Machine Unlearning through Information Theory and Optimal Transport

Fairness as Feature Unlearning and Auditable Data Unlearning

Shizhou Xu

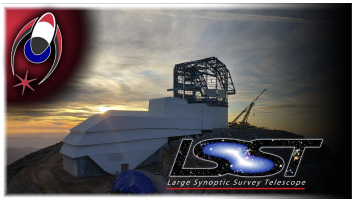
Stanford University & SLAC National Accelerator Laboratory

AIM Fairness and Foundations in Machine Learning Workshop @ Caltech, July 16, 2026

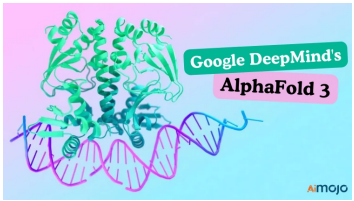
Motivation: AI for scientific discovery

AI is increasingly used throughout modern scientific workflows.

FourCastNet



LSST



AlphaFold 3

Experimental data may contain corrupted measurements, systematic biases, or mislabeled samples. Once incorporated into a learning system, these errors can propagate through downstream scientific inference.

Central question

How can we remove undesirable information while preserving useful knowledge?

A common mathematical question

Feature and data unlearning appear different, but they share the same underlying challenge:

Shared problem

How can we remove undesirable information while preserving task-relevant structure and utility?

This talk develops two instances of this principle:

I **Fairness as feature unlearning**

II **Data unlearning via marginal information**

Fairness as Feature Unlearning

[I] Fairness as feature unlearning

- **Difficulty**: In many learning systems, sensitive attributes are not used explicitly, but their *influence* remains through correlated features.
- **Goal**: remove the influence of a sensitive variable while preserving predictive utility.

Question to address:

Can we characterize the *optimal* way to enforce fairness constraint, and the best possible utility-fairness trade-off?

Problem 1: Optimization Problems with Sensitive Variable Independence Constraint

Let (X, Y, Z) be respectively independent, dependent, and sensitive variables. Let $\hat{Y} = f(X, Z)$ be the released predictor. We focus on **statistical parity** [4]:

$$\hat{Y} \perp Z.$$

Problem (Optimal fair L^2-objective learning outcome)

$$\underbrace{\inf_{f \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})} \{||Y - f(X, Z)||_2^2 : f(X, Z) \perp Z\}}_{V^2 := \textit{(squared) minimum loss for statistical parity}} \qquad (*)$$

- Utility = L^2 loss,
- Fairness = prediction is independent from the sensitive variable.

Three Open Challenges

Listed in NeurIPS and numerous surveys [3]:

- **Open challenge 1:** characterize the *optimal fair predictor*.
  Wasserstein Barycenter on output space $\mathcal{Y} = \mathbb{R}^d$.
- **Open challenge 2:** characterize the *Pareto frontier* between utility loss and statistical disparity.
  Wasserstein Geodesics on output space $\mathcal{Y} = \mathbb{R}$.
- **Open challenge 3:** characterize the *fair data representations* family that automatically results in the Pareto frontier under regular algorithms.
  Wasserstein Barycenter on input space $\mathcal{X} = \mathbb{R}^d$.

Result 1: optimal fair learning outcome

Let $\mu_z := \mathcal{L}(\mathbb{E}[Y \mid X, Z]_z)$ denote the conditional expectation distribution on group z.

Theorem: Optimal fair L^2-objective supervised learning characterization [8]

Assume that the conditional expectation marginals $\{\mu_z\}_{z\in\mathcal{Z}} \subset \mathcal{P}_{2,ac}(\mathcal{Y})$, then

$$\overline{\mathbb{E}(Y|X,Z)} := T(\mathbb{E}(Y|X,Z),Z) := \{T(\mathbb{E}(Y|X,Z)_z, z)\}_{z\in\mathcal{Z}}$$

is the unique solution to Problem 1. Furthermore, we have $(*) =$

$$||Y - \overline{\mathbb{E}(Y|X,Z)}||_2^2 = \underbrace{||Y - \mathbb{E}(Y|X,Z)||_2^2}_{\text{orthogonal projection loss}} + \underbrace{||\mathbb{E}(Y|X,Z) - \overline{\mathbb{E}(Y|X,Z)}||_2^2}_{\text{independence projection loss}}.$$

Takeaway: the optimal fair predictor is a Wasserstein barycenter projection.

Problem 2: Pareto frontier via Wasserstein geodesics

To quantify residual disparity, define

$$D(\hat{Y}, Z) := \Big(\int \mathcal{W}_2^2(\mathcal{L}(\hat{Y}_{z_1}), \mathcal{L}(\hat{Y}_{z_2}))\, d\lambda(z_1) d\lambda(z_2) \Big)^{1/2}.$$

Then the relaxed problem

Problem (Optimal L^2-objective learning Pareto frontier)

$$\underbrace{\inf_{f \in L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})} \{||Y - f(X,Z)||_2^2 : D(f(X,Z),Z) \leq d\}}_{V_d^2 := \text{(squared) minimum loss at Wasserstein disparity tolerance level } d}$$

characterizes the utility–fairness Pareto frontier.

Result 2: Pareto frontier via Wasserstein geodesics

Theorem (Pareto optimal fair L^2-objective learning (with $\mathcal{Y} = \mathbb{R}$) [8])

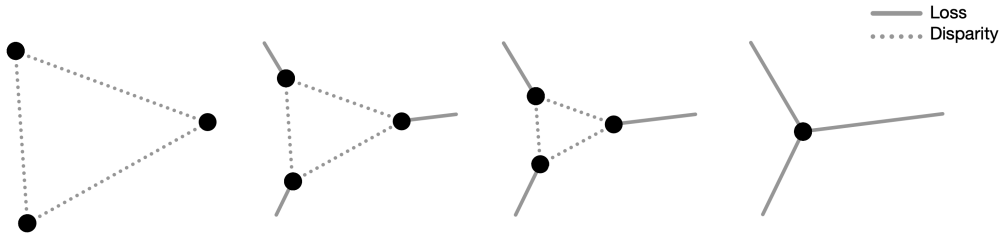
Given (X, Y, Z) satisfying $\mu_z \in \mathcal{P}_{ac}$, λ-a.e., then

$$f_d(X,Z) := \begin{cases} T(1 - \frac{d}{\sqrt{2}V})(\mathbb{E}(Y|X,Z), Z), & \text{if } d \in [0, \sqrt{2}V) \\ \mathbb{E}(Y|X,Z), & \text{if } d \in [\sqrt{2}V, \infty) \end{cases}$$

is the unique solution to Problem 2 for $d \in [0, \infty)$.

- $T(\cdot, z)$ is the optimal transport map from μ_z to $\bar{\mu}$,
- $T(t)(\cdot, z) := (1-t)Id + tT(\cdot, z)$ denotes the McCann interpolation,
- $V := (\inf_{f \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})} \{||Y - f(X,Z)||_2^2 : f(X,Z) \perp Z\})^{\frac{1}{2}}$ is the minimum work required to achieve statistical parity in $(*)$.

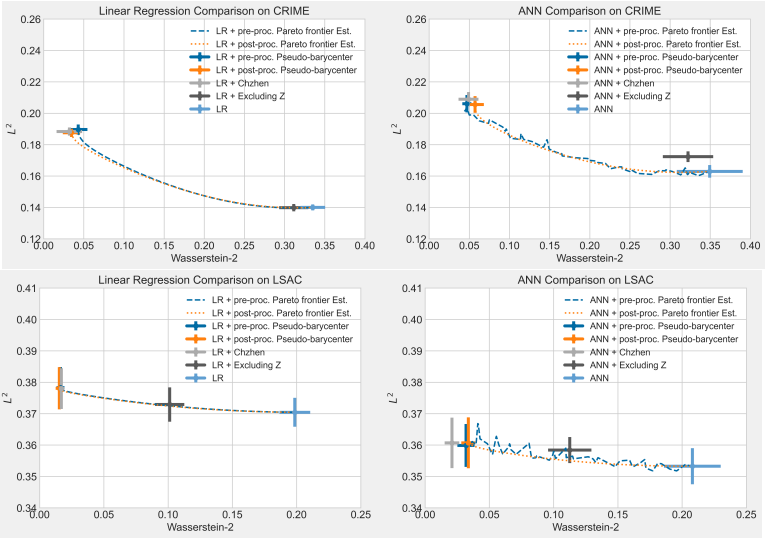
Result 2: Intuition with Euclidean Analog of Linear Optimal Trade-off via Geodesics



$$(\frac{1}{N}\sum_{i=1}^{N}||x_i - m_x||^2)^{\frac{1}{2}} \leq \underbrace{\text{Solid(t)}}_{\text{utility loss}} + \frac{1}{\sqrt{2}}\underbrace{\text{Dotted(t)}}_{\text{disparity}}, \forall t \in [0,1]$$

- $x_i(t) = (1-t)x_i + tm_x$,
- Dotted(t) $:= (\frac{1}{N^2}\sum_{i,j=1}^{N}||x_i(t) - x_j(t)||^2)^{\frac{1}{2}}$,
- Solid(t) $:= (\frac{1}{N}\sum_{i=1}^{N}||x_i - x_i(t)||^2)^{\frac{1}{2}}$.

Numerical Experiments: 1-dimensional Regression



Problem 3: General Cost Feature Unlearning

Problem (Feature Unlearning)

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

where C quantifies the integrated utility loss of using S to predict Y on each Z-slices. Viewing unlearning as compressing (S,Z) giving the following rate-distortion relaxation:

$$\inf_{f:\mathcal{X}\times\mathcal{Z}\to\mathcal{S}} C(Y;S,Z) + \gamma\, \mathrm{I}(S;Z), \qquad \gamma \geq 0.$$

Cost function choice

We focus on cost functions such as:

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

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

Result 3: Feature unlearning and Wasserstein barycenter

Notation: let $\sigma(X, Z)$ denote the sigma-algebra generated by (X, Z), and define $X_z := \{x_i : (x_i, z_i) \in (X, Z),\ z_i = z\}$.

Theorem (Optimal Feature Unlearning Solution [9])

Let $\{\mathcal{L}(X_z)\}_{z\in\mathcal{Z}} \subset \mathcal{P}_{2,ac}(\mathcal{X})$, and define the barycentric representation $\bar{X} := T_Z(X)$ with $T_z : P_z \to \bar{P}$ being the Brenier map. Then the following are equivalent for any admissible $\hat{X} = f(X, Z)$:

- $\sigma(\hat{X}, Z) = \sigma(\bar{X}, Z)$,
- $\hat{X} \in \arg\min\{H(Y \mid \hat{X}, Z) :\ \hat{X} \perp Z\}$ for all $Y : \Omega \to \mathcal{Y}$,
- $\hat{X} \in \arg\max\{I(Y; \hat{X}, Z) :\ \hat{X} \perp Z\}$ for all $Y : \Omega \to \mathcal{Y}$,
- $\hat{X} \in \arg\max\{\|\mathbb{P}(Y \in A \mid \hat{X}, Z)\|_{L^2}^2 :\ \hat{X} \perp Z\}$ for all $A \in \mathcal{B}_\mathcal{Y}$ and $Y : \Omega \to \mathcal{Y}$.
- $\hat{X} \in \arg\max\{\|\mathbb{E}(Y \mid \hat{X}, Z)\|_{L^2}^2 :\ \hat{X} \perp Z\}$ for all $Y \in L^2(\Omega; \mathcal{Y})$.

Proof idea: the W_2 barycenter generates the finest sigma-algebra: $\sigma(\hat{X}, Z) \subset \sigma(\overline{X}, Z)$ for all admissible $\hat{X} = f(X, Z)$.

From theory to practice

We use the CelebFaces Attributes dataset (CelebA), 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*.

Let $Z \in \{0, 1\}$ denote the attribute to unlearn (e.g., $Z = 1$ for smiling, $Z = 0$ for non-smiling). Let μ_0 and μ_1 be 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

1. **minimally distorted** from the originals, and
2. **uninformative about** Z.

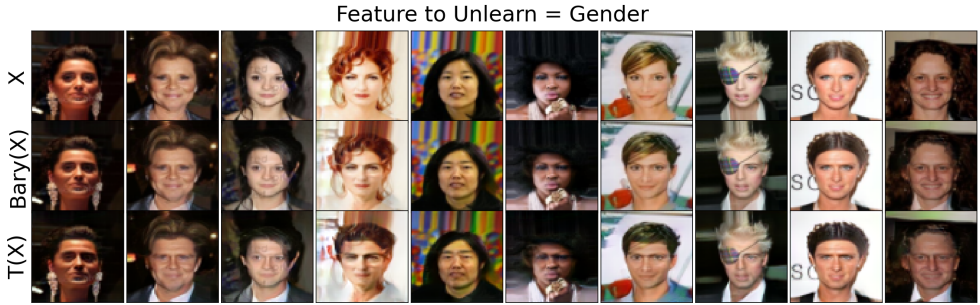
Who wants to smile in times like these?

We (applied neural OT [6]) but now use our own deep learning optimal transport solver to approximate the optimal transport map from smile faces to non-smile faces, where “smile” is the feature to unlearn.

Then we generate the Wasserstein barycenter $\overline{X}$ using McCann interpolation with $t = 0.5$ $(X : t = 0,\ T(X) : t = 1)$.

Unlearning gender via Wasserstein barycenter

Same setup as before, but now unlearning the feature “gender”.



Visualization for how the optimal fair data representation works for different objective functions. It even works for our eyes.

Beyond demographic parity: separation

Everything so far concerns **demographic parity**,

$$\hat{Y} \perp Z.$$

Many applications instead require **separation** (equalized odds),

$$\hat{Y} \perp Z \mid Y.$$

Why not use conditional Wasserstein barycenters?

Unlike demographic parity, the conditional variable Y is unavailable at prediction time. Consequently, the optimal transport construction no longer produces a practical prediction rule.

Instead, we relax the conditional independence constraint using information theory:

$$\min_{\theta} -I(Y;\hat{Y}) + \lambda I(\hat{Y}; Z \mid Y).$$

Open questions

Although the geometric picture is now much clearer, several fundamental questions remain.

- **Beyond the L_2–W_2 frontier for independence:** Can explicit optimal frontiers be derived for more general utility functions and transport geometries?
- **Beyond the MI-CMI for separation:** Can other dependence measures produce tighter or more computationally efficient characterizations of separation?
- **Toward sufficiency:** What is the optimal solution or Pareto frontier for sufficiency: $Y \perp Z \mid \hat{Y}$?

Data Unlearning via Marginal Information

[II] Machine Unlearning via Marginal Information

Goal: Data unlearning 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.

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.

Anchor model = retrain-from-scratch model.

If the retrain-from-scratch model is known, then one would not need machine unlearning in the first place.

While intuitive, the "retrain on the retain set" heuristic depends on an counterfactual anchor model that is not auditable.

Human analogy

Consider helping a person forget a text they once read.

Existing anchor-based unlearning demands an unverifiable counterfactual: the subject should behave as if they had never encountered the text.

Why is machine unlearning difficult?

Many existing approaches seek to make the learned model less sensitive to changes in the training data, often using ideas related to differential privacy.

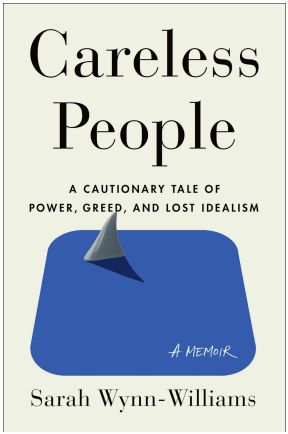
In real-world applications, however, they face several limitations:

- **Utility loss:** stronger forgetting often leads to substantial performance degradation.
- **Computational cost:** retraining or repeated optimization can be prohibitively expensive.
- **Limited compatibility:** many methods depend on specific learning objectives, model architectures, or training procedures.
- **Guarantee–practicality trade-off:** methods with strong theoretical guarantees often require restrictive assumptions, such as access to a retrain-from-scratch model, convexity, or Lipschitz continuity of the Hessian.

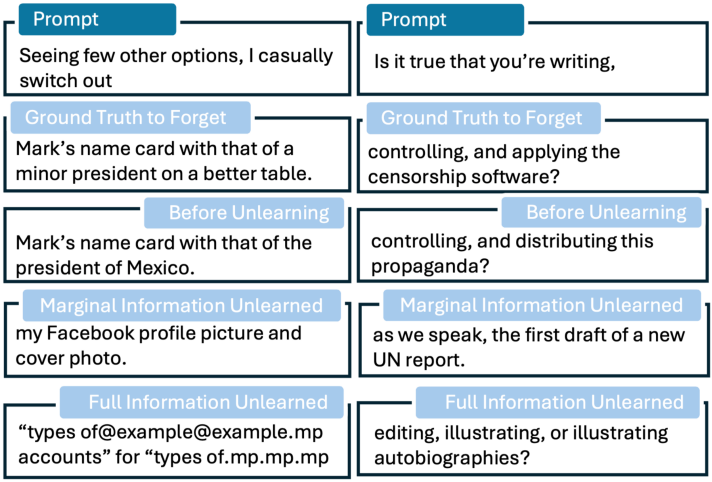
Machine unlearning: Application to Large Language Models (LLM)

Meta's Llama-3.2-1B model with 1.23 billion parameters.

Data to unlearn:



Machine unlearning: Application to Large Language Models (LLM)



Marginal Unlearning

Consider a copyright unlearning scenario where we have an LLM pretrained on an article from *The Washington Post* and on one from *The New York Times*, but only the former is legally authorized for use by the LLM.

Both outlets report on an identical event, yet their articles differ in narrative style, phrasing, and editorial perspective.

Marginal Unlearning

Usual data point unlearning: Erase all content associated with the NYT article, including factual information independently supported by the retained WP article.

Marginal Information Unlearning: Remove only the stylistic elements and content unique to the NYT article, while retaining shared factual content that also appears in the WP article.

Marginal Unlearning

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 refers to the unique information in the model output contributed by the unlearn set in addition to the existing information represented by the retain set.

Intuitively, marginal unlearning requires that the model output be **invariant to whether training examples were drawn purely from the retain set or from a mixture that also includes the unlearn source.**

This invariance is auditable from the output or model bahvior.

Anchor-based vs. marginal unlearning

Marginal unlearning follows a two-phase **blur + reinforce** mechanism:

- **Blur (auditable invariance):** suppress the diagnostic signal of the to-be-forgotten content so responses from the “retain” vs. “retain+unlearn” pools become statistically indistinguishable.
- **Reinforce (utility preservation):** strengthen desired knowledge so useful behavior is maintained.

This corresponds to the optimization problem

$$\inf_{f:\mathcal{X}\to\mathcal{S}} \underbrace{C(Y;S)}_{\text{utility}} + \lambda \underbrace{\mathrm{I}(S';Z)}_{\text{unlearning}} .$$

Marginal Unlearning

We quantify marginal information by measuring the distinguishability between the retain distribution p^r and the mixture distribution

$$p^m := (1-\alpha)p^r + \alpha p^u,$$

where p^u is the distribution of the unlearn set and $\alpha \in (0, 1]$ controls the mixing ratio.

The marginal effect is substantial only when p^m deviates significantly from p^r, indicating that the unlearn source contributes a significant amount of unique information not already represented in the retain set.

Marginal Information – MarI

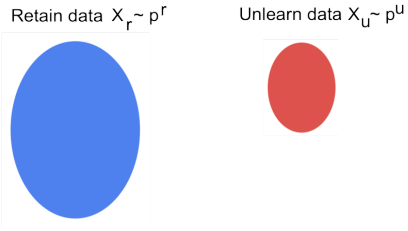
We formalize this distinction as a binary inference problem by introducing a latent variable $Z \sim \mathrm{Bernoulli}(\frac{1}{2})$ that governs the sampling source.

Conditional on Z, define X_{MarI} by

$$\mathrm{Law}(X_{\mathrm{MarI}} \mid \{Z = 1\}) = p^r, \qquad \mathrm{Law}(X_{\mathrm{MarI}} \mid \{Z = 0\}) = p^m.$$

Marginal unlearning

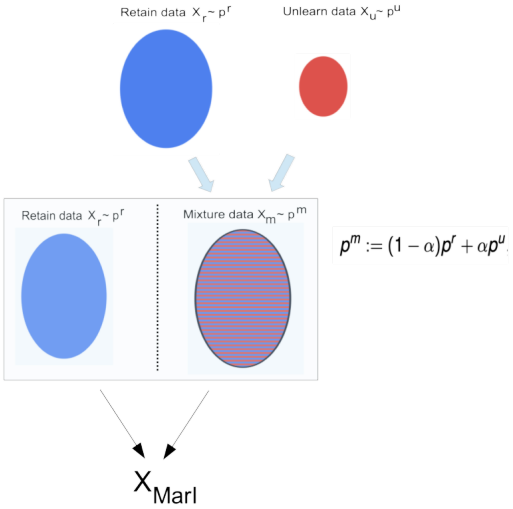
Suppose we are given $X_r \sim p^r, X_u \sim p^u$, where X_r denotes the retain set and X_u the data to be forgotten.



Key idea

The unlearn set may contain information already represented in the retain set, together with information that is unique to the unlearn data. Our goal is to remove only this *marginal information*.

Marginal Information



$$\mathrm{Law}(X_{\mathrm{MarI}} \mid \{Z=1\}) = p^r, \qquad \mathrm{Law}(X_{\mathrm{MarI}} \mid \{Z=0\}) = p^m.$$

ε-Marginal Unlearning

Let $\hat{Y}_{\mathrm{MarI}} := f(X_{\mathrm{MarI}})$ denote the model output. The pair $(\hat{Y}_{\mathrm{MarI}}, Z)$ encodes the model's behavior under the baseline $(Z = 1)$ versus the shifted $(Z = 0)$ distribution.

A natural quantification of marginal information is the ability to infer the inclusion label Z from the observed output $\hat{Y}_{\mathrm{MarI}}$.

Definition (ϵ-marginal unlearning)

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

$$\sup_{D \in \mathcal{B}_{\hat{y}}} \left| \log \left(\frac{\mathbb{P}\left(Z = 0 \,\middle|\, \hat{Y}_{\mathrm{margin}} \in D \right)}{\mathbb{P}\left(Z = 1 \,\middle|\, \hat{Y}_{\mathrm{margin}} \in D \right)} \right) \right| \leq \epsilon, \tag{1}$$

where the supremum is taken over measurable events D for which the posterior probabilities are well-defined.

Problem 5: Marginal Information Unlearning

Problem (Forgetting-MarI)

Let C be a loss function, (X_{MarI}, Z) be defined as above, $\hat{Y}_{\mathrm{MarI}} := f(X_{\mathrm{MarI}})$ for measurable $f : \mathcal{X} \to \mathcal{Y}$, and Y a target variable. Optimal marginal unlearning solves

$$\inf_f \Big\{ C(Y; \hat{Y}) : \ \hat{Y}_{\mathrm{MarI}} \perp Z \Big\},$$

with relaxed form

$$\inf_f C(Y; \hat{Y}) + \gamma \, \mathrm{I}(\hat{Y}_{\mathrm{MarI}}; Z).$$

We call this unlearning framework **Forgetting-MarI**.

Result 5: ε-marginal unlearning guarantee

The following guarantee translates a small compression rate into an auditable inference bound for ε-marginal unlearning.

Theorem (High-probability marginal unlearning via compression rate[9])

Let $Z \sim \mathrm{Bernoulli}(1/2)$. *If* $I(\hat{X}_{\mathrm{margin}}; Z) \leq \mu < \frac{1}{2}(\frac{e^\epsilon - 1}{e^\epsilon + 1})^2$, *then for every* $\epsilon > 0$, f *satisfies high-probability* ϵ*-marginal unlearning with probability at least* $1 - \delta_\epsilon$*:*

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

Result 6: Theoretical guarantees for LLM

Let

$$S_\theta(x, y) = \frac{1}{T} \sum_{t=1}^{T} \big(-\log p_\theta(x_t \mid y_{<t})\big)$$

be the standard cross-entropy (per-token negative log-likelihood).

State-of-the-art detectors flag membership of x in training by testing whether $S_\theta(x, x)$ is suspiciously low.

Theorem

Fix $u = (u_1, \ldots, u_T) \in V^T$ and assume the pathwise non-vanishing condition

$$\min\{p_t^u(u_t), p_t^r(u_t)\} \geq \gamma \in (0, 1] \qquad \forall t.$$

Then

$$\big|S_\theta(u, u) - S_\theta(u, r)\big| \leq \frac{2\sqrt{2}}{\gamma(1-\alpha)} \sqrt{\mathrm{I}(X_{\mathrm{MarI}}; Z)}.$$

Sufficiency

Theorem (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 $I(\hat{Y}_{\mathrm{margin}}; Z) \le \varepsilon_u$, where $\hat{Y}_{\mathrm{margin}} := f_{\Theta_u}(X_{\mathrm{margin}})$ and $\Theta_u \perp (X_{\mathrm{margin}}, Z)$. Then,

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

Necessity

Lemma (Marginal unlearning as a condition for "good" retraining)

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

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

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

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

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

Necessity

Thus, the lemma yields the following incompatibility among:

1. **Inclusion leakage**: the retrained model truthfully reveals the unlearn record, $f(x_u) = y_u$, at a positive-density point of the unlearn-source distribution;
2. **Marginal unlearning**: the output density ratio between $f_{\#}p^d$ and $f_{\#}p^r$ satisfies the ϵ log-ratio bound of Definition 5;
3. **Good regularity/generalizability**: the retrained model has small Lipschitz constant, so pointwise leakage propagates only through controlled local neighborhoods rather than arbitrary discontinuous spikes.

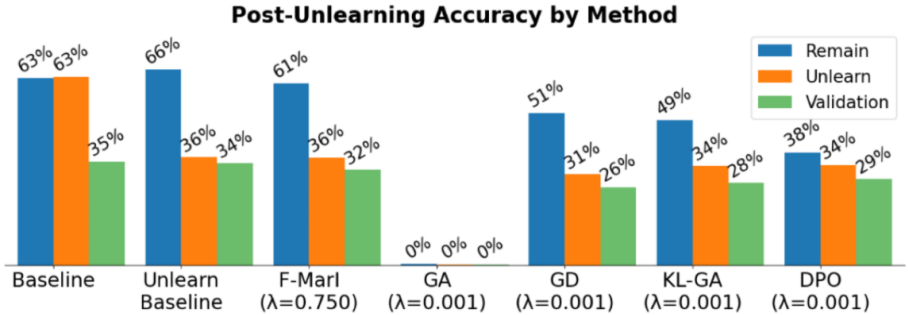
Indeed, under (1) and (3), inclusion leakage propagates to a local output-level signal near y_u. If this signal is specific to the unlearn source, then it produces a density-ratio gap and violates (2). Equivalently, if (1) and (2) both hold, then the model cannot remain regular in the sense quantified by Lemma 9.

Numerical experiments for Forgetting-MarI

Our experiments are designed to address two questions:

1. **Utility-Forgetting Trade-off:** Does Forgetting-MarI balance performance preservation on D_r while removing D_u?
2. **Detectability & Capacity:** Consistent with our theory, does the model defeat perplexity-based detectors while retaining general capabilities?

Machine unlearning for Large Language Models

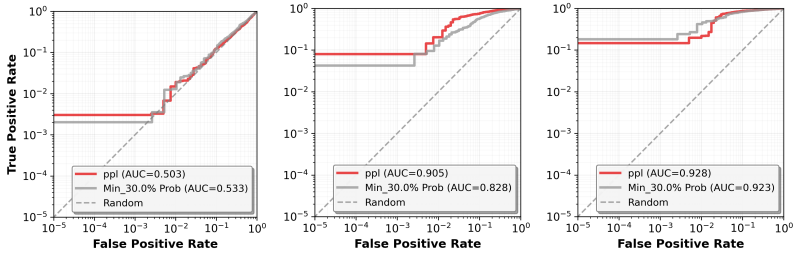


Next-token prediction accuracies for different unlearning methods compared to the baselines.

Next-token prediction accuracies for different unlearning methods compared to the baselines.

Experimental validation of theory

Our theory implies that after Forgetting-MarI, the mutual information between logits and the "seen/unseen" bit Z is negligible; hence any confidence-based test (perplexity, cross-entropy, log-likelihood, etc.) should fail to separate forgotten from genuinely unseen text.



Detector performance for a GPT2-LG model without unlearning (left), unlearned with Forgetting-MarI (middle), and the gold-standard unlearn baseline (right). (ppl = perplexity)

Open Challenges

Several fundamental questions remain open:

- **Unlearning through the inverse problem lens.** Separate parameter- and output-level unlearning. What are the limitation of auditing in each case?
- **Unlearning mechanisms.** Is invariance principle sufficient for output-level unlearning, or is randomness still required to achieve the optimal utility even for output-level unlearning? In the case of parameter-level Is randomness is needed to fill the holes of parameter-level unlearning.
- **Selecting the retain set.** Can we develop principled methods for identifying the minimal retain information needed to preserve downstream utility?

Machine unlearning for Large Language Models

Fix a language model p_θ over a finite vocabulary V and a length $T \geq 1$. For $y \in V^T$, let $p_\theta(\cdot \mid y_{<t})$ be the next-token distribution.

For a subset $s \subseteq V^T$, let μ_s be the uniform law on s and define averaged next-token marginals

$$(p_\theta)_t^s(v) := \mathbb{E}_{Y \sim \mu_s}\big[p_\theta(v \mid Y_{<t})\big], \qquad t \in [T],\ v \in V.$$

Write $p^r := \{(p_\theta)_t^r\}_{t \in [T]}$ and $p^u := \{(p_\theta)_t^u\}_{t \in [T]}$. For $m := r \cup u$,

$$p_t^m = \alpha p_t^r + (1-\alpha) p_t^u, \qquad \alpha := \frac{|r|}{|r| + |u|} \in (0,1).$$

Let $T^* \sim \mathrm{Uniform}([T])$ and $Z \sim \mathrm{Bernoulli}(1/2)$ be independent. Conditioned on $(T^* = t, Z)$, draw $X \sim p_t^m$ if $Z = 0$ and $X \sim p_t^r$ if $Z = 1$, and set $X_{\mathrm{MarI}} = (T^*, X)$.

$$\mathrm{I}(X_{\mathrm{MarI}}; Z) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{JSD}(p_t^m, p_t^r).$$

References i

Laurent Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot.
Machine unlearning.
In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159, Oxford, UK, 2021. IEEE.

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

References ii

- Alexandra Chouldechova and Aaron Roth.
  The frontiers of fairness in machine learning.
  CoRR, abs/1810.08810, 2018.
- Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel.
  Fairness through awareness.
  In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
- Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten.
  Certified data removal from machine learning models.
  arXiv:1911.03030, 2019.

References iii

- Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev.
  Neural optimal transport.
  arXiv preprint arXiv:2201.12220, 2022.
- Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh.
  Remember what you want to forget: Algorithms for machine unlearning.
  Advances in Neural Information Processing Systems, 34:18075–18086, 2021.
- Shizhou Xu and Thomas Strohmer.
  Fair data representation for machine learning at the Pareto frontier.
  Journal of Machine Learning Research, 24(331):1–63, 2023.
- Shizhou Xu and Thomas Strohmer.
  Machine unlearning via information theoretic regularization, 2025.

Existing Anchor-based Methods

- *(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 $\mathcal{H}$ [2, 1];
- *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)[5];
- *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 $W \subseteq \mathcal{H}$ [7].

ε-Marginal Unlearning vs. ε-Differential Privacy

This looks like differential privacy, no?

No.

Differential privacy (invented by Cynthia Dwork in 2006) requires

$$\sup_{D} \left| \log \left(\frac{\mathbb{P}(\hat{Y} \in D \mid X = X_0)}{\mathbb{P}(\hat{Y} \in D \mid X = X_1)} \right) \right| \leq \varepsilon,$$

where the datasets X_0 and X_1 differ by one data point.

ε-Marginal Unlearning vs. ε-Differential Privacy

- Differential privacy requires that the learning outcome (via a randomized algorithm) does not change significantly if any unknown individual's data is added to or removed from the dataset.
- In machine unlearning, we are asked to unlearn a *specific* data point or feature.
- Marginal unlearning addresses a *remedial* objective, whereas differential privacy addresses a *preventive* objective.

DP has been used to define differential unlearning, but this is often overkill and can give inferior unlearning performance.