CERTIFIED UNLEARNING FOR DISTILLED TEACHER-STUDENT MODELS WITH A BLACK-BOX TEACHER

Anonymous authors
Paper under double-blind review

ABSTRACT

We study certified machine unlearning for student models distilled from a teacher accessible only through black-box prediction queries. After deployment, the student must unlearn requested samples, while the teacher is unavailable for retraining and accessible only through black-box prediction queries. We introduce a student-side certification framework that re-distills the student on the retained data using only black-box teacher access, and releases a randomized model by carefully calibrating the perturbation to certify indistinguishability against the ideal but counterfactual retrain-from-scratch reference, where the latter would require retraining the (inaccessible) teacher without the forget samples. Our certificate builds on a student sensitivity bound for distillation, which we obtain by characterizing how the teacher's predictions would change under retraining without the forget samples, and then propagating this bound through the distillation objective to bound the student's parameter shift. The final student sensitivity is used to perturb the released student parameters with calibrated Gaussian noise to ensure the target certification guarantees. In doing so, our work provides a first study of unlearning certification for black-box teacher-student distillation. While our work is primarily theoretical and relies on several regularity assumptions for tractability, we hope our results will provide interesting insights for unlearning in this increasingly common regime.

1 INTRODUCTION

Machine unlearning studies how to efficiently remove the influence of designated training samples from an already trained model, often due to privacy or copyright requirements (GDPR, 2016; CCPA, 2018; CPRA, 2021). At a minimum, these regulations require that, after a deletion request, the released model should be indistinguishable from a model that has been trained only on the retained data. A standard formalization of this objective is the *retraining* reference where, given a dataset and a set of samples to delete, the reference model is obtained by retraining from scratch on the retained data. While this provides a clear target for unlearning, it is also infeasible to perform retraining after each deletion request. Certified unlearning aims to address this challenge by developing efficient unlearning mechanisms, while providing formal data removal guarantees in the form of statistical indistinguishability between the unlearned and retrained-from-scratch models (Guo et al., 2020; Sekhari et al., 2021; Neel et al., 2021; Chien et al., 2024).

Existing certified unlearning formulations are designed for single-stage training pipelines, where the retraining-from-scratch reference is well-defined. On the other hand, deployed models in practice are often trained not in isolation, but are produced through multi-stage pipelines in which compact downstream (student) models inherit information from larger upstream (teacher) models. On-device models, for instance, are typically distilled from larger proprietary models that cannot be stored, retrained, or updated during unlearning, and are only accessible through (often budgeted) black-box prediction queries. In such settings, unlearning must account for the entire training pipeline, as certified unlearning methods developed for single stage training pipelines may fail to remove the information that persists through the upstream model.

In this work, we study this problem for knowledge distillation with black-box teacher access, where the teacher parameters are unavailable, but the teacher can be queried. In a typical distillation setting, the teacher is trained first on the original dataset, and the student is trained using a distillation loss to match the teacher's predictions (Hinton et al., 2015). In the black-box setting, the teacher is only

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.

accessible through prediction queries. The goal is to unlearn designated training samples from the student model after deployment. For this problem, the ideal condition is *retrain-from-scratch* which retrains the teacher on retained data and then redistills the student using this updated teacher. While correct, this is computationally expensive and infeasible under black-box teacher access. A cheaper alternative is *redistill-on-retain-data* using the original teacher, which redistills the student on retained data only but by using the original teacher. While in this case the student never observes the forget samples directly, it can still preserve deleted-sample information through the teacher's predictions.

We illustrate this on a simple example in Fig. 1, where we consider a classification task with three well-separated clusters, and demonstrate the student predictions for each sample. Fig. 1 (a) shows the original student, where both the student and teacher are trained on the full dataset. Suppose we wish to forget the samples belonging to class 3 from the student model. In the following, we show the predictions for these samples in red (for evaluation only, not used during student training). For a fair comparison, in each case we allow the predictor to choose from one of the three classes. Fig. 1 (b) shows the predictions for the correct *retrain-from-scratch* reference, where both the teacher and student are retrained on the retained data only (classes 1 and 2). In comparison, Fig. 1 (c) demonstrates *redistill-on-retain-data*, where the student is distilled on the retained data (classes 1-2), however by using the original teacher (which was trained on all 3 classes). We observe that in this case, although the student never directly sees the deleted samples (class 3), the teacher's soft predictions still enable correct classification for the deleted class, hence retraining the student alone is not sufficient by itself for unlearning. These observations are also consistent with prior findings on memorization (Jagielski et al., 2023; Zhang et al., 2025; Dankers & Raunak, 2025). The details of this example are provided in App. F.

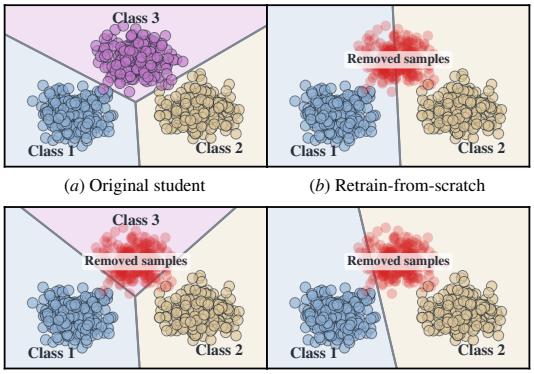


Figure 1: **Illustrative example** with three classes, where Class 3 is to be unlearned. (a) Original model trained on all three classes. (b) Teacher and student retrained from scratch on the retained data only. (c) Student re-distilled on the retained data using the original teacher. (d) CKDU (our method): student distilled on the retained data using the original teacher, released after perturbation.

In contrast, our approach CKDU (Certified Knowledge Distillation Unlearning) redistills the student with calibrated randomization, with the goal of indistinguishability between the released student and the retrain-from-scratch reference. Our approach is simple. Since teacher parameters cannot be observed or modified under black-box access, CKDU first performs redistill-on-retain-data producing an intermediate distilled model, and then releases the student after adding calibrated Gaussian noise to mask the mismatch induced by using the original teacher rather than the retrained-from-scratch teacher. The key intuition is that removing samples from the teacher would alter the teacher predictions used for student supervision in the ideal setting (retrained-from-scratch teacher and student). Since retraining the teacher is not possible in the black-box setting, we instead use stability arguments to bound this counterfactual prediction shift, propagate it through the student's distillation objective to bound the resulting student-parameter mismatch, and release a randomized version of the intermediate student with Gaussian noise calibrated to this bound. In Fig. 1 CKDU corresponds to panel (d), which closely aligns with the ideal retraining reference in Fig. 1(b).

To the best of our knowledge, CKDU provides a first study toward student-side certified unlearning in distillation pipelines under black-box teacher access. Our main contributions are the formal characterization of the teacher influence on the student model before and after unlearning, and formalizing machine unlearning under black-box teacher influence. While our contributions are primarily theoretical and rely on several regularity conditions for tractability, we hope our results will provide insights and primitives for more complex multi-stage settings. Our contributions can be summarized as follows.

- We formalize student-side certified unlearning for distilled models under black-box teacher access. The teacher can only be queried for predictions hence the retraining reference is a counterfactual that requires both teacher retraining and student re-distillation.

- We develop a theoretical sensitivity analysis that propagates shifts in the teacher-output predictions due to data deletion through the student training trajectory. Our main result applies to smooth non-convex student objectives, while stronger curvature conditions lead to tighter sensitivity guarantees.
- We provide CKDU, a student-side noise calibration rule for the perturbation required to achieve target certified unlearning guarantees with respect to (counterfactual) retraining-from-scratch both the teacher and student.
- We evaluate CKDU across multiple ablation studies and demonstrate that our approach closely matches the retraining reference on both forgetting and utility metrics.

2 Related Works

Machine unlearning can be broadly categorized as exact, heuristic, and certified. Exact unlearning aims to recover the model retrained from scratch under the same training procedure, including update-efficient formulations and partitioned pipelines such as SISA (Cao & Yang, 2015; Bourtoule et al., 2021), while heuristic methods approximate retraining via lightweight post-hoc updates (Golatkar et al., 2020; Fan et al., 2023; Kurmanji et al., 2023). Recently, PURGE has extended SISA to the distillation setting Quan et al. (2025). Unlike PURGE, we study the black-box teacher setting where the teacher is only accessible through black-box queries and cannot be re-trained during unlearning.

Certified unlearning provides unlearning guarantees between the unlearned and retrained-from-scratch models in the form of statistical indistinguishability (Guo et al., 2020; Sekhari et al., 2021; Neel et al., 2021; Liu et al., 2023; Zhang et al., 2024a; Allouah et al., 2024; Georgiev et al., 2024; Chien et al., 2024; Koloskova et al., 2025; Lu et al., 2025; Sepahvand et al., 2025; Mu & Klabjan, 2025). Existing approaches largely consider single-stage supervised training. Several works study unlearning with limited information, e.g., without access to retain data (Cha et al., 2024; Chen et al., 2024; Foster et al., 2024; Ahmed et al., 2025; Basaran et al., 2025), or after pruning or quantization (Jia et al., 2023; Zhang et al., 2024b). Other works use knowledge distillation as a tool for unlearning through teacher design (Kurmanji et al., 2023; Chundawat et al., 2023). In contrast, we study certified unlearning for models trained via knowledge distillation, without control over the teacher.

Knowledge distillation (Hinton et al., 2015) is widely used for compression and knowledge transfer. Recent works show that students can inherit memorization and privacy risks from their teachers (Jagielski et al., 2023; Ojha et al., 2023; Dankers & Raunak, 2025; Zhang et al., 2025). Motivated by these findings, we explore certified unlearning for distillation, to provide formal guarantees on removing training data influence from the student model.

3 Problem formulation

We study student-side certified unlearning under black-box teacher access in a teacher-student distillation setup (Hinton et al., 2015). A teacher is trained on a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \{1, \dots, K\}$, where $\mathcal{X}$ is the input space and K is the number of classes. The teacher is a predictor $f_{\boldsymbol{\phi}} : \mathcal{X} \to \mathcal{P}_K$ parameterized by $\boldsymbol{\phi} \in \mathcal{H}_T \subseteq \mathbb{R}^{d_T}$, where $\mathcal{P}_K = \{p \in [0,1]^K : \sum_{k=1}^K p_k = 1\}$ and $\mathcal{H}_T$ is the hypothesis space for the teacher. The teacher is trained on D by a learning algorithm $\mathcal{A}_T$ that outputs parameters $\boldsymbol{\phi}^* = \mathcal{A}_T(\mathcal{D})$, and we denote the trained teacher as $f_{\boldsymbol{\phi}^*}$.

The student is trained using the teacher as an additional source of supervision, typically in a setting where the teacher is more expressive than the student, and both use the same training inputs. The student is a predictor $g_{\boldsymbol{w}} : \mathcal{X} \to \mathcal{P}_K$ parameterized by $\boldsymbol{w} \in \mathcal{H}_S \subseteq \mathbb{R}^{d_S}$, where $\mathcal{H}_S$ is the hypothesis space for the student. Let $\mathcal{A}_S$ denote the student training algorithm for distillation. The student is obtained by minimizing an objective that combines supervised cross-entropy with a distillation term,

$$\mathcal{L}(\boldsymbol{w}; f_{\boldsymbol{\phi}^*}) := \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \left(-(1-\lambda)\log g_{\boldsymbol{w}}(\boldsymbol{x})_y + \lambda\, D_{\mathrm{KL}}(f_{\boldsymbol{\phi}^*}(\boldsymbol{x}) \parallel g_{\boldsymbol{w}}(\boldsymbol{x}))\right), \ \boldsymbol{w}^* := \mathcal{A}_S(\mathcal{D}, f_{\boldsymbol{\phi}^*}) \tag{1}$$

where $\lambda \in [0,1]$ is a weighting parameter, $-\log g_{\boldsymbol{w}}(\boldsymbol{x})_y$ denotes the cross-entropy loss for the true label y, and the Kullback-Leibler (KL) divergence term enforces alignment between the student and teacher output distributions.

An unlearning request specifies a subset $\mathcal{D}_u \subseteq \mathcal{D}$ of m training samples to be forgotten from the student model. The corresponding retained dataset is $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_u$ of size $|\mathcal{D}_r| = n - m$. During unlearning, only the student parameters can be updated, whereas the teacher is fixed and accessible only through black-box predictions. In this setting, the ideal retraining-from-scratch pipeline requires first retraining the teacher on the retained dataset $\mathcal{D}_r$, which produces a retrained teacher $\boldsymbol{\phi}_r^* = \mathcal{A}_T(\mathcal{D}_r)$, and then distilling the student on $\mathcal{D}_r$ using the retrained teacher, which produces a retrained student $\boldsymbol{w}_r^* = \mathcal{A}_S(\mathcal{D}_r, f_{\boldsymbol{\phi}_r^*})$. Accordingly, we treat $(\boldsymbol{\phi}_r^*, \boldsymbol{w}_r^*)$ as the ideal *retrain-from-scratch* reference for training without the forget samples $\mathcal{D}_u$. Our goal is to design student-side unlearning mechanisms that approximate the target student predictor without having access to the teacher parameters. Formally, given $\mathcal{D}$, $\mathcal{D}_u$, the learning algorithm $\mathcal{A}_S$ and query access to the teacher $f_{\boldsymbol{\phi}^*}$, our goal is to design a certified unlearning mechanism $\mathcal{U}$ that outputs unlearned student parameters $\widetilde{\boldsymbol{w}}_r := \mathcal{U}(\mathcal{A}_S(\mathcal{D}, f_{\boldsymbol{\phi}^*}), \mathcal{D}, \mathcal{D}_u)$, where the mechanism has access only to the black-box teacher predictions $f_{\boldsymbol{\phi}^*}$ but not the parameters $\boldsymbol{\phi}^*$. Our unlearning definition is adapted from (Koloskova et al., 2025, Definition 2.1) to knowledge distillation.

Definition 3.1 ((ϵ, δ)-unlearning for knowledge distillation)**.** An unlearning mechanism $\mathcal{U}$ provides (ε, δ)-certified unlearning for the distilled student if there exists a certifying algorithm $\overline{A}$ such that for any forget set $\mathcal{D}_u \subseteq \mathcal{D}$ and any measurable set $\mathcal{S} \subseteq \mathcal{H}_S$,

$$\mathbb{P}(\mathcal{U}(\mathcal{A}_S(\mathcal{D}, f_{\boldsymbol{\phi}^*}), \mathcal{D}, \mathcal{D}_u) \in \mathcal{S}) \leq e^{\varepsilon}\mathbb{P}(\overline{\mathcal{A}}(\mathcal{D}\backslash\mathcal{D}_u, f_{\boldsymbol{\phi}_r^*}) \in \mathcal{S}) + \delta$$
$$\mathbb{P}(\overline{\mathcal{A}}(\mathcal{D}\backslash\mathcal{D}_u, f_{\boldsymbol{\phi}_r^*}) \in \mathcal{S}) \leq e^{\varepsilon}\mathbb{P}(\mathcal{U}(\mathcal{A}_S(\mathcal{D}, f_{\boldsymbol{\phi}^*}), \mathcal{D}, \mathcal{D}_u) \in \mathcal{S}) + \delta,$$

To achieve (ϵ, δ)-unlearning, one can consider redistilling the student on the retained data $\mathcal{D}_r$ using the original teacher, and release the resulting student $\widehat{\boldsymbol{w}}_r = \mathcal{A}_S(\mathcal{D}_r, f_{\boldsymbol{\phi}^*})$. This way both $\boldsymbol{w}_r^*$ and $\widehat{\boldsymbol{w}}_r$ are obtained by running the same student learning algorithm $\mathcal{A}_S$ on the retrained set $\mathcal{D}_r$. On the other hand, the two models use different teachers, where $\boldsymbol{w}_r^*$ uses the retrained teacher and $\widehat{\boldsymbol{w}}_r$ uses the original teacher, respectively. Since the predictions of the original teacher $f_{\boldsymbol{\phi}^*}$ on the retrained set $\mathcal{D}_r$ are still influenced by the forget samples $\mathcal{D}_u$, $\widehat{\boldsymbol{w}}_r$ does not in general coincide with the retraining reference $\boldsymbol{w}_r^*$. In Fig. 1, $\boldsymbol{w}_r^*$ corresponds to the retrain-from-scratch reference in panel (b), whereas $\widehat{\boldsymbol{w}}_r$ corresponds to the redistill-on-retain-data baseline in panel (c). To erase the influence of black-box teacher predictions from the student model, we instead release a perturbed student $\widetilde{\boldsymbol{w}}_r = \widehat{\boldsymbol{w}}_r + \boldsymbol{\xi}$ where the noise $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{d_S})$ is calibrated according to the student sensitivity on the teacher predictions. Motivated by (Neel et al., 2021; Sekhari et al., 2021; Zhang et al., 2024a; Lu et al., 2025; Basaran et al., 2025), we define the certification algorithm $\overline{\mathcal{A}}$ as a randomized release $\widetilde{\boldsymbol{w}}_r^* := \boldsymbol{w}_r^* + \boldsymbol{\xi}'$ of the retrained from scratch model $\boldsymbol{w}_r^*$, where $\boldsymbol{\xi}' \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{d_S})$. The certifying algorithm $\overline{\mathcal{A}}$ is used only to define the reference distribution for statistical indistinguishability and is not executed during unlearning. We provide the details of the certification procedure in Thm. 5.5.

4 METHODOLOGY

This section describes our student-side unlearning mechanism, CKDU. Since we operate in the black-box distillation setting where the teacher is a fixed predictor that can only be queried, unlearning is performed by updating the student parameters. Given the dataset D and a forget set $\mathcal{D}_u$, we form the retained set $\mathcal{D}_r = \mathcal{D}\backslash\mathcal{D}_u$ and proceed in two stages: 1) (Distillation) We consider student distillation on $\mathcal{D}_r$ using the available teacher $f_{\boldsymbol{\phi}^*}$, which produces an intermediate student model $\widehat{\boldsymbol{w}}_r$, 2) (Randomized release) We then release a randomized version of $\widehat{\boldsymbol{w}}_r$ by adding calibrated Gaussian noise, which is calibrated to satisfy the certified unlearning guarantee (Def. 3.1). Before providing details, we state a few useful technical assumptions (please see App. A for a detailed statement).

Assumption 4.1 (Teacher side)**.** The teacher is trained with SGD for T iterations using a nonnegative learning-rate schedule $\{\alpha_t\}_{t=1}^{T}$. The per-example teacher loss is L-Lipschitz and β-smooth in the teacher parameters. These conditions will be utilized in the teacher perturbation analysis in Thm. 5.3.

Assumption 4.2 (Student side)**.** The student is distilled using the objective in equation 1 for T_S iterations with a nonincreasing learning rate schedule $\{\eta_t\}_{t=1}^{T_S-1}$. We assume that the student loss is β_S-smooth and its logit Jacobian is bounded by some B_J. These two conditions will be used in the student sensitivity analysis in Thm. 5.1.

Optional: While our main result in Thm. 5.1 applies to non-convex teacher and student losses, if the student loss additionally satisfies μ-strong convexity, we can obtain a tighter contraction-based student sensitivity bound which will be discussed in Cor. 5.2.

We next describe the individual stages of our mechanism.

Student Distillation on the Retained Dataset. The first step applies the deletion request on the student objective by restricting the student dataset to $\mathcal{D}_r$. In particular, we apply the distillation procedure on $\mathcal{D}_r$ using only query (prediction) access to the teacher, $\widehat{\boldsymbol{w}}_r := \mathcal{A}_S(\mathcal{D}_r, f_{\boldsymbol{\phi}^*})$. This is the closest analogue of retraining that is compatible with black-box teacher access: it uses the same student learning rule $\mathcal{A}_S$ as in equation 1 and removes $\mathcal{D}_u$ from the optimization entirely. On the other hand, the resulting model $\widehat{\boldsymbol{w}}_r$ is generally not identical to the retraining-reference student $\boldsymbol{w}_r^*$, because $\boldsymbol{w}_r^*$ is obtained by distillation on $\mathcal{D}_r$ under the retrained teacher $f_{\boldsymbol{\phi}_r^*}$. In other words, even after removing $\mathcal{D}_u$ from student training, the supervision signal provided by the original teacher can differ from the one induced by a retrained (ideal) teacher, which is not available in the black-box setting. To that end, we refer to $\widehat{\boldsymbol{w}}_r$ as the intermediate student model. Our analysis in Sec. 5 will quantify the impact of this teacher mismatch on the student model.

Algorithm 1 CKDU: Student-side $(\varepsilon, \delta{+}\gamma)$ certified unlearning for knowledge distillation with black-box teacher.

Require: Dataset $\mathcal{D}$, forget set $\mathcal{D}_u$, teacher predictor $f_{\boldsymbol{\phi}^*}$, distillation algorithm $\mathcal{A}_S$, certification budget (ε, δ)

Ensure: Released unlearned student parameters $\widetilde{\boldsymbol{w}}_r$

1: $\mathcal{D}_r \leftarrow \mathcal{D} \setminus \mathcal{D}_u$
2: $\widehat{\boldsymbol{w}}_r \leftarrow \mathcal{A}_S(\mathcal{D}_r, f_{\boldsymbol{\phi}^*})$
3: Set σ using Thm. 5.5 (after bounding $\Delta \le \overline{\Delta}$ as given in equation 11)
4: Sample $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d_S})$
5: $\widetilde{\boldsymbol{w}}_r \leftarrow \widehat{\boldsymbol{w}}_r + \boldsymbol{\xi}$
6: **return** $\widetilde{\boldsymbol{w}}_r$

Randomized Release of the Student Model. The certification requirement in Def. 3.1 compares the distributions of released student parameters with the retraining reference (ideal). To satisfy this constraint, we introduce controlled randomness by releasing a perturbed version of the intermediate student model $\widetilde{\boldsymbol{w}}_r = \widehat{\boldsymbol{w}}_r + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d_S})$ and, the noise level σ is calibrated using an upper bound on the end-to-end parameter deviation from the retraining reference $\boldsymbol{w}_r^*$,

$$\Delta := \|\widehat{\boldsymbol{w}}_r - \boldsymbol{w}_r^*\|_2. \tag{2}$$

which we refer to as the student sensitivity to unlearning. In Thm. 5.4, we demonstrate that for any $\gamma \in (0, 1)$, we have $\Delta \le \overline{\Delta}$, with probability at least $1 - \gamma$, where $\overline{\Delta}$ is as given in equation 11. Once we establish this bound, we calibrate the noise level σ using the analytic Gaussian mechanism Balle & Wang (2018), which provides us with the $(\varepsilon, \delta{+}\gamma)$-unlearning guarantees as will be discussed in Thm. 5.5. Alg. 1 summarizes our mechanism. In the following section, we provide the formal derivations for our student sensitivity bound and end-to-end certification guarantees.

5 Theoretical Principles

This section develops the bounds required for the unlearning guarantees under Def. 3.1. We first proceed by bounding the sensitivity of the student model parameters to the discrepancies between teacher predictive distributions, and then characterize how much a teacher's predictions on retained examples change if the teacher model was retrained from scratch by using the retain samples $\mathcal{D}_r$ only. To that end, our first result is a student-side sensitivity bound that quantifies the discrepancy between the student parameters obtained via two different teacher distributions, as a function of the discrepancy between the teacher distributions. For this, we analyze the finite training trajectories for the student parameters induced by different teacher predictions, and then obtain an upper bound on the student parameter variations under smoothness and bounded logit Jacobian conditions.

We then characterize the shift in the teacher predictions that would result from removing the forget samples $\mathcal{D}_u$, and obtain a teacher-side bound in Thm. 5.3 on the discrepancy between the original and retrained-from-scratch teacher distributions. Since in the black-box setting the retrained-from-scratch teacher is a counterfactual which cannot be observed directly, to bound the teacher discrepancy we leverage uniform-stability of SGD (Hardt et al., 2016). We then combine the student-side sensitivity bound with the teacher-side bound to obtain a high-probability upper bound on the deviation of the student parameters $\Delta = \|\widehat{\boldsymbol{w}}_r - \boldsymbol{w}_r^*\|_2$. Once we have a bound on Δ, the final certification step follows from Gaussian perturbation. Our main result applies to general non-convex student objectives, and if the student objective further satisfies strong convexity, we can provide tighter bounds.

In the following, we will use the empirical total variation distance (TV),

$$\widehat{D}_{\mathrm{TV}}(p_1, p_0) \coloneqq \frac{1}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} D_{\mathrm{TV}}(p_1(\boldsymbol{x}), p_0(\boldsymbol{x}))\,. \tag{3}$$

where the average is taken across the retain samples $\boldsymbol{x} \in \mathcal{D}_r$, and $p_1(\boldsymbol{x}), p_0(\boldsymbol{x})$ are two predictive distributions over the labels $\{1, \ldots, K\}$. We will first characterize the sensitivity of the student objective to variations in the teacher predictions. Recall that $\widehat{\boldsymbol{w}}_r$ is the model obtained by distilling on $\mathcal{D}_r$ using the original teacher trained on D, while the retraining reference $\boldsymbol{w}_r^*$ uses the (counterfactual) retrained teacher on $\mathcal{D}_r$. Hence, both students are trained on the same inputs, but they see different supervision signals through the teacher predictions. We measure this mismatch by $\widehat{D}_{\mathrm{TV}}$ and bound the induced change in the student parameters Δ.

Theorem 5.1 (Non-convex distillation sensitivity to teacher discrepancy)**.** *Let $p_1, p_0 : \mathcal{X} \to \mathcal{P}_K$ be two teacher predictive distributions, and recall $\widehat{D}_{\mathrm{TV}}(p_1, p_0)$ from equation 3. Consider two student runs on $\mathcal{D}_r$, using teacher predictions p_i for $i \in \{0, 1\}$. Under Assumption 4.2, suppose that the two student runs start from the same initialization and execute T_S gradient-descent steps in $\mathcal{D}_r$. Then, the following upper bound holds for the student parameter deviation Δ,*

$$\Delta \leq 2\lambda B_J G_{T_S} \widehat{D}_{\mathrm{TV}}(p_1, p_0) \tag{4}$$

where

$$G_{T_S} := \sum_{t=0}^{T_S-1} \eta_t \prod_{k=t+1}^{T_S-1} (1 + \eta_k \beta_S) \tag{5}$$

such that β_S is the smoothness constant of $\mathcal{L}(\cdot; p_0)$ and $\eta_t > 0$ is the student learning rate at round t.

Proof. The proof is provided in App. B. □

Thm. 5.1 quantifies the shift in the student parameters induced by the changes in the teacher predictions. As $\widehat{D}_{\mathrm{TV}}(p_1, p_0)$ decreases, the right hand side of equation 4 gradually decreases. When $p_1 = p_0$, which is satisfied when $\mathcal{D}_u = \emptyset$, we have $\Delta = 0$.

The above sensitivity bound applies to non-convex student loss functions. If the student loss additionally satisfies stronger curvature conditions, one can obtain a tighter sensitivity bound that replaces the expansion factor $(1 + \eta_k \beta_S)$ in equation 5 with a contraction factor $(1 - \eta_k \mu)$ as follows.

Corollary 5.2 (Tighter sensitivity under strong convexity)**.** *Let, in addition to Assumption 4.2, the student loss $\mathcal{L}(\cdot; p_0)$ be μ-strongly convex and $0 < \eta_t \leq 1/\beta_S$ for all $t \in 0, \ldots, T_S - 1$. Then,*

$$\Delta \leq 2\lambda B_J \widetilde{G}_{T_S} \widehat{D}_{\mathrm{TV}}(p_1, p_0) \tag{6}$$

where

$$\widetilde{G}_{T_S} := \sum_{t=0}^{T_S-1} \eta_t \prod_{k=t+1}^{T_S-1} (1 - \eta_k \mu) \tag{7}$$

For constant learning rate $\eta_t = \eta$, we further observe,

$$\Delta \leq \frac{2\lambda B_J}{\mu}(1 - (1 - \eta\mu)^{T_S}) \widehat{D}_{\mathrm{TV}}(p_1, p_0) \leq \frac{2\lambda B_J}{\mu} \widehat{D}_{\mathrm{TV}}(p_1, p_0) \tag{8}$$

Proof. The proof is available in App. B.4. □

We next upper bound the discrepancy between the two teacher predictions on $\mathcal{D}_r$, where the first teacher is trained using the original teacher $p_1 = f_{\phi^*}$ (i.e., trained on the entire dataset $\mathcal{D}$) and the second teacher is trained using the retrained-from-scratch teacher $p_0 = f_{\phi_r^*}$ (i.e., trained only on the retain dataset $\mathcal{D}_r$). Since we do not have access to the retrained-from-scratch teacher in the black-box setting, we instead leverage algorithmic stability for SGD (Hardt et al., 2016; Zhang et al., 2022) to obtain a high-probability bound on this discrepancy. Specifically, classical uniform stability controls how much the learned predictor changes when one training example is replaced, and we adapt this argument to the deletion setting and then interpret the resulting stability guarantee as a bound on the discrepancy between the corresponding teacher output distributions.

Theorem 5.3 (Teacher output discrepancy under m deletions)**.** *Let f_{ϕ^*} and $f_{\phi_r^*}$ denote the predictions of the original teacher trained on $\mathcal{D}$ and retrained-from-scratch teacher trained on $\mathcal{D}_r = \mathcal{D}\backslash\mathcal{D}_u$, respectively, where $|\mathcal{D}| = n$ and $|\mathcal{D}_u| = m$. Fix any $\gamma \in (0,1)$. With probability at least $1-\gamma$ over the randomness of the teacher training procedures used to obtain f_{ϕ^*} and $f_{\phi_r^*}$,*

$$\widehat{D}_{\mathrm{TV}}(f_{\phi^*}, f_{\phi_r^*}) \le \varepsilon_{\mathrm{TV}} := \tanh\left(\frac{\varepsilon_{\mathrm{stab}}}{2}\right) \tag{9}$$

where

$$\varepsilon_{\mathrm{stab}} := \min_{\substack{t_0\in\{0,\dots,T\}:\\ q_{t_0}<\gamma}} \left\{ \sum_{t=t_0+1}^{T} 2L^2\alpha_t,\ \inf_{\theta>0} \frac{\log\left(\frac{1-q_{t_0}}{\gamma-q_{t_0}}\right) + \sum_{t=t_0+1}^{T}\log\left(1-\frac{m}{n}+\frac{m}{n}e^{\theta a_t}\right)}{\theta} \right\} \tag{10}$$

such that $a_t := 2L^2\alpha_t\prod_{k=t+1}^{T}(1+\alpha_k\beta)$ and $q_{t_0} := 1-\left(1-\frac{m}{n}\right)^{t_0}$, and $\alpha_t > 0$ is the learning rate at round t. Moreover, under a decaying step-size schedule $\alpha_t \le c_{\mathrm{lr}}/t$, the expansion term can be bounded as $a_t \le \frac{2c_{lr}L^2}{t}\left(\frac{T}{t}\right)^{\beta c_{lr}}$.

Proof. The proof follows trajectory coupling and high-probability analysis for SGD stability Hardt et al. (2016); Kuzborskij & Lampert (2018); Madden et al. (2020), and is deferred to App. C. □

By combining Thm. 5.1 with Thm. 5.3, we now obtain an end-to-end upper bound on Δ.

Theorem 5.4 (End-to-end bound on student sensitivity Δ)**.** *Fix $\gamma \in (0,1)$ and suppose Assumptions 4.1 and 4.2 hold. Let $\varepsilon_{\mathrm{TV}}$ denote the teacher discrepancy bound in equation 9. Then, with probability at least $1-\gamma$, the student sensitivity satisfies,*

$$\Delta \le \overline{\Delta} := 2\lambda B_J G_{T_S}\varepsilon_{\mathrm{TV}} \tag{11}$$

If, in addition, the student loss $\mathcal{L}(\cdot; p_0)$ is μ-strongly convex and $0 < \eta_t \le 1/\beta_S$, then G_{T_S} can be replaced with $\widetilde{G}_{T_S}$ from equation 7.

Proof. The proof is available in App. D.1. □

Thm. 5.4 provides the L_2 bound $\overline{\Delta}$ needed to calibrate the noise for (ε,δ)-certified unlearning. After establishing this bound, we calibrate the noise level σ for randomization using the analytic Gaussian mechanism (Balle & Wang, 2018, Thm. 9), and release $\widetilde{\boldsymbol{w}}_r = \widehat{\boldsymbol{w}}_r + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{d_S})$. The next theorem gives sufficient conditions on σ to satisfy Def. 3.1.

Theorem 5.5 ($(\varepsilon, \delta{+}\gamma)-$certified unlearning)**.** *Fix $\gamma \in (0,1)$ and assume that $\Delta \le \overline{\Delta}$ holds with probability at least $1-\gamma$. Let $\widetilde{\boldsymbol{w}}_r = \widehat{\boldsymbol{w}}_r + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{d_S})$. Then, for any measurable set $\mathcal{S} \subseteq \mathcal{H}_S$, Def. 3.1 holds with parameters $(\varepsilon, \delta{+}\gamma)$ if,*

$$\Phi\left(\frac{\overline{\Delta}}{2\sigma} - \frac{\epsilon\sigma}{\overline{\Delta}}\right) - e^{\epsilon}\Phi\left(-\frac{\overline{\Delta}}{2\sigma} - \frac{\epsilon\sigma}{\overline{\Delta}}\right) \le \delta \tag{12}$$

where Φ denotes the cumulative distribution function (CDF) of the standard normal distribution.

Proof. The proof is provided in App. D.2. □

In practice, we numerically compute the smallest admissible noise scale σ using the calibration procedure in (Balle & Wang, 2018, Alg. 1) with sensitivity $\overline{\Delta}$ and the target (ϵ, δ).

6 Experiments

Setup. We evaluate CKDU along two main directions. The first is empirical unlearning, which measures whether the released student matches the retraining reference in utility and in the membership information it retains about the forget set. The second is the validity of our analysis, which tests whether the bounds of Sec. 5 hold empirically and scale with the forget-set size as the observed discrepancies do. Both evaluations require the retrained teacher, which is unavailable under black-box access. We therefore train all distillation pipelines end-to-end, where the retrained teacher serves solely as an offline evaluation oracle. Details of our experimental setup are provided in App. E.

The data settings progress from a fully controlled synthetic task to natural images. We start with synthetic XOR, where points are drawn from four overlapping clusters in an XOR arrangement, with

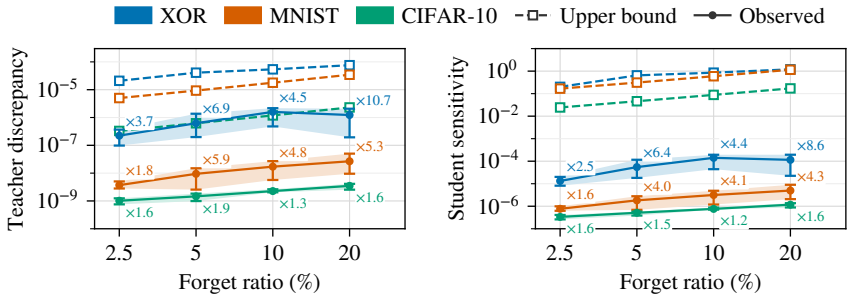


Figure 2: Bounds (dashed) against observed values (solid) for the convex student across different forget ratios m/n. The left panel compares $\varepsilon_{\mathrm{TV}}$ with $\widehat{D}_{\mathrm{TV}}(p_1, p_0)$, and the right panel $\overline{\Delta}$ with Δ. Observed curves should be read against the bound of the same color. The label next to each observed point gives the spread across three seeds (the ratio of the largest seed value to the mean value).

diagonally opposite clusters sharing a label. No linear classifier separates these classes, so teacher and student must learn nonlinear decision boundaries. We then turn to real images with binary MNIST (digits 3 and 8) (LeCun et al., 1998), the task on which Guo et al. (2020) evaluate certified data removal, placing our results next to prior certified unlearning work. Finally, CIFAR-10 (Krizhevsky, 2009) brings natural images and ten classes. Following the transfer-learning setup of Koloskova et al. (2025), teacher and student operate on features of a frozen ResNet-18 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009). This mirrors common practice, where a public backbone is shared and only task-specific models are trained on private data, so the backbone holds nothing to unlearn. The forget samples are selected uniformly at random according to the given forget ratio m/n.

Initially, teacher and student are one hidden layer tanh networks, which keeps every quantity in our bounds computable. The tanh activation is smooth and bounded, so with bounded inputs the Lipschitz and smoothness constants required by Assumptions 4.1 and 4.2 admit closed forms, while a single hidden layer captures nonlinear decision boundaries. The teacher trains its hidden weights and biases under a fixed, balanced output layer, which keeps the scale of its logits and constants under control. It is also wider than the student, so a compact student learns from a more expressive teacher. We study two student regimes that mirror Sec. 5 and differ only in which of the student's layers are trained: 1) The convex student freezes a randomly initialized hidden layer and trains only its output layer, so the distillation objective is strongly convex and Cor. 5.2 applies. 2) The non-convex student instead trains the first layer while fixing the final layer, for which the general non-convex setting of Thm. 5.1 applies, where the student sensitivity is expanding and can grow with every step. Beyond these two regimes, App. H further explores sensitivity at scale, in a deep network whose constants are empirically estimated along training rather than derived in closed form.

Evaluation metrics. We compare the released student against the retraining reference targeted by Def. 3.1. Recall that the original student $\boldsymbol{w}^*$ (Original) is the deployed model before deletion, and the reference $\boldsymbol{w}_r^*$ (Retrain) is distilled on the retained data $\mathcal{D}_r$ from a teacher retrained on $\mathcal{D}_r$. The student redistilled on the retain data $\mathcal{D}_r$ using original teacher produces the intermediate model $\widehat{\boldsymbol{w}}_r$ (Redistill-only). CKDU takes the intermediate student $\widehat{\boldsymbol{w}}_r$ and releases $\widetilde{\boldsymbol{w}}_r = \widehat{\boldsymbol{w}}_r + \boldsymbol{\xi}$. We also report $\widetilde{\boldsymbol{w}}_r^* = \widehat{\boldsymbol{w}}_r^* + \boldsymbol{\xi}$ (Retrain + noise), the retrained from scratch model released by introducing the same amount of calibrated noise as a correct target illustration to our Def. 3.1. While existing unlearning heuristics carry no certificate in our setting, we also compare against well-known methods solely focusing on the empirical performance and not the theoretical certification in App. G.

We measure utility by accuracy on the retained data and on test data. On the forget set, successful unlearning does not mean low accuracy, since a model that never saw these samples still generalizes to them, but accuracy close to that of the reference. Accuracy alone, however, cannot reveal whether a model still remembers individual samples. We therefore also run the loss-threshold membership inference attack (MIA) of Kurmanji et al. (2023), which separates forgotten samples from test data by their loss and reports balanced accuracy, where 50% corresponds to chance.

Finally, we test the theory directly by setting each bound against the quantity it controls. The teacher bound $\varepsilon_{\mathrm{TV}}$ of Thm. 5.3 is compared with the observed teacher discrepancy $\widehat{D}_{\mathrm{TV}}(p_1, p_0)$ of equation 3, and the sensitivity bound $\overline{\Delta}$ of Thm. 5.4 with the observed student distance Δ of equation 2. Both observed quantities require the retrained teacher and never enter the calibration.

Table 1: Unlearning performances of both the convex and non-convex students at a forget ratio of 20%. For the convex student, each entry is the mean over three seeds with min/max as subscript.

	Method	XOR				MNIST				CIFAR-10			
		Test	Forget	Retain	MIA	Test	Forget	Retain	MIA	Test	Forget	Retain	MIA
Convex	Original	$88.4_{\pm 1.2}$	$91.7_{\pm 4.0}$	$94.8_{\pm 1.7}$	$53.2_{\pm 8.2}$	$96.1_{\pm 0.2}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$58.5_{\pm 4.1}$	$74.9_{\pm 0.2}$	$99.3_{\pm 0.3}$	$99.8_{\pm 0.1}$	$69.3_{\pm 1.0}$
	Retrain	$88.5_{\pm 0.8}$	$84.6_{\pm 1.9}$	$96.6_{\pm 1.3}$	$54.6_{\pm 4.4}$	$96.0_{\pm 0.4}$	$95.4_{\pm 1.0}$	$100.0_{\pm 0.0}$	$49.6_{\pm 2.2}$	$73.6_{\pm 0.3}$	$70.0_{\pm 0.9}$	$100.0_{\pm 0.0}$	$50.3_{\pm 3.0}$
	Retrain + noise	$88.1_{\pm 0.6}$	$82.8_{\pm 2.9}$	$95.6_{\pm 1.5}$	$52.1_{\pm 6.8}$	$95.7_{\pm 0.3}$	$94.8_{\pm 0.7}$	$100.0_{\pm 0.0}$	$49.7_{\pm 1.5}$	$73.5_{\pm 0.3}$	$70.4_{\pm 0.9}$	$100.0_{\pm 0.0}$	$50.0_{\pm 2.6}$
	CKDU	$\mathbf{88.1_{\pm 0.6}}$	$\mathbf{82.6_{\pm 3.2}}$	$\mathbf{95.5_{\pm 1.6}}$	$\mathbf{52.4_{\pm 6.6}}$	$\mathbf{95.6_{\pm 0.2}}$	$\mathbf{94.8_{\pm 0.7}}$	$\mathbf{100.0_{\pm 0.0}}$	$\mathbf{49.9_{\pm 1.3}}$	$\mathbf{73.5_{\pm 0.4}}$	$\mathbf{70.5_{\pm 0.9}}$	$\mathbf{100.0_{\pm 0.0}}$	$\mathbf{49.9_{\pm 2.4}}$
	Redistill-only	$88.3_{\pm 0.9}$	$87.3_{\pm 1.9}$	$94.4_{\pm 1.2}$	$50.1_{\pm 5.9}$	$95.8_{\pm 0.7}$	$98.0_{\pm 0.9}$	$100.0_{\pm 0.1}$	$55.6_{\pm 2.4}$	$74.8_{\pm 0.1}$	$89.3_{\pm 0.9}$	$99.7_{\pm 0.1}$	$59.4_{\pm 1.1}$
non-convex	Original	$88.7_{\pm 0.5}$	$87.9_{\pm 2.1}$	$87.9_{\pm 1.4}$	$53.2_{\pm 2.8}$	$96.8_{\pm 0.2}$	$96.0_{\pm 0.8}$	$98.6_{\pm 0.5}$	$56.5_{\pm 2.1}$	$72.9_{\pm 0.3}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$72.5_{\pm 2.4}$
	Retrain	$88.7_{\pm 0.4}$	$86.7_{\pm 1.8}$	$88.2_{\pm 1.2}$	$52.0_{\pm 2.5}$	$96.8_{\pm 0.2}$	$96.4_{\pm 0.7}$	$98.8_{\pm 0.4}$	$50.3_{\pm 1.8}$	$72.2_{\pm 0.3}$	$75.0_{\pm 1.2}$	$100.0_{\pm 0.0}$	$52.0_{\pm 2.1}$
	Retrain + noise	$88.3_{\pm 0.5}$	$84.9_{\pm 2.0}$	$87.6_{\pm 1.3}$	$52.1_{\pm 2.4}$	$96.6_{\pm 0.3}$	$96.0_{\pm 0.8}$	$98.4_{\pm 0.6}$	$50.2_{\pm 1.7}$	$71.9_{\pm 0.3}$	$73.7_{\pm 1.1}$	$99.3_{\pm 0.4}$	$51.9_{\pm 2.0}$
	CKDU	$\mathbf{87.5_{\pm 0.5}}$	$\mathbf{83.9_{\pm 2.2}}$	$\mathbf{87.8_{\pm 1.3}}$	$\mathbf{52.6_{\pm 2.6}}$	$\mathbf{96.3_{\pm 0.3}}$	$\mathbf{96.0_{\pm 0.8}}$	$\mathbf{97.6_{\pm 0.6}}$	$\mathbf{51.2_{\pm 1.9}}$	$\mathbf{70.3_{\pm 0.4}}$	$\mathbf{72.9_{\pm 1.3}}$	$\mathbf{99.3_{\pm 0.4}}$	$\mathbf{53.0_{\pm 2.2}}$
	Redistill-only	$87.6_{\pm 0.4}$	$86.8_{\pm 2.0}$	$87.2_{\pm 1.0}$	$51.5_{\pm 3.5}$	$96.4_{\pm 0.3}$	$96.0_{\pm 1.0}$	$97.6_{\pm 0.8}$	$51.6_{\pm 2.1}$	$70.5_{\pm 0.5}$	$77.8_{\pm 1.3}$	$99.1_{\pm 0.5}$	$57.9_{\pm 3.4}$

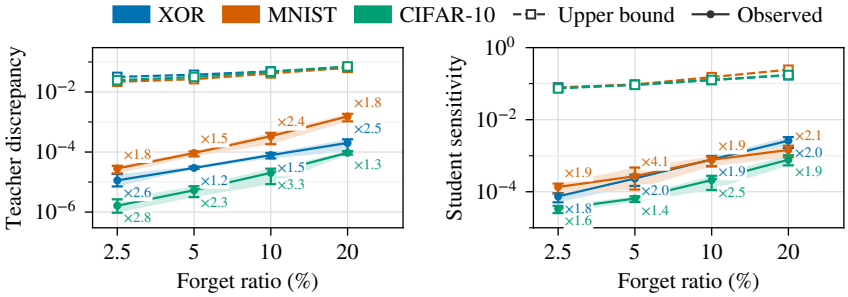


Figure 3: Bounds (dashed) against observed values (solid) for the non-convex student across different forget ratios m/n. The left panel compares $\varepsilon_{\mathrm{TV}}$ with $\widehat{D}_{\mathrm{TV}}(p_1, p_0)$, and the right panel compares $\overline{\Delta}$ with Δ. Each observed curve should be read against the bound of the same color.

The noise level σ is the smallest value satisfying Thm. 5.5, computed with the analytic Gaussian mechanism of Balle & Wang (2018), and thus depends only on the configuration and the forget-set size. Unless stated otherwise, we use $(\varepsilon, \delta + \gamma) = (1, 0.002)$ with $\delta = \gamma = 10^{-3}$ and $\lambda = 0.5$.

Experiments with convex student. We train the convex student's output layer with a small ridge penalty and a constant step size below $1/\beta_S$, as Cor. 5.2 requires. Table 1 reports the results at a forget ratio of 20%. Across all settings, CKDU matches Retrain + noise within 0.3 points on every metric, and the noise costs less than half a point of test accuracy. On MNIST/CIFAR-10, the original student fits the forget set almost perfectly and the attack detects these samples with 58% and 69% accuracy, whereas against CKDU it falls to chance. Redistill-only shows that without the noise, the student still retains some memory of the forget set, highlighting the importance of the calibrated noise in achieving unlearning. Fig. 2 compares the bounds with the observed quantities across different forget ratios. Every observed value lies below its bound, and both bounds grow with the forget ratio. As Thm. 5.3 suggests, more deletions make the two teachers draw different samples more often, and on MNIST/CIFAR-10 the observed discrepancies indeed rise with the forget ratio. The student inherits the teacher discrepancy in proportion, in line with the linear dependence in Cor. 5.2.

Experiments with non-convex student. The non-convex student trains only its hidden layer, while the output layer is fixed, which satisfies the general setting of Thm. 5.1. Since its sensitivity bound can grow with every step, we certify $T_S = 32$ gradient steps after a pretrained warm-start (warm-start samples are independent of the training data, see App. E). With 20% forget ratio, CKDU again stays close to Retrain + noise (Table 1) and brings the MIA on MNIST and CIFAR-10 from 56% and 72.5% down to near chance. Across forget ratios from 2.5% to 20%, every observed quantity lies below its bound and grows with the forget ratio (Fig. 3).

7 Conclusion

We study certified unlearning for a student model distilled from a black-box teacher, and develop a formal certification methodology by characterizing the student sensitivity against teacher perturbations due to data removal. Our analysis leads to a provable sensitivity bound and a noise calibration mechanism to randomize the student model re-distilled on the retain data, achieving distributional indistinguishability guarantees relative to the counterfactual retraining-from-scratch reference.

Reproducibility Statement

We provide the details of the theoretical assumptions and proofs, including the certification and noise calibration procedure underlying our approach CKDU in the main text with details in the appendices. Our experiments evaluate synthetic XOR, binary MNIST, and CIFAR-10 datasets using convex and non-convex student architectures with reproducible constants used in our theoretical analysis. We compare CKDU against retraining and re-distillation baselines using classification accuracy, membership-inference attacks, and empirical sensitivity measurements. All experimental details are provided in the experiments section and appendices.

Ethics and Limitations Statement

This work investigates certified unlearning to help address data-deletion requests in teacher–student learning pipelines. Our experiments use synthetic data and publicly available benchmarks without collecting additional personal information. Our guarantees apply to the randomized student relative to a randomized retraining reference under stated assumptions. Although we evaluate membership-inference attacks, we state that empirical forgetting does not establish the absence of all privacy risks or guarantee regulatory compliance. Removal of information from the original teacher is also beyond our scope. Our theoretical guarantees depend on regularity assumptions, and specific teacher and student-optimization procedures. Our bounds also utilize certain training metadata from the teacher training procedure (specifically the constants from Assumption A.1). Our method does not aim to protect against information leaked from such metadata and we leave the investigation of these to future work. In certain regimes, such as the non-convex student setup, the resulting sensitivity bounds may be conservative, potentially requiring large Gaussian noise that reduces predictive utility. Our experiments use controlled architectures and selected theoretically tractable classification benchmarks to verify our theoretical findings, and extending CKDU to larger models, including more complex training pipelines and black-box teacher settings remains as future work.

AI Use Statement

In this work, we used AI tools to assist in reviewing our paper and identifying gaps in the theoretical assumptions and proofs. We used AI as a strictly diagnostic tool to identify the technical gaps and help us search/identify the relevant literature for refinements. We have not used AI tools for conceptualizing our research idea, problem formulation and methodology. We have carefully reviewed and verified each claim for correctness (and only accounted for the correct ones), and manually searched and verified the literature for relevant work and provided references to the relevant literature whenever our proofs utilize a previous result. No claims were directly adopted from the AI tool without verification, and all technical claims were manually formalized and verified/validated rigorously by us. We accept full responsibility for the final content of our work.

References

Sk Miraj Ahmed, Umit Yigit Basaran, Dripta S. Raychaudhuri, Arindam Dutta, Rohit Kundu, Fahim Faisal Niloy, Basak Guler, and Amit K. Roy-Chowdhury. Towards source-free machine unlearning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4948–4957, June 2025.

Youssef Allouah, Joshua Kazdan, Rachid Guerraoui, and Sanmi Koyejo. The Utility and Complexity of In- and Out-of-Distribution Machine Unlearning. In *The Thirteenth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HVFMooKrHX.

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

Umit Yigit Basaran, Sk Miraj Ahmed, Amit Roy-Chowdhury, and Basak Guler. A certified unlearning approach without access to source data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8lt5776GLB.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019. URL https://ieeexplore.ieee.org/abstract/document/9519428.

Yinzhi Cao and Junfeng Yang. Towards Making Systems Forget with Machine Unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35. URL https://ieeexplore.ieee.org/document/7163042.

CCPA. California Consumer Privacy Act (CCPA), 2018. URL https://oag.ca.gov/privacy/ccpa.

Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to Unlearn: Instance-Wise Unlearning for Pre-trained Classifiers. 38(10):11186–11194, 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i10.28996. URL https://ojs.aaai.org/index.php/AAAI/article/view/28996.

Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Score Forgetting Distillation: A Swift, Data-Free Method for Machine Unlearning in Diffusion Models. 2024. URL https://openreview.net/forum?id=gjwhDHeAsz.

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, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9161a3e810cc0542718abddb791e8aed-Abstract-Conference.html.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Can Bad Teaching Induce Forgetting? Unlearning in Deep Networks Using an Incompetent Teacher. 37(6):7210–7217, 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i6.25879. URL https://ojs.aaai.org/index.php/AAAI/article/view/25879.

CPRA. *The California Privacy Rights Act (CPRA) – An implementation and compliance guide*. IT Governance Publishing, 2021. ISBN 9781787782860. URL http://www.jstor.org/stable/j.ctv1kv1d14.

Verna Dankers and Vikas Raunak. Memorization Inheritance in Sequence-Level Knowledge Distillation for Neural Machine Translation, 2025. URL http://arxiv.org/abs/2502.01491.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X, 1551-3068. doi: 10.1561/0400000042. URL https://www.nowpublishers.com/article/Details/TCS-042.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. 2023. URL https://openreview.net/forum?id=gn0mIhQGNM.

Jack Foster, Kyle Fogarty, Stefan Schoepf, Zack Dugue, Cengiz Oztireli, and Alexandra Brintrup. An Information Theoretic Approach to Machine Unlearning. 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=t1utIThKHD.

GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016. URL http://data.europa.eu/eli/reg/2016/679/oj.

Kristian Georgiev, Roy Rinberg, Sung Min Park, Shivam Garg, Andrew Ilyas, Aleksander Madry, and Seth Neel. Machine Unlearning via Simulated Oracle Matching. In *The Thirteenth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3vXpZpOn29.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. pp. 9304–9312, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Golatkar_Eternal_Sunshine_of_the_Spotless_Net_Selective_Forgetting_in_Deep_CVPR_2020_paper.html.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified Data Removal from Machine Learning Models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3832–3842. PMLR, 2020. URL https://proceedings.mlr.press/v119/guo20c.html.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1225–1234. PMLR, 2016. URL https://proceedings.mlr.press/v48/hardt16.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2:1–9, 2015. URL https://arxiv.org/abs/1503.02531.

Matthew Jagielski, Milad Nasr, Katherine Lee, Christopher A. Choquette-Choo, Nicholas Carlini, and Florian Tramer. Students Parrot Their Teachers: Membership Inference on Model Distillation. *Advances in Neural Information Processing Systems*, 36:44382–44397, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8b07d224a643b02e7571e083578a86d2-Abstract-Conference.html.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model Sparsity Can Simplify Machine Unlearning. 36:51584–51605, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a204aa68ab4e970e1ceccfb5b5cdc5e4-Abstract-Conference.html.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pp. 1376–1385. PMLR, 2015.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified Unlearning for Neural Networks. In *Forty-Second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=3rWQlV3s1I.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards Unbounded Machine Unlearning. 36:1957–1987, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/062d711fb777322e2152435459e6e9d9-Abstract-Conference.html.

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c69465280855cfe25d566e359da140c1-Abstract-Conference.html.

Linda Lu, Ayush Sekhari, and Karthik Sridharan. System-Aware Unlearning Algorithms: Use Lesser, Forget Faster. In *Forty-Second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DnuuFjCsDo.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2020.

Siqiao Mu and Diego Klabjan. Rewind-to-Delete: Certified Machine Unlearning for Nonconvex Functions, 2025. URL http://arxiv.org/abs/2409.09778.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-Delete: Gradient-Based Methods for Machine Unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021. URL https://proceedings.mlr.press/v132/neel21a.html.

Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What Knowledge Gets Distilled in Knowledge Distillation? 36:11037–11048, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2433fec2144ccf5fea1c9c5ebdbc3924-Abstract-Conference.html.

Yijun Quan, Zushu Li, and Giovanni Montana. Efficient verified unlearning for distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ofyy0Dzk6k.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember What You Want to Forget: Algorithms for Machine Unlearning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18075–18086. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9627c45df543c816a3ddf2d8ea686a99-Abstract.html.

Nazanin Mohammadi Sepahvand, Anvith Thudi, Berivan Isik, Ashmita Bhattacharyya, Nicolas Papernot, Eleni Triantafillou, Daniel M. Roy, and Gintare Karolina Dziugaite. Leveraging Per-Instance Privacy for Machine Unlearning. In *Forty-Second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0A4Y9qRnu9.

Xiaotong Yuan and Ping Li. l_2-uniform stability of randomized learning algorithms: Sharper generalization bounds and confidence boosting. *Advances in Neural Information Processing Systems*, 36:78580–78592, 2023.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards Certified Unlearning for Deep Neural Networks. In *Forty-First International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=1mf1ISuyS3.

Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of SGD: Tightness analysis and improved bounds. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 2364–2373. PMLR, 2022. URL https://proceedings.mlr.press/v180/zhang22b.html.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic Failure of LLM Unlearning via Quantization. 2024b. URL https://openreview.net/forum?id=lHSeDYamnz.

Ziqi Zhang, Ali Shahin Shamsabadi, Hanxiao Lu, Yifeng Cai, and Hamed Haddadi. Membership and Memorization in LLM Knowledge Distillation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20085–20095. Association for Computational Linguistics, 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1015. URL https://aclanthology.org/2025.emnlp-main.1015/.

APPENDIX

A ASSUMPTIONS

This section provides the details of the regularity assumptions used in our analysis. We first re-state the teacher-side assumptions from Assumption 4.1, which will later be used in the teacher-side stability analysis in Appendix C.

Assumption A.1 (Teacher-side regularity for SGD stability)**.** Let $\mathcal{H}_T \subseteq \mathbb{R}^{d_T}$ be the hypothesis set of the teacher model and $\mathcal{Z} = \mathcal{X} \times \{1, \dots, K\}$ be the input and label space with K number of classes. Define $\ell(\boldsymbol{\phi}; z)$ as the per-sample teacher training loss where $\boldsymbol{\phi} \in \mathcal{H}_T$, $z \in \mathcal{Z}$ and $\ell : \mathcal{H}_T \times \mathcal{Z} \to \mathbb{R}$ is differentiable in $\boldsymbol{\phi}$ for all $z \in \mathcal{Z}$. Then, assume the following.

(T1) **Lipschitzness.** Teacher loss $\ell(\cdot; z)$ is L-Lipschitz in $\boldsymbol{\phi}$ and differentiable at z. For all $\boldsymbol{\phi}, \boldsymbol{\phi}'$ and all z,

$$|\ell(\boldsymbol{\phi}; z) - \ell(\boldsymbol{\phi}'; z)| \leq L \left\|\boldsymbol{\phi} - \boldsymbol{\phi}'\right\|_2 . \tag{13}$$

and equivalently $\left\|\nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}; z)\right\|_2 \leq L$.

(T2) **Smoothness.** For all $\boldsymbol{\phi}, \boldsymbol{\phi}'$ and all z,

$$\left\|\nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}; z) - \nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}'; z)\right\|_2 \leq \beta \left\|\boldsymbol{\phi} - \boldsymbol{\phi}'\right\|_2 \tag{14}$$

for some $\beta > 0$.

(T3) **SGD dynamics.** The teacher is trained via SGD for T iterations with a nonincreasing learning-rate schedule $(\alpha_t)_{t=1}^{T}$ satisfying $\alpha_t \leq c_{\text{lr}}/t$ for some $c_{\text{lr}} > 0$,

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t - \alpha_t \nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}_t; z_{i_t}), \qquad i_t \sim \text{Unif}(\{1, \dots, n\}), \tag{15}$$

where n is the dataset size. Our theoretical analysis is focused on the cross-entropy loss.

We next provide the details of the student-side regularity assumptions from Assumption 4.2, which will be useful for the student-side sensitivity analysis in Appendix B.

Assumption A.2 (Student-side regularity for distillation sensitivity)**.** Fix the retained set $\mathcal{D}_r$ and define $p_0(\boldsymbol{x}) \coloneqq f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})$, where $f_{\boldsymbol{\phi}_r^*}$ denotes the teacher model if retrained on $\mathcal{D}_r$. Consider the student distillation objective from equation 1 when both the teacher and student are trained using only the retained set $\mathcal{D}_r$,

$$\mathcal{L}(\boldsymbol{w}; p_0) \coloneqq \frac{1}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_r} \Big((1-\lambda)\big(-\log g_{\boldsymbol{w}}(\boldsymbol{x})_y \big) + \lambda \, D_{\text{KL}}(p_0(\boldsymbol{x}) \,\|\, g_{\boldsymbol{w}}(\boldsymbol{x})) \Big) , \tag{16}$$

and a student model distilled for T_S training rounds,

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla \mathcal{L}(\boldsymbol{w}_t; p_0), \qquad t = 0, \dots, T_S - 1 \tag{17}$$

starting from a random initialization $\boldsymbol{w}_0$, where η_t denotes the student learning rate at round t.

Then, for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{H}_S$, the following conditions hold.

(S1) **Student smoothness.** The distillation loss $\mathcal{L}(\boldsymbol{w}; p_0)$ from equation 16 is β_S-smooth,

$$\left\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}; p_0) - \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}'; p_0)\right\|_2 \leq \beta_S \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|_2 \tag{18}$$

for some $\beta_S > 0$.

(S2) **Bounded logit Jacobian.** For all retained inputs $(\boldsymbol{x}, y) \in \mathcal{D}_r$, the logit Jacobian $\mathbf{J}_{\boldsymbol{w}}$ is bounded,

$$\left\|\mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x})\right\|_2 \leq B_J, \tag{19}$$

for some $B_J \geq 0$.

(S3) **(Optional) Strong convexity.** For twice differentiable losses, the Hessian $\mathbf{H}(\boldsymbol{w}; p_0) \coloneqq \nabla^2_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}; p_0)$ is uniformly positive definite for all $\boldsymbol{w} \in \mathcal{H}_S$,

$$\mathbf{H}(\boldsymbol{w}; p_0) \succeq \mu \mathbf{I}_{d_S} \tag{20}$$

for some $\mu > 0$.

We note that condition (S3) represents optional stronger geometry used in Corollary 5.2, and is not required for the general result in Theorem 5.1.

B DISTILLATION SENSITIVITY ANALYSIS

This appendix provides the proof of Theorem 5.1 and Corollary 5.2. Our goal is to quantify how a discrepancy between teacher predictive distributions on the retained set $\mathcal{D}_r$ propagates through the distillation objective and shifts the distilled student trajectories. To do so, we compare two student trajectories over a finite horizon, initialized identically but distilled using different teacher predictive distributions. We first bound the gradient perturbation caused by changing the teacher distribution and then propagate this perturbation through the student optimization trajectory.

B.1 SETUP AND NOTATION

Recall that the student model is parameterized by $\boldsymbol{w} \in \mathcal{H}_S$ and produces predictive distributions

$$g_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{s}_{\boldsymbol{w}}(\boldsymbol{x})) \in \mathcal{P}_K, \tag{21}$$

where $\boldsymbol{s}_{\boldsymbol{w}}(\boldsymbol{x}) \in \mathbb{R}^K$ denotes the (pre-softmax) logits. The corresponding logit Jacobian is given as,

$$\mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x}) = \nabla_{\boldsymbol{w}} \boldsymbol{s}_{\boldsymbol{w}}(\boldsymbol{x}) \in \mathbb{R}^{K \times d_S}. \tag{22}$$

Given a teacher distribution $p : \mathcal{X} \to \mathcal{P}_K$, the distillation objective restricted to the retained set $\mathcal{D}_r$ is

$$\mathcal{L}(\boldsymbol{w}; p) = \frac{1}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \big((1-\lambda)\big(-\log g_{\boldsymbol{w}}(\boldsymbol{x})_y\big) + \lambda\, D_{\mathrm{KL}}(p(\boldsymbol{x}) \,\|\, g_{\boldsymbol{w}}(\boldsymbol{x}))\big)\,. \tag{23}$$

Consider two teachers with predictive distributions $p_0(\boldsymbol{x})$ and $p_1(\boldsymbol{x})$ such that,

$$p_0(\boldsymbol{x}) = f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x}), \qquad p_1(\boldsymbol{x}) = f_{\boldsymbol{\phi}^*}(\boldsymbol{x}), \tag{24}$$

hence p_1 represents the original teacher which is available during unlearning, and p_0 is the ideal retrained-from-scratch teacher which is not available during unlearning. Let

$$\Delta p(\boldsymbol{x}) \coloneqq p_1(\boldsymbol{x}) - p_0(\boldsymbol{x}) \tag{25}$$

Define two distilled student models,

$$\boldsymbol{w}_{t+1}^{(i)} = \boldsymbol{w}_t^{(i)} - \eta_t \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t^{(i)}; p_i) \quad \text{for } i \in \{0,1\}, \tag{26}$$

trained using the same initialization,

$$\boldsymbol{w}_0^{(0)} = \boldsymbol{w}_0^{(1)} = \boldsymbol{w}_0 \tag{27}$$

over T_S gradient-descent steps $t \in \{0,\dots,T_S-1\}$. In the following, we let,

$$\boldsymbol{w}_{T_S}^{(0)} = \boldsymbol{w}_r^*, \qquad \boldsymbol{w}_{T_S}^{(1)} = \widehat{\boldsymbol{w}}_r \tag{28}$$

hence $\boldsymbol{w}_r^*$ and $\widehat{\boldsymbol{w}}_r$ are the outputs of the two T_S-step student trajectories on $\mathcal{D}_r$, using p_0 and p_1, respectively, where the two student trajectories start from the same initialization. Note that we do not assume convergence to a local or global minimizer, and allow both the teacher and student loss functions to be non-convex. Our analysis then compares the two finite-step trajectories. Finally, we define the corresponding student deviation as,

$$\Delta = \|\widehat{\boldsymbol{w}}_r - \boldsymbol{w}_r^*\|_2\,. \tag{29}$$

B.2 GRADIENT MAP AND TEACHER DEPENDENCE

This section makes explicit how the teacher distribution affects the student gradient, by deriving an explicit form of the gradient of the loss function defined in equation 23 in order to use in the sensitivity proof.

Define the gradient function

$$\boldsymbol{F}(\boldsymbol{w}, p) \coloneqq \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}; p). \tag{30}$$

Next, fix a retained sample $(\boldsymbol{x}, y) \in \mathcal{D}_r$ and write

$$\boldsymbol{q} \coloneqq g_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{s}_{\boldsymbol{w}}(\boldsymbol{x})) \in \mathcal{P}_K, \qquad \boldsymbol{p} \coloneqq p(\boldsymbol{x}) \in \mathcal{P}_K. \tag{31}$$

Then, expand the KL term in equation 23 as follows,

$$D_{\mathrm{KL}}(\boldsymbol{p} \,\|\, \boldsymbol{q}) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} = \sum_{k=1}^{K} p_k \log p_k - \sum_{k=1}^{K} p_k \log q_k.$$

Since $\boldsymbol{p}$ is independent of $\boldsymbol{w}$, the term $\sum_k p_k \log p_k$ is constant with respect to the logits $\boldsymbol{s}_{\boldsymbol{w}}(\boldsymbol{x})$. For $\boldsymbol{q} = \mathrm{softmax}(\boldsymbol{s})$ we have

$$q_k = \frac{e^{s_k}}{\sum_{j=1}^{K} e^{s_j}} \quad \Longrightarrow \quad \log q_k = s_k - \log\Big(\sum_{j=1}^{K} e^{s_j}\Big). \tag{32}$$

Therefore, for any $i \in \{1, \ldots, K\}$,

$$\frac{\partial}{\partial s_i} \log q_k = \frac{\partial}{\partial s_i} s_k - \frac{\partial}{\partial s_i} \log\Big(\sum_{j=1}^{K} e^{s_j}\Big) \tag{33}$$

$$= \delta_{ik} - \frac{1}{\sum_{j=1}^{K} e^{s_j}} \frac{\partial}{\partial s_i} \Big(\sum_{j=1}^{K} e^{s_j}\Big) \tag{34}$$

$$= \delta_{ik} - \frac{e^{s_i}}{\sum_{j=1}^{K} e^{s_j}} = \delta_{ik} - q_i. \tag{35}$$

where $\delta_{ik} = 1$ if and only if $i = k$ and 0 otherwise. Differentiating the KL expansion then yields,

$$\frac{\partial}{\partial s_i} D_{\mathrm{KL}}(\boldsymbol{p} \,\|\, \boldsymbol{q}) = -\sum_{k=1}^{K} p_k \frac{\partial}{\partial s_i} \log q_k = -\sum_{k=1}^{K} p_k(\delta_{ik} - q_i) = -p_i + q_i \sum_{k=1}^{K} p_k = q_i - p_i, \tag{36}$$

using $\sum_{k=1}^{K} p_k = 1$. Hence,

$$\nabla_{\boldsymbol{s}} \, D_{\mathrm{KL}}(p(\boldsymbol{x}) \,\|\, g_{\boldsymbol{w}}(\boldsymbol{x})) = g_{\boldsymbol{w}}(\boldsymbol{x}) - p(\boldsymbol{x}). \tag{37}$$

Similarly, writing $-\log g_{\boldsymbol{w}}(\boldsymbol{x})_y = -s_y + \log\big(\sum_{j=1}^{K} e^{s_j}\big)$, we obtain

$$\frac{\partial}{\partial s_i}\big(-\log g_{\boldsymbol{w}}(\boldsymbol{x})_y\big) = -\delta_{iy} + q_i = q_i - (\mathbf{e}_y)_i, \tag{38}$$

or equivalently,

$$\nabla_{\boldsymbol{s}}\big(-\log g_{\boldsymbol{w}}(\boldsymbol{x})_y\big) = g_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbf{e}_y, \tag{39}$$

where $\mathbf{e}_y \in \mathbb{R}^K$ denotes the one-hot vector for class y. Then, applying the chain rule $\nabla_{\boldsymbol{w}}(\cdot) = \mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x})^\top \nabla_{\boldsymbol{s}}(\cdot)$ and averaging over $\mathcal{D}_r$ yields,

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{w}, p) &= \frac{1}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x})^\top \Big((1-\lambda)\big(g_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbf{e}_y\big) + \lambda\big(g_{\boldsymbol{w}}(\boldsymbol{x}) - p(\boldsymbol{x})\big)\Big) \\ &= \frac{1-\lambda}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x})^\top \big(g_{\boldsymbol{w}}(\boldsymbol{x}) - \mathbf{e}_y\big) + \frac{\lambda}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x})^\top \big(g_{\boldsymbol{w}}(\boldsymbol{x}) - p(\boldsymbol{x})\big), \end{aligned} \tag{40}$$

where the first term denotes the supervised component of the gradient, which is independent of the teacher. Hence, for fixed $\boldsymbol{w}$,

$$\boldsymbol{F}(\boldsymbol{w}, p_1) - \boldsymbol{F}(\boldsymbol{w}, p_0) = -\frac{\lambda}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x})^\top \Delta p(\boldsymbol{x}). \tag{41}$$

where Δp is as defined in equation 25, from which we observe,

$$\|\boldsymbol{F}(\boldsymbol{w}, p_1) - \boldsymbol{F}(\boldsymbol{w}, p_0)\|_2 \leq \frac{\lambda}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \left\|\mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x})^\top \Delta p(\boldsymbol{x})\right\|_2 \tag{42}$$

$$\leq \frac{\lambda}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \|\mathbf{J}_{\boldsymbol{w}}(\boldsymbol{x})\|_2 \|\Delta p(\boldsymbol{x})\|_2 \tag{43}$$

$$\leq \frac{\lambda B_J}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} \|\Delta p(\boldsymbol{x})\|_1 \tag{44}$$

$$\leq \frac{2\lambda B_J}{|\mathcal{D}_r|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_r} D_{\mathrm{TV}}(p_1(\boldsymbol{x}), p_0(\boldsymbol{x})) \tag{45}$$

$$= 2\lambda B_J \widehat{D}_{\mathrm{TV}}(p_1, p_0) \tag{46}$$

where $D_{\mathrm{TV}}(p_1(\boldsymbol{x}), p_0(\boldsymbol{x})) := \frac{1}{2}\|p_1(\boldsymbol{x}) - p_0(\boldsymbol{x})\|_1$, and equation 44 follows from equation 19 and the fact that $\|\Delta p(\boldsymbol{x})\|_2 \leq \|\Delta p(\boldsymbol{x})\|_1$.

B.3 Proof of Theorem 5.1

Using the upper bound obtained in equation 46, which quantifies the gradient deviation between the two students resulting from the two teachers, one corresponding to the (available) original teacher and the other one corresponding to the (unavailable) retrained-from-scratch teacher, this section proves the main result in Theorem 5.1.

For each iteration t, define the student update corresponding to the reference teacher p_0 as follows,

$$\boldsymbol{\Phi}_t(\boldsymbol{w}_t; p_0) := \boldsymbol{w}_t - \eta_t \boldsymbol{F}(\boldsymbol{w}_t, p_0). \tag{47}$$

Then,

$$\|\boldsymbol{\Phi}_t(\boldsymbol{w}_t; p_0) - \boldsymbol{\Phi}_t(\boldsymbol{w}'_t; p_0)\|_2 = \|(\boldsymbol{w}_t - \boldsymbol{w}'_t) - \eta_t(\boldsymbol{F}(\boldsymbol{w}_t, p_0) - \boldsymbol{F}(\boldsymbol{w}'_t, p_0))\|_2 \tag{48}$$

$$\leq \|\boldsymbol{w}_t - \boldsymbol{w}'_t\|_2 + \eta_t \|\boldsymbol{F}(\boldsymbol{w}_t, p_0) - \boldsymbol{F}(\boldsymbol{w}'_t, p_0)\|_2 \tag{49}$$

$$\leq (1 + \eta_t \beta_S)\|\boldsymbol{w}_t - \boldsymbol{w}'_t\|_2. \tag{50}$$

which follows from the triangle inequality along with equation 18. Let

$$\Delta_t := \|\boldsymbol{w}_t^{(1)} - \boldsymbol{w}_t^{(0)}\|_2, \tag{51}$$

where $\boldsymbol{w}_t^{(1)}$ and $\boldsymbol{w}_t^{(0)}$ are as defined in equation 26. Then,

$$\boldsymbol{w}_{t+1}^{(1)} - \boldsymbol{w}_{t+1}^{(0)} = \boldsymbol{w}_t^{(1)} - \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_1) - \boldsymbol{w}_t^{(0)} + \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(0)}, p_0) \tag{52}$$

$$= \boldsymbol{w}_t^{(1)} - \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_1) - \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_0) + \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_0) - \boldsymbol{w}_t^{(0)} + \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(0)}, p_0) \tag{53}$$

$$= \boldsymbol{w}_t^{(1)} - \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_0) - \boldsymbol{w}_t^{(0)} + \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(0)}, p_0) - \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_1) + \eta_t \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_0) \tag{54}$$

$$= (\boldsymbol{\Phi}_t(\boldsymbol{w}_t^{(1)}; p_0) - \boldsymbol{\Phi}_t(\boldsymbol{w}_t^{(0)}; p_0)) - \eta_t(\boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_1) - \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_0)) \tag{55}$$

and from triangle inequality

$$\Delta_{t+1} \leq \left\|\boldsymbol{\Phi}_t(\boldsymbol{w}_t^{(1)}; p_0) - \boldsymbol{\Phi}_t(\boldsymbol{w}_t^{(0)}; p_0))\right\|_2 + \eta_t \left\|\boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_1) - \boldsymbol{F}(\boldsymbol{w}_t^{(1)}, p_0)\right\|_2 \tag{56}$$

$$\leq (1 + \eta_t \beta_S)\Delta_t + 2\eta_t \lambda B_J \widehat{D}_{\mathrm{TV}}(p_1, p_0). \tag{57}$$

where the last inequality follows by combining equation 50 and equation 46. Since the student trajectories have the same initialization, we have $\Delta_0 = 0$. Then, by unrolling equation 57 we observe,

$$\Delta_{T_S} \leq 2\lambda B_J \widehat{D}_{\mathrm{TV}}(p_1, p_0) \sum_{t=0}^{T_S-1} \eta_t \prod_{k=t+1}^{T_S-1} (1 + \eta_k \beta_S) \tag{58}$$

$$= 2\lambda B_J G_{T_S} \widehat{D}_{\mathrm{TV}}(p_1, p_0), \tag{59}$$

where G_{T_S} is as defined in equation 5. Since $\boldsymbol{w}_{T_S}^{(0)} = \boldsymbol{w}_r^*$ and $\boldsymbol{w}_{T_S}^{(1)} = \widehat{\boldsymbol{w}}_r$ from equation 27, we have $\Delta_{T_S} = \Delta$, from which Theorem 5.1 follows.

B.4 Proof of Corollary 5.2

This section provides the details of the tighter student sensitivity bound that can be obtained if the student loss satisfies stronger curvature conditions. To that end, in addition to Assumption A.2 (S1)–(S2), assume that the student loss function $\mathcal{L}(\cdot; p_0)$ is twice differentiable and μ-strongly convex as detailed in Assumption A.2 (S3). Define

$$\boldsymbol{A}(\boldsymbol{w}, \boldsymbol{w}') := \int_{u=0}^{1} \nabla_w^2 \mathcal{L}(\boldsymbol{w}' + u(\boldsymbol{w} - \boldsymbol{w}'); p_0)\, du. \tag{60}$$

Then, from the fundamental theorem of calculus,

$$\boldsymbol{F}(\boldsymbol{w}, p_0) - \boldsymbol{F}(\boldsymbol{w}', p_0) = \boldsymbol{A}(\boldsymbol{w}, \boldsymbol{w}')(\boldsymbol{w} - \boldsymbol{w}'). \tag{61}$$

From equation 18 and equation 20, one can observe that $\|\boldsymbol{H}(\boldsymbol{w}; p_0)\|_2 \preceq \beta_S$ and $\boldsymbol{H}(\boldsymbol{w}; p_0) \succeq \mu \mathbf{I}_{d_S}$, hence,

$$\mu \mathbf{I}_{d_S} \preceq \boldsymbol{A}(\boldsymbol{w}, \boldsymbol{w}') \preceq \beta_S \mathbf{I}_{d_S}. \tag{62}$$

Therefore, whenever $0 < \eta_t \leq 1/\beta_S$, all eigenvalues of $\mathbf{I} - \eta_t \boldsymbol{A}(\boldsymbol{w}, \boldsymbol{w}')$ lie in $[1 - \eta_t\beta_S, 1 - \eta_t\mu] \subseteq [0, 1]$, and

$$\|\boldsymbol{\Phi}_t(\boldsymbol{w}; p_0) - \boldsymbol{\Phi}_t(\boldsymbol{w}'; p_0)\|_2 \leq (1 - \eta_t\mu)\|\boldsymbol{w} - \boldsymbol{w}'\|_2. \tag{63}$$

Then, by repeating the same decomposition as in equation 57 but instead using equation 63 in the first term of equation 56, we observe,

$$\Delta_{t+1} \leq (1 - \eta_t\mu)\Delta_t + 2\eta_t\lambda B_J \widehat{D}_{\mathrm{TV}}(p_1, p_0). \tag{64}$$

Hence, by unrolling equation 64 and using the fact that $\Delta_0 = 0$, we have,

$$\Delta = \Delta_{T_S} \leq 2\lambda B_J \widehat{D}_{\mathrm{TV}}(p_1, p_0) \sum_{t=0}^{T_S-1} \eta_t \prod_{k=t+1}^{T_S-1} (1 - \eta_k\mu) \tag{65}$$

$$= 2\lambda B_J \widetilde{G}_{T_S} \widehat{D}_{\mathrm{TV}}(p_1, p_0), \tag{66}$$

which completes the proof for equation 6.

Finally, for a constant learning rate $\eta_t = \eta$, we have that,

$$\widetilde{G}_{T_S} = \eta \sum_{j=0}^{T_S-1} (1 - \eta\mu)^j \tag{67}$$

$$= \frac{1 - (1 - \eta\mu)^{T_S}}{\mu} \leq \frac{1}{\mu}, \tag{68}$$

where the last inequality follows from $0 \leq 1 - \eta\mu \leq 1$ since $\mu \leq \beta_S$ as can be observed in equation 62. Therefore,

$$\Delta \leq \frac{2\lambda B_J}{\mu} \left(1 - (1 - \eta\mu)^{T_S}\right) \widehat{D}_{\mathrm{TV}}(p_1, p_0) \leq \frac{2\lambda B_J}{\mu} \widehat{D}_{\mathrm{TV}}(p_1, p_0), \tag{69}$$

which completes the proof.

C Proof of Theorem 5.3

This appendix provides a detailed proof of Theorem 5.3. Recall our notation from Section 3: the full training set is $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, an unlearning request specifies a forget subset $\mathcal{D}_u \subseteq \mathcal{D}$ with $|\mathcal{D}_u| = m$, and the retained set is $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_u$ with $|\mathcal{D}_r| = n - m$. The teacher training procedure is the randomized algorithm $\mathcal{A}_T$ instantiated as SGD under Assumption 4.1 (or equivalently, Assumption A.1), producing $\boldsymbol{\phi}^* = \mathcal{A}_T(\mathcal{D})$ and $\boldsymbol{\phi}_r^* = \mathcal{A}_T(\mathcal{D}_r)$, with corresponding predictors $f_{\boldsymbol{\phi}^*}$ and $f_{\boldsymbol{\phi}_r^*}$.

In the following, we measure the teacher output discrepancy under m deletions. Specifically, we study the teacher output discrepancy (in terms of $\widehat{D}_{\mathrm{TV}}$) between the original teacher $f_{\boldsymbol{\phi}^*}$ trained on $\mathcal{D}$ and the (counterfactual) retrained from scratch teacher $f_{\boldsymbol{\phi}_r^*}$ trained on $\mathcal{D}_r$. Our argument proceeds as follows. We first couple the SGD trajectories on $\mathcal{D}$ and $\mathcal{D}_r$ to derive stability bounds for m-deletion, by building on the trajectory-coupling analysis for SGD stability Hardt et al. (2016). We then apply an exponential-moment concentration bound to obtain a high-probability uniform-stability guarantee on the change in the teacher loss. We then convert this loss bound into a total-variation bound on teacher predictive distributions.

C.1 STABILITY OF NON-CONVEX SGD UNDER m DELETIONS

Before providing our proof, we state a few useful preliminaries. For a dataset $\mathcal{D} = (z_1, \ldots, z_n)$, consider the empirical teacher objective given by the average of per-example losses,

$$F_S(\boldsymbol{\phi}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\phi}; z_i) \tag{70}$$

and let $\mathcal{A}_T$ run SGD for T iterations,

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t - \alpha_t \nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}_t; Z_t) \tag{71}$$

where Z_t is sampled independently and uniformly at random from $\mathcal{D}$ across the iterations $t = 1, \ldots, T$. The following analysis demonstrates a high-probability bound for the teacher discrepancy under m deletions. This proof closely follows the trajectory coupling arguments and high-probability analysis for SGD stability Hardt et al. (2016); Kuzborskij & Lampert (2018); Madden et al. (2020); Yuan & Li (2023), and is summarized below for completeness.

Proof. To compare SGD on the full dataset $\mathcal{D}$ and on the reduced dataset $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_u$, we follow a similar approach to the trajectory coupling of (Hardt et al., 2016, Theorem 3.12), and analyze the two iterations over $\mathcal{D}$ and $\mathcal{D}_r$ under a coupled random trajectory. Specifically, let $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T$ denote the SGD trajectory on $\mathcal{D}$, and $\boldsymbol{\phi}'_1, \ldots, \boldsymbol{\phi}'_T$ denote the SGD trajectory on $\mathcal{D}_r$, through T training rounds, both starting from the same initialization $\boldsymbol{\phi}_0 = \boldsymbol{\phi}'_0$. At iteration t, let Z_t denote the sample drawn uniformly at random from $\mathcal{D}$. Then, define,

$$Z'_t = \begin{cases} Z_t & \text{if } Z_t \in \mathcal{D}_r \\ \widetilde{Z}_t & \text{if } Z_t \in \mathcal{D}_u \end{cases} \tag{72}$$

where $\widetilde{Z}_t$ is an auxiliary sample drawn uniformly from $\mathcal{D}_r$, and independently from Z_t. Note that Z'_t satisfies a uniform marginal distribution across $\mathcal{D}_r$. Specifically, for any $z' \in \mathcal{D}_r$,

$$\Pr(Z'_t = z') = \sum_{z \in \mathcal{D}_r} \Pr(Z'_t = z' | Z_t = z) \Pr(Z_t = z) + \sum_{z \in \mathcal{D}_u} \Pr(Z'_t = z' | Z_t = z) \Pr(Z_t = z) \tag{73}$$

$$= \Pr(Z_t = z') + \Pr(\widetilde{Z}_t = z') \Pr(Z_t \in \mathcal{D}_u) \tag{74}$$

$$= \frac{1}{n - m} \tag{75}$$

hence each sample is i.i.d. from the uniform distribution over $\mathcal{D}_r$. The corresponding update functions for the two coupled trajectories can then be written as,

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t - \alpha_t \nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}_t; Z_t) \tag{76}$$

$$\boldsymbol{\phi}'_{t+1} = \boldsymbol{\phi}'_t - \alpha_t \nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}'_t; Z'_t) \tag{77}$$

where we also define the L_2-distance between the updates produced by the two trajectories,

$$d_t := \|\boldsymbol{\phi}_t - \boldsymbol{\phi}'_t\|_2. \tag{78}$$

Now, define a Bernoulli random variable to indicate whether a forget sample is selected at iteration t,

$$X_t := \mathbf{1}[Z_t \in \mathcal{D}_u] \tag{79}$$

hence $\Pr(X_t = 1) = m/n$, where $\mathbf{1}[\cdot]$ is an indicator function that takes a value of 1 if $Z_t \in \mathcal{D}_u$ and 0 otherwise. Then, using a loss function satisfying Assumption A.1, one can observe the following "growth recursion" (Hardt et al., 2016, Lemma 2.5),

$$d_{t+1} \leq \begin{cases} (1 + \alpha_t \beta) d_t & \text{if } X_t = 0 \quad \text{iterate over the same sample} \\ d_t + 2\alpha_t L & \text{if } X_t = 1 \quad \text{otherwise} \end{cases} \tag{80}$$

where the case for $X_t = 0$ follows from β-smoothness,

$$d_{t+1} = \|\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}'_{t+1}\|_2 \tag{81}$$

$$= \|\boldsymbol{\phi}_t - \boldsymbol{\phi}'_t - \alpha_t (\nabla_{\boldsymbol{\phi}_t} \ell(\boldsymbol{\phi}_t; Z_t) - \nabla_{\boldsymbol{\phi}'_t} \ell(\boldsymbol{\phi}'_t; Z_t))\|_2 \tag{82}$$

$$\leq \|\boldsymbol{\phi}_t - \boldsymbol{\phi}'_t\|_2 + \alpha_t \|\nabla_{\boldsymbol{\phi}_t} \ell(\boldsymbol{\phi}_t; Z_t) - \nabla_{\boldsymbol{\phi}'_t} \ell(\boldsymbol{\phi}'_t; Z_t)\|_2 \tag{83}$$

$$\leq \|\boldsymbol{\phi}_t - \boldsymbol{\phi}'_t\|_2 + \alpha_t \beta \|\boldsymbol{\phi}_t - \boldsymbol{\phi}'_t\|_2 \tag{84}$$

and the case for $X_t = 1$ follows from L-Lipschitzness,

$$d_{t+1} = \|\boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}'_{t+1}\|_2 \tag{85}$$
$$\leq \|\boldsymbol{\phi}_t - \boldsymbol{\phi}'_t\|_2 + \alpha_t(\|\nabla_{\boldsymbol{\phi}_t}\ell(\boldsymbol{\phi}_t; Z_t)\|_2 + \|\nabla_{\boldsymbol{\phi}'_t}\ell(\boldsymbol{\phi}'_t; Z'_t)\|_2) \tag{86}$$
$$\leq d_t + 2\alpha_t L \tag{87}$$

From equation 80, one can then observe,

$$d_{t+1} \leq (1 + \alpha_t\beta)d_t + 2\alpha_t L X_t \tag{88}$$

Next, for each $t_0 \in \{0, \ldots, T\}$, define the event that the SGD trajectories coincide up to time t_0,

$$E_{t_0} \coloneqq \mathbf{1}[d_t = 0 \text{ for all } \ t \leq t_0] \tag{89}$$

and note that when $t_0 = 0$, by definition $d_{t_0} = 0$ hence $\Pr(E^c_{t_0}) = 0$, and when $t_0 \in \{1, \ldots, T\}$,

$$q_{t_0} \coloneqq \Pr(E^c_{t_0}) = 1 - \left(1 - \frac{m}{n}\right)^{t_0} \tag{90}$$

since X_t is distributed i.i.d. Bernoulli with parameter m/n.

Under the event E_{t_0}, the two trajectories are identical up to round t_0 and the first encounter of a sample from the forget set $\mathcal{D}_u$ occurs after t_0. Then, unrolling equation 88 from iteration $t_0 + 1$ through T leads to,

$$d_{T+1} \leq 2L \sum_{t=t_0+1}^{T} \alpha_t \left(\prod_{k=t+1}^{T}(1 + \alpha_k\beta)\right) X_t \tag{91}$$

and since the loss function ℓ is also L-Lipschitz, for every z the final deviation (conditioned on E_{t_0}) between the two trajectories can be bounded as,

$$\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| \leq L\|\boldsymbol{\phi}_{T+1} - \boldsymbol{\phi}'_{T+1}\|_2 \leq \sum_{t=t_0+1}^{T} a_t X_t \tag{92}$$

where

$$a_t \coloneqq 2L^2\alpha_t \prod_{k=t+1}^{T}(1 + \alpha_k\beta) \tag{93}$$

Recall that $q_{t_0} = \Pr(E^c_{t_0})$. Next, consider any $q_{t_0} < \gamma$, noting that by definition $q_{t_0} = 0 < \gamma$ hence there is at least one admissible t_0. Then, for any given $\varepsilon_{t_0} > 0$, the failure probability for any t_0 satisfying $q_{t_0} < \gamma$ can be characterized as follows,

$$\Pr(\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| > \varepsilon_{t_0}) \tag{94}$$
$$= \Pr(\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| > \varepsilon_{t_0}|E_{t_0}) \Pr(E_{t_0})$$
$$+ \Pr(\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| > \varepsilon_{t_0}|E^c_{t_0}) \Pr(E^c_{t_0}) \tag{95}$$
$$\leq (1 - q_{t_0}) \Pr(\sum_{t=t_0+1}^{T} a_t X_t > \varepsilon_{t_0}|E_{t_0}) + q_{t_0} \tag{96}$$

where equation 95 follows from equation 92.

Since the failure probability has a total budget of γ, from equation 96 we further observe that,

$$(1 - q_{t_0}) \Pr(\sum_{t=t_0+1}^{T} a_t X_t > \varepsilon_{t_0}|E_{t_0}) + q_{t_0} \leq \gamma \tag{97}$$

or equivalently,

$$\Pr(\sum_{t=t_0+1}^{T} a_t X_t > \varepsilon_{t_0}|E_{t_0}) \leq \frac{\gamma - q_{t_0}}{1 - q_{t_0}} \tag{98}$$

Note that conditioned on E_{t_0}, the random variables X_t for $t > t_0$ are distributed independently from a Bernoulli distribution with parameter m/n. Then, from the moment generating function of the sum of independent Bernoulli random variables, we have for every θ,

$$\mathbb{E}\left[\exp\left(\theta \sum_{t_0+1}^{T} a_t X_t\right)\Big| E_{t_0}\right] = \prod_{t=t_0+1}^{T}\left(1 - \frac{m}{n} + \frac{m}{n}e^{\theta a_t}\right) \tag{99}$$

using which and by applying the Chernoff bound, for every $\varepsilon_{t_0} \geq 0$,

$$\Pr\left(\sum_{t_0+1}^{T} a_t X_t > \varepsilon_{t_0}\Big| E_{t_0}\right) \leq \inf_{\theta>0} \exp\left(-\theta\varepsilon_{t_0} + \sum_{t=t_0+1}^{T} \log\left(1 - \frac{m}{n} + \frac{m}{n}e^{\theta a_t}\right)\right) \tag{100}$$

which quantifies the failure probability conditioned on the event E_{t_0}.

Then, note that setting,

$$\varepsilon_{t_0} = \inf_{\theta>0} \frac{\log\left(\frac{1-q_{t_0}}{\gamma - q_{t_0}}\right) + \sum_{t=t_0+1}^{T} \log\left(1 - \frac{m}{n} + \frac{m}{n}e^{\theta a_t}\right)}{\theta} \tag{101}$$

in equation 100 satisfies the right hand side of equation 98. This satisfies the second term in the right hand side of equation 9.

The first term follows from the fact that the teacher loss ℓ is L-Lipschitz. Note that conditioned on E_{t_0}, the two trajectories satisfy $\boldsymbol{\phi}_{t_0+1} = \boldsymbol{\phi}'_{t_0+1}$. Hence, the final distance between the two trajectories can be bounded by using the triangle inequality,

$$\begin{aligned}
\|\boldsymbol{\phi}_{T+1} - \boldsymbol{\phi}'_{T+1}\|_2 &= \|\boldsymbol{\phi}_{T+1} - \boldsymbol{\phi}_{t_0+1} + \boldsymbol{\phi}_{t_0+1} - \boldsymbol{\phi}'_{t_0+1} + \boldsymbol{\phi}'_{t_0+1} - \boldsymbol{\phi}'_{T+1}\|_2 && (102)\\
&\leq \|\boldsymbol{\phi}_{T+1} - \boldsymbol{\phi}_{t_0+1}\| + \|\boldsymbol{\phi}'_{t_0+1} - \boldsymbol{\phi}'_{T+1}\|_2 && (103)\\
&\leq 2L \sum_{t=t_0+1}^{T} \alpha_t && (104)
\end{aligned}$$

where the second inequality follows from $\|\boldsymbol{\phi}_{t_0+1} - \boldsymbol{\phi}'_{t_0+1}\| = 0$. Also from L-Lipschitzness, the loss difference can be bounded for every z as,

$$\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| \leq 2L^2 \sum_{t=t_0+1}^{T} \alpha_t := \overline{\varepsilon}_{t_0} \tag{105}$$

Note that (conditioned on E_{t_0}) equation 105 is a deterministic bound. Then, the overall failure probability satisfies,

$$\begin{aligned}
&\Pr(\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| > \overline{\varepsilon}_{t_0}) && (106)\\
&\quad = \Pr(\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| > \overline{\varepsilon}_{t_0} | E_{t_0}) \Pr(E_{t_0}) \\
&\qquad\qquad + \Pr(\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| > \overline{\varepsilon}_{t_0} | E^c_{t_0}) \Pr(E^c_{t_0}) && (107)\\
&\quad \leq 0 \times (1 - q_{t_0}) + q_{t_0} && (108)\\
&\quad < \gamma && (109)
\end{aligned}$$

where the last inequality follows from the fact that $q_{t_0} < \gamma$.

Since both equation 97 and equation 109 hold for any t_0 such that $q_{t_0} < \gamma$, we define the final stability constant as,

$$\varepsilon_{\text{stab}} = \min_{t_0 \in \{0,\ldots,T\}: q_{t_0} < \gamma} \{\overline{\varepsilon}_{t_0}, \varepsilon_{t_0}\} \tag{110}$$

which leads to the final teacher stability bound,

$$\Pr(\|\ell(\boldsymbol{\phi}_{T+1}; z) - \ell(\boldsymbol{\phi}'_{T+1}; z)\| \leq \varepsilon_{\text{stab}}) \geq 1 - \gamma \tag{111}$$

under m deletions.

□

C.2 From stability to total variation distance of teacher outputs

We now show the relation between the stability of the teacher training loss and the teacher-output discrepancy $\varepsilon_{\mathrm{TV}}$ in equation 9. Starting from the cross entropy loss,

$$\ell(\boldsymbol{\phi}; (\boldsymbol{x}, y)) \coloneqq -\log\big(f_{\boldsymbol{\phi}}(\boldsymbol{x})_y\big) \tag{112}$$

where $f_{\boldsymbol{\phi}}(\boldsymbol{x})_y$ is the y^{th} index of the simplex $f_{\boldsymbol{\phi}}(\boldsymbol{x})$. In the sequel we assume $f_{\boldsymbol{\phi}}(\boldsymbol{x})_y > 0$ for all classes, as is the case for softmax teacher outputs. Under the event from equation 111, for any fixed input $x \in \mathcal{D}_r$ and label $y \in \{1, \dots, K\}$,

$$\big|\ell(\boldsymbol{\phi}^*; (\boldsymbol{x}, y)) - \ell(\boldsymbol{\phi}_r^*; (\boldsymbol{x}, y))\big| = \left|\log \frac{f_{\boldsymbol{\phi}^*}(\boldsymbol{x})_y}{f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})_y}\right| \leq \varepsilon_{\mathrm{stab}} \tag{113}$$

from which one can observe,

$$e^{-\varepsilon_{\mathrm{stab}}} \leq \frac{f_{\boldsymbol{\phi}^*}(\boldsymbol{x})_y}{f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})_y} \leq e^{\varepsilon_{\mathrm{stab}}} \tag{114}$$

We next relate the bound from equation 114 into a bound on the TV distance between the two teacher predictions. The following inequality is motivated by the TV conversion in (Kairouz et al., 2015, Appendix A, Remark A.1).

$$\frac{|f_{\boldsymbol{\phi}^*}(\boldsymbol{x})_y - f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})_y|}{|f_{\boldsymbol{\phi}^*}(\boldsymbol{x})_y + f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})_y|} \leq \frac{e^{\varepsilon_{\mathrm{stab}}} - 1}{e^{\varepsilon_{\mathrm{stab}}} + 1} = \tanh\Big(\frac{\varepsilon_{\mathrm{stab}}}{2}\Big) \tag{115}$$

Then, from the definition of total variation distance,

$$D_{\mathrm{TV}}(f_{\boldsymbol{\phi}^*}(\boldsymbol{x}), f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})) = \frac{1}{2}\sum_{y=1}^{K} |f_{\boldsymbol{\phi}^*}(\boldsymbol{x})_y - f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})_y| \tag{116}$$

$$\leq \frac{1}{2}\tanh\Big(\frac{\varepsilon_{\mathrm{stab}}}{2}\Big)\sum_{y=1}^{K}(f_{\boldsymbol{\phi}^*}(\boldsymbol{x})_y + f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})_y) \tag{117}$$

$$= \tanh\Big(\frac{\varepsilon_{\mathrm{stab}}}{2}\Big) \tag{118}$$

which holds since $\sum_{y=1}^{K} f_{\boldsymbol{\phi}^*}(\boldsymbol{x})_y = \sum_{y=1}^{K} f_{\boldsymbol{\phi}_r^*}(\boldsymbol{x})_y = 1$ by definition of the teacher predictive probabilities.

Finally, averaging equation 118 over the retained set $\mathcal{D}_r$ leads to the final TV bound for the teacher discrepancy under m deletions,

$$\Pr\Big(\widehat{D}_{\mathrm{TV}}(f_{\boldsymbol{\phi}^*}, f_{\boldsymbol{\phi}_r^*}) \leq \tanh\Big(\frac{\varepsilon_{\mathrm{stab}}}{2}\Big)\Big) \geq 1 - \gamma \tag{119}$$

which completes the proof for Theorem 5.3.

D Noise calibration and end-to-end certified unlearning

This appendix provides the detailed arguments behind the final step of Section 5: given an L_2 deviation bound between the student obtained from the original teacher, which is the only one available during unlearning, and the inaccessible retraining reference, which represents the ideal (counterfactual) student model obtained from a teacher retrained using the retained set only, we calibrate the Gaussian parameter perturbation to obtain the certified unlearning guarantee in Definition 3.1. To that end, we first provide the proof of Theorem 5.4, which provides an end-to-end bound for the student parameters. Then, we demonstrate how we can achieve the certified unlearning guarantees stated in Theorem 5.5, by quantifying the Gaussian output perturbation required for a given deviation.

D.1 End-to-end bound on Δ (Proof of Theorem 5.4)

Theorem 5.1 quantifies the impact of teacher discrepancy between any two predictive distributions p_0 and p_1, on the parameter deviation Δ of the corresponding student models. This deviation is characterized in terms of the total variation distance $\widehat{D}_{\mathrm{TV}}(p_1, p_0)$ between the predictive distributions of the two teacher models. Theorem 5.3 further provides a high-probability bound on the teacher discrepancy $\widehat{D}_{\mathrm{TV}}(p_1, p_0)$ where the two teacher models correspond to the ones before and after removing m samples. We now combine these two ingredients to obtain an end-to-end upper bound on the student parameter deviation Δ for the removal of m samples.

From Theorem 5.3, with probability at least $1-\gamma$, the teacher discrepancy (before and after the removal of m samples) satisfies,

$$\widehat{D}_{\mathrm{TV}}(f_{\boldsymbol{\phi}^*}, f_{\boldsymbol{\phi}_r^*}) \leq \varepsilon_{\mathrm{TV}} \tag{120}$$

where $\varepsilon_{\mathrm{TV}} = \tanh\left(\frac{\varepsilon_{\mathrm{stab}}}{2}\right)$ On the other hand, Theorem 5.1 shows that, for any two teacher distributions p_1 and p_0,

$$\Delta \leq 2\lambda B_J G_{T_S} \widehat{D}_{\mathrm{TV}}(p_1, p_0) \tag{121}$$

The result then follows by substituting equation 120 in the right hand side of equation 121, and defining,

$$\overline{\Delta} = 2\lambda B_J G_{T_S} \varepsilon_{\mathrm{TV}} \tag{122}$$

from which it can be observed that, with probability at least $1-\gamma$,

$$\Delta \leq \overline{\Delta} \tag{123}$$

If in addition the strong convexity condition of Corollary 5.2 holds, G_{T_S} can be replaced with $\widetilde{G}_{T_S}$.

D.2 Proof of Theorem 5.5

This is a direct application of the Gaussian mechanism (Dwork & Roth, 2014). Specifically, consider two vectors $\widehat{\boldsymbol{w}}_r, \boldsymbol{w}_r^* \in \mathbb{R}^{d_S}$ such that

$$\|\widehat{\boldsymbol{w}}_r - \boldsymbol{w}_r^*\|_2 \leq \overline{\Delta}. \tag{124}$$

We then calibrate the noise variance σ using the analytic Gaussian mechanism Balle & Wang (2018). By the analytic Gaussian mechanism, if two outputs differ by at most $\overline{\Delta}$ in L_2 norm (i.e., have L_2 sensitivity at most $\overline{\Delta}$), then adding isotropic Gaussian noise with σ ensures (ε, δ)-indistinguishability between the resulting output distributions whenever,

$$\Phi\left(\frac{\overline{\Delta}}{2\sigma} - \frac{\epsilon\sigma}{\overline{\Delta}}\right) - e^{\epsilon}\Phi\left(-\frac{\overline{\Delta}}{2\sigma} - \frac{\epsilon\sigma}{\overline{\Delta}}\right) \leq \delta \tag{125}$$

such that Φ is the cumulative distribution function (CDF) of the standard normal distribution, which leads to Definition 3.1.

Since the bound holds with probability at least $1-\gamma$, applying the Gaussian mechanism on this event and accounting for its complement gives the stated $(\varepsilon, \delta+\gamma)$ guarantee in Theorem 5.5.

E Experimental setup

This appendix gives the details of the experiments in Section 6.

E.1 Data

Synthetic XOR. The data consist of four two-dimensional clusters in an XOR arrangement. Before scaling, the cluster centers are $(\pm 0.6, \pm 0.6)$, and each sample adds noise drawn uniformly from a disk of radius 0.75 around its center. Since this radius exceeds the distance 0.6 from each center to the coordinate axes, neighboring clusters overlap. The overlap keeps the task noisy, so that no model separates the two classes perfectly. The label of a sample is the label of the cluster that generated it, which is generated according to an XOR condition, such that diagonally opposite clusters share a label. Specifically, the samples generated around clusters centered at $(+0.6, +0.6)$, $(-0.6, -0.6)$ have a label $+1$, whereas the samples generated around clusters centered at $(-0.6, +0.6)$, $(+0.6, -0.6)$ have a label -1. A single global scale maps every sample into the unit ball, which places the centers at $(\pm 0.375, \pm 0.375)$ with disk radius 0.469.

MNIST. We use the digits 3 and 8 at their native resolution of 28×28. Each image is centered and then scaled to norm at most 4. As for XOR and CIFAR-10, we bound the input norm because it bounds the gradients of both the teacher and the student, which keeps their regularity constants finite.

CIFAR-10. Each image is resized to 224×224, normalized with the ImageNet statistics, and passed through the ImageNet-pretrained ResNet-18 of He et al. (2016). The 512-dimensional pooled feature is scaled to unit norm. Features are extracted once, and the backbone is never updated, so it holds no information about $\mathcal{D}$ and nothing in it needs to be unlearned.

Splits and forget sets. The forget set is a uniformly random subset of the training set $\mathcal{D}$. For each seed, the forget sets of different sizes are nested prefixes of one fixed random permutation, so the data and the original models are shared across forget ratios, and differences across forget ratios reflect only the number of deleted samples.

E.2 Models and training

Teacher. The teacher is a one-hidden-layer tanh network of width M_T whose hidden weights and biases are trained, while its output layer is fixed. For the binary tasks, the first half of the hidden units has output weight $+1$ and the second half -1, so the teacher computes the score $h(\boldsymbol{x}) = \frac{\rho_T}{\sqrt{M_T}} \sum_{j=1}^{M_T} s_j \tanh(\boldsymbol{\theta}_j^\top \boldsymbol{x} + b_j)$ with these fixed signs $s_j \in \{\pm 1\}$, and its logits are the centered pair $(h/2, -h/2)$. For CIFAR-10, the hidden units are split into ten equal blocks, one per class, the logit of each class sums the units of its block, and the ten sums are centered by subtracting their mean before the same factor $\rho_T/\sqrt{M_T}$ is applied. The output scale ρ_T is chosen such that $\rho_T\sqrt{M_T} = 0.177$. Since each $\tanh$ lies in $[-1, 1]$, this bounds the magnitude of the binary score h. A fixed and small output layer keeps the teacher's logits and their gradients small, which keeps the constants L and β of Assumption 4.1, and hence $\varepsilon_{\mathrm{TV}}$, small, while the trained hidden layer still lets the teacher learn nonlinear decision boundaries. The teacher is trained by single-sample SGD with step size α_t as in Assumption 4.1, without momentum, weight decay, or clipping. The teacher is trained for $T = 32{,}768$ steps by single-sample SGD with the decaying step size $\alpha_t = 1/t$ on XOR and $\alpha_t = 0.3/(t + 1000)$ on MNIST and CIFAR-10, as in Assumption 4.1, without momentum, weight decay, or clipping.

Coupled teachers. The original and retrained teachers start from the same initialization and share one sample sequence. At every step, both use the same draw, except that whenever the original teacher draws a forgotten sample, the retrained teacher draws a uniformly random retained sample instead. Each teacher is therefore an ordinary SGD run on its own data, and the coupling only aligns their randomness. It is the same coupling as in the stability argument behind Theorem 5.3, so the observed discrepancy measures exactly the quantity that the theorem bounds.

E.3 Convex student

The student has a hidden tanh layer of width M_S whose weights and biases are drawn once per seed, independently of the data, and then frozen. Each entry of the hidden weight matrix is drawn independently from $\mathcal{N}(0, s_W^2)$ and each bias from $\mathcal{N}(0, s_b^2)$, with $(s_W, s_b) = (75.3, 64)$ for XOR, $(18.3, 1)$ for MNIST, and $(22.6, 1)$ for CIFAR-10. Only its output layer, multiplied by an output scale $\rho_S = 1/8$ as for the teacher, is trained, starting from zero. Freezing the hidden layer makes the logits linear in the trained weights, so the objective is convex, and ρ_S sets the size of the student's logits and hence its constants B_J and β_S. It minimizes the distillation objective of Equation (1) plus the ridge penalty $\frac{\mu}{2}\|\boldsymbol{w}\|_2^2$ by full-batch gradient descent with a constant step size $\eta < 1/\beta_S$. The ridge penalty makes the objective strongly convex, and the step size makes every gradient step a contraction, which is what Corollary 5.2 requires. All compared students of a seed share the same frozen layer and initialization. The noise is added to the trained output layer only, since the frozen layer does not depend on the data.

E.3.1 Constants and noise calibration

All constants follow from the configuration alone, before any training, and are listed in Table 3. Table 4 gives, for each forget size, the teacher bound $\varepsilon_{\mathrm{TV}}$ of Theorem 5.3, the sensitivity bound $\overline{\Delta}$

Table 2: Teacher and convex student configurations.

	XOR	MNIST	CIFAR-10
Teacher width M_T	1024	1024	1000
Teacher steps T	32768	32768	32768
Student width M_S	512	1000	600
Trained student parameters d_S	512	1000	600
Ridge weight μ	3×10^{-7}	10^{-7}	10^{-7}
Student steps T_S	4096	8192	8192
Learning rate η	255	127	127

of Theorem 5.4, and the noise level σ, which we calibrate with the analytic Gaussian mechanism of Balle & Wang (2018) at $(\varepsilon, \delta + \gamma) = (1, 0.002)$ with $\delta = \gamma = 10^{-3}$. The observed discrepancies never enter the calibration.

Table 3: Regularity constants of the teacher and the convex student.

	XOR	MNIST	CIFAR-10
Teacher Lipschitz constant L	0.0042506	0.012393	0.0031971
Teacher smoothness β	1.5987×10^{-4}	1.3589×10^{-3}	3.3084×10^{-4}
Student Jacobian bound B_J	0.088388	0.088388	0.125
Student smoothness β_S	0.00390655	0.00390635	0.0078126

Table 4: Bounds and noise level for each forget ratio. The values are identical across seeds.

Setting	Forget ratio	$\varepsilon_{\mathrm{TV}}$	$\overline{\Delta}$	σ
XOR	5%	4.088×10^{-5}	0.6480	1.6683
	10%	5.391×10^{-5}	0.8546	2.2002
	20%	7.596×10^{-5}	1.2041	3.1001
MNIST	5%	9.291×10^{-6}	0.3096	0.7972
	10%	1.780×10^{-5}	0.5933	1.5276
	20%	3.449×10^{-5}	1.1493	2.9591
CIFAR-10	5%	6.203×10^{-7}	0.04597	0.11835
	10%	1.189×10^{-6}	0.08809	0.22680
	20%	2.303×10^{-6}	0.17065	0.43936

E.4 Non-convex student

The model configuration is similar to the convex teacher-student setting, except that the student is non-convex because its hidden layer is trained instead. The student is first warm-started (with Adam with a learning rate of $1/16$ across 100 iterations) and pretrained on an independently sampled dataset with the same number of samples as the training set (in Table 5) with the same geometry/task, and the resulting parameters are used to initialize the certified student trajectory. The warm up is not part of the certified trajectory. The teacher uses $T = 32768$ SGD steps, while the certified student uses $T_S = 32$ full-batch gradient descent steps with learning rate $\eta = 1/16$. The short certified trajectory keeps the finite-step amplification factor in the non-convex sensitivity bound controlled. All compared students of a seed use the same initialization and the trajectories are coupled. The calibrated noise is added to the resulting trained student parameters.

E.4.1 Constants and noise calibration

In Table 5, the dataset size, along with the teacher and non-convex student training configurations are listed. The regularity constants are listed in Table 6. Table 7 presents, for each forget ratio, the teacher bound $\varepsilon_{\mathrm{TV}}$ and the student sensitivity bound $\overline{\Delta}$.

Table 5: Dataset size, and teacher and non-convex student configurations.

	XOR	MNIST	CIFAR-10
Training set $\lvert\mathcal{D}\rvert$	8192	8192	8200
Forget set $\lvert\mathcal{D}_f\rvert$	410	410	410
Retain set $\lvert\mathcal{D}_r\rvert$	7782	7782	7790
Evaluation set	4096	1800	2000
Teacher steps T	32768	32768	32768
Student steps T_S	32	32	32
Student learning rate η	$1/16$	$1/16$	$1/16$

Table 6: Regularity constants of the teacher and the non-convex student.

	XOR	MNIST	CIFAR-10
Teacher Lipschitz constant L	0.03266407412	0.01767766953	0.02905694150
Teacher smoothness β	0.01428274476	0.00418331909	0.01142913817
Student Jacobian bound B_J	0.02309698831	0.01250000000	0.02090169944
Student smoothness β_S	0.02589813518	0.00758538817	0.01723327635

After obtaining the CKDU intermediate student, Gaussian noise is added to the student parameters. We use the analytic Gaussian calibration of Balle & Wang (2018).

E.5 Evaluation

Membership inference. We use the MIA attack of Kurmanji et al. (2023). The members are the forgotten samples, and the nonmembers are held-out samples, taken from the test split for XOR and from the validation split for MNIST and CIFAR-10. Each group is split in half. The attack fits a single threshold on the cross-entropy of the true label on one half and reports its balanced accuracy on the other half, so that 50% corresponds to chance. For a noisy model, the threshold is refitted for every noise draw.

Seeds and noise draws. Each configuration runs on three seeds, each fixing the data, the forget set, and the training randomness. The seeds (data, forget, training) are $(11, 1, 1)$, $(12, 1, 2)$, and $(13, 1, 3)$ for XOR, $(21, 21, 1)$, $(22, 22, 2)$, and $(23, 23, 3)$ for MNIST, and $(51, 51, 1)$, $(52, 52, 2)$, and $(53, 53, 3)$ for CIFAR-10. Every noisy model is evaluated on 100 independent noise draws, which are first averaged within a seed. Tables report the mean and the sample standard deviation of these per-seed values across the three seeds. The draws only estimate the distribution of a release, and the guarantee covers a single release.

F Details of the motivating example

This appendix describes the synthetic example of Figure 1, which illustrates how a redistilled student can inherit information about deleted samples from its teacher. We keep the example deliberately simple. Teacher and student are both linear classifiers, so every training objective is convex, and the leakage in panel (c) can be attributed neither to excess model capacity nor to nonconvex training.

Data and deletion request. The data consist of three well-separated clusters in the plane, one per class, with 240 points each, so that $n = 720$. A point of class k is drawn as $\boldsymbol{x} = \boldsymbol{c}_k + 0.12\,\boldsymbol{z}$, where $\boldsymbol{z}$ is a standard Gaussian vector in $\mathbb{R}^2$ truncated to $\|\boldsymbol{z}\|_2 \le 3$ by rejection sampling, and the class centers are

$$\boldsymbol{c}_1 = (-0.40,\, 0.00), \qquad \boldsymbol{c}_2 = (0.50,\, 0.04), \qquad \boldsymbol{c}_3 = (0.04,\, 0.52).$$

The truncation radius does not depend on the data and ensures $\|\boldsymbol{x}\|_2 \le 1$ for every input, as our bounds require. The deletion request removes the entire Class 3, so the forget set $\mathcal{D}_u$ contains all $m = 240$ points of this class and the retained set $\mathcal{D}_r$ contains the 480 points of Classes 1 and 2. We fix the forget set before training any model. Removing a whole class makes the influence of the

Table 7: Bounds for each forget ratio. The theoretical bounds are identical across seeds.

Setting	Forget ratio	$\varepsilon_{\mathrm{TV}}$	$\overline{\Delta}$	σ
XOR	5%	3.75×10^{-2}	9.29×10^{-2}	0.31
	10%	4.88×10^{-2}	1.26×10^{-1}	0.42
	20%	6.35×10^{-2}	1.71×10^{-1}	0.57
MNIST	5%	2.67×10^{-2}	9.48×10^{-2}	0.27
	10%	4.14×10^{-2}	1.52×10^{-1}	0.46
	20%	6.42×10^{-2}	2.44×10^{-1}	0.73
CIFAR-10	5%	3.16×10^{-2}	9.27×10^{-2}	0.21
	10%	4.74×10^{-2}	1.27×10^{-1}	0.34
	20%	7.11×10^{-2}	1.74×10^{-1}	0.51

teacher easy to see, since a student trained only on $\mathcal{D}_r$ has no reason to predict Class 3 unless its teacher suggests it.

Models, training and observations. Both models act on the augmented input $\boldsymbol{u} = (\boldsymbol{x}, 1) \in \mathbb{R}^3$ and output a distribution over all three classes. The teacher is $f_{\boldsymbol{\phi}}(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{\Theta}\boldsymbol{u})$ and the student is $g_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{u})$, where $\boldsymbol{\phi}$ and $\boldsymbol{w}$ collect the class weights and biases in $\boldsymbol{\Theta}, \boldsymbol{W} \in \mathbb{R}^{3\times 3}$, so that $d_T = d_S = 9$. We keep all three outputs after the deletion, so any region that a student still assigns to the deleted class remains visible. The original teacher $f_{\boldsymbol{\phi}^*}$ is trained on $\mathcal{D}$ by single-sample SGD on the cross-entropy with hard labels, with learning rate $\alpha_t = 3/(t+10)$ for $T = 20{,}000$ steps from zero initialization. The retrained teacher $f_{\boldsymbol{\phi}_r^*}$ uses the same schedule and initialization on $\mathcal{D}_r$, and its sample sequence is coupled to that of the original teacher as in Appendix C. Both runs use the same draws, except that each draw of a forget point is replaced by an independent, uniformly random point of $\mathcal{D}_r$.

Each student minimizes the distillation objective in equation 1 on its own training set with $\lambda = 0.9$, plus the ridge term $(\mu/2)\|\boldsymbol{W}\|_F^2$ with $\mu = 10^{-3}$. The ridge term does not depend on the teacher and makes the objective μ-strongly convex, as Corollary 5.2 requires. We run full-batch gradient descent from zero initialization for $T_S = 5{,}000$ steps with the constant step size $\eta = 0.9$, which satisfies $\eta \le 1/\beta_S$ for the smoothness constant β_S. The students in panels (a) to (c) differ only in their training set and their teacher. The original student $\boldsymbol{w}^*$ uses $\mathcal{D}$ and $f_{\boldsymbol{\phi}^*}$, the retrain-from-scratch student $\boldsymbol{w}_r^*$ uses $\mathcal{D}_r$ and $f_{\boldsymbol{\phi}_r^*}$, and the redistilled student $\widehat{\boldsymbol{w}}_r$ uses $\mathcal{D}_r$ and $f_{\boldsymbol{\phi}^*}$. We choose a large distillation weight so that the soft labels of the teacher dominate the hard labels, which makes the leakage in panel (c) clearly visible.

Panel (d) shows one draw of the release $\widetilde{\boldsymbol{w}}_r = \widehat{\boldsymbol{w}}_r + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbf{I}_{d_S})$ and $\sigma = 0.8$. Among $1{,}000$ independent noise draws, we show the first one in sampling order that leaves no Class 3 region in the plotted window and classifies at least 99% of the retained points correctly. This is the sixth draw, and it classifies every retained point correctly. On a 240×240 grid over the window, 5.2% of all draws satisfy both conditions, compared with 53.8% for the noisy retraining reference $\boldsymbol{w}_r^* + \boldsymbol{\xi}'$ with the same σ, where $\boldsymbol{\xi}'$ is an independent copy of $\boldsymbol{\xi}$. Averaged over all draws, the released student still assigns 53.3% of the forget points to Class 3, so panel (d) shows a favorable draw rather than a typical one.

The decision regions in Figure 1 are the argmax labels of each model on a 900×900 grid that covers the range of the data, extended by 0.2 in each direction. Since the logits are affine in $\boldsymbol{x}$, the boundaries between regions are straight line segments. Retained points are colored by class, and forget points are shown in purple in panel (a) and in red in panels (b) to (d), where no student is trained on them. Table 8 summarizes the models. Both teachers classify every retained point correctly, but only the original teacher places substantial probability on Class 3 at the forget points. Even on the retained points, it assigns Class 3 a mean probability of 0.21, compared with 0.02 for the retrained teacher, and this probability grows toward the Class 3 cluster. The redistilled student never sees a forget point, yet it fits these soft labels and extrapolates them linearly, so it assigns 94.2% of the forget points to Class 3. The retrain-from-scratch student assigns none of them to Class 3, and neither does the released draw in panel (d).

Table 8: Models in the motivating example of Figure 1. The last two columns report the fraction of forget points assigned to Class 3 and the mean probability of Class 3 on the forget points.

Model	Training data, teacher	Accuracy on $\mathcal{D}_r$	Class 3 rate on $\mathcal{D}_u$	Mean p_3 on $\mathcal{D}_u$
Original teacher	$\mathcal{D}$	100%	99.2%	0.598
Retrained teacher	$\mathcal{D}_r$	100%	0.0%	0.034
(a) Original student	$\mathcal{D}$, original	100%	99.2%	0.619
(b) Retrain-from-scratch	$\mathcal{D}_r$, retrained	100%	0.0%	0.034
(c) Redistill-on-retain-data	$\mathcal{D}_r$, original	100%	94.2%	0.485
(d) CKDU (shown draw)	$\mathcal{D}_r$, original, $\sigma = 0.8$	100%	0.0%	0.160

Implementation. The example is implemented in PyTorch in double precision. The data, the SGD sampling, and the release noise use separate fixed random streams, so rerunning the same configuration on the same hardware reproduces the figure exactly.

G Comparison with unlearning baselines on deep networks

The experiments in Section 6 use one-hidden-layer tanh networks, for which every constant in our bounds has a closed form. This appendix asks how CKDU behaves with deep convolutional teachers and students, and how it compares with existing unlearning methods that apply directly to the student. For such networks, the regularity constants of Assumptions 4.1 and 4.2 have no closed form. Following prior practice for deep networks (Zhang et al., 2024a), we treat them as hyperparameters, so the resulting release carries no formal certificate, and this appendix evaluates empirical unlearning only.

Setup. We use a class-balanced subset of $n = 10{,}000$ CIFAR-10 training images (Krizhevsky, 2009) and delete $m = 1{,}000$ of them uniformly at random, a forget ratio of 10%. A further 2,000 training images form a validation set for hyperparameter selection, and all test metrics use the full CIFAR-10 test set. The teacher is a ResNet-50 and the student a ResNet-18 (He et al., 2016), both trained with random cropping and horizontal flipping. The teacher is trained with SGD (momentum 0.9, weight decay 5×10^{-4}, batch size 128) for 60 epochs, that is $T = 4{,}740$ steps, with the decaying schedule $\alpha_t = 100/(t + 1000)$. The original and retrained teachers share their initialization and minibatch sequence, and the retrained teacher simply drops the samples of $\mathcal{D}_u$ from each minibatch, so both teachers take the same T steps. The student minimizes equation 1 with $\lambda = 0.5$ by SGD with the same momentum, weight decay, and batch size for 50 epochs, with a cosine schedule that starts at $\eta = 0.05$, which gives $T_S = 3{,}550$ steps on $\mathcal{D}_r$. All students share one initialization, and Retrain and CKDU also share minibatches and augmentations, so they differ only in the teacher. We report the mean and standard deviation over three seeds, where each seed redraws the subset, the forget set, and all initializations.

Noise calibration. Since the constants are declared rather than derived, we evaluate each bound with declared values. Theorem 5.3 with the actual teacher schedule, $m, n, \gamma = 10^{-3}$, and the declared constants $L = 0.1$ and $\beta = 10^{-3}$ gives $\varepsilon_{\mathrm{TV}} = 0.224$. For the student, we evaluate G_{T_S} in equation 5 with the peak learning rate $\eta = 0.05$ at every step, which upper bounds its value under the cosine schedule, together with the declared $\beta_S = 10^{-2}$, which gives $G_{T_S} = 489.8$. Rather than fixing the Jacobian constant B_J, we sweep it over $\{3, 6, 15, 30, 60\} \times 10^{-6}$. Since $\overline{\Delta}$ in equation 11 and σ scale linearly with B_J, this sweep traces the trade-off between noise and utility. The noise level σ follows from Theorem 5.5 with the analytic Gaussian mechanism at $(\varepsilon, \delta) = (1, 10^{-3})$, so the full budget is $(\varepsilon, \delta + \gamma) = (1, 2 \times 10^{-3})$.

Baselines. We compare with standard unlearning methods that apply directly to the student. All of them start from the original student $\boldsymbol{w}^*$, and whenever a baseline trains on $\mathcal{D}_r$, it uses the distillation objective with the original teacher, so it has the same black-box teacher access as CKDU.

- **D2D** (Neel et al., 2021) continues gradient descent on $\mathcal{D}_r$ for $\{2, 5, 10, 20\}$ epochs.

Table 9: Comparison with unlearning baselines on a CIFAR-10 subset with a ResNet-50 teacher and ResNet-18 student, at a forget ratio of 10%. Mean and standard deviation over three seeds. MIA is the confidence ROC-AUC (%), where 50 is chance.

Method	Test	Forget	Retain	MIA
Original	80.25 ± 0.05	94.13 ± 0.39	94.80 ± 0.37	55.50 ± 0.84
Retrain	79.37 ± 0.27	78.53 ± 0.33	95.26 ± 0.54	49.00 ± 0.66
Retrain + noise	78.28 ± 0.64	77.09 ± 0.62	93.52 ± 1.31	48.85 ± 0.66
CKDU	78.47 ± 0.07	78.49 ± 0.63	93.46 ± 0.39	49.71 ± 0.92
Redistill-only	79.32 ± 0.69	82.93 ± 0.71	95.03 ± 0.50	53.72 ± 0.82
D2D	78.35 ± 0.60	85.73 ± 0.97	94.02 ± 0.67	52.56 ± 0.95
R2D-5	80.49 ± 0.13	93.93 ± 0.29	94.89 ± 0.34	55.36 ± 0.83
R2D-20	80.23 ± 0.26	87.27 ± 0.12	95.05 ± 0.41	52.50 ± 0.82
NegGrad	78.10 ± 0.07	80.97 ± 0.29	92.58 ± 0.13	51.83 ± 1.34
RandRelabel	78.69 ± 0.29	89.70 ± 0.73	94.18 ± 0.53	53.52 ± 0.93
SalUn	78.10 ± 0.35	87.80 ± 0.80	93.82 ± 0.24	53.53 ± 1.13
SCRUB	77.26 ± 1.68	84.37 ± 3.30	88.99 ± 3.27	52.20 ± 0.86

- **R2D** (Mu & Klabjan, 2025) rewinds the student to an earlier checkpoint and continues the original schedule on $\mathcal{D}_r$. R2D-5 and R2D-20 rewind to epochs 45 and 30 of 50 and train for 5 and 20 epochs, respectively.
- **NegGrad** (Golatkar et al., 2020) minimizes the retained-set loss minus w times the cross-entropy on $\mathcal{D}_u$, with $w \in \{0.05, 0.1, 0.2\}$, for 5 epochs.
- **RandRelabel** (Golatkar et al., 2020) adds to the retained-set loss the cross-entropy toward fixed random wrong labels on $\mathcal{D}_u$, weighted by $m/|\mathcal{D}_r|$ to match one pass over the union of both sets, for 5 epochs with learning rate in $\{0.001, 0.003, 0.01, 0.03\}$.
- **SalUn** (Fan et al., 2023) applies RandRelabel only to the half of the weights with the largest forget-loss gradient magnitude, with the same grid.
- **SCRUB** (Kurmanji et al., 2023) uses the original student as a frozen teacher, alternating max-steps on $\mathcal{D}_u$ in the first 2 of 5 epochs with min-steps on $\mathcal{D}_r$, with learning rate in $\{0.002, 0.005, 0.01, 0.02\}$.

All fine-tuning baselines use SGD with momentum 0.9, weight decay 5×10^{-4}, and gradient clipping at norm 5.

Results. Table 9 summarizes the comparison. CKDU matches Retrain on the forget set, with forget accuracy 78.49 against 78.53 and MIA 49.71 against 49.00, and it stays close to Retrain + noise on every accuracy. Every baseline leaves the forget accuracy well above that of Retrain, from 80.97 for NegGrad to 93.93 for R2D-5, and D2D, R2D, and SCRUB also keep the MIA between 52 and 56.

H Deep teacher with monitored regularity constants

The one-hidden-layer tanh networks of Section 6 admit closed-form regularity constants, while Appendix G treats the constants of deep networks as hyperparameters that are set apriori. This appendix takes a middle path and estimates the constants of a deep teacher along its own training trajectory. We note, however, that since these estimations are evaluated using the empirical training trajectory, they are data-dependent, hence we do not claim formal certification in the following ablations. Rather, these ablations are performed to provide further insight on the teacher discrepancy and student deviation behavior at scale. Here, the teacher is a ResNet-56 (He et al., 2016) fine-tuned end-to-end on CIFAR-10 images (Krizhevsky, 2009), and we evaluate Theorem 5.3 at the monitored values of its constants. The student keeps the convex construction of Appendix E.2 on top of frozen public features, so all of its constants remain exact. Since the monitored constants can underestimate the worst-case constants of Assumption 4.1, the resulting release simulates the guarantee of Theorem 5.5 rather than certifying it. As in Section 6, the retrained teacher and student serve solely as offline evaluation oracles.

Data. We split the CIFAR-10 training set into three disjoint, class-balanced pools. A public pool of 30,000 images is used only for pretraining, a validation set of 10,000 images supplies the nonmembers of the MIA, and a pool of 10,000 images provides the training set $\mathcal{D}$ of $n = 1{,}000$ images. The forget set $\mathcal{D}_u$ is a uniformly random subset of $m = 100$ training images, a forget ratio of 10%, and all test metrics use the CIFAR-10 test set. Images are normalized with the CIFAR-100 channel statistics of the pretrained networks below, without augmentation.

Teacher. The teacher is a ResNet-56 initialized from a public CIFAR-100 checkpoint with a new ten-class output layer. We replace each ReLU by the softplus $x \mapsto \frac{1}{10}\log(1 + e^{10x})$, which is twice differentiable, and freeze the batch-normalization statistics, so that each per-sample loss is a fixed and smooth function of the parameters, as Assumption 4.1 requires. The teacher is first pretrained on the public pool for 8 epochs with SGD (momentum 0.9, learning rate 10^{-3}, batch size 128), which gives between 84.7% and 86.5% validation accuracy across seeds. Since the public pool is disjoint from $\mathcal{D}$, this pretrained network is a data-independent initialization shared by the original and retrained teachers. Both teachers are then fine-tuned end to end, updating all 855,770 parameters, with the coupled single-sample SGD of Appendix E.2 for $T = 1{,}000$ steps.

We monitor three constants along the original teacher's run only, at 18 checkpoints, namely the initialization, steps $1, 2, 4, \ldots, 64$, every 100 steps, and step 1,000. The Lipschitz constant $\hat{L}$ is the largest per-sample gradient norm of the loss $-\log p_y(\boldsymbol{x})$, where $p = f_{\boldsymbol{\phi}}(\boldsymbol{x})$, and it also covers the sample used at every SGD step. The smoothness constant $\hat{\beta}$ is the largest absolute eigenvalue of a per-sample Hessian, estimated by power iteration with Hessian-vector products. Since the total variation bound requires every log-probability to be Lipschitz, not only the loss, we also monitor $\hat{L}_{\mathrm{out}} = \max_k \|\nabla_{\boldsymbol{\phi}} \log p_k(\boldsymbol{x})\|_2$ from the Jacobian of the logits. The per-sample quantities are evaluated on a fixed monitor set, which here consists of the forget set, and we use their running maxima over all checkpoints. Since monitored maxima can miss the worst case, we also report a conservative variant that doubles all three constants.

We evaluate Theorem 5.3 with the actual step sizes and the monitored constants, replacing $2L^2$ by $2\hat{L}\hat{L}_{\mathrm{out}}$ in a_t and in the deterministic term of $\varepsilon_{\mathrm{stab}}$, so that $\varepsilon_{\mathrm{stab}}$ bounds the change of every log-probability rather than only that of the loss. Using $\hat{L}$ in place of $\hat{L}_{\mathrm{out}}$ would change $\varepsilon_{\mathrm{TV}}$ by between -21% and $+4\%$ across seeds. On every seed, the minimum in $\varepsilon_{\mathrm{stab}}$ is attained by the Chernoff term.

Student. The student applies the convex construction of Appendix E.2 to frozen public features. A ResNet-20 from the same collection of CIFAR-100 checkpoints, with the same softplus and frozen normalization, maps each image to its 64-dimensional pooled feature, which we scale to unit norm. It is followed by a frozen random tanh layer of width $M_S = 1{,}000$, with weight standard deviation 32 and bias standard deviation 1, and only the output layer, with $d_S = 10{,}000$ parameters and $\rho_S = 1/8$, is trained, starting from zero. The student minimizes equation 1 with $\lambda = 0.3$ plus the ridge penalty $\frac{\mu}{2}\|\boldsymbol{w}\|_2^2$ with $\mu = 10^{-7}$ by full-batch gradient descent for $T_S = 8{,}192$. Its constants are therefore exact, $B_J = 0.125$ and $\beta_S = 0.0078126$. The frozen ResNet-20 has about a third of the teacher's parameters and is never trained on $\mathcal{D}$, so the student stays smaller than the teacher and learns about $\mathcal{D}$ only through its output layer.

Calibration and seeds. We set $\overline{\Delta}$ by equation 11 with $\widetilde{G}_{T_S}$ as in Theorem 5.4, and calibrate σ with Theorem 5.5 and the analytic Gaussian mechanism at $(\varepsilon, \delta + \gamma) = (1, 0.002)$ with $\delta = \gamma = 10^{-3}$, which gives $\overline{\Delta}/\sigma = 0.388$. We fixed the configuration on a development seed using the validation split only, and chose the teacher's step size by screening the bound rather than the observed discrepancies which is 10^{-4}. We report three fresh seeds (data, forget, training), $(71, 71, 1)$, $(72, 72, 2)$, and $(73, 73, 3)$, each of which redraws the data pools, the forget set, the public pretraining, and the SGD sample order.

Results. Table 10 summarizes the results. The original student fits the forget set far better than unseen data, with 96.67 forget accuracy against 67.66 test accuracy, and the attack detects the forget samples with 66.01% balanced accuracy. CKDU matches Retrain + noise on every metric, with seed means within 0.01 points and single seeds within 0.36 points, and the attack falls to chance.

Table 10: Monitored simulation with a ResNet-56 teacher and a convex student on frozen ResNet-20 features, at a forget ratio of 10%. Mean and standard deviation over three seeds. The last two rows double the monitored teacher constants.

Method	Test	Forget	Retain	MIA
Original	67.66 ± 0.19	96.67 ± 0.58	97.22 ± 0.78	66.01 ± 2.18
Retrain	67.43 ± 0.17	67.33 ± 1.53	98.00 ± 0.78	50.02 ± 0.73
Retrain + noise	67.22 ± 0.20	67.42 ± 2.10	97.64 ± 0.63	50.08 ± 0.52
CKDU	67.21 ± 0.22	67.43 ± 1.82	97.64 ± 0.65	50.07 ± 0.56
Retrain + noise, $2\times$ constants	64.67 ± 1.04	65.61 ± 1.47	93.14 ± 2.05	49.93 ± 0.70
CKDU, $2\times$ constants	64.56 ± 1.05	65.37 ± 1.66	93.06 ± 2.11	49.97 ± 0.77

Table 11: Monitored teacher constants, bounds, and observed discrepancies for each seed. The teacher test accuracy is identical for the original and retrained teachers. The last column doubles the monitored constants.

Seed	Teacher test	$\hat{L}$	$\hat{L}_{\mathrm{out}}$	$\hat{\beta}$	$\varepsilon_{\mathrm{TV}}$	$\widehat{D}_{\mathrm{TV}}(p_1, p_0)$	$\overline{\Delta}$	Δ	σ	$\sigma, 2\times$
71	83.86	490.0	520.3	32,056	6.99×10^{-6}	6.41×10^{-10}	0.517	6.15×10^{-6}	1.33	5.32
72	85.85	410.9	517.7	20,295	5.83×10^{-6}	1.82×10^{-12}	0.431	4.06×10^{-8}	1.11	4.44
73	85.97	567.6	545.8	28,623	8.49×10^{-6}	2.60×10^{-11}	0.628	5.23×10^{-7}	1.62	6.47

Table 11 compares the bounds with the observed discrepancies, and every observed value stays below its bound on every seed.