Machine Unlearning via Information Theoretic Regularization

Shizhou Xu SHZXU@UCDAVIS.EDU
Department of Mathematics
University of California Davis
Davis, CA 95616-5270, USA

Thomas Strohmer STROHMER@MATH.UCDAVIS.EDU
Department of Mathematics
Center of Data Science and Artificial Intelligence Research
University of California Davis
Davis, CA 95616-5270, USA

Abstract

How can we effectively remove or "unlearn" undesirable information, such as specific features or the influence of individual data points, from a learning outcome while minimizing utility loss and ensuring rigorous guarantees? We introduce a unified mathematical framework based on information-theoretic regularization to address both data-point unlearning and feature unlearning. For data-point unlearning, we introduce the *Marginal Unlearning Principle*, an auditable and provable framework. Moreover, we provide an information-theoretic unlearning definition based on the proposed principle and provable guarantees on sufficiency and necessity of marginal unlearning. We then show that the proposed framework provides a natural solution to the marginal unlearning problem and yields auditable high-probability marginal-unlearning guarantees. For feature unlearning, the framework applies to deep learning with flexible training objectives. By combining flexibility in learning objectives with simplicity in regularization design, our approach is highly adaptable and practical for a wide range of machine learning and AI applications. From a mathematical perspective, we provide a unified analytic solution to the optimal feature unlearning problem with a variety of information-theoretic training objectives. Our theoretical analysis reveals intriguing connections between machine unlearning, information theory, optimal transport, and extremal sigma algebras. Numerical simulations support our theoretical findings.

Keywords: Machine Unlearning, Feature Unlearning, Data Privacy, Information-Theoretic Regularization, Optimal Transport, Marginal Unlearning

© Shizhou Xu.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/.

Contents

1 Introduction **4**
1.1 Paper organization and contribution 5
1.2 Related work 6
1.3 Tools and notation 8

2 Unlearning Principles, Definitions, & Justifications **9**
2.1 Feature unlearning 10
2.2 Data unlearning 11
2.3 Sufficiency for existing definitions: an auditable criterion 15
2.4 Necessity for any "robust" retrained model 17

3 Information Theoretic Unlearning Framework **18**
3.1 An information-theoretic lens 18
3.2 Problem statements 19
3.3 Interpreting feature unlearning and marginal data unlearning as a generalization of rate-distortion theory 21

4 Theoretical Guarantees **22**
4.1 Feature unlearning guarantee 22
4.2 Data unlearning guarantee 25
4.3 An analytic feature unlearning solution independent of the downstream tasks 26

5 Algorithm Design **29**
5.1 General cost feature unlearning: solution to Definition 2.2 30
5.2 General cost marginal data unlearning: solution to Definition 2.5 30
5.3 Analytic optimal feature unlearning for specific utilities via Theorem 4.3 . . 31

6 Numerical Experiments **32**
6.1 Feature Unlearning via Regularization: Algorithm 1 33
6.2 Marginal Data-Point Unlearning via Regularization: Algorithm 2 34
6.2.1 Forget Gaussian Experiment 34
6.2.2 Forget MNIST 38
6.3 Optimal Feature Unlearning via Analytic Solution: Algorithm 3 42

A Appendix: Supplementary material for Section 1 **52**
A.1 Feature Unlearning 52
A.2 Data Unlearning 53

B Appendix: Supplementary material for Section 2 **54**
B.1 Proof of Lemma 2.1 54
B.2 Proof of Theorem 2.1 54
B.3 Proof of Lemma 2.2 55

C Appendix: Supplementary material for Section 3 **56**
C.1 Utility motivation 56
C.2 Admissibility 57
C.3 Details on considered utility quantifications 58
C.4 Formulation of constrained optimization problems 59

D Appendix: Supplementary material for Section 4 **61**
D.1 Wasserstein distance and barycenter 61
D.2 Proof of Lemma 4.1 62
D.3 Proof of Lemma 4.2 63
D.4 Sigma-algebra and information 63
D.5 Proof of Lemma 4.3 64
D.6 Intuitive insights into Theorem 4.3 65

1. Introduction

As machine learning systems are deployed in sensitive and scientific domains, it becomes essential to remove the influence of designated attributes or individual records from *trained* models while preserving utility and providing rigorous guarantees. Naively deleting attributes or data points from the raw data, or masking them in model outputs, rarely suffices: their effects often persist through correlated proxies, latent representations, and model parameters. Fully retraining on the corrected training data set becomes impractical at scale, especially considering the increasing size of AI models and the potential sequence of unlearning requests. *Machine unlearning* offers a principled alternative, enabling models to "forget" specified information for privacy, fairness, legal compliance, or scientific correction. The central challenge is to design provable, auditable procedures that achieve the desired removal with minimal utility loss, while avoiding unnecessary end-to-end retraining when possible.

In this work, we focus on two complementary problem settings, motivated by distinct practical needs:

- **Feature unlearning (removing attribute influence).** Redacting explicit tokens rarely suffices when proxies remain. For example, a recruiting model favored male applicants because the training data were predominantly male; removing explicit gender markers failed due to correlated features (e.g., all-women colleges) [20, 16]. Feature unlearning aims to neutralize the *influence* of a designated attribute Z on what the model releases, denoted generically by S, even through proxies.
- **Marginal data-point unlearning (removing a record's footprint).** Under GDPR's "right to be forgotten" [27] and California's CCPA [10], individuals may request deletion of their data. Naively deleting records and fine-tuning can leave a detectable footprint in the parameters. Retraining from scratch eliminates a record's *unique contribution* while preserving performance attributable to retained data [11], but is costly. Marginal data-point unlearning seeks efficient, auditable post-hoc procedures that approximate the retrain-on-retain behavior while maintaining utility.

Although originally motivated by privacy and fairness, machine unlearning now plays a broader role in various domains, such as AI-assisted scientific discovery and autonomous experimentation. In these settings, unlearning is used to remove the influence of corrupted or drifting measurements (e.g., broken sensors), instrument-specific artifacts, and other unwanted dependencies (e.g., sensor or batch effects), thereby correcting models and pipelines without costly retraining and enabling reliable, on-the-fly scientific discovery.

For intuition, it is helpful (but not necessary) to think of X as a data matrix (e.g., an EHR table or consumer database), where rows are users and columns are attributes. Feature unlearning removes the information or influence of the columns $X_{[:,j\in z]}$[1] corresponding to Z, whereas data-point unlearning removes the information contributed by rows X_u to the remainder $X_r := X_{[i\notin u,:]}$.

Problem statement. We introduce a generic information-theoretic framework that covers both unlearning settings:

1. If the feature Z to be unlearned is not explicitly present in X, one can still remove information in X that is *correlated* with Z by working with the joint pair (X, Z) as the effective input.

Given data X, an unlearning variable Z, and a target variable Y, how can we construct a released outcome S that exposes minimal information about Z while preserving as much task-relevant information about Y as possible?

Here, S may live in the data space, an intermediate representation space, or the output space. In feature unlearning, Z denotes the feature, source, or intervention label to be neutralized, and the released object takes the form $S = f(X, Z)$. In marginal data-point unlearning, Z denotes the latent source label distinguishing the retain distribution from the retain-plus-unlearn mixture, and the released object takes the form $S = f(X)$. Section 2 formalizes these two cases.

1.1 Paper organization and contribution

This paper makes the following contributions:

- **Marginal unlearning principle.** We introduce the *Marginal Unlearning Principle* and a corresponding formal unlearning definition (Definition 2.3) in Section 2.2. Inspired by "blur + reinforce" memory suppression mechanisms in cognitive neuroscience, this principle offers an inference-based, post-hoc guarantee for data point unlearning. We show that the proposed principle and the resulting definition guarantee (i) *practically checkable*, since it depends only on the released model and the original dataset in Section 4.2, (ii) *sufficient* for a "good-utility" model retrained from scratch on the retain data set, under utility requirements in Section 2.3, and (iii) *necessary* for a "robust" model retrained from scratch on the retain dataset unless the retrained model violates smoothness/Lipschitz regularity, which would harm generalization (Lemma 2.2) in Section 2.4.
- **Unified information-theoretic framework and practical algorithms.** We formulate unlearning as an information-theoretic trade-off, analogous to a rate–distortion objective (Section 3.2):

$$\inf_{f} \quad (1-\lambda) \underbrace{\mathcal{C}(Y;S,Z)}_{\text{utility: memory reinforcement}} + \lambda \underbrace{I(S';Z)}_{\text{unlearning: memory blur}}, \qquad \lambda \in [0,1], \tag{1}$$

  where S denotes the released learning outcome and S' denotes the auditing object on which unlearning is certified. The term $I(S';Z)$ quantifies the information about the unlearning variable Z exposed by the audited object. In feature unlearning, Z is an observed attribute, source, or intervention label, and typically $S' = S = f(X, Z)$. The utility $\mathcal{C}(Y;S,Z)$ measures task-relevant information preserved within each Z-slice, while the exposure penalty $I(S;Z)$ controls what S reveals about Z. In marginal data-point unlearning, Z is instead a latent source label distinguishing the retain-only reference from the retain-plus-unlearn mixture. It is used only for indistinguishability certification, and utility is evaluated from the released model output alone. In this case, the utility term specializes to a non-sliced cost $\mathcal{C}(Y;S)$. We provide practical regularization algorithms for both settings (Algorithms 1 and 2, Section 5) and evaluate their empirical behavior on tabular and image data.
- **Analytic solution and theoretical guarantees.** We first show that bounding the "compression rate" $I(S';Z)$ directly implies high-probability guarantees for feature unlearning (Theorem 4.1) and marginal data-point unlearning (Lemma 4.1, Theorem 4.2).

This provides actionable guidance for tuning λ to meet a target leakage level while preserving learning utility. Furthermore, in the hard-independence limit $S' \perp Z$, we identify an analytic $\mathcal{W}_2$-barycenter solution for feature unlearning in the data-space setting under a class of utility objectives. The barycentric representation generates a finest lifted sigma-algebra among admissible outcomes (Lemma 4.2), yielding maximal lifted utility while enforcing $S' \perp Z$ (Theorem 4.3). We also provide a practical implementation of the analytic solution as Algorithm 3 in Section 5.3.

- **Exemplary numerical experiments.** We evaluate both feature unlearning and datapoint unlearning across tabular and image modalities. Section 6.2 reports data-unlearning results on tabular data and MNIST; Sections 6.1 and 6.3 report feature-unlearning results on tabular benchmarks and CelebA, respectively.
- **Intriguing mathematical connections.** Our proposed framework uncovers compelling connections between machine unlearning, information theory, optimal transport, and extremal sigma algebras.

Organization. The remainder of the paper is organized as follows. Section 1 reviews related work and establishes notation. Section 2 formulates the principles for feature and datapoint unlearning and provides definition-related theoretical guarantees. Section 3 presents the unified information-theoretic formulation that underpins both settings. Section 4 develops the main framework-related theoretical guarantees and the analytic barycenter solution for optimal feature unlearning. Section 6 provides empirical evidence across different data modalities and datasets.

1.2 Related work

Research on unlearning comprises two main threads: *data unlearning* (removing the influence of specific training points) and *feature unlearning* (removing the influence of sensitive attributes). We briefly connect to privacy and fairness where relevant, but the focus of this paper is unlearning.

Data unlearning (machine unlearning). The canonical *exact* unlearning standard requires the released model to match the distribution of a model retrained from scratch on the retain data set (the "anchor" or "retain set"). Systems-oriented approaches realize this exactly via sharded pipelines or statistical caching [12, 11, 29], though often at significant infrastructure cost. To avoid full retraining, recent literature adopts *approximate* guarantees, relaxing the requirement to distributional proximity (e.g., in total variation or KL-divergence) between the unlearned model and the retrained anchor. A diverse array of algorithmic solutions has emerged to meet this criterion, including influence-based updates [32, 60], performance-preserving objectives like PUMA [70], and parameter scrubbing [30]. While these methods offer novel optimization perspectives, framing unlearning as error minimization, targeted repair, or information erasure, they ultimately operate under the shadow of the anchor-based definition. Consequently, verifying their guarantees requires comparing the unlearned model against a stochastic, *unknown* retrained model. This dependency renders these methods difficult to audit or certify [64].

In contrast, our *marginal unlearning* framework breaks this dependency. We propose an inference-based criterion that is checkable solely via observable outputs and the origi-

nal dataset, providing a practical path to auditable unlearning while maintaining rigorous population-level guarantees. In parallel, a companion work Forgetting-MarI [73] applies the marginal unlearning principle to LLMs (Large Language Models), providing large-scale empirical support. This work focuses on providing the mathematical foundations of the principle, such as the sufficiency and necessity guarantees.

Feature unlearning. Feature unlearning removes a sensitive attribute's information from predictions or representations, via adversarial or information-theoretic training [68, 33]. This connects to outcome-independence in fair ML, such as *statistical parity* [24]. Optimal-transport formulations characterize (under regularity) optimally fair predictors/representations via Wasserstein barycenters [17, 31, 72]. In computer vision, related *concept removal* methods aim to erase targeted concepts from generative or recognition models [28].

Information-theoretic formulations. Classical information-bottleneck-style compression and rate-distortion frameworks target *all* information about a point or attribute (e.g., minimizing $I(X;\hat{Y})$) [67, 33]. Our formulation instead targets the *marginal* effect of adding/removing the point via an information-theoretic regularizer on an auxiliary pair (S', Z) defined in Section 3.2. Because our probabilistic guarantee (Definition 2.3) is inference-based and does not rely on an unknown retraining oracle, it yields practical, auditable conditions while remaining compatible with diverse utilities.

Data & AI privacy Data-point unlearning is a post hoc complement to preventative privacy techniques such as anonymization and *differential privacy* (DP) [22, 23]. Privacy protection techniques, such as anonymization (e.g., k-anonymity) and DP, offer training-time protection by bounding the effect of any single record on learned parameters (e.g., DP-SGD) [1] or learned outcomes. In contrast, unlearning provides *remediation* after training, when sensitive data were already incorporated or when tail-event leakages are discovered via auditing (e.g., membership inference and data extraction [62, 76, 13]). Thus, unlearning *complements* DP by enabling targeted removal after deployment, rather than replacing it.

ML & Algorithmic fairness. Feature unlearning aligns with *statistical parity* (outcome independence from a protected attribute) [24]. Related notions include *equalized odds* (conditional independence given the true label) [34] and counterfactual fairness, but here we focus on *unconditional* feature unlearning; conditional variants are left to future work.

Fine-tuning & robustness Unlearning complements fine-tuning in improving model reliability and longevity: fine-tuning *adds/adapts* knowledge (full or parameter-efficient updates such as adapters, prompt tuning, LoRA/QLoRA) [39, 55, 46, 40, 21], while unlearning *removes* misinformation, outdated facts, or undesired influences without full retraining. In practice, combining fine-tuning with unlearning allows AI models to stay current and responsible without rebuilding from scratch. This perspective is complementary (orthogonal) to continual-learning methods that mitigate *catastrophic forgetting* during *addition* of tasks [43]. Here, the goal is principled *removal*.

Moreover, the proposed marginal-information idea can guide robust fine-tuning: update only on the *unique* signal in new data (i.e., the information contributed beyond what is already captured by the retain set). By focusing on this marginal information, the fine-tuned model is less prone to over-training and catastrophic interference with existing knowledge.

Conceptually, our marginal approach parallels the residual learning in ResNet [35]: just as residuals allow a model to learn only "what is new," rather than destabilizing what has already been learned.

Memory suppression studies in cognitive psychology and neuroscience. The proposed marginal-information unlearning principle aligns with established mechanisms in cognitive psychology and neuroscience that implement a two-phase "blur + reinforce" process to reduce the diagnostic influence of specific memories while preserving alternatives. In the *Think/No-Think* paradigm, repeated suppression of cue-elicited retrieval engages inhibitory control and reliably diminishes later accessibility, effectively lowering the discriminability of responses that would otherwise reveal the suppressed trace [6, 7]. *Directed forgetting* (list/item methods) achieves a comparable reduction via context change and selective rehearsal, making earlier material less accessible at test and thus functionally "blurring" its contribution to observed behavior [58]. *Reconsolidation-update* procedures (retrieval–extinction) transiently destabilize reactivated memories and incorporate non-reinforcing information, attenuating the unique predictive impact of the original association [59, 54, 75]. Complementarily, *retrieval practice* strengthens desired knowledge and improves later performance on targeted content [42]. Viewed through our lens, these blur mechanisms reduce the "membership signal", analogous to lowering $I(S';Z)$, while reinforcement preserves task utility, providing intuitive and empirical support for our auditable, information-theoretic formulation of unlearning.

1.3 Tools and notation

We collect the basic notation used throughout the paper.

Probability. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ a measurable state space. A (measurable) random variable $X : \Omega \to \mathcal{X}$ has *law* (distribution) $\mathcal{L}(X) := \mathbb{P} \circ X^{-1}$, i.e., $\mathcal{L}(X)(A) = \mathbb{P}(X \in A)$ for all $A \in \mathcal{B}_{\mathcal{X}}$. For random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$, we write $X \perp Y$ if

$$\mathbb{P}(X \in A,\, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B) \quad \text{for all } A \in \mathcal{B}_{\mathcal{X}},\ B \in \mathcal{B}_{\mathcal{Y}}.$$

For information-theoretic quantities, $H(\cdot)$ denotes (Shannon) entropy, $I(\cdot;\cdot)$ mutual information (MI), and $D_{\mathrm{KL}}(\cdot\|\cdot)$ Kullback–Leibler (KL) divergence. If X is discrete with pmf p_X, then $H(X) = -\sum_x p_X(x) \log p_X(x)$; for absolutely continuous X with pdf p_X (w.r.t. a reference measure μ), the differential entropy is $h(X) = -\int p_X(x) \log p_X(x)\, \mathrm{d}\mu(x)$. For probability measures P, Q on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$,

$$D_{\mathrm{KL}}(P\|Q) = \begin{cases} \displaystyle\int \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\, \mathrm{d}P, & P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Given a pair (X, Y) with joint law $\mathbb{P}_{XY}$ and marginals $\mathbb{P}_X, \mathbb{P}_Y$,

$$I(X;Y) = D_{\mathrm{KL}}\big(\mathbb{P}_{XY} \,\|\, \mathbb{P}_X \otimes \mathbb{P}_Y\big),$$

and, in the discrete case, $I(X;Y) = H(X) + H(Y) - H(X,Y)$. Unless stated otherwise, logarithms are natural, so information quantities are measured in nats. The total variation

distance is defined as

$$\mathrm{TV}(P,Q) := \sup_{A\in\mathcal{B}_{\mathcal{X}}} |P(A)-Q(A)| = \tfrac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|.$$

Optimal transport. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a Polish metric space and $\mathcal{P}(\mathcal{X})$ the set of Borel probability measures on $\mathcal{X}$. For $\mu,\nu\in\mathcal{P}(\mathcal{X})$, their p-Wasserstein distance is [66]

$$\mathcal{W}_{d_{\mathcal{X}},p}(\mu,\nu) := \left(\inf_{\lambda\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,x')^p \,\mathrm{d}\lambda(x,x')\right)^{1/p},$$

where $\Pi(\mu,\nu)$ is the set of couplings with $\lambda(\cdot,\mathcal{X})=\mu$ and $\lambda(\mathcal{X},\cdot)=\nu$. We write $\mathcal{P}_2(\mathcal{X})$ for Borel probability measures with finite second moment, and $\mathcal{P}_{2,ac}(\mathcal{X})\subset\mathcal{P}_2(\mathcal{X})$ for measures that are absolutely continuous (w.r.t. Lebesgue when $\mathcal{X}\subset\mathbb{R}^d$). For random variables X_1, X_2 taking values in $\mathcal{X}$, we use the shorthand $\mathcal{W}_{d_{\mathcal{X}},p}(X_1,X_2) := \mathcal{W}_{d_{\mathcal{X}},p}(\mathcal{L}(X_1),\mathcal{L}(X_2))$. When the metric is clear, we write $\mathcal{W}_p$; in particular, $\mathcal{W}_2 := \mathcal{W}_{d_{\mathcal{X}},2}$.

Wasserstein barycenters. Given a family $\{\mu_z\}_{z\in\mathcal{Z}}\subset\mathcal{P}_{2,ac}(\mathcal{X})$ and weights $\lambda\in\mathcal{P}(\mathcal{Z})$, a (squared) $\mathcal{W}_2$ barycenter is any solution [66]

$$\bar{\mu} \;\in\; \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathcal{X})} \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z,\mu)\,\mathrm{d}\lambda(z).$$

When Brenier maps $T_z : \mathcal{X}\to\mathcal{X}$ from μ_z to $\bar{\mu}$ exist (e.g., under absolute continuity and quadratic cost), one may represent a barycentric random variable $\bar{X}\sim\bar{\mu}$ as $\bar{X} = T_z(X_z)$ with $X_z\sim\mu_z$. Additional details and sufficient conditions (existence/uniqueness, regularity) are deferred to Appendix D.1.

2. Unlearning Principles, Definitions, & Justifications

In this section, we present the unlearning principles, formal definitions, and provable guarantees for both feature and data unlearning. Section 2.1 states the feature-unlearning principle, gives its formal definition, and motivates its information-theoretic regularization. Section 2.2 introduces the proposed marginal-unlearning principle for data-point unlearning, together with its formal and information-theoretic definitions. Section 2.3 establishes *sufficiency*: when the retrain-on-retain model obtains small utility loss, marginal unlearning implies existing approximate unlearning guarantees. Section 2.4 establishes a complementary *necessity* statement: a well-regularized retrain-on-retain model cannot exhibit detectable inclusion leakage while satisfying marginal unlearning.

For clarity and intuition, we state the unlearning principles, definitions, and justifications in the canonical prediction setting where the outcome space coincides with the target space, i.e., $\mathcal{S}=\mathcal{Y}$ and $S=\hat{Y}$. The same principles apply to the generic setting $\mathcal{S}\in\{\mathcal{X},\mathcal{V},\mathcal{Y}\}$, where $\mathcal{X}$ is the input data space, as in synthetic data generation or self-supervised learning with $\mathcal{Y}=\mathcal{X}$; $\mathcal{V}$ is a latent representation space in the representation learning; and $\mathcal{Y}$ is the target space in supervised learning.

2.1 Feature unlearning

Feature unlearning aims to erase the influence of a designated *feature* Z (e.g., a protected attribute, a spurious concept, or a source/domain indicator) from what the model *exposes*, while preserving task-relevant information on the retain distribution p^r. In this setting, Z is observed and available to the unlearning procedure or auditor, but the released output should not reveal Z. Because exposure occurs through the released predictor's output, a natural target is the *output law* of $\hat{Y} = f_\theta(X, Z)$. This suggests the following **feature unlearning principle:**

> *No observer should be able to infer anything about Z from the model's outputs beyond prior knowledge.*

We now formalize this principle. Let $(X, Y, Z) \sim p^r$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, with released output $\hat{Y} = f_\theta(X, Z) \in \hat{\mathcal{Y}}$, where f_θ is the ML/AI model parameterized by model parameter θ. Define the lifted retain loss

$$L(f_\theta) := \mathcal{C}(Y; \hat{Y}, Z), \qquad L^\star := \inf_\theta L(f_\theta).$$

We quantify exposure of Z via the mutual information $I(\hat{Y}; Z)$ under p^r.

Here, we evaluate utility within the observed Z-slices through losses of the form

$$\mathcal{C}(Y; \hat{Y}, Z) = \mathbb{E}_Z\Big[\mathcal{C}_z(Y; \hat{Y})\Big].$$

This lifted loss isolates the utility preservation from the independence constraint to avoid unnecessary conflict between the two, by guiding how the released output $\hat{Y}$ is deformed within each Z-slice, while the exposure penalty $I(\hat{Y}; Z)$ controls how much $\hat{Y}$ reveals about Z after the slices are mixed. For information-based utilities, this interpretation becomes exact: For a general target Y, the non-lifted information utility satisfies

$$I(Y; \hat{Y}) = I(Y; \hat{Y} \mid Z) + I(Z; \hat{Y}) - I(Z; \hat{Y} \mid Y).$$

Thus, when Y is statistically dependent on Z, the non-lifted utility can partially reward the same Z-dependence that the exposure penalty $I(\hat{Y}; Z)$ is designed to remove. More specifically, using equation (1), we have

$$-(1-\lambda)I(Y; \hat{Y}) + \lambda I(\hat{Y}; Z) = (2\lambda - 1)I(\hat{Y}; Z) + (1-\lambda)I(Z; \hat{Y} \mid Y) - (1-\lambda)I(Y; \hat{Y} \mid Z).$$

The first two terms quantify marginal and conditional dependence involving Z, whereas the third term is the within-slice task information. Therefore, using $I(Y; \hat{Y})$ would result in a mixture of utility and (conditional) dependence. In contrast,

$$I(Y; (\hat{Y}, Z)) = I(Y; Z) + I(Y; \hat{Y} \mid Z),$$

and the first term is fixed with respect to the learned representation. Hence, the lifted utility isolates the within-Z-slice information $I(Y; \hat{Y} \mid Z)$, which serves as the genuine utility guidance for deforming $\hat{Y}$ under the exposure constraint $I(\hat{Y}; Z)$.

Now, we provide a formal definition of feature unlearning.

Definition 2.1 (Feature unlearning (independence)) *We say that* $\hat{Y}$ unlearns *the feature* Z *(with respect to* p^r*) if* $\hat{Y} \perp Z$*:*

$$\mathbb{P}\big(\hat{Y} \in A,\, Z \in B\big) \;=\; \mathbb{P}(\hat{Y} \in A)\,\mathbb{P}(Z \in B) \quad \textit{for all measurable events } A \in \mathcal{B}_{\hat{\mathcal{Y}}},\; B \in \mathcal{B}_{\mathcal{Z}}.$$

Equivalently, $I(\hat{Y}; Z) = 0$*.*

The exact independence (or zero mutual information) requirement can be relaxed by an information-theoretic tolerance while enforcing utility.

Definition 2.2 (Mutual-information feature unlearning with utility budget) *Given tolerances* $(\varepsilon, \delta) \in [0, \infty)^2$*, we say that* f_θ *achieves* (ε, δ)-feature unlearning of Z (w.r.t. p^r) *if*

$$I(\hat{Y}; Z) \;\le\; \varepsilon \qquad \textit{and} \qquad L(f_\theta) - L^\star \;\le\; \delta. \tag{2}$$

When $\varepsilon = 0$ *we have* exact feature unlearning*:* $\hat{Y} \perp Z$*.*

The quantity $I(\hat{Y}; Z)$ rigorously captures the no-inference feature unlearning principle above. Let $\pi := \mathbb{P}(\{Z{=}1\})$ and let P_{acc} be the Bayes (optimal) accuracy for predicting Z from $\hat{Y}$. By the binary Fano inequality (see [18]), the Bayes error $P_e := 1 - P_{\text{acc}}$ satisfies

$$H_2(P_e) \;\ge\; H_2(\pi) - I(\hat{Y}; Z), \quad \text{so} \quad P_{\text{acc}} \;\le\; 1 - H_2^{-1}\big(H_2(\pi) - I(\hat{Y}; Z)\big), \tag{3}$$

with H_2 denoting binary entropy and H_2^{-1} understood on $[0, \frac{1}{2}]$. Thus, imposing $I(\hat{Y}; Z) \le \varepsilon$ quantitatively limits the best-possible inference of Z from the released output $\hat{Y}$. Moreover, for any (possibly randomized) measurable post-processing T, the data-processing inequality yields $I(T(\hat{Y}); Z) \le I(\hat{Y}; Z)$: downstream transformations cannot *increase* Z-dependence. Finally, the utility budget δ in (2) rules out trivial "unlearn-by-collapse" solutions (e.g., constant predictors).

2.2 Data unlearning

Data unlearning (often equated with "machine unlearning" in the existing literature) seeks to remove the influence of a designated subset of training data (the *unlearn set*) on a learned predictor, so that an observer cannot detect whether those records participated in training.

A straightforward baseline is the model obtained by fully retraining on the *retain* set alone, which yields the **anchor-based unlearning principle**:

> *No observer should be able to distinguish the released model from the model that would have been obtained by training* from scratch *on the retain set alone.*

While intuitive, the "retrain-on-retain" heuristic depends on a retrained anchor model that is either unknown or not auditable [64]. In particular, if the anchor-based requirement implicitly allows fresh internal randomness on each verification attempt, and there is no mechanism that binds the curator to the original randomness and training procedure, then a curator can, at least in principle, synthesize or sample outputs that pass an indistinguishability check without performing the intended deletion. In effect, the definition presupposes an unlearned anchor and merely shifts the burden to verifying that the anchor is

unlearned. Moreover, existing verification schemes either rely on auxiliary instrumentation (e.g., seeds, logs, gradient commitments, hash attestations) or are brittle to minor pipeline changes, making them easy to spoof and difficult to deploy at scale [78, 63, 69, 14]; see Appendix A.2. These limitations motivate an auditable criterion expressed solely in terms of observable outputs.

To address the lack of auditability above, we propose a novel marginal invariance perspective on the unique information of the unlearn dataset in the model output, which is named **marginal unlearning principle**:

> *No observer should be able to infer anything about the marginal information contributed by the unlearn set (beyond the retain set) from the model's outputs, beyond prior knowledge.*

Here, *marginal information* denotes the unique information (or marginal effect) on the model output contributed by the unlearn set, in addition to the output information already contributed by the retain set. Informally, *marginal unlearning* requires that the model output be *invariant* to whether training examples were drawn purely from p^r (retain) or from a mixture that also includes the unlearn source. This invariance is *auditable* from output. A separate *utility* constraint preserves performance on p^r.

Remark 2.1 (Human analogy: direct vs. marginal unlearning) *The principle aligns with evidence from cognitive psychology and neuroscience. Consider helping a person forget a text they once read. Existing* ***direct (anchor-based) unlearning*** *demands an unverifiable counterfactual: the subject should behave as if they had* never *encountered the text. By contrast,* ***marginal unlearning*** *follows a two-phase* blur + reinforce *mechanism:*

- Blur *(auditable invariance): suppress the diagnostic signal of the to-be-forgotten content so responses to prompts/tests from the "retain" vs. "retain+unlearn" pools become statistically indistinguishable. In our framework, this corresponds to the regularizer in Eq.* (1) *and is formalized by Def. 2.3. In the cognitive psychology and neuroscience literature, it corresponds to the memory destabilization step such as Think/No-Think [6] or reconsolidation update [54].*
- Reinforce *(utility preservation): strengthen desired knowledge so useful behavior is maintained. In our setup this is the first term in Eq.* (1) *(learning on the retain set). Behaviorally, targeted retrieval practice robustly enhances retention and transfer [42].*

We quantify marginal information by measuring the distinguishability between the pure retain distribution p^r and the mixture distribution $p^d := (1-\alpha)p^r + \alpha p^u$, where p^u is the distribution of the unlearn set and $\alpha \in (0, 1]$ controls the mixing ratio and should reflect the cardinality ratio between the retain and unlearn sets. The marginal effect is substantial only when p^d deviates significantly from p^r, indicating that the unlearning source p^u contributes a significant amount of unique information not already represented in the retain set. We formalize this distinction as a binary inference problem by introducing a latent variable $Z \sim \text{Bernoulli}(\pi)$, $\pi \in (0, 1)$, that governs the sampling source. Conditional on Z, define X_{margin} as follows:

$$\mathcal{L}(X_{\text{margin}} \mid \{Z{=}1\}) = p^r, \qquad \mathcal{L}(X_{\text{margin}} \mid \{Z{=}0\}) = p^d.$$

Let $\hat{Y}_{\text{margin}} := f(X_{\text{margin}})$ denote the model output. If the released mechanism uses internal randomness, then $\hat{Y}_{\text{margin}}$ denotes the output law after averaging over that internal randomness; the randomness is not treated as revealed to the observer. The pair $(\hat{Y}_{\text{margin}}, Z)$ thus encodes the model's behavior under the baseline $(Z = 1)$ versus the shifted $(Z = 0)$ distribution. Therefore, the natural quantification of marginal information is the ability to infer the inclusion label Z from the observed output $\hat{Y}_{\text{margin}}$:

Definition 2.3 (ε-marginal unlearning) *Assume the balanced auditing prior* $\mathbb{P}(Z = 0) = \mathbb{P}(Z = 1) = \frac{1}{2}$. *A model f satisfies* ε-marginal unlearning *if*

$$\sup_{D \in \mathcal{B}_{\hat{\mathcal{Y}}}} \left| \log \left(\frac{\mathbb{P}\Big(Z = 0 \,\Big|\, \hat{Y}_{\text{margin}} \in D\Big)}{\mathbb{P}\Big(Z = 1 \,\Big|\, \hat{Y}_{\text{margin}} \in D\Big)} \right) \right| \leq \varepsilon, \tag{4}$$

where the supremum is taken over measurable events D for which the posterior probabilities are well-defined. Equivalently, for every such event D, the posterior odds of "excluded" versus "included" are bounded by $e^{\pm\varepsilon}$ given $\{\hat{Y}_{\text{margin}} \in D\}$.

More generally, if the auditing prior is $\pi := \mathbb{P}(Z = 1) \in (0, 1)$, one can use the prior-normalized posterior odds

$$\frac{\mathbb{P}\Big(Z = 0 \,\Big|\, \hat{Y}_{\text{margin}} \in D\Big) / \mathbb{P}(Z = 0)}{\mathbb{P}\Big(Z = 1 \,\Big|\, \hat{Y}_{\text{margin}} \in D\Big) / \mathbb{P}(Z = 1)}.$$

The corresponding generalized ε-marginal-unlearning condition is

$$\sup_{D \in \mathcal{B}_{\hat{\mathcal{Y}}}} \left| \log \left(\frac{\mathbb{P}\Big(Z = 0 \,\Big|\, \hat{Y}_{\text{margin}} \in D\Big) / \mathbb{P}(Z = 0)}{\mathbb{P}\Big(Z = 1 \,\Big|\, \hat{Y}_{\text{margin}} \in D\Big) / \mathbb{P}(Z = 1)} \right) \right| \leq \varepsilon,$$

again over events for which the conditional probabilities are well-defined. This reduces to Definition 2.3 when $\pi = \frac{1}{2}$.

Definition 2.3 is a worst-case posterior-odds guarantee across all measurable events in the learning-output space. The balanced prior $\pi = \frac{1}{2}$ reflects the default auditing setting in which the observer has no prior preference between the retain-only source and the retain-plus-unlearn source. Under this prior, posterior odds coincide with likelihood ratios, and 0-marginal unlearning is equivalent to

$$\hat{Y}_{\text{margin}} \perp Z,$$

or equivalently

$$I(\hat{Y}_{\text{margin}}; Z) = 0.$$

Furthermore, inequalities such as (3) show that larger $I(\hat{Y}_{\text{margin}}; Z)$ allows more accurate inference of the source label Z, and hence greater distinguishability between the output laws induced by the retain-only and retain-plus-unlearn distributions. Therefore, $I(\hat{Y}_{\text{margin}}; Z)$ provides a natural quantification of marginal information and motivates the following information-theoretic relaxation of Definition 2.3.

Definition 2.4 (High-probability marginal unlearning) *Assume the balanced auditing prior* $\mathbb{P}(Z=0)=\mathbb{P}(Z=1)=\frac{1}{2}$. *A model* f *satisfies* high-probability (ε,δ)-marginal unlearning *if there exists a measurable good set* $G\in\mathcal{B}_{\hat{\mathcal{Y}}}$ *such that*

$$\mathbb{P}(\hat{Y}_{\text{margin}}\in G)\geq 1-\delta,$$

and

$$\sup_{\substack{D\in\mathcal{B}_{\hat{\mathcal{Y}}}\\ D\subseteq G}}\left|\log\left(\frac{\mathbb{P}\Big(Z=0\,\Big|\,\hat{Y}_{\text{margin}}\in D\Big)}{\mathbb{P}\Big(Z=1\,\Big|\,\hat{Y}_{\text{margin}}\in D\Big)}\right)\right|\leq\varepsilon, \tag{5}$$

where the supremum is taken over measurable events $D\subseteq G$ *for which the posterior probabilities are well-defined.*

Definition 2.4 differs from Definition 2.3 only by allowing a small exceptional set G^c. On the good set G, the guarantee remains worst-case over all measurable events $D\subseteq G$. The high-probability definition is naturally controlled by mutual information. Indeed, if $I(\hat{Y}_{\text{margin}};Z)\leq\mu$, then as shown in Lemma 4.1 below, for every $\varepsilon>0$, the good set

$$G_\varepsilon:=\left\{\hat{y}:\left|\log\frac{\mathbb{P}(Z=0\mid\hat{Y}_{\text{margin}}=\hat{y})}{\mathbb{P}(Z=1\mid\hat{Y}_{\text{margin}}=\hat{y})}\right|\leq\varepsilon\right\}$$

satisfies

$$\mathbb{P}(\hat{Y}_{\text{margin}}\in G_\varepsilon)\geq 1-\frac{\sqrt{2\mu}}{\tanh(\varepsilon/2)}.$$

Consequently, mutual information provides an implementation-friendly information-theoretic relaxation of the high-probability marginal-unlearning definition.

Definition 2.5 (Mutual-information marginal unlearning with utility budget) *For tolerances* $\mu,\rho\geq 0$, *with retain risk* $L(f):=\mathbb{E}_{(X,Y)\sim p^r}[\ell(f(X),Y)]$ *and* $L^\star:=\inf_h L(h)$, *we say* f *achieves* MI-regularized (μ,ρ)-marginal unlearning *if*

$$I\big(\hat{Y}_{\text{margin}};Z\big)\leq\mu,\qquad\text{and}\qquad L(f)-L^\star\leq\rho. \tag{6}$$

When $\mu=\rho=0$, *we obtain* exact marginal unlearning with retain optimality*:* $\hat{Y}_{\text{margin}}\perp Z$ *and* f *is retain-optimal.*

The above definition is implementation friendly and scalable in practice, hence particularly applicable in deep learning. See Section 6 for its direct implementation on tabular data analysis and image classification. Also, see [73] for its implementation on LLMs.

We now reflect on the main motivations behind the definition:

- *Interpretability/practicality:* The posterior-odds form (4) gives a worst-case event-wise indistinguishability guarantee, while the high-probability form (5) allows a small exceptional set of outputs. The mutual-information criterion (6) aggregates leakage across outputs and yields the high-probability guarantee in Lemma 4.1.

- *Auditability:* Both depend only on samples of $(X_{\text{margin}}, Z, \hat{Y}_{\text{margin}})$; no anchor or training transcript is needed.
- *Sufficiency:* As shown in Section 2.3 below, with a utility guarantee, marginal unlearning implies the usual anchor-based guarantees (Theorem 2.1).
- *Necessity for "robust" models:* In Section 2.4, we show that a well-generalizing retrained model cannot leak inclusion while also satisfying marginal unlearning without sacrificing regularity (Lemma 2.2).

2.3 Sufficiency for existing definitions: an auditable criterion

Classical *anchor-based* unlearning reads: *after removing X_u, the released model should behave as if it were retrained on $X_r := X \setminus \{X_u\}$.* Formally, let $\mathcal{H}$ be the model/hypothesis space. A (possibly randomized) training algorithm A maps a dataset D to a random model $\Theta \in \mathcal{H}$ with law $A(D)$. Given a delete set $U \subseteq D$, the *anchor (retrain)* distribution is

$$\Theta_r \sim \mathcal{L}\big(A(D \setminus U)\big) \text{ or sometimes } \sim \mathcal{L}\big(M(A(D \setminus U), D \setminus U, \emptyset))\big).$$

An unlearning mechanism M takes (Θ, D, U) and outputs an updated (random) model

$$\Theta_u \sim \mathcal{L}\big(M(A(D), D, U)\big).$$

The existing data (machine) unlearning definitions can be summarized into the following three categories:

- *(Anchored) exact unlearning:* (A, M) is said to satisfy (anchored) *exact unlearning* on (S, U) if $\mathcal{L}(\Theta_u) = \mathcal{L}(\Theta_r)$ as probability measures on $\mathcal{H}$ [12, 11];
- *Divergence relaxation:* (A, M) is said to satisfy (anchored) approximate unlearning if $D\big(\mathcal{L}(\Theta_u), \mathcal{L}(\Theta_r)\big) \leq \varepsilon$ for some divergence D (e.g., total variation, KL, JS)[32];
- *High-probability relaxation:* (A, M) is said to satisfy (anchored) (ε, δ)–indistinguishable unlearning if $\mathbb{P}\big(\Theta_u \in W\big) \leq e^{\varepsilon}\mathbb{P}\big(\Theta_r \in W\big) + \delta$ and $\mathbb{P}\big(\Theta_r \in W\big) \leq e^{\varepsilon}\mathbb{P}\big(\Theta_u \in W\big) + \delta$ for all measurable $W \subseteq \mathcal{H}$ [60].

These criteria hinge on the *unknown* stochastic anchor Θ_r in the model parameter space. Consequently, they are *inauditable* from the released predictor alone. Verification requires privileged access to training artifacts (e.g., checkpoints, witness models) which are fragile, easily spoofed, or unavailable in practical settings [64, 78]. To resolve this, we translate these parameter-space definitions into verifiable *output-space* criteria.

Let p^{test} be a test distribution. We define anchored unlearning on the output level by comparing the predictive laws $\mathcal{L}(f_\Theta(X))$ for $X \sim p^{\text{test}}$, where $f_\Theta(X)$ denotes the (randomized) model output with parameter Θ. Specifically, we adopt the mixture distribution $p^{\text{test}} = p^d$ (the mixture of retain and unlearn sets, which represents the original training data) as the auditing ground. For a divergence D and $\varepsilon \geq 0$, a predictor f_{Θ_u} (model parametrized by Θ_u) satisfies *output-level anchored approximate unlearning* if:

$$D\big(\mathcal{L}(f_{\Theta_r}(X)), \mathcal{L}(f_{\Theta_u}(X))\big) \leq \varepsilon, \quad \text{for } X \sim p^d. \tag{7}$$

We now show that our *marginal unlearning* criterion (Def. 2.3) is sufficient to guarantee Eq. (7), provided the model maintains high utility. We first establish a lemma bounding the output drift via mutual information.

Lemma 2.1 (MI controls output drift) *Let Θ be an $\mathcal{H}$-valued random variable independent of (X_{margin}, Z). Define the marginal output $\hat{Y}_{\text{margin}} := f_\Theta(X_{margin})$, and the model output distribution tested on p^d and p^r respectively by*

$$\mu_\Theta^{p^d} := \mathcal{L}(f_\Theta(X_{margin}) \mid Z = 0), \quad \mu_\Theta^{p^r} := \mathcal{L}(f_\Theta(X_{margin}) \mid Z = 1).$$

Then,

$$\mathrm{TV}\left(\mu_\Theta^{p^d}, \mu_\Theta^{p^r}\right) \;\le\; \sqrt{\frac{I(\hat{Y}_{\text{margin}}; Z)}{2\pi(1-\pi)}}.$$

See Appendix B.1 for a proof.

Using this lemma, we derive the main sufficiency theorem. We utilize the log-loss regret (e.g., cross-entropy in deep learning): for a (possibly randomized) model f, define the (population) log-loss regret under $X \sim p^r$ by

$$\mathrm{reg}_{\log}(f) \;:=\; \mathbb{E}_{X\sim p^r}\Big[\,\mathrm{D}_{\mathrm{KL}}\left(\mathcal{L}(Y \mid X) \,\|\, \mathcal{L}(f(X) \mid X)\right)\Big].$$

For a random f_Θ, write $\overline{\mathrm{reg}}_{\log}(f_\Theta) := \mathbb{E}_\Theta\big(\mathrm{reg}_{\log}(f_\Theta)\big)$.

Theorem 2.1 (Marginal unlearning + utility $\Rightarrow$ approximate unlearning) *Let $\Theta_u \sim \mathcal{L}(M(A(D), D, U))$ be the unlearned model and $\Theta_r \sim \mathcal{L}(A(D \setminus U))$ (or $\mathcal{L}(M(A(D \setminus U), D \setminus U, \varnothing))$) be the retrained anchor. Assume: $\overline{\mathrm{reg}}_{\log}(f_{\Theta_u}) \le \delta$, $\overline{\mathrm{reg}}_{\log}(f_{\Theta_r}) \le \delta_g$ under $X \sim p^r$, and $I(\hat{Y}_{\text{margin}}; Z) \le \varepsilon_u$, where $\hat{Y}_{\text{margin}} := f_{\Theta_u}(X_{\text{margin}})$ and $\Theta_u \perp (X_{\text{margin}}, Z)$. Then,*

$$\mathrm{TV}\Big(\mu_{\Theta_u}^{p^d}, \mu_{\Theta_r}^{p^d}\Big) \;\le\; \underbrace{\sqrt{\tfrac{1}{2}}\left(\sqrt{\delta} + \sqrt{\delta_g}\right)}_{\textit{Utility alignment}} + \underbrace{\sqrt{\tfrac{\varepsilon_u}{2\pi(1-\pi)}}}_{\textit{Membership independence}} + \underbrace{\mathrm{TV}\Big(\mu_{\Theta_r}^{p^r}, \mu_{\Theta_r}^{p^d}\Big)}_{\text{Anchor generalization}}.$$

Here, the second term comes from Lemma 2.1. For the proof see Appendix B.2.

Anchor-based formulations capture the intuition "retrain on the retain set" but are generally not auditable. Marginal unlearning replaces the unknown anchor with a statistically verifiable independence criterion and, when paired with utility control, *recovers* the anchored guarantees.

What can we say about the choice of loss function in this context? Log-loss regret *is* an expected KL divergence, and Pinsker's inequality converts KL to TV on general measurable spaces: $\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(P\|Q)}$. This enables a dimension- and output-space-agnostic bound. Many standard AI/ML objectives instantiate log-loss (classification cross-entropy, GLMs, likelihood-based deep models, CRFs, etc.), making the theorem broadly applicable.

2.4 Necessity for any "robust" retrained model

We now establish a complementary *necessity* statement: a model cannot simultaneously exhibit (1) inclusion leakage, (2) good generalizability, and (3) marginal unlearning. Intuitively, a well-regularized retrained model that fits the remaining data cannot leak inclusion without violating the marginal-unlearning criterion.

Assume equal priors $\mathbb{P}(Z{=}0) = \mathbb{P}(Z{=}1) = \frac{1}{2}$ and that $\mathcal{L}(f(X_0))$ and $\mathcal{L}(f(X_1))$ are mutually absolutely continuous with densities w.r.t. a common reference. Then, for every measurable D with positive probability under both output laws, Bayes' rule gives

$$\log \frac{\mathbb{P}(Z = 0 \mid \hat{Y}_{\text{margin}} \in D)}{\mathbb{P}(Z = 1 \mid \hat{Y}_{\text{margin}} \in D)} = \log \frac{\mathcal{L}(f(X_0))(D)}{\mathcal{L}(f(X_1))(D)}.$$

Hence, Definition 2.3 is equivalent to the pair of set-wise inequalities

$$\mathcal{L}\big(f(X_0)\big)(D) \;\le\; e^{\varepsilon}\,\mathcal{L}\big(f(X_1)\big)(D), \qquad \mathcal{L}\big(f(X_1)\big)(D) \;\le\; e^{\varepsilon}\,\mathcal{L}\big(f(X_0)\big)(D) \quad \forall\, D,$$

which imply mutual absolute continuity. By the Radon–Nikodym theorem, there exists a density ratio $r := \frac{\mathrm{d}\mathcal{L}(f(X_0))}{\mathrm{d}\mathcal{L}(f(X_1))}$ with $e^{-\varepsilon} \le r \le e^{\varepsilon}$ a.s., and conversely this bound implies the set-wise inequalities. Equivalently,

$$\operatorname*{ess\,sup}_y \left|\log \frac{p_1(y)}{p_0(y)}\right| \le \varepsilon \quad \Longleftrightarrow \quad \varepsilon\text{-marginal unlearning},$$

where p_i denotes the density of $\mathcal{L}(f(X_i))$, $i = 0, 1$, with respect to the common reference measure.

Lemma 2.2 (Marginal unlearning as a condition for "good" retraining) *Assume f is a retrained model, $P_i := \mathcal{L}(f(X_i))$, $i \in \{0, 1\}$, admit continuous densities p_i on an open set $\widehat{\mathcal{Y}} \subset \mathbb{R}^m$, f is L-Lipschitz from $(\mathcal{X}, d_{\mathcal{X}})$ to $(\widehat{\mathcal{Y}}, \|\cdot\|)$, and $\mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0) > 0$. Then either*

$$\operatorname*{ess\,sup}_y \left|\log \frac{p_1(y)}{p_0(y)}\right| \le \varepsilon,$$

or there exists a point $y^* \in \widehat{\mathcal{Y}}$, at which the likelihood-ratio bound fails, and a local radius $\rho_*(y^*) > 0$ such that

$$L \ge L^*(\varepsilon, y^*),$$

where

$$L^*(\varepsilon, y^*) := \frac{\gamma(y^*)\,\omega_m\,\rho_*(y^*)^{m+1}}{(m+1)\,\mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0)}, \qquad \gamma(y^*) := \frac{1}{2}|p_1(y^*) - p_0(y^*)| > 0. \tag{8}$$

Here, $X_1 \sim p^r$, $X_0 \sim p^d$, ω_m is the volume of the unit ball in $\mathbb{R}^m$, and $\mathcal{W}_{d_{\mathcal{X}}}$ is the 1-Wasserstein distance on $(\mathcal{X}, d_{\mathcal{X}})$.

The proof can be found in Appendix B.3. The lemma makes precise the sense in which pointwise inclusion leakage becomes detectable at the level of output distributions. The equality $f(x_u) = y_u$ alone is a pointwise statement. However, when x_u is a non-zero-density point of the unlearn-source distribution p^u, every sufficiently small neighborhood

of x_u carries positive p^u-mass. If f is Lipschitz, this local input mass is transported into a controlled neighborhood of $y_u = f(x_u)$. Hence, pointwise inclusion leakage propagates to a local output-level signal. When this signal is not already represented by the retain output law $f_{\#}p^r$, it becomes a detectable density-ratio gap between $f_{\#}p^d$ and $f_{\#}p^r$, contradicting ε-marginal unlearning unless the Lipschitz constant satisfies the lower bound in Lemma 2.2.

Thus, the lemma yields the following incompatibility among:

1. **Inclusion leakage**: the retrained model truthfully reveals the unlearn record, $f(x_u) = y_u$, at a positive-density point of the unlearn-source distribution;
2. **Marginal unlearning**: the output density ratio between $f_{\#}p^d$ and $f_{\#}p^r$ satisfies the ε log-ratio bound of Definition 2.3;
3. **Good regularity/generalizability**: the retrained model has small Lipschitz constant, so pointwise leakage propagates only through controlled local neighborhoods rather than arbitrary discontinuous spikes.

Indeed, under (1) and (3), inclusion leakage propagates to a local output-level signal near y_u. If this signal is specific to the unlearn source, then it produces a density-ratio gap and violates (2). Equivalently, if (1) and (2) both hold, then the model cannot remain regular in the sense quantified by Lemma 2.2.

Consequently, marginal unlearning does not require uniform degradation of model utility. Rather, under regularity, it rules out output behavior that is locally diagnostic of the unlearn source. Thus, the effect of marginal unlearning is concentrated on information unique to the unlearn set and not already supported by the retain distribution, while utility on the retain distribution is preserved through the separate retain optimization objective.

3. Information Theoretic Unlearning Framework

This section formalizes our unlearning framework and explains the rate-distortion, or compression, motivation behind it. We first present the general information-theoretic formulation in Section 3.1. Section 3.2 then state the optimization problems for feature unlearning and marginal data unlearning separately, based on the unlearning definitions introduced in Section 2. Finally, Section 3.3 explains the compression intuition behind the framework through the lens of rate-distortion theory.

3.1 An information-theoretic lens

We connect unlearning to classical rate–distortion and compression ideas [61, 9]. The key quantity is the *mutual information* $I(S'; Z)$ between a chosen exposure or auditing variable S' and the signal to forget Z. We interpret compression as reducing $I(S'; Z)$, thereby compressing away information about Z, while retaining task utility through a user-chosen cost functional. In the unified notation, this gives constrained and Lagrangian formulations of the form

$$\min_f \ \mathcal{C}(Y; S, Z) \quad \text{s.t.} \quad I(S'; Z) \le \tau, \qquad \text{or} \qquad \min_f \ (1-\lambda)\,\mathcal{C}(Y; S, Z) + \lambda\, I(S'; Z).$$

Here $\mathcal{C}(Y; S, Z)$ denotes the lifted utility used in feature unlearning; in marginal data unlearning, where Z is only an auditing label, this specializes to the non-sliced utility $\mathcal{C}(Y; S)$.

To start, we note that the learning outcome space $\mathcal{S}$ is not necessarily the data space $\mathcal{X}$ or the target space $\mathcal{Y}$. More specifically, the outcome space $\mathcal{S}$ can be the data space $\mathcal{X}$ in synthetic data or self-supervised learning (i.e. $\mathcal{Y} = \mathcal{X}$) setting, some new latent or embedding variable space $\mathcal{V}$ in the representation learning setting, or the target space $\mathcal{Y}$ in the supervised learning setting. Therefore, we use $\mathcal{S} \in \{\mathcal{X}, \mathcal{V}, \mathcal{Y}\}$ here to denote a generic learning outcome space.

- **Feature unlearning ($S = S' = f(X, Z)$ for $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{S}$).** Z denotes feature(s) to forget and X the remaining features. We construct $S = f(X, Z)$ whose *outputs* reveal little about Z while preserving utility for the task(s) Y. As explained in Section 2.1, the *feature* influence in model outputs is quantified by $\ell_{\text{feature}}(f) := I(S; Z)$, utility loss is given by $\ell_{\text{utility}}(f) := \mathcal{C}(Y; S, Z)$, yielding the training objective

$$\inf_f (1 - \lambda)\, \ell_{\text{utility}}(f) + \lambda\, \ell_{\text{feature}}(f).$$

- **Marginal data unlearning ($S = f(X)$, $S' = S_{\text{margin}} := f(X_{\text{margin}})$).** Let p^r be the retain distribution and p^u the unlearn source distribution, let $\alpha \in (0, 1]$ be the cardinality or mixture weight of the unlearn source, define $p^d := (1 - \alpha)p^r + \alpha p^u$. Define $(X_{\text{margin}}, Z, S_{\text{margin}})$ as the following: $\mathcal{L}(Z) = \text{Bernoulli}(\frac{1}{2})$,

$$\mathcal{L}(X_{\text{margin}} \mid Z{=}1) = p^r, \qquad \mathcal{L}(X_{\text{margin}} \mid Z{=}0) = p^d,$$

  and $S_{\text{margin}} := f(X_{\text{margin}})$. As explained in Section 2.2, the *marginal* information of p^u in model outputs is quantified by $\ell_{\text{margin}}(f) := I(S_{\text{margin}}; Z)$, utility loss is given by $\ell_{\text{utility}}(f) := \mathcal{C}(Y; S)$, yielding the training objective

$$\inf_f (1 - \lambda)\, \ell_{\text{utility}}(f) + \lambda\, \ell_{\text{margin}}(f).$$

  As shown in Section 2, this formulation not only suffices classical (anchor-based) exact/approximate notions, but also avoids their unauditability by *replacing* an unknown anchor with an observable independence criterion.

3.2 Problem statements

We now state the Pareto-optimal feature-unlearning and marginal-unlearning problems over a generic outcome space $\mathcal{S} \in \{\mathcal{X}, \mathcal{V}, \mathcal{Y}\}$, allowing practitioners to choose the outcome space appropriate for a given application.

Definition 3.1 (Pareto optimal feature unlearning) *Given (X, Z) with Z the feature(s) to forget and X the remaining features, and target Y, solve*

$$\inf_{f:\, \mathcal{X}\times\mathcal{Z}\to\mathcal{S}} \Big\{ \mathcal{C}(Y; S, Z) \;:\; S \perp Z \Big\}, \qquad S := f(X, Z), \tag{9}$$

where $\mathcal{C}$ quantifies the utility loss of using S to predict Y within the observed Z-slices. To trade off utility and forgetting, relax independence via a compression-rate *penalty:*

$$\inf_{f:\, \mathcal{X}\times\mathcal{Z}\to\mathcal{S}} (1 - \lambda)\, \mathcal{C}(Y; S, Z) \;+\; \lambda\, I(S; Z), \qquad \lambda \in [0, 1]. \tag{10}$$

For data unlearning, direct anchor-based definitions are generally not auditable from the released model output alone; see Section 2.2. Section 2.3 shows that marginal unlearning, together with an appropriate utility guarantee, recovers output-level anchored guarantees. This motivates marginal unlearning as a constructive and auditable approach to data unlearning.

Definition 3.2 (Pareto-optimal marginal unlearning) *Let (X_{margin}, Z) be defined as above, let $S_{\text{margin}} := f(X_{\text{margin}})$ for a measurable map $f : \mathcal{X} \to \mathcal{S}$, let $S := f(X)$ denote the released outcome used to evaluate utility, and let Y be a target variable. The optimal marginal unlearning problem is*

$$\inf_{f:\,\mathcal{X}\to\mathcal{S}} \Big\{ \mathcal{C}(Y;S) \;:\; S_{\text{margin}} \perp Z \Big\}, \tag{11}$$

with relaxed form

$$\inf_{f:\,\mathcal{X}\to\mathcal{S}} \; (1-\lambda)\,\mathcal{C}(Y;S) \;+\; \lambda\, I(S_{\text{margin}};Z), \qquad \lambda \in [0,1]. \tag{12}$$

Why mutual information (MI) rather than a one–way KL. The reader might wonder why we did not use the KL divergence between the unlearning information and the retaining information which would reflect the asymmetry of unlearning, rather than mutual information. Let us explore this potential alternative.

That is, we might penalize a directional KL between the "to-unlearn" and "to-retain" output laws. But instead we chose to penalize $I(S_{\text{margin}};Z)$, which equals a *generalized Jensen–Shannon divergence* between the conditional output laws. Writing $P_1 := \mathbb{P}_{S_{\text{margin}}|Z=1}$, $P_0 := \mathbb{P}_{S_{\text{margin}}|Z=0}$, $\pi := \Pr(Z{=}1)$, and $P := \pi P_1 + (1-\pi)P_0$, we obtain

$$I(S_{\text{margin}};Z) = \mathbb{E}_Z\big[D_{\text{KL}}(\mathbb{P}_{S_{\text{margin}}|Z} \,\|\, \mathbb{P}_{S_{\text{margin}}})\big] = \pi\, D_{\text{KL}}(P_1\|P) + (1-\pi)\, D_{\text{KL}}(P_0\|P).$$

Again, $\pi := \Pr(Z{=}1)$ is the prior probability/belief one has before observing anything, which we set to be $\frac{1}{2}$ by default. MI offers three advantages:

- *flexibility*: the reference P adapts online, avoiding commitment to a fixed anchor;
- *stability*: the generalized JS divergence is bounded by $H(Z) \le \log 2$ (binary Z), which typically leads to more stable gradients even with partial support mismatch;
- *guarantee under post-processing*: for any downstream $\hat{Y}_{\text{margin}} = g(S_{\text{margin}})$ for $g : \mathcal{S} \to \mathcal{Y}$, the data-processing inequality gives $I(\hat{Y}_{\text{margin}};Z) \le I(S_{\text{margin}};Z)$, so suppressing MI at exposure controls leakage after any measurable post-processing.

A one-way KL is appropriate only when a *fixed* ideal reference must be matched exactly; it can be unstable when the supports of the distributions are disjoint.

3.3 Interpreting feature unlearning and marginal data unlearning as a generalization of rate-distortion theory

In this subsection, we explain how the proposed unlearning framework in Definitions 3.1 and 3.2 generalizes the classical rate-distortion framework for data compression.

To start, we first specialize the generic outcome space to the data space itself, $\mathcal{S} = \mathcal{X}$, as in synthetic-data design or self-supervised learning. In this setting, we write $S = \hat{X}$. We consider unlearning as compressing $(\hat{X}, Z)$ (or $(\hat{X}_{\text{margin}}, Z)$) so that the output preserves task-relevant information in $\hat{X}$ while removing information about Z. A transparent special case recovers classical rate-distortion. Take $\mathcal{S} = \mathcal{X}$, set the task variable equal to the source, $Y = X$, and take the exposure signal to be the source itself, $Z = X$. Then $S = \hat{X} = f(X, Z) = f(X, X)$ can be written simply as $\hat{X} = f(X)$. If the utility loss is chosen to be an expected distortion,

$$\mathcal{C}(X; \hat{X}, X) := \mathbb{E}_X\Big[\mathbb{E}\big[d(X, \hat{X}) \mid X\big]\Big] = \mathbb{E}[d(X, \hat{X})],$$

then the relaxed objective becomes

$$\inf_{f:\mathcal{X}\to\mathcal{X}} \mathbb{E}[d(X, \hat{X})] + \lambda I(X; \hat{X}), \qquad \hat{X} = f(X), \tag{13}$$

which is the Lagrangian form of the classical rate-distortion problem. The classical rate-distortion formulation optimizes over stochastic kernels $P_{\hat{X}|X}$. Under standard measurable-space assumptions, such kernels can be represented by randomized mechanisms $\hat{X} = f(X, \Theta)$, where Θ is internal randomness independent of X. Thus the deterministic map formulation above is the deterministic subclass, while the stochastic rate-distortion formulation can be recovered by allowing internal randomness and evaluating the induced output law after averaging over Θ.

Here, $I(\hat{X}; X) = H(X) - H(X \mid \hat{X})$ measures preserved information. Since $2^{H(X)}$ and $2^{H(X|\hat{X})}$ are the effective description lengths of X before and after compression, the ratio $2^{I(X;\hat{X})}$ is the expected partition cardinality induced by $\hat{X}$; maximizing it preserves fidelity (Appendix C.1). In classic data compression, the original data set X is considered "unwanted" in data compression because it is often too redundant for the task variable Y.

Therefore, feature unlearning generalizes the classical rate-distortion viewpoint by allowing the unwanted information to be a designated feature, source, or attribute Z, rather than the entire source variable X. The feature Z may be explicitly present in X, or may be provided as an auxiliary variable correlated with X. Marginal data unlearning further generalizes this viewpoint by constructing the auditing pair (X_{margin}, Z), where Z labels whether the input is drawn from the retain-only distribution or the retain-plus-unlearn mixture, and then compressing away the information about this constructed source label.

Admissibility (no artificial information creation). Requiring $S = f(X, Z)$ in feature unlearning or $S = f(X)$ in marginal data unlearning ensures

$$\sigma(S) \subset \sigma(X, Z) \qquad \text{or} \qquad \sigma(S) \subset \sigma(X),$$

respectively. Thus, the released outcome is a measurable transformation of the available information and does not create information ex nihilo. See Appendix C.2 for a detailed explanation.

4. Theoretical Guarantees

This section establishes theoretical guarantees for both *feature unlearning* and *marginal unlearning* as defined in Definitions 3.1 and 3.2. For clarity, we carry out all proofs in the canonical setting where the outcome space is the same as the input data space, i.e., $\mathcal{S} = \mathcal{X}$ and $S = \hat{X}$. This data-space setting also yields leakage guarantees for generic outcome spaces via the data-processing inequality. See Remark 4.1.

For *feature unlearning*, we show that driving the mutual information $I(\hat{X}; Z)$ small forces conditional law $\mathbb{P}(\hat{X} \mid Z = z)$ to concentrate around the mixture $\mathbb{P}(\hat{X})$ in standard statistical distances (KL, TV), thereby suppressing Z–signal in the released output. For *marginal unlearning*, we obtain a probabilistic, *auditable* guarantee by penalizing $I(\hat{X}_{\text{margin}}; Z)$, where Z denotes the latent source label distinguishing the retain-only distribution from the retain-plus-unlearn mixture. Finally, for feature unlearning in the data-space setting, we use optimal transport to derive an analytic lifted solution for particular utility criteria.

Remark 4.1 ($\mathcal{S} = \mathcal{X}$ suffices leakage guarantee in generic setting) *The guarantees we derive for $S = \hat{X}$ immediately extend to the generic case $S \in \{\mathcal{X}, \mathcal{V}, \mathcal{Y}\}$, where $\mathcal{V}$ denotes an intermediate representation space (e.g., an encoder output) and $\mathcal{Y}$ the prediction/output space. Indeed, let $T : \mathcal{X} \to \mathcal{S}'$ be any measurable map representing downstream post-processing (e.g., $T = h : \mathcal{X} \to \mathcal{V}$ or $T = g : \mathcal{X} \to \mathcal{Y}$), and define $S' := T(\hat{X})$.*

- Feature unlearning: *If $I(\hat{X}; Z) \le \varepsilon$, then by the data-processing inequality,*

$$I\big(T(\hat{X}); Z\big) \;\le\; I(\hat{X}; Z) \;\le\; \varepsilon,$$

  so any bound proved for $S = \hat{X}$ propagates to $S' = T(\hat{X}) \in \{\mathcal{V}, \mathcal{Y}\}$.

- Marginal unlearning: *Writing $S'_{\text{margin}} := T(\hat{X}_{\text{margin}})$, the odds bound in Definition 2.3 is preserved under measurable pushforwards: for every measurable $B \subseteq \mathcal{S}'$,*

$$\frac{\mathbb{P}\big(Z = 0 \,\big|\, S'_{\text{margin}} \in B\big)}{\mathbb{P}\big(Z = 1 \,\big|\, S'_{\text{margin}} \in B\big)} \;=\; \frac{\mathbb{P}\big(Z = 0 \,\big|\, \hat{X}_{\text{margin}} \in T^{-1}(B)\big)}{\mathbb{P}\big(Z = 1 \,\big|\, \hat{X}_{\text{margin}} \in T^{-1}(B)\big)},$$

  so the same ε-bound holds with $D := T^{-1}(B)$. Likewise, $I(S'_{\text{margin}}; Z) \le I(\hat{X}_{\text{margin}}; Z)$ by the data-processing inequality.

Consequently, leakage bounds proved for the data-space release $\hat{X}$ automatically apply to any representation or prediction obtained from $\hat{X}$ by measurable post-processing. We adopt the notation $S = \hat{X}$ for the remainder of the section.

4.1 Feature unlearning guarantee

In practice we optimize the relaxed problem

$$\inf_f \; (1 - \lambda)\mathcal{C}(Y; \hat{X}, Z) \;+\; \lambda\, I(\hat{X}; Z),$$

and consider solutions for which $I(\hat{X}; Z) \le \varepsilon$ for some small $\varepsilon > 0$. The next bounds convert this constraint into quantitative control of the deviation between $\mathbb{P}(\hat{X} \mid Z = z)$ and its center $\mathbb{P}(\hat{X})$ in KL and TV. We write $\mathrm{TV}(P, Q) := \frac{1}{2}\int |dP - dQ| = \sup_A |P(A) - Q(A)|$.

Proposition 4.1 (Discrete Z: MI controls conditional–marginal drift) *Let $(\hat{X}, Z)$ be jointly distributed with Z supported on a finite set $\mathcal{Z}$ and $p(z) := \mathbb{P}(Z{=}z) > 0$ for all $z \in \mathcal{Z}$. Then*

$$\mathbb{E}_Z\Big[\, D_{\mathrm{KL}}\big(\mathbb{P}(\hat{X} \mid Z) \,\|\, \mathbb{P}(\hat{X})\big)\,\Big] = I(\hat{X}; Z), \tag{14}$$

$$D_{\mathrm{KL}}\big(\mathbb{P}(\hat{X} \mid Z{=}z) \,\|\, \mathbb{P}(\hat{X})\big) \le \frac{I(\hat{X}; Z)}{p(z)}, \qquad \forall\, z \in \mathcal{Z}, \tag{15}$$

and

$$\mathbb{E}_Z\Big[\, \mathrm{TV}\big(\mathbb{P}(\hat{X} \mid Z), \mathbb{P}(\hat{X})\big)\,\Big] \le \sqrt{\tfrac{1}{2}\, I(\hat{X}; Z)}, \tag{16}$$

$$\mathrm{TV}\big(\mathbb{P}(\hat{X} \mid Z{=}z), \mathbb{P}(\hat{X})\big) \le \sqrt{\tfrac{1}{2\,p(z)}\, I(\hat{X}; Z)}, \qquad \forall\, z \in \mathcal{Z}. \tag{17}$$

Proof Since

$$I(\hat{X}; Z) \;=\; \sum_{z \in \mathcal{Z}} p(z)\, D_{\mathrm{KL}}\big(\,\mathbb{P}(\hat{X} \mid Z{=}z) \,\|\, \mathbb{P}(\hat{X})\big),$$

and Pinsker's inequality gives, for each z,

$$\mathrm{TV}\big(\,\mathbb{P}(\hat{X} \mid Z{=}z),\, \mathbb{P}(\hat{X})\big) \;\le\; \sqrt{\tfrac{1}{2}\, D_{\mathrm{KL}}\big(\,\mathbb{P}(\hat{X} \mid Z{=}z) \,\|\, \mathbb{P}(\hat{X})\big)}.$$

It then follows from Jensen's inequality and concavity of square root that

$$\begin{aligned}\sum_z p(z)\, \mathrm{TV}\big(\mathbb{P}(\hat{X}\,|\,z), \mathbb{P}(\hat{X})\big) \;&\le\; \sum_z p(z) \sqrt{\tfrac{1}{2}\, D_{\mathrm{KL}}(\mathbb{P}(\hat{X}\,|\,z) \| \mathbb{P}(\hat{X}))} \\ &\le\; \sqrt{\tfrac{1}{2} \sum_z p(z)\, D_{\mathrm{KL}}(\mathbb{P}(\hat{X}\,|\,z) \| \mathbb{P}(\hat{X}))},\end{aligned}$$

which establishes the bound in (16).

For the pointwise bound, note that all terms $D_{\mathrm{KL}}(\mathbb{P}(\hat{X}\,|\,z) \| \mathbb{P}(\hat{X}))$ are nonnegative and satisfy

$$p(z)\, D_{\mathrm{KL}}(\mathbb{P}(\hat{X}\,|\,z) \| \mathbb{P}(\hat{X})) \;\le\; \sum_{z'} p(z')\, D_{\mathrm{KL}}(\mathbb{P}(\hat{X}\,|\,z') \| \mathbb{P}(\hat{X})) \;=\; I(\hat{X}; Z),$$

hence $D_{\mathrm{KL}}(\mathbb{P}(\hat{X}\,|\,z) \| \mathbb{P}(\hat{X})) \le I(\hat{X}; Z)/p(z)$. Combining this with Pinsker's inequality gives (17). That completes the proof. ■

The case where Z is not restricted to discrete values is a bit more involved.

Theorem 4.1 (General Z: mutual information controls KL and TV) *Let $(\hat{X}, Z)$ be jointly distributed on standard Borel spaces $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$. Let $\mathbb{P}_{\hat{X}}$ denote the marginal law of $\hat{X}$, and let $\{\mathbb{P}_{\hat{X}|Z=z}\}_{z \in \mathcal{Z}}$ be a regular conditional distribution, defined for $\mathbb{P}_Z$-a.e. z. Assume $I(\hat{X}; Z) < \infty$. Then:*

1. KL expectation identity:

$$\mathbb{E}_Z\Big[\, \mathrm{D_{KL}}\left(\mathbb{P}_{\hat{X}|Z} \| \, \mathbb{P}_{\hat{X}}\right)\Big] = I(\hat{X};Z). \tag{18}$$

2. KL probability tail bound: For all $\tau > 0$,

$$\mathbb{P}\Big(\, \mathrm{D_{KL}}\left(\mathbb{P}_{\hat{X}|Z} \| \, \mathbb{P}_{\hat{X}}\right) \geq \tau\Big) \leq \frac{I(\hat{X};Z)}{\tau}. \tag{19}$$

3. TV expectation bound:

$$\mathbb{E}_Z\Big[\, \mathrm{TV}\big(\mathbb{P}_{\hat{X}|Z}, \, \mathbb{P}_{\hat{X}}\big)\Big] \;\leq\; \sqrt{\frac{1}{2}\, I(\hat{X};Z)}\,. \tag{20}$$

4. TV probability tail bound: For all $\tau > 0$,

$$\mathbb{P}\,\Big(\, \mathrm{TV}\big(\mathbb{P}_{\hat{X}|Z}, \, \mathbb{P}_{\hat{X}}\big) \geq \tau\Big) \;\leq\; \frac{I(\hat{X};Z)}{2\,\tau^2}\,. \tag{21}$$

Proof By disintegration of relative entropy,

$$I(\hat{X};Z) = D_{\mathrm{KL}}\big(\mathbb{P}_{\hat{X},Z} \| \mathbb{P}_{\hat{X}} \otimes \mathbb{P}_Z\big) = \int D_{\mathrm{KL}}\big(\mathbb{P}_{\hat{X}|Z=z} \| \mathbb{P}_{\hat{X}}\big)\, \mathbb{P}_Z(dz).$$

This proves the KL expectation identity. Since $I(\hat{X};Z) < \infty$, the integrand is finite for $\mathbb{P}_Z$-a.e. z.

Step 1 (pointwise Pinsker). For $\mathbb{P}_Z$-a.e. z, Pinsker's inequality yields

$$\mathrm{TV}\big(\mathbb{P}_{\hat{X}|Z=z}, \, \mathbb{P}_{\hat{X}}\big) \;\leq\; \sqrt{\tfrac{1}{2}\, D_{\mathrm{KL}}\big(\mathbb{P}_{\hat{X}|Z=z} \,\|\, \mathbb{P}_{\hat{X}}\big)} < \infty.$$

Step 2 (average TV bound). Taking expectation with respect to Z, it then follows from Jensen's inequality and concavity of the square root function that

$$\mathbb{E}_Z\, \mathrm{TV}\big(\mathbb{P}_{\hat{X}|Z}, \mathbb{P}_{\hat{X}}\big) \;\leq\; \mathbb{E}_Z \sqrt{\tfrac{1}{2}\, D_{\mathrm{KL}}(\mathbb{P}_{\hat{X}|Z} \| \mathbb{P}_{\hat{X}})} \;\leq\; \sqrt{\tfrac{1}{2}\, \mathbb{E}_Z\, D_{\mathrm{KL}}(\mathbb{P}_{\hat{X}|Z} \| \mathbb{P}_{\hat{X}})}.$$

Step 3 (tail bound). The KL tail bound follows directly from Markov's inequality:

$$\mathbb{P}\Big(\mathrm{D_{KL}}(\mathbb{P}_{\hat{X}|Z} \| \mathbb{P}_{\hat{X}}) \geq \tau\Big) \leq \frac{\mathbb{E}_Z\, \mathrm{D_{KL}}(\mathbb{P}_{\hat{X}|Z} \| \mathbb{P}_{\hat{X}})}{\tau} = \frac{I(\hat{X};Z)}{\tau}.$$

From Pinsker inequality we have $\mathrm{TV}(\mathbb{P}_{\hat{X}|Z}, \mathbb{P}_{\hat{X}})^2 \leq \frac{1}{2}\, D_{\mathrm{KL}}(\mathbb{P}_{\hat{X}|Z} \| \mathbb{P}_{\hat{X}})$, $\mathbb{P}_Z$-a.s.. It follows from Markov's inequality that, for any $\tau > 0$,

$$\mathbb{P}\Big(\mathrm{TV}(\mathbb{P}_{\hat{X}|Z}, \mathbb{P}_{\hat{X}}) \geq \tau\Big) = \mathbb{P}\Big(\mathrm{TV}(\mathbb{P}_{\hat{X}|Z}, \mathbb{P}_{\hat{X}})^2 \geq \tau^2\Big) \leq \frac{I(\hat{X};Z)}{2\tau^2},$$

which is (21). That completes the proof. ■

Remark 4.2 (On pointwise bounds without discreteness) *In the discrete case with* $\mathbb{P}(Z = z) = p(z) > 0$ *one has the pointwise estimate*

$$\mathrm{TV}(\mathbb{P}_{\hat{X}|z}, \mathbb{P}_{\hat{X}}) \le \sqrt{I(\hat{X}; Z)/(2p(z))},$$

because $I(\hat{X}; Z) = \sum_z p(z)\, D_{\mathrm{KL}}(\mathbb{P}_{\hat{X}|z} \| \mathbb{P}_{\hat{X}})$ *and each summand is nonnegative. For general (non-discrete)* Z*, no analogous pointwise inequality can hold without additional assumptions (e.g., uniform lower bounds on the density of* Z *over sets of positive reference measure), since an expectation constraint* $\int p_Z(z)\, K(z)\, dz = I$ *does not control the essential sup of* $K(z)$*. Therefore, the average and tail bounds* (20)–(21) *are the natural general theoretical guarantees.*

4.2 Data unlearning guarantee

For data unlearning we operate on the data space and penalize

$$I\big(\hat{X}_{\mathrm{margin}}; Z\big),$$

with $\hat{X}_{\mathrm{margin}} := f(X_{\mathrm{margin}})$, $X_{\mathrm{margin}} \mid (Z{=}1) \sim p^r$, $X_{\mathrm{margin}} \mid (Z{=}0) \sim p^d$, and $Z \sim$ Bernoulli(1/2), as in Section 2.2. That is, we penalize the information that $\hat{X}_{\mathrm{margin}}$ carries about the latent source label Z, which distinguishes the retain-only distribution from the retain-plus-unlearn mixture:

$$\inf_{f:\, \mathcal{X} \to \mathcal{X}} \Big\{ (1-\lambda)\mathcal{C}(Y; \hat{X}) \;+\; \lambda\, I(\hat{X}_{\mathrm{margin}}; Z) \Big\}. \tag{22}$$

The next lemma shows that controlling $I(\hat{X}_{\mathrm{margin}}; Z)$ directly yields an inference guarantee.

Lemma 4.1 (MI controls high-probability posterior odds) *Let* $\hat{X}_{\mathrm{margin}} = f(X_{\mathrm{margin}})$ *defined as above with* $X_0 \sim p^d$*,* $X_1 \sim p^r$*, and* $Z \in \{0, 1\}$ *indicating the source. Then for any* $\eta \in (0, 1)$*,*

$$\mathbb{P}\left(\left| \log \frac{\mathbb{P}\big(Z{=}0 \mid \hat{X}_{\mathrm{margin}}\big)}{\mathbb{P}\big(Z{=}1 \mid \hat{X}_{\mathrm{margin}}\big)} \right| \;\le\; \log \frac{1+\eta}{1-\eta} \right) \;\ge\; 1 \;-\; \frac{2}{\eta} \sqrt{\tfrac{1}{2}\, I\big(\hat{X}_{\mathrm{margin}}; Z\big)}.$$

See Appendix D.2 for the proof. Setting

$$\eta_\varepsilon := \frac{e^\varepsilon - 1}{e^\varepsilon + 1} = \tanh(\varepsilon/2),$$

Lemma 4.1 implies

$$\mathbb{P}\left(\left| \log \frac{\mathbb{P}(Z = 0 \mid \hat{X}_{\mathrm{margin}})}{\mathbb{P}(Z = 1 \mid \hat{X}_{\mathrm{margin}})} \right| \le \varepsilon \right) \ge 1 - \frac{\sqrt{2I(\hat{X}_{\mathrm{margin}}; Z)}}{\tanh(\varepsilon/2)}.$$

Thus, if $I(\hat{X}_{\mathrm{margin}}; Z) = 0$, then $\hat{X}_{\mathrm{margin}} \perp Z$. More generally, if $I(\hat{X}_{\mathrm{margin}}; Z)$ is small, then the set of outputs with posterior odds larger than e^ε has small marginal probability. Equivalently, except on a small-probability set of outputs, observing $\hat{X}_{\mathrm{margin}}$ does not

substantially change the posterior odds of whether the input came from the retain-only distribution or the retain-plus-unlearn mixture.

The following guarantee translates a small "compression rate" into an auditable inference bound for high-probability marginal unlearning as defined in Definition 2.4:

Theorem 4.2 (High-probability marginal unlearning via compression rate) *Let $Z \sim$ Bernoulli(1/2). If $I(\hat{X}_{\text{margin}}; Z) \le \mu < \frac{1}{2}(\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1})^2$, then for every $\varepsilon > 0$, f satisfies high-probability $(\varepsilon, \delta_\varepsilon)$-marginal unlearning in the sense of Definition 2.4, with*

$$\delta_\varepsilon = \frac{\sqrt{2\mu}}{\tanh(\varepsilon/2)} = \frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}\sqrt{2\mu}.$$

Hence, one can select the regularization weight λ to drive the observable $I(\hat{X}_{\text{margin}}; Z)$ below a target μ aligned with a desired $(\varepsilon, \text{confidence})$ pair (ε is the marginal difference allowed in the ε-marginal unlearning definition, confidence is the guarantee probability provided by Theorem 4.2), directly yielding an *auditable* post hoc certificate from samples of $(X_{\text{margin}}, Z, \hat{X}_{\text{margin}})$.

Finally, if one wants to apply the marginal unlearning guarantee together with utility to provide an *anchor-based unlearning* (also known as approximate unlearning) guarantee, Theorem 4.2 together with Theorem 2.1 provide an auditable/tractable guarantee (based on the observed μ) under the assumption of small log-loss for both unlearned model and the anchor (retain from scratch) model.

4.3 An analytic feature unlearning solution independent of the downstream tasks

Here, we show that for specific types of cost functions, we can obtain an analytic feature-unlearning solution using optimal transport and extreme sigma-algebras. In particular, we focus on the following cost functions:

- *Mutual information maximization:* $\mathcal{C}(Y; \hat{X}, Z) = -I(Y; \hat{X}, Z)$ maximizes informativeness about Y.
- *Posterior concentration:* $\mathcal{C}(Y; \hat{X}, Z) = -\mathbb{E}\Big[D_{\text{KL}}\big(\mathbb{P}(Y \mid \hat{X}, Z) \,\|\, \mathbb{P}(Y)\big)\Big]$, which coincides with $-I(Y; \hat{X}, Z)$ provided the KL-divergence is well-defined.
- *Conditional–probability energy:*

$$\mathcal{C}(Y; \hat{X}, Z) = \begin{cases} -\mathbb{E}\Big[\mathbb{P}(Y \in A \mid \hat{X}, Z)^2\Big], & \text{for classification events } A \in \mathcal{B}_{\mathcal{Y}}, \\ -\|\mathbb{E}(Y \mid \hat{X}, Z)\|_{L^2}^2, & \text{for regression with } Y \in L^2. \end{cases}$$

Although these criteria are different, they are all monotone with respect to the lifted sigma-algebra generated by $(\hat{X}, Z)$. Under the mild assumptions below, the Wasserstein barycenter construction produces a canonical representation $\bar{X} \perp Z$ whose lifted sigma-algebra $\sigma(\bar{X}, Z)$ is maximal among admissible lifted representations. Consequently, $\bar{X}$ simultaneously optimizes the lifted monotone utility criteria above, subject to the hard feature-unlearning constraint $\hat{X} \perp Z$.

We start with the following result, which establishes that the Wasserstein-2 barycenter generates the finest sigma-algebra among all admissible outcomes:

Lemma 4.2 (Wasserstein barycenter $\implies$ lifted finest sigma-algebra) *Let $P_z := \mathcal{L}(X \mid Z = z) \in \mathcal{P}_{2,ac}(\mathcal{X})$, and let $\bar{P}$ denote their $\mathcal{W}_2$-barycenter. Let $T_z : P_z \to \bar{P}$ be the Brenier map and define the barycentric representation $\bar{X} := T_Z(X)$. Under the Brenier a.e.-invertibility guaranteed by the absolute-continuity assumptions, one has $\sigma(X, Z) = \sigma(\bar{X}, Z)$ up to null sets. Consequently, for every admissible $\hat{X} = f(X, Z)$,*

$$\sigma(\hat{X}, Z) \subset \sigma(\bar{X}, Z).$$

See Appendix D.3 for a proof. Now, the importance of the above result lies in the monotonicity of the objective functions listed earlier w.r.t. the sigma-algebra generated by random variables. That is, the fineness of the sigma-algebra is equivalent to the amount of information (in a probability-theoretic sense) contained by the random variable:

Lemma 4.3 (Monotonicity w.r.t. sigma–algebra) *If $\sigma(X_1) \subset \sigma(X_2)$, then the following hold whenever the quantities are well-defined:*

- *$H(Y \mid X_2) \leq H(Y \mid X_1)$ for any $Y : \Omega \to \mathcal{Y}$ with finite-entropy in discrete setting or with the absolutely continuous conditional laws $\mathcal{L}(Y \mid X_s)$ and finite conditional differential entropies in continuous setting.*
- *$I(Y; X_1) \leq I(Y; X_2)$ for any $Y : \Omega \to \mathcal{Y}$.*
- *$\left\|\mathbb{P}(Y \in A \mid X_1)\right\|_{L^2}^2 \leq \left\|\mathbb{P}(Y \in A \mid X_2)\right\|_{L^2}^2$; for any $A \in \mathcal{B}_\mathcal{Y}$, any $Y : \Omega \to \mathcal{Y}$.*
- *$\left\|\mathbb{E}(Y \mid X_1)\right\|_{L^2}^2 \leq \left\|\mathbb{E}(Y \mid X_2)\right\|_{L^2}^2$. for any $Y \in L^2$.*

See a proof in Appendix D.5. As a result, the information quantifications discussed above are naturally monotone w.r.t. the fineness of the generated sigma-algebra: See Lemma 4.3 for the formal result. Informally, Lemma 4.3 shows that conditioning on a finer sigma-algebra reduces conditional uncertainty and increases information-based or energy-based measures of predictive informativeness.

Theorem 4.3 (Unified optimal feature unlearning solution) *Under the assumptions of Lemma 4.2, the following are equivalent for any admissible $\hat{X} = f(X, Z)$:*

- *$\sigma((\hat{X}, Z)) = \sigma((\bar{X}, Z))$.*
- *$\hat{X} \in \arg\min\{H(Y \mid \tilde{X}, Z) : \ \tilde{X} \perp Z\}$ for all $Y : \Omega \to \mathcal{Y}$ with finite-entropy in discrete setting or with the absolutely continuous conditional laws $\mathcal{L}(Y \mid X_s)$ and finite conditional differential entropies in continuous setting.*
- *$\hat{X} \in \arg\max\{I(Y; \tilde{X}, Z) : \ \tilde{X} \perp Z\}$ for all $Y : \Omega \to \mathcal{Y}$.*
- *$\hat{X} \in \arg\max\{\|\mathbb{P}(Y \in A \mid \tilde{X}, Z)\|_2^2 : \ \tilde{X} \perp Z\}$ for all $A \in \mathcal{B}_\mathcal{Y}$ and $Y : \Omega \to \mathcal{Y}$.*
- *$\hat{X} \in \arg\max\{\|\mathbb{E}(Y \mid \tilde{X}, Z)\|_2^2 : \ \tilde{X} \perp Z\}$ for all $Y \in L^2(\Omega; \mathcal{Y})$.*

Proof First, by Lemma 4.2, for every admissible $\hat{X} = f(X, Z)$, we have $\sigma(\hat{X}, Z) \subset \sigma(\bar{X}, Z)$. Now, suppose that

$$\sigma(\hat{X}, Z) = \sigma(\bar{X}, Z),$$

then conditioning on $(\hat{X}, Z)$ is the same as conditioning on $(\bar{X}, Z)$. Since $\bar{X} \perp Z$, the barycenter representation is feasible for the hard feature-unlearning constraint. Moreover, by Lemma 4.3, for every feasible admissible $\tilde{X}$, we have

- $H(Y \mid \bar{X}, Z) \le H(Y \mid \tilde{X}, Z)$,
- $I(Y; \tilde{X}, Z) \le I(Y; \bar{X}, Z)$,
- $\|\mathbb{P}(Y \in A \mid \tilde{X}, Z)\|_{L^2}^2 \le \|\mathbb{P}(Y \in A \mid \bar{X}, Z)\|_{L^2}^2$, and
- $\|\mathbb{E}(Y \mid \tilde{X}, Z)\|_{L^2}^2 \le \|\mathbb{E}(Y \mid \bar{X}, Z)\|_{L^2}^2$,

whenever the corresponding quantities are well-defined. Since $\sigma(\hat{X}, Z) = \sigma(\bar{X}, Z)$, the same optimal values are attained by $\hat{X}$. Therefore, $\hat{X}$ is one optimal solution to all the above objectives under the hard independence constraint.

Conversely, suppose that one of the stated optimality conditions holds for $\hat{X}$. We show that $\sigma(\bar{X}, Z) \subset \sigma(\hat{X}, Z)$. More specifically, let $B \in \sigma(\bar{X}, Z)$ be arbitrary, we show that $B \in \sigma(\hat{X}, Z)$. Indeed, since the above four optimality are true for all measurable Y, let $Y = \mathbb{1}_B$.

If the conditional-entropy optimality holds, then because $Y = \mathbb{1}_B$ is $\sigma(\bar{X}, Z)$-measurable, we have $H(Y \mid \bar{X}, Z) = 0$. Since $\bar{X}$ is feasible and $\hat{X}$ is optimal, we also obtain

$$0 \le H(Y \mid \hat{X}, Z) \le H(Y \mid \bar{X}, Z) = 0.$$

Hence, $Y = \mathbb{1}_B$ is $\sigma(\hat{X}, Z)$-measurable, and therefore $B \in \sigma(\hat{X}, Z)$.

If the mutual-information optimality holds, take again $Y = \mathbb{1}_B$, where $B \in \sigma(\bar{X}, Z)$. Since Y is $\sigma(\bar{X}, Z)$-measurable, we have $I(Y; \bar{X}, Z) = H(Y)$. Because $\bar{X}$ is feasible and $\hat{X}$ is optimal for this target,

$$I(Y; \hat{X}, Z) \ge I(Y; \bar{X}, Z) = H(Y).$$

On the other hand, $I(Y; \hat{X}, Z) \le H(Y)$. Hence, $I(Y; \hat{X}, Z) = H(Y)$, or equivalently,

$$H(Y \mid \hat{X}, Z) = 0.$$

Therefore $Y = \mathbb{1}_B$ is $\sigma(\hat{X}, Z)$-measurable, and so $B \in \sigma(\hat{X}, Z)$.

If the conditional-probability-energy optimality holds, take $A = \{1\}$. Since $Y = \mathbb{1}_B$ is $\sigma(\bar{X}, Z)$-measurable, $\mathbb{P}(Y \in A \mid \bar{X}, Z) = \mathbb{1}_B$. Hence,

$$\|\mathbb{P}(Y \in A \mid \bar{X}, Z)\|_{L^2}^2 = \mathbb{P}(B).$$

By Lemma 4.3,

$$\|\mathbb{P}(Y \in A \mid \hat{X}, Z)\|_{L^2}^2 \le \mathbb{P}(B).$$

On the other hand, since $\bar{X} \perp Z$, the barycenter representation is feasible. Therefore, optimality of $\hat{X}$ for the conditional-probability energy implies

$$\|\mathbb{P}(Y \in A \mid \hat{X}, Z)\|_{L^2}^2 \ge \|\mathbb{P}(Y \in A \mid \bar{X}, Z)\|_{L^2}^2 = \mathbb{P}(B).$$

Thus, the equality holds. Let

$$q_B := \mathbb{P}(B \mid \hat{X}, Z).$$

Then $q_B \in [0, 1]$ satisfies

$$\mathbb{E}[q_B] = \mathbb{P}(B), \qquad \mathbb{E}[q_B^2] = \mathbb{P}(B).$$

Therefore,

$$\mathbb{E}[q_B(1-q_B)] = 0.$$

Since $q_B(1-q_B) \geq 0$, we have $q_B \in \{0,1\}$ a.s. Moreover, because $q_B = \mathbb{E}(\mathbb{1}_B \mid \hat{X}, Z)$,

$$\begin{aligned}\mathbb{E}[(\mathbb{1}_B - q_B)^2] &= \mathbb{P}(B) - 2\mathbb{E}[\mathbb{1}_B q_B] + \mathbb{E}[q_B^2] \\ &= \mathbb{P}(B) - 2\mathbb{E}[q_B^2] + \mathbb{E}[q_B^2] \\ &= \mathbb{P}(B) - \mathbb{E}[q_B^2] = 0.\end{aligned}$$

Thus, $\mathbb{1}_B = q_B$ a.s.. Since q_B is $\sigma(\hat{X}, Z)$-measurable, $B \in \sigma(\hat{X}, Z)$.

Finally, if the conditional-mean-energy optimality holds, take again $Y = \mathbb{1}_B \in L^2$. Since $B \in \sigma(\bar{X}, Z)$,

$$\mathbb{E}(Y \mid \bar{X}, Z) = \mathbb{1}_B,$$

and hence $\|\mathbb{E}(Y \mid \bar{X}, Z)\|_{L^2}^2 = \mathbb{P}(B)$. By Lemma 4.3,

$$\|\mathbb{E}(Y \mid \hat{X}, Z)\|_{L^2}^2 \leq \mathbb{P}(B).$$

Optimality forces equality. Writing

$$q_B := \mathbb{E}(Y \mid \hat{X}, Z) = \mathbb{P}(B \mid \hat{X}, Z),$$

the same argument as above gives

$$\mathbb{1}_B = q_B \quad \text{a.s.}$$

Thus, $B \in \sigma(\hat{X}, Z)$.

Therefore, under any of the stated optimality conditions, every $B \in \sigma(\bar{X}, Z)$ belongs to $\sigma(\hat{X}, Z)$. Hence,

$$\sigma(\bar{X}, Z) \subset \sigma(\hat{X}, Z).$$

Combining this with the inclusion $\sigma(\hat{X}, Z) \subset \sigma(\bar{X}, Z)$, we obtain

$$\sigma(\hat{X}, Z) = \sigma(\bar{X}, Z).$$

This proves the equivalence and completes the proof. ■

5. Algorithm Design

In this section we describe the practical realization of our framework in terms of numerical algorithms. We first give general marginal information regularization algorithms that apply broadly to *feature unlearning* defined by Definition 3.1 and *marginal (data–point) unlearning* by Definition 3.2. We then present the analytic (closed-form) solver available for feature unlearning under the structured utility class in Theorem 4.3.

5.1 General cost feature unlearning: solution to Definition 2.2

For general utility costs, we solve the relaxed feature-unlearning problem in Definition 3.1 by penalizing the mutual information between the released outcome and the feature to be forgotten. Given a parametrized measurable map

$$f_\theta : \mathcal{X} \times \mathcal{Z} \to \mathcal{S},$$

we set

$$S = f_\theta(X, Z)$$

and optimize the regularized objective

$$(1-\lambda)\mathcal{C}(Y; S, Z) + \lambda I(S; Z).$$

Here $I(S; Z)$ may be estimated using a differentiable plug-in, variational, or adversarial mutual-information estimator, depending on the model class and data modality.

Algorithm 1: Feature Unlearning on Model via Regularization

Require: Dataset $D = (X, Y, Z) = \{(x_i, y_i, z_i)\}_{i=1}^N$, loss function $\mathcal{C}$, learning rate η, batch size B, number of epochs T, regularization parameter λ.
(Optional:) Pre-trained neural network $f_{\theta_{\text{origin}}}$ with parameters θ_{origin}.
Ensure: Unlearned model parameters θ.

Initialize: $\theta \leftarrow$ random initialization
if pre-trained model available **then**
  Load $\theta \leftarrow \theta_{\text{origin}}$
end if
for $t = 1$ to T **do**
  Shuffle dataset D
  for each mini-batch $d \subset D$ with $d = (X_d, Y_d, Z_d)$ **do**
    Compute predictions: $S_d = f_\theta(X_d)$
    Compute loss: $\mathcal{L}_{\text{reg}} = (1-\lambda)\mathcal{C}(Y_d; S_d, Z_d) + \lambda I(S_d; Z_d)$
    Compute gradients: $\nabla_\theta \mathcal{L}_{\text{reg}}$
    Update parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{reg}}$
  end for
end for
Return Unlearned model parameters θ

5.2 General cost marginal data unlearning: solution to Definition 2.5

For marginal unlearning, we operate in prediction space and penalize $I(S_{\text{margin}}; Z)$ with the paired construction $X_{\text{margin}}|_{\{Z=1\}} \sim p^r$, $X_{\text{margin}}|_{\{Z=0\}} \sim p^d$ as in Definition 3.2. We train by minimizing

$$(1-\lambda)\mathcal{C}(Y; S) + \lambda\, I(S_{\text{margin}}; Z).$$

The population guarantee follows from Lemma 4.1 and Theorem 4.2.

Algorithm 2: Marginal Data Unlearning via Regularization

Require: Remaining dataset $R = \{(x_i, y_i)\}_{i=1}^{N}$, Unlearning dataset $U = \{(x_i, y_i)\}_{i=N+1}^{N+K}$, trained neural network $f_{\theta_{\text{origin}}}$ with parameters θ_{origin}, loss function $\mathcal{C}$, learning rate η, batch size B, number of epochs T, regularization parameter λ.

Ensure: Unlearned model parameters θ and empirical leakage estimate $\widehat{\mu} \approx I(S_{\text{margin}}; Z)$. If $I(S_{\text{margin}}; Z) \leq \mu$, then Theorem 4.2 certifies high-probability $(\varepsilon, \delta_\varepsilon)$-marginal unlearning with

$$\delta_\varepsilon = \frac{e^\varepsilon + 1}{e^\varepsilon - 1}\sqrt{2\mu}.$$

Initialize: Load pre-trained parameters $\theta \leftarrow \theta_{\text{origin}}$.
for $t = 1$ to T **do**
  Shuffle datasets R and U.
  for each mini-batch $r \subset R$ and $u \subset U$, where $r = (X_r, Y_r)$ and $u = (X_u, Y_u)$ **do**
    Compute predictions: $S_r = f_\theta(X_r)$ and $S_u = f_\theta(X_u)$.
    Construct $S_0 = \text{concat}(S_r, S_u, \text{dim} = 0)^a$ and $S_1 = S_r$.
    Define joint distribution (S_{margin}, Z) where: $S_{\text{margin}}|_{Z=0} = S_0$ and $S_{\text{margin}}|_{Z=1} = S_1$.
    Compute regularized loss: $\mathcal{L}_{\text{reg}} = (1 - \lambda)\mathcal{C}(Y_r; S_r) + \lambda I(S_{\text{margin}}; Z)$.
    Compute gradients: $\nabla_\theta \mathcal{L}_{\text{reg}}$.
    Update model parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{reg}}$.
  end for
end for
Estimate final leakage $\widehat{\mu} \approx I(S_{\text{margin}}; Z)$.
Return unlearned model parameters θ and certificate estimate $\widehat{\mu}$.

a. dim = 0 here means row-wise concatenation.

5.3 Analytic optimal feature unlearning for specific utilities via Theorem 4.3

When specializing to the case where $\mathcal{S} = \mathcal{X}$ (hence $S = \hat{X}$) and the utility is one of the monotone information criteria in Theorem 4.3. Examples of such utilities include $-I(Y; \hat{X}, Z)$, the equivalent posterior-concentration objective $-\mathbb{E}[D_{\text{KL}}(\mathbb{P}(Y \mid \hat{X}, Z) \,\|\, \mathbb{P}(Y))]$, and the classification or regression energy criteria in Theorem 4.3. The barycentric representation $\bar{X} = T_Z(X)$ pushes every conditional law P_z to a common barycenter law $\bar{P}$, thereby enforcing $\bar{X} \perp Z$. In such cases, feature unlearning admits a closed-form/analytic optimal solution: the $\mathcal{W}_2$ barycenter of $\{X \mid Z = z\}_z$, which produces the finest admissible σ-algebra and enforces $\hat{X} \perp Z$.[2] More specifically, the Algorithm 3 below contains a pseudocode for the analytic solution of the feature unlearning problems by calculating the Wasserstein-2 barycenter of marginal distributions w.r.t. the feature to remove Z.

2. This analytic solution applies to feature unlearning only; marginal (data–point) unlearning does not admit a comparable barycenter form in general.

Algorithm 3: Analytic Unlearning Solution for Feature Unlearning via Wasserstein Barycenter

Require: Dataset $D = (X, Z) = \{(x_i, z_i)\}_{i=1}^N$, Maximum number of iterations T, Convergence threshold ε.
Ensure: Estimated Wasserstein barycenter $\bar{X}$.

Initialize: $\bar{X} \leftarrow \bar{X}_0$ {Random initialization of barycenter}
for $t = 1$ to T **do**
  for each unique value of Z: $z \in \text{unique}(Z)$ **do**
    Compute the optimal transport map T_z that maps $\bar{X}$ to $X_z := X|_{Z=z}$.
  end for
  Compute updated barycenter: $\bar{X}_{\text{new}} = \sum_{z \in \text{unique}(Z)} \frac{|X_z|}{|X|} T_z(\bar{X})$
  Compute convergence criterion: $\varepsilon_t = \mathcal{W}_2(\bar{X}, \bar{X}_{\text{new}})$.
  Update barycenter: $\bar{X} \leftarrow \bar{X}_{\text{new}}$.
  if $\varepsilon_t < \varepsilon$ **then**
    break {Terminate loop if convergence threshold is met}
  end if
end for
Return Estimated Wasserstein barycenter $\bar{X}$.

The computation of the Wasserstein-2 barycenter is known to be NP-hard in general [4]. Our proposed algorithm applies the fixed point estimation mechanism which guarantees convergence to the true barycenter [2], when provided true optimal transport maps. Therefore, one can apply different computational methods to estimate the optimal transport maps T_z's, e.g. linear maps or neural networks, and then follow the iterative method in Algorithm 3 to find an estimation of the true barycenter. In the experiment 6.3 below, we use neural optimal transport [44], which formulates the Kantorovich Duality as an adversarial generative network (GAN) structure to estimate the true optimal transport maps, together with the proposed iterative method (which becomes McCann interpolation in binary marginals setting) to estimate the Wasserstein-2 barycenter. We defer a detailed study of different estimation methods to future work.

We note here again that the above analytic solution only applies when the cost function is among the listed ones. Otherwise, the assumption of Theorem 4.3 is not satisfied and one should apply information-theoretic relaxation to approach the optimal solution for each of the unlearning problems with general cost functions.

6. Numerical Experiments

We conduct numerical experiments on synthetic and real-world datasets to validate the proposed framework, with emphasis on its theoretical guarantees, diagnostic behavior, and interpretability rather than exhaustive benchmarking against all existing unlearning methods. A more comprehensive evaluation, including improved barycenter estimation for feature unlearning, principled regularization selection, and refined mutual-information estimation for marginal data-point unlearning, is left for future work. The code is available at `https://github.com/xushizhou/Machine_Unlearn_via_Info_Reg`.

6.1 Feature Unlearning via Regularization: Algorithm 1

Setting We adopt a fixed 5-fold validation. In particular, for each fold $f \in \{1, \ldots, 5\}$, the data are split into train/validation/test as follows: we first take the provided train/test split for the fold, then partition the training split into (train, validation) using a stratified 80/20 split with a fixed seed. Models are trained only on the train portion, all thresholds and tradeoff knobs are selected using the validation portion if needed, and final metrics are reported on the test portion. We aggregate results as mean ± standard deviation over the five folds. For comparability purpose, all methods share the same classifier backbone: a two-layer MLP (width 128, ReLU, dropout 0.2), and the same training: Adam with learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, and 60 epochs.

Datasets For indirect comparison purpose, we evaluate four benchmark datasets that are popular choices in feature unlearning:

Dataset	**Target** Y	**Sensitive** Z	**Dim**(X)	**#Samples**
UCI Adult	Income (> \$50k)	Gender	16	48842
COMPAS	Two-year recidivism	Race	8	5300
UCI Bank	Subscription	Marital	51	45211
CelebA	Smiling	Gender	512	202599

Table 1: Benchmarks used in the feature unlearning experiments. For each fold, we train on the fold's training partition (with an internal 20% validation split, select thresholds/tradeoff knobs on validation, and report metrics on the fold's test partition. *Feature dimension Dim(X)* equals the post-encoding input dimensionality used by the classifier backbone (excluding Z; some methods concatenate Z during training/evaluation as specified in the method section).

Methods compared We include six representative approaches spanning pre-processing, post-processing, in-processing, representation learning, and two ERM baselines (include and exclude Z). Each exposes a single monotone knob (denoted $\lambda \in [0, 1]$), so that we can trace a consistent Pareto frontier.

Method	**Type**	**Tradeoff knob**	**Uses** Z **(train/test)**
Barycenter	Post-proc	λ (toward barycenter)	no / **yes**
DIR (quantile repair)	Pre-proc	λ (repair strength)	**yes** / no
MI (ours)	In-proc	λ in $(1-\lambda)\,\mathrm{CE} + \lambda\, I(\hat{Y}; Z)$	**yes** / **yes**
Zemel LFR	Representation	λ (monotone map to λ)	**yes** / no
ERM_deep(X)	Baseline	N/A	no / no
ERM_deep($[Z \mid X]$)	Baseline	N/A	**yes** / **yes**

Table 2: Comparison methods included in the feature unlearning experiment.

Policies and metrics Each method yields either a score $s_i \in [0, 1]$ or a per-sample probability mass function $\pi_i = (p_{i,0}, p_{i,1})$. To minimize the influence an additional step of generating hard label prediction, we directly evaluate via randomized policy which treats π_i as a randomized classifier and compute:

$$\mathrm{acc}_{\mathrm{rand}} \;=\; \frac{1}{n}\sum_{i=1}^{n} p_{i,Y_i}, \qquad \mathrm{dp_gap} \;=\; \left|\, \mathbb{E}[p_{i,1} \mid Z{=}1] - \mathbb{E}[p_{i,1} \mid Z{=}0] \,\right|,$$

where $p_{i,1}$ is the probability of predicting class 1. AUROC is computed from the underlying scalar score s_i or $p_{i,1}$ (e.g., softmax logit for ERM/MI/LFR, OT-mapped score for Barycenter).

Results. Figure 1 shows that the proposed mutual information regularization method (MI) is the best method in reducing the influence of Z on learning outcome (quantified by dp-gap) while preserving utility accuracy and the area under curve on Adult and Compas datasets, and only slightly behind Chzhen's Wasserstein barycenter method (which only works for binary variable with significantly worse computational cost) on Bank and CelebA datasets. Furthermore, it offers the best smooth parameterization of the Pareto frontier, equivalently the optimal trade-off, between utility (quantified by accuracy and AUROC) and feature influence (quantified by dp-gap).

6.2 Marginal Data-Point Unlearning via Regularization: Algorithm 2

6.2.1 Forget Gaussian Experiment

This synthetic experiment illustrates the mechanism of marginal data-point unlearning in a one-dimensional setting where the retain and unlearn distributions have visibly different structure. The retain distribution is nearly flat on an interval, while the unlearn distribution is concentrated near the origin. Thus, the information to be forgotten is the Gaussian concentration around 0, while the information to be preserved is the approximately uniform retain density on $[-L, L]$.

Data and notation. We consider two distributions supported on the same interval $[-L, L]$, with $L = 3$:

$$X_r \sim \mathrm{Unif}([-L, L]), \qquad X_u \sim \mathcal{N}(\mu, \sigma^2) \quad \text{truncated to } [-L, L].$$

We draw N_R i.i.d. retain samples $\{x_i^{(r)}\}_{i=1}^{N_R}$ from X_r, and N_U i.i.d. unlearn samples $\{x_j^{(u)}\}_{j=1}^{N_U}$ from X_u. Let

$$\rho := \frac{N_R}{N_R + N_U}$$

denote the empirical retain fraction. The retain-plus-unlearn mixture is therefore $p^d = \rho p^r + (1-\rho)p^u$. Equivalently, if the theoretical notation in Section 3.2 uses $p^d = (1-\alpha)p^r + \alpha p^u$, then here $\rho = 1 - \alpha$.

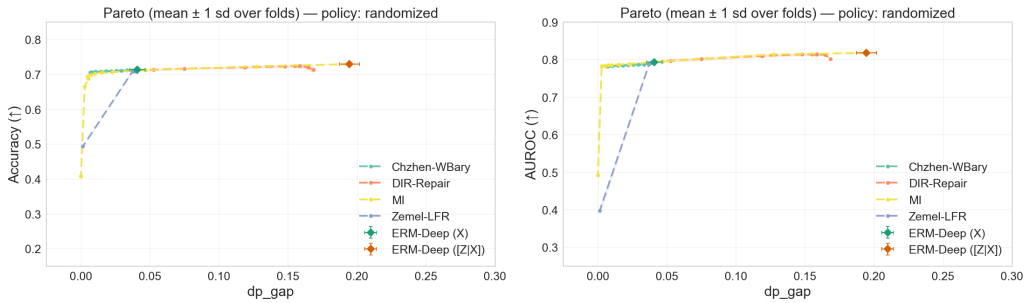
We use a residual network $f_\theta : \mathbb{R} \to \mathbb{R}$, initialized near the identity, and first pretrain it so that the retain output density mimics the retain-plus-unlearn mixture:

$$\widehat{p}_{f_{\theta_0}(X_r)} \approx \rho\, \widehat{p}_{X_r} + (1-\rho)\, \widehat{p}_{X_u}.$$

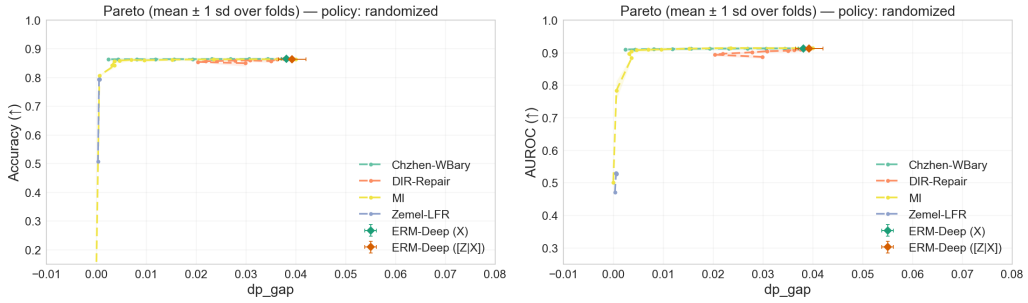
This initialization models the situation in which the original model has absorbed the unlearn-set signal into its output behavior.

All densities are estimated on a uniform grid $\mathcal{G} = \{x_k\}_{k=1}^{N_X}$ with spacing Δx, using Gaussian KDE with bandwidths h_x for inputs and h_y for outputs:

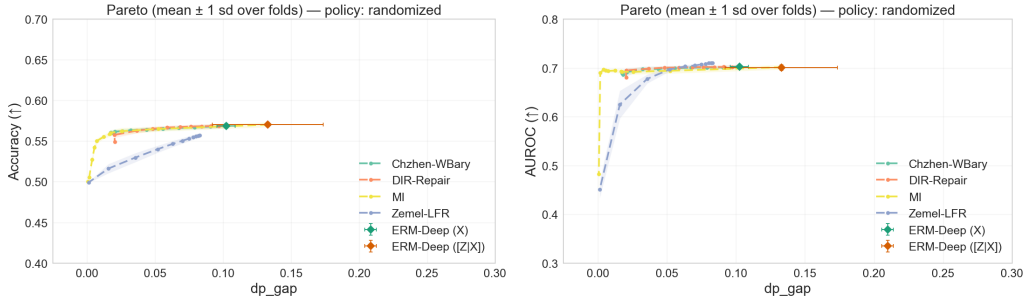
$$\widehat{p}_{S,h}(x_k) := \frac{1}{m} \sum_{\ell=1}^{m} \frac{\exp\big(-\frac{(x_k - s_\ell)^2}{2h^2}\big)}{\sum_{j=1}^{N_X} \exp\big(-\frac{(x_j - s_\ell)^2}{2h^2}\big)\Delta x}.$$



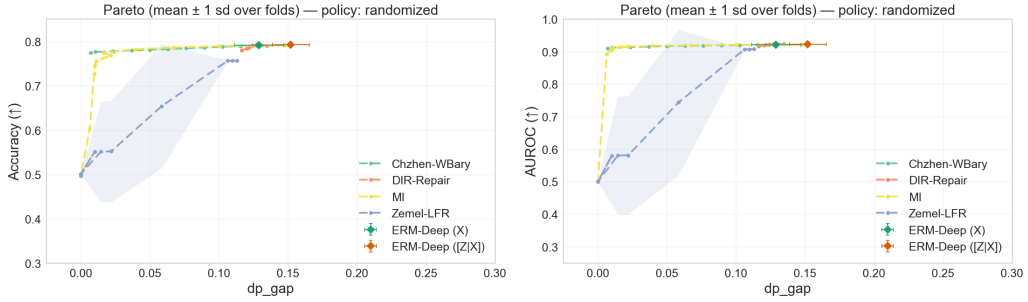
(a) Results on UCI Adult



(b) Comparison result on UCI Bank



(c) Comparison result on COMPAS



(d) Comparison result on CelebA

Figure 1: Feature–unlearning frontiers: Each row reports utility versus feature influence for a dataset: the left panel shows accuracy (↑) and the right panel shows AUROC (↑) against the demographic-parity gap (DP-gap, ↓), defined as $|\,\mathbb{P}[\hat{Y}{=}1 \mid Z{=}1] - \mathbb{P}[\hat{Y}{=}1 \mid Z{=}0]\,|$. Curves trace each method's trade-off as its trade-off parameter varies. Points denote the mean and bands denote ±1 s.d. over 5 folds. Lower DP-gap at comparable or higher utility indicates a better frontier.

For grid densities P, Q, we approximate

$$H(P\|Q) := -\sum_k P(x_k) \log Q(x_k)\, \Delta x,$$

and

$$\mathrm{D}_{\mathrm{KL}}(P\|Q) := \sum_k P(x_k) \log \frac{P(x_k)}{Q(x_k)} \, \Delta x.$$

For a binary auditing label $Z \in \{0,1\}$ with priors p_0, p_1 and conditional output densities P_0, P_1, the mutual information estimator is

$$I(S_{\mathrm{margin}}; Z) = p_0 \, \mathrm{D}_{\mathrm{KL}}(P_0\|P) + p_1 \, \mathrm{D}_{\mathrm{KL}}(P_1\|P), \qquad P := p_0 P_0 + p_1 P_1.$$

Compared objectives. We compare three objectives. Unless otherwise stated, the auditing prior is balanced, $p_0 = p_1 = \frac{1}{2}$.

- **Marginal-MI (ours).** The marginal objective compares the retain output law against the retain-plus-unlearn mixture output law. Define

  $$P_1 := \widehat{p}_{f_\theta(X_r)}, \qquad P_0 := \rho \, \widehat{p}_{f_\theta(X_r)} + (1-\rho) \, \widehat{p}_{f_\theta(X_u)}.$$

  The loss is

  $$\mathcal{L}_{\mathrm{MARGINAL}}(\theta) = (1-\lambda) \, H\big(\widehat{p}_{X_r} \| \widehat{p}_{f_\theta(X_r)}\big) + \lambda \, I(S_{\mathrm{margin}}; Z).$$

  Minimizing the MI term drives $P_0 = P_1$. Since $P_0 = \rho P_1 + (1-\rho)\widehat{p}_{f_\theta(X_u)}$, this forces

  $$\widehat{p}_{f_\theta(X_u)} = \widehat{p}_{f_\theta(X_r)}$$

  whenever $1-\rho > 0$. The utility term simultaneously anchors $\widehat{p}_{f_\theta(X_r)}$ to the retain density $\widehat{p}_{X_r}$.

- **Grad-Diff.** This baseline replaces the MI term by a negative cross-entropy term that pushes the unlearn output away from the original unlearn density:

  $$\mathcal{L}_{\mathrm{GRAD_DIFF}}(\theta) = (1-\lambda) \, H\big(\widehat{p}_{X_r} \| \widehat{p}_{f_\theta(X_r)}\big) - \lambda \, H\big(\widehat{p}_{X_u} \| \widehat{p}_{f_\theta(X_u)}\big).$$

  This discourages resemblance to the unlearn density, but does not prescribe where the unlearn mass should move.

- **KL+utility.** This baseline anchors the retain output to the pretrained model while applying the same negative unlearn-density term:

  $$\mathcal{L}_{\mathrm{KL}}(\theta) = (1-\lambda) \, \mathrm{D}_{\mathrm{KL}}\big(\widehat{p}_{f_{\theta_0}(X_r)} \| \widehat{p}_{f_\theta(X_r)}\big) - \lambda \, H\big(\widehat{p}_{X_u} \| \widehat{p}_{f_\theta(X_u)}\big).$$

  The first term keeps $f_\theta(X_r)$ close to the pretrained retain output, whereas the second term discourages resemblance to X_u.

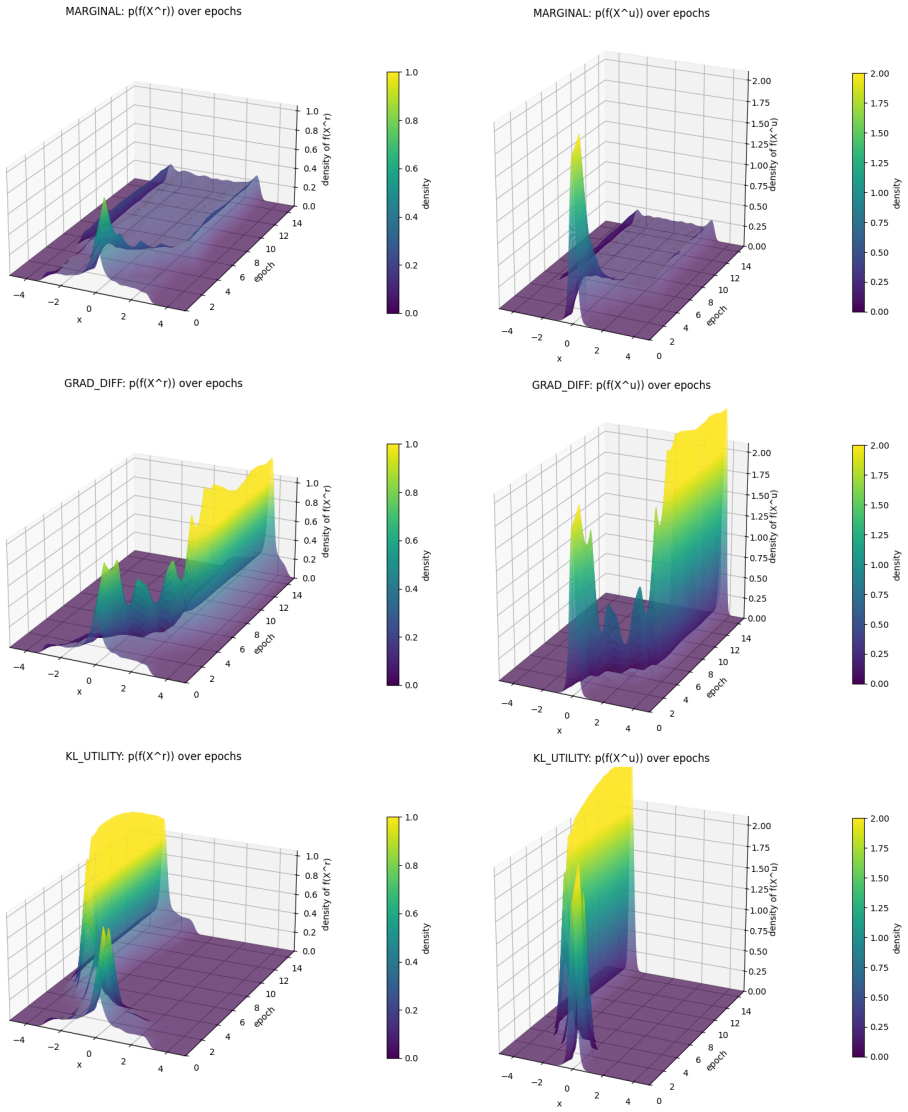


Figure 2: Evolution of the output densities $p(f_\theta(X_r))$ and $p(f_\theta(X_u))$ over epochs for the three objectives. Left column: retain outputs. Right column: unlearn outputs. Marginal-MI suppresses the unlearn-specific concentration near zero while keeping the retain output close to the uniform retain density. In contrast, the gradient-ascent-based baselines tend to relocate the unlearn mass toward other regions, such as the boundaries, rather than matching it to the retain distribution.

Outcomes. Figure 2 shows how the retain and unlearn output densities evolve from the pretrained model. Under MARGINAL, the MI penalty equalizes the retain-only and retain-plus-unlearn output conditionals. Algebraically, this drives

$$p(f_\theta(X_u)) \approx p(f_\theta(X_r)),$$

while the utility term keeps

$$p(f_\theta(X_r)) \approx p(X_r).$$

Thus, the unlearn-specific Gaussian concentration is removed by matching it to the retain output distribution, rather than by pushing it to an arbitrary location.

In contrast, GRAD_DIFF lacks a target distribution for the unlearn mass. Its negative cross-entropy term can move the unlearn density toward the boundaries or other regions, and shared parameters may also distort the retain output. The KL baseline stabilizes the retain trajectory by anchoring to the pretrained output, but its unlearning term still does not specify where the unlearn mass should go. Consequently, it can also push the forget mass toward the edges.

Summary. The Gaussian experiment illustrates the qualitative difference between marginal unlearning and gradient-ascent-style forgetting. Marginal-MI removes the unlearn-specific signal by matching the unlearn output law to the retain output law, while retaining the uniform structure through the utility term. The baselines can reduce resemblance to the unlearn density, but they do so without a principled replacement distribution for the removed mass.

6.2.2 Forget MNIST

Dataset and retain/unlearn split. We use MNIST with the standard 60,000/10,000 train/test split and normalize images using mean 0.1307 and standard deviation 0.3081. Let U_{all} denote the set of training images whose label is digit 3. We randomly select 99.5% of U_{all} as the unlearn set U, and define the retain set as

$$R := \text{train} \setminus U.$$

Thus, the retain set contains almost no digit-3 examples, while the FineTune-All model is trained on the full original training set.

For evaluation and trajectory aggregation, we form five random folds independently on R and on U. For each fold i, this gives

$$(R_i^{\text{tr}}, R_i^{\text{val}}) \qquad \text{and} \qquad (U_i^{\text{tr}}, U_i^{\text{val}}).$$

These folds are used to train and validate the unlearning trajectories. Test accuracy is computed on the original MNIST test set. Unless otherwise stated, the nominal retain and unlearn batch sizes are 128, and the test batch size is 512; on MPS devices, the implementation reduces the retain and unlearn batch sizes to 96 for stability.

Shared backbone and optimization for comparability. We use a lightweight CNN for MNIST (28×28 grayscale) to ensure strict cross-method comparability. All convolutions use 3×3 kernels, stride 1, and padding 1 ("same"), so spatial resolution is preserved between pooling layers. The architecture is:

$$\begin{aligned} &\underbrace{\text{Conv}(32) \to \text{ReLU} \to \text{Conv}(64) \to \text{ReLU} \to \text{MaxPool}(2) \to \text{Dropout}(0.25)}_{\textbf{Block 1}} \\ \to\ &\underbrace{\text{Conv}(128) \to \text{ReLU} \to \text{MaxPool}(2) \to \text{Dropout}(0.25)}_{\textbf{Block 2}} \\ \to\ &\underbrace{\text{Flatten}(128{\times}7{\times}7) \to \text{FC}(256) \to \text{ReLU} \to \text{Dropout}(0.5) \to \text{FC}(10)}_{\textbf{Head}}. \end{aligned}$$

Max-pooling layers are 2×2 (stride 2), yielding spatial sizes 28×28 → 14×14 → 7×7. ReLU follows every convolution and the first fully connected layer; dropout is applied after each pooling block and before the classifier head.

We use Adam with learning rate 10^{-3} and weight decay 10^{-4}, and fix the random seed to 1337. FineTune-All and Retrain-on-Retain each run for 10 epochs. Unlearning runs for at most 30 epochs and may stop early according to the rules below.

Baselines and compared methods. We compare five models.

- **FineTune-All (FT).** Train the backbone on the full MNIST training set for 10 epochs. This is the initial model that has seen the unlearn data.
- **Retrain-on-Retain (RT).** Train the same backbone from scratch on R only for 10 epochs. This is the anchor model that never trains on the selected unlearn set U.
- **Marginal-MI (ours).** Starting from the FT weights, minimize

$$\mathcal{L}_{\mathrm{MI}} = (1-\gamma)\,\mathrm{CE}\big(f_\theta(x), y;\ x \in R_i^{\mathrm{tr}}\big) + \gamma\,\widehat{\mathrm{MI}}_{\mathrm{soft}}.$$

  The estimator $\widehat{\mathrm{MI}}_{\mathrm{soft}}$ is a softmax-marginal plug-in proxy for $I(S_{\mathrm{margin}}; Z)$. Let

$$P_r := \mathbb{E}_{x \in R_i^{\mathrm{tr}}}[\hat{p}_\theta(\cdot \mid x)], \qquad P_u := \mathbb{E}_{x \in U_i^{\mathrm{tr}}}[\hat{p}_\theta(\cdot \mid x)].$$

  With $\rho := \frac{|R_i^{\mathrm{tr}}|}{|R_i^{\mathrm{tr}}|+|U_i^{\mathrm{tr}}|}$, the retain-plus-unlearn mixture output is $P_d := \rho P_r + (1-\rho)P_u$. Using balanced auditing priors, the estimator is

$$\widehat{\mathrm{MI}}_{\mathrm{soft}} = \frac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(P_r \| P) + \frac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(P_d \| P), \qquad P := \frac{1}{2}P_r + \frac{1}{2}P_d.$$

  This is not the full empirical output-law mutual information. Instead, it is a stable low-dimensional proxy based on average class-probability vectors.

- **Grad-Diff.** Starting from the FT weights, optimize

$$\mathcal{L}_{\mathrm{GD}} = (1-\gamma)\,\mathrm{CE}\big(f_\theta(x), y;\ x \in R_i^{\mathrm{tr}}\big) - \gamma\,\mathrm{CE}\big(f_\theta(x), y;\ x \in U_i^{\mathrm{tr}}\big).$$

- **KL+CE.** With the FT model frozen as teacher $f_{\theta^\star}$, optimize from the FT weights

$$\mathcal{L}_{\mathrm{KL}} = (1-\gamma)\mathbb{E}_{x \in R_i^{\mathrm{tr}}}\left[\mathrm{D}_{\mathrm{KL}}\left(\hat{p}_{\theta^\star}(\cdot \mid x) \| \hat{p}_\theta(\cdot \mid x)\right)\right] - \gamma\,\mathrm{CE}\big(f_\theta(x), y;\ x \in U_i^{\mathrm{tr}}\big).$$

Trade-off grids and initialization. Every unlearning method is initialized from the same FT checkpoint. We sweep method-specific γ-grids:

$$\gamma_{\mathrm{MI}} \in \{0.0020,\ 0.0055,\ 0.0100\},$$

$$\gamma_{\mathrm{KL}} \in \{0.0003,\ 0.0006,\ 0.0012\}, \qquad \gamma_{\mathrm{GD}} \in \{0.0002,\ 0.00035,\ 0.0007\}.$$

The grids are chosen so that each method ranges from mild unlearning to a substantial decrease in unlearn-set accuracy.

Early stopping. We use early stopping only for Marginal-MI and Grad-Diff. For KL+CE, the student is initialized from the same FT checkpoint as the frozen teacher, so the initial teacher–student retain-validation KL is numerically zero. A relative KL stopping rule based on this initial value is therefore inactive in this experiment, and KL+CE is run for the full unlearning budget of 30 epochs.

- **Marginal-MI:** compute $\mathrm{MI}_0^{\mathrm{val}}$ before unlearning on $(R_i^{\mathrm{val}}, U_i^{\mathrm{val}})$, and stop when

$$\mathrm{MI}^{\mathrm{val}} \le 0.85\,\mathrm{MI}_0^{\mathrm{val}}$$

  with minimum epoch 1 and patience 1.

- **Grad-Diff:** infer the number of classes C from the logits ($C = 10$ for MNIST), and stop when the unlearn-validation accuracy satisfies

$$\mathrm{Acc}_U^{\mathrm{val}} \le \frac{1}{C} + 0.02$$

  for two consecutive checks, with minimum epoch 1 and patience 2.

Results. Figure 3 records the training-fold accuracy on R_i^{tr} and U_i^{tr}, together with the softmax-marginal MI proxy, over the unlearning trajectory. Curves are aggregated over five folds using mean ± standard deviation. The horizontal bands indicate the FineTune-All and Retrain-on-Retain accuracy references.

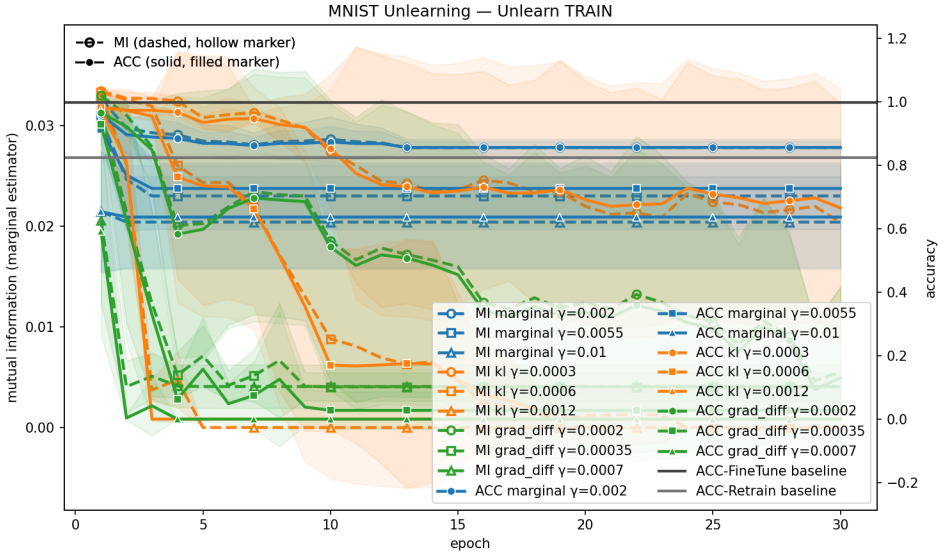
The experiment shows two main trends:

- *The softmax-marginal MI proxy tracks the unlearning trajectory.* For Marginal-MI, decreasing the MI proxy moves the unlearn-set behavior away from the FineTune-All baseline and toward the Retrain-on-Retain reference. This is consistent with the theory: reducing $I(S_{\mathrm{margin}}; Z)$ reduces the distinguishability between the retain-only output law and the retain-plus-unlearn output law. Since digit 3 is almost entirely removed from the retain set, the retrain-on-retain model provides the natural reference behavior for the unlearn subset.

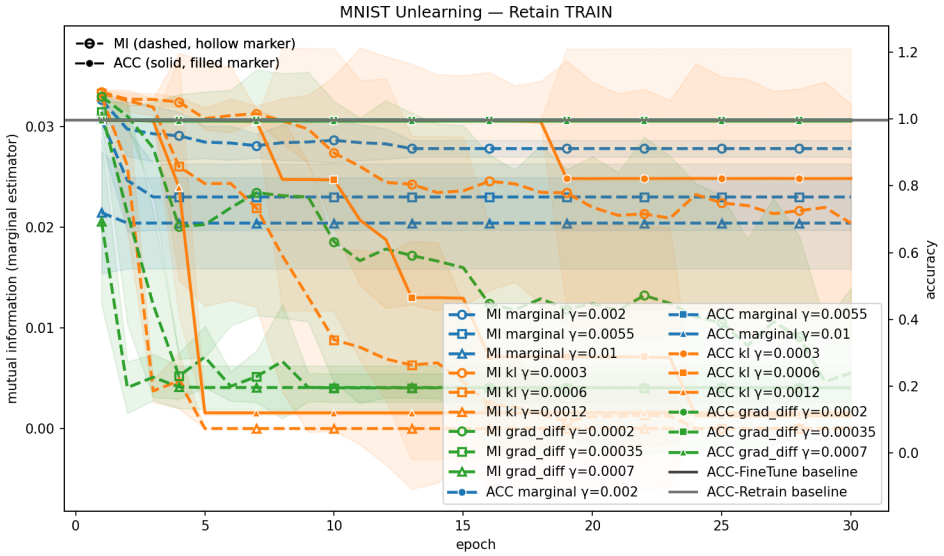
- *Marginal-MI most stably approaches the retrain-on-retain reference in this experiment.* Compared with the gradient-ascent-based baselines, Marginal-MI stays closest to the Retrain-on-Retain reference across the tested trade-off parameters and training trajectories. This stability can be seen from four perspectives: (1) robustness to the trade-off parameter, (2) robustness across folds, (3) stability of the training trajectory, and (4) simultaneous retain- and unlearn-set behavior.

  From the *parameter-robustness* perspective, different choices of γ for Marginal-MI lead to only moderate changes in the unlearn-set trajectory, and the resulting curves remain near the Retrain-on-Retain reference. In contrast, the Grad-Diff and KL+CE baselines are more sensitive to γ: depending on the value of the trade-off parameter, they may either remain too close to the FineTune-All model or move too aggressively toward chance-level behavior.

  From the *cross-validation robustness* perspective, Marginal-MI exhibits smaller fold-to-fold variation in the plotted trajectories. The shaded regions in Figure 3 are visibly



(a) Unlearn set performance trajectories



(b) Retain set performance trajectories

Figure 3: MNIST marginal data-unlearning trajectories on the training folds. The top panel reports unlearn-train accuracy together with the softmax-marginal MI proxy, and the bottom panel reports retain-train accuracy together with the same MI proxy. Solid filled markers show accuracy, while dashed hollow markers show MI. Lines denote means and shaded regions denote ± 1 standard deviation over five folds. Horizontal bands show the FineTune-All and Retrain-on-Retain accuracy references.

narrower for Marginal-MI than for the baselines, indicating more stable behavior across the five folds. If the corresponding table is included, this observation can be quantified by reporting the standard deviations of retain and unlearn accuracies at the stopping epoch.

From the *training-trajectory* perspective, Marginal-MI produces a smooth and stable movement of the unlearn-set behavior toward the Retrain-on-Retain reference, while re-

taining high performance on the retain set. By contrast, Grad-Diff and KL+CE display larger fluctuations and sharper drops for some trade-off choices, reflecting the instability caused by directly maximizing an unlearn-set loss.

From the *retain/unlearn accuracy* perspective, Marginal-MI better balances the two desiderata: it moves the unlearn-set behavior toward the Retrain-on-Retain reference while preserving retain-set accuracy. The gradient-ascent-based baselines are more vulnerable to under- or over-unlearning: they can either fail to move far enough from the FineTune-All behavior, or push the unlearn-set accuracy too far toward chance while also causing larger changes in retain-set performance.

These four empirical observations are consistent with the objective-level interpretation of marginal unlearning. Marginal-MI directly penalizes the source-label distinguishability $I(S_{\text{margin}}; Z)$, while its cross-entropy term preserves retain-set utility. Thus, the two terms are aligned with the marginal-unlearning goal: preserve the retain behavior while making the retain-only and retain-plus-unlearn output laws difficult to distinguish. In contrast, Grad-Diff and KL+CE use a gradient-ascent term on the unlearn set. This term indirectly seeks forgetting by pushing against the unlearn labels, but it does not specify the target behavior that should replace the removed information. As a result, the utility-preservation term and the unlearn-set gradient-ascent term can act like competing forces, producing greater sensitivity to γ, fold variation, and training dynamics.

6.3 Optimal Feature Unlearning via Analytic Solution: Algorithm 3

Dataset. We use the CelebFaces Attributes dataset (CelebA) [49], a large-scale collection of 200K celebrity face images annotated with 40 binary attributes (e.g., `Smiling`, `Female`). We focus on two binary features, *Smile* and *Gender*, and consider the standard image resolution pre-processing used in face-generation benchmarks.

Problem setup. Let $Z \in \{0, 1\}$ denote the attribute to unlearn (e.g., Z=1 for "smiling", Z=0 for "non-smiling"). Denote by μ_0 and μ_1 the empirical image distributions conditioned on Z=0 and Z=1, respectively. Our goal is to transform each image (x, z) to an *attribute-neutral* counterpart $\bar{x}$ whose distribution $\bar{\mu}$ is (i) minimally distorted from the originals and (ii) uninformative about Z.

For a complete feature-unlearning map, one applies slice-wise maps from each conditional law to the common barycenter:

$$R_0(x) = \frac{1}{2}x + \frac{1}{2}T_{0\to1}(x), \qquad R_1(y) = \frac{1}{2}y + \frac{1}{2}T_{1\to0}(y),$$

so that ideally $R_{0\#}\mu_0 = R_{1\#}\mu_1 = \bar{\mu}$. In the visualization below, we display one learned transport direction and its midpoint interpolation.

Barycentric characterization (theory). Theorem 4.3 shows that the 2-Wasserstein barycenter of μ_0 and μ_1 with equal weights characterizes the optimal Z-neutral distribution:

$$\bar{\mu} \;\in\; \arg\min_{\mu} \; \tfrac{1}{2}W_2^2(\mu, \mu_0) + \tfrac{1}{2}W_2^2(\mu, \mu_1).$$

For two measures with equal weights, the unique minimizer equals the midpoint of the W_2 geodesic connecting μ_0 and μ_1 [52]. Let $T_{0\to1}$ be the Brenier map pushing μ_0 to μ_1 (i.e.,

$T_{0\to1}{=}\nabla\psi$ for a convex potential ψ). The McCann *displacement interpolation* induces a geodesic

$$\mu_t \;=\; \big((1{-}t)\,\mathrm{Id} + t\,T_{0\to1}\big)_{\#}\mu_0, \qquad t\in[0,1],$$

and the equal-weight barycenter is precisely the midpoint $\bar{\mu} = \mu_{1/2}$. Intuitively, projecting each sample to the geodesic midpoint makes the two attribute-conditional populations equidistant in W_2, attenuating attribute information while minimizing transport cost.

Neural OT implementation (practice). Direct computation of $T_{0\to1}$ in image space is intractable. We therefore use a neural OT approach in the spirit of Korotin et al. [44]: a *critic* network approximates the Kantorovich potential (dual), and a *generator* parameterizes the transport map. Concretely, we learn a forward map $T_{0\to1}$ using adversarial objectives that enforce the dual constraints and pushforward consistency. Given a source image x with attribute z, we produce its barycentric counterpart by a midpoint displacement step along the learned map for its group:

$$\bar{x} \;= \frac{1}{2}\,x + \frac{1}{2}\,T_{0\to1}(x)$$

which realizes the McCann interpolation at $t{=}0.5$.

Results. Figure 4 illustrates the unlearning outcomes for two target attributes: smile and gender. In each panel, the first row X shows source images from one attribute-conditioned group. The last row $T(X)$ shows the corresponding push-forward images under the learned OT map. The middle row,

$$\mathrm{Bary}(X) := \frac{1}{2}X + \frac{1}{2}T(X),$$

shows the midpoint displacement interpolation, which corresponds to the two-marginal barycenter in the ideal population setting.

In theory, $\mathrm{Bary}(X)$ approximates the optimal feature-unlearned representation. The target attribute is visually attenuated, while other image features, such as lighting, pose, and idiosyncratic facial traits, are largely preserved.

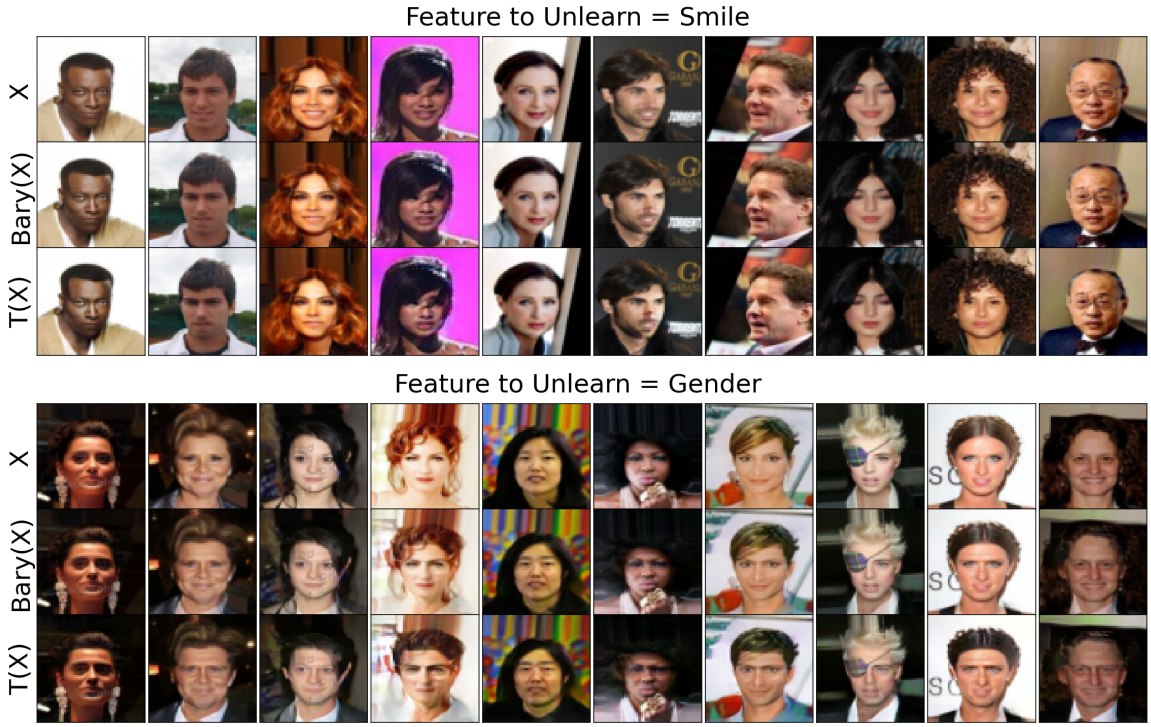


Figure 4: illustrates the unlearning outcomes for both tasks: The above penal shows the unlearn results for smile feature and the bottom shows the results for gender feature. In each of the panel, the first row, denoted by X, represents samples from the original data set with the chosen feature value (smile and female, respectively). The last row, denoted by $T(X)$ represents the push-forward image of the corresponding sample in X by the generated optimal transport map T. The middle row, denoted by $\text{Bary}(X) := [0.5Id + 0.5T](x)$, represents the corresponding sample generated by the McCann interpolation at $t = 0.5$, which coincides with the barycenter in the our two-marginal cases.

Remarks and limitations. The W_2 barycenter in pixel space gives a principled, utility-aware neutralization target, but the pixel ℓ_2 ground cost is in general not perfectly aligned with the perceptual face manifold. In practice, neural OT ameliorates this by learning structured maps, yet alternative ground costs or feature-space OT (e.g., in a perceptual embedding) may further improve realism. Extending to multi-class or continuous attributes is straightforward via multi-marginal barycenters with prior weights.

Conclusion

This paper advances *machine unlearning* by shifting the foundation from an *anchor-based* retrain-on-retain paradigm, which is difficult to verify in practice, to an *auditable, inference-based* paradigm grounded in the **Marginal Unlearning Principle**. Inspired by "blur + reinforce" mechanisms from cognitive neuroscience, our principle targets the *marginal information* contributed by the unlearn set beyond the retain set and provides provable guarantees about what an observer can infer from model outputs. We further implement the principle through a unified information-theoretic framework (a rate–distortion formulation) that trades off utility against an information regularizer $I(S'; Z)$. We also devise practical algorithms for both feature and data-point unlearning, and provide guarantees: (i) *auditability* from directly from model inputs-outputs, (ii) *sufficiency* for the prevailing (approximate) anchored guarantees under utility control, and (iii) a complementary *necessity* statement for robust retrain-on-retain models. In the hard-independence limit ($S' \perp Z$), we identify a

single analytic solution, the $\mathcal{W}_2$ barycenter, that solves feature unlearning under standard learning utility objectives and yields maximal utility among admissible outcomes. Experiments across tabular and image modalities corroborate the capability of the framework and illustrate its practicality. Collectively, these results pave a principled and testable road for replacing anchor-based requirements with neuroscience-inspired, information-theoretic *marginal unlearning*.

Open problems. The results here open several directions of research for theory, algorithms, and practice:

1. **Sharper quantification of marginal information.** Beyond mutual information, identify and analyze alternative leakage measures (e.g., general f-divergence, Rényi–MI, χ^2-divergence, or other integral probability metrics (IPMs) over task-relevant test functions) that (a) admit tighter utility-unlearning tradeoffs, (b) are easier to estimate from finite samples, and (c) align with specific downstream utilities.

2. **From MI to auditable stopping rules.** Empirically, unlearn accuracy tracks $I(S'; Z)$ closely. Developing provable theoretical results on the connection between marginal information quantification (not necessarily the proposed mutual information) and unlearn accuracy may yield principled schedules for tuning λ and *stopping criteria* that certify target leakage levels.

3. **Information–transport synthesis in high dimensions.** Our results demonstrate that information-theoretical regularization methods potentially can outperform brute force matching methods in estimating optimal transport maps, plans, and Wasserstein barycenter. Can this be extended beyond $\mathcal{W}_2$ to task-aligned costs and multi-marginal barycenters for complex, high-dimensional unlearning?

4. **Broader modalities and multi-modal systems.** Instantiate marginal unlearning for LLMs, speech, graphs, time-series, and multi-modal models. This includes modality-specific (S', Z) designs, direct output-level audit protocols for generative systems, and integration with continual learning/fine-tuning so that updates learn only the *unique* signal of new data while preserving retained knowledge.

Acknowledgement

The authors acknowledge support from NSF DMS-2208356, NIH R01HL16351, P41EB032840, and DE-SC0023490.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 308–318, New York, NY, USA, 2016. ACM.

[2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[3] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[4] J. M. Altschuler and E. Boix-Adsera. Wasserstein barycenters are NP-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

[5] I. Amara, A. I. Humayun, I. Kajic, Z. Parekh, N. Harris, S. Young, C. Nagpal, N. Kim, J. He, C. Nader Vasconcelos, et al. EraseBench: Understanding the ripple effects of concept erasure techniques. *arXiv:2501.09833*, 2025.

[6] M. C. Anderson and C. Green. Suppressing unwanted memories by executive control. *Nature*, 410(6826):366–369, 2001.

[7] M. C. Anderson and S. Hanslmayr. Neural mechanisms of motivated forgetting. *Trends in cognitive sciences*, 18(6):279–292, 2014.

[8] N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman. LEACE: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023.

[9] T. Berger. *Rate-distortion theory*. Wiley, Hoboken, NJ, 2003.

[10] R. Bonta. California consumer privacy act (ccpa). *Retrieved from State of California Department of Justice: https://oag. ca. gov/privacy/ccpa*, 2022.

[11] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159, Oxford, UK, 2021. IEEE.

[12] Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy (SP)*, pages 463–480, Oxford, UK, 2015. IEEE.

[13] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, U. Erlingsson, A. Oprea, and N. Papernot. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2021.

[14] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*, pages 896–911, 2021.

doi: 10.1145/3460120.3484756. URL https://yangzhangalmo.github.io/papers/CCS21-UnlearningLeaks.pdf.

[15] S. Chiappa. Path-specific counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7801–7808, Jul. 2019. doi: 10.1609/aaai.v33i01.33017801. URL https://ojs.aaai.org/index.php/AAAI/article/view/4777.

[16] B. Christian. *The Alignment Problem: Machine Learning and Human Values.* W.W Norton & Company, New York, 2020.

[17] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:7321–7331, 2020.

[18] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, Hoboken, NJ, 2nd edition, 2006.

[19] A. Criminisi, J. Shotton, E. Konukoglu, et al. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and trends® in computer graphics and vision*, 7(2–3):81–227, 2012.

[20] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, New York, NY, 2022.

[21] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[22] C. Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[23] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[24] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[25] H. Edwards and A. Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR)*, 2016.

[26] Y. Elazar, S. Ravfogel, Y. Goldberg, and Y. Belinkov. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

[27] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL https://data.europa.eu/eli/reg/2016/679/oj.

[28] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436, 2023.

[29] A. A. Ginart, M. Y. Guan, G. Valiant, and J. Zou. Making ai forget you: data deletion in machine learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[30] A. Golatkar, A. Achille, and S. Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.

[31] T. L. Gouic, J.-M. Loubes, and P. Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

[32] C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten. Certified data removal from machine learning models. *arXiv:1911.03030*, 2019.

[33] L. Han, H. Huang, D. Scheinost, M.-A. Hartley, and M. R. Martínez. Unlearning information bottleneck: Machine unlearning of systematic patterns and biases. *arXiv preprint arXiv:2405.14020*, 2024.

[34] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[35] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[36] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[37] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[38] S. Hong, J. Lee, and S. S. Woo. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *AAAI*, 2024.

[39] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[40] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[41] C.-P. Huang, K.-P. Chang, C.-T. Tsai, Y.-H. Lai, F.-E. Yang, and Y.-C. F. Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pages 360–376. Springer, 2024.

[42] J. D. Karpicke and J. R. Blunt. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018):772–775, 2011.

[43] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwińska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.

[44] A. Korotin, D. Selikhanovych, and E. Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

[45] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *NeurIPS*, 2017.

[46] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL `https://aclanthology.org/2021.emnlp-main.243/`.

[47] Z. Li, A. Perez-Suay, G. Camps-Valls, and D. Sejdinovic. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *arXiv:1911.04322*, 2019.

[48] J. Liu, Y. Yao, J. Chen, B. Wang, and H. Shen. Fair representation learning: An alternative to mutual information. In *KDD*, 2022.

[49] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[50] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. In *International Conference on Learning Representations (ICLR)*, 2016.

[51] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3384–3393. PMLR, 2018.

[52] R. J. McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[53] R. Nabi and I. Shpitser. Fair inference on outcomes. AAAI'18/IAAI'18/EAAI'18, New Orleans, Louisiana, USA, 2018. AAAI Press. ISBN 978-1-57735-800-8.

[54] K. Nader, G. E. Schafe, and J. E. LeDoux. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797):722–726, 2000.

[55] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 487–503, 2021.

[56] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7237–7256, 2020.

[57] S. Ravfogel, Y. Moskovitch, Y. Goldberg, and R. Cotterell. Linear Adversarial Concept Erasure. In *ICML (PMLR v162)*, 2022.

[58] L. Sahakyan and C. M. Kelley. A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6): 1064, 2002.

[59] D. Schiller, M.-H. Monfils, C. M. Raio, D. C. Johnson, J. E. Ledoux, and E. A. Phelps. Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, 463(7277):49–53, 2010.

[60] A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

[61] C. E. Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

[62] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, Oxford, UK, 2017. IEEE.

[63] D. M. Sommer, L. Song, S. Wagh, and P. Mittal. Athena: Probabilistic verification of machine unlearning. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2022 (3):268–290, 2022.

[64] A. Thudi, H. Jia, I. Shumailov, and N. Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security)*, pages 4007–4022, 2022.

[65] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3.

[66] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., Providence, RI, 2021.

[67] W. Wang, C. Zhang, Z. Tian, and S. Yu. Machine unlearning via representation forgetting with parameter self-sharing. *IEEE Transactions on Information Forensics and Security*, 19:1099–1111, 2023.

[68] A. Warnecke, L. Pirch, C. Wressnegger, and K. Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

[69] J. Weng, S. Yao, Y. Du, J. Huang, J. Weng, and C. Wang. Proof of unlearning: Definitions and instantiation. *IEEE Transactions on Information Forensics and Security*, 19:3309–3323, 2024.

[70] G. Wu, M. Hashemi, and C. Srinivasa. Puma: Performance unchanged model augmentation for training data removal. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8675–8682, 2022.

[71] J. Xu, Z. Wu, C. Wang, and X. Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[72] S. Xu and T. Strohmer. Fair data representation for machine learning at the Pareto frontier. *Journal of Machine Learning Research*, 24(331):1–63, 2023.

[73] S. Xu, Y. Ni, S. Broecker, and T. Strohmer. Forgetting-MarI: LLM unlearning via marginal information regularization. *arXiv preprint arXiv:2511.11914*, 2025.

[74] L. Xue, S. Hu, W. Lu, Y. Shen, D. Li, P. Guo, Z. Zhou, M. Li, Y. Zhang, and L. Y. Zhang. Towards reliable forgetting: A survey on machine unlearning verification. *arXiv preprint arXiv:2506.15115*, 2025.

[75] Y.-X. Xue, Y.-X. Luo, P. Wu, H.-S. Shi, L.-F. Xue, C. Chen, W.-L. Zhu, Z.-B. Ding, Y. Bao, J. Shi, D. H. Epstein, Y. Shaham, and L. Lu. A memory retrieval–extinction procedure to prevent drug craving and relapse. *Science*, 336(6078):241–245, 2012.

[76] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, Oxford, UK, 2018. IEEE.

[77] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[78] B. Zhang, Z. Chen, C. Shen, and J. Li. Verification of machine unlearning is fragile. In *International Conference on Machine Learning (ICML)*, 2024.

[79] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Definitions	Target	Enforcement mechanism	Notes/guarantees
Fairness: outcome statistical parity			
Statistical Parity [24, 77]	$\hat{Y} \perp Z$	Constraint at training or post-hoc calibration of selection rates	Distributional parity; potential utility drop if base rates differ.
Fairness: representation statistical parity			
Adversarial debiasing [25, 51, 79]	$h(X) \perp Z$	Predictor vs. adversary predicting Z from $h(X)$ or $\hat{Y}$	Empirical proxy to reduce $I(\cdot; Z)$; architecture-agnostic.
VFAE [50]	$h(X) \perp Z$	VAE with MMD penalty to match $p(h \mid Z{=}z)$ across z	Produces reusable "purged" latents.
HSIC / dCov regs. [47, 48]	$\mathrm{dCov}\big(h(X), Z\big) = 0$	Add HSIC or distance-covariance penalties during training	Nonparametric dependence control; task-agnostic.
Concept erasure: representation independence			
INLP [56]	$\Sigma_{hZ} = \mathbf{0}$	Iterate: train linear Z-probe, project $h(X)$ to probe nullspace	No linear predictor of Z from $h(X)$; nonlinear leakage may remain.
LACE / R–LACE [57]	$\Sigma_{hZ} = \mathbf{0}$	Linear minimax/convex program to erase concept subspace	Improves optimality/control vs. heuristic projections.
LEACE [8]	$\Sigma_{hZ} = \mathbf{0}$	Closed-form least-squares projection	Blocks all *linear* Z-probes; nonlinear leakage may remain.
Concept erasure: generative independence			
ESD [28]	$f_\theta(X) \perp C$	Fine-tune to a negative-guidance teacher (subtract conditional–unconditional score)	Local removal of concept c with utility preservation.
All-but-One [38]	$f_\theta(X) \perp C$	Surgical update of guidance term with preservation constraints	Improves fidelity vs. naive fine-tuning.
Receler [41]	$f_\theta(X) \perp C$	Lightweight eraser adapters + regularization	Robust/local erasure; preserves base weights.

Table 3: Representative methods grouped by *unlearning/removal definition* (mathematical form shown) and *enforcement*. Here $h(X)$ denotes a learned representation. Linear erasure is expressed as cross-covariance independence $\Sigma_{hZ} = \mathbf{0}$ (with a centered encoding of Z); generative concept removal aims at independence in the model distribution $p_\theta(\cdot)$.

A. Appendix: Supplementary material for Section 1

A.1 Feature Unlearning

Relation to other notions. Equalized Odds post-processing [34] requires conditional independence $\hat{Y} \perp Z \mid Y$ (parity of errors). This may still permit $I(\hat{Y}; Z) > 0$, so it is insufficient for unlearning at the point of exposure. Causal notions (e.g., counterfactual or path-specific fairness) formalize invariance under interventions on Z within a structural causal model and are not equivalent to observational independence; they are typically enforced via functional restrictions or constraints on path-specific effects [45, 53, 15]. For

auditing, *amnesic probing* assesses whether removing targeted information from representations changes behavior [26]; recent diffusion benchmarks stress-test leakage and collateral damage after erasure [5].

A.2 Data Unlearning

Why anchor-based unlearning is unauditable, and why a verifiable definition is needed. Anchor-based frameworks define unlearning via equivalence to a theoretical retrain-on-retain *anchor*. However, because this anchor is typically unobserved, strict auditing based solely on black-box behavior is theoretically impossible [64]. In practice, existing verification schemes have proven *fragile*: Zhang et al. [78] demonstrate that adversarial training procedures can satisfy both backdoor-based and reproducing verification metrics while retaining target information. While early frameworks establish feasibility, they rely heavily on *auxiliary instrumentation*, such as seeded randomness, gradient traces, provenance metadata, or planted canaries, rather than observable properties of the released model [63]. Similarly, cryptographic and TEE-based proofs target verifiability but necessitate trusted hardware and impose significant operational overhead [69]. Recent surveys underscore that robust verification remains a central open challenge [74, 71]. Moreover, naive workflows (e.g., comparing pre- and post-unlearning models) can inadvertently leak membership information [14]. These systemic limitations motivate black-box, output-level definitions, such as our proposed marginal unlearning principle, that are directly auditable from observable model behavior.

B. Appendix: Supplementary material for Section 2

B.1 Proof of Lemma 2.1

Proof Since $\mathcal{L}(\widehat{Y}_{\text{margin}}) = M$, the chain rule for KL gives

$$I(\widehat{Y}_{\text{margin}}; Z) = \mathrm{D}_{\mathrm{KL}}\big(\mathcal{L}(\widehat{Y}_{\text{margin}}, Z) \parallel M \otimes \mathcal{L}(Z)\big) = \mathbb{E}_Z\Big[\mathrm{D}_{\mathrm{KL}}\big(\mathcal{L}(\widehat{Y}_{\text{margin}} \mid Z) \parallel M\big)\Big],$$

which equals $(1-\pi)\,\mathrm{D}_{\mathrm{KL}}(P_0\|M) + \pi\,\mathrm{D}_{\mathrm{KL}}(P_1\|M) = \mathrm{JSD}_\pi(P_0, P_1)$, proving the identity.

For TV, use triangle inequality and Pinsker on a general measurable space:

$$\mathrm{TV}(P_0, P_1) \;\le\; \mathrm{TV}(P_0, M) + \mathrm{TV}(P_1, M) \;\le\; \sqrt{\tfrac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(P_0\|M)} + \sqrt{\tfrac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(P_1\|M)}.$$

Let $a = (1-\pi)\,\mathrm{D}_{\mathrm{KL}}(P_0\|M)$ and $b = \pi\,\mathrm{D}_{\mathrm{KL}}(P_1\|M)$ so that $a + b = I(\widehat{Y}_{\text{margin}}; Z)$. It then follows from Cauchy–Schwarz that

$$\begin{aligned}
\sqrt{\mathrm{D}_{\mathrm{KL}}(P_0\|M)} + \sqrt{\mathrm{D}_{\mathrm{KL}}(P_1\|M)} &= \frac{\sqrt{a}}{\sqrt{1-\pi}} + \frac{\sqrt{b}}{\sqrt{\pi}} \\
&\le \sqrt{\Big(\frac{1}{1-\pi} + \frac{1}{\pi}\Big)(a+b)} \\
&= \sqrt{\frac{I(\widehat{Y}_{\text{margin}}; Z)}{\pi(1-\pi)}}.
\end{aligned}$$

Therefore, we obtain

$$\mathrm{TV}(P_0, P_1) \le \sqrt{\frac{I(\widehat{Y}_{\text{margin}}; Z)}{2\pi(1-\pi)}}.$$

■

B.2 Proof of Theorem 2.1

Proof Write $\mu_\Theta^p := \mathcal{L}(\Theta(X))$ for $X \sim p$ independent of Θ. Let $Q_h(\cdot \mid x)$ denote the predictive distribution of model h at input x, and let $\eta_x := \mathcal{L}(Y \mid X = x)$ be the true label distribution under the retain set.

Step 1: Triangle decomposition at p^d. By the triangle inequality for total variation distance:

$$\mathrm{TV}\big(\mu_{\Theta_u}^{p^d}, \mu_{\Theta_r}^{p^d}\big) \;\le\; \underbrace{\mathrm{TV}\big(\mu_{\Theta_u}^{p^d}, \mu_{\Theta_u}^{p^r}\big)}_{\text{(I)}} + \underbrace{\mathrm{TV}\big(\mu_{\Theta_u}^{p^r}, \mu_{\Theta_r}^{p^r}\big)}_{\text{(II)}} + \underbrace{\mathrm{TV}\big(\mu_{\Theta_r}^{p^r}, \mu_{\Theta_r}^{p^d}\big)}_{\text{(III)}}.$$

Step 2: Utility alignment (Term II). We bound the distance between the unlearned model and the anchor on the retain distribution p^r. By the joint convexity of TV and the triangle inequality relative to the ground truth η_x:

$$\begin{aligned}
\mathrm{TV}\big(\mu_{\Theta_u}^{p^r}, \mu_{\Theta_r}^{p^r}\big) &\le \mathbb{E}_{\Theta_u, \Theta_r} \int \mathrm{TV}\big(Q_{\Theta_u}(\cdot \mid x),\, Q_{\Theta_r}(\cdot \mid x)\big)\, p^r(dx) \\
&\le \mathbb{E}_{\Theta_u, \Theta_r} \int \Big[\mathrm{TV}\big(Q_{\Theta_u}(\cdot \mid x), \eta_x\big) + \mathrm{TV}\big(Q_{\Theta_r}(\cdot \mid x), \eta_x\big)\Big]\, p^r(dx).
\end{aligned}$$

Applying Pinsker's inequality pointwise, $\mathrm{TV}(Q,\eta) \le \sqrt{\frac{1}{2}\,\mathrm{D_{KL}}(\eta\|Q)}$. Using Jensen's inequality (concavity of the square root) to move the expectation inside:

$$\begin{aligned}\mathbb{E}_{\Theta_u}\int \mathrm{TV}\big(Q_{\Theta_u},\eta_x\big)\,p^r(dx) &\le \sqrt{\tfrac{1}{2}\,\mathbb{E}_{\Theta_u}\int \mathrm{D_{KL}}\big(\eta_x\|Q_{\Theta_u}(\cdot\mid x)\big)\,p^r(dx)}\\ &= \sqrt{\tfrac{1}{2}\,\overline{\mathrm{reg}}_{\log}(\Theta_u)} \;\le\; \sqrt{\tfrac{\delta}{2}}.\end{aligned}$$

A symmetric bound applies to Θ_r with δ_g. Summing these yields:

$$\text{(II)} \;\le\; \sqrt{\tfrac{1}{2}}\,\big(\sqrt{\delta}+\sqrt{\delta_g}\big).$$

Step 3: Membership independence (Term I). Since

$$\mathcal{L}(X_{\mathrm{margin}}\mid Z=0)=p^d, \qquad \mathcal{L}(X_{\mathrm{margin}}\mid Z=1)=p^r,$$

and $\widehat{Y}_{\mathrm{margin}} = f_{\Theta_u}(X_{\mathrm{margin}})$, we have

$$\mathcal{L}(\widehat{Y}_{\mathrm{margin}}\mid Z=0)=\mu_{\Theta_u}^{p^d}, \qquad \mathcal{L}(\widehat{Y}_{\mathrm{margin}}\mid Z=1)=\mu_{\Theta_u}^{p^r}.$$

These are mixture laws obtained after averaging over the internal randomness Θ_u. Applying Lemma 2.1 directly to $\widehat{Y}_{\mathrm{margin}}$ gives

$$\text{(I)} = \mathrm{TV}\big(\mu_{\Theta_u}^{p^d},\mu_{\Theta_u}^{p^r}\big) \le \sqrt{\frac{I(\widehat{Y}_{\mathrm{margin}};Z)}{2\pi(1-\pi)}} \le \sqrt{\frac{\varepsilon_u}{2\pi(1-\pi)}}.$$

Step 4: Anchor generalization (Term III). The term $\text{(III)} = \mathrm{TV}\big(\mu_{\Theta_r}^{p^r},\mu_{\Theta_r}^{p^d}\big)$ is the inherent distribution shift of the anchor model and is left explicit.

Finally, combining Steps 2, 3, and 4 into Step 1 completes the proof. ■

B.3 Proof of Lemma 2.2

Proof Assume that the first alternative fails. Then there exists $y^* \in \widehat{\mathcal{Y}}$ such that

$$\left|\log\frac{p_1(y^*)}{p_0(y^*)}\right| > \varepsilon.$$

In particular, the two densities differ at y^*, so

$$|p_1(y^*)-p_0(y^*)|>0.$$

Define

$$\gamma(y^*) := \frac{1}{2}|p_1(y^*)-p_0(y^*)|>0.$$

Since $\widehat{\mathcal{Y}}$ is open and $y^* \in \widehat{\mathcal{Y}}$, there exists $r_0 > 0$ such that

$$B_{r_0}(y^*) \subset \widehat{\mathcal{Y}}.$$

By continuity of p_0 and p_1, equivalently of $p_1 - p_0$, there exists $\rho_*(y^*) \in (0, r_0]$ such that the sign of $p_1 - p_0$ is constant on $B_{\rho_*(y^*)}(y^*)$, and

$$|p_1(y) - p_0(y)| \geq \gamma(y^*), \qquad y \in B_{\rho_*(y^*)}(y^*).$$

For notational simplicity, write $\rho_* := \rho_*(y^*)$. Now define the 1-Lipschitz test function

$$\varphi(y) := (\rho_* - \|y - y^*\|)_+.$$

By Kantorovich-Rubinstein duality,

$$\begin{aligned}
\mathcal{W}_{d_{\widehat{\mathcal{Y}}}}(f(X_0), f(X_1)) &\geq \left| \int_{\widehat{\mathcal{Y}}} \varphi(y)(p_1(y) - p_0(y))\, dy \right| \\
&\geq \gamma(y^*) \int_{B_{\rho_*}(y^*)} (\rho_* - \|y - y^*\|)\, dy \\
&= \gamma(y^*) \frac{\omega_m \rho_*^{m+1}}{m+1}.
\end{aligned}$$

On the other hand, since f is L-Lipschitz,

$$\mathcal{W}_{d_{\widehat{\mathcal{Y}}}}(f(X_0), f(X_1)) \leq L\, \mathcal{W}_{d_{\mathcal{X}}}(X_0, X_1).$$

Combining the two estimates and using $\mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0) > 0$, we obtain

$$L \geq \frac{\gamma(y^*)\, \omega_m\, \rho_*^{m+1}}{(m+1)\, \mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0)} = L^*(\varepsilon, y^*).$$

Therefore, if the likelihood-ratio bound fails, then the Lipschitz constant must satisfy $L \geq L^*(\varepsilon, y^*)$. ∎

C. Appendix: Supplementary material for Section 3

C.1 Utility motivation

Mutual information is a widely used quantification of the common information shared by two random variables, with useful coding interpretation in the discrete finite-entropy setting. In particular, given a data set X with a goal to compress X by an encoding $\hat{X}$, the volume of code needed to encode X is $2^{H(X)}$ where $H(X)$ is the entropy of X. Furthermore, after observing $\hat{X}$, the average volume of X mapped to individual realizations of $\hat{X}$ is equal to $2^{H(X|\hat{X})}$. Here,

$$H(X \mid \hat{X}) := -\sum_{\hat{x}} p(\hat{x}) \sum_{x} p(x \mid \hat{x}) \log p(x \mid \hat{x}).$$

is the conditional entropy of X on $\hat{X}$. Intuitively, a higher conditional entropy means more volume of X is expected to be mapped to individual realizations of $\hat{X}$, which implies more

randomness of X remains given the observation of $\hat{X}$. In other words, less X is explained by $\hat{X}$.

Since the volume of code for X is $2^{H(X)}$ and the average volume of code mapped to each $\hat{X}$ is $2^{H(X|\hat{X})}$, the average cardinality of the partition generated by the values of $\hat{X}$ on the values of X is the ratio:

$$\frac{2^{H(X)}}{2^{H(X|\hat{X})}} = 2^{I(X;\hat{X})}. \tag{23}$$

Here, $I(X;\hat{X}) = H(X) - H(X|\hat{X})$ is the mutual information between X and $\hat{X}$. On the one hand, higher mutual information implies that $\hat{X}$ generates a partition with higher cardinality (or usually finer partition) on X, which further implies more common information is shared between X and $\hat{X}$. On the other hand, from a data compression perspective, lower mutual information means $\hat{X}$ generates a partition on X with lower cardinality, which further implies a better data compression rate, because $\hat{X}$ can compress X into a partition of smaller cardinality.

Based on the motivation provided by coding interpretation from the discrete finite-entropy setting, as discussed in Section 2, we adopt mutual information to quantify the common information and compression rate between random variables. For unlearning quality purposes, we hope to maintain as much information of X as possible in generating $\hat{X}$. Therefore, to maximize utility, we should maximize mutual information $I(X;\hat{X})$ or, equivalently, minimize the information loss from from X to $\hat{X}$.

C.2 Admissibility

Since we are unlearning the information of Z by compressing dataset or variable pair (X, Z), it is natural to require the resulting compressed data $\hat{X}$ to be measurable with respect to (X, Z). Intuitively, the compression output $\hat{X}$ should have its "root" from (X, Z) without introducing additional randomness by the compression map f itself. Technically speaking, the "root" here means that for every event or observation A of the compressed $\hat{X}$, the information or pre-image represented by the observation $\hat{X}^{-1}(A) := \{\omega : \hat{X}(\omega) \in A\}$ comes from the knowledge of $f^{-1}(A) := \{(x, z) : f(x, z) \in A\}$ based on (X, Z):

$$\begin{aligned} \{\omega : \hat{X}(\omega) \in A\} &= \hat{X}^{-1}(A) \\ &= (X, Z)^{-1}(f^{-1}(A)) \\ &= \{\omega : (X(\omega), Z(\omega)) \in f^{-1}(A)\}. \end{aligned} \tag{24}$$

Here, $\omega \in \Omega$ is the smallest unit of information we can have from the measure space $(\Omega, \mathcal{F}, \mathbb{P})$. From a probability-theoretical perspective, since $\hat{X}$ is a deterministic compression of (X, Z), it generates a coarser partition (or, more technically, sigma-algebra) than the original information (X, Z) and we say $\hat{X}$ is measurable with respect to (X, Z), denoted by

$$\sigma(\hat{X}) \subset \sigma(X, Z). \tag{25}$$

This is equivalent to the existence of a $\mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Z}}/\mathcal{B}_{\mathcal{X}}$-measurable map, denoted by f, such that $\hat{X} = f(X, Z)$. That is, our admissibility is equivalent to the assumption that the data compression process does not create information or randomness by itself. Therefore, we

define the admissible unlearning outcome in our framework as follows:

$$\mathcal{A}(X,Z) \;:=\; \Big\{\hat{X} = f(X,Z) : f \text{ is } \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Z}}/\mathcal{B}_{\mathcal{X}}\text{-measurable}\Big\}, \tag{26}$$

and we use $\hat{X} = f(X,Z)$ and $\hat{X} \in \mathcal{A}(X,Z)$ interchangeably. This is the deterministic admissibility class. If randomized mechanisms are allowed, the deterministic class can be enlarged to mechanisms of the form $\hat{X} = f(X,Z,\Theta)$, where the internal seed $\Theta \perp (X,Z)$ is not released, and all information quantities are computed after averaging over Θ. In this current work, we focus on the deterministic admissible compressions to better demonstrate the theoretical intuition behind the framework, and defer a detailed study of the randomized compression to future work.

C.3 Details on considered utility quantifications

In practical machine unlearning, the utility of unlearning may need to be evaluated with respect to a different target variable Y rather than the original dataset X. Moreover, mutual information, while often a natural choice, is not the only metric for quantifying the relationship between the unlearning outcome $\hat{X}$ and the target variable Y. To accommodate diverse objectives, we extend our framework to the general formulation:

$$\inf_{\hat{X}=f(X,Z)} \big\{\mathcal{C}(Y;\hat{X},Z) : \hat{X} \perp Z\big\}. \tag{27}$$

where $\mathcal{C}(Y;\hat{X},Z)$ represents a utility cost or risk quantification, and the constraint $\hat{X} \perp Z$ ensures that the unwanted information Z is fully removed from $\hat{X}$. Below, we introduce several commonly used cost objectives and their corresponding constrained optimization problems, for which we provide a unified analytic feature unlearning solution in Section 4.3.

Mutual information maximization: The utility is defined as $\mathcal{C}(Y;\hat{X},Z) = -I(Y;\hat{X},Z)$, measuring the shared information between the target variable Y and the unlearning outcome $\hat{X}$. This objective is widely applied in classification methods such as decision trees [19] and in deep learning techniques, including Deep InfoMax [37] and information bottleneck methods [3]. The corresponding optimization problem, *Mutual-Information-Maximized Feature Unlearning*, is given by

$$\sup_{\hat{X}=f(X,Z)} \big\{I(Y;\hat{X},Z) : \hat{X} \perp Z\big\}.$$

Since $I(Y;\hat{X},Z) = H(Y) - H(Y \mid \hat{X},Z)$, this problem is equivalent to minimizing the conditional entropy $H(Y|\hat{X},Z)$.

Posterior KL-divergence cost: The cost is

$$\mathcal{C}(Y;\hat{X},Z) = -\mathbb{E}\Big[D_{\mathrm{KL}}\big(\mathbb{P}(Y \mid \hat{X},Z) \,\|\, \mathbb{P}(Y)\big)\Big],$$

where D_{KL} measures the divergence between the posterior and prior distributions of Y. KL-divergence objectives are widely used in information-theoretic learning and generative modeling, including variational methods such as VAEs [36]. The corresponding optimization problem, *Posterior-KL-Minimized Feature Unlearning*, is formulated as

$$\inf_{\hat{X}=f(X,Z)} \Big\{-\mathbb{E}\Big[D_{\mathrm{KL}}\big(\mathbb{P}(Y \mid \hat{X},Z) \,\|\, \mathbb{P}(Y)\big)\Big] : \hat{X} \perp Z\Big\}.$$

This problem seeks to make $\mathbb{P}(Y|\hat{X}, Z)$ as deterministic as possible relative to the prior $\mathbb{P}(Y)$, thereby enhancing the predictive power of $(\hat{X}, Z)$ for Y. Provided the quantities are well-defined, the expected posterior KL equals $I(Y; \hat{X}, Z)$. Thus, minimizing this cost is equivalent to maximizing the informativeness of the representation $(\hat{X}, Z)$ about Y.

Conditional probability energy maximization: The utility is defined as the L^2-norm of the conditional probability $\mathbb{P}(Y \in A|\hat{X}, Z)$ for classification or the negative mean squared error (MSE) for regression:

$$\mathcal{C}(Y; \hat{X}, Z) = \begin{cases} -\left\|\mathbb{P}(Y \in A \mid \hat{X}, Z)\right\|_{L^2}^2 & \text{for classification-type events } A \in \mathcal{B}_{\mathcal{Y}}, \\ -\left\|\mathbb{E}(Y \mid \hat{X}, Z)\right\|_{L^2}^2 & \text{regression-type targets } Y \in L^2(\Omega; \mathcal{Y}). \end{cases}$$

The corresponding optimization problem, *Energy-Maximized Feature Unlearning*, is formulated as $\sup_{\hat{X}=f(X,Z)}\{||\mathbb{P}(Y \in A \mid \hat{X}, Z)||_{L^2}^2 : \hat{X} \perp Z\}$. A higher L^2-norm indicates a more precise prediction of the event $\{Y \in A\}$ based on $(\hat{X}, Z)$, leading to reduced Bayes error and improved decision boundaries. Similarly, in the regression setting: $\sup_{\hat{X}=f(X,Z)}\{||\mathbb{E}(Y \mid \hat{X}, Z)||_{L^2}^2 : \hat{X} \perp Z\}$, a smaller conditional-mean-energy cost corresponds to larger explained variance and smaller mean-squared prediction error.

As we show in Section 4.3, when the above-listed objectives are applied, there exists a universal optimal feature unlearning solution to the general formulation: equation (27) for arbitrary target variables Y.

C.4 Formulation of constrained optimization problems

Our goal here is to provide theoretical solutions to the following constrained optimization problems under mild assumptions, thereby developing a unified mathematical framework for machine unlearning of features and labels under various utility objectives:

Problem 1 (Conditional-entropy-minimized feature unlearning)

$$\inf_{\hat{X} \in \mathcal{A}(X,Z)} \left\{H(Y \mid \hat{X}, Z) : \hat{X} \perp Z\right\}.$$

In many cases, an unlearning output $\hat{X}$ may be used to generate inferences or predictions for some random variable Y. Thus, it is desirable to solve Problem 1 for such a target variable Y. Notice that, due to $I(Y; \hat{X}, Z) = H(Y) - H(Y|\hat{X}, Z)$, the above problem shares the same solution as the maximization of mutual information between Y and $(\hat{X}, Z)$:

Problem 2 (Mutual-information-maximized feature unlearning)

$$\sup_{\hat{X} \in \mathcal{A}(X,Z)} \left\{I(Y; \hat{X}, Z) : \hat{X} \perp Z\right\}.$$

Notably, the optimal solution to Problems 1 and 2 does not depend on the specific choice of Y due to the monotonicity of the functional $H(Y|\cdot)$ with respect to the sigma-algebra generated by $(\hat{X}, Z)$. Thus, despite the explicit presence of Y in Problem 1, it provides a generalized solution for any choice of Y.

Problem 3 (Posterior-divergence cost for feature unlearning)

$$\inf_{\hat{X}\in\mathcal{A}(X,Z)} \left\{ -\mathbb{E}\Big[D_{\mathrm{KL}}\big(\mathbb{P}(Y \mid \hat{X}, Z) \| \mathbb{P}(Y)\big)\Big] : \hat{X} \perp Z \right\}.$$

Given a variable of interest denoted by Y, a general downstream machine learning or AI task may aim to estimate the conditional probability using the unlearning outcome $(\hat{X}, Z)$. Therefore, it is desirable to make $\mathbb{P}(Y|\hat{X}, Z)$ as deterministic as possible relative to the original distribution of Y. To quantify this determinism, we use the KL-divergence of $\mathbb{P}(Y|\hat{X}, Z)$ relative to $\mathbb{P}(Y)$, leading to the optimization problem above. Intuitively, a more accurate prediction of $\mathbb{P}(Y|\hat{X}, Z)$ implies less randomness relative to $\mathbb{P}(Y)$, which increases the KL-divergence of $\mathbb{P}(Y|\hat{X}, Z)$ relative to $\mathbb{P}(Y)$.

Problem 4 (Energy-maximized feature unlearning) *For a measurable target $Y : \Omega \to \mathcal{Y}$ and event $A \in \mathcal{B}_{\mathcal{Y}}$, consider*

$$\sup_{\hat{X}\in\mathcal{A}(X,Z)} \left\{ \big\|\mathbb{P}(Y \in A \mid \hat{X}, Z)\big\|_{L^2}^2 : \hat{X} \perp Z \right\}.$$

For regression-type targets $Y \in L^2(\Omega; \mathcal{Y})$, with $\mathcal{Y}$ a Hilbert space, one may instead consider

$$\sup_{\hat{X}\in\mathcal{A}(X,Z)} \left\{ \big\|\mathbb{E}(Y \mid \hat{X}, Z)\big\|_{L^2}^2 : \hat{X} \perp Z \right\}.$$

Finally, from the perspective of conditional probability estimation, for a given Y and event $A \in \mathcal{B}_{\mathcal{Y}}$, it is natural to maximize the energy (or equivalently, the L^2 norm) of the conditional probability $\mathbb{P}(\{Y \in A\}|\hat{X}, Z)$. Here, a larger L^2 norm of the conditional probability indicates a more precise prediction of the event $\{Y \in A\}$ based on the information provided by $\hat{X}$.

D. Appendix: Supplementary material for Section 4

D.1 Wasserstein distance and barycenter

Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of all the probability measures on $\mathbb{R}^d$,

$$\mathcal{W}_2(\mu,\nu) := \left(\inf_{\lambda \in \Pi(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} ||x_1 - x_2||^2 d\lambda(x_1, x_2) \right\} \right)^{\frac{1}{2}}.$$

Here, $\Pi(\mu,\nu) := \{\pi \in \mathcal{P}((\mathbb{R}^d)^2) : \int_{\mathbb{R}^d} d\pi(\cdot, v) = \mu, \int_{\mathbb{R}^d} d\pi(u, \cdot) = \nu\}$. $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is called the Wasserstein space, where $\mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} ||x||^2 d\mu < \infty \right\}$. Also, we use $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ to denote the set of probability measures with finite second moments and are absolute continuous w.r.t. the Lebesgue measure. To simplify notation, we often denote

$$\mathcal{W}_2(X_1, X_2) := \mathcal{W}_2(\mathcal{L}(X_1), \mathcal{L}(X_2)),$$

where $\mathcal{L}(X) := \mathbb{P} \circ X^{-1} \in \mathcal{P}(\mathbb{R}^d)$ is the law or distribution of X, $X : \Omega \to \mathcal{X} := \mathbb{R}^d$ is a random variable (or vector) with an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Intuitively, one can consider the Wasserstein distance as L^2 distance after optimally coupling two random variables whose distributions are μ and ν. That is, if the pair (X_1, X_2) is an optimal coupling [65], then

$$\mathcal{W}_2(X_1, X_2) = ||X_1 - X_2||_{L^2} = \left(\int_\Omega \|X_1(\omega) - X_2(\omega)\|^2 \, d\mathbb{P}(\omega) \right)^{1/2}.$$

Given $\{\mu_z\}_{z \in \mathcal{Z}} \subset \mathcal{P}_2(\mathbb{R}^d)$ and weights $\lambda \in \mathcal{P}(\mathcal{Z})$, their Wasserstein barycenter [2] is any

$$\bar{\mu} \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mu) \, d\lambda(z) \right\}.$$

If there is no danger of confusion, we will refer to the Wasserstein barycenter simply as barycenter. In the feature-unlearning setting, let

$$\mu_z := \mathcal{L}(X \mid Z = z),$$

and let $\bar{\mu}$ be a barycenter of $\{\mu_z\}_{z \in \mathcal{Z}}$. When $\mu_z \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, Brenier's theorem gives an optimal transport map $T_z : \mu_z \to \bar{\mu}$. We define the barycentric representation by

$$\bar{X} := T_Z(X).$$

Then, conditionally on $Z = z$,

$$\mathcal{L}(\bar{X} \mid Z = z) = (T_z)_{\#}\mu_z = \bar{\mu}.$$

Hence $\bar{X} \perp Z$. Moreover, under the absolute-continuity assumptions used below, the Brenier maps are a.e. invertible on their supports, with inverse maps given by the corresponding Brenier maps from $\bar{\mu}$ back to μ_z.

D.2 Proof of Lemma 4.1

Proof To start, let $\hat{X} := \hat{X}_{\text{margin}}$ to simplify notation, and notice that it follows from the construction of dataset or variable pair (X_{margin}, Z) that $\{Z = 0\} = \{X_{\text{margin}} = X_0\}$ and $\{Z = 1\} = \{X_{\text{margin}} = X_1\}$. Also, we have

$$|\mathbb{P}(Z = 0|\hat{X}) - \mathbb{P}(Z = 1|\hat{X})| = |2\mathbb{P}(Z = 0|\hat{X}) - 1| \tag{28}$$

$$= 2|\mathbb{P}(Z = 0|\hat{X}) - \frac{1}{2}| \tag{29}$$

$$= 2||\mathbb{P}(Z|\hat{X}) - \mathbb{P}(Z)||_{TV}, \tag{30}$$

where the third line follows from the definition of total variation distance and the prior information $\mathbb{P}(Z = 0) = \frac{1}{2}$. By taking the expectation over $\hat{X}$, we have

$$\begin{aligned}
\mathbb{E}_{\hat{X}}(|\mathbb{P}(Z = 0|\hat{X}) - \mathbb{P}(Z = 1|\hat{X})|) &= 2\mathbb{E}_{\hat{X}}(||\mathbb{P}(Z|\hat{X}) - \mathbb{P}(Z)||_{TV}) \\
&\leq 2\mathbb{E}_{\hat{X}}\left(\sqrt{\frac{1}{2}D_{\text{KL}}(\mathbb{P}(Z|\hat{X})||\mathbb{P}(Z))}\right) \\
&\leq 2\sqrt{\frac{1}{2}\mathbb{E}_{\hat{X}}\left(D_{\text{KL}}(\mathbb{P}(Z|\hat{X})||\mathbb{P}(Z))\right)} \\
&= 2\sqrt{\frac{1}{2}I(Z;\hat{X})}.
\end{aligned}$$

Here, the second line follows from Pinsker's inequality, the third from Jensen's inequality, and the fourth from the definition of mutual information. Now, for any fixed $\varepsilon \in (0, 1)$, it follows from Markov's inequality that

$$\mathbb{P}\left(\{|\mathbb{P}(Z = 0|\hat{X}) - \mathbb{P}(Z = 1|\hat{X})| \leq \varepsilon\}\right) \geq 1 - \frac{2}{\varepsilon}\left(\sqrt{\frac{1}{2}I(Z;\hat{X})}\right).$$

Finally, it follows from

$$|\mathbb{P}(Z = 0|\hat{X}) - \mathbb{P}(Z = 1|\hat{X})| \leq \varepsilon \Longrightarrow \log\left(\frac{\mathbb{P}(Z = 0 \mid \hat{X})}{\mathbb{P}(Z = 1 \mid \hat{X})}\right) \leq \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)$$

that

$$\left\{|\mathbb{P}(Z = 0|\hat{X}) - \mathbb{P}(Z = 1|\hat{X})| \leq \varepsilon\right\} \subset \left\{|\log\left(\frac{\mathbb{P}(Z = 0 \mid \hat{X})}{\mathbb{P}(Z = 1 \mid \hat{X})}\right)| \leq \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\right\},$$

and

$$\begin{aligned}
\mathbb{P}\left(\{|\log\left(\frac{\mathbb{P}(Z = 0 \mid \hat{X})}{\mathbb{P}(Z = 1 \mid \hat{X})}\right)| \leq \log(\frac{1+\varepsilon}{1-\varepsilon})\}\right) &\geq \mathbb{P}\left(\{|\mathbb{P}(Z = 0|\hat{X}) - \mathbb{P}(Z = 1|\hat{X})| \leq \varepsilon\}\right) \\
&\geq 1 - \frac{2}{\varepsilon}\left(\sqrt{\frac{1}{2}I(Z;\hat{X})}\right).
\end{aligned}$$

That completes the proof. ■

D.3 Proof of Lemma 4.2

Proof We provide the argument for the lifted statement. By the absolute-continuity assumptions, the barycenter $\bar{P}$ is absolutely continuous, and Brenier's theorem yields optimal maps $T_z : P_z \to \bar{P}$. Moreover, the corresponding Brenier maps are a.e. invertible on their supports. Denote the a.e. inverse of T_z by $S_z : \bar{P} \to P_z$. Thus,

$$S_z(T_z(x)) = x \qquad P_z\text{-a.e.}$$

Since $\bar{X} = T_Z(X)$, the pair $(\bar{X}, Z)$ is a measurable function of (X, Z). Hence,

$$\sigma(\bar{X}, Z) \subset \sigma(X, Z).$$

Conversely, by the a.e. inverse relation, we also have

$$X = S_Z(\bar{X}), \qquad \mathbb{P}\text{-a.s..}$$

Therefore (X, Z) is a measurable function of $(\bar{X}, Z)$, up to null sets, and hence

$$\sigma(X, Z) \subset \sigma(\bar{X}, Z).$$

Combining the two inclusions gives

$$\sigma(X, Z) = \sigma(\bar{X}, Z), \qquad \mathbb{P}\text{-a.s..}$$

Finally, let $\hat{X} = f(X, Z)$ be any admissible deterministic unlearning outcome. Then $(\hat{X}, Z)$ is a measurable function of (X, Z), so that

$$\sigma(\hat{X}, Z) \subset \sigma(X, Z).$$

Using the equality above, we obtain

$$\sigma(\hat{X}, Z) \subset \sigma(X, Z) = \sigma(\bar{X}, Z).$$

That completes the proof. ■

D.4 Sigma-algebra and information

In probability theory, a probability space is often represented as a triple $(\Omega, \Sigma, \mathbb{P})$, where Ω is the sample space, Σ is the sigma-algebra (a collection of subsets of Ω closed under complements and countable unions), and $\mathbb{P} : \Sigma \to [0, 1]$ is a probability measure that assigns probabilities to each event in Σ.

The same sample space can be associated with different sigma-algebras, resulting in different measurable spaces, and, after specifying a probability measure, different probability spaces. We say that a sigma-algebra Σ_1 is finer than Σ_2, denoted $\Sigma_2 \subset \Sigma_1$, if Σ_1 contains all events in Σ_2. Conversely, we say Σ_1 is coarser than Σ_2 if Σ_1 contains fewer events than Σ_2.

A random variable or random vector X is a measurable function from the probability space to $\mathbb{R}^d$ (or $\mathbb{C}^d$), $X : \Omega \to \mathbb{R}^d$. The sigma-algebra generated by X, denoted by $\sigma(X)$,

comprises all possible events in Ω that could be defined based on the image of X in $\mathbb{R}^d$ (or $\mathbb{C}^d$). Thus, if X generates a finer sigma-algebra than another variable X', denoted $\sigma(X') \subset \sigma(X)$, then X contains more events and, therefore, more information than X'.

In modern probability theory, sigma-algebras facilitate the construction of probability measures, especially in countably or uncountably infinite spaces (as the concept is trivial in finite spaces). They satisfy certain axioms, including countable additivity, that link the set algebra of events in the space to the algebra of their probabilities, particularly through continuity properties.

D.5 Proof of Lemma 4.3

Proof Assume that

$$\sigma(X_1) \subset \sigma(X_2).$$

Then X_1 is a measurable post-processing of X_2. Equivalently, under the standard Borel assumptions, there exists a measurable map g such that

$$X_1 = g(X_2) \qquad \mathbb{P}\text{-a.s.}$$

For mutual information, this implies the Markov chain $Y \to X_2 \to X_1$. Hence, by the data-processing inequality,

$$I(Y; X_1) \leq I(Y; X_2),$$

for arbitrary measurable Y, whenever the mutual informations are well-defined, possibly taking the value $+\infty$.

For the conditional entropy statement, the discrete finite-entropy case follows from

$$I(Y; X_s) = H(Y) - H(Y \mid X_s), \qquad s = 1, 2,$$

together with the mutual-information monotonicity above. In the continuous case, it follows from assumption that the conditional laws $\mathcal{L}(Y \mid X_s)$ admit densities $p_{Y|X_s}$, and the conditional differential entropies are finite. Since $\sigma(X_1) \subset \sigma(X_2)$, we have

$$I(Y; X_2 \mid X_1) \geq 0.$$

By the differential-entropy chain rule, we have

$$I(Y; X_2 \mid X_1) = h(Y \mid X_1) - h(Y \mid X_2).$$

Therefore,

$$h(Y \mid X_2) \leq h(Y \mid X_1).$$

Next, fix a measurable target $Y : \Omega \to \mathcal{Y}$ and an event $A \in \mathcal{B}_{\mathcal{Y}}$. Let $U := \mathbb{1}_{\{Y \in A\}}$. Since $\sigma(X_1) \subset \sigma(X_2)$, the tower property gives

$$\mathbb{E}(U \mid X_1) = \mathbb{E}(\mathbb{E}(U \mid X_2) \,|\, X_1)\,.$$

Using the L^2-contraction property of conditional expectation, we obtain

$$\big\|\mathbb{E}(U \mid X_1)\big\|_{L^2}^2 \leq \big\|\mathbb{E}(U \mid X_2)\big\|_{L^2}^2.$$

Since

$$\mathbb{E}(U \mid X_s) = \mathbb{P}(Y \in A \mid X_s), \qquad s = 1, 2,$$

it follows that

$$\big\|\mathbb{P}(Y \in A \mid X_1)\big\|_{L^2}^2 \le \big\|\mathbb{P}(Y \in A \mid X_2)\big\|_{L^2}^2.$$

Finally, let $Y \in L^2(\Omega; \mathbb{R}^d)$. Again, by the tower property,

$$\mathbb{E}(Y \mid X_1) = \mathbb{E}(\mathbb{E}(Y \mid X_2) \,|\, X_1)\,.$$

Applying the L^2-contraction property of conditional expectation gives

$$\big\|\mathbb{E}(Y \mid X_1)\big\|_{L^2}^2 \le \big\|\mathbb{E}(Y \mid X_2)\big\|_{L^2}^2.$$

That completes the proof. ■

D.6 Intuitive insights into Theorem 4.3

In feature unlearning, the released representation $\hat{X}$ is required to satisfy $\hat{X} \perp Z$, so that $\hat{X}$ itself does not expose the sensitive or unlearned feature Z. At the same time, the predictive relationship between the downstream target Y and the representation may be different across Z-slices. Therefore, measuring utility only through $I(Y; \hat{X})$ can underestimate the task information preserved within each slice. We instead evaluate utility through the lifted quantity $I(Y; \hat{X}, Z)$. Since

$$I(Y; \hat{X}, Z) = I(Y; Z) + I(Y; \hat{X} \mid Z),$$

and $I(Y; Z)$ is fixed by the data distribution, maximizing $I(Y; \hat{X}, Z)$ is equivalent to maximizing the within-slice task information $I(Y; \hat{X} \mid Z)$. Thus, the lifted formulation separates the two roles of Z: the released representation $\hat{X}$ must not reveal Z, while utility is evaluated slice-wise through $(\hat{X}, Z)$, as discussed in Section 2.1.

Under the assumptions of Lemma 4.2, the barycentric representation $\bar{X} = T_Z(X)$ satisfies

$$\sigma(\hat{X}, Z) \subset \sigma(\bar{X}, Z)$$

for every admissible representation $\hat{X} = f(X, Z)$. Thus, among all admissible feature-unlearning representations, $(\bar{X}, Z)$ preserves the finest lifted information available from (X, Z), while $\bar{X}$ itself is independent of Z. This lifted maximality has several consequences.

- *Utility is evaluated with the task context retained.* The constraint $\hat{X} \perp Z$ ensures that the released representation does not reveal Z. At the same time, the lifted pair $(\hat{X}, Z)$ retains the task context needed to evaluate utility. The theorem therefore identifies $\bar{X}$ as a representation whose release removes marginal information about Z, while whose lifted version $(\bar{X}, Z)$ preserves maximal task-relevant information.

- *Conditional uncertainty about the target is minimized.* For finite-valued targets Y, the sigma-algebra inclusion implies

$$H(Y \mid \bar{X}, Z) \le H(Y \mid \hat{X}, Z)$$

for every admissible $\hat{X} = f(X, Z)$. Thus, after the slice variable Z is included for utility evaluation, $\bar{X}$ leaves the least residual uncertainty about Y.

- *Lifted mutual information is maximized.* Since $(\bar{X}, Z)$ generates the finest lifted sigma-algebra among admissible representations, data processing gives

$$I(Y; \hat{X}, Z) \leq I(Y; \bar{X}, Z).$$

Hence, $\bar{X}$ preserves the largest amount of information about any downstream target Y when utility is evaluated in the lifted sense.

- *Posterior concentration is maximized.* The expected posterior divergence

$$\mathbb{E}\Big[D_{\mathrm{KL}}\Big(\mathbb{P}(Y \mid \hat{X}, Z) \,\|\, \mathbb{P}(Y)\Big)\Big]$$

equals $I(Y; \hat{X}, Z)$ whenever the quantities are finite. Therefore, the same lifted maximality means that the posterior law of Y given $(\bar{X}, Z)$ is, on average, the most concentrated relative to the prior law of Y.

- *Conditional probability and conditional mean energies are maximized.* For any event $A \in \mathcal{B}_{\mathcal{Y}}$,

$$\big\|\mathbb{P}(Y \in A \mid \hat{X}, Z)\big\|_{L^2}^2 \leq \big\|\mathbb{P}(Y \in A \mid \bar{X}, Z)\big\|_{L^2}^2.$$

Thus events generated by Y are predicted at least as sharply from $(\bar{X}, Z)$ as from any other admissible lifted representation. Similarly, for $Y \in L^2$,

$$\big\|\mathbb{E}(Y \mid \hat{X}, Z)\big\|_{L^2}^2 \leq \big\|\mathbb{E}(Y \mid \bar{X}, Z)\big\|_{L^2}^2.$$

Hence, $(\bar{X}, Z)$ also maximizes explained conditional-mean energy.

In summary, the theorem separates the two roles of Z: the released representation $\hat{X}$ must remove marginal information about Z, while the lifted pair $(\hat{X}, Z)$ is used to measure how much task-relevant information is preserved on each of the Z-slices. The barycentric representation $\bar{X}$ is optimal because it satisfies both requirements: $\bar{X} \perp Z$, and $(\bar{X}, Z)$ is maximal in the lifted sigma-algebra order.