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 individual data points, from a learning outcome while minimizing utility loss and ensuring rigorous guarantees? We introduce a mathematical framework based on information-theoretic regularization to address both feature and data point unlearning. For feature unlearning, we derive a unified solution that simultaneously optimizes diverse learning objectives, including entropy, conditional entropy, KL-divergence, and the energy of conditional probability. For data point unlearning, we first propose a novel definition that serves as a practical condition for unlearning via retraining, is easy to verify, and aligns with the principles of differential privacy from an inference perspective. Then, we provide provable guarantees for our framework on data point unlearning. 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.

Keywords: Machine Unlearning, Feature Unlearning, Data Privacy, Mutual Information, Optimal Transport

Contents

1. Introduction

As machine learning models become more common in sensitive domains, removing specific features or data points from trained models has become increasingly important. Sensitive information such as gender, ethnicity, or private data can perpetuate biases and lead to unfair results. Simply deleting these attributes in raw data is often insufficient, since their influence may persist through correlated or latent variables. Retraining models without these attributes can be impractical due to high computational costs. Machine unlearning provides a systematic way to "forget" specific features or data points, ensuring legal compliance and ethical alignment. It is critical to develop provable methods so that models can adapt to evolving demands for privacy and fairness.

- **Feature Unlearning:** In 2014, Amazon built a machine learning based recruitment tool [11, 7]. The system favored male candidates because the training data came predominantly from men, penalizing words commonly associated with women. Removing explicit gender references failed, as gender correlated with other features (e.g., all-female colleges). Unable to resolve these biases, Amazon discontinued the tool. This example shows that merely dropping a sensitive feature is not enough if correlated information remains.

- **Data Point Unlearning:** The General Data Protection Regulation (GDPR) [14] introduced the "right to be forgotten," letting individuals request deletion of their personal data. Although removing raw data is straightforward, it does not eliminate that data's effect on a trained model's parameters. The challenge is how to unlearn specific data points while preserving overall model performance, highlighting the need for robust frameworks offering provable compliance and minimal utility loss.

In this work, we introduce a framework for machine unlearning within a probabilistic and information-theoretic setting, addressing the following question:

> *Given a dataset denoted by X, an undesirable attribute Z, and a target variable Y, how can we optimally modify the information in (X, Z) to produce an unlearning outcome S which retains minimal information about Z while preserving as much information in X about Y as possible?*

For machine unlearning, we consider $S = \hat{X}$ if the unlearning happens on the data space $\mathcal{X}$ and $S = \hat{Y}$ if the unlearning happens on the learning outcome space $\mathcal{Y}$ (via unlearning on the model parameter space). Note that the feature or label Z may also represent identity information, which blurs the line between feature unlearning and data point unlearning, making the framework applicable to both. Further discussion on applying the framework to data point unlearning can be found in Section 1.2 below. In the remainder of this work, we focus on unlearning in settings where relational data (X, Z) and target data Y are available.

To address this problem, we connect modern machine unlearning to classic rate-distortion theory and data compression [34]. Here, we use data compression in the sense of minimizing $I(S; Z)$, thereby reducing shared information and effectively removing unwanted details of Z. We introduce two core concepts inspired by this information-theoretic framework:

- *Unlearning*, achieved by compressing the relational data (X, Z) into an unlearning outcome S such that the compression rate (quantified by mutual information in this work) between S and the undesirable information Z is minimized.

- *Utility*, preserved by minimizing the distortion (or maximizing the utility quantification) guided by Y, such as the mutual information between S and Y when $S = \hat{Y}$.

Following the above information-theoretic perspective, we propose a unified machine unlearning framework applicable to both feature and data point unlearning for a variety of downstream tasks.

1.1 Application to Feature Unlearning

For feature unlearning, we define Z as the feature(s) to be unlearned, X as the remaining available data features, and Y as the target variable to estimate using the information in (X, Z). Since feature unlearning often involves removing a feature from a dataset intended for multiple downstream tasks, it is more practical to modify (X, Z) directly in the data space, producing an unlearning outcome $S = \hat{X}$. The objective is to generate $\hat{X}$ that maximizes utility (considering multiple Y), while minimizing or compressing the information related to Z to achieve effective unlearning.

Alternatively, if feature unlearning is performed for a specific Y or a specific model, one may set $S = \hat{Y}$. See Algorithm 2 in Section 4.

1.2 Application to Data Point Unlearning

For data point unlearning, the application is less straightforward and requires a more detailed explanation. Here, we discuss the key distinctions between feature unlearning and data point unlearning and clarify how the proposed unlearning method can be applied to data point unlearning by defining a probabilistic data point unlearning guarantee.

First, since data point unlearning operates on the learning outcome space (through unlearning on the parameter space), we define $S = \hat{Y}$ for data point unlearning.

It is crucial to emphasize that data point unlearning typically aims to "forget" the marginal effect of adding/removing a data point on the rest of the dataset, rather than removing all the information associated with that data point. For example, the baseline *unlearning via retraining approach* aims to remove the marginal effect (on training) by directly removing the data to unlearn from training. Here, we apply information theoretic regularization and the existing model to directly estimate the marginal effect on learning outcome and thereby penalize it during training.

A straightforward application of an information theoretic framework, such as via the information bottleneck, might minimize $I(X; \hat{Y})$, the mutual information between X and $\hat{Y}$ (see Appedix A.1 for more details), while maximizing $I(X \setminus \{x_u\}; \hat{Y})$ [21], where x_u represents the data belonging to the individual requiring unlearning. However, enforcing perfect unlearning by setting $I(X; \hat{Y}) = 0$ would leave almost no utility in $\hat{Y}$. This approach unnecessarily targets removing all data information rather than focusing solely on unlearning the marginal effect of x_u, leading to an excessive loss of information and failing to balance unlearning with utility.

It is thus more reasonable to pursue a probabilistic, information-theoretic approach which constructs relational data to indicate whether a data point is included in the training step: I.e., we consider (X_{train}, Z), where $X_{train}|_{Z=1} := X_1 := \{x_i\}_{i=1}^N$ with $x_i \sim \text{Uniform}(X \backslash \{x_u\})$ and $X_{train}|_{Z=0} := X_0 := \{x_j\}_{j=N+1}^{2N-1} \cup \{x_u\}$ with $x_j \sim \text{Uniform}(X \backslash \{x_u\})$. In this formulation, perfect unlearning of the marginal effect of x_u can be achieved by finding a measurable function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes $I(Z; \hat{Y}_{train})$, where $\hat{Y}_{train} = f(X_{train})$, without directly conflicting with utility objectives $\mathcal{U}(Y; \hat{Y})$, such as $I(Y; \hat{Y})$ or $||Y - \hat{Y}||_{\ell^2}$.

It is important to highlight that the data point unlearning we propose is fundamentally distinct from existing approaches (exact or approximate unlearning, see more details in Section 1.3) that aim to achieve the same result as a retrained model on a dataset excluding the individual record. Instead, our approach will provide a probabilistic guarantee that aligns with the principles of differential privacy from an inference perspective. Specifically, we formalize the data point unlearning guarantee as follows:

Definition 1.1 (ε-Differential Unlearning) *Given a dataset X and an individual record $\{x_i\}$ that requires unlearning, an unlearning output $f_{\theta(X,Y,Z)}$ satisfies* ε-differential unlearning *if*

$$\sup_{D \in \mathcal{B}_{\hat{Y}}} \left| \log \left(\frac{\mathbb{P}(\{X_{train} = X_0\} \mid \hat{Y}_{train} \in D)}{\mathbb{P}(\{X_{train} = X_1\} \mid \hat{Y}_{train} \in D)} \right) \right| \leq \varepsilon. \tag{1}$$

That is, given any observation of the unlearning output $\hat{Y}_{train} = f_\theta(X_{train})$ and the knowledge that the training process uses either the original dataset X or the dataset excluding x_u, it is impossible (up to ε inference capability) to determine whether the individual record x_u was included in the training. We briefly summarize the advantages of the proposed definition:

- **Necessary Condition for Exact Unlearning**: Under mild assumptions, one can show that any unlearned model that is assumed to estimate an exact unlearning outcome (via retraining) can fail the ε-differential unlearning definition only if it is non-smooth (in particular, violates a Lipschitz condition) and hence suffers from low generalizability for unseen data (See Lemma 2.2 and its discussion below). This establishes ε-differential unlearning as a foundational requirement for exact unlearning.

- **Ease of Verification**: The definition relies solely on the unlearned model and the original dataset X, without requiring access to the exact retrained outcome. As a result, ε-differential unlearning can serve as an efficient preliminary test to assess unlearning outcomes.

This unlearning concept is inspired by differential privacy [12]. It is instructive to compare ε-differential unlearning to differential privacy to highlight their differences:

- *ε-Differential Unlearning (Remediation via Inference Guarantee)*: Requires that the unlearning outcome (via data compression) provides no inference capability about whether the training process included the individual record requiring unlearning. Since unlearning is a *remedial* process, it is more reasonable to consider it from an inference perspective: $\sup_{D \in \mathcal{B}_\mathcal{Y}} |\log \left(\frac{\mathbb{P}(\{X_{train}=X_1\} | \hat{Y}_{train} \in D)}{\mathbb{P}(\{X_{train}=X_0\} | \hat{Y}_{train} \in D)} \right)| \leq \varepsilon$. Here, $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ is

a measurable space for the learning outcome, where $\mathcal{B}_\mathcal{Y}$ denotes the sigma-algebra containing all possible events D that can occur in the outcome space $\mathcal{Y}$.

- *ε-Differential Privacy (Prevention via Generative Guarantee)*: Requires that the learning outcome (via a randomized algorithm) does not change significantly if an unknown individual's data is added to or removed from the training dataset. Since privacy focuses on *prevention*, it is more reasonable to provide guarantees from a generative perspective: $\sup_{D\in\mathcal{B}_\mathcal{Y}} |\log\left(\frac{\mathbb{P}(\hat{Y}\in D|X)}{\mathbb{P}(\hat{Y}\in D|X')}\right)| \leq \varepsilon$, where X' is the dataset that differs from X by one (unknown) data point.

It is worth noting that differential unlearning can also be framed as a straightforward extension of the original differential privacy definition from a generative perspective: $|\log\left(\frac{\mathbb{P}(\hat{Y}|X)}{\mathbb{P}(\hat{Y}|X\setminus\{x_u\})}\right)| < \varepsilon$ implies that the unlearning outcome $\hat{Y}$ should remain largely unchanged if X is replaced with $X\setminus\{x_i\}$ during training or vice versa. While this perspective offers valuable insights and practical tools for achieving unlearning, we argue that the generative approach is better understood as a *means* to unlearning rather than a *definition*, because a definition of unlearning should capture the remedial nature inherent to machine unlearning.

We will demonstrate in Section 2 that the proposed framework naturally produces unlearning outputs that satisfy the ε-differential unlearning guarantee when applied to data point unlearning tasks.

1.3 Related Work

The existing body of work on unlearning can be separated into feature unlearning and machine unlearning:

1.3.1 Machine Unlearning

Machine Unlearning focuses on removing data points' influence during training on the model and thereby the learning outcome, mainly to comply with regulations such as GDPR's "right to be forgotten." The current study can be separated into exact unlearning and approximate unlearning:

Exact unlearning [3, 4, 44, 19, 32, 15, 6, 5] approaches ensure the removal of data influence from the model by separating the model into sub-models (via subsets of training data in SISA training, parameters, or training steps in amnesiac unlearning) which allow fast retraining by only retraining the affected sub-models or training steps. While effective, these methods are computationally expensive, particularly in scenarios with multiple data removal requests.

Approximate unlearning methods try to keep the probabilistic similarity quantification between the unlearning outcome and the retraining outcome (also known as data point influence quantification). The main current approaches include: (1) using the influence function [40] to estimate the influence of the data point and thus remove its effect on the model [20, 33, 35, 29, 41, 36, 39]; (2) using the scrubbing function to enforce similarity between the network weights and the retraining weights [16, 17].

While these methods are effective in approximating the retrained model, each has practical limitation(s), such as significant performance degradation, high computational cost, limited compatibility with learning objectives, or restricted evaluation capability on simple datasets [30, 38]. More importantly, due to the aim to estimate the unknown (due to the complexity of algorithms, objective functions, and data influence) retraining outcome, the provable unlearning guarantees rely heavily on impractical assumptions such as convexity of the objective function and Lipschitz condition on Hessian matrices. [42]

In comparison, the proposed method can be considered as using mutual information to quantify the marginal effect of adding or removing the data point to unlearn in training, thereby more effectively removing the influence of the data to unlearn by diminishing the marginal effect. Furthermore, since the proposed unlearning definition (Definition 1.1) does not explicitly depend on the unknown retraining outcome, its provable guarantees avoid strong assumptions and become more practical to achieve. Finally, we establish a connection to retraining by demonstrating that Definition 1.1 serves as a necessary condition for a "good" retrained model (Lemma 2.2 and the discussion below).

Recent unlearning methods have leveraged the information bottleneck [26, 21] to avoid utility degradation by addressing both the utility and influence function. While at first glance these information bottleneck-based methods may seem similar to our proposed method, there is a fundamental difference. In particular, these information bottleneck methods aim to unlearn the information of that data point itself, such as minimizing the mutual information between the remaining target variable and the model parameter [21] or the mutual information between the data to unlearn and the unlearning latent representations [26]. In comparison, our regularization tries to help unlearn (or learn) the marginal effect of adding (removing) the data point.

1.3.2 Feature Unlearning and Statistical Parity in Machine Learning Fairness

Feature Unlearning and Statistical Parity in Machine Learning Fairness aim to remove the influence of a feature in the learning outcome [39, 21]. It is closely related to machine learning fairness when fairness is defined as statistical parity [13]. In particular, our framework when applied to feature unlearning is closely related to the Wasserstein barycenter characterization of both the optimal statistical parity learning outcome [8, 18] and the optimally fair data representation [43].

1.4 Our Contributions

Our contributions in this paper are as follows:

- We propose *ε-Differential Unlearning*, a novel probabilistic guarantee for data point(s) unlearning that reflects the remedial nature of machine unlearning, inspired by the protection guarantees of differential privacy.
- We propose a unified machine unlearning framework, inspired by *rate-distortion theory* in *data compression*, for both feature and data point unlearning. The framework is compatible with diverse target information and utility quantifications, and hence with a range of downstream tasks for the unlearned output.

- For feature unlearning, we demonstrate the existence of a unified analytic solution across multiple utility objectives and arbitrary target variables.
- For data point unlearning, we provide a provable ε-differential unlearning guarantee using compression rate and, thereby, provide a sufficient condition for the proposed unlearning framework to generate an ε-differential unlearning guaranteed result.

1.5 Some Tools and Notation

Before we proceed, we introduce some tools used later and fix some notation. Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on a metric space $(\mathcal{X}, d)$. The *p-Wasserstein distance* between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is defined as

$$\mathcal{W}_{d,p}(\mu,\nu) = \Big(\inf_{\lambda\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x_1,x_2)^p \, d\lambda(x_1,x_2)\Big)^{\frac{1}{p}},$$

where $\Pi(\mu,\nu)$ denotes the set of all couplings of μ and ν. Let $\mathcal{P}_{2,ac}(\mathcal{X})$ be the subset of measures with finite second moments that are absolutely continuous w.r.t. the Lebesgue measure. For random variables X_1, X_2, we write $\mathcal{W}_{d,p}(X_1,X_2) := \mathcal{W}_{d,p}(\mathcal{L}(X_1),\mathcal{L}(X_2))$, where $\mathcal{L}(X)$ denotes the law of X. For simplicity, we let $\mathcal{W}_d := \mathcal{W}_{d,1}$. Also, $\mathcal{W}_2 := \mathcal{W}_{l^2,2}$ when $d(x,x') := ||x-x'||_{l^2}$. Given a family of measures $\{\mu_z\}_{z\in\mathcal{Z}} \subset \mathcal{P}_{2,ac}(\mathcal{X})$ with weights λ, their *Wasserstein barycenter* $\bar{\mu}$ is the minimizer of $\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z,\mu)\, d\lambda(z)$. We define $\bar{X}$ as the random variable distributed according to $\bar{\mu}$ and satisfy $\bar{X}_z = T_z(X_z)$, where T_z is the optimal transport map from $\mathcal{L}(X_z)$ to $\bar{\mu}$. More details can be found in Appendix B.1.

2. A Unified Unlearning Framework

We first introduce our machine unlearning framework, followed by a concrete example illustrating the data compression motivation behind this approach and a brief review of key technical tools underlying the definitions of utility, unlearning, and admissibility. Next, we extend the motivational example from feature unlearning (Section 2.1) to the data point unlearning (Section 2.2) setting. Finally, we generalize the framework to accommodate different choices of target information and utility quantification for both feature and data point unlearning.

Definition 2.1 (Optimal Feature Unlearning) *Given relational data (X, Z) and a target variable Y, optimal feature unlearning is defined as the solution, if it exists, to:*

$$\sup_{f:\mathcal{X}\times\mathcal{Z}\to\mathcal{X}} \{\mathcal{U}(Y;\hat{X}) : \hat{X} \perp Z\}, \tag{2}$$

where $\hat{X} := f(X,Z)$ is the unlearning outcome, and $\mathcal{U} : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ quantifies the utility retained in $\hat{X}$ relevant to Y. The optimization is over all measurable functions f that remove the information of Z while preserving the information of Y from X.

While setting $S = \hat{X}$ instead of $\hat{Y}$ may seem unconventional, this choice ensures compatibility with multiple target variables Y, making it practical for feature unlearning. Moreover, Theorem 3.1 shows that under a wide range of utility functions, the optimal solution is independent of Y.

To relax the strict independence constraint in (2), we introduce a soft probabilistic constraint using mutual information as a compression rate:

$$\sup_{f:\mathcal{X}\times\mathcal{Z}\to\mathcal{X}} \mathcal{U}(Y;\hat{X}) - \gamma I(\hat{X};Z), \tag{3}$$

for some $\gamma \geq 0$, allowing a trade-off between utility preservation and unlearning effectiveness.

For data point unlearning, we follow the setting developed in Section 1.2:

Definition 2.2 (Optimal Differential Unlearning) *Given data X, a data point $\{x_u\}$ to unlearn, and a target variable Y, the optimal differential unlearning is defined as the optimal solution, if it exists, to the following problem:*

$$\sup_{f:\mathcal{X}\to\mathcal{Y}} \{\mathcal{U}(Y;\hat{Y}) : \hat{Y}_{train} \perp Z\}, \tag{4}$$

where $\hat{Y}_{train} := f(X_{train})$ is the unlearning outcome. The relational data (X_{train}, Z) is defined by $X_{train}|_{Z=0} = X_0$ and $X_{train}|_{Z=1} = X_1$. The optimization is over all measurable functions $f : \mathcal{X} \to \mathcal{Y}$ of which the goal is to leverage the relational data (X_{train}, Z) to remove the information of the training data set indicator Z but retain information of Y based on X.

Similarly to above, we can also relax equation (4) to the following version:

$$\sup_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{U}(Y;\hat{Y}) - \gamma I(\hat{Y}_{train};Z).$$

Remark 2.1 (Alternative Regularization to Mutual Information) *The regularization term using mutual information in our framework serves to relax the strict independence constraint through a soft penalization. While mutual information is a natural choice due to its strong theoretical foundations in information theory and its connection to data compression that motivates the proposed framework (See Section 2.1 below), it is by no means the only option. More generally, any quantification of statistical dependence between two random variables can serve as an alternative regularizer.*

For instance, in the binary Z case, one could use the Kullback-Leibler (KL) divergence between $\hat{Y}_0$ and $\hat{Y}_1$, i.e., $D_{KL}(\hat{Y}_0||\hat{Y}_1)$. Alternatively, if the objective is to account for both distributional differences and geometric distance in the learning outcome space, the Wasserstein distance, $\mathcal{W}_{d,p}(\hat{Y}_0, \hat{Y}_1)$, provides a viable alternative. A systematic investigation into the optimal choice of regularizer for specific utility objectives is an interesting direction for future research.

2.1 Motivation: Feature Unlearning & Data Compression

To illustrate our framework's link to data compression, consider the special case where the target variable is the original data ($Y = X$) and utility is quantified by mutual information ($\mathcal{U}(X;\hat{X}) := I(X;\hat{X})$). We then solve:

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

Here, $I(X;\hat{X})$ measures unlearning quality: larger values indicate $\hat{X}$ retains more of X. Meanwhile, $\hat{X} \perp Z$ enforces complete removal of Z by making $I(Z;\hat{X}) = 0$. Next, we discuss how utility, unlearning, and admissibility fit into this context.

- **Compression Rate via Mutual Information**: Mutual information measures shared information between two variables and is fundamental in data compression [9]. For dataset X encoded as $\hat{X}$, the information contained in X is quantified by $H(X)$ where $H(X)$ is the entropy of X. The conditional entropy $H(X|\hat{X})$ quantifies the remaining uncertainty in X given $\hat{X}$, with higher values indicating less explanatory power of $\hat{X}$. The compression ratio, $2^{H(X)-H(X|\hat{X})} = 2^{I(X;\hat{X})}$, reflects how well $\hat{X}$ retains information about X. Higher mutual information implies $\hat{X}$ generates a finer partition on X, while lower values indicate better compression. See Appendix A.1 for details.

- **Feature Unlearning**: To unlearn a feature Z from data X, we minimize $I(Z;\hat{X})$, measuring how much information $\hat{X}$ retains about Z. Perfect unlearning occurs if $I(Z;\hat{X}) = 0$, i.e., $\hat{X} \perp Z$. In practice, partial unlearning is more realistic, balancing the rate-distortion-like trade-off between removing Z and preserving utility. By bounding $I(Z;\hat{X}) < \varepsilon$ (or equivalently introducing a Lagrange multiplier γ), we obtain $\sup_{\hat{X}=f(X,Z)}\{\, I(X;\hat{X}) - \gamma\, I(Z;\hat{X})\}$, where $\gamma > 0$ governs the trade-off between minimizing unwanted information and retaining the utility in $\hat{X}$.

- **Admissibility**: Since we unlearn the information of Z by compressing relational data (X, Z), it is natural to require the resulting compressed data $\hat{X}$ to be measurable with respect to (X, Z). In particular, for every event or observation A of the compressed $\hat{X}$, the information 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):

  $$\{\omega : \hat{X}(\omega) \in A\} = \hat{X}^{-1}(A) = (X, Z)^{-1}(f^{-1}(A)) = \{\omega : (X(\omega), Z(\omega)) \in f^{-1}(A)\}.$$

  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 $\sigma(\hat{X}) \subset \sigma((X, Z))$. 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: $\mathcal{A}(X, Z) := \{\hat{X} = f(X, Z) : f \text{ is } \mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Z}/\mathcal{B}_\mathcal{X}\text{-measurable}\}$ and we use $\hat{X} = f(X, Z)$ and $\hat{X} \in \mathcal{A}(X, Z)$ interchangeably.

2.2 Motivation: Data Point Unlearning & Data Compression

We now connect the feature unlearning constraint using mutual information to ε-differential unlearning (Definition 1.1) and demonstrate that the proposed feature unlearning framework with a mutual information constraint can directly ensure ε-differential unlearning for data points.

In particular, we apply the following framework to estimate the optimal differential unlearning solution:

$$\sup_{\hat{Y}=f(X)} \{I(Y;\hat{Y}) - \gamma I(\hat{Y}_{train}; Z)\}, \tag{5}$$

and then use the following result to provide a provable unlearning guarantee:

Lemma 2.1 (Mutual Information Bound on Unlearning Inference Log Ratio) *Let the proposed unlearning framework (equation (5)) compress $(\hat{Y}_{train}, Z)$. Then, for any given $\varepsilon > 0$, we have*

$$\mathbb{P}\Big(\Big|\log\Big(\frac{\mathbb{P}(\{X_{train} = X_0\} \mid \hat{Y}_{train})}{\mathbb{P}(\{X_{train} = X_1\} \mid \hat{Y}_{train})}\Big)\Big| \;\le\; \log\Big(\frac{1+\varepsilon}{1-\varepsilon}\Big)\Big) \;\ge\; 1 \;-\; \frac{1}{\varepsilon}\left(\sqrt{\frac{1}{2}\, I(Z;\hat{Y}_{train})}\right).$$

See proof in Appendix A.3. That is, if the proposed unlearning framework is applied to penalize the mutual information term $I(Z;\hat{Y}_{train})$ such that it remains relatively small compared to $\frac{\exp(\varepsilon)-1}{\exp(\varepsilon)+1}$, then the probability of observing an event in $\hat{Y}_{train}$ that grants more than $\frac{1+\varepsilon}{1-\varepsilon}$ inference capability becomes extremely low. By the construction of $\hat{Y}_{train}$, it is clear that even if we observe the learning outcomes from both the data set including data point to unlearn and the one excluding the point, one can still not tell whether or not the model is trained using $X_0 \sim X$ or $X_1 \sim X \setminus \{x_u\}$. Therefore, to ensure ε-differential unlearning with high probability, it suffices to adjust the regularization weight γ such that $\sqrt{I(Z;\hat{Y}_{train})}$ is small relative to $\frac{\exp(\varepsilon)-1}{\exp(\varepsilon)+1}$.

Lastly, we connect the proposed ε-differential unlearning definition to the unlearning via retraining by showing that it serves as a practical necessary condition for a "good" retraining model. In particular, one can show that if a model, denoted by $f : \mathcal{X} \to \mathcal{Y}$, can minimize training loss on the training data, then either the unlearned model satisfies the ε-differential unlearning definition or it suffers from low generalizability by violating a Lipschitz condition as described below.

Lemma 2.2 (ε-DU as a Condition for "Good" Retrained Model) *Assume f is a retrained model with $f(X_0)$ and $f(X_1)$ both absolutely continuous, then either f violates ε-DU: $\sup_y |\log(\frac{\mathcal{L}(f(X_1))(y)}{\mathcal{L}(f(X_0))(y)})| \le \varepsilon$ or f violates the L-Lipschitz condition for any $L \le L^*$ where*

$$L^*(\varepsilon) := \frac{\delta\,\big|f(X_1)(y^*) - f(X_0)(y^*)\big|}{\mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0)} = \frac{\delta\,\big(\exp(\varepsilon) - 1\big)\,\min\{f(X_0)(y^*),\, f(X_1)(y^*)\}}{\mathcal{W}_{d_{\mathcal{X}}}(X_1, X_0)}. \tag{6}$$

Here, y^* is the point where the ε-differential unlearning definition is violated, $\mathcal{W}_{d_{\mathcal{X}}}(X_0, X_1)$ is the Wasserstein distance between X_0 and X_1 on the metric space $(\mathcal{X}, d_{\mathcal{X}})$, and δ is the radius around y^* where $\operatorname{sign}(\log(\frac{\mathcal{L}(f(X_0))(y)}{\mathcal{L}(f(X_1))(y)})) = \operatorname{sign}(\log(\frac{\mathcal{L}(f(X_0))(y^*)}{\mathcal{L}(f(X_1))(y^*)}))$.

See proof in Appendix A.4. The above result implies that, if we assume that there is a significant marginal effect of adding/removing (x_u, y_u) relative to the original dataset (otherwise there is no significant need for unlearning) and the retrained model f achieves a low training loss on the remaining dataset, then one cannot simultaneously have all of the following: (1) truthful revealing of (x_u, y_u): $f(x_u) = y_u$; (2) ε-DU, (3) good generalizability of f with low Lipschitz constant. That is, if we require (2), then a retrained model with good generalizability cannot reveal (x_u, y_u).

Proposition 2.1 (ε-DU Reduces Utility on the Data to Unlearn) *Assume $f(X_1) = Y_1$ and $\sup_y |\log(\frac{\mathcal{L}(f(X_0))(y)}{\mathcal{L}(f(X_1))(y)})| \leq \varepsilon$ such that $L^*(\varepsilon) > 1$, but Y_0 satisfies $\delta|Y_0(y) - Y_1(y)| > \mathcal{W}_{d_\mathcal{X}}(X_0, X_1)$ for some $y \in \mathcal{Y}$, then $f(X_0) \neq Y_0$.*

Proof This is a direct corollary of Lemma 2.2. ■

In other words, if a retrained model achieves good training performance on the remaining data set ($f(X_1) = Y_1$) and satisfies the proposed ε-differential unlearning for small enough ε ($L^*(\varepsilon) > 1$), but there is a significant marginal training signal ($\delta|Y_0(y) - Y_1(y)| > 2\mathcal{W}_{d_\mathcal{X}}(X_0, X_1)$) resulting from adding/removing the data to unlearn (to the existing data set), then $f(X_0)$ cannot reveal Y_0 truthfully. That further implies $f(x_u)$ cannot reveal y_u truthfully by the construction of (X_0, Y_0).

Finally, by enforcing ε-DU by mutual information regularization while preserving utility on the remaining data, the proposed framework achieves unlearning by only diminishing the marginal utility at the data point to be unlearned. But we further notice that the utility via our approach can decrease further than the retrained (from scratch) model in practice.

2.3 Versatile Utility

In our feature unlearning setting, often the target variable Y lives in a different space from X, and mutual information may not suffice to capture how $\hat{X}$ relates to Y. Thus, we consider

$$\sup_{\hat{X}=f(X,Z)} \{\mathcal{U}(Y;\hat{X}) : \hat{X} \perp Z\}, \tag{7}$$

where $\mathcal{U}(Y;\hat{X})$ is a user-defined utility. Below are some widely used objectives, each with a short motivation. For more detailed explanation of each utility below, see Appendix A.5 and Appendix A.6.

- **Entropy Maximization:** $\mathcal{U}(Y;\hat{X}) = H(\hat{X})$ preserves the total uncertainty in $\hat{X}$.
- **Mutual Information Maximization:** $\mathcal{U}(Y;\hat{X}) = I(Y;\hat{X})$ maximizes how much $\hat{X}$ reveals about Y.
- **KL-Divergence Maximization:** $\mathcal{U}(Y;\hat{X}) = D_{KL}\big(\mathbb{P}(Y \mid \hat{X}) \,\|\, \mathbb{P}(Y)\big)$ makes $\mathbb{P}(Y \mid \hat{X})$ more deterministic relative to $\mathbb{P}(Y)$.
- **Conditional Probability Energy Maximization:**

$$\mathcal{U}(Y;\hat{X}) = \begin{cases} \sqrt{\mathbb{E}[\,\mathbb{P}(Y \in A \mid \hat{X})^2\,]}, & \text{classification,} \\ -\|\,Y - \mathbb{E}(Y \mid \hat{X})\|_2, & \text{regression,} \end{cases} \tag{8}$$

  improves classification boundaries or reduces mean squared error.

We note that the proposed (feature and data point) unlearning framework supports a wide range of utility and objective functions beyond the four listed here. The four listed objectives are selected for their shared analytical optimal solution. As shown in Theorem 3.1, this optimal solution is of particular interest for its independence from the choice of Y, even though three of the objectives explicitly involve Y.

3. Theoretical Guarantees

In this section, we provide theoretical guarantees for both feature unlearning and data point unlearning. Specifically, we leverage optimal transport to derive a unified analytic solution for feature unlearning under all the listed objectives. For data point unlearning, we establish a provable guarantee by ensuring that the unlearning outcome satisfies ε-differential unlearning through the compression rate quantified by mutual information.

3.1 Feature Unlearning: Unified Optimal Solution for Multiple Objectives

The utility objectives outlined earlier are commonly used across fields such as biology, physics, and AI. See Appendix A.5 for more details. Despite their diverse forms, these objectives can be unified within a single framework by focusing on the sigma-algebra generated by the unlearning outcome $\hat{X}$. This approach allows for a cohesive solution to seemingly distinct optimization problems, demonstrating the versatility and practicality of the proposed framework.

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

Lemma 3.1 (Wasserstein-2 Barycenter Generates the Finest Sigma-Algebra) *Let* $\{X_z\}_{z\in\mathcal{Z}} \subset \mathcal{P}_{2,ac}(\mathcal{X})$. *We have* $\sigma(\hat{X}) \subset \sigma(\bar{X})$ *for all* $\hat{X} = f(X, Z)$.

The proof is in the Appendix B.2. The above result demonstrates one feasible optimal solution. Under the assumption of absolute continuity of marginals, the barycenters of convex costs also satisfy 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 contained by the random variable in probability theory. See the following remark for a more detailed explanation.

Remark 3.1 (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.

As a result, the information quantification discussed in Section 2.3 are naturally monotone w.r.t. the fineness of the generated sigma-algebra. In particular,

Lemma 3.2 (Monotonicity of Information Measures w.r.t. Sigma-Algebra) *If* $\sigma(X_1) \subset \sigma(X_2)$*, then:*

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

Informally, Lemma 3.2 shows entropy increases as the sigma-algebra becomes finer, while conditional entropy decreases, indicating reduced uncertainty. Similarly, mutual information and conditional probability energy increase, reflecting enhanced informativeness and predictive utility. Furthermore, as the sigma-algebra generated by X becomes finer, the conditional prediction of Y given X becomes more deterministic, resulting in a larger KL-divergence between the conditional distribution and the original distribution of Y.

Combining Lemma 3.1 with Lemma 3.2, we deduce that $\bar{X}$ is the optimal solution to all utility objectives discussed in Section 2 and further specified in Problems 1–5 in the Appendix.

Theorem 3.1 (Unified Optimal Feature Unlearning Solutions) *Assume* $\{X_z\}_{z\in\mathcal{Z}} \subset \mathcal{P}_{2,ac}(\mathcal{X})$*. Then the following statements are equivalent:*

- $\sigma(\hat{X}) = \sigma(\bar{X})$,
- $\hat{X} \in \arg\max_{\hat{X}=f(X,Z)}\{H(\hat{X}) : \hat{X} \perp Z\}$,
- $\hat{X} \in \arg\min_{\hat{X}=f(X,Z)}\{H(Y|\hat{X}) : \hat{X} \perp Z\}$ *for all* Y,
- $\hat{X} \in \arg\max_{\hat{X}=f(X,Z)}\{I(Y; \hat{X}) : \hat{X} \perp Z\}$ *for all* Y,
- $\hat{X} \in \arg\max_{\hat{X}=f(X,Z)}\{D_{KL}(\mathbb{P}(Y|\hat{X})||\mathbb{P}(Y)) : \hat{X} \perp Z\}$ *for all* Y,
- $\hat{X} \in \arg\max_{\hat{X}=f(X,Z)}\{\|\mathbb{P}(Y \in A|\hat{X})\|_2^2 : \hat{X} \perp Z\}$ *for all* A *and* Y.

Proof This follows from Lemma 3.1 and Lemma 3.2. ■

3.2 Data Point Unlearning: Fine-tuning Guided by Compression Rate

For data point unlearning, we adopt the following framework to fine-tune the trained model parameters under the guidance of mutual information regularization:

$$\sup_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{U}(Y;\hat{Y}) - \gamma I(\hat{Y}_{train};Z)$$

The following theorem is now a direct consequence of Lemma 2.1.

Theorem 3.2 (ε-Differential Unlearning Guarantee via Compression Rate) *Assume the unlearning outcome f_θ satisfies $I(f_\theta(X_{train});Z) \le \mu$, then f_θ satisfies ε-differential unlearning with probability at least $1 - \frac{\exp(\varepsilon)+1}{\exp(\varepsilon)-1}\sqrt{\frac{\mu}{2}}$.*

Therefore, one can achieve ε-differential unlearning based on the compression rate and choose the hyperparameter γ in the framework according to the required ε and the resulting trade-off between utility and mutual information.

4. Algorithm Design

Here, we summarize the pseudo-code implementation of the feature and data point unlearning framework we propose:

4.1 Feature Unlearning

4.1.1 Feature Unlearning on Data via Theorem 3.1

Algorithm 1: Feature Unlearning via Wasserstein Barycenter

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

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

4.1.2 Feature Unlearning on Modal via Regularization

Algorithm 2: Feature Unlearning on Model via Regularization

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

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

4.2 Data Point Unlearning on the Model Parameter Space

Algorithm 3: Data Point Unlearning via Regularization

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

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: $\hat{Y}_r = f_\theta(X_r)$ and $\hat{Y}_u = f_\theta(X_u)$.
    Construct $\hat{Y}_0 = \text{concat}(\hat{Y}_r, \hat{Y}_u, \text{dim} = 0)$ and $\hat{Y}_1 = \hat{Y}_r$.
    Define relational dataset $(\hat{Y}_{train}, Z)$ where: $\hat{Y}_{train}|_{Z=0} = \hat{Y}_0$ and $\hat{Y}_{train}|_{Z=1} = \hat{Y}_1$.
    Compute regularized loss: $\mathcal{L}_{\text{reg}} = -\mathcal{U}(\hat{Y}_r, Y_r) + \gamma I(\hat{Y}_{train}; Z)$.
    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 θ.

5. Numerical Experiments

We conduct numerical experiments on 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 data point unlearning, is left for future work. The code is available at `https://github.com/xushizhou/Machine_Unlearn_via_Info_Reg`.

5.1 Feature Unlearning

Here, we demonstrate the unified solution using the Wasserstein barycenter on image data (Celeba [27]) with the following features to unlearn: *gender* and *smile*.

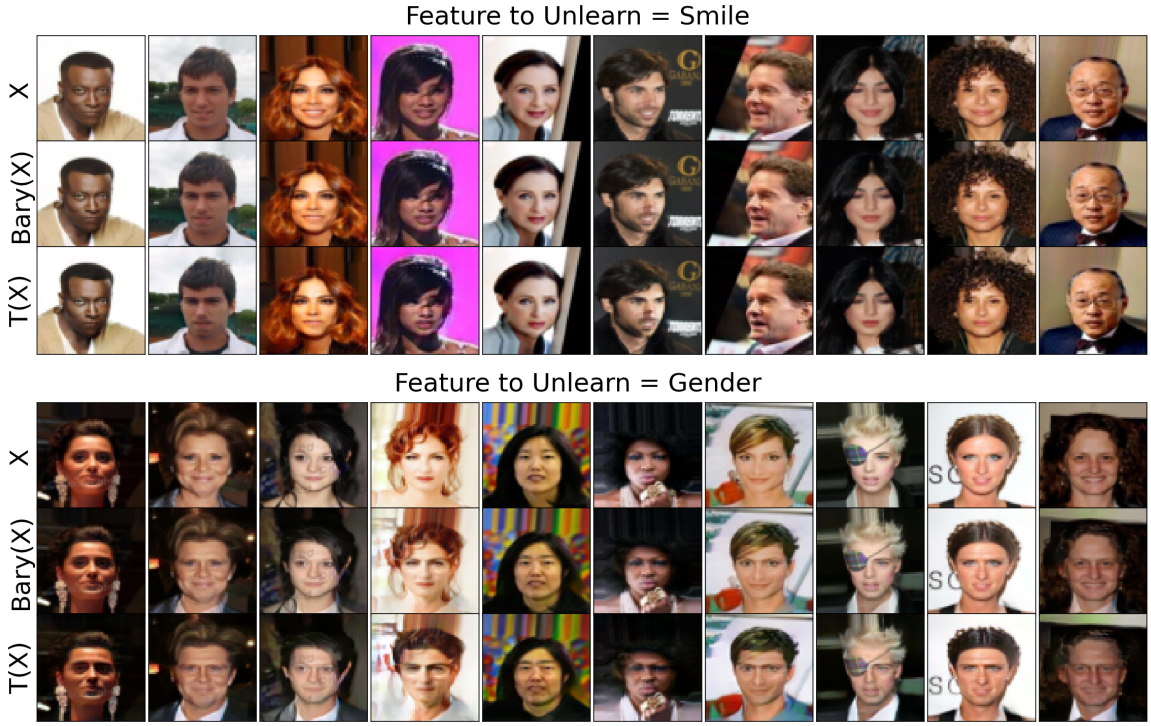

Figure 1: The above plot shows the Wasserstein Barycenter (Bary(X)) characterization of the optimal feature unlearning result. In particular, we apply neural optimal transport [24] to learn the optimal transport map from the smile (female) faces to non-smile (male) faces if smile (gender) is the feature to unlearn, then generate the barycenter using McCann interpolation [28] with $t = 0.5$.

We note that the Wasserstein-2 barycenter provides one solution to the optimal feature unlearning under the utility objectives outlined in Section 2.3. But alternative solutions can be obtained by selecting a more appropriate metric space, instead of the $\ell^2(\mathbb{R}^{d\times d})$ space used here, to better align the Wasserstein geodesic path with the natural data manifold (e.g., the face manifold).

For tabular data, the solution and corresponding experiments closely align with those in machine learning fairness. We refer interested readers to fairness studies such as [8, 43] for detailed numerical evaluations of the Wasserstein barycenter approach on datasets including

Method	Epochs	Avg. (Std.) Acc. Remain	Acc. Unlearn
Original Model	100	0.9334 (0.0017)	0.9676 (0.0300)
Retrain from Scratch	100	0.9334 (0.0018)	0.7848 (0.1901)
Retrain on Original Model	100	0.9452 (0.0000)	0.9999 (0.0000)
Mutual Info Reg. ($\gamma = 4.2$)	**5**	0.9244 (0.0000)	0.7805 (0.0000)

Table 1: **Gaussian Mixture Model (GMM) Unlearning.** The proposed method effectively removes the marginal effect of the unlearned data while maintaining high accuracy on the remaining dataset. In contrast, retraining on the original model fails to remove the effect even after extensive training.

Method	Epochs	Avg. (Std.) Acc. Remain	Acc. Unlearn	Acc. Test
Original Model	100	0.9989 (0.0001)	0.9998 (0.0004)	0.9781 (0.0011)
Retrain from Scratch	100	0.9986 (0.0007)	0.9389 (0.0112)	0.9764 (0.0016)
Retrain on Original Model	100	0.9996 (0.0003)	0.9702 (0.0131)	0.9797 (0.0016)
Mutual Info Reg. ($\gamma = 3.5$)	**5**	0.9976 (0.0013)	0.9418 (0.0309)	0.9736 (0.0045)

Table 2: **MNIST Data Point Unlearning.** The proposed method effectively removes the marginal effect of the unlearned dataset while maintaining strong generalization (high test accuracy). In contrast, retraining from the original model treats the dataset to unlearn as test data after extensive training.

UCI Adult (unlearn gender), COMPAS (race), LSAC (race), and CRIME (race), along with comparisons to other fairness-driven feature unlearning methods.

5.2 Data Point Unlearning

To assess the effectiveness and efficiency of the proposed framework for data point unlearning, we conduct experiments on both synthetic and real-world datasets. First, we generate synthetic data of size 5000 from a Gaussian mixture model (GMM) with two distinct means and train a neural network for classification. We then select a data point for unlearning and compare the predictions of four models: (1) the original classifier, (2) the classifier retrained from scratch, (3) the classifier retained from the original one, (4) the unlearned model via our mutual information regularization. See Table 1 for the results averaged over 100 repetitions.

Next, we extend the evaluation to the MNIST [25] dataset. To ensure a significant difference between the predictions of the original model and those of the retrained-from-scratch model, we construct the unlearning dataset such that 75% of the samples belong to the digit 3 class. See Table 2 for the results averaged over 10 repetitions.

As shown in Figure 2, to match the performance of retraining from scratch, careful tuning of the relaxation parameter γ and proper monitoring of mutual information are essential.

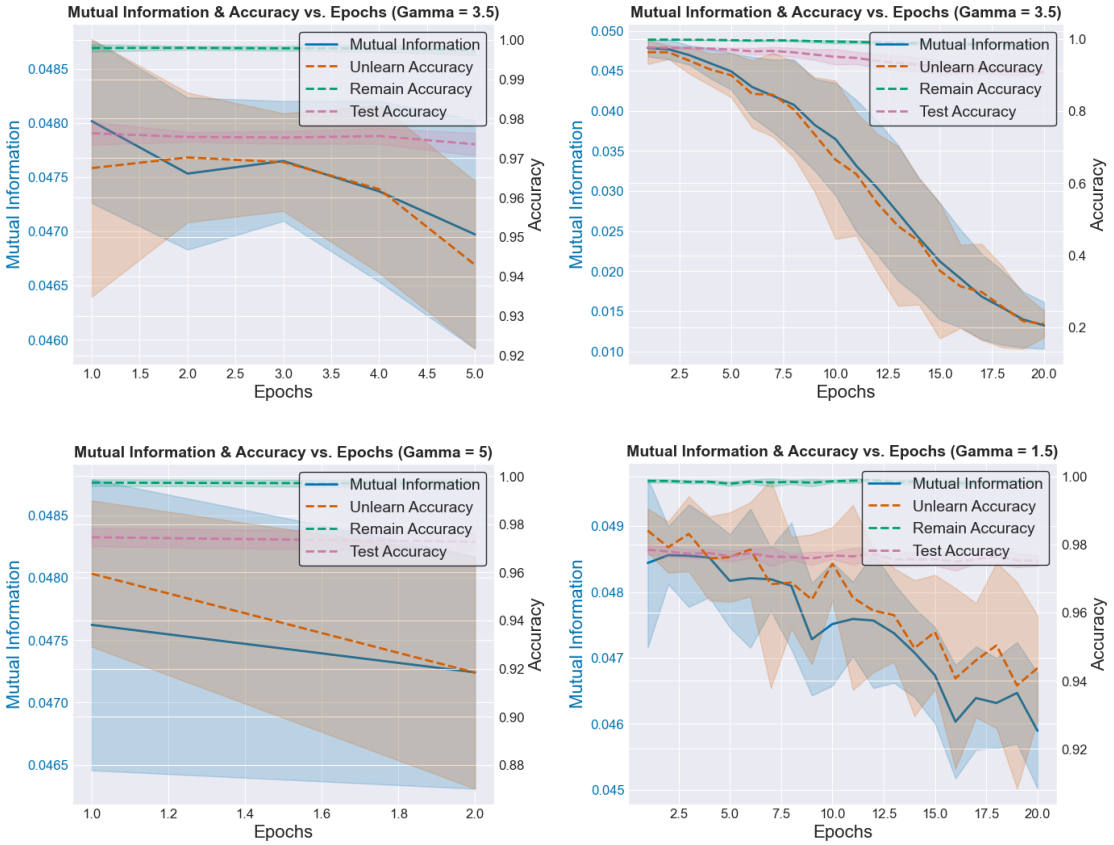

Figure 2: **Mutual Information vs. Utility in MNIST Unlearning.** Mutual information and model utility on the MNIST unlearning dataset exhibit a strong correlation. The rate of utility reduction can be precisely controlled by adjusting the regularization parameter γ, allowing for flexible unlearning speed while maintaining accuracy.

Acknowledgement

The authors are supported by NSF DMS-2208356, NIH R01HL16351, P41EB032840, and DE-SC0023490. T.S. wants to thank Kate Kaye for inspiring conversations on topics related to this paper.

References

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

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

[3] 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. IEEE, 2021.

[4] J. Brophy and D. Lowd. Machine unlearning for random forests. In *International Conference on Machine Learning*, pages 1092–1104. PMLR, 2021.

[5] C. Chen, F. Sun, M. Zhang, and B. Ding. Recommendation unlearning. In *Proceedings of the ACM Web Conference 2022*, pages 2768–2777, 2022.

[6] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pages 499–513, 2022.

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

[8] 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.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 2005.

[10] 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.

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

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

[13] 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.

[14] 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`.

[15] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou. Making AI forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

[16] 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.

[17] A. Golatkar, A. Achille, and S. Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 383–398. Springer, 2020.

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

[19] L. Graves, V. Nagisetty, and V. Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524, 2021.

[20] G. T. Guo, Cheng and N. Pustelnik. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.07252*, 2020.

[21] 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.

[22] 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.

[23] 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.

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

[25] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[26] X. Li, L. Feng, J. Liu, and D. Hu. Machine unlearning via representation forgetting with parameter self-sharing. *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[27] 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. URL `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`.

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

[29] R. Mehta, S. Pal, V. Singh, and S. N. Ravi. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10422–10431, 2022.

[30] T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

[31] 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.

[32] S. Schelter, S. Grafberger, and T. Dunning. Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1545–1557, 2021.

[33] 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.

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

[35] V. Suriyakumar and A. C. Wilson. Algorithms that approximate data removal: New results and limitations. *Advances in Neural Information Processing Systems*, 35:18892–18903, 2022.

[36] R. Tanno, M. F Pradier, A. Nori, and Y. Li. Repairing neural networks by leaving the right past behind. *Advances in Neural Information Processing Systems*, 35:13132–13145, 2022.

[37] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL `https://books.google.com/books?id=hV8o5R7_5tkC`.

[38] W. Wang, Z. Tian, C. Zhang, and S. Yu. Machine unlearning: A comprehensive survey. *arXiv preprint arXiv:2405.07406*, 2024.

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

[40] S. Weisberg and R. D. Cook. Residuals and influence in regression. 1982.

[41] 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.

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

[43] 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.

[44] H. Yan, X. Li, Z. Guo, H. Li, F. Li, and X. Lin. Arcane: An efficient architecture for exact machine unlearning. In *IJCAI*, volume 6, page 19, 2022.

A. Appendix: Supplementary Material for Section 2

A.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,

$$H(X|\hat{Y}) := -\sum_x p(x) \sum_{\hat{Y}} p(\hat{Y}|x) \log(p(\hat{Y}|x)) \tag{9}$$

is the conditional entropy of X on $\hat{Y}$. Intuitively, a higher conditional entropy means more volume of x are expected to be mapped to individual $\hat{Y}$, which implies more randomness of X remained 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:

$$\frac{2^{H(X)}}{2^{H(X|\hat{Y})}} = 2^{I(X;\hat{Y})}. \tag{10}$$

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.

A.2 Admissibility

Since we are unlearning the information of Z by compressing relational data (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):

$$\begin{aligned} \{\omega : \hat{X}(\omega) \in A\} &= \hat{X}^{-1}(A) \\ &= (X, Z)^{-1}(f^{-1}(A)) \\ &= \{\omega : (X(\omega), Z(\omega)) \in f^{-1}(A)\}. \end{aligned} \tag{11}$$

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

$$\sigma(\hat{X}) \subset \sigma((X, Z)). \tag{12}$$

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:

$$\begin{aligned} \mathcal{A}(X, Z) \ := \ \Big\{ \hat{X} = f(X, Z) : f \text{ is } \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Z}}/\mathcal{B}_{\mathcal{X}} \\ \text{-measurable} \Big\}. \end{aligned} \tag{13}$$

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

A.3 Proof of Lemma 2.1

Proof First, notice that it follows from the construction of relational data (X_{train}, Z) that $\{Z = 0\} = \{X_{train} = X_0\}$ and $\{Z = 1\} = \{X_{train} = X_1\}$. Also, we have

$$\begin{aligned} |\mathbb{P}(Z = 0|\hat{Y}) - \mathbb{P}(Z = 1|\hat{Y})| &= |2\mathbb{P}(Z = 0|\hat{Y}) - 1| && (14) \\ &\leq 2|\mathbb{P}(Z = 0|\hat{Y}) - \mathbb{P}(Z = 0)| + 2|\mathbb{P}(Z = 0) - 0.5| && (15) \\ &\leq ||\mathbb{P}(Z|\hat{Y}) - \mathbb{P}(Z)||_{TV}, && (16) \end{aligned}$$

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

$$\begin{aligned}\mathbb{P}\left(\{|\log\left(\frac{\mathbb{P}(Z=0\mid\hat{Y})}{\mathbb{P}(Z=1\mid\hat{Y})}\right)|\leq\log(\frac{1+\varepsilon}{1-\varepsilon})\}\right)&\geq\mathbb{P}\left(\{|\mathbb{P}(Z=0|\hat{Y})-\mathbb{P}(Z=1|\hat{Y})|\leq\varepsilon\}\right)\\&\geq1-\frac{1}{\varepsilon}\left(\sqrt{\frac{1}{2}I(Z;\hat{Y})}\right).\end{aligned}$$

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

A.4 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 $\text{sign}(\log(\frac{f(X_0)(y)}{f(X_1)(y)}))=\text{sign}(\log(\frac{f(X_0)(y^*)}{f(X_1)(y^*)}))$, we have

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

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

$$d_\mathcal{Y}(f(x),f(x'))\leq Ld_\mathcal{X}(x,x')\leq L^*(\varepsilon)d_\mathcal{X}(x,x'). \tag{17}$$

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 L-Lipschitz for any $L\leq L^*(\varepsilon)$. ■

A.5 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:

$$\sup_{\hat{X}=f(X,Z)} \{\mathcal{U}(Y;\hat{X}) : \hat{X} \perp Z\}, \tag{18}$$

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 [31]. 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 [10] and in deep learning techniques, including Deep InfoMax [23] and information bottleneck methods [2]. 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) [22]. 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 (18) for arbitrary target variable Y.

A.6 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 Problem 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 $\mathbb{P}(Y)$. Thus, maximizing the KL-divergence enhances the predictive power of $\hat{Y}$ for Y. We later show that Problem 4 is also independent of the choice of 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 $\hat{Y}$.

B. Appendix: Supplementary Material for Section 3

B.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 $\mathbb{R}^d$,

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

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 $\mathcal{P}_2(\mathbb{R}^d) := \Big\{\mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} ||x||^2 d\mu < \infty\Big\}$. 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 [37], 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 [1] with weights $\lambda \in \mathcal{P}(\mathcal{Z})$ is

$$\bar{\mu} := \text{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \Big\{ \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mu) d\lambda(z) \Big\}. \tag{19}$$

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.

B.2 Proof of Lemma 3.1

Proof See Lemma 5.2 in [43] 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. ■

B.3 Proof of Lemma 3.2

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 [9] 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 $\sigma(X_1) \subset \sigma(X_2)$,

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

■

B.4 Intuitive Insights into Theorem 3.1

- 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.