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. For feature unlearning, the framework applies to deep learning with arbitrary 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 finding.

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

Contents

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 based 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, allowing models to adapt without costly end-to-end retraining.

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 model outputs, even through proxies.
- **Marginal data-point unlearning (removing a record's footprint).** Under GDPR's "right to be forgotten" [26] 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 $X_r := X_{[i\notin u,:]}$.

[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.

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* as 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 retained 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 solution.** We formulate unlearning as a generalized version of the classic rate–distortion compression framework (see Section 3.2):

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

  Here, f is an ML/AI model that maps from the data space $\mathcal{X}$ to an output space[2] $\mathcal{S}$, Y is the target variable (which can be the dataset X itself in self-supervised learning), $S := f(X)$ is the learning outcome (e.g., predictions or a representation), $\mathcal{C}$ is a cost function which quantifies the learning loss via $\mathcal{C}(Y;S)$. S' denotes a particular form of the learning outcome which we will introduce later separately in the feature and data unlearning settings in Section 2, and Z denotes a feature attached to S' that one wants to forget. More specifically, the designed $I(S';Z)$ encodes unique feature information that we want the machine to forget in feature unlearning and the unique information contributed by the data set one wants to unlearn in data unlearning. We provide practical regularization algorithms for both settings (Algorithms 1 and 2, Section 5), with experiments on tabular and image data validating the approach.

- **Analytic solution and theoretical guarantee.** We first show that bounding the "compression rate" $I(S';Z)$ *directly* implies a high-probability guarantee for feature unlearning (Theorem 4.1) and data unlearning (Lemma 4.1, Theorem 4.2). This provides actionable guidance for tuning λ to meet a target ε while preserving learning utility. Furthermore, in the extreme case of a hard constraint $S' \perp Z$, we prove that under mild assumptions a *single* analytic solution solves the feature and data unlearning problems for certain utility cost functions: namely, the Wasserstein–2 barycenter of the conditional distributions $\{X \mid Z = z\}_z$ (Theorem 4.3). It generates the finest sigma-algebra among the admissible outcomes (Lemma 4.2), yielding maximal utility while enforcing $S' \perp Z$. Furthermore, we provide practical implementation of the analytic solution as Algorithm 3 in Section 5.3.

[2] Here, the output space can vary for different learning task, such as data space $\mathcal{X}$ for synthetic data or self-supervised learning, some latent space $\mathcal{V}$ for representation, or target space $\mathcal{Y}$ for supervised learning.

- **Exemplary numerical experiments.** We evaluate both feature unlearning and data-point 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 paper is organized as the following: Section 1 reviews related work and establishes notation. Section 2 formulates the principles for feature and data-point 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 we focus on 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, 28], 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 [31, 59], performance-preserving objectives like PUMA [69], and parameter scrubbing [29]. 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 [63].

In contrast, our *marginal unlearning* framework breaks this dependency. We propose an inference-based criterion that is checkable solely via observable outputs and the original dataset, providing a practical path to auditable unlearning while maintaining rigorous population-level guarantees. In parallel, a companion work Forgetting-MarI [72] 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 [67, 32].

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, 30, 71]. In computer vision, related *concept removal* methods aim to erase targeted concepts from generative or recognition models [27].

Information-theoretic formulations. Classical information-bottleneck-style compression or rate-distortion framework target *all* information about a point or attribute (e.g., minimizing $I(X;\hat{Y})$) [66, 32]. 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 [61, 75, 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) [33] 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 responsibility and longevity: fine-tuning *adds/adapts* knowledge (full or parameter-efficient updates such as adapters, prompt tuning, LoRA/QLoRA) [38, 54, 45, 39, 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 [42]. 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 [34]: 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 [57]. *Reconsolidation-update* procedures (retrieval–extinction) transiently destabilize reactivated memories and incorporate non-reinforcing information, attenuating the unique predictive impact of the original association [58, 52, 74]. Complementarily, *retrieval practice* strengthens desired knowledge and improves later performance on targeted content [41]. 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 base 2. The total variation distance is define 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 [65]

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

where $\Pi(\mu, \nu)$ is the set of couplings with $\lambda(\cdot, \mathcal{X}) = \mu$ and $\lambda(\mathcal{X}, \cdot) = \nu$. We write $\mathcal{P}_{2,ac}(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$ for measures that are absolutely continuous (w.r.t. Lebesgue when $\mathcal{X} \subset \mathbb{R}^d$) and have finite second moment. 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 [65]

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

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 provides an algorithmic implementation. Section 2.2 introduces the proposed marginal-unlearning principle for data-point unlearning, together with its formal definition and a practical implementation. 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 *necessity*: marginal unlearning is required for any well-regularized retrain-on-retain model.

For purposes of clarity and intuition, we state all the unlearning principles, definitions, and justifications in the canonical setting where the outcome space is the same as the target space, i.e., $\mathcal{S} = \mathcal{Y}$ and $S = \hat{Y}$. But we note that the principle, definitions, and justifications all apply to the generic setting $\mathcal{S} \in \{\mathcal{X}, \mathcal{V}, \mathcal{Y}\}$, where $\mathcal{X}$ is the the input data space in the setting of synthetic data generation or for self-supervised learning (i.e. $\mathcal{Y} = \mathcal{X}$) setting, $\mathcal{V}$ is some new latent variable space in the representation learning setting, and $\mathcal{Y}$ is the target space in the supervised learning setting.

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 membership/domain indicator) from what the model *exposes*, while preserving task utility on the distribution p^r of the retain set. Because exposure occurs through the released predictor's outputs, 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 retain loss $L(f_\theta) := \mathcal{C}(Y; \hat{Y})$ and $L^\star := \inf_\theta L(f_\theta)$. We quantify exposure of Z via the mutual information $I(\hat{Y}; Z)$ under p^r.

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 outputs. 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 [63]. 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 [76, 62, 68, 14]; see Appendix A.2. These limitations motivate an auditable criterion expressed solely in terms of observable outputs.

To address the inaudibility issue 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 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 [52].*
- 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 [41].*

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 set. The marginal effect is substantial only when p^d deviates significantly from p^r, indicating that the unlearning source p^u contributes 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

$Z \sim \text{Bernoulli}(\pi)$ 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. 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) *A model f satisfies ε-*marginal unlearning *if*

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

i.e., for every measurable event D, the posterior odds of "excluded" vs. "included" are bounded by $e^{\pm\varepsilon}$ given $\{\hat{Y}_{\text{margin}} \in D\}$.

Definition 2.3 is a worst-case (odds) guarantee across all the possible events in the learning output space. Notice that $I(\hat{Y}_{\text{margin}}; Z) = 0$ if and only if we have 0-marginal unlearning. Furthermore, it follows from the sharp inequality (3) that a larger $I(\hat{Y}_{\text{margin}}; Z)$ implies more distributional shift on the learning outcome resulted from adding unlearn set to retain set. Therefore, $I(\hat{Y}_{\text{margin}}; Z)$ forms a natural quantification of marginal information, and a natural information-theoretic relaxation of Definition 2.3 a constraint on $I(\hat{Y}_{\text{margin}}; Z)$:

Definition 2.4 (Mutual-information marginal unlearning with utility budget) *For tolerances $\varepsilon, \delta \ge 0$, 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 (ε, δ)-*marginal unlearning *if*

$$I\big(\hat{Y}_{\text{margin}}; Z\big) \;\le\; \varepsilon, \qquad \textit{and} \qquad L(f) - L^\star \;\le\; \delta. \tag{5}$$

When $\varepsilon = \delta = 0$, we obtain exact marginal unlearning*: $\hat{Y}_{\text{margin}} \perp Z$ 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 [72] for its implementation on LLMs.

We now reflect on the main motivations behind the definition:

- *Interpretability/practicality:* The posterior odds form (4) and the mutual information form (5) directly upper-bound an optimal adversary's ability to infer inclusion; the MI criterion aggregates leakage across events.
- *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];
- *The 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)[31];
- *The 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 $W \subseteq \mathcal{H}$ [59].

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 [63, 76]. 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:

$$D\big(\mathcal{L}(f_{\Theta_r}(X)), \mathcal{L}(f_{\Theta_u}(X))\big) \leq \varepsilon, \quad \text{for } X \sim p^d. \tag{6}$$

We now show that our *marginal unlearning* criterion (Def. 2.3) is sufficient to guarantee Eq. (6), 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

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

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 ⇒ 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 $Y_{\text{margin}} := f_{\Theta_u}(X_{\text{margin}})$ 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 B.2.

Anchor-based formulations capture the intuition "retrain on the retain set" but are 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: $\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2}\, \mathrm{D_{KL}}(P \| Q)}$. 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 by Bayes' rule,

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

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

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

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 $e^{-\varepsilon} \leq r \leq e^{\varepsilon}$ a.s., and conversely this bound implies the set-wise inequalities. Equivalently,

$$\sup_y \Big|\log \frac{\mathcal{L}(f(X_1))(y)}{\mathcal{L}(f(X_0))(y)}\Big| \;\leq\; \varepsilon \quad\iff\quad \varepsilon\text{-marginal unlearning.}$$

Lemma 2.2 (Marginal unlearning as a condition for "good" retraining) *Assume f is a retrained model, $\mathcal{L}(f(X_0))$ and $\mathcal{L}(f(X_1))$ admit continuous densities, and f is L-Lipschitz from $(\mathcal{X}, d_{\mathcal{X}})$ to $(\hat{\mathcal{Y}}, \|\cdot\|)$. Then either*

$$\sup_y \Big|\log \frac{\mathcal{L}(f(X_1))(y)}{\mathcal{L}(f(X_0))(y)}\Big| \;\leq\; \varepsilon,$$

or $L \geq L^*(\varepsilon)$, where for some y^* and all sufficiently small $\delta > 0$,

$$L^*(\varepsilon) \;:=\; \frac{\delta\,\big|f(X_1)(y^*) - f(X_0)(y^*)\big|}{\mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0)} \;=\; \frac{\delta\,\big(e^{\varepsilon} - 1\big)\,\min\{f(X_0)(y^*), f(X_1)(y^*)\}}{\mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0)}. \tag{7}$$

Here $X_1 \sim p^r$, $X_0 \sim p^d$, 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 yields a three-way incompatibility: a model cannot simultaneously have

1. **Leakage**: truthful revelation of the unlearn record (x_u, y_u) (i.e., $f(x_u) = y_u$);
2. **Marginal unlearning**: the ε log–ratio bound (Def. 2.3);
3. **Good generalizability**: small Lipschitz constant (regular f).

Thus, once (2) is required, any well-regularized retrained model cannot also exhibit (1).

Proposition 2.1 (Marginal unlearning reduces utility on the unlearned record) *Assume $f(X_1) = Y_1$, $\sup_y\big|\log\big(\frac{\mathcal{L}(f(X_0))(y)}{\mathcal{L}(f(X_1))(y)}\big)\big| \leq \varepsilon$ with $L^*(\varepsilon) > 1$, and there exists $y \in \hat{\mathcal{Y}}$ such that $\delta\,|Y_0(y) - Y_1(y)| > 2\,\mathcal{W}_{d_{\mathcal{X}}}(X_0, X_1)$. Then $f(X_0) \neq Y_0$.*

If a retrained model fits the remaining data ($f(X_1) = Y_1$) and satisfies ε-marginal unlearning with small ε ($L^*(\varepsilon) > 1$), but the unlearn record induces a significant marginal change ($\delta|Y_0(y) - Y_1(y)| > 2\mathcal{W}_{d_{\mathcal{X}}}(X_0, X_1)$), then the model cannot also fit the unlearned record ($f(X_0) \neq Y_0$). In particular $f(x_u) \neq y_u$. Enforcing marginal unlearning via an information-theoretic regularizer therefore concentrates utility loss where it should (on the unlearned portion), while preserving retain performance.

3. Information Theoretic Unlearning Framework

This section formalizes our unlearning framework and explains the rate–distortion (compression) motivation behind our framework. We first look at the proposed unlearning framework in its general format via an information-theoretic lens in Section 3.1, then state the optimization problems for feature unlearning and (marginal) data unlearning separately based on the respective unlearning definitions introduced in Section 3.2, and finally provide the intuition and motivation from data compression, more specifically rate-distortion theory, in Section 3.3.

3.1 An information-theoretic lens

We connect unlearning to classical rate-distortion/compression ideas [60, 9]. The key quantity is the *mutual information* $I(S';Z)$ between a chosen model-exposure variable S' and the signal to forget Z (see Section 1.3). We interpret "compression" as reducing $I(S';Z)$, thereby "compressing away" Z while retaining task utility via a user-chosen functional $\mathcal{C}(Y;S)$. This yields constrained and Lagrangian-style formulations:

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

Different choices of (S,S') instantiate different unlearning types:

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 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 $\ell_{\text{utility}}(f) := \mathcal{C}(Y;S)$, 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 (to-be-unlearned) source; let $\alpha\in(0,1]$ to be the cardinality ratio between the retain and 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 $\ell_{\text{margin}}(f) := I(S_{\text{margin}};Z)$, 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 provide a formal definition of the (Pareto) optimal feature and marginal unlearning problems with a generic outcome space $\mathcal{S} \in \{\mathcal{X}, \mathcal{V}, \mathcal{Y}\}$, so that practitioners can choose the suitable outcome space for particular application purpose.

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*

$$\inf_{f\colon \mathcal{X}\times\mathcal{Z}\to\mathcal{S}} \Big\{ \mathcal{C}(Y;S) \;:\; S \perp Z \Big\}, \qquad S := f(X,Z), \tag{8}$$

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

$$\inf_{f\colon \mathcal{X}\times\mathcal{Z}\to\mathcal{S}} (1-\lambda)\,\mathcal{C}(Y;S) \;+\; \lambda\, I(S;Z), \qquad \lambda \in [0,1]. \tag{9}$$

For data unlearning, the fact that direct (anchor-based) unlearning definitions are inauditable (Section 2.2), combined with the fact that marginal unlearning (Definition 2.4) together with utility is sufficient for anchor-based unlearning (Section 2.3), implies that marginal unlearning provides a constructive approach to data unlearning.

Definition 3.2 (Pareto Optimal marginal unlearning) *Let (X_{margin}, Z) be defined as above, $S_{\mathrm{margin}} := f(X_{\mathrm{margin}})$ for measurable $f : \mathcal{X} \to \mathcal{S}$, Y a given target variable, the optimal marginal unlearning solves*

$$\inf_{f\colon \mathcal{X}\to\mathcal{S}} \Big\{ \mathcal{C}(Y;S) \;:\; S_{\mathrm{margin}} \perp Z \Big\}, \tag{10}$$

with relaxed form

$$\inf_{f\colon \mathcal{X}\to\mathcal{S}} (1-\lambda)\,\mathcal{C}(Y;S) \;+\; \lambda\, I(S_{\mathrm{margin}};Z), \qquad \lambda \in [0,1]. \tag{11}$$

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 $I(S_{\mathrm{margin}}; Z)$, which equals a *generalized Jensen–Shannon divergence* between the conditional output laws. Writing $P_1 := \mathbb{P}_{S_{\mathrm{margin}}|Z=1}$, $P_0 := \mathbb{P}_{S_{\mathrm{margin}}|Z=0}$, $\pi := \Pr(Z{=}1)$, and $P := \pi P_1 + (1-\pi)P_0$, we obtain

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

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*: JS is bounded by $H_2(Z) \leq \log 2$ (binary Z), so gradients remain well behaved 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 $I(\hat{Y}_{\text{margin}}; Z) \leq I(S_{\text{margin}}; Z)$, so suppressing MI at exposure controls leakage everywhere.

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 show that our proposed unlearning framework, Definitions 3.1 and 3.2, are generalizations of the classic rate-distortion framework in data compression.

To start, we specialize the generic outcome space to be the data space itself: $\mathcal{S} = \mathcal{X}$, e.g. synthetic data design or self-supervised learning. So that we denote S by $\hat{X}$ to better reflect the specific setting. 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 is feature unlearning with $Z = X$:

$$\inf_{f:\, \mathcal{X}\times\mathcal{X}\to\mathcal{X}} \; \mathcal{C}(Y; \hat{X}) \; + \; \lambda\, I(\hat{X}; X) \tag{12}$$

is precisely a rate-distortion tradeoff [60]: $I(\hat{X}; X)$ is the *rate* allocated to X, while $\mathcal{C}(Y; \hat{X})$ quantifies distortion for the task defined by Y.

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). 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 (or specifically Definition 3.1) is a generalization of the classic rate-distortion type of data compression by allowing unwanted information to be particular features Z that might not in X. Furthermore, marginal unlearning (or specifically Definition 3.2) is a further generalization by allowing the compression objective to utilize subsets of the original data X to construct (X_{margin}, Z) and then compress the constructed information Z.

Admissibility (no artificial information creation). Requiring $S = f(X, Z)$ or $S = f(X)$ to be measurable with respect to (X, Z) and X, respectively, ensures $\sigma(S) \subset \sigma(X, Z)$, i.e., the compressor does not create information ex nihilo (out of nothing) (See more detailed explanation of the importance in Appendix C.2).

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 $S = \hat{X}$. But we note that the setting suffices to also provide theoretical guarantees for the generic setting via the data-processing inequality. See details in Remark 4.1 below.

For *feature unlearning*, we show that driving the mutual information $I(\hat{X}; Z)$ small forces each 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 indicates training inclusion. Finally, we leverage optimal transport to derive analytic solution for both unlearning with particular cost functions.

Remark 4.1 ($\mathcal{S} = \mathcal{X}$ suffices genetic 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 $\hat{Y}_{\text{margin}} := f(\hat{X}_{\text{margin}})$, the odds bound in Definition 3.2 is preserved under measurable pushforwards: for every measurable $B \subseteq \mathcal{S}'$,*

$$\frac{\mathbb{P}\big(Z = 0 \,\big|\, T(\hat{Y}_{\text{margin}}) \in B\big)}{\mathbb{P}\big(Z = 1 \,\big|\, T(\hat{Y}_{\text{margin}}) \in B\big)} \;=\; \frac{\mathbb{P}\big(Z = 0 \,\big|\, \hat{Y}_{\text{margin}} \in T^{-1}(B)\big)}{\mathbb{P}\big(Z = 1 \,\big|\, \hat{Y}_{\text{margin}} \in T^{-1}(B)\big)},$$

so the same ε-bound holds with $D := T^{-1}(B)$. Likewise, MI relaxations obey data processing, i.e., $I\big(T(\hat{Y}_{\text{margin}}); Z\big) \le I(\hat{Y}_{\text{margin}}; Z)$.

Consequently, it suffices to analyze $S = \hat{X}$ without loss of generality. We adopt this notation ($S = \hat{X}$) for the remainder of the section.

4.1 Feature unlearning guarantee

In practice we optimize the relaxed problem

$$\inf_{f} \; \mathcal{C}(Y; \hat{X}) \;+\; \lambda\, I(\hat{X}; Z),$$

so that $I(\hat{X}; Z) \le \varepsilon$ 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

$$\mathbb{E}_Z\Big[\, D_{\mathrm{KL}}\big(\mathbb{P}(\hat{X} \mid Z) \,\|\, \mathbb{P}(\hat{X})\big)\,\Big] = I(\hat{X}; Z), \tag{13}$$

$$D_{\mathrm{KL}}\big(\mathbb{P}(\hat{X} \mid Z{=}z) \,\|\, \mathbb{P}(\hat{X})\big) \le \frac{I(\hat{X}; Z)}{p(z)}, \qquad \forall\, z \in \mathcal{Z}, \tag{14}$$

and

$$\mathbb{E}_Z\Big[\, \mathrm{TV}\big(\mathbb{P}(\hat{X} \mid Z),\, \mathbb{P}(\hat{X})\big)\,\Big] \le \sqrt{\tfrac{1}{2}\, I(\hat{X}; Z)}, \tag{15}$$

$$\mathrm{TV}\big(\mathbb{P}(\hat{X} \mid Z{=}z),\, \mathbb{P}(\hat{X})\big) \le \sqrt{\tfrac{1}{2\, p(z)}\, I(\hat{X}; Z)}, \qquad \forall\, z \in \mathcal{Z}. \tag{16}$$

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 (15).

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 (16). 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 $\mathbb{R}^{d_{\mathcal{X}} + d_{\mathcal{Z}}}$ with Borel sigma algebra $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ 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 version (define $\mathbb{P} \circ Z^{-1}$ almost surely and may differ on null sets) of the regular conditional distributions. Assume $P_{\hat{X}} \ll \mathcal{L}$ (Lebesgue measure) and $I(\hat{X}; Z) < \infty$. Then:*

1. KL expectation bound:

$$\mathbb{E}_Z\Big[\; \mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_{\hat{X}|Z} || \, \mathbb{P}_{\hat{X}}\right) \Big] \;\le\; \sqrt{I(\hat{X}; Z)}\,. \tag{17}$$

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

$$\mathbb{P}\Big(\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_{\hat{X}|Z} || \, \mathbb{P}_{\hat{X}}\right) \geq \tau \Big) \;\leq\; \frac{I(\hat{X};Z)}{\tau^2}\,. \tag{18}$$

3. TV expectation bound:

$$\mathbb{E}_Z\Big[\,\mathrm{TV}\big(\mathbb{P}_{\hat{X}|Z},\, \mathbb{P}_{\hat{X}}\big)\,\Big] \;\leq\; \sqrt{\frac{1}{2}\,I(\hat{X};Z)}\,. \tag{19}$$

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

$$\mathbb{P}\,\Big(\,\mathrm{TV}\big(\mathbb{P}_{\hat{X}|Z},\, \mathbb{P}_{\hat{X}}\big) \geq \tau\Big) \;\leq\; \frac{I(\hat{X};Z)}{2\,\tau^2}\,. \tag{20}$$

Proof By the assumption $P_{\hat{X}} \ll \mathcal{L}$ and $I(\hat{X};Z) < \infty$, we have $D_{\mathrm{KL}}\big(P_{\hat{X}|Z=z} \,\|\, P_{\hat{X}}\big) < \infty$ for $\mathbb{P}_Z$-a.e. z. Indeed, if not, then there exists a set $B \subseteq \mathcal{Z}$ with $P_Z(B) > 0$ such that for all $z \in B$, $P_{\hat{X}|Z=z} \not\ll P_{\hat{X}}$, which further implies $D_{\mathrm{KL}}(P_{\hat{X}|Z=z} \| P_{\hat{X}}) = \infty$ for all $z \in B$. It follows from

$$I(\hat{X};Z) = D_{\mathrm{KL}}\big(P_{\hat{X},Z} \,\|\, P_{\hat{X}} \otimes P_Z\big) = \int D_{\mathrm{KL}}\big(P_{\hat{X}|Z=z} \,\|\, P_{\hat{X}}\big)\, P_Z(dz)$$

that $I(\hat{X};Z) = \infty$ which contradicts the assumption $I(\hat{X};Z) < \infty$.
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). 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 $\tau > 0$,

$$\mathbb{P}(D_{\mathrm{KL}}(\mathbb{P}_{\hat{X}|Z} \| \mathbb{P}_{\hat{X}}) \geq \tau^2) \;\leq\; \frac{\frac{1}{2}\, \mathbb{E}_Z\, D_{\mathrm{KL}}(\mathbb{P}_{\hat{X}|Z} \| \mathbb{P}_{\hat{X}})}{\tau^2} \;=\; \frac{I(\hat{X};Z)}{2\,\tau^2},$$

which is (20). 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*

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

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* (19)–(20) *are the natural general theoretical guarantees.*

4.2 Data unlearning guarantee

For data unlearning we operate on the data space and penalize

$$I\big(\hat{X}_{\text{margin}}; Z\big), \qquad \hat{X}_{\text{margin}} := f(X_{\text{margin}}), \quad X_{\text{margin}} \mid (Z{=}1) \sim p^r, \; X_{\text{margin}} \mid (Z{=}0) \sim p^d,$$

as in Section 2.2. That is, we penalize the information that $\hat{X}_{\text{margin}}$ carries about inclusion/exclusion information encoded by Z:

$$\inf_{f:\,\mathcal{X}\to\mathcal{X}} \Big\{ (1-\lambda)\mathcal{C}(Y;\hat{X}) \;+\; \lambda\, I(\hat{X}_{\text{margin}};Z) \Big\}. \tag{21}$$

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

Lemma 4.1 (MI controls odds-inference) *Let* $\hat{X}_{\text{margin}} = \big[f(X_0),\, f(X_1)\big]$ *with* $X_0 \sim p^d$, $X_1 \sim p^r$, *and* $Z \in \{0,1\}$ *indicating the source. Then for any* $\varepsilon \in (0,1)$,

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

(See Appendix D.2 for the proof.) Thus, either $\hat{X}_{\text{margin}} \perp Z$ (i.e., $I(\hat{X}_{\text{margin}};Z) = 0$), or if $I(\hat{X}_{\text{margin}};Z)$ is small relative to $\frac{e^\varepsilon-1}{e^\varepsilon+1}$, the probability of any observation that confers more than that inference strength is small. Since $\hat{Y}_{\text{margin}}$ aggregates predictions *with and without* X_u, even access to both outputs does not reveal whether training used X_0 or X_1 with probability exceeding the bound. Practically, one tunes λ so that $I(\hat{X}_{\text{margin}};Z)$ is small enough for the desired ε.

The following guarantee translates a small "compression rate" into an auditable inference bound for ε-marginal unlearning (Definition 2.4):

Theorem 4.2 (Marginal unlearning via compression rate) *Let* $Z \sim \text{Bernoulli}(1/2)$. *If* $I(\hat{X}_{\text{margin}};Z) \le \mu < 2(\frac{e^\varepsilon-1}{e^\varepsilon+1})^2$, *then* f *satisfies* ε*–marginal unlearning with probability at least*

$$1 \;-\; \frac{e^\varepsilon+1}{e^\varepsilon-1}\sqrt{\frac{\mu}{2}}\,.$$

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 solution using optimal transport and extreme sigma-algebras. In particular, we focus on the following cost functions:

- *Entropy maximization:* $\mathcal{C}(Y;\hat{X}) = -H(\hat{X})$ preserves variability in $\hat{X}$.
- *Mutual information maximization:* $\mathcal{C}(Y;\hat{X}) = -I(Y;\hat{X})$ maximizes informativeness about Y.
- *Posterior concentration:* $\mathcal{C}(Y;\hat{X}) = -D_{\mathrm{KL}}\big(\mathbb{P}(Y \mid \hat{X}) \,\|\, \mathbb{P}(Y)\big)$.
- *Conditional–probability energy:*

$$\mathcal{C}(Y;\hat{X}) = \begin{cases} \sqrt{\mathbb{E}\big[\mathbb{P}(Y \in A \mid \hat{X})^2\big]}, & \text{classification,} \\ -\|Y - \mathbb{E}(Y \mid \hat{X})\|_2, & \text{regression.} \end{cases}$$

Despite the fact that these cost functions are all different, they do admit the *same* analytic solution in the feature/full–data setting (Theorem 4.3): the $\mathcal{W}_2$ barycenter of $\{X \mid Z = z\}_z$, which generates the finest sigma–algebra among admissible outcomes and therefore maximizes all monotone information criteria.

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 $\implies$ finest sigma–algebra [71, Lemma 5.2]) *Let $\{X_z\}_{z\in\mathcal{Z}} \subset \mathcal{P}_{2,ac}(\mathcal{X})$ and let $\bar{X}$ denote their $\mathcal{W}_2$ barycenter. Then $\sigma(\hat{X}) \subset \sigma(\bar{X})$ for all admissible $\hat{X} = f(X, Z)$.*

See Appendix D.1 for a sketch of the proof; a full proof is given in [71, Lemma 5.2]. The above result demonstrates one feasible optimal solution. Under the assumption of absolute continuity of the marginals, the barycenter of convex costs also satisfies the invertibility of transport maps, ensuring optimality.

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 for any target Y and event A,*

- $H(X_1) \leq H(X_2)$,
- $H(Y \mid X_2) \leq H(Y \mid X_1)$ *for any* $Y : \Omega \to \mathcal{Y}$,
- $I(Y; X_1) \leq I(Y; X_2)$ *for any* $Y : \Omega \to \mathcal{Y}$,
- $\|\mathbb{P}(Y \in A \mid X_1)\|_2^2 \leq \|\mathbb{P}(Y \in A \mid X_2)\|_2^2$ *for any* $A \in \mathcal{B}_{\mathcal{Y}}$.

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 how entropy increases as the sigma-algebra becomes finer, while conditional entropy decreases, indicating reduced uncertainty. Similarly, mutual information, KL-divergence, and conditional probability energy increase, reflecting enhanced informativeness and predictive utility.

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

- $\sigma(\hat{X}) = \sigma(\bar{X})$,
- $\hat{X} \in \arg\max\{H(\hat{X}) : \ \hat{X} \perp Z\}$,
- $\hat{X} \in \arg\min\{H(Y \mid \hat{X}) : \ \hat{X} \perp Z\}$ *for all* Y,
- $\hat{X} \in \arg\max\{I(Y; \hat{X}) : \ \hat{X} \perp Z\}$ *for all* Y,
- $\hat{X} \in \arg\max\{\|\mathbb{P}(Y \in A \mid \hat{X})\|_2^2 : \ \hat{X} \perp Z\}$ *for all* A, Y.

Proof This follows directly from Lemma 4.2 and Lemma 4.3. ■

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 cost functions, we present an algorithm for *feature unlearning* (Definition 3.1) based on information–theoretic regularization and an arbitrary cost function. The procedure accommodates any generic cost function $\mathcal{C}$ and enforces forgetting via a penalty on the mutual information between the learned representation and the sensitive feature Z. By default, we instantiate unlearning in the model–output (or parameter) space: given a (parametrized by θ) measurable map $f_\theta : \mathcal{X} \times \mathcal{Z} \to \mathcal{S}$, we form $S = f_\theta(X, Z)$ and optimize a regularized objective of the form $(1 - \lambda)\mathcal{C}(S, Y) + \lambda\, I(S; Z)$.

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 $\eta$, batch
    size $B$, number of epochs $T$, regularization parameter $\lambda$.
    (Optional:) Pre-trained neural network $f_{\theta_{\text{origin}}}$ with parameters $\theta_{\text{origin}}$.
Ensure: Unlearned model parameters $\theta$.
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: $\mathcal{L}_{\text{reg}} = (1 - \lambda)\mathcal{C}(Y_d; S_d) + \lambda I(S_d; Z_d)$
        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 $\theta$
```

5.2 General cost marginal data unlearning: solution to Definition 2.4

For marginal unlearning, we operate in prediction space and penalize $I(S_{\text{margin}}; Z)$ with the paired construction $X_{\text{margin}} \mid_{\{Z=1\}} \sim p^r$, $X_{\text{margin}} \mid_{\{Z=0\}} \sim p^d$ 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 Thm. 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 $\theta_{\text{origin}}$, loss function $\mathcal{C}$, learning rate $\eta$,
      batch size $B$, number of epochs $T$, regularization parameter $\lambda$.
  Ensure: Unlearned model parameters $\theta$. The model $f_\theta$ satisfies $\varepsilon$-marginal unlearning
      (Definition 3.2) guarantee with probability at least $1 - \frac{\exp \varepsilon + 1}{\exp \varepsilon - 1}\sqrt{\frac{\mu}{2}}$ with
      $\mu := I(S_{\text{margin}}; Z)$ at training termination.
  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 $(S_{\text{margin}}, Z)$ 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
  Return unlearned model parameters $\theta$.
```

[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 (e.g., $-H(\hat{X})$, $-I(Y; \hat{X})$, $-D_{\text{KL}}(\mathbb{P}(Y \mid \hat{X}) \| \mathbb{P}(Y))$, the classification energy, or the regression energy), 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 $\hat{X} \perp Z$.[3] 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.

[3] 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 $\varepsilon$.
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 [43], 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 both synthetic and real-world datasets to validate the proposed framework, focusing on its theoretical guarantees and explainability rather than benchmarking against other methods. A more comprehensive evaluation, including improved barycenter estimation for feature unlearning, optimal 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	45211
UCI Bank	Subscription	Marital	51	5300
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

In this simple experiment, we demonstrate how marginal unlearning effectively helps the model forget the information specific to the unlearn set (a zero-mean Gaussian), while retaining what is supported by the retain set (a uniform distribution). Concretely, the information to forget is the concentration around 0, and the information to retain is the nearly flat density over $[-L, L]$.

Data and Notation We consider two datasets supported (initially) on the same interval $[-L, L]$ with $L = 3$:

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

We draw N_R i.i.d. samples $\{x_i^{(r)}\}_{i=1}^{N_R}$ from $\mathrm{Unif}([-L, L])$ (retain set) and N_U i.i.d. samples $\{x_j^{(u)}\}_{j=1}^{N_U}$ from $\mathcal{N}(\mu, \sigma^2)$ (unlearn set).

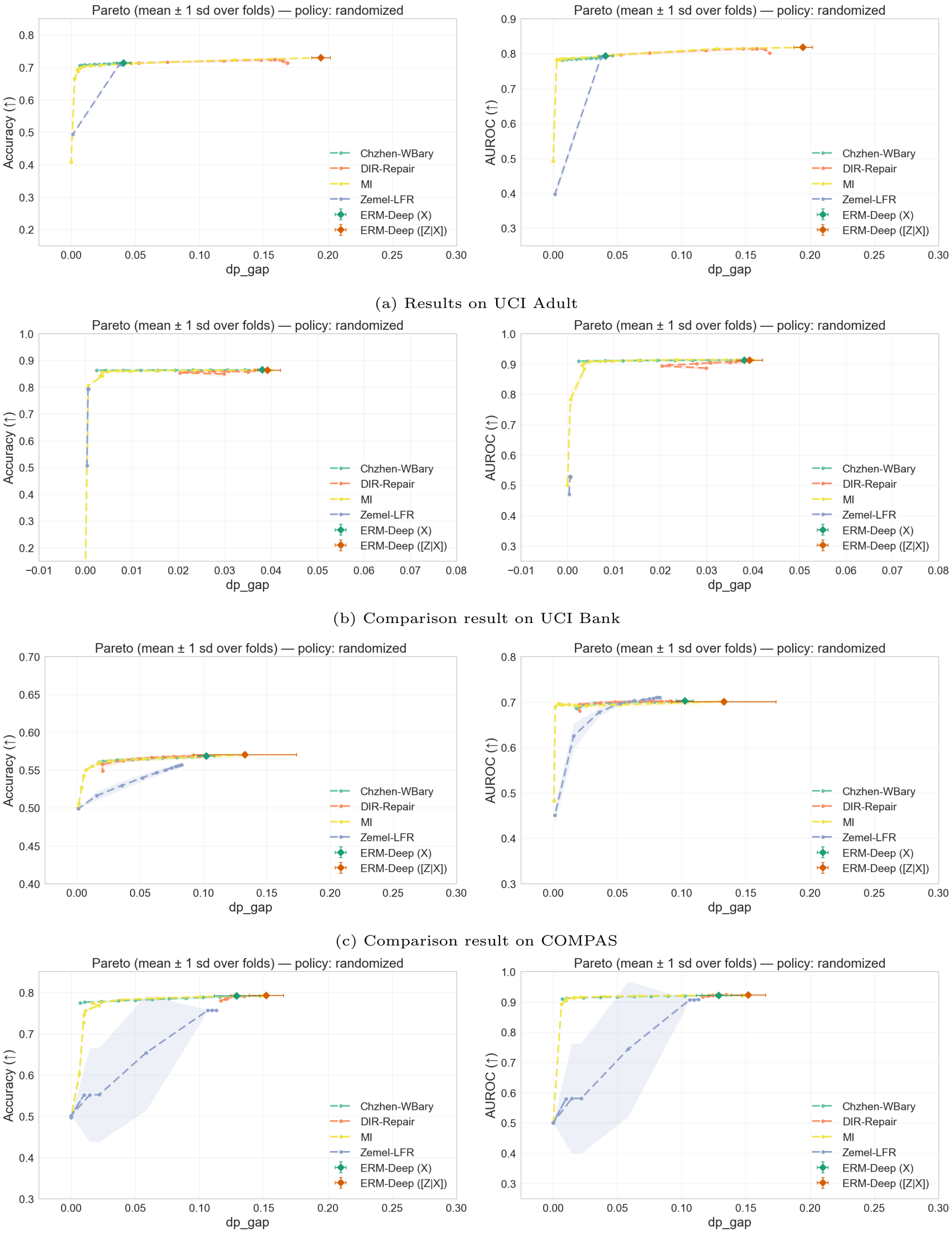
We use a residual network $f_\theta : \mathbb{R} \to \mathbb{R}$, initialized near the identity, and first pretrain it so that the retain output density approximates the mixture to mimic the situation where the model has seen the unlearn set even when the input is the retain set:

$$\widehat{p}_{f_{\theta_0}(X_r)} \;\approx\; \alpha\, \widehat{p}_{X_r} + (1 - \alpha)\, \widehat{p}_{X_u} \quad (\alpha \in (0, 1)).$$

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 (inputs) and h_y (outputs):

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

For densities P, Q on $\mathcal{G}$, 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 label $Z \in \{0, 1\}$ with priors p_0, p_1 and conditionals P_0, P_1, the MI is $I(Y; Z) = p_0\ \mathrm{D}_{\mathrm{KL}}(P_0\|P) + p_1\ \mathrm{D}_{\mathrm{KL}}(P_1\|P)$ for $P := p_0 P_0 + p_1 P_1$.



(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.

Unlearning Methods in Comparison We fix a tradeoff $\lambda \in [0,1]$ and (unless stated) $p_0 = p_1 = \frac{1}{2}$. We compare three objectives:

- **Marginal (ours).** The MI compares the retain output to the *mixture* output. With $P_0 = \widehat{p}_{f_\theta(X_r)}$ and $P_1 = \widehat{p}_{f_\theta(X_r \cup X_u)} = \alpha\, \widehat{p}_{f_\theta(X_r)} + (1-\alpha)\, \widehat{p}_{f_\theta(X_u)}$, the loss is

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

  Minimizing the MI drives $P_1 = P_0$, which algebraically forces $\widehat{p}_{f_\theta(X_u)} = \widehat{p}_{f_\theta(X_r)}$ pointwise, while the CE term anchors $\widehat{p}_{f_\theta(X_r)}$ to the uniform $\widehat{p}_{X_r}$.

- **Grad_diff (gradient difference).** Replace MI with a "push away from X_u" term via negative cross entropy:

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

  This discourages resemblance to the forget density but does not prescribe where the X_u mass should move, so $f_\theta(X_u)$ can be pushed to arbitrary regions (e.g., boundaries), potentially perturbing $f_\theta(X_r)$ through parameter sharing.

- **KL (retain-anchored + grad_diff regularization).** Here the utility anchors the unlearning trajectory to the *fine-tuned* baseline on the retain set:

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

  The first term keeps $f_\theta(X_r)$ close to the pretrained $f_{\theta_0}(X_r)$ (which already matches the mixture), while the second term reduces resemblance to X_u.

Outcomes. Figure 2 shows how the output densities evolve (starting from the fine-tuned model) under the three methods. Under MARGINAL, the mechanism *equalizes* the retain and mixture conditionals, driving $p\big(f_\theta(X_u)\big)$ to coincide with $p\big(f_\theta(X_r)\big)$ while the retain cross-entropy term keeps $p\big(f_\theta(X_r)\big) \approx p(X_r)$ to stay at the uniform distribution. This simultaneously removes X_u-specific signal and preserves the retain support. In contrast, GRAD_DIFF lacks a principled target for the "unlearn" mass: the term $-H\big(\widehat{p}_{X_u} \| \widehat{p}_{f_\theta(X_u)}\big)$ can push density toward the boundaries or other arbitrary regions, and shared parameters can slightly distort $p\big(f_\theta(X_r)\big)$. Finally, KL anchors the retain output near the pretrained baseline via $\mathrm{D}_{\text{KL}}\big(p(f_{\theta_0}(X_r)) \| p(f_\theta(X_r))\big)$ and discourages resemblance to X_u, but, similar to grad_diff, the gradient ascent does not enforce how to distribute the mass of the unlearn and, hence, can likewise push the forget mass toward the edges.

Summary. Among the three, MARGINAL most directly penalizes X_u-specific information while preserving the retain distribution: minimizing its MI term yields $p(f_\theta(X_u)) = p(f_\theta(X_r))$, and the utility keeps this shared density aligned with the uniform support. The KL objective is a stable, anchor-based alternative; GRAD_DIFF provides a simple regularizer but can relocate X_u mass to undesirable regions.

6.2.2 Forget MNIST

Dataset and retain/unlearn split. We use MNIST (60,000 train / 10,000 test) with the standard normalization $\mu = 0.1307$, $\sigma = 0.3081$ applied channelwise. Let U denote the set of training images whose label is digit 3. We mark a (99.5%) fraction of U as the

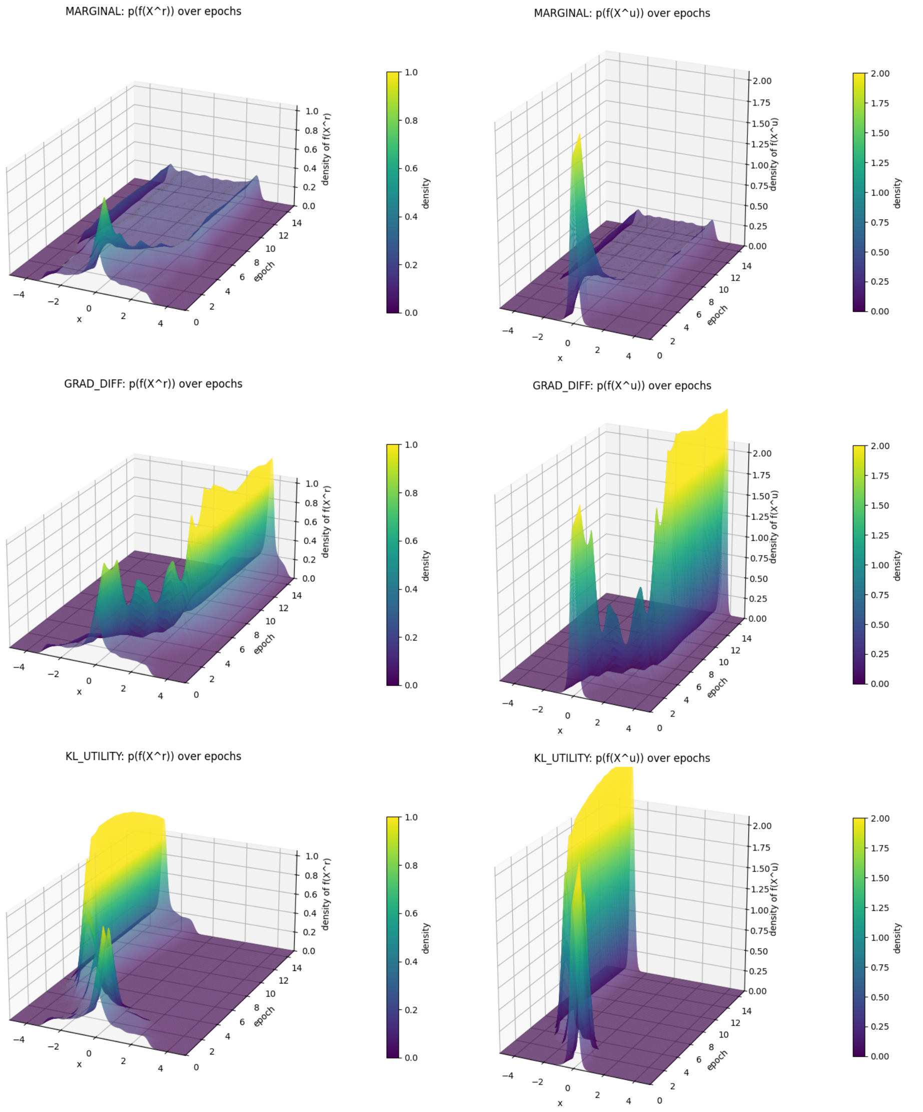


Figure 2: Evolution of output densities $p(f_\theta(X_r))$ (left column) and $p(f_\theta(X_u))$ (right column) over epochs for the three objectives. It is clear that marginal unlearning suppresses the unlearn signal (concentration around zero) while preserving the uniform density supported by the retain signal. In comparison, the method based on gradient ascent all push the mass concentration to somewhere else, even though they have different utility preservation that try to counter the destructive gradient ascent.

unlearn set and take the complement $R :=$ train $\setminus$ unlearn as the *retain set*. For evaluation we form 5-fold CV *independently* on R and on the unlearn subset: for each fold i, we obtain $(R_i^{\text{tr}}, R_i^{\text{val}})$ and $(U_i^{\text{tr}}, U_i^{\text{val}})$. All results reported for validation use the R_i^{val} and U_i^{val} splits; test accuracy is computed on the original MNIST test set. Unless otherwise stated, batch sizes are 128 for retain/unlearn and 512 for test.

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:

$$\begin{aligned}
&\underbrace{\mathrm{Conv}(32)\to\mathrm{ReLU}\to\mathrm{Conv}(64)\to\mathrm{ReLU}\to\mathrm{MaxPool}(2)\to\mathrm{Dropout}(0.25)}_{\textbf{Block 1}}\\
\to\ &\underbrace{\mathrm{Conv}(128)\to\mathrm{ReLU}\to\mathrm{MaxPool}(2)\to\mathrm{Dropout}(0.25)}_{\textbf{Block 2}}\\
\to\ &\underbrace{\mathrm{Flatten}(128\times7\times7)\to\mathrm{FC}(256)\to\mathrm{ReLU}\to\mathrm{Dropout}(0.5)\to\mathrm{FC}(10)}_{\textbf{Head}}.
\end{aligned}$$

Max-pooling layers are 2×2 (stride 2), yielding spatial sizes $28\times28\to14\times14\to7\times7$. 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 in our optimization with learning rate 10^{-3} and weight decay 10^{-4}. We fix the random seed to 1337. Fine-tuning and retraining run for 10 epochs each; unlearning runs for at most 30 epochs (may stop early by the rules below).

Baselines and compared methods.

- **(i) FineTune-All (FT).** Train the backbone on the entire training set ($R \cup U$) for 10 epochs; this reflects the initial model that "remembers everything."
- **(ii) Retrain-on-Retain (RT).** Train the same backbone from scratch on R only for 10 epochs; this is the anchor that never saw the unlearn data.
- **(iii) Marginal-MI (ours).** Starting from the FT weights, minimize

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

  The marginal MI estimator uses the model softmax averages $P_0 = \mathbb{E}_{x\in R}[\hat{p}(y|x)]$, $P_u = \mathbb{E}_{x\in U}[\hat{p}(y|x)]$, $P_1 = \alpha P_0 + (1-\alpha)P_u$ with $\alpha = \frac{|R|}{|R|+|U|}$, equal priors $p_0 = p_1 = \frac{1}{2}$, and $\widehat{\mathrm{MI}}_{\mathrm{marg}} = \frac{1}{2}\mathrm{KL}(P_0\|P) + \frac{1}{2}\mathrm{KL}(P_1\|P)$ where $P = \frac{1}{2}P_0 + \frac{1}{2}P_1$.
- **(iv) Grad-Diff.** From the FT weights, optimize

  $$\mathcal{L}_{\mathrm{GD}} = (1-\lambda)\,\mathrm{CE}\big(f_\theta(x), y;\, x\in R_i^{\mathrm{tr}}\big) \;-\; \lambda\,\mathrm{CE}\big(f_\theta(x), y;\, x\in U_i^{\mathrm{tr}}\big).$$
- **(v) KL+CE (teacher–student).** With the *FT model frozen as teacher* $f_{\theta^\star}$, optimize from the FT weights

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

Trade-off grids and initialization. Every unlearning method is initialized *from the same FT checkpoint* (weights cloned at runtime). We sweep per-method λ on small grids:

$$\lambda_{\mathrm{MI}} \in \{0.0020,\, 0.0055,\, 0.0100\},\quad \lambda_{\mathrm{KL}} \in \{0.0003,\, 0.0006,\, 0.0012\},\quad \lambda_{\mathrm{GD}} \in \{0.0002,\, 0.00035,\, 0.0007\}.$$

Here, the trade-off parameter is chosen so that each method range from slight decrease in unlearning accuracy to significant drop in unlearning accuracy.

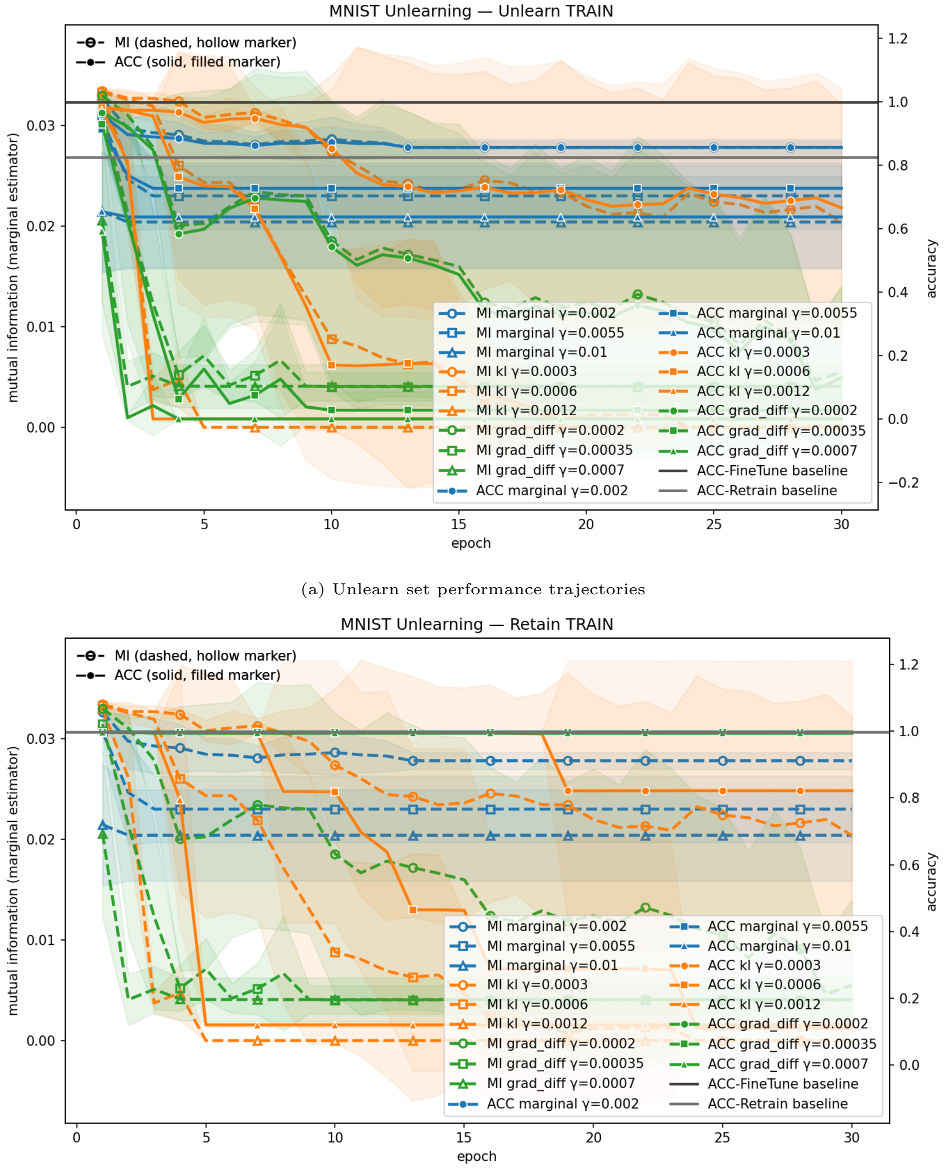
Early stopping for each method. We also adopt the following early stopping criterion for each of the method to avoid over-unlearning:

- **Marginal-MI:** compute $\mathrm{MI}_0^{\mathrm{val}}$ before training on $(R_i^{\mathrm{val}}, U_i^{\mathrm{val}})$ and stop when $\mathrm{MI}^{\mathrm{val}} \leq 0.85\,\mathrm{MI}_0^{\mathrm{val}}$ (min-epochs = 1, patience = 1).
- **KL+CE:** compute the teacher→student KD on retain-val, $\mathrm{KD}_0^{\mathrm{val}}$, and stop when $\mathrm{KD}^{\mathrm{val}} \leq 0.85\,\mathrm{KD}_0^{\mathrm{val}}$ (min-epochs = 1, patience = 1).
- **Grad-Diff:** infer the number of classes C from the logits ($C{=}10$ for MNIST) and stop when the unlearn-val accuracy $\leq 1/C + 0.02$ *for two consecutive checks* (min-epochs = 1, patience = 2).

Results. In Figure 3, we record accuracy on R_i^{tr}, R_i^{val}, U_i^{tr}, U_i^{val} and the marginal MI on both train/val splits for each epoch. Each method's run stops according to the criterion above; curves are aggregated across the 5 folds by mean±std. For tabular summaries, we report the value at the stopping epoch (padding after early stop keeps the last value constant), averaged over folds. Plots show MI (dashed, hollow markers) and accuracy (solid, filled markers) with a shared color per method and shaded mean±std; we overlay horizontal bands for the FT and RT validation accuracies (mean±std across folds) as references.

Figure 3 demonstrates two important observations:

- *Mutual information is an accurate unlearn regularizer for unlearn accuracy:* The mutual information regularization term has nearly perfect correlation with the unlearn accuracy, mostly regardless of the trade-off parameter chosen and unlearn method adopted. This is expected from the theoretical result because mutual information represents the distribution shift. Since we want to unlearn the label 3 from the model output, the retain set should not predict 3 if it want to minimize that distributional shift. This observation also points out another stopping criterion or trade-off parameter tuning policy: It is recommended that the practitioner picks a goal of unlearn accuracy that one would expect the retrain model to achieve on the unlearn set by generalization from its training on the retain set. In particular, unlearn accuracy = $g*$retain accuracy with $g \in [0,1]$ Then, the practitioner can set the stopping criterion to be the $MI(f_\theta(X_{margin}); Z) \leq g * MI(f_{\theta_0}(X_{margin}); Z)$. So that the correlation (between the MI value and the unlearning accuracy) can make sure that the resulting model approximates the desired unlearning accuracy.
- *Marginal unlearning is the most stable and accurate method in estimating the retrain on retain standard:* Compared with other methods, marginal unlearning is the only method that can constantly stay near the retrain on retain set standard. One can see it from four perspectives: (1) parameter robustness, (2) cross-validation robustness, (3) training trajectory robustness, and (4) accuracy preservation on both retain and unlearn set. From parameter choosing perspective, it is clear that marginal unlearning has more robustness in trade-off parameter choice: A different parameter choice leads to slightly different unlearning accuracy performance but still steadily around the retrain on retain standard. In comparison, the other methods are very sensitive to parameter choice. From a cross-valid perspective, the standard deviation of both retain and unlearn learn



(a) Unlearn set performance trajectories

(b) Retain set performance trajectories

Figure 3: MNIST Unlearning trajectories on train (left panel) and test (right panel) folders with line representing the mean and shade width representing 1 standard deviation.

accuracy is at least 3 times smaller compared to the other methods over the 5-fold cross validation. From a training trajectory perspective, marginal unlearning also provide monotone penalization of unlearning accuracy and then stabilizes without performance fluctuation. In comparison, other methods all fluctuate and demonstrate volant instability in training trajectory. Finally, from a retain and unlearn accuracy perspective, marginal unlearning is able to demonstrate unlearning while preserving high retain and unlearn accuracy. In comparison, the other methods are very vulnerable in unlearn accuracy drop. That is, either the method does not unlearn at all or it tends to unlearn too much. All four advantages we observe from the MNIST experiment in Figure 3 are consistent with theory: Marginal unlearning has no necessary conflict between utility objective and its direct marginal unlearning objective. In comparison, the other methods require conflicting utility and full information unlearn objectives to fight each other to indirectly search for marginal information. Considering the training as a dynamical system guided by two forces or gradient fields, marginal unlearning with non-conflicting forces can achieve an equilibrium much easier and more robust than other methods with conflicting forces.

6.3 Optimal Feature Unlearning via Analytic Solution: Algorithm 3

Dataset. We use the CelebFaces Attributes dataset (CelebA) [48], 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 to an *attribute-neutral* counterpart $\bar{x}$ whose distribution $\bar{\mu}$ is (i) minimally distorted from the originals and (ii) uninformative about Z.

Barycentric characterization (theory). Theorem 4.3 shows that the 2-Wasserstein barycenter of μ_0 and μ_1 with equal weights as 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 [50]. 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. [43]: 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 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 $\mathrm{Bary}(X) := [0.5Id + 0.5T](x)$, represents the corresponding sample generated by the McCann interpolation at $t = 0.5$, which coincides with the barycenter in the our two-marginal cases.

In theory, $\mathrm{Bary}(X)$ is (an estimation of) the optimal feature unlearning solution. Therefore, one should not be confident in telling whether or not the face is smiling/non-smiling or belonging to male/female. Notably, it is clear from Figure 4 that the generated image reduces the targeted attribute while preserving the other information on the images, such as lighting, and idiosyncratic facial traits.

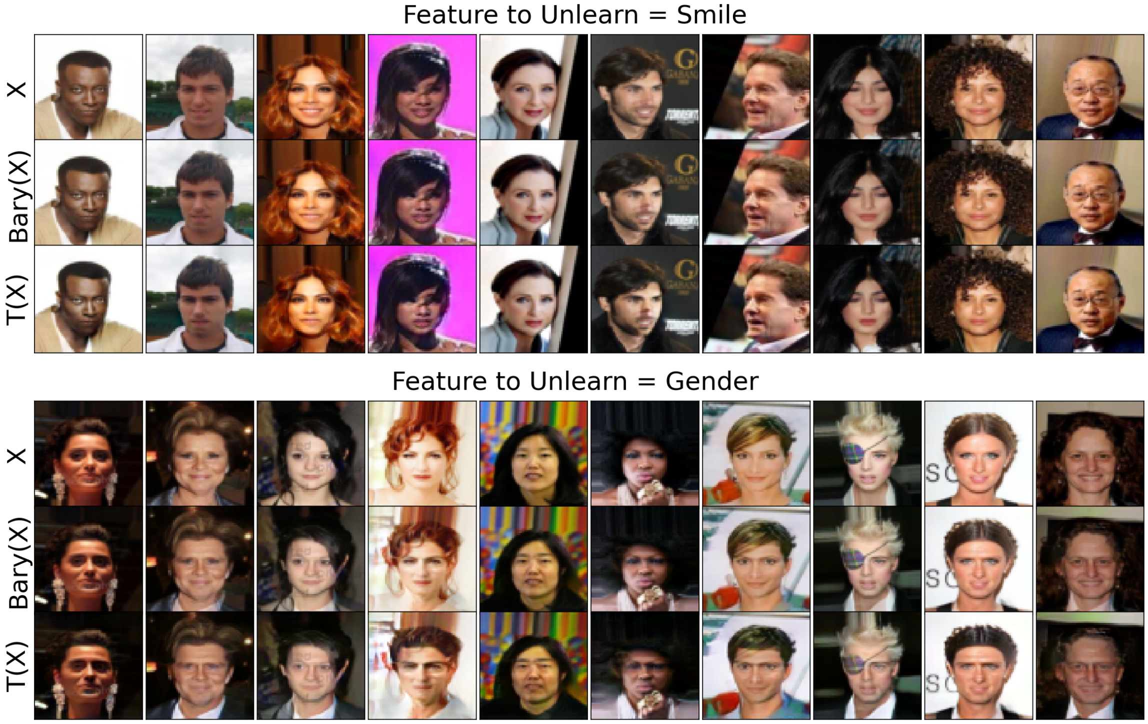


Figure 4: illustrates the unlearning outcomes for both tasks: The above penal shows the unlearn results for smile feature and the bottom shows the results for gender feature. In each of the panel, the first row, denoted by X, represents samples from the original data set with the chosen feature value (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 $\mathrm{Bary}(X) := [0.5Id + 0.5T](x)$, 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 simultaneously solves both feature and data unlearning under standard utilities 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] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

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

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

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

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

[45] 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/`.

[46] 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.

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

[48] 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.

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

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

[51] 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.

[52] 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.

[53] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

[63] 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.

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

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

[66] 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.

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

[68] 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.

[69] 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.

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

[71] 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.

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

[73] 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.

[74] 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.

[75] 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.

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

[77] 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]	$\hat{Y} \perp Z$	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 [77]	$h(X) \perp Z$	Predictor vs. adversary predicting Z from $h(X)$ or $\hat{Y}$	Empirical proxy to reduce $I(\cdot; Z)$; architecture-agnostic.
VFAE [49]	$h(X) \perp Z$	VAE with MMD penalty to match $p(h \mid Z{=}z)$ across z	Produces reusable "purged" latents.
HSIC / dCov regs. [46, 47]	$\mathrm{dCov}\big(h(X), Z\big) = 0$	Add HSIC or distance-covariance penalties during training	Nonparametric dependence control; task-agnostic.
Concept erasure: representation independence			
INLP [55]	$\Sigma_{hZ} = \mathbf{0}$	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 [56]	$\Sigma_{hZ} = \mathbf{0}$	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 [27]	$f_\theta(X) \perp C$	Fine-tune to a negative-guidance teacher (subtract conditional–unconditional score)	Local removal of concept c with utility preservation.
All-but-One [37]	$f_\theta(X) \perp C$	Surgical update of guidance term with preservation constraints	Improves fidelity vs. naive fine-tuning.
Receler [40]	$f_\theta(X) \perp C$	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 [33] 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 [44, 51, 15]. For auditing, *amnesic probing* assesses whether removing targeted information from representations changes behavior [25]; 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 [63]. In practice, existing verification schemes have proven *fragile*: Zhang et al. [76] 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 [62]. Similarly, cryptographic and TEE-based proofs target verifiability but necessitate trusted hardware and impose significant operational overhead [68]. Recent surveys underscore that robust verification remains a central open challenge [73, 70]. 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

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

Finally, let ν be the law of Θ and write $P_\theta := \mu_\theta^{p^d}$, $Q_\theta := \mu_\theta^{p^r}$. By convexity of total variation under mixing,

$$\mathrm{TV}\big(\mu_\Theta^{p^d}, \mu_\Theta^{p^r}\big) = \mathrm{TV}\Big(\int P_\theta\,\nu(d\theta), \int Q_\theta\,\nu(d\theta)\Big) \;\le\; \int \mathrm{TV}(P_\theta, Q_\theta)\,\nu(d\theta).$$

It follows from the pointwise bound from above and Jensen's inequality for the concave square root,

$$\int \mathrm{TV}(P_\theta, Q_\theta)\,\nu(d\theta) \;\le\; \int \sqrt{\frac{I(\widehat{Y}_{\text{margin}}; Z \mid \Theta{=}\theta)}{2\pi(1-\pi)}}\,\nu(d\theta) \;\le\; \sqrt{\frac{\int I(\widehat{Y}_{\text{margin}}; Z \mid \Theta{=}\theta)\,\nu(d\theta)}{2\pi(1-\pi)}}.$$

The integral in the numerator is $I(\widehat{Y}_{\text{margin}}; Z)$. ■

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:

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

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:

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

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:

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

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). We apply Lemma 2.1 to the random variable Θ_u. Given the assumption $I(\widehat{Y}_{\text{margin}}; Z \mid \Theta_u) \le \varepsilon_u$ almost surely, and following the convexity argument in Lemma 2.1:

$$\text{(I)} = \mathrm{TV}\big(\mu^{p^d}_{\Theta_u}, \mu^{p^r}_{\Theta_u}\big) \;\le\; \mathbb{E}_{\Theta_u}\Big[\mathrm{TV}\big(\mu^{p^d}_{\Theta_u}, \mu^{p^r}_{\Theta_u}\big)\Big] \;\le\; \mathbb{E}_{\Theta_u}\left[\sqrt{\frac{I(\widehat{Y}; Z \mid \Theta_u)}{2\pi(1-\pi)}}\right] \;\le\; \sqrt{\frac{\varepsilon_u}{2\pi(1-\pi)}}.$$

Step 4: Anchor generalization (Term III). The term $\text{(III)} = \mathrm{TV}\big(\mu^{p^r}_{\Theta_r}, \mu^{p^d}_{\Theta_r}\big)$ 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 for contradiction that there exists a $y^* \in \mathcal{Y}$ such that $|\log(\frac{f(X_0)(y^*)}{f(X_0)(y^*)})| > \varepsilon$, then let δ be the radius around y^* that satisfies $\mathrm{sign}(\log(\frac{f(X_0)(y)}{f(X_1)(y)})) = \mathrm{sign}(\log(\frac{f(X_0)(y^*)}{f(X_1)(y^*)}))$. We have

$$\begin{aligned} \mathcal{W}_{d_\mathcal{Y}}(f(X_0), f(X_1)) &> \delta \int_{B_\delta(y^*)} |f(X_0) - f(X_1)|(y)dy \\ &\ge \delta |f(X_0) - f(X_1)|(y^*). \end{aligned}$$

Now, if there exists a $L \leq L^*(\varepsilon)$ such that f is L-Lipschitz, then we have

$$d_{\mathcal{Y}}(f(x), f(x')) \leq L d_{\mathcal{X}}(x, x') \leq L^*(\varepsilon) d_{\mathcal{X}}(x, x'). \tag{22}$$

But that implies

$$\begin{aligned}
\mathcal{W}_{d_{\mathcal{Y}}}(f(X_0), f(X_1)) &\leq L^*(\varepsilon)\mathcal{W}_{d_{\mathcal{X}}}(X_0, X_1) \\
&= \frac{\delta \left| f(X_1)(y^*) - f(X_0)(y^*) \right|}{\mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0)} \mathcal{W}_{d_{\mathcal{X}}}(X_0, X_1) \\
&= \delta \left| f(X_1)(y^*) - f(X_0)(y^*) \right|,
\end{aligned}$$

which contradicts $\mathcal{W}_{d_{\mathcal{Y}}}(f(X_0), f(X_1)) > \delta |f(X_0) - f(X_1)|(y^*)$. Therefore, f must violate the L-Lipschitz condition for any $L \leq L^*(\varepsilon)$. ■

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. In particular, given a data set X with a goal to compress X by an encoding $\hat{Y}$, the volume of code needed to encode X is $2^{H(X)}$ where $H(X)$ is the entropy of X. Furthermore, from the Chapman-Kolmogorov equation $p(\hat{Y}) = \sum_x p(\hat{Y}|x)p(x)$, the average volume of x mapped to individual $\hat{Y}$ is equal to $2^{H(X|\hat{Y})}$. Here,

$$H(X|\hat{Y}) := -\sum_x p(x) \sum_{\hat{Y}} p(\hat{Y}|x) \log(p(\hat{Y}|x)) \tag{23}$$

is the conditional entropy of X on $\hat{Y}$. Intuitively, a higher conditional entropy means more volume of x is expected to be mapped to individual $\hat{Y}$, which implies more randomness of X remains given the observation of $\hat{Y}$. In other words, less X is explained by $\hat{Y}$.

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

$$\frac{2^{H(X)}}{2^{H(X|\hat{Y})}} = 2^{I(X;\hat{Y})}. \tag{24}$$

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

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 $\hat{Y}$. Therefore, to maximize utility, we should maximize mutual information $I(X; \hat{Y})$ or, equivalently, minimize the compression rate.

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):

$$\begin{aligned}\{\omega : \hat{X}(\omega) \in A\} &= \hat{X}^{-1}(A) \\ &= (X, Z)^{-1}(f^{-1}(A)) \\ &= \{\omega : (X(\omega), Z(\omega)) \in f^{-1}(A)\}. \end{aligned} \tag{25}$$

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 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

$$\sigma(\hat{X}) \subset \sigma((X, Z)). \tag{26}$$

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:

$$\mathcal{A}(X, Z) \;:=\; \Big\{\hat{X} = f(X, Z) : f \text{ is } \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Z}}/\mathcal{B}_{\mathcal{X}}\text{-measurable}\Big\}, \tag{27}$$

and we use $\hat{X} = f(X, Z)$ and $\hat{X} \in \mathcal{A}(X, Z)$ interchangeably.

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:

$$\sup_{\hat{X} = f(X, Z)} \{\mathcal{U}(Y; \hat{X}) : \hat{X} \perp Z\}, \tag{28}$$

where $\mathcal{U}(Y; \hat{X})$ represents a utility 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 utility objectives and their corresponding constrained optimization problems, for which we provide a unified analytic feature unlearning solution in the next section.

Entropy maximization: The utility is defined as the entropy of the unlearning output, $\mathcal{U}(Y; \hat{X}) = H(\hat{X})$, which quantifies the information in $\hat{X}$. Entropy is commonly used to balance exploitation and exploration, such as in classifier training [53]. The optimization problem, *Entropy-Maximized Feature Unlearning*, is given by $\sup_{\hat{X}=f(X,Z)}\{H(\hat{X}) : \hat{X} \perp Z\}$. Alternatively, it can be interpreted as entropy-regularized mutual information minimization: $\sup_{\hat{X}=f(X,Z)}\{-I(Z;\hat{X}) + \frac{1}{\beta}H(\hat{X})\}$, where β controls the trade-off.

Mutual information maximization: The utility is defined as $\mathcal{U}(Y;\hat{X}) = I(Y;\hat{X})$, 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 [36] and information bottleneck methods [3]. The corresponding optimization problem, *Mutual-Information-Maximized Feature Unlearning*, is given by $\sup_{\hat{X}=f(X,Z)}\{I(Y;\hat{X}) : \hat{X} \perp Z\}$. Since $I(Y;\hat{X}) = H(Y) - H(Y|\hat{X})$, this problem is equivalent to minimizing the conditional entropy $H(Y|\hat{X})$. As a result, it provides an optimal solution for utility preservation with respect to any target variable Y.

KL-divergence maximization: The utility is $\mathcal{U}(Y;\hat{X}) = -D_{KL}(\mathbb{P}(Y|\hat{X})||\mathbb{P}(Y))$, where D_{KL} measures the divergence between the predicted and prior distributions of Y. This objective is commonly applied in generative models such as Variational Autoencoders (VAEs) [35]. The corresponding optimization problem, *KL-Divergence-Maximized Feature Unlearning*, is formulated as $\sup_{\hat{X}=f(X,Z)}\{D_{KL}(\mathbb{P}(Y|\hat{X})||\mathbb{P}(Y)) : \hat{X} \perp Z\}$. This problem seeks to make $\mathbb{P}(Y|\hat{X})$ as deterministic as possible relative to the prior $\mathbb{P}(Y)$, thereby enhancing the predictive power of $\hat{X}$ for Y.

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

$$\mathcal{U}(Y;\hat{X}) = \begin{cases} \sqrt{\mathbb{E}_{\hat{X}}\left[\mathbb{P}(Y \in A|\hat{X})^2\right]} & \text{for classification,} \\ -||Y - \mathbb{E}(Y|\hat{X})||_2 & \text{for regression.} \end{cases}$$

The corresponding optimization problem, *Energy-Maximized Feature Unlearning*, is formulated as $\sup_{\hat{X}=f(X,Z)}\{||\mathbb{P}(\{Y \in A_Y\}|\hat{X})||_2^2 : \hat{X} \perp Z\}$. A higher L^2-norm indicates a more precise prediction of the event $\{Y \in A_Y\}$ based on $\hat{X}$, leading to reduced Bayes error and improved decision boundaries.

As we show in the next section, when the above-listed objectives are applied, there exists a universal optimal feature unlearning solution to the general formulation: equation (28) for arbitrary target variable 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 (Entropy-maximized feature unlearning)

$$\sup_{\hat{Y} \in \mathcal{A}(X,Z)} \{H(\hat{Y}) : \hat{Y} \perp Z\}.$$

Here, H denotes the entropy (or differential entropy), which measures the information contained in the unlearning outcome $\hat{Y}$. As discussed, entropy is a fundamental metric in information theory and probability for quantifying information and randomness. Thus, Problem 1 seeks to optimally compress (X, Z) to produce $\hat{Y}$ with the information about Z effectively removed.

Problem 2 (Conditional-entropy-minimized feature unlearning)

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

In many cases, an unlearning output $\hat{Y}$ may be used to generate inferences or predictions for some random variable Y. Thus, it is also desirable to solve Problem 2 for some dependent variable Y. Notice that, due to $I(X;Y) = H(Y) - H(Y|X)$, the above problem shares the same solution as the maximization of mutual information between Y and $\hat{Y}$:

Problem 3 (Mutual-information-maximized feature unlearning)

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

Notably, the optimal solution to Problems 2 and 3 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 $\hat{Y}$. Thus, despite the explicit presence of Y in Problem 2, it provides a generalized solution for any choice of Y.

Problem 4 (KL-divergence-maximized feature unlearning)

$$\sup_{\hat{Y} \in \mathcal{A}(X,Z)} \{D_{KL}(\mathbb{P}(Y|\hat{Y})||\mathbb{P}(Y)) : \hat{Y} \perp Z\}.$$

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 $\hat{Y}$. Therefore, it is desirable to make $\mathbb{P}(Y|\hat{Y})$ as deterministic as possible relative to the original distribution of Y. To quantify this determinism, we use the KL-divergence of $\mathbb{P}(Y|\hat{Y})$ relative to $\mathbb{P}(Y)$, leading to the optimization problem above. Intuitively, a more accurate prediction of $\mathbb{P}(Y|\hat{Y})$ implies less randomness relative to $\mathbb{P}(Y)$, which increases the KL-divergence of $\mathbb{P}(Y|\hat{Y})$ relative to $\mathbb{P}(Y)$.

Problem 5 (Energy-maximized feature unlearning)

$$\sup_{\hat{Y} \in \mathcal{A}(X,Z)} \{||\mathbb{P}(\{Y \in A_Y\}|\hat{Y})||_2^2 : \hat{Y} \perp Z\}.$$

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

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 $\mathbb{R}^d$,

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

Here, $\prod(\mu,\nu) := \{\pi \in \mathcal{P}((\mathbb{R}^d)^2) : \int_{\mathbb{R}^d} d\pi(\cdot, v) = \mu, \int_{\mathbb{R}^d} d\pi(u,\cdot) = \nu\}$. $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ 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 [64], then

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

Given $\{\mu_z\}_{z \in \mathcal{Z}} \subset (\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ for some index set $\mathcal{Z}$, their Wasserstein barycenter [2] with weights $\lambda \in \mathcal{P}(\mathcal{Z})$ is

$$\bar{\mu} := \text{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mu) d\lambda(z) \right\}. \tag{29}$$

If there is no danger of confusion, we will refer to the Wasserstein barycenter simply as barycenter. Also, we use $\bar{X}$ to denote the random variable such that $\mathcal{L}(\bar{X}_z) = (T_z)_\sharp \mathcal{L}(X_z)$ where T_z is the optimal transport map from $\mathcal{L}(X_z)$ to $\bar{\mu}$. $\bar{\mu}$ is the barycenter of $\mu_z = \mathcal{L}(X_z)$ for all z.

D.2 Proof of Lemma 4.1

Proof First, 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

$$\begin{aligned} |\mathbb{P}(Z=0|\hat{Y}) - \mathbb{P}(Z=1|\hat{Y})| &= |2\mathbb{P}(Z=0|\hat{Y}) - 1| &(30)\\ &\leq 2|\mathbb{P}(Z=0|\hat{Y}) - \mathbb{P}(Z=0)| + 2|\mathbb{P}(Z=0) - 0.5| &(31)\\ &\leq ||\mathbb{P}(Z|\hat{Y}) - \mathbb{P}(Z)||_{TV}, &(32) \end{aligned}$$

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 $\hat{Y}$, we have

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

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 $\varepsilon>0$, it follows from Markov's inequality that

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

Finally, it follows from

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

that

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

and

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

Since $\{Z=0\}=\{X_{\text{margin}}=X_0\}$ and $\{Z=1\}=\{X_{\text{margin}}=X_1\}$ by construction, the proof is complete. ■

D.3 Proof of Lemma 4.2

Proof See Lemma 5.2 in [71] for the full proof. We provide a sketch here: Under the absolute continuity assumption of X_z's, we have the Wasserstein-2 barycenter is also absolute continuous. Therefore, the optimal transport maps between the barycenter and each of the X_z are invertible (almost everywhere). This implies that there exists invertible measurable maps between $(\bar{X}, Z)$ and (X, Z) and hence $\sigma((\hat{Y}, Z)) \subset \sigma((X, Z)) = \sigma((\bar{X}, Z))$. Now, by the independence constraint, we also have $\sigma((\bar{X}, Z)) = \sigma(\bar{X}) \otimes \sigma(Z)$ and $\sigma((\hat{Y}, Z)) = \sigma(\hat{Y}) \otimes \sigma(Z)$. Therefore, it follows from $\sigma(\hat{Y}) \otimes \sigma(Z) \subset \sigma(\bar{X}) \otimes \sigma(Z) \implies \sigma(\hat{Y}) \subset \sigma(\bar{X})$ that $\sigma(\hat{Y}) \subset \sigma(\bar{X})$ for all admissible $\hat{Y}$. 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), 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 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 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 $\sigma(X_1) \subset \sigma(X_2)$. For entropy, we have $H(X_2) - H(X_1) = H(X_2|X_1) \geq 0$. Therefore, $H(X_2) - H(X_1)$.

For mutual information, it follows from the assumption $\sigma(X_1) \subset \sigma(X_2)$ that there exists a measurable function $g(X_2) = X_1$. Therefore, given any Y, we have Y is conditionally independent of X_1 given X_2 because X_1 is a constant given X_2. That is, $Y \to X_2 \to X_1$ forms a Markov chain. It follows from the data-processing inequality [18] that $I(Y; X_1) \leq I(Y; X_2)$.

For the conditional entropy, we have $H(Y|X_s) = H(Y) - I(Y; X_s)$ for $s \in \{1, 2\}$. Therefore, it follows from $I(Y; X_1) \leq I(Y; X_2)$ that $H(Y|X_2) \leq H(Y|X_1)$.

For KL-divergence, it follows from the convexity of the divergence in the first argument, the $\mathbb{P}(Y|X_1) = \mathbb{E}(\mathbb{P}(Y|X_2)|X_1)$, and Jensen's inequality that

$$D_{KL}(\mathbb{P}(Y|X_1)||\mathbb{P}(Y)) \leq \mathbb{E}_{X_2}(D_{KL}(\mathbb{P}(Y|X_2)||\mathbb{P}(Y))|X_1).$$

Finally, by taking expectation w.r.t. X_1 on both sides, we have

$$D_{KL}(\mathbb{P}(Y|X_1)||\mathbb{P}(Y)) \leq D_{KL}(\mathbb{P}(Y|X_2)||\mathbb{P}(Y)).$$

For the conditional probability energy, it follows directly from $\mathbb{P}(\{Y \in A_{\mathcal{Y}}|X_1\}|X) = \mathbb{E}(\mathbb{1}_{Y \in A_{\mathcal{Y}}}|X)$ and the tower property of conditional expectation: If $\sigma(X_1) \subset \sigma(X_2)$,

$$\mathbb{E}(\mathbb{1}_{Y \in A_{\mathcal{Y}}}|X_1) = \mathbb{E}(\mathbb{E}(\mathbb{1}_{Y \in A_{\mathcal{Y}}}|X_2)|X_1).$$

■

D.6 Intuitive insights into Theorem 4.3

- Among all admissible $\hat{Y}$ outcomes, $\bar{X}$ maximizes randomness or information when quantified by entropy.

- Given $\bar{X}$, the conditional probability distribution of Y retains the least randomness among all admissible $\hat{Y}$ outcomes, with randomness measured by conditional entropy.

- From the information-theoretical perspective, since $\bar{X}$ contains the most information among the admissible, it contains the most mutual information to Y compared to other admissible.

- The conditional probability $\mathbb{P}(Y|\bar{X})$ provides the greatest certainty (or least randomness) relative to $\mathbb{P}(Y)$, with the reduction in randomness measured by KL-divergence.

- Assuming sufficiently regularized sensitive distributions with a density function, $\bar{X}$ achieves higher or equal energy (or L^2-norm) in $\mathbb{P}(Y|\bar{X})$ for any random variable Y and any event generated by Y, compared to other admissible $\hat{Y}$ outcomes.