WHY FORGET-ONLY UNLEARNING NEEDS MEMORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning asks for a deletion algorithm whose output is close to retraining from scratch without the selected forget examples. In this work, we study forget-only unlearning, where the deletion algorithm receives only the trained model and the examples to forget, with no retained data or extra training information. We ask whether forget-only unlearning is always possible. We first show that this depends on the learning method: different datasets can produce the same trained model but require very different outputs after the same examples are removed. Using this observation, we derive lower bounds on how accurately unlearning can match retraining and instantiate it for several standard learning algorithms. We then ask what must be true when forget-only unlearning succeeds. To this end, we derive lower bounds on what an algorithm must memorize about the training data to handle arbitrary deletion requests. For simple threshold learners, the required information can be as large as the entire dataset, even though ordinary training keeps only one boundary point. Overall, our results show that information discarded during ordinary learning may be needed later for deletion, so models designed for forget-only unlearning may need to retain more information than standard training does.

1 INTRODUCTION

As machine learning models are more widely deployed, the ability to retroactively remove specific data from their training sets has become important. This need arises from privacy regulations, such as the EU GDPR's right to be forgotten (GDPR (2016)), and from the accidental inclusion of poisoned (Schoepf et al., 2024), copyrighted (Dou et al., 2025), or unlawfully collected (Thiel, 2023) data. While retraining a model on the remaining data is the gold standard, doing so is often too expensive or practically impossible. Certified machine *unlearning*, therefore, asks for a cheaper procedure whose output is close in a distributional sense to the model obtained by retraining from scratch on the retained data (Cao & Yang, 2015).

An unlearning scheme consists of a learning algorithm $\mathcal{A}$ and an unlearning algorithm $\bar{\mathcal{A}}$. Given a training dataset S, the learner outputs a model $M \coloneqq \mathcal{A}(S)$. The unlearner receives a forget set $U \subseteq S$ and the trained model, and outputs the unlearned model $\overline{W} \coloneqq \bar{\mathcal{A}}(M, U)$, with the goal of being distributionally close to the retraining target $W \coloneqq \mathcal{A}(S \setminus U)$. In this paper we study the *forget-only* setting: at deletion time, $\bar{\mathcal{A}}$ has access only to M and U, with no retained data or other training information except what is already stored in M. This raises a difficulty: information discarded during training could be necessary to determine the correct retraining target after deletion.

We adopt a Rényi definition of unlearning that holds pointwise over datasets and deletion requests. A pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (α, ε)-Rényi unlearning (RU) (Definition 1) if, for every S, U, $d_\alpha\big(\bar{\mathcal{A}}(M, U),\, \bar{\mathcal{A}}(W, \emptyset)\big) \leq \varepsilon$, where d_α denotes the symmetric α-Rényi divergence. We note that Rényi unlearning implies the standard ε', δ-unlearning for appropriate choice of ε'. Similar to Sekhari et al. (2021) we compare the unlearned output with $\bar{\mathcal{A}}(W, \emptyset)$ to allow for deterministic learners. However, indistinguishability alone does not guarantee utility: an unlearner could output the same distribution regardless of its inputs and still satisfy RU. We therefore also require the empty-request unlearning to preserve utility by remaining close to its input, with high probability in the chosen metric (see Definition 2).

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.

Several empirical unlearning methods are forget-only or close to forget-only, including commonly used methods based on gradient ascent on the forget set (Zhang et al., 2024; Fan et al., 2024; Jang et al., 2023). These methods can work well in some settings, but fail in others (Mavrothalassitis et al., 2025; Yu et al., 2025). This leads to our first question:

Question 1. *Is forget-only unlearning always possible? If not, what conditions make it impossible?*

Our first result in Section 3 shows that forget-only unlearning is not always possible. Intuitively, the obstruction arises when many (say two) datasets S_1, S_2 produce the same trained model, $\mathcal{A}(S_1) = \mathcal{A}(S_2)$, but deleting the same forget set U pushes their corresponding retraining targets far apart. For example, consider a threshold learner that returns the leftmost positive point. Two datasets may share this point but have different second-leftmost positive points, which become the respective retraining targets when the shared point is deleted. A forget-only unlearner receives the same input in both cases, namely $(\mathcal{A}(S_1), U) = (\mathcal{A}(S_2), U)$ and must therefore produce the same output distribution. However, that output cannot be close to two far apart retraining targets if they are required to preserving utility on empty unlearning requests. We generalize these ideas to lower bound the unlearning parameter ε Theorem 3.1, and instantiate it for several common learning algorithms (Examples 2 to 4). One consequence is the following.

Corollary 1.1 (Informal: hard-margin SVM)**.** *Let $\mathcal{A}$ be hard-margin linear SVM in dimension $d \geq 2$. Then, no forget-only unlearner can satisfy (α, ε)-Rényi unlearning with a fixed finite ε uniformly over all datasets and deletion requests.*

The hardness above arises because the trained SVM has discarded information needed after deletion. If the model does not memorize enough information to distinguish datasets that look identical before deletion but different after deletion, then the unlearner has no way to recover the correct retraining target. This leads to the second question:

Question 2. *How much information about the training dataset must the learned model memorize in order to support future forget-only unlearning requests?*

Following Brown et al. (2021); Feldman et al. (2025), we measure memorization through the *mutual information* $I(M; S)$, where S is drawn from an analysis distribution over datasets. Our main contribution is a technique for lower-bounding $(I(M; S))$ using the retraining targets induced by possible deletion requests, given the unlearning and utility requirements. We prove the bound for discrete model spaces Section 4.1, extend it to continuous spaces Appendix E, and instantiate it for concrete learning problems.

Since the unlearning guarantee is pointwise, it must hold for every S and every nonempty request $U_i \subseteq S$. The model M must therefore retain enough information to support each retraining target $W_i := \mathcal{A}(S \setminus U_i)$. Memorization $I(M; S)$ might seem to add across requests, but requests can reveal overlapping information: e.g., for a threshold learner, many deletions can expose the same next positive point. Thus the bound must count only new, nonredundant information per request. We formalize this via a permutation π of the requests. The bound holds for every ordering, so we can choose the one giving the strongest lower bound. The following informal statement summarizes our result for finite model spaces.

Theorem 1.2. *Let S be a training dataset, $M = \mathcal{A}(S)$, $(\mathcal{A}, \bar{A})$ satisfy (α, ε)-Rényi unlearning, and $U_1, \ldots, U_K$ be the unlearning requests. For each request, let $W_i = \mathcal{A}(S \setminus U_i)$ be the corresponding retrained model. Then, for any ordering π of the requests,*

$$I(M; S) \gtrsim \sum_{i=1}^{K} H\big(W_{\pi(i)} \mid W_{\pi(<i)}, U_{\pi(\leq i)}\big) - K \cdot \mathrm{err}(\alpha, \varepsilon).$$

The first term captures the genuinely new information in the respective retraining targets while the second term gives some slack for the approximation error. We instantiate this bound for several learning algorithms and hypothesis classes in Section 4.2. For canonical threshold learners, this bound gives a concrete separation between the information retained by ordinary learning and the information required for forget-only unlearning. We show that over datasets of size n from a domain of size N, the ordinary learner keeps only one boundary point and carries $\log(N/n)$ bits of information about the dataset. However, supporting unlearning requests of size at most m, with $n \geq 2m$, requires memorizing $\Omega(m \log(N/n))$ bits of information.

Closest to our work, Cherapanamjeri et al. (2025) lower bounds the space complexity of learning–unlearning algorithms in terms of the eluder dimension of the hypothesis class. Our bounds depend

on the particular learning algorithm and the retraining targets induced by the chosen deletion requests. Additionally, their bounds are not for forget-only algorithms and mainly apply to the pure unlearning setting. Taken together, our results give, to our knowledge, one of the first formal accounts of why memorization is necessary for forget-only unlearning. Thus, forget-only unlearning entails a basic tradeoff: either the learner preserves enough information for future deletions, or the unlearner must rely on additional retained data, training state, or surrogate information.

2 PRELIMINARIES

Datasets and deletion requests. Let $\mathcal{X}$ be the data space and let $\mathcal{W}$ be the model space. A dataset of size n is an ordered tuple $S = (x_1, \dots, x_n) \in \mathcal{X}^n$. A deletion request (or forget set) is denoted by $U \subseteq S$, with $|U| \le m$, where $0 \le m \le n$ is the deletion budget. We write $S \setminus U$ for the retained dataset.

The unlearning guarantee is pointwise: S and U are fixed, and probabilities are taken only over algorithmic randomness. In Section 4, where we lower-bound $I(M; S)$, we let S be random under an analysis distribution $\mathcal{P}_S$ over datasets. No i.i.d. assumption is made unless stated explicitly.

Learning and forget-only unlearning. A learning algorithm is a possibly randomized map $\mathcal{A} : \mathcal{X}^\star \to \mathcal{W}$, where $\mathcal{X}^\star$ is the set of finite datasets. An unlearning algorithm is a randomized map $\bar{\mathcal{A}} : \mathcal{W} \times \mathcal{X}^\star \to \mathcal{W}$. For fixed $S \in \mathcal{X}^n$ and $U \subseteq S$, define $M := \mathcal{A}(S)$, $W := \mathcal{A}(S \setminus U)$, $\overline{W} := \bar{\mathcal{A}}(M, U)$, and $W^\emptyset := \bar{\mathcal{A}}(W, \emptyset)$. We call these the pretrained model, retraining target, unlearned model, and empty-request output, respectively.

The unlearner $\bar{\mathcal{A}}$ is *forget-only*: at deletion time, its only inputs are the pretrained model M and the forget set U. It has no access to the retained data, training logs, checkpoints, gradients, or auxiliary statistics beyond any information stored in M. This differs from settings where the unlearner may also receive second-order information or the retained dataset (Sekhari et al., 2021; Neel et al., 2021).

Rényi unlearning. For distributions P, Q and $\alpha > 1$, we use the symmetric α-Rényi divergence $d_\alpha(P, Q) := \max\{D_\alpha(P \,\|\, Q), D_\alpha(Q \,\|\, P)\}$, where $D_\alpha(P \,\|\, Q) := \frac{1}{\alpha-1} \log \int \left(\frac{dP}{dQ}\right)^\alpha dQ$, with the convention $D_\alpha(P \,\|\, Q) = \infty$ if P is not absolutely continuous with respect to Q. We write $\mathrm{Law}(X)$ for the law of a random variable X and use the shorthand $d_\alpha(X, Y) := d_\alpha(\mathrm{Law}(X), \mathrm{Law}(Y))$.

Definition 1 (Rényi unlearning)**.** Fix $\alpha > 1$ and $\varepsilon \ge 0$. We say that $\left(\mathcal{A}, \bar{\mathcal{A}}\right)$ satisfies (α, ε)-Rényi unlearning (RU) over deletion sets of size at most m if, for every dataset $S \in \mathcal{X}^n$ and every request $U \subseteq S$ with $|U| \le m$,

$$d_\alpha\left(\bar{\mathcal{A}}\left(\mathcal{A}(S), U\right), \bar{\mathcal{A}}\left(\mathcal{A}(S \setminus U), \emptyset\right)\right) \le \varepsilon.$$

When m is clear from context, we simply say that $\left(\mathcal{A}, \bar{\mathcal{A}}\right)$ satisfies (α, ε)-Rényi unlearning.

The reference $W^\emptyset = \bar{\mathcal{A}}(W, \emptyset)$ is obtained by applying unlearning the empty set to the retrained model. This allows us to have randomized unlearning outputs even when $\mathcal{A}$ is deterministic (Sekhari et al., 2021). However, an unlearner that always returns the same distribution would satisfy RU regardless of its inputs. We therefore impose an empty-request utility requirement: $\bar{\mathcal{A}}(w, \emptyset)$ should remain close to w in model space.

Definition 2 (Γ-utility on empty requests)**.** Fix a metric $d_{\mathcal{W}}$ on $\mathcal{W}$ and a non-increasing function $\Gamma : [0, \infty) \to [0, 1]$. For each $w \in \mathcal{W}$, let $Q_w := \mathrm{Law}(\bar{\mathcal{A}}(w, \emptyset))$. We say that $\bar{\mathcal{A}}$ has $(d_{\mathcal{W}}, \Gamma)$-utility on empty requests if, for every $w \in \mathcal{W}$ and every radius $r \ge 0$,

$$\mathbb{P}_{W^\emptyset \sim Q_w}\left[d_{\mathcal{W}}(W^\emptyset, w) > r\right] \le \Gamma(r).$$

For discrete $\mathcal{W}$, we use the zero-one metric $d_{\mathcal{W}}(w, w') = \mathbb{I}\{w \ne w'\}$. At $r = 0$, writing $\eta := \Gamma(0)$, the utility condition gives $\mathbb{P}_{W^\emptyset \sim Q_w}[W^\emptyset \ne w] \le \eta$ for every $w \in \mathcal{W}$. For a vector space $\mathcal{W}$ with norm ϕ, we use $d_{\mathcal{W}}(w, w') = \phi(w - w')$ and write (ϕ, Γ)-utility. The example below shows that the Gaussian mechanism satisfies this utility condition.

Example 1 (Gaussian empty-request map)**.** Suppose $\mathcal{W} = \mathbb{R}^d$ and $\bar{\mathcal{A}}(w, \emptyset) = w + Z$, where $Z \sim \mathcal{N}(0, \sigma^2 I_d)$ and $\sigma > 0$. Then $\bar{\mathcal{A}}$ has $(\|\cdot\|_2, \Gamma)$-utility with $\Gamma(r) = 1$ for $r < \sigma\sqrt{d}$, and $\Gamma(r) = \exp\left(-(r - \sigma\sqrt{d})^2/(2\sigma^2)\right)$ otherwise.

3 IMPOSSIBILITY OF FORGET-ONLY UNLEARNING

In this section, we show that forget-only unlearning cannot, in general, achieve both small ε and nontrivial utility. Such an unconditional lower bound cannot hold for every learner: a learner that memorizes the full dataset, e.g., by storing an encoding of the entire training dataset in the model, can recover the retained dataset from (M, U). Any such lower bound must therefore necessarily depend on the learner. Our lower bound depends on a learner-specific condition, that we define as a *shared-deletion example*. This also connects to Section 4, where we quantify the amount of memorization required for successful unlearning.

Our construction is based on the idea that there are multiple datasets whose differences are invisible while learning, but become clear after deletion. Specifically, N datasets produce identical or close distributions over pretrained models, yet their retraining outputs are separated after deleting the same forget set. We formalize this condition below, with Δ measuring the separation between the retraining outputs.

Definition 3 (Shared-deletion example)**.** Fix a learning algorithm $\mathcal{A}$, a norm ϕ, integers $N \geq 2$ and $m \geq 0$, and parameters $\alpha > 1$, $\Delta > 0$, $\delta_\alpha \geq 0$, and $\kappa \in [0, 1]$. We call $(S_1, \ldots, S_N, U)$, with $S_i \in \mathcal{X}^n$, an $(N, m, \Delta, \phi; \delta_\alpha, \kappa)$-*shared-deletion example* (s.d.e.) for $\mathcal{A}$ if:

1. $U \subseteq \bigcap_{i=1}^{N} S_i$ and $|U| \leq m$.
2. For every distinct $i, j \in [N]$, $d_\alpha\big(\mathcal{A}(S_i), \mathcal{A}(S_j)\big) \leq \delta_\alpha$.
3. There exist measurable sets $E_1, \ldots, E_N \subseteq \mathcal{W}$ such that, for every distinct $i, j \in [N]$, $\phi(x - z) \geq \Delta$ for all $x \in E_i,\ z \in E_j$, and, for every $i \in [N]$, $\Pr_{W \sim \mathcal{A}(S_i \setminus U)}[W \in E_i] \geq 1 - \kappa$.

An N-sized s.d.e. consists of N distinct datasets whose pretrained-model distributions are close, but whose retraining distributions concentrate on well-separated sets after deleting the same forget set. Similar packing conditions appear in several lower bound proofs like for Differential Privacy (Hardt & Talwar, 2010). We use this condition to prove our main impossibility result.

Theorem 3.1. *Fix $\alpha > 1$. Let $\mathcal{A}, \bar{\mathcal{A}}$ be (α, ε)-RU for all deletion requests of size at most m and let $\bar{\mathcal{A}}$ have (ϕ, Γ)-utility on empty requests. If $\mathcal{A}$ admits an $(N, m, \Delta, \phi; \delta_\alpha, \kappa)$-s.d.e., then the following bounds hold.*

Exact collision ($\delta_\alpha = 0$):

$$\varepsilon \geq \sup_{r \in R} \max \left\{ \log N + \frac{1}{\gamma} \log(1 - u(r)), \frac{1}{\gamma} \log \frac{N-1}{N} - \log u(r) \right\}.$$

Approximate collision ($\delta_\alpha > 0$):

$$\varepsilon \geq \sup_{r \in R} \max \left\{ \gamma \left(\log N - \delta_\alpha\right) + \frac{1}{\gamma} \log(1 - u(r)), \frac{1}{\gamma^2} \log \frac{N-1}{N} - \frac{1}{\gamma}\delta_\alpha - \log u(r) \right\},$$

where $R := \{r \in (0, \Delta/2) : 0 < u(r) < 1\}$, $u(r) = \kappa + \Gamma(r)$, and $\gamma = \frac{\alpha - 1}{\alpha}$.

We provide a proof sketch for the exact-collision below. The full proof is deferred to Appendix C.

Proof sketch. Write $M_i = \mathcal{A}(S_i)$, $W_i = \mathcal{A}(S_i \setminus U)$, and $Q_i = \mathrm{Law}(\bar{\mathcal{A}}(W_i, \emptyset))$. Since $\delta_\alpha = 0$, the pretrained models M_i have the same distribution. The forget-only unlearner therefore has the same output law $T = \mathrm{Law}(\bar{\mathcal{A}}(M_i, U))$ for every i. The unlearning guarantee gives $d_\alpha(T, Q_i) \leq \varepsilon$.

Fix $r \in R$. Due to property (3) of shared deletion examples (using Lemma C.1), there are pairwise disjoint sets B_i with $Q_i(B_i) \geq 1 - u(r)$. Applying the event transfer lemma Lemma B.1 gives

$$1 - u(r) \leq Q_i(B_i) \leq e^{\gamma\varepsilon} T(B_i)^\gamma.$$

As all B_i are disjoint, summing over them gives us

$$1 \geq \sum_{i=1}^{N} T(B_i) \geq N e^{-\varepsilon} (1 - u(r))^{1/\gamma},$$

$$\implies \varepsilon \geq \log N + \frac{1}{\gamma} \log(1 - u(r)).$$

The second term in the max follows by applying the same steps to B_i^c. Taking the supremum over $r \in R$ completes the proof. □

Theorem 3.1 provides a bound on the unlearning guarantee whenever shared-deletion examples exist. We next show that shared deletion examples arise in simple, standard learning problems. We provide a stronger result in Proposition C.2 showing that any empirical risk minimizer for zero-one loss on a hypothesis class with at least two distinct hypotheses and finite teaching dimension (Goldman & Kearns, 1995) is guaranteed to have an s.d.e. for sufficiently large dataset sizes. We relegate this result to Appendix C.1 to avoid introducing additional notation here. The examples below illustrate the range of possible Δ in shared deletion examples: a constant fraction of the model-space diameter, the full diameter, or an unbounded gap. The examples are stated for $N = 2$, but their constructions extend to larger N, with suitable choices of the separation and domain parameters.

Example 2 (Canonical thresholds)**.** Fix a dataset size $n \geq 3$, a deletion budget $m \in [n-2]$, and an integer $c \in \mathbb{Z}_+$ with $c > 1$. Let the data space be $\mathcal{X}_c \times \{-1, +1\}$, where $\mathcal{X}_c = \{-c, -c+1, \ldots, c\}$. On a realizable sample containing both labels, the canonical threshold learner $\mathcal{A}_c^{\text{th}}$ returns the midpoint between the largest negative example and the smallest positive example. This learner admits a $(2, m, c - \frac{1}{2}, |\cdot|\,; 0, 0)$-s.d.e.

Example 3 (Empirical median)**.** Fix a dataset size $n \geq 3$, a deletion budget $m \in [n-1]$, and $c > 0$. Let $\mathcal{A}_c^{\text{med}}$ be the empirical median learner on samples supported on $[-c, c]$. The largest Δ for which this learner admits a $(2, m, \Delta, |\cdot|\,; 0, 0)$-s.d.e. is as follows. If $n = 2r$ for some $r \in \mathbb{N}$, then $\Delta = c$ for $m \in \{1, 2\}$, and $\Delta = 2c$ for $m \geq 3$. If $n = 2r - 1$ for some $r \in \mathbb{N}$, then $\Delta = c$ for $m = 1$, and $\Delta = 2c$ for $m \geq 2$.

Example 4 (Hard-margin linear classification)**.** Fix a dataset size $n \geq 3$, a deletion budget $m \in [n-1]$, and $d \geq 2$. Let $\mathcal{A}_d^{\text{svm}}$ be the hard-margin homogeneous linear SVM on realizable samples supported on $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\} \times \{-1, +1\}$. That is, $\mathcal{A}_d^{\text{svm}}$ returns the minimum-norm separator w satisfying $y \langle w, x \rangle \geq 1$ for all training examples (x, y). For every $\Delta > 0$, this learner admits a $(2, 1, \Delta, \|\cdot\|_2\,; 0, 0)$-s.d.e.

The learner returns the unnormalized SVM vector, so its model space is unbounded and the separation Δ can be arbitrarily large. If the learner instead returns the normalized direction $w/\|w\|_2$, the same construction gives a $(2, 1, \sqrt{2}, \|\cdot\|_2\,; 0, 0)$-s.d.e.

Combining these examples with Theorem 3.1 gives lower bounds on ε for the corresponding separations Δ, with $N = 2$ and $\delta_\alpha = \kappa = 0$. For the SVM learner, shared-deletion examples exist for every finite separation $\Delta > 0$. Under the Gaussian empty-request unlearning from Example 1 with $\sigma > 0$, taking $r = \Delta/4$ in the second exact-collision bound gives

$$\varepsilon \geq \frac{(\Delta/4 - \sigma\sqrt{d})^2}{2\sigma^2} - \frac{\log 2}{\gamma}$$

for $\Delta > 4\sigma\sqrt{d}$. Since Δ can be arbitrarily large, this lower bound grows arbitrarily large. Thus, for this SVM learner, no forget-only unlearner with fixed-variance Gaussian empty-unlearning post-processing can satisfy a uniform (α, ε)-Rényi unlearning guarantee with finite ε, even for single-example deletions.

4 Memorization for Forget-Only Unlearning

The previous section identified an obstruction to forget-only unlearning: a uniform (α, ε)-RU guarantee is controlled by the hardest shared-deletion instance of the learner $\mathcal{A}$. In the examples above, two datasets share a common part that *masks* a difference between them, so that they collide under $\mathcal{A}$; deleting that shared part removes the mask, exposing the difference and separating the retraining targets, which makes forget-only unlearning impossible. We now ask the complementary question: when forget-only unlearning *is* possible, how much information about S must $\mathcal{A}(S)$ store to support it?

Forget-only unlearning can succeed only if the retraining target is determined by the pair $(\mathcal{A}(S), U)$, for every admissible deletion U. We quantify this required information through $I(\mathcal{A}(S); S)$, where $S \sim \mathcal{P}_S$ is an analysis distribution over datasets. For any collection of admissible requests $\mathcal{U} =$

$\{U_1, \ldots, U_K\} \subseteq 2^S$ we choose, the trained model must then retain enough information to determine each retraining target $\mathcal{A}(S \setminus U)$, giving the guiding intuition for our memorization lower bounds: $I(\mathcal{A}(S); S) \geq I\big(\mathcal{A}(S); \{\mathcal{A}(S \setminus U)\}_{U \in \mathcal{U}}\big)$.

Setup. Draw $S \sim \mathcal{P}_S$ and allow $\mathcal{A}$ to be randomized. The requests $\mathcal{U} = \{U_1, \ldots, U_K\}$ are deterministic functions of S, with $U_i \subseteq S$ [1]. For each $i \in [K]$, define $M := \mathcal{A}(S)$, $W_i := \mathcal{A}(S \setminus U_i)$, $\overline{W}_i := \bar{\mathcal{A}}(\mathcal{A}(S), U_i)$, $W_i^\emptyset := \bar{\mathcal{A}}(\mathcal{A}(S \setminus U_i), \emptyset)$. Thus, M is the pretrained model, W_i is the retraining target for the request U_i, $\overline{W}_i$ is the unlearning output, while $W_i^\emptyset$ is the empty-request output of W_i. For a tuple $A^K := (A_1, \ldots, A_K)$ and a permutation $\pi : [K] \mapsto [K]$, write $A_{\pi(\leq i)} := \big(A_{\pi(1)}, \ldots, A_{\pi(i)}\big)$ and $A_{\pi(<i)} := \big(A_{\pi(1)}, \ldots, A_{\pi(i-1)}\big)$. When $i = 1$, $A_{\pi(<1)}$ is the empty tuple. Figure 1 records the conditional independences used below: M is independent of (U_i, W_i) given S, $\overline{W}_i$ depends on the rest only through M, U_i), and $W_i^\emptyset$ depends only on W_i.

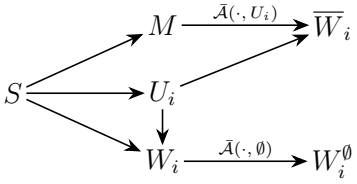

Figure 1: Markov graph for a single deletion request U_i.

The following proposition serves as a basis for deriving concrete memorization lower bounds in both the discrete and continuous settings.

Proposition 4.1. *Let $S \sim \mathcal{P}_S$, and let $U_1, \ldots, U_K$ be requests generated as above. For each $i \in [K]$, define $M, W_i, \overline{W}_i$ as above. Then, for every permutation $\pi : [K] \to [K]$,*

$$I(M; S) \geq \sum_{i=1}^{K} I\big(W_{\pi(i)}; \overline{W}_{\pi(i)} \,\big|\, W_{\pi(<i)}, U_{\pi(\leq i)}\big)\,.$$

4.1 Discrete model space

We consider a discrete, finite model space $\mathcal{W}$. Two sources of error affect whether $\overline{W}$ can recover the retraining target W: the noise from comparing against $W^\emptyset = \bar{\mathcal{A}}(W, \emptyset)$ rather than W itself, and the unlearning error itself. We measure this recovery via deterministic decoders g, since the relevant question is how much information about W is recoverable from the observed output: $g(z) = z$ gives direct comparison, while the optimal g gives the best possible recovery error, the natural quantity for the Fano-type bounds below. The decoder observes only $W^\emptyset$ for empty requests, and also u for unlearning requests; the definitions below take the optimal g in each case.

Definition 4 (Discrete recovery errors)**.** The *empty-request recovery error* is defined as,

$$\beta_\emptyset^\star := \inf_{g: \mathcal{W} \to \mathcal{W}} \sup_{w \in \mathcal{W}} \mathbb{P}\left(g\left(W^\emptyset\right) \neq w \mid W = w\right),$$

where $W^\emptyset = \bar{\mathcal{A}}(W, \emptyset)$ is the empty-request output, and the randomness in the probability is over the algorithmic randomness of $\bar{\mathcal{A}}$ only.

Next, recall that the retraining target is $W = \mathcal{A}(S \setminus U)$ and the unlearning output is $\overline{W} = \bar{\mathcal{A}}(M, U)$. The runs of $\mathcal{A}$ for fixed $S = s, U = u$ are independent when $\mathcal{A}$ is randomized. The *unlearning recovery error* is defined as,

$$\beta_{\text{un}}^\star := \inf_{g: \mathcal{W} \times \mathcal{X}^\star \to \mathcal{W}} \sup_{\substack{s, u:\, u \subseteq s \\ |u| \leq m}} \mathbb{P}\left(g\left(\overline{W}, u\right) \neq W \mid S = s, U = u\right)$$

Here, the randomness in the probability is over the algorithmic randomness of $\mathcal{A}$ and $\bar{\mathcal{A}}$, conditional on fixed $S = s$ and $U = u$.

The same map g must work uniformly over all inputs, or the recovery error becomes vacuous: allowed to depend on W it could simply output the true target, and allowed to depend on (S, U) it

[1] They can also be randomized; the distribution of U_i must be independent of the algorithmic randomness.

could compute $W = \mathcal{A}(S \setminus U)$ exactly for deterministic $\mathcal{A}$, so the error would no longer measure recovery from the unlearning output alone.

As discussed in Section 2, for the discrete model space $\mathcal{W}$, take the zero-one metric $d_0(w, w') := \mathbb{I}\{w \neq w'\}$. If $\bar{\mathcal{A}}$ has (d_0, Γ)-utility on empty requests and $\eta := \Gamma(0)$, then using the identity reconstruction map $g(y) = y$ gives the following upper bound on $\beta_Y^\star$,

$$\beta_\emptyset^\star \leq \sup_{w \in \mathcal{W}} \mathbb{P}\left(Y \neq w \mid W = w\right) = \sup_{w \in \mathcal{W}} \mathbb{P}\left(d_0(Y, w) > 0 \mid W = w\right) \leq \Gamma(0) = \eta.$$

We use this to further upper bound the unlearning recovery error.

Lemma 4.2. *Assume that $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (α, ε)-Rényi unlearning for all deletion requests of size at most m and $\bar{\mathcal{A}}$ has (d_0, Γ)-utility on empty requests with $\eta := \Gamma(0)$. Define $\gamma := \frac{\alpha-1}{\alpha}$., then*

$$\beta_{\mathrm{un}}^\star \leq \bar{\beta}, \quad \bar{\beta} := \min\left\{1, e^{\gamma\varepsilon}\left(\eta + \Delta_{\mathcal{A}}\right)^\gamma\right\}, \quad \Delta_{\mathcal{A}} := \sup_{\substack{s,u:\, u \subseteq s \\ |u| \leq m}} \mathbb{P}\left(W_{s,u}^{(1)} \neq W_{s,u}^{(2)}\right),$$

where $W_{s,u}^{(1)}$ and $W_{s,u}^{(2)}$ are independent runs of $\mathcal{A}(s \setminus u)$.

Now, we state our main lower bound in the discrete setting.

Theorem 4.3. *Let $S \sim \mathcal{P}_S$, and $U_1, \ldots, U_K$ be a set of unlearning requests generated as above, each of size at most m. Assume $\mathcal{W}$ is finite and that $\mathcal{A}, \bar{\mathcal{A}}, \bar{\beta}$ satisfy the same assumptions as in Lemma 4.2. Then, for any permutation $\pi : [K] \to [K]$,*

$$I(M; S) \geq \sum_{i=1}^{K} H\left(W_{\pi(i)} \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right) - K\left(\log 2 + \bar{\beta} \log\left(|\mathcal{W}| - 1\right)\right).$$

Proof sketch. Apply Proposition 4.1. It remains to lower bound each of the terms $I\left(W_{\pi(i)}; \overline{W}_{\pi(i)} \mid \mathcal{H}_i\right)$, where $\mathcal{H}_i := \left(W_{\pi(<i)}, U_{\pi(\leq i)}\right)$. Let $g_{\mathrm{un}}^\star$ be an (almost-)optimal decoder in the definition of $\beta_{\mathrm{un}}^\star$. Use the definition of mutual information, the fact that conditioning reduces entropy, and the data processing inequality, to get, $I\left(W_{\pi(i)}, \overline{W}_{\pi(i)} \mid \mathcal{H}_i\right) \geq H\left(W_{\pi(i)} \mid \mathcal{H}_i\right) - H\left(W_{\pi(i)} \mid g_{\mathrm{un}}^\star\left(\overline{W}_{\pi(i)}, U_{\pi(i)}\right)\right)$. Use Lemma 4.2 to bound the average recovery error $\mathbb{P}\left(g_{\mathrm{un}}^\star\left(\overline{W}_{\pi(i)}, U_{\pi(i)}\right) \neq W_{\pi(i)}\right) \leq \bar{\beta}$. Apply Fano's inequality to the latter term to obtain, $H\left(W_{\pi(i)} \mid g_{\mathrm{un}}^\star\left(\overline{W}_{\pi(i)}, U_{\pi(i)}\right)\right) \leq \log 2 + \bar{\beta} \log\left(|\mathcal{W}| - 1\right)$. Plugging this back into the mutual information bound and summing over i proves the theorem. □

The bound is informative when the entropy terms are large and the penalty is small. Each entropy term measures the new information a retraining target reveals beyond the previously accounted requests and targets, in the spirit of an inclusion-exclusion principle; since the unlearner is not run sequentially, π is only an accounting device that decomposes this information across requests handled in parallel, and, as the bound holds for every ordering, we choose the one giving the strongest guarantee. Entropy terms grow with the size of the model space and the algorithmic randomness of $\mathcal{A}$, since a fixed input then yields a more uncertain output. The penalty term captures the loss from approximate unlearning: it grows when the unlearning requirement is weaker (e.g. a larger ε) or the empty-request map is noisier, allowing a looser approximation of the retraining target; it also scales with the model-space size and $\mathcal{A}$'s randomness, normalizing it against the entropy terms.

4.2 Instantiations of Theorem 4.3

In this subsection, m denotes the deletion budget: the unlearning algorithm satisfies (α, ε)-Rényi unlearning for every deletion set of size at most m. Let $\gamma := (\alpha - 1)/\alpha$, and assume the discrete empty-request utility condition $\mathbb{P}\left(\bar{\mathcal{A}}(w, \emptyset) \neq w \mid W = w\right) \leq \eta$ for each retraining target w. Put $\delta := e^{\gamma\varepsilon}\eta^\gamma$. For $q \geq 2$, $F(q) := \log(q/2) - \beta \log(q - 1)$, where $\beta := \min\{1, \delta\}$,.

For the deterministic learners below, the recovery error is at most β for every dataset, so this bound also holds conditional on $\mathcal{H}_i = (W_{<i}, U_{\leq i})$. If W_i has at most $q_i \geq 2$ possible values given this history, any estimate outside that set can be mapped into it without increasing the error. Applying Fano's inequality conditional on $\mathcal{H}_i$ therefore gives the penalty $\log 2 + \beta \log(q_i - 1)$, even when the full model space is infinite. If only one target is possible, the conditional entropy is zero.

The propositions below are necessary conditions for the existence of a forget-only unlearning algorithm satisfying the assumptions above. Thus, if the usual learner output contains less information than the displayed lower bound, the conclusion is that the usual output alone cannot support such forget-only unlearning; a successful implementation would have to store an additional state. All proofs are deferred to Appendix D.2.

Canonical threshold ERM learner. Let $\mathcal{Z} = \mathcal{X} \times \{0,1\}$, where $\mathcal{X} = [N]$, and consider $\mathcal{H} = \{h_a : h_a(x) = \mathbb{1}\{x \geq a\},\ a \in [N+1]\}$. The canonical ERM tie-breaking rule returns the leftmost point labeled 1, or $N+1$ if there is no positive example.

Proposition 4.4 (Canonical threshold ERM)**.** *Fix deletion budget $m \geq 1$. Let $n \geq 2m$, let $q \geq 2$, and set $N = nq$. Then there exists a distribution P_S over realizable threshold samples of size n, and there are deletion requests $U_1, \ldots, U_m$, each of size exactly m, such that any forget-only unlearning algorithm satisfying the assumptions above must obey*

$$I(M;S) \geq m\left[\log\left(\frac{N}{n}\right) - \log 2 - \beta \log\left(\frac{N}{n} - 1\right)\right].$$

For the bare canonical ERM output on thresholds, one has $I(M;S) = \log(N/n)$.

Coordinate PCA. Here the learner returns the coordinate direction with largest raw empirical second moment among the axis-aligned directions. Without unlearning, storing the selected coordinate direction requires only $O(\log d)$ bits.

Proposition 4.5 (Coordinate PCA)**.** *Consider top-coordinate PCA in dimension d. Let n be the dataset size and let m be the deletion budget. Assume $d = nq$ for some integer $q \geq 2$, and assume $2m \leq n$. Then there exists a distribution P_S over datasets of size n, and deletion requests $U_1, \ldots, U_m$, each of size exactly m, such that any forget-only unlearning algorithm satisfying the assumptions above must obey*

$$I(M;S) \geq m\left[\log\left(\frac{d}{n}\right) - \log 2 - \beta \log\left(\frac{d}{n} - 1\right)\right].$$

For the bare top-coordinate PCA output, $M = e_{a_1}$, one has $I(M_{\mathrm{bare}};S) = \log q$.

Sparse subset least-squares regression. Consider least squares with a one-per-block sparsity constraint. The d features are partitioned into disjoint blocks $G_1, \ldots, G_b$, and the learner minimizes $\sum_i (\langle w, x_i\rangle - y_i)^2$ subject to $|\mathrm{supp}\,(w) \cap G_j| \leq 1$ for every $j \in [b]$. Best-subset regression is classical (Garside, 1965; Hocking & Leslie, 1967; Furnival & Wilson, 1974), and the one-per-block constraint is the standard multiple-choice programming constraint (Healy, 1964). Without unlearning, specifying one selected coordinate per block costs $O(b \log(d/b))$ bits.

Proposition 4.6 (One-per-block sparse least-squares regression)**.** *Consider one-per-block sparse least-squares regression in dimension d. Let n be the dataset size and let $m \geq 1$ be the deletion budget. Assume $b := \frac{n}{m+1} \in \mathbb{N}$, where $d = b(q+1)$ for some integer $q \geq 2$. Then there exists a distribution P_S over datasets of size n, and deletion requests $U_1, \ldots, U_b$, each of size exactly m, such that any forget-only unlearning algorithm satisfying the assumptions above must obey*

$$I(M;S) \geq \frac{n}{m+1}\left[(1-\beta)\log\left(\frac{(m+1)d}{n} - 1\right) - \log(2)\right]$$

On P_S, the bare ERM output is deterministic and thus has $I\,(M;S) = 0$.

Row-wise factorized affine matrix completion. Fix r and view the first r^2 coordinates as an $r \times r$ matrix X with rows $x_i \in \mathbb{R}^r$. A data point (i, a, y), with $i \in [r]$, $a \in \mathbb{R}^r$, and $y \in \mathbb{R}$, imposes the row-wise affine constraint $\langle a, x_i\rangle = y$. Thus $(i, \mathbf{1_r}, \mathbf{r})$ imposes the row-sum constraint, and $(i, e_\ell, 1)$ imposes $(x_i)_\ell = 1$. The learner uses a row-wise factorization $x_i = u_i v_i$, with $u_i \in \mathbb{R}$ and $v_i \in \mathbb{R}^r$, and minimizes $\frac{1}{2}\sum_{(i,a,y)\in S}(\langle a, u_i v_i\rangle - y)^2 + \frac{\lambda}{2}\sum_{i=1}^r \left(u_i^2 + \|v_i\|_2^2\right)$. The algorithm $\mathcal{A}_{\mathrm{fac}}$ returns the fitted matrix in the limit $\lambda \to 0$. On the consistent datasets below, this limit is the interpolating matrix minimizing $\sum_{i=1}^r \|x_i\|_2$, because $\inf_{u_i v_i = x_i} \frac{1}{2}(u_i^2 + \|v_i\|_2^2) = \|x_i\|_2$.

This is best viewed as a row-wise affine matrix-completion problem with a factorized norm bias. The construction below shows that forget-only unlearning may force the trained object to store a hidden permutation even though the ordinary fitted matrix on the full sample is deterministic.

Proposition 4.7 (Row-wise factorized affine matrix completion)**.** *Let $d > 4$ be the ambient number of scalar matrix entries, let $m \geq 1$ be the deletion budget, and let $n > 3m$ be the dataset size. Define $r_\star \coloneqq \min\left\{\left\lfloor\sqrt{d}\right\rfloor, \left\lfloor\frac{n}{m+1}\right\rfloor\right\}$. Then there exists a distribution P_S over datasets of size n, and deletion requests $U_1, \ldots, U_{r_\star}$, each of size exactly m, such that any forget-only unlearning algorithm satisfying the assumptions above must obey*

$$I(M; S) \geq \log(r_\star!) - \sum_{s=2}^{r_\star} \left[\log 2 + \beta \log(s-1)\right].$$

On P_S, the basic fitted matrix is deterministic and thus $I(M; S) = 0$ when no unlearning is required.

5 Related Work and Discussion

Machine unlearning asks for a cheaper alternative to retraining whose output behaves like a model trained from scratch without the deleted data (Cao & Yang, 2015; Guo et al., 2020). We study the stricter forget-only setting: at deletion time, the unlearner receives only the trained model $M = A(S)$ and the forget set U, not the retained set $S \setminus U$, training logs, checkpoints, gradients, or auxiliary statistics. Several empirical methods are forget-only, or close to forget-only, in this operational sense. Some zero-shot unlearning methods update the trained model using only the forget examples (Chundawat et al., 2023; Foster et al., 2024). Related single-point analyses of SGD show that approximate unlearning guarantee depends on the stability of the training dynamics (Thudi et al., 2022). Recent work has also developed nearly forget-only unlearning methods for language models. However, these works use notions of unlearning that differ from ours; we discuss them, together with other empirical evaluations, in Appendix A. The empirical literature shows that forget-only and retain-free unlearning can be useful, but our theory explains why this does not imply universal feasibility. Existing successes typically rely on at least one of a few common forms of structure: the request is benign, the model is locally stable, retain information is approximated by a proxy, or the unlearning definition is weaker than retraining-style indistinguishability. This is consistent with our results: the lower bounds do not rule out forget-only unlearning, but they require the retraining target to be recoverable from the information available in (M, U).

A related line removes access to the original retain data but reconstructs a proxy. Source-free and retain-free methods synthesize adversarial, energy-guided, or generator-based proxy data to preserve utility while unlearning (Mu & Klabjan, 2025; Chen et al., 2025; Wang et al., 2025a; Lee et al., 2025). Other works obtain unlearning guarantees using a surrogate distribution or model-side curvature information (Basaran et al., 2025; Ahmed et al., 2025). Several theoretical works allow the unlearner to use training summaries, gradients, or other auxiliary local information in order to obtain certified unlearning algorithms (Ginart et al., 2019; Neel et al., 2021; Sekhari et al., 2021; Heinzler et al., 2026). These methods are close in motivation to ours, but the missing retain information is replaced by a proxy, a surrogate, a local approximation, or a state deliberately preserved during training. Our memorization lower bound intuitively captures the same intuition: the trained model must memorize enough recoverable information to reconstruct the retraining target.

The closest theoretical comparison is the work of Cherapanamjeri et al. (2025), who lower bound the memory complexity of learning–unlearning algorithms in terms of the *eluder dimension*. Their result is complementary to ours: it studies realizability testing, with extensions to ERM using exact minimizers as the retraining reference, and is governed by the hypothesis class, whereas our bounds are learner-dependent and unlearning request-dependent, and apply directly to strict forget-only unlearning without auxiliary information for any (α, ε). In particular, their results do not distinguish between learning algorithms with the same hypothesis class, and do not directly cover tasks such as regression, PCA, and factorization.

Our lower bounds quantify how much recoverable information must be memorized in the trained model for forget-only unlearning to succeed, but they do not identify the form that this information should take. In particular, they do not specify which features or statistics should be memorized, nor how such information should be stored. This leaves open a natural question: whether the memorization already observed in overparameterized models can be accessed in a way that supports certified forget-only unlearning. Finally, instantiating these lower bounds for contemporary architectures and large-scale learning pipelines remains an important direction for understanding when empirical failures of forget-only unlearning are algorithmic and when they are information-theoretic.

AI USE STATEMENT

In this work, we used generative AI tools for editing the text, such as correcting grammar and improving the flow of the document, for searching relevant literature, and also for elaborating proof ideas into full proofs. All proofs and text are written and verified by the authors. Additionally, we used generative AI tools for writing some code for simulations and visualising the bounds but we did not include them in the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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 *Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Umit Yigit Basaran, Sk Miraj Ahmed, Amit Roy-Chowdhury, and Basak Guler. A certified unlearning approach without access to source data. In *International Conference on Machine Learning (ICML)*, 2025.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Symposium on Theory of Computing (STOC)*, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (IEEE)*, 2015.

Huiqiang Chen, Tianqing Zhu, Xin Yu, and Wanlei Zhou. Zero-shot machine unlearning with proxy adversarial data generation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Yeshwanth Cherapanamjeri, Sumegha Garg, Nived Rajaraman, Ayush Sekhari, and Abhishek Shetty. The space complexity of learning-unlearning algorithms. In *Conference on Learning Theory (COLT)*, 2025.

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

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. Avoiding copyright infringement via large language model unlearning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *CoRR*, 2023.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *CoRR*, 2024.

Vitaly Feldman, Guy Kornowski, and Xin Lyu. Trade-offs in data memorization via strong data processing inequalities. In *Conference on Learning Theory (COLT)*, 2025.

Jack Foster, Kyle Fogarty, Stefan Schoepf, Cengiz Öztireli, and Alexandra Brintrup. Zero-shot machine unlearning at scale via lipschitz regularization. *CoRR*, 2024.

George M. Furnival and Robert W. Wilson. Regressions by leaps and bounds. *Technometrics*, 16 (4):499–511, 1974. doi: 10.1080/00401706.1974.10489231.

M. J. Garside. The best sub-set in multiple regression analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 14(2–3):196–200, 1965. doi: 10.2307/2985341.

GDPR. Regulation (eu) 2016/679 (general data protection regulation). Official Journal of the European Union, L 119, 2016. European Parliament and Council of the European Union. Article 17: Right to erasure ("right to be forgotten"). Available via EUR-Lex.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Sally A Goldman and Michael J Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *International Conference on Machine Learning (ICML)*, 2020.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 705–714, 2010.

W. C. Healy, Jr. Multiple choice programming (a procedure for linear programming with zero-one variables). *Operations Research*, 12(1):122–138, 1964. doi: 10.1287/opre.12.1.122.

Carolin Heinzler, Kasra Malihi, and Amartya Sanyal. Less noise, same certificate: Retain sensitivity for unlearning. *arXiv preprint arXiv:2603.03172*, 2026.

R. R. Hocking and R. N. Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967. doi: 10.1080/00401706.1967.10490502.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Yongwoo Kim, Sungmin Cha, and Donghyun Kim. Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. *Engineering Applications of Artificial Intelligence*, 2026.

SangYong Lee, Sangjun Chung, and Simon S. Woo. RUAGO: Effective and practical retain-free unlearning via adversarial attack and OOD generator. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for llms. In *Conference on Language Modeling*, 2024.

Ioannis Mavrothalassitis, Pol Puigdemont, Noam Itzhak Levi, and Volkan Cevher. Ascent fails to forget. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Siqiao Mu and Diego Klabjan. Rewind-to-delete: Certified machine unlearning for nonconvex functions. *Neural Information Processing Systems (NeurIPS)*, 2025.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Conference on Algorithmic Learning Theory (ALT)*, 2021.

Stefan Schoepf, Jack Foster, and Alexandra Brintrup. Potion: Towards poison unlearning. *Data-Centric Machine Learning Research (DMLR)*, 2024.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember What You Want to Forget: Algorithms for Machine Unlearning. *Neural Information Processing Systems (NeurIPS)*, 2021.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. Position: Llm unlearning benchmarks are weak measures of progress. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2025.

David Thiel. Identifying and eliminating csam in generative ml training data and models. *Stanford Internet Observatory, Cyber Policy Center, December*, 23:3, 2023.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding factors influencing machine unlearning. In *IEEE Symposium on Security and Privacy (IEEE)*, 2022.

Xiuyuan Wang, Chaochao Chen, Weiming Liu, Xinting Liao, Fan Wang, and Xiaolin Zheng. Efficient source-free unlearning via energy-guided data synthesis and discrimination-aware multitask optimization. In *International Conference on Machine Learning (ICML)*, 2025a.

Yaxuan Wang, Jiaheng Wei, Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data. In *International Conference on Learning Representations (ICLR)*, 2025b.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Neural Information Processing Systems (NeurIPS)*, 2024.

Jiatong Yu, Yinghui He, Anirudh Goyal, and Sanjeev Arora. On the impossibility of retrain equivalence in machine unlearning. *CoRR*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling*, 2024.

A Additional Related Works

Recent work on language-model unlearning has produced several forget-only, or nearly forget-only, procedures. Early work on knowledge unlearning uses gradient-ascent or unlikelihood objectives on target sequences to reduce memorization risk (Jang et al., 2023). Yao et al. (2024) view LLM unlearning as a post-hoc alignment operation driven by negative examples, while Eldan & Russinovich (2023) fine-tune a large model so that it stops recalling Harry Potter content while largely preserving standard benchmark performance. More recent methods refine this optimization perspective: Zhang et al. (2024) introduce negative preference optimization to avoid the catastrophic collapse often caused by plain gradient ascent, and Wang et al. (2025b) propose FLAT, a forget-data-only objective that uses neither retain data nor a reference model. These works optimize empirical forgetting criteria rather than the stronger indistinguishability requirement studied here. They therefore leave open whether weaker, task-level notions of forgetting can be achieved with substantially less memorization.

Another line of work shows that the difficulty of unlearning depends on the deletion distribution and on the evaluation criterion. Allouah et al. (2025) distinguish in-distribution from out-of-distribution deletion requests: simple certified procedures can achieve tight tradeoffs for in-distribution deletions, whereas out-of-distribution deletions can be harder than retraining in some regimes. For language models, benchmarks such as TOFU and MUSE show that forgetting synthetic biographies, books, or news articles is already difficult under empirical metrics (Maini et al., 2024; Shi et al., 2025). Thaker et al. (2025) argue further that the usual split into independent forget and retain queries can overstate progress, since realistic queries may couple forgotten and retained information. Representation-based evaluations lead to a similar conclusion: methods that appear successful under logit- or accuracy-based metrics may leave internal representations close to those of the original model, or may destroy the representation altogether (Kim et al., 2026). These empirical observations are consistent with our lower bound: unlearning becomes harder when the retraining target is more sensitive to retained data, and when the criterion used to compare against retraining is more stringent. A related distinction between deletion-specific and distribution-free sensitivity is made in the certified setting by Heinzler et al. (2026).

B Supporting Lemmas

Lemma B.1 (Event transfer)**.** *Let $\alpha > 1$, let $\gamma = \frac{\alpha-1}{\alpha}$, and suppose $D_\alpha(P\|Q) \le \varepsilon$. Then, for every measurable event E, $P(E) \le e^{\gamma\varepsilon} Q(E)^\gamma$.*

Proof of Lemma B.1. Let μ be a common dominating measure for P and Q, meaning that P and Q are both absolutely continuous with respect to μ: if $\mu(A) = 0$, then $P(A) = Q(A) = 0$. Such a common dominating measure always exists; for example, one may take $\mu = 1/2(P + Q)$. Define

$$p \coloneqq \frac{dP}{d\mu}, \qquad q \coloneqq \frac{dQ}{d\mu}.$$

The Rényi divergence can be written as,

$$D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log \int p(x)^\alpha q(x)^{1-\alpha} d\mu(x). \tag{1}$$

Define the Radon-Nikodym derivative of P w.r.t. Q as,

$$L(x) \coloneqq \frac{dP}{dQ}(x) = \frac{p(x)}{q(x)}. \tag{2}$$

Then we have,

$$\int L(x)^\alpha dQ(x) = \int \left(\frac{p(x)}{q(x)}\right)^\alpha q(x) d\mu(x) = \int p(x)^\alpha q(x)^{1-\alpha} d\mu(x), \tag{3}$$

which implies that,

$$\begin{aligned} D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log \int L(x)^\alpha dQ(x) = \frac{1}{\alpha - 1} \log \mathbb{E}_Q(L^\alpha) \le \varepsilon \\ \implies \mathbb{E}_Q(L^\alpha) \le e^{(\alpha-1)\varepsilon}. \end{aligned} \tag{4}$$

For any event A, we have,

$$P(A) = \int_A dP(x) = \int_A L(x)\, dQ(x) = \mathbb{E}_Q(L\, \mathbb{1}_A). \tag{5}$$

Using Hölder's inequality with $m = \alpha$ and $n = \frac{\alpha}{\alpha-1}$ ($\frac{1}{m} + \frac{1}{n} = 1$), we finally get,

$$\begin{aligned} P(A) &= \mathbb{E}_Q(L\, \mathbb{1}_A) \le (\mathbb{E}_Q L^m)^{1/m}\, (\mathbb{E}_Q \mathbb{1}_A^n)^{1/n} \\ &= (\mathbb{E}_Q L^\alpha)^{1/\alpha}\, (\mathbb{E}_Q \mathbb{1}_A)^{(\alpha-1)/\alpha} \le e^{(\alpha-1)/\alpha\, \varepsilon}\, Q(A)^{(\alpha-1)/\alpha}. \end{aligned} \tag{6}$$

This completes the proof. □

C Deferred Proofs from Section 3

Lemma C.1. *Suppose $\bar{\mathcal{A}}$ has (ϕ, Γ)-utility on empty requests. Consider two outputs of $\mathcal{A}$ W_1, W_2 with laws P_1, P_2. Fix two disjoint sets E_1, E_2 with $\mathrm{dist}_\phi(E_1, E_2) > 2r$ such that $\mathbb{P}_{W_i \sim P_i}(W_i \in E_i) \ge 1 - \kappa$ for $i \in \{1, 2\}$. Define the sets $B_i \coloneqq \{w \in \mathcal{W} : \mathrm{dist}_\phi(w, E_i) \le r\}$ and $Q_i \coloneqq \mathrm{Law}\left(\bar{\mathcal{A}}(W_i, \emptyset)\right)$ for $i \in \{1, 2\}$. Then, it holds that $B_1 \cap B_2 = \emptyset$, and,*

$$Q_i(B_i) \ge 1 - \kappa - \Gamma(r).$$

Proof. First, note that B_1 and B_2 are disjoint. Suppose $y \in B_1 \cap B_2$. Then for every $\varepsilon' > 0$ there are $w_i \in E_i$ with $\phi(y - w_i) \le r + \varepsilon'$. Hence $\phi(w_1 - w_2) \le 2r + 2\varepsilon'$, which contradicts $\mathrm{dist}_\phi(E_1, E_2) > 2r$ once ε' is small enough.

For $w \in E_i$, the ball $\{y : \phi(y - w) \le r\}$ is contained in B_i. By Definition 2, it holds that $Q_w(B_i) \ge 1 - \Gamma(r)$ for $i \in \{1, 2\}$. Integrating over $w \sim P_i$ and keeping only $w \in E_i$, we have,

$$\begin{aligned} Q_i(B_i) &\ge \int_{E_i} Q_w(B_i)\, P_i(dw) \ge (1 - \Gamma(r)) \int_{E_i} P_i(dw) \\ &= (1 - \Gamma(r))\, P_i(E_i) \ge (1 - \Gamma(r))\, (1 - \kappa) \ge 1 - \kappa - \Gamma(r). \end{aligned}$$

□

Proof of Theorem 3.1. For $i \in [N]$, write $M_i := \mathcal{A}(S_i)$, $Q_i := \mathrm{Law}(\bar{\mathcal{A}}(W_i, \emptyset))$, and $T_i := \mathrm{Law}(\bar{\mathcal{A}}(M_i, U))$, Define $B_i := \{w \in \mathcal{W} : \mathrm{dist}_\phi(w, E_i) \le r\}$, and $\gamma = \frac{\alpha-1}{\alpha}$.

Recall that $u(r) := \kappa + \Gamma(r)$, and fix $r \in R$, so $0 < r < \Delta/2$ and $0 < u(r) < 1$. For each pair $i \neq j$, the sets E_i, E_j from the s.d.e. satisfy $\mathrm{dist}_\phi(E_i, E_j) \ge \Delta > 2r$ and $\mathbb{P}_{W_k \sim P_k}(W_k \in E_k) \ge 1 - \kappa$ for $k \in \{i, j\}$, where $P_k := \mathrm{Law}(\mathcal{A}(S_k \setminus U))$. Applying Lemma C.1 to the laws P_i, P_j and the sets E_i, E_j, we get that $B_i \cap B_j = \emptyset$ and $Q_i(B_i) \ge 1 - u(r)$, implying $Q_i(B_i^c) \le u(r)$. Since this holds for every pair $i \neq j$, the sets $B_1, \ldots, B_N$ are pairwise disjoint.

The unlearning assumption gives $d_\alpha(Q_i, T_i) \le \varepsilon$ for all $i \in [N]$. As T_i is the image of M_i under the map $w \to \bar{\mathcal{A}}(w, U)$, by the data processing inequality, we have that, for $i \neq j$,

$$d_\alpha(T_i, T_j) \le d_\alpha(M_i, M_j) \le \delta_\alpha.$$

Specifically, we will use $d_\alpha(T_1, T_i) \le \delta_\alpha$.

As the sets $B_1, \ldots, B_N$ are disjoint, it holds that, $\sum_{i=1}^N T_1(B_i) \le 1$.

We first analyze the exact collision case, which also covers the deterministic learners. The zero-collision condition forces $T_i = T_1$, which implies $d_\alpha(Q_i, T_1) \le \varepsilon$ for all $i \in [N]$. Using Lemma B.1 in two directions, together with the unlearning assumption, we have,

$$\begin{aligned} 1 - u(r) \le Q_i(B_i) \le e^{\gamma\varepsilon} T_1(B_i)^\gamma &\implies T_1(B_i) \ge e^{-\varepsilon} \left(1 - u(r)\right)^{1/\gamma}, \\ 1 - T_1(B_i) = T_1(B_i^c) \le e^{\gamma\varepsilon} Q_i(B_i^c)^\gamma \le e^{\gamma\varepsilon} u(r)^\gamma &\implies T_1(B_i) \ge 1 - e^{\gamma\varepsilon} u(r)^\gamma \end{aligned}$$

Summing each lower bound over i and using $\sum_{i=1}^N T_1(B_i) \le 1$, we get the final two terms depending on r.

In the approximate collision case, we use Lemma B.1 twice to get,

$$\begin{aligned} &1 - u(r) \le Q_i(B_i) \le e^{\gamma\varepsilon} T_i(B_i)^\gamma \le e^{\gamma\varepsilon + \gamma^2\delta_\alpha} T_1(B_i)^{\gamma^2} \implies T_1(B_i) \ge \left(1 - u(r)\right)^{1/\gamma^2} e^{-\varepsilon/\gamma - \delta_\alpha}, \\ &T_1(B_i^c) \le e^{\gamma\delta_\alpha} T_i(B_i^c)^\gamma \le e^{\gamma^2\varepsilon + \gamma\delta_\alpha} Q_i(B_i^c)^{\gamma^2} \le e^{\gamma^2\varepsilon + \gamma\delta_\alpha} u(r)^{\gamma^2} \implies T_1(B_i) \ge 1 - e^{\gamma^2\varepsilon + \gamma\delta_\alpha} u(r)^{\gamma^2}, \end{aligned}$$

Here we used that $x \mapsto x^\gamma$ is increasing. Summing the first lower bound over the disjoint sets B_i gives

$$1 \ge \sum_{i=1}^N T_1(B_i) \ge N \left(1 - u(r)\right)^{1/\gamma^2} e^{-\varepsilon/\gamma - \delta_\alpha}.$$

Taking logarithms and rearranging,

$$\begin{aligned} \frac{\varepsilon}{\gamma} &\ge \log N - \delta_\alpha + \frac{1}{\gamma^2} \log(1 - u(r)), \\ \varepsilon &\ge \gamma \left(\log N - \delta_\alpha\right) + \frac{1}{\gamma} \log(1 - u(r)). \end{aligned}$$

Similarly, summing the second lower bound gives

$$1 \ge \sum_{i=1}^N T_1(B_i) \ge N \left(1 - e^{\gamma^2\varepsilon + \gamma\delta_\alpha} u(r)^{\gamma^2}\right),$$

hence

$$e^{\gamma^2\varepsilon + \gamma\delta_\alpha} u(r)^{\gamma^2} \ge \frac{N-1}{N}.$$

Taking logarithms and dividing by γ^2 yields

$$\varepsilon \ge \frac{1}{\gamma^2} \log \frac{N-1}{N} - \frac{\delta_\alpha}{\gamma} - \log u(r).$$

In both cases, taking the supremum over $\{r \in (0, \Delta/2) : 0 < u(r) < 1\}$ gives the final bounds.

□

C.1 TEACHING DIMENSION

Proposition C.2. *Let $\mathcal{H}$ be a hypothesis class with $|\mathcal{H}| \geq N > 1$, $\mathcal{A}$ be any empirical risk minimizer for zero-one loss, $d_{\mathcal{H}}$ be any metric on $\mathcal{H}$, and $h_1, \ldots, h_N \in \mathcal{H}$ be distinct. Let the teaching dimension of $\mathcal{H}$ be finite $\tau < \infty$. Then, there exist $(S_1, \ldots, S_N, U)$ with $|U| = O(\tau^2)$ and $|S_i| = o(\tau^3)$ for $i \in [N]$ such that,*

$$\mathcal{A}(S_i) = \mathcal{A}(S_j), \qquad \mathcal{A}(S_i \setminus U) = h_i,$$

for $i, j \in [N], i \neq j$. Hence, there exists an N-sized shared-deletion example with $m = O(\tau^2)$, $\delta_\alpha = \kappa = 0$, and $\Delta = \min_{i \neq j} d_{\mathcal{H}}(h_i, h_j)$.

Proof. Let $h_0 \in \mathcal{H}$ have a teaching set T_0. A teaching set size is upper-bounded by the teaching dimension, hence $|T_0| \leq \tau$. Similarly, let R_i be teaching sets for h_i, with $|R_i| \leq \tau$ as well. Construct the common unlearning request as $\tau + 1$ copies of T_0: $U = \uplus_{i=1}^{\tau+1} T_0$, so $|U| \leq \tau(\tau + 1)$. Set $S_i = R_i \cup U$, so $|S_i| \leq \tau(\tau + 2)$.

Consider first the hypothesis h_0. On S_i, it errs only on the retain part R_i. U is consisting of copies of T_0, which is perfectly fitted by h_0 by construction. As $|R_i| \leq \tau$, h_0 can make at most τ mistakes on it. On the other hand, $h_i \neq h_0$ perfectly fits the retain part R_i. However, on each of the $\tau + 1$ copies of T_0 in U, it must make at least one mistake, as h_0 is the only hypothesis that perfectly fits U. Therefore, h_i makes at least $\tau + 1$ mistakes on S_i. Combining these two results, we obtain that $\mathcal{A}(S_i) = h_0$ for any $i \in [N]$. The retrained targets are $\mathcal{A}(S_i \setminus U) = \mathcal{A}(R_i) = h_i$, again by the teaching set property.

Therefore, the above implies that N different datasets exactly collide under any ERM $\mathcal{A}$, but are strictly separated under retraining.

□

C.2 PROOFS OF THE EXAMPLES

Proof of Example 2. For a realizable sample S, write $n(S) \coloneqq \max\{x : (x, -1) \in S\}$ and $p(S) \coloneqq \min\{x : (x, +1) \in S\}$. Thus $\mathcal{A}_c^{\text{th}}(S) = (n(S) + p(S))/2$.

We first prove the lower bound. Let $U = \cup_{i=1}^{m} \{(-c + 1, +1)\}$ be the deletion multiset. Define the following multisets:

$$S_1 = U \cup \{(-c, -1), (-c + 1, +1)\} \qquad S_2 = U \cup \{(-c, -1), (c, +1)\}.$$

Pad each dataset with $n - m - 2$ additional copies of $(-c, -1)$. Both full samples have largest negative example $-c$ and smallest positive example $-c + 1$, so $\mathcal{A}_c^{\text{th}}(S_1) = \mathcal{A}_c^{\text{th}}(S_2) = -c + \frac{1}{2}$. After deleting U, however, $\mathcal{A}_c^{\text{th}}(S_1 \setminus U) = -c + \frac{1}{2}$ and $\mathcal{A}_c^{\text{th}}(S_2 \setminus U) = 0$. Hence, (S_1, S_2, U) is a $(2, m, c - \frac{1}{2}, |\cdot|; 0, 0)$-shared-deletion example.

It remains to show that no larger separation is possible. Consider any shared-deletion example (S_1, S_2, U) with $|U| \leq m$. Put $a_i = n(S_i)$, $b_i = p(S_i)$, $a_i' = n(S_i \setminus U)$, and $b_i' = p(S_i \setminus U)$ for $i \in \{1, 2\}$. Since $\mathcal{A}_c^{\text{th}}(S_1) = \mathcal{A}_c^{\text{th}}(S_2)$, we have $a_1 + b_1 = a_2 + b_2$. Assume without loss of generality that $\mathcal{A}_c^{\text{th}}(S_2 \setminus U) \geq \mathcal{A}_c^{\text{th}}(S_1 \setminus U)$. Then

$$\mathcal{A}_c^{\text{th}}(S_2 \setminus U) - \mathcal{A}_c^{\text{th}}(S_1 \setminus U) \leq \frac{c - b_2}{2} + \frac{a_1 + c}{2}.$$

If $b_2' = b_2$, then the first term is unnecessary and the bound is at most $(a_1 + c)/2 \leq c - \frac{1}{2}$, since $a_1 \leq c - 1$. Similarly, if $a_1' = a_1$, then the bound is at most $(c - b_2)/2 \leq c - \frac{1}{2}$, since $b_2 \geq -c + 1$.

It remains to consider the case $b_2' > b_2$ and $a_1' < a_1$. Then U must contain $(b_2, +1)$, so $(b_2, +1) \in S_1$ and hence $b_2 \geq b_1$. Also U must contain $(a_1, -1)$, so $(a_1, -1) \in S_2$ and hence $a_1 \leq a_2$. Together with $a_1 + b_1 = a_2 + b_2$, these inequalities imply $a_1 = a_2$ and $b_1 = b_2$. Thus, writing $a = a_1 = a_2$ and $b = b_1 = b_2$,

$$\mathcal{A}_c^{\text{th}}(S_2 \setminus U) - \mathcal{A}_c^{\text{th}}(S_1 \setminus U) \leq \frac{a + c}{2} + \frac{c - b}{2} = c - \frac{b - a}{2} \leq c - \frac{1}{2},$$

because $a, b \in \mathcal{X}_c$ and $a < b$. Thus the largest admissible separation in a $(2, m, \Delta, |\cdot|\,; 0, 0)$-s.d.e. is $\Delta = c - \frac{1}{2}$.

□

Proof of Example 3. For each $i \in \{1, 2\}$, write the multiset S_i in sorted order as $x_1^{(i)} \le \cdots \le x_n^{(i)}$. We first consider the even case $n = 2r$, with $r \in \mathbb{N}$. Since (S_1, S_2, U) is a shared-deletion example, the original medians coincide; write $M = \mathcal{A}_c^{\mathrm{med}}(S_i) = (x_r^{(i)} + x_{r+1}^{(i)})/2$, so $x_r^{(i)} \le M \le x_{r+1}^{(i)}$, for $i \in \{1, 2\}$.

1. Case $m = 1$. Write $U = \{u\}$. Suppose first that $u < M$. Then deletion removes a point from the left half of each dataset, so $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) = x_{r+1}^{(i)}$. Since $x_{r+1}^{(i)} = 2M - x_r^{(i)}$, $x_r^{(i)} \ge -c$, and $x_{r+1}^{(i)} \le c$, we have $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) \in [M, \min\{c, 2M + c\}]$, an interval of length at most c. The case $u > M$ is analogous and gives $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) \in [\max\{-c, 2M - c\}, M]$, again an interval of length at most c. Finally, if $u = M$, then necessarily $x_r^{(i)} = x_{r+1}^{(i)} = M$, so $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) = M$. Hence, in all cases, $\left|\mathcal{A}_c^{\mathrm{med}}(S_1 \setminus U) - \mathcal{A}_c^{\mathrm{med}}(S_2 \setminus U)\right| \le c$.

   This bound is tight. Take $S_1 = \{-c, 0, \ldots, 0\}$, with $2r - 1$ copies of 0, and $S_2 = \{-c, \ldots, -c, c, \ldots, c\}$, with r copies of $-c$ and r copies of c. Both medians are 0. Removing the common multiset $U = \{-c\}$ gives post-deletion medians 0 and c, so the separation is c. Therefore the largest admissible separation in a $(2, 1, \Delta, |\cdot|\,; 0, 0)$-s.d.e. is $\Delta = c$.

2. Case $m = 2$. Write $U = \{u, v\}$, with $u \le v$. Consider first $u \le v < M$. Then both deleted points lie below the common median, so $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) = (x_{r+1}^{(i)} + x_{r+2}^{(i)})/2 = M + (x_{r+2}^{(i)} - x_r^{(i)})/2$. Since $-c \le x_r^{(i)} \le x_{r+2}^{(i)} \le c$, we get $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) \in [M, \min\{c, M + c\}]$, an interval of length at most c. The case $M < u \le v$ is analogous.

   It remains to consider $u \le M \le v$. Write the remaining points as $y_1^{(i)} \le \cdots \le y_{2r-2}^{(i)}$, so that $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) = (y_{r-1}^{(i)} + y_r^{(i)})/2$. If $u < M < v$, one point is deleted from each side of the median, so $y_{r-1}^{(i)} \le M \le y_r^{(i)}$. If a deleted value equals M, then M is a sample value, hence $x_r^{(i)} = x_{r+1}^{(i)} = M$, and the same bracketing still holds. Therefore $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) \in [(M - c)/2, (M + c)/2]$, again an interval of length c. Thus, in all cases, $\left|\mathcal{A}_c^{\mathrm{med}}(S_1 \setminus U) - \mathcal{A}_c^{\mathrm{med}}(S_2 \setminus U)\right| \le c$.

   The bound is tight. For $r \ge 2$, take $S_1 = \{-c, \ldots, -c, 0, \ldots, 0\}$, with $r - 1$ copies of $-c$ and $r + 1$ copies of 0, and $S_2 = \{0, \ldots, 0, c, \ldots, c\}$, with $r + 1$ copies of 0 and $r - 1$ copies of c. Both medians are 0. Deleting $U = \{0, 0\}$ gives post-deletion medians $-c/2$ and $c/2$, so the separation is c. Hence the largest admissible separation in a $(2, 2, \Delta, |\cdot|\,; 0, 0)$-s.d.e. is $\Delta = c$.

3. Case $3 \le m \le n - 1$. Consider again the datasets from the previous case, and take $U = \{0, 0, 0\}$. Then $\mathcal{A}_c^{\mathrm{med}}(S_1 \setminus U) = -c$ and $\mathcal{A}_c^{\mathrm{med}}(S_2 \setminus U) = c$, so the separation is $2c$. This achieves the trivial upper bound $2c$. Since an s.d.e. allows $|U| \le m$, this construction is admissible for every $m \ge 3$. Hence the largest admissible separation in a $(2, m, \Delta, |\cdot|\,; 0, 0)$-s.d.e. is $\Delta = 2c$ for all $3 \le m \le n - 1$.

Now consider the odd case $n = 2r - 1$, with $r \in \mathbb{N}$. For each $i \in \{1, 2\}$, the median is $M = \mathcal{A}_c^{\mathrm{med}}(S_i) = x_r^{(i)}$.

1. Case $m = 1$. Write $U = \{u\}$. If $u < M$, then $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) = (x_r^{(i)} + x_{r+1}^{(i)})/2 = (M + x_{r+1}^{(i)})/2$, and hence $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) \in [M, (M + c)/2]$. Similarly, if $u > M$, then $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) \in [(M - c)/2, M]$. Finally, if $u = M$, then $\mathcal{A}_c^{\mathrm{med}}(S_i \setminus U) = (x_{r-1}^{(i)} + x_{r+1}^{(i)})/2 \in [(M - c)/2, (M + c)/2]$. In all cases, the two post-deletion medians lie in a common interval of length at most c, so their separation is at most c.

This bound is tight. For $r \geq 2$, take $S_1 = \{-c, \ldots, -c, c, \ldots, c\}$, with $r-1$ copies of $-c$ and r copies of c, and $S_2 = \{c, \ldots, c\}$, with $2r-1$ copies of c. Both medians are c. Removing the common multiset $U = \{c\}$ gives post-deletion medians 0 and c, so the separation is c. Therefore the largest admissible separation in a $(2, 1, \Delta, |\cdot|\,; 0, 0)$-s.d.e. is $\Delta = c$.

2. Case $2 \leq m \leq n-1$. Take $S_1 = \{-c, \ldots, -c, 0, \ldots, 0\}$, with $r-1$ copies of $-c$ and r copies of 0, and $S_2 = \{0, \ldots, 0, c, \ldots, c\}$, with r copies of 0 and $r-1$ copies of c. Both medians are 0. Deleting $U = \{0, 0\}$ gives post-deletion medians $-c$ and c, so the separation is $2c$, which is the trivial upper bound. Since an s.d.e. allows $|U| \leq m$, this construction is admissible for every $2 \leq m \leq n-1$. Hence the largest admissible separation in a $(2, m, \Delta, |\cdot|\,; 0, 0)$-s.d.e. is $\Delta = 2c$ for all such m.

□

Proof of Example 4. We write the construction first for odd n. Let $q = (n-1)/2$, and define

$$U_\rho = \{(\rho e_1, +1)\},$$
$$S_{1,\rho} = U_\rho \biguplus_{j=1}^{q} \{(-\rho(e_1 + e_2), -1), (\rho(e_1 + e_2), +1)\},$$
$$S_{2,\rho} = U_\rho \biguplus_{j=1}^{q} \{(-\rho(e_1 - e_2), -1), (\rho(e_1 - e_2), +1)\},$$

where $\uplus$ denotes multiset union. Thus each dataset contains one copy of $(\rho e_1, +1)$, together with q copies of each of the two remaining labeled points, and therefore has size $1 + 2q = n$. For even n, the same construction is obtained by taking $q = (n-2)/2$ and adding one extra duplicate of any nondeleted point; duplicates do not change the hard-margin SVM constraints. All points in $S_{1,\rho} \cup S_{2,\rho}$ lie in the unit ball because $\|\rho e_1\|_2 = \rho \leq 1$ and $\|\rho(e_1 \pm e_2)\|_2 = \rho\sqrt{2} \leq 1$.

We first compute the SVM on the full datasets. For $S_{1,\rho}$, the margin constraints are

$$\langle w, \rho e_1 \rangle \geq 1, \qquad \langle w, \rho(e_1 + e_2) \rangle \geq 1,$$

or equivalently $w_1 \geq 1/\rho$ and $w_1 + w_2 \geq 1/\rho$. The first inequality implies $\|w\|_2 \geq |w_1| \geq 1/\rho$, while $\rho^{-1} e_1$ satisfies both inequalities with norm $1/\rho$. Thus $\mathcal{A}_d^{\mathrm{svm}}(S_{1,\rho}) = \rho^{-1} e_1$. The same argument for $S_{2,\rho}$, whose nonredundant constraints are $w_1 \geq 1/\rho$ and $w_1 - w_2 \geq 1/\rho$, gives $\mathcal{A}_d^{\mathrm{svm}}(S_{2,\rho}) = \rho^{-1} e_1$.

After deleting U_ρ, the first retained sample consists of the two labeled points $(-\rho(e_1 + e_2), -1)$ and $(\rho(e_1 + e_2), +1)$. Its hard-margin SVM is the minimum-norm vector w such that $\langle w, \rho(e_1 + e_2) \rangle \geq 1$. By Cauchy–Schwarz, the unique minimizer is parallel to $e_1 + e_2$ and satisfies the constraint with equality, hence

$$\mathcal{A}_d^{\mathrm{svm}}(S_{1,\rho} \setminus U_\rho) = \frac{1}{2\rho}(e_1 + e_2).$$

Similarly,

$$\mathcal{A}_d^{\mathrm{svm}}(S_{2,\rho} \setminus U_\rho) = \frac{1}{2\rho}(e_1 - e_2).$$

Therefore

$$\|\mathcal{A}_d^{\mathrm{svm}}(S_{1,\rho} \setminus U_\rho) - \mathcal{A}_d^{\mathrm{svm}}(S_{2,\rho} \setminus U_\rho)\|_2 = \frac{1}{\rho}.$$

Since $\rho > 0$ can be taken arbitrarily small, for every $\Delta > 0$ the construction gives a $(2, 1, \Delta, \|\cdot\|_2\,; 0, 0)$-s.d.e. by choosing ρ sufficiently small. Since $|U_\rho| = 1 \leq m$, the same construction gives a $(2, m, \Delta, \|\cdot\|_2\,; 0, 0)$-s.d.e. for every $1 \leq m \leq n-1$. □

D DEFERRED PROOFS FROM SECTION 4

Proof of Proposition 4.1. Since $M \perp\!\!\!\perp (W^K, U^K) \mid S$, using the data processing inequality in (a), we have that,

$$\begin{aligned} I(M;S) &\overset{(a)}{\geq} I(M;W^K,U^K) = I\left(M;(W_{\pi(1)},U_{\pi(1)}),\dots(W_{\pi(K)},U_{\pi(K)})\right) \\ &\overset{(b)}{=} \sum_{i=1}^{K} I\left(M;W_{\pi(i)},U_{\pi(i)}|W_{\pi(<i)},U_{\pi(<i)}\right) \\ &\overset{(c)}{=} \sum_{i=1}^{K} \left[I\left(M;U_{\pi(i)}|W_{\pi(<i)},U_{\pi(<i)}\right) + I\left(M;W_{\pi(i)}|W_{\pi(<i)},U_{\pi(<i)},U_{\pi(i)}\right)\right] \\ &\overset{(d)}{\geq} \sum_{i=1}^{K} I\left(M;W_{\pi(i)}|W_{\pi(<i)},U_{\pi(<i)},U_{\pi(i)}\right) \\ &\overset{(e)}{\geq} \sum_{i=1}^{K} I\left(\overline{W}_{\pi(i)};W_{\pi(i)}|W_{\pi(<i)},U_{\pi(<i)},U_{\pi(i)}\right), \end{aligned} \tag{7}$$

where (b) and (c) follow from the chain rule of mutual information, while (d) follows from its non-negativity; (e) uses the data processing inequality again: since $\overline{W}_{\pi(i)} = \bar{\mathcal{A}}(M, U_{\pi(i)})$ is generated from $(M, U_{\pi(i)})$, it holds that $\overline{W}_{\pi(i)} \perp\!\!\!\perp W_{\pi(i)} \mid (M, U_{\pi(i)}, W_{\pi(<i)}, U_{\pi(<i)})$. □

D.1 DISCRETE MODEL SPACE

Proof of Lemma 4.2. Let $h^\star$ be an optimal decoder for W given $W^\emptyset$ (if the infimum is not attained, one can use an δ-optimal decoder and let $\delta \downarrow 0$), i.e.,

$$\beta_\emptyset^\star = \sup_{w\in\mathcal{W}} \mathbb{P}\left(h^\star(W^\emptyset) \neq w \mid W = w\right).$$

Define an admissible decoder for W from $(\overline{W}, U)$ by

$$g_\emptyset(\overline{w}, u) := h^\star(\overline{w}),$$

which ignores the u-coordinate. For fixed $w \in \mathcal{W}$ and $u \in \mathcal{X}^\star$, define

$$A_{g,w,u} := \{x \in \mathcal{W} : g_\emptyset(x,u) \neq w\} = \{x \in \mathcal{W} : h^\star(x) \neq w\}.$$

By the fact that $(\mathcal{A}, \bar{\mathcal{A}})$ are (α,ε)-RU, and Lemma B.1, for every fixed s, u, w,

$$\mathbb{P}\left(g_\emptyset(\overline{W},u) \neq w \mid S = s, U = u\right) \leq e^{\gamma\varepsilon}\mathbb{P}\left(h^\star(W^\emptyset) \neq w \mid S = s, U = u\right)^\gamma. \tag{8}$$

Denote

$$p(w \mid s,u) := \mathbb{P}\left(W = w \mid S = s, U = u\right).$$

Multiplying equation 8 by $p(w \mid s,u)$, summing over w, and using Jensen's inequality since x^γ is concave for $0 < \gamma < 1$, gives

$$\begin{aligned} &\sum_{w\in\mathcal{W}} p(w \mid s,u)\mathbb{P}\left(g_\emptyset(\overline{W},u) \neq w \mid S = s, U = u\right) \\ &\leq e^{\gamma\varepsilon} \sum_{w\in\mathcal{W}} p(w \mid s,u)\mathbb{P}\left(h^\star(W^\emptyset) \neq w \mid S = s, U = u\right)^\gamma \\ &\leq e^{\gamma\varepsilon} \left(\sum_{w\in\mathcal{W}} p(w \mid s,u)\mathbb{P}\left(h^\star(W^\emptyset) \neq w \mid S = s, U = u\right)\right)^\gamma. \end{aligned} \tag{9}$$

We now bound the term inside the parentheses. Expanding over the conditional law of $W \mid S = s, U = u$,

$$\begin{aligned} &\sum_{w\in\mathcal{W}} p(w \mid s,u)\mathbb{P}\left(h^\star(W^\emptyset) \neq w \mid S = s, U = u\right) \\ &= \sum_{w,w'\in\mathcal{W}} p(w \mid s,u)p(w' \mid s,u)\mathbb{P}\left(h^\star(W^\emptyset) \neq w \mid W = w', S = s, U = u\right). \end{aligned} \tag{10}$$

Let

$$W_1, W_2 \mid S = s, U = u \overset{\text{i.i.d.}}{\sim} P_{W|S=s,U=u},$$

and let $W_2^{\emptyset}$ be the empty-set unlearning output generated from W_2, namely

$$W_2^{\emptyset} \mid W_2 = w', S = s, U = u \sim P_{W^{\emptyset}|W=w',S=s,U=u}.$$

Then the double sum in equation 10 equals

$$\mathbb{P}\left(h^\star(W_2^{\emptyset}) \neq W_1 \mid S = s, U = u\right).$$

Moreover,

$$\{h^\star(W_2^{\emptyset}) \neq W_1\} \subseteq \{h^\star(W_2^{\emptyset}) \neq W_2\} \cup \{W_1 \neq W_2\},$$

and hence

$$\mathbb{P}\left(h^\star(W_2^{\emptyset}) \neq W_1 \mid S = s, U = u\right) \leq \mathbb{P}\left(h^\star(W_2^{\emptyset}) \neq W_2 \mid S = s, U = u\right) + \mathbb{P}\left(W_1 \neq W_2 \mid S = s, U = u\right). \tag{11}$$

It remains to bound the matched term. Since $W^{\emptyset} \perp\!\!\!\perp (S, U) \mid W$,

$$\mathcal{L}(W_2^{\emptyset} \mid W_2 = w, S = s, U = u) = \mathcal{L}(W^{\emptyset} \mid W = w).$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(h^\star(W_2^{\emptyset}) \neq W_2 \mid S = s, U = u\right) &= \sum_{w \in \mathcal{W}} p(w \mid s, u) \mathbb{P}\left(h^\star(W^{\emptyset}) \neq w \mid W = w\right) \\ &\leq \sup_{w \in \mathcal{W}} \mathbb{P}\left(h^\star(W^{\emptyset}) \neq w \mid W = w\right) = \beta_{\emptyset}^\star.\end{aligned} \tag{12}$$

Combining equation 9, equation 10, equation 11, and equation 12, we obtain, for every fixed s, u,

$$\sum_{w \in \mathcal{W}} p(w \mid s, u) \mathbb{P}\left(g_{\emptyset}(\overline{W}, u) \neq w \mid S = s, U = u\right) \leq e^{\gamma\varepsilon}\left(\beta_{\emptyset}^\star + \mathbb{P}\left(W_1 \neq W_2 \mid S = s, U = u\right)\right)^{\gamma}.$$

Since $\overline{W} \perp\!\!\!\perp W \mid S, U$, the left-hand side is exactly

$$\mathbb{P}\left(g_{\emptyset}(\overline{W}, u) \neq W \mid S = s, U = u\right).$$

Thus,

$$\mathbb{P}\left(g_{\emptyset}(\overline{W}, u) \neq W \mid S = s, U = u\right) \leq e^{\gamma\varepsilon}\left(\beta_Y^\star + \mathbb{P}\left(W_1 \neq W_2 \mid S = s, U = u\right)\right)^{\gamma}.$$

Finally, since $g_{\emptyset}$ is an admissible decoder in the definition of $\beta_{\text{un}}^\star$,

$$\beta_{\text{un}}^\star \leq \sup_{s,u} \mathbb{P}\left(g_{\emptyset}(\overline{W}, u) \neq W \mid S = s, U = u\right).$$

Therefore,

$$\beta_{\text{un}}^\star \leq e^{\gamma\varepsilon}\left(\beta_{\emptyset}^\star + \Delta\right)^{\gamma},$$

where

$$\Delta := \sup_{s,u} \mathbb{P}\left(W_1 \neq W_2 \mid S = s, U = u\right), \qquad W_1, W_2 \mid S = s, U = u \overset{\text{i.i.d.}}{\sim} P_{W|S=s,U=u}.$$

The proof is completed by noting that $\beta_{\emptyset}^\star \leq \eta$ and $\beta_{\text{un}}^\star \leq 1$. □

Proof of Theorem 4.3. To prove the memorization lower bound, we start from the result of Proposition 4.1 and will bound the term,

$$I\left(W_{\pi(i)}; \overline{W}_{\pi(i)} \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right).$$

Assuming the model space $\mathcal{W}$ is discrete and finite, we have that,

$$\begin{aligned}I\left(W_{\pi(i)}; \overline{W}_{\pi(i)} \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right) &= H\left(W_{\pi(i)} | W_{\pi(<i)}, U_{\pi(\leq i)}\right) - H\left(W_{\pi(i)} | \overline{W}_{\pi(i)}, W_{\pi(<i)}, U_{\pi(\leq i)}\right) \\ &\overset{(a)}{\geq} H\left(W_{\pi(i)} | W_{\pi(<i)}, U_{\pi(\leq i)}\right) - H\left(W_{\pi(i)} | g_{\text{un}}^\star(\overline{W}_{\pi(i)}, U_{\pi(i)})\right),\end{aligned} \tag{13}$$

where $g^\star_{\mathrm{un}}$ is the optimal decoder of W from $(\overline{W}, U)$ (or almost-optimal if the infimum is not attained), i.e., $\beta^\star_{\mathrm{un}} = \sup_{s,u} \mathbb{P}\left(g^\star_{\mathrm{un}}(\overline{W}, u) \neq W \mid S = s, U = u\right)$, and (a) follows from the following chain of inequalities,

$$H\left(W_{\pi(i)}|\overline{W}_{\pi(i)}, W_{\pi(<i)}, U_{\pi(\leq i)}\right) \leq H\left(W_{\pi(i)}|\overline{W}_{\pi(i)}, U_{\pi(i)}\right) \leq H\left(W_{\pi(i)}|\, g^\star_{\mathrm{un}}(\overline{W}_{\pi(i)}, U_{\pi(i)})\right).$$

The first inequality uses that conditioning reduces entropy, while the second uses data processing. By Fano's inequality applied to the pair

$$\left(W_{\pi(i)},\, g^\star_{\mathrm{un}}(\overline{W}_{\pi(i)}, U_{\pi(i)})\right),$$

we get

$$H\left(W_{\pi(i)} \mid g^\star_{\mathrm{un}}(\overline{W}_{\pi(i)}, U_{\pi(i)})\right) \leq h(P_{e,\pi(i)}) + P_{e,\pi(i)} \log(|\mathcal{W}| - 1) \leq \log 2 + P_{e,\pi(i)} \log(|\mathcal{W}| - 1),$$

where

$$P_{e,\pi(i)} := \mathbb{P}\left(g^\star_{\mathrm{un}}(\overline{W}_{\pi(i)}, U_{\pi(i)}) \neq W_{\pi(i)}\right).$$

Moreover,

$$P_{e,\pi(i)} = \mathbb{E}_{S,U_{\pi(i)}}\left[\mathbb{P}\left(g^\star_{\mathrm{un}}(\overline{W}_{\pi(i)}, U_{\pi(i)}) \neq W_{\pi(i)} \mid S, U_{\pi(i)}\right)\right] \leq \mathbb{E}_{S,U_{\pi(i)}}[\beta^\star_{\mathrm{un}}] = \beta^\star_{\mathrm{un}} \leq \bar{\beta}. \quad (14)$$

Hence, Fano's inequality gives

$$H\left(W_{\pi(i)} \mid g^\star_{\mathrm{un}}(\overline{W}_{\pi(i)}, U_{\pi(i)})\right) \leq \log 2 + \bar{\beta} \log(|\mathcal{W}| - 1).$$

Therefore,

$$I\left(W_{\pi(i)}; \overline{W}_{\pi(i)} \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right) \geq H\left(W_{\pi(i)} \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right) - \log 2 - \bar{\beta} \log(|\mathcal{W}| - 1).$$

Summing over $i = 1, \ldots, K$ and using Proposition 4.1 gives the claim. □

D.2 Deferred proofs from Section 4.2

Proof of Proposition 4.4. Partition $[N]$ into n consecutive bins

$$B_j = \{(j-1)q + 1, \ldots, jq\}, \qquad j \in [n],$$

where $q = N/n$. Draw $x_j \sim \mathrm{Unif}(B_j)$ independently, and set

$$y_j = \mathbb{I}\{j \geq m\}.$$

Let

$$S = ((x_1, y_1), \ldots, (x_n, y_n)).$$

Since the bins are ordered, every negative example lies to the left of every positive example, so S is realizable by a threshold. By the canonical tie-breaking rule, the ERM output on the full sample is

$$\mathcal{A}_{\mathrm{ERM}}(S) = h_{x_m}.$$

For each $i \in [m]$, define the deletion request

$$U_i = \{(x_j, 0) : 1 \leq j \leq m - i\} \cup \{(x_j, 1) : m \leq j \leq m + i - 1\}.$$

The first set has size $m - i$, the second has size i, and hence $|U_i| = m$. Since $n \geq 2m$, the point x_{m+i} exists. After deleting U_i, the leftmost remaining positive example is x_{m+i}. Therefore

$$W_i \coloneqq \mathcal{A}_{\mathrm{ERM}}(S \setminus U_i) = h_{x_{m+i}}.$$

Let $H_i = (W_{<i}, U_{\leq i})$. The variables in H_i reveal only points from bins other than B_{m+i}. In particular, the previous retraining targets $W_1, \ldots, W_{i-1}$ reveal $x_{m+1}, \ldots, x_{m+i-1}$, all of which are already deleted by U_i. Since the x_j's are independent across bins,

$$x_{m+i} \mid H_i \sim \mathrm{Unif}(B_{m+i}).$$

Distinct threshold locations define distinct hypotheses, so

$$H(W_i \mid H_i) = \log q.$$

Applying $Theorem$ 4.3 and substituting q for each $i \in [m]$, gives the result. Finally, for the bare ERM output, $M_{\mathrm{bare}} = h_{x_m}$, and $x_m \sim \mathrm{Unif}(B_m)$. Hence

$$I(M_{\mathrm{bare}}; S) = H(M_{\mathrm{bare}}) = \log q.$$

□

Proof of Proposition 4.5. Partition $[d]$ into n disjoint blocks

$$B_j = \{(j-1)q+1, \dots, jq\}, \qquad j \in [n].$$

Let $e_a \in \mathbb{R}^d$ denote the a-th standard basis vector. Fix strictly decreasing positive numbers

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0.$$

Draw $a_j \sim \mathrm{Unif}(B_j)$ independently and set

$$z_j = \sqrt{\lambda_j}\, e_{a_j}.$$

Let $S = (z_1, \dots, z_n)$.

The learner returns

$$\mathcal{A}_{\mathrm{PCA}}(D) = e_{\widehat{a}(D)}, \qquad \widehat{a}(D) \in \arg\max_{a \in [d]} \sum_{z \in D} \langle e_a, z \rangle^2,$$

with an arbitrary fixed tie-breaking rule. On the full sample, because the λ_j's are strictly decreasing,

$$M_{\mathrm{bare}} = \mathcal{A}_{\mathrm{PCA}}(S) = e_{a_1}.$$

For $i \in [m]$, define

$$U_i = \{z_1, \dots, z_i\} \cup \{z_{n-m+i+1}, \dots, z_n\},$$

where the second set is empty when $i = m$. The first set has size i and the second has size $m - i$, so $|U_i| = m$. The assumption $2m \le n$ ensures that these deletions do not remove z_{i+1}.

After deleting U_i, the largest remaining empirical second moment is λ_{i+1}, attained uniquely at coordinate a_{i+1}. Hence

$$W_i \coloneqq \mathcal{A}_{\mathrm{PCA}}(S \setminus U_i) = e_{a_{i+1}}.$$

Let $H_i = (W_{<i}, U_{\le i})$. The variables in H_i reveal $a_1, \dots, a_i$, possibly some tail coordinates, and $a_2, \dots, a_i$ through the previous retraining targets. They reveal no information about a_{i+1}. Since the blocks are independent,

$$a_{i+1} \mid H_i \sim \mathrm{Unif}(B_{i+1}).$$

Therefore

$$H(W_i \mid H_i) = \log q.$$

Applying the conditional-support Fano argument from Section 4.2, with q possible targets per request, gives the result. For the bare PCA output, $M_{\mathrm{bare}} = e_{a_1}$, so

$$I(M_{\mathrm{bare}}; S) = H(e_{a_1}) = \log q.$$

□

Proof of Proposition 4.6. Partition the coordinates into b blocks. Block j contains one shortcut coordinate s_j and q candidate coordinates $v_{j,1}, \dots, v_{j,q}$. Thus

$$G_j = \{s_j, v_{j,1}, \dots, v_{j,q}\}, \qquad d = b(q+1).$$

The model class is

$$\mathcal{W}_{\ell_0} = \left\{ w \in \mathbb{R}^d : |\mathrm{supp}\,(w) \cap G_j| \le 1 \text{ for every } j \in [b] \right\},$$

and the learner is

$$\mathcal{A}_{\ell_0}(D) = \arg\min_{w \in \mathcal{W}_{\ell_0}} \sum_{(x,y) \in D} (\langle w, x \rangle - y)^2,$$

with a fixed tie-breaking rule. In the construction below the minimizer is unique, so the tie-breaking rule is irrelevant.

Fix $R > 1$. For each block j, draw

$$A_j \sim \mathrm{Unif}\,([q])$$

independently. The dataset contains m indexed copies of the shortcut example

$$c_j^{(r)} = (e_{s_j}, R), \qquad r \in [m],$$

and one candidate example

$$z_j = (e_{v_{j,A_j}}, 1).$$

Thus

$$|S| = b(m+1) = n.$$

The copies are treated as distinct indexed examples, so a deletion request may delete all m shortcut copies in a block.

The loss separates over blocks. Fix block j, and condition on $A_j = a$. If the model chooses the shortcut coordinate s_j, the optimal coefficient is R, the shortcut examples are fit perfectly, and the candidate example contributes loss 1. If the model chooses the active candidate coordinate $v_{j,a}$, the optimal coefficient is 1, the candidate example is fit perfectly, and the m shortcut examples contribute loss mR^2. If it chooses an inactive candidate coordinate or no coordinate in the block, the loss is at least $mR^2 + 1$. Since $R > 1$ and $m \geq 1$, the unique blockwise minimizer chooses the shortcut coordinate with coefficient R.

Therefore the full trained bare model is deterministic:

$$M_{\text{bare}} = \mathcal{A}_{\ell_0}(S) = R \sum_{j=1}^{b} e_{s_j}.$$

For each $i \in [b]$, define

$$U_i = \left\{ c_i^{(1)}, \ldots, c_i^{(m)} \right\}.$$

This request has size m. After deleting U_i, block i contains only the candidate example $z_i = (e_{v_{i,A_i}}, 1)$. Hence the unique minimizer in block i sets the coefficient of v_{i,A_i} equal to 1. Every other block still chooses its shortcut coordinate. Thus

$$W_i \coloneqq \mathcal{A}_{\ell_0}(S \setminus U_i) = R \sum_{j \neq i} e_{s_j} + e_{v_{i,A_i}}.$$

The map $A_i \mapsto W_i$ is injective.

Let $H_i = (W_{<i}, U_{\leq i})$. The requests $U_1, \ldots, U_i$ reveal only deterministic shortcut examples. The previous retraining targets $W_1, \ldots, W_{i-1}$ reveal $A_1, \ldots, A_{i-1}$, but reveal no information about A_i. Since the A_j's are independent and uniform,

$$A_i \mid H_i \sim \text{Unif}\left([q]\right).$$

Therefore

$$H(W_i \mid H_i) = H(A_i) = \log q.$$

Applying Theorem 4.3 and substituting q gives the result. Since M_{bare} is deterministic on this distribution, $I(M_{\text{bare}}; S) = 0$. □

Proof of Proposition 4.7. Set $r = r_\star$. By definition of $r_\star$, we have $r^2 \leq d$ and $r(m+1) \leq n$. Since $d > 4$ and $n > 3m$, we have $r \geq 2$.

We use the first r^2 ambient coordinates as an $r \times r$ matrix. If $n > r(m+1)$, we add $n - r(m+1)$ deterministic dummy measurements of the form $(1, 0_r, 0)$. These impose no constraint and do not affect the fitted matrix.

Draw a uniformly random permutation σ of $[r]$. For each row $i \in [r]$, include m indexed copies of the row-sum measurement

$$c_i^{(a)} = (i, \mathbf{1}_r, r), \qquad a \in [m],$$

and one coordinate measurement

$$z_i = (i, e_{\sigma(i)}, 1).$$

Thus the non-dummy part of the dataset has size $r(m+1)$. The copies are treated as distinct indexed examples, so a deletion request may delete all m row-sum copies in a row.

We first compute the fitted matrix on the full dataset. Because the data are consistent and $\lambda \to 0$, the learner selects, among interpolating matrices, one minimizing

$$\sum_{i=1}^{r} \|x_i\|_2 .$$

The rows decouple. Fix row i. The constraints are

$$\sum_{\ell=1}^{r} (x_i)_\ell = r, \qquad (x_i)_{\sigma(i)} = 1.$$

Therefore the remaining $r-1$ coordinates must sum to $r-1$. By Cauchy–Schwarz,

$$\sum_{\ell \neq \sigma(i)} (x_i)_\ell^2 \geq \frac{(r-1)^2}{r-1} = r-1,$$

with equality iff all remaining coordinates are equal to 1. Hence the unique minimum-norm interpolating row is

$$x_i = \mathbf{1}_r.$$

This holds for every row, so the full fitted matrix is deterministic:

$$M_{\text{bare}} = \mathcal{A}_{\text{fac}}(S) = \mathbf{1}_r \mathbf{1}_r^\top .$$

For each row $i \in [r]$, define

$$U_i = \left\{ c_i^{(1)}, \ldots, c_i^{(m)} \right\}.$$

This deletion request has size m. After deleting U_i, row i has only the coordinate constraint

$$(x_i)_{\sigma(i)} = 1.$$

The unique minimum-ℓ_2-norm row satisfying this constraint is

$$x_i = e_{\sigma(i)}.$$

All other rows still have both their row-sum and coordinate measurements, so they remain equal to $\mathbf{1}_r$. Thus $W_i = \mathcal{A}_{\text{fac}}(S \setminus U_i)$ is the matrix whose i-th row is $e_{\sigma(i)}$ and whose other rows are $\mathbf{1}_r$. In particular, W_i reveals exactly $\sigma(i)$.

Let $H_i = (W_{<i}, U_{\leq i})$. The deletion requests themselves are deterministic functions of the row index and reveal no information about σ. After observing $W_1, \ldots, W_{i-1}$, one knows

$$\sigma(1), \ldots, \sigma(i-1).$$

Since σ is a uniformly random permutation, conditional on this information, $\sigma(i)$ is uniform over the remaining $r-i+1$ columns. Therefore

$$H(W_i \mid H_i) = \log(r-i+1).$$

Applying $Theorem\ 4.3$ and using $\sum_{s=2}^{r} \log s = \log(r!)$, we get

$$I(M;S) \geq \log(r!) - \sum_{s=2}^{r} \left[h(\rho_s) + \rho_s \log(s-1) \right].$$

Finally, M_{bare} is deterministic on this distribution, so $I(M_{\text{bare}}; S) = 0$. □

E CONTINUOUS MODEL SPACE

For the case where the model space $\mathcal{W}$ is continuous, we follow a similar proof strategy as in the discrete case, but with several important modifications. In this setting, asking whether W is exactly recoverable from the empty-request output, or from the unlearned model together with the unlearning set, is no longer the right notion, since exact equality is typically a probability-zero event. Instead, we use a natural approximate recovery criterion: given a norm ϕ and a radius $\rho > 0$, we ask whether W lies in the ϕ-ball of radius ρ centered at the recovered model. Namely, for $w_0 \in \mathbb{R}^d$, we define the ϕ-ball of radius ρ centered at w_0, restricted to $\mathcal{W}$, as

$$B^{\phi}_{\mathcal{W},\rho}(w_0) \coloneqq \{ w \in \mathcal{W} : \phi(w - w_0) \leq \rho \}.$$

Definition 5 (Continuous recovery errors)**.** The *empty-request recovery error* is defined as,

$$\beta_{\emptyset}^{\star}(\rho) \coloneqq \inf_{g:\mathcal{W}\to\mathcal{W}} \sup_{w\in\mathcal{W}} \mathbb{P}\left(w \notin B_{\mathcal{W},\rho}^{\phi}\left(g\left(W^{\emptyset}\right)\right) \mid W = w\right),$$

where $W^{\emptyset} = \bar{\mathcal{A}}(W, \emptyset)$ is the empty-request output, and the randomness in the probability is over the algorithmic randomness of $\bar{\mathcal{A}}$ only.

Next, recall that the retraining target is $W = \mathcal{A}(S \setminus U)$ and the unlearning output is $\overline{W} = \bar{\mathcal{A}}(W, U)$. The runs of $\mathcal{A}$ for fixed $S = s, U = u$ are independent when $\mathcal{A}$ is randomized. The *unlearning recovery error* is defined as,

$$\beta_{\mathrm{un}}^{\star}(\rho) \coloneqq \inf_{g:\mathcal{W}\times\mathcal{X}^{\star}\to\mathcal{W}} \sup_{\substack{s,u:\, u\subseteq s \\ |u|\le m}} \mathbb{P}\left(W \notin B_{\mathcal{W},\rho}^{\phi}\left(g\left(\overline{W}, u\right)\right) \mid S = s, U = u\right)$$

Here, the randomness in the probability is over the algorithmic randomness of $\mathcal{A}$ and $\bar{\mathcal{A}}$, conditional on fixed $S = s$ and $U = u$.

As in the discrete case, the same recovery map g must work uniformly over all inputs. Otherwise, the recovery error could depend on information that is not available from the unlearning output itself. If $\bar{\mathcal{A}}$ has (ϕ, Γ)-utility on empty requests, then for any $\rho > 0$, using the identity recovery map $g(y) = y$ gives the following upper bound on $\beta_{\emptyset}^{\star}(\rho)$,

$$\beta_{\emptyset}^{\star}(\rho) \le \sup_{w\in\mathcal{W}} \mathbb{P}\left(w \notin B_{\mathcal{W},\rho}^{\phi}\left(g\left(W^{\emptyset}\right)\right) \mid W = w\right) = \sup_{w\in\mathcal{W}} \mathbb{P}\left(w \notin B_{\mathcal{W},\rho}^{\phi}\left(W^{\emptyset}\right) \mid W = w\right) \le \Gamma(\rho). \tag{15}$$

Now, as in the discrete case, we use this result to bound the unlearning recovery error.

Lemma E.1. *Assume that $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (α, ε)-Rényi unlearning for all deletion requests of size at most m and $\bar{\mathcal{A}}$ has (ϕ, Γ)-utility on empty requests. Define $\gamma := \frac{\alpha-1}{\alpha}$, and fix $\rho > 0$, $0 \le r < \rho$. Then,*

$$\begin{aligned} &\beta_{\mathrm{un}}^{\star}(\rho) \le \bar{\beta}_{\rho,r}, \quad \bar{\beta}_{\rho,r} \coloneqq \min\left\{1, e^{\gamma\varepsilon}\left(\Gamma(\rho - r) + \Delta_{\mathcal{A}}(r)\right)^{\gamma}\right\}, \\ &\Delta_{\mathcal{A}}(r) := \sup_{\substack{s,u:\, u\subseteq s \\ |u|\le m}} \mathbb{P}\left(W_{s,u}^{(1)} \notin B_{\mathcal{W},r}^{\phi}\left(W_{s,u}^{(2)}\right) \mid S = s, U = u\right), \end{aligned} \tag{16}$$

and $W_{s,u}^{(1)}$ and $W_{s,u}^{(2)}$ are independent runs of $\mathcal{A}(s \setminus u)$.

Proof of Lemma E.1. Fix $\rho > 0$ and $0 \le r < \rho$. Let $h^{\star}$ be an optimal $(\rho - r)$-decoder for W given Y (if the infimum is not attained, one can use an δ-optimal decoder and let $\delta \downarrow 0$), and write

$$\beta_{\emptyset}^{\star}(\rho - r) = \sup_{w\in\mathcal{W}} \mathbb{P}\left(w \notin B_{\mathcal{W},\rho-r}^{\phi}(h^{\star}(W^{\emptyset})) \mid W = w\right).$$

Define the corresponding decoder from $(\overline{W}, U)$ by

$$g_{\emptyset}(\overline{w}, u) := h^{\star}(\overline{w}),$$

which ignores the u-coordinate. For fixed $w \in \mathcal{W}$ and $u \in \mathcal{X}^{\star}$, define the bad event

$$A_{g,w,u} := \left\{x \in \mathcal{W} : w \notin B_{\mathcal{W},\rho}^{\phi}(g_{\emptyset}(x, u))\right\} = \left\{x \in \mathcal{W} : w \notin B_{\mathcal{W},\rho}^{\phi}(h^{\star}(x))\right\}.$$

By (α, ε)-RU and Lemma B.1, for every fixed s, u, w,

$$\mathbb{P}\left(w \notin B_{\mathcal{W},\rho}^{\phi}(g_{\emptyset}(\overline{W}, u)) \mid S = s, U = u\right) \le e^{\gamma\varepsilon}\mathbb{P}\left(w \notin B_{\mathcal{W},\rho}^{\phi}(h^{\star}(W^{\emptyset})) \mid S = s, U = u\right)^{\gamma}. \tag{17}$$

Let $P_{W|s,u}$ denote the conditional law of $W \mid S = s, U = u$. Integrating equation 17 with respect to $w \sim P_{W|s,u}$, and using Jensen's inequality since x^{γ} is concave for $0 < \gamma < 1$, gives

$$\begin{aligned} &\int \mathbb{P}\left(w \notin B_{\mathcal{W},\rho}^{\phi}(g_{\emptyset}(\overline{W}, u)) \mid S = s, U = u\right) dP_{W|s,u}(w) \\ &\le e^{\gamma\varepsilon}\left(\int \mathbb{P}\left(w \notin B_{\mathcal{W},\rho}^{\phi}(h^{\star}(W^{\emptyset})) \mid S = s, U = u\right) dP_{W|s,u}(w)\right)^{\gamma}. \end{aligned} \tag{18}$$

We now bound the term inside the parentheses. Let

$$W_1, W_2 \mid S = s, U = u \overset{\text{i.i.d.}}{\sim} P_{W|S=s,U=u},$$

and let $W_2^{\emptyset}$ be generated from W_2, namely

$$W_2^{\emptyset} \mid W_2 = w', S = s, U = u \sim P_{Y|W=w',S=s,U=u}.$$

Then

$$\int \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho}(h^\star(W^{\emptyset})) \mid S = s, U = u\right) dP_{W|s,u}(w) = \mathbb{P}\left(W_1 \notin B^{\phi}_{\mathcal{W},\rho}(h^\star(W_2^{\emptyset})) \mid S = s, U = u\right).$$

By the triangle inequality of the norm ϕ, if

$$W_2 \in B^{\phi}_{\mathcal{W},\rho-r}(h^\star(W_2^{\emptyset})) \qquad \text{and} \qquad W_1 \in B^{\phi}_{\mathcal{W},r}(W_2),$$

then

$$W_1 \in B^{\phi}_{\mathcal{W},\rho}(h^\star(W_2^{\emptyset})).$$

Equivalently,

$$\left\{W_1 \notin B^{\phi}_{\mathcal{W},\rho}(h^\star(W_2^{\emptyset}))\right\} \subseteq \left\{W_2 \notin B^{\phi}_{\mathcal{W},\rho-r}(h^\star(W_2^{\emptyset}))\right\} \cup \left\{W_1 \notin B^{\phi}_{\mathcal{W},r}(W_2)\right\}.$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(W_1 \notin B^{\phi}_{\mathcal{W},\rho}(h^\star(W_2^{\emptyset})) \mid S = s, U = u\right) \\
&\leq \mathbb{P}\left(W_2 \notin B^{\phi}_{\mathcal{W},\rho-r}(h^\star(W_2^{\emptyset})) \mid S = s, U = u\right) + \mathbb{P}\left(W_1 \notin B^{\phi}_{\mathcal{W},r}(W_2) \mid S = s, U = u\right). \quad (19)
\end{aligned}$$

It remains to bound the matched term. Since $W^{\emptyset} \perp\!\!\!\perp (S, U) \mid W$,

$$\mathcal{L}(W_2^{\emptyset} \mid W_2 = w, S = s, U = u) = \mathcal{L}(Y \mid W = w).$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(W_2 \notin B^{\phi}_{\mathcal{W},\rho-r}(h^\star(W_2^{\emptyset})) \mid S = s, U = u\right) \\
&= \int \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho-r}(h^\star(W^{\emptyset})) \mid W = w\right) dP_{W|s,u}(w) \\
&\leq \sup_{w \in \mathcal{W}} \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho-r}(h^\star(W^{\emptyset})) \mid W = w\right) = \beta^\star_{\emptyset}(\rho - r). \quad (20)
\end{aligned}$$

Combining equation 18, equation 19, and equation 20, we obtain, for every fixed s, u,

$$\begin{aligned}
&\int \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho}(g_{\emptyset}(\overline{W}, u)) \mid S = s, U = u\right) dP_{W|s,u}(w) \\
&\leq e^{\gamma\varepsilon}\left(\beta^\star_{\emptyset}(\rho - r) + \mathbb{P}\left(W_1 \notin B^{\phi}_{\mathcal{W},r}(W_2) \mid S = s, U = u\right)\right)^{\gamma}.
\end{aligned} \quad (21)$$

Since $T \perp\!\!\!\perp W \mid S, U$, the left-hand side equals

$$\mathbb{P}\left(W \notin B^{\phi}_{\mathcal{W},\rho}(g_{\emptyset}(\overline{W}, u)) \mid S = s, U = u\right).$$

Therefore,

$$\mathbb{P}\left(W \notin B^{\phi}_{\mathcal{W},\rho}(g_{\emptyset}(\overline{W}, u)) \mid S = s, U = u\right) \leq e^{\gamma\varepsilon}\left(\beta^\star_{\emptyset}(\rho - r) + \mathbb{P}\left(W_1 \notin B^{\phi}_{\mathcal{W},r}(W_2) \mid S = s, U = u\right)\right)^{\gamma}.$$

Finally, since $g_{\emptyset}$ is an admissible decoder in the definition of $\beta^\star_{\mathrm{un}}(\rho)$,

$$\beta^\star_{\mathrm{un}}(\rho) \leq \sup_{s,u} \mathbb{P}\left(W \notin B^{\phi}_{\mathcal{W},\rho}(g_{\emptyset}(\overline{W}, u)) \mid S = s, U = u\right).$$

Hence,

$$\beta^\star_{\mathrm{un}}(\rho) \leq e^{\gamma\varepsilon}\left(\beta^\star_{\emptyset}(\rho - r) + \Delta(r)\right)^{\gamma},$$

where

$$\Delta(r) \coloneqq \sup_{s,u} \mathbb{P}\left(W_1 \notin B^{\phi}_{\mathcal{W},r}(W_2) \mid S = s, U = u\right), \qquad W_1, W_2 \mid S = s, U = u \overset{\text{i.i.d.}}{\sim} P_{W|S=s,U=u}.$$

The claim then follows from $\beta^\star_{\emptyset}(\rho - r) \leq \Gamma(\rho - r)$ and $\beta^\star_{\mathrm{un}}(\rho) \leq 1$. □

The upper bound on the unlearning recovery error depends on two coupled radii, ρ and r, which enter through the empty-request Γ-utility condition and the learner's randomness for a fixed input. Choosing r close to ρ reduces the contribution from learner randomness, but it also makes $\Gamma(\rho - r)$ approach $\Gamma(0)$, which can make the bound vacuous. Conversely, taking $r = 0$ minimizes this utility penalty, but for a randomized learner the probability that two independent runs agree exactly may be zero, again making the bound vacuous. Thus, r must be chosen to balance empty-request noise against the intrinsic randomness of the learner.

We now state the memorization lower bound in the continuous setting.

Theorem E.2 (Continuous memorization lower bound)**.** *Let $S \sim \mathcal{P}_S$, and $U_1, \ldots, U_K$ be a set of unlearning requests generated as above, each of size at most m. Assume $\mathcal{W}$ is continuous and that $\mathcal{A}, \bar{\mathcal{A}}$ satisfy the same assumptions as in Lemma E.1. Then, for any permutation $\pi : [K] \to [K]$,*

$$I(\mathcal{A}(S); S) \geq \sum_{i=1}^{K} \left(1 - \bar{\beta}_{\rho,r}\right) \mathbb{E}\left(\log \frac{1}{\alpha_\rho^\phi(W_{\pi(<i)}, U_{\pi(\leq i)})}\right) - K \log 2,$$

where $\bar{\beta}_{\rho,r}$ is defined in Lemma E.1, and

$$\alpha_\rho^\phi(w_{\pi(<i)}, u_{\pi(\leq i)}) \coloneqq \sup_{w_0 \in \mathcal{W}} \mathbb{P}\left(W \in B_{\mathcal{W},\rho}^\phi(w_0) \mid W_{\pi(<i)} = w_{\pi(<i)}, U_{\pi(\leq i)} = u_{\pi(\leq i)}\right).$$

Proof sketch. Apply Proposition 4.1 and define $\mathcal{H}_i := \left(W_{\pi(<i)}, U_{\pi(\leq i)}\right)$ and $\widehat{W}_{\pi(i)} := g_{\mathrm{un}}^\star\left(\overline{W}_{\pi(i)}, U_{\pi(i)}\right)$, where $g_{\mathrm{un}}^\star$ is an (almost-)optimal decoder in the definition of $\beta_{\mathrm{un}}^\star(\rho)$. Fix $\mathcal{H}_i = h_i$. By the data processing inequality and the definition of mutual information through KL-divergence, we get

$$I\left(W_{\pi(i)}; \overline{W}_{\pi(i)} \mid h_i\right) \geq I\left(W_{\pi(i)}; \widehat{W}_{\pi(i)} \mid h_i\right) = D_{\mathrm{KL}}\left(P \,\|\, Q\right),$$

where $P := \mathrm{Law}\left(W_{\pi(i)}, \widehat{W}_{\pi(i)}\right)$ and $Q := \mathrm{Law}\left(W_{\pi(i)}\right) \otimes \mathrm{Law}\left(\widehat{W}_{\pi(i)}\right)$. Moreover,

$$D_{\mathrm{KL}}\left(P \,\|\, Q\right) \geq D_{\mathrm{KL}}\left(\mathrm{Ber}\left(p\right) \,\|\, \mathrm{Ber}\left(q\right)\right) \geq -h(p) - p \log q,$$

where $p := P(E)$, $q := Q(E)$, $E := \left\{(w, w') \in \mathcal{W}^2 \,:\, w \in B_{\mathcal{W},\rho}^\phi\left(w'\right)\right\}$, and $h(\cdot)$ denotes the binary entropy. In other words, p and q are the correct recovery probabilities under the conditional law given $\mathcal{H}_i = h_i$. The Markov structure gives $p \geq 1 - \bar{\beta}_{\rho,r}$ and $q \leq \alpha_\rho^\phi(h_i)$, and therefore $-h(p) - p \log q \geq \left(1 - \bar{\beta}_{\rho,r}\right) \log \frac{1}{\alpha_\rho^\phi(h_i)} - \log 2$. Substituting this bound above, taking expectation over $\mathcal{H}_i$, and summing over i proves the theorem. □

Proof of Theorem E.2. We start from the result of Proposition 4.1. By the data processing inequality, for any decoding function $g : \mathcal{W} \times \mathcal{X}^\star \to \mathcal{W}$, we have

$$I\left(W_{\pi(i)}; T_{\pi(i)} \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right) \geq I\left(W_{\pi(i)}; g(T_{\pi(i)}, U_{\pi(i)}) \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right),$$

since, conditional on $U_{\pi(i)}$, the quantity $g(T_{\pi(i)}, U_{\pi(i)})$ is a function of $T_{\pi(i)}$. For the ease of notation, we define,

$$W := W_{\pi(i)}, \qquad T := T_{\pi(i)}, \qquad U := U_{\pi(i)}, \qquad Z := (W_{\pi(<i)}, U_{\pi(<i)}),$$

so we have that the variables conditioning the mutual information can be expressed as,

$$W_{\pi(<i)}, U_{\pi(\leq i)} = (Z, U).$$

We additionally denote with $g_{\mathrm{un}}^\star$ the optimal decoder for W from T, U, so that,

$$\beta_{\mathrm{un}}^\star(\rho) = \sup_{s,u} \mathbb{P}\left(W \notin B_{\mathcal{W},\rho}^\phi(g_{\mathrm{un}}^\star(T, u)) \mid S = s, U = u\right).$$

Furthermore, we define the maximum mass contained in a radius-ρ ball according to the law of $W \mid U = u, Z = z$,

$$\alpha_\rho^\phi(u, z) = \sup_{w_0 \in \mathcal{W}} \mathbb{P}\left(W \in B_{\mathcal{W},\rho}^\phi(w_0) \mid U = u, Z = z\right),$$

as well as the following shorthands,

$$P_{u,z} := \mathcal{L}(W, \widehat{W} \mid U = u, Z = z),$$
$$Q_{u,z} := \mathcal{L}(W \mid U = u, Z = z) \otimes \mathcal{L}(\widehat{W} \mid U = u, Z = z),$$

where $\widehat{W} := g^\star_{\mathrm{un}}(\overline{W}, U)$. Fix $U = u, Z = z$ and consider the event,

$$E = \{(w, w') \in \mathcal{W}^2 : w \in B^\phi_{\mathcal{W},\rho}(w')\},$$

with,

$$p := P_{u,z}(E), \qquad q := Q_{u,z}(E).$$

We have that the conditional mutual information,

$$\begin{aligned} I(W; \widehat{W} \mid U = u, Z = z) &= D_{\mathrm{KL}}(P_{u,z} \,\|\, Q_{u,z}) \\ &\geq D_{\mathrm{KL}}(\mathrm{Ber}(P_{u,z}(E)) \,\|\, \mathrm{Ber}(Q_{u,z}(E))) \\ &= D_{\mathrm{KL}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)) \\ &= -h(p) - p \log q - (1-p)\log(1-q), \end{aligned} \tag{22}$$

where D_{KL} denotes the KL-divergence, and Ber is the Bernoulli law. We further have that,

$$\begin{aligned} p &= \mathbb{P}\left(W \in B^\phi_{\mathcal{W},\rho}(\widehat{W}) \mid U = u, Z = z\right) \\ &= 1 - \mathbb{P}\left(W \notin B^\phi_{\mathcal{W},\rho}(\widehat{W}) \mid U = u, Z = z\right) \\ &= 1 - \sum_{s \in \mathcal{S}} \mathbb{P}\left(S = s \mid U = u, Z = z\right) \, \mathbb{P}\left(W \notin B^\phi_{\mathcal{W},\rho}(\widehat{W}) \mid S = s, U = u, Z = z\right) \end{aligned} \tag{23}$$

The Markov structure states that, conditional on S and U, the target W is independent of all other variables, and T is independent of the history Z. The latter implies that $\widehat{W} \perp\!\!\!\perp Z \mid S, U$, as $\widehat{W}$ is a function of T given U. Consequently, the pair $(\widehat{W}, W)$ is independent of Z given (S, U). Hence,

$$\begin{aligned} &\mathbb{P}\left(W \notin B^\phi_{\mathcal{W},\rho}(\widehat{W}) \mid S = s, U = u, Z = z\right) \\ &= \mathbb{P}\left(W \notin B^\phi_{\mathcal{W},\rho}(\widehat{W}) \mid S = s, U = u\right) \leq \beta^\star_{\mathrm{un}}(\rho), \end{aligned}$$

which implies,

$$p \geq 1 - \beta^\star_{\mathrm{un}}(\rho).$$

On the other hand, under the law $Q_{u,z}$, we have that W' and $\widehat{W}'$ are independently sampled as,

$$W' \sim \mathcal{L}(W \mid U = u, Z = z), \qquad \widehat{W}' \sim \mathcal{L}(\widehat{W} \mid U = u, Z = z).$$

Then, using this independence, we can express q as,

$$\begin{aligned} q &= \mathbb{E}_{\widehat{W}' \mid U=u, Z=z}\left[\mathbb{P}\left(W' \in B^\phi_{\mathcal{W},\rho}(\widehat{W}') \mid U = u, Z = z, \widehat{W}'\right)\right] \\ &= \mathbb{E}_{\widehat{W}' \mid U=u, Z=z}\left[\mathbb{P}\left(W' \in B^\phi_{\mathcal{W},\rho}(\widehat{W}') \mid U = u, Z = z\right)\right] \\ &\leq \mathbb{E}_{\widehat{W}' \mid U=u, Z=z}\left[\alpha^\phi_\rho(u, z)\right] \\ &= \alpha^\phi_\rho(u, z). \end{aligned} \tag{24}$$

Since $-(1-p)\log(1-q) \geq 0$, we can write equation 22 as,

$$I(W; g^\star_{\mathrm{un}}(T, u) \mid U = u, Z = z) \geq -h(p) - p \log q \geq (1 - \beta^\star_{\mathrm{un}}(\rho)) \log \frac{1}{\alpha^\phi_\rho(u, z)} - \log 2 \tag{25}$$

where the last inequality uses the fact that $h(p) \leq \log 2$ and $-p \log q$ is increasing in p and decreasing in q. Taking the expectation over the law of (U, Z) and substituting the previously redefined variables we finally obtain,

$$I\left(W_{\pi(i)}; g^\star_{\mathrm{un}}(T_{\pi(i)}, U_{\pi(i)}) \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right) \geq (1 - \beta^\star_{\mathrm{un}}(\rho)) \, \mathbb{E}\left(\log \frac{1}{\alpha^\phi_\rho(W_{\pi(<i)}, U_{\pi(\leq i)})}\right) - \log 2.$$

Summing over $i = 1, \ldots, K$ and using Proposition 4.1 gives the claim. □