RETAIN-FORGET LIMITS AND TRADEOFFS IN MACHINE UNLEARNING WITH PARTIAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning is commonly framed as an approximation to retraining on the retained data. While prior literature typically characterizes unlearning as a tradeoff between computational efficiency and accuracy, this work reframes the task as an optimization problem constrained by *partial information* about the training set, such as the original model and local Hessian of the training loss, but not the full retained data. We formalize the resulting ambiguity through the set of feasible datasets consistent with the same partial information, and define *retain-set excess risk* and *forget-set unlearning risk*, which measure the gap between an unlearned model and the retraining oracle on the retained and forgetting data respectively. For least-squares regression, we decompose both risks into algorithm-dependent fitting terms and algorithm-independent ambiguity floors. We then construct a two-layer ReLU example, showing that both ambiguity floors scale as $\Omega(k)$, where k is the forget-set size, even when the algorithm utilizes the original model, the forget set, and local Hessian. We further show that the retain and forget fitting objectives need not share a minimizer, yielding a Pareto tradeoff between the two goals. Finally, for representative first- and second-order unlearning methods, we derive matching $O(k)$ risk upper bounds under suitable regularity conditions and analyze their tradeoff between retain and forget.

1 INTRODUCTION

Modern machine learning systems are trained on large-scale datasets that often contain sensitive or user-specific information. Regulations such as GDPR (Voigt & Bussche, 2017) and CCPA (Goldman, 2020) require models to remove the influence of specific samples upon request, motivating the rapidly growing field of *machine unlearning* (Bourtoule et al., 2021; Xu et al., 2023; Nguyen et al., 2025), whose goal is to eliminate the influence of a forgetting dataset from a trained model.

The natural gold standard is *retraining from scratch* on the retained dataset, which provably removes the deleted data but is computationally prohibitive and requires full access to the retained data. Consequently, the literature has largely framed machine unlearning as a three-way tradeoff between efficiency, efficacy, and accuracy (Hayes et al., 2025), and has focused on *approximate unlearning* algorithms (Guo et al., 2020; Golatkar et al., 2020; Izzo et al., 2021; Graves et al., 2021; Thudi et al., 2022) that approximate the retrained model at much lower computational cost.

We argue that this efficiency-centric view is incomplete and propose a complementary perspective: unlearning should be regarded as an *optimization problem constrained by partial information*. This limitation originates from two main sources: (1) **Computational efficiency.** Approximate unlearners such as Newton's method (Guo et al., 2020; Izzo et al., 2021) and Fisher forgetting (Golatkar et al., 2020) bypass the cost of optimizing over the full retained dataset by using local second-order approximations at the original model, which require only the local Hessian rather than the data itself. (2) **Incomplete data access.** In large language model unlearning (Liu et al., 2025) and federated learning (Liu et al., 2024), unlearning algorithms often lack the full retained dataset due to privacy or storage constraints. More extreme settings include zero-shot unlearning, where only the original weights and the target class are available (Kravets & Namboodiri, 2025), and in-context unlearning, which has only black-box access to the model (Pawelczyk et al., 2024). We defer a detailed discussion of related works to Appendix A.

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.

In this work, we rigorously investigate the information-theoretic hardness of unlearning with partial information. For a dataset $D \in \mathcal{D}$, a collection of datasets, we define

$$\mathcal{I}_{\mathcal{F}}(D) := \{f(D) : f \in \mathcal{F}\}$$

as the information extracted from D by a class of functions $\mathcal{F} := \{f : \mathcal{D} \to \mathbb{R}\}$. Under this definition, we notice that incomplete information can create ambiguity. Specifically, different datasets $D \neq D'$ may yield identical information features, i.e., $\mathcal{I}_{\mathcal{F}}(D) = \mathcal{I}_{\mathcal{F}}(D')$, when observed through a restricted function class $\mathcal{F}$. To formalize this ambiguity, for given information $I_{\mathcal{F}}$, we define $S(I_{\mathcal{F}})$ as the set of all feasible datasets D that are consistent with $I_{\mathcal{F}}$, i.e.,

$$S(I_{\mathcal{F}}) := \{D : \mathcal{I}_{\mathcal{F}}(D) = I_{\mathcal{F}}\}.$$

Hence, an unlearning algorithm $\mathcal{U}(\cdot)$ that only takes the partial information $I_{\mathcal{F}}$ as input is unable to tell the difference between the feasible datasets $D \in S(I_{\mathcal{F}})$. On the other hand, recall that the fundamental objective of unlearning is to produce a model that is indistinguishable from the retrained model on the retained dataset $D_r \subset D$. However, different feasible datasets may correspond to different minimizers of the loss on the retained data, thus making it impossible for any unlearning algorithm to approximate all retrained models of every feasible dataset simultaneously.

We formally characterize the cost of partial information by splitting the approximation of retrained model into two complementary aspects: (1) **Retained empirical excess risk**. Since a retrained model achieves the empirical loss $\min_w L(w; D_r)$ on D_r, an unlearning algorithm should match this loss as closely as possible. For a randomized unlearning algorithm $\mathcal{U}$ that takes information $I_{\mathcal{F}}$ as input and outputs a model distribution, we formalize it through the *retain-set excess risk* (retain risk):

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \;:=\; \min_{\mathcal{U}} \max_{D \in S(I_{\mathcal{F}})} \mathbb{E}_{w_u \sim \mathcal{U}(I_{\mathcal{F}})} \big[L(w_u; D_r) - \min_{w} L(w; D_r)\big],$$

(2) **Forget-set retraining discrepancy**. Since the ideal unlearned model should behave like the retrained reference model, we measure the prediction discrepancy between them on the forget data inputs. We formalize it through the *forget-set unlearning risk* (forget risk):

$$\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \;:=\; \min_{\mathcal{U}} \max_{D \in S(I_{\mathcal{F}})} \mathbb{E}_{w_u \sim \mathcal{U}(I_{\mathcal{F}})} \sum_{(X_i, Y_i) \in D_f} \ell\big(\phi_{w_u}(X_i); \phi_{w_r(D_r)}(X_i)\big),$$

where $w_r(D_r) := \arg\min_w L(w; D_r)$ denotes the retrained model[1] and $\ell(\phi_{w_1}(X); \phi_{w_2}(X))$ measures the discrepancy between the predictions of two models on input X. We note that both risks are the worst-case losses by choosing an adversarial feasible dataset consistent with the information $I_{\mathcal{F}}$.

1.1 Our Contributions

Bias–variance decomposition of retain and forget risks (Section 3). For least-squares regression, we apply Yao's minimax principle to convert the min-max formulation into a maximin over distributions of feasible datasets, and decompose each risk into an algorithm-dependent *fitting loss* and an algorithm-independent *ambiguity floor* (Propositions 1 and 2). The fitting loss measures how well the unlearned model tracks the conditional mean over feasible datasets, while the ambiguity floor captures the irreducible label/prediction variance induced by partial information.

$\Omega(k)$ **lower bounds via a hardness construction (Section 4).** Denote $k := |D_f|$. We construct a two-layer ReLU least-squares example (Example 1) in which all feasible datasets share the same original model, the same forgetting set, and the same local curvature at the original optimum, yet correspond to genuinely different retrained models. For this construction, we prove in Theorem 1 that

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \;\geq\; \tfrac{23}{8}k, \qquad \mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \;\geq\; 25k,$$

showing that both ambiguity floors scale as $\Omega(k)$ even when the unlearning algorithm has access to the original weights, the forget set, and the full Hessian, thus exact unlearning is impossible.

[1] We slightly abuse the notation here as $\arg\min_w L(w; D_r)$ is a non-empty set of minimizers. We can focus on any of the optimal models and apply any tie-breaker to formally define the retrained model as a deterministic function of D_r. In the examples of this work, it suffices to use the minimum L_2-norm as the tie-breaker.

Pareto tradeoff between retain and forget (Section 5). We show that for nonlinear models, the retain and forget fitting losses generally do *not* share a common minimizer. We illustrate this on a constant-ReLU toy example and on the hardness construction in Example 1, where the retain-optimal and forget-optimal weights differ and the Pareto tradeoffs exist. This suggests that single-objective unlearning is insufficient under partial information and motivates multi-objective optimization as a natural algorithmic framework (Jin et al., 2025; Zhou et al., 2025; Reisizadeh et al., 2026).

Matching $O(k)$ upper bounds for canonical algorithms (Section 6). We complement the lower bounds with a generic $O(k)$ upper bound for the trivial unlearner that returns the original model (Theorem 2). Together with the $\Omega(k)$ lower bound, these results establish that the linear-in-k risk is order-tight for unlearning under partial information. For the original model, ℓ_2-regularized Negative Gradient and Newton's method, we also analyze their tradeoff between retain and forget.

2 Preliminaries

Let $D = \{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^{n(d+1)}$ be a dataset consisting of n training samples $x_i \in \mathbb{R}^d$ and associated labels $y_i \in \mathbb{R}$. Let $D_f \subset D$ denote the forgetting dataset whose influence we wish to remove from a trained model, and let $D_r := D \setminus D_f$ denote the retained dataset.

Consider a model $\phi_w(\cdot)$ parameterized by weights $w \in \mathcal{W}$ in a compact hypothesis class $\mathcal{W}$, and a loss function $\ell(\phi_w(x); y)$. We define the ReLU function $\mathrm{ReLU}(x) = \max\{0, x\}$. The cumulative empirical loss on the dataset D is defined as $L(w; D) := \sum_{(x,y)\in D} \ell(\phi_w(x); y)$. Note that the loss function is a cumulative sum and is not normalized by the dataset size. We denote the optimal weight for the full loss by $w^\star \in \arg\min L(w; D)$, and the retrained model $w_r(D_r) := \arg\min_w L(w; D_r)$ (as discussed in Footnote 1). We further assume that $D \mapsto \phi_{w_r(D_r)}(X)$ is continuous on $S(I_\mathcal{F})$ for all fixed X. For least squares regression, we use the squared loss:

$$\ell(\phi_w(x); y) = (\phi_w(x) - y)^2.$$

We assume an unlearning algorithm $\mathcal{U}$ receives only partial information from the dataset D, denoted by $I_\mathcal{F} = \mathcal{I}_\mathcal{F}(D)$. We allow $\mathcal{U}$ to be randomized, thus $\mathcal{U}(I_\mathcal{F})$ denotes a distribution over weights, and $w_u \sim \mathcal{U}(I_\mathcal{F})$ denotes the model weights sampled from this distribution. A weight $w_u \in \mathcal{W}$ is *Pareto-optimal* for the retain and forget objectives $\big(\varepsilon_\mathrm{r}, \varepsilon_\mathrm{f}\big)$ if no $w'_u \in \mathcal{W}$ satisfies $\varepsilon_\mathrm{r}(w'_u) \le \varepsilon_\mathrm{r}(w_u)$ and $\varepsilon_\mathrm{f}(w'_u) \le \varepsilon_\mathrm{f}(w_u)$ with at least one strict inequality, and the *Pareto front* is the set of all such Pareto-optimal points.

We assume the number of training samples n and the size of the forgetting dataset $k := |D_f|$ are fixed. We assume the set of feasible datasets $S(I_\mathcal{F}) \subset \mathbb{R}^{n(d+1)}$ is compact, and $L(w; D)$ is jointly continuous on $\mathcal{W} \times S(I_\mathcal{F})$. We denote $\Delta(S(I_\mathcal{F}))$ as the set of probability distributions over this set, and $P \in \Delta(S(I_\mathcal{F}))$ as a probability distribution of feasible datasets. For all $1 \le i \le n$, let P_{X_i} and P_{Y_i} denote the marginal distributions of the i-th sample x_i and label y_i, respectively. We denote the Dirac point mass by $\delta(\cdot)$. We denote the ℓ_2 norm by $\|\cdot\|$ and employ standard asymptotic notation.

3 Decomposition of Retain and Forget Risks

In this section, we decompose the retain and forget risks respectively into (i) a non-negative fitting loss on the unlearned model, and (ii) an unlearning-independent ambiguity floor caused by partial information. First, we invoke Yao's Minimax Principle (Yao, 1977) to convert the min-max problems over randomized unlearning algorithms $\mathcal{U}$ in retain and forget risks into tractable max-min problems over distributions of feasible datasets and deterministic models. Lemma 1 is a direct application of Yao's Minimax Principle and the proof is deferred to Appendix B.1.

Lemma 1 (Min-max to max-min reformulation)**.** *Let $\Delta(S(I_\mathcal{F}))$ denote the set of probability distributions over the feasible set $S(I_\mathcal{F})$. Then, the retain risk and the forget risk satisfy*

$$\mathsf{R}_\mathrm{r}(I_\mathcal{F}) = \max_{P\in\Delta(S(I_\mathcal{F}))} \min_{w_u} \mathbb{E}_{D\sim P}\big[L(w_u; D_r) - \min_w L(w; D_r)\big],$$

$$\mathsf{R}_\mathrm{f}(I_\mathcal{F}) = \max_{P\in\Delta(S(I_\mathcal{F}))} \min_{w_u} \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \ell\big(\phi_{w_u}(X_i); \phi_{w_r(D_r)}(X_i)\big).$$

We begin with decomposing the retain risk. Since partial information creates ambiguity over the datasets, the distribution P can be supported on multiple feasible datasets, all of which are consistent with the same partial information. To build intuition, consider two feasible datasets $D^{(1)}, D^{(2)} \in S(I_{\mathcal{F}})$ with the same retained data x_i but distinct labels $y_i^{(1)}$ and $y_i^{(2)}$. Any unlearned model w_u must produce the same prediction on x_i, therefore cannot fit both labels $y_i^{(1)}$ and $y_i^{(2)}$. In such a scenario, the optimal unlearned model should approximate the mean value of the labels, while the variance is irreducible. We formalize this intuition in Proposition 1, and the proof is deferred to Appendix B.2.

Proposition 1 (Retain risk decomposition)**.** *In least squares regression, for every $P \in \Delta(S(I_{\mathcal{F}}))$ and each retained index i, define the conditional mean of the label as $\mu_i(x|P) := \mathbb{E}_{D\sim P}[Y_i \mid X_i = x]$. Then, the retain risk can be decomposed as*

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \;=\; \max_{P\in\Delta(S(I_{\mathcal{F}}))} \left(\min_{w_u} \varepsilon_{\mathrm{r}}(w_u; P) + \mathsf{A}_{\mathrm{r}}(P) - \mathsf{B}_{\mathrm{r}}(P) \right),$$

where the three terms are defined respectively as

$$\varepsilon_{\mathrm{r}}(w_u; P) \;:=\; \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_r} \left[\big(\phi_{w_u}(X_i) - \mu_i(X_i|P)\big)^2 \right],$$

$$\mathsf{A}_{\mathrm{r}}(P) \;:=\; \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_r} \left[\big(Y_i - \mu_i(X_i|P)\big)^2 \right], \quad \mathsf{B}_{\mathrm{r}}(P) \;:=\; \mathbb{E}_{D\sim P}\Big[\min_{w} L(w; D_r) \Big].$$

Here, $\varepsilon_{\mathrm{r}}(w_u; P)$ is an unlearning-dependent fitting loss of the unlearned model $\phi_{w_u}(\cdot)$ to the mean label $\mu_i(\cdot|P)$; while $\mathsf{A}_{\mathrm{r}}(P)$ and $\mathsf{B}_{\mathrm{r}}(P)$ are unlearning-independent: the former is the label variance under dataset distribution P and the latter is the expected retrained approximation error. We show that forget risk accepts a similar decomposition, with the proof deferred to Appendix B.3.

Proposition 2 (Forget risk decomposition)**.** *In least squares regression, for every $P \in \Delta(S(I_{\mathcal{F}}))$ and each forgetting index i, denote the retrained model $w_r(D_r) = \arg\min_w L(w; D_r)$ and define the conditional mean prediction of the retrained model as $\nu_i(x|P) := \mathbb{E}_{D\sim P}[\phi_{w_r(D_r)}(x) \mid X_i = x]$. Then, the forget risk can be decomposed as*

$$\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \;=\; \max_{P\in\Delta(S(I_{\mathcal{F}}))} \left(\min_{w_u} \varepsilon_{\mathrm{f}}(w_u; P) + \mathsf{A}_{\mathrm{f}}(P) \right),$$

where the two terms are defined respectively as

$$\varepsilon_{\mathrm{f}}(w_u; P) \;:=\; \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \left[\big(\phi_{w_u}(X_i) - \nu_i(X_i|P)\big)^2 \right],$$

$$\mathsf{A}_{\mathrm{f}}(P) \;:=\; \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \left[\big(\phi_{w_r(D_r)}(X_i) - \nu_i(X_i|P)\big)^2 \right].$$

Similar to Proposition 1, $\varepsilon_{\mathrm{f}}(w_u; P)$ is an unlearning-dependent fitting loss of the unlearned model $\phi_{w_u}(\cdot)$ to the mean prediction of the retrained model $\nu(\cdot|P)$; while $\mathsf{A}_{\mathrm{f}}(P)$ is the prediction variance of the retrained model $w_r(D_r)$ on the forgetting data, under dataset distribution P.

Combining Propositions 1 and 2, we note that for any fixed dataset distribution $P \in \Delta(S(I_{\mathcal{F}}))$, the retain and forget risks can be decomposed into two parts:

- Fitting losses ε_{r}/ε_{f}, which formulate a multi-objective optimization of unlearning;
- Irreducible unlearning-independent floors, which consist of the label/prediction variance.

Lower bounds on the retain and forget risks can be immediately derived from the non-negativity of the fitting losses, as shown in the following Corollary 1:

Corollary 1. *In least squares regression, the retain risk and forget risk can be lower bounded by*

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \geq \max_{P\in\Delta(S(I_{\mathcal{F}}))} \left(\mathsf{A}_{\mathrm{r}}(P) - \mathsf{B}_{\mathrm{r}}(P)\right), \qquad \mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \geq \max_{P\in\Delta(S(I_{\mathcal{F}}))} \mathsf{A}_{\mathrm{f}}(P).$$

We defer additional results to Appendix B.4, which establishes basic properties of R_{r} and R_{f} as functionals of $I_{\mathcal{F}}$ (monotonicity, non-negativity, and optimality under full retain access), and Appendix F, which discusses the connections between these forget-risk lower bounds and privacy definitions for unlearning, e.g., certified removal (Guo et al., 2020).

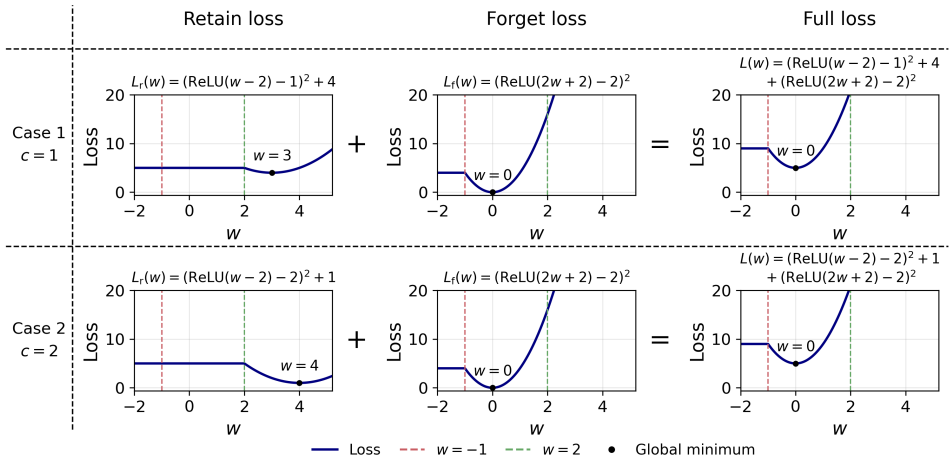


Figure 1: Illustration of the loss functions defined in equation 1. Despite having identical local geometry at the full-loss optimum $w^\star = 0$, the two cases imply distinct retrained optima.

4 Lower Bounds on the Ambiguity Floors

To concretize the risk decompositions, in this section we construct a hardness example to show that the ambiguity floors of both risks can scale as $\Omega(k)$ even when $I_{\mathcal{F}}$ includes local statistics widely used in practical unlearning algorithms. This gives strictly positive lower bounds on both retain and forget risks, which implies the impossibility of exactly recovering the retrained model given insufficient information.

We begin by elucidating the intuition behind the hardness example. Many approximate unlearning algorithms, such as Newton's method (Guo et al., 2020) and Fisher forgetting (Golatkar et al., 2020), rely on a local second-order approximation of the retained loss around the original optimal weight $w^\star$. However, such approximations often fail to capture critical information outside the immediate neighborhood of $w^\star$. To demonstrate this, we consider an illustrative example with a model parameter $w \in \mathbb{R}$ and a constant $c \in \{1, 2\}$:

$$L(w) = \underbrace{(\mathrm{ReLU}(w-2) - c)^2 + (c-3)^2}_{\text{Retain loss}} + \underbrace{(\mathrm{ReLU}(2w+2) - 2)^2}_{\text{Forget loss}}. \tag{1}$$

As illustrated in Figure 1, the full loss is minimized at $w^\star = 0$ for both $c \in \{1, 2\}$, and the forget loss is identical in both cases. We notice that the gradient of the retain loss vanishes at $w^\star = 0$ due to the inactive $\mathrm{ReLU}(\cdot)$ function. Consequently, the local curvature (Hessian) at $w^\star$ is identical for both values of c. This implies that any algorithm relying solely on local gradient information at $w^\star$ cannot distinguish between the two retained loss scenarios. As a result, such algorithms are unable to recover the distinct retrained optima, given by $w_r = c + 2$, for the respective cases. We formalize this intuition in Example 1.

Example 1 (Hardness example for least squares regression)**.** *Consider the feasible set* $S(I_{\mathcal{F}}) = \{D^{(c)} : c \in [-1, 0] \cup [1, 3]\}$*, where each dataset* $D^{(c)} = \{(x_i, y_i)\}_{i=1}^n$ *is defined as*

$$(x_i, y_i) = \begin{cases} ((0, 1 + \frac{i}{n-2k}),\ 1 + \frac{i}{n-2k}) & \text{for } 1 \le i \le n-2k \\ ((1, -2 - \frac{i-(n-2k)}{2k}),\ c - \frac{i-(n-2k)}{2k}) & \text{for } n-2k+1 \le i \le n-k \\ ((2, 2 + \frac{i-(n-k)}{k}),\ 2 + \frac{i-(n-k)}{k}) & \text{for } n-k+1 \le i \le n \end{cases}$$

We assume the forgetting dataset D_f *consists of the last* k *samples, where* $n \ge 162k$*.*

We consider a two-layer MLP with a single hidden neuron and ReLU activation, defined as $\phi_w(x) = \theta_1 \mathrm{ReLU}(\theta_2^\top x)$, where $\theta_1 \in \mathbb{R}$ and $\theta_2 \in \mathbb{R}^2$. Obviously any minimizer must satisfy $\theta_1 \ge 0$, otherwise the first $n - 2k$ samples have loss $> n - 2k$ while $\theta_1 = 1$ and $\theta_2 = (0, 1)$ result in a smaller loss $\le 9k$. Thus, without loss of generality, we reparameterize the model as $\phi_w(x) = \mathrm{ReLU}(w^\top x) = \mathrm{ReLU}(w_1 x^{(1)} + w_2 x^{(2)})$, where $w := \theta_1 \theta_2 \in \mathbb{R}^2$.

To provide intuition for Example 1, we partition the samples into three blocks: (1) The first $n-2k$ samples where $x^{(1)}=0$ and $x^{(2)}=y$ constrain the second weight component to $w_2^\star=1$. (2) The samples indexed $n-2k+1\le i\le n-k$ correspond to the retain loss in Equation (1). While for $c\in[1,3]$ these samples are perfectly fitted by $w_1=c+2$, they are positioned such that they are ignored by local gradient information at the global minimizer. (3) The final k samples constitute the forget dataset, corresponding to the forget loss in Equation (1). This block forces the full loss to be optimized at $w_1^\star=0$, thereby putting the second block into the ReLU's inactive region and masking its gradient information. Next, we prove some properties of Example 1 and establish a lower bound on the ambiguity floors of retain and forget risks. The proof of Theorem 1 is deferred to Appendix C.

Theorem 1 (Lower bounds for the hardness example)**.** *For any dataset $D^{(c)}\in S(I_{\mathcal{F}})$ in Example 1 with $c\in[-1,0]\cup[1,3]$, the least squares regression with $\phi_w(x)=$ ReLU$(w^\top x)$ satisfies:*

1. ***Identical Original Model**: They share the same full-loss minimizer $w^\star=(0,1)$;*
2. ***Identical Forgetting Set**: They share the same forgetting dataset D_f;*
3. ***Identical Curvature**: They exhibit the same local curvature information at $w^\star$;*
4. ***Label Ambiguity**: The labels y_i for $n-2k+1\le i\le n-k$ differ across datasets.*
5. ***Retrained Model Ambiguity:** For $c\in[1,3]$, the retained dataset is perfectly realizable by $w_r(D_r^{(c)})=(c+2,1)$; while for $c\in[-1,0]$, the retrained model is $w_r(D_r^{(c)})=(0,1)$.*[2]

Then, upon considering the dataset distribution $P=\frac{1}{2}\delta(D^{(-1)})+\frac{1}{2}\delta(D^{(3)})$, from Corollary 1 the retain risk and forget risk can be lower bounded by

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}})\ge\frac{23}{8}k,\qquad \mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}})\ge 25k.$$

Properties 1–3 ensure that all datasets $D^{(c)}$ are indistinguishable to any unlearning algorithm that only observes these statistics. Then, across different values of c, Property 4 shows the label variance on the retained data, and Property 5 shows the retrained model ambiguity, which further propagates to prediction variance on the forgetting data. Finally, combining Properties 4-5 with Corollary 1 yields $\Omega(k)$ lower bounds on both retain and forget risks for this hardness example.

5 Pareto Tradeoff in Retain and Forget Losses Minimization

In this section, we investigate the Pareto tradeoff for the unlearned model in retain and forget fitting losses minimization. Intuitively, the overall target of unlearning, *approximating the retrained model*, can be interpreted in two ways: (i) matching the empirical loss of the retrained model on the retained dataset; (ii) matching the prediction of the retrained model on the forgetting data to prevent privacy leakage. However, due to the ambiguity of the retained dataset and retrained model induced by incomplete information, the two targets may not align. Recall the unlearning-dependent terms in the retain and forget risks decomposition:

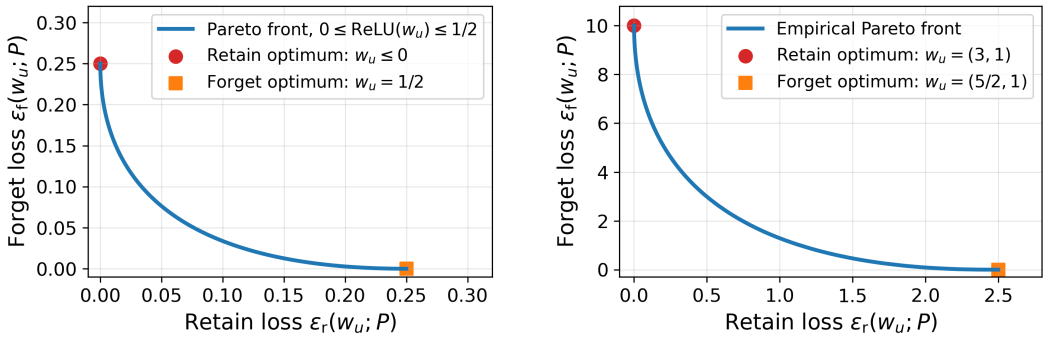
$$\min_{w_u}\begin{cases}\varepsilon_{\mathrm{r}}(w_u;P)=\mathbb{E}_{D\sim P}\sum_{(X_i,Y_i)\in D_r}\left[\left(\phi_{w_u}(X_i)-\mu_i(X_i|P)\right)^2\right]\\ \varepsilon_{\mathrm{f}}(w_u;P)=\mathbb{E}_{D\sim P}\sum_{(X_i,Y_i)\in D_f}\left[\left(\phi_{w_u}(X_i)-\nu_i(X_i|P)\right)^2\right]\end{cases}$$

where $\mu_i(x|P)=\mathbb{E}_{D\sim P}[Y_i\mid X_i=x]$ and $\nu_i(x|P)=\mathbb{E}_{D\sim P}[\phi_{w_r(D_r)}(x)\mid X_i=x]$. Assume the retained and forgetting inputs are shared across all $D\in S(I_{\mathcal{F}})$ and only the retained labels Y_r vary. For linear models, $w_u=\mathbb{E}_{D\sim P}[w_r(D_r)]$ achieves zero forget fitting loss due to the linearity, and in Appendix D.1 we show that $w_u=\mathbb{E}_{D\sim P}[w_r(D_r)]$ also minimizes the retain fitting loss, i.e., the two losses share the same minimizer:

$$\mathbb{E}_{D\sim P}[w_r(D_r)]\in\arg\min_{w_u}\varepsilon_{\mathrm{r}}(w_u;P)\cap\arg\min_{w_u}\varepsilon_{\mathrm{f}}(w_u;P).$$

On the other hand, for non-linear models, we demonstrate that they may not be minimized by the same weight, thus arguing the necessity of applying multi-objective optimization to fully explore the Pareto tradeoff between the two fitting losses.

[2]For $c\in[-1,0]$, any model $w=(\alpha,1)$ with $\alpha\le 2+\frac{1}{2k}$ is a minimizer to the retrained loss. According to Footnote 1, we use minimum L_2-norm as a tie-breaker and define $w_r(D_r^{(c)})=(0,1)$ as the retrained model.



(a) Toy nonlinear example in Eq. 2.

(b) Example 1 with $n = 2000$ and $k = 10$ in Eq. 3.

Figure 2: Pareto frontiers of the retain–forget fitting losses for nonlinear examples.

Toy Non-Linear Example. We begin with an illustrative example. Consider the hypothesis class of constant-ReLU functions $\{\phi_w(x) = \mathrm{ReLU}(w) : w \in \mathbb{R}\}$, and two feasible datasets that share the same forget dataset but differ only in the retained label:

$$D^{(+)} = \{(X_r, +1), (X_f, -1)\}, \qquad D^{(-)} = \{(X_r, -1), (X_f, -1)\}.$$

Then, the full losses are:

$$L(w; D^{(+)}) = 2\mathrm{ReLU}(w)^2 + 2, \qquad L(w; D^{(-)}) = 2(\mathrm{ReLU}(w) + 1)^2.$$

Both attain their minimum value 2 at $\mathrm{ReLU}(w) = 0$, so as discussed in Footnote 1, both datasets share the same original model $w^\star = 0$. Consequently, the two datasets are indistinguishable by any unlearning algorithm with access to only the original model and the forgetting dataset, thus the adversary's mixture distribution over datasets can be defined as

$$P = \frac{1}{2}\delta(D^{(+)}) + \frac{1}{2}\delta(D^{(-)}).$$

The retrained model, on the other hand, is distinct for the two retained datasets:

$$w_r(D_r) = \arg\min_w (\mathrm{ReLU}(w) - Y_r)^2 = \begin{cases} 1 & \text{for } D_r = D_r^{(+)}, \\ 0 & \text{for } D_r = D_r^{(-)}. \end{cases}$$

Since the conditional mean label $\mu_r(X_r|P) = 0$ and the conditional mean prediction of the retrained model $\nu_f(X_f|P) = \frac{1}{2}$, the retain and forget losses are given by

$$\min_{w_u} \begin{cases} \varepsilon_{\mathrm{r}}(w_u; P) = \mathrm{ReLU}(w_u)^2 \\ \varepsilon_{\mathrm{f}}(w_u; P) = \left(\mathrm{ReLU}(w_u) - \frac{1}{2}\right)^2 \end{cases} \tag{2}$$

Clearly the retain loss is minimized at $w_u \le 0$, while the forget loss is minimized at $w_u = \frac{1}{2}$, which cannot be satisfied simultaneously. The Pareto front is achieved by $\mathrm{ReLU}(w_u) \in [0, \frac{1}{2}]$, as shown in Fig. 2a, indicating that the two objectives need not share the same minimizer.

Pareto Tradeoff in Example 1. Following from the same setup of Theorem 1, we consider the dataset distribution $P = \frac{1}{2}\delta(D^{(-1)}) + \frac{1}{2}\delta(D^{(3)})$. On the retained dataset, the sample x_i and conditional mean label $\mu_i(x_i|P)$ for $1 \le i \le n - k$ are

$$(x_i,\, \mu_i(x_i|P)) = \begin{cases} ((0, 1 + \frac{i}{n-2k}),\, 1 + \frac{i}{n-2k}) & \text{for } 1 \le i \le n - 2k, \\ ((1, -2 - \frac{i-(n-2k)}{2k}),\, 1 - \frac{i-(n-2k)}{2k}) & \text{for } n - 2k + 1 \le i \le n - k. \end{cases}$$

Due to Property 5 in Theorem 1, for $n - k + 1 \le i \le n$, the conditional mean prediction of the retrained model on the forgetting sample $x_i = (2, 2 + \frac{i-(n-k)}{k})$ is

$$\nu_i(x_i|P) = \frac{1}{2}\mathrm{ReLU}(w_r^\top(D_r^{(-1)})x_i) + \frac{1}{2}\mathrm{ReLU}(w_r^\top(D_r^{(3)})x_i) = 7 + \frac{i - (n-k)}{k}.$$

Therefore, the retain and forget losses are given by

$$\min_{w_u} \begin{cases} \varepsilon_{\mathrm{r}}(w_u; P) = \sum_{i=1}^{n-k} (\mathrm{ReLU}(w_u^\top x_i) - \mu_i(x_i|P))^2 \\ \varepsilon_{\mathrm{f}}(w_u; P) = \sum_{i=n-k+1}^{n} (\mathrm{ReLU}(w_u^\top x_i) - \nu_i(x_i|P))^2 \end{cases} \tag{3}$$

One can easily verify that the retain loss $\varepsilon_{\mathrm{r}}(w_u; P)$ admits a unique minimizer $w_u^{\mathrm{r}} = (3, 1)$; while for $k \geq 2$, the forget loss $\varepsilon_{\mathrm{f}}(w_u; P)$ admits a different unique minimizer $w_u^{\mathrm{f}} = (\frac{5}{2}, 1)$. More precisely, the retain and forget losses at the two optima are

1. At $w_u^{\mathrm{r}} = (3, 1)$, $\varepsilon_{\mathrm{r}}(w_u^{\mathrm{r}}; P) = 0$ and $\varepsilon_{\mathrm{f}}(w_u^{\mathrm{r}}; P) = k$;
2. At $w_u^{\mathrm{f}} = (\frac{5}{2}, 1)$, $\varepsilon_{\mathrm{r}}(w_u^{\mathrm{f}}; P) = \frac{1}{4}k$ and $\varepsilon_{\mathrm{f}}(w_u^{\mathrm{f}}; P) = 0$.

A natural guess is whether $w = (\lambda, 1)$ with $\lambda \in [\frac{5}{2}, 3]$ can explore the Pareto front, but it is rejected by our analysis in Appendix D.2 which shows that the partial derivatives of both losses with respect to w_2 are strictly positive at $w = (\lambda, 1)$ for $\lambda \in (\frac{5}{2}, 3)$, finding closed-form solutions to the Pareto front is therefore non-trivial. Instead, we numerically solve the Pareto front for $n = 2000$ and $k = 10$, and the result is plotted in Fig. 2b.

6 Upper Bounds: Analysis of Existing Algorithms

We complement the lower bounds established in Section 4 by analyzing the retain and forget risks of representative practical unlearning algorithms within the limited-information regime. Recall that Theorem 1 establishes lower bounds for the hardness Example 1:

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \geq \frac{23}{8}k, \qquad \mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \geq 25k.$$

Notice that the forget risk is summed over the forgetting dataset of size k, which means that the lower bound $\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) = \Omega(k)$ is already order-wise optimal for bounded label predictions. Therefore, in this section our analysis primarily focuses on the retain risk $\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}})$, which is summed over the retained dataset of size $n - k$.

A generic $O(k)$ upper bound for the original model. We begin by establishing an algorithm-independent upper bound on the retain risk, assuming access to the full-loss minimizer $w^\star$. This bound is derived from a trivial unlearning strategy: simply returning $w^\star$ without modification. While this estimate is conservative, it demonstrates that merely retaining $w^\star$ is sufficient to bound the worst-case retain risk. The proof of Theorem 2 is deferred to Appendix E.1.

Theorem 2 (Upper bound)**.** *Let the information set $I_{\mathcal{F}}$ contains the full-dataset minimizer $w^\star \in \arg\min_w L(w; D)$ (i.e., $\mathcal{F}$ contains a set of functions that map to each entry of $w^\star$). Assuming the loss function ℓ is non-negative, the retain risk satisfies*

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \leq Ck,$$

where C denotes the maximum loss of an optimal retrained model on any forgetting sample, i.e.,

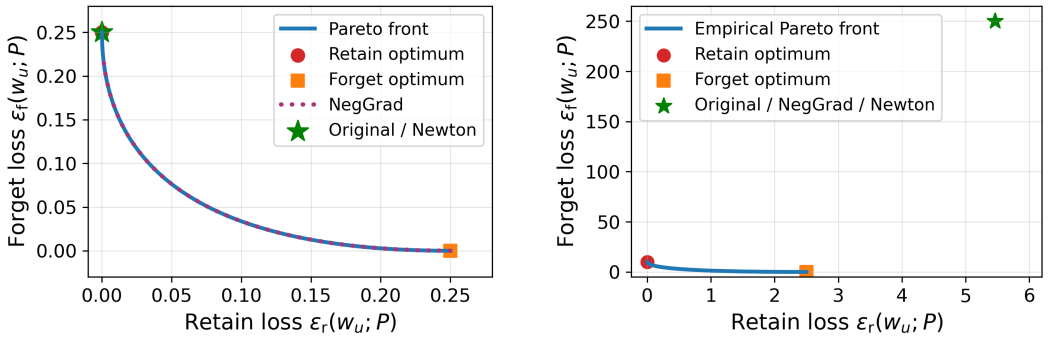
$$C := \max_{D \in S(I_{\mathcal{F}})} \max_{\substack{(x,y) \in D_f, \\ w_r \in \arg\min_w L(w; D_r)}} \ell(\phi_{w_r}(x); y). \tag{4}$$

Theorem 2 establishes that, provided $w^\star$ is available and the loss is bounded, the retain risk $\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}})$ is upper-bounded by $O(k)$. This scaling matches the $\Omega(k)$ lower bound derived in Example 1 and Theorem 1, confirming that the linear dependence on k is order-wise tight. This result demonstrates that a linear retain risk is both theoretically unavoidable and practically order-wise achievable.

Negative Gradient with ℓ_2 regularization. We next examine a standard first-order unlearning baseline: the Negative Gradient method with ℓ_2 regularization (Golatkar et al., 2020; Yao et al., 2024). This approach seeks to maximize the loss on the forgetting dataset while maintaining proximity to the original solution $w^\star$ via regularization:

$$\mathcal{U}_{\text{NegGrad}}: \quad w_u = \arg\min_w -L(w; D_f) + \tfrac{\lambda n}{2} \left\| w - w^\star \right\|^2,$$

where the regularization parameter $\lambda > 0$ constrains the deviation from $w^\star$, thereby preventing the model from catastrophically forgetting retained knowledge. When $\lambda \geq \beta$, Proposition 3 in Appendix E.2 establishes the same upper bound Ck on the retain risk of $\mathcal{U}_{\text{NegGrad}}$ under smoothness and full-loss stationarity assumptions.



(a) Toy nonlinear example in Eq. 2.

(b) Example 1 with $n = 2000$ and $k = 10$ in Eq. 3.

Figure 3: Pareto frontiers and representative unlearning methods for nonlinear examples. In the toy example, $\mathcal{U}_{\text{NegGrad}}$ with suitable regularization traces the Pareto frontier, while the original model $w^\star$ and $\mathcal{U}_{\text{Newton}}$ stay at the retain-optimal endpoint. In Example 1, the original model $w^\star$ and $\mathcal{U}_{\text{Newton}}$ collapse to an off-front point, and $\mathcal{U}_{\text{NegGrad}}$ either collapses to the same point or diverges.

Newton's Method. Finally, we consider Newton's Method $\mathcal{U}_{\text{Newton}}$ (Guo et al., 2020), which utilizes Hessian to approximate the retrained model. This method performs a single Newton step starting from $w^\star$ to estimate the minimizer of the retained loss. The update rule for $\mathcal{U}_{\text{Newton}}$ is given by:

$$\mathcal{U}_{\text{Newton}}: \quad w_u = w^\star - \left[H(w^\star; D_r) + \lambda(n-k)I\right]^{-1} \nabla L(w^\star; D_r).$$

Proposition 4 in Appendix E.2 also establishes the same upper bound Ck on the retain risk of $\mathcal{U}_{\text{Newton}}$ under smoothness assumptions.

Pareto Tradeoff of Existing Algorithms. In Fig. 3, we demonstrate the retain/forget fitting losses tradeoff of the original model, $\mathcal{U}_{\text{NegGrad}}$, and $\mathcal{U}_{\text{Newton}}$. In the toy example from equation 2, the original model and $\mathcal{U}_{\text{Newton}}$ converge to the retain optimum, while $\mathcal{U}_{\text{NegGrad}}$ with different regularization coefficient λ can explore the Pareto front. However, in Example 1, the original model and $\mathcal{U}_{\text{Newton}}$ converge to an off-front point, while $\mathcal{U}_{\text{NegGrad}}$ either collapses to the same point when $\lambda > \frac{1}{n}\lambda_{\max}(2X^\top X)$ or becomes unstable and diverges when $\lambda \leq \frac{1}{n}\lambda_{\max}(2X^\top X)$.

7 Discussion and Conclusion

This work establishes an information-theoretic foundation for machine unlearning with incomplete access to the retained dataset. We treat unlearning as an optimization problem under partial knowledge, in which the algorithm operates with an information budget $I_{\mathcal{F}}$ that may not uniquely identify the retained dataset, and introduce the *retain-set excess risk* and the *forget-set unlearning risk*.

For least-squares regression, each risk decomposes into an algorithm-dependent fitting loss and an algorithm-independent ambiguity floor. A two-layer ReLU hardness construction shows that both floors scale as $\Omega(k)$ even when the algorithm has access to the original weights, forgetting set, and local Hessian. The retain and forget fitting losses further fail to share a minimizer for nonlinear models, exposing an inherent Pareto tradeoff that no single-objective formulation can resolve. The original weights, ℓ_2-regularized Negative Gradient and Newton's method all attain a matching $O(k)$ retain risk under suitable regularity conditions, giving an order-wise tight dependence on k.

Future directions. Our framework opens several questions. (i) Can the bias–variance decomposition and the $\Omega(k)$ ambiguity floors be extended beyond least-squares regression to classification losses such as binary or multi-class cross-entropy? (ii) Which side information, e.g., D_r subsamples or training-trajectory checkpoints, can effectively shrink the ambiguity floors, and how does the bound depend on the shared information between D_r and $I_{\mathcal{F}}$? (iii) How do the retain/forget bounds behave on realistic tasks, and how does the prediction-space forget risk relate to empirical attacks such as membership inference? (iv) Can multi-objective unlearning algorithms attain the Pareto front with provable guarantees? (v) Can the framework be extended to population risk?

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the writing and finding related work. We did not use AI for research ideation or execution, drafting sections of the paper, or proving mathematical claims, and using AI for generating synthetic datasets is not applicable to this work. We have reviewed all AI-assisted work. For example, we checked LLM-generated references through a manual literature review. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies theoretical limits of machine unlearning through mathematical analysis and synthetic examples. Our results aim to support accurate assessments of unlearning capabilities and limitations.

REPRODUCIBILITY STATEMENT

Our assumptions and problem formulation are specified in Section 2, with detailed proofs provided in Appendices B–F. The numerical settings used in the illustrations are described in Sections 4 and 5.

REFERENCES

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

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023.

Chongyu Fan, Jiancheng Liu, Alfred Hero, and Sijia Liu. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In *European Conference on Computer Vision*, pp. 278–297. Springer, 2024.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition*, pp. 9304–9312, 2020.

Eric Goldman. An introduction to the california consumer privacy act (CCPA). *Santa Clara Univ. Legal Studies Research Paper*, 2020.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning*, pp. 3832–3842. PMLR, 2020.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 497–519. IEEE, 2025.

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

Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11278–11294, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN

979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.563. URL https://aclanthology.org/2025.naacl-long.563/.

Alexey Kravets and Vinay P Namboodiri. Zero-shot class unlearning in clip with synthetic samples. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6456–6464. IEEE, 2025.

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

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194, 2025.

Ziyao Liu, Yu Jiang, Jiyuan Shen, Minyi Peng, Kwok-Yan Lam, Xingliang Yuan, and Xiaoning Liu. A survey on federated unlearning: Challenges, methods, and future directions. *ACM Computing Surveys*, 57(1):1–38, 2024.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *International Conference on Machine Learning*, pp. 40034–40050. PMLR, 2024.

Hadi Reisizadeh, Jinghan Jia, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, Sijia Liu, and Mingyi Hong. BLUR: A bi-level optimization approach for LLM unlearning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7043–7058, 2026.

Hammad Rizwan, Mahtab Sarvmaili, Hassan Sajjad, and Ga Wu. Instance-level difficulty: A missing perspective in machine unlearning. *arXiv preprint arXiv:2410.03043*, 2024.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2023.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Paul Voigt and Axel Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Cham, 01 2017. ISBN 978-3-319-57958-0. doi: 10.1007/978-3-319-57959-7.

Martin Van Waerebeke, Marco Lorenzi, Giovanni Neglia, and Kevin Scaman. When to forget? complexity trade-offs in machine unlearning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=uEUIeIrRPv.

Yinjun Wu, Edgar Dobriban, and Susan Davidson. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, pp. 10355–10366. PMLR, 2020.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), August 2023. ISSN 0360-0300. doi: 10.1145/3603620. URL https://doi.org/10.1145/3603620.

Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science*, pp. 222–227. IEEE Computer Society, 1977.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37:12293–12333, 2024.

Shiji Zhou, Tianbai Yu, Zhi Zhang, Heng Chang, Xiao Zhou, Dong Wu, and Han Zhao. Efficient utility-preserving machine unlearning with implicit gradient surgery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=vzcVDwLtwA.

A Related Works

Machine unlearning algorithms with different information access. Approximate unlearners can be categorized by their information budgets. The first category requires full access to the retained samples, e.g., SCRUB (Kurmanji et al., 2023), which alternates between maximizing the forgetting loss and minimizing the retained loss, and the zero-glance approach (Tarun et al., 2023), which uses retained but not forgetting data. The second category bypasses retained data and instead relies on local curvature or gradient information: Newton's method (Guo et al., 2020) and Fisher forgetting (Golatkar et al., 2020) use the original weights, the forgetting set, and the full Hessian, while Negative Gradient methods (Golatkar et al., 2020; Kurmanji et al., 2023; Yao et al., 2024) simply ascend the forgetting loss. The most restrictive category is zero-shot unlearning (Chundawat et al., 2023), which uses only the original model and target labels. Methods leveraging training metadata such as trajectories (Wu et al., 2020; Graves et al., 2021; Thudi et al., 2022) are out of scope for this work.

Difficulty of machine unlearning. Recent work has examined the intrinsic difficulty of unlearning from several angles: embedding entanglement and memorization (Zhao et al., 2024), adversarially selected forgetting sets (Fan et al., 2024), instance-level factors such as decision-boundary distance (Rizwan et al., 2024), and computational complexity bounds on feasible regimes (Waerebeke et al., 2025). These works characterize *computational* obstacles. In contrast, our analysis is information-theoretic: we show that without sufficient information about the retained data, exactly matching the retrained model is impossible regardless of computational power.

Multi-objective optimization in machine unlearning. A growing line of work casts unlearning as a multi-objective problem balancing retained-data utility against forgetting, resolving the conflict via normalized gradient-difference updates with adaptive learning rates (Jin et al., 2025), implicit gradient surgery (Zhou et al., 2025), or bi-level optimization (Reisizadeh et al., 2026). These works focus on algorithmic design, while we instead ask what is fundamentally achievable under partial information, and justify the multi-objective view by proving that no single unlearned model can simultaneously minimize retain and forget losses in our hardness examples.

B Decomposition of Retain and Forget Risks

B.1 Proof of Lemma 1

Lemma 1 (Min-max to max-min reformulation)**.** *Let $\Delta(S(I_{\mathcal{F}}))$ denote the set of probability distributions over the feasible set $S(I_{\mathcal{F}})$. Then, the retain risk and the forget risk satisfy*

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) = \max_{P\in\Delta(S(I_{\mathcal{F}}))} \min_{w_u} \mathbb{E}_{D\sim P}\big[L(w_u; D_r) - \min_{w} L(w; D_r)\big],$$

$$\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) = \max_{P\in\Delta(S(I_{\mathcal{F}}))} \min_{w_u} \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \ell\big(\phi_{w_u}(X_i); \phi_{w_r(D_r)}(X_i)\big).$$

Proof. We first recall Yao's minimax principle (Yao, 1977), which provides a duality between randomized algorithms against worst-case inputs and deterministic algorithms against the worst-case input distribution.

Lemma 2 (Yao's Minimax Principle, (Yao, 1977))**.** *Let $\mathcal{A}$ be a compact set of deterministic algorithms, $\mathcal{X}$ a compact set of inputs, and $c : \mathcal{A}\times\mathcal{X}\to\mathbb{R}$ a continuous cost function. Let $\Delta(\mathcal{A})$ and $\Delta(\mathcal{X})$ denote the set of probability distributions over $\mathcal{A}$ and $\mathcal{X}$, respectively. Then*

$$\min_{\mu\in\Delta(\mathcal{A})} \max_{x\in\mathcal{X}} \mathbb{E}_{a\sim\mu}[c(a,x)] \;=\; \max_{P\in\Delta(\mathcal{X})} \min_{a\in\mathcal{A}} \mathbb{E}_{x\sim P}[c(a,x)].$$

We now apply Lemma 2 to derive Lemma 1. For the retain risk, identify $\mathcal{A}$ with the set of deterministic unlearned weights $w_u \in \mathcal{W}$, $\mathcal{X}$ with the feasible dataset set $S(I_\mathcal{F})$, and the cost function with $c(w_u, D) := L(w_u; D_r) - \min_w L(w; D_r)$. A randomized unlearning algorithm $\mathcal{U}$ taking input $I_\mathcal{F}$ corresponds to a distribution $\mu \in \Delta(\mathcal{W})$, since $I_\mathcal{F}$ is a deterministic function of D and is fixed across all $D \in S(I_\mathcal{F})$. The compactness of $\mathcal{W}$ and $S(I_\mathcal{F})$ together with the continuity of L assumed in Section 2 ensure the conditions of Lemma 2 are satisfied. Applying the principle yields the first equality in Lemma 1.

The forget risk follows analogously by setting $c(w_u, D) := \sum_{(X_i,Y_i)\in D_f} \ell(\phi_{w_u}(X_i); \phi_{w_r(D_r)}(X_i))$ and noting that D_f and $\phi_{w_r(D_r)}$ are deterministic functions of D. □

B.2 Proof of Proposition 1

Proposition 1 (Retain risk decomposition)**.** *In least squares regression, for every $P \in \Delta(S(I_\mathcal{F}))$ and each retained index i, define the conditional mean of the label as $\mu_i(x|P) := \mathbb{E}_{D\sim P}[Y_i \mid X_i = x]$. Then, the retain risk can be decomposed as*

$$\mathsf{R}_\mathrm{r}(I_\mathcal{F}) \;=\; \max_{P\in\Delta(S(I_\mathcal{F}))} \left(\min_{w_u} \varepsilon_\mathrm{r}(w_u; P) + \mathsf{A}_\mathrm{r}(P) - \mathsf{B}_\mathrm{r}(P) \right),$$

where the three terms are defined respectively as

$$\begin{aligned}
\varepsilon_\mathrm{r}(w_u; P) \;&:=\; \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_r} \Big[\big(\phi_{w_u}(X_i) - \mu_i(X_i|P)\big)^2 \Big], \\
\mathsf{A}_\mathrm{r}(P) \;&:=\; \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_r} \Big[\big(Y_i - \mu_i(X_i|P)\big)^2 \Big], \quad \mathsf{B}_\mathrm{r}(P) \;:=\; \mathbb{E}_{D\sim P}\Big[\min_w L(w; D_r) \Big].
\end{aligned}$$

Proof. In least squares regression, by fixing $P \in \Delta(S(I_\mathcal{F}))$ and w_u, the standard bias–variance decomposition for each $(X_i, Y_i) \in D_r$ gives

$$\begin{aligned}
\mathbb{E}_{D\sim P}\big[(\phi_w(X_i) - Y_i)^2\big] \;&=\; \mathbb{E}_{D\sim P}\big[(\phi_w(X_i) - \mu_i(X_i|P))^2\big] + \mathbb{E}_{D\sim P}\big[(\mu_i(X_i|P) - Y_i)^2\big] \\
&\quad + 2\,\mathbb{E}_{D\sim P}[(\phi_w(X_i) - \mu_i(X_i|P))(\mu_i(X_i|P) - Y_i)]\,.
\end{aligned}$$

The cross term vanishes, since conditioning on X_i yields

$$\mathbb{E}_{D\sim P}\big[\mu_i(X_i|P) - Y_i \,\big|\, X_i\big] \;=\; \mu_i(X_i|P) - \mathbb{E}_{D\sim P}[Y_i \mid X_i] \;=\; 0.$$

Summing the remaining terms over D_r yields

$$\begin{aligned}
\mathbb{E}_{D\sim P}[L(w; D_r)] &= \sum_{(X_i,Y_i)\in D_r} \mathbb{E}_{D\sim P}(\phi_w(X_i) - Y_i)^2 \\
&= \sum_{(X_i,Y_i)\in D_r} \mathbb{E}_{D\sim P}\left[(\phi_w(X_i) - \mu_i(X_i|P))^2 + (\mu_i(X_i|P) - Y_i)^2\right] \\
&= \varepsilon_\mathrm{r}(w_u; P) + \mathsf{A}_\mathrm{r}(P).
\end{aligned}$$

Therefore, by Lemma 1, we have

$$\begin{aligned}
\mathsf{R}_\mathrm{r}(I_\mathcal{F}) &= \min_{\mathcal{U}} \max_{D\in S(I_\mathcal{F})} \mathbb{E}_{w_u\sim\mathcal{U}(I_\mathcal{F})}\big[L(w_u; D_r) - \min_w L(w; D_r)\big] \\
&= \max_{P\in\Delta(S(I_\mathcal{F}))} \min_{w_u} \mathbb{E}_{D\sim P}\big[L(w_u; D_r) - \min_w L(w; D_r)\big] \\
&= \max_{P\in\Delta(S(I_\mathcal{F}))} \left(\min_{w_u} \varepsilon_\mathrm{r}(w_u; P) + \mathsf{A}_\mathrm{r}(P) - \mathsf{B}_\mathrm{r}(P) \right),
\end{aligned}$$

where the last equation is because $\mathsf{A}_\mathrm{r}(P)$ and $\mathsf{B}_\mathrm{r}(P)$ are independent from w_u. Hence we finish the proof of Proposition 1. □

B.3 Proof of Proposition 2

Proposition 2 (Forget risk decomposition). *In least squares regression, for every $P \in \Delta(S(I_{\mathcal{F}}))$ and each forgetting index i, denote the retrained model $w_r(D_r) = \arg\min_w L(w; D_r)$ and define the conditional mean prediction of the retrained model as $\nu_i(x|P) := \mathbb{E}_{D\sim P}[\phi_{w_r(D_r)}(x) \mid X_i = x]$. Then, the forget risk can be decomposed as*

$$\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \;=\; \max_{P\in\Delta(S(I_{\mathcal{F}}))} \left(\min_{w_u} \varepsilon_{\mathrm{f}}(w_u; P) + \mathsf{A}_{\mathrm{f}}(P) \right),$$

where the two terms are defined respectively as

$$\varepsilon_{\mathrm{f}}(w_u; P) \;:=\; \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \left[\big(\phi_{w_u}(X_i) - \nu_i(X_i|P)\big)^2 \right],$$
$$\mathsf{A}_{\mathrm{f}}(P) \;:=\; \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \left[\big(\phi_{w_r(D_r)}(X_i) - \nu_i(X_i|P)\big)^2 \right].$$

Proof. In least squares regression, by fixing $P \in \Delta(S(I_{\mathcal{F}}))$ and w_u, the standard bias–variance decomposition for each $(X_i, Y_i) \in D_f$ gives

$$\begin{aligned}\mathbb{E}_{D\sim P}\big[(\phi_w(X_i) - \phi_{w_r(D_r)}(X_i))^2\big] \;=\; & \mathbb{E}_{D\sim P}\big[(\phi_w(X_i) - \nu_i(X_i|P))^2\big] + \mathbb{E}_{D\sim P}\big[(\nu_i(X_i|P) - \phi_{w_r(D_r)}(X_i))^2\big] \\ & + 2\,\mathbb{E}_{D\sim P}\big[(\phi_w(X_i) - \nu_i(X_i|P))(\nu_i(X_i|P) - \phi_{w_r(D_r)}(X_i))\big].\end{aligned}$$

The cross term vanishes, since conditioning on X_i yields

$$\mathbb{E}_{D\sim P}\big[\nu_i(X_i|P) - \phi_{w_r(D_r)}(X_i) \,\big|\, X_i\big] \;=\; \nu_i(X_i|P) - \mathbb{E}_{D\sim P}[\phi_{w_r(D_r)}(X_i) \mid X_i] \;=\; 0.$$

Summing the remaining terms over D_f yields

$$\begin{aligned}&\mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \ell\big(\phi_{w_u}(X_i); \phi_{w_r(D_r)}(X_i)\big) \\ &= \sum_{(X_i,Y_i)\in D_f} \mathbb{E}_{D\sim P}(\phi_w(X_i) - \phi_{w_r(D_r)}(X_i))^2 \\ &= \sum_{(X_i,Y_i)\in D_f} \mathbb{E}_{D\sim P}\left[(\phi_{w_u}(X_i) - \nu_i(X_i|P))^2 + (\nu_i(X_i|P) - \phi_{w_r(D_r)}(X_i))^2\right] \\ &= \varepsilon_{\mathrm{f}}(w_u; P) + \mathsf{A}_{\mathrm{f}}(P).\end{aligned}$$

Therefore, by Lemma 1, we have

$$\begin{aligned}\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) &= \min_{\mathcal{U}} \max_{D\in S(I_{\mathcal{F}})} \mathbb{E}_{w_u\sim\mathcal{U}(I_{\mathcal{F}})} \sum_{(X_i,Y_i)\in D_f} \ell\big(\phi_{w_u}(X_i); \phi_{w_r(D_r)}(X_i)\big) \\ &= \max_{P\in\Delta(S(I_{\mathcal{F}}))} \min_{w_u} \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \ell\big(\phi_{w_u}(X_i); \phi_{w_r(D_r)}(X_i)\big) \\ &= \max_{P\in\Delta(S(I_{\mathcal{F}}))} \left(\min_{w_u} \varepsilon_{\mathrm{f}}(w_u; P) + \mathsf{A}_{\mathrm{f}}(P) \right),\end{aligned}$$

where the last equation is because $\mathsf{A}_{\mathrm{f}}(P)$ is independent from w_u. Hence we finish the proof of Proposition 2. □

B.4 Basic Properties of Retain/Forget Risk

We collect three structural properties that the retain and forget risks satisfy as functionals of the information budget $I_{\mathcal{F}}$. These properties are independent of the loss function or the model class and follow from the min-max definition alone.

Lemma 3 (Basic properties of retain/forget risk). *Assume the loss ℓ is non-negative. For $\mathsf{R} \in \{\mathsf{R}_{\mathrm{r}}, \mathsf{R}_{\mathrm{f}}\}$, the following hold:*

1. ***Monotonicity:*** *If $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then $\mathsf{R}\big(\mathcal{I}_{\mathcal{F}_1}(D)\big) \geq \mathsf{R}\big(\mathcal{I}_{\mathcal{F}_2}(D)\big)$ for every dataset D.*

2. ***Non-negativity:*** $\mathsf{R}(I_{\mathcal{F}}) \geq 0$ *for every* $I_{\mathcal{F}}$.

3. ***Optimality:*** *If the retained dataset is fully accessible by* $I_{\mathcal{F}}$ *(i.e.,* $\mathcal{F}$ *contains a set of functions that map to the* i*-th data point for all* i*), then* $\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) = 0$*. If in addition* $\ell(\phi_w(x); \phi_w(x)) = 0$ *for all* w, x*, then* $\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) = 0$.

Proof. For simplicity, we treat both risks via the unified objective

$$J(w_u; D) := \begin{cases} L(w_u; D_r) - \min_w L(w; D_r), & \text{(for retain risk)}, \\ \sum_{(X_i, Y_i) \in D_f} \ell\big(\phi_{w_u}(X_i); \phi_{w_r(D_r)}(X_i)\big), & \text{(for forget risk)}, \end{cases}$$

so that $\mathsf{R}(I_{\mathcal{F}}) = \min_{\mathcal{U}} \max_{D \in S(I_{\mathcal{F}})} \mathbb{E}_{w_u \sim \mathcal{U}(I_{\mathcal{F}})}[J(w_u; D)]$. Both versions satisfy $J(w_u; D) \geq 0$: the retain version by the definition of the minimum, and the forget version by the assumed non-negativity of ℓ.

Monotonicity. Let $I_{\mathcal{F}2} := \mathcal{I}_{\mathcal{F}_2}(D)$ and $I_{\mathcal{F}1} := \mathcal{I}_{\mathcal{F}_1}(D)$. Since $\mathcal{F}_1 \subseteq \mathcal{F}_2$, the value set $I_{\mathcal{F}1}$ is a deterministic function of $I_{\mathcal{F}2}$ — explicitly, restrict the function-value set in $I_{\mathcal{F}2}$ to the indices in $\mathcal{F}_1$. Consequently any $D' \in S(I_{\mathcal{F}2})$ also satisfies $\mathcal{I}_{\mathcal{F}_1}(D') = I_{\mathcal{F}1}$, so $S(I_{\mathcal{F}2}) \subseteq S(I_{\mathcal{F}1})$. Let $\mathcal{U}_1^\star \in \arg\min_{\mathcal{U}} \max_{D' \in S(I_{\mathcal{F}1})} \mathbb{E}_{w_u \sim \mathcal{U}(I_{\mathcal{F}1})}[J(w_u; D')]$ and define $\mathcal{U}_2(I_{\mathcal{F}2}) := \mathcal{U}_1^\star(I_{\mathcal{F}1})$. Then

$$\begin{aligned} \mathsf{R}(I_{\mathcal{F}2}) &\leq \max_{D' \in S(I_{\mathcal{F}2})} \mathbb{E}_{w_u \sim \mathcal{U}_2(I_{\mathcal{F}2})}[J(w_u; D')] && \text{(taking } \mathcal{U}_2 \text{ as a feasible algorithm)} \\ &\leq \max_{D' \in S(I_{\mathcal{F}1})} \mathbb{E}_{w_u \sim \mathcal{U}_1^\star(I_{\mathcal{F}1})}[J(w_u; D')] && \text{(since } S(I_{\mathcal{F}2}) \subseteq S(I_{\mathcal{F}1}) \text{ and } \mathcal{U}_2 = \mathcal{U}_1^\star) \\ &= \mathsf{R}(I_{\mathcal{F}1}). \end{aligned}$$

Non-negativity. Since $J(w_u; D) \geq 0$ pointwise, $\mathsf{R}(I_{\mathcal{F}}) = \min_{\mathcal{U}} \max_{D \in S(I_{\mathcal{F}})} \mathbb{E}_{w_u}[J(w_u; D)] \geq 0$.

Optimality. Suppose $\mathcal{F}$ contains the per-sample retain-set projections. Then $I_{\mathcal{F}} = \mathcal{I}_{\mathcal{F}}(D)$ uniquely determines D_r: every $D' \in S(I_{\mathcal{F}})$ must agree with D on each retained sample $(X_i, Y_i) \in D_r$, hence $D'_r = D_r$. Consider the deterministic algorithm $\mathcal{U}^\star(I_{\mathcal{F}})$ that outputs the retrained model $w_r(D_r)$. For any $D' \in S(I_{\mathcal{F}})$:

- *Retain*: $L(w_r(D_r); D'_r) - \min_w L(w; D'_r) = L(w_r(D_r); D_r) - \min_w L(w; D_r) = 0$.
- *Forget*: $\ell\big(\phi_{w_r(D'_r)}(X_i); \phi_{w_r(D_r)}(X_i)\big) = \ell\big(\phi_{w_r(D_r)}(X_i); \phi_{w_r(D_r)}(X_i)\big) = 0$ under the assumption $\ell(\phi_w(x); \phi_w(x)) = 0$ for all w, x.

In both cases $J(\mathcal{U}^\star(I_{\mathcal{F}}); D') = 0$, so $\mathsf{R}(I_{\mathcal{F}}) \leq 0$. Combined with the non-negativity property, $\mathsf{R}(I_{\mathcal{F}}) = 0$. □

C Lower Bounds: Proof of Theorem 1

Theorem 1 (Lower bounds for the hardness example)**.** *For any dataset* $D^{(c)} \in S(I_{\mathcal{F}})$ *in Example 1 with* $c \in [-1, 0] \cup [1, 3]$*, the least squares regression with* $\phi_w(x) = \mathit{ReLU}(w^\top x)$ *satisfies:*

1. ***Identical Original Model****: They share the same full-loss minimizer* $w^\star = (0, 1)$*;*
2. ***Identical Forgetting Set****: They share the same forgetting dataset* D_f*;*
3. ***Identical Curvature****: They exhibit the same local curvature information at* $w^\star$*;*
4. ***Label Ambiguity****: The labels* y_i *for* $n - 2k + 1 \leq i \leq n - k$ *differ across datasets.*
5. ***Retrained Model Ambiguity:*** *For* $c \in [1, 3]$*, the retained dataset is perfectly realizable by* $w_r(D_r^{(c)}) = (c + 2, 1)$*; while for* $c \in [-1, 0]$*, the retrained model is* $w_r(D_r^{(c)}) = (0, 1)$*.*[3]

Then, upon considering the dataset distribution $P = \frac{1}{2}\delta(D^{(-1)}) + \frac{1}{2}\delta(D^{(3)})$*, from Corollary 1 the retain risk and forget risk can be lower bounded by*

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \geq \frac{23}{8}k, \qquad \mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \geq 25k.$$

[3]For $c \in [-1, 0]$, any model $w = (\alpha, 1)$ with $\alpha \leq 2 + \frac{1}{2k}$ is a minimizer to the retrained loss. According to Footnote 1, we use minimum L_2-norm as a tie-breaker and define $w_r(D_r^{(c)}) = (0, 1)$ as the retrained model.

Proof. We first show the global minimizer $w^\star = (0, 1)$ for any dataset $D^{(c)} \in S(I_{\mathcal{F}})$. For simplicity, denote $w = (w_1, w_2)$ and

$$S_1 := \sum_{i=1}^{n-2k} \left(\text{ReLU}\left(\left(1 + \frac{i}{n-2k}\right) w_2 \right) - \left(1 + \frac{i}{n-2k}\right) \right)^2,$$
$$S_2 := \sum_{j=1}^{k} \left(\text{ReLU}\left(w_1 + \left(-2 - \frac{j}{2k}\right) w_2 \right) - \left(c - \frac{j}{2k}\right) \right)^2,$$
$$S_3 := \sum_{j=1}^{k} \left(\text{ReLU}\left(2w_1 + \left(2 + \frac{j}{k}\right) w_2 \right) - \left(2 + \frac{j}{k}\right) \right)^2.$$

The optimization objective is

$$\min_{w_1, w_2} S_1 + S_2 + S_3.$$

Notice that when $w_1 = 0$, $w_2 = 1$, we have $S_1 = S_3 = 0$ and

$$S_2 = \sum_{j=1}^{k} \left(c - \frac{j}{2k}\right)^2 < 9k.$$

We now show that this is the only optimal solution.

Case 1: $w_2 \le 0$. We have

$$S_1 + S_2 + S_3 \ge S_1 = \sum_{i=1}^{n-2k} \left(1 + \frac{i}{n-2k}\right)^2 > 2n - 4k > 9k.$$

Case 2: $w_1 \ge 2w_2 > 0$. Notice that all samples in S_1 and S_3 are activated. If $0 < w_2 \le \frac{1}{2}$, then

$$S_1 + S_2 + S_3 \ge S_1 \ge \sum_{i=1}^{n-2k} \frac{1}{4} \left(1 + \frac{i}{n-2k}\right)^2 > \frac{2n-4k}{4} \ge 9k.$$

If $w_2 > \frac{1}{2}$, we have

$$\text{ReLU}\left(2w_1 + \left(2 + \frac{j}{k}\right) w_2\right) - \left(2 + \frac{j}{k}\right) \ge 4w_2 + \left(2 + \frac{j}{k}\right) w_2 - \left(2 + \frac{j}{k}\right) > 0.$$

Therefore,

$$S_1 + S_3 \ge \sum_{i=1}^{n-2k} \left(1 + \frac{i}{n-2k}\right)^2 (w_2 - 1)^2 + \sum_{j=1}^{k} \left(\left(6 + \frac{j}{k}\right) w_2 - \left(2 + \frac{j}{k}\right) \right)^2.$$

The right-hand side is minimized when

$$w_2 = 1 - \frac{4\sum_{j=1}^{k}(6 + \frac{j}{k})}{\sum_{i=1}^{n-2k}(1 + \frac{i}{n-2k})^2 + \sum_{j=1}^{k}(6 + \frac{j}{k})^2} =: 1 - \delta,$$

where δ can be bounded by

$$\delta = \frac{4\sum_{j=1}^{k}(6 + \frac{j}{k})}{\sum_{i=1}^{n-2k}(1 + \frac{i}{n-2k})^2 + \sum_{j=1}^{k}(6 + \frac{j}{k})^2} \le \frac{28k}{(n-2k) + 36k} \le \frac{1}{7}.$$

Plugging in $w_2 = 1 - \delta$ we obtain

$$S_1 + S_2 + S_3 \ge S_1 + S_3 \ge \sum_{j=1}^{k} \left(\left(6 + \frac{j}{k}\right) w_2 - \left(2 + \frac{j}{k}\right) \right)^2 \ge \sum_{j=1}^{k} (4 - 7\delta)^2 \ge 9k.$$

Case 3: $w_2 > 0$ and $w_1 < 2w_2$. Then $\text{ReLU}(w_1 + (-2 - \frac{j}{2k})w_2) = 0$ in the second block is always non-activated, which means $S_2 = \sum_{j=1}^{k}(c - \frac{j}{2k})^2$. To minimize $S_1 + S_3$, obviously it takes the min value 0 only when $w_1 = 0$ and $w_2 = 1$.

Combining the three cases, we obtain that $w^\star = (0, 1)$ is the only global minimizer, with the full loss $L(w^\star; D^{(c)}) = \sum_{j=1}^{k}\left(c - \frac{j}{2k}\right)^2$.

Secondly, the forgetting dataset D_f is consistent since the last k samples are the same for all $D^{(c)}$.

Thirdly, since $w^{\star\top} x_i \le -1$ for all $n-2k+1 \le i \le n-k$, $\text{ReLU}(w^\top x_i) = 0$ is locally non-activated at $w^\star$. Therefore, any local statistics of the loss gradient $\nabla L(w; D^{(c)})$ on the full dataset evaluated at the optimal solution $w^\star$ satisfies $\nabla L(w; D^{(c)}) = \nabla(S_1 + S_3)$, which does not depend on the parameter c and thus is consistent for all $D^{(c)} \in S(I_{\mathcal{F}})$.

Fourth, the labels y_i are obviously different for $n - 2k + 1 \le i \le n - k$ across datasets.

Finally, for $c \in [1, 3]$, the retained loss equals 0 only when $w_r(D_r^{(c)}) = (c + 2, 1)$; while for $c \in [-1, 0]$, since the labels y_i for $n - 2k + 1 \le i \le n - k$ are negative, the retain loss is lower-bounded by $\sum_{i=n-2k+1}^{n-k}\left(-c + \frac{i-(n-2k)}{2k}\right)^2$, which can be achieved by any $w = (\alpha, 1)$ with $\alpha \le 2 + \frac{1}{2k}$. According to Footnote 1, we use minimum L_2-norm as a tie-breaker and define $w_r(D_r^{(c)}) = (0, 1)$ as the retrained model.

Next, we consider the dataset distribution $P = \frac{1}{2}\delta(D^{(-1)}) + \frac{1}{2}\delta(D^{(3)})$. For the retain risk, from the definition of the conditional mean label we have

$$\mu_i(X_i|P) = \mathbb{E}_{D\sim P}[Y_i \mid X_i] = \begin{cases} 1 + \frac{i}{n-2k} & \text{for } 1 \le i \le n - 2k \\ 1 - \frac{i-(n-2k)}{2k} & \text{for } n - 2k + 1 \le i \le n - k \end{cases}$$

Hence the label variance under P can be computed by

$$\mathsf{A}_{\mathrm{r}}(P) = \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_r} [(Y_i - \mu(X_i|P))^2] = \sum_{i=n-2k+1}^{n-k} 2^2 = 4k.$$

On the other hand, the expected retrained approximation error $\mathsf{B}_{\mathrm{r}}(P)$ is upper bounded by

$$\mathsf{B}_{\mathrm{r}}(P) = \mathbb{E}_{D\sim P}\left[\min_w L(w; D_r)\right] = \frac{1}{2}\cdot 0 + \frac{1}{2}\sum_{i=n-2k+1}^{n-k}\left(1 + \frac{i-(n-2k)}{2k}\right)^2 \le \frac{9}{8}k.$$

Combining the two parts with Corollary 1 derives a lower bound on the retain risk:

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \ge \max_{P\in\Delta(S(I_{\mathcal{F}}))}(\mathsf{A}_{\mathrm{r}}(P) - \mathsf{B}_{\mathrm{r}}(P)) \ge \frac{23}{8}k.$$

For the forget risk, from the definition of conditional mean prediction of retrained model we have

$$\nu_i(X_i|P) = \mathbb{E}_{D\sim P}[\phi_{w_r(D_r)}(X_i) \mid X_i] = 7 + \frac{i-(n-k)}{k} \quad \text{for } n - k + 1 \le i \le n.$$

Hence the prediction variance of the retrained model is

$$\mathsf{A}_{\mathrm{f}}(P) = \mathbb{E}_{D\sim P} \sum_{(X_i,Y_i)\in D_f} \left[\left(\phi_{w_r(D_r)}(X_i) - \nu_i(X_i|P)\right)^2\right] = 25k.$$

Combining this with Corollary 1 derives a lower bound on the forget risk:

$$\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \ge \max_{P\in\Delta(S(I_{\mathcal{F}}))} \mathsf{A}_{\mathrm{f}}(P) \ge 25k.$$

Hence we finish the proof of Theorem 1. □

D Pareto Tradeoff in Retain and Forget Losses Minimization

D.1 Linear Models

We first show that for linear models, the retain and forget fitting objectives share a common minimizer. Consider a linear model $\phi_w(x) = x^\top w$. Assume that, under the feasible dataset distribution P, the retained and forgetting input matrices X_r and X_f are fixed across all $D \in S(\mathcal{I}_F)$, while the retained labels Y_r may vary. We define the retrained model using the minimum-ℓ_2-norm least-squares solution, $w_r(D_r) = X_r^\dagger Y_r$, where $X_r^\dagger$ denotes the Moore–Penrose pseudoinverse.

For the retain fitting loss, the conditional mean label vector is $\bar{Y}_r := \mathbb{E}_{D\sim P}[Y_r]$. Thus,

$$\varepsilon_{\mathrm{r}}(w_u; P) = \|X_r w_u - \bar{Y}_r\|_2^2.$$

The minimum-ℓ_2-norm minimizer is therefore

$$w_u^\star = X_r^\dagger \bar{Y}_r = X_r^\dagger \mathbb{E}_{D\sim P}[Y_r] = \mathbb{E}_{D\sim P}[X_r^\dagger Y_r] = \mathbb{E}_{D\sim P}[w_r(D_r)].$$

Hence $\mathbb{E}_{D\sim P}[w_r(D_r)]$ minimizes $\varepsilon_{\mathrm{r}}(w_u; P)$.

For the forget fitting loss, by linearity of the model and the fixed forgetting input matrix X_f, the conditional mean prediction of the retrained model on the forgetting inputs is $\mathbb{E}_{D\sim P}[X_f w_r(D_r)] = X_f \mathbb{E}_{D\sim P}[w_r(D_r)]$. Therefore,

$$\varepsilon_{\mathrm{f}}(w_u; P) = \|X_f w_u - X_f \mathbb{E}_{D\sim P}[w_r(D_r)]\|_2^2.$$

Taking $w_u^\star = \mathbb{E}_{D\sim P}[w_r(D_r)]$ makes the above loss equal to zero, and hence minimizes $\varepsilon_{\mathrm{f}}(w_u; P)$. Consequently, for linear models,

$$\mathbb{E}_{D\sim P}[w_r(D_r)] \in \arg\min_{w_u} \varepsilon_{\mathrm{r}}(w_u; P) \cap \arg\min_{w_u} \varepsilon_{\mathrm{f}}(w_u; P).$$

D.2 Sub-Optimality of Affine Combination of Retain and Forget Loss Minimizers

We now show that the solutions $w = (w_1, w_2)$ with $w_1 = a \in (\frac{5}{2}, 3)$ are not Pareto-optimal in equation 3. For the hard distribution

$$P = \frac{1}{2}\delta(D^{(-1)}) + \frac{1}{2}\delta(D^{(3)}),$$

the retain fitting loss is

$$\varepsilon_{\mathrm{r}}(w; P) = \sum_{i=1}^{n-2k} \left[\mathrm{ReLU}\left(w_2\left(1 + \frac{i}{n-2k}\right)\right) - \left(1 + \frac{i}{n-2k}\right)\right]^2 + \sum_{j=1}^{k} \left[\mathrm{ReLU}\left(w_1 - \left(2 + \frac{j}{2k}\right) w_2\right) - \left(1 - \frac{j}{2k}\right)\right]^2.$$

At $w = (a, 1)$, since $a > \frac{5}{2}$, for every $j \in [k]$,

$$a - \left(2 + \frac{j}{2k}\right) \geq a - \frac{5}{2} > 0.$$

Thus all middle-block ReLUs are active. The first block is also active and perfectly fitted at $w_2 = 1$, so its derivative contribution is zero. Therefore,

$$\begin{aligned}
\left.\frac{\partial \varepsilon_{\mathrm{r}}(w; P)}{\partial w_2}\right|_{w=(a,1)} &= \sum_{j=1}^{k} 2\left[a - \left(2 + \frac{j}{2k}\right) - \left(1 - \frac{j}{2k}\right)\right]\left[-\left(2 + \frac{j}{2k}\right)\right] \\
&= \sum_{j=1}^{k} 2(a-3)\left[-\left(2 + \frac{j}{2k}\right)\right] \\
&= 2(3-a)\sum_{j=1}^{k}\left(2 + \frac{j}{2k}\right) > 0,
\end{aligned}$$

where the strict positivity follows from $a < 3$.

Similarly, the forget fitting loss is

$$\varepsilon_{\mathrm{f}}(w; P) = \sum_{j=1}^{k} \left[\mathrm{ReLU}\left(2w_1 + \left(2 + \frac{j}{k}\right) w_2\right) - \left(7 + \frac{j}{k}\right) \right]^2.$$

At $w = (a, 1)$, all forgetting-block ReLUs are active because

$$2a + 2 + \frac{j}{k} > 0.$$

Hence,

$$\begin{aligned} \left.\frac{\partial \varepsilon_{\mathrm{f}}(w; P)}{\partial w_2}\right|_{w=(a,1)} &= \sum_{j=1}^{k} 2 \left[2a + \left(2 + \frac{j}{k}\right) - \left(7 + \frac{j}{k}\right)\right] \left(2 + \frac{j}{k}\right) \\ &= 2(2a - 5) \sum_{j=1}^{k} \left(2 + \frac{j}{k}\right) > 0, \end{aligned}$$

where the strict positivity follows from $a > \frac{5}{2}$.

Therefore, for every $a \in (\frac{5}{2}, 3)$,

$$\left.\frac{\partial \varepsilon_{\mathrm{r}}(w; P)}{\partial w_2}\right|_{w=(a,1)} > 0, \qquad \left.\frac{\partial \varepsilon_{\mathrm{f}}(w; P)}{\partial w_2}\right|_{w=(a,1)} > 0.$$

Consequently, decreasing w_2 slightly below 1 strictly decreases both fitting losses. Thus the interior solutions $(a, 1)$ with $a \in (\frac{5}{2}, 3)$ are dominated and cannot belong to the Pareto frontier.

E Upper Bounds

E.1 Proof of Theorem 2

Theorem 2 (Upper bound). *Let the information set $I_{\mathcal{F}}$ contains the full-dataset minimizer $w^\star \in \arg\min_w L(w; D)$ (i.e., $\mathcal{F}$ contains a set of functions that map to each entry of $w^\star$). Assuming the loss function ℓ is non-negative, the retain risk satisfies*

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \le Ck,$$

where C denotes the maximum loss of an optimal retrained model on any forgetting sample, i.e.,

$$C := \max_{D \in S(I_{\mathcal{F}})} \max_{\substack{(x,y) \in D_f, \\ w_r \in \arg\min_w L(w; D_r)}} \ell(\phi_{w_r}(x); y). \tag{4}$$

Proof. Under the assumption that information $I_{\mathcal{F}}$ includes the global minimizer $w^\star$ of the full loss $L(w; D)$, we consider an unlearning algorithm that directly outputs $w^\star$, then we obtain

$$\begin{aligned} \mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) &= \min_{\mathcal{U}} \max_{D \in S(I_{\mathcal{F}})} \left(\mathbb{E}_{w_u \sim \mathcal{U}(I_{\mathcal{F}})}[L(w_u; D_r)] - \min_w L(w; D_r) \right) \\ &\le \max_{D \in S(I_{\mathcal{F}})} \left(L(w^\star; D_r) - \min_w L(w; D_r) \right). \end{aligned}$$

For every $D \in S(I_{\mathcal{F}})$, denote $w_r \in \arg\min_w L(w; D_r)$. Since $w^\star \in \arg\min_w L(w; D)$, we have

$$\begin{aligned} L(w^\star; D_r) - L(w_r; D_r) &= L(w^\star; D) - L(w^\star; D_f) - L(w_r; D) + L(w_r; D_f) \\ &\le L(w_r; D_f) = \sum_{(x,y) \in D_f} \ell(\phi_{w_r}(x); y) \le Ck, \end{aligned} \tag{5}$$

where the last inequality is derived from the definition of C. Hence, we obtain the desired result

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}) \le \max_{D \in S(I_{\mathcal{F}})} \left(L(w^\star; D_r) - \min_w L(w; D_r) \right) \le Ck.$$

□

E.2 Upper Bounds on the Retain Risk of Negative Gradient and Newton's Method

For a twice-differentiable loss function ℓ, the Hessian matrix on dataset D is defined as $H(w; D) := \nabla_w^2 L(w; D)$. We say that a loss function ℓ is β-smooth if its gradient with respect to w is Lipschitz continuous, i.e., for all w_1, w_2, x, y,

$$\|\nabla_w \ell(\phi_{w_1}(x); y) - \nabla_w \ell(\phi_{w_2}(x); y)\| \leq \beta \|w_1 - w_2\|.$$

For an unlearning algorithm $\mathcal{U}$, we define its retain risk as

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}; \mathcal{U}) = \max_{D \in S(I_{\mathcal{F}})} \Big(\mathbb{E}_{w_u \sim \mathcal{U}(I_{\mathcal{F}})}[L(w_u; D_r)] - \min_w L(w; D_r) \Big).$$

Negative Gradient.

Proposition 3. *Let the differentiable loss function $\ell(\phi_w(x); y)$ be non-negative and β-smooth with respect to w for all x, y. Assume additionally that $\nabla_w L(w^\star; D) = 0$ for every $D \in S(I_{\mathcal{F}})$. Let C be defined as in equation 4. If $\lambda \geq \beta$, then*

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}; \mathcal{U}_{\textit{NegGrad}}) \leq Ck.$$

Proof. Due to the optimality of w_u, we have

$$-L(w_u; D_f) + \tfrac{\lambda n}{2} \|w_u - w^\star\|^2 \leq -L(w^\star; D_f).$$

Since $\ell(\phi_w(x); y)$ is β-smooth, $L(w; D_f) = \sum_{(x,y) \in D_f} \ell(\phi_w(x); y)$ is βk-smooth, we have

$$-L(w_u; D_f) \geq -L(w^\star; D_f) + \langle -\nabla L(w^\star; D_f), w_u - w^\star \rangle - \tfrac{\beta k}{2} \|w_u - w^\star\|^2.$$

Combining the above two inequalities derives

$$\langle -\nabla L(w^\star; D_f), w_u - w^\star \rangle \leq -\tfrac{\lambda_n - \beta k}{2} \|w_u - w^\star\|^2.$$

Since $\ell(\phi_w(x); y)$ is β-smooth, $L(w; D_r) = \sum_{(x,y) \in D_r} \ell(\phi_w(x); y)$ is $\beta(n-k)$-smooth, thus we can bound

$$\begin{aligned}
L(w_u; D_r) &\leq L(w^\star; D_r) + \langle \nabla L(w^\star; D_r), w_u - w^\star \rangle + \tfrac{\beta(n-k)}{2} \|w^\star - w_u\|^2 \\
&= L(w^\star; D_r) + \langle -\nabla L(w^\star; D_f), w_u - w^\star \rangle + \tfrac{\beta(n-k)}{2} \|w^\star - w_u\|^2 \\
&\leq L(w^\star; D_r) - \tfrac{\lambda_n - \beta k}{2} \|w_u - w^\star\|^2 + \tfrac{\beta(n-k)}{2} \|w^\star - w_u\|^2 \\
&= L(w^\star; D_r) + \tfrac{(\beta - \lambda)n}{2} \|w^\star - w_u\|^2 \\
&\leq \min_w L(w; D_r) + Ck,
\end{aligned}$$

where the second line is due to the stationarity assumption $\nabla_w L(w^\star; D) = 0$ and the decomposition $L(w; D) = L(w; D_r) + L(w; D_f)$, and the last line is due to the assumption $\lambda \geq \beta$ and equation 5 that $L(w^\star; D_r) - \min_w L(w; D_r) \leq Ck$. Since this bound holds for any $D \in S(I_{\mathcal{F}})$, we obtain the desired result

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}; \mathcal{U}_{\mathrm{NegGrad}}) = \max_{D \in S(I_{\mathcal{F}})} \Big(L(w_u; D_r) - \min_w L(w; D_r) \Big) \leq Ck.$$

This completes the proof of Proposition 3. □

Newton's Method

Proposition 4. *Let the differentiable loss function $\ell(\phi_w(x); y)$ be non-negative and β-smooth with respect to w for all x, y. Let C be defined as in equation 4. If $\lambda \geq \frac{3}{2}\beta$, then*

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}; \mathcal{U}_{\textit{Newton}}) \leq Ck.$$

Proof. Denote $M(w) := H(w; D_r) + \lambda(n-k)I$ for simplicity. Since the loss function $\ell(\phi_w(x); y)$ is β-smooth with respect to w for all x, y, we know $L(w; D_r)$ is $\beta(n-k)$-smooth and $M(w) \succeq (\lambda - \beta)(n-k)I$. Therefore, we have

$$\begin{aligned} L(w_u; D_r) &\le L(w^\star; D_r) + \langle \nabla L(w^\star; D_r), w_u - w^\star \rangle + \frac{\beta(n-k)}{2} \|w^\star - w_u\|^2 \\ &= L(w^\star; D_r) + \nabla L(w^\star; D_r)^\top \left(-(M(w^\star))^{-1} + \tfrac{\beta(n-k)}{2} (M(w^\star))^{-2} \right) \nabla L(w^\star; D_r). \end{aligned}$$

Since $M(w) \succeq (\lambda - \beta)(n-k)I$ and $\lambda \ge \frac{3}{2}\beta$, we have

$$-I + \tfrac{\beta(n-k)}{2} (M(w^\star))^{-1} \preceq -I + \tfrac{\beta}{2(\lambda-\beta)} I \preceq 0.$$

Combining this with equation 5 that $L(w^\star; D_r) - \min_w L(w; D_r) \le Ck$ derives

$$L(w_u; D_r) \le L(w^\star; D_r) \le \min_w L(w; D_r) + Ck$$

Hence we obtain the desired result

$$\mathsf{R}_{\mathrm{r}}(I_{\mathcal{F}}; \mathcal{U}_{\text{Newton}}) = \max_{D \in S(I_{\mathcal{F}})} \left(L(w_u; D_r) - \min_w L(w; D_r) \right) \le Ck.$$

This completes the proof of Proposition 4. □

At a high level, our information-theoretic lower bounds demonstrate why excess risk scaling linearly in k is often unavoidable given incomplete access to D_r, while the upper bounds show that simple first- and second-order methods under regularity conditions achieve this order-wise optimal scaling.

F Connections to Privacy Definitions

The forget risk measures forget-set retraining discrepancy, rather than a general privacy guarantee. We use certified removal (Guo et al., 2020), a privacy definition based on retraining indistinguishability, as a concrete example of how forget-risk lower bounds can translate into necessary conditions on privacy parameters. Let $Q_{I_{\mathcal{F}}}$ be the output distribution of an unlearning algorithm using only $I_{\mathcal{F}}$, and let P_D be the output distribution of the retraining procedure on D_r. For $\epsilon \ge 0$ and $0 \le \delta < 1$, (ϵ, δ)-certified removal requires, for every feasible dataset D and measurable $E \subseteq \mathcal{W}$,

$$Q_{I_{\mathcal{F}}}(E) \le e^\epsilon P_D(E) + \delta, \qquad P_D(E) \le e^\epsilon Q_{I_{\mathcal{F}}}(E) + \delta.$$

We specialize this definition to the deterministic retraining reference used in this paper.

Proposition 5 (From forget risk to certified removal)**.** *In least squares regression, suppose P_D is the point mass at $w_r(D_r)$, and, for some $B > 0$, $\left(\phi_w(X_i) - \phi_{w_r(D_r)}(X_i)\right)^2 \le B$ for every $D \in S(I_{\mathcal{F}})$, $w \in \mathcal{W}$, and $(X_i, Y_i) \in D_f$. If $\mathcal{U}(I_{\mathcal{F}})$ satisfies (ϵ, δ)-certified removal for every $D \in S(I_{\mathcal{F}})$, then*

$$\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \le \mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}; \mathcal{U}) \le \delta k B,$$

where $\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}; \mathcal{U})$ denotes the worst-case forget risk of algorithm $\mathcal{U}$. Specifically, a lower bound $\mathsf{R}_{\mathrm{f}}(I_{\mathcal{F}}) \ge ck$ for some constant $c > 0$ implies $\delta \ge c/B$.

Proof. For each $D \in S(I_{\mathcal{F}})$, take $E = \mathcal{W} \setminus \{w_r(D_r)\}$. Since $P_D(E) = 0$, certified removal gives $Q_{I_{\mathcal{F}}}(E) \le \delta$. The total squared prediction discrepancy on D_f vanishes at $w_r(D_r)$ and is at most kB elsewhere. Its expectation is therefore at most $\delta k B$, which yields the claim. □

If $\operatorname{diam}(\mathcal{W}) \le d_{\mathcal{W}}$ and each forget-sample prediction map is uniformly L_0-Lipschitz in w, we may take $B = L_0^2 d_{\mathcal{W}}^2$. Furthermore, in the hardness example 1, applying Theorem 1 then gives $\delta \ge \frac{25}{B}$, a constant lower bound independent of ϵ, n, and k when B is uniformly bounded independently of n, k.