Why Forget-Only Unlearning Needs Memorization

Anonymous Authors[1]

Abstract

Machine unlearning asks for a deletion procedure whose output is close to retraining from scratch on the retained data. We study the strict forget-only setting, where the unlearner receives only the trained model $M = \mathcal{A}(S)$ and forget set U, with no retained data or auxiliary training state. We show that forget-only unlearning is not uniformly possible: if two datasets yield the same trained model but a common deletion sends their retraining targets far apart, no forget-only unlearner can satisfy Rényi unlearning while preserving nontrivial utility. Conversely, we prove mutual-information lower bounds showing that supporting many deletion requests requires the trained model to memorize enough dataset information to recover the corresponding retraining targets. We instantiate these results on standard learners, including thresholds, medians, SVMs, PCA, sparse regression, and factorized matrix completion. Overall, strict forget-only unlearning can require retaining far more information than ordinary learning: deletion requests may expose information the learner would otherwise discard.

1. Introduction

Privacy regulations, such as the GDPR right to erasure (Art. 17; GDPR (2016)), and the accidental inclusion of poisoned (Schoepf et al., 2024), copyrighted (Dou et al., 2025), or unlawfully collected (Thiel, 2023) training data call for efficient procedures that remove data influence from a model without full retraining. Certified machine *unlearning* formalizes this requirement by asking that the unlearned model to be close in distribution to one retrained from scratch without the removed data (Cao & Yang, 2015). We measure this closeness using a pointwise Rényi definition, which implies standard (ε, δ) notions of unlearning (Mironov, 2017).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

In this work, we study *forget-only* unlearning, a strictly harder unlearning setting in which the unlearning algorithm $\bar{A}$ receives only the trained model $M = \mathcal{A}(S)$ and forget set U, accessing the original dataset S only implicitly through M. Several empirical methods that are forget-only (or close to), including gradient-ascent methods on the forget set, perform well in some settings (Zhang et al., 2024; Fan et al., 2024; Jang et al., 2023), but fail in others (Mavrothalassitis et al., 2025; Yu et al., 2025). This motivates our first question: *Is forget-only unlearning always possible, and if not, what conditions make it possible?*

Our first result shows that forget-only unlearning is not uniformly possible. If two datasets S_1, S_2 yield the same trained model $M = \mathcal{A}(S_1) = \mathcal{A}(S_2)$, but a common deletion U produces well-separated retraining targets, then the unlearner receives the same input (M, U) in both worlds and must output the same distribution, which cannot be close to both targets. We formalize this obstruction through a necessary condition in Theorem 3.2, illustrated for standard learners. Thus, forget-only unlearning can succeed only when the trained model retains enough information to determine the relevant retraining target. This raises our second question: *How much dataset information must the learned model memorize to support forget-only unlearning requests?*

Following Brown et al. (2021); Feldman et al. (2025), we measure memorization through *mutual information* $I(M; S)$, where S is drawn from an analysis distribution and $M = \mathcal{A}(S)$. Our main contribution is a technique for lower bounding $I(M; S)$ in terms of the hypothesis class, the learning and unlearning algorithms, and the Rényi unlearning parameters ε, α. In Section 4, we prove the lower bound for discrete model spaces, and extend it to continuous spaces in Section G.

Theorem 1.1 (Informal memorization lower bound)**.** *Let $M = \mathcal{A}(S)$, and let $W_i = \mathcal{A}(S \setminus U_i)$ be the retrained targets for possible forget requests $U_1, \ldots, U_K$. If $\bar{A}$ is an (α, ε)-unlearning algorithm, then for any ordering π,*

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

Each summand measures the *new* information in the next retraining target beyond what is revealed by the current

request and previous requests and targets, preventing double counting (similar to the inclusion-exclusion principle). The bound holds for any number of requests, any choice of requests, and any ordering, hence the applications choose these objects to make the conditional entropy terms large. The error term captures approximate unlearning: weaker approximation requirements reduce the lower bound, while small error and informative retraining targets force large memorization $I(M;S)$. We instantiate this framework for various learning problems in Sections 4 and F.

Taken together, our results give, to our knowledge, one of the first formal accounts of why memorization is necessary in forget-only unlearning. If (M,U) alone does not determine the appropriate retraining target, unlearning is impossible; conversely, supporting many deletion requests requires the trained model to memorize enough information to recover those targets. In short, forget-only unlearning entails a basic tradeoff: either the learner preserves enough information for future deletions, or some requests require additional retained data, training state, or surrogate information.

2. Preliminaries

Let $\mathcal{X}$ be the data space and $\mathcal{W}$ the model space. A dataset of size n is an ordered tuple $S \in \mathcal{X}^n$; we write $U \subseteq S$ for an unlearning request (forget set) and $S \setminus U$ for the retained set, with both operations interpreted in the multiset sense. Unless stated otherwise, definitions are pointwise in fixed S, U, with probability over algorithmic randomness. In Section 4, when bounding $I(M;S)$, $S \sim P_S$, where P_S is an analysis distribution over datasets.

Unless stated otherwise, the learning algorithm $\mathcal{A} : \mathcal{X}^\star \to \mathcal{W}$ is a deterministic map. The unlearning algorithm is a randomized[1] map $\bar{\mathcal{A}} : \mathcal{W} \times \mathcal{X}^\star \to \mathcal{W}$. For fixed S and $U \subseteq S$, define $M = \mathcal{A}(S)$, $W = \mathcal{A}(S \setminus U)$, $T = \bar{\mathcal{A}}(M, U)$, and $Y = \bar{\mathcal{A}}(W, \emptyset)$, called the pretrained model, retraining target, unlearned model, and empty-request output, respectively. The unlearner $\bar{\mathcal{A}}$ is *forget-only*: at deletion time, it receives only the trained model M and forget set U, with no access to side information such as S, $S \setminus U$, training logs, checkpoints, gradients, or auxiliary statistics. This contrasts with settings where the unlearner may use additional information (Sekhari et al., 2021; Neel et al., 2021).

Rényi divergence. For distributions P, Q and $\alpha > 1$, let $d_\alpha(P,Q) = \max\{D_\alpha(P\|Q), D_\alpha(Q\|P)\}$ denote the symmetric α-Rényi divergence, with the standard definition of D_α recalled in Section B. For random variables X, Y, write $d_\alpha(X,Y) = d_\alpha(\mathcal{L}(X), \mathcal{L}(Y))$. We use Rényi divergence because its symmetric form implies the usual two-sided (ε, δ)-style indistinguishability (Mironov, 2017, Proposition 3) and yields the event-transfer and weak triangle inequalities used in our proofs; see Section B.

[1] The randomness of $\bar{\mathcal{A}}$, and of $\mathcal{A}$ when randomized, is independent of all other randomness.

Definition 2.1. 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 all S and all $U \subseteq S$ with $|U| \le m$,

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

We omit m from the terminology when clear from context.

Throughout the paper, we assume $\varepsilon < \infty$. The comparison target is $\bar{\mathcal{A}}(\mathcal{A}(S \setminus U), \emptyset)$, rather than $\mathcal{A}(S \setminus U)$, so deterministic learners are compared after the same empty-request randomized post-processing, as in Sekhari et al. (2021). A natural utility requirement is that this empty-request unlearning should not change its input too much.

Definition 2.2. Fix a metric $d_{\mathcal{W}}$ on the model space $\mathcal{W}$ and a non-increasing function $\Gamma : [0,\infty) \to [0,1]$. We say that $\bar{\mathcal{A}}$ has $(d_{\mathcal{W}}, \Gamma)$-*utility on empty requests* if, for all $w \in \mathcal{W}$ and all $r \ge 0$: $\mathbb{P}\left(Y \sim Q_w\right)[d_{\mathcal{W}}(Y,w) > r] \le \Gamma(r), Q_w := \mathcal{L}(\bar{\mathcal{A}}(w,\emptyset))$.

When $\mathcal{W}$ is a vector space equipped with the norm ϕ, we use the induced metric $d_{\mathcal{W}}\left(w,w'\right) = \phi\left(w - w'\right)$ and the notation (ϕ,Γ)-utility. When $\mathcal{W}$ is a discrete set, we will use the 0-1 metric $d_{\mathcal{W}}\left(w,w'\right) = \mathbb{I}\left\{w \ne w'\right\}$.

Example 2.3 (Gaussian empty-request map)**.** *Suppose* $\mathcal{W} = \mathbb{R}^d$ *and* $\bar{\mathcal{A}}(w,\emptyset) = w + Z$*, where* $Z \sim \mathcal{N}(0, \sigma^2 I_d)$*. 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 a forget-only unlearner cannot, in general, achieve both small ε and nontrivial utility. An unconditional lower bound over all learners is impossible: a learner could encode the entire dataset S in M, making $S \setminus U$ fully recoverable. Our lower bound therefore depends on the learner-specific quantity termed the *deletion gap*. This also connects to Section 4, which quantifies the memorization needed for successful unlearning. The main structural condition is a learner collision exposed by deletion: two datasets collide under $\mathcal{A}$ before deletion but yield separated retraining targets after removing a common forget set.

Definition 3.1. Fix $\mathcal{A}$, a norm ϕ, and $m \ge 0$. A triple (S_1, S_2, U) is a *shared-deletion example* for $\mathcal{A}$ with m deletions and separation Δ, if $U \subseteq S_1 \cap S_2$, $|U| = m$, and $\mathcal{A}(S_1) = \mathcal{A}(S_2)$, and $\phi(\mathcal{A}(S_1 \setminus U) - \mathcal{A}(S_2 \setminus U)) \ge \Delta$. The *deletion gap* of $\mathcal{A}$ at deletion size m is

$$\Delta_{\mathcal{A},\phi}(m) := \sup_{(S_1,S_2,U)} \phi(\mathcal{A}(S_1 \setminus U) - \mathcal{A}(S_2 \setminus U)),$$

with value 0 if the feasible set is empty.

The second ingredient is that empty-request utility turns separation in model space into separation between probability laws. For $\alpha > 1, u \in (0, 1/2)$, define $\ell_\alpha(u) := \frac{1}{\alpha-1} \log\left((1-u)^\alpha u^{1-\alpha}\right)$, with $\ell_\alpha(0) = +\infty$. For $\Delta > 0$, define

$$\Lambda_{\alpha,\Gamma}(\Delta) := \sup_{0<r<\Delta/2,\ \Gamma(r)<1/2} \ell_\alpha(\Gamma(r)), \qquad (1)$$

with value 0 if the feasible set is empty. By Lemma C.1, if $\bar{\mathcal{A}}$ has (ϕ, Γ)-utility on empty requests and $\phi(w_1 - w_2) \geq \Delta$, then the empty-request laws must be separated: $d_\alpha(\bar{\mathcal{A}}(w_1, \emptyset), \bar{\mathcal{A}}(w_2, \emptyset)) \geq \Lambda_{\alpha,\Gamma}(\Delta)$. Combining the learner-dependent deletion gap with the unlearner-dependent separation $\Lambda_{\alpha,\Gamma}$, we obtain the main result of this section.

Theorem 3.2. *Let $\alpha > 1$, and set $\beta := \frac{\alpha^2}{2\alpha-1}$. Fix $\mathcal{A}$, $\bar{\mathcal{A}}$, a norm ϕ, and $m \geq 0$. Suppose that $\bar{\mathcal{A}}$ has (ϕ, Γ)-utility on empty requests and that $(\mathcal{A}, \bar{\mathcal{A}})$ is (α, ε)-RU. If $\Delta_{\mathcal{A},\phi}(m) > 0$, then for every $0 < \Delta < \Delta_{\mathcal{A},\phi}(m)$,*

$$\varepsilon \geq \frac{\alpha - 1}{2\alpha - 1} \Lambda_{\beta,\Gamma}(\Delta).$$

The full proof is deferred to Section C.

Proof sketch. Fix $0 < \Delta \leq \Delta_{\mathcal{A},\phi}(m)$. By Definition 3.1, there exist S_1, S_2, U with $\mathcal{A}(S_1) = \mathcal{A}(S_2) = M$ and Δ-separated retraining targets $W_i = \mathcal{A}(S_i \setminus U)$. By Lemma C.1, the empty-request laws $\bar{\mathcal{A}}(W_i, \emptyset)$ are $\Lambda_{\beta,\Gamma}(\Delta)$-separated. But the forget-only input is the same, (M, U), so the unlearned output has one common law. The weak triangle inequality with the two empty-request laws gives the bound. □

Theorem 3.2 is informative only when the deletion gap is nonzero. We show that this already occurs in standard problems, using hard-margin SVMs as an example. Further examples in Section D show the range of possible gaps: they can be a constant fraction of the model-space diameter, the full diameter, or even unbounded.

Example 3.3. *Let $d \geq 2$, and let $\mathcal{A}_d^{\mathrm{svm}}$ return the minimum-norm homogeneous separator w satisfying $y \langle w, x \rangle \geq 1$ on realizable samples from $\{x : \|x\|_2 \leq 1\} \times \{-1, +1\}$. Then $\Delta_{\mathcal{A}_d^{\mathrm{svm}}, \|\cdot\|_2}(2) = \infty$.*

Note that, with normalized outputs $w/\|w\|_2$, the gap becomes 2, the diameter of the unit sphere and hence the largest possible gap in that model space. Since the unnormalized SVM gap is infinite, Theorem 3.2 applies at arbitrarily large Δ. For Gaussian empty-request post-processing, where $\Lambda_{\alpha,\Gamma}(\Delta) = \Omega(\Delta^2/\sigma^2)$, this rules out any uniform finite (α, ε)-Rényi unlearning guarantee over all datasets and two-point deletions.

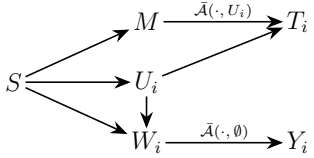

Figure 1. Markov graph for a single deletion request U_i.

4. Memorization for Forget-Only Unlearning

The previous section identified an obstruction to forget-only unlearning: after deletion, the retraining target may depend on dataset information no longer visible from the trained model. We now ask the complementary question: when forget-only unlearning succeeds for many possible requests, how much information about the dataset must the trained model have memorized?

We measure this memorization by mutual information $I(M; S)$, with $S \sim P_S$ drawn from an analysis distribution over datasets and $M = \mathcal{A}(S)$. The unlearning guarantee remains pointwise and worst-case; P_S is used only to measure how much information the trained model carries about the data. The key idea is to analyze several possible deletion requests $U_1, \ldots, U_K$ for the same dataset S in parallel. The unlearner is not run sequentially and does not see previous answers; rather, the single trained model M must contain enough information to support any one request. We account for the required information one request at a time, using the ordering only as an analysis device: for the i-th request, we condition on the requests and retraining targets already counted, preventing double counting.

We use the following setup. Draw $S \sim P_S$ and allow $\mathcal{A}$ to be randomized. Let $U_1, \ldots, U_K$ be requests with $U_k \subseteq S$[2], define $M = \mathcal{A}(S)$, and, for each $k \in [K]$, $W_k = \mathcal{A}(S \setminus U_k)$, $T_k = \bar{\mathcal{A}}(M, U_k)$, and $Y_k = \bar{\mathcal{A}}(W_k, \emptyset)$. For $A^K = (A_1, \ldots, A_K)$ and a permutation π, write $A_{\pi(<i)} = (A_{\pi(1)}, \ldots, A_{\pi(i-1)})$ and $A_{\pi(\leq i)} = (A_{\pi(1)}, \ldots, A_{\pi(i)})$, with $A_{\pi(<1)}$ empty. Figure 1 summarizes the conditional independences used below: $M \to S \to (U_i, W_i)$, T_i depends only on (M, U_i), and Y_i only on W_i.

We consider a finite model space $\mathcal{W}$, and extend to continuous in Section G. The lower bound uses a recovery error for identifying the retraining target from the unlearned model together with the request.

Definition 4.1. The unlearning *recovery error* is

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

The randomness is over the algorithmic randomness of $\mathcal{A}$ and $\bar{\mathcal{A}}$, conditional on fixed s and u.

[2] If randomized, their randomness is independent of the algorithmic randomness.

A single reconstruction map g must work uniformly over all inputs. g observes the unlearned model and forget set, but not the original dataset s, and must recover the model retrained on $s \setminus u$.

Lemma 4.2. *Assume $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (α, ε)-Rényi unlearning, and that $\bar{\mathcal{A}}$ has (d_0, Γ)-utility on empty requests, where $d_0(w, w') = \mathbb{I}\{w \neq w'\}$ and $\eta := \Gamma(0)$. Then, for $\gamma = (\alpha - 1)/\alpha$,*

$$\beta_T^\star \leq \bar{\beta}, \quad \beta := \min\{1,\, e^{\gamma\varepsilon}\,(\eta + \Delta_{\mathcal{A}})^\gamma\}\,,$$

where $\Delta_{\mathcal{A}}$ is the worst-case disagreement probability between two independent runs of $\mathcal{A}(s \setminus u)$.

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

Theorem 4.3. *Let $S \sim P_S$, $M = \mathcal{A}(S)$, and let $U_1, \ldots, U_K$ be requests of size at most m. Suppose $\mathcal{W}$ is finite and the assumptions of Lemma 4.2 hold, with $\bar{\beta}$ as defined there. Then, for any permutation π of $[K]$,*

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

Proof sketch. Apply Proposition B.4; it remains to lower-bound $I(W_{\pi(i)}; T_{\pi(i)} \mid \mathcal{H}_i)$, where $\mathcal{H}_i = (W_{\pi(<i)}, U_{\pi(\leq i)})$. Let g be an almost-optimal decoder for $\beta_T^\star$ and set $\widehat{W}_{\pi(i)} = g(T_{\pi(i)}, U_{\pi(i)})$. By data processing, $I(W_{\pi(i)}; T_{\pi(i)} \mid \mathcal{H}_i) \geq I(W_{\pi(i)}; \widehat{W}_{\pi(i)} \mid \mathcal{H}_i)$. We use Lemma 4.2 to get the uniform error bound $\mathbb{P}\left(\widehat{W}_{\pi(i)} \neq W_{\pi(i)}\right) \leq \bar{\beta}$. Applying Fano's inequality gives $H(W_{\pi(i)} \mid \widehat{W}_{\pi(i)}, \mathcal{H}_i) \leq \log 2 + \bar{\beta}\log(|\mathcal{W}| - 1)$. Hence $I(W_{\pi(i)}; \widehat{W}_{\pi(i)} \mid \mathcal{H}_i) \geq H(W_{\pi(i)} \mid \mathcal{H}_i) - \log 2 - \bar{\beta}\log(|\mathcal{W}| - 1)$. Summing over i proves the theorem. □

The lower bound is informative when the conditional entropy terms are large and the penalty is small. The entropy terms measure the new information in each retraining target beyond previously accounted-for requests and targets, while the penalty captures approximate unlearning. Thus, accurate unlearning with little empty-request noise forces M to retain enough information to reconstruct many possible retraining targets. Next, we provide the necessary conditions for forget-only unlearning under the setup of Theorem 4.3, with $m, \alpha, \eta, \gamma, \bar{\beta}$ as above, for several learning examples. If the usual learner output carries less information than the displayed lower bound, forget-only unlearning requires additional stored state. Full constructions, further examples, and proofs are deferred to Section F.

Canonical threshold ERM. Let $\mathcal{Z} = [N] \times \{0, 1\}$ and $\mathcal{H} = \{h_a : h_a(x) = \mathbb{I}\{x \geq a\}, a \in [N + 1]\}$, with canonical ERM returning the leftmost positive point, or $N + 1$ if none exists. Fix $m \geq 1$, $n \geq 2m$, $q \geq 2$, and $N = nq$. Then there exist a distribution P_S over realizable samples of size n and requests $U_1, \ldots, U_m$, each of size m, such that any forget-only unlearner satisfying the assumptions above obeys

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

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

Factorized affine matrix completion. Consider the row-wise factorized affine matrix-completion learner defined in Section F, which returns the $\lambda \to 0$ fitted matrix under factorized weight decay. Let $d > 4$, $m \geq 1$, $n > 3m$, and $r_\star := \min\left\{\left\lfloor\sqrt{d}\right\rfloor, \left\lfloor\frac{n}{m+1}\right\rfloor\right\}$. Then there exist a distribution P_S over datasets of size n and requests $U_1, \ldots, U_{r_\star}$, each of size m, such that any forget-only unlearner satisfying the assumptions above obeys

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

On P_S, the basic fitted matrix is deterministic, so $I(M;S) = 0$.

5. Related Work and Discussion

Existing empirical forget-only successes (Chundawat et al., 2023; Foster et al., 2024; Thudi et al., 2022) typically rely on benign requests, local stability, proxy retained-data information, or weaker unlearning notions. This is consistent with our bounds: they just require the information needed for retraining to be recoverable from (M, U), whatever it may be. Other guarantees use surrogate distributions or model-side curvature information (Basaran et al., 2025; Ahmed et al., 2025), which either degrades utility or implicitly assumes a near-zero deletion gap. See Section A for related work.

The closest theoretical comparison is Cherapanamjeri et al. (2025), who lower bound memory for learning–unlearning via the *eluder dimension*. Their result is complementary: it treats exact unlearning for realizability testing and is hypothesis-class dependent, whereas our bounds are learner- and request-dependent, apply to strict forget-only unlearning for any (α, ε), and cover regression, PCA, and factorization.

Our bounds quantify how much recoverable information the trained model must memorize, but not which features or statistics to store, or how to store them. Whether memorization already present in overparameterized models can support certified forget-only unlearning remains open. Instantiating our bounds for modern architectures and large-scale pipelines may help distinguish algorithmic failures from information-theoretic ones.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmed, S. M., Basaran, U. Y., Raychaudhuri, D. S., Dutta, A., Kundu, R., Niloy, F. F., Guler, B., and Roy-Chowdhury, A. K. Towards source-free machine unlearning. In *Computer Vision and Pattern Recognition (CVPR)*, 2025.

Allouah, Y., Kazdan, J., Guerraoui, R., and Koyejo, S. The utility and complexity of in- and out-of-distribution machine unlearning. In *International Conference on Learning Representations (ICLR)*, 2025.

Basaran, U. Y., Ahmed, S. M., Roy-Chowdhury, A., and Guler, B. A certified unlearning approach without access to source data. In *International Conference on Machine Learning (ICML)*, 2025.

Brown, G., Bun, M., Feldman, V., Smith, A., and Talwar, K. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Symposium on Theory of Computing (STOC)*, 2021.

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

Chen, H., Zhu, T., Yu, X., and Zhou, W. Zero-shot machine unlearning with proxy adversarial data generation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Cherapanamjeri, Y., Garg, S., Rajaraman, N., Sekhari, A., and Shetty, A. The space complexity of learning-unlearning algorithms. In *Conference on Learning Theory (COLT)*, 2025.

Chundawat, V. S., Tarun, A. K., Mandal, M., and Kankanhalli, M. S. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 2023.

Dou, G., Liu, Z., Lyu, Q., Ding, K., and Wong, E. Avoiding copyright infringement via large language model unlearning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

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

Fan, C., Liu, J., Lin, L., Jia, J., Zhang, R., Mei, S., and Liu, S. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *CoRR*, 2024.

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

Foster, J., Fogarty, K., Schoepf, S., Öztireli, C., and Brintrup, A. Zero-shot machine unlearning at scale via lipschitz regularization. *CoRR*, 2024.

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

Garside, M. J. 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.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. Making ai forget you: Data deletion in machine learning. *Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. *International Conference on Machine Learning (ICML)*, 2020.

Healy, Jr., W. C. 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.

Heinzler, C., Malihi, K., and Sanyal, A. Less noise, same certificate: Retain sensitivity for unlearning. *CoRR*, 2026.

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

Jang, J., Yoon, D., Yang, S., Cha, S., Lee, M., Logeswaran, L., and Seo, M. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

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

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

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. TOFU: A task of fictitious unlearning for llms. In *Conference on Language Modeling*, 2024.

Mavrothalassitis, I., Puigdemont, P., Levi, N. I., and Cevher, V. Ascent fails to forget. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Mironov, I. Rényi differential privacy. In *IEEE Computer security foundations symposium (CSF)*, 2017.

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

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

Schoepf, S., Foster, J., and Brintrup, A. Potion: Towards poison unlearning. *Data-Centric Machine Learning Research (DMLR)*, 2024.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember What You Want to Forget: Algorithms for Machine Unlearning. *Neural Information Processing Systems (NeurIPS)*, 2021.

Shi, W., Lee, J., Huang, Y., Malladi, S., Zhao, J., Holtzman, A., Liu, D., Zettlemoyer, L., Smith, N. A., and Zhang, C. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Thaker, P., Hu, S., Kale, N., Maurya, Y., Wu, Z. S., and Smith, V. Position: Llm unlearning benchmarks are weak measures of progress. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2025.

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

Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. Unrolling SGD: Understanding factors influencing machine unlearning. In *IEEE Symposium on Security and Privacy (IEEE)*, 2022.

Wang, X., Chen, C., Liu, W., Liao, X., Wang, F., and Zheng, X. Efficient source-free unlearning via energy-guided data synthesis and discrimination-aware multitask optimization. In *International Conference on Machine Learning (ICML)*, 2025a.

Wang, Y., Wei, J., Liu, Y., Pang, J., Liu, Q., Shah, A. P., Bao, Y., Liu, Y., and Wei, W. Llm unlearning via loss adjustment with only forget data. In *International Conference on Learning Representations (ICLR)*, 2025b.

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

Yu, J., He, Y., Goyal, A., and Arora, S. On the impossibility of retrain equivalence in machine unlearning. *CoRR*, 2025.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling*, 2024.

A. Additional Related Works

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 at the end of the section. 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.

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. Additional Technical Details

Definition B.1. Let P and Q be probability measures, and $\alpha > 1$. The α-Rényi divergence is,

$$D_\alpha(P\|Q) \coloneqq \frac{1}{\alpha-1}\log\int\left(\frac{dP}{dQ}\right)^\alpha dQ, \tag{2}$$

with the convention $D_\alpha(P\|Q) = \infty$ if P is not absolutely continuous with respect to Q.

Lemma B.2 (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.2. 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,

$$D_\alpha(P\,\|\,Q) = \frac{1}{\alpha-1}\log\int p(x)^\alpha q(x)^{1-\alpha}d\mu(x). \tag{3}$$

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

$$L(x) \coloneqq \frac{dP}{dQ}(x) = \frac{p(x)}{q(x)}. \tag{4}$$

Then we have,

$$\int L(x)^\alpha dQ(x) = \int\left(\frac{p(x)}{q(x)}\right)^\alpha q(x)d\mu(x) = \int p(x)^\alpha q(x)^{1-\alpha}d\mu(x), \tag{5}$$

which implies that,

$$\begin{aligned} D_\alpha(P\,\|\,Q) = \frac{1}{\alpha-1}\log\int L(x)^\alpha dQ(x) = \frac{1}{\alpha-1}\log\mathbb{E}_Q(L^\alpha) \le \varepsilon \\ \implies \mathbb{E}_Q(L^\alpha) \le e^{(\alpha-1)\varepsilon}. \end{aligned} \tag{6}$$

For any event A, we have,

$$P(A) = \int_A dP(x) = \int_A L(x)\,dQ(x) = \mathbb{E}_Q(L\,\mathbb{I}_A). \tag{7}$$

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

$$\begin{aligned} P(A) &= \mathbb{E}_Q(L\,\mathbb{I}_A) \le (\mathbb{E}_Q L^m)^{1/m}\,(\mathbb{E}_Q \mathbb{I}_A^n)^{1/n} \\ &= (\mathbb{E}_Q L^\alpha)^{1/\alpha}\,(\mathbb{E}_Q \mathbb{I}_A)^{(\alpha-1)/\alpha} \le e^{(\alpha-1)/\alpha\,\varepsilon}\,Q(A)^{(\alpha-1)/\alpha}. \end{aligned} \tag{8}$$

This completes the proof. □

Lemma B.3 (Weak triangle inequality)**.** *Let $\alpha > 1$ and set $\widetilde{\alpha} = \frac{\alpha^2}{2\alpha-1}$. For probability measures P, Q, R: $D_{\widetilde{\alpha}}(P\|Q) \le \frac{\alpha}{\alpha-1}D_\alpha(P\|R) + D_\alpha(R\|Q)$, whenever the right-hand side is finite.*

Proof of Lemma B.3. This is Proposition 11 of Mironov (2017) with

$$p = \frac{2\alpha-1}{\alpha}, \qquad q = \frac{2\alpha-1}{\alpha-1}.$$

Indeed, $p\widetilde{\alpha} = \alpha$ and $q(\widetilde{\alpha} - 1/p) = \alpha$, which gives the displayed inequality. □

Proposition B.4. *Let $S \sim P_S$, let $M := \mathcal{A}(S)$, and let $U_1, \dots, U_K$ be requests generated as above. For each $i \in [K]$, define W_i, T_i as in Section 4. Then, for every permutation $\pi : [K] \to [K]$,*

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

Proof of Proposition B.4. Since $M \perp\!\!\!\perp (W^K, U^K) \mid S$, using the data processing inequality in (a), we have that,

$$\begin{aligned}
I(M;S) &\overset{(a)}{\geq} I(M; W^K, U^K) = I\left(M; (W_{\pi(1)}, U_{\pi(1)}), \dots (W_{\pi(K)}, U_{\pi(K)})\right) \\
&\overset{(b)}{=} \sum_{i=1}^{K} I\left(M; W_{\pi(i)}, U_{\pi(i)} | W_{\pi(<i)}, U_{\pi(<i)}\right) \\
&\overset{(c)}{=} \sum_{i=1}^{K} \left[I\left(M; U_{\pi(i)} | W_{\pi(<i)}, U_{\pi(<i)}\right) + I\left(M; W_{\pi(i)} | W_{\pi(<i)}, U_{\pi(<i)}, U_{\pi(i)}\right)\right] \\
&\overset{(d)}{\geq} \sum_{i=1}^{K} I\left(M; W_{\pi(i)} | W_{\pi(<i)}, U_{\pi(<i)}, U_{\pi(i)}\right) \\
&\overset{(e)}{\geq} \sum_{i=1}^{K} I\left(T_{\pi(i)}; W_{\pi(i)} | W_{\pi(<i)}, U_{\pi(<i)}, U_{\pi(i)}\right),
\end{aligned} \tag{9}$$

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 $T_{\pi(i)} = \bar{\mathcal{A}}(M, U_{\pi(i)})$ is generated from $(M, U_{\pi(i)})$, it holds that $T_{\pi(i)} \perp\!\!\!\perp W_{\pi(i)} \mid (M, U_{\pi(i)}, W_{\pi(<i)}, U_{\pi(<i)})$. □

C. Deferred Proofs from Section 3

Lemma C.1. *Let $\alpha > 1$, and suppose $\bar{\mathcal{A}}$ has (ϕ, Γ)-utility on empty requests. Fix two models W_1, W_2, and let $r > 0$ satisfy $\phi(W_1 - W_2) > 2r$. If $\Gamma(r) < 1/2$, then $d_\alpha(Q_{W_1}, Q_{W_2}) \geq \ell_\alpha(\Gamma(r))$.*

Proof. For $i \in \{1, 2\}$, let $B_i := \{y : \phi(y - W_i) \leq r\}$. Since $\phi(W_1 - W_2) > 2r$, the balls B_1 and B_2 are disjoint.

By Definition 2.2, $Q_{W_i}(B_i) \geq 1 - \Gamma(r)$ for $i \in \{1, 2\}$. Since the balls are disjoint, we also have $Q_{W_2}(B_1) \leq \Gamma(r)$ and $Q_{W_1}(B_2) \leq \Gamma(r)$.

Now apply data processing to the indicator map $y \mapsto \mathbf{1}\{y \in B_1\}$. This gives

$$D_\alpha(Q_{W_1} \,\|\, Q_{W_2}) \;\geq\; D_\alpha(\mathrm{Ber}(Q_{W_1}(B_1)) \,\|\, \mathrm{Ber}(Q_{W_2}(B_1)))\,.$$

Using the closed form for Rényi divergence between Bernoulli laws and keeping only the first term,

$$D_\alpha(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)) = \frac{1}{\alpha - 1} \log\big(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\big) \geq \frac{1}{\alpha - 1} \log\big(p^\alpha q^{1-\alpha}\big)\,.$$

Setting $p = Q_{W_1}(B_1)$ and $q = Q_{W_2}(B_1)$, and using $p \geq 1 - \Gamma(r)$ and $q \leq \Gamma(r)$, we obtain

$$D_\alpha(Q_{W_1} \,\|\, Q_{W_2}) \geq \frac{1}{\alpha - 1} \log\Big((1 - \Gamma(r))^\alpha \Gamma(r)^{1-\alpha}\Big) = \ell_\alpha(\Gamma(r))\,.$$

The reverse bound is identical, using the event B_2. □

Proof of Theorem 3.2. Fix any $0 < \Delta < \Delta_{\mathcal{A},\phi}(m)$. By the definition of the deletion gap, there exists a shared-deletion example (S_1, S_2, U) with m deletions and separation at least Δ. Write $M := \mathcal{A}(S_1) = \mathcal{A}(S_2)$, $W_i := \mathcal{A}(S_i \setminus U)$, and $Q_i := Q_{W_i} = \mathcal{L}\left(\bar{\mathcal{A}}(W_i, \emptyset)\right)$ for $i \in \{1, 2\}$. Also write $T := \mathcal{L}\left(\bar{\mathcal{A}}(M, U)\right)$.

Since $(\mathcal{A}, \bar{\mathcal{A}})$ is (α, ε)-RU, we have $d_\alpha(T, Q_i) \leq \varepsilon < \infty$ for $i \in \{1, 2\}$. Hence Q_1, Q_2, and T have the same support.

By the shared-deletion property, $\phi(W_1 - W_2) \geq \Delta$. Thus, by Lemma C.1, for every $r < \Delta/2$ with $\Gamma(r) < 1/2$,

$$D_\beta\left(Q_1 \,\|\, Q_2\right) \geq \ell_\beta\left(\Gamma\left(r\right)\right), \qquad D_\beta\left(Q_2 \,\|\, Q_1\right) \geq \ell_\beta\left(\Gamma\left(r\right)\right).$$

Taking the supremum over such r, we obtain

$$D_\beta\left(Q_1 \,\|\, Q_2\right) \geq \Lambda_{\beta,\Gamma}\left(\Delta\right), \qquad D_\beta\left(Q_2 \,\|\, Q_1\right) \geq \Lambda_{\beta,\Gamma}\left(\Delta\right). \tag{10}$$

Choose $p = \frac{2\alpha-1}{\alpha}$, $q = \frac{2\alpha-1}{\alpha-1}$, and $\beta = \frac{\alpha^2}{2\alpha-1}$. Then $p, q, \beta > 1$, $1/p + 1/q = 1$, $p\beta = \alpha$, and $q\left(\beta - 1/p\right) = \alpha$.

Applying Lemma B.3 with $P = Q_1$, $Q = Q_2$, and $R = T$, we get

$$D_\alpha\left(T \,\|\, Q_2\right) \;\geq\; D_\beta\left(Q_1 \,\|\, Q_2\right) - \frac{\alpha}{\alpha-1} D_\alpha\left(Q_1 \,\|\, T\right).$$

Since $d_\alpha(T, Q_1) \leq \varepsilon$, we have $D_\alpha\left(Q_1 \,\|\, T\right) \leq \varepsilon$. Using (10), we obtain

$$D_\alpha\left(T \,\|\, Q_2\right) \;\geq\; \Lambda_{\beta,\Gamma}\left(\Delta\right) - \frac{\alpha}{\alpha-1}\varepsilon.$$

Since $d_\alpha(T, Q_2) \leq \varepsilon$, we also have $D_\alpha\left(T \,\|\, Q_2\right) \leq \varepsilon$. Therefore

$$\varepsilon \;\geq\; \Lambda_{\beta,\Gamma}\left(\Delta\right) - \frac{\alpha}{\alpha-1}\varepsilon.$$

Rearranging gives

$$\frac{2\alpha-1}{\alpha-1}\,\varepsilon \;\geq\; \Lambda_{\beta,\Gamma}\left(\Delta\right),$$

or equivalently

$$\varepsilon \;\geq\; \frac{\alpha-1}{2\alpha-1}\Lambda_{\beta,\Gamma}\left(\Delta\right).$$

Taking the supremum over $0 < \Delta < \Delta_{\mathcal{A},\phi}(m)$ gives

$$\varepsilon \;\geq\; \frac{\alpha-1}{2\alpha-1} \sup_{0<\Delta<\Delta_{\mathcal{A},\phi}(m)} \Lambda_{\beta,\Gamma}\left(\Delta\right).$$

Finally, the last supremum is $\Lambda_{\beta,\Gamma}\left(\Delta_{\mathcal{A},\phi}(m)\right)$: indeed, $\Lambda_{\beta,\Gamma}$ is nondecreasing, and every $r < \Delta_{\mathcal{A},\phi}(m)/2$ is also below $\Delta/2$ for some $\Delta < \Delta_{\mathcal{A},\phi}(m)$. This proves Theorem 3.2. □

Proof of Example 3.3. Let e_1, e_2 denote the first two standard basis vectors in $\mathbb{R}^d$, and fix $\rho \in (0, 1/\sqrt{2}]$. Define

$$\begin{aligned} U_\rho &= \{(-\rho e_1, -1), (\rho e_1, +1)\}, \\ S_{1,\rho} &= U_\rho \cup \{(-\rho(e_1+e_2), -1), (\rho(e_1+e_2), +1)\}, \\ S_{2,\rho} &= U_\rho \cup \{(-\rho(e_1-e_2), -1), (\rho(e_1-e_2), +1)\}. \end{aligned}$$

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, it follows that $\Delta_{\mathcal{A}_d^{\mathrm{svm}},\|\cdot\|_2}(2) = \infty$. The lower bound on ε follows from Theorem 3.2. □

D. Further examples for Section 3

Example D.1 (Canonical thresholds)**.** *Fix $c \in \mathbb{Z}_+, c > 1$ and let the data space be $\mathcal{X}_c \times \{-1, +1\}$, where $\mathcal{X}_c = \{-c, -c+1, \ldots, c\}$. Let $\mathcal{A}_c$ be the canonical threshold learner on samples supported on $\mathcal{X}_c$ that, on a realizable sample containing both labels, returns the midpoint between the largest negative example and the smallest positive example. Then $\Delta_{\mathcal{A}_c,|\cdot|}(2) = c - \frac{1}{2}$.*

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

We first prove the lower bound. Let $U = \{(-c, -1), (-c+1, +1)\}$, and define the following multisets:

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

Both full samples have largest negative example $-c$ and smallest positive example $-c+1$, so $\mathcal{A}_c(S_1) = \mathcal{A}_c(S_2) = -c + \frac{1}{2}$. After deleting U, however, $\mathcal{A}_c(S_1 \setminus U) = -c + \frac{1}{2}$ and $\mathcal{A}_c(S_2 \setminus U) = 0$. Hence (S_1, S_2, U) is a shared-deletion example with two deletions and separation $c - \frac{1}{2}$. Therefore $\Delta_{\mathcal{A}_c,|\cdot|}(2) \geq c - \frac{1}{2}$.

It remains to show that no larger separation is possible. Consider any shared-deletion example (S_1, S_2, U) with $|U| = 2$. 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 $\mathcal{A}_c(S_1) = \mathcal{A}_c(S_2)$, we have $a_1 + b_1 = a_2 + b_2$. Assume without loss of generality that $\mathcal{A}_c(S_2 \setminus U) \geq \mathcal{A}_c(S_1 \setminus U)$. Then

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

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 $b = b_1 = b_2$,

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

because $a, b \in \mathcal{X}_c$ and $a < b$. This proves $\Delta_{\mathcal{A}_c,|\cdot|}(2) \leq c - \frac{1}{2}$, and hence the claimed equality.

The stated lower bound on ε follows from Theorem 3.2. □

Example D.2 (Empirical median)**.** *Fix $c > 0$, and let $\mathcal{A}_c^{\mathrm{med}}$ be the empirical median learner on samples supported on $[-c, c]$, with midpoint tie-breaking for even sample size. Then $\Delta_{\mathcal{A}_c^{\mathrm{med}},|\cdot|}(3) = 2c$.*

Proof of Example D.2. Let

$$S_1 = \{0, 0, 0, c, c\}, \qquad S_2 = \{-c, -c, 0, 0, 0\}, \qquad U = \{0, 0, 0\}.$$

Since both multisets have size 5, the empirical median is the third order statistic, so $\mathcal{A}_c^{\mathrm{med}}(S_1) = \mathcal{A}_c^{\mathrm{med}}(S_2) = 0$. After deleting U, we obtain $S_1 \setminus U = \{c, c\}$ and $S_2 \setminus U = \{-c, -c\}$. By midpoint tie-breaking for even sample size, $\mathcal{A}_c^{\mathrm{med}}(S_1 \setminus U) = c$ and $\mathcal{A}_c^{\mathrm{med}}(S_2 \setminus U) = -c$. Hence (S_1, S_2, U) is a shared-deletion example with three deletions and separation $2c$, and therefore $\Delta_{\mathcal{A}_c^{\mathrm{med}},|\cdot|}(3) \geq 2c$.

The reverse inequality is immediate because every retained empirical median is supported on $[-c, c]$, so the distance between any two retraining targets is at most $2c$. Thus $\Delta_{\mathcal{A}_c^{\mathrm{med}},|\cdot|}(3) = 2c$. The lower bound on ε follows from Theorem 3.2. □

E. Deferred proofs from Section 4

Proof of Lemma 4.2. Define the empty-request recovery error as,

$$\beta_Y^\star := \inf_{h:\mathcal{W}\mapsto\mathcal{W}} \sup_{w\in\mathcal{W}} \mathbb{P}\left(h(Y) \neq w \mid W = w\right),$$

where the randomness in the probability is over the algorithmic randomness of $\bar{\mathcal{A}}$ only. Since $\bar{\mathcal{A}}$ has (d_0, Γ)-utility with $\eta := \Gamma(0)$, for a decoder $h_I(y) = y$, we have

$$\beta_Y^\star \leq \sup_w \mathbb{P}\left(h_I(Y) \neq w \mid W = w\right) \leq \eta.$$

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

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

Define an admissible decoder for W from (T, U) by

$$g_Y(t, u) := h^\star(t),$$

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

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

By the fact that $(\mathcal{A}, \bar{\mathcal{A}})$ are (α, ε)-RU, and Lemma B.2, for every fixed s, u, w,

$$\mathbb{P}\left(g_Y(T, u) \neq w \mid S = s, U = u\right) \leq e^{\gamma\varepsilon}\mathbb{P}\left(h^\star(Y) \neq w \mid S = s, U = u\right)^\gamma. \tag{11}$$

Denote

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

Multiplying (11) by $p(w \mid s, u)$, summing over w, and using Jensen's inequality since x^γ is concave for $0 < \gamma < 1$, gives

$$\begin{aligned}
&\sum_{w\in\mathcal{W}} p(w \mid s, u)\mathbb{P}\left(g_Y(T, u) \neq w \mid S = s, U = u\right) \\
&\leq e^{\gamma\varepsilon} \sum_{w\in\mathcal{W}} p(w \mid s, u)\mathbb{P}\left(h^\star(Y) \neq w \mid S = s, U = u\right)^\gamma \\
&\leq e^{\gamma\varepsilon} \left(\sum_{w\in\mathcal{W}} p(w \mid s, u)\mathbb{P}\left(h^\star(Y) \neq w \mid S = s, U = u\right)\right)^\gamma.
\end{aligned} \tag{12}$$

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

$$\begin{aligned}
&\sum_{w\in\mathcal{W}} p(w \mid s, u)\mathbb{P}\left(h^\star(Y) \neq w \mid S = s, U = u\right) \\
&= \sum_{w,w'\in\mathcal{W}} p(w \mid s, u)p(w' \mid s, u)\mathbb{P}\left(h^\star(Y) \neq w \mid W = w', S = s, U = u\right).
\end{aligned} \tag{13}$$

Let

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

and let Y_2 be the empty-set unlearning output generated from W_2, namely

$$Y_2 \mid W_2 = w', S = s, U = u \sim P_{Y|W=w',S=s,U=u}.$$

Then the double sum in (13) equals

$$\mathbb{P}\left(h^\star(Y_2) \neq W_1 \mid S = s, U = u\right).$$

Moreover,

$$\{h^\star(Y_2) \neq W_1\} \subseteq \{h^\star(Y_2) \neq W_2\} \cup \{W_1 \neq W_2\},$$

and hence

$$\mathbb{P}\left(h^\star(Y_2) \neq W_1 \mid S=s, U=u\right) \leq \mathbb{P}\left(h^\star(Y_2) \neq W_2 \mid S=s, U=u\right) + \mathbb{P}\left(W_1 \neq W_2 \mid S=s, U=u\right). \tag{14}$$

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

$$\mathcal{L}(Y_2 \mid W_2 = w, S=s, U=u) = \mathcal{L}(Y \mid W = w).$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(h^\star(Y_2) \neq W_2 \mid S=s, U=u\right) &= \sum_{w\in\mathcal{W}} p(w \mid s,u)\mathbb{P}\left(h^\star(Y) \neq w \mid W=w\right) \\ &\leq \sup_{w\in\mathcal{W}} \mathbb{P}\left(h^\star(Y) \neq w \mid W=w\right) = \beta_Y^\star.\end{aligned} \tag{15}$$

Combining (12), (13), (14), and (15), we obtain, for every fixed s, u,

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

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

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

Thus,

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

Finally, since g_Y is an admissible decoder in the definition of $\beta_T^\star$,

$$\beta_T^\star \leq \sup_{s,u} \mathbb{P}\left(g_Y(T,u) \neq W \mid S=s, U=u\right).$$

Therefore,

$$\beta_T^\star \leq e^{\gamma\varepsilon}\left(\beta_Y^\star + \Delta_{\mathcal{A}}\right)^\gamma,$$

where

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

The proof is completed by noting that $\beta_Y^\star \leq \eta$ and $\beta_T^\star \leq 1$. □

Proof of Theorem 4.3. To prove the memorization lower bound, we start from the result of Proposition B.4 and will bound the term,

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

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

$$\begin{aligned}I\left(W_{\pi(i)}; T_{\pi(i)} \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right) &= H\left(W_{\pi(i)} | W_{\pi(<i)}, U_{\pi(\leq i)}\right) - H\left(W_{\pi(i)} | T_{\pi(i)}, W_{\pi(<i)}, U_{\pi(\leq i)}\right) \\ &\overset{(a)}{\geq} H\left(W_{\pi(i)} | W_{\pi(<i)}, U_{\pi(\leq i)}\right) - H\left(W_{\pi(i)} | g_T^\star(T_{\pi(i)}, U_{\pi(i)})\right),\end{aligned} \tag{16}$$

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

$$H\left(W_{\pi(i)} | T_{\pi(i)}, W_{\pi(<i)}, U_{\pi(\leq i)}\right) \leq H\left(W_{\pi(i)} | T_{\pi(i)}, U_{\pi(i)}\right) \leq H\left(W_{\pi(i)} |\, g_T^\star(T_{\pi(i)}, U_{\pi(i)})\right).$$

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

$$\left(W_{\pi(i)},\, g_T^\star(T_{\pi(i)}, U_{\pi(i)})\right),$$

we get

$$H\left(W_{\pi(i)} \mid g_T^\star(T_{\pi(i)}, U_{\pi(i)})\right) \le h(P_{e,\pi(i)}) + P_{e,\pi(i)} \log(|\mathcal{W}| - 1) \le \log 2 + P_{e,\pi(i)} \log(|\mathcal{W}| - 1),$$

where

$$P_{e,\pi(i)} := \mathbb{P}\left(g_T^\star(T_{\pi(i)}, U_{\pi(i)}) \neq W_{\pi(i)}\right).$$

Moreover,

$$P_{e,\pi(i)} = \mathbb{E}_{S,U_{\pi(i)}}\left[\mathbb{P}\left(g_T^\star(T_{\pi(i)}, U_{\pi(i)}) \neq W_{\pi(i)} \mid S, U_{\pi(i)}\right)\right] \le \mathbb{E}_{S,U_{\pi(i)}}[\beta_T^\star] = \beta_T^\star \le \bar{\beta}. \tag{17}$$

Hence, Fano's inequality gives

$$H\left(W_{\pi(i)} \mid g_T^\star(T_{\pi(i)}, U_{\pi(i)})\right) \le \log 2 + \bar{\beta} \log(|\mathcal{W}| - 1).$$

Therefore,

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

Summing over $i = 1, \dots, K$ and using Proposition B.4 gives the claim. □

F. Full formulations, further examples and their proofs – Section 4

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

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

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

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

Proof of Proposition F.1. Partition $[N]$ into n consecutive bins

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

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

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

Let

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

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

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

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

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

□

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 $\frac{1}{2} \sum_{(i,a,y) \in S} \left(\langle a, u_i v_i \rangle - y\right)^2 + \frac{\lambda}{2} \sum_{i=1}^r \left(u_i^2 + \|v_i\|_2^2\right)$. 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 F.2 (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.

Proof of Proposition F.2. 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 $I(M_{\text{bare}}; S) = 0$. □

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 F.3 (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 $I(M_{\text{bare}}; S) = \log q$.

Proof of Proposition F.3. Partition $[d]$ into n disjoint blocks

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

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

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

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

The learner returns

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

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

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

For $i \in [m]$, define

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

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 $2m \leq n$ 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

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

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

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

Therefore

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

Applying *Theorem* 4.3 for each $i \in [m]$ and substuting q gives the result. For the bare PCA output, $M_{\text{bare}} = e_{a_1}$, so

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

□

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 $|\text{supp}\,(w) \cap G_j| \leq 1$ 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 F.4 (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 $I(M;S) = 0$.

Proof of Proposition F.4. Partition the coordinates into b blocks. Block j contains one shortcut coordinate s_j and q candidate coordinates $v_{j,1}, \ldots, v_{j,q}$. Thus

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

The model class is

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

and the learner is

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

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:

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

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

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

The map $A_i \mapsto W_i$ is injective.

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

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

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, $I(M_{\mathrm{bare}}; S) = 0$. □

G. 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

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

Definition G.1 (Continuous recovery errors)**.** Recall that $Y = \bar{\mathcal{A}}(W, \emptyset)$. The empty-request recovery error is

$$\beta^{\star}_Y(\rho) \coloneqq \inf_{g:\mathcal{W}\to\mathcal{W}} \sup_{w\in\mathcal{W}} \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho}(g(Y)) \mid W = w\right).$$

where the randomness in the probability is over the algorithmic randomness of $\bar{\mathcal{A}}$ only. Next, recall $M \sim \mathcal{A}(S), W \sim \mathcal{A}(S \setminus U), T \sim \bar{\mathcal{A}}(M, U)$, where the runs of $\mathcal{A}$ on the same dataset are independent when $\mathcal{A}$ is randomized. The unlearning recovery error is

$$\beta^{\star}_T(\rho) \coloneqq \inf_{g:\mathcal{W}\times\mathcal{X}^{\star}\to\mathcal{W}} \sup_{\substack{s,u:\, u\subseteq s\\ |u|\le m}} \mathbb{P}\left(W \notin B^{\phi}_{\mathcal{W},\rho}(g(T,u)) \mid S = s, U = u\right).$$

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

Similarly to the discrete case, 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 $\beta^{\star}_Y(\rho)$:

$$\beta^{\star}_Y(\rho) \le \sup_{w\in\mathcal{W}} \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho}(g(Y)) \mid W = w\right) = \sup_{w\in\mathcal{W}} \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho}(Y) \mid W = w\right) \le \Gamma(\rho). \tag{18}$$

Now, analogous to the discrete case, we bound the recovery error $\beta^{\star}_T(\rho)$.

Lemma G.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 (ϕ, Γ)-utility on empty requests. Define $\gamma := \frac{\alpha-1}{\alpha}$, and fix $\rho > 0$, $0 \le r < \rho$. Then,*

$$\begin{aligned} \beta^{\star}_T(\rho) &\le \bar{\beta}_{\rho,r}, \quad \bar{\beta}_{\rho,r} \coloneqq \min\left\{1, e^{\gamma\varepsilon}\left(\Gamma(\rho - r) + \Delta_{\mathcal{A}}(r)\right)^{\gamma}\right\}, \\ \Delta_{\mathcal{A}}(r) &:= \sup_{\substack{s,u:\, u\subseteq s\\ |u|\le m}} \mathbb{P}\left(W^{(1)}_{s,u} \notin B^{\phi}_{\mathcal{W},r}\left(W^{(2)}_{s,u}\right) \mid S = s, U = u\right), \end{aligned} \tag{19}$$

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

Proof of Lemma G.2. 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

$$\beta^{\star}_Y(\rho - r) = \sup_{w\in\mathcal{W}} \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho-r}(h^{\star}(Y)) \mid W = w\right).$$

Define the corresponding decoder from (T, U) by

$$g_Y(t, u) := h^\star(t),$$

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

$$A_{g,w,u} := \left\{ x \in \mathcal{W} : w \notin B^\phi_{\mathcal{W},\rho}(g_Y(x, u)) \right\} = \left\{ x \in \mathcal{W} : w \notin B^\phi_{\mathcal{W},\rho}(h^\star(x)) \right\}.$$

By (α, ε)-RU and Lemma B.2, for every fixed s, u, w,

$$\mathbb{P}\left(w \notin B^\phi_{\mathcal{W},\rho}(g_Y(T, u)) \mid S = s, U = u \right) \le e^{\gamma\varepsilon} \mathbb{P}\left(w \notin B^\phi_{\mathcal{W},\rho}(h^\star(Y)) \mid S = s, U = u \right)^\gamma. \tag{20}$$

Let $P_{W|s,u}$ denote the conditional law of $W \mid S = s, U = u$. Integrating (20) with respect to $w \sim P_{W|s,u}$, and using Jensen's inequality since x^γ is concave for $0 < \gamma < 1$, gives

$$\begin{aligned} &\int \mathbb{P}\left(w \notin B^\phi_{\mathcal{W},\rho}(g_Y(T, u)) \mid S = s, U = u \right) dP_{W|s,u}(w) \\ &\le e^{\gamma\varepsilon} \left(\int \mathbb{P}\left(w \notin B^\phi_{\mathcal{W},\rho}(h^\star(Y)) \mid S = s, U = u \right) dP_{W|s,u}(w) \right)^\gamma. \end{aligned} \tag{21}$$

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 Y_2 be generated from W_2, namely

$$Y_2 \mid W_2 = w', S = s, U = u \sim P_{Y|W=w',S=s,U=u}.$$

Then

$$\int \mathbb{P}\left(w \notin B^\phi_{\mathcal{W},\rho}(h^\star(Y)) \mid S = s, U = u \right) dP_{W|s,u}(w) = \mathbb{P}\left(W_1 \notin B^\phi_{\mathcal{W},\rho}(h^\star(Y_2)) \mid S = s, U = u \right).$$

By the triangle inequality of the norm ϕ, if

$$W_2 \in B^\phi_{\mathcal{W},\rho-r}(h^\star(Y_2)) \qquad \text{and} \qquad W_1 \in B^\phi_{\mathcal{W},r}(W_2),$$

then

$$W_1 \in B^\phi_{\mathcal{W},\rho}(h^\star(Y_2)).$$

Equivalently,

$$\left\{ W_1 \notin B^\phi_{\mathcal{W},\rho}(h^\star(Y_2)) \right\} \subseteq \left\{ W_2 \notin B^\phi_{\mathcal{W},\rho-r}(h^\star(Y_2)) \right\} \cup \left\{ W_1 \notin B^\phi_{\mathcal{W},r}(W_2) \right\}.$$

Hence,

$$\begin{aligned} &\mathbb{P}\left(W_1 \notin B^\phi_{\mathcal{W},\rho}(h^\star(Y_2)) \mid S = s, U = u \right) \\ &\le \mathbb{P}\left(W_2 \notin B^\phi_{\mathcal{W},\rho-r}(h^\star(Y_2)) \mid S = s, U = u \right) + \mathbb{P}\left(W_1 \notin B^\phi_{\mathcal{W},r}(W_2) \mid S = s, U = u \right). \end{aligned} \tag{22}$$

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

$$\mathcal{L}(Y_2 \mid W_2 = w, S = s, U = u) = \mathcal{L}(Y \mid W = w).$$

Therefore,

$$\begin{aligned} &\mathbb{P}\left(W_2 \notin B^\phi_{\mathcal{W},\rho-r}(h^\star(Y_2)) \mid S = s, U = u \right) \\ &= \int \mathbb{P}\left(w \notin B^\phi_{\mathcal{W},\rho-r}(h^\star(Y)) \mid W = w \right) dP_{W|s,u}(w) \\ &\le \sup_{w \in \mathcal{W}} \mathbb{P}\left(w \notin B^\phi_{\mathcal{W},\rho-r}(h^\star(Y)) \mid W = w \right) = \beta^\star_Y(\rho - r). \end{aligned} \tag{23}$$

Combining (21), (22), and (23), we obtain, for every fixed s, u,

$$\begin{aligned}&\int \mathbb{P}\left(w \notin B^{\phi}_{\mathcal{W},\rho}(g_Y(T,u)) \mid S=s, U=u\right) dP_{W|s,u}(w)\\ &\leq e^{\gamma\varepsilon}\left(\beta^{\star}_Y(\rho-r)+\mathbb{P}\left(W_1 \notin B^{\phi}_{\mathcal{W},r}(W_2) \mid S=s, U=u\right)\right)^{\gamma}.\end{aligned} \tag{24}$$

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

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

Therefore,

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

Finally, since g_Y is an admissible decoder in the definition of $\beta^{\star}_T(\rho)$,

$$\beta^{\star}_T(\rho) \leq \sup_{s,u} \mathbb{P}\left(W \notin B^{\phi}_{\mathcal{W},\rho}(g_Y(T,u)) \mid S=s, U=u\right).$$

Hence,

$$\beta^{\star}_T(\rho) \leq e^{\gamma\varepsilon}\left(\beta^{\star}_Y(\rho-r)+\Delta(r)\right)^{\gamma},$$

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 $\beta^{\star}_Y(\rho-r) \leq \Gamma(\rho-r)$ and $\beta^{\star}_T(\rho) \leq 1$. □

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

Theorem G.3 (Continuous memorization lower bound)**.** *Let $S \sim P_S$, let $M \coloneqq \mathcal{A}(S)$, and let $U_1, \ldots, U_K$ be possible 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 Lemma G.2. Then, for any permutation $\pi : [K] \to [K]$, we have*

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

where $\bar{\beta}_{\rho,r}$ is defined in Lemma G.2, and

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

Proof of Theorem G.3. We start from the result of Proposition B.4. 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,

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

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 $g_T^\star$ the optimal decoder for W from T, U, so that,

$$\beta_T^\star(\rho) = \sup_{s,u} \mathbb{P}\left(W \notin B^\phi_{\mathcal{W},\rho}(g_T^\star(T,u)) \mid S=s, U=u\right).$$

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

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

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 $\widehat{W} := g_T^\star(T,U)$. 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,

$$\begin{aligned} I(W;\widehat{W} \mid U=u, Z=z) &= D_{\mathrm{KL}}(P_{u,z} \,\|\, Q_{u,z}) \\ &\geq D_{\mathrm{KL}}(\mathrm{Ber}(P_{u,z}(E)) \,\|\, \mathrm{Ber}(Q_{u,z}(E))) \\ &= D_{\mathrm{KL}}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(q)) \\ &= -h(p) - p\log q - (1-p)\log(1-q), \end{aligned} \tag{25}$$

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

$$\begin{aligned} p &= \mathbb{P}\left(W \in B^\phi_{\mathcal{W},\rho}(\widehat{W}) \mid U=u, Z=z\right) \\ &= 1 - \mathbb{P}\left(W \notin B^\phi_{\mathcal{W},\rho}(\widehat{W}) \mid U=u, Z=z\right) \\ &= 1 - \sum_{s \in \mathcal{S}} \mathbb{P}(S=s \mid U=u, Z=z)\ \mathbb{P}\left(W \notin B^\phi_{\mathcal{W},\rho}(\widehat{W}) \mid S=s, U=u, Z=z\right) \end{aligned} \tag{26}$$

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,

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

which implies,

$$p \geq 1 - \beta_T^\star(\rho).$$

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,

$$\begin{aligned} q &= \mathbb{E}_{\widehat{W}' \mid U=u, Z=z}\left[\mathbb{P}\left(W' \in B^\phi_{\mathcal{W},\rho}(\widehat{W}') \mid U=u, Z=z, \widehat{W}'\right)\right] \\ &= \mathbb{E}_{\widehat{W}' \mid U=u, Z=z}\left[\mathbb{P}\left(W' \in B^\phi_{\mathcal{W},\rho}(\widehat{W}') \mid U=u, Z=z\right)\right] \\ &\leq \mathbb{E}_{\widehat{W}' \mid U=u, Z=z}\left[\alpha_\rho^\phi(u,z)\right] \\ &= \alpha_\rho^\phi(u,z). \end{aligned} \tag{27}$$

Since $-(1-p)\log(1-q) \geq 0$, we can write (25) as,

$$I(W; g_T^\star(T,u) \mid U=u, Z=z) \geq -h(p) - p\log q \geq (1-\beta_T^\star(\rho))\log\frac{1}{\alpha_\rho^\phi(u,z)} - \log 2 \tag{28}$$

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,

$$I\left(W_{\pi(i)}; g_T^\star(T_{\pi(i)}, U_{\pi(i)}) \mid W_{\pi(<i)}, U_{\pi(\leq i)}\right) \geq (1-\beta_T^\star(\rho))\,\mathbb{E}\left(\frac{1}{\alpha_\rho^\phi(W_{\pi(<i)}, U_{\pi(\leq i)})}\right) - \log 2.$$

Summing over $i = 1, \ldots, K$ and using Proposition B.4 gives the claim. □