacing $2r_{T_u,p}$ by $P^{\mathrm{TV}}(\epsilon, \delta)$.

4.2 THE PL STABILITY BOUND GIVES AN EXCESS-RISK BOUND

Let ϵ_{opt} denote the learner's optimization and publication error. For an L-Lipschitz loss, the standard stability-to-generalization lemma for learner stability (Def. 7) applies the previous theorem directly. For a nonnegative smooth loss, a self-bounding inequality gives a sharper form when the losses along the relevant paths are controlled.

Theorem 5 (Excess risk under PL and quadratic growth (Def. 7))**.** *Under the hypotheses of Theorem 4,*

$$\mathbb{E}[R(\theta^{\mathrm{out}})] - R^\star \le \frac{8L^2}{\mu_{\mathrm{PL}} n} + 2Lr_{T_u,p} + \epsilon_{\mathrm{opt}}. \tag{5}$$

If, in addition, $\mathcal{L} \ge 0$ is β-smooth, the loss on the relevant paths is at most Λ, and

$$\Lambda \le c_\Lambda \, \mathbb{E}[R(\theta^{\mathrm{out}})], \qquad n \ge \frac{32 c_\Lambda \beta}{\mu_{\mathrm{PL}}},$$

then

$$\mathbb{E}[R(\theta^{\mathrm{out}})] - R^\star \le \frac{32 c_\Lambda \beta (R^\star + \epsilon_{\mathrm{opt}})}{\mu_{\mathrm{PL}} n} + 4\sqrt{2\beta\Lambda}\, r_{T_u,p} + \epsilon_{\mathrm{opt}}. \tag{6}$$

The first bound has a clear two-part interpretation: the $1/n$ term is the statistical contribution from replacing one record, and $2Lr_{T_u,p}$ is the contribution from the two comparisons with the corresponding retrained models. If only the CU-implied distance is known, replace this latter term by $LP^{\mathrm{TV}}(\epsilon, \delta)$.

In the interpolation case $R^\star = 0$, choosing $T \ge 2 \ln n/(\eta\mu_{\mathrm{PL}})$ makes the deterministic optimization error $O(n^{-2})$. Equation equation 6 then gives

$$\mathbb{E}[R(\theta^{\mathrm{out}})] - R^\star = O(n^{-2}) + 4\sqrt{2\beta\Lambda}\, r_{T_u,p} + \epsilon_{\mathrm{opt}}.$$

Thus the faster risk rate requires both the PL/QG control of learner stability (Def. 7) and a sufficiently small direct distance to the corresponding retrained model. The PL condition alone does not provide the latter.

The condition involving Λ is a separate assumption about the loss values along the two deletion and retraining paths. It is not implied by PL or quadratic growth. If that condition is unavailable but $|\mathcal{L}(\theta, z)| \le \rho$ uniformly, the same argument gives the unconditional PL/QG risk bound

$$\mathbb{E}[R(\theta^{\mathrm{out}})] - R^\star \le \frac{16\beta\rho}{\mu_{\mathrm{PL}} n} + 2\sqrt{2\beta\rho}\, r_{T_u,p} + \epsilon_{\mathrm{opt}}.$$

This version keeps the statistical term explicit and does not claim the faster interpolation rate without the additional loss-control condition. In every display in this section, if only the CU-implied distance is known, replace the term involving $r_{T_u,p}$ by the corresponding term involving $P^{\mathrm{TV}}(\epsilon, \delta)$.

5 When Unlearning Uses Fewer Updates Than Retraining

The reverse direction of the Conversion Theorem gives a necessary condition:

$$\Delta \ge \tau - 2r_{T_u,p}.$$

If the distance to retraining is not directly available, the same argument uses the CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$ and yields a lower bound on the certificate parameters.

Proposition 2 (Contrapositive of the Conversion Theorem)**.** *For the Wasserstein learner-stability parameter τ (Def. 7) and the Wasserstein unlearning-stability parameter Δ (Def. 10), every (ϵ, δ)-CU unlearner satisfies $\Delta \ge \tau - 2r_{T_u,p}$. With the CU-implied distance,*

$$\epsilon/2 + \delta \ge \big((\tau - \Delta)_+/2D\big)^p.$$

To exclude vacuous releases, consider learner–unlearner sequences for which (C1) the certificate parameters vanish, (C2) $\Delta \to 0$, (C3) $T_u/T \to 0$, and (C4) the publication noise stays bounded relative to the parameter domain. The following theorem applies these conditions only to the explicit anchored constructions in Appendix P.

Theorem 6 (Learner-specific impossibility)**.** *Fix $p \ge 1$, $T_u \ge 1$, and a constant step size $\eta > 0$ independent of n. Under the learner-family assumptions in Appendix P and (C1)–(C4): **(i)** the nonsmooth convex family has $\Delta \ge c_1(\eta L\sqrt{T_u} + \eta L T_u/n)$; **(ii)** the expansive smooth family has $s_{T_u} \ge (2\eta L/N)(1+\eta\beta)^{T-1}$ and violates (C4) when $T \ge T^\dagger$; and **(iii)** the centered bounded-domain quadratic family has a stability floor when $\beta/\gamma = \Omega(n)$, while $\beta/\gamma = \Theta(n^a)$ with $a \in (0,1)$ forces $T + T_u = \Omega(n^a \log n)$ under the interior and output-perturbation assumptions of Lem. P.3.*

These conclusions are conditional and family-specific: each construction lower-bounds one term in $\tau \le \Delta + 2r_{T_u,p}$.

Nonsmooth convex family. For the high-dimensional full-batch subgradient construction of Lemma P.1, successive subgradients can be nearly orthogonal. The two deletion runs therefore separate by at least $\eta L\sqrt{T_u} + \eta L T_u/n$. The first term is independent of n, so a fixed positive step size violates (C2); taking almost no deletion updates avoids the floor only by giving up a useful reduction in the distance to retraining.

Expansive smooth nonconvex family. For the quadratic concave loss in Appendix P, the retained-data update has linear part EI with $E := 1 + \eta\beta > 1$. Two deletion runs separate at rate E^{T_u}, while the rewind construction leaves the pre-publication distance to retraining equal to

$$s_{T_u} = \frac{2L}{\beta N}\big(E^T - E^{T_u}\big) \ge \frac{2\eta L}{N} E^{T-1} \qquad (T_u < T),$$

as shown in Lemma P.2. Increasing T_u reduces this distance but increases the deletion-run separation; decreasing T_u leaves the E^T factor. The conversion therefore retains a training-horizon factor, and (C4) fails once $T \ge T^\dagger$.

Ill-conditioned strongly convex family. For the centered quadratic family in Lemma P.3, $\chi = \beta/\gamma = \Omega(n)$ moves the exact empirical minimizer by a constant amount inside the bounded domain, giving a learner-stability floor. When $\chi = \Theta(n^a)$ with $a \in (0, 1)$, the minimizer shift vanishes but the slow eigendirection requires $\Theta(\chi \log n)$ updates. The certificate and interior-trajectory conditions then force $T + T_u = \Omega(n^a \log n)$, so the unlearner cannot use fewer updates while retaining the stated utility. The endpoint $a = 0$ is logarithmic, while $a \geq 1$ is in the stability-floor regime. These lower bounds identify the quantities a positive result must control: contractive deletion dynamics, a small distance to retraining, or a training-scale update cost.

6 Experiments and Conclusion

Experiments. *Experimental setting.* We evaluate the quantities in the Conversion Theorem on two feature pipelines. For CIFAR-10, a convolutional encoder is frozen and its 2048-dimensional features are cached; for SST-2, a frozen DistilBERT encoder supplies 768-dimensional [CLS] features. Most experiments train a linear head, while the scale experiment fine-tunes LoRA adapters. For each selected deletion index, we run the learner, the unlearner, and the corresponding retraining from the same initialization and with the same data order. We record the distance between the unlearned model and its retrained reference, the leave-one-out sensitivity between full-data and deleted-data outputs, validation accuracy, and excess risk. The deletion indices form a fixed finite pool: the reported quantities are medians or maxima over that pool, with bands across the listed random seeds. These finite-pool measurements are empirical diagnostics, not worst-case certificates or tests of every theorem assumption. The first comparison varies the training horizon on the same frozen feature head. With no weight decay, the leave-one-out sensitivity keeps increasing through $T = 100$ on both datasets. With weight decay 1, the sensitivity rises initially and then levels off (Figures B.1 and B.2). This pattern is consistent with the difference between a non-contractive update and a contractive update. The second comparison separates the two endpoint contributions in the conversion. For a strongly convex head trained for $T = 300$ updates, we compare the bound obtained from fixed CU parameters with the bound obtained from the measured worst-case distance to retraining. On CIFAR-10 the median distance over the deletion pool has log-log slope about -0.96; on SST-2 the slope is about -1.01. These median-pool slopes are empirical diagnostics rather than estimates of the worst-case $r_{T_u,p}$. The CU-implied distance stays essentially flat when its CU parameters are held fixed, whereas the direct distance decreases with n; the measured risk follows the decreasing curve (Figures B.3 and B.4). We also test whether the contractive signature persists beyond linear heads. In the LoRA comparison, the noised arm has an approximately flat sensitivity curve while the un-noised arm grows with n; the reported slopes are -0.02 versus $+0.63$ on SST-2 and 0.00 versus $+0.22$ on CIFAR-10, with validation accuracy reported alongside sensitivity (Figures B.5 and B.6). A sequential-deletion experiment then compares $T_u = 2$ with $T_u = 20$. Both curves approach a plateau. Dividing the late-stream sensitivity for $T_u = 2$ by that for $T_u = 20$ gives 4.90 on CIFAR-10 and 5.38 on SST-2 (Figures B.9 and B.10), close to the ratio predicted by Cor. Q.1. Finally, the supporting experiments examine the two qualitative risk and obstruction statements. The high-dimensional nonsmooth head develops a sensitivity floor and grows with T_u at the $\sqrt{T_u}$ scale described in Thm. 6(i), while a one-dimensional smooth control continues to decrease with n; this pattern appears on both feature pipelines (Figures B.11 and B.12). When the publication-noise scale is recomputed from the measured sensitivity at each n, excess risk decreases with n instead of following the risk curve obtained with a fixed noise scale (Figures B.13 and B.14).

Conclusion. Under the Conversion Theorem assumptions, CU certification and unlearning stability control learner stability through the deletion-run distance and the two distances to the corresponding retrained models. The direct endpoint term $2r_{T_u,p}$ has an unavoidable coefficient two, while the CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$ is sharp when only a common finite support diameter is known. Shared- checkpoint contraction gives $O(1/n)$ stability uniformly in T; under every condition of Corollary N.2, the non-contractive trade-off is the conditional $O(n^{-1/3})$ rate. PL and quadratic growth control objective gaps and unlearning stability but not the endpoint distance to finite-step retraining, and the reverse inequality yields family-specific obstructions. The finite-pool CIFAR-10 and SST-2 experiments illustrate the resulting sensitivity, distance, and risk patterns.

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 and all experimental results are submitted to the Supplementary Material on OpenReview.

REFERENCES

Jason Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *Advances in Neural Information Processing Systems*, 35:3788–3800, 2022.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Raef Bassily, Om Thakkar, and Abhradeep Guha Thakurta. Model-agnostic private learning. *Advances in neural information processing systems*, 31, 2018.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 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.

Hanna Benarroch, Jamal Atif, and Olivier Cappé. Certified per-instance unlearning using individual sensitivity bounds. *arXiv preprint arXiv:2602.15602*, 2026.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626. PMLR, 2020.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *2015 IEEE 56th annual symposium on foundations of computer science*, pp. 634–649. IEEE, 2015.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International conference on machine learning*, pp. 745–754. PMLR, 2018.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 155–186. JMLR Workshop and Conference Proceedings, 2011.

Eli Chien and Pan Li. Convergent privacy loss of noisy-sgd without convexity and smoothness. In *International Conference on Learning Representations*, volume 2025, pp. 12238–12260, 2025.

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, 2024a.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. *Advances in neural information processing systems*, 37: 79666–79703, 2024b.

Leello Tadesse Dadi and Volkan Cevher. Generalization of noisy sgd in unbounded non-convex settings. In *Forty-second International Conference on Machine Learning*, 2025.

Damek Davis and Dmitriy Drusvyatskiy. Graphical convergence of subgradients in nonconvex optimization and learning. *Mathematics of Operations Research*, 47(1):209–231, 2022.

Polina Dolgova and Sebastian U Stich. Sequential subspace noise injection prevents accuracy collapse in certified unlearning. *arXiv preprint arXiv:2601.05134*, 2026.

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

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3):851–886, 2016.

Murat Bilgehan Ertan and Marten van Dijk. Fundamental limitations of favorable privacy-utility guarantees for dp-sgd. *arXiv preprint arXiv:2601.10237*, 2026.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on learning theory*, pp. 1270–1279. PMLR, 2019.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532. IEEE, 2018.

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

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Fengxiang He, Bohan Wang, and Dacheng Tao. Tighter generalization bounds for iterative differentially private learning algorithms. In *Uncertainty in Artificial Intelligence*, pp. 802–812. PMLR, 2021.

Yiyang Huang and Clément L Canonne. Tight bounds for machine unlearning via differential privacy. *arXiv preprint arXiv:2309.00886*, 2023.

Yilin Kang, Jian Li, Yong Liu, and Weiping Wang. Stability and generalization of differentially private minimax problems. *Neurocomputing*, pp. 134363, 2026.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34:5065–5076, 2021.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *international conference on machine learning*, pp. 2815–2824. PMLR, 2018.

Christoph H Lampert and Hossein Zakerinia. From privacy to generalization: Linear max-information bounds for dp-sgd. *arXiv preprint arXiv:2605.26222*, 2026.

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 191–227. PMLR, 2023.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *international conference on machine learning*, pp. 5809–5819. PMLR, 2020.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021.

Xuechen Li, Daogao Liu, Tatsunori B Hashimoto, Huseyin A Inan, Janardhan Kulkarni, Yin-Tat Lee, and Abhradeep Guha Thakurta. When does differentially private learning not suffer in high dimensions? *Advances in Neural Information Processing Systems*, 35:28616–28630, 2022.

Chong Liu, Yuqing Zhu, Kamalika Chaudhuri, and Yu-Xiang Wang. Revisiting model-agnostic private learning: Faster rates and active learning. *Journal of Machine Learning Research*, 22 (262):1–44, 2021.

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.

Yi-An Ma, Teodor Vanislavov Marinov, and Tong Zhang. Dimension independent generalization of dp-sgd for overparameterized smooth convex optimization. *arXiv preprint arXiv:2206.01836*, 2022.

Siqiao Mu and Diego Klabjan. Descend or rewind? stochastic gradient descent unlearning. *arXiv preprint arXiv:2511.15983*, 2025.

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.

Konstantinos E Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios S Kalogerias. Beyond lipschitz: Sharp generalization and excess risk bounds for full-batch gd. *arXiv preprint arXiv:2204.12446*, 2022.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–494. IEEE, 2016.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for markov chains via wasserstein distance. *Bernoulli*, pp. 2610–2639, 2018.

Tom Sander, Pierre Stock, and Alexandre Sablayrolles. Tan without a burn: Scaling laws of dp-sgd. In *International Conference on Machine Learning*, pp. 29937–29949. PMLR, 2023.

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.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pp. 2638–2646. PMLR, 2021.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pp. 4126–4142. PMLR, 2021.

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.

Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Differentially private stochastic gradient descent with low-noise. *Neurocomputing*, 585:127557, 2024.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *Journal of Machine Learning Research*, 17(183):1–40, 2016.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Jiaming Zhang, Huanyi Xie, Meng Ding, Shaopeng Fu, Jinyan Liu, and Di Wang. Understanding the impact of differentially private training on memorization of long-tailed data. *arXiv preprint arXiv:2602.03872*, 2026.

Bowei Zhu, Shaojie Li, Mingyang Yi, and Yong Liu. Stability and sharper risk bounds with convergence rate $\tilde{O}(1/n^2)$, 2025. URL https://arxiv.org/abs/2410.09766.

Lingjiong Zhu, Mert Gurbuzbalaban, Anant Raj, and Umut Simsekli. Uniform-in-time wasserstein stability bounds for (noisy) stochastic gradient descent. *Advances in Neural Information Processing Systems*, 36:1794–1832, 2023.

A Notation and Roadmap

The main text uses abbreviated notation, whereas the appendix gives the full notation and definitions. Notation matches Section 2; the contraction rate is γ under strong convexity. PL is used only to bound the empirical objective gap and the distance from an unlearned model to the corresponding empirical minimizer; PL alone does not contract the parameter transition rule. The main text diameter D is written $D_\Theta = \operatorname{diam}(\Theta)$ here, and the expansion coefficient E is written $\mathcal{E}$, as in equation 21. The term $2r_{T_u,p}$ uses the distance to retraining; $P^{\mathrm{TV}}(\epsilon, \delta)$ instead follows from the CU guarantee and D_Θ.

Every main text claim is restated here with explicit constants before its proof. Theorem 1 is Theorem I.2 in Appendix I; the sharpness result for the CU bound using only D_Θ is Theorem Q.2, and the direct-bound sharpness result is Proposition Q.1; Theorem 3 is Theorem N.4. Proposition N.1 is Proposition N.1. The metric-contraction theorem uses Theorems N.1 and N.3; the PL minimizer-distance result is Theorem N.6. The non-contractive counterparts are Theorem N.2 and Corollary N.2. Theorem N.5, the risk results, the impossibility theorem, experiments, related work, matching algorithms, and the elevation map are restated in their corresponding appendix sections.

B Experimental Settings

B.1 Experimental Scope and Limitations

Thm. 6 and the two tightness theorems concern worst-case instances. Experiments provide empirical checks and cannot establish these theorems. Each experimental setting gives a prediction, a measurable proxy, and a refutation condition. Agreement on CIFAR-10 and SST-2 suggests that the signature is not limited to one substrate under these experimental settings.

(a) Boundary location. Table 1 predicts growth of sensitivity of the worst-case distance to retraining for non-contractive dynamics, while its contractive rows predict saturation for a contractive mechanism. The experimental setting in B.2 compares these two regimes on the same frozen head.

(b) Budget choice. Thm. 2 gives an example in which $P^{\mathrm{TV}}(\epsilon, \delta)$ does not improve with n when the CU parameters and diameter are fixed. Prop. 1 shows that the coefficient 2 in $2r_{T_u,p}$ is exact. Under contraction, Thm. 3 gives $2r_{T_u,p} = O(1/n)$. The experimental setting in B.3 compares both bounds for one published model: it tests whether the measured distance to retraining decays near slope -1, whether the bound from fixed CU parameters stays flat, and which curve the measured risk follows.

(c) Attainable contraction. Rows A–D of Table 1 use the contraction, smoothness, dissipativity, or KL assumptions listed for those rows. The PL–QG risk statement is given separately in Thm. 5 for row E. The experimental setting in B.4 measures noisy and un-noised LoRA controls. The experimental setting in B.5 measures the geometric compounding in Cor. Q.1. The experimental setting in B.6 measures the dimensional pattern in Thm. 6(i). The experimental setting in B.7 tests whether the fixed point in Prop. Q.2 remains a risk floor after recomputing σ from the measured sensitivity.

Every headline panel is reported on both CIFAR-10 and SST-2. Notes B–E give closed-form slopes, plateau ratios, and accountant gaps before the same measurements are repeated on frozen feature heads and LoRA adapters. We measure both the distance to the corresponding retrained model and the leave-one-out distance by explicit retraining on a fixed deletion pool with a shared seed and data order. A privacy accountant is not used as evidence for a dynamical property. The measurements in experimental settings B.2–B.7 are empirical checks under those experimental settings; they do not verify the worst-case assumptions of the theorems.

B.2 Headline Experimental Setting 1. Locating the Boundary Between the Contractive and Non-Contractive Regimes

Objective. Check the non-contractive versus contractive dichotomy in Table 1 on real data. The theory predicts growth of sensitivity of the worst-case distance to retraining with T in the unregularized convex head and saturation under the contractive assumptions in row A.

Design. Train a linear head on frozen features under an unregularized arm with weight decay 0 and a weight-decay arm with weight decay 1, which supplies strong convexity near the solution as in Combination A of Table 1. Record leave-one-out sensitivity $\hat{s} = \max_i \|u^{(i)} - \mathcal{A}(\mathcal{D}_N^{-i})\|$ against the number of training updates T on both CIFAR-10 and SST-2.

Critical comparison. The unregularized control represents the non-contractive $E = 1$ regime and should keep increasing in $\hat{s}$ with T. The weight-decay control from Combination A should rise briefly and then plateau. The same comparison is made on CIFAR-10 and SST-2. Note B gives the corresponding closed-form prediction.

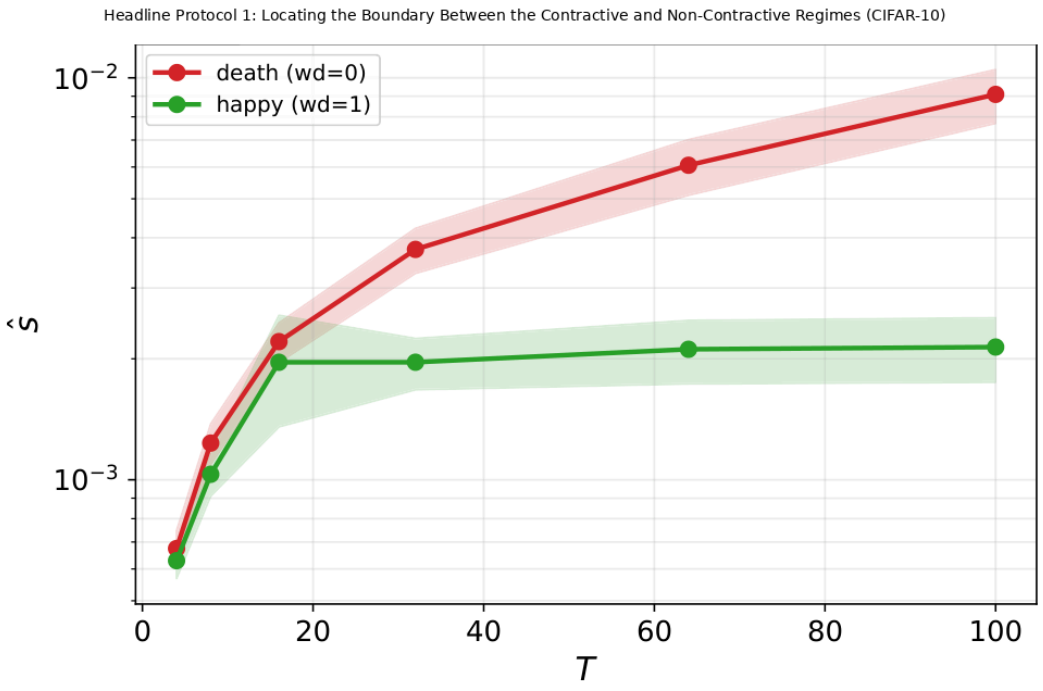


Figure B.1: Experimental setting B.2 on CIFAR-10 frozen features. The unregularized arm uses weight decay 0; the weight-decay arm uses 1. Curves show mean leave-one-out sensitivity against number of training updates T with seed bands.

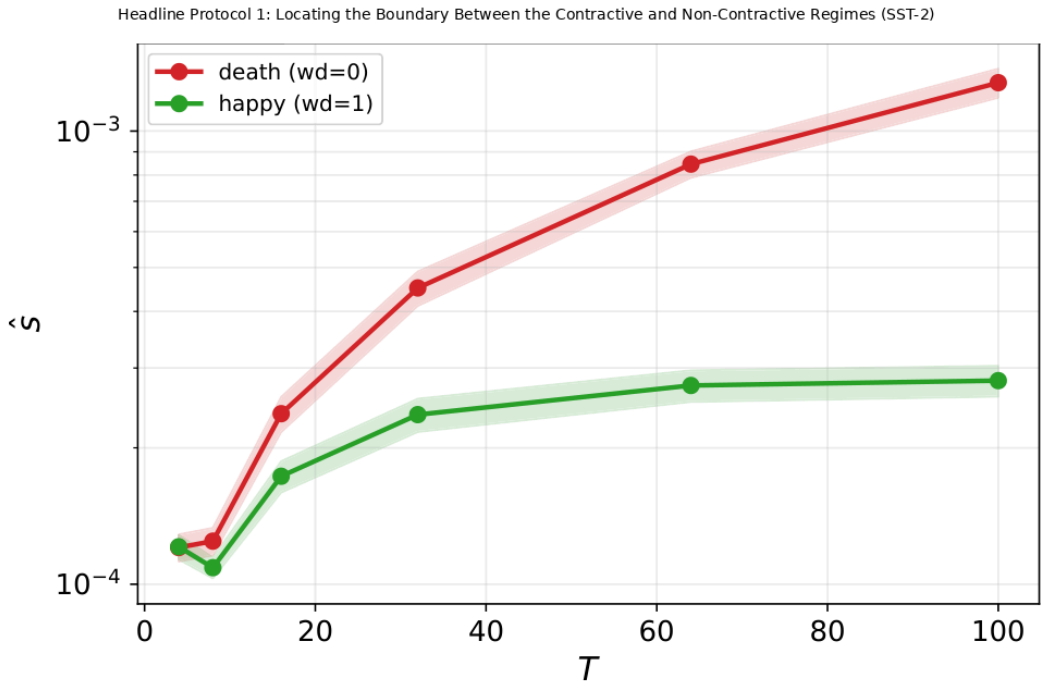


Figure B.2: Experimental setting B.2 on SST-2 DistilBERT frozen [CLS] features under the same unregularized and weight-decay arms as Figure B.1.

Outcome. Figure B.1 and Figure B.2 show a similar qualitative pattern. The unregularized arm keeps increasing in $\hat{s}$ through $T = 100$, so sensitivity of the worst-case distance to retraining does not decrease with longer training. The weight-decay arm rises briefly and then plateaus, consistent with the contractive behavior in row A of Table 1. The two substrates are consistent with the controlled quadratic in Note B under these experimental settings.

Setup. On CIFAR-10 a CNN encoder is frozen and its 2048-dimensional features are cached once. On SST-2 a DistilBERT encoder is frozen and its 768-dimensional [CLS] features are cached once with sequence length 64. A linear head is then trained on the cache for each horizon $T \in \{4, 8, 16, 32, 64, 100\}$ at $n = 10^4$ with learning rate 0.05. The unregularized arm sets weight decay to 0; the weight-decay arm sets it to 1. For every horizon and arm we delete each of 16 fixed indices in turn, retrain from the same initialization and data order, and record the leave-one-out sensitivity $\hat{s}$ over seeds $\{0, 1\}$.

Refutation. An unregularized arm that saturates would be consistent with contraction rather than the non-contractive prediction. A weight-decay arm that keeps climbing would refute the contractive control.

B.3 Headline Experimental Setting 2. Comparing the Two Endpoint-Distance Bounds

Objective. Convert the distinction between the training-level and worst-case-distance comparisons of Sec. 3.2 into a measured utility gap on one published model. Thm. 2 predicts a flat training-level distance bound. Thm. 3 predicts an $O(1/n)$ bound from the worst-case distance to retraining under contraction. The contractive risk result in Thm. O.1 then predicts that measured risk should follow the curve based on that distance rather than the flat one.

Design. Freeze the encoder and train a strongly convex linear head on both substrates. Compare $L\, P^{\mathrm{TV}}(\epsilon, \delta)$, calculated from fixed training-level CU parameters, with the bound from the measured distance to retraining in Thm. N.4, while sweeping n and ϵ.

Critical comparison. The fixed curve should stay flat in n, as in Thm. 2. The curve for the worst-case distance to retraining should decay near slope -1, as in Thm. 3 and Combination A of Table 1.

Measured risk should track the curve for the worst-case distance to retraining. Note D gives the corresponding closed-form ordering.

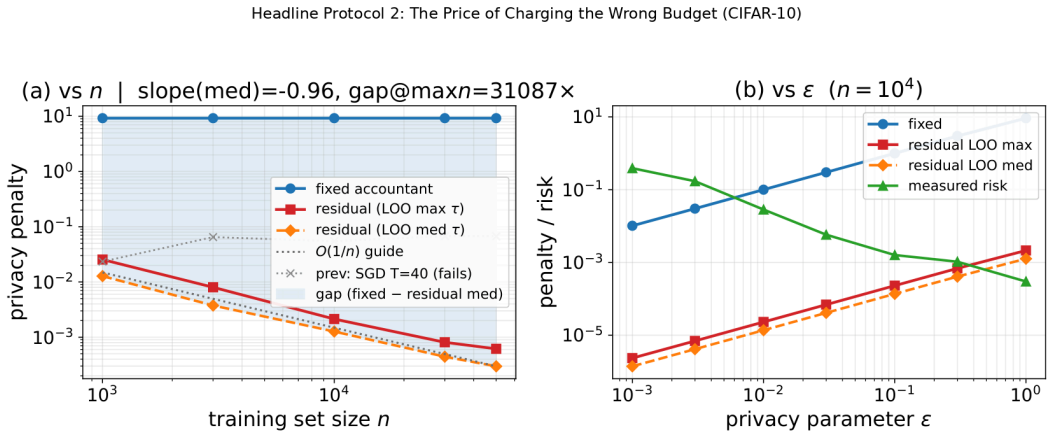


Figure B.3: Experimental setting B.3 on CIFAR-10 frozen features with a strongly convex head. Left is the distance bound against n at $\epsilon = 1$. Right is the distance bound and measured risk against ϵ at $n = 10^4$.

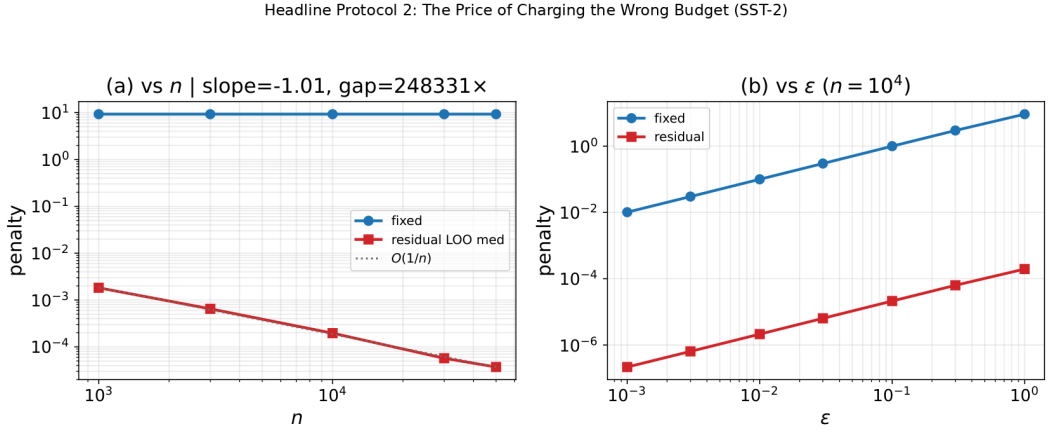


Figure B.4: Experimental setting B.3 on SST-2 DistilBERT frozen features under the same accountants as Figure B.3.

Outcome. Figure B.3 shows a worst-case distance to retraining median LOO slope of about -0.96 on CIFAR-10 and a gap of about $3.1 \cdot 10^4$ between fixed distance bounds and bounds based on the worst-case distance to retraining at $n = 5 \cdot 10^4$. Figure B.4 shows a worst-case distance to retraining slope of about -1.01 on SST-2 and a larger gap at the same n. The fixed curves stay flat, while the slopes of the worst-case distance to retraining are close to the $O(1/n)$ prediction of Thm. 3. Measured risk follows the curve for the worst-case distance to retraining under these experimental settings. The two substrates are consistent with the comparison in Sec. 3.2; Note D gives the closed-form ordering.

Setup. Both substrates freeze the encoder and train a linear head by full batch gradient descent for $T = 300$ steps with weight decay $\lambda = 1$ and gradient tolerance 10^{-4}. The sample sizes are $n \in \{10^3, 3{\cdot}10^3, 10^4, 3{\cdot}10^4, 5{\cdot}10^4\}$ at privacy budget $(\epsilon, \delta) = (1, 10^{-5})$, and an ϵ sweep at $n = 10^4$ uses $\epsilon \in \{10^{-3}, 3{\cdot}10^{-3}, 10^{-2}, 3{\cdot}10^{-2}, 10^{-1}, 3{\cdot}10^{-1}, 1\}$. Unlearning depth is $T_u = \lceil T/2 \rceil$. Leave-one-out τ is measured over 24 forget indices on CIFAR-10 and 16 on SST-2 with seeds $\{0, 1\}$. The fixed distance bound uses the training-level sensitivity. The accountant using the bound based on the worst-case distance to retraining recomputes σ from the measured median leave-one-out sensitivity at each n, as in Thm. N.4.

Refutation. Measured risk that tracks the fixed curve would indicate that the noise scale was not recomputed at each n.

B.4 Headline Experimental Setting 3. Contraction Is Attainable at Scale

Objective. Check that dissipativity with nondegenerate noise from row C of Table 1 is attainable on modern adapters. The contractive row and Theorem N.3 predict that a contractive noisy transition rule keeps leave-one-out $\hat{\tau}$ from growing with n. Row $C^{\dagger}$ predicts that the same template without noise is expansive and unstable.

Design. Fine tune LoRA adapters on SST-2 DistilBERT and CIFAR-10 ViT-Base under a shared projected noisy SGD template. Compare a noised arm with Rem. 2 faithful isotropic noise against an un-noised control, and plot leave-one-out $\hat{\tau}$ against n together with validation accuracy.

Critical comparison. The noised arm is expected to stay flat or fall while remaining above chance accuracy. The un-noised arm is expected to rise. This comparison is an empirical proxy for row C versus row $C^{\dagger}$. Rem. 2 predicts that dimension can mask the classical slope -1, so the measured slope is reported directly. Persistence of the comparison at full SST-2 size is the scale check.

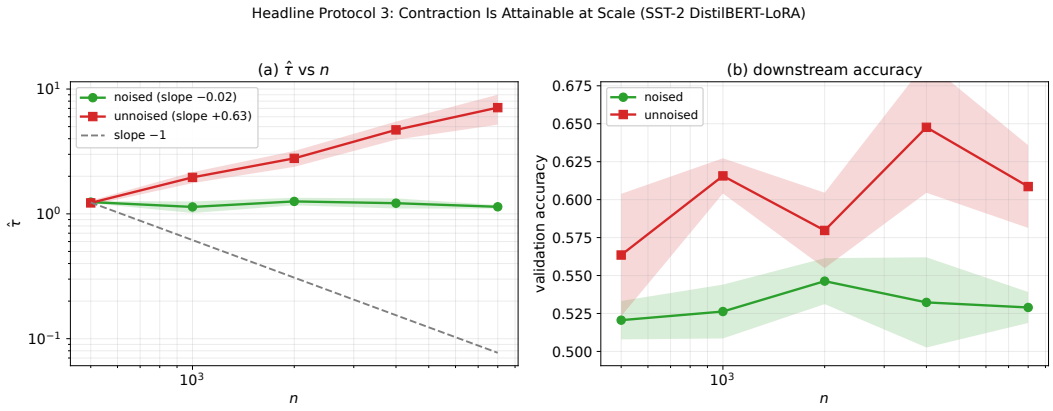


Figure B.5: Experimental setting B.4 on SST-2 DistilBERT-LoRA. Left is $\hat{\tau}$ against n. Right is validation accuracy.

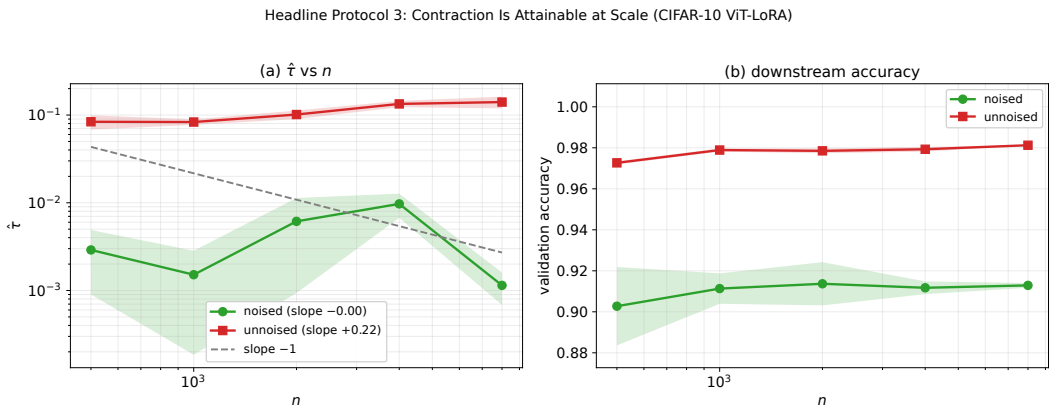


Figure B.6: Experimental setting B.4 on CIFAR-10 ViT-LoRA under the same projected noisy SGD template as Figure B.5.

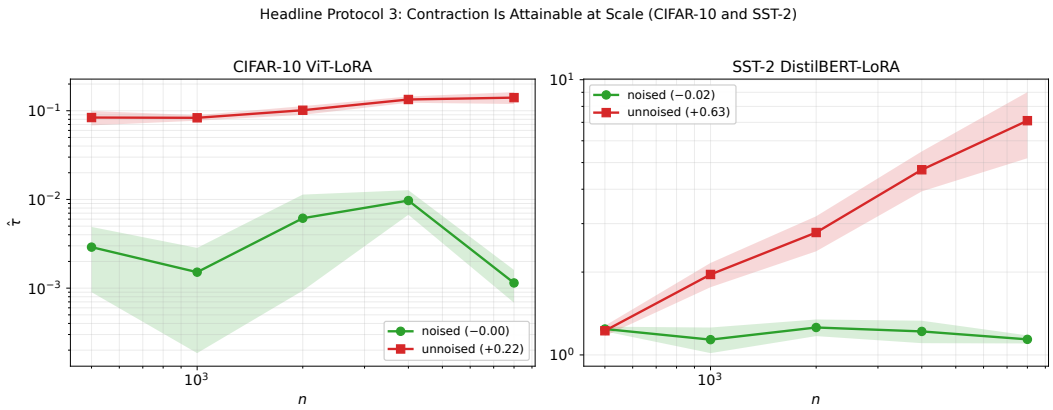


Figure B.7: Side by side $\hat{\tau}$ against n for CIFAR-10 ViT-LoRA and SST-2 DistilBERT-LoRA from experimental setting B.4.

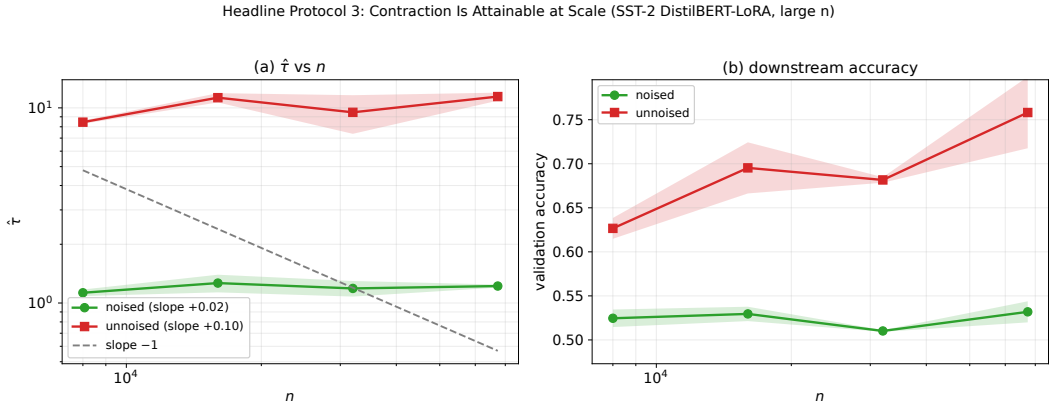


Figure B.8: Experimental setting B.4 large n companion on full SST-2 DistilBERT-LoRA with n up to 67349.

Outcome. Figure B.5 shows an SST-2 noised slope of about -0.02 at mean accuracy about 0.53 and an un-noised slope of about $+0.63$ at mean accuracy about 0.60. Figure B.6 shows CIFAR-10 slopes of about 0.00 and $+0.22$ with mean accuracies about 0.91 and 0.98. Figure B.7 displays the comparison on one row, and Figure B.8 shows the same ordering to $n = 67349$. These measurements are consistent with the row C and row C† predictions under the experimental setting. Accuracy remains above chance. The 2-step noised slopes are compatible with the dimension effect in Rem. 2.

Setup. On SST-2 we fine tune DistilBERT with LoRA rank $r = 2$ on the query and value projections, learning rate $\eta = 0.1$, L2 ball radius 10, batch size 16, and 4 epochs. The noised arm injects isotropic Gaussian noise with Euclidean norm $\|\xi\|_2 = \sigma\sqrt{2\eta}$ at $\sigma = 2.2$. The un-noised arm is identical with $\sigma = 0$. We sweep $n \in \{500, 1000, 2000, 4000, 8000\}$ with 3 seeds and 8 forget indices, and a large n companion extends to $n \in \{8000, 16000, 32000, 67349\}$ with 2 seeds. On CIFAR-10 we fine tune ViT-Base with patch size 16, LoRA rank $r = 4$ on the query and value projections, and a jointly trained classification head, using learning rate 0.1, radius 10, batch size 32, and a fixed budget of 400 steps. The noised arm uses the same L2 noise law at $\sigma = 7.0$ on the small n grid with 2 seeds and 6 forget indices. Every run starts from a shared initialization and data order. Leave-one-out $\hat{\tau}$ is the mean parameter distance between full and deleted runs, and validation accuracy is recorded so that a flat curve is not purchased by collapse to chance.

Refutation. A slope near -1 in the un-noised arm would indicate contraction without injected noise. Accuracy near chance in the noised arm would indicate that the chosen noise scale fails the utility condition.

B.5 Headline Experimental Setting 4. Compounding over a Real Deletion Stream

Objective. Check Cor. Q.1 on a sequential deletion stream. The theory predicts geometric approach to a plateau and a ratio between 2-step and 20-step unlearning controlled by contraction depth, so 2-step unlearning must pay a larger compounding cost.

Design. Run a sequential deletion stream on a strongly convex frozen head and compare 2-step unlearning with 20-step unlearning. Notes C and E give the closed-form stream and condition-number predictions used for comparison.

Critical comparison. Geometric convergence predicts horizontal plateaus. The late stream 2-step-to-20-step ratio is compared with the Note E target of 6.64 at condition number 10. A ratio that does not change with T_u would refute the compounding model. CIFAR-10 and SST-2 provide two substrate checks.

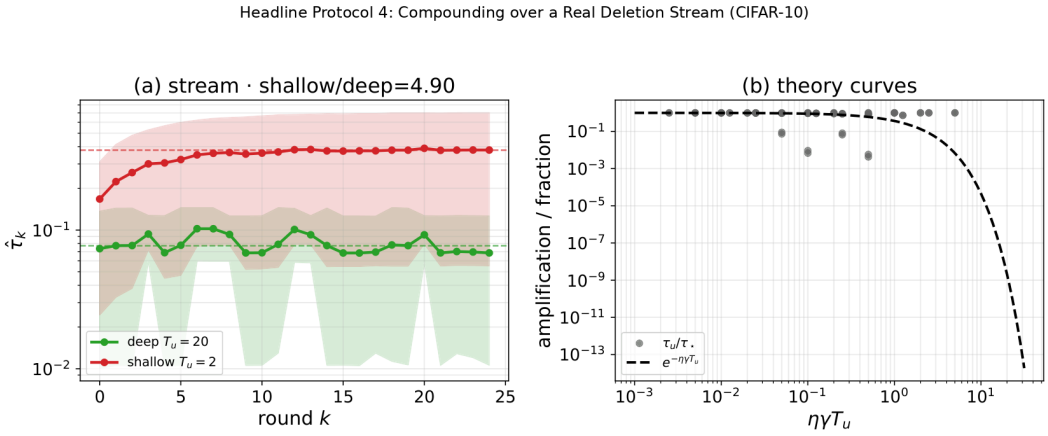


Figure B.9: Experimental setting B.5 on CIFAR-10. Left is $\hat{\tau}_k$ against round index for 2-step and 20-step T_u. Right is the collapse companion.

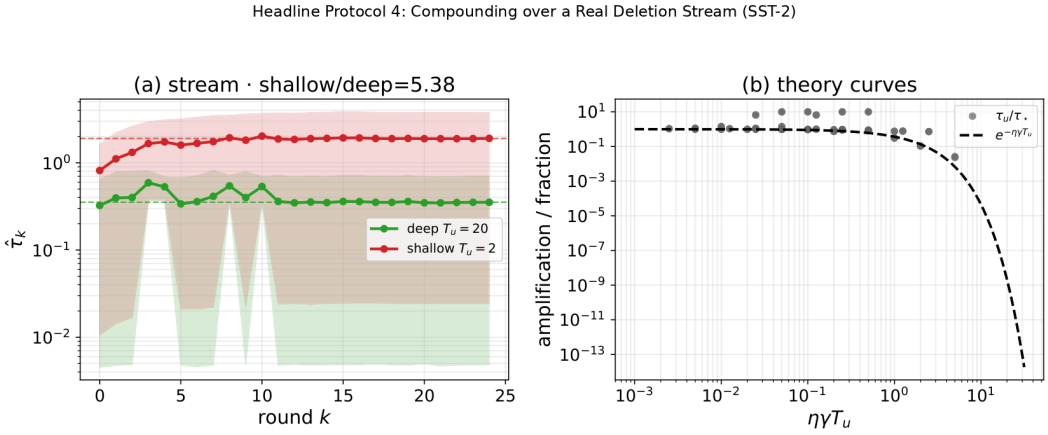


Figure B.10: Experimental setting B.5 on SST-2 under the same 2-step and 20-step depths as Figure B.9.

Outcome. Figure B.9 gives a late stream 2-step-to-20-step plateau ratio of about 4.90 on CIFAR-10; Figure B.10 gives about 5.38 on SST-2. Both arms form approximate plateaus. The 2-step arm is above the 20-step arm, and the ratios are near the Note E target 6.64. These measurements are consistent with the compounding prediction for B.5.

Setup. A strongly convex linear head is warm started for $T_{\text{init}} = 120$ steps at $n = 10^4$ with learning rate 0.05 and weight decay 1.0 on the frozen CIFAR-10 or SST-2 feature cache. We then

run $K = 25$ sequential deletions. Each round re-anchors at the previous published model, applies Gaussian publish noise chosen for $(\epsilon, \delta) = (1, 10^{-5})$, and compares the unlearned model against an explicit retained set reference. 2-step unlearning uses $T_u = 2$ updates. 20-step unlearning uses $T_u = 20$ updates. The late stream plateau ratio between 2-step and 20-step unlearning is the reported statistic, averaged over seeds $\{0, 1\}$.

Refutation. Plateaus that do not move when T_u changes would mean compounding is not the operative mechanism.

B.6 Supporting Experimental Setting 5. The Nonsmooth Barrier Is Dimensional

Objective. Check Thm. 6(i) on feature space heads. The theorem attributes the nonsmooth barrier to dimension through the orthogonal subgradient construction of (Bassily et al., 2020), so a floor must appear at direct feature dimension and must vanish at $d = 1$.

Design. Compare a high dimensional nonsmooth hinge style head with a one dimensional smooth strongly convex control on the same frozen features. The floor predicted by Thm. 6(i) should appear only in high dimension.

Critical comparison. The high-d arm is expected to approach a floor and to increase with T_u near a $\sqrt{T_u}$ reference, following the geometry used in Thm. 6(i). The $d = 1$ control is expected to continue near slope -1 without a floor. This contrast tests the dimensional attribution.

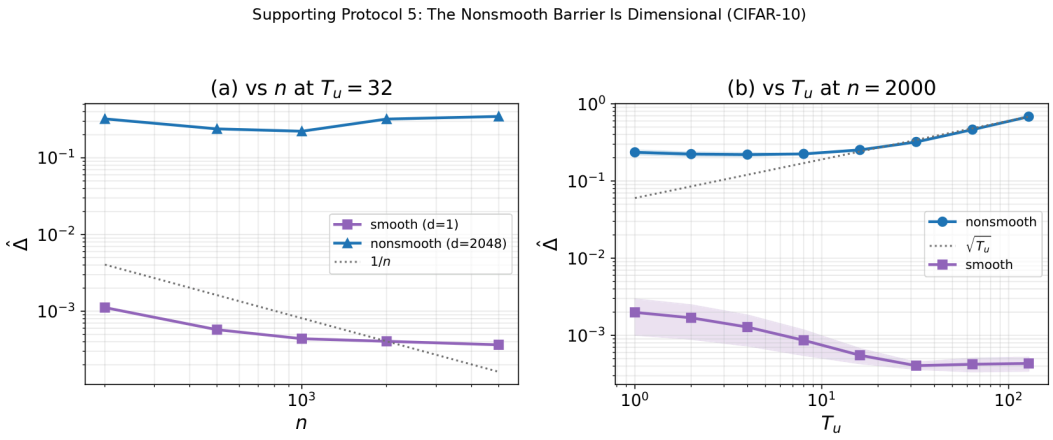


Figure B.11: Experimental setting B.6 on CIFAR-10. Smooth $d = 1$ control versus nonsmooth high d head, with a T_u companion.

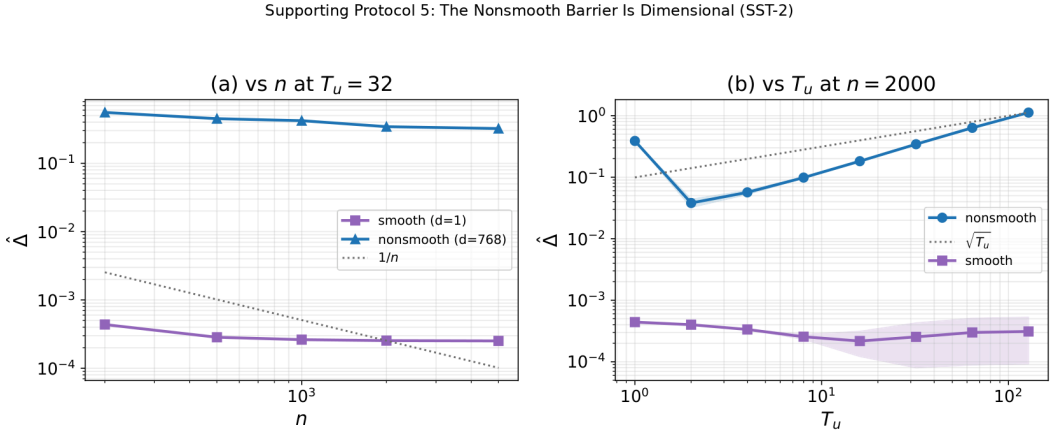


Figure B.12: Experimental setting B.6 on SST-2 with $d = 1$ smooth control versus $d = 768$ nonsmooth head.

Outcome. Figures B.11 and B.12 show the $d = 1$ control near slope -1 without a floor. The high-d arm approaches a higher floor and increases with T_u near the $\sqrt{T_u}$ reference. The two substrates are consistent with a dimensional effect under this experimental setting; they do not establish the worst-case theorem.

Setup. The nonsmooth arm is a hinge style head in the direct feature dimension, namely $d = 2048$ on CIFAR-10 and $d = 768$ on SST-2. The smooth control is a one dimensional strongly convex head on the same cache. Both arms use projected steps with learning rate 0.05 and radius 10 for $T = 150$ steps. We estimate pairwise $\hat{\Delta}$ over 12 index pairs while sweeping $n \in \{200, 500, 1000, 2000, 5000\}$ at fixed $T_u = 32$ and sweeping $T_u \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ at fixed $n = 2000$, with seeds $\{0, 1\}$.

Refutation. A floor at $d = 1$ would mean the obstruction is not the dimensional one claimed in Thm. 6(i).

B.7 Supporting Experimental Setting 6. A Falsification Arm Against Our Own Fixed Point

Objective. Check Prop. Q.2 against a natural misreading of equation 87. The fixed point bounds the stream, but under noise chosen from the worst-case distance it is not a floor on excess risk. The contractive risk result in Thm. O.1 still predicts decay with n at fixed (ϵ, δ).

Design. Using the same strongly convex head substrate as experimental setting B.3, plot excess risk against n at a fixed privacy budget with noise recomputed from the measured sensitivity. A floor would contradict Prop. Q.2. Note C(b) gives the corresponding closed-form prediction.

Critical comparison. Excess risk should decay on log-log axes and remain below the flat saturation reference. The accountant using the worst-case distance should remain below the fixed accountant across ϵ and δ scans, as in experimental setting B.3. Saturation independent of n is the refuted interpretation.

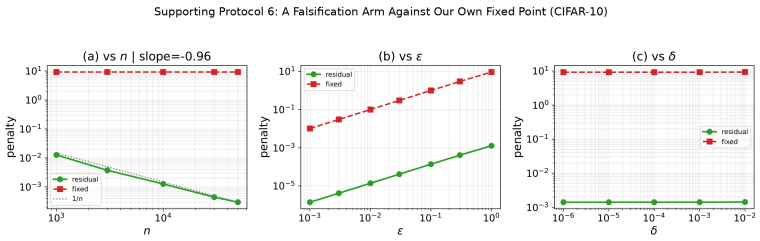


Figure B.13: Experimental setting B.7 on CIFAR-10. Excess risk against n at fixed (ϵ, δ) with noise recomputed from the measured sensitivity. The flat line is the refuted saturation prediction.

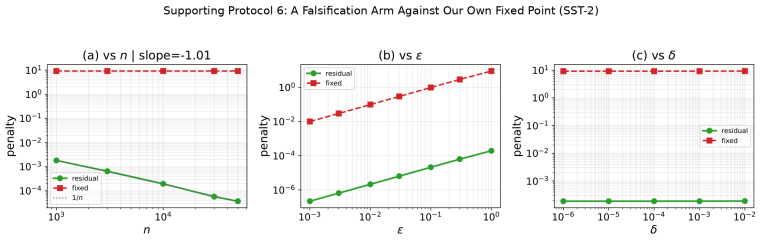


Figure B.14: Experimental setting B.7 on SST-2 under the same fixed budget and noise-setting rule as Figure B.13.

Outcome. Figures B.13 and B.14 show excess risk decreasing on log-log axes. The curves remain below the flat saturation reference when σ is recomputed from the measured sensitivity. The accountant using the worst-case distance is below the fixed accountant across the ϵ and δ scans. These

measurements are consistent with Prop. Q.2 and the private optimization baseline of (Bassily et al., 2019) under experimental setting B.7.

Setup. The substrate and head match experimental setting B.3. Excess risk is plotted against n at fixed $(\epsilon, \delta) = (1, 10^{-5})$ with σ recomputed from the measured sensitivity from measured sensitivity at each n. Companion scans vary ϵ and δ at fixed $n = 10^4$. Each substrate contributes 12 risk points, 16 ϵ points, and 10 δ points. A flat reference line marks the saturation reading denied by Prop. Q.2.

Refutation. Saturation independent of n should first be checked as a frozen σ bug before it is read as evidence against Prop. Q.2.

B.8 What the Panels Test and What They Do Not Test

The six experimental settings provide empirical checks of different signatures. The experimental setting in B.2 compares non-contractive and contractive curves on frozen CIFAR-10 and SST-2 heads. The experimental setting in B.3 compares fixed training-level penalties and bounds based on the worst-case distance to retraining. The experimental setting in B.4 compares noisy and un-noised LoRA controls while recording accuracy. The experimental setting in B.5 measures sequential-deletion plateaus. The experimental setting in B.6 compares direct and one-dimensional feature heads. The experimental setting in B.7 measures risk after recomputation of the noise scale. The panels are consistent with the predictions specified in experimental settings B.2–B.7; they do not verify the worst-case assumptions of the theorems.

B.9 Implementation Notes and Synthetic Precursors

The notes below give the closed-form reference calculations and implementation details. Notes B–E predict slopes, plateau ratios, and accountant gaps. Note A records the frozen-head recipe used in experimental setting B.2. The intended order is to compare the closed-form prediction with the frozen-head and LoRA results.

Note A (used in experimental setting B.2). The reference calculation targets sensitivity growth without contraction. On CIFAR-10 we freeze CNN features of dimension 2048 and train a linear head with learning rate 0.05 at $n = 10^4$ for $T \in \{4, 8, 16, 32, 64, 100\}$. The unregularized arm uses weight decay 0; the weight-decay arm uses 1. Sensitivity $\hat{s}$ is measured over 16 deletions and seeds $\{0, 1\}$. On SST-2 the same schedule runs on DistilBERT [CLS] features of dimension 768 with sequence length 64. The experimental setting in B.2 repeats this comparison.

Note B (used in experimental setting B.2). The reference calculation is the exact quadratic family of Lem. P.3 with curvature parameters β and γ and condition number $\chi = \beta/\gamma$. Full-batch gradient descent at $\eta = 1/\beta$ is paired with projected noisy SGD unlearning from a shared anchor. The non-contractive prediction has no geometric saturation in T; the weight-decay prediction collapses the ratio T_u/T^{re}, where T^{re} is the retraining-update count needed to reach the target empirical error (defined in Lem. P.3), onto $1 - T/T^{\mathrm{re}}$. The experimental setting in B.2 repeats these measurements on both substrates.

Note C (used in experimental settings B.5 and B.7). The reference calculation uses the strongly convex quadratic $\mathcal{L}(\theta, z) = \frac{\gamma}{2}\|\theta - z\|^2$ with $\gamma = 0.1$ and $\beta = 1$. Cor. Q.1 and Prop. Q.2 are then closed form. The calculation predicts geometric plateaus with amplification $(1 - c)^{-1}$ and no excess-risk saturation at fixed (ϵ, δ) when σ is recomputed from the measured sensitivity. The experimental settings in B.5 and B.7 repeat these measurements on frozen heads with per-round DP publication noise chosen from the stated sensitivity.

Note D (used in experimental setting B.3). The reference calculation compares fixed and accountant using the worst-case distance on the Note C quadratic and predicts an $O(n)$ utility gap. The experimental setting in B.3 repeats the comparison on CIFAR-10 with a full-batch convex head of depth $T = 300$ and on SST-2 with the DistilBERT head. Measured slopes of the worst-case distance to retraining are about -0.96 and -1.01. Gaps at $n = 5 \cdot 10^4$ are about $3.1 \cdot 10^4$ and larger on SST-2.

Note E (used in experimental setting B.5). The reference calculation identifies the condition number as the 2-step-to-20-step amplification factor under Cor. Q.1 and predicts collapse onto $(1 - e^{-x})^{-1}$, where $x := T_u\gamma/\beta$ is the contraction-normalized unlearning depth. The experimental setting in B.5 reports stream ratios 4.90 on CIFAR-10 and 5.38 on SST-2.

Shared real data stack. CIFAR-10 uses the standard training split with a frozen CNN feature cache of dimension 2048. SST-2 uses DistilBERT with a frozen encoder and a cached [CLS] feature of dimension 768. LoRA experiments fine tune query and value projections with explicit leave-one-out retraining from a shared initialization. Every panel is regenerated from the measured leave-one-out distances and risk values described in the experimental setting setups above.

C Assumption Elevation Map

Figure C.1 summarizes the revised scope of Table 1. The horizontal axis records update geometry. The vertical position records qualitative proof burden. The low region requires a verified metric contraction and distance to the corresponding retrained model; PL and dissipativity appear there only with their additional endpoint or transition rule assumptions. The middle region is the non-expansive $n^{-1/3}$ trade-off. The high region contains expansive dynamics and learner-specific stability floors (Def. 7). The drawing is ordinal, not a numerical metric.

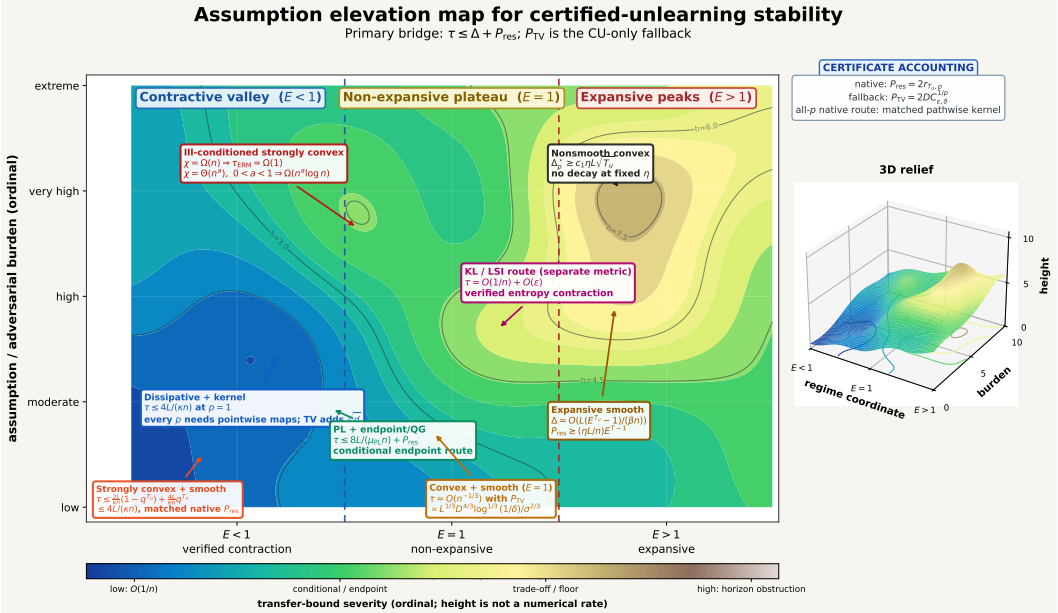


Figure C.1: Revised terrain-style assumption map. Surface height and contours are schematic and encode ordinal proof burden, not numerical rates. The low region is conditional on verified metric contraction and a distance to the corresponding retrained model; the PL and KL annotations mark separate endpoint and entropy routes.

D Formal Definitions

The CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$ described in Section 3.2 uses the domain diameter. Only the supports that carry mass matter for this CU-implied distance. Writing P_X for the law of X, one may replace the diameter by the *effective diameter*

$$D_{\mathrm{eff}} := \sup_{\mathcal{D}_N, i} \operatorname{diam}\left(\operatorname{supp} P_{u^{(i)}} \cup \operatorname{supp} P_{\mathcal{A}(\mathcal{D}_N^{-i})}\right) \le D_\Theta, \tag{7}$$

The construction in Lemma I.2 moves mass only between points of the two supports. Replacing D_Θ by D_{eff} removes unused parts of the parameter domain, but it does not remove the dimension dependence of the CU-implied distance; see Remark 2.

Definition 3 (Sampler.). *Given an index set $\mathcal{I}_n = \{1, \dots, n\}$, a sampler $\mathcal{S}(\mathcal{I}_n, T)$ outputs a sequence of indices $(i_t)_{t=1}^T$ according to one of the following four modes, where $[T] = \{1, \dots, T\}$ and $\mathfrak{S}_n$ is the set of all permutations of $\mathcal{I}_n$:*

1. ***i.i.d. sampler*** $\mathcal{S}_1$. $i_t \overset{iid}{\sim} \mathrm{Unif}(\mathcal{I}_n)$ *for each* $t \in [T]$.
2. ***Fixed sampler*** $\mathcal{S}_2$. $i_t = \pi\big(((t-1) \bmod n) + 1\big)$, *with fixed* $\pi \in \mathfrak{S}_n$.
3. ***Shuffling sampler*** $\mathcal{S}_3$. $i_t = \pi_{\lfloor (t-1)/n \rfloor}\big(((t-1) \bmod n) + 1\big)$, *where* $\pi_e \overset{iid}{\sim} \mathrm{Unif}(\mathfrak{S}_n)$ *for each epoch* $e \geq 0$.
4. ***Sequential sampler*** $\mathcal{S}_4$. $i_t = ((t-1) \bmod n) + 1$.

The modulo operation is understood to return values in $\{1, \dots, n\}$.

Definition 4 ((ϵ, δ)-Certified Unlearning.). *Let $\epsilon > 0$, $\delta \in [0, 1/2]$. We say that $\mathcal{U}$ is a (randomized) (ϵ, δ)-certified unlearning (CU) algorithm for an algorithm $\mathcal{A}$ if for any forget set $\mathcal{D}_f \subset \mathcal{D}$ and any Borel region $O \subseteq \Theta$,*

$$\Pr[\mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{D}_f) \in O] \leq e^{\epsilon} \Pr[\mathcal{A}(\mathcal{D} \backslash \mathcal{D}_f) \in O] + \delta, \tag{8}$$

$$\Pr[\mathcal{A}(\mathcal{D} \backslash \mathcal{D}_f) \in O] \leq e^{\epsilon} \Pr[\mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{D}_f) \in O] + \delta. \tag{9}$$

Definition 5 (ϵ-Pointwise Certified Unlearning.). *Let $\epsilon \geq 0$. We say that $\mathcal{U}$ is an ϵ-pointwise certified unlearning (PCU) algorithm for $\mathcal{A}$ if for any forget set $\mathcal{D}_f \subset \mathcal{D}$ the two laws in Def. 4 are mutually absolutely continuous and*

$$\left\| \ln \frac{dP_{\mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{D}_f)}}{dP_{\mathcal{A}(\mathcal{D} \backslash \mathcal{D}_f)}} \right\|_{\infty} \leq \epsilon. \tag{10}$$

Definition 6 ((τ, n)-Averaged Stable Learning.). *Let $\tau > 0$, $\mathcal{D}_n^{'} = \{z_1^{'}, ..., z_n^{'}\}$ which is i.i.d. to the initial dataset $\mathcal{D}_n$, we say that $\mathcal{A}$ is (randomized) (τ, n)-averaged stable learning (ASL) algorithm (Kuzborskij & Lampert, 2018) if for neighbor datasets $\mathcal{D}_n^{(1)}, ..., \mathcal{D}_n^{(n)}$ where $\mathcal{D}_n^{(i)} = \{z_1, ..., z_i^{'}, ..., z_n\}$, $1 \leq i \leq n$,*

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\mathcal{L}\left(\mathcal{A}\left(\mathcal{D}_n^{(i)} \right), z_i^{'} \right) - \mathcal{L}\left(\mathcal{A}\left(\mathcal{D}_n \right), z_i^{'} \right) \right] \right| \leq \tau, \tag{11}$$

the expectation is taken over the randomness of $\mathcal{A}, \mathcal{D}_n, \mathcal{D}_n^{'}$.

Definition 7 ((τ, n)-Model Stable Learning). *Let $\tau > 0$, using $i, \mathcal{D}_n^{'}, \mathcal{D}_n, \mathcal{D}_n^{(i)}$ in Def. 6, we say that $\mathcal{A}$ is (τ, n)-l_1-model stable learning (MSL) algorithm (Lei & Ying, 2020) if*

$$\frac{1}{n} \sum_{i} \mathbb{E}\left[\left\| \mathcal{A}(\mathcal{D}_n) - \mathcal{A}(\mathcal{D}_n^{(i)}) \right\| \right] \leq \tau, \tag{12}$$

and (τ, n)-l_2-model stable if

$$\frac{1}{n} \sum_{i} \mathbb{E}[\|\mathcal{A}(\mathcal{D}_n) - \mathcal{A}(\mathcal{D}_n^{(i)})\|^2] \leq \tau^2. \tag{13}$$

Let $p \in [1, \infty)$, we say that $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-uniformly stable learning (USL) algorithm (Zhu et al., 2023) if

$$\sup_{\mathcal{D}_n, \mathcal{D}_n^{(i)}, i} \mathcal{W}_p^p\left(\mathcal{A}\left(\mathcal{D}_n \right), \mathcal{A}(\mathcal{D}_n^{(i)}) \right) \leq \tau^p. \tag{14}$$

Definition 8 ((τ, n)-KL Stable Learning.). *Let $\tau > 0$, using $\mathcal{D}_n^{'}, \mathcal{D}_n$ in Def. 6, we say that $\mathcal{A}$ is (τ, n)-KL stable learning (KLSL) algorithm (Dadi & Cevher, 2025) if*

$$\mathbb{E}_{\mathcal{D}_n, \mathcal{D}_n^{'}}\left[KL\left(P_{\mathcal{A}(\mathcal{D}_n)} \| P_{\mathcal{A}(\mathcal{D}_n^{'})} \right) \right] \leq \tau. \tag{15}$$

Definition 9 ((Δ, n)-Stable Unlearning.). *Let $\Delta > 0$, $n \geq 2$, $\mathcal{D}_{n,f}^{(i)} = \{z_i\} \subset \mathcal{D}_n$, $\mathcal{D}_{n,f}^{(j)} = \{z_j\} \subset \mathcal{D}_n, i \neq j$ are i.i.d. 1-entry forget sets, and $u^{(i)} = \mathcal{U}(\mathcal{A}(\mathcal{D}_n), \mathcal{D}_{n,f}^{(i)})$, $u^{(j)} = \mathcal{U}(\mathcal{A}(\mathcal{D}_n), \mathcal{D}_{n,f}^{(j)})$. We*

say that $\mathcal{U}$ is a (randomized) (Δ, n)-stable unlearning (SU) algorithm if for any initial dataset $\mathcal{D}_n$ that $|\mathcal{D}_n| = n$,

$$\frac{1}{n(n-1)} \sum_{i \neq j} \left| \mathbb{E}\left[\mathcal{L}\left(u^{(i)}, z_j\right) - \mathcal{L}\left(u^{(j)}, z_j\right) \right] \right| \leq \Delta, \tag{16}$$

the expectation is taken over the randomness of $\mathcal{U}, \mathcal{A}, \mathcal{D}_n$.

Definition 10 ((Δ, n)-$\mathcal{W}_p$-Uniformly Stable Unlearning.)**.** *Let $N = n + 1$ and use the notation $i, j, u^{(i)}, u^{(j)}, \mathcal{D}_N$ from Def. 9 with the trained dataset enlarged by one record. We say that $\mathcal{U}$ is (Δ, n)-$\mathcal{W}_p$ uniformly stable unlearning (USU) if*

$$\sup_{\mathcal{D}_N,\, i \neq j} \mathcal{W}_p^p(u^{(i)}, u^{(j)}) \leq \Delta^p. \tag{17}$$

Definition 11 ((Δ, n)-Model Stable Unlearning)**.** *Let $N = n + 1$ and use the notation from Def. 9. We say that $\mathcal{U}$ is (Δ, n)-l_p-model stable unlearning (MSU) if*

$$\sup_{\mathcal{D}_N,\, i \neq j} \mathbb{E}_{\mathcal{U},\mathcal{A}} \left[\|u^{(i)} - u^{(j)}\|^p \right] \leq \Delta^p. \tag{18}$$

Definition 12 ($(\Delta, 2n)$-KL Stable Unlearning.)**.** *Let $\mathcal{D}_{2n}$ be i.i.d., $\mathcal{I}_{2n}$ from $\mathcal{S}_3$, $\mathcal{D}^{(1)} := \mathcal{D}_{2n}[\mathcal{I}_{2n}[: n]]$, $\mathcal{D}^{(2)} := \mathcal{D}_{2n}[\mathcal{I}_{2n}[n :]]$, $u^{(1)} = \mathcal{U}(\mathcal{A}(\mathcal{D}_{2n}), \mathcal{D}^{(2)})$, $u^{(2)} = \mathcal{U}(\mathcal{A}(\mathcal{D}_{2n}), \mathcal{D}^{(1)})$. We say that $\mathcal{U}$ is $(\Delta, 2n)$-KL stable unlearning (KLSU) if*

$$\mathbb{E}_{\mathcal{D}_{2n}, \mathcal{S}_3} \left[KL\left(P_{u^{(1)}} \| P_{u^{(2)}}\right) \right] \leq \Delta. \tag{19}$$

Definition 13 (ρ-Total-Variation Stable Unlearning.)**.** *For $\rho \in [0, 1]$, an unlearner is ρ-TV-stable if*

$$\sup_{\mathcal{D}_{n+1}, i \neq j} \mathrm{TV}(P_{u^{(i)}}, P_{u^{(j)}}) \leq \rho,$$

with $u^{(i)}$ and $u^{(j)}$ as in Def. 10. This is the total-variation version of unlearning stability used in the convex noisy-SGD comparison.

Definition 14 ($(\theta^{\mathrm{a}}, T_u)$-Anchored Unlearner, restates Def. 2 with the sampler made explicit.)**.** *Let $\mathcal{A}$ generate iterates $(\theta_t)_{t \leq T}$ on $\mathcal{D}_N$. A random variable θ^{a} is an* anchor *if it is measurable with respect to the training run and does not depend on the forget set. We say that $\mathcal{U}$ is* anchored with depth T_u *if $u^{(i)} = \theta_{T_u}^{(i)}$, where $\theta_0^{(i)} = \theta^{\mathrm{a}}$ and*

$$\theta_{k+1}^{(i)} = \Pi_\Theta\Big(\theta_k^{(i)} - \eta_k g_k^{(i)} + \sqrt{2\eta_k}\, \sigma\, \xi_k\Big), \quad \xi_k \overset{iid}{\sim} \mathcal{N}(0, I_d), \tag{20}$$

and $g_k^{(i)}$ is either $\nabla R_{\mathcal{D}_N^{-i}}(\theta_k^{(i)})$ or a stochastic gradient drawn by a sampler $\mathcal{S}$ (Def. 3) on $\mathcal{I}_N \backslash \{i\}$.

The one-step expansion inequality used in Prop. N.1 is

$$\left\| u - v - \eta\big(\nabla \mathcal{L}(u, z) - \nabla \mathcal{L}(v, z)\big) \right\| \leq \mathcal{E}\|u - v\| + \varsigma. \tag{21}$$

The coefficient $\mathcal{E}$ is written E in the main text, Lemma N.1 lists the pairs $(\mathcal{E}, \varsigma)$ realized by each loss class, and Section 5 exhibits instances in which the expansive value is attained.

E Related Work and Classical Stability Baselines

This appendix expands the summary of Section 1. It has four parts. First, we review the five literatures connected to this paper. Second, we review the known implication from differential privacy to learner stability. Third, we state the missing implication from a CU certificate to learner stability and explain how the Conversion Theorem supplies it. Fourth, we compare the classical stability baselines recovered by Table 1 row by row.

E.1 Five Literatures

(a) Algorithmic stability (Defs. 6, 7) and generalization. The subject begins with (Bousquet & Elisseeff, 2002), who defined uniform stability and proved that a τ-uniformly stable algorithm (Defs. 6, 7) generalizes at rate $\tau + O(n^{-1/2})$, instantiating it for regularized ERM. Stability became

algorithmic with (Hardt et al., 2016), who observed that the whole question reduces to how a single gradient step transforms the distance between two runs. Their expansiveness calculus, contractive at $1 - \eta\gamma$ under strong convexity, nonexpansive under convex smoothness, $(1 + \eta\beta)$-expansive in general, is the direct ancestor of our Lemma N.1, and the trichotomy of Table 1 is theirs, transplanted from the training recursion to the unlearning recursion (Def. 10).

The following decade refined every ingredient. The *worst-case Lipschitz constant* was replaced by data-dependent surrogates: (Kuzborskij & Lampert, 2018) substitute the expected gradient norm along the trajectory, paying the risk at initialization, and (Lei & Ying, 2020) introduce on-average model stability (Def. 7) together with the self-bounding inequality $\|\nabla\mathcal{L}\|^2 \leq 2\beta\mathcal{L}$ (our Lem. J.1), which removes the Lipschitz hypothesis for nonnegative smooth losses. (Nikolakakis et al., 2022) push this idea further for full-batch gradient descent by telescoping the descent lemma along the path; we use their resulting inequality as equation 37. The *concentration* was sharpened from the $\sqrt{n}$-lossy exponential bounds of (Bousquet & Elisseeff, 2002) to nearly optimal high-probability rates (Feldman & Vondrak, 2019; Bousquet et al., 2020) and finally to deviation-optimal $O(1/n)$ (Klochkov & Zhivotovskiy, 2021). The *rate* was improved under curvature: (Charles & Papailiopoulos, 2018) for algorithms converging to global optima, (Lei & Ying, 2021) for gradient-dominated objectives, and (Zhu et al., 2025) to $\tilde{O}(1/n^2)$ under low noise. Certified unlearning also has minimax generalization and deletion-capacity lower bounds (Liu et al., 2023); our lower bounds instead isolate the recursion and conditions using the corresponding retrained model. The *geometry* was generalized from Euclidean argument stability to Wasserstein stability (Def. 7), uniform in time, by (Zhu et al., 2023) using Markov-chain perturbation theory (Rudolf & Schweizer, 2018), and to relative entropy by (Dadi & Cevher, 2025). Finally, the *limits* were mapped: (Bassily et al., 2020) prove that subgradient descent on nonsmooth convex losses is unstable at $\Omega(\eta L\sqrt{T})$ in dimension $d \geq \min\{T, \eta^{-2}\}$, with a matching upper bound, and (Lei, 2023) extend the positive theory to weakly convex nonsmooth objectives measured through the Moreau envelope, where subgradients converge only graphically (Davis & Drusvyatskiy, 2022). The nonsmooth lower bound from (Bassily et al., 2020) is the input to Theorem 6(i).

We introduce no new stability notion. Each learner notion (Defs. 6, 7, 8) has an *unlearning analogue* (Defs. 9, 10, 12) whose comparison starts at zero discrepancy and runs for T_u rather than T steps. The CU certificate converts an unlearning-stability bound into a learner-stability bound, with the direct worst-case-distance term or the total-variation bound obtained from the CU inequalities derived in Section 3.2.

(b) Machine unlearning and certified removal. The formal notion of a deletion certificate is due to (Guo et al., 2019), who defined (ϵ, δ)-certified removal against the retrained law and realized it for convex objectives by a Newton correction followed by objective perturbation. (Sekhari et al., 2021) separated two questions that had previously been conflated: how many deletions a scheme supports and what population risk the scheme retains. They gave the first deletion-capacity bounds. Algorithmically, later work separates loss classes. Under strong convexity, (Neel et al., 2021) study descent-to-delete: warm-started projected gradient descent runs on the updated objective and then perturbs the published output. Their Lemma 8 gives sensitivity $2L/(\gamma n)$, exactly the $T_u = 0$ value of our Lem. L.2(i), for both their secret-state and perfect variants. (Ullah et al., 2021) take a different route: their noisy accelerated SGD learner has total-variation stability (Def. 13), and rejection sampling plus partial retraining produces exact unlearning. For nonconvex losses, (Chien et al., 2024a) analyse projected noisy SGD with a Rényi accountant, (Chien et al., 2024b) obtain iteration-independent certificates for Langevin dynamics through a log-Sobolev privacy-recuperation argument, and (Mu & Klabjan, 2026; 2025) introduce rewind-to-delete. The rewind sensitivity $\frac{2L}{\beta N}[(1 + \eta\beta)^T - (1 + \eta\beta)^{T_u}]$ is the quantity that Theorem 6(ii) shows cannot be made small in the expansive construction. Recent work continues to study two phenomena explained by our analysis: accuracy collapse under repeated deletion (Dolgova & Stich, 2026) and heavy-tailed per-example sensitivity that makes worst-case certificates wasteful (Benarroch et al., 2026). On the negative side, (Huang & Canonne, 2023) and (Van Waerebeke et al., 2025) establish complexity trade-offs and phase boundaries for first-order unlearning. Theorem 6(iii) recovers a corresponding phase boundary from a stability argument rather than a computational argument, so the conclusion does not depend on a particular unlearning implementation.

These papers use learner stability to obtain a small CU privacy budget; they do not use a CU certificate to bound the stability of the learner. The Conversion Theorem addresses this missing direction.

(c) Differential privacy and privacy amplification. The certificate of Def. 1 has the same two-sided probability-law form as differential privacy, with the retrained output law serving as the reference law. We therefore use the Gaussian mechanism and the total-variation characterization of (ϵ, δ)-indistinguishability from (Dwork & Roth, 2014). The constant $C_{\epsilon,\delta}$ in Section 3.2 is the extremal total-variation distance allowed by that characterization, and it controls only the CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$; the direct term $2r_{T_u,p}$ instead uses the distance to the corresponding retrained model. The rates we compare against come from private optimization: (Bassily et al., 2014) for DP-ERM and (Bassily et al., 2019) for DP stochastic convex optimization. The $\sqrt{d}/(\epsilon n)$ term in the latter is the same dimension factor that appears in the CU-implied distance (Rem. 2); additional spectral structure can remove that factor (Ma et al., 2022; Li et al., 2022; Song et al., 2021). Privacy amplification by iteration (Feldman et al., 2018) and its contractive refinement (Altschuler & Talwar, 2022) explain how a certificate changes with the number of updates. The two cases in equation 31 give geometric improvement for contractive maps and $T_u^{-1/2}$ improvement for merely nonexpansive maps, which is exactly the boundary between the first and second blocks of Table 1. (Chien & Li, 2025) show that the convergent regime persists without convexity or smoothness. The utility side of this trade-off and its limits is studied in (Sander et al., 2023; Ertan & van Dijk, 2026; Zhang et al., 2026). The sample-complexity separations between private and non-private learning are due to (Chaudhuri & Hsu, 2011; Bun et al., 2015), with model-agnostic constructions in (Bassily et al., 2018; Liu et al., 2021).

(d) Information-theoretic generalization. Our KL route terminates in the mutual-information bound of (Xu & Raginsky, 2017): an algorithm whose output carries I nats about its input generalizes at $\rho\sqrt{2I/n}$ for ρ-sub-Gaussian losses. This is what turns the KL stability of Theorem N.5 (Def. 8) into a risk bound (Lem. J.6), and it is the only one of our three routes that never mentions the diameter of Θ. The quantitative form of "privacy implies generalization" in this language is max-information (Rogers et al., 2016), recently sharpened for DP-SGD by (Lampert & Zakerinia, 2026).

(e) Optimization theory. Two facts from optimization theory are used. The Polyak-Łojasiewicz condition implies quadratic growth (Karimi et al., 2016); this converts the anchor's objective error into a distance to the minimizer set in Theorem N.6, so the PL rows of Table 1 need no convexity. Reflection construction for diffusions (Eberle, 2016) gives a positive rate $\kappa(m, K, \beta, \sigma)$ for dissipative nonconvex losses with nondegenerate noise. Row C additionally assumes that the discretized transition rule satisfies the metric contraction in Def. 26; without that verification, the row reduces to C†.

E.2 The One Existing Implication in Our Direction

Differential privacy implies stability (Def. 7) and generalization; this was formalized for learnability by (Wang et al., 2016) and is used in private optimization (Bassily et al., 2014; 2019; Wang et al., 2024). In our framework, take $\mathcal{U}$ to be the identity map, so the published law remains $P_{\mathcal{A}(\mathcal{D})}$ after deletion. Then $\Delta = 0$, Def. 1 becomes the differential-privacy comparison between $P_{\mathcal{A}(\mathcal{D})}$ and $P_{\mathcal{A}(\mathcal{D}\setminus\mathcal{D}_f)}$, and Conversion Theorem 1(2) gives $\tau \le 2r_{T_u,p}$ (Theorem Q.1). The identity unlearner uses the training-level distance between those two laws, so Theorem 2 gives a constant contribution obtained from the CU inequalities. An anchored unlearner instead uses the distance between each unlearned law and its corresponding retrained law, which can be $O(1/n)$ under Theorem N.4. These are two distinct sufficient conditions because they use the CU guarantee for different pairs of probability laws.

E.3 The Gap and How the Conversion Closes It

Literature (a) bounds learner stability (Def. 7) by comparing two training runs for T updates. Literature (b) bounds the privacy budget needed to make an unlearned output law indistinguishable from the output law of the learner retrained without the forgotten record. Conversion Theorem 1 combines two CU comparisons, one for deletion index i and one for deletion index j, with the unlearning-stability bound for $P_{u^{(i)}}$ and $P_{u^{(j)}}$. The theorem then bounds the distance between $P_{\mathcal{A}(\mathcal{D}_N^{-i})}$ and $P_{\mathcal{A}(\mathcal{D}_N^{-j})}$. Table 1 lists the resulting rates; Theorem 2 gives the CU-implied distance, Theorem N.4

gives the rate obtained from the worst-case distance to retraining, and the contrapositive bound transfers learner-stability lower bounds to unlearning lower bounds.

E.4 How to Read Tables E.1–E.3

The three tables below collect the classical learning-stability results (Def. 7) that our combinations reproduce. They are background rather than contributions. The point of listing them is to make the comparison checkable. Each caption states which combination of Table 1 matches it and under what additional cost.

Table E.1 is the strongly convex baseline. The uniform stability bound $2L^2/(\gamma n)$ of (Hardt et al., 2016) and the $O(1/n)$ Wasserstein bound of (Zhu et al., 2023) are both attained by Combination A, whose argument-stability parameter (Def. 7) is $2L/(\gamma n)$ up to the direct term based on the worst-case distance to retraining. Two differences are worth noting. Our bound needs no step-size restriction beyond $\eta \le 2/(\beta+\gamma)$ and no ergodicity hypothesis, because the comparison between anchored runs starts at zero discrepancy rather than at stationarity. Under a shared transition rule, Theorem N.4 bounds the additive $2r_{T_u,p}$ directly by the worst-case distance and makes it $O(1/n)$ for every p.

Table E.2 is the convex baseline, and it is where our route is weakest. The classical stability bounds (Def. 7) of (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Lei & Ying, 2020) scale with the *training* horizon T, and Combination B replaces that horizon by the *unlearning* horizon T_u, which is a genuine improvement whenever $T_u \ll T$. The CU certificate, however, improves only as $T_u^{-1/2}$ in this regime (Lemma L.3). Consequently, the sum of the unlearning-stability term and the CU-implied distance has an interior optimum, and the CU-implied distance is never absorbed completely. This is why the corresponding row of Table 1 reads "Interior optimum" rather than "Always." The nonsmooth entry of (Bassily et al., 2020) in this table is not matched by our route; Theorem 6(i) explains the obstruction.

Table E.3 is the nonconvex baseline. Combination C matches the $\mathcal{W}_1$ bound of (Zhu et al., 2023) under dissipativity, and Combination D matches the KL bound of (Dadi & Cevher, 2025) for SGLD. The classical proofs for those two baselines need a Lyapunov drift or a stationarity assumption on the learner; our corresponding proofs need only contraction of the *unlearning* transition rule. The comparison also isolates what the certificate route cannot do: for bare β-smooth nonconvexity, with no dissipativity and no log-Sobolev inequality, neither the classical tables nor our table has a valid rate. This limitation is formalized in Theorem 6(ii).

Research	Stability	$\mathcal{A}$	Bound	Assumption	Requirement
(Hardt et al., 2016)	Uniform	PSGD	$\dfrac{2L^2}{\gamma n}$	Def. 16, Def. 21	$\eta_t \le \dfrac{1}{\beta}$, $\mathcal{S}_1/\mathcal{S}_2$
(Zhu et al., 2023)	$\mathcal{W}_1$-USL	SGD	$O\left(\dfrac{1}{n}\right)$	Def. 17, $\|\nabla\mathcal{L}(0,z)\| \le E,\ \|z\| \le D$	$\eta \le \min\{\frac{1}{\gamma}, \frac{\gamma}{K_1^2+64D^2K_2^2}\}$, $\mathcal{S}_1$, ergodicity

Table E.1: Classical stability baselines in the γ-strongly convex setting (attained by Combination A).

F Deferred Remarks on Certificates

Remark 1 (Why the KL route needs a pointwise certificate.)**.** *Under Def. 4 with $\delta > 0$, the two laws may be mutually singular on a set of mass δ. Then $KL(P_{\mathcal{A}(\mathcal{D}^{(1)})} \| P_{\mathcal{A}(\mathcal{D}^{(2)})}) = +\infty$ no matter how small δ is. Relative entropy is not continuous in total variation, so no additive or multiplicative correction of equation 26 gives a KL bound. The δ slack cannot be treated as a failure event.*

We state the alternatives explicitly. (i) Strengthen the certificate to Def. 5. Then equation 26 is exact and unconditional, and Thm. O.2 gives an $O(1/n)$ excess-risk bound. This is the route used in the paper and is supplied by pure or ∞-order certificates. (ii) Keep Def. 4 and replace the KL

Research	Stability	$\mathcal{A}$	Bound	Assumptions	Requirements
(Hardt et al., 2016)	Uniform	SGD	$\frac{2L^2}{n}\sum_{t=1}^{T}\eta_t$	Def. 16, Def. 21	$\eta_t \le \frac{2}{\beta}$, $\mathcal{S}_1/\mathcal{S}_2$
(Kuzborskij & Lampert, 2018)	ASL	SGD	$O\left(\sqrt{c(R(\theta_0)-R(\theta^*))}\,\frac{T^{1/4}}{n}+\frac{\sqrt{T}}{n}\right)$	Def. 16, Def. 21, Def. 24	$\eta_t=\frac{c}{\sqrt{t}}\le\frac{1}{\beta}$, $\mathcal{S}_3$
(Bassily et al., 2020)	Uniform	PSGD	$O\left(\min\{1,\frac{T}{n}\}\sqrt{T}+\frac{T}{n}\right)$	Def. 16	$\eta_{t+1}\le\eta_t$, $\mathcal{S}_1/\mathcal{S}_2$
(Lei & Ying, 2020)	l_1-MSL	PSGD	$O\left(\frac{\sum_t \eta_t \mathbb{E}[\sqrt{R_{\mathcal{D}_n}(\theta_t)}]}{n}\right)$	Def. 21, Def. 16*	$\eta_t \le \frac{2}{\beta}$, $\mathcal{S}_1$
	l_2-MSL	PSGD	$O\left(\frac{\sum_t \eta_t^2 \mathbb{E}[\sqrt{R_{\mathcal{D}_n}(\theta_t)}]}{n}\right)$	Def. 21	$\eta_t \le \frac{2}{L}$, $\mathcal{S}_1$

Table E.2: Classical stability baselines in the (non-strongly) convex setting (attained by Combinations B, F and G). The assumption with * is needed only when converting stability into a generalization bound.

Research	Stability	$\mathcal{A}$	Bound	Assumptions	Requirements
(Hardt et al., 2016)	Uniform	SGD	$O\left(\frac{T^{\frac{\beta c}{1+\beta c}}}{n-1}\right)$	Def. 16, Def. 21, Def. 22	$\eta_t \le \frac{c}{t}$, $\mathcal{S}_1$
(Kuzborskij & Lampert, 2018)	ASL	SGD	$O\left(\frac{(R(\theta_0)\,T)^{\frac{c\zeta}{1+c\zeta}}}{n}\right)$	Def. 16, Def. 21, Def. 22, Def. 24	$\eta_t=\frac{c}{t}$, $\mathcal{S}_3$
(Zhu et al., 2023)	$\mathcal{W}_1$-USL	NSGD	$O\left(\frac{1}{n}\right)$	Def. 17, Def. 19, Def. 24, $\|\nabla\mathcal{L}(0,z)\|\le E,\ \|z\|\le D$	$\eta \le \min\{\frac{1}{m}, \frac{m}{K_1^2+64D^2K_2^2}\}$, $\mathcal{S}_1$, ergodicity
	$\mathcal{W}_2$-USL	SGD	$O\left(\sqrt{\frac{1}{n}+\frac{2K}{m}}\right)$	Def. 17, Def. 19 $\|\nabla\mathcal{L}(0,z)\|\le E,\ \|z\|\le D$	$\eta \le \min\{\frac{1}{m}, \frac{m}{K_1^2+64D^2K_2^2}\}$, $\mathcal{S}_1$
(Dadi & Cevher, 2025)	KLSL	SGLD	$O\left(\frac{1}{n}\right)$	Def. 21, Def. 19, Def. 18	$\eta_t \in \left(\frac{31}{32m}, \frac{m}{2\beta^2}\right]$

Table E.3: Classical stability baselines in the nonconvex setting (attained by Combinations C and D). The $\mathcal{W}_2$ entry of (Zhu et al., 2023) controls a squared distance of order $O(1/n + K/m)$, hence the displayed $O(\sqrt{\cdot})$ rate for the $\mathcal{W}_2$-USL parameter.

conclusion by equation 24. For bounded losses this route is unconditional. Pinsker's inequality gives $\Delta_{\mathrm{SU}} \le \rho\sqrt{2\Delta}$, so the statistical term changes from $O(1/n)$ to $O(n^{-1/2})$. (iii) Apply equation 26 only to a δ-trimmed pair of laws. This yields no bound on Def. 8 itself and is not used here.

Remark 2 (Dimension dependence of the endpoint-distance term.)**.** *The CU-implied distance $P^{\mathrm{TV}}(\epsilon,\delta)$ in part (2) is linear in the domain diameter. The direct term $2r_{T_u,p}$, which is based on the worst-case distance to retraining, is determined by the distance to the corresponding retrained model and has no diameter or $\sqrt{d}$ factor when the two branches use a common 1-Lipschitz transition rule. The quantity D_Θ is not dimension free. A Euclidean ball of constant per-coordinate radius has $D_\Theta = \Theta(\sqrt{d})$. Three points should be kept in view. First, the diameter can be localized for the CU-implied distance. The proof of Lem. I.2 moves mass only between points of $\operatorname{supp}\mu \cup \operatorname{supp}\nu$, so D_Θ may everywhere be replaced by D_{eff} of equation 7, and more generally by $2r$ up to an additive $2D_\Theta\delta_0^{1/p}$ if a ball of radius r carries at least $1-\delta_0$ of the mass of both laws. Second, localization does not remove the dimension from the CU-implied distance. For a Gaussian-mechanism certificate the published law is the unlearner output convolved with $\mathcal{N}(0,\sigma^2 I_d)$, whose typical norm is $\sigma\sqrt{d}$, so $D_{\mathrm{eff}} = O(\eta L T_u + \sigma\sqrt{d})$. Substituting $\sigma = s\sqrt{2\ln(1.25/\delta)}/\epsilon$ from Lem. L.1 into $2D_{\mathrm{eff}}C_{\epsilon,\delta}$ at $p=1$, the noise scale cancels and the CU-implied distance becomes $O\big(s\sqrt{d\ln(1/\delta)}\big)$. With the contractive sensitivity $s = 2L/(\kappa n)$ of Lem. L.2(i), this is $O\big(\frac{L}{\kappa n}\sqrt{d\ln(1/\delta)}\big)$. Thus the $\sqrt{d}$ factor belongs to the CU-implied distance; the direct shared-transition-rule bound remains $O(1/n)$ for every p. Third, the KL route of part (3) has no diameter factor at all and is the only conversion that applies on an unbounded Θ. The dimension nevertheless re-enters its* input *through the covering constant $c_\Theta = O(d\ln(\beta D_\Theta n))$ of Lem. L.4. The dimension factor is absent only when the pointwise rather than the uniform gradient bound is used.*

G Function Assumptions

Definition 15 (Convexity and Non-convexity.)**.** *Let $f : \Theta \times \mathcal{Z} \to \mathbf{R}$ be differentiable in its first argument. The function f is convex if for all $u, v \in \Theta$ and every $z \in \mathcal{Z}$,*

$$\langle \nabla f(u, z) - \nabla f(v, z), u - v \rangle \geq 0.$$

We call f γ-strongly convex, for $\gamma > 0$, if for all $u, v \in \Theta$ and every $z \in \mathcal{Z}$,

$$f(u, z) - f(v, z) \geq \langle \nabla f(v, z), u - v \rangle + \frac{\gamma}{2}\|u - v\|^2.$$

A differentiable function f is μ-weakly convex if for all $u, v \in \Theta$ and every $z \in \mathcal{Z}$,

$$f(u, z) - f(v, z) \geq \langle \nabla f(v, z), u - v \rangle - \frac{\mu}{2}\|u - v\|^2.$$

Definition 16. *A function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ is L-Lipschitz if, for all $u, v \in \Theta$ and $z \in \mathcal{Z}$,*

$$\|f(u, z) - f(v, z)\| \leq L\|u - v\|.$$

Definition 17. *Equip $\mathcal{Z}$ with a metric $d_{\mathcal{Z}}$. Given constants $K_1, K_2 > 0$, a differentiable function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ is pseudo-Lipschitz if for any $u, v \in \Theta$ and every $z_1, z_2 \in \mathcal{Z}$, we have,*

$$\begin{aligned}\|\nabla f(u, z_1) - \nabla f(v, z_2)\| \leq K_1\|u - v\| \\ + K_2 d_{\mathcal{Z}}(z_1, z_2)(\|u\| + \|v\| + 1).\end{aligned}$$

Definition 18. *Let $K_3 > 0$ and $M > 0$. A differentiable function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ is sample-pseudo-Lipschitz if, for any $z_1, z_2 \in \mathcal{Z}$ and any $u \in \Theta$,*

$$\|\nabla f(u, z_1) - \nabla f(u, z_2)\| \leq K_3\|u\| + M.$$

Definition 19. *Given constants $m > 0$ and $K > 0$, a differentiable function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ satisfies the (m, K)-dissipativity condition if for any $u, v \in \Theta$ and every $z \in \mathcal{Z}$, we have,*

$$\langle \nabla f(u, z) - \nabla f(v, z), u - v \rangle \geq m\|u - v\|^2 - K.$$

Definition 20. *Let $L > 0, \alpha \in [0, 1]$, and let $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ be differentiable. We call f (α, L)-Hölder gradient-continuous if for all $u, v \in \Theta$ and $z \in \mathcal{Z}$,*

$$\|\nabla f(u, z) - \nabla f(v, z)\| \leq L\|u - v\|^{\alpha}.$$

Definition 21. *A function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ is β-smooth if it is differentiable in its first argument and, for all $u, v \in \Theta$ and $z \in \mathcal{Z}$,*

$$\|\nabla f(u, z) - \nabla f(v, z)\| \leq \beta\|u - v\|.$$

Definition 22. *A function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ is ρ-bounded if for all $u \in \Theta$ and $z \in \mathcal{Z}$, $|f(u, z)| \leq \rho$.*

Definition 23. *A function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ is H-Lipschitz Hessian if for all points $u, v \in \Theta$ and $z \in \mathcal{Z}$, f is twice-differentiable and,*

$$\|\nabla^2 f(u, z) - \nabla^2 f(v, z)\| \leq H\|u - v\|.$$

Definition 24. *A function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ is (σ, n)-variance bounded if for every $\theta \in \Theta$,*

$$\mathbb{E}_{\mathcal{D}_n, z}\left[\|\nabla f(\theta, z) - \nabla R_{\mathcal{D}_n}(\theta)\|^2\right] \leq \sigma^2.$$

H Gradient Stability

Definition 25 ((τ, n)-Gradient Stable Learning.)**.** *Let $\tau > 0$. Using the notation $i, \mathcal{D}_n^{'}, \mathcal{D}_n, \mathcal{D}_n^{(i)}$ from Def. 6, we say that $\mathcal{A}$ is a (τ, n)-gradient stable learning (GSL) algorithm (Lei, 2023) if, for all $\mathcal{D}_n, \mathcal{D}_n^{(i)}$,*

$$\sup_{z \in \mathcal{Z}} \mathbb{E}_{\mathcal{A}}\left[\left\|\nabla \mathcal{L}(\mathcal{A}(\mathcal{D}_n), z) - \nabla \mathcal{L}(\mathcal{A}(\mathcal{D}_n^{(i)}), z)\right\|^2\right] \leq \tau^2. \tag{22}$$

I PROOFS OF CONVERSION THEOREMS

Lemma I.1 (Sampler Invariance (Hardt et al., 2016)). *If an algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \Theta$ takes a sampler as $\mathcal{S}_1$, $\mathcal{S}_2$, or $\mathcal{S}_3$, where there exists a sub-algorithm $\hat{\mathcal{A}}$ such that $\mathcal{A}(\mathcal{D}_n) = \hat{\mathcal{A}}(\mathcal{D}_n[\mathcal{S}_k(\mathcal{D}_n)]), 1 \leq k \leq 3$, we have that for any $1 \leq i \leq n, j \neq i$:*

$$\mathbb{E}_{\mathcal{A}}[\mathcal{A}(\mathcal{D}_n) - \mathcal{A}(\mathcal{D}_n^{(i)})] = \mathbb{E}_{\mathcal{A}}[\mathcal{A}(\mathcal{D}_n) - \mathcal{A}(\mathcal{D}_n^{(i)\mapsto(j)})],$$

where $\mathcal{D}_n^{(i)\mapsto(j)} = \{z_1, ..., z_{i-1}, z_{i+1}, ..., z_j, z_i^{'}, ..., z_n\}$ if $j > i$, and $\mathcal{D}_n^{(i)\mapsto(j)} = \{z_1, ..., z_{j-1}, z_i, z_j, ..., z_{i-1}, z_{i+1}, ..., z_n\}$ if $j < i$. Besides, for any function $f : \Theta \times \mathcal{Z} \mapsto \mathbf{R}$ and any $z \in \mathcal{Z}$, we have,

$$\mathbb{E}_{\mathcal{A}}[f(\mathcal{A}(\mathcal{D}_n), z) - f(\mathcal{A}(\mathcal{D}_n^{(i)}), z)] = \mathbb{E}_{\mathcal{A}}[f(\mathcal{A}(\mathcal{D}_n), z) - f(\mathcal{A}(\mathcal{D}_n^{(i)\mapsto(j)}), z)].$$

Proof. For $k = 1$, let $\mathbf{S} = \{1, \ldots, n\}^n$ be the set of all n-tuples of sampler indices. For $i < j$, let $\mathbf{S}'$ be the same set with the i- and j-coordinates relabeled; the two finite sums are identical after this relabeling.

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}[\mathcal{A}(\mathcal{D}_n) - \mathcal{A}(\mathcal{D}_n^{(i)})] &= \mathbb{E}_{\mathcal{S}_1,\hat{\mathcal{A}}}[\hat{\mathcal{A}}(\mathcal{D}_n[\mathcal{S}_1(\mathcal{D}_n)]) - \hat{\mathcal{A}}(\mathcal{D}_n^{(i)}[\mathcal{S}_1(\mathcal{D}_n^{(i)})])] \\ &= \mathbb{E}_{\hat{\mathcal{A}}}\left[\frac{1}{n^n}\sum_{q=1}^{n^n}\left(\hat{\mathcal{A}}(\mathcal{D}_n\left[\mathbf{S}\left[q\right]\right]) - \hat{\mathcal{A}}(\mathcal{D}_n^{(i)}\left[\mathbf{S}\left[q\right]\right])\right)\right].\end{aligned}$$

After this coordinate relabeling (without loss of generality, assume $i < j$), we have:

$$\begin{aligned}&\frac{1}{n^n}\sum_{q=1}^{n^n}\left(\hat{\mathcal{A}}(\mathcal{D}_n\left[\mathbf{S}\left[q\right]\right]) - \hat{\mathcal{A}}(\mathcal{D}_n^{(i)}\left[\mathbf{S}\left[q\right]\right])\right) \\ &= \frac{1}{n^n}\sum_{q=1}^{n^n}\left(\hat{\mathcal{A}}(\mathcal{D}_n\left[\mathbf{S}\left[q\right]\right]) - \hat{\mathcal{A}}\left(\mathcal{D}_n^{(i)\mapsto(j)}\left[\mathbf{S}^{'}\left[q\right]\right]\right)\right) \\ &= \frac{1}{n^n}\sum_{q=1}^{n^n}\left(\hat{\mathcal{A}}(\mathcal{D}_n\left[\mathbf{S}\left[q\right]\right]) - \hat{\mathcal{A}}\left(\mathcal{D}_n^{(i)\mapsto(j)}\left[\mathbf{S}\left[q\right]\right]\right)\right) \\ &= \mathbb{E}_{\mathcal{S}_1}[\hat{\mathcal{A}}(\mathcal{D}_n) - \hat{\mathcal{A}}(\mathcal{D}_n^{(i)\mapsto(j)})].\end{aligned}$$

For $k = 2, 3$, the same relabeling argument applies after replacing $\mathbf{S}$ by the set of all permutations of $\{1, \ldots, n\}$; the corresponding $\mathbf{S}'$ relabels the i- and j-coordinates.

□

Lemma I.2 (Metric Interpolation Lemma). *The following three bounds govern how certified-unlearning budgets transfer across metrics.*

(i) TV certificate constant ((Dwork & Roth, 2014)). *Let μ, ν be probability measures on Θ. Suppose μ, ν are (ϵ, δ)-indistinguishable in both directions, i.e., for every measurable $O \subseteq \Theta$,*

$$\mu(O) \leq e^{\epsilon}\nu(O) + \delta, \qquad \nu(O) \leq e^{\epsilon}\mu(O) + \delta.$$

Then

$$\mathrm{TV}(\mu, \nu) \leq C_{\epsilon,\delta} := \min\left\{1,\ \frac{e^{\epsilon} - 1 + 2\delta}{e^{\epsilon} + 1}\right\} = \min\left\{1,\ \delta + (1 - \delta)\tanh\frac{\epsilon}{2}\right\}.$$

Moreover, this bound is sharp. For every $\epsilon \geq 0$ and $\delta \in [0, 1]$, there exist such μ, ν attaining equality in the displayed expression whenever $(e^{\epsilon} - 1 + 2\delta)/(e^{\epsilon} + 1) \leq 1$.

(ii) KL under multiplicative bounds ((Dwork & Roth, 2014; Dadi & Cevher, 2025)). *Let μ, μ_U, ν, ν_U be probability measures on Θ with $\mu \ll \mu_U \ll \nu_U$ and $\nu \ll \nu_U$. Suppose*

$$e^{-\epsilon} \leq \frac{d\mu}{d\mu_U} \leq e^{\epsilon}, \qquad e^{-\epsilon} \leq \frac{d\nu}{d\nu_U} \leq e^{\epsilon}$$

hold μ_U-a.e. and ν_U-a.e., respectively. Then

$$KL(\mu\|\nu) \le e^{\epsilon}\, KL(\mu_U\|\nu_U) + \epsilon\Big(1 + \tanh\frac{\epsilon}{2}\Big).$$

(iii) TV controls Wasserstein ((Zhu et al., 2023)). *For any probability measures μ, ν on Θ with $\mathrm{diam}(\Theta) = D_\Theta < \infty$, and any $p \ge 1$,*

$$\mathcal{W}_p^p(\mu,\nu) \le D_\Theta^p\ \mathrm{TV}(\mu,\nu).$$

Proof. **(i)** The algebraic identity

$$\frac{e^\epsilon - 1 + 2\delta}{e^\epsilon + 1} = \delta + (1-\delta)\tanh\frac{\epsilon}{2}$$

is elementary. If $(e^\epsilon - 1 + 2\delta)/(e^\epsilon + 1) > 1$, the claim is trivial, so assume otherwise.

Upper bound. Let $O^\star$ be a Hahn set for $\mu - \nu$, so $\mathrm{TV}(\mu,\nu) = \mu(O^\star) - \nu(O^\star)$. Two-sided (ϵ,δ)-indistinguishability yields

$$\mu(O^\star) - \nu(O^\star) \le (e^\epsilon - 1)\nu(O^\star) + \delta,$$
$$\nu((O^\star)^c) - \mu((O^\star)^c) \le (e^\epsilon - 1)\mu((O^\star)^c) + \delta.$$

Adding these and using $\nu((O^\star)^c) - \mu((O^\star)^c) = \mu(O^\star) - \nu(O^\star)$ together with

$$\nu(O^\star) + \mu((O^\star)^c) = 1 - \big(\mu(O^\star) - \nu(O^\star)\big) = 1 - \mathrm{TV}(\mu,\nu)$$

gives $2\ \mathrm{TV}(\mu,\nu) \le (e^\epsilon - 1)\big(1 - \mathrm{TV}(\mu,\nu)\big) + 2\delta$, which rearranges to $\mathrm{TV}(\mu,\nu) \le C_{\epsilon,\delta}$.

Sharpness. On $\{0,1\}$, set

$$\mu(\{1\}) = \delta + (1-\delta)\frac{e^\epsilon}{e^\epsilon+1}, \qquad \nu(\{1\}) = (1-\delta)\frac{1}{e^\epsilon+1}$$

(clipped to $[0,1]$ if needed), so that ν is the reflection of μ and $\mu(\{0\}) = \nu(\{1\})$. Both defining inequalities hold with equality on $O = \{1\}$ and on $O = \{0\}$ respectively, and hold strictly on the complements, so (μ,ν) is (ϵ,δ)-indistinguishable in both directions, moreover

$$\mathrm{TV}(\mu,\nu) = \mu(\{1\}) - \nu(\{1\}) = \delta + (1-\delta)\frac{e^\epsilon - 1}{e^\epsilon + 1} = C_{\epsilon,\delta}.$$

When $\delta = 0$ this reduces to $\mathrm{TV} = \tanh(\epsilon/2)$.

(ii) Write $r := d\mu/d\mu_U$ and $s := d\nu/d\nu_U$. Then

$$KL(\mu\|\nu) = \int \log\frac{d\mu}{d\nu}\, d\mu = \int \log\frac{d\mu}{d\nu_U}\, d\mu + \int \log\frac{d\nu_U}{d\nu}\, d\mu$$
$$= KL(\mu\|\nu_U) + \mathbb{E}_\mu[-\log s].$$

Since $-\log s \le \epsilon$ pointwise, $\mathbb{E}_\mu[-\log s] \le \epsilon$. For the remaining term,

$$KL(\mu\|\nu_U) = \int r\log\Big(r\frac{d\mu_U}{d\nu_U}\Big)\, d\mu_U = \underbrace{\int r\log r\, d\mu_U}_{T_1} + \underbrace{\int r\log\frac{d\mu_U}{d\nu_U}\, d\mu_U}_{T_2}.$$

Under the constraints $\int r\, d\mu_U = 1$ and $r \in [e^{-\epsilon}, e^\epsilon]$, the convex functional $r \mapsto r\log r$ is maximized by a two-point extremal, giving

$$T_1 \le \epsilon\tanh\frac{\epsilon}{2}.$$

For T_2, let $f := \log(d\mu_U/d\nu_U)$. Splitting into positive/negative parts and using $e^{-\epsilon} \le r \le e^\epsilon$,

$$T_2 = \int r f_+\, d\mu_U - \int r(-f)_+\, d\mu_U \le e^\epsilon \int f_+\, d\mu_U - e^{-\epsilon}\int (-f)_+\, d\mu_U$$
$$= e^\epsilon KL(\mu_U\|\nu_U) + (e^{-\epsilon} - e^\epsilon)\int (-f)_+\, d\mu_U \le e^\epsilon KL(\mu_U\|\nu_U),$$

where the last step uses $e^{-\epsilon} - e^\epsilon \le 0$. Adding the bounds on T_1, T_2, and $\mathbb{E}_\mu[-\log s]$ proves the inequality in equation I.2.

(iii) Let $\lambda = \mu \wedge \nu$, $\mu_s = \mu - \lambda$, $\nu_s = \nu - \lambda$. Then $\mu_s(\Theta) = \nu_s(\Theta) = \mathrm{TV}(\mu,\nu)$. Match λ identically and pair μ_s with ν_s arbitrarily, only mass TV moves, each at distance $\le D_\Theta$. Hence $\mathcal{W}_p^p \le D_\Theta^p\ \mathrm{TV}$. □

Theorem I.1 (Sharp High-Probability $\mathcal{W}_p$ Bound)**.** *Let μ_i, ν_i be probability measures on Θ with diameter $D_\Theta < \infty$. Suppose for each $i = 1, 2$, the pair (μ_i, ν_i) is (ϵ, δ)-indistinguishable in both directions (as in Lemma I.2). Then for any $p \geq 1$,*

$$\mathcal{W}_p^p(\nu_1, \nu_2) \leq \big(\mathcal{W}_p(\mu_1, \mu_2) + P^{\mathrm{TV}}(\epsilon, \delta)\big)^p,$$

where $P^{\mathrm{TV}}(\epsilon, \delta) := 2D_\Theta\, C_{\epsilon,\delta}^{1/p}$ and $C_{\epsilon,\delta}$ is defined in Lemma I.2. Let $a_1 \sim \mu_1$ and $a_2 \sim \mu_2$. If additionally $M_q := \mathbb{E}\|a_1 - a_2\|^q < \infty$ for some $q > 0$, then

$$\mathcal{W}_p^p(\nu_1, \nu_2) \leq \begin{cases} \Big(M_q^{1/q} + P^{\mathrm{TV}}(\epsilon, \delta)\Big)^p, & q \geq p, \\ \Big(D_\Theta^{\frac{p-q}{p}}\, M_q^{1/p} + P^{\mathrm{TV}}(\epsilon, \delta)\Big)^p, & q \leq p. \end{cases}$$

The sharpest bound among all choices of q is obtained at $q = p$, yielding

$$\mathcal{W}_p^p(\nu_1, \nu_2) \leq \Big(M_p^{1/p} + P^{\mathrm{TV}}(\epsilon, \delta)\Big)^p.$$

Proof. By triangle inequality for $\mathcal{W}_p$,

$$\mathcal{W}_p(\nu_1, \nu_2) \leq \mathcal{W}_p(\nu_1, \mu_1) + \mathcal{W}_p(\mu_1, \mu_2) + \mathcal{W}_p(\mu_2, \nu_2).$$

From Lemmas I.2–I.2, $\mathcal{W}_p(\mu_i, \nu_i) \leq D_\Theta\ \mathrm{TV}(\mu_i, \nu_i)^{1/p} \leq D_\Theta\, C_{\epsilon,\delta}^{1/p}$ for $i = 1, 2$. This proves the first bound. The moment-based bounds follow from $\mathcal{W}_p(\mu_1, \mu_2) \leq (\mathbb{E}\|a_1 - a_2\|^p)^{1/p}$ and the standard norm monotonicity or diameter interpolation. The optimality of $q = p$ follows since $M_q^{1/q} \geq M_p^{1/p}$ for $q > p$ and $D_\Theta^{\frac{p-q}{p}}\, M_q^{1/p} \geq M_p^{1/p}$ for $q < p$. □

Theorem I.2 (Conversion Theorem (restated).)**.** *Let $\mathcal{U}$ be an (ϵ, δ)-CU algorithm (Def. 4) for $\mathcal{A}$. In all three parts, the unlearned output law for each deleted record is (ϵ, δ)-indistinguishable from the learner's output law after retraining on the retained dataset (Def. 4); each part adds the stability assumption stated below. Assume $\mathcal{L} \geq 0$ in (1).*

(1) SU $\Rightarrow$ ASL. *Let $\mathcal{U}$ be additionally $(\Delta, n+1)$-SU (Def. 9). If $\mathcal{U}$ is ϵ-PCU (Def. 5), then $\mathcal{A}$ is (τ, n)-ASL (Def. 6) with*

$$\tau \leq e^\epsilon \Delta + 2\sinh(\epsilon)\mathbb{E}_{\mathcal{U},\mathcal{A},\mathcal{D}_{n+1}}\left[R^{\mathrm{un}}[\mathcal{D}_{n+1}] \vee R^{\mathrm{fg}}[\mathcal{D}_{n+1}]\right], \tag{23}$$

where $R^{\mathrm{un}}[\mathcal{D}_N] := \mathbb{E}_{\mathcal{U},\mathcal{A}}[\mathcal{L}(u^{(i)}, z_i)]$ and $R^{\mathrm{fg}}[\mathcal{D}_N] := \mathbb{E}_{\mathcal{A}}[\mathcal{L}(\mathcal{A}(\mathcal{D}_N^{-i}), z_i)]$, averaged over the relevant indices. If only Def. 4 holds and $|\mathcal{L}| \leq \rho$ (Def. 22), then $\mathcal{A}$ is (τ, n)-ASL with

$$\tau \leq \Delta + 4\rho C_{\epsilon,\delta}. \tag{24}$$

(2) USU/MSU $\Rightarrow$ $\mathcal{W}_p$-USL. *Let $p \geq 1$, $\mathrm{diam}(\Theta) \leq D_\Theta < \infty$, and let $C_{\epsilon,\delta}$ be as in Lemma I.2. Let $r_{T_u,p}$ be the uniform $\mathcal{W}_p$ radius between each unlearned output law and the output law of the learner retrained without the same deleted record, as defined in Section 2. If $\mathcal{U}$ is additionally $(\Delta_1, n+1)$-$\mathcal{W}_p$-USU (Def. 10), then $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-USL (Def. 7) with*

$$\tau \leq \Delta_1 + 2r_{T_u,p} \leq \Delta_1 + P^{\mathrm{TV}}(\epsilon, \delta). \tag{25}$$

If $\mathcal{U}$ is additionally $(\Delta_2, n+1)$-l_p-MSU (Def. 11), then the learner output laws satisfy (τ, n)-$\mathcal{W}_p$-USL with $\tau \leq \Delta_2 + 2r_{T_u,p} \leq \Delta_2 + P^{\mathrm{TV}}(\epsilon, \delta)$.

(3) KLSU $\Rightarrow$ KLSL. *If $\mathcal{U}$ is additionally $(\Delta, 2n)$-KLSU (Def. 12), and no further stability hypothesis such as SU/USU/MSU, and if $\mathcal{U}$ is ϵ-PCU (Def. 5), then $\mathcal{A}$ is (τ, n)-KLSL (Def. 8) with*

$$\tau \leq e^\epsilon \Delta + \epsilon\big(1 + \tanh(\epsilon/2)\big). \tag{26}$$

In particular, the only stability input is KLSU, the certificate is used solely to replace $(P_{u^{(1)}}, P_{u^{(2)}})$ by $(P_{\mathcal{A}(\mathcal{D}^{(1)})}, P_{\mathcal{A}(\mathcal{D}^{(2)})})$ in Lemma I.2. Each conclusion is unconditional once the assumptions stated in its part hold. In particular, Def. 5 is required for parts (1) and (3) because it supplies the two-sided density-ratio bounds used by Lemma I.2; Def. 4 with $\delta > 0$ supplies no such bound and, being an inequality between measures, defines no event on which one could condition (Rem. 1).

Proof. **Part (1):** Write $N := n+1$ and fix an N-sample dataset $\mathcal{D}_N = \{z_1, \dots, z_N\}$. Let $\mathcal{D}_N^{-i} := \mathcal{D}_N \backslash \{z_i\}$ and $u^{(i)} := \mathcal{U}(\mathcal{A}(\mathcal{D}_N), \{z_i\})$. By Lemma I.1, the ASL gap equals

$$G := \mathbb{E}_{\mathcal{D}_N}\left[\frac{1}{N(N-1)}\sum_{i\neq j}\mathbb{E}_{\mathcal{A}}\left[\mathcal{L}\left(\mathcal{A}(\mathcal{D}_N^{-i}), z_i\right) - \mathcal{L}\left(\mathcal{A}(\mathcal{D}_N^{-j}), z_i\right)\right]\right], \tag{27}$$

and it suffices to bound $|G|$. For each ordered pair (i,j) with $i \neq j$, write

$$a := \mathbb{E}_{\mathcal{A}}\big[\mathcal{L}(\mathcal{A}(\mathcal{D}_N^{-i}), z_i)\big], \quad b := \mathbb{E}_{\mathcal{A}}\big[\mathcal{L}(\mathcal{A}(\mathcal{D}_N^{-j}), z_i)\big],$$
$$a_u := \mathbb{E}_{\mathcal{U},\mathcal{A}}\big[\mathcal{L}(u^{(i)}, z_i)\big], \quad b_u := \mathbb{E}_{\mathcal{U},\mathcal{A}}\big[\mathcal{L}(u^{(j)}, z_i)\big].$$

By Def. 5 the density ratio between $P_{\mathcal{A}(\mathcal{D}_N^{-i})}$ and $P_{u^{(i)}}$ lies in $[e^{-\epsilon}, e^{\epsilon}]$ almost everywhere, and likewise for j, both statements are deterministic. Since $\mathcal{L} \geq 0$, integrating the ratio bounds gives

$$e^{-\epsilon}a_u \leq a \leq e^{\epsilon}a_u, \qquad e^{-\epsilon}b_u \leq b \leq e^{\epsilon}b_u,$$

hence $a-b \leq e^{\epsilon}(a_u-b_u)+2\sinh(\epsilon)\, b_u$ and $b-a \leq e^{\epsilon}(b_u-a_u)+2\sinh(\epsilon)\, a_u$. Taking expectations over pairs and $\mathcal{D}_N$ gives

$$G \leq e^{\epsilon}\,\mathbb{E}[a_u - b_u] + 2\sinh(\epsilon)\,\mathbb{E}[b_u],$$
$$-G \leq e^{\epsilon}\,\mathbb{E}[b_u - a_u] + 2\sinh(\epsilon)\,\mathbb{E}[a_u],$$

and therefore

$$|G| \leq e^{\epsilon}\left|\mathbb{E}[a_u - b_u]\right| + 2\sinh(\epsilon)\max\big\{\mathbb{E}[a_u],\, \mathbb{E}[b_u]\big\}.$$

The index swap $(i,j) \mapsto (j,i)$ identifies $\mathbb{E}[|a_u - b_u|]$ with the (Δ, N)-SU average, so $\left|\mathbb{E}[a_u - b_u]\right| \leq \mathbb{E}[|a_u - b_u|] \leq \Delta$. Moreover, averaging b_u over $i \neq j$ recovers $\mathbb{E}[R^{\mathrm{un}}[\mathcal{D}_N]]$, while averaging a_u recovers $\mathbb{E}[R^{\mathrm{fg}}[\mathcal{D}_N]]$. Hence

$$|G| \leq e^{\epsilon}\Delta + 2\sinh(\epsilon)\max\big\{\mathbb{E}[R^{\mathrm{un}}],\, \mathbb{E}[R^{\mathrm{fg}}]\big\} \leq e^{\epsilon}\Delta + 2\sinh(\epsilon)\,\mathbb{E}\big[R^{\mathrm{un}} \vee R^{\mathrm{fg}}\big],$$

which proves the ASL inequality in equation 23.

For the second form, assume only Def. 4 and $|\mathcal{L}| \leq \rho$. Lemma I.2 gives $\mathrm{TV}\big(P_{\mathcal{A}(\mathcal{D}_N^{-k})}, P_{u^{(k)}}\big) \leq C_{\epsilon,\delta}$ for $k = i, j$, a deterministic inequality. Since $\theta \mapsto \mathcal{L}(\theta, z_i)$ is bounded by ρ, the elementary estimate $|\mathbb{E}_{\mu}f - \mathbb{E}_{\nu}f| \leq 2\|f\|_{\infty}\,\mathrm{TV}(\mu,\nu)$ yields $|a-a_u| \leq 2\rho C_{\epsilon,\delta}$ and $|b-b_u| \leq 2\rho C_{\epsilon,\delta}$, whence

$$|a-b| \leq |a_u - b_u| + 4\rho C_{\epsilon,\delta}.$$

Averaging over pairs and $\mathcal{D}_N$ and using (Δ, N)-SU gives $|G| \leq \Delta + 4\rho C_{\epsilon,\delta}$. No density-ratio bound and no conditioning are used, so the statement is unconditional and valid for every $\delta \in [0, 1/2]$.

Part (2): We first prove the $\mathcal{W}_p$-USU implication. The l_p-MSU implication uses the triangle inequality in the preceding display, with the middle term bounded by the L^p coupling parameter Δ_2; the substitution appears below. Let $\mathcal{D}_N$ be any dataset of size $N = n+1$. For any $i \neq j$, let $u^{(i)} = \mathcal{U}(\mathcal{A}(\mathcal{D}_N), \{z_i\})$ and $u^{(j)} = \mathcal{U}(\mathcal{A}(\mathcal{D}_N), \{z_j\})$.

The sets $\mathcal{D}_N^{-i}$ and $\mathcal{D}_N^{-j}$ each contain n records. The set $\mathcal{D}_N^{-i}$ contains z_j but not z_i, the set $\mathcal{D}_N^{-j}$ contains z_i but not z_j, and the two sets share the remaining $n-1$ records; therefore they are neighboring datasets in the replacement sense of Def. 7. Their tuple order can differ, but the exchangeable sampler assumption in Section 2 makes $P_{\mathcal{A}(S)}$ invariant to permutations of S, so Lemma I.1 applies to the output laws. Conversely, any replacement pair $(\mathcal{D}_n, \mathcal{D}_n^{(i)})$ from Def. 7 can be represented by $\mathcal{D}_N = \mathcal{D}_n \cup \{z_i'\}$ with deleted index N, for which $\mathcal{D}_N^{-N} = \mathcal{D}_n$ and $\mathcal{D}_N^{-i} = \mathcal{D}_n^{(i)}$. Thus the supremum over $(\mathcal{D}_N, i, j)$ covers every pair in Def. 7.

By (ϵ,δ)-CU, the pairs $\big(P_{\mathcal{A}(\mathcal{D}_N^{-i})}, P_{u^{(i)}}\big)$ and $\big(P_{\mathcal{A}(\mathcal{D}_N^{-j})}, P_{u^{(j)}}\big)$ are each (ϵ,δ)-indistinguishable in both directions. Lemma I.2 therefore gives

$$\mathrm{TV}\left(P_{\mathcal{A}(\mathcal{D}_N^{-k})},\, P_{u^{(k)}}\right) \leq C_{\epsilon,\delta}, \quad k = i, j.$$

Triangle inequality for $\mathcal{W}_p$ and Lemma I.2 yield

$$\mathcal{W}_p\big(P_{\mathcal{A}(\mathcal{D}_N^{-i})},\, P_{\mathcal{A}(\mathcal{D}_N^{-j})}\big) \leq \mathcal{W}_p(P_{\mathcal{A}(\mathcal{D}_N^{-i})}, P_{u^{(i)}}) + \mathcal{W}_p(P_{u^{(i)}}, P_{u^{(j)}}) + \mathcal{W}_p(P_{u^{(j)}}, P_{\mathcal{A}(\mathcal{D}_N^{-j})})$$
$$\leq r_{T_u,p} + \Delta_1 + r_{T_u,p} = \Delta_1 + 2r_{T_u,p} \leq \Delta_1 + P^{\mathrm{TV}}(\epsilon,\delta),$$

where the middle term uses (Δ_1, N)-$\mathcal{W}_p$-USU. Since this holds for all $i \neq j$ and all $\mathcal{D}_N$, and since by the first paragraph every replacement pair of size n is realized as some $(\mathcal{D}_N^{-i}, \mathcal{D}_N^{-j})$,

$$\sup_{\mathcal{D}_n, i} \mathcal{W}_p\big(P_{\mathcal{A}(\mathcal{D}_n)}, P_{\mathcal{A}(\mathcal{D}_n^{(i)})}\big) \leq \Delta_1 + 2r_{T_u,p} \leq \Delta_1 + P^{\mathrm{TV}}(\epsilon, \delta),$$

i.e., $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-USL with the direct endpoint term $\tau = \Delta_1 + 2r_{T_u,p}$.

If instead $\mathcal{U}$ is (Δ_2, N)-l_p-MSU, then $\mathcal{W}_p(P_{u^{(i)}}, P_{u^{(j)}}) \leq (\mathbb{E}\|u^{(i)} - u^{(j)}\|^p)^{1/p} \leq \Delta_2$, and the triangle inequality in the preceding display gives $\tau \leq \Delta_2 + 2r_{T_u,p} \leq \Delta_2 + P^{\mathrm{TV}}(\epsilon, \delta)$. This proves Part (2).

Part (3) (KLSU $\Rightarrow$ KLSL): The only stability hypothesis used here is KLSU, CU enters only as a bridge from unlearned models to retrained models. Let $\mathcal{D}_{2n}$ be an *i.i.d.* dataset of size $2n$, let $\mathcal{I}_{2n} \sim \mathcal{S}_3$, and write $\mathcal{D}^{(1)} := \mathcal{D}_{2n}[\mathcal{I}_{2n}[:n]]$, $\mathcal{D}^{(2)} := \mathcal{D}_{2n}[\mathcal{I}_{2n}[n:]]$. For *i.i.d.* data, the random half-split $(\mathcal{D}^{(1)}, \mathcal{D}^{(2)})$ has the same law as a pair of independent n-sample datasets $(\mathcal{D}_n, \mathcal{D}'_n)$. Set

$$\mu := P_{\mathcal{A}(\mathcal{D}^{(1)})}, \qquad \nu := P_{\mathcal{A}(\mathcal{D}^{(2)})},$$
$$\mu_U := P_{u^{(1)}}, \qquad \nu_U := P_{u^{(2)}},$$

where $u^{(1)} = \mathcal{U}(\mathcal{A}(\mathcal{D}_{2n}), \mathcal{D}^{(2)})$ and $u^{(2)} = \mathcal{U}(\mathcal{A}(\mathcal{D}_{2n}), \mathcal{D}^{(1)})$.

- **Stability input (KLSU only).** By $(\Delta, 2n)$-KLSU,

  $$\mathbb{E}_{\mathcal{D}_{2n}, \mathcal{S}_3}\big[KL(\mu_U \| \nu_U)\big] \leq \Delta.$$

  No SU/USU/MSU (or any other stability notion) is invoked.

- **Bridge (certificate only).** By Def. 5 applied to the forget sets $\mathcal{D}^{(2)}$ and $\mathcal{D}^{(1)}$,

  $$e^{-\epsilon} \leq \frac{d\mu}{d\mu_U} \leq e^{\epsilon}, \qquad e^{-\epsilon} \leq \frac{d\nu}{d\nu_U} \leq e^{\epsilon}$$

  almost everywhere. These density-ratio inequalities are deterministic statements about (μ, μ_U) and (ν, ν_U). This density-ratio step is the only step that Def. 4 with $\delta > 0$ cannot provide, because the additive slack permits $\mu \not\ll \mu_U$ and therefore can make $KL(\mu\|\nu) = +\infty$.

Lemma I.2 therefore yields

$$KL(\mu\|\nu) \leq e^{\epsilon}\, KL(\mu_U\|\nu_U) + \epsilon\Big(1 + \tanh\frac{\epsilon}{2}\Big).$$

Taking expectation over $\mathcal{D}_{2n}$ and $\mathcal{S}_3$ and using $(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}) \overset{d}{=} (\mathcal{D}_n, \mathcal{D}'_n)$ gives

$$\mathbb{E}_{\mathcal{D}_n, \mathcal{D}'_n}\big[KL(P_{\mathcal{A}(\mathcal{D}_n)}\|P_{\mathcal{A}(\mathcal{D}'_n)})\big] \leq e^{\epsilon}\Delta + \epsilon\Big(1 + \tanh\frac{\epsilon}{2}\Big),$$

i.e., (τ, n)-KLSL, with no exceptional event. (If $\epsilon = 0$, the certificate is exact retrain and the bound collapses to $\tau \leq \Delta$, so KLSU alone determines the rate.)

□

Remark 3 (On the role of δ in the three parts.)**.** *The three Conversion Theorem parts use the CU certificate in different ways, and only some parts tolerate a positive δ slack. Part (2) tolerates positive δ completely: Lemma I.2 is an exact inequality between probability laws and returns the deterministic constant $C_{\epsilon,\delta}$, in which δ appears additively inside a magnitude rather than as a failure probability. Part (1) tolerates positive δ at a constant cost through $|\mathbb{E}_\mu f - \mathbb{E}_\nu f| \leq 2\|f\|_\infty \mathrm{TV}(\mu, \nu)$, which is why equation 24 carries $4\rho C_{\epsilon,\delta}$ and needs Def. 22. Part (3) does not tolerate positive δ, because relative entropy is not a bounded functional of total variation. An earlier form of this conversion attached probability $1 - 2\delta$ to parts (1) and (3), but Def. 4 is an inequality between probability laws and does not define an event over the random coins of $\mathcal{U}$ on which one could condition. The valid statements are therefore: PCU gives the KL conclusion without an exceptional event, while CU with positive δ gives the total-variation conclusion.*

J AUXILIARY LEMMAS FROM STABILITY TO GENERALIZATION

This appendix collects the conversion lemmas used in the subsection on excess population risk. None is new; each is stated in the exact form used in the main text.

Write $\theta^{\mathrm{out}} = \mathcal{A}(\mathcal{D}_n)$, choose $\theta^\star \in \arg\min_{\theta\in\Theta} R(\theta)$, and set $R^\star := R(\theta^\star)$. Then decompose

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \le \mathbb{E}\big[R(\theta^{\mathrm{out}}) - R_{\mathcal{D}_n}(\theta^{\mathrm{out}})\big] + \epsilon_{\mathrm{opt}}, \tag{28}$$

where $\epsilon_{\mathrm{opt}} := \mathbb{E}[R_{\mathcal{D}_n}(\theta^{\mathrm{out}}) - R_{\mathcal{D}_n}(\theta^\star)]$. Since $\theta^\star$ is fixed independently of $\mathcal{D}_n$, $\mathbb{E}[R_{\mathcal{D}_n}(\theta^\star)] = R^\star$.

Lemma J.1 (Self-Bounding Property (Lei & Ying, 2020)). *Let $f : \Theta \mapsto \mathbf{R}$ be nonnegative, differentiable and β-smooth on $\mathbf{R}^d$. Then $\|\nabla f(\theta)\|^2 \le 2\beta f(\theta)$ for all θ.*

Proof. By β-smoothness, $0 \le f(\theta - \beta^{-1}\nabla f(\theta)) \le f(\theta) - \beta^{-1}\|\nabla f(\theta)\|^2 + \frac{1}{2\beta}\|\nabla f(\theta)\|^2 = f(\theta) - \frac{1}{2\beta}\|\nabla f(\theta)\|^2$. □

Lemma J.2 (The Generalization Gap Depends Only on the Two Output Laws (Lei & Ying, 2020; Zhu et al., 2023)). *Let P_1, P_2 be Borel probability measures on a Polish space Θ and let $\phi : \Theta \mapsto \mathbf{R}$ be measurable and P_k-integrable. Then $\mathbb{E}_{P_1}[\phi] - \mathbb{E}_{P_2}[\phi] = \mathbb{E}_\pi[\phi(\vartheta_1) - \phi(\vartheta_2)]$ for every pairing $\pi \in \Pi(P_1, P_2)$. Consequently, a finite $\mathcal{W}_p$-USL bound may be used with an optimal pairing of each neighbouring pair when bounding the generalization gap. For a squared-distance argument one must use $p = 2$ (or a stronger moment bound). This statement does not bound the distance under an arbitrary prescribed pairing of the algorithm's random seeds.*

Proof. The first identity holds because the marginals of π are P_1 and P_2. For probability laws with finite pth moments on a Polish metric space, the lower-semicontinuous distance cost attains its infimum. Choose an optimal pairing for each neighbouring dataset pair, or use pairings that approach the optimum and pass to the limit. The expected loss difference depends only on the marginals, so this choice leaves the generalization gap unchanged. The distance between paired outputs does depend on how the two runs are paired. □

Lemma J.3 (Generalization under Def. 16 (Bousquet & Elisseeff, 2002; Hardt et al., 2016)). *Let $\mathcal{L}(\cdot, z)$ be L-Lipschitz and let $\mathcal{A}$ be (τ, n)-$\mathcal{W}_p$-USL for some $p \ge 1$. Then*

$$\Big|\mathbb{E}\big[R(\mathcal{A}(\mathcal{D}_n)) - R_{\mathcal{D}_n}(\mathcal{A}(\mathcal{D}_n))\big]\Big| \le L\tau.$$

Proof. By the leave-one-out symmetrization identity for *i.i.d.* data,

$$\mathbb{E}\big[R(\mathcal{A}(\mathcal{D}_n)) - R_{\mathcal{D}_n}(\mathcal{A}(\mathcal{D}_n))\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\mathcal{L}(\mathcal{A}(\mathcal{D}_n^{(i)}), z_i) - \mathcal{L}(\mathcal{A}(\mathcal{D}_n), z_i)\big].$$

For fixed $\mathcal{D}_n, \mathcal{D}_n'$ and z_i, the summand is a difference of expectations of the L-Lipschitz function $\theta \mapsto \mathcal{L}(\theta, z_i)$ under two laws, hence bounded by $L\mathcal{W}_1$ by Kantorovich duality. Finally $\mathcal{W}_1 \le \mathcal{W}_p$ by Jensen's inequality. □

Lemma J.4 (Generalization without Def. 16 (Lei & Ying, 2020)). *Let $\mathcal{L} \ge 0$ be β-smooth in θ and let $\mathcal{A}$ be (τ, n)-$\mathcal{W}_2$-USL. Then for every $c_0 > 0$,*

$$\mathbb{E}\big[R - R_{\mathcal{D}_n}\big] \le c_0\,\mathbb{E}\big[R_{\mathcal{D}_n}(\theta^{\mathrm{out}})\big] + \frac{\beta(1 + c_0)}{2c_0}\tau^2. \tag{29}$$

Proof. Write $u = \mathcal{A}(\mathcal{D}_n^{(i)})$ and $v = \mathcal{A}(\mathcal{D}_n)$. By smoothness and Lem. J.1,

$$\mathcal{L}(u, z_i) - \mathcal{L}(v, z_i) \le \sqrt{2\beta\mathcal{L}(v, z_i)}\|u - v\| + \frac{\beta}{2}\|u - v\|^2.$$

Young's inequality $xy \le \frac{x^2}{2\lambda} + \frac{\lambda y^2}{2}$ with $\lambda = \beta/c_0$ gives $\sqrt{2\beta\mathcal{L}(v, z_i)}\|u - v\| \le c_0\mathcal{L}(v, z_i) + \frac{\beta}{2c_0}\|u - v\|^2$. Averaging over i, using the identity in the proof of Lem. J.3 and then Lem. J.2 to evaluate $\frac{1}{n}\sum_i \mathbb{E}\|u - v\|^2$ under the $\mathcal{W}_2$-optimal pairing, yields the claim. This is the form of Theorem 2 of (Lei & Ying, 2020). □

Lemma J.5 (Conditional fast risk from the distance to retraining)**.** *Let $\mathcal{L} \geq 0$ be β-smooth on $\mathbb{R}^d$. Suppose that the learner has finite second moments and that*

$$\gamma_{\mathrm{ms}}^2 := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\, \mathcal{W}_2^2\Big(P_{\mathcal{A}(\mathcal{D}_n)}, P_{\mathcal{A}(\mathcal{D}_n^{(i)})}\Big) < \infty.$$

If $\mathbb{E}R_{\mathcal{D}_n}(\theta^{\mathrm{out}}) \leq R^\star + \epsilon_{\mathrm{opt}}$, with $\epsilon_{\mathrm{opt}} \geq 0$ including the effect of publication, then

$$\mathbb{E}R(\theta^{\mathrm{out}}) - R^\star \leq \epsilon_{\mathrm{opt}} + \sqrt{2\beta(R^\star + \epsilon_{\mathrm{opt}})}\,\gamma_{\mathrm{ms}} + \frac{\beta}{2}\gamma_{\mathrm{ms}}^2. \tag{30}$$

In particular, the $p = 2$ worst-case-distance transfer bound gives $\gamma_{\mathrm{ms}} \leq \Delta_2 + 2r_{T_u,2}$ whenever both terms are uniform. If $R^\star = 0$, $\epsilon_{\mathrm{opt}} = O(n^{-2})$, and $\Delta_2 + 2r_{T_u,2} = O(n^{-1})$, the excess risk is $O(n^{-2})$.

Proof. Apply the smoothness and self-bounding inequality in the proof of Lemma J.4 under a $\mathcal{W}_2$-optimal pairing for each replacement pair. Cauchy–Schwarz, over the sample, replacement index, and paired outputs, bounds the linear term by $\sqrt{2\beta\,\mathbb{E}R_{\mathcal{D}_n}(\theta^{\mathrm{out}})}\,\gamma_{\mathrm{ms}}$; the quadratic term is $\beta\gamma_{\mathrm{ms}}^2/2$. Add equation 28 and the assumed empirical-risk bound. The uniform bound based on the worst-case distance bounds each Wasserstein distance by $\Delta_2 + 2r_{T_u,2}$. This proves the displayed rate under the assumptions stated in the lemma. □

Lemma J.6 (Generalization under KL Stability (Def. 8) (Xu & Raginsky, 2017))**.** *Let $|\mathcal{L}(\theta, z)| \leq \rho$ (Def. 22) and let $\mathcal{A}$ be (τ, n)-KLSL (Def. 8). Then $I(\mathcal{A}(\mathcal{D}_n), \mathcal{D}_n) \leq \tau$ and*

$$\Big|\mathbb{E}\big[R(\mathcal{A}(\mathcal{D}_n)) - R_{\mathcal{D}_n}(\mathcal{A}(\mathcal{D}_n))\big]\Big| \leq \rho\sqrt{\frac{2\tau}{n}}.$$

Proof. Let $\bar{P} = \mathbb{E}_{\mathcal{D}_n'}[P_{\mathcal{A}(\mathcal{D}_n')}]$ be the marginal law of the output. By convexity of KL in its second argument, $KL(P_{\mathcal{A}(\mathcal{D}_n)}\|\bar{P}) \leq \mathbb{E}_{\mathcal{D}_n'} KL(P_{\mathcal{A}(\mathcal{D}_n)}\|P_{\mathcal{A}(\mathcal{D}_n')})$, taking $\mathbb{E}_{\mathcal{D}_n}$ gives $I(\mathcal{A}(\mathcal{D}_n), \mathcal{D}_n) \leq \tau$. A loss bounded by ρ is ρ-sub-Gaussian, so the information-theoretic generalization bound of (Xu & Raginsky, 2017) applies. □

Lem. J.6 is an unconditional implication, and it is exact only if the KLSL parameter τ fed into it is itself unconditional. This is the reason Conversion Theorem 1(3) is stated under Def. 5 rather than Def. 4: were τ available only after discarding an exceptional set, the conclusion of Lem. J.6 would be a conditional statement, and since the generalization gap of a ρ-bounded loss can be as large as 2ρ off that set, the resulting unconditional bound would carry an additive 2ρ term that no choice of δ makes small relative to $\rho\sqrt{2\tau/n}$. Under Def. 5 the question does not arise. Under Def. 4 with $\delta > 0$ one abandons this lemma and uses Conversion Theorem 1(1) in the form equation 24 instead, which bounds the same generalization gap directly and unconditionally.

Lemma J.7 (Quadratic-growth condition for the PL route)**.** *Let $F : \Theta \mapsto \mathbf{R}$ be differentiable with nonempty minimizer set $\mathcal{X}^\star$. Assume that F satisfies the quadratic-growth condition*

$$F(\theta) - \min F \geq \frac{\mu_{\mathrm{PL}}}{2}\mathrm{dist}(\theta, \mathcal{X}^\star)^2 \quad \textit{for every } \theta \in \Theta.$$

The PL–QG theorem used in the paper requires this condition (for example, it follows from the PL condition under the additional hypotheses stated in the citepd equivalence results (Karimi et al., 2016)); PL alone is not used as a substitute for QG.

Proof. The displayed inequality is the assumed quadratic-growth condition. □

Lemma J.8 (Minimizer Stability under PL (Charles & Papailiopoulos, 2018))**.** *Let $\mathcal{L}(\cdot, z)$ be L-Lipschitz and suppose $R_{\mathcal{D}_N^{-i}}$ and $R_{\mathcal{D}_N^{-j}}$ both satisfy the μ_{PL}-PL condition on Θ, where $\mathcal{D}_N^{-i}$ and $\mathcal{D}_N^{-j}$ each contain n points and differ by the swap $z_j \leftrightarrow z_i$. Assume that the minimizer sets are singletons, written $\{\theta_i^\star\}$ and $\{\theta_j^\star\}$, or, more generally, that the selected minimizers are mutually nearest-point selections. Then For $\ell \in \{i, j\}$, define $\mathcal{X}_{\setminus \ell}^\star := \arg\min_{\theta\in\Theta} R_{\mathcal{D}_N^{-\ell}}(\theta)$.*

$$\mathrm{dist}\big(\theta_i^\star, \mathcal{X}_{\setminus j}^\star\big) \leq \frac{4L}{\mu_{\mathrm{PL}} n}.$$

Proof. Write $R_{\mathcal{D}_N^{-j}} = R_{\mathcal{D}_N^{-i}} + \frac{1}{n}(\mathcal{L}(\cdot, z_i) - \mathcal{L}(\cdot, z_j))$, let $\theta_j^\star$ be the projection of $\theta_i^\star$ onto $\mathcal{X}_{\setminus j}^\star$, and set $d := \|\theta_i^\star - \theta_j^\star\|$. Then

$$\begin{aligned} R_{\mathcal{D}_N^{-j}}(\theta_i^\star) - R_{\mathcal{D}_N^{-j}}(\theta_j^\star) &= \big[R_{\mathcal{D}_N^{-i}}(\theta_i^\star) - R_{\mathcal{D}_N^{-i}}(\theta_j^\star)\big] \\ &\quad + \tfrac{1}{n}\big[\big(\mathcal{L}(\theta_i^\star, z_i) - \mathcal{L}(\theta_j^\star, z_i)\big) - \big(\mathcal{L}(\theta_i^\star, z_j) - \mathcal{L}(\theta_j^\star, z_j)\big)\big] \\ &\leq 0 + \tfrac{2L}{n} d, \end{aligned}$$

The term $R_{\mathcal{D}_N^{-i}}(\theta_i^\star) - R_{\mathcal{D}_N^{-i}}(\theta_j^\star)$ is nonpositive because $\theta_i^\star$ minimizes $R_{\mathcal{D}_N^{-i}}$. Lemma J.7 applied to $R_{\mathcal{D}_N^{-j}}$ gives $\frac{\mu_{\mathrm{PL}}}{2} d^2 \leq \frac{2L}{n} d$, i.e. $d \leq \frac{4L}{\mu_{\mathrm{PL}} n}$. This calculation proves the displayed minimizer-distance bound and is the argument used in (Charles & Papailiopoulos, 2018). □

K Certified Unlearning Algorithms with Shared References

Each row of Tables E.1 to E.3 describes a learning algorithm under a set of assumptions. To apply Conversion Theorem 1, we must pair each row with a certified unlearner that operates under the same assumptions. This appendix records the three pairings used in the main text. For each pairing, we state the update mechanism, the exact privacy or indistinguishability statement proved by that mechanism, and the stability conclusions available in the citepd source. Notation is as follows. The symbol L is the Lipschitz constant of Def. 16. The symbol β is the smoothness of Def. 21. The symbols γ or m are the strong convexity or dissipativity constants. The symbol D_Θ is the domain diameter. The symbol T is the number of training updates. The symbol T_u is the number of unlearning updates. The symbol σ is the injected noise scale. A caution on notation is needed. The sources conflict. In particular L denotes smoothness in (Chien et al., 2024a;b; Mu & Klabjan, 2026; 2025; Ullah et al., 2021) while it denotes the Lipschitz constant in (Neel et al., 2021) and here. All formulas below have been translated.

Pairing 1: strongly convex learners (Table E.1) with PNSGD and descent to delete. *Mechanism.* Projected noisy SGD (Chien et al., 2024a) draws a single cyclic mini-batch partition of $[n]$ once and reuses it for both learning and unlearning. The unlearner is the identical recursion run on the retained dataset for T_u epochs, initialized at the final learning iterate. Descent to delete (Neel et al., 2021) instead runs warm-started projected gradient descent on the updated objective and perturbs the published output. It has a secret-state regime, where the internal iterate is passed forward unnoised, and a perfect regime, where the published noisy iterate is passed forward.

Guarantee type. PNSGD proves a Rényi differential privacy bound and converts it to (ϵ, δ) by numerical optimization over the Rényi order. The central device is the shifted Rényi divergence of (Feldman et al., 2018), so the relevant notion of sensitivity is a $\mathcal{W}_\infty$ distance rather than a Euclidean one. Descent to delete proves (ϵ, δ) indistinguishability directly, choosing the noise scale from the concentrated-privacy inversion of $\epsilon = \frac{s^2}{2\sigma^2} + \frac{s}{\sigma}\sqrt{2\ln(1/\delta)}$, which is valid for $\epsilon = O(\ln(1/\delta))$ and is tighter than the classical form of Lem. L.1. Both imply Def. 4.

Known stability conclusions. Let $\alpha > 1$ denote the Rényi order and let $\varsigma_0 \in (0, 1)$ denote the failure probability in the empirical-risk statement below. The one-step map contracts with coefficient $c = 1 - \eta\gamma$ for $\eta \leq \frac{1}{\beta}$. The $\mathcal{W}_\infty$ sensitivity of the learning process saturates at

$$Z = \min\Big\{\frac{2\eta L}{b\,(1 - c^{n/b})},\ 2D_\Theta\Big\},$$

where b is the batch size. At full batch, the sensitivity is $\min\{\frac{2L}{\gamma n}, 2D_\Theta\}$, which coincides with the empirical risk minimizer sensitivity $\frac{2L}{\gamma n}$ proved as Lemma 8 of (Neel et al., 2021) and with the $T_u = 0$ value of our Lem. L.2(i). The certificate satisfies $\epsilon_\alpha \leq \frac{\alpha Z^2}{2\eta\sigma^2} c^{2T_u n/b}$. Thus the CU-implied distance decreases geometrically in the unlearning depth and is uniform in T; longer training only helps because the distance-to-stationarity term carries $c^{2Tn/b}$. Neither citepd method derives a population-risk bound from its CU certificate. In (Neel et al., 2021), accuracy is measured against the empirical minimizer and the excess empirical risk is $O\big(\frac{\beta L^2 c^{2T_u} d \ln(1/\delta) \ln^2(d/\varsigma_0)}{(1-c^{T_u})^2 \gamma^2 \epsilon^2 n^2}\big)$ with probability $1 - \varsigma_0$. The direct contractive bound in Theorem N.1 first converts the CU certificate into learner stability, and Theorem O.1 then converts learner stability into a population-risk bound.

Pairing 2: convex learners (Table E.2) with TV-stable noisy SGD. *Mechanism.* The learner of (Ullah et al., 2021) is accelerated mini-batch SGD with Gaussian gradient noise. The unlearner uses the same mini-batch index sequence in the two runs and then performs, at each step, rejection sampling on the ratio of the two conditional Gaussian densities, falling back on a reflection map and a partial retrain when rejection occurs. A sub-sampled full-batch variant obtains the same effect from sub-sampling alone, with no injected noise.

Guarantee type. The method of (Ullah et al., 2021) directly guarantees exact equality between the unlearned and retrained output laws; it does not first state an (ϵ, δ) indistinguishability bound. The same method also proves ρ-total variation stability (Def. 13), which yields a pairing with small expected distance. Since ρ-TV stability implies $(\rho\sqrt{\ln(1/\delta)}, \delta)$ differential privacy, Def. 4 still applies, and Conversion Theorem 1 may be used with $\epsilon = \rho\sqrt{\ln(1/\delta)}$. In the exact case $\epsilon = \delta = 0$, the conversion is lossless because $C_{\epsilon,\delta} = 0$.

Known stability conclusions. Setting $\sigma = \frac{8\sqrt{T}L}{n\rho}$ yields $\min\{\rho, 1\}$-TV stability (Def. 13); the source states that this rate is tight in all parameters. For k edits, the recomputation probability is $O(k\rho\sqrt{T})$, and the regime $\rho \gtrsim 1/n$ is necessary by a matching lower bound. The population-risk bound is $O\big(\frac{LD_\Theta}{\sqrt{n}} + LD_\Theta \min\{\frac{1}{\sqrt{\rho n}}, \frac{\sqrt{d}}{\rho n}\}\big)$. That bound is obtained from the classical uniform-stability parameter $\frac{L^2\eta T}{n}$ (Def. 7), rather than from the CU certificate. Two consequences matter here. First, the required noise grows as $\sigma \propto \sqrt{T}$, so the T-dependence removed from the learner-stability bound by Thm. N.2 reappears in the utility of the noisy mechanism. Second, the convex gradient map is nonexpansive but not contractive, so the certificate does not improve geometrically with T_u. The convex corollary of (Chien et al., 2024a) gives $\epsilon_\alpha \le \frac{\alpha Z^2}{2\eta\sigma^2} \cdot \frac{b}{T_u n}$ with $Z = \min\{2\eta LT/b, 2D_\Theta\}$. This is a $1/T_u$ improvement, and Z grows with T until it saturates; these two facts produce the interior optimum in Thm. N.2.

Pairing 3: nonconvex learners (Table E.3) with rewind to delete and Langevin unlearning. *Mechanism.* Rewind to delete (Mu & Klabjan, 2026) saves the checkpoint θ_{T-T_u} during training, reruns T_u gradient steps on the retained data, and publishes with Gaussian noise. When no checkpoint was stored it is recovered post hoc by proximal point steps, which invert a gradient step exactly when $\eta < 1/\beta$. The stochastic versions (Mu & Klabjan, 2025) compare this with the descend variant anchored at θ_T. Langevin unlearning (Chien et al., 2024b) runs projected noisy gradient descent to stationarity and continues the same dynamics on the retained data.

Guarantee type. Rewind to delete uses the plain (ϵ, δ) Gaussian mechanism of Lem. L.1. Langevin unlearning proves a Rényi bound of the privacy recuperation form $\epsilon \le \exp\big(-\frac{1}{\alpha}\sum_{k<T_u} R_k\big)\epsilon_0$, where ϵ_0 is the Rényi order of the learning process and each R_k is a per-step entropy decay driven by the log-Sobolev constant. Under Def. 27 the R_k are constant in k and the certificate decays geometrically. For stochastic gradients the sensitivity is random, and (Mu & Klabjan, 2025) convert a moment bound by Markov's inequality, paying an extra δ^{-1} or $\delta^{-1/2}$ in σ. Our full-batch statements avoid this cost, which is why Prop. N.1 is stated pathwise.

Known stability conclusions. For general smooth nonconvex losses, the rewind sensitivity is $\frac{2L}{\beta n}\big[(1+\eta\beta)^T - (1+\eta\beta)^{T_u}\big]$. This sensitivity is exponential in $T - T_u$ and vanishes only at $T_u = T$, when rewinding becomes retraining. The source reports that additional rewinding reduces the noise requirement only slowly, in contrast with the convex case. The stochastic analysis of (Mu & Klabjan, 2025) separates the update rule into expansive, semi-contracting, and contracting regimes; Lem. L.2(ii) records the corresponding expressions. Langevin unlearning has a certificate uniform in T even without convexity, but the constant in its nonconvex log-Sobolev recursion is exponential in $D_\Theta^2/(\eta\sigma^2)$, which the authors describe as impractical. Under strong convexity, the composite certificate is $\epsilon \le \exp(-\frac{\gamma\eta T_u}{\alpha})\frac{4\alpha L^2}{\gamma\sigma^2 n^2}$. The required rewind depth then converges to a constant as $T \to \infty$, so the computational saving $T - T_u$ is unbounded. Table L.1 therefore separates two cases: a certificate uniform in T requires either a verified contraction mechanism or a separate endpoint-decay argument. Theorem N.3 uses a verified contraction of the transition rule under dissipativity, whereas Theorem 4 uses the PL objective-gap condition to control the endpoint distance.

L Certified Unlearning Mechanisms and Certificate Relations

The bounds of the main text depend on the CU algorithm through two quantities. One is its unlearning stability Δ (Def. 10), computed in Appendix N. The other is its *budget relation*, the map $(T_u, \sigma) \mapsto \epsilon$ that says which certificate the mechanism can issue. This appendix records the profiles used in the main text and the two structural hypotheses used in Thms. N.3 and N.5.

Definition 26 (κ-Wasserstein-Contractive Transition Rule.)**.** *Let $P^S_{\eta,\sigma}$ be the one-step Markov transition rule of equation 20 on dataset S. We say it is κ-Wasserstein contractive if $\mathcal{W}_1(\mu P^S_{\eta,\sigma}, \nu P^S_{\eta,\sigma}) \leq (1-\kappa\eta)\mathcal{W}_1(\mu,\nu)$ for all probability measures μ, ν on Θ and all S.*

Def. 26 holds with $\kappa = \gamma$ for γ-strongly convex losses and $\eta \leq \frac{2}{\beta+\gamma}$, by Lem. N.1 and same-randomness comparison. PL gives linear objective-gap decay and, with minimizer-set stability, endpoint bounds; it does not imply Def. 26 because flat directions can preserve parameter distances. For β-smooth losses satisfying Def. 19, reflection construction can give a strictly positive rate $\kappa = \kappa(m, K, \beta, \sigma)$ when the nondegenerate-noise discretized transition rule satisfies Def. 26; this transition rule verification is an additional assumption.

Definition 27 (λ-KL-Contractive Transition Rule.)**.** *A Markov transition rule Q on Θ is λ-KL contractive if $KL(\mu Q \| \nu Q) \leq \lambda KL(\mu\|\nu)$ for all μ, ν with $KL(\mu\|\nu) < \infty$.*

Data processing always gives $\lambda = 1$. A strict contraction $\lambda < 1$ is available for the Langevin transition rule when the stationary measure $\pi \propto e^{-R_S/\sigma^2}$ satisfies a log-Sobolev inequality with constant m_{LSI}, in which case $1-\lambda \geq \eta m_{\mathrm{LSI}}$ up to discretization error. Defs. 21 and 19 may support such an inequality under additional domain and measure conditions, but the value of m_{LSI} must be verified. This is the hypothesis under which (Chien et al., 2024b) obtain iteration-independent unlearning certificates and under which (Dadi & Cevher, 2025) obtain their KLSL bound.

Lemma L.1 (Gaussian Mechanism Profile (Dwork & Roth, 2014))**.** *Let $\mathcal{U}_0$ be a deterministic anchored unlearner with sensitivity of the worst-case distance to retraining*

$$s := \sup_{\mathcal{D}_N, i} \big\| \mathcal{U}_0(\mathcal{A}(\mathcal{D}_N), \{z_i\}) - \mathcal{A}(\mathcal{D}_N^{-i}) \big\|,$$

and let $\mathcal{U}$ publish $\mathcal{U}_0 + \mathcal{N}(0, \sigma^2 I_d)$ while $\mathcal{A}$ publishes its own output perturbed by the same law. Then for every $\delta \in (0,1)$ and every $\epsilon \leq 1$ with $\sigma \geq s\sqrt{2\ln(1.25/\delta)}/\epsilon$, the algorithm $\mathcal{U}$ is (ϵ,δ)-CU for $\mathcal{A}$ in the sense of Def. 4. Equivalently, if the deterministic sensitivity in the display is denoted by s_{T_u}, the budget relation is $\epsilon(T_u, \sigma) = s_{T_u}\sqrt{2\ln(1.25/\delta)}/\sigma$.

Proof. This is the Gaussian mechanism analysis of (Dwork & Roth, 2014) applied to the pair of output laws appearing in Def. 4. □

Lemma L.2 (Sensitivity of the Worst-Case Distance to Retraining of Anchored Unlearners (Neel et al., 2021; Mu & Klabjan, 2026; 2025; Chien et al., 2024a))**.** *Consider a full-batch anchored unlearner as in Def. 14 with constant step η and one-step expansion coefficient $\mathcal{E}$ as in equation 21 with $\varsigma = 0$.*

- *(i)* ***Descend to delete**, $\theta^{\mathrm{a}} = \theta_T$, with retraining defined as running the same recursion on $\mathcal{D}_N^{-i}$ to its fixed point. If $\mathcal{E} = 1-\kappa\eta < 1$, then $s_{T_u} \leq (1-\kappa\eta)^{T_u}\|\theta_T - \mathcal{A}(\mathcal{D}_N^{-i})\| \leq (1-\kappa\eta)^{T_u} D_\Theta$, and more sharply $s_{T_u} \leq (1-\kappa\eta)^{T_u}\frac{2L}{\kappa N}$ when both runs start from a common initialization.*
- *(ii)* ***Rewind to delete**, $\theta^{\mathrm{a}} = \theta_{T-T_u}$, with retraining defined as T steps on $\mathcal{D}_N^{-i}$ from the shared initialization. Then*

$$s_{T_u} \leq \frac{2L}{\beta N}\big[(1+\eta\beta)^T - (1+\eta\beta)^{T_u}\big]$$

under Def. 21, $s_{T_u} \leq \frac{2\eta L(T-T_u)}{N}$ under convexity with $\eta \leq \frac{2}{\beta}$, and $s_{T_u} \leq \frac{2L}{\kappa N}(1-\kappa\eta)^{T_u}$ under Def. 26. In each case $s_{T_u} \leq D_\Theta$ may be used instead when Θ is bounded, so the operative bound is the minimum of the two.

Proof. (i) Both the unlearned run and the retrained run apply the same map $G_{\mathcal{D}_N^{-i}}$, if that map is $\mathcal{E}$-Lipschitz, the distance between them shrinks by the factor $\mathcal{E}$ per step, and the initial distance is at most D_Θ. The sharper form follows from $\|\theta_T - \mathcal{A}(\mathcal{D}_N^{-i})\| \le \frac{2\eta L}{N}\sum_{s\ge 0}\mathcal{E}^s = \frac{2L}{\kappa N}$, which is the recursion of Prop. N.1 with the drift gap $\frac{2L}{N}$ between $R_{\mathcal{D}_N}$ and $R_{\mathcal{D}_N^{-i}}$.

(ii) Let θ_t and θ'_t denote the full-batch runs on $\mathcal{D}_N$ and on $\mathcal{D}_N^{-i}$ from the shared initialization. The unlearned model is $G^{T_u}_{\mathcal{D}_N^{-i}}(\theta_{T-T_u})$ and the retrained model is $G^{T_u}_{\mathcal{D}_N^{-i}}(\theta'_{T-T_u})$, so $s_{T_u} \le \mathcal{E}^{T_u}\|\theta_{T-T_u} - \theta'_{T-T_u}\| \le \mathcal{E}^{T_u}\frac{2\eta L}{N}\sum_{t<T-T_u}\mathcal{E}^t$. Summing the geometric series with $\mathcal{E} = 1+\eta\beta$ gives the first display, and with $\mathcal{E} = 1-\kappa\eta$ gives the second after bounding the sum by $(\kappa\eta)^{-1}$. □

The three regimes of part (ii) are the expansive, semi-contracting and contracting trichotomy established for stochastic rewind-to-delete in Theorem 4.5 of (Mu & Klabjan, 2025). Our smooth nonconvex and convex expressions coincide with the corresponding expressions in that theorem. The full-batch analysis of (Mu & Klabjan, 2026) carries an extra inflation factor inside the base because that analysis compares datasets of different sizes, whereas Def. 10 compares datasets related by a replacement. The contracting expression coincides with the quantity $Z_{\mathcal{B}}$ of (Chien et al., 2024a) at full batch. Its value at $T_u = 0$, namely $\frac{2L}{\kappa N}$, coincides with the empirical risk minimizer sensitivity $\frac{2L}{\gamma n}$ of Lemma 8 of (Neel et al., 2021).

Two noise-setting details matter. First, (Neel et al., 2021) do not use the $\sqrt{2\ln(1.25/\delta)}$ form of Lem. L.1; they use the tighter concentrated-privacy inversion $\epsilon = \frac{s^2}{2\sigma^2} + \frac{s}{\sigma}\sqrt{2\ln(1/\delta)}$, which is valid for $\epsilon = O(\ln(1/\delta))$ and only improves the constants in our bounds. Second, when the sensitivity is random, as for stochastic gradients, (Mu & Klabjan, 2025) convert a moment bound by Markov's inequality and pay an extra factor δ^{-1} or $\delta^{-1/2}$ in σ. Our full-batch statements avoid this cost, which is one reason Prop. N.1 is stated pathwise for full-batch unlearners.

Part (ii) also makes the rewind trade-off explicit: the sensitivity vanishes as $T_u \uparrow T$, where rewinding becomes retraining, and it saturates at $\frac{2L}{\beta N}[(1+\eta\beta)^T - 1]$ as $T_u \downarrow 0$. Under Def. 21 alone the relation is therefore exponential in T, and combining it with the exponential growth of Δ in Prop. N.1 shows that no choice of T_u removes the factor $(1+\eta\beta)^T$: the sum of the two terms in Conversion Theorem 1(2) is monotone in T_u, so the optimum sits at an endpoint and both endpoints carry $(1+\eta\beta)^T$. Each additional rewind step buys only a factor of roughly $\frac{1+\eta\beta}{1+\eta\beta N/(N-1)}$, which is negligible for a single deletion, and (Mu & Klabjan, 2026) report the same phenomenon empirically. Thm. N.3 therefore requires a verified contractive transition rule, whereas Thm. 4 uses the PL endpoint condition; the CU route inherits this limitation rather than removing it.

Remark 4 (What assumptions imply a bound uniform in T.)**.** *The unlearning stability Δ (Def. 10) is uniform in T whenever Prop. N.1 is non-expansive or contracting, since $\Delta_0 = 0$ and the recursion runs for T_u steps only. The CU-implied distance is uniform in T only when unlearning dynamics contract, because sensitivity of the worst-case distance to retraining is the distance between the anchor and the retrained model. Under Def. 19 with $\sigma > 0$, geometric sensitivity decay requires a verified contractive discretized transition rule; under strong convexity it follows from the stated step-size condition, while PL supplies geometric objective-gap decay only with the endpoint and minimizer-set conditions used in Thm. 4. Under Def. 21 alone the relation grows as $(1+\eta\beta)^T$ and no choice of T_u removes that factor.*

CU algorithm	Anchor	Certificate	Profile $\epsilon(T_u, \sigma)$	Unif. T
D2D (Neel et al., 2021)	θ_T	(ϵ, δ)	$\frac{2L}{\kappa\sigma n}(1-\kappa\eta)^{T_u}\sqrt{\ln\frac{1}{\delta}}$	yes
PNSGD (Chien et al., 2024a)	θ_T	Rényi DP	$\frac{2\alpha L^2(1-\eta\gamma)^{2T_u}}{\eta\gamma^2\sigma^2 n^2}$	yes
PNSGD, convex	θ_T	Rényi DP	$\frac{\alpha s_0^2}{2\eta\sigma^2 T_u}$, $s_0 = \frac{2\eta LT}{n} \wedge D_\Theta$	saturates
Langevin (Chien et al., 2024b)	θ_T	Rényi DP	$e^{-m\eta T_u/\alpha}\frac{4\alpha L^2}{m\sigma^2 n^2}$	yes
R2D (Mu & Klabjan, 2026)	θ_{T-T_u}	(ϵ, δ)	$\frac{2L[(1+\eta\beta)^T - (1+\eta\beta)^{T_u}]}{\beta\sigma n}\sqrt{2\ln\frac{1.25}{\delta}}$	no
TV-stable (Ullah et al., 2021)	θ_T	TV, exact	$\rho\sqrt{\ln(1/\delta)}$, $\sigma = \frac{8\sqrt{T}L}{n\rho}$	no

Table L.1: Certificate relations for certified unlearners with shared references. Here L is the Lipschitz constant, β the smoothness, γ or m the strong convexity, κ the contraction rate in Def. 26, α the Rényi order, and ρ the TV-stability parameter. The last column records whether the certificate is uniform in the number of training updates T. Only the certificates with a contraction mechanism are uniform in T, which is precisely the dichotomy inherited by Table 1.

Lemma L.3 (Privacy Amplification by Iteration (Feldman et al., 2018; Altschuler & Talwar, 2022))**.** *Let $(\Phi_k)_{k<T_u}$ be maps on Θ that are c-Lipschitz with $c \le 1$, let $\varpi^2 = 2\eta\sigma^2$, and define $\mu_{k+1} = (\Phi_k)_\sharp \mu_k * \mathcal{N}(0, \varpi^2 I)$ and $\nu_{k+1} = (\Phi_k)_\sharp \nu_k * \mathcal{N}(0, \varpi^2 I)$. If $\mathcal{W}_\infty(\mu_0, \nu_0) \le s_0$, then for every $\alpha > 1$,*

$$\mathcal{R}_\alpha(\mu_{T_u} \| \nu_{T_u}) \le \frac{\alpha s_0^2}{2\varpi^2} \times \begin{cases} c^{2T_u}, & c < 1, \\ 1/T_u, & c = 1. \end{cases} \tag{31}$$

Consequently, optimizing the conversion $\epsilon = \mathcal{R}_\alpha + \frac{\ln(1/\delta)}{\alpha - 1}$ over $\alpha > 1$, the budget relation is

$$\epsilon(T_u, \sigma) = O\Big(\frac{s_0}{\sigma}\sqrt{\frac{\ln(1/\delta)}{\eta}}\Big) \times \begin{cases} c^{T_u}, & c < 1, \\ T_u^{-1/2}, & c = 1. \end{cases}$$

Proof. The proof of equation 31 uses the shifted Rényi divergence argument of (Feldman et al., 2018). The contractive-map form is given by (Altschuler & Talwar, 2022) and appears as Lemma 2.6 of (Chien et al., 2024a). When $c = 1$, the initial discrepancy s_0 is removed in equal increments s_0/T_u over the T_u steps, and each increment costs $\frac{\alpha(s_0/T_u)^2}{2\varpi^2}$ in shifted divergence. When $c < 1$, the maps contract the discrepancy and the final increment carries the remaining cost. □

The $c < 1$ branch of equation 31 describes contractive updates, whereas the $c = 1$ branch describes non-expansive updates. Contractive updates make the CU privacy parameter decay geometrically with T_u, so logarithmic unlearning depth suffices. Non-expansive updates make the privacy parameter decay as $T_u^{-1/2}$, producing the interior optimum in Theorem N.2. Substituting $c = 1 - \eta m$ and the source-specific initial-gap expression $s_0 = \frac{2\eta M}{b(1-c^{n/b})} \wedge 2R$ into equation 31 recovers Corollary 3.7 of (Chien et al., 2024a) verbatim. Here M is the source's smoothness or gradient-sensitivity constant and R is its parameter-radius bound; the rest of this paper uses the source-independent bound $s_0 \le D_\Theta$. Setting $m = 0$ recovers their Corollary 3.10. In our notation, the strongly convex instance reads $\epsilon \le \frac{2\alpha L^2}{\eta\gamma^2\sigma^2 n^2}(1 - \eta\gamma)^{2T_u}$.

Lemma L.4 (Uniform Gradient Concentration (Dadi & Cevher, 2025))**.** *Let $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$ be the two halves of $\mathcal{D}_{2n}$ as in Def. 12 and assume Def. 24 with constant σ_{g}. Then, for each fixed θ, $\mathbb{E}\|\nabla R_{\mathcal{D}^{(1)}}(\theta) - \nabla R_{\mathcal{D}^{(2)}}(\theta)\|^2 \le \frac{2\sigma_{\mathrm{g}}^2}{n}$. If in addition $\nabla\mathcal{L}(\cdot, z)$ is β-Lipschitz and $\mathrm{diam}(\Theta) \le D_\Theta$, a covering argument upgrades this to*

$$\bar{s}_n^2 = \mathbb{E}\Big[\sup_{\theta\in\Theta} \|\nabla R_{\mathcal{D}^{(1)}}(\theta) - \nabla R_{\mathcal{D}^{(2)}}(\theta)\|^2\Big] \le \frac{2c_\Theta\sigma_{\mathrm{g}}^2}{n},$$

with $c_\Theta = O(d\ln(\beta D_\Theta n))$.

Proof. The two halves are independent samples of size n, so the pointwise bound is the variance of a difference of two independent empirical means. The uniform version follows from a covering of Θ at scale $1/(\beta n)$, whose cardinality is $(\beta D_\Theta n)^d$, combined with a union bound and the β-Lipschitz continuity of the gradient field. □

M Imported Stability Techniques and Their Adaptation

Our six main theorems each borrow one technique from the stability literature. The citepd papers compare two learner runs on neighboring training datasets, whereas our proofs compare two unlearning runs that start from one shared checkpoint and delete different records. The citepd inequalities therefore cannot be copied without modification. This appendix states the modification for each source and identifies the single inequality used in each theorem.

Shared-randomness comparisons and transition-rule perturbations (Zhu et al., 2023). The source's overarching framework is Markov-chain perturbation theory (Rudolf & Schweizer, 2018). It combines geometric ergodicity, a Lyapunov drift, and a one-step transition-rule distance. The source divides the one-step distance between its two transition rules by a Lyapunov function; this normalized distance between transition rules supplies the factor $1/n$. The source uses the same random choices to establish contraction in the strongly convex case. Under dissipativity, the source cannot use that contraction argument, so the Lyapunov drift controls the comparison. Let θ_t and θ'_t denote the parameter iterates of the two runs on $\mathcal{D}_n$ and $\mathcal{D}_n^{(i)}$, using the same random choices. In our proof, we import the shared-index and shared-noise construction and use it to upper-bound the Wasserstein distance between the two learner output laws through

$$\mathcal{W}_p\big(\mathcal{A}(\mathcal{D}_n), \mathcal{A}(\mathcal{D}_n^{(i)})\big) \le \big\|\theta_T - \theta'_T\big\|_{L^p}. \tag{32}$$

The coupling also supplies the one-step estimate that we reuse as equation 21. Three changes are needed. First, the source compares runs that read different datasets from the first update, so its uniform-in-time result needs an initial-condition term and geometric ergodicity. In our construction, the two deletion runs start from the same checkpoint, so $\Delta_0 = 0$ and uniformity in T_u follows directly; ergodicity is needed only for the sharper $\mathcal{W}_1$ statement in equation 47. Second, the source controls the sample-difference term through the pseudo-Lipschitz condition in Def. 17, which introduces $\|z - z'\|(\|u\| + \|v\| + 1)$ and requires uniform moment bounds on the iterates. When two unlearning runs are compared, the sample-difference term is bounded by either the crude bound $2L$ or the self-bounding surrogate, so no iterate-moment assumption is needed. Third, the factor $1/n$ in our comparison comes from the drift gap between $R_{\mathcal{D}_N^{-i}}$ and $R_{\mathcal{D}_N^{-j}}$, namely $\frac{1}{n}(\nabla\mathcal{L}(\cdot, z_j) - \nabla\mathcal{L}(\cdot, z_i))$, rather than from the probability of drawing one differing index. Therefore equation 38 is an L^∞ statement for full-batch unlearners and holds simultaneously for all p. For the dissipative row, the source controls $\mathcal{W}_2^2 = O(1/n + K/m)$ rather than $\mathcal{W}_2$, so its induced generalization rate is $O(n^{-1/2})$ whenever K/m does not vanish; our equation 46 has the same squared-distance form.

On-average model stability (Def. 7) and self-bounding (Lei & Ying, 2020). Two inequalities are imported. The first is the self-bounding property $\|\nabla\mathcal{L}(\theta, z)\| \le \sqrt{2\beta\mathcal{L}(\theta, z)}$ of Lem. J.1 (their Lemma A.1 at $\alpha = 1$). The second is their Theorem 2(b): for nonnegative β-smooth losses and any $\gamma > 0$. For a neighboring pair, let u and v denote the two outputs under the chosen coupling.

$$\mathbb{E}\big[R - R_{\mathcal{D}_n}\big] \le \frac{\beta}{\gamma}\mathbb{E}\big[R_{\mathcal{D}_n}\big] + \frac{\beta + \gamma}{2n}\sum_i \mathbb{E}\|u - v\|^2, \tag{33}$$

which becomes Lem. J.4 after the reparametrization $\gamma = \beta/c_0$. Inequality equation 33 assumes the ℓ_2 on-average model-stability coupling in Def. 7. Conversion Theorem 1(2) instead provides a $\mathcal{W}_2$ parameter, which is the infimum over all such pairings. Lemma J.2 allows the optimal $\mathcal{W}_2$ pairing because the generalization gap depends only on the two output laws. We apply the self-bounding inequality along the unlearning trajectory and evaluate the loss at the trained anchor.

Sampling-determined algorithms and Moreau envelopes (Lei, 2023). Two facts are imported, and they must be kept separate. First, the source analyzes *sampling-determined* algorithms (plain SGD and AdaGrad-Norm), for which the output depends only on sampled indices. Under source-specific boundedness, parameter-radius, and gradient-size constants, the source obtains uniform stability of function values $O(BT/n)$, uniform argument stability $O(RT/n)$, and uniform gradient stability $O(G\sqrt{T/n})$. The corresponding function-value and model-stability notions used here are in Defs. 6 and 7; the latter controls a squared discrepancy of order $O(T/n)$. These source conclusions do not use the PL condition. Second, for weakly convex nonsmooth objectives the source uses the Moreau envelope as a *performance measure*, and notes explicitly that neither the Moreau envelope nor the proximal map of the empirical objective R_S appears in the implementation of SGD. What we adapt for Combination G is therefore the elementary proximal Lipschitz estimate, rather than the source algorithm itself. If $\mathcal{L}(\cdot, z)$ is μ-weakly convex and $\eta < 1/\mu$, then

$$\big\|\mathrm{prox}_{\eta\mathcal{L}(\cdot,z)}(u) - \mathrm{prox}_{\eta\mathcal{L}(\cdot,z)}(v)\big\| \le \frac{1}{1 - \eta\mu}\|u - v\|, \tag{34}$$

with the constant equal to one when $\mathcal{L}(\cdot, z)$ is convex. Replacing the gradient step in equation 20 by a proximal step makes equation 21 hold with $\varsigma = 0$ even without Def. 21, so Prop. N.1 applies

and the additive term responsible for the $\sqrt{T_u}$ in equation 57 disappears. A secondary benefit of the sampling-determined viewpoint is that horizon restrictions of the form $T \leq n$, which appear in classical sampling-determined analyses (Lei, 2023), become $T_u \leq n$ for us, a condition on the unlearner rather than on the training run.

Data-dependent stability (Kuzborskij & Lampert, 2018). The imported step replaces the worst-case Lipschitz constant in the recursion for the distance between the runs by the expected gradient norm along the trajectory. For two coupled unlearning runs on $\mathcal{D}_N^{-i}$ and $\mathcal{D}_N^{-j}$, let $\theta_t^{(i)}$ and $\theta_t^{(j)}$ denote their step-t iterates, and let $z \sim \mathcal{P}$ be an independent sample used in the gradient-norm estimate. Write $\delta_t := \|\theta_t^{(i)} - \theta_t^{(j)}\|$.

$$\mathbb{E}\big[\delta_{t+1}\big] \leq (1 + \eta_t\beta)\mathbb{E}\big[\delta_t\big] + \frac{2\eta_t}{n}\mathbb{E}\big\|\nabla\mathcal{L}(\theta_t^{(i)}, z)\big\|, \tag{35}$$

after which the source controls $\mathbb{E}\|\nabla\mathcal{L}(\theta_t^{(i)}, z)\|$ using the risk at the initialization together with Def. 24, producing bounds that carry $\sqrt{R(\theta_0)}$. The adaptation is conceptually the cleanest of the six. In our framework the role of the initialization is played by the anchor, and the anchor is by construction a *trained* model, so the data-dependent quantity is the training loss rather than the loss at a random starting point. This gives Cor. M.1, which is the data-dependent form of Thm. N.7. The source pays $\sqrt{R(\theta_0)}$ because it must start where training starts. We pay $\sqrt{\Lambda_0}$ because we start where training ends.

Comparing subgradient runs for nonsmooth convex losses (Bassily et al., 2020). The imported inequality is Lem. N.1(v). For a convex L-Lipschitz loss, let $g_u \in \partial f(u)$ and $g_v \in \partial f(v)$ be the selected subgradients. Monotonicity of the subdifferential removes the cross term but leaves the quadratic term:

$$\big\|u - v - \eta(g_u - g_v)\big\|^2 \leq \|u - v\|^2 + 4\eta^2L^2, \tag{36}$$

so that even steps taken on the *same* function accumulate divergence, at a rate that integrates to $\eta L\sqrt{T}$ on top of the $\eta LT/n$ contributed by the differing steps. The adaptation is the substitution $T \mapsto T_u$ in equation 57. What does not adapt is the obstruction. The source proves matching lower bounds in high dimension $d \geq \min\{T, 1/\eta^2\}$. For sampling SGD the expectation lower bound carries $\min\{1, T/n\}$ on the $\sqrt{T}$ term. Since an unlearner performing T_u nonsmooth steps is itself subject to that lower bound, the CU route inherits the barrier and can only shrink its argument from T to T_u.

Path-dependent risk control without Lipschitzness (Nikolakakis et al., 2022). The imported step is that the optimization error controls the whole gradient path. For a retained dataset S, let $\theta_{t+1} = \theta_t - \eta\nabla R_S(\theta_t)$ be its full-batch gradient-descent iterates. For $\eta \leq 1/\beta$ the descent lemma telescopes into

$$\eta\sum_{t<T}\big\|\nabla R_S(\theta_t)\big\|^2 \leq 2\big(R_S(\theta_0) - \inf R_S\big), \tag{37}$$

so that by Cauchy-Schwarz the total movement satisfies $\|\theta_T - \theta_0\| \leq \sqrt{2T\eta(R_S(\theta_0) - \inf R_S)}$, and combining this with self-bounding removes every appearance of L. This is what we use to prove equation 54, and hence Thm. N.7. The difference in scope is worth recording. The source obtains $O(1/n)$ generalization for full-batch gradient descent run for $T \asymp n$ iterations, so its conclusion is a statement about a particular learner. By contrast equation 55 constrains any learner admitting a certified unlearner, needs only T_u unlearning iterations, and imposes no condition on T.

Corollary M.1 (Data-Dependent Form of Theorem N.7)**.** *In the setting of Thm. N.7 with $\mathcal{E} = 1$, let $\Lambda_0 = \max_{z\in\mathcal{D}_N}\mathcal{L}(\theta^{\mathrm{a}}, z)$ be the largest per-sample loss at the anchor and Ψ_0 the optimization error there. Then $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-USL with*

$$\tau \leq \frac{2\eta T_u}{n}\Big(2\sqrt{\beta\Lambda_0} + \beta\sqrt{2\eta T_u\Psi_0}\Big) + 2r_{T_u,p},$$

and in particular $\tau \leq 2r_{T_u,p}$ in the interpolation limit $\Lambda_0, \Psi_0 \to 0$.

Proof. By equation 53 with $\mathcal{E} = 1$ and equation 54, $\Delta \leq \frac{2\eta T_u}{n}\sqrt{2\beta(2\Lambda_0 + 2\beta\eta T_u\Psi_0)}$. Subadditivity of the square root gives $\sqrt{2\beta(2\Lambda_0 + 2\beta\eta T_u\Psi_0)} \leq 2\sqrt{\beta\Lambda_0} + \beta\sqrt{2\eta T_u\Psi_0}$, and Conversion Theorem 1(2) supplies the endpoint-distance term. □

N Proofs of Stability Bounds via Stable Certified Unlearning (Defs. 7, 10)

Lemma N.1 (Expansion Coefficients (Hardt et al., 2016; Zhu et al., 2023; Bassily et al., 2020)). *Let $f : \Theta \mapsto \mathbf{R}$ and consider the map $G(\theta) = \theta - \eta\nabla f(\theta)$. Inequality equation 21 holds with the following pairs $(\mathcal{E}, \varsigma)$.*

(i) If f is γ-strongly convex and β-smooth and $\eta \le \frac{2}{\beta+\gamma}$, then $(\mathcal{E}, \varsigma) = (1 - \eta\gamma, 0)$.

(ii) If f is convex and β-smooth and $\eta \le \frac{2}{\beta}$, then $(\mathcal{E}, \varsigma) = (1, 0)$.

(iii) If f is β-smooth, then $(\mathcal{E}, \varsigma) = (1 + \eta\beta, 0)$.

(iv) If f is β-smooth and satisfies Def. 19 with constants (m, K), then $\|G(u) - G(v)\|^2 \le (1 - 2\eta m + \eta^2\beta^2)\|u - v\|^2 + 2\eta K$, hence equation 21 with $(\sqrt{1 - 2\eta m + \eta^2\beta^2}, \sqrt{2\eta K})$.

(v) If f is convex and L-Lipschitz but possibly nonsmooth, and G uses any subgradient selection, then $\|G(u) - G(v)\|^2 \le \|u - v\|^2 + 4\eta^2 L^2$.

The projection Π_Θ onto a convex Θ is nonexpansive, so all statements survive projection.

Proof. (i) Since f is γ-strongly convex and β-smooth, $h := \nabla f - \gamma I$ is the gradient of a convex $(\beta - \gamma)$-smooth function and is therefore $\frac{1}{\beta-\gamma}$-cocoercive. Writing $I - \eta\nabla f = (1 - \eta\gamma)I - \eta h$ and expanding,

$$\|G(u) - G(v)\|^2 \le (1 - \eta\gamma)^2\|u - v\|^2 + \Big(\eta^2 - \frac{2\eta(1 - \eta\gamma)}{\beta - \gamma}\Big)\|h(u) - h(v)\|^2,$$

and the second coefficient is nonpositive precisely when $\eta(\beta + \gamma) \le 2$.

(ii) Set $\gamma = 0$ in (i) and use the cocoercivity of ∇f directly.

(iii) Immediate from $\|\nabla f(u) - \nabla f(v)\| \le \beta\|u - v\|$ and the triangle inequality.

(iv) Expand $\|u - v - \eta(\nabla f(u) - \nabla f(v))\|^2$, apply Def. 19 to the inner product term and β-smoothness to the squared gradient term. The conversion to equation 21 uses $\sqrt{a^2x^2 + b^2} \le ax + b$.

(v) Expand and use monotonicity of the subdifferential of a convex function together with $\|g_u - g_v\| \le 2L$. □

Proposition N.1 (Two Anchored Deletion Runs with Shared Randomness, restates Proposition N.1). *Let $\mathcal{U}$ be anchored with depth T_u and constant step η. Suppose there are $\mathcal{E} \ge 0$ and $\varsigma \ge 0$ such that for all $u, v \in \Theta$ and $z \in \mathcal{Z}$,*

$$\big\|u - v - \eta\big(\nabla\mathcal{L}(u, z) - \nabla\mathcal{L}(v, z)\big)\big\| \le \mathcal{E}\|u - v\| + \varsigma.$$

If $\mathcal{L}(\cdot, z)$ is L-Lipschitz, then $\mathcal{U}$ is (Δ, N)-$\mathcal{W}_p$-USU (Def. 10) for every $p \ge 1$ with

$$\Delta \le \big(\varsigma + 2\eta L\, c_p(n)\big) \sum_{k=0}^{T_u - 1} \mathcal{E}^k. \tag{38}$$

where $c_p(n) = 1/n$ for full-batch gradients and $c_p(n) = n^{-1/p}$ for the i.i.d. sampler $\mathcal{S}_1$. The displayed unlearning-stability bound also holds for l_p-MSU (Def. 11).

Proof. Fix $\mathcal{D}_N$ and $i \ne j$. The proof couples the two unlearning runs by using the same anchor, Gaussian noises, and sampler indices; this coupling supplies an admissible joint law for the Wasserstein distance. Both runs use the identical anchor θ^{a}, which is legitimate because Def. 14 requires the anchor to be independent of the forget set, and they use the identical Gaussian sequence (ξ_k). For stochastic gradients, identify $\mathcal{D}_N^{-i}$ with $\mathcal{D}_N^{-j}$ through the bijection that is the identity on the $n - 1$ common points and maps z_j to z_i, and let both runs read the same slot. By Lemma I.1, this reindexing does not change the laws of $u^{(i)}$ and $u^{(j)}$ for samplers $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$.

Write $\Delta_k := \|\theta_k^{(i)} - \theta_k^{(j)}\|$, so $\Delta_0 = 0$. Since Π_Θ is nonexpansive and the noise terms are identical and hence cancel,

$$\Delta_{k+1} \le \big\|\theta_k^{(i)} - \theta_k^{(j)} - \eta_k\big(g_k^{(i)} - g_k^{(j)}\big)\big\|.$$

Adding and subtracting the gradient of the i-run objective evaluated at $\theta_k^{(j)}$,

$$\Delta_{k+1} \le \big\|\theta_k^{(i)} - \theta_k^{(j)} - \eta_k\big(\nabla\mathcal{L}(\theta_k^{(i)}, w) - \nabla\mathcal{L}(\theta_k^{(j)}, w)\big)\big\| + \eta_k J_k,$$

where w, w' are the samples read by the two runs at step k and $J_k := \|\nabla\mathcal{L}(\theta_k^{(j)}, w) - \nabla\mathcal{L}(\theta_k^{(j)}, w')\|$. The first term is at most $\mathcal{E}\Delta_k + \varsigma$ by equation 21. For full-batch gradients the same decomposition applies with $R_{\mathcal{D}_N^{-i}}$ and $R_{\mathcal{D}_N^{-j}}$ in place of the per-sample losses, and since these two empirical risks differ by exactly $\frac{1}{n}(\mathcal{L}(\cdot, z_j) - \mathcal{L}(\cdot, z_i))$, we obtain $J_k \le \frac{2L}{n}$ pathwise. For the *i.i.d.* sampler $\mathcal{S}_1$, $w = w'$ unless the differing slot is drawn, an event of probability $1/n$, so $J_k \le 2LB_k$ with $B_k \sim \mathrm{Bern}(1/n)$ independent of Δ_k.

Unrolling $\Delta_{k+1} \le \mathcal{E}\Delta_k + \varsigma + \eta J_k$ from $\Delta_0 = 0$ gives $\Delta_{T_u} \le \sum_{k<T_u} \mathcal{E}^{T_u-1-k}(\varsigma + \eta J_k)$. In the full-batch case this is a pathwise bound, so Proposition N.1 gives the stated inequality with $c_p(n) = 1/n$. In the $\mathcal{S}_1$ case, Minkowski's inequality in L^p gives

$$\big\|\Delta_{T_u}\big\|_{L^p} \le \sum_{k<T_u} \mathcal{E}^{T_u-1-k}\big(\varsigma + 2\eta L\|B_k\|_{L^p}\big), \qquad \|B_k\|_{L^p} = n^{-1/p}.$$

Since the construction is a valid pairing of the laws of $u^{(i)}$ and $u^{(j)}$, we have $\mathcal{W}_p^p(u^{(i)}, u^{(j)}) \le \mathbb{E}\Delta_{T_u}^p$, and taking the supremum over $\mathcal{D}_N, i, j$ yields Def. 10. The moment $\mathbb{E}\Delta_{T_u}^p$ also bounds the left-hand side of equation 18, giving the MSU claim. □

Theorem N.1 (Direct contractive stability bound)**.** *Let $\mathcal{L}(\cdot, z)$ be γ-strongly convex, L-Lipschitz (Def. 16) and β-smooth (Def. 21), let $\mathrm{diam}(\Theta) \le D_\Theta$, and let $\mathcal{U}$ be a full-batch anchored unlearner with depth T_u and step $\eta \le \frac{2}{\beta+\gamma}$ that is (ϵ, δ)-CU for $\mathcal{A}$. Assume a shared checkpoint and shared full-batch randomness with the retraining reference, and use the same data-independent publication mechanism on both branches, with that transition rule nonexpansive in $\mathcal{W}_p$. Then the retained-data recursion is γ-contractive in the sense of Def. 26 (Hardt et al., 2016, Lem. 3.4), $\mathcal{U}$ is (Δ, N)-$\mathcal{W}_p$-USU for every $p \ge 1$ with*

$$\Delta \le \frac{2L}{\gamma n}\big(1 - (1-\eta\gamma)^{T_u}\big) \wedge \frac{2\eta L T_u}{n}, \tag{39}$$

and $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-USL with

$$\tau \le \frac{2L}{\gamma n}\big(1 - (1-\eta\gamma)^{T_u}\big) + 2r_{T_u,p}. \tag{40}$$

Proof of Theorem N.1. Lem. N.1(i) gives $(\mathcal{E}, \varsigma) = (1-\eta\gamma, 0)$ for each per-sample loss, hence for the average $R_{\mathcal{D}_N^{-i}}$. Prop. N.1 with $c_p(n) = 1/n$ therefore gives $\Delta \le \frac{2\eta L}{n}\sum_{k<T_u}(1-\eta\gamma)^k$. Summing the geometric series gives $\frac{2L}{\gamma n}(1-(1-\eta\gamma)^{T_u})$, and bounding each of the T_u summands by 1 gives $\frac{2\eta L T_u}{n}$, taking the minimum yields equation 39. Thm. I.2(2) then adds the endpoint-distance term and gives equation 40. □

The bound equation 40 leaves the endpoint-distance term unspecified. In the contractive construction, the worst-case distance to retraining is at most $q^{T_u} b_0$, with $b_0 = O(1/n)$. Substituting this estimate into the conversion bound gives an explicit rate and shows how the endpoint term decreases with T_u.

Corollary N.1 (Combination A, Explicit Distance Bound)**.** *In the setting of Thm. N.1, suppose the certificate is issued by the Gaussian mechanism at sensitivity of the worst-case distance to retraining $s_{T_u} \le (1-\eta\gamma)^{T_u} D_\Theta$ with noise σ, so that $\epsilon = s_{T_u}\sqrt{2\ln(1.25/\delta)}/\sigma$ (Lem. L.1), and suppose both branches use the same data-independent nonexpansive publication mechanism. Then, for $p =$*

1, *the bound* $2r_{T_u,p}$ *from the worst-case distance to retraining and the CU-implied distance give, respectively*

$$\tau \le \frac{2L}{\gamma n}\big(1-(1-\eta\gamma)^{T_u}\big) + \min\left\{2D_\Theta(1-\eta\gamma)^{T_u},\ \frac{D_\Theta^2\sqrt{2\ln(1.25/\delta)}}{\sigma}(1-\eta\gamma)^{T_u} + 2D_\Theta\delta\right\}. \tag{41}$$

In particular, the direct route gives $\tau \le \frac{2L}{\gamma n}(1-c)+2D_\Theta c$ *with* $c=(1-\eta\gamma)^{T_u}$*, while the CU-implied distance gives* $\tau \le \frac{4L}{\gamma n}$ *whenever* $\delta \le \frac{L}{2D_\Theta\gamma n}$ *and*

$$T_u \ge \frac{1}{\eta\gamma}\ln\frac{D_\Theta^2\gamma n\sqrt{2\ln(1.25/\delta)}}{L\sigma} = O\Big(\frac{\ln n}{\eta\gamma}\Big). \tag{42}$$

Proof. Substitute the relation between Lems. L.1 and L.2(i) into the CU-implied distance in equation 40; use $\tanh(x/2)\le x/2$ so that $C_{\epsilon,\delta}\le\delta+\epsilon/2$, and bound $1-(1-\eta\gamma)^{T_u}\le 1$ in the distance term. The shared-transition-rule route based on the worst-case distance to retraining gives the first term inside the minimum in equation 41, while the CU-implied distance gives the second. For the final claim, $\delta \le \frac{L}{2D_\Theta\gamma n}$ makes $2D_\Theta\delta \le \frac{L}{\gamma n}$, and the middle term in the CU-implied distance is at most $\frac{L}{\gamma n}$ exactly when $(1-\eta\gamma)^{T_u} \le \frac{L\sigma}{D_\Theta^2\gamma n\sqrt{2\ln(1.25/\delta)}}$, taking logarithms and using $\ln\frac{1}{1-\eta\gamma}\ge\eta\gamma$ gives the sufficient threshold equation 42, and the two terms in the CU-implied distance then sum to at most $\frac{4L}{\gamma n}$. □

Theorem N.2 (Combination B: Convex, Lipschitz, Smooth)**.** *Let* $\mathcal{L}(\cdot,z)$ *be convex, L-Lipschitz and* β*-smooth, let* $\operatorname{diam}(\Theta)\le D_\Theta$*, and let* $\mathcal{U}$ *be a full-batch anchored unlearner with depth* T_u*, steps* $\eta_k\le\frac{2}{\beta}$*, that is* (ϵ,δ)*-CU for* $\mathcal{A}$*. Let* θ_k^{a} *denote the unlearning iterate after* k *updates from the anchor* θ^{a}*. Then* $\mathcal{U}$ *is* (Δ,N)*-*$\mathcal{W}_p$*-USU with* $\Delta\le\frac{2L}{n}\sum_{k=0}^{T_u-1}\eta_k$*, and* $\mathcal{A}$ *is* (τ,n)*-*$\mathcal{W}_p$*-USL with*

$$\tau \le \frac{2L}{n}\sum_{k=0}^{T_u-1}\eta_k + 2r_{T_u,p}. \tag{43}$$

If in addition $\mathcal{L}\ge 0$*, the constant* L *may be replaced by* $\sqrt{2\beta\Lambda}$*, where* $\Lambda := \max_{k<T_u}\max_{z\in\mathcal{D}_N}\mathcal{L}(\theta_k^{\mathrm{a}},z)$ *is the path-maximal training loss of the unlearning run.*

Proof. Apply Lem. N.1(ii) in the proof of Thm. N.1; because $\mathcal{E}=1$, the recursion contributes one additive term per unlearning step, which gives the first bound. For the nonnegative-loss refinement, replace $J_k\le\frac{2L}{n}$ in the proof of Prop. N.1 by

$$J_k \le \tfrac{1}{n}\big(\|\nabla\mathcal{L}(\theta_k^{\mathrm{a}},z_i)\| + \|\nabla\mathcal{L}(\theta_k^{\mathrm{a}},z_j)\|\big) \le \tfrac{2}{n}\sqrt{2\beta\Lambda_k}$$

using Lem. J.1, and bound Λ_k by Λ. □

The bound equation 43 is stated at a fixed depth. The certificate improves as T_u grows, while the distance between the two deletion runs worsens, so T_u is a free parameter. Optimizing the bound over T_u gives the $O(n^{-1/3})$ rate quoted in the main text; the following corollary gives the optimizer and its constants.

Corollary N.2 (Combination B at the optimal unlearning depth)**.** *Let* $\mathcal{L}(\cdot,z)$ *be convex, L-Lipschitz and* β*-smooth, let* $\operatorname{diam}(\Theta)\le D_\Theta$*, let* $\mathcal{U}$ *be a full-batch anchored unlearner with constant step* $\eta\le\frac{2}{\beta}$ *and depth* $T_u\ge 1$*, and let its certificate be issued by privacy amplification by iteration (Lem. L.3, case* $c=1$*) at an initial gap* $s_0\le D_\Theta$ *and noise* $\sigma>0$*, so that for some absolute constant* $c_0>0$

$$\epsilon(T_u,\sigma) \ \le\ \frac{c_0 s_0}{\sigma}\sqrt{\frac{\ln(1/\delta)}{\eta}}\cdot T_u^{-1/2}.$$

Put $a := \frac{2\eta L}{n}$ *and* $b := \frac{c_0 D_\Theta s_0}{\sigma}\sqrt{\frac{\ln(1/\delta)}{\eta}}$*. Then, at* $p=1$*, for every* T_u*,*

$$\tau \ \le\ a\,T_u + b\,T_u^{-1/2} + 2D_\Theta\delta, \tag{44}$$

and the upper bound in equation 44 is minimized at $T_u^\star = \big(b/(2a)\big)^{2/3}$*, where the optimized upper bound equals*

$$\tau \;\le\; 3\cdot 2^{-2/3}\,a^{1/3}b^{2/3} + 2D_\Theta\delta \;\le\; \frac{3\,c_0^{2/3}\,(2L)^{1/3}\,D_\Theta^{4/3}\,\ln^{1/3}(1/\delta)}{2^{2/3}\,\sigma^{2/3}\,n^{1/3}} + 2D_\Theta\delta, \tag{45}$$

uniformly in the number of training updates T*. In particular* $\tau = O\big(L^{1/3}D_\Theta^{4/3}\ln^{1/3}(1/\delta)\,\sigma^{-2/3}n^{-1/3}\big)$ *once* $\delta = O(n^{-1/3})$*, and the step size* η *has cancelled.*

Proof. By Thm. N.2 with constant steps, $\Delta \le \frac{2\eta L T_u}{n} = aT_u$. For the endpoint-distance term, $\tanh(x/2) \le x/2$ gives $C_{\epsilon,\delta} \le \delta + \epsilon/2$, so at $p = 1$

$$P^{\mathrm{TV}}(\epsilon,\delta) \;\le\; 2D_\Theta\delta + D_\Theta\epsilon \;\le\; 2D_\Theta\delta + \frac{c_0 D_\Theta s_0}{\sigma}\sqrt{\frac{\ln(1/\delta)}{\eta}}\,T_u^{-1/2},$$

which is $2D_\Theta\delta + bT_u^{-1/2}$. Adding the two and invoking Conversion Theorem 1(2) gives equation 44. The function $\varphi(t) = at + bt^{-1/2}$ on $(0,\infty)$ has $\varphi'(t) = a - \frac{b}{2}t^{-3/2}$, which vanishes only at $t^\star = (b/(2a))^{2/3}$ and is negative before and positive after, so $t^\star$ is the unique minimizer. Substituting,

$$\varphi(t^\star) = a\Big(\frac{b}{2a}\Big)^{2/3} + b\Big(\frac{2a}{b}\Big)^{1/3} = \big(2^{-2/3} + 2^{1/3}\big)a^{1/3}b^{2/3} = 3\cdot 2^{-2/3}a^{1/3}b^{2/3},$$

using $2^{-2/3} + 2^{1/3} = 2^{-2/3}(1+2) = 3\cdot 2^{-2/3}$. Finally

$$a^{1/3}b^{2/3} = \Big(\frac{2\eta L}{n}\Big)^{1/3}\Big(\frac{c_0 D_\Theta s_0}{\sigma}\Big)^{2/3}\frac{\ln^{1/3}(1/\delta)}{\eta^{1/3}} = \frac{(2L)^{1/3}c_0^{2/3}(D_\Theta s_0)^{2/3}\ln^{1/3}(1/\delta)}{\sigma^{2/3}n^{1/3}},$$

in which $\eta^{1/3}$ cancels, and $s_0 \le D_\Theta$ gives $(D_\Theta s_0)^{2/3} \le D_\Theta^{4/3}$. Since T_u must be an integer, using $\lceil t^\star \rceil$ inflates the first term of equation 44 by at most a, which is $O(n^{-1})$ and hence absorbed. Uniformity in T holds because neither a nor b involves T once s_0 is bounded by the diameter. □

Two consequences explain the non-contractive rows of Table 1. First, the exponent is $-1/3$ because Δ is linear in T_u by Thm. N.2, while the $T_u^{-1/2}$ branch of equation 31 is tight for nonexpansive maps (Feldman et al., 2018; Altschuler & Talwar, 2022); both sides of the trade-off are therefore attained. Second, equation 45 is uniform in T. The classical uniform-stability bound (Def. 6) of (Hardt et al., 2016) for convex smooth SGD is $\frac{2L^2\eta T}{n}$, which becomes $\Theta(LD_\Theta\sqrt{T}/n)$ at the standard $\eta = D_\Theta/(L\sqrt{T})$. Our term is $\Theta(LD_\Theta T_u/(n\sqrt{T}))$, so it is smaller by the factor T/T_u when $T_u \ll T$. The conversion bound remains uniform in T; under the stated noise choice, the two terms cross when $T \gtrsim n^{4/3}$.

Theorem N.3 (Combination C: Dissipative and Smooth)**.** *Let* $\mathcal{L}(\cdot, z)$ *satisfy Def. 19 with constants* (m, K)*, be* β*-smooth and L-Lipschitz, let* $\mathrm{diam}(\Theta) \le D_\Theta$*, and let* $\mathcal{U}$ *be a full-batch anchored unlearner with* $\eta \le \min\{\frac{1}{m}, \frac{m}{\beta^2}\}$ *that is* (ϵ,δ)*-CU for* $\mathcal{A}$*. Then, uniformly in* T_u *and* T*,* $\mathcal{U}$ *is* (Δ, N)*-*$\mathcal{W}_2$*-USU with*

$$\Delta^2 \le \frac{6K}{m} + \frac{24L^2}{m^2n^2}, \tag{46}$$

and $\mathcal{A}$ *is* (τ, n)*-*$\mathcal{W}_2$*-USL with* $\tau \le \Delta + 2r_{T_u,p}$*. If, in addition, the discretized noisy transition rule of equation 20 can be verified to be* κ*-Wasserstein contractive in* $\mathcal{W}_1$ *(Def. 26), with reflection construction supplying a candidate rate under suitable step-size and domain conditions, then the additive dissipativity term* $6K/m$ *in equation 46 is replaced by the contractive* $p = 1$ *bounds in equation 47. If the same update is pointwise* q*-contractive under shared randomness and the publication mechanism is nonexpansive in* $\mathcal{W}_p$*, the three inequalities in equation 47 hold in* $\mathcal{W}_p$ *with the same constants (and hence for every* p *under the corresponding pathwise hypothesis):*

$$\Delta \le \frac{2L}{\kappa n}\big(1-(1-\kappa\eta)^{T_u}\big), \qquad 2r_{T_u,p} \le \frac{4L}{\kappa n}(1-\kappa\eta)^{T_u}, \qquad \tau \le \frac{2L}{\kappa n}\big(1-(1-\kappa\eta)^{T_u}\big) + 2r_{T_u,p} \le \frac{4L}{\kappa n}. \tag{47}$$

Both displays require Def. 19. The contractive display also requires Def. 26. Uniformity in T *follows from these hypotheses; it does not follow from the CU route alone. If Def. 19 and the*

contractive assumption are both removed, so that the loss is only β-smooth and nonconvex, Lem. N.1(iii) gives $\mathcal{E} = 1 + \eta\beta$ and Prop. N.1 yields only $\Delta \le \frac{2L}{\beta n}\big[(1+\eta\beta)^{T_u} - 1\big]$. The worst-case distance for the rewind unlearner, and therefore the privacy budget required by its CU certificate, grows as $(1+\eta\beta)^T$ (Lem. L.2(ii) and Table L.1). In this regime neither Δ nor the certificate is uniform in T, and the bound is vacuous once $(1+\eta\beta)^T \gtrsim \beta n D_\Theta / L$.

Proof of Theorem N.3. Let $\Delta_k := \|\theta_k^{(i)} - \theta_k^{(j)}\|$ be the distance between the two deletion iterates at step k. For equation 46, Lem. N.1(iv) applied to $R_{\mathcal{D}_N^{-i}}$, which inherits Def. 19 with the same constants because it is an average of functions satisfying it, gives

$$\Delta_{k+1} \le \sqrt{\varrho^2 \Delta_k^2 + 2\eta K} + \frac{2\eta L}{n}, \qquad \varrho^2 := 1 - 2\eta m + \eta^2\beta^2,$$

where ϱ is written with a distinct symbol to avoid confusion with the loss bound ρ of Def. 22. Squaring and applying $2ab \le \zeta a^2 + \zeta^{-1} b^2$ with $\zeta = \frac{\eta m}{2}$,

$$\Delta_{k+1}^2 \le \Big(1 + \tfrac{\eta m}{2}\Big)\big(\varrho^2 \Delta_k^2 + 2\eta K\big) + \Big(1 + \tfrac{2}{\eta m}\Big)\frac{4\eta^2 L^2}{n^2}.$$

The condition $\eta \le \frac{m}{\beta^2}$ gives $\varrho^2 \le 1 - \eta m$, hence $(1 + \frac{\eta m}{2})\varrho^2 \le 1 - \frac{\eta m}{2}$, while $\eta \le \frac{1}{m}$ gives $1 + \frac{\eta m}{2} \le \frac{3}{2}$ and $1 + \frac{2}{\eta m} \le \frac{3}{\eta m}$. Since $\Delta_0 = 0$, the iteration is dominated by its fixed point,

$$\Delta^2 \le \frac{3\eta K + \frac{12\eta L^2}{m n^2}}{\eta m / 2} = \frac{6K}{m} + \frac{24 L^2}{m^2 n^2},$$

uniformly in T_u, Conversion Theorem 1(2) with $p = 2$ concludes.

For the contractive display equation 47, Def. 26 states that the one-step transition rule contracts $\mathcal{W}_1$ at rate $1 - \kappa\eta$, so the display is first a $p = 1$ statement. Run the recursion of Prop. N.1 at the level of laws rather than realizations: writing μ_k, ν_k for the laws of the two unlearning runs and P, Q for the one-step transition rules driven by $R_{\mathcal{D}_N^{-i}}$ and $R_{\mathcal{D}_N^{-j}}$,

$$\mathcal{W}_1(\mu_{k+1}, \nu_{k+1}) \le \mathcal{W}_1(\mu_k P, \nu_k P) + \mathcal{W}_1(\nu_k P, \nu_k Q) \le (1 - \kappa\eta)\mathcal{W}_1(\mu_k, \nu_k) + \tfrac{2\eta L}{n}.$$

The first inequality uses the triangle inequality. The first term on the right contracts by Def. 26; the second term is at most $2\eta L/n$ because the two transition rules applied to the same point differ by the deterministic translation $\eta(\nabla R_{\mathcal{D}_N^{-i}} - \nabla R_{\mathcal{D}_N^{-j}})$. Since $\mu_0 = \nu_0$, unrolling gives $\mathcal{W}_1(u^{(i)}, u^{(j)}) \le \frac{2\eta L}{n}\sum_{k<T_u}(1-\kappa\eta)^k = \frac{2L}{\kappa n}(1 - (1-\kappa\eta)^{T_u})$. Conversion Theorem 1(2) at $p = 1$ then adds the endpoint-distance term. Under the additional pointwise hypothesis in the theorem statement, the same recursion holds pathwise in $\mathcal{W}_p$ and the common publication mechanism preserves the pathwise distance. □

The endpoint-distance term in equation 47 and in every contractive row of Table 1 is determined by the same contraction assumption (Def. 26) that bounds the unlearning-stability parameter. The contraction assumption determines the sensitivity used by the Gaussian certificate. Substituting that sensitivity into the CU inequalities replaces the fixed training-level endpoint term in Theorem 2 by the worst-case-distance term in Theorem N.4.

Theorem N.4 (Worst-Case Distance to Retraining, restates Theorem 3)**.** *Let $\mathcal{L}(\cdot, z)$ be L-Lipschitz. Assume a shared full-batch checkpoint, pointwise q-contractive retained-data maps under shared randomness with $q = 1 - \kappa\eta \in [0, 1)$, and a common data-independent Gaussian publication mechanism followed by an optional common nonexpansive projection. If the initial mismatch under the shared setup obeys*

$$b_0 \le \frac{2L}{\kappa N}(1 - q^{T - T_u}) \le \frac{2L}{\kappa n},$$

then the sensitivity of the worst-case distance to retraining and the direct endpoint term satisfy

$$s_{T_u} \le q^{T_u} b_0, \qquad 2r_{T_u,p} \le 2s_{T_u} \le \frac{4L}{\kappa n} q^{T_u}.$$

For $a_\delta = \sqrt{2\ln(1.25/\delta)}$, the Gaussian privacy relation $\epsilon = a_\delta q^{T_u} b_0 / \sigma$ is sufficient. If only distribution-level $\mathcal{W}_1$ contraction is known, the bound on s_{T_u} and the direct worst-case-distance

bound is valid only for $p = 1$; this argument does not provide a worst-case-distance sensitivity bound for $p > 1$. The CU-implied distance at $p = 1$ is

$$P^{\mathrm{TV}}(\epsilon, \delta) \le \frac{2LD_\Theta a_\delta}{\kappa\sigma n} q^{T_u} + 2D_\Theta\delta. \tag{48}$$

Proof. Under the pointwise shared-map hypothesis, the worst-case distance to retraining recursion is $s_{k+1} \le q s_k$ with $s_0 \le 2L(1 - q^{T-T_u})/(\kappa N)$. Hence $s_{T_u} \le q^{T_u} s_0 \le 2Lq^{T_u}/(\kappa n)$. Lem. L.1 converts a sensitivity s and a noise scale σ into the budget $\epsilon = s\sqrt{2\ln(1.25/\delta)}/\sigma$, so

$$\epsilon \;\le\; \frac{2L\sqrt{2\ln(1.25/\delta)}}{\kappa\sigma n}(1 - \kappa\eta)^{T_u} \;\le\; \frac{2L\sqrt{2\ln(1.25/\delta)}}{\kappa\sigma n}.$$

At $p = 1$, $C_{\epsilon,\delta} \le \delta + \tanh(\epsilon/2) \le \delta + \epsilon/2$, whence

$$P^{\mathrm{TV}}(\epsilon, \delta) \;\le\; 2D_\Theta\delta + D_\Theta\epsilon \;\le\; \frac{2LD_\Theta\sqrt{2\ln(1.25/\delta)}}{\kappa\sigma n}(1 - \kappa\eta)^{T_u} + 2D_\Theta\delta,$$

Equation equation 48 follows, and both displayed terms are $O(1/n)$ when $\delta = O(1/n)$. For every p, the direct term is at most $2s_{T_u}$. The CU-implied distance instead scales as $C_{\epsilon,\delta}^{1/p} = \Theta((\delta + \epsilon/2)^{1/p})$ and can therefore be $\Theta(D_\Theta(\sigma n)^{-1/p})$. Replacing D_Θ by the effective diameter equation 7 of the two published laws gives the dimensional form used in the main text: for a Gaussian mechanism, $D_{\mathrm{eff}} = O(\eta L T_u + \sigma\sqrt{d})$, and substituting $\sigma = s\sqrt{2\ln(1.25/\delta)}/\epsilon$ cancels the noise scale and leaves $O(s\sqrt{d\ln(1/\delta)}) = O\big(\frac{L}{\kappa n}\sqrt{d\ln(1/\delta)}\big)$, as detailed in Remark 2. □

Theorem N.5 (Verified KL stability result)**.** *Let $\mathcal{U}$ be a full-batch anchored unlearner of the form equation 20 with $\sigma > 0$, run on the two halves of $\mathcal{D}_{2n}$ as in Def. 12, and set*

$$\bar{s}_n^2 := \mathbb{E}_{\mathcal{D}_{2n},\mathcal{S}_3}\Big[\sup_{\theta\in\Theta}\big\|\nabla R_{\mathcal{D}^{(1)}}(\theta) - \nabla R_{\mathcal{D}^{(2)}}(\theta)\big\|^2\Big].$$

Then $\mathcal{U}$ is $(\Delta, 2n)$-KLSU (Def. 12) with

$$\Delta \le \frac{\eta\,\bar{s}_n^2}{4\sigma^2}\min\Big\{T_u,\ \frac{1}{1-\lambda}\Big\}, \tag{49}$$

where $\lambda < 1$ is an assumed or verified KL contraction coefficient of the unlearning update rule (Def. 27). A log-Sobolev inequality for the Gibbs measure $\pi \propto e^{-R_S/\sigma^2}$ can imply $1 - \lambda \ge 2\eta\sigma^2 m_{\mathrm{LSI}}$, with m_{LSI} depending on the chosen measure, domain, and discretization; Defs. 21 and 19 alone do not verify this coefficient. If $\mathcal{U}$ is ϵ-PCU for $\mathcal{A}$ (Def. 5), then $\mathcal{A}$ is (τ, n)-KLSL with

$$\tau \le \frac{e^\epsilon\,\bar{s}_n^2}{8\sigma^4 m_{\mathrm{LSI}}} + \epsilon\big(1 + \tanh(\epsilon/2)\big).$$

Under Def. 24 and the uniform gradient concentration of Lem. L.4, $\bar{s}_n^2 \le 2c_\Theta\sigma_{\mathrm{g}}^2/n$, so

$$\tau = O\Big(\frac{1}{n}\Big) + O(\epsilon), \tag{50}$$

uniformly in T and T_u. The KLSU inequality itself holds under Def. 4. To transfer that KLSU inequality to KLSL for $\mathcal{A}$, Def. 5 is required; the resulting KLSL bound is unconditional. If the certificate is genuinely (ϵ, δ) with $\delta > 0$, no KLSL conclusion is available (Remark 1), and one uses equation 24 instead, at the cost of the slower statistical term recorded in equation 62.

Proof. Let μ_k, ν_k be the laws of the two unlearning runs, which share the anchor, so $\mu_0 = \nu_0$ and $KL(\mu_0\|\nu_0) = 0$. Let P, Q be the one-step transition rules driven by $R_{\mathcal{D}^{(1)}}$ and $R_{\mathcal{D}^{(2)}}$. The chain rule for relative entropy, followed by data processing under marginalization, gives

$$KL(\mu_{k+1}\|\nu_{k+1}) \le KL(\mu_k\|\nu_k) + \mathbb{E}_{\mu_k}\big[KL\big(P(\theta,\cdot)\|Q(\theta,\cdot)\big)\big].$$

Both transition rules are Gaussian with covariance $2\eta\sigma^2 I$. Their means differ by $\eta(\nabla R_{\mathcal{D}^{(1)}}(\theta) - \nabla R_{\mathcal{D}^{(2)}}(\theta))$, so the one-step divergence is $\frac{\eta\|\nabla R_{\mathcal{D}^{(1)}}(\theta) - \nabla R_{\mathcal{D}^{(2)}}(\theta)\|^2}{4\sigma^2} \le \frac{\eta s^2}{4\sigma^2}$, where $s := \sup_\theta \|\nabla R_{\mathcal{D}^{(1)}}(\theta) - \nabla R_{\mathcal{D}^{(2)}}(\theta)\|$. Iterating this recurrence for T_u steps and taking $\mathbb{E}_{\mathcal{D}_{2n},\mathcal{S}_3}$ gives

the first branch of equation 49. If Q is also λ-KL contractive (Def. 27), the recurrence becomes $KL(\mu_{k+1}\|\nu_{k+1}) \le \lambda KL(\mu_k\|\nu_k) + \frac{\eta s^2}{4\sigma^2}$, whose fixed point gives the second branch. The log-Sobolev interpretation in (Dadi & Cevher, 2025) gives $1-\lambda \ge 2\eta\sigma^2 m_{\mathrm{LSI}}$ for a Gibbs target $\pi \propto e^{-R_S/\sigma^2}$, and therefore $\Delta \le \frac{\bar{s}_n^2}{8\sigma^4 m_{\mathrm{LSI}}}$. Conversion Theorem 1(3) transfers the KLSU inequality to KLSL, Lem. L.4 bounds $\bar{s}_n^2$, and $\epsilon(1+\tanh(\epsilon/2)) \le 2\epsilon$ gives the stated final rate. □

Theorem N.6 (Restatement of Theorem 4)**.** *Let $\mathcal{L}(\cdot, z)$ be L-Lipschitz and β-smooth, let $\mathrm{diam}(\Theta) \le D_\Theta$, and suppose that each retained empirical risk satisfies the μ_{PL}-Polyak–Lojasiewicz condition*

$$\frac{1}{2}\|\nabla R_{\mathcal{D}_N^{-i}}(\theta)\|^2 \ge \mu_{\mathrm{PL}}\big(R_{\mathcal{D}_N^{-i}}(\theta) - \min R_{\mathcal{D}_N^{-i}}\big) \quad \text{for every } \theta \in \Theta.$$

Assume in addition the matching quadratic-growth condition from Lem. J.7, where $\mathcal{X}^\star_{\setminus i} := \arg\min_{\theta\in\Theta} R_{\mathcal{D}_N^{-i}}(\theta)$, deterministic full-batch updates, a unique minimizer $\theta_i^\star$ in the interior of Θ for every retained empirical risk, and the minimizer-stability conclusion of Lem. J.8. The PL condition controls objective descent, while the QG condition controls distance to $\mathcal{X}^\star_{\setminus i}$; neither condition by itself implies contraction of the transition rule in Def. 26. Let $\mathcal{U}$ be a full-batch anchored unlearner with depth T_u and step $\eta \le \frac{1}{\beta}$ that is (ϵ,δ)-CU for $\mathcal{A}$, and let $\Psi_0 := \max_i\big(R_{\mathcal{D}_N^{-i}}(\theta^{\mathrm{a}}) - \min R_{\mathcal{D}_N^{-i}}\big)$. Then $\mathcal{U}$ is (Δ, N)-$\mathcal{W}_p$-USU with

$$\Delta \le \frac{4L}{\mu_{\mathrm{PL}} n} + 2\sqrt{\frac{2\Psi_0}{\mu_{\mathrm{PL}}}}\,(1-\eta\mu_{\mathrm{PL}})^{T_u/2}, \tag{51}$$

and, as soon as $T_u \ge \frac{1}{\eta\mu_{\mathrm{PL}}} \ln \frac{2\Psi_0\mu_{\mathrm{PL}} n^2}{4L^2}$, $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-USL with

$$\tau \le \frac{8L}{\mu_{\mathrm{PL}} n} + 2r_{T_u,p}. \tag{52}$$

If $\mathcal{L} \ge 0$, then L may be replaced in the endpoint comparison by $\sqrt{2\beta\Lambda}$ when the required self-bounding extension holds. The learner-stability bound in equation 52 is conditional on the PL–QG minimizer-distance argument and on the shared-anchor, common-transition, and worst-case-distance assumptions used in Theorem 4; smoothness alone does not make the learner-stability term or the endpoint-distance term uniform in T (Remark 4).

Proof. Under the PL condition and $\eta \le \frac{1}{\beta}$, full-batch gradient descent on $R_{\mathcal{D}_N^{-i}}$ satisfies $R_{\mathcal{D}_N^{-i}}(\theta_{k+1}) - \min R_{\mathcal{D}_N^{-i}} \le (1-\eta\mu_{\mathrm{PL}})(R_{\mathcal{D}_N^{-i}}(\theta_k) - \min R_{\mathcal{D}_N^{-i}})$, so after T_u steps from the anchor the optimization error is at most $(1-\eta\mu_{\mathrm{PL}})^{T_u}\Psi_0$. Lem. J.7 converts this into

$$r_i := \mathrm{dist}\big(u^{(i)}, \mathcal{X}^\star_{\setminus i}\big) \le \sqrt{\frac{2(1-\eta\mu_{\mathrm{PL}})^{T_u}\Psi_0}{\mu_{\mathrm{PL}}}}.$$

Let $\theta_i^\star \in \mathcal{X}^\star_{\setminus i}$ attain r_i and let $\theta_j^\star$ be its projection onto $\mathcal{X}^\star_{\setminus j}$. Lem. J.8 gives $\|\theta_i^\star - \theta_j^\star\| \le \frac{4L}{\mu_{\mathrm{PL}} n}$, so the triangle inequality yields $\|u^{(i)} - u^{(j)}\| \le r_i + \frac{4L}{\mu_{\mathrm{PL}} n} + r_j$, which is equation 51 since the selected minimizers are unique and hence the distance-to-set bounds give $\|u^{(i)} - \theta_i^\star\| \le r_i$ and $\|u^{(j)} - \theta_j^\star\| \le r_j$. The stated threshold on T_u makes the second term of equation 51 at most $\frac{4L}{\mu_{\mathrm{PL}} n}$, and Conversion Theorem 1(2) gives equation 52. The self-bounding variant replaces L by $\sqrt{2\beta\Lambda}$ in Lem. J.8, which is legitimate because the only use of the Lipschitz property there is to bound $|\mathcal{L}(\theta_i^\star, z) - \mathcal{L}(\theta_j^\star, z)|$ for $z \in \{z_i, z_j\}$ along a segment on which the loss does not exceed Λ. □

Theorem N.7 (Combination F: Smooth without a Lipschitz Bound)**.** *Let $\mathcal{L} \ge 0$ be β-smooth in θ, with no Lipschitz assumption, let $\mathrm{diam}(\Theta) \le D_\Theta$, and let $\mathcal{U}$ be a full-batch anchored unlearner with depth T_u, step $\eta \le \frac{1}{\beta}$, that is (ϵ,δ)-CU for $\mathcal{A}$. Let θ_k^{a} denote the unlearning iterate after k updates from the anchor θ^{a}. Write $\Lambda_k := \max_{z\in\mathcal{D}_N} \mathcal{L}(\theta_k^{\mathrm{a}}, z)$ for the largest per-sample training loss along the unlearning path. Then $\mathcal{U}$ is (Δ, N)-$\mathcal{W}_p$-USU with*

$$\Delta \le \frac{2\eta\sqrt{2\beta}}{n} \sum_{k=0}^{T_u-1} \mathcal{E}^{T_u-1-k}\sqrt{\Lambda_k}, \tag{53}$$

and $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-USL with $\tau \le \Delta + 2r_{T_u,p}$. Moreover the path-maximal loss is itself controlled by the optimization error at the anchor,

$$\max_{k<T_u} \Lambda_k \le 2\Lambda_0 + 2\beta\eta T_u \Psi_0, \qquad \Psi_0 := \max_i \big(R_{\mathcal{D}_N^{-i}}(\theta^{\mathrm{a}}) - \min R_{\mathcal{D}_N^{-i}}\big), \tag{54}$$

so that in the convex case $\mathcal{E} = 1$ and

$$\tau \le \frac{2\eta T_u \sqrt{2\beta(2\Lambda_0 + 2\beta\eta T_u \Psi_0)}}{n} + 2r_{T_u,p}. \tag{55}$$

The bounds in equation 53 and equation 54 depend on the dataset and algorithm. The bounds are informative when the shared anchor has small per-sample loss and small optimization error; Theorem N.7 does not imply either property for a general nonconvex empirical risk. The two bounds replace the global Lipschitz constant by quantities measured along the training run, while Corollary N.3 gives a worst-case substitute.

Proof. Run the proof of Prop. N.1 with the injection bounded by Lem. J.1, as in the proof of Thm. N.2; this gives equation 53. For equation 54, fix i and set $F := R_{\mathcal{D}_N^{-i}}$. Let $r_k := \|\theta_k^{\mathrm{a}} - \theta^{\mathrm{a}}\|$. Since $\eta \le \frac{1}{\beta}$, the descent lemma gives $F(\theta_{l+1}^{\mathrm{a}}) \le F(\theta_l^{\mathrm{a}}) - \frac{\eta}{2}\|\nabla F(\theta_l^{\mathrm{a}})\|^2$. Hence $\eta \sum_{l<k} \|\nabla F(\theta_l^{\mathrm{a}})\|^2 \le 2\Psi_0$, and by Cauchy–Schwarz

$$r_k \le \eta \sum_{l<k} \|\nabla F(\theta_l^{\mathrm{a}})\| \le \sqrt{k\eta \cdot \eta \sum_{l<k} \|\nabla F(\theta_l^{\mathrm{a}})\|^2} \le \sqrt{2k\eta\Psi_0}.$$

Smoothness and Lem. J.1 then give, for every $z \in \mathcal{D}_N$,

$$\mathcal{L}(\theta_k^{\mathrm{a}}, z) \le \mathcal{L}(\theta^{\mathrm{a}}, z) + \sqrt{2\beta\mathcal{L}(\theta^{\mathrm{a}}, z)}\, r_k + \tfrac{\beta}{2} r_k^2 \le 2\mathcal{L}(\theta^{\mathrm{a}}, z) + \beta r_k^2$$

by Young's inequality, and substituting $r_k^2 \le 2T_u\eta\Psi_0$ yields equation 54. Combining with $\mathcal{E} = 1$ and Conversion Theorem 1(2) gives equation 55. □

Corollary N.3 (Combination F, Worst-Case Surrogate)**.** *In the setting of Thm. N.7 with $\mathcal{E} = 1$, suppose each $\min R_{\mathcal{D}_N^{-i}}$ is attained at a point interior to Θ. Then $\Psi_0 \le \frac{\beta}{2} D_\Theta^2$, and if in addition $|\mathcal{L}| \le \rho$ (Def. 22) so that $\Lambda_0 \le \rho$,*

$$\tau \le \frac{2\eta T_u}{n}\Big(2\sqrt{\beta\rho} + \beta D_\Theta \sqrt{2\beta\eta T_u}\Big) + 2r_{T_u,p}, \tag{56}$$

with no reference to the quality of the anchor. If the minimizer is on the boundary, the bound remains valid after replacing Ψ_0 by $D_\Theta \sqrt{2\beta \min R_{\mathcal{D}_N^{-i}}} + \frac{\beta}{2} D_\Theta^2$.

Proof. Fix i and write $F := R_{\mathcal{D}_N^{-i}}$, which is nonnegative and β-smooth as an average of nonnegative β-smooth functions. Let $\theta^\star \in \arg\min F$. Smoothness gives

$$F(\theta^{\mathrm{a}}) - F(\theta^\star) \le \big\langle \nabla F(\theta^\star), \theta^{\mathrm{a}} - \theta^\star \big\rangle + \tfrac{\beta}{2}\big\|\theta^{\mathrm{a}} - \theta^\star\big\|^2.$$

If $\theta^\star$ is interior then $\nabla F(\theta^\star) = 0$ and the first term vanishes, so $\Psi_0 \le \frac{\beta}{2} D_\Theta^2$. Otherwise bound the first term by $\|\nabla F(\theta^\star)\| D_\Theta$ and apply Lem. J.1 to F, giving $\|\nabla F(\theta^\star)\| \le \sqrt{2\beta \min F}$ and hence the stated variant. For equation 56, insert $\Psi_0 \le \frac{\beta}{2} D_\Theta^2$ and $\Lambda_0 \le \rho$ into equation 54 to get $\max_{k<T_u} \Lambda_k \le 2\rho + \beta^2\eta T_u D_\Theta^2$, then use equation 53 with $\mathcal{E} = 1$ together with the subadditivity of the square root,

$$\sqrt{2\beta\big(2\rho + \beta^2\eta T_u D_\Theta^2\big)} \le 2\sqrt{\beta\rho} + \beta D_\Theta \sqrt{2\beta\eta T_u},$$

and add the endpoint-distance term from Conversion Theorem 1(2). □

Proposition N.2 (Combination G: Convex, Lipschitz, Nonsmooth)**.** *For convex L-Lipschitz losses that need not be smooth, the same-randomness comparison argument of (Bassily et al., 2020) gives equation 21 only in the squared sense, with an additive $4\eta^2 L^2$ per step, and Prop. N.1 then yields*

$$\Delta \le 2\eta L\sqrt{T_u} + \frac{2\eta L T_u}{n}, \qquad \tau \le \Delta + 2r_{T_u,p}. \tag{57}$$

Proof. By Lem. N.1(v), $\Delta_{k+1} \le \sqrt{\Delta_k^2 + 4\eta^2 L^2} + \eta J_k$ with $J_k \le \frac{2L}{n}$. Using $\sqrt{(a+b)^2 + c^2} \le \sqrt{a^2 + c^2} + b$ for $a, b, c \ge 0$, an induction shows $\Delta_k \le a_k + b_k$ with $a_{k+1} = \sqrt{a_k^2 + 4\eta^2 L^2}$, $a_0 = 0$, and $b_{k+1} = b_k + \frac{2\eta L}{n}$, $b_0 = 0$, so that $a_{T_u} = 2\eta L\sqrt{T_u}$ and $b_{T_u} = \frac{2\eta L T_u}{n}$. □

O PROOFS OF EXCESS POPULATION RISK BOUNDS

Table O.1 collects the excess-risk conclusions for Combinations A–F of Table 1 and records, for each row, when the bound improves on the classical stability-to-risk baseline of Tables E.1–E.3. Rows C† and G are omitted because the corresponding stability bounds are vacuous (Theorem 6). The comparison is for the *statistical* term obtained by feeding τ into Lemmas J.3–J.6, optimization error ϵ_{opt} is included identically on both sides unless noted.

Theorem O.1 (Excess Risk for Combinations A and B)**.** *In the setting of Theorem N.1, with $\mathcal{A}$ run for T steps with $\eta_t = \frac{1}{\gamma t}$,*

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \le \frac{2L^2}{\gamma n} + 2L r_{T_u,p} + O\Big(\frac{L^2 \ln T}{\gamma T}\Big). \tag{58}$$

In the setting of Thm. N.2, with constant step $\eta = \frac{D_\Theta}{L\sqrt{T}}$,

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \le \frac{2LD_\Theta T_u}{n\sqrt{T}} + 2L r_{T_u,p} + \frac{2LD_\Theta}{\sqrt{T}}. \tag{59}$$

Proof. Insert equation 40 into Lem. J.3 to obtain the generalization term, bound $1-(1-\eta\gamma)^{T_u} \le 1$, and add the optimization error appearing in equation 28. For the strongly convex case with $\eta_t = \frac{1}{\gamma t}$ the classical rate $\epsilon_{\mathrm{opt}} = O(\frac{L^2 \ln T}{\gamma T})$ applies. For the convex case with constant $\eta = \frac{D_\Theta}{L\sqrt{T}}$ the standard bound $\epsilon_{\mathrm{opt}} \le \frac{D_\Theta^2}{2\eta T} + \frac{\eta L^2}{2} = \frac{LD_\Theta}{\sqrt{T}}$ applies, and substituting η into equation 43 gives $\frac{2LD_\Theta T_u}{n\sqrt{T}}$. □

Theorem O.2 (Excess Risk for Combinations C and D)**.** *In the setting of Thm. N.3 with a κ-contractive transition rule (Def. 26; (Eberle, 2016; Zhu et al., 2023)),*

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \le \frac{2L^2}{\kappa n} + 2L r_{T_u,p} + \epsilon_{\mathrm{opt}}, \tag{60}$$

uniformly in T. In the setting of Thm. N.5 with $|\mathcal{L}| \le \rho$ and an ϵ-PCU certificate (Def. 5),

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \le \rho\sqrt{\frac{2}{n}\Big(\frac{e^\epsilon c_\Theta \sigma_{\mathrm{g}}^2}{4\sigma^4 m_{\mathrm{LSI}} n} + 2\epsilon\Big)} + \epsilon_{\mathrm{opt}}, \tag{61}$$

uniformly in T and T_u. If the certificate satisfies only Def. 4 with $\delta > 0$, then equation 61 is replaced by

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \le \rho\sqrt{2\Delta} + 4\rho\, C_{\epsilon,\delta} + \epsilon_{\mathrm{opt}}. \tag{62}$$

Both bounds are unconditional expectation bounds.

Proof. For equation 60, insert equation 47 into Lem. J.3; neither the unlearning-stability parameter Δ nor the CU-implied distance depends on T, so the population-risk bound is uniform in T. For equation 61, insert equation 50 into Lem. J.6. The KLSU-to-KLSL conversion and the KL-to-risk lemma are both unconditional because Def. 5 makes Conversion Theorem 1(3) unconditional.

For equation 62, drop Def. 5 and keep Def. 4. The generalization gap is now bounded through Conversion Theorem 1(1) in the form equation 24 rather than through Lem. J.6, since Def. 6 is by definition the leave-one-out generalization gap: $|\mathbb{E}[R - R_{\mathcal{D}_n}]| \le \tau \le \Delta_{\mathrm{SU}} + 4\rho C_{\epsilon,\delta}$, where Δ_{SU} is the SU parameter of the same unlearner. It remains to express Δ_{SU} through the KL stability already computed in equation 49. For each pair of output laws entering Def. 9, Pinsker's inequality gives $\mathrm{TV}(P_{u^{(i)}}, P_{u^{(j)}}) \le \sqrt{KL(P_{u^{(i)}} \| P_{u^{(j)}})/2}$, so a loss bounded by ρ satisfies

$$\big|\mathbb{E}[\mathcal{L}(u^{(i)}, z) - \mathcal{L}(u^{(j)}, z)]\big| \le 2\rho\ \mathrm{TV} \le \rho\sqrt{2KL(P_{u^{(i)}} \| P_{u^{(j)}})},$$

and averaging over pairs together with concavity of the square root gives $\Delta_{\mathrm{SU}} \le \rho\sqrt{2\Delta}$. Adding ϵ_{opt} as in equation 28 yields equation 62. Since $\Delta = O(1/n)$ by Lem. L.4, the term $\rho\sqrt{2\Delta}$ is $O(n^{-1/2})$. Compared with equation 61, the rate loses a square root because Pinsker's inequality converts KL divergence to total variation. This loss is unavoidable for this route that uses the CU inequalities alone because total variation does not inherit the $1/n$ scaling of the empirical gradient average. □

Theorem O.3 (Conditional Excess Risk Bound for Combination E)**.** *In the setting of Thm. 4 with $\mathcal{L} \geq 0$, let ϵ_{opt} include the optimization and publication-noise error of the output law, let $\epsilon_{\mathrm{opt}} \leq (1-\eta\mu_{\mathrm{PL}})^T\Psi_0$ for the deterministic pre-publication iterate, and assume a deterministic global bound $\Lambda \geq \mathcal{L}(\theta, z)$ on the relevant path together with the dispersion condition*

$$\Lambda \leq c_\Lambda\, \mathbb{E}\big[R(\theta^{\mathrm{out}})\big] \quad \textit{for some } c_\Lambda \geq 1. \tag{63}$$

If $n \geq \frac{32c_\Lambda\beta}{\mu_{\mathrm{PL}}}$, then

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \leq \frac{32c_\Lambda\beta\big(R^\star+\epsilon_{\mathrm{opt}}\big)}{\mu_{\mathrm{PL}}n} + 4\sqrt{2\beta\Lambda}\, r_{T_u,p} + \epsilon_{\mathrm{opt}}. \tag{64}$$

In the interpolation regime $R^\star = 0$, taking $T = \frac{2\ln n}{\eta\mu_{\mathrm{PL}}}$ so that the deterministic optimization error is at most $\Psi_0 n^{-2}$,

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star = O\Big(\frac{1}{n^2}\Big) + 4\sqrt{2\beta\Lambda}\, r_{T_u,p} + \epsilon_{\mathrm{opt}}. \tag{65}$$

Remark 5 (When the dispersion condition fails.)**.** *Condition equation 63 is the only hypothesis of Thm. O.3 that goes beyond those of Thm. N.6, and we do not claim it holds in applications. Two endpoint-distance bounds are available, both proved alongside equation 64. First, if the loss is merely bounded, $|\mathcal{L}| \leq \rho$ (Def. 22), then substituting $\Lambda \leq \rho$ requires no fixed-point step and gives the honest default rate*

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \leq \frac{16\beta\rho}{\mu_{\mathrm{PL}}n} + \sqrt{2\beta\rho}\, 2r_{T_u,p} + \epsilon_{\mathrm{opt}}, \tag{66}$$

which is $O((\mu_{\mathrm{PL}}n)^{-1})$ and carries neither the condition on n nor the factor 2 in front of $R^\star + \epsilon_{\mathrm{opt}}$. Second, the relaxation

$$\mathbb{E}\big[\Lambda\big] \leq c_\Lambda\, \mathbb{E}\big[R(\theta^{\mathrm{out}})\big] + \Lambda_\perp, \tag{67}$$

with an additive dispersion floor $\Lambda_\perp \geq 0$, interpolates between the two regimes and yields, under the same condition $n \geq 32c_\Lambda\beta/\mu_{\mathrm{PL}}$,

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \leq \frac{32c_\Lambda\beta\big(R^\star+\epsilon_{\mathrm{opt}}\big)}{\mu_{\mathrm{PL}}n} + \frac{32\beta\Lambda_\perp}{\mu_{\mathrm{PL}}n} + 4\sqrt{2\beta\Lambda}\, r_{T_u,p} + \epsilon_{\mathrm{opt}}. \tag{68}$$

Setting $\Lambda_\perp = 0$ recovers equation 64, and letting $c_\Lambda \to 0$ with $\Lambda_\perp = \rho$ recovers equation 66 up to constants.

Proof of Theorem O.3 and Remark 5. Let $\hat{R} := \mathbb{E}[R(\theta^{\mathrm{out}})]$. By the self-bounding variant of Thm. 4 and Lem. J.3 applied with the effective Lipschitz constant $\sqrt{2\beta\Lambda}$,

$$\mathbb{E}\big[R - R_{\mathcal{D}_n}\big] \leq \sqrt{2\beta\Lambda}\Big(\frac{8\sqrt{2\beta\Lambda}}{\mu_{\mathrm{PL}}n} + 2r_{T_u,p}\Big) = \frac{16\beta\Lambda}{\mu_{\mathrm{PL}}n} + \sqrt{2\beta\Lambda}\, 2r_{T_u,p}.$$

Taking expectations and applying the dispersion condition equation 63 to the first term gives $\mathbb{E}[R - R_{\mathcal{D}_n}] \leq \frac{16c_\Lambda\beta\hat{R}}{\mu_{\mathrm{PL}}n} + \sqrt{2\beta\Lambda}\, 2r_{T_u,p}$. Since $\hat{R} \leq R^\star + \epsilon_{\mathrm{opt}} + \mathbb{E}[R - R_{\mathcal{D}_n}]$ by equation 28,

$$\hat{R} \leq R^\star + \epsilon_{\mathrm{opt}} + \frac{16c_\Lambda\beta\hat{R}}{\mu_{\mathrm{PL}}n} + \sqrt{2\beta\Lambda}\, 2r_{T_u,p}.$$

For $n \geq \frac{32c_\Lambda\beta}{\mu_{\mathrm{PL}}}$, the coefficient of $\hat{R}$ in the third term is at most $1/2$, so moving that term to the left gives $\hat{R} \leq 2(R^\star + \epsilon_{\mathrm{opt}}) + 4\sqrt{2\beta\Lambda}\, r_{T_u,p}$. Substituting this bound on $\hat{R}$ into the generalization inequality introduces a additional term involving the CU comparison; the same coefficient condition absorbs at most half of that cross term and yields equation 64. Under PL with $\eta \leq \frac{1}{\beta}$, the choice $T = \frac{2\ln n}{\eta\mu_{\mathrm{PL}}}$ makes $\epsilon_{\mathrm{opt}} \leq (1-\eta\mu_{\mathrm{PL}})^T\Psi_0 \leq \Psi_0 n^{-2}$, which gives equation 65 when $R^\star = 0$. Condition equation 63 holds with $c_\Lambda = O(1)$ whenever the per-sample losses along the path are uniformly comparable to their average. This condition is the only place where our fast-rate claim goes beyond the hypotheses of Thm. 4; it plays the role of the low-noise condition in (Lei & Ying, 2021).

The two variants of Rem. 5 follow from the same display. For equation 66, substitute $\Lambda \le \rho$ into $\mathbb{E}[R - R_{\mathcal{D}_n}] \le \frac{16\beta\Lambda}{\mu_{\mathrm{PL}} n} + \sqrt{2\beta\Lambda}\, 2r_{T_u,p}$ and add ϵ_{opt}. No fixed-point step is needed because the first term no longer involves $\hat{R}$. This is why equation 66 has neither the condition on n nor the factor 2 multiplying $R^\star + \epsilon_{\mathrm{opt}}$. For equation 68, replace equation 63 by equation 67. The generalization bound becomes $\frac{16\beta c_\Lambda \hat{R}}{\mu_{\mathrm{PL}} n} + \frac{16\beta\Lambda_\perp}{\mu_{\mathrm{PL}} n} + \sqrt{2\beta\Lambda}\, 2r_{T_u,p}$. The additive floor passes unchanged through the absorption step under $n \ge \frac{32 c_\Lambda \beta}{\mu_{\mathrm{PL}}}$; doubling the floor once gives the stated $\frac{32\beta\Lambda_\perp}{\mu_{\mathrm{PL}} n}$. □

Theorem O.4 (Excess Risk for Combination F)**.** *In the setting of Thm. N.7 with $\mathcal{E} = 1$, use its same-randomness comparison to instantiate the $p = 2$ worst-case-distance transfer bound and apply equation 29 with $c_0 = 1$,*

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \le \mathbb{E}\big[R_{\mathcal{D}_n}(\theta^{\mathrm{out}})\big] + \beta\Big(\frac{2\eta T_u\sqrt{2\beta\Lambda}}{n} + 2r_{T_u,p}\Big)^2 + \epsilon_{\mathrm{opt}}. \tag{69}$$

If the anchor is arbitrary rather than well trained, substituting the worst-case surrogate of Cor. N.3 gives, under the additional hypotheses $|\mathcal{L}| \le \rho$ (Def. 22) and interiority of each $\arg\min R_{\mathcal{D}_N^{-i}}$*,*

$$\mathbb{E}\big[R(\theta^{\mathrm{out}})\big] - R^\star \le \mathbb{E}\big[R_{\mathcal{D}_n}(\theta^{\mathrm{out}})\big] + \beta\Big(\frac{2\eta T_u}{n}\big(2\sqrt{\beta\rho} + \beta D_\Theta\sqrt{2\beta\eta T_u}\big) + 2r_{T_u,p}\Big)^2 + \epsilon_{\mathrm{opt}}. \tag{70}$$

Proof. Apply Lem. J.4 with $c_0 = 1$ to the $p = 2$ instance of the same-randomness comparison in Thm. N.7, and use τ given by equation 55; then add ϵ_{opt} as in equation 28. For equation 70, use equation 56 in place of equation 55. □

P Proofs of the Impossibility Results

Definition 28 (Optimal Stability Profiles and Admissible Schemes.)**.** *For $p \ge 1$ write*

$$\tau_p^\star(\mathcal{A}, n) := \sup_{\mathcal{D}_n, i} \mathcal{W}_p\big(\mathcal{A}(\mathcal{D}_n), \mathcal{A}(\mathcal{D}_n^{(i)})\big), \qquad \Delta_p^\star(\mathcal{U}, N) := \sup_{\mathcal{D}_N, i, j} \mathcal{W}_p\big(u^{(i)}, u^{(j)}\big).$$

An admissible CU scheme $\{(\mathcal{A}_n, \mathcal{U}_n)\}_{n\ge 2}$ *satisfies, as $n \to \infty$, the following. (C1) $\epsilon_n + \delta_n = o(1)$, (C2) $\Delta_p^\star(\mathcal{U}_n, n+1) = o(1)$, (C3) $T_u(n) = o(T(n))$, (C4) $\sigma(n)\sqrt{d} \le c_0 D_\Theta$.*

Condition (C4) formalizes utility preservation. Without the noise-scale restriction in (C4), an unlearner could add arbitrarily large noise and satisfy the CU inequalities without preserving a useful model. Condition (C1) makes the CU privacy parameters vanish, (C2) prevents the published unlearned output law from revealing the deleted index, (C3) requires the unlearning procedure to use asymptotically fewer updates than retraining, and (C4) limits the publication-noise scale so that the released model remains useful. Together, (C1)–(C4) define the non-vacuous unlearning regime studied below.

For the Gaussian depth statements below, let Φ be the standard normal cdf and define the exact equal-covariance Gaussian profile

$$\delta_G(\epsilon, \mu) := \Phi\left(\frac{\mu}{2} - \frac{\epsilon}{\mu}\right) - e^\epsilon \Phi\left(-\frac{\mu}{2} - \frac{\epsilon}{\mu}\right), \qquad \mu_{\epsilon,\delta} := \sup\{\mu \ge 0 : \delta_G(\epsilon, \mu) \le \delta\}.$$

For two Gaussian laws with common covariance $\sigma^2 I$ and mean gap s, exact (ϵ, δ) indistinguishability is equivalent to $s/\sigma \le \mu_{\epsilon,\delta}$. This replaces the classical sufficient noise choice whenever a lower bound is needed.

Proposition P.1 (Contrapositive of the Conversion Theorem, restates Proposition 2)**.** *Let $p \ge 1$, let $\mathrm{diam}(\Theta) \le D_\Theta < \infty$, and let $\mathcal{U}$ be (ϵ, δ)-CU for $\mathcal{A}$ (Def. 4) on single-point deletions. Then, with $\tau_p^\star$ and $\Delta_p^\star$ as in Def. 28,*

$$\Delta_p^\star(\mathcal{U}, n+1) \;\ge\; \tau_p^\star(\mathcal{A}, n) - 2r_{T_u,p}, \tag{71}$$

and the CU-implied distance also gives

$$\Delta_p^\star(\mathcal{U}, n+1) \;\ge\; \tau_p^\star(\mathcal{A}, n) - P^{\mathrm{TV}}(\epsilon, \delta), \tag{72}$$

and consequently

$$\frac{\epsilon}{2} + \delta \;\ge\; \Big(\frac{\big(\tau_p^\star(\mathcal{A}, n) - \Delta_p^\star(\mathcal{U}, n+1)\big)_+}{2D_\Theta}\Big)^p. \tag{73}$$

Proof. Theorem I.2(2) with $\Delta_1 = \Delta_p^\star(\mathcal{U}, n+1)$ gives $\tau_p^\star(\mathcal{A}, n) \le \Delta_p^\star(\mathcal{U}, n+1) + 2r_{T_u,p}$; rearranging proves equation 71. Applying the CU-implied distance in Conversion Theorem I.2(2) gives equation 72. For equation 73, assume $\tau_p^\star > \Delta_p^\star$; otherwise the lower bound in equation 73 is zero. Then equation 72 gives $P^{\mathrm{TV}}(\epsilon,\delta) \ge \tau_p^\star - \Delta_p^\star > 0$, so $C_{\epsilon,\delta} \ge \big((\tau_p^\star - \Delta_p^\star)/(2D_\Theta)\big)^p$. Finally, $C_{\epsilon,\delta} = \min\{1, \delta + (1-\delta)\tanh(\epsilon/2)\} \le \delta + \tanh(\epsilon/2) \le \delta + \epsilon/2$ because $\tanh(x) \le x$ for $x \ge 0$. □

Equation equation 73 is the key step in this appendix. It converts any lower bound on the learner-stability parameter $\tau_p^\star(\mathcal{A}, n)$ into a lower bound on the CU privacy parameters of the unlearner $\mathcal{U}$, using only the CU inequalities and no additional property of $\mathcal{U}$.

Corollary P.1 (Certified Unlearning Does Not Manufacture Stability.)**.** *The nonsmooth obstruction and the contrapositive budget bound are mechanism-independent. The depth obstructions below are stated for the specified rewind/descend architectures and Gaussian output perturbation; changing either architecture or accountant requires a separate lower-bound argument.*

Proof. Part (i) derives its contradiction from equation 73, whose only inputs are Def. 4 and a learner-stability lower bound. Parts (ii) and (iii) additionally assume the rewind/descend architecture and Gaussian accountant; their lower bounds use the exact Gaussian profile above and the worst-case-distance identities in Lemmata P.2 and P.3, so they do not apply to arbitrary unlearning architectures or privacy accountants. □

The lower-bound formulas used in Theorem 6 are:

$$\Delta_p^\star(\mathcal{U}, N) \;\ge\; c_1\Big(\eta L\sqrt{T_u} + \frac{\eta L T_u}{n}\Big), \tag{74}$$

$$s_{T_u} = \frac{2L}{\beta N}\big[(1+\eta\beta)^T - (1+\eta\beta)^{T_u}\big] \;\ge\; \frac{2\eta L}{N}(1+\eta\beta)^{T-1} \quad (T_u < T), \tag{75}$$

$$T \;\ge\; T^\dagger := 1 + \frac{1}{\ln(1+\eta\beta)}\ln\frac{c_0 D_\Theta N \mu_{\epsilon,\delta}}{2\eta L\sqrt{d}} \;=\; O\Big(\frac{\ln(nD_\Theta)}{\ln(1+\eta\beta)}\Big), \tag{76}$$

$$T + T_u \;\ge\; \Big(\frac{\beta}{\gamma} - 1\Big)\ln\frac{D_\Theta^2 \gamma n\sqrt{2\ln(1.25/\delta)}}{4B\sigma_{\mathrm{pub}}} \;=\; \Omega\big((\beta/\gamma)\ln(\gamma n)\big). \tag{77}$$

Equation equation 74 is the nonsmooth unlearner-instability floor (Def. 10) proved in Lemma P.1; equation equation 75 is the exact rewind sensitivity proved in Lemma P.2; equation equation 76 gives the training horizon beyond which the expansive construction violates utility condition (C4); and equation equation 77 gives the total update count required by the ill-conditioned quadratic construction in Lemma P.3. The three lemmas are proved next, followed by the impossibility theorem.

Lemma P.1 (Nonsmooth Instability of the Anchored Pair (Bassily et al., 2020))**.** *Let $\eta > 0$, $T_u \ge 1$ and $d \ge \min\{T_u, \eta^{-2}\}$, fix $\theta^{\mathrm{a}} \in \mathbf{R}^d$ and suppose $\Theta \supseteq B(\theta^{\mathrm{a}}, 4\eta L\sqrt{T_u})$. There exist a convex L-Lipschitz family $\{\mathcal{L}(\cdot, z)\}_{z\in\mathcal{Z}}$, a dataset $\mathcal{D}_N$ and indices $i \ne j$ such that the two full-batch subgradient runs*

$$\theta_{k+1}^{(i)} = \Pi_\Theta\big(\theta_k^{(i)} - \eta\nabla R_{\mathcal{D}_N^{-i}}(\theta_k^{(i)})\big), \qquad \theta_0^{(i)} = \theta^{\mathrm{a}},$$

The j-run is defined by the same recursion with i replaced by j. The two runs satisfy $\big\|\theta_{T_u}^{(i)} - \theta_{T_u}^{(j)}\big\| \ge c_1\big(\eta L\sqrt{T_u} + \eta L T_u/N\big)$ for a universal $c_1 > 0$.

Proof. The lower bound of (Bassily et al., 2020) for full-batch subgradient descent produces, for every η, every T and every $d \ge \min\{T, \eta^{-2}\}$, a convex L-Lipschitz family and a pair of datasets differing in one entry on which T steps started at the origin separate by $\Omega(\eta L\sqrt{T} + \eta L T/N)$. Their construction is a maximum of linear functions plus a linear drift controlled by one sample. The drift makes successive subgradient steps nearly orthogonal, which yields the $\sqrt{T}$ contribution, and the iterates remain inside a ball of radius $4\eta L\sqrt{T}$.

Translate every loss by $\theta \mapsto \theta - \theta^{\mathrm{a}}$. Convexity, the Lipschitz constant, and the subgradient oracle are preserved, and the construction starts at the anchor. Def. 14 allows this anchor because it is independent of the forget set. The sets $\mathcal{D}_N^{-i}$ and $\mathcal{D}_N^{-j}$ differ in one entry, so they form the required replacement pair. Finally, the projection is inactive because the trajectory stays in $B(\theta^{\mathrm{a}}, 4\eta L\sqrt{T_u}) \subseteq \Theta$. Applying Lemma P.1 with $T = T_u$ yields the displayed lower bound. □

Lemma P.2 (The Rewind Sensitivity Bound Is Attained)**.** *Let $\beta > 0$, $\eta \le 1/\beta$ and $\mathcal{L}(\theta, z) = -\frac{\beta}{2}\|\theta\|^2 + \langle b(z), \theta\rangle$ with $\|b(z)\| = L$, which is β-smooth and, on Θ, Lipschitz with constant $L + \beta D_\Theta$. Let $\mathcal{D}_N$ satisfy $b(z_k) = b$ for $k \neq i$ and $b(z_i) = -b$. Let θ_t and θ'_t be the full-batch runs of equation 20 at $\sigma = 0$ on $\mathcal{D}_N$ and on $\mathcal{D}_N^{-i}$ from a shared θ_0. Then, as long as the iterates remain interior to Θ,*

$$\|\theta_t - \theta'_t\| = \frac{2L}{\beta N}\big[(1+\eta\beta)^t - 1\big], \tag{78}$$

and the rewind unlearner with $\theta^{\mathrm{a}} = \theta_{T-T_u}$ has, for every $T_u \le T$,

$$s_{T_u} = \frac{2L}{\beta N}\big[(1+\eta\beta)^T - (1+\eta\beta)^{T_u}\big] \;\ge\; \frac{2\eta L}{N}(1+\eta\beta)^{T-1} \quad (T_u < T), \tag{79}$$

with equality in Lem. L.2(ii). The descend anchor $\theta^{\mathrm{a}} = \theta_T$ gives $s_{T_u} = (1+\eta\beta)^{T_u}\frac{2L}{\beta N}[(1+\eta\beta)^T - 1]$, which is increasing in T_u.

Proof. The gradient is $\nabla R_S(\theta) = -\beta\theta + \bar{b}_S$, so one step at $\sigma = 0$ is $G_S(\theta) = (1+\eta\beta)\theta - \eta\bar{b}_S$. Its linear part is $(1+\eta\beta)I$ and is independent of S. Hence

$$\theta_t - \theta'_t = (1+\eta\beta)(\theta_{t-1} - \theta'_{t-1}) - \eta g, \qquad g = \bar{b}_{\mathcal{D}_N} - \bar{b}_{\mathcal{D}_N^{-i}}.$$

For the dataset with $b(z_k) = b$ for $k \neq i$ and $b(z_i) = -b$, $\|g\| = 2L/N$. Unrolling from the shared initialization and summing the geometric series gives equation 78.

The unlearned model is $G^{T_u}_{\mathcal{D}_N^{-i}}(\theta_{T-T_u})$ and the retrained model is $G^{T_u}_{\mathcal{D}_N^{-i}}(\theta'_{T-T_u})$. The affine retained-data map multiplies the distance between these two iterates by $(1+\eta\beta)^{T_u}$, yielding the first equality in equation 79. Writing $T_u = T - k$ and minimizing over $k \ge 1$ gives $k = 1$ and the stated lower bound. With the descend anchor θ_T, the same calculation gives a sensitivity that increases with T_u. □

Lemma P.3 (Ill-Conditioned Quadratics: Stability Floor (Def. 7) and Depth Requirement)**.** *Let $\Theta = [-1,1]^2$, $H_n = \mathrm{diag}(\gamma_n, \beta)$ with $0 < \gamma_n < \beta$, $\eta = 1/\beta$, and $\chi_n := \beta/\gamma_n$. For a fixed $B > 0$ and $z \in [-1,1]$, set*

$$\mathcal{L}_n(\theta, z) = B + \tfrac{1}{2}\langle\theta, H_n\theta\rangle - Bz\theta_1.$$

This loss is nonnegative on Θ, γ_n-strongly convex, β-smooth, and L_ℓ-Lipschitz with $L_\ell := B + \beta D_\Theta$. Its empirical minimizer is $c_S = (B\bar{z}_S/\gamma_n)e_1$; write $c_{\setminus i} := c_{\mathcal{D}_N^{-i}}$ for the retained-data minimizer. For the learner-stability floor (Def. 7), take a size-n sample with one $z = -1$ and $n-1$ samples $z = 1/(n-1)$, and its replacement with that record changed to $z = 1$; then $\bar{z}_S = 0$, $\bar{z}_{S'} = 2/n$, and choosing $\gamma_n = 2B/n$ puts the two minimizers one unit apart inside Θ. For the depth statement below, assume the compared minimizers and iterates remain in the interior of Θ.

(i) Stability floor. *For the centered construction above, exact ERM has $\tau_p^\star(\mathcal{A}_n, n) \ge 1$ for every n, while $\chi_n = \beta n/(2B) = \Omega(n)$.*

(ii) Depth requirement. *For an interior quadratic pair with eigenvalues γ and β, if $\theta_0 - c_{\mathcal{D}_N}$ lies in the γ-eigenspace and points in the same direction as $c_{\mathcal{D}_N} - c_{\setminus i}$, then*

$$s_{T_u} = (1-\eta\gamma)^{T_u}\Big[(1-\eta\gamma)^T\|\theta_0 - c_{\mathcal{D}_N}\| + \|c_{\mathcal{D}_N} - c_{\setminus i}\|\Big] \ge (1-\eta\gamma)^{T+T_u}\|\theta_0 - c_{\mathcal{D}_N}\|, \tag{80}$$

and if moreover $\|\theta_0 - c_{\mathcal{D}_N}\| \ge D_\Theta/2$, the certificate is issued by output perturbation at noise σ_{pub}, and the budget based on the worst-case distance of Thm. N.4, in the form $\epsilon \le \frac{2B}{D_\Theta\gamma n}$, is imposed, then

$$T + T_u \;\ge\; \Big(\frac{\beta}{\gamma} - 1\Big)\,\ln\frac{D_\Theta}{2\sigma_{\mathrm{pub}}\,\mu_{2B/(D_\Theta\gamma n),\delta}}. \tag{81}$$

(iii) Retraining baseline. *The same recursion applied to $R_{\mathcal{D}_N^{-i}}$ from θ_0 reaches optimization error e after exactly $T^{\mathrm{re}} = \frac{\ln(e_0/e)}{2\ln(1-\eta\gamma)^{-1}} = \Theta(\chi_n \ln(e_0/e))$ steps, where $e_0 = R_{\mathcal{D}_N^{-i}}(\theta_0) - \min R_{\mathcal{D}_N^{-i}}$ and $\chi_n = \beta/\gamma_n$ is the condition number.*

Proof. (i) For the two centered size-n samples, the empirical means are 0 and $2/n$. Their exact ERM solutions are therefore $c_S = 0$ and $c_{S'} = e_1$ when $\gamma_n = 2B/n$. Both points lie in the fixed domain Θ, and the output laws are Dirac measures, so their $\mathcal{W}_p$ distance is one for every $p \geq 1$. The construction has loss Lipschitz constant $L_\ell = B + \beta D_\Theta$ and condition number $\chi_n = \beta n/(2B)$.

(ii) The map $\theta \mapsto \theta - \eta \nabla R_S(\theta)$ equals $c_S + (I - \eta H)(\theta - c_S)$, so $\theta_T - c_{\mathcal{D}_N} = (I - \eta H)^T(\theta_0 - c_{\mathcal{D}_N})$ and $u^{(i)} - c_{\setminus i} = (I - \eta H)^{T_u}(\theta_T - c_{\setminus i})$. Decomposing $\theta_T - c_{\setminus i} = (\theta_T - c_{\mathcal{D}_N}) + (c_{\mathcal{D}_N} - c_{\setminus i})$ and using the shared γ eigendirection gives equation 80. Exact Gaussian indistinguishability gives $s_{T_u}/\sigma_{\mathrm{pub}} \leq \mu_{\epsilon,\delta}$. Since $\mu_{\epsilon,\delta}$ is nondecreasing in ϵ, the imposed budget $\epsilon \leq 2B/(D_\Theta \gamma n)$ gives $s_{T_u}/\sigma_{\mathrm{pub}} \leq \mu_{2B/(D_\Theta \gamma n),\delta}$. Combining this with $\|\theta_0 - c_{\mathcal{D}_N}\| \geq D_\Theta/2$ gives $(1 - \eta\gamma)^{T+T_u} \leq 2\sigma_{\mathrm{pub}} \mu_{2B/(D_\Theta \gamma n),\delta}/D_\Theta$, and taking logarithms with $\eta = 1/\beta$ yields equation 81.

(iii) Along the γ-eigendirection the excess risk after t steps is exactly $\frac{\gamma}{2}(1 - \eta\gamma)^{2t}\|\theta_0 - c_{\setminus i}\|^2 = (1 - \eta\gamma)^{2t} e_0$, so the displayed T^{re} is an equality rather than an upper bound, and $\ln(1 - \eta\gamma)^{-1} = \Theta(1/\chi_n)$. □

Theorem P.1 (Restatement of Theorem 6)**.** *Fix $p \geq 1$ and a constant step size $\eta > 0$ independent of n, and let $\{(\mathcal{A}_n, \mathcal{U}_n)\}_{n \geq 2}$ be a family in which each $\mathcal{U}_n$ is an anchored unlearner (Def. 14) of depth $T_u = T_u(n)$ that is (ϵ_n, δ_n)-CU for $\mathcal{A}_n$ (Def. 4), and whose publication noise is an additive $\mathcal{N}(0, \sigma(n)^2 I_d)$ applied to both $\mathcal{U}_n$ and the retraining reference. The following are conditional obstructions for the stated loss families; in part (ii) the architecture is the rewind or descend construction of Lem. P.2.*

- ***(i)*** Nonsmooth convex. *$\mathcal{L}(\cdot, z)$ ranges over the convex L-Lipschitz family of Lem. P.1, which is not β-smooth for any finite β, in dimension $d \geq \min\{T_u, \eta^{-2}\}$, with $\Theta \supseteq B(\theta^{\mathrm{a}}, 4\eta L\sqrt{T_u})$. Quantitatively, equation 74 holds, so $\Delta_p^\star(\mathcal{U}_n, n+1) \geq c_1 \eta L$ for a universal $c_1 > 0$ and every n, and (C2) fails outright.*
- ***(ii)*** Expansive smooth nonconvex. *$\mathcal{L}(\theta, z) = -\frac{\beta}{2}\|\theta\|^2 + \langle b(z), \theta\rangle$ with $\|b(z)\| = L$ and $\eta \leq 1/\beta$, so that the retained-data recursion is expansive with $\mathcal{E} = 1 + \eta\beta > 1$. Then the sensitivity of the worst-case distance obeys equation 75 with equality for every $T_u \leq T$. Exact Gaussian noise setting and (C4) imply that any admissible scheme in this architecture must satisfy $T < T^\dagger$ with $T^\dagger$ as in equation 76, which is $O(\ln(nD_\Theta)/\ln(1 + \eta\beta))$; hence no such scheme exists once $T \geq T^\dagger$.*
- ***(iii)*** Ill-conditioned centered quadratic. *$\mathcal{L}_n$ is the bounded-domain family of Lem. P.3, with condition number $\chi_n = \beta/\gamma_n$. If $\chi_n = \Omega(n)$, exact ERM has the constant learner-stability floor (Def. 7) constructed in Lem. P.3(i), so (C1) and (C2) are incompatible. If $\chi_n = \Theta(n^a)$ with $a \in (0, 1)$, the interior trajectory and the output-perturbation certificate assumptions of Lem. P.3(ii) imply equation 77, namely $T + T_u = \Omega(n^a \ln n)$ at the target accuracy specified in that lemma.*

Proof. Part (i). Because the publication noise is a common additive translation applied to both branches, the two published laws are translates of the pair of deterministic iterates, so $\mathcal{W}_p(u^{(i)}, u^{(j)}) = \|\theta_{T_u}^{(i)} - \theta_{T_u}^{(j)}\|$ for every $p \geq 1$. Lem. P.1 then gives equation 74, and since $T_u \geq 1$,

$$\Delta_p^\star(\mathcal{U}_n, n+1) \;\geq\; c_1 \eta L \sqrt{T_u} \;\geq\; c_1 \eta L.$$

The lower bound $c_1 \eta L$ involves neither n nor the CU certificate, so it does not vanish at fixed η and condition (C2) already fails. This is the strongest conclusion among the three parts, and the assumptions (C1), (C3), and (C4) are not needed for this nonsmooth obstruction.

If $\eta = \eta(n) \to 0$, the Gaussian noise setting branch is outside the fixed-step statement above and requires the exact profile $\mu_{\epsilon,\delta}$; the mechanism-independent route remains the contrapositive bound. Finally, the lower bound $\tau_p^\star(\mathcal{A}_n, n) \geq c_1(\eta L\sqrt{T} \wedge D_\Theta)$ feeds equation 73 directly and gives a privacy-parameter lower bound independent of n. This is the mechanism-independent lower bound used in Theorem 6(i).

Part (ii). Lemma P.2 gives equation equation 75 with equality. The lower bound $\frac{2\eta L}{N}(1 + \eta\beta)^{T-1}$ does not depend on T_u, so additional rewind steps cannot reduce it. The descend construction is worse: its worst-case-distance sensitivity increases with T_u. Exact Gaussian noise setting gives

$s_{T_u}/\sigma \le \mu_{\epsilon,\delta}$. Imposing utility condition (C4), namely $\sigma\sqrt{d} \le c_0 D_\Theta$, gives $(1+\eta\beta)^{T-1} \le \frac{c_0 D_\Theta N \mu_{\epsilon,\delta}}{2\eta L\sqrt{d}}$, which is equation equation 76. Since $T^\dagger$ grows logarithmically with nD_Θ, the utility violation occurs after only logarithmically many training updates, before the projection Π_Θ becomes active in the construction of Lemma P.2; therefore equation equation 79 remains valid over the relevant range. The unlearning-stability term has the complementary bound from Theorem N.3, with $\mathcal{E} = 1+\eta\beta$,

$$\Delta \le \frac{2L}{\beta n}\big[(1+\eta\beta)^{T_u} - 1\big].$$

This upper bound exceeds the domain diameter D_Θ once $(1+\eta\beta)^{T_u} \ge 1 + \frac{\beta n D_\Theta}{2L}$ and then provides no useful control. The worst-case-distance term decreases with T_u, whereas the displayed unlearning-stability bound increases with T_u. The sum in Conversion Theorem 1(2) is therefore minimized at either the smallest or the largest admissible unlearning depth, and both endpoint values retain a factor of order $(1+\eta\beta)^T$. This proves the expansive-case obstruction in Theorem 6(ii).

Part (iiia). Lem. P.3(i) gives $\tau_p^\star(\mathcal{A}_n, n) \ge 1$ for the centered exact-ERM pair. Since $\chi_n = \Omega(n)$, the floor is independent of n, and Prop. 2 rules out simultaneous vanishing budget and unlearning stability.

Part (iiib). For $\chi_n = \Theta(n^a)$, Lem. P.3(ii) gives equation 81 and hence equation 77; the update-count lower bound in equation 77 is $\Omega(n^a \ln n)$ at the target empirical error e used in Lem. P.3(ii). The conclusion is learner-specific and uses the interior trajectory and the corresponding retraining reference assumptions in that lemma. □

Q Proofs for the Self-Referential Analysis

This appendix applies the Conversion Theorem in two directions. First, it shows that a differentially private learner can be viewed as an identity unlearner, which recovers the usual privacy-to-stability implication. Second, it compares the privacy budget required for a training-level output law with the smaller worst-case distance between an unlearned output law and its retrained reference. Finally, it iterates the conversion inequality over a stream of deletion requests and shows that the resulting stream expression is an upper bound on learner stability; the stream expression does not give a lower bound on excess risk.

Remark 6 (Comparison with DP-SCO lower bounds.)**.** *The contribution $L\,P^{\mathrm{TV}}(\epsilon,\delta)$ from the CU-implied distance and the DP-SCO lower bound $\Omega(LD_\Theta\sqrt{d\ln(1/\delta)}/(\epsilon n))$ of (Bassily et al., 2019) measure different quantities. After localization (Remark 2), both contain a $\sqrt{d}$ factor. The direct endpoint-distance term $2Lr_{T_u,p}$, based on the distance from an unlearned output law to its corresponding retrained output law, avoids this factor under a shared transition rule. The two terms also have different dependence on n and ϵ.*

Remark 7 (Relation to Thm. 6.)**.** *Theorem 6 concerns existence, whereas Sec. 3.2 concerns cost. Some algorithms cannot satisfy the CU conditions. For algorithms that do satisfy them, each deletion invocation still contributes the direct endpoint term $r_{T_u,p}$ at fixed certificate budget.*

Algorithm 1: Two-point extremal learner for tightness of the conversion

1: **Input:** dataset $\mathcal{D}_N$, distinguished sample $z^\star$, parameters $\epsilon \ge 0$, $\delta \in [0, 1/2]$, two points $\theta_+, \theta_- \in \Theta$ with $\|\theta_+ - \theta_-\| = D_\Theta$
2: Define

$$q_+ := \delta + (1-\delta)\frac{e^\epsilon}{e^\epsilon+1}, \qquad q_- := (1-\delta)\frac{1}{e^\epsilon+1}$$

3: **if** $z^\star \in \mathcal{D}_N$ **then**
4: sample $\theta \sim q_+\delta_{\theta_+} + (1-q_+)\delta_{\theta_-}$
5: **else**
6: sample $\theta \sim q_-\delta_{\theta_+} + (1-q_-)\delta_{\theta_-}$
7: **end if**
8: **return** θ

Theorem Q.1 (Differential Privacy Is the Degenerate Unlearner). *Let $p \geq 1$, let $\operatorname{diam}(\Theta) \leq D_\Theta < \infty$, and let $\mathcal{A}$ be (ϵ, δ)-differentially private with respect to the removal of one record.*

(i) The identity unlearner $\mathcal{U}_0(\mathcal{A}(\mathcal{D}), \mathcal{D}_f) := \mathcal{A}(\mathcal{D})$ is (ϵ, δ)-CU for $\mathcal{A}$ (Def. 4) and is $(0, N)$-$\mathcal{W}_p$-USU (Def. 10). Consequently $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-USL with

$$\tau \;\leq\; 2r_{T_u,p}. \tag{82}$$

(ii) Assume instead the hypotheses of Thm. N.1: $\mathcal{L}(\cdot, z)$ is γ-strongly convex, L-Lipschitz and β-smooth, and $\mathcal{U}$ is a full-batch anchored unlearner of depth T_u and step $\eta \leq \frac{2}{\beta+\gamma}$ whose certificate is issued by the Gaussian mechanism at sensitivity of the worst-case distance to retraining with noise σ. Then $\mathcal{A}$ is (τ, n)-$\mathcal{W}_p$-USL with

$$\tau \;\leq\; \frac{2L}{\gamma n}\big(1 - (1 - \eta\gamma)^{T_u}\big) + 2r_{T_u,p}, \tag{83}$$

where $\epsilon(T_u, \sigma) = \frac{2L}{\gamma N}(1 - \eta\gamma)^{T_u}\sqrt{2\ln(1.25/\delta)}/\sigma$.

The two bounds use different probability-law comparisons. Equation equation 82 uses the training-level distance between the full-data output law and the retrained output law; Theorem Q.2 shows that this contribution obtained from the CU inequalities cannot generally be reduced. Equation equation 83 uses the worst-case distance between each unlearned output law and its corresponding retrained output law; Theorem N.4 shows that this distance can be $O(1/n)$ under the shared contractive update assumptions.

Proof. (i) Instantiate Def. 4 with $\mathcal{U} = \mathcal{U}_0$ and $|\mathcal{D}_f| = 1$. Since $\mathcal{U}_0(\mathcal{A}(\mathcal{D}), \mathcal{D}_f) = \mathcal{A}(\mathcal{D})$, the two required inequalities read $\Pr[\mathcal{A}(\mathcal{D}) \in O] \leq e^\epsilon \Pr[\mathcal{A}(\mathcal{D}\backslash\mathcal{D}_f) \in O] + \delta$ and its reverse, which is exactly the differential privacy of $\mathcal{A}$ at the removal of one record. For the unlearning stability, $u^{(i)} = \mathcal{A}(\mathcal{D}_N)$ for every i, so the two laws compared in Def. 10 are identical and $\mathcal{W}_p(u^{(i)}, u^{(j)}) = 0$, the identity map is the $T_u = 0$ member of Def. 14. Conversion Theorem 1(2) with $\Delta_1 = 0$ gives equation 82.

(ii) Theorem N.1 gives $\Delta \leq \frac{2L}{\gamma n}(1 - (1 - \eta\gamma)^{T_u})$ and $\tau \leq \Delta + 2r_{T_u,p}$. It remains to identify which ϵ is admissible. The certificate compares $u^{(i)}$ with $\mathcal{A}(\mathcal{D}_N^{-i})$, so its sensitivity is the sensitivity of the worst-case distance to retraining in Lem. L.2(i), namely $s_{T_u} \leq (1 - \eta\gamma)^{T_u}\frac{2L}{\gamma N}$ when both runs start from a common initialization, and Lem. L.1 converts it into $\epsilon(T_u, \sigma) = s_{T_u}\sqrt{2\ln(1.25/\delta)}/\sigma$. Substituting gives the stated profile and hence equation 83. □

Theorem Q.2 (Restatement of Theorem 2). *Let Θ satisfy $\operatorname{diam}(\Theta) = D_\Theta < \infty$ with the diameter attained by some pair of points, let $p \geq 1$, $\epsilon \geq 0$ and $\delta \in [0, 1]$ with $C_{\epsilon,\delta} \leq 1$. There exist a randomized learner $\mathcal{A} : \mathcal{Z}^N \mapsto \mathcal{M}(\Theta)$ and an unlearner $\mathcal{U}$ such that $\mathcal{U}$ is (ϵ, δ)-CU for $\mathcal{A}$ (Def. 4), $\mathcal{U}$ is $(0, N)$-$\mathcal{W}_p$-USU (Def. 10), and For the constructed pair, define $P^{\mathrm{TV}}(\epsilon, \delta) := 2D_\Theta C_{\epsilon,\delta}^{1/p}$.*

$$\tau_p^\star(\mathcal{A}, n) \;=\; P^{\mathrm{TV}}(\epsilon, \delta)/2 \qquad \textit{for every } n \geq 2. \tag{84}$$

Consequently, if φ is any function such that $\tau_p^\star(\mathcal{A}, n) \leq \Delta_p^\star(\mathcal{U}, n+1) + \varphi(\epsilon, \delta, D_\Theta, p)$ holds for every pair $(\mathcal{A}, \mathcal{U})$ satisfying Def. 4, then $\varphi(\epsilon, \delta, D_\Theta, p) \geq P^{\mathrm{TV}}(\epsilon, \delta)/2$. The CU-implied distance $P^{\mathrm{TV}}(\epsilon, \delta)$ is therefore optimal within a factor of two under Def. 4 alone.

Proof. Fix $\theta_+, \theta_- \in \Theta$ with $\|\theta_+ - \theta_-\| = D_\Theta$, fix a distinguished record $z^\star \in \mathcal{Z}$, and set

$$q_+ := \delta + (1 - \delta)\frac{e^\epsilon}{e^\epsilon + 1}, \qquad q_- := (1 - \delta)\frac{1}{e^\epsilon + 1},$$

the two masses that attain equality in Lemma I.2; they satisfy $q_+ - q_- = C_{\epsilon,\delta}$ and $q_+ + q_- = 1$. Define the randomized algorithm

$$\mathcal{A}(S) := \begin{cases} \theta_+ \text{ w.p. } q_+,\ \theta_- \text{ w.p. } q_-, & z^\star \in S, \\ \theta_+ \text{ w.p. } q_-,\ \theta_- \text{ w.p. } q_+, & z^\star \notin S, \end{cases}$$

and take $\mathcal{U} = \mathcal{U}_0$, the identity unlearner. We verify (a) the CU inequalities, (b) unlearning stability, and (c) learner stability.

The certificate. Θ carries only the two atoms in the range of $\mathcal{A}$, so it suffices to check Def. 4 on $O \in \{\{\theta_+\}, \{\theta_-\}\}$. Writing μ and ν for the laws of $\mathcal{A}(\mathcal{D}_N)$ and $\mathcal{A}(\mathcal{D}_N^{-i})$ with $z_i = z^\star$, we have $\mu(\{\theta_+\}) = q_+ = e^\epsilon q_- + \delta = e^\epsilon \nu(\{\theta_+\}) + \delta$ by the definition of $q_\pm$, and symmetrically on $\{\theta_-\}$, the two remaining inequalities are strict. Hence $\mathcal{U}_0$ is (ϵ, δ)-CU for $\mathcal{A}$, with equality in both constraints, so no smaller budget certifies this pair. If $z_i \neq z^\star$ the two laws coincide and the constraints are trivial.

The unlearning stability (Def. 10). $u^{(i)} = \mathcal{A}(\mathcal{D}_N)$ for every i, so $\Delta_p^\star(\mathcal{U}_0, N) = 0$.

The learning stability. Let $\mathcal{D}_n$ contain $z^\star$ in slot i and let $\mathcal{D}_n^{(i)}$ replace it by any $z' \neq z^\star$. The two output laws are μ and ν above. A pairing of two laws supported on $\{\theta_-, \theta_+\}$ is determined by the mass it moves between the atoms, and it must move at least $|q_+ - q_-| = C_{\epsilon,\delta}$, each unit of moved mass travels the distance D_Θ. Hence $\mathcal{W}_p^p(\mu, \nu) = D_\Theta^p C_{\epsilon,\delta}$ exactly, and taking the supremum over neighbouring pairs gives equation 84, the supremum being attained. Since $C_{\epsilon,\delta} \leq 1$ by hypothesis, $\theta_\pm \in \Theta$ and no clipping is needed.

If a universal inequality $\tau \leq \Delta + \varphi$ held, the constructed pair would force $\varphi \geq P^{\mathrm{TV}}(\epsilon, \delta)/2$. Because the construction uses only two parameter values, it applies in every ambient dimension, uniformly in n and T. □

Proposition Q.1 (Restatement of Proposition 1)**.** *For every $p \geq 1$, $s > 0$, $\epsilon \in (0, 1]$, and $\delta \in (0, 1/2)$, let*

$$K_\sigma(x) := x + \sigma Z, \qquad Z \sim \mathcal{N}(0, 1), \qquad \sigma \geq \frac{s\sqrt{2\ln(1.25/\delta)}}{\epsilon}.$$

There exist a learner family and an anchored unlearner with deterministic core $u_0 \equiv 0$, $T_u = 0$ (or zero retained-data updates), followed by one common data-independent transition rule K_σ, such that the unlearner is (ϵ, δ)-CU, $\Delta_p^\star = 0$, $r_{T_u,p} = s$, and

$$\tau_p^\star = 2s = \Delta_p^\star + 2r_{T_u,p}. \tag{85}$$

Thus the coefficient 2 in the conversion based on the worst-case distance to retraining is exact, for every $p \geq 1$.

Proof. Let $\Theta = \mathbb{R}$ and choose two neighbouring retained datasets S_-, S_+ that differ in one record. Define the retraining learner on the two marker datasets by

$$\mathcal{A}(S_-) = K_\sigma(-s), \qquad \mathcal{A}(S_+) = K_\sigma(+s),$$

and set $\mathcal{A}(S) = K_\sigma(0)$ for every other retained dataset. For every full dataset and every deletion index, let the anchored unlearner ignore the index, use anchor 0, perform no retained-data update, and apply the single publication mechanism $u^{(i)} = K_\sigma(0)$. Each retrained reference law has mean in $[-s, s]$, so the Gaussian shift profile proves the stated (ϵ, δ)-CU inequalities. Because $u^{(i)}$ is independent of i, its unlearning-stability parameter is zero. The common additive Gaussian transition rule is an isometry for Wasserstein distance, so $\mathcal{W}_p(K_\sigma(0), K_\sigma(\pm s)) = s$ and $r_{T_u,p} = s$.

For the marker pair, comparing the two outputs with the same Gaussian noise gives $\mathcal{W}_p \leq 2s$. Conversely, for any pairing (X, Y), Jensen's inequality gives $\mathbb{E}|X - Y|^p \geq |\mathbb{E}(X - Y)|^p = (2s)^p$, because the two means differ by $2s$. Hence $\mathcal{W}_p(K_\sigma(-s), K_\sigma(+s)) = 2s$, which proves equation 85. □

Corollary Q.1 (Compounding under Repeated Deletion.)**.** *In the setting of Thm. N.1, consider a stream of deletion requests in which round k re-anchors at the model published in round $k-1$, runs T_u steps of equation 20 on the current retained set, and publishes with a fresh certificate. Let $r_{k,p}$ be the round-k radius for the worst-case distance to retraining, defined as in Section 2; set $P_{k,p} := 2r_{k,p}$, write τ_k for the $\mathcal{W}_p$-USL parameter of the round-k output, and $c := (1 - \eta\gamma)^{T_u} \in (0, 1)$. Then*

$$\tau_k \;\leq\; c\,\tau_{k-1} + \frac{2L}{\gamma n}(1 - c) + P_{k,p}, \tag{86}$$

and hence

$$\tau_k \le c^k D_\Theta + \frac{2L}{\gamma n}(1-c^k) + \sum_{\ell=1}^{k} c^{k-\ell} P_{\ell,p}.$$

If $P_{\ell,p} \le P_{\max,p}$ for all ℓ, then

$$\limsup_{k\to\infty} \tau_k \;\le\; \frac{2L}{\gamma n} + \frac{P_{\max,p}}{1-(1-\eta\gamma)^{T_u}}. \tag{87}$$

The displayed limsup is an upper envelope under a uniform round-wise bound on the worst-case distance to retraining; for constant $P_{k,p} = P$ it reduces to the corresponding geometric-series expression. The amplification factor is $\big(1-(1-\eta\gamma)^{T_u}\big)^{-1}$, which at $\eta = 1/\beta$ and $T_u = 1$ equals the condition number β/γ and at $T_u \ge 1/(\eta\gamma)$ is at most $(1-e^{-1})^{-1} < 1.6$.

Proof. Compared with Proposition N.1, the repeated-deletion proof changes the initial distance between the two runs. In that proposition, the discrepancy obeys $\Delta_{m+1} \le \mathcal{E}\Delta_m + \varsigma + \eta J_m$ with $J_m \le 2L/n$ for full-batch gradients, and Lem. N.1(i) gives $(\mathcal{E}, \varsigma) = (1-\eta\gamma, 0)$ under the contractive hypotheses of Theorem N.1. At round k the two runs are anchored at the two round-$(k-1)$ publications rather than at a common point, so $\Delta_0 = \tau_{k-1}$ instead of 0, and unrolling T_u steps gives

$$\Delta \le (1-\eta\gamma)^{T_u}\tau_{k-1} + \frac{2\eta L}{n}\sum_{m<T_u}(1-\eta\gamma)^m = c\,\tau_{k-1} + \frac{2L}{\gamma n}(1-c).$$

Conversion Theorem 1(2), applied afresh because the round-k publication carries its own certificate, adds the round-dependent $P_{k,p}$ and yields equation 86. Iterating the affine inequality gives the displayed weighted sum. Since $\tau_0 \le \operatorname{diam}(\Theta) = D_\Theta$ and the weighted sum of the term based on the worst-case distance to retraining is at most $P_{\max,p}(1-c^k)/(1-c)$ under the uniform bound, substituting these two bounds into the weighted recursion gives the displayed limit. At $\eta = 1/\beta$ and $T_u = 1$, $1-c = \eta\gamma = \gamma/\beta = 1/\kappa$. □

Proposition Q.2 (The Stream Upper Bound Is Not an Excess-Risk Lower Bound.)**.** *The stream expression in equation 87 is an upper bound on learner stability, not a lower bound. In particular, no lower bound on the excess risk of DP-SGD that stays positive uniformly in n at fixed (ϵ,δ) follows from this expression.*

Proof. (i) The recursion equation 86 is a chain of upper bounds. When the round-wise worst-case distance is uniformly bounded, the geometric envelope in equation 87 follows from the affine recursion; it is not a lower bound. No lower bound on τ follows, and none can, since the trivial learner $\mathcal{A} \equiv \theta_0$ is 0-differentially private and 0-stable while satisfying every hypothesis of Theorem N.1.

(ii) In equation 83, the budget certifies $u^{(i)}$ against $\mathcal{A}(\mathcal{D}_N^{-i})$, the model retrained without the deleted record. Theorem N.4 gives the direct bound $2r_{T_u,p} \le 2s_{T_u}$ and, through the CU-implied distance $P^{\mathrm{TV}}(\epsilon,\delta) \le \frac{2LD_\Theta\sqrt{2\ln(1.25/\delta)}}{\gamma\sigma n}(1-\eta\gamma)^{T_u} + 2D_\Theta\delta$, which is equation 48. Thus at $p=1$ and $\delta = O(1/n)$ both equation 83 and $\tau^\star$ are $O(1/n)$; for every p, the direct term inherits the factor from the worst-case distance to retraining $(1-\eta\gamma)^{T_u}$. The larger constant obtained by using the training-level budget in Proposition Q.1(i) certifies the full-data output law against the retrained output law, whereas the budget based on the worst-case distance certifies each unlearned output law against its corresponding retrained output law.

(iii) Suppose a bound $\mathbb{E}[R(\theta^{\mathrm{out}})] - R^\star \ge c_0 > 0$ held for every n at fixed (ϵ,δ) for some (ϵ,δ)-differentially private learner on γ-strongly convex L-Lipschitz losses. The algorithms of (Bassily et al., 2014; 2019) achieve excess population risk of order $\frac{L^2}{\gamma n} + \frac{L^2 d\ln(1/\delta)}{\gamma\epsilon^2 n^2}$ up to logarithmic factors, which is below c_0 for all n large enough, a contradiction. The displayed excess-risk rates show that the two terms of equation 58 behave differently in the limit: the statistical term $\frac{2L^2}{\gamma n}$ and the optimization term both vanish, and so does the round-wise direct cost $LP_{k,p}$ whenever sensitivities of the worst-case distance vanish. Under a fixed training-level budget the CU-implied distance remains constant, while the direct term based on the worst-case distance to retraining decays with the sensitivity of the worst-case distance to retraining; Thm. 2 and Prop. 1 give the corresponding sharpness statements. □

Loss class	Our excess risk	vs. classical (when better / by how much)
Contractive.		
A strongly convex, Lip., smooth	$\frac{2L^2}{\gamma n} + 2Lr_{T_u,p} + O\big(\frac{L^2 \ln T}{\gamma T}\big)$	Matches Hardt/Zhu rate $O(L^2/(\gamma n))$ (Hardt et al., 2016; Zhu et al., 2023). *Not* a sharper leading term. Gain: same rate *without* learner ergodicity (Zhu et al., 2023); the direct contribution from the worst-case distance to retraining is $2Lr_{T_u,p} = O(L^2/(\gamma n))$, while only the localized CU-implied distance can add $L\,P^{\mathrm{TV}}(\epsilon,\delta) = O(L^2\sqrt{d\ln(1/\delta)}/(\gamma n))$ (Rem. 2), so the sum stays $O(1/n)$. Fixed training-level budget makes $P^{\mathrm{TV}}(\epsilon,\delta)$ constant and can dominate.
E PL, $\mathcal{L} \geq 0$, dispersion equation 63	$\frac{32c_\Lambda\beta(R^\star+\epsilon_{\mathrm{opt}})}{\mu_{\mathrm{PL}}n} + O(r_{T_u,p}) + \epsilon_{\mathrm{opt}}$, $\tilde{O}(n^{-2})$ if $R^\star = 0$	Matches fast PL rates of (Lei & Ying, 2021; Zhu et al., 2025) under interpolation. Gain: no minimizer stability (Charles & Papailiopoulos, 2018). Without dispersion: only $O(1/(\mu_{\mathrm{PL}}n))$ (Rem. 5), same order as the slow classical baseline.
C dissipative, smooth, noisy	$\frac{2L^2}{\kappa n} + 2Lr_{T_u,p} + \epsilon_{\mathrm{opt}}$	Matches Zhu $\mathcal{W}_1$ rate $O(1/n)$ (Zhu et al., 2023). Gain. The population-risk proof needs contractivity of the *unlearning* transition rule only; it does not require learner ergodicity or Lyapunov stationarity. Leading term not sharper, $+2r_{T_u,p}$ as in A.
D LSI, $\lvert\mathcal{L}\rvert \leq \rho$, ϵ-PCU	$O\big(n^{-1} + \sqrt{\epsilon/n}\big) + \epsilon_{\mathrm{opt}}$	Matches Dadi KL→risk rate $O(1/n)$ at small ϵ (Dadi & Cevher, 2025). Gain. The KL population-risk bound has no learner stationarity requirement and remains valid on an unbounded parameter space Θ. Under (ϵ,δ)-CU with $\delta > 0$, Pinsker gives an $O(n^{-1/2})$ statistical term in equation 62, which is slower than Dadi's $O(1/n)$ rate.
Non-contractive.		
B convex, Lip., smooth	$\frac{2LD_\Theta T_u}{n\sqrt{T}} + 2Lr_{T_u,p} + \frac{2LD_\Theta}{\sqrt{T}}$	Hardt with $\eta = D_\Theta/(L\sqrt{T})$ pays gen. term $\Theta(LD_\Theta\sqrt{T}/n)$ (Hardt et al., 2016). Our learner-stability contribution is $\Theta(LD_\Theta T_u/(n\sqrt{T}))$, whereas the classical contribution is $\Theta(LD_\Theta\sqrt{T}/n)$. Therefore the statistical term improves by a factor $\Theta(T/T_u)$ whenever $T_u \ll T$. Opt. term identical $\Theta(LD_\Theta/\sqrt{T})$. Net win only if the saved gen. term exceeds the worst-case distance to retraining bound $2r_{T_u,p}$ (interior T_u, Table 1).
F smooth, no Lip. bound	$\mathbb{E}[R_{\mathcal{D}_n}] + \beta\big(\frac{2\eta T_u\sqrt{2\beta\Lambda}}{n} + 2r_{T_u,p}\big)^2 + \epsilon_{\mathrm{opt}}$	Kuzborskij/Lei path bounds use loss-at-initialization surrogates (Kuzborskij & Lampert, 2018; Lei & Ying, 2020). Our bound evaluates Λ at the trained anchor. If the trained-anchor loss is below the initialization loss, the squared path term is smaller by the factor $\Lambda_{\mathrm{anchor}}/\Lambda_{\mathrm{init}}$ inside the square. As in B, $2r_{T_u,p}$ is not absorbed.

Table O.1: Excess population risk for Combinations A–F, with an explicit comparison to the classical baselines of Tables E.1–E.3. "Matches" means the same n-rate for the statistical term, quantitative gains are either a factor in T/T_u (row B), a smaller path surrogate (row F), a fast $\tilde{O}(n^{-2})$ under PL interpolation (row E), or different conditional hypotheses at the same order (rows A,C,D). The direct shared-transition-rule certificate is dimension-free; only the TV CU-implied distance may add $\sqrt{d}$ (Remark 2).