# CERTIFIED FEATURE UNLEARNING UNDER SELECTIVE FEATURE REVOCATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Feature unlearning seeks to remove a model's predictive dependence on specified input features while preserving the utility of the remaining features. Existing feature-unlearning methods typically re-access the original values of the very feature being removed, which is often impermissible once that feature has been revoked for legal, privacy, or governance reasons. We study this *selective feature revocation* setting, in which the retained features and labels remain available but the per-example values of the revoked feature do not, and propose Bounded Mechanism-Covered Replacement (BMCR). BMCR replaces the revoked feature with values drawn from a declared bounded range and fine-tunes the model so that its predictions stay stable across these replacements; the declared range is the only information BMCR uses about the revoked feature, and its coverage defines the scope of the guarantee. We prove that the BMCR objective penalizes residual direct dependence on the revoked coordinate and that its near-minimizers are prediction-level close to models retrained without that feature. We further show that, at any finite training checkpoint, the effect of off-manifold replacements on in-distribution risk and the model-assisted attribute-inference advantage over the retained features alone are both controlled by the same residual dependence. Experiments on seven tabular datasets, three architectures, and single-, multi-, and sequential-deletion settings show that BMCR matches retraining in utility, outperforms replacement and structural-masking baselines, and is about $40\times$ faster than retraining in wall-clock time.

## 1 INTRODUCTION

In real-world deployments of machine learning models, some training data may later prove sensitive, mislabeled, legally restricted, poisoned, or outdated, while the deployed model may still retain and expose their influence through memorization, bias, or shortcut-based decisions Act (1996); Voigt & Von dem Bussche (2017); McCoy & Auret (2019). For example, a credit-risk model may have been trained with a medical or demographic attribute that must later be removed, while non-sensitive financial covariates and labels remain available for compliance-preserving model editing. Retraining from scratch for every such request is costly, making feature unlearning a practical alternative.

Feature unlearning differs from data-point unlearning because the removed feature is entangled across all training samples, so the model must erase a global predictive dependency rather than forget a finite set of examples. It is also typically requested precisely when the feature can no longer be used: a feature may become prohibited, turn out to have been unlawfully collected, or fall under a shorter retention period than the rest of the record Act (1996); Voigt & Von dem Bussche (2017). Regulators have recognized that deleting raw data does not remove its influence from a trained model, and have required the deletion or remediation of models derived from such data European Data Protection Board (2024); Federal Trade Commission (2021); U.S. District Court for the Northern District of California (2022). Yet most feature-unlearning methods re-access the original training set, including the removed feature Yang et al. (2025); Guo et al. (2022); Moon et al. (2024); Xu et al. (2024); Warnecke et al. (2023), and approximating or replaying its distribution may itself leak the information to be removed Chundawat et al. (2023); Tarun et al. (2023).

**Setting.** We study *selective feature revocation*: access is withdrawn for specific features (e.g., a medical attribute in a credit model) while the retained features and labels remain lawfully usable



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(e.g., income and repayment history). This is exactly the information available to retraining from scratch without the feature, the standard oracle for feature unlearning; our goal is to approach that oracle cheaply, with a certificate, and without touching the revoked values.

**Our approach.** Bounded Mechanism-Covered Replacement (BMCR) overwrites the revoked feature with random values from a declared range and fine-tunes the model until its predictions no longer change across those values (Fig. 1). Unlike random labeling Graves et al. (2021), which randomizes labels and pushes predictions toward noise, BMCR keeps the true labels and randomizes only the revoked input, so its optimum is the best predictor that uses the retained features alone. The declared range is the only information BMCR uses about the revoked feature, and it sets the scope of the guarantee: a range covering the feature's valid domain yields a full-domain certificate, and a sub-interval yields a certificate on that interval. Our contributions are summarized as follows.

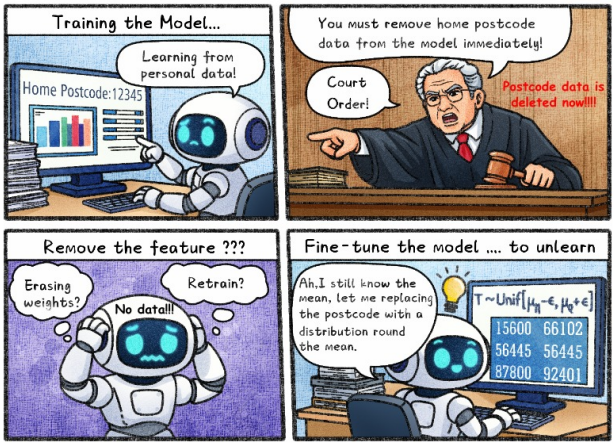


Figure 1: BMCR replaces the revoked feature with draws from a declared range and fine-tunes on retained features and true labels until predictions are invariant to it.

**1. Certified unlearning under selective revocation.** A fine-tuning procedure that needs only retained features and labels, alongside a declared support for the revoked feature, never its per-example values.

**2. Theory.** The BMCR objective decomposes into a retained-feature risk plus a residual-dependence penalty, so near-minimizers have small direct dependence on the revoked feature and are prediction-level close to retraining. We also bound, at any finite checkpoint, the effect of off-manifold replacements and the model-assisted attribute-inference advantage, and extend the guarantee to sequential deletions.

**3. Comprehensive empirical evaluation.** Seven datasets, three architectures, single-, multi-, and sequential-feature unlearning, structural-masking baselines, correlated-feature and dependence-controlled studies, attribute-inference and relearning audits, and wall-clock costs.

## 2 Related Work

Machine unlearning has become a key topic in machine learning, aiming to let models remove the influence of specific data points or subsets efficiently, without full retraining Nguyen et al. (2022); Gupta et al. (2021); Bourtoule et al. (2021); Guo et al. (2019); Graves et al. (2021); Baumhauer et al. (2022). Most existing machine unlearning methods focus on data-level unlearning, where the goal is to remove the influence of specific training examples (e.g., user data) rather than features Bourtoule et al. (2021); Guo et al. (2019); Gupta et al. (2021). Compared to data-level unlearning, substantially fewer studies Yang et al. (2025); Guo et al. (2022); Moon et al. (2024); Xu et al. (2024); Warnecke et al. (2023) have investigated how to selectively erase the impact of individual features in trained models. Early work by Warnecke et al. (2023) tackled feature unlearning by selectively recalibrating network weights linked to the target feature, followed by fine-tuning to preserve performance on retained features and labels. Later, Guo et al. (2022) introduced a more computationally efficient strategy that removes attribute influence via a mutual-information–based representation detachment loss. However, this approach provides no formal theoretical guarantee, and the mutual information term is estimated online during training, which may introduce bias and instability. More recently, Moon et al. (2024) extended feature unlearning to pre-trained generative models by modifying generator and encoder–decoder pathways, but the method is specialized to Generative Adversarial Network architectures and can incur substantial computational overhead. Xu et al. (2024) introduced a method using adversarial learning when feature annotations are available and interpretability-based

filtering otherwise. Yang et al. (2025) proposed a simple yet theoretically grounded method that randomizes and shuffles the values of target features in the training data for feature unlearning.

All of the aforementioned methods read the per-example values of the removed feature during unlearning. Once a feature has been revoked, such re-processing may be restricted by regulation or internal governance Act (1996); Voigt & Von dem Bussche (2017); De Min et al. (2024); Basaran et al. (2025); Chen et al. (2024), and it means that the deletion procedure itself processes the data it is intended to remove. BMCR instead needs only a declared support for the revoked feature, which is schema-level information that describes no individual record, and ties the scope of its guarantee to that support: the certificate covers exactly the declared range.

*Remark* 2.1 (Differences with randomization and masking). Random labeling Graves et al. (2021) relabels forget samples at random, and Shuffle Yang et al. (2025) permutes the observed values of the target feature, which requires those values. BMCR differs in what it randomizes (the input, with labels fixed), in what it needs (a declared support rather than observed values), and in what it certifies (a support-scoped bound on residual dependence and a prediction-level distance to retraining). Structural masking, which zeroes or removes the feature's input weights, deletes the input pathway but leaves the rest of the network tuned to inputs that contained the revoked feature; we compare against such baselines in Section 5.

# 3 METHOD

## 3.1 SETTING, ACCESS MODEL, AND SCOPE

Let $x = (x_{-i}, x_i)$, where $x_i$ is the revoked feature, and let $f_{\theta_0}$ be a model trained on the full feature vector. At unlearning time we observe $(x_{-i}, y)$ pairs but not the per-example values of $x_i$. We seek $\theta$ such that 1) predictions become insensitive to $x_i$ and 2) utility from $x_{-i}$ is retained.

**Unlearning target (invariance to $x_i$ while preserving $x_{-i} \to y$).** Our unlearning goal is to remove the influence of the missing coordinate $x_i$ while preserving the predictive relationship between the retained features $x_{-i}$ and the label $y$. We measure residual dependence on $x_i$ under the replacement distribution $T \sim q$ by

$$\mathcal{D}_q(\theta) := \mathbb{E}_{x_{-i}}\big[\mathrm{Var}_{T\sim q}\left(f_\theta(x_{-i}, T)\right)\big]. \tag{1}$$

By construction, $D_q(\theta) = 0$ if and only if, for $P_{x_{-i}}$-almost every $x_{-i}$, the random variable $f_\theta(x_{-i}, T)$ is $q$-almost surely constant in $T$. If, in addition, $t \mapsto f_\theta(x_{-i}, t)$ is continuous on $\mathrm{supp}(q)$, then this $q$-almost-sure invariance extends to pointwise invariance over $\mathrm{supp}(q)$. We write $D_q(\theta)$ as shorthand for the function-level residual-dependence functional evaluated at the predictor represented by $\theta$, namely $D_q(\theta) \equiv D_q(f_\theta) \equiv D_q(f)$ when the same quantity with $f = f_\theta$.

To assess whether unlearning preserves predictive utility from the remaining features $x_{-i}$ while removing dependence on the missing coordinate, we evaluate the *replacement risk* defined as

$$\mathcal{R}_q(\theta) := \mathbb{E}_{(x_{-i},y)}\,\mathbb{E}_{T\sim q}\big[\ell\big(f_\theta(x_{-i}, T), y\big)\big], \tag{2}$$

which depends only on $(x_{-i}, y)$ and the injected replacement noise; Section 4 shows that near-minimizing $\mathcal{R}_q$ yields invariance certificates while retaining the Bayes-optimal predictor based on $x_{-i}$. We also write $R_q(f)$ for the same quantity with $f = f_\theta$.

**Definition 3.1** (Certified unlearning domain and covered replacement)**.** Let $\mathcal{I}_i \subset \mathbb{R}$ denote the target interval of values over which the $i$-th feature coordinate must be unlearned. A replacement distribution $q_i$ is said to *cover* $\mathcal{I}_i$ if $\mathcal{I}_i \subseteq \mathrm{supp}(q_i)$. All invariance guarantees in this paper are with respect to $\mathrm{supp}(q_i)$. Therefore, whenever $q_i$ covers $I_i$, the average residual-dependence certificate is evaluated under a replacement distribution whose support covers $I_i$. Pointwise unlearning over $I_i$ follows in the ideal case $D_{q_i}(f) = 0$ together with continuity. Finite-$\delta$ range-wide guarantees require an additional regularity condition, such as Lipschitzness.

**Definition 3.2** ($(\epsilon, \delta)$ functional feature unlearning)**.** Let $f^{\mathrm{edit}} : \mathcal{X}_{-i} \times \mathbb{R} \to \mathbb{R}$ be an edited model obtained after unlearning coordinate $i$, and let $g^{\mathrm{rt}} : \mathcal{X}_{-i} \to \mathbb{R}$ be a model retrained from scratch without access to coordinate $i$. For a discrepancy function $\Delta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, we say that $f^{\mathrm{edit}}$ satisfies $(\epsilon, \delta)$ functional feature unlearning with respect to $g^{\mathrm{rt}}$ if $\mathbb{P}_{(X,Y)}\left(\Delta\big(f^{\mathrm{edit}}(X_{-i}, X_i), g^{\mathrm{rt}}(X_{-i})\big) > \epsilon\right) \le \delta$. For regression we use $\Delta(a, b) = |a - b|$. For binary classification we apply the same definition to bounded effective logits or predicted probabilities.

This definition follows the prediction-level feature-unlearning criterion: an edited model is considered feature-unlearned when its predictions are close to those of a model retrained from scratch without the target feature. Our formal results below establish this equivalence at the functional level. They do not imply parameter-level indistinguishability from retraining.

### 3.2 Core Idea: Bounded Replacement Unlearning

The unlearning guarantee is determined by the support of the replacement distribution. Let $\mathcal{I}_i$ be the target interval over which feature $i$ should be unlearned. We choose a bounded replacement distribution $q_i$ satisfying $\mathcal{I}_i \subseteq \mathrm{supp}(q_i)$. For a general bounded uniform replacement, $T_i \sim q_i := \mathrm{Unif}[a_i, b_i], \mathcal{I}_i \subseteq [a_i, b_i], \mu_{q_i} := \mathbb{E}[T_i] = \frac{a_i+b_i}{2}, \sigma^2_{q_i} := \mathrm{Var}(T_i) = \frac{(b_i-a_i)^2}{12}$. The mean-centered replacement used in BMCR is a special case. Given a reference center $c_i$, define the replacement radius $\rho_i(\mathcal{I}_i; c_i) := \sup_{t\in\mathcal{I}_i} |t - c_i|$. Then $T_i \sim q_i := \mathrm{Unif}[c_i - \rho_i,\ c_i + \rho_i], \mu_{q_i} = c_i, \sigma^2_{q_i} = \frac{\rho_i^2}{3}$, and by construction $\mathcal{I}_i \subseteq [c_i - \rho_i,\ c_i + \rho_i] = \mathrm{supp}(q_i)$. Thus, $\rho_i$ determines the certified unlearning interval, while $c_i$ only chooses the center of the replacement distribution.

**Known-bounds versus mean-only settings.** If valid feature bounds are known, say $\mathcal{S}_i = [\ell_i, u_i]$, then full-domain coordinate-level unlearning is obtained by choosing the full-support replacement $T_i \sim q_i := \mathrm{Unif}[\ell_i, u_i]$. This choice does not require the training mean of the removed feature. Equivalently, it can be written in centered form with $c_i = \frac{\ell_i+u_i}{2}, \rho_i = \frac{u_i-\ell_i}{2}$. If only a reference mean $\mu_i$ is available, then the guarantee is interval-specific rather than full-domain. For any declared target interval $\mathcal{I}_i = [a_i, b_i]$, we set $c_i = \mu_i, \rho_i = \sup_{t\in\mathcal{I}_i} |t - \mu_i| = \max\{|a_i - \mu_i|,\ |b_i - \mu_i|\}$, and use $T_i \sim \mathrm{Unif}[\mu_i - \rho_i,\ \mu_i + \rho_i]$. Then $I_i \subseteq \mathrm{supp}(q_i)$, so the average certificate is computed under a replacement support that covers $I_i$. Range-wide pointwise control over all $t \in I_i$ follows only under the continuity or Lipschitz conditions stated in Section 4 and the appendix.

**Fine-tuning objective:** We fine-tune starting from $\theta_0$ by minimizing the expected loss over the feature-replaced distribution:

$$\mathcal{R}_q(\theta) := \mathbb{E}_{(x_{-i},y)}\, \mathbb{E}_{T\sim q}\Big[\ell\big(f_\theta(x_{-i}, T), y\big)\Big]. \tag{3}$$

In practice, we optimize a weight-decayed variant

$$\widetilde{\mathcal{R}}_q(\theta) := \mathcal{R}_q(\theta) + \frac{\lambda}{2}\|\theta\|_2^2, \tag{4}$$

where $\ell$ is the task loss (squared loss for regression or cross-entropy for classification) and $\lambda > 0$ is the weight decay coefficient. We optimize Eq. (4) by standard SGD/AdamW initialized at $\theta_0$, using mini-batches $\{(x^{(j)}_{-i}, y^{(j)})\}_{j=1}^B$ and i.i.d. draws $\{T^{(j)}\}_{j=1}^B$ with $T^{(j)} \sim q$ ($B$ is the batch size):

$$\theta \leftarrow \theta - \eta\nabla_\theta \left[\frac{1}{B}\sum_{j=1}^{B} \ell\big(f_\theta(x^{(j)}_{-i}, T^{(j)}), y^{(j)}\big) + \frac{\lambda}{2}\|\theta\|_2^2\right]. \tag{5}$$

Section 4 analyzes the unregularized risk $\mathcal{R}_q$ in Eq. (3). In practice we optimize $\widetilde{\mathcal{R}}_q = \mathcal{R}_q + \frac{\lambda}{2}\|\theta\|_2^2$. If optimization is restricted to $\{\theta : \|\theta\|_2 \le B\}$ (e.g., by early stopping or norm clipping), then any minimizer of $\widetilde{\mathcal{R}}_q$ is a $\frac{\lambda}{2}B^2$-approximate minimizer of $\mathcal{R}_q$, so our certificates bound $D_q$ up to $O(\lambda B^2)$.

## 4 Theoretical Analysis

**Proof sketch.** Our analysis studies how fine-tuning with feature replacement drives functional invariance to the removed coordinate. The key quantity is the residual dependence $D_q(f) = \mathbb{E}_{x_{-i}} \mathrm{Var}_{T\sim q}(f(x_{-i}, T))$, which vanishes if and only if $f(x_{-i}, T)$ is $q$-almost surely constant in the replacement coordinate for almost every $x_{-i}$. Under squared loss, the replacement risk decomposes as $\mathbb{E}_{x_{-i},y}\mathbb{E}_{T\sim q}\left[(f(x_{-i}, T) - y)^2\right] = \mathbb{E}_{x_{-i},y}\left[(\bar{f}(x_{-i}) - y)^2\right] + D_q(f)$, where $\bar{f}(x_{-i}) = \mathbb{E}_{T\sim q}[f(x_{-i}, T)]$. Thus, dependence on the removed coordinate cannot improve the replacement objective. Moreover, a near-minimizer of the replacement risk is close in prediction space to the

Bayes predictor $h^\star(x_{-i}) = \mathbb{E}[y \mid x_{-i}]$, and therefore close to a near-optimal model retrained without $x_i$. For classification, bounded effective logits give an analogous dependence-penalized lower bound through the uniform strong convexity of logistic cross-entropy on $[-K, K]$.

**Lemma 4.1** (Exact decomposition of replacement MSE risk)**.** *Assume $T \sim q$ is independent of $(x_{-i}, y)$ and $R_q(f) < \infty$ with $\ell(\hat{y}, y) = (\hat{y} - y)^2$. Then*

$$R_q(f) = \mathbb{E}_{(x_{-i},y)}\left[(\bar{f}(x_{-i}) - y)^2\right] \; + \; D_q(f). \tag{6}$$

*Proof.* A detailed proof is provided in Appendix D.2. □

**Theorem 4.2** (Excess-risk certificate for coordinate-level feature unlearning under MSE)**.** *Assume $Y \in L_2$, $T_i \sim q_i$ is sampled independently of $(X_{-i}, Y)$, and $R_{q_i}(f) < \infty$ under squared loss. Define $\bar{f}(X_{-i}) := \mathbb{E}_{T_i \sim q_i}[f(X_{-i}, T_i)]$, $D_{q_i}(f) := \mathbb{E}_{X_{-i}} \operatorname{Var}_{T_i \sim q_i}(f(X_{-i}, T_i))$. Let $h^\star(X_{-i}) := \mathbb{E}[Y \mid X_{-i}]$, and define the $T_i$-invariant predictor $f^\star(X_{-i}, t) \equiv h^\star(X_{-i})$, $t \in \operatorname{supp}(q_i)$. Then $f^\star$ is a global minimizer of $R_{q_i}$ over all measurable predictors with finite second moment, and $D_{q_i}(f^\star) = 0$. Moreover, for any such $f$, $R_{q_i}(f) - R_{q_i}(f^\star) = \mathbb{E}_{X_{-i}}\left[\left(\bar{f}(X_{-i}) - h^\star(X_{-i})\right)^2\right] + D_{q_i}(f)$. Consequently, $D_{q_i}(f) \le R_{q_i}(f) - \inf_g R_{q_i}(g)$, where the infimum is over the same class of measurable finite-second-moment predictors. In particular, if $\widehat{f}$ is a $\delta$-near minimizer, $R_{q_i}(\widehat{f}) - \inf_g R_{q_i}(g) \le \delta$, then $D_{q_i}(\widehat{f}) \le \delta$.*

*Proof.* A detailed proof is provided in Appendix D.3. □

Theorem 4.2 controls average residual dependence under the replacement distribution $q_i$. If a target interval $I_i$ satisfies $I_i \subseteq \operatorname{supp}(q_i)$, then the replacement mechanism covers the interval on which coordinate $i$ is intended to be unlearned. In the ideal case $D_{q_i}(\widehat{f}) = 0$, $\widehat{f}(X_{-i}, T_i)$ is $q_i$-almost surely constant in $T_i$ for $P_{X_{-i}}$-almost every $X_{-i}$. If, in addition, $t \mapsto \widehat{f}(X_{-i}, t)$ is continuous on $\operatorname{supp}(q_i)$, then this implies exact coordinate-level unlearning over every interval $I_i \subseteq \operatorname{supp}(q_i)$.

An immediate consequence of Theorem 4.2 is that the replacement objective has the same population optimum as retraining with only $X_{-i}$; a formal statement and proof are given in Appendix D.15.

**Theorem 4.3.** *Let $Z := X_{-i}$, and let $T \sim q$ be sampled independently of $(Z, Y)$. Let $h^\star(Z) := \mathbb{E}[Y \mid Z]$ be the Bayes predictor using only the retained features. Suppose the BMCR-edited predictor $\widehat{f}$ and the retrained-without-feature predictor $\widehat{g}$ satisfy $R_q(\widehat{f}) - R_q(h^\star) \le \gamma$, $R_{-i}(\widehat{g}) - R_{-i}(h^\star) \le \eta$, where $h^\star$ is viewed as the $t$-invariant predictor $h^\star(Z, t) \equiv h^\star(Z)$ inside $R_q$. Then $\left\|\widehat{f}(Z, T) - \widehat{g}(Z)\right\|_{L_2(P_Z \otimes q)} \le \sqrt{\gamma} + \sqrt{\eta}$. Consequently, for every $\epsilon > 0$, $\mathbb{P}_{Z,T}\left(\left|\widehat{f}(Z, T) - \widehat{g}(Z)\right| > \epsilon\right) \le \frac{(\sqrt{\gamma} + \sqrt{\eta})^2}{\epsilon^2}$.*

See Appendix D.8 for detailed proof. Theorem 4.3 shows that BMCR is prediction-level equivalent to retraining under the replacement distribution: a near-minimizer of the replacement risk is close to a near-optimal retrained model that only observes $X_{-i}$. This is a functional guarantee, not a parameter-level indistinguishability guarantee.

**Assumption 4.4** (Bounded effective logits)**.** For binary classification, $Y \in \{0, 1\}$, and the loss is $\ell(z, y) = \log(1 + \exp(z)) - yz$. For a fixed $K < \infty$, classification risks are computed on the bounded effective logit $\widetilde{f}_\theta(x_{-i}, t) = \phi_K(f_\theta(x_{-i}, t))$, $\phi_K : \mathbb{R} \to [-K, K]$.

Throughout this subsection, define $m_K := \min_{|z| \le K} \sigma(z)(1 - \sigma(z)) = \sigma(K)(1 - \sigma(K)) > 0$, where $\sigma(z) = 1/(1 + \exp(-z))$. Since $\partial^2_{zz} \ell(z, y) = \sigma(z)(1 - \sigma(z))$, the binary logistic cross-entropy loss is $m_K$-strongly convex in the logit on $[-K, K]$.

We use the standard fact that if $\phi'' \ge m$, then $\mathbb{E}[\phi(Z)] \ge \phi(\mathbb{E}[Z]) + \frac{m}{2} \operatorname{Var}(Z)$, with proof provided in Appendix D.4.

**Theorem 4.5** (Bounded-logit replacement certificate)**.** *Assume $T \sim q$ is sampled independently of $(X_{-i}, Y)$, and suppose Assumption 4.4 holds. For any raw-logit predictor $f$, let $\widetilde{f} = \phi_K \circ f$ be its bounded effective logit, and define $\overline{\widetilde{f}}(x_{-i}) := \mathbb{E}_{T \sim q}\left[\widetilde{f}(x_{-i}, T)\right]$, and*

$D_q(\widetilde{f}) := \mathbb{E}_{x_{-i}} \operatorname{Var}_{T\sim q}\left(\widetilde{f}(x_{-i}, T)\right)$. *Then* $R_q(\widetilde{f}) \geq \mathbb{E}_{(x_{-i},y)}\left[\ell\left(\overline{\widetilde{f}}(x_{-i}), y\right)\right] + \frac{m_K}{2} D_q(\widetilde{f})$. *Consequently,* $D_q(\widetilde{f}) \leq \frac{2}{m_K}\left(R_q(\widetilde{f}) - \inf_{g\in\mathcal{G}_K} R_q(g)\right)$, *where* $\mathcal{G}_K$ *is the class of measurable predictors* $g : \mathcal{X}_{-i} \times \operatorname{supp}(q) \to [-K, K]$. *In particular, if* $R_q(\widehat{\widetilde{f}}) - \inf_{g\in\mathcal{G}_K} R_q(g) \leq \delta$, *then* $D_q(\widehat{\widetilde{f}}) \leq \frac{2\delta}{m_K}$.

*Proof.* A detailed proof is provided in Appendix D.12. □

Theorem 4.5 controls the residual dependence of the bounded effective logit $\widetilde{f}$, and hence of the deployed probability $\sigma(\widetilde{f})$. The theorem does not require the raw logit $f$ itself to be bounded or invariant outside the effective-logit map. Thus, for classification, the certified quantity is $D_q(\widetilde{f})$, not necessarily $D_q(f)$.

**Off-manifold replacements and proxy information.** Independent replacement can create feature combinations that never occur in data (e.g., a 30-year-old with 30 years of work experience), and information about $x_i$ may remain recoverable from correlated retained features. The next result addresses both at *any* finite checkpoint, without convexity, optimality, or independence between $X_i$ and $X_{-i}$ (proof in Appendix E).

**Proposition 4.6** (Finite-checkpoint transfer)**.** *Let* $q = \operatorname{Unif}[a, b]$ *with* $X_i \in [a, b]$ *a.s., let* $t \mapsto f(z, t)$ *be* $L_f$*-Lipschitz on* $[a, b]$ *for every* $z$*, and let* $\kappa(f) := 2\,(L_f(b-a)\,D_q(f))^{1/3}$. *For any predictor* $f$*: (i)* Realistic inputs: *for any conditional law of* $X_i$ *given* $X_{-i}$ *supported in* $[a, b]$, $\mathbb{E}_{X_{-i}} \operatorname{Var}_{X_i|X_{-i}}[f(X_{-i}, X_i)] \leq \kappa(f)^2$. *(ii)* In-distribution risk: *if* $\ell(\cdot, y)$ *is* $L_\ell$*-Lipschitz on the range of* $f$*, then* $|R_{\mathrm{ID}}(f) - R_q(f)| \leq 2L_\ell\,\kappa(f)$, *where* $R_{\mathrm{ID}}(f) := \mathbb{E}[\ell(f(X_{-i}, X_i), Y)]$. *(iii)* Attribute inference: *let an attacker observe* $X_{-i}$ *and* $f(X_{-i}, t_1), \ldots, f(X_{-i}, t_k)$ *for queries* $t_j = t_j(X_{-i}) \in [a, b]$*, and output an estimate of* $X_i$ *that is* $L_A$*-Lipschitz in the responses (in* $\ell_\infty$*) under a* $\lambda$*-Lipschitz loss. Its expected loss is at least that of the best attacker using* $X_{-i}$ *alone, minus* $\lambda L_A\,\kappa(f)$.

with Theorem 4.2, a $\delta$-near minimizer has $\kappa \leq 2(L_f(b-a)\delta)^{1/3}$; for classification the statements apply to $\widetilde{f}$, on which cross-entropy is 1-Lipschitz. A finite-sample version that separates optimization, sampling, and off-manifold error is given in Appendix E.

*Remark* 4.7 (What BMCR does and does not remove). $D_q$ measures *direct* functional dependence on coordinate $i$. BMCR does not, and is not meant to, remove information about $X_i$ that is statistically recoverable from $X_{-i}$: that information is equally available to the retrained oracle $g^{\mathrm{rt}}$, the target in Definition 3.2, which offers an attacker no advantage over $X_{-i}$ alone. Proposition 4.6(iii) shows that BMCR's *excess* leakage over this data-inherent baseline vanishes with $D_q$. Guarantees are support-scoped (queries outside $\operatorname{supp}(q)$ are not covered), and the Lipschitz constants are audited empirically rather than certified.

*Remark* 4.8 (*Extension to multiple features*). The above argument extends verbatim when replacing a set $S$ of features with $T_S = (T_j)_{j\in S} \sim q_S := \bigotimes_{j\in S} q_j$ independent of $(x_{-S}, y)$: one simply replaces the scalar $T$ by the vector $T_S$ and interprets $\bar{f}(x_{-S}) = \mathbb{E}_{T_S}[f(x_{-S}, T_S)]$, $D_{q_S}(f) = \mathbb{E}_{x_{-S}} \operatorname{Var}_{T_S}\big(f(x_{-S}, T_S)\big)$. Moreover, by the law of total variance, for any $j \in S$, $\operatorname{Var}_{T_S}\big(f(x_{-S}, T_S)\big) \geq \mathbb{E}_{T_{S\setminus\{j\}}}\big[\operatorname{Var}_{T_j}\big(f(x_{-S}, T_S) \mid T_{S\setminus\{j\}}\big)\big]$, so controlling $D_{q_S}(f)$ also controls dependence on each individual replaced feature. For classification, the same statement applies to the bounded effective logit $\widetilde{f}$, so controlling $D_{q_S}(\widetilde{f})$ controls joint and coordinate-wise residual dependence of the deployed classifier on the replaced features.

**Sequential requests.** When deletion requests arrive over time, BMCR is applied cumulatively: after the $k$-th request, all revoked features $S_k$ are replaced jointly with $q_{S_k}$. Each run therefore re-certifies every earlier deletion, and because the population optimum $\mathbb{E}[Y \mid X_{-S_k}]$ does not depend on the order of requests, no importance-based schedule is required (Appendix U).

## 5 Experiments and Results

**Datasets.** Seven widely-used tabular datasets, abbreviated as *Blood*, *Cali*, *Credit*, *Diabetes*, *House*, *Magic*, and *Pol*, from the OpenML library Feurer et al. (2021) are used. Details are provided in Appendix G.

Table 1: Information about the revoked feature $x_i$ used by BMCR and all baselines.

| Tier | Information on $x_i$ | Methods | Support-scoped certificate |
|---|---|---|---|
| A0 | None beyond the deployed pipeline | $\text{FA}_{\{-1,0,1\}}$, SFI, FPE, BMCR-$\varepsilon$ | BMCR-$\varepsilon$: $[-\varepsilon, \varepsilon]$ (standardized) |
| A1 | Declared or public support $[\ell_i, u_i]$ | BMCR-full | full domain $[\ell_i, u_i]$ |
| A2 | Statistics computed from data | GS (mean, std), SAM (mean), BMCR-MM (min/max) | BMCR-MM: empirical range |
| A3 | Per-example values | Shuffle, FOUL, MIUL | – |

Table 2: BMCR (ResNet, $\varepsilon = 0.1$): epoch-wise averages over all datasets and both the most- and least-influential unlearned features. Standard deviations pool heterogeneity *across* datasets and features, not run-to-run instability (per-dataset results: Appendix T). "Orig." is the model before unlearning; only MIA is defined for it. Other backbones and ranges: Appendix Q.

| Metric | Orig. | 1 | 10 | 50 | 100 | 200 | 600 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|
| RIPV | – | 0.0268 ±0.0215 | 0.0186 ±0.0148 | 0.0139 ±0.0112 | 0.0114 ±0.0096 | 0.0098 ±0.0083 | 0.0081 ±0.0070 | 0.0075 ±0.0065 | 0.0069 ±0.0061 |
| TRI | – | 0.931 ±0.116 | 0.974 ±0.092 | 0.970 ±0.047 | 0.984 ±0.031 | 0.985 ±0.033 | 0.986 ±0.029 | 0.985 ±0.028 | 0.987 ±0.031 |
| EI | – | 2213.8 ±1203.2 | 343.0 ±104.9 | 78.8 ±12.2 | 40.3 ±5.3 | 20.4 ±2.6 | 6.9 ±0.8 | 4.1 ±0.5 | 2.7 ±0.3 |
| RASI | – | 0.723 ±0.167 | 0.781 ±0.109 | 0.828 ±0.077 | 0.858 ±0.065 | 0.868 ±0.072 | 0.886 ±0.068 | 0.891 ±0.066 | 0.923 ±0.078 |
| SRI | – | 0.118 ±0.089 | 0.130 ±0.117 | 0.127 ±0.094 | 0.121 ±0.091 | 0.107 ±0.079 | 0.133 ±0.095 | 0.089 ±0.060 | 0.116 ±0.087 |
| SDI | – | 0.033 ±0.033 | 0.032 ±0.029 | 0.033 ±0.033 | 0.034 ±0.035 | 0.031 ±0.031 | 0.031 ±0.033 | 0.032 ±0.031 | 0.033 ±0.033 |
| MIA | 0.682 ±0.024 | 0.555 ±0.041 | 0.529 ±0.032 | 0.515 ±0.026 | 0.518 ±0.027 | 0.530 ±0.031 | 0.532 ±0.032 | 0.519 ±0.027 | 0.518 ±0.026 |
| $\Delta_{\text{DP}}$ | – | 0.0835 ±0.0642 | 0.0382 ±0.0315 | 0.0405 ±0.0331 | 0.0182 ±0.0164 | 0.0321 ±0.0276 | 0.0178 ±0.0161 | 0.0291 ±0.0243 | 0.0145 ±0.0132 |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | – | 0.0445 ±0.0386 | 0.0335 ±0.0281 | 0.0368 ±0.0304 | 0.0205 ±0.0182 | 0.0332 ±0.0275 | 0.0225 ±0.0195 | 0.0245 ±0.0208 | 0.0238 ±0.0202 |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | – | 0.1385 ±0.0964 | 0.0521 ±0.0435 | 0.0442 ±0.0372 | 0.0212 ±0.0191 | 0.0315 ±0.0278 | 0.0172 ±0.0160 | 0.0385 ±0.0324 | 0.0195 ±0.0178 |

**Methods.** Methods differ in what they need to know about the revoked feature $x_i$; Table 1 groups every method we evaluate into four access tiers, and we report results by tier (examples and discussion in Appendix B; per-method details in Table 5). BMCR is run with five standardized-space ranges $[-\varepsilon, \varepsilon]$, $\varepsilon \in \{0.1, 0.5, 1, 5, 10\}$ (tier A0), and with uniform sampling over the feature's min–max range (BMCR-MM; A1 if the bounds are public, A2 if computed from data). Baselines follow the tiers of Table 1: fixed-value assignment (FA, value $-1$, 0, or 1) and two structural-masking methods (A0), namely *SFI*, which zeroes and freezes the input-projection column of $x_i$ and fine-tunes, and *FPE*, which rebuilds the network without the $x_i$ input, initializes all compatible weights from $f_{\theta_0}$, and fine-tunes; Gaussian sampling (GS) with the training mean and standard deviation, and sampling around the mean (SAM) (A2); and the full-access methods Shuffle Yang et al. (2025), FOUL Warnecke et al. (2023), and MIUL Guo et al. (2022) (A3; Appendix J). Dataset statistics are in Appendix H.

**Neural Architecture.** MLP, ResNet He et al. (2016), and FT-Transformer (FTformer) Gorishniy et al. (2021) are used as neural models for all unlearning algorithms.

**Evaluation Metrics.** To directly evaluate the residual-dependence certificate in Eq. (1), we report Replacement-Induced Prediction Variance (RIPV). For an audit set $\{x_{-i}^{(j)}\}_{j=1}^{n}$ and $m$ replacement samples $T_{jr} \sim q$, we compute $\text{RIPV} = \widehat{D}_q(f) := \frac{1}{n}\sum_{j=1}^{n} \text{Var}_{r=1,\dots,m}\left[f(x_{-i}^{(j)}, T_{jr})\right]$. For classification, we compute RIPV on the bounded effective logit $\widetilde{f}$ used in Theorem 4.5, or on the deployed probability $\sigma(\widetilde{f})$. Lower RIPV indicates smaller direct residual dependence on the removed coordinate. Further, following Yang et al. (2025), we utilize additional nine key metrics to quantify unlearning performance. *Test Retention Index* (TRI) and *Efficiency Index* (EI): TRI measures accuracy alignment with a retrain-from-scratch model (where a value of one represents parity), while EI assesses the speedup relative to retrain-from-scratch, where a value greater than one means time-efficient unlearning; *Robustness Against Shuffling Index* (RASI): It measures how well the model has decoupled from the unlearned feature by shuffling its values during testing. A value near one suggests the model's direct sensitivity to the target coordinate is substantially reduced under the evaluated metrics. *SHAP-based Metrics (supplementary)*: Using SHapley Additive exPlanations (SHAP) Lundberg & Lee (2017), we define the *SHAP Retention Index* (SRI) to compare feature importance before and after unlearning, and the *SHAP Distance-to-Zero Index* (SDI) to measure the absolute deviation of the unlearned feature's importance from zero. Values near zero for both metrics indicate effective unlearning from the SHAP perspective. In addition, we also evaluate privacy and fairness

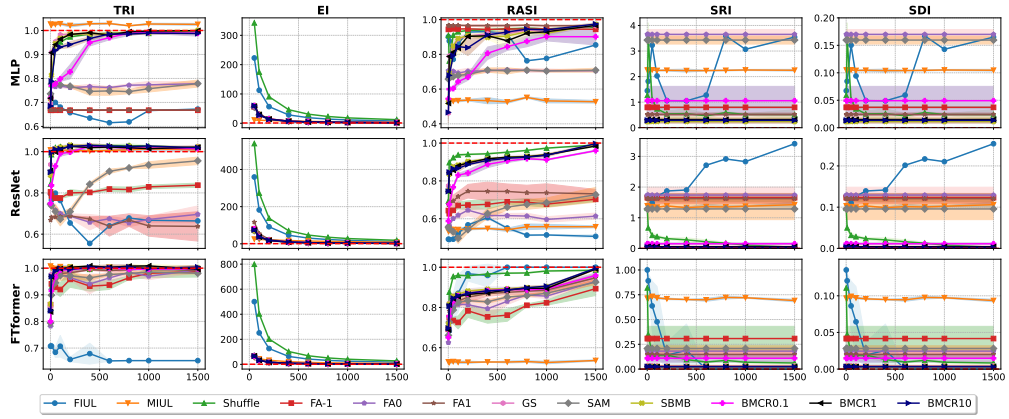


Figure 2: Single-feature unlearning evaluation results of the most influential feature in the *Magic* tabular dataset. For visual clarity, only five metrics and three variants of BMCR are shown. The full results are provided in Tables 17 to 19.

behavior after unlearning. For privacy, we use a membership-inference attack (MIA) metric (Shokri et al., 2017) to test whether the edited model can be distinguished from the retrain-from-scratch reference. An attack accuracy close to $50\%$ indicates near-random guessing under this audit. This is an empirical indistinguishability audit, not a formal privacy guarantee. For fairness, when the removed coordinate is treated as protected, we report demographic parity and equalized odds gaps. Demographic parity is measured by the absolute difference in favorable prediction rates, denoted $\Delta_{\mathrm{DP}}$. Equalized odds is measured by the absolute differences in true positive rates and false positive rates, denoted $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ and $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$. Lower values indicate smaller group-wise prediction disparities. Details of MIA, DP, and EO are provided in Appendix N.

**Training and cost protocol.** Original and retrained models are trained for up to 1500 and 2000 epochs; checkpoints are saved every 100 epochs and the one with the lowest held-out loss is used. EI uses the time needed to reach the *selected* checkpoint, not the epoch cap, so the larger retraining cap only ensures convergence (Appendix K). BMCR is run once for 1500 epochs; the reported epochs are checkpoints of that single run, used to characterize the trajectory, and the $\varepsilon$ values are ablations fixed in advance rather than a per-request search. All trials are repeated three times on one NVIDIA V100 GPU; other hyperparameters are in Appendix L.

**Selection on Unlearned Features.** To test the adaptability of our unlearning method, we target features at opposite ends of the importance spectrum: the most influential and the least influential features. Importance is determined by the Shapley values calculated from the originally trained models. Detailed rankings for each dataset can be found in Appendix O.

**Single-feature unlearning.** Table 2 and Fig. 2 summarize single-feature results (all datasets, features, and $\varepsilon \in \{0.5, 5\}$: Appendices P, I, and Q). RIPV, the empirical certificate, falls from $0.0268$ to $0.0069$ while TRI approaches parity with retraining and RASI rises, as Theorem 4.2 predicts. Most of the gain occurs early: at 100 epochs BMCR already reaches RIPV $0.0114$ and TRI $0.984$ at EI $40.3$. MIA accuracy drops from $0.682$ for the original model to $0.515$–$0.555$, near the $0.5$ chance level, and fairness gaps shrink. Aggregate SRI/SDI changes look small because Table 2 also averages least-influential features whose SHAP importance is near zero from the start; for the most influential *Cali* feature, SRI falls from $99.1\%$ to $3.1\%$ (Appendix T). Across architectures, BMCR matches or exceeds the A0 and A2 baselines, which use the same or more information about $x_i$, and is competitive with the A3 methods, some of which are unstable on FT-Transformer.

**Multi-feature and sequential unlearning.** Jointly removing the two or three most influential features gives results that qualitatively match the single-feature case, with slightly slower convergence (three features at 1500 epochs: RIPV $0.0083$, TRI $0.923$; Appendix R). When the three features are deleted one at a time in all six orders, RIPV stays within $0.0067$–$0.0085$ and TRI within $0.974$–$0.994$, with no consistent advantage for importance-based ordering (Appendix U).

**Structural-masking baselines.** A natural alternative to replacement is to cut the input pathway of $x_i$ directly, which needs no information about $x_i$ at all. SFI zeroes the first-layer weights attached to $x_i$ and keeps them frozen while fine-tuning on $(x_{-i}, y)$; FPE builds a network without the $x_i$ input,

Table 3: **(a) Structural masking vs. BMCR** on the most influential feature of each dataset, averaged over seven datasets and three architectures (per-dataset mean±std in Appendix W). **(b) Attribute inference of $x_i$.** An attacker predicts $x_i$ from $x_{-i}$ alone (data-only floor, shared by every model) or additionally queries the model at candidate values of $x_i$; $R^2$ is higher for more leakage, and "Gain" is the increase over the data-only floor. "+200 ep." fine-tunes the model for 200 further epochs on retained data only.

(a)

| | RIPV↓ | TRI→ 1 | RASI↑ | SRI↓ | MIA→ .5 | $\Delta_{\mathrm{DP}}$ ↓ |
|---|---|---|---|---|---|---|
| SFI | 0.033 | 0.780 | 0.868 | 0.805 | 0.560 | 0.070 |
| FPE | 0.042 | 0.896 | 0.804 | 1.575 | 0.546 | 0.076 |
| $\mathrm{BMCR}_1$ | 0.011 | 1.009 | 0.974 | 0.147 | 0.515 | 0.013 |
| $\mathrm{BMCR}_{10}$ | 0.007 | 1.003 | 0.987 | 0.121 | 0.509 | 0.016 |

(b)

| Attacker input | $R^2$ | Gain | +200 ep. |
|---|---|---|---|
| $x_{-i}$ only (floor) | 0.669 | – | – |
| + Original model | 0.813 | $+0.144$ | – |
| + Retrained model | 0.678 | $+0.009$ | 0.686 |
| + BMCR model | 0.686 | $+0.017$ | 0.697 |

copies every compatible weight from $f_{\theta_0}$, and fine-tunes it. Table 3(a) shows that both lose substantial utility (TRI $0.78$ and $0.90$ vs. $1.01$ for $\mathrm{BMCR}_1$) and leave larger fairness gaps ($\Delta_{\mathrm{DP}}$ $0.070$ and $0.076$ vs. $0.013$). Removing the pathway does not re-fit the downstream layers, which were trained jointly with $x_i$; BMCR re-fits them toward the retained-feature optimum $\mathbb{E}[Y \mid X_{-i}]$ (Theorem 4.2).

**Correlated features.** Correlation between forget and retain data can prevent gradient ascent from reaching the retrained model Mavrothalassitis et al. (2025); we test whether a retained near-duplicate of $x_i$ similarly hinders BMCR. We take all 10 feature pairs with Pearson $r > 0.8$ (in *Cali*, *Credit*, *Magic*, *Blood*, and *House*), unlearn one feature of each pair while keeping the other. RIPV falls from $0.028$ to $0.008$, $\Delta_{\mathrm{DP}}$ falls from $0.077$ to $0.015$, and TRI reaches $0.92$ (Appendix V). Correlation creates no conflicting objective, since BMCR and retraining share the same retained-feature optimum.

**Off-manifold replacement.** BMCR samples $T$ independently of $x_{-i}$, which can create unrealistic inputs when $x_i$ depends on $x_{-i}$. To isolate this effect, we build a synthetic task in which $x_i$ and a retained feature $x_{\mathrm{corr}}$ are both $\mathrm{Unif}[-1, 1]$ with dependence $\rho \in \{0, 0.5, 0.9, 0.99\}$, while the label's dependence on $x_i$ stays fixed. We compare independent BMCR with an oracle that samples $T$ from the true $P(x_i \mid x_{-i})$ (a reference that needs information BMCR lacks), reporting the utility gap as the percentage increase in test loss over retraining. At $\rho = 0$ the two coincide; at $\rho = 0.99$, independent replacement is behind after 200 epochs ($0.78\%$ vs. $0.43\%$), but the difference shrinks from $1.68$ points at 10 epochs to $0.07$ at 1500. On real data, grouping removed features by how well $x_{-i}$ predicts them (cross-fitted $R^2$), the high-dependence group ($R^2 \geq 0.6$) shows a gap of only $0.26$ points at 200 epochs (Appendix E). This matches Proposition 4.6(ii): the off-manifold effect is transient and shrinks with $D_q$.

**Proxy leakage and relearning.** Because $x_i$ may be predictable from $x_{-i}$, we measure leakage relative to a data-only floor: a random forest predicting $x_i$ from $x_{-i}$ alone reaches $R^2 = 0.669$, a level that any model, including the retrained oracle, cannot reduce. A model-assisted attacker additionally queries the model at candidate values of $x_i$ within the support. In Table 3(b), this adds $0.144$ to $R^2$ for the original model but only $0.017$ for BMCR, close to the $0.009$ of retraining, as Proposition 4.6(iii) predicts. After 200 further epochs of fine-tuning on retained data only, BMCR's attack $R^2$ stays close to retraining ($0.697$ vs. $0.686$) and RIPV barely changes ($0.0075 \to 0.0089$), so direct dependence does not reappear (Appendix X).

**Wall-clock cost.** BMCR and retraining use the same data, so the advantage is purely computational. In end-to-end time on one V100, retraining on *Magic*/MLP reaches its selected checkpoint (epoch 1600) in $329.7$ s ($0.092$ GPU-hours), whereas 100 BMCR epochs take $8.3$ s ($0.002$ GPU-hours), a $39.7\times$ speedup. Sampling adds $2.4\%$ overhead, and peak memory is similar (Appendix S).

## 6 Conclusion

We introduced BMCR, a feature-unlearning method for selective feature revocation that edits a trained tabular model using only the retained features, the labels, and a declared support for the revoked feature. Its replacement objective decomposes into a retained-feature risk plus a residual-dependence penalty. This yields support-scoped certificates, prediction-level closeness to retraining, finite-checkpoint control of off-manifold effects and of model-assisted attribute inference, and cumulative guarantees for sequential deletions. Empirically, BMCR matches retraining in utility at a fraction of its cost and outperforms baselines that use the same or less information about the revoked feature.

### AI USE STATEMENT

Generative AI tools were used to edit the manuscript for clarity and readability, to assist with routine computational workflows such as preparing and submitting Slurm jobs for the unlearning and retraining runs, and to assist with writing and checking the proofs in the Appendix. Generative AI was not used to develop the core research ideas of this work, including the selective feature revocation setting, the BMCR procedure, the residual-dependence certificate and its connection to retraining, the access-tier taxonomy, the design of the experiments, or the interpretation of the results. The remaining disclosure categories, such as dataset reformatting and translation, do not apply to this work. All AI-assisted outputs were reviewed by the authors. For AI-assisted code, we inspected the generated code, verified that it matched the intended implementation, and checked that the resulting experimental outputs were consistent with our experimental design. For the theory, the main results, arguments, and proof strategies were developed by the authors; generative AI was used to help write out and polish proof steps, check mathematical rigor, and flag potential gaps or unclear steps. Any AI-assisted mathematical content was independently verified by the authors before being included. The authors take full responsibility for the final content of this work, including all text, claims, code, mathematical arguments, and other artifacts produced with the assistance of generative AI.

### REPRODUCIBILITY STATEMENT

We provide the complete source code as supplementary material alongside this submission. The code implements BMCR, all baselines (FA, GS, SAM, BMCR-MM, SFI, FPE, Shuffle, FOUL, and MIUL), the original and retrain-from-scratch training pipelines, and all evaluation metrics (RIPV, TRI, EI, RASI, SRI, SDI, MIA, and the fairness metrics), together with configuration files and scripts for reproducing the reported experiments. All seven datasets are publicly available from OpenML Feurer et al. (2021); their descriptions, statistical summaries, and SHAP-based feature-importance rankings are given in Appendices G, H, and O. Architectures, hyperparameters, epoch budgets, checkpoint-selection rules, and the replacement ranges are specified in Section 5 and Appendix L, with training-loss curves in Appendix K. The data-generating process of the synthetic dependence study is given in Appendix E, and the hardware and wall-clock measurement protocol are given in Appendix S. All experiments were run on a single NVIDIA V100 GPU and repeated three times; we report means and standard deviations. Complete proofs of all theoretical results, with their assumptions stated explicitly, are provided in the Appendix.

## REFERENCES

Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.

Umit Yigit Basaran, Sk Miraj Ahmed, Amit Roy-Chowdhury, and Basak Guler. A certified unlearning approach without access to source data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8lt5776GLB.

Thomas Baumhauer, Pascal Schöttle, and Matthias Zeppelzauer. Machine unlearning: Linear filtration for logit-based classifiers. *Machine Learning*, 111(9):3203–3226, 2022.

Lucas 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)*, pp. 141–159. IEEE, 2021.

Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. *arXiv preprint arXiv:2409.11219*, 2024.

Zhi Chen, Melvyn Sim, and Huan Xu. Distributionally robust optimization with infinitely constrained ambiguity sets. *Operations Research*, 67(5):1328–1344, 2019.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023.

Thomas De Min, Massimiliano Mancini, Stéphane Lathuilière, Subhankar Roy, and Elisa Ricci. Unlearning personal data from a single image. *arXiv preprint arXiv:2407.12069*, 2024.

European Data Protection Board. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 2024. Adopted 17 December 2024, para. 102(d).

Federal Trade Commission. In the matter of Everalbum, Inc., docket no. C-4743, decision and order, 2021. Issued 6 May 2021.

Matthias Feurer, Jan N van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Mueller, Joaquin Vanschoren, and Frank Hutter. Openml-python: an extensible python api for openml. *Journal of Machine Learning Research*, 22(100):1–5, 2021. URL http://jmlr.org/papers/v22/19-920.html.

Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. 2019.

Tao Guo, Song Guo, Jiewei Zhang, Wenchao Xu, and Junxiao Wang. Efficient attribute unlearning: Towards selective removal of input attributes from feature representations, 2022. URL https://arxiv.org/abs/2202.13295.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Ioannis Mavrothalassitis, Pol Puigdemont, Noam Levi, and Volkan Cevher. Ascent fails to forget. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

John T McCoy and Lidia Auret. Machine learning applications in minerals processing: A review. *Minerals Engineering*, 132:95–109, 2019.

Saemi Moon, Seunghyuk Cho, and Dongwoo Kim. Feature unlearning for pre-trained gans and vaes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21420–21428, 2024.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models . In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, Los Alamitos, CA, USA, May 2017. IEEE Computer Society. doi: 10.1109/SP.2017.41. URL https://doi.ieeecomputersociety.org/10.1109/SP.2017.41.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2023.

U.S. District Court for the Northern District of California. United states v. Kurbo Inc. and WW International, Inc., no. 3:22-cv-00946-TSH, stipulated order for permanent injunction, civil penalty judgment, and other relief, 2022. 3 March 2022.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *Proc. of the 30th Network and Distributed System Security (NDSS)*, 2023.

Heng Xu, Tianqing Zhu, Wanlei Zhou, and Wei Zhao. Don't forget too much: Towards machine unlearning on feature level. *arXiv preprint arXiv:2406.10951*, 2024.

Yue Yang, Jinhao Li, and Hao Wang. Feature unlearning: Theoretical foundations and practical applications with shuffling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=JMtVNZ8Np7.

# A PLAIN-LANGUAGE GLOSSARY

Table 4: Key terms, in the order they appear.

| Term | Meaning |
|---|---|
| Selective feature revocation | Access to one or more features is withdrawn, while the retained features $x_{-i}$ and labels $y$ remain usable. |
| Replacement distribution $q$ | The distribution from which BMCR draws substitute values for the revoked feature during fine-tuning. |
| Replacement support $\mathrm{supp}(q)$ | The range of values $q$ can produce; every guarantee is stated over this range. |
| Residual dependence $D_q(f)$ | Average variance of the prediction when only the revoked coordinate is resampled from $q$; zero means the prediction ignores that coordinate on $\mathrm{supp}(q)$. |
| RIPV | The Monte Carlo estimate of $D_q(f)$ on an audit set; the empirical certificate. |
| Functional equivalence | The edited model's predictions are close to those of a model retrained without the feature; no claim is made about parameters. |
| Bounded effective logit $\widetilde{f}$ | The classifier logit squashed to $[-K, K]$ (e.g., $K\tanh(f/K)$), on which the classification certificate holds. |
| Direct vs. proxy dependence | Direct: the model's output changes when $x_i$ changes. Proxy: information about $x_i$ contained in correlated retained features; also present in the retrained oracle. |

# B Access Model and Threat Model

**Who holds what.** The model owner holds the trained model $f_{\theta_0}$, the retained records $(x_{-i}, y)$, and the deployed preprocessing pipeline. After revocation, the owner may no longer read, store, or process the per-example values of $x_i$. Typical cases include removing a medical attribute from a credit model while keeping income and repayment history, or restricting employee medical information while keeping ordinary personnel data; in both, different fields of the same record have different legal bases, retention periods, or ownership Voigt & Von dem Bussche (2017); Act (1996). Regulators have also required that models trained on unlawfully held data be deleted or remediated European Data Protection Board (2024); Federal Trade Commission (2021); U.S. District Court for the Northern District of California (2022), which motivates both a certificate and minimal access to the revoked data.

**Why a declared support is not re-access.** BMCR never reads per-example values of $x_i$. It needs only a range to sample from. Such ranges are often public or schema-level (binary indicators, ages within legal limits, percentages, sensor ranges), are routinely treated as support information in robust optimization Goh & Sim (2010); Chen et al. (2019), and can be declared by the requester without consulting the data. We distinguish this (tier A1) from statistics that must be computed from the revoked column (tier A2: mean, standard deviation, empirical min/max) and from per-example access (tier A3); Table 5 in Appendix B.1 lists what each method uses. When the model consumes standardized inputs, BMCR-$\varepsilon$ samples $[-\varepsilon, \varepsilon]$ in the model's own input space and computes nothing new (tier A0); the corresponding raw-unit interval is $[\mu_i - \varepsilon\sigma_i, \mu_i + \varepsilon\sigma_i]$, where $(\mu_i, \sigma_i)$ are the scaler parameters already embedded in the deployed pipeline. If those parameters are themselves deemed sensitive, the practitioner should use a declared support (A1).

**What is not assumed.** We do not address the fully data-free setting in which $(x_{-i}, y)$ is also unavailable; there, neither retraining nor BMCR applies. We also do not claim to remove information about $x_i$ contained in $x_{-i}$ (Remark 4.7).

## B.1 Taxonomy of BMCR and Baselines

Table 5 refines the access tiers of Table 1 by listing, for each method, exactly which information about the removed feature $x_i$ it uses during unlearning.

Table 5: **Baseline taxonomy vs. information required about the removed feature $x_i$.** Tiers follow Table 1: A0, nothing beyond the deployed pipeline; A1, a declared or public support; A2, statistics computed from the revoked column; A3, per-example values. "Scaler" means the standardization parameters already embedded in the deployed model's preprocessing; no new statistic is computed during unlearning.

| Method | Tier | What it does (for removed feature $x_i$) | min/max | mean | std |
|---|---|---|---|---|---|
| **BMCR-$\varepsilon$ (ours)** | A0 | Uniform replacement $T \sim \mathrm{Unif}[-\varepsilon, \varepsilon]$ in the standardized input space (raw units: $[\mu_i - \varepsilon\sigma_i, \mu_i + \varepsilon\sigma_i]$ via the scaler). | No | Scaler | Scaler |
| **BMCR (z-score Gaussian)** | A0 | Sample in normalized space, $T \sim \mathcal{N}(0, \Delta^2)$ (mean 0 by construction; Lemma D.14). | No | No | No |
| **BMCR-full (ours)** | A1 | Uniform replacement over a declared or public support, $T \sim \mathrm{Unif}[\ell_i, u_i]$. | Declared | No | No |
| **BMCR-MM (ours)** | A2 | Uniform replacement over the empirical range, $T \sim \mathrm{Unif}[\min(x_i), \max(x_i)]$ (A1 if these bounds are public). | Yes | No | No |
| BMCR (raw-space mean-centered) | A2 | $T \sim \mathrm{Unif}[\mu_i - \rho_i, \mu_i + \rho_i]$ with $\mu_i$ computed from the training data. | No | Yes | No |
| FA | A0 | Set $x_i$ to a constant in the standardized space ($-1$, 0, or 1). | No | No | No |
| SFI | A0 | Zero and freeze the input weights of $x_i$, then fine-tune on $(x_{-i}, y)$. | No | No | No |
| FPE | A0 | Rebuild the network without the $x_i$ input, copy compatible weights, fine-tune on $(x_{-i}, y)$. | No | No | No |
| GS | A2 | Gaussian replacement with training-data mean and std, $T \sim \mathcal{N}(\mu, \sigma^2)$. | No | Yes | Yes |
| SAM | A2 | Sampling around the training-data mean of $x_i$. | No | Yes | No |
| FOUL | A3 | Uses gradient differences between perturbed and original inputs on the training set. | — | — | — |
| MIUL | A3 | Reduces mutual information between target feature and label; trains extra networks on original data. | — | — | — |
| Shuffle | A3 | Shuffles $x_i$ across examples and fine-tunes on the shuffled dataset. | — | — | — |

# C Relation to Random Labeling, Shuffle, and Structural Masking

**Random labeling (RL).** RL minimizes $\mathbb{E}[\ell(f(X), \widetilde{Y})]$ on forget samples with random labels $\widetilde{Y}$. Its population optimum on those samples is $\mathbb{E}[\widetilde{Y}]$, which carries no task information, and it does not target a specific input coordinate. BMCR minimizes $\mathbb{E}_{X_{-i},Y}\mathbb{E}_{T\sim q}[\ell(f(X_{-i}, T), Y)]$ with *true* labels. Under squared loss (Theorem 4.2),

$$R_q(f) - R_q(f^\star) = \mathbb{E}\big[(\bar{f}(X_{-i}) - \mathbb{E}[Y \mid X_{-i}])^2\big] + D_q(f),$$

so the optimum is the retained-feature Bayes predictor, exactly the population target of retraining without $x_i$, and any excess risk bounds the residual dependence on $x_i$.

**Shuffle.** Shuffle Yang et al. (2025) permutes observed values of $x_i$ and therefore requires them (tier A3). BMCR replaces the permutation distribution by a declared $q$, which removes the need for those values and makes the certified scope explicit through $\mathrm{supp}(q)$.

**Input-space randomization in general.** If "input-space RL" is taken to mean replacing $x_i$ by noise, then FA, GS, SAM, and BMCR are all instances. Our contribution is not the replacement operation itself but (i) the support-controlled formulation under partial access, (ii) the certificate

and prediction-level connection to retraining, and (iii) the finite-checkpoint, attribute-inference, and sequential extensions. Empirically, BMCR outperforms the other instances (Sec. 5).

**Structural masking.** Zeroing and freezing the input weights of $x_i$ (SFI), or rebuilding the network without that input (FPE), removes the input pathway without using any information about $x_i$. However, the downstream layers were fitted jointly with $x_i$; after masking they receive a shifted input distribution and must be re-fitted to $x_{-i}$. BMCR performs this re-fitting directly, and the risk decomposition above shows it targets the correct retained-feature optimum. Results are in Appendix W.

# D PROOFS

## D.1 DEFINITIONS AND ASSUMPTIONS REFERENCED IN THE MAIN TEXT

**Definition D.1** (Coordinate-level feature unlearning over an interval)**.** Let $\mathcal{I}_i \subset \mathbb{R}$ be a target unlearning interval. A predictor $f$ unlearns coordinate $i$ over $\mathcal{I}_i$ if, for $\mathbb{P}_{x_{-i}}$-almost every $x_{-i}$, $f(x_{-i}, t) = f(x_{-i}, t')$ for all $t, t' \in \mathcal{I}_i$. More generally, $f$ $(\tau, \alpha)$-approximately unlearns coordinate $i$ over $\mathcal{I}_i$ if $\mathbb{P}_{x_{-i}}\big(\sup_{t,t' \in \mathcal{I}_i} |f(x_{-i}, t) - f(x_{-i}, t')| \le \tau\big) \ge 1 - \alpha$.

**Assumption D.2.** For all $\theta \in \mathcal{N}$ and all $x_{-i}$ in the support, $f_\theta(x_{-i}, T)$ is square-integrable under $T \sim q$ and satisfies $\mathbb{E}_{T\sim q}[f_\theta(x_{-i}, T)^2] \le C_2$ and $\mathbb{E}_{T\sim q}[(f_\theta(x_{-i}, T) - \mathbb{E}_{T\sim q}[f_\theta(x_{-i}, T)])^4] \le C_4$. Moreover, $T$ is independent of $(x_{-i}, y)$, with mean $\mu_q$ and variance $\sigma_q^2 > 0$.

**Assumption D.3.** There exist $\bar{\theta} \in \mathcal{N}$ and a measurable $g : \mathcal{X}_{-i} \to \mathbb{R}$ such that $f_{\bar{\theta}}(x_{-i}, t) = g(x_{-i})$ for all $t \in \mathrm{supp}(q)$ and $\mathbb{E}[(g(x_{-i}) - y)^2] < \infty$; i.e., the architecture can ignore the replaced feature.

## D.2 PROOF FOR LEMMA 4.1

*Proof.* Fix an arbitrary pair $(x_{-i}, y)$ in the support. The only randomness below comes from the replacement draw $T \sim q$. Define the scalar random variable

$$Z \;:=\; f(x_{-i}, T).$$

By Assumption D.2 (square-integrability), $\mathbb{E}_T[Z^2] < \infty$ and hence $\mathbb{E}_T[Z]$ and $Var_T(Z)$ are well-defined and finite. By the definition of the $q$-averaged predictor,

$$\bar{f}(x_{-i}) \;:=\; \mathbb{E}_{T\sim q}[f(x_{-i}, T)] \;=\; \mathbb{E}_T[Z].$$

We now expand the squared error around the mean $\bar{f}(x_{-i})$. Write

$$Z - y \;=\; \big(Z - \bar{f}(x_{-i})\big) \;+\; \big(\bar{f}(x_{-i}) - y\big).$$

Squaring both sides gives

$$(Z - y)^2 = \big(Z - \bar{f}(x_{-i})\big)^2 + \big(\bar{f}(x_{-i}) - y\big)^2 + 2\big(Z - \bar{f}(x_{-i})\big)\big(\bar{f}(x_{-i}) - y\big).$$

Take expectation with respect to $T \sim q$ (equivalently, over $Z$) and use linearity of expectation:

$$\mathbb{E}_T[(Z - y)^2] = \mathbb{E}_T\Big[\big(Z - \bar{f}(x_{-i})\big)^2\Big] + \big(\bar{f}(x_{-i}) - y\big)^2 + 2\big(\bar{f}(x_{-i}) - y\big)\mathbb{E}_T\big[Z - \bar{f}(x_{-i})\big].$$

The final term vanishes because $\bar{f}(x_{-i}) = \mathbb{E}_T[Z]$, hence

$$\mathbb{E}_T\big[Z - \bar{f}(x_{-i})\big] = \mathbb{E}_T[Z] - \bar{f}(x_{-i}) = 0.$$

Therefore,

$$\mathbb{E}_T[(Z - y)^2] = \mathbb{E}_T\Big[\big(Z - \bar{f}(x_{-i})\big)^2\Big] + \big(\bar{f}(x_{-i}) - y\big)^2. \tag{7}$$

Finally, recognize the first term as the conditional variance of $Z$ under $T \sim q$:

$$\mathbb{E}_T\Big[\big(Z - \bar{f}(x_{-i})\big)^2\Big] = \mathbb{E}_T\Big[\big(Z - \mathbb{E}_T[Z]\big)^2\Big] = Var_T(Z) = Var_{T\sim q}\big(f(x_{-i}, T)\big).$$

Substituting this identity into (7) yields, for every fixed $(x_{-i}, y)$,

$$\mathbb{E}_{T\sim q}\big[(f(x_{-i}, T) - y)^2\big] = Var_{T\sim q}\big(f(x_{-i}, T)\big) + \big(\bar{f}(x_{-i}) - y\big)^2.$$

Taking expectation over $(x_{-i}, y)$ on both sides and using that $T$ is independent of $(x_{-i}, y)$ (Assumption D.2) gives

$$\mathbb{E}_{(x_{-i},y)}\mathbb{E}_{T\sim q}\big[(f(x_{-i}, T) - y)^2\big] = \mathbb{E}_{(x_{-i},y)}\big[(\bar{f}(x_{-i}) - y)^2\big] + \mathbb{E}_{x_{-i}}\big[Var_{T\sim q}\big(f(x_{-i}, T)\big)\big],$$

which is exactly (6). □

### D.3 Proof for Theorem 4.2

*Proof.* Recall the definitions

$$\bar{f}(x_{-i}) := \mathbb{E}_{T\sim q}[f(x_{-i}, T)], \qquad D_q(f) := \mathbb{E}_{x_{-i}}\big[Var_{T\sim q}\big(f(x_{-i}, T)\big)\big],$$

and the replacement risk under squared loss

$$R_q(f) := \mathbb{E}_{(x_{-i},y)}\mathbb{E}_{T\sim q}\big[(f(x_{-i}, T) - y)^2\big].$$

All expectations below are assumed finite by the moment condition in Assumption D.2.

By Lemma 4.1, we have the identity

$$R_q(f) = \mathbb{E}_{(x_{-i},y)}\big[(\bar{f}(x_{-i}) - y)^2\big] + D_q(f). \tag{8}$$

Since $D_q(f)$ is an expectation of a variance, it is nonnegative:

$$D_q(f) \geq 0. \tag{9}$$

Thus, for any predictor $f$,

$$R_q(f) \;\geq\; \mathbb{E}\big[(\bar{f}(x_{-i}) - y)^2\big],$$

with equality if and only if $D_q(f) = 0$, i.e., if and only if $Var_{T\sim q}(f(x_{-i}, T)) = 0$ for almost every $x_{-i}$, which means $f(x_{-i}, T)$ is $q$-almost surely constant in $T$. If $t \mapsto f(x_{-i}, t)$ is continuous on $\mathrm{supp}(q)$, then this $q$-almost-sure constancy extends to pointwise constancy on $\mathrm{supp}(q)$.

Fix $f$. Define a $t$-invariant predictor $\tilde{f}$ by

$$\tilde{f}(x_{-i}, t) := \bar{f}(x_{-i}) \qquad \forall t \in \mathrm{supp}(q).$$

Then $\tilde{f}$ is $t$-invariant and satisfies

$$\overline{\tilde{f}}(x_{-i}) = \mathbb{E}_{T\sim q}[\tilde{f}(x_{-i}, T)] = \mathbb{E}_{T\sim q}[\bar{f}(x_{-i})] = \bar{f}(x_{-i}),$$

and also $D_q(\tilde{f}) = 0$ because $\tilde{f}(x_{-i}, T)$ is almost surely constant in $T$. Applying (8) to both $f$ and $\tilde{f}$ yields

$$R_q(f) = \mathbb{E}[(\bar{f} - y)^2] + D_q(f) \;\geq\; \mathbb{E}[(\bar{f} - y)^2] = R_q(\tilde{f}).$$

Therefore, for any $f$ there exists a $t$-invariant predictor $\tilde{f}$ with no larger replacement risk. Consequently, *a global minimizer of $R_q$ can be chosen to be $t$-invariant*, and minimizing $R_q$ reduces to minimizing the $x_{-i}$-only risk over functions $h : \mathcal{X}_{-i} \to \mathbb{R}$:

$$\inf_f R_q(f) = \inf_h \; \mathbb{E}_{(x_{-i},y)}\big[(h(x_{-i}) - y)^2\big]. \tag{10}$$

Define the Bayes regressor

$$h^\star(x_{-i}) := \mathbb{E}[y \mid x_{-i}].$$

It is well known (and we show below for completeness) that $h^\star$ minimizes $\mathbb{E}[(h(x_{-i}) - y)^2]$ over all measurable $h$ with finite second moment. Indeed, fix $x_{-i}$ and consider the conditional risk

$$\mathcal{L}(h; x_{-i}) := \mathbb{E}\big[(h(x_{-i}) - y)^2 \mid x_{-i}\big],$$

where $h(x_{-i})$ is a scalar (deterministic) value when conditioning on $x_{-i}$. Expand around $h^\star(x_{-i})$:

$$\begin{aligned}
\mathbb{E}\big[(h(x_{-i}) - y)^2 \mid x_{-i}\big] &= \mathbb{E}\big[(h(x_{-i}) - h^\star(x_{-i}) + h^\star(x_{-i}) - y)^2 \mid x_{-i}\big] \\
&= (h(x_{-i}) - h^\star(x_{-i}))^2 + \mathbb{E}\big[(h^\star(x_{-i}) - y)^2 \mid x_{-i}\big] \\
&\quad + 2(h(x_{-i}) - h^\star(x_{-i}))\mathbb{E}\big[h^\star(x_{-i}) - y \mid x_{-i}\big].
\end{aligned}$$

The cross term vanishes because $\mathbb{E}[y \mid x_{-i}] = h^\star(x_{-i})$, hence $\mathbb{E}[h^\star(x_{-i}) - y \mid x_{-i}] = 0$. Therefore,

$$\mathbb{E}\big[(h(x_{-i}) - y)^2 \mid x_{-i}\big] = \mathbb{E}\big[(h^\star(x_{-i}) - y)^2 \mid x_{-i}\big] + (h(x_{-i}) - h^\star(x_{-i}))^2, \tag{11}$$

which shows $h^\star$ minimizes the conditional risk pointwise in $x_{-i}$. Taking expectation over $x_{-i}$ yields that $h^\star$ minimizes $\mathbb{E}[(h(x_{-i}) - y)^2]$ globally, and the minimizer is unique up to a.s. equality (because the excess term $(h - h^\star)^2$ must vanish a.s. at optimum). Define the $t$-invariant predictor

$f^\star(x_{-i}, t) \equiv h^\star(x_{-i})$ on $\mathrm{supp}(q)$. This establishes part (1): $f^\star$ is a global minimizer of $R_q$ and $D_q(f^\star) = 0$.

Apply Lemma 4.1 to both $f$ and $f^\star$:

$$R_q(f) - R_q(f^\star) = \Big(\mathbb{E}[(\bar{f} - y)^2] - \mathbb{E}[(\overline{f^\star} - y)^2]\Big) + \Big(D_q(f) - D_q(f^\star)\Big).$$

Since $f^\star$ is $t$-invariant, $\overline{f^\star}(x_{-i}) = h^\star(x_{-i})$ and $D_q(f^\star) = 0$. Thus,

$$R_q(f) - R_q(f^\star) = \Big(\mathbb{E}[(\bar{f} - y)^2] - \mathbb{E}[(h^\star - y)^2]\Big) + D_q(f). \tag{12}$$

Now apply the conditional decomposition (11) with $h = \bar{f}$ and then take expectation over $x_{-i}$:

$$\mathbb{E}[(\bar{f} - y)^2] - \mathbb{E}[(h^\star - y)^2] = \mathbb{E}\big[(\bar{f} - h^\star)^2\big].$$

Substituting into (12) yields the claimed exact decomposition

$$R_q(f) - R_q(f^\star) = \mathbb{E}_{x_{-i}}\big[(\bar{f}(x_{-i}) - h^\star(x_{-i}))^2\big] + D_q(f),$$

Because $f^\star$ is a global minimizer, $\inf_g R_q(g) = R_q(f^\star)$. Dropping the nonnegative term $\mathbb{E}[(\bar{f} - h^\star)^2] \geq 0$ gives

$$D_q(f) \leq R_q(f) - R_q(f^\star) = R_q(f) - \inf_g R_q(g),$$

. This completes the proof. □

### D.4 Variance lower bound for strongly convex losses

**Lemma D.4.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be twice differentiable and satisfy $\phi''(z) \geq m > 0$ for all $z$ in an interval containing the support of a square-integrable random variable $Z$. Then*

$$\mathbb{E}[\phi(Z)] \geq \phi(\mathbb{E}[Z]) + \frac{m}{2}\mathrm{Var}(Z). \tag{13}$$

### D.5 Proof for Lemma D.4

*Proof.* Let $Z$ be a square-integrable random variable, so $\mathbb{E}[Z^2] < \infty$. In particular, $\mu := \mathbb{E}[Z]$ is finite and $Var(Z) = \mathbb{E}[(Z - \mu)^2]$ is well-defined. By assumption, the support of $Z$ is contained in an interval $I \subset \mathbb{R}$ on which $\phi$ is twice differentiable and satisfies the uniform curvature lower bound

$$\phi''(z) \;\geq\; m \;>\; 0 \qquad \forall z \in I. \tag{14}$$

Fix $\mu \in I$ and any $z \in I$. Consider the scalar function

$$g(t) := \phi\big(\mu + t(z - \mu)\big), \qquad t \in [0, 1].$$

Since $\phi$ is twice differentiable on $I$, $g$ is twice differentiable on $[0, 1]$ with

$$g'(t) = \phi'\big(\mu + t(z - \mu)\big)(z - \mu), \qquad g''(t) = \phi''\big(\mu + t(z - \mu)\big)(z - \mu)^2.$$

By the fundamental theorem of calculus,

$$\phi(z) - \phi(\mu) = g(1) - g(0) = \int_0^1 g'(t)\, dt. \tag{15}$$

Decompose $g'(t)$ around $g'(0)$:

$$g'(t) = g'(0) + \int_0^t g''(s)\, ds.$$

Substitute into (15) and change the order of integration:

$$\begin{aligned}\phi(z) - \phi(\mu) &= \int_0^1 g'(0)\, dt + \int_0^1 \int_0^t g''(s)\, ds\, dt \\ &= g'(0) + \int_0^1 (1 - s)\, g''(s)\, ds.\end{aligned}$$

Now plug in $g'(0) = \phi'(\mu)(z-\mu)$ and $g''(s) = \phi''(\mu + s(z-\mu))(z-\mu)^2$ to obtain the *integral remainder representation*

$$\phi(z) = \phi(\mu) + \phi'(\mu)(z-\mu) + (z-\mu)^2 \int_0^1 (1-s)\,\phi''\big(\mu + s(z-\mu)\big)\,ds. \tag{16}$$

Using the curvature bound (14), the integral term is bounded below by

$$\int_0^1 (1-s)\,\phi''(\mu + s(z-\mu))\,ds \;\geq\; \int_0^1 (1-s)\,m\,ds = m\int_0^1 (1-s)\,ds = \frac{m}{2}.$$

Substituting back into (16) yields the pointwise inequality

$$\phi(z) \;\geq\; \phi(\mu) + \phi'(\mu)(z-\mu) + \frac{m}{2}(z-\mu)^2, \qquad \forall z \in I. \tag{17}$$

Apply (17) with $z = Z(\omega)$ (for each outcome $\omega$). Since $Z$ takes values in $I$ almost surely, the inequality holds almost surely:

$$\phi(Z) \;\geq\; \phi(\mu) + \phi'(\mu)(Z-\mu) + \frac{m}{2}(Z-\mu)^2.$$

Take expectation on both sides. By linearity of expectation and $\mathbb{E}[Z-\mu] = 0$,

$$\begin{aligned}\mathbb{E}[\phi(Z)] &\geq \phi(\mu) + \phi'(\mu)\mathbb{E}[Z-\mu] + \frac{m}{2}\mathbb{E}[(Z-\mu)^2] \\ &= \phi(\mu) + \frac{m}{2}Var(Z).\end{aligned}$$

Recalling $\mu = \mathbb{E}[Z]$ gives exactly

$$\mathbb{E}[\phi(Z)] \;\geq\; \phi(\mathbb{E}[Z]) + \frac{m}{2}Var(Z),$$

which is (13). □

### D.6 Proof for Proposition D.13

*Detailed proof.* Recall the replacement risk under squared loss:

$$R_q(f) := \mathbb{E}_{(x_{-i},y)}\mathbb{E}_{T\sim q}\Big[(f(x_{-i},T)-y)^2\Big], \qquad \bar{f}(x_{-i}) := \mathbb{E}_{T\sim q}[f(x_{-i},T)], \qquad D_q(f) := \mathbb{E}_{x_{-i}}\mathrm{Var}_{T\sim q}(f(x_{-i},T)).$$

By Lemma 4.1 (exact MSE decomposition),

$$R_q(f) = \mathbb{E}\big[(\bar{f}(x_{-i}) - y)^2\big] + D_q(f). \tag{18}$$

Since $D_q(f) \geq 0$ by definition of variance, (18) immediately implies

$$R_q(f) \;\geq\; \mathbb{E}\big[(\bar{f}(x_{-i}) - y)^2\big]. \tag{19}$$

Let $h : \mathcal{X}_{-i} \to \mathbb{R}$ be any measurable function, and define a $t$-invariant predictor

$$g(x_{-i}, t) \equiv h(x_{-i}) \qquad \text{for all } t \in \mathrm{supp}(q).$$

Then for every $x_{-i}$,

$$\bar{g}(x_{-i}) = \mathbb{E}_{T\sim q}[g(x_{-i},T)] = \mathbb{E}_{T\sim q}[h(x_{-i})] = h(x_{-i}),$$

and

$$\mathrm{Var}_{T\sim q}(g(x_{-i},T)) = \mathrm{Var}_{T\sim q}(h(x_{-i})) = 0,$$

hence $D_q(g) = \mathbb{E}_{x_{-i}}[0] = 0$. Substituting into (18) yields

$$R_q(g) = \mathbb{E}\big[(h(x_{-i}) - y)^2\big]. \tag{20}$$

Assume the hypothesis class $\mathcal{H}$ contains all $t$-invariant predictors of the form $g(x_{-i}, t) \equiv h(x_{-i})$ for measurable $h$. Then combining (19) and (20) gives the two-sided bound

$$\inf_{f\in\mathcal{H}} R_q(f) \;\geq\; \inf_{f\in\mathcal{H}} \mathbb{E}\big[(\bar{f}(x_{-i}) - y)^2\big] \;\geq\; \inf_{h:\mathcal{X}_{-i}\to\mathbb{R}} \mathbb{E}\big[(h(x_{-i}) - y)^2\big],$$

while the reverse inequality holds because for every measurable $h$, the corresponding $t$-invariant $g \in \mathcal{H}$ satisfies $R_q(g) = \mathbb{E}[(h(x_{-i}) - y)^2]$ by (20). Therefore,

$$\inf_{f \in \mathcal{H}} R_q(f) = \inf_{h: \mathcal{X}_{-i} \to \mathbb{R}} \mathbb{E}\big[(h(x_{-i}) - y)^2\big]. \tag{21}$$

Define the regression functional

$$\mathcal{J}(h) := \mathbb{E}\big[(h(x_{-i}) - y)^2\big].$$

By the law of total expectation, we can write

$$\mathcal{J}(h) = \mathbb{E}_{x_{-i}}\Big[\mathbb{E}\big[(h(x_{-i}) - y)^2 \mid x_{-i}\big]\Big].$$

For a fixed $x_{-i}$, the inner term is a one-dimensional quadratic in the scalar $a := h(x_{-i})$:

$$\mathbb{E}\big[(a - y)^2 \mid x_{-i}\big] = (a - \mathbb{E}[y \mid x_{-i}])^2 + \mathrm{Var}(y \mid x_{-i}),$$

where we used the identity $\mathbb{E}[(a - y)^2] = (a - \mathbb{E}[y])^2 + \mathrm{Var}(y)$ applied conditionally. Since $\mathrm{Var}(y \mid x_{-i})$ does not depend on $a$, the unique minimizer in $a$ is $a^\star(x_{-i}) = \mathbb{E}[y \mid x_{-i}]$. Thus the (a.s.) unique minimizer of $\mathcal{J}(h)$ over measurable $h$ is

$$h^\star(x_{-i}) = \mathbb{E}[y \mid x_{-i}].$$

Combining with (21) proves the claim. □

### D.7 Remark on the role of centering.

The invariance certificate is support-based rather than mean-based. The exact MSE decomposition and the resulting bound on $D_q(f)$ hold for any replacement distribution $q$ that is sampled independently of $(x_{-i}, y)$ and has the required finite moments. Centering the replacement distribution at a reference value such as the training mean is a practical design choice: it avoids introducing a systematic numerical shift in the replaced inputs and often stabilizes fine-tuning. However, the domain over which unlearning is certified is determined by $\mathrm{supp}(q)$, not by the mean of $q$. Therefore, full-domain coordinate-level unlearning requires the replacement support to cover the full target feature domain, whereas mean-centered replacement certifies unlearning over the interval covered by its chosen radius.

### D.8 Proof of Theorem 4.3

**Corollary D.5** (Original-input functional equivalence under Lipschitzness)**.** *Let the assumptions of Theorem 4.3 hold. Assume $q = \mathrm{Unif}[a, b]$, $X_i \in [a, b]$ almost surely, and for $P_Z$-almost every $Z$, the map $t \mapsto \widehat{f}(Z, t)$ is L-Lipschitz on $[a, b]$. Then, for any $\alpha_1, \alpha_2, \alpha_3 \in (0, 1)$, with probability at least $1 - \alpha_1 - \alpha_2 - \alpha_3$ over $Z$,*

$$\sup_{t \in [a,b]} \left|\widehat{f}(Z, t) - \widehat{g}(Z)\right| \le \left(\frac{8L(b-a)\gamma}{\alpha_1}\right)^{1/3} + \sqrt{\frac{\gamma}{\alpha_2}} + \sqrt{\frac{\eta}{\alpha_3}}.$$

*In particular, setting $\alpha_1 = \alpha_2 = \alpha_3 = \delta/3$ gives*

$$\mathbb{P}_{(X,Y)}\left(\left|\widehat{f}(X_{-i}, X_i) - \widehat{g}(X_{-i})\right| > \epsilon_\delta\right) \le \delta,$$

*where*

$$\epsilon_\delta = \left(\frac{24L(b-a)\gamma}{\delta}\right)^{1/3} + \sqrt{\frac{3\gamma}{\delta}} + \sqrt{\frac{3\eta}{\delta}}.$$

*Proof.* Let $Z := X_{-i}$. Throughout the proof, all expectations are assumed finite; in particular, $Y$ is square-integrable. Recall that

$$R_q(f) = \mathbb{E}_{Z,Y,T}\left[(f(Z, T) - Y)^2\right],$$

where $T \sim q$ is sampled independently of $(Z, Y)$, and

$$R_{-i}(g) = \mathbb{E}_{Z,Y}\left[(g(Z) - Y)^2\right].$$

Let

$$h^\star(Z) := \mathbb{E}[Y \mid Z]$$

be the Bayes predictor using only the retained features.

We first show that the excess BMCR replacement risk is exactly the squared $L_2(P_Z \otimes q)$ distance to $h^\star$. Since $T$ is independent of $(Z, Y)$, we have

$$\mathbb{E}[Y \mid Z, T] = \mathbb{E}[Y \mid Z] = h^\star(Z).$$

Therefore,

$$\mathbb{E}[h^\star(Z) - Y \mid Z, T] = 0.$$

For any predictor $f$, write

$$f(Z,T) - Y = (f(Z,T) - h^\star(Z)) + (h^\star(Z) - Y).$$

Squaring and taking expectations gives

$$\begin{aligned} R_q(f) &= \mathbb{E}_{Z,Y,T}\left[(f(Z,T) - Y)^2\right] \\ &= \mathbb{E}_{Z,T}\left[(f(Z,T) - h^\star(Z))^2\right] + \mathbb{E}_{Z,Y}\left[(h^\star(Z) - Y)^2\right] \\ &\quad + 2\,\mathbb{E}_{Z,Y,T}\left[(f(Z,T) - h^\star(Z))\,(h^\star(Z) - Y)\right]. \end{aligned}$$

The cross term is zero because

$$\begin{aligned} &\mathbb{E}_{Z,Y,T}\left[(f(Z,T) - h^\star(Z))\,(h^\star(Z) - Y)\right] \\ &\quad = \mathbb{E}_{Z,T}\left[(f(Z,T) - h^\star(Z))\,\mathbb{E}[h^\star(Z) - Y \mid Z, T]\right] = 0. \end{aligned}$$

Hence

$$R_q(f) = \mathbb{E}_{Z,T}\left[(f(Z,T) - h^\star(Z))^2\right] + \mathbb{E}_{Z,Y}\left[(h^\star(Z) - Y)^2\right].$$

Viewing $h^\star$ as the $t$-invariant predictor $h^\star(Z, t) \equiv h^\star(Z)$, we have

$$R_q(h^\star) = \mathbb{E}_{Z,Y}\left[(h^\star(Z) - Y)^2\right].$$

Therefore,

$$R_q(f) - R_q(h^\star) = \mathbb{E}_{Z,T}\left[(f(Z,T) - h^\star(Z))^2\right].$$

Applying this identity to the BMCR-edited predictor $\widehat{f}$ and using the assumption

$$R_q(\widehat{f}) - R_q(h^\star) \le \gamma$$

yields

$$\left\|\widehat{f}(Z,T) - h^\star(Z)\right\|_{L_2(P_Z \otimes q)} \le \sqrt{\gamma}.$$

Next, we apply the standard squared-loss Bayes decomposition to the retrained predictor. For any retained-feature predictor $g$,

$$R_{-i}(g) = \mathbb{E}_{Z,Y}\left[(g(Z) - Y)^2\right].$$

Writing

$$g(Z) - Y = (g(Z) - h^\star(Z)) + (h^\star(Z) - Y)$$

and using

$$\mathbb{E}[h^\star(Z) - Y \mid Z] = 0,$$

we obtain

$$R_{-i}(g) - R_{-i}(h^\star) = \mathbb{E}_Z\left[(g(Z) - h^\star(Z))^2\right].$$

Applying this identity to $\widehat{g}$ and using

$$R_{-i}(\widehat{g}) - R_{-i}(h^\star) \le \eta$$

gives

$$\|\widehat{g}(Z) - h^\star(Z)\|_{L_2(P_Z)} \le \sqrt{\eta}.$$

Since $\widehat{g}(Z) - h^\star(Z)$ does not depend on $T$, the same norm is also its $L_2(P_Z \otimes q)$ norm:

$$\|\widehat{g}(Z) - h^\star(Z)\|_{L_2(P_Z \otimes q)} = \|\widehat{g}(Z) - h^\star(Z)\|_{L_2(P_Z)} \le \sqrt{\eta}.$$

Finally, by the triangle inequality in $L_2(P_Z \otimes q)$,

$$\begin{aligned}\left\|\widehat{f}(Z,T) - \widehat{g}(Z)\right\|_{L_2(P_Z \otimes q)} &\le \left\|\widehat{f}(Z,T) - h^\star(Z)\right\|_{L_2(P_Z \otimes q)} \\ &\quad + \|h^\star(Z) - \widehat{g}(Z)\|_{L_2(P_Z \otimes q)} \\ &\le \sqrt{\gamma} + \sqrt{\eta}.\end{aligned}$$

This proves the claimed $L_2$ bound.

For the high-probability statement, Markov's inequality applied to the nonnegative random variable

$$\left(\widehat{f}(Z,T) - \widehat{g}(Z)\right)^2$$

gives, for every $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}_{Z,T}\left(\left|\widehat{f}(Z,T) - \widehat{g}(Z)\right| > \epsilon\right) &= \mathbb{P}_{Z,T}\left(\left(\widehat{f}(Z,T) - \widehat{g}(Z)\right)^2 > \epsilon^2\right) \\ &\le \frac{\mathbb{E}_{Z,T}\left[\left(\widehat{f}(Z,T) - \widehat{g}(Z)\right)^2\right]}{\epsilon^2} \\ &\le \frac{\left(\sqrt{\gamma} + \sqrt{\eta}\right)^2}{\epsilon^2}.\end{aligned}$$

This completes the proof. □

### D.9 Corollary about certificate

**Corollary D.6** (Full-domain coordinate unlearning with known feature bounds)**.** *Suppose the valid domain of feature $i$ is known and bounded: $\mathcal{S}_i = [\ell_i, u_i]$. Choose the full-support replacement distribution $T_i \sim q_i := \mathrm{Unif}[\ell_i, u_i]$. Then $q_i$ covers the entire feature domain: $\mathcal{S}_i = \mathrm{supp}(q_i)$. Under the assumptions of Theorem 4.2, any $\delta$-near minimizer $\hat{f}$ of the replacement risk satisfies $D_{q_i}(\hat{f}) \le \delta$. In particular, if $\delta = 0$, then for $\mathbb{P}_{x_{-i}}$-almost every $x_{-i}$, the map $t \mapsto \hat{f}(x_{-i}, t)$ is $q_i$-almost surely constant on $[\ell_i, u_i]$. If $t \mapsto \hat{f}(x_{-i}, t)$ is continuous on $[\ell_i, u_i]$, then $\hat{f}(x_{-i}, t) = \hat{f}(x_{-i}, t') \quad \forall\, t, t' \in [\ell_i, u_i]$, for $\mathbb{P}_{x_{-i}}$-almost every $x_{-i}$. Therefore, $\hat{f}$ exactly unlearns coordinate $i$ over the full bounded feature domain.*

**Corollary D.7** (Mean-centered interval certificate)**.** *Suppose a reference center $\mu_i$ is available, but the full feature domain is not known. Let $\mathcal{I}_i = [a_i, b_i]$ be any declared target interval over which coordinate $i$ should be unlearned. Define $\rho_i := \sup_{t \in \mathcal{I}_i} |t - \mu_i| = \max\{|a_i - \mu_i|,\ |b_i - \mu_i|\}$, and sample $T_i \sim q_i := \mathrm{Unif}[\mu_i - \rho_i,\ \mu_i + \rho_i]$. Then $\mathcal{I}_i \subseteq \mathrm{supp}(q_i)$. Therefore, under the assumptions of Theorem 4.2, any $\delta$-near minimizer $\hat{f}$ satisfies $D_{q_i}(\hat{f}) \le \delta$, and the average residual-dependence certificate is evaluated under a replacement distribution whose support covers $I_i$. Exact pointwise unlearning over $I_i$ follows when $\delta = 0$ and $t \mapsto \widehat{f}(x_{-i}, t)$ is continuous. For finite $\delta$, range-wide control requires an additional regularity condition such as Lipschitzness. The guarantee does not extend to feature values outside $\mathrm{supp}(q_i)$ unless the unlearning procedure is run with a replacement distribution covering those values.*

### D.10 Lemma about variation over the entire range satisfies

**Lemma D.8.** *Fix $x_{-i}$ and define $g(t) := f_\theta(x_{-i}, t)$ for $t \in [a, b]$. Assume $g$ is $L$-Lipschitz on $[a, b]$, i.e., $|g(t) - g(t')| \le L|t - t'| \quad \forall\, t, t' \in [a, b]$. Let $T \sim \mathrm{Unif}[a, b]$, and write $\mu(x_{-i}) := \mathbb{E}[g(T)], \quad V(x_{-i}) := \mathrm{Var}(g(T))$. Then the maximum deviation from the replacement-mean is bounded by $\sup_{t \in [a,b]} |g(t) - \mu(x_{-i})| \ \le\ \big(8L(b-a)\,V(x_{-i})\big)^{1/3}$, and hence the worst-case prediction variation over the entire range satisfies $\sup_{t,t' \in [a,b]} |g(t) - g(t')| \ \le\ 2\big(8L(b-a)\,V(x_{-i})\big)^{1/3}$.*

*Proof.* Let $\mu := \mathbb{E}[g(T)]$ and $M := \sup_{t\in[a,b]} |g(t) - \mu|$. Pick $t_0 \in [a, b]$ such that $|g(t_0) - \mu| \geq M - \eta$ for any $\eta > 0$. By Lipschitzness, for all $t$ with $|t - t_0| \leq \frac{M-\eta}{2L}$ we have

$$|g(t) - \mu| \;\geq\; |g(t_0) - \mu| - |g(t) - g(t_0)| \;\geq\; (M - \eta) - L|t - t_0| \;\geq\; \frac{M - \eta}{2}.$$

Therefore the set $A := \{t \in [a, b] : |g(t) - \mu| \geq \frac{M-\eta}{2}\}$ has Lebesgue measure at least $\frac{M-\eta}{2L}$, and since $T$ is uniform,

$$\mathbb{P}(T \in A) \;\geq\; \frac{M - \eta}{2L(b - a)}.$$

Using $V = \mathbb{E}\big[(g(T) - \mu)^2\big] \geq \mathbb{E}\big[(g(T) - \mu)^2 \mathbf{1}\{T \in A\}\big]$, we obtain

$$V \;\geq\; \Big(\frac{M - \eta}{2}\Big)^2 \cdot \frac{M - \eta}{2L(b - a)} \;=\; \frac{(M - \eta)^3}{8L(b - a)}.$$

Letting $\eta \downarrow 0$ yields $M^3 \leq 8L(b - a)V$, i.e., $M \leq (8L(b - a)V)^{1/3}$. Finally, for any $t, t'$, $|g(t) - g(t')| \leq |g(t) - \mu| + |g(t') - \mu| \leq 2M$, giving the second inequality. □

### D.11 Range-wide approximate unlearning for bounded effective logits

**Corollary D.9** (Range-wide approximate unlearning for bounded effective logits)**.** *Let $T_i \sim q_i = \mathrm{Unif}[a_i, b_i]$, and suppose $\mathcal{I}_i \subseteq [a_i, b_i]$. Under Assumption 4.6, let $\tilde{f} := \phi_K \circ f$ be the bounded effective logit used inside the cross-entropy loss, and let $m_K := \sigma(K)(1 - \sigma(K))$. Assume that, for every $x_{-i}$, the map $t \mapsto \tilde{f}(x_{-i}, t)$ is $L$-Lipschitz on $[a_i, b_i]$. If $\tilde{f}$ is a $\delta$-near minimizer of the replacement cross-entropy risk, namely $R_{q_i}(\tilde{f}) - \inf_g R_{q_i}(g) \leq \delta$, then for any $\alpha \in (0, 1)$, $\mathbb{P}_{x_{-i}}\left(\sup_{t,t'\in[a_i,b_i]} |\tilde{f}(x_{-i}, t) - \tilde{f}(x_{-i}, t')| \leq 2\left(\frac{16L(b_i-a_i)\delta}{m_K\alpha}\right)^{1/3}\right) \geq 1 - \alpha$. In particular, the same guarantee holds over any declared interval $\mathcal{I}_i \subseteq [a_i, b_i]$.*

### D.12 Proof for Theorem 4.5

*Proof.* Fix an arbitrary pair $(x_{-i}, y)$ in the support of the data distribution. Let $T \sim q$ and define the scalar random variable

$$Z \;:=\; \tilde{f}(x_{-i}, T).$$

Since $\phi_K$ maps into $[-K, K]$, we have $Z \in [-K, K]$ almost surely, and therefore also $\mathbb{E}_T[Z] \in [-K, K]$. By definition,

$$\mathbb{E}_T[Z] \;=\; \mathbb{E}_{T\sim q}[\tilde{f}(x_{-i}, T)] \;=\; \overline{\tilde{f}}(x_{-i}).$$

Now define $\varphi(z) := \ell(z, y)$, viewed as a univariate function of $z$ with $y$ fixed. Under Assumption 4.4, $\varphi$ is twice differentiable and $\varphi''(z) \geq m_K > 0$ for all $z \in [-K, K]$. Hence $\varphi$ is $m_K$-strongly convex on the interval $[-K, K]$.

We claim that for any random variable $Z$ supported on an interval on which $\varphi'' \geq m_K$, the following variance lower bound holds:

$$\mathbb{E}_T[\varphi(Z)] \;\geq\; \varphi(\mathbb{E}_T[Z]) \;+\; \frac{m_K}{2}\,\mathrm{Var}_T(Z). \tag{22}$$

*Derivation of* (22). Fix any $z_0 \in [-K, K]$. By the second-order Taylor theorem with integral remainder, for every $z \in [-K, K]$,

$$\begin{aligned}
\varphi(z) &= \varphi(z_0) + \varphi'(z_0)(z - z_0) + \int_{z_0}^{z} (z - u)\,\varphi''(u)\,du \\
&\geq \varphi(z_0) + \varphi'(z_0)(z - z_0) + m_K \int_{z_0}^{z} (z - u)\,du \\
&= \varphi(z_0) + \varphi'(z_0)(z - z_0) + \frac{m_K}{2}(z - z_0)^2.
\end{aligned} \tag{23}$$

Now set $z_0 := \mathbb{E}_T[Z]$ and substitute $z = Z$ into (23):

$$\varphi(Z) \;\geq\; \varphi(\mathbb{E}_T[Z]) + \varphi'(\mathbb{E}_T[Z])\big(Z - \mathbb{E}_T[Z]\big) + \frac{m_K}{2}\big(Z - \mathbb{E}_T[Z]\big)^2.$$

Taking expectation over $T$ and using $\mathbb{E}_T[Z - \mathbb{E}_T[Z]] = 0$ yields

$$\mathbb{E}_T[\varphi(Z)] \;\geq\; \varphi(\mathbb{E}_T[Z]) + \frac{m_K}{2}\,\mathbb{E}_T\big[(Z - \mathbb{E}_T[Z])^2\big] \;=\; \varphi(\mathbb{E}_T[Z]) + \frac{m_K}{2}\,\mathrm{Var}_T(Z),$$

which proves (22).

Applying (22) with $\varphi(z) = \ell(z, y)$ and $Z = \tilde{f}(x_{-i}, T)$ gives

$$\begin{aligned}\mathbb{E}_{T\sim q}\big[\ell(\tilde{f}(x_{-i}, T), y)\big] &\geq \ell\Big(\mathbb{E}_{T\sim q}[\tilde{f}(x_{-i}, T)], y\Big) + \frac{m_K}{2}\,\mathrm{Var}_{T\sim q}\big(\tilde{f}(x_{-i}, T)\big) \\ &= \ell(\overline{\tilde{f}}(x_{-i}), y) + \frac{m_K}{2}\,\mathrm{Var}_{T\sim q}\big(\tilde{f}(x_{-i}, T)\big). \qquad (24)\end{aligned}$$

Taking expectation of (24) over $(x_{-i}, y)$ gives

$$\begin{aligned}R_q(\tilde{f}) &= \mathbb{E}_{(x_{-i},y)}\mathbb{E}_{T\sim q}\big[\ell(\tilde{f}(x_{-i}, T), y)\big] \\ &\geq \mathbb{E}_{(x_{-i},y)}\big[\ell(\overline{\tilde{f}}(x_{-i}), y)\big] + \frac{m_K}{2}\,\mathbb{E}_{(x_{-i},y)}\mathrm{Var}_{T\sim q}\big(\tilde{f}(x_{-i}, T)\big).\end{aligned}$$

Since $\mathrm{Var}_{T\sim q}(\tilde{f}(x_{-i}, T))$ depends only on $x_{-i}$ (not on $y$), we can drop the conditioning on $y$:

$$\mathbb{E}_{(x_{-i},y)}\mathrm{Var}_{T\sim q}\big(\tilde{f}(x_{-i}, T)\big) = \mathbb{E}_{x_{-i}}\mathrm{Var}_{T\sim q}\big(\tilde{f}(x_{-i}, T)\big) = D_q(\tilde{f}).$$

we have

$$\frac{m_K}{2}\,D_q(\tilde{f}) \;\leq\; R_q(\tilde{f}) - \mathbb{E}_{(x_{-i},y)}\big[\ell(\overline{\tilde{f}}(x_{-i}), y)\big].$$

Next, note that the infimum in $\inf_g R_q(g)$ ranges over all measurable predictors $g : \mathcal{X}_{-i} \times \mathbb{R} \to \mathbb{R}$ and in particular includes $t$-invariant predictors of the form $g(x_{-i}, t) \equiv h(x_{-i})$. For any such $t$-invariant $g(x_{-i}, t) \equiv h(x_{-i})$, we have

$$R_q(g) = \mathbb{E}_{(x_{-i},y)}\big[\ell(h(x_{-i}), y)\big].$$

Therefore,

$$\inf_g R_q(g) \;\leq\; \inf_{h:\mathcal{X}_{-i}\to\mathbb{R}} \mathbb{E}_{(x_{-i},y)}\big[\ell(h(x_{-i}), y)\big] \;\leq\; \mathbb{E}_{(x_{-i},y)}\big[\ell(\overline{\tilde{f}}(x_{-i}), y)\big],$$

where the last inequality holds by choosing $h(x_{-i}) = \overline{\tilde{f}}(x_{-i})$.

Combining the last two displays yields

$$\frac{m_K}{2}\,D_q(\tilde{f}) \;\leq\; R_q(\tilde{f}) - \inf_g R_q(g),$$

. □

### D.13 Gaussian Replacement

**Lemma D.10** (Validity of Gaussian replacement for the MSE certificate)**.** *Let*

$$T \sim \mathcal{N}(\mu_q, \sigma_q^2), \qquad \sigma_q^2 > 0,$$

*be sampled independently of $(x_{-i}, y)$. Assume that $R_q(f) < \infty$ under squared loss and that*

$$\mathbb{E}_{T\sim q}\big[f(x_{-i}, T)^2\big] < \infty$$

*for $\mathbb{P}_{x_{-i}}$-almost every $x_{-i}$. Then the exact replacement-risk decomposition remains valid:*

$$R_q(f) = \mathbb{E}_{(x_{-i},y)}\Big[\big(\bar{f}(x_{-i}) - y\big)^2\Big] + D_q(f),$$

*where*

$$\bar{f}(x_{-i}) := \mathbb{E}_{T\sim q}[f(x_{-i}, T)]$$

*and*

$$D_q(f) := \mathbb{E}_{x_{-i}} \operatorname{Var}_{T\sim q}\big(f(x_{-i}, T)\big).$$

*Consequently, any $\delta$-near minimizer $\hat{f}$ of $R_q$ satisfies*

$$D_q(\hat{f}) \le R_q(\hat{f}) - \inf_g R_q(g) \le \delta.$$

*Thus Gaussian replacement yields the same residual-dependence certificate as any other independent replacement distribution with finite second moment.*

*Proof.* Fix $(x_{-i}, y)$ and define

$$Z := f(x_{-i}, T).$$

By the assumed square-integrability, $\mathbb{E}_T[Z]$ and $\operatorname{Var}_T(Z)$ are finite. Let

$$\bar{f}(x_{-i}) := \mathbb{E}_T[Z] = \mathbb{E}_{T\sim q}[f(x_{-i}, T)].$$

Then

$$Z - y = \big(Z - \bar{f}(x_{-i})\big) + \big(\bar{f}(x_{-i}) - y\big).$$

Squaring and taking expectation over $T \sim q$ gives

$$\begin{aligned}\mathbb{E}_T[(Z-y)^2] = \mathbb{E}_T\Big[\big(Z - \bar{f}(x_{-i})\big)^2\Big] + \big(\bar{f}(x_{-i}) - y\big)^2 \\ + 2\big(\bar{f}(x_{-i}) - y\big)\mathbb{E}_T\big[Z - \bar{f}(x_{-i})\big].\end{aligned}$$

The cross term vanishes because

$$\mathbb{E}_T[Z - \bar{f}(x_{-i})] = 0.$$

Therefore,

$$\mathbb{E}_T[(Z-y)^2] = \operatorname{Var}_{T\sim q}\big(f(x_{-i}, T)\big) + \big(\bar{f}(x_{-i}) - y\big)^2.$$

Taking expectation over $(x_{-i}, y)$ yields

$$R_q(f) = \mathbb{E}_{(x_{-i},y)}\Big[\big(\bar{f}(x_{-i}) - y\big)^2\Big] + D_q(f).$$

The excess-risk certificate then follows exactly as in Theorem 4.2. □

*Remark* D.11 (Local slope-penalty interpretation). The certificate above does not require bounded support or a Taylor expansion. However, when $t \mapsto f_\theta(x_{-i}, t)$ is sufficiently smooth near $\mu_q$, Gaussian replacement admits a useful local interpretation. For fixed $x_{-i}$, write

$$B_\theta(x_{-i}) := \partial_t f_\theta(x_{-i}, \mu_q).$$

A first-order expansion gives

$$f_\theta(x_{-i}, T) = f_\theta(x_{-i}, \mu_q) + B_\theta(x_{-i})(T - \mu_q) + r_\theta(x_{-i}, T),$$

where $r_\theta(x_{-i}, T)$ is a higher-order remainder. Hence

$$\begin{aligned}\operatorname{Var}_{T\sim q}\big(f_\theta(x_{-i}, T)\big) = \sigma_q^2 B_\theta(x_{-i})^2 + \operatorname{Var}_{T\sim q}\big(r_\theta(x_{-i}, T)\big) \\ + 2B_\theta(x_{-i})\operatorname{Cov}_{T\sim q}\big(T - \mu_q,\ r_\theta(x_{-i}, T)\big).\end{aligned}$$

Thus, for locally linear models or sufficiently small $\sigma_q$, the leading term of the residual-dependence certificate is

$$\sigma_q^2 B_\theta(x_{-i})^2,$$

which penalizes local sensitivity to the removed coordinate. In general, the covariance and remainder terms need not vanish, so the exact certificate should be understood through $D_q(f)$, not solely through the local slope term.

### D.14 GAUSSIAN REPLACEMENT CERTIFICATE

**Corollary D.12** (Gaussian replacement certificate)**.** *Under the assumptions of Lemma D.10, Gaussian replacement satisfies the same excess-risk certificate as bounded replacement:*

$$D_q(\hat{f}) \le R_q(\hat{f}) - \inf_g R_q(g).$$

*Therefore, if $\hat{f}$ is a $\delta$-near minimizer of the Gaussian replacement risk, then*

$$D_q(\hat{f}) \le \delta.$$

**Remark.** Gaussian replacement certifies invariance in an average sense under $T \sim \mathcal{N}(\mu_q, \sigma_q^2)$. Unlike bounded uniform replacement, it does not directly provide a full-domain certificate over a finite interval unless the Gaussian is truncated or the desired interval is otherwise covered with an explicit high-probability statement. In practice, a truncated Gaussian may be used to preserve numerical stability and to align the replacement support with a declared unlearning interval.

### D.15 REPLACEMENT-RISK OPTIMUM EQUALS RETAINED-FEATURE RETRAINING OPTIMUM

**Proposition D.13.** *Assume $T \sim q$ is independent of $(x_{-i}, y)$ and $\ell(\hat{y}, y) = (\hat{y} - y)^2$. Let $\mathcal{H}$ be any hypothesis class that contains all $t$-invariant predictors $g(x_{-i}, t) \equiv h(x_{-i})$ for measurable $h$. Then any minimizer $f^\star \in \arg\min_{f \in \mathcal{H}} R_q(f)$ satisfies*

$$\bar{f}^\star(x_{-i}) = \mathbb{E}[y \mid x_{-i}] \quad \textit{a.s.}, \tag{25}$$

*and in particular the minimum value equals the minimum achievable by predictors that only observe $x_{-i}$:*

$$\inf_{f \in \mathcal{H}} R_q(f) = \inf_{h : \mathcal{X}_{-i} \to \mathbb{R}} \mathbb{E}\big[(h(x_{-i}) - y)^2\big]. \tag{26}$$

### D.16 RANGE-WIDE APPROXIMATE UNLEARNING FOR BOUNDED EFFECTIVE LOGITS

*Proof.* By Theorem 4.5, if

$$R_{q_i}(\widetilde{f}) - \inf_{g \in \mathcal{G}_K} R_{q_i}(g) \le \delta,$$

then

$$D_{q_i}(\widetilde{f}) = \mathbb{E}_{x_{-i}} \operatorname{Var}_{T_i \sim q_i}\left(\widetilde{f}(x_{-i}, T_i)\right) \le \frac{2\delta}{m_K}.$$

Let

$$V(x_{-i}) := \operatorname{Var}_{T_i \sim q_i}\left(\widetilde{f}(x_{-i}, T_i)\right).$$

By Markov's inequality, with probability at least $1 - \alpha$ over $x_{-i}$,

$$V(x_{-i}) \le \frac{2\delta}{m_K \alpha}.$$

For such $x_{-i}$, Lemma D.4 applied to $g(t) = \widetilde{f}(x_{-i}, t)$ on $[a_i, b_i]$ gives

$$\sup_{t, t' \in [a_i, b_i]} \left|\widetilde{f}(x_{-i}, t) - \widetilde{f}(x_{-i}, t')\right| \le 2\left(8L(b_i - a_i)V(x_{-i})\right)^{1/3}.$$

Substituting the bound on $V(x_{-i})$ yields

$$\sup_{t, t' \in [a_i, b_i]} \left|\widetilde{f}(x_{-i}, t) - \widetilde{f}(x_{-i}, t')\right| \le 2\left(\frac{16L(b_i - a_i)\delta}{m_K \alpha}\right)^{1/3}.$$

Since $I_i \subseteq [a_i, b_i]$, the same bound holds over $I_i$. □

### D.17 LEMMA ABOUT Z-SCORE

With full-support replacement $q = \mathrm{Unif}[a, b]$ and the Lipschitz condition $|\partial_t f_\theta(x_{-i}, t)| \le L_1$, Lemma D.8 turns a small certificate $D_q(\theta) = \mathbb{E}_{x_{-i}} \mathrm{Var}_{T\sim q}\big(f_\theta(x_{-i}, T)\big)$ into a range-wide unlearning guarantee, i.e., $f_\theta(x_{-i}, t)$ is (approximately) invariant to the removed feature for all $t \in [a, b]$.

Suppose the $i$-th feature of the training data was normalized using z-score normalization, i.e., for each raw feature value $x_i$ in the training set, the stored value is

$$\tilde{x}_i = \frac{x_i - \mu_i}{\sigma_i}, \tag{27}$$

where $\mu_i$ and $\sigma_i$ denote the (empirical) mean and standard deviation of the $i$-th feature, respectively.

**Lemma D.14.** *Let $T \sim \mathcal{N}(0, \Delta^2)$ be a normal distribution centered at 0 with user-specified variance $\Delta^2$. Then: Replacing $\tilde{x}_i$ by samples from $T$ yields a mean-zero replacement distribution in the normalized space, requiring no access to the original mean $\mu_i$.This replacement strategy is equivalent to drawing from $\mathcal{N}(\mu_i, \Delta^2\sigma_i^2)$ in the original (unnormalized) feature space. Therefore, fine-tuning the model using random draws from $\mathcal{N}(0, \Delta^2)$ in the z-score normalized space is sufficient to implement feature unlearning without requiring access to the raw feature mean or training data.*

*Proof.* By the definition of z-score normalization, the transformed variable $\tilde{x}_i$ has empirical mean 0 and standard deviation 1. When we sample $T \sim \mathcal{N}(0, \Delta^2)$ in this space, it is already centered at the mean of the normalized data.

To understand its correspondence in the original feature space, note that if we un-normalize a draw $t \sim \mathcal{N}(0, \Delta^2)$, we obtain

$$x_i' = \sigma_i t + \mu_i \sim \mathcal{N}(\mu_i, \Delta^2\sigma_i^2).$$

Thus, sampling from $\mathcal{N}(0, \Delta^2)$ in the z-score space is equivalent to sampling from a distribution centered at $\mu_i$ in the original space.

As a result, this replacement preserves the desired properties (bounded variance, centered replacement) without requiring explicit knowledge of $\mu_i$—it is implicitly encoded in the normalization. □

Lemma D.14 shows that our unlearning procedure requires no additional information when the original model is trained on normalized features (a standard practice in tabular learning), since mean-centered bounded replacement reduces to sampling around zero in the normalized space.

# E FINITE-CHECKPOINT ANALYSIS AND DEPENDENCE STUDIES

We use Lemma D.8 (proved above): if $g$ is $L$-Lipschitz on $[a,b]$ and $T \sim \mathrm{Unif}[a,b]$, then $\sup_{t\in[a,b]} |g(t) - \mathbb{E}g(T)| \le (8L(b-a)\operatorname{Var} g(T))^{1/3}$.

*Proof of Proposition 4.6.* Write $Z = X_{-i}$, $V(z) = \mathrm{Var}_{T\sim q} f(z,T)$, and $M(z) = \sup_{t\in[a,b]} |f(z,t) - \bar{f}(z)|$. By Lemma D.8, $M(z) \le (8L_f(b-a)V(z))^{1/3}$. Since $x \mapsto x^{1/3}$ and $x \mapsto x^{2/3}$ are concave, Jensen's inequality gives $\mathbb{E}M(Z) \le (8L_f(b-a)D_q(f))^{1/3} = \kappa(f)$ and $\mathbb{E}M(Z)^2 \le \kappa(f)^2$.

(i) Since $X_i \in [a,b]$, $\mathrm{Var}_{X_i|Z} f(Z,X_i) \le \mathbb{E}[(f(Z,X_i) - \bar{f}(Z))^2 \mid Z] \le M(Z)^2$; take expectations.

(ii) For every $t \in [a,b]$, $|\ell(f(z,t),y) - \ell(\bar{f}(z),y)| \le L_\ell M(z)$. Therefore $|R_{\mathrm{ID}}(f) - R_q(f)| \le \big|\mathbb{E}[\ell(f(Z,X_i),Y) - \ell(\bar{f}(Z),Y)]\big| + \big|\mathbb{E}[\ell(f(Z,T),Y) - \ell(\bar{f}(Z),Y)]\big| \le 2L_\ell\, \mathbb{E}M(Z) \le 2L_\ell\kappa(f)$.

(iii) Let $A(z,r)$ be the attacker's output from responses $r = (f(z,t_1(z)),\dots,f(z,t_k(z)))$ and define the data-only attacker $A'(z) := A(z,(\bar{f}(z),\dots,\bar{f}(z)))$. Then $|A(z,r) - A'(z)| \le L_A \max_j |f(z,t_j(z)) - \bar{f}(z)| \le L_A M(z)$, so for the $\lambda$-Lipschitz attack loss $\ell_{\mathrm{att}}$, $\mathbb{E}\,\ell_{\mathrm{att}}(A(Z,r),X_i) \ge \mathbb{E}\,\ell_{\mathrm{att}}(A'(Z),X_i) - \lambda L_A\kappa(f) \ge \inf_h \mathbb{E}\,\ell_{\mathrm{att}}(h(Z),X_i) - \lambda L_A\kappa(f)$. □

For the retrained model $g^{\mathrm{rt}}(z)$, every model-assisted attacker is a data-only attacker, so the retrained oracle has zero advantage; (iii) states that BMCR's advantage is at most $\lambda L_A\kappa(f)$. Queries outside $[a,b]$ are not covered, and attackers that are not Lipschitz in the responses (e.g., hard thresholds) can amplify small differences; we therefore complement (iii) with the empirical audit in Appendix X.

**Corollary E.1** (Finite-sample decomposition)**.** *Let $\widehat{R}_q$ be the empirical replacement risk on $n$ retained examples with $m$ replacement draws each, let $\Delta_{n,m} := \sup_{f\in\mathcal{F}} |\widehat{R}_q(f) - R_q(f)|$ over the model class $\mathcal{F}$, let $g^{\mathrm{rt}} \in \mathcal{F}$ be invariant on $[a,b]$, and let $\widehat{f} \in \mathcal{F}$ satisfy $\widehat{R}_q(\widehat{f}) \le \widehat{R}_q(g^{\mathrm{rt}}) + \varepsilon_{\mathrm{opt}}$. Under the conditions of Proposition 4.6(ii),*

$$R_{\mathrm{ID}}(\widehat{f}) - R_{\mathrm{ID}}(g^{\mathrm{rt}}) \le \varepsilon_{\mathrm{opt}} + 2\Delta_{n,m} + 2L_\ell\,\kappa(\widehat{f}).$$

*Proof.* $R_{\mathrm{ID}}(\widehat{f}) \le R_q(\widehat{f}) + 2L_\ell\kappa(\widehat{f}) \le \widehat{R}_q(\widehat{f}) + \Delta_{n,m} + 2L_\ell\kappa(\widehat{f}) \le \widehat{R}_q(g^{\mathrm{rt}}) + \varepsilon_{\mathrm{opt}} + \Delta_{n,m} + 2L_\ell\kappa(\widehat{f}) \le R_q(g^{\mathrm{rt}}) + \varepsilon_{\mathrm{opt}} + 2\Delta_{n,m} + 2L_\ell\kappa(\widehat{f})$, and $R_q(g^{\mathrm{rt}}) = R_{\mathrm{ID}}(g^{\mathrm{rt}})$ because $g^{\mathrm{rt}}$ is invariant on $[a,b] \ni X_i$. □

The three terms separate finite non-convex optimization, finite data and replacement sampling, and off-manifold replacement. The $m$-draw estimator of $R_q$ is unbiased, with replacement-induced variance of order $1/m$. On an independent audit set, $D_q(\widehat{f})$ can be replaced by RIPV (with the unbiased $1/(m-1)$ normalization) plus a standard concentration term for bounded outputs.

## E.1 DEPENDENCE-CONTROLLED SYNTHETIC STUDY

**Data.** Let $X_i$ be the removed feature and $X_{\mathrm{corr}}$ a retained feature whose dependence on $X_i$ is controlled; the retained vector is $X_{-i} = (X_{\mathrm{corr}}, Z_1, \dots, Z_4)$. For $\rho \in \{0, 0.5, 0.9, 0.99\}$, draw $V, R \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and set

$$U = \rho V + \sqrt{1-\rho^2}\, R, \qquad X_i = 2\Phi(U) - 1, \qquad X_{\mathrm{corr}} = 2\Phi(V) - 1,$$

so $X_i, X_{\mathrm{corr}} \sim \mathrm{Unif}[-1,1]$ for every $\rho$. With $Z_1,\dots,Z_4 \sim \mathrm{Unif}[-1,1]$ independent,

$$Y = 1.5\sin(\pi X_{\mathrm{corr}}) + 0.5Z_1^2 - 0.75Z_2 + 0.5Z_3Z_4 + R + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 0.1^2).$$

The component $R = \big(\Phi^{-1}(\frac{X_i+1}{2}) - \rho\,\Phi^{-1}(\frac{X_{\mathrm{corr}}+1}{2})\big)/\sqrt{1-\rho^2}$ is recoverable from $(X_i, X_{\mathrm{corr}})$ but not from $X_{\mathrm{corr}}$ alone, so the marginal, support, label mechanism, and incremental predictive contribution of $X_i$ are fixed across $\rho$; only the dependence changes. We use 10,000/2,000/5,000 train/validation/test examples and three seeds.

**Methods.** Independent BMCR samples $T \sim \mathrm{Unif}[-1, 1]$. The oracle conditional diagnostic computes $v = \Phi^{-1}(\frac{x_{\mathrm{corr}}+1}{2})$, draws $U' \mid V = v \sim \mathcal{N}(\rho v, 1-\rho^2)$, and sets $T = 2\Phi(U') - 1$, an exact draw from $P(X_i \mid X_{-i})$. It uses information unavailable to BMCR and serves only as a reference. Architecture, pretrained model, optimizer, learning rate, batch size, and budget are matched. The utility gap is the percentage increase in test loss relative to retraining without $X_i$; Dist-RT is the normalized RMS distance to the retrained predictions; $D_{\mathrm{ind}}$ and $D_{\mathrm{cond}}$ are residual prediction variances under independent and true conditional replacement.

Table 6: Synthetic dependence sweep (200 epochs, one replacement draw per example, $K = 1$).

| $\rho$ | Replacement | Utility gap (%) | Dist-RT | $D_{\mathrm{ind}}$ | $D_{\mathrm{cond}}$ |
|---|---|---|---|---|---|
| 0.00 | Independent BMCR | $0.21 \pm 0.08$ | $0.010 \pm 0.002$ | $0.0014 \pm 0.0004$ | $0.0015 \pm 0.0003$ |
| | Oracle conditional | $0.17 \pm 0.07$ | $0.009 \pm 0.002$ | $0.0016 \pm 0.0003$ | $0.0015 \pm 0.0004$ |
| 0.50 | Independent BMCR | $0.29 \pm 0.11$ | $0.013 \pm 0.003$ | $0.0018 \pm 0.0005$ | $0.0020 \pm 0.0005$ |
| | Oracle conditional | $0.23 \pm 0.08$ | $0.010 \pm 0.002$ | $0.0023 \pm 0.0006$ | $0.0018 \pm 0.0004$ |
| 0.90 | Independent BMCR | $0.54 \pm 0.15$ | $0.019 \pm 0.005$ | $0.0022 \pm 0.0006$ | $0.0031 \pm 0.0008$ |
| | Oracle conditional | $0.32 \pm 0.11$ | $0.013 \pm 0.003$ | $0.0047 \pm 0.0010$ | $0.0024 \pm 0.0006$ |
| 0.99 | Independent BMCR | $0.78 \pm 0.21$ | $0.023 \pm 0.006$ | $0.0029 \pm 0.0007$ | $0.0048 \pm 0.0010$ |
| | Oracle conditional | $0.43 \pm 0.14$ | $0.017 \pm 0.004$ | $0.0088 \pm 0.0017$ | $0.0031 \pm 0.0008$ |

At $\rho = 0$, $P(X_i \mid X_{-i}) = P(X_i)$ and the two methods coincide, as expected. As $\rho$ grows, independent replacement shows a measurable finite-budget disadvantage (utility gap $0.54\%$ vs. $0.32\%$ at $\rho = 0.9$; $0.78\%$ vs. $0.43\%$ at $\rho = 0.99$), but it stays within $0.78\%$ of retraining and achieves $D_{\mathrm{cond}} = 0.0048$ even when $X_i$ is almost determined by $X_{\mathrm{corr}}$. The oracle has a larger $D_{\mathrm{ind}}$ at high $\rho$ because it concentrates its samples near the value predicted by $X_{\mathrm{corr}}$ and rarely visits the rest of the marginal support.

Table 7: Finite-training analysis at $\rho = 0.99$; $K$ is the number of replacement draws averaged per example.

| Method | $K$ | Epochs | Utility gap (%) | Dist-RT | $D_{\mathrm{ind}}$ | $D_{\mathrm{cond}}$ |
|---|---|---|---|---|---|---|
| Independent BMCR | 1 | 10 | $4.74 \pm 0.69$ | $0.060 \pm 0.010$ | $0.0352 \pm 0.0058$ | $0.0403 \pm 0.0068$ |
| Independent BMCR | 1 | 50 | $2.16 \pm 0.39$ | $0.039 \pm 0.007$ | $0.0145 \pm 0.0031$ | $0.0184 \pm 0.0038$ |
| Independent BMCR | 1 | 200 | $0.78 \pm 0.21$ | $0.023 \pm 0.006$ | $0.0029 \pm 0.0007$ | $0.0048 \pm 0.0010$ |
| Independent BMCR | 1 | 1500 | $0.35 \pm 0.12$ | $0.014 \pm 0.004$ | $0.0012 \pm 0.0003$ | $0.0019 \pm 0.0005$ |
| Independent BMCR | 16 | 10 | $3.14 \pm 0.55$ | $0.047 \pm 0.008$ | $0.0235 \pm 0.0042$ | $0.0266 \pm 0.0048$ |
| Independent BMCR | 16 | 50 | $1.13 \pm 0.26$ | $0.030 \pm 0.006$ | $0.0072 \pm 0.0015$ | $0.0093 \pm 0.0021$ |
| Independent BMCR | 16 | 200 | $0.47 \pm 0.15$ | $0.018 \pm 0.004$ | $0.0019 \pm 0.0005$ | $0.0029 \pm 0.0007$ |
| Independent BMCR | 16 | 1500 | $0.30 \pm 0.11$ | $0.012 \pm 0.003$ | $0.0011 \pm 0.0003$ | $0.0016 \pm 0.0004$ |
| Oracle conditional | 1 | 10 | $3.06 \pm 0.48$ | $0.046 \pm 0.007$ | $0.0285 \pm 0.0048$ | $0.0206 \pm 0.0039$ |
| Oracle conditional | 1 | 50 | $0.93 \pm 0.24$ | $0.025 \pm 0.005$ | $0.0174 \pm 0.0032$ | $0.0068 \pm 0.0015$ |
| Oracle conditional | 1 | 200 | $0.43 \pm 0.14$ | $0.017 \pm 0.004$ | $0.0088 \pm 0.0017$ | $0.0031 \pm 0.0008$ |
| Oracle conditional | 1 | 1500 | $0.28 \pm 0.10$ | $0.011 \pm 0.003$ | $0.0056 \pm 0.0012$ | $0.0015 \pm 0.0004$ |

Most of the disadvantage of independent replacement is transient: with $K = 1$, the independent–conditional utility-gap difference falls from $1.68$ points at 10 epochs to $0.07$ at 1500, while $D_{\mathrm{cond}}$ of independent BMCR falls from $0.0403$ to $0.0019$ (oracle: $0.0015$). More draws help early (at 50 epochs, $K = 16$ halves the utility gap), and by 1500 epochs $K = 1$ and $K = 16$ are nearly identical.

### E.2 Dependence-stratified validation on real data

For every removed feature in the single-feature setting (most and least influential, all seven datasets), we estimate its dependence on $X_{-i}$ as the five-fold cross-fitted out-of-fold $R^2$ of a gradient-boosted regressor of $X_i$ on $X_{-i}$, and group features as low ($< 0.2$), medium (0.2–0.6), or high ($\geq 0.6$) dependence. The estimated conditional diagnostic uses cross-fitted local donor sampling: for each $x_{-i}$, it samples $X_i$ from one of the 20 nearest training-fold neighbours in standardized retained-feature space. This requires the removed feature and is only a reference. Both methods share the pretrained model, architecture, optimizer, hyperparameters, and a 200-epoch budget with $K = 1$. Results are first averaged over architectures per dataset–feature case and then within groups.

The real-data results reproduce the synthetic trend: the finite-budget penalty of independent replacement grows with dependence but remains below $0.3$ points, and independent BMCR has low residual

Table 8: Real-data validation by dependence group (200 epochs).

| Group | Cases | Mean dep. | Indep. gap (%) | Cond. gap (%) | Difference | Indep. Dist-RT | Indep. $\widehat{D}_{\mathrm{cond}}$ |
|---|---|---|---|---|---|---|---|
| Low | 3 | 0.08 | $0.29 \pm 0.10$ | $0.24 \pm 0.09$ | $0.05 \pm 0.05$ | $0.013 \pm 0.004$ | $0.0023 \pm 0.0006$ |
| Medium | 7 | 0.44 | $0.44 \pm 0.15$ | $0.35 \pm 0.13$ | $0.09 \pm 0.06$ | $0.017 \pm 0.005$ | $0.0035 \pm 0.0010$ |
| High | 4 | 0.84 | $0.74 \pm 0.23$ | $0.48 \pm 0.18$ | $0.26 \pm 0.11$ | $0.022 \pm 0.006$ | $0.0052 \pm 0.0012$ |

variance under conditionally realistic replacements. These are finite-training empirical findings; the population-level guarantee is Theorem 4.2.

### E.3 Reaching the Theoretical Regime with Non-Convex Fine-Tuning

For MLP, ResNet, and FT-Transformer, BMCR fine-tuning is non-convex in $\theta$, so we do not rely on optimizer convergence. The certificate is instead a property of the *trained predictor*: for any $\theta$, with $\widetilde{f}_\theta = K \tanh(f_\theta/K)$ and independent $T \sim q$, Theorem 4.5 gives

$$D_q(\widetilde{f}_\theta) \le \frac{2}{m_K}\Big(R_q(\widetilde{f}_\theta) - \mathbb{E}_{X_{-i},Y}\,\ell\big(\mathbb{E}_{T\sim q}[\widetilde{f}_\theta(X_{-i},T)],Y\big)\Big),$$

which requires neither convexity in $\theta$ nor global optimality; strong convexity is used only in the scalar logit. Both conditions (bounded logits and independent replacement) hold by construction. In practice we estimate $D_q$ directly on an independent audit set via RIPV with repeated draws (Sec. 5), and Proposition 4.6 and Corollary E.1 hold for any checkpoint.

# F LIMITATIONS

**Scope of the guarantee.** BMCR provides a coordinate-level functional guarantee. It controls direct prediction dependence on the removed coordinate under the replacement distribution. It does not guarantee parameter-level indistinguishability from retraining, nor does it imply that all information about the removed feature has been eliminated from correlated retained features. If $X_i$ is predictable from $X_{-i}$, then proxy information remains, exactly as it does for the retrained oracle. What BMCR controls is the *excess* model-assisted leakage beyond this data-inherent baseline (Proposition 4.6(iii)), and only for attackers that query within $\mathrm{supp}(q)$ and are Lipschitz in the model's responses. Privacy and fairness results should therefore be interpreted as empirical audits (Appendix X) rather than formal privacy guarantees.

**Access assumptions.** BMCR assumes that the retained features and labels remain accessible and that a valid support for the revoked feature can be declared (Appendix B). It does not apply to the fully data-free setting, in which retraining is also impossible.

**Finite training.** The certificate holds for near-minimizers; at a finite checkpoint it is audited through RIPV (Appendix E.3), and range-wide statements rely on Lipschitz constants that are not certified by training. Independent replacement incurs a small finite-budget penalty that grows with the dependence between removed and retained features (Appendix E).

**Beyond tabular data.** While our framework offers a theoretically well-founded, data-agnostic framework for feature unlearning in tabular or structured data domains, its direct application to CV tasks is subject to a number of operational difficulties. In the CV domain, the features of the image are often a convoluted mix of spatial information, and the convolutional network backbone learns the features implicitly without clear feature boundaries. Moreover, the raw data in the CV domain is a high-dimensional tensor of pixel information, in which the semantically meaningful features of the image (for example, “watermark," “sky texture," or “logo") are not immediately available or disentanglement-friendly without additional annotation or segmentation mechanisms. Thus, although the framework is theoretically data-agnostic, the operationalization of feature unlearning for CV tasks is likely to necessitate additional structure, such as interpretable attention mechanisms or learned bottlenecks for feature disentanglement (see also Appendix F.1).

## F.1 EXTENSION BEYOND TABULAR DATA

BMCR applies whenever the unlearning target is an explicit input variable with a declarable domain. For a conditional generative model $p_\theta(y \mid x_{-i}, x_i)$ trained with negative log-likelihood, the replacement objective $\mathbb{E}_{X_{-i},Y}\mathbb{E}_{T\sim q}[-\log p_\theta(Y \mid X_{-i}, T)]$ upper-bounds the loss of the mixture $\bar{p}_\theta(y \mid x_{-i}) = \mathbb{E}_T\, p_\theta(y \mid x_{-i}, T)$ by Jensen's inequality, with equality if and only if $p_\theta(Y \mid X_{-i}, T)$ is $q$-a.s. constant in $T$; near-minimizers thus have a small Jensen gap, a likelihood analogue of $D_q$. Turning this gap into an output-distribution distance, and handling concepts distributed across pixels, tokens, or latent features where no explicit coordinate or support exists, are left for future work. Tabular models remain widely used in credit, healthcare, insurance, and risk assessment, where individual attributes are explicit and may need to be removed for legal or governance reasons.

## G DATASET INFORMATION

A summary of used dataset from OpenML Feurer et al. (2021) in this paper is presented in Table 9.

Table 9: Summary of used tabular datasets.

| Dataset Name | Number of Samples | Number of Features |
|---|---|---|
| *Blood* | 748 | 4 |
| *Cali* | 20634 | 8 |
| *Credit* | 16714 | 10 |
| *Diabetes* | 768 | 8 |
| *House* | 13488 | 16 |
| *Magic* | 13376 | 10 |
| *Pol* | 10082 | 26 |

- *Blood*: The link to this dataset is `https://www.openml.org/d/1464`.
- *Cali*: The link to this dataset is `https://www.openml.org/d/44090`.
- *Credit*: The link to this dataset is `https://www.openml.org/d/44089`.
- *Diabetes*: The link to this dataset is `https://www.openml.org/d/37`.
- *House*: The link to this dataset is `https://www.openml.org/d/43957`
- *Magic*: The link to this dataset is `https://www.openml.org/d/44125`.
- *Pol*: The link to this dataset is `https://www.openml.org/d/43956`.

## H Dataset Statistical Summary

Table 10 shows the statistical summary for the *Blood* dataset.

Table 10: Summary statistics for *Blood* dataset

| Statistic \ Feature | V1 | V2 | V3 | V4 |
|---|---|---|---|---|
| mean | 9.51 | 5.51 | 1378.68 | 34.28 |
| std | 8.1 | 5.84 | 1459.83 | 24.38 |
| min | 0.0 | 1.0 | 250.0 | 2.0 |
| 25% | 2.75 | 2.0 | 500.0 | 16.0 |
| 50% | 7.0 | 4.0 | 1000.0 | 28.0 |
| 75% | 14.0 | 7.0 | 1750.0 | 50.0 |
| max | 74.0 | 50.0 | 12500.0 | 98.0 |

Table 11 shows the statistical summary for the *Cali* dataset.

Table 11: Summary statistics for *Cali* dataset

| Statistic \ Feature | MedInc | HouseAge | AveRooms | AveBedrms | Population | AveOccup | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|
| mean | 3.87 | 28.64 | 5.43 | 1.1 | 1425.4 | 3.07 | 35.63 | -119.57 |
| std | 1.9 | 12.58 | 2.47 | 0.47 | 1132.14 | 10.39 | 2.14 | 2.0 |
| min | 0.5 | 1.0 | 0.85 | 0.33 | 3.0 | 0.69 | 32.54 | -124.35 |
| 25% | 2.56 | 18.0 | 4.44 | 1.01 | 787.0 | 2.43 | 33.93 | -121.8 |
| 50% | 3.53 | 29.0 | 5.23 | 1.05 | 1166.0 | 2.82 | 34.26 | -118.49 |
| 75% | 4.74 | 37.0 | 6.05 | 1.1 | 1725.0 | 3.28 | 37.71 | -118.01 |
| max | 15.0 | 52.0 | 141.91 | 34.07 | 35682.0 | 1243.33 | 41.95 | -114.31 |

Table 12 shows the statistical summary for the *Credit* dataset.

Table 12: Summary statistics for *Credit* dataset.

| Feature \ Statistic | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|
| RevolvingUtilizationOfUnsecuredLines | 4.8 | 204.06 | 0.0 | 0.08 | 0.44 | 0.93 | 22000.0 |
| age | 48.8 | 13.91 | 21.0 | 38.0 | 48.0 | 58.0 | 101.0 |
| NumberOfTime30-59DaysPastDueNotWorse | 1.11 | 7.17 | 0.0 | 0.0 | 0.0 | 1.0 | 98.0 |
| DebtRatio | 30.98 | 719.69 | 0.0 | 0.16 | 0.32 | 0.53 | 61106.5 |
| MonthlyIncome | 6118.12 | 5931.84 | 0.0 | 3128.5 | 5000.0 | 7573.0 | 250000.0 |
| NumberOfOpenCreditLinesAndLoans | 8.5 | 5.37 | 0.0 | 5.0 | 8.0 | 11.0 | 57.0 |
| NumberOfTimes90DaysLate | 0.86 | 7.17 | 0.0 | 0.0 | 0.0 | 0.0 | 98.0 |
| NumberRealEstateLoansOrLines | 1.05 | 1.27 | 0.0 | 0.0 | 1.0 | 2.0 | 29.0 |
| NumberOfTime60-89DaysPastDueNotWorse | 0.73 | 7.14 | 0.0 | 0.0 | 0.0 | 0.0 | 98.0 |
| NumberOfDependents | 0.94 | 1.2 | 0.0 | 0.0 | 0.0 | 2.0 | 8.0 |

Table 13 shows the statistical summary for the *Diabetes* dataset.

Table 13: Summary statistics for *Diabetes* dataset

| Statistic \ Feature | preg | plas | pres | skin | insu | mass | pedi | age |
|---|---|---|---|---|---|---|---|---|
| mean | 3.85 | 120.89 | 69.11 | 20.54 | 79.8 | 31.99 | 0.47 | 33.24 |
| std | 3.37 | 31.97 | 19.36 | 15.95 | 115.24 | 7.88 | 0.33 | 11.76 |
| min | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.08 | 21.0 |
| 25% | 1.0 | 99.0 | 62.0 | 0.0 | 0.0 | 27.3 | 0.24 | 24.0 |
| 50% | 3.0 | 117.0 | 72.0 | 23.0 | 30.5 | 32.0 | 0.37 | 29.0 |
| 75% | 6.0 | 140.25 | 80.0 | 32.0 | 127.25 | 36.6 | 0.63 | 41.0 |
| max | 17.0 | 199.0 | 122.0 | 99.0 | 846.0 | 67.1 | 2.42 | 81.0 |

Table 14 shows the statistical summary for the *House* dataset.

Table 14: Summary statistics for *House* dataset

| Statistic \ Feature | P1 | P5p1 | P6p2 | P11p4 | P14p9 | P15p1 | P15p3 | P16p2 | P18p2 | P27p4 | H2p2 | H8p2 | H10p1 | H13p1 | H18pA | H40p4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 10012.98 | 0.48 | 0.06 | 0.16 | 0.11 | 0.85 | 0.02 | 0.72 | 0.0 | 0.04 | 0.11 | 0.05 | 0.96 | 0.3 | 0.1 | 0.46 |
| std | 80659.44 | 0.03 | 0.14 | 0.08 | 0.06 | 0.08 | 0.06 | 0.09 | 0.0 | 0.03 | 0.12 | 0.13 | 0.1 | 0.14 | 0.09 | 0.32 |
| min | 4.0 | 0.23 | 0.0 | 0.0 | 0.0 | 0.12 | 0.0 | 0.24 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 25% | 587.0 | 0.47 | 0.0 | 0.1 | 0.07 | 0.82 | 0.0 | 0.66 | 0.0 | 0.02 | 0.04 | 0.0 | 0.98 | 0.21 | 0.05 | 0.22 |
| 50% | 1973.5 | 0.48 | 0.01 | 0.15 | 0.11 | 0.87 | 0.0 | 0.72 | 0.0 | 0.03 | 0.07 | 0.0 | 0.99 | 0.29 | 0.08 | 0.46 |
| 75% | 6662.5 | 0.5 | 0.03 | 0.19 | 0.15 | 0.9 | 0.02 | 0.78 | 0.01 | 0.05 | 0.12 | 0.03 | 1.0 | 0.38 | 0.13 | 0.67 |
| max | 7322564.0 | 0.84 | 1.0 | 0.92 | 0.51 | 1.0 | 0.85 | 1.0 | 0.12 | 0.71 | 0.98 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

Table 15 shows the statistical summary for the *Magic* dataset.

Table 15: Summary statistics for *Magic* dataset

| Statistic \ Feature | fLength: | fWidth: | fSize: | fConc: | fConc1: | fAsym: | fM3Long: | fM3Trans: | fAlpha: | fDist: |
|---|---|---|---|---|---|---|---|---|---|---|
| mean | 57.21 | 23.67 | 2.84 | 0.38 | 0.21 | -7.61 | 7.57 | 0.27 | 31.35 | 195.22 |
| std | 46.89 | 20.89 | 0.48 | 0.18 | 0.11 | 65.52 | 56.31 | 23.23 | 26.92 | 76.52 |
| min | 4.28 | 0.0 | 1.94 | 0.01 | 0.0 | -457.92 | -331.78 | -205.89 | 0.0 | 1.28 |
| 25% | 24.74 | 11.7 | 2.5 | 0.23 | 0.13 | -24.57 | -14.85 | -11.0 | 7.33 | 142.25 |
| 50% | 38.62 | 17.11 | 2.75 | 0.36 | 0.2 | 2.76 | 14.72 | 0.42 | 23.42 | 193.75 |
| 75% | 75.68 | 26.1 | 3.12 | 0.51 | 0.29 | 23.76 | 35.09 | 11.14 | 52.35 | 242.62 |
| max | 334.18 | 256.38 | 5.32 | 0.89 | 0.64 | 575.24 | 238.32 | 179.85 | 90.0 | 495.56 |

Table 16 shows the statistical summary for the *Pol* dataset.

Table 16: Summary statistics for *Pol* dataset

| Feature \ Statistic | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|
| f5 | 74.77 | 31.77 | 13.0 | 57.0 | 74.0 | 92.0 | 200.0 |
| f6 | 92.06 | 19.49 | 77.0 | 77.0 | 83.0 | 101.0 | 200.0 |
| f7 | 82.43 | 14.98 | 76.0 | 76.0 | 76.0 | 79.0 | 200.0 |
| f8 | 74.69 | 11.97 | 71.0 | 71.0 | 71.0 | 72.0 | 200.0 |
| f9 | 95.54 | 7.44 | 94.0 | 94.0 | 94.0 | 94.0 | 172.0 |
| f13 | 5.64 | 13.62 | 0.0 | 0.0 | 0.0 | 2.0 | 113.0 |
| f14 | 11.69 | 18.92 | 0.0 | 0.0 | 1.0 | 18.0 | 128.0 |
| f15 | 11.89 | 19.0 | 0.0 | 0.0 | 1.0 | 19.0 | 150.0 |
| f16 | 11.78 | 18.76 | 0.0 | 0.0 | 1.0 | 19.0 | 127.0 |
| f17 | 11.69 | 18.78 | 0.0 | 0.0 | 1.0 | 19.0 | 127.0 |
| f18 | 11.63 | 19.2 | 0.0 | 0.0 | 0.0 | 19.0 | 143.0 |
| f19 | 7.2 | 15.37 | 0.0 | 0.0 | 0.0 | 4.0 | 111.0 |
| f20 | 3.29 | 10.54 | 0.0 | 0.0 | 0.0 | 0.0 | 117.0 |
| f21 | 2.98 | 9.94 | 0.0 | 0.0 | 0.0 | 0.0 | 113.0 |
| f22 | 3.05 | 10.43 | 0.0 | 0.0 | 0.0 | 0.0 | 131.0 |
| f23 | 2.17 | 8.67 | 0.0 | 0.0 | 0.0 | 0.0 | 104.0 |
| f24 | 1.19 | 5.87 | 0.0 | 0.0 | 0.0 | 0.0 | 60.0 |
| f25 | 1.08 | 5.92 | 0.0 | 0.0 | 0.0 | 0.0 | 72.0 |
| f26 | 0.78 | 4.88 | 0.0 | 0.0 | 0.0 | 0.0 | 108.0 |
| f27 | 0.79 | 4.82 | 0.0 | 0.0 | 0.0 | 0.0 | 66.0 |
| f28 | 0.84 | 5.05 | 0.0 | 0.0 | 0.0 | 0.0 | 74.0 |
| f29 | 0.65 | 4.25 | 0.0 | 0.0 | 0.0 | 0.0 | 67.0 |
| f30 | 0.43 | 2.87 | 0.0 | 0.0 | 0.0 | 0.0 | 43.0 |
| f31 | 0.42 | 2.83 | 0.0 | 0.0 | 0.0 | 0.0 | 43.0 |
| f32 | 0.4 | 2.69 | 0.0 | 0.0 | 0.0 | 0.0 | 41.0 |
| f33 | 0.21 | 1.98 | 0.0 | 0.0 | 0.0 | 0.0 | 39.0 |

Table 17: Single-feature unlearning evaluation results of the most influential feature in the *Magic* tabular dataset using MLP neural backbone.

| Metric | FOUL | MIUL | Shuffle | FA-1 | FA0 | FA1 | GS | SAM | SBMB | BMCR0.1 | BMCR0.5 | BMCR1 | BMCR5 | BMCR10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | 0.063 ±0.018 | 0.058 ±0.018 | 0.012 ±0.006 | 0.032 ±0.011 | 0.067 ±0.021 | 0.026 ±0.010 | 0.010 ±0.005 | 0.061 ±0.019 | 0.009 ±0.005 | 0.022 ±0.012 | 0.015 ±0.008 | 0.011 ±0.006 | 0.009 ±0.005 | 0.009 ±0.005 |
| TRI | 0.672 ±0.000 | 1.026 ±0.004 | 0.998 ±0.003 | 0.667 ±0.000 | 0.778 ±0.007 | 0.669 ±0.001 | 0.994 ±0.004 | 0.779 ±0.015 | 1.000 ±0.002 | 0.989 ±0.012 | 0.994 ±0.010 | 0.999 ±0.009 | 0.994 ±0.005 | 0.987 ±0.002 |
| EI | 7.619 ±0.325 | 3.741 ±0.482 | 12.372 ±3.332 | 1.909 ±0.143 | 1.957 ±0.162 | 1.962 ±0.155 | 1.871 ±0.139 | 1.896 ±0.148 | 1.889 ±0.136 | 1.973 ±0.146 | 1.913 ±0.142 | 1.844 ±0.135 | 1.871 ±0.139 | 1.897 ±0.143 |
| RASI | 0.854 ±0.002 | 0.527 ±0.008 | 0.957 ±0.007 | 0.944 ±0.000 | 0.705 ±0.004 | 0.966 ±0.001 | 0.964 ±0.004 | 0.709 ±0.011 | 0.973 ±0.001 | 0.901 ±0.047 | 0.945 ±0.026 | 0.966 ±0.007 | 0.970 ±0.007 | 0.974 ±0.007 |
| SRI | 3.545 ±0.000 | 2.249 ±0.036 | 0.513 ±0.190 | 0.798 ±0.022 | 3.653 ±0.187 | 0.518 ±0.057 | 0.272 ±0.080 | 3.429 ±0.156 | 0.241 ±0.066 | 1.058 ±0.560 | 0.543 ±0.276 | 0.320 ±0.074 | 0.301 ±0.088 | 0.288 ±0.102 |
| SDI | 0.165 ±0.000 | 0.105 ±0.002 | 0.024 ±0.009 | 0.037 ±0.001 | 0.170 ±0.009 | 0.024 ±0.003 | 0.013 ±0.004 | 0.160 ±0.007 | 0.011 ±0.003 | 0.049 ±0.026 | 0.025 ±0.013 | 0.015 ±0.003 | 0.014 ±0.004 | 0.013 ±0.005 |
| MIA | 0.612 ±0.041 | 0.598 ±0.036 | 0.512 ±0.021 | 0.579 ±0.033 | 0.586 ±0.034 | 0.575 ±0.032 | 0.535 ±0.025 | 0.548 ±0.027 | 0.521 ±0.022 | 0.519 ±0.026 | 0.516 ±0.024 | 0.514 ±0.023 | 0.515 ±0.024 | 0.516 ±0.024 |
| $\Delta_{\mathrm{DP}}$ | 0.162 ±0.084 | 0.096 ±0.052 | 0.020 ±0.014 | 0.082 ±0.044 | 0.118 ±0.063 | 0.076 ±0.039 | 0.032 ±0.021 | 0.095 ±0.051 | 0.018 ±0.013 | 0.014 ±0.013 | 0.013 ±0.012 | 0.012 ±0.012 | 0.013 ±0.012 | 0.013 ±0.012 |
| $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | 0.118 ±0.069 | 0.088 ±0.049 | 0.025 ±0.018 | 0.063 ±0.037 | 0.098 ±0.056 | 0.058 ±0.034 | 0.027 ±0.019 | 0.082 ±0.045 | 0.022 ±0.016 | 0.022 ±0.019 | 0.021 ±0.018 | 0.021 ±0.018 | 0.021 ±0.018 | 0.021 ±0.018 |
| $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ | 0.186 ±0.104 | 0.113 ±0.067 | 0.027 ±0.019 | 0.095 ±0.054 | 0.137 ±0.078 | 0.089 ±0.050 | 0.035 ±0.025 | 0.108 ±0.062 | 0.021 ±0.015 | 0.018 ±0.017 | 0.017 ±0.016 | 0.017 ±0.015 | 0.017 ±0.016 | 0.017 ±0.016 |

Table 18: Single-feature unlearning evaluation results of the most influential feature in the *Magic* tabular dataset using ResNet neural backbone.

| Metric | FOUL | MIUL | Shuffle | FA-1 | FA0 | FA1 | GS | SAM | SBMB | BMCR0.1 | BMCR0.5 | BMCR1 | BMCR5 | BMCR10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | 0.072 ±0.022 | 0.053 ±0.016 | 0.007 ±0.004 | 0.047 ±0.015 | 0.052 ±0.017 | 0.044 ±0.014 | 0.006 ±0.004 | 0.039 ±0.013 | 0.005 ±0.003 | 0.010 ±0.006 | 0.007 ±0.005 | 0.006 ±0.004 | 0.005 ±0.004 | 0.005 ±0.003 |
| TRI | 0.664 ±0.000 | 1.008 ±0.004 | 1.020 ±0.004 | 0.837 ±0.010 | 0.696 ±0.041 | 0.638 ±0.071 | 1.025 ±0.005 | 0.955 ±0.018 | 1.023 ±0.008 | 1.014 ±0.000 | 1.018 ±0.001 | 1.022 ±0.002 | 1.020 ±0.002 | 1.019 ±0.002 |
| EI | 12.240 ±1.990 | 9.108 ±1.224 | 18.849 ±3.048 | 2.469 ±0.188 | 2.379 ±0.172 | 2.471 ±0.181 | 2.461 ±0.176 | 2.320 ±0.164 | 2.251 ±0.158 | 2.510 ±0.182 | 2.482 ±0.178 | 2.454 ±0.173 | 2.452 ±0.171 | 2.449 ±0.170 |
| RASI | 0.507 ±0.009 | 0.557 ±0.005 | 0.990 ±0.002 | 0.703 ±0.011 | 0.614 ±0.016 | 0.732 ±0.023 | 0.990 ±0.001 | 0.727 ±0.039 | 0.992 ±0.001 | 0.960 ±0.003 | 0.973 ±0.003 | 0.984 ±0.003 | 0.991 ±0.002 | 0.997 ±0.001 |
| SRI | 3.413 ±0.000 | 1.426 ±0.087 | 0.038 ±0.012 | 1.655 ±0.128 | 1.739 ±0.256 | 1.623 ±0.169 | 0.058 ±0.021 | 1.287 ±0.349 | 0.039 ±0.014 | 0.151 ±0.029 | 0.091 ±0.019 | 0.054 ±0.011 | 0.042 ±0.007 | 0.034 ±0.002 |
| SDI | 0.254 ±0.000 | 0.106 ±0.006 | 0.003 ±0.001 | 0.123 ±0.009 | 0.129 ±0.019 | 0.121 ±0.013 | 0.004 ±0.002 | 0.096 ±0.026 | 0.003 ±0.001 | 0.011 ±0.002 | 0.007 ±0.001 | 0.004 ±0.001 | 0.003 ±0.000 | 0.003 ±0.000 |
| MIA | 0.624 ±0.044 | 0.596 ±0.035 | 0.508 ±0.020 | 0.583 ±0.034 | 0.590 ±0.035 | 0.581 ±0.033 | 0.514 ±0.021 | 0.548 ±0.028 | 0.507 ±0.019 | 0.516 ±0.024 | 0.513 ±0.022 | 0.511 ±0.021 | 0.510 ±0.020 | 0.509 ±0.020 |
| $\Delta_{\mathrm{DP}}$ | 0.175 ±0.089 | 0.092 ±0.050 | 0.015 ±0.011 | 0.093 ±0.050 | 0.125 ±0.067 | 0.088 ±0.047 | 0.018 ±0.013 | 0.073 ±0.040 | 0.013 ±0.010 | 0.015 ±0.013 | 0.013 ±0.012 | 0.012 ±0.011 | 0.013 ±0.012 | 0.013 ±0.012 |
| $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | 0.127 ±0.073 | 0.083 ±0.046 | 0.019 ±0.014 | 0.072 ±0.040 | 0.103 ±0.058 | 0.067 ±0.038 | 0.021 ±0.015 | 0.064 ±0.037 | 0.018 ±0.013 | 0.024 ±0.020 | 0.022 ±0.019 | 0.020 ±0.018 | 0.021 ±0.018 | 0.021 ±0.018 |
| $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ | 0.198 ±0.112 | 0.108 ±0.064 | 0.022 ±0.016 | 0.113 ±0.064 | 0.147 ±0.083 | 0.105 ±0.059 | 0.024 ±0.018 | 0.085 ±0.049 | 0.017 ±0.013 | 0.020 ±0.018 | 0.018 ±0.016 | 0.016 ±0.015 | 0.016 ±0.015 | 0.017 ±0.015 |

Table 19: Single-feature unlearning evaluation results of the most influential feature in the *Magic* tabular dataset using FTformer neural backbone

| Metric | FOUL | MIUL | Shuffle | FA-1 | FA0 | FA1 | GS | SAM | SBMB | BMCR0.1 | BMCR0.5 | BMCR1 | BMCR5 | BMCR10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | 0.001 ±0.000 | 0.047 ±0.014 | 0.008 ±0.004 | 0.025 ±0.009 | 0.019 ±0.006 | 0.016 ±0.006 | 0.006 ±0.003 | 0.020 ±0.007 | 0.005 ±0.003 | 0.011 ±0.006 | 0.008 ±0.005 | 0.006 ±0.004 | 0.005 ±0.003 | 0.005 ±0.003 |
| TRI | 0.653 ±0.000 | 0.998 ±0.007 | 0.992 ±0.005 | 0.987 ±0.021 | 0.994 ±0.010 | 0.990 ±0.006 | 1.001 ±0.008 | 0.981 ±0.003 | 1.001 ±0.009 | 0.997 ±0.012 | 0.997 ±0.012 | 0.996 ±0.012 | 1.000 ±0.008 | 1.004 ±0.005 |
| EI | 16.880 ±0.819 | 7.304 ±0.689 | 27.341 ±0.526 | 2.254 ±0.161 | 2.348 ±0.174 | 2.317 ±0.169 | 2.297 ±0.166 | 2.280 ±0.163 | 2.259 ±0.159 | 2.242 ±0.156 | 2.244 ±0.157 | 2.245 ±0.157 | 2.263 ±0.159 | 2.280 ±0.163 |
| RASI | 1.000 ±0.000 | 0.536 ±0.003 | 0.985 ±0.003 | 0.894 ±0.034 | 0.929 ±0.010 | 0.948 ±0.015 | 0.991 ±0.003 | 0.926 ±0.012 | 0.990 ±0.003 | 0.959 ±0.012 | 0.975 ±0.007 | 0.991 ±0.003 | 0.993 ±0.002 | 0.995 ±0.002 |
| SRI | 0.000 ±0.000 | 0.692 ±0.013 | 0.061 ±0.016 | 0.308 ±0.123 | 0.181 ±0.035 | 0.148 ±0.033 | 0.034 ±0.006 | 0.208 ±0.027 | 0.026 ±0.001 | 0.109 ±0.039 | 0.058 ±0.022 | 0.025 ±0.006 | 0.024 ±0.006 | 0.024 ±0.006 |
| SDI | 0.000 ±0.000 | 0.094 ±0.002 | 0.008 ±0.002 | 0.042 ±0.017 | 0.024 ±0.005 | 0.020 ±0.004 | 0.005 ±0.001 | 0.028 ±0.004 | 0.004 ±0.000 | 0.015 ±0.005 | 0.008 ±0.003 | 0.003 ±0.001 | 0.003 ±0.001 | 0.003 ±0.001 |
| MIA | 0.501 ±0.008 | 0.591 ±0.034 | 0.510 ±0.020 | 0.556 ±0.030 | 0.544 ±0.027 | 0.538 ±0.026 | 0.512 ±0.020 | 0.545 ±0.027 | 0.508 ±0.019 | 0.515 ±0.023 | 0.512 ±0.021 | 0.510 ±0.020 | 0.511 ±0.020 | 0.512 ±0.021 |
| $\Delta_{\mathrm{DP}}$ | 0.005 ±0.003 | 0.087 ±0.047 | 0.016 ±0.012 | 0.055 ±0.032 | 0.046 ±0.026 | 0.039 ±0.023 | 0.017 ±0.012 | 0.052 ±0.029 | 0.013 ±0.010 | 0.016 ±0.014 | 0.015 ±0.014 | 0.014 ±0.013 | 0.014 ±0.013 | 0.015 ±0.013 |
| $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | 0.004 ±0.003 | 0.078 ±0.044 | 0.021 ±0.015 | 0.047 ±0.028 | 0.041 ±0.024 | 0.035 ±0.021 | 0.019 ±0.013 | 0.046 ±0.027 | 0.018 ±0.013 | 0.024 ±0.021 | 0.023 ±0.020 | 0.022 ±0.019 | 0.023 ±0.020 | 0.023 ±0.020 |
| $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ | 0.005 ±0.004 | 0.102 ±0.060 | 0.022 ±0.016 | 0.061 ±0.036 | 0.051 ±0.030 | 0.045 ±0.027 | 0.021 ±0.015 | 0.056 ±0.032 | 0.017 ±0.012 | 0.021 ±0.019 | 0.019 ±0.018 | 0.018 ±0.017 | 0.019 ±0.017 | 0.019 ±0.017 |

# I Detailed Results

# J Descriptions on Unlearning Algorithms with Full Access to Original Dataset

- *FOUL* Warnecke et al. (2023): This method introduces small perturbations to the target feature and updates the neural network by exploiting first-order gradient differences between the perturbed and original inputs.

- *MIUL* Guo et al. (2022): This approach aims to diminish the mutual information between the target feature and the label while preserving other learned information. It relies on mutual information estimation and requires training three additional neural networks.
- *Shuffle* Yang et al. (2025): This method randomly shuffles the target feature and fine-tunes the original model using the resulting shuffled dataset to achieve unlearning.

## K TRAINING LOSS CURVES FOR RETRAIN-FROM-SCRATCH MODELS

Figure 3 shows the training loss curves for the *Blood* dataset.

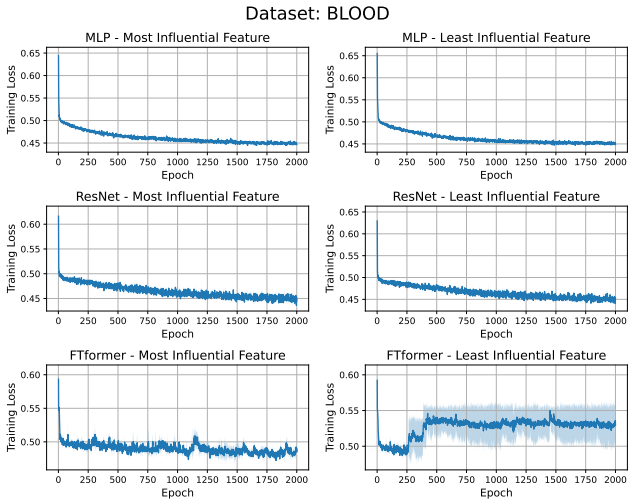


Figure 3: Training loss curves of the retrain-from-scratch model for the *Blood* dataset.

Figure 4 shows the training loss curves for the *Cali* dataset.

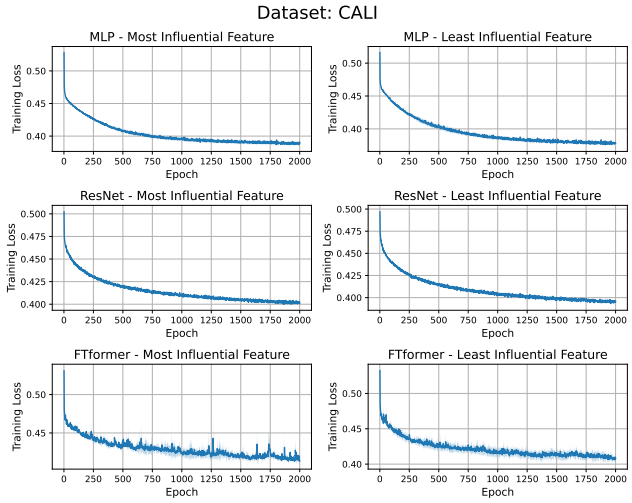


Figure 4: Training loss curves of the retrain-from-scratch model for the *Cali* dataset.

Figure 5 shows the training loss curves for the *Credit* dataset.

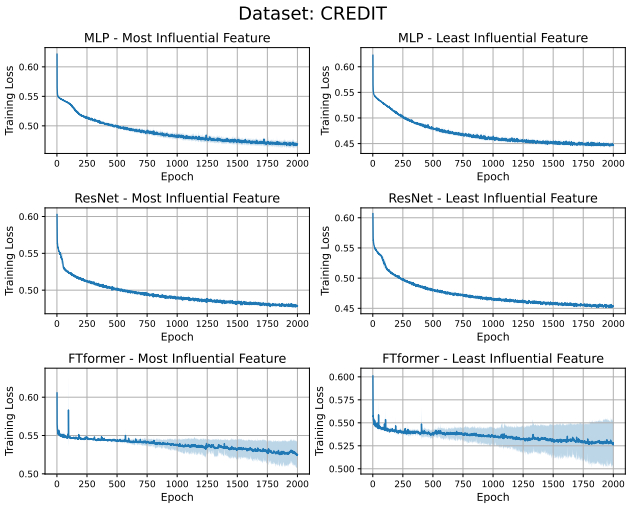


Figure 5: Training loss curves of the retrain-from-scratch model for the *Credit* dataset.

Figure 6 shows the training loss curves for the *Diabetes* dataset.

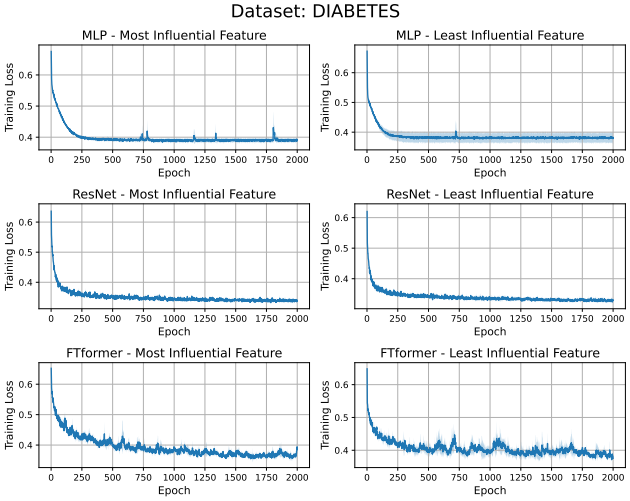


Figure 6: Training loss curves of the retrain-from-scratch model for the *Diabetes* dataset.

Figure 7 shows the training loss curves for the *House* dataset.

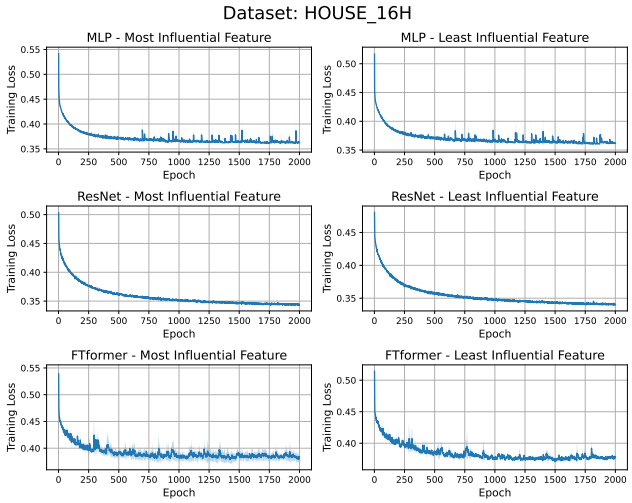


Figure 7: Training loss curves of the retrain-from-scratch model for the *House* dataset.

Figure 8 shows the training loss curves for the *Magic* dataset.

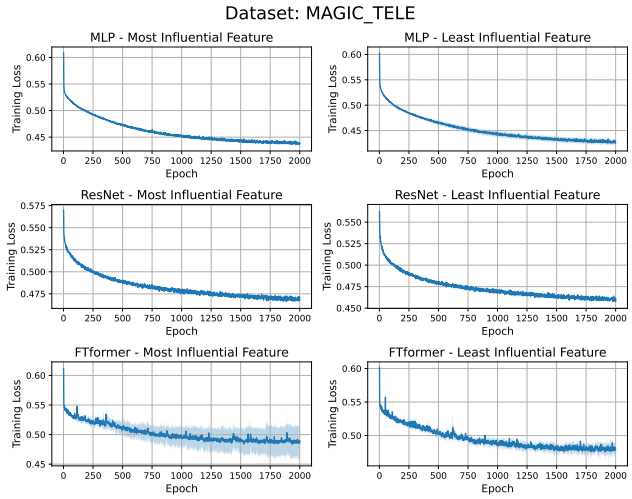


Figure 8: Training loss curves of the retrain-from-scratch model for the *Magic* dataset.

Figure 9 shows the training loss curves for the *Pol* dataset.

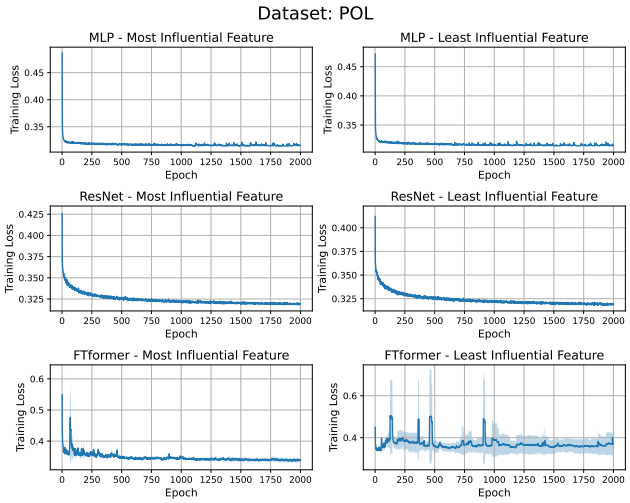


Figure 9: Training loss curves of the retrain-from-scratch model for the *Pol* dataset.

## L HYPERPARAMETER CONFIGURATION

The batch size is set as 64 and the learning rate is set as 0.001. The optimizer used for neural network update is the Adam optimizer and the loss function is the cross-entropy loss function.

For all classification experiments, we compute binary cross-entropy on the bounded effective logit

$$\widetilde{f}_\theta(x) = K \tanh(f_\theta(x)/K),$$

matching Assumption 4.4 and Theorem 4.5. Unless otherwise stated, we use $K = 5$. We additionally report sensitivity results for $K \in \{3, 5, 10\}$. All classification probabilities are computed as

$$p_\theta(Y = 1 \mid x) = \sigma(\widetilde{f}_\theta(x)).$$

For direct certificate audits, we report $D_q(\widetilde{f}_\theta)$ and the probability-level analogue $D_q(\sigma \circ \widetilde{f}_\theta)$.

The MLP model has two hidden layers with the number of neurons of $128$ and $64$. The ResNet model has two residual blocks and a final fully-connected layer for classification outputs. The FTformer model has two transformer blocks with a two-layer MLP head (whose hidden layer dimension is $64$) for classification.

For the *FOUL* baseline Warnecke et al. (2023), the scale of permutation on the unlearned feature is set as $0.01$.

For the *MIUL* baseline Guo et al. (2022), the training epochs to estimate mutual information is set as $500$. The weight coefficients balancing different mutual information estimations, denoted as $\lambda_1$, $\lambda_2$, and $\lambda_3$, are set as $5$, $5$, and $1$, respectively.

All codes are implemented in Python 3.12.12 and PyTorch 2.10.0.

## M IMPACT STATEMENT

Feature unlearning is motivated by legitimate needs to comply with privacy, regulatory, contractual, and organizational requirements to remove reliance on sensitive or invalid attributes from deployed models. BMCR provides a lightweight post-hoc editing procedure for settings where the per-example values of the removed feature cannot be re-accessed.

However, BMCR should not be interpreted as a complete privacy or fairness solution. The formal guarantee controls direct coordinate-level dependence under the replacement distribution; it does not guarantee parameter-level deletion, remove all information about the target feature from correlated retained features, or ensure demographic fairness. In sensitive deployments, BMCR should therefore be combined with leakage audits, fairness evaluations, and domain-specific governance procedures.

## N CALCULATIONS OF PRIVACY AND FAIRNESS METRICS

To evaluate privacy, we adopt an MIA-based metric. The purpose of MIA is to examine whether an attack model can distinguish the unlearned model from the retrained-from-scratch model based on their output behaviors. Specifically, given the same input samples, we collect the prediction outputs from both the unlearned model and the retrained-from-scratch model, and train an attack classifier, i.e., a Random Forest binary classier used in our study, to identify which model produces each output.

The attack accuracy is used as the privacy metric. It is calculated as

$$\text{MIA Accuracy} = \frac{\text{Number of correctly classified model outputs}}{\text{Total number of evaluated model outputs}}. \tag{28}$$

If the attack accuracy is close to $50\%$, the attack classifier performs similarly to random guessing in the binary setting. This indicates that the unlearned model is difficult to distinguish from the retrained-from-scratch model, suggesting that the unlearning process has effectively removed observable information associated with the target feature.

Demographic parity (DP) evaluates whether different sensitive groups receive the favorable prediction outcome at similar rates. We first identify several key components for DP calculation:

- **Privileged and unprivileged groups**: The privileged group refers to the group that has historically received advantages, while the unprivileged group refers to the group that has historically faced disadvantages, such as groups defined by sensitive attributes including "Race" or "Gender".
- **Favorable outcome**: The desired model prediction, such as "loan approved" in a loan approval task.
- **Model predictions**: The predicted labels generated by either the unlearned model or the retrained-from-scratch model.

For each group, we calculate the rate at which samples receive the favorable outcome:

$$\text{Rate}(g) = \frac{\text{Number of individuals in group } g \text{ receiving the favorable outcome}}{\text{Total number of individuals in group } g}, \tag{29}$$

where $g$ denotes either the privileged group or the unprivileged group. Accordingly, the favorable outcome rates for the privileged and unprivileged groups are denoted as

$$\text{DP}_{\text{priv}} = P(\hat{Y} = 1 \mid A = \text{privileged}), \tag{30}$$

and

$$\text{DP}_{\text{unpriv}} = P(\hat{Y} = 1 \mid A = \text{unprivileged}), \tag{31}$$

where $\hat{Y} = 1$ denotes the favorable predicted outcome and $A$ denotes the sensitive attribute.

The DP difference is then calculated as

$$\Delta_{\text{DP}} = \text{DP}_{\text{unpriv}} - \text{DP}_{\text{priv}}. \tag{32}$$

A DP difference closer to zero indicates a smaller disparity between the unprivileged and privileged groups.

Equalized odds (EO) evaluates whether the model has similar error behavior across sensitive groups. Unlike DP, which only considers the overall favorable prediction rate, EO compares the true positive rate (TPR) and false positive rate (FPR) between the privileged and unprivileged groups.

For each group $g$, the true positive rate is calculated as

$$\text{TPR}(g) = P(\hat{Y} = 1 \mid Y = 1, A = g), \tag{33}$$

where $Y = 1$ denotes the ground-truth favorable label. The false positive rate is calculated as

$$\text{FPR}(g) = P(\hat{Y} = 1 \mid Y = 0, A = g), \tag{34}$$

where $Y = 0$ denotes the ground-truth unfavorable label.

The TPR and FPR differences between the unprivileged and privileged groups are defined as

$$\Delta_{\text{TPR}} = \text{TPR}_{\text{unpriv}} - \text{TPR}_{\text{priv}}, \tag{35}$$

and

$$\Delta_{\text{FPR}} = \text{FPR}_{\text{unpriv}} - \text{FPR}_{\text{priv}}. \tag{36}$$

A value close to zero for both $\Delta_{\text{TPR}}$ and $\Delta_{\text{FPR}}$ indicates that the model exhibits similar predictive behavior across groups.

## O Feature Importance Ranking for Each Dataset

The background data size and test data size as 1000 and 100 for all datasets except *Blood* and *Diabetes*. For these two datasets, as their dataset sizes are lower than 1000, the background data size is set to 500.

The feature importance rankings (in the descending order) for each dataset are summarized below.

- *Blood*: `V2`, `V1`, `V4`, `V3`.
- *Cali*: `Latitude`, `Population`, `Longitude`, `AveRooms`, `HouseAge`, `AveBedrms`, `MedInc`, `AveOccup`.

- *Diabetes*: plas, age, pres, insu, mass, pedi, preg, skin.
- *Credit*: NumberOfOpenCreditLinesAndLoans, NumberOfDependents, NumberOfTime30-59DaysPastDueNotWorse, NumberOfTimes90DaysLate, NumberOfTime60-89DaysPastDueNotWorse, age, NumberRealEstateLoansOrLines, MonthlyIncome, RevolvingUtilizationOfUnsecuredLines, DebtRatio
- *House*: P27p4, P16p2, P1, P5p1, P11p4, P14p9, P15p1, P15p3, P18p2, H2p2, H10p1, H13p1, H18pA, H40p4, H8p2, P6p2.
- *Magic*: fConc:, fConc1:, fWidth:, fLength:, fSize:, fAsym:.
- *Pol*: f5, f6, f7, f8, f9, f13, f14, f15, f16, f17, f18, f19, f20, f21, f22, f23, f24, f25, f26, f27, f28, f29, f32, f31, f30, f33.

## P FULL RESULTS OF SINGLE-FEATURE UNLEARNING

Figure 10 shows the evaluation results of the most influential feature in the *Blood* dataset.

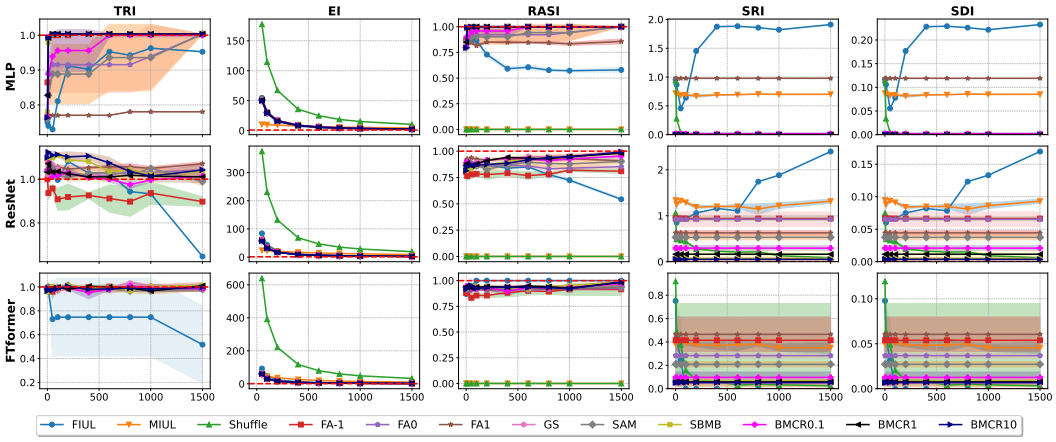


Figure 10: Evaluation results of the most influential feature in the *Blood* dataset.

Figure 11 shows the evaluation results of the least influential feature in the *Blood* dataset.

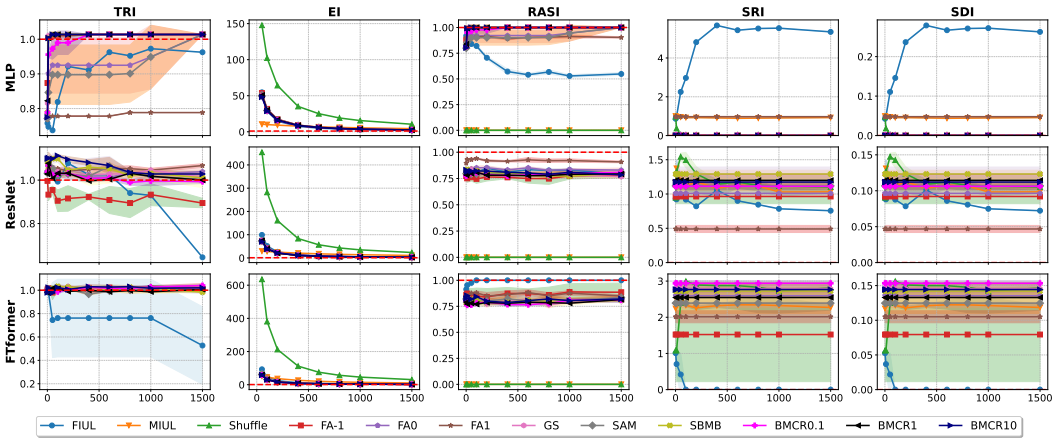


Figure 11: Evaluation results of the least influential feature in the *Blood* dataset.

Figure 12 shows the evaluation results of the most influential feature in the *Cali* dataset.

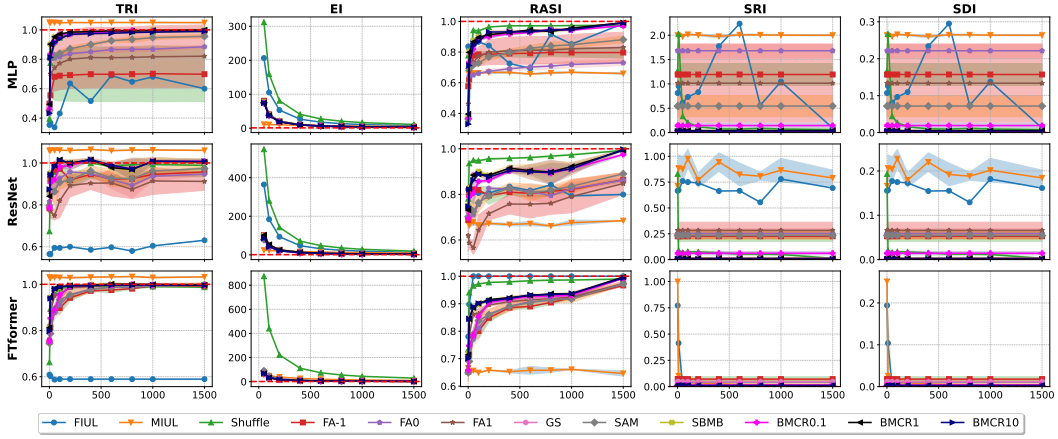


Figure 12: Evaluation results of the most influential feature in the *Cali* dataset.

Figure 13 shows the evaluation results of the least influential feature in the *Cali* dataset.

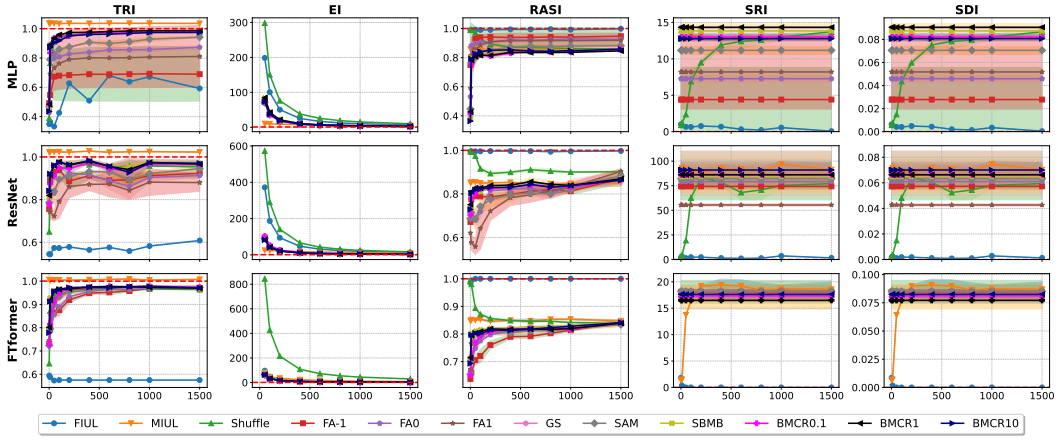


Figure 13: Evaluation results of the least influential feature in the *Cali* dataset.

Figure 14 shows the evaluation results of the most influential feature in the *Credit* dataset.

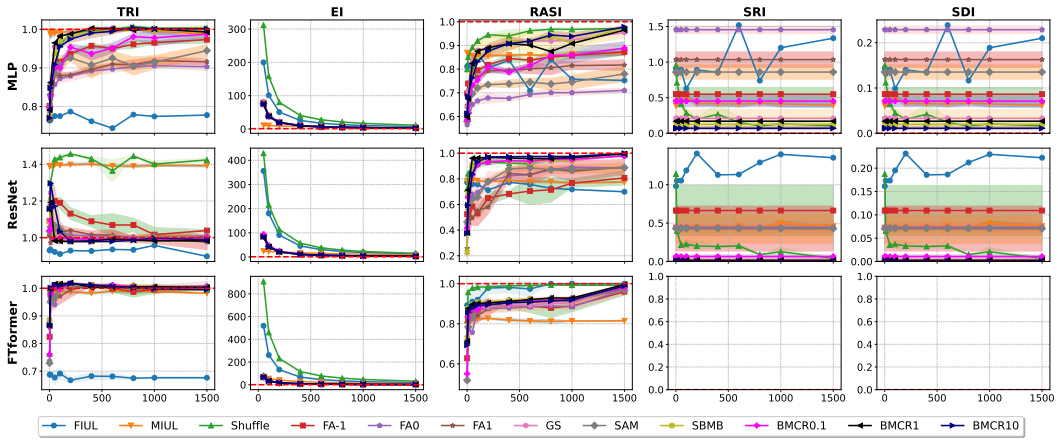


Figure 14: Evaluation results of the most influential feature in the *Credit* dataset.

Figure 15 shows the evaluation results of the least influential feature in the *Credit* dataset.

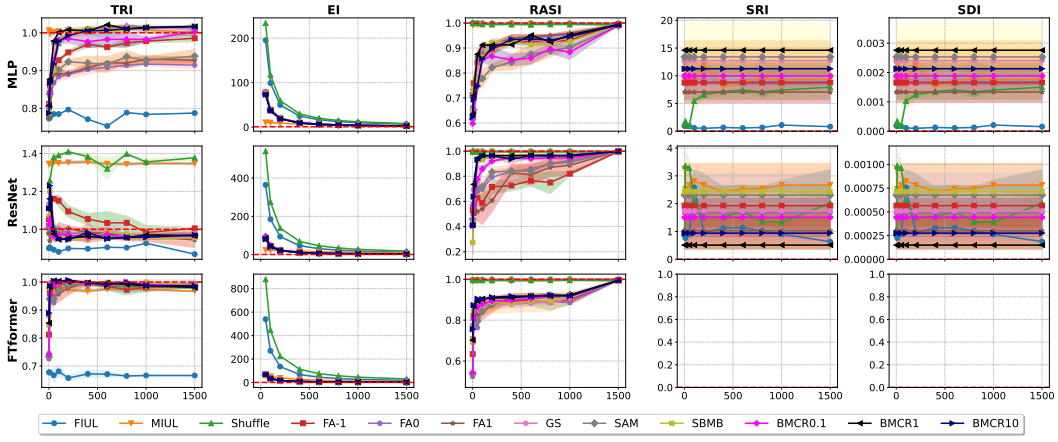


Figure 15: Evaluation results of the least influential feature in the *Credit* dataset.

Figure 16 shows the evaluation results of the most influential feature in the *Diabetes* dataset.

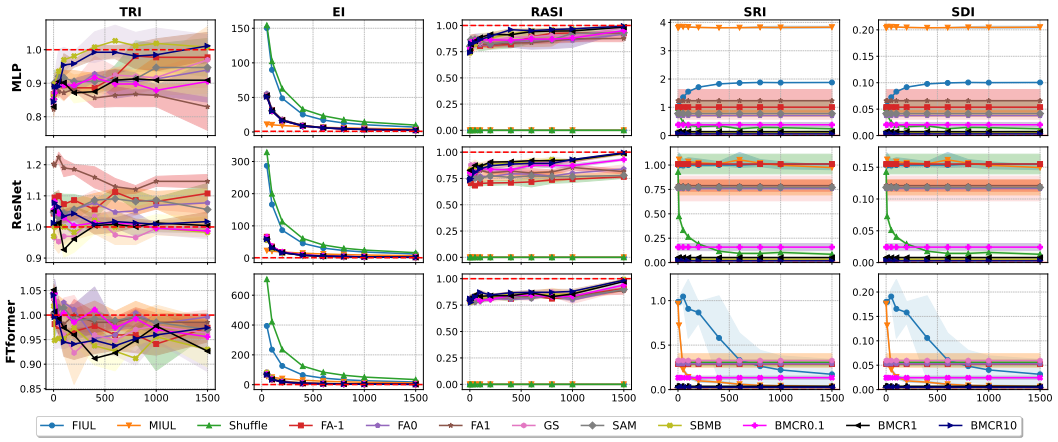


Figure 16: Evaluation results of the most influential feature in the *Diabetes* dataset.

Figure 17 shows the evaluation results of the least influential feature in the *Diabetes* dataset.

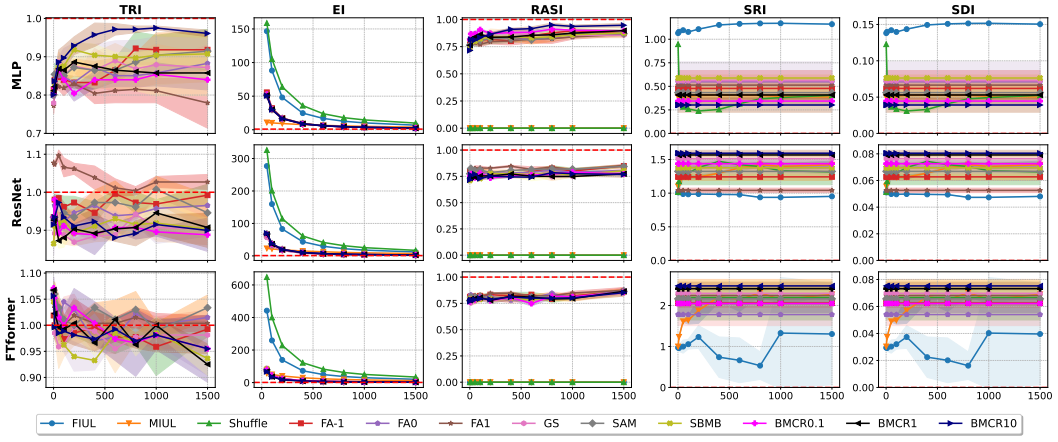


Figure 17: Evaluation results of the least influential feature in the *Diabetes* dataset.

Figure 18 shows the evaluation results of the most influential feature in the *House* dataset.

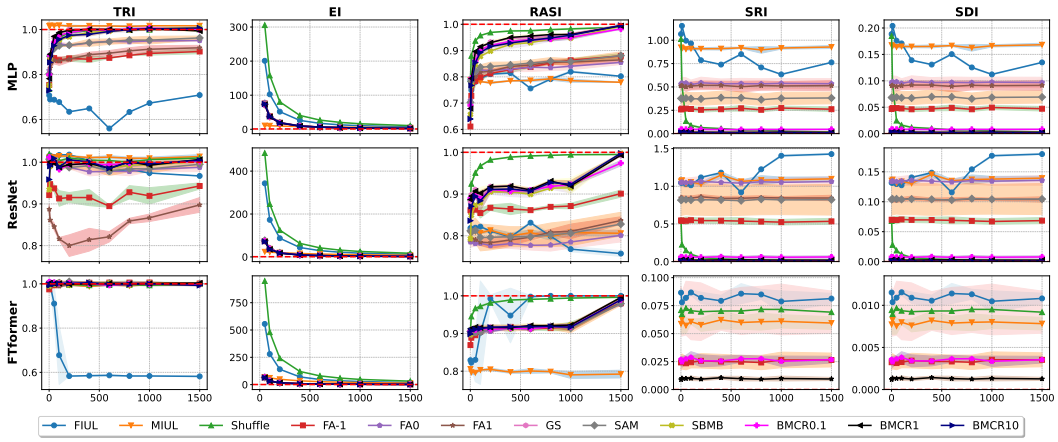


Figure 18: Evaluation results of the most influential feature in the *House* dataset.

Figure 19 shows the evaluation results of the least influential feature in the *House* dataset.

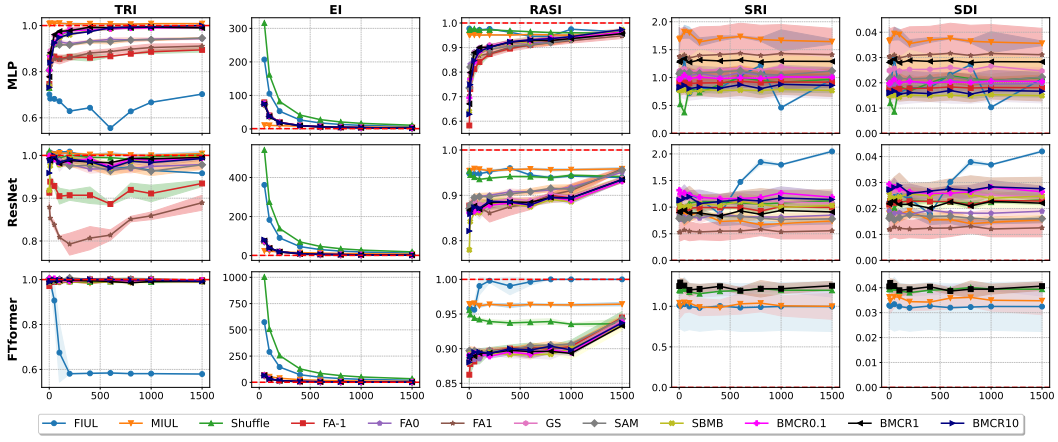


Figure 19: Evaluation results of the least influential feature in the *House* dataset.

Figure 20 shows the evaluation results of the least influential feature in the *Magic* dataset.

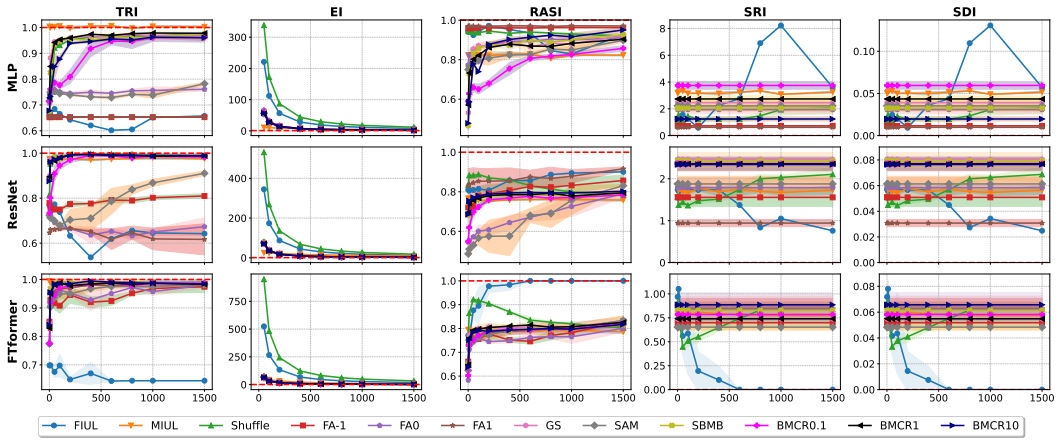


Figure 20: Evaluation results of the least influential feature in the *Magic* dataset.

Figure 2 shows the evaluation results of the most influential feature in the *Pol* dataset.

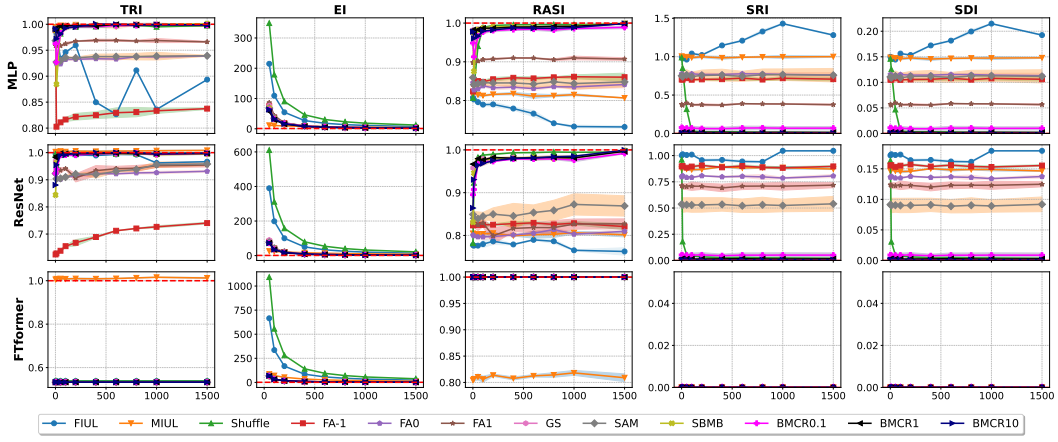


Figure 21: Evaluation results of the most influential feature in the *Pol* dataset.

Figure 22 shows the evaluation results of the least influential feature in the *Pol* dataset.

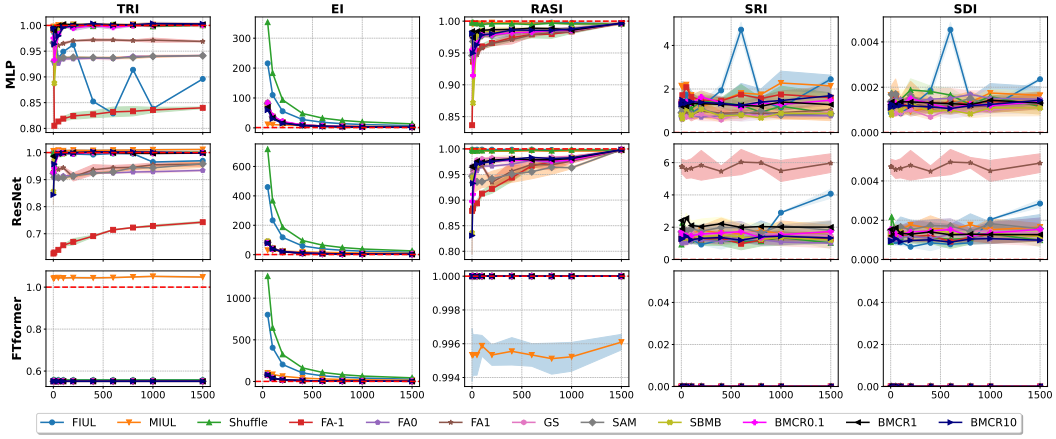


Figure 22: Evaluation results of the least influential feature in the *Pol* dataset.

# Q FULL RESULTS OF SINGLE-FEATURE UNLEARNING PERFORMANCE BY USING BMCR

Table 20 shows the evaluation results of FTformer when using BMCR with the range of $[-0.1, 0.1]$.

Table 20: Evaluation results of FTformer when using BMCR with the range of $[-0.1, 0.1]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0295 \pm 0.0238$ | $0.0212 \pm 0.0169$ | $0.0168 \pm 0.0136$ | $0.0142 \pm 0.0117$ | $0.0125 \pm 0.0104$ | $0.0113 \pm 0.0095$ | $0.0106 \pm 0.0090$ | $0.0101 \pm 0.0086$ | $0.0097 \pm 0.0083$ | $0.0088 \pm 0.0076$ |
| TRI | $0.840 \pm 0.173$ | $0.901 \pm 0.160$ | $0.914 \pm 0.156$ | $0.923 \pm 0.156$ | $0.932 \pm 0.160$ | $0.930 \pm 0.159$ | $0.928 \pm 0.158$ | $0.930 \pm 0.159$ | $0.929 \pm 0.158$ | $0.926 \pm 0.158$ |
| EI | $2314.0 \pm 1009.0$ | $308.7 \pm 37.1$ | $67.9 \pm 6.0$ | $34.5 \pm 3.8$ | $17.5 \pm 2.1$ | $8.9 \pm 1.1$ | $6.0 \pm 0.7$ | $4.5 \pm 0.6$ | $3.6 \pm 0.4$ | $2.4 \pm 0.3$ |
| RASI | $0.768 \pm 0.156$ | $0.833 \pm 0.094$ | $0.855 \pm 0.080$ | $0.867 \pm 0.073$ | $0.876 \pm 0.071$ | $0.881 \pm 0.067$ | $0.882 \pm 0.075$ | $0.889 \pm 0.068$ | $0.892 \pm 0.066$ | $0.936 \pm 0.069$ |
| SRI | $0.146 \pm 0.220$ | $0.150 \pm 0.265$ | $0.134 \pm 0.172$ | $0.155 \pm 0.266$ | $0.154 \pm 0.272$ | $0.148 \pm 0.269$ | $0.167 \pm 0.270$ | $0.125 \pm 0.186$ | $0.142 \pm 0.269$ | $0.142 \pm 0.245$ |
| SDI | $0.038 \pm 0.046$ | $0.039 \pm 0.045$ | $0.036 \pm 0.045$ | $0.035 \pm 0.044$ | $0.037 \pm 0.047$ | $0.039 \pm 0.047$ | $0.037 \pm 0.045$ | $0.036 \pm 0.043$ | $0.036 \pm 0.047$ | $0.037 \pm 0.044$ |
| $\Delta_{\text{DP}}$ | $0.0795 \pm 0.0610$ | $0.0412 \pm 0.0341$ | $0.0385 \pm 0.0318$ | $0.0215 \pm 0.0189$ | $0.0302 \pm 0.0264$ | $0.0208 \pm 0.0185$ | $0.0189 \pm 0.0168$ | $0.0245 \pm 0.0212$ | $0.0276 \pm 0.0235$ | $0.0158 \pm 0.0142$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0472 \pm 0.0401$ | $0.0358 \pm 0.0304$ | $0.0345 \pm 0.0288$ | $0.0221 \pm 0.0195$ | $0.0316 \pm 0.0268$ | $0.0272 \pm 0.0235$ | $0.0234 \pm 0.0202$ | $0.0283 \pm 0.0241$ | $0.0256 \pm 0.0217$ | $0.0241 \pm 0.0208$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1326 \pm 0.0945$ | $0.0552 \pm 0.0461$ | $0.0468 \pm 0.0392$ | $0.0234 \pm 0.0208$ | $0.0338 \pm 0.0297$ | $0.0232 \pm 0.0209$ | $0.0185 \pm 0.0169$ | $0.0281 \pm 0.0246$ | $0.0402 \pm 0.0341$ | $0.0208 \pm 0.0189$ |

Table 21 shows the evaluation results of FTformer when using BMCR with the range of $[-0.5, 0.5]$.

Table 21: Evaluation results of FTformer when using BMCR with the range of $[-0.5, 0.5]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0281 \pm 0.0227$ | $0.0201 \pm 0.0163$ | $0.0159 \pm 0.0130$ | $0.0134 \pm 0.0112$ | $0.0118 \pm 0.0101$ | $0.0107 \pm 0.0092$ | $0.0100 \pm 0.0086$ | $0.0096 \pm 0.0082$ | $0.0092 \pm 0.0080$ | $0.0083 \pm 0.0072$ |
| TRI | $0.856 \pm 0.166$ | $0.912 \pm 0.158$ | $0.924 \pm 0.158$ | $0.927 \pm 0.158$ | $0.930 \pm 0.159$ | $0.927 \pm 0.159$ | $0.927 \pm 0.158$ | $0.927 \pm 0.158$ | $0.928 \pm 0.158$ | $0.922 \pm 0.157$ |
| EI | $2307.7 \pm 1003.4$ | $307.9 \pm 38.8$ | $66.8 \pm 4.7$ | $33.8 \pm 2.8$ | $17.2 \pm 1.5$ | $8.7 \pm 0.7$ | $5.9 \pm 0.5$ | $4.4 \pm 0.4$ | $3.5 \pm 0.3$ | $2.3 \pm 0.2$ |
| RASI | $0.792 \pm 0.132$ | $0.850 \pm 0.082$ | $0.868 \pm 0.074$ | $0.875 \pm 0.071$ | $0.881 \pm 0.070$ | $0.885 \pm 0.066$ | $0.889 \pm 0.069$ | $0.890 \pm 0.069$ | $0.892 \pm 0.067$ | $0.941 \pm 0.070$ |
| SRI | $0.126 \pm 0.224$ | $0.126 \pm 0.252$ | $0.124 \pm 0.214$ | $0.116 \pm 0.223$ | $0.109 \pm 0.213$ | $0.129 \pm 0.263$ | $0.138 \pm 0.265$ | $0.118 \pm 0.226$ | $0.111 \pm 0.236$ | $0.131 \pm 0.255$ |
| SDI | $0.034 \pm 0.044$ | $0.036 \pm 0.043$ | $0.033 \pm 0.042$ | $0.032 \pm 0.041$ | $0.034 \pm 0.042$ | $0.036 \pm 0.043$ | $0.034 \pm 0.043$ | $0.035 \pm 0.043$ | $0.033 \pm 0.043$ | $0.036 \pm 0.045$ |
| $\Delta_{\text{DP}}$ | $0.0772 \pm 0.0594$ | $0.0386 \pm 0.0318$ | $0.0364 \pm 0.0301$ | $0.0195 \pm 0.0173$ | $0.0292 \pm 0.0253$ | $0.0192 \pm 0.0169$ | $0.0176 \pm 0.0160$ | $0.0237 \pm 0.0209$ | $0.0267 \pm 0.0230$ | $0.0145 \pm 0.0133$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0454 \pm 0.0385$ | $0.0338 \pm 0.0288$ | $0.0337 \pm 0.0282$ | $0.0205 \pm 0.0182$ | $0.0308 \pm 0.0260$ | $0.0265 \pm 0.0226$ | $0.0224 \pm 0.0192$ | $0.0272 \pm 0.0232$ | $0.0246 \pm 0.0209$ | $0.0231 \pm 0.0199$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1299 \pm 0.0929$ | $0.0524 \pm 0.0440$ | $0.0446 \pm 0.0375$ | $0.0217 \pm 0.0197$ | $0.0316 \pm 0.0278$ | $0.0214 \pm 0.0193$ | $0.0172 \pm 0.0159$ | $0.0262 \pm 0.0232$ | $0.0384 \pm 0.0324$ | $0.0192 \pm 0.0175$ |

Table 22 shows the evaluation results of FTformer when using BMCR with the range of $[-1, 1]$.

Table 22: Evaluation results of FTformer when using BMCR with the range of $[-1, 1]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0272 \pm 0.0221$ | $0.0194 \pm 0.0158$ | $0.0152 \pm 0.0125$ | $0.0129 \pm 0.0108$ | $0.0114 \pm 0.0097$ | $0.0102 \pm 0.0088$ | $0.0095 \pm 0.0082$ | $0.0091 \pm 0.0079$ | $0.0088 \pm 0.0076$ | $0.0079 \pm 0.0069$ |
| TRI | $0.868 \pm 0.159$ | $0.919 \pm 0.156$ | $0.930 \pm 0.159$ | $0.930 \pm 0.159$ | $0.927 \pm 0.158$ | $0.923 \pm 0.158$ | $0.926 \pm 0.158$ | $0.924 \pm 0.157$ | $0.926 \pm 0.157$ | $0.920 \pm 0.156$ |
| EI | $2303.6 \pm 999.6$ | $307.4 \pm 39.9$ | $66.0 \pm 3.9$ | $33.4 \pm 2.2$ | $17.0 \pm 1.1$ | $8.6 \pm 0.5$ | $5.8 \pm 0.4$ | $4.4 \pm 0.3$ | $3.5 \pm 0.2$ | $2.3 \pm 0.1$ |
| RASI | $0.808 \pm 0.116$ | $0.862 \pm 0.075$ | $0.877 \pm 0.069$ | $0.880 \pm 0.069$ | $0.884 \pm 0.069$ | $0.888 \pm 0.066$ | $0.894 \pm 0.065$ | $0.890 \pm 0.070$ | $0.892 \pm 0.068$ | $0.944 \pm 0.071$ |
| SRI | $0.113 \pm 0.226$ | $0.110 \pm 0.244$ | $0.117 \pm 0.246$ | $0.090 \pm 0.194$ | $0.079 \pm 0.175$ | $0.117 \pm 0.261$ | $0.119 \pm 0.263$ | $0.113 \pm 0.246$ | $0.090 \pm 0.223$ | $0.123 \pm 0.261$ |
| SDI | $0.032 \pm 0.043$ | $0.035 \pm 0.042$ | $0.031 \pm 0.041$ | $0.030 \pm 0.039$ | $0.032 \pm 0.040$ | $0.034 \pm 0.041$ | $0.033 \pm 0.042$ | $0.035 \pm 0.043$ | $0.032 \pm 0.042$ | $0.035 \pm 0.045$ |
| $\Delta_{\text{DP}}$ | $0.0758 \pm 0.0584$ | $0.0369 \pm 0.0305$ | $0.0354 \pm 0.0292$ | $0.0188 \pm 0.0167$ | $0.0286 \pm 0.0248$ | $0.0184 \pm 0.0163$ | $0.0169 \pm 0.0155$ | $0.0232 \pm 0.0206$ | $0.0261 \pm 0.0228$ | $0.0139 \pm 0.0129$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0441 \pm 0.0376$ | $0.0324 \pm 0.0276$ | $0.0331 \pm 0.0278$ | $0.0195 \pm 0.0174$ | $0.0302 \pm 0.0256$ | $0.0261 \pm 0.0222$ | $0.0218 \pm 0.0187$ | $0.0265 \pm 0.0226$ | $0.0239 \pm 0.0205$ | $0.0224 \pm 0.0193$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1282 \pm 0.0918$ | $0.0505 \pm 0.0424$ | $0.0431 \pm 0.0365$ | $0.0206 \pm 0.0189$ | $0.0304 \pm 0.0268$ | $0.0209 \pm 0.0188$ | $0.0165 \pm 0.0153$ | $0.0251 \pm 0.0222$ | $0.0372 \pm 0.0315$ | $0.0181 \pm 0.0166$ |

Table 23 shows the evaluation results of FTformer when using BMCR with the range of $[-5, 5]$.

Table 23: Evaluation results of FTformer when using BMCR with the range of $[-5, 5]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0276 \pm 0.0224$ | $0.0198 \pm 0.0161$ | $0.0155 \pm 0.0128$ | $0.0132 \pm 0.0110$ | $0.0117 \pm 0.0099$ | $0.0105 \pm 0.0090$ | $0.0098 \pm 0.0084$ | $0.0093 \pm 0.0080$ | $0.0089 \pm 0.0077$ | $0.0081 \pm 0.0070$ |
| TRI | $0.867 \pm 0.158$ | $0.918 \pm 0.156$ | $0.927 \pm 0.158$ | $0.929 \pm 0.159$ | $0.928 \pm 0.159$ | $0.926 \pm 0.159$ | $0.927 \pm 0.158$ | $0.926 \pm 0.158$ | $0.926 \pm 0.157$ | $0.923 \pm 0.157$ |
| EI | $2335.1 \pm 1068.7$ | $309.1 \pm 44.1$ | $66.3 \pm 4.6$ | $33.4 \pm 2.4$ | $16.9 \pm 1.1$ | $8.6 \pm 0.5$ | $5.8 \pm 0.4$ | $4.4 \pm 0.3$ | $3.5 \pm 0.2$ | $2.3 \pm 0.1$ |
| RASI | $0.807 \pm 0.116$ | $0.862 \pm 0.075$ | $0.878 \pm 0.070$ | $0.883 \pm 0.066$ | $0.884 \pm 0.069$ | $0.889 \pm 0.067$ | $0.893 \pm 0.065$ | $0.893 \pm 0.066$ | $0.894 \pm 0.066$ | $0.944 \pm 0.072$ |
| SRI | $0.123 \pm 0.260$ | $0.136 \pm 0.285$ | $0.135 \pm 0.274$ | $0.120 \pm 0.241$ | $0.111 \pm 0.242$ | $0.130 \pm 0.288$ | $0.131 \pm 0.279$ | $0.126 \pm 0.278$ | $0.112 \pm 0.262$ | $0.130 \pm 0.275$ |
| SDI | $0.036 \pm 0.046$ | $0.037 \pm 0.045$ | $0.035 \pm 0.043$ | $0.034 \pm 0.043$ | $0.036 \pm 0.044$ | $0.038 \pm 0.047$ | $0.035 \pm 0.043$ | $0.036 \pm 0.045$ | $0.036 \pm 0.045$ | $0.037 \pm 0.046$ |
| $\Delta_{\text{DP}}$ | $0.0766 \pm 0.0590$ | $0.0376 \pm 0.0312$ | $0.0366 \pm 0.0302$ | $0.0193 \pm 0.0172$ | $0.0291 \pm 0.0253$ | $0.0188 \pm 0.0166$ | $0.0172 \pm 0.0157$ | $0.0235 \pm 0.0208$ | $0.0264 \pm 0.0230$ | $0.0143 \pm 0.0132$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0448 \pm 0.0380$ | $0.0331 \pm 0.0282$ | $0.0337 \pm 0.0284$ | $0.0199 \pm 0.0178$ | $0.0305 \pm 0.0259$ | $0.0265 \pm 0.0225$ | $0.0220 \pm 0.0189$ | $0.0268 \pm 0.0229$ | $0.0242 \pm 0.0207$ | $0.0228 \pm 0.0195$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1293 \pm 0.0925$ | $0.0513 \pm 0.0431$ | $0.0438 \pm 0.0370$ | $0.0212 \pm 0.0192$ | $0.0312 \pm 0.0273$ | $0.0213 \pm 0.0191$ | $0.0168 \pm 0.0156$ | $0.0256 \pm 0.0227$ | $0.0378 \pm 0.0321$ | $0.0186 \pm 0.0170$ |

Table 24 shows the evaluation results of FTformer when using BMCR with the range of $[-10, 10]$.

Table 24: Evaluation results of FTformer when using BMCR with the range of $[-10, 10]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0280 \pm 0.0228$ | $0.0201 \pm 0.0163$ | $0.0158 \pm 0.0130$ | $0.0134 \pm 0.0112$ | $0.0119 \pm 0.0101$ | $0.0107 \pm 0.0092$ | $0.0100 \pm 0.0086$ | $0.0094 \pm 0.0081$ | $0.0090 \pm 0.0078$ | $0.0082 \pm 0.0071$ |
| TRI | $0.866 \pm 0.157$ | $0.917 \pm 0.156$ | $0.923 \pm 0.156$ | $0.927 \pm 0.158$ | $0.928 \pm 0.159$ | $0.929 \pm 0.159$ | $0.927 \pm 0.158$ | $0.928 \pm 0.158$ | $0.926 \pm 0.157$ | $0.925 \pm 0.157$ |
| EI | $2366.6 \pm 1137.7$ | $310.7 \pm 48.3$ | $66.5 \pm 5.2$ | $33.4 \pm 2.5$ | $16.9 \pm 1.2$ | $8.6 \pm 0.6$ | $5.8 \pm 0.4$ | $4.3 \pm 0.3$ | $3.5 \pm 0.2$ | $2.3 \pm 0.1$ |
| RASI | $0.806 \pm 0.115$ | $0.861 \pm 0.075$ | $0.878 \pm 0.070$ | $0.885 \pm 0.063$ | $0.883 \pm 0.068$ | $0.889 \pm 0.067$ | $0.892 \pm 0.065$ | $0.895 \pm 0.063$ | $0.895 \pm 0.064$ | $0.943 \pm 0.072$ |
| SRI | $0.132 \pm 0.294$ | $0.162 \pm 0.325$ | $0.152 \pm 0.303$ | $0.150 \pm 0.295$ | $0.142 \pm 0.310$ | $0.143 \pm 0.314$ | $0.142 \pm 0.295$ | $0.142 \pm 0.310$ | $0.142 \pm 0.312$ | $0.138 \pm 0.297$ |
| SDI | $0.040 \pm 0.049$ | $0.039 \pm 0.048$ | $0.038 \pm 0.045$ | $0.038 \pm 0.047$ | $0.039 \pm 0.047$ | $0.041 \pm 0.053$ | $0.036 \pm 0.043$ | $0.037 \pm 0.046$ | $0.039 \pm 0.047$ | $0.038 \pm 0.048$ |
| $\Delta_{\text{DP}}$ | $0.0772 \pm 0.0596$ | $0.0381 \pm 0.0317$ | $0.0378 \pm 0.0312$ | $0.0199 \pm 0.0175$ | $0.0294 \pm 0.0257$ | $0.0191 \pm 0.0168$ | $0.0174 \pm 0.0158$ | $0.0238 \pm 0.0209$ | $0.0268 \pm 0.0232$ | $0.0146 \pm 0.0134$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0453 \pm 0.0384$ | $0.0336 \pm 0.0287$ | $0.0342 \pm 0.0289$ | $0.0202 \pm 0.0181$ | $0.0308 \pm 0.0262$ | $0.0268 \pm 0.0229$ | $0.0223 \pm 0.0191$ | $0.0271 \pm 0.0232$ | $0.0244 \pm 0.0209$ | $0.0231 \pm 0.0198$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1304 \pm 0.0932$ | $0.0521 \pm 0.0438$ | $0.0445 \pm 0.0375$ | $0.0218 \pm 0.0196$ | $0.0319 \pm 0.0279$ | $0.0216 \pm 0.0193$ | $0.0171 \pm 0.0158$ | $0.0262 \pm 0.0231$ | $0.0384 \pm 0.0326$ | $0.0192 \pm 0.0174$ |

Table 25 shows the evaluation results of MLP when using BMCR with the range of $[-0.1, 0.1]$.

Table 25: Evaluation results of MLP when using BMCR with the range of $[-0.1, 0.1]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0346 \pm 0.0275$ | $0.0258 \pm 0.0206$ | $0.0204 \pm 0.0168$ | $0.0169 \pm 0.0141$ | $0.0143 \pm 0.0121$ | $0.0120 \pm 0.0102$ | $0.0107 \pm 0.0091$ | $0.0098 \pm 0.0084$ | $0.0092 \pm 0.0079$ | $0.0084 \pm 0.0072$ |
| TRI | $0.767 \pm 0.140$ | $0.859 \pm 0.055$ | $0.911 \pm 0.061$ | $0.927 \pm 0.072$ | $0.935 \pm 0.068$ | $0.960 \pm 0.044$ | $0.967 \pm 0.045$ | $0.972 \pm 0.046$ | $0.973 \pm 0.046$ | $0.976 \pm 0.045$ |
| EI | $1398.6 \pm 803.0$ | $283.9 \pm 114.1$ | $65.3 \pm 10.8$ | $34.0 \pm 3.9$ | $17.8 \pm 1.8$ | $8.9 \pm 0.8$ | $5.9 \pm 0.5$ | $4.5 \pm 0.4$ | $3.6 \pm 0.3$ | $2.4 \pm 0.2$ |
| RASI | $0.686 \pm 0.173$ | $0.780 \pm 0.119$ | $0.832 \pm 0.110$ | $0.861 \pm 0.107$ | $0.870 \pm 0.097$ | $0.891 \pm 0.075$ | $0.906 \pm 0.066$ | $0.918 \pm 0.059$ | $0.921 \pm 0.057$ | $0.946 \pm 0.052$ |
| SRI | $0.214 \pm 0.209$ | $0.317 \pm 0.392$ | $0.180 \pm 0.178$ | $0.268 \pm 0.343$ | $0.153 \pm 0.143$ | $0.258 \pm 0.277$ | $0.240 \pm 0.239$ | $0.268 \pm 0.294$ | $0.176 \pm 0.167$ | $0.180 \pm 0.163$ |
| SDI | $0.030 \pm 0.030$ | $0.025 \pm 0.024$ | $0.028 \pm 0.028$ | $0.030 \pm 0.027$ | $0.028 \pm 0.027$ | $0.031 \pm 0.031$ | $0.027 \pm 0.028$ | $0.027 \pm 0.025$ | $0.031 \pm 0.029$ | $0.030 \pm 0.029$ |
| $\Delta_{\text{DP}}$ | $0.0868 \pm 0.0665$ | $0.0435 \pm 0.0358$ | $0.0398 \pm 0.0327$ | $0.0206 \pm 0.0182$ | $0.0314 \pm 0.0271$ | $0.0192 \pm 0.0169$ | $0.0171 \pm 0.0154$ | $0.0228 \pm 0.0198$ | $0.0255 \pm 0.0217$ | $0.0136 \pm 0.0125$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0508 \pm 0.0432$ | $0.0382 \pm 0.0326$ | $0.0357 \pm 0.0301$ | $0.0218 \pm 0.0191$ | $0.0328 \pm 0.0279$ | $0.0264 \pm 0.0226$ | $0.0212 \pm 0.0183$ | $0.0254 \pm 0.0219$ | $0.0235 \pm 0.0201$ | $0.0219 \pm 0.0187$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1425 \pm 0.1008$ | $0.0584 \pm 0.0487$ | $0.0475 \pm 0.0398$ | $0.0228 \pm 0.0204$ | $0.0332 \pm 0.0291$ | $0.0221 \pm 0.0198$ | $0.0175 \pm 0.0161$ | $0.0258 \pm 0.0226$ | $0.0368 \pm 0.0312$ | $0.0182 \pm 0.0167$ |

Table 26 shows the evaluation results of MLP when using BMCR with the range of $[-0.5, 0.5]$.

Table 26: Evaluation results of MLP when using BMCR with the range of $[-0.5, 0.5]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0337 \pm 0.0270$ | $0.0242 \pm 0.0194$ | $0.0181 \pm 0.0148$ | $0.0151 \pm 0.0125$ | $0.0128 \pm 0.0108$ | $0.0110 \pm 0.0094$ | $0.0100 \pm 0.0085$ | $0.0093 \pm 0.0080$ | $0.0087 \pm 0.0075$ | $0.0079 \pm 0.0069$ |
| TRI | $0.772 \pm 0.140$ | $0.887 \pm 0.056$ | $0.936 \pm 0.050$ | $0.949 \pm 0.054$ | $0.956 \pm 0.055$ | $0.970 \pm 0.044$ | $0.974 \pm 0.043$ | $0.976 \pm 0.043$ | $0.978 \pm 0.044$ | $0.979 \pm 0.044$ |
| EI | $1382.5 \pm 811.0$ | $271.7 \pm 105.9$ | $64.7 \pm 10.9$ | $33.8 \pm 4.1$ | $17.6 \pm 1.9$ | $8.8 \pm 1.0$ | $5.9 \pm 0.7$ | $4.4 \pm 0.5$ | $3.5 \pm 0.4$ | $2.3 \pm 0.3$ |
| RASI | $0.689 \pm 0.176$ | $0.806 \pm 0.111$ | $0.860 \pm 0.091$ | $0.884 \pm 0.085$ | $0.894 \pm 0.077$ | $0.909 \pm 0.064$ | $0.919 \pm 0.060$ | $0.925 \pm 0.056$ | $0.931 \pm 0.052$ | $0.955 \pm 0.049$ |
| SRI | $0.169 \pm 0.178$ | $0.218 \pm 0.269$ | $0.143 \pm 0.150$ | $0.203 \pm 0.255$ | $0.144 \pm 0.158$ | $0.201 \pm 0.222$ | $0.185 \pm 0.191$ | $0.191 \pm 0.214$ | $0.144 \pm 0.145$ | $0.141 \pm 0.154$ |
| SDI | $0.025 \pm 0.028$ | $0.023 \pm 0.025$ | $0.024 \pm 0.027$ | $0.025 \pm 0.026$ | $0.025 \pm 0.027$ | $0.026 \pm 0.028$ | $0.023 \pm 0.026$ | $0.023 \pm 0.025$ | $0.025 \pm 0.028$ | $0.025 \pm 0.027$ |
| $\Delta_{\text{DP}}$ | $0.0842 \pm 0.0647$ | $0.0411 \pm 0.0339$ | $0.0370 \pm 0.0306$ | $0.0191 \pm 0.0169$ | $0.0296 \pm 0.0256$ | $0.0182 \pm 0.0161$ | $0.0164 \pm 0.0148$ | $0.0222 \pm 0.0194$ | $0.0247 \pm 0.0212$ | $0.0130 \pm 0.0120$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0487 \pm 0.0414$ | $0.0362 \pm 0.0309$ | $0.0337 \pm 0.0285$ | $0.0203 \pm 0.0180$ | $0.0312 \pm 0.0265$ | $0.0256 \pm 0.0219$ | $0.0209 \pm 0.0180$ | $0.0248 \pm 0.0214$ | $0.0229 \pm 0.0197$ | $0.0213 \pm 0.0183$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1397 \pm 0.0992$ | $0.0558 \pm 0.0467$ | $0.0451 \pm 0.0378$ | $0.0212 \pm 0.0192$ | $0.0315 \pm 0.0278$ | $0.0209 \pm 0.0188$ | $0.0166 \pm 0.0154$ | $0.0246 \pm 0.0217$ | $0.0357 \pm 0.0302$ | $0.0174 \pm 0.0159$ |

Table 27 shows the evaluation results of MLP when using BMCR with the range of $[-1, 1]$.

Table 27: Evaluation results of MLP when using BMCR with the range of $[-1, 1]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0328 \pm 0.0264$ | $0.0226 \pm 0.0182$ | $0.0165 \pm 0.0135$ | $0.0136 \pm 0.0113$ | $0.0117 \pm 0.0098$ | $0.0101 \pm 0.0086$ | $0.0093 \pm 0.0080$ | $0.0088 \pm 0.0076$ | $0.0083 \pm 0.0072$ | $0.0075 \pm 0.0066$ |
| TRI | $0.776 \pm 0.140$ | $0.910 \pm 0.056$ | $0.959 \pm 0.039$ | $0.970 \pm 0.040$ | $0.975 \pm 0.042$ | $0.978 \pm 0.043$ | $0.981 \pm 0.041$ | $0.981 \pm 0.041$ | $0.982 \pm 0.043$ | $0.982 \pm 0.042$ |
| EI | $1366.3 \pm 818.9$ | $259.5 \pm 97.8$ | $64.1 \pm 11.0$ | $33.6 \pm 4.3$ | $17.3 \pm 2.0$ | $8.7 \pm 1.1$ | $5.8 \pm 0.8$ | $4.4 \pm 0.6$ | $3.5 \pm 0.4$ | $2.3 \pm 0.3$ |
| RASI | $0.691 \pm 0.178$ | $0.832 \pm 0.103$ | $0.887 \pm 0.071$ | $0.907 \pm 0.062$ | $0.918 \pm 0.057$ | $0.926 \pm 0.052$ | $0.931 \pm 0.054$ | $0.931 \pm 0.053$ | $0.940 \pm 0.048$ | $0.964 \pm 0.046$ |
| SRI | $0.124 \pm 0.146$ | $0.119 \pm 0.146$ | $0.105 \pm 0.122$ | $0.138 \pm 0.169$ | $0.135 \pm 0.173$ | $0.144 \pm 0.167$ | $0.129 \pm 0.143$ | $0.113 \pm 0.133$ | $0.112 \pm 0.123$ | $0.101 \pm 0.145$ |
| SDI | $0.020 \pm 0.025$ | $0.020 \pm 0.026$ | $0.020 \pm 0.026$ | $0.019 \pm 0.025$ | $0.021 \pm 0.026$ | $0.020 \pm 0.025$ | $0.019 \pm 0.024$ | $0.018 \pm 0.024$ | $0.019 \pm 0.026$ | $0.020 \pm 0.024$ |
| $\Delta_{\text{DP}}$ | $0.0815 \pm 0.0628$ | $0.0388 \pm 0.0319$ | $0.0342 \pm 0.0284$ | $0.0175 \pm 0.0156$ | $0.0278 \pm 0.0241$ | $0.0172 \pm 0.0153$ | $0.0156 \pm 0.0142$ | $0.0215 \pm 0.0189$ | $0.0238 \pm 0.0206$ | $0.0124 \pm 0.0115$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0465 \pm 0.0395$ | $0.0341 \pm 0.0291$ | $0.0316 \pm 0.0268$ | $0.0188 \pm 0.0169$ | $0.0295 \pm 0.0251$ | $0.0248 \pm 0.0213$ | $0.0205 \pm 0.0177$ | $0.0241 \pm 0.0208$ | $0.0224 \pm 0.0193$ | $0.0206 \pm 0.0178$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1368 \pm 0.0975$ | $0.0532 \pm 0.0446$ | $0.0426 \pm 0.0358$ | $0.0195 \pm 0.0179$ | $0.0298 \pm 0.0264$ | $0.0197 \pm 0.0178$ | $0.0158 \pm 0.0147$ | $0.0234 \pm 0.0208$ | $0.0345 \pm 0.0292$ | $0.0165 \pm 0.0151$ |

Table 28 shows the evaluation results of MLP when using BMCR with the range of $[-5, 5]$.

Table 28: Evaluation results of MLP when using BMCR with the range of $[-5, 5]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0342 \pm 0.0274$ | $0.0235 \pm 0.0189$ | $0.0171 \pm 0.0140$ | $0.0140 \pm 0.0116$ | $0.0119 \pm 0.0100$ | $0.0102 \pm 0.0087$ | $0.0092 \pm 0.0079$ | $0.0086 \pm 0.0074$ | $0.0081 \pm 0.0071$ | $0.0073 \pm 0.0064$ |
| TRI | $0.760 \pm 0.148$ | $0.894 \pm 0.066$ | $0.945 \pm 0.043$ | $0.962 \pm 0.039$ | $0.974 \pm 0.034$ | $0.981 \pm 0.031$ | $0.985 \pm 0.028$ | $0.986 \pm 0.029$ | $0.988 \pm 0.029$ | $0.989 \pm 0.030$ |
| EI | $1313.2 \pm 778.4$ | $259.2 \pm 97.0$ | $63.7 \pm 10.6$ | $33.3 \pm 4.3$ | $17.2 \pm 2.0$ | $8.7 \pm 1.0$ | $5.8 \pm 0.7$ | $4.4 \pm 0.5$ | $3.5 \pm 0.4$ | $2.3 \pm 0.3$ |
| RASI | $0.673 \pm 0.186$ | $0.816 \pm 0.114$ | $0.873 \pm 0.078$ | $0.897 \pm 0.068$ | $0.917 \pm 0.055$ | $0.932 \pm 0.047$ | $0.937 \pm 0.047$ | $0.941 \pm 0.046$ | $0.946 \pm 0.043$ | $0.970 \pm 0.042$ |
| SRI | $0.139 \pm 0.199$ | $0.140 \pm 0.195$ | $0.126 \pm 0.171$ | $0.148 \pm 0.207$ | $0.140 \pm 0.191$ | $0.145 \pm 0.194$ | $0.139 \pm 0.189$ | $0.148 \pm 0.218$ | $0.118 \pm 0.160$ | $0.135 \pm 0.219$ |
| SDI | $0.018 \pm 0.024$ | $0.017 \pm 0.024$ | $0.017 \pm 0.024$ | $0.016 \pm 0.023$ | $0.018 \pm 0.024$ | $0.018 \pm 0.024$ | $0.017 \pm 0.023$ | $0.017 \pm 0.023$ | $0.017 \pm 0.024$ | $0.017 \pm 0.023$ |
| $\Delta_{\text{DP}}$ | $0.0824 \pm 0.0635$ | $0.0397 \pm 0.0326$ | $0.0354 \pm 0.0293$ | $0.0182 \pm 0.0162$ | $0.0284 \pm 0.0247$ | $0.0176 \pm 0.0156$ | $0.0159 \pm 0.0144$ | $0.0218 \pm 0.0192$ | $0.0242 \pm 0.0209$ | $0.0128 \pm 0.0117$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0473 \pm 0.0403$ | $0.0349 \pm 0.0296$ | $0.0322 \pm 0.0274$ | $0.0192 \pm 0.0172$ | $0.0299 \pm 0.0254$ | $0.0251 \pm 0.0215$ | $0.0207 \pm 0.0179$ | $0.0245 \pm 0.0211$ | $0.0227 \pm 0.0196$ | $0.0209 \pm 0.0180$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1382 \pm 0.0983$ | $0.0544 \pm 0.0457$ | $0.0434 \pm 0.0365$ | $0.0202 \pm 0.0184$ | $0.0305 \pm 0.0269$ | $0.0201 \pm 0.0181$ | $0.0161 \pm 0.0149$ | $0.0238 \pm 0.0211$ | $0.0352 \pm 0.0298$ | $0.0170 \pm 0.0155$ |

Table 29 shows the evaluation results of MLP when using BMCR with the range of $[-10, 10]$.

Table 29: Evaluation results of MLP when using BMCR with the range of $[-10, 10]$

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0355 \pm 0.0283$ | $0.0244 \pm 0.0197$ | $0.0178 \pm 0.0146$ | $0.0145 \pm 0.0120$ | $0.0121 \pm 0.0102$ | $0.0102 \pm 0.0087$ | $0.0091 \pm 0.0078$ | $0.0084 \pm 0.0073$ | $0.0079 \pm 0.0069$ | $0.0071 \pm 0.0063$ |
| TRI | $0.744 \pm 0.156$ | $0.877 \pm 0.077$ | $0.930 \pm 0.046$ | $0.954 \pm 0.039$ | $0.972 \pm 0.025$ | $0.983 \pm 0.020$ | $0.989 \pm 0.016$ | $0.990 \pm 0.017$ | $0.993 \pm 0.015$ | $0.995 \pm 0.018$ |
| EI | $1260.1 \pm 738.0$ | $258.8 \pm 96.2$ | $63.3 \pm 10.3$ | $33.0 \pm 4.2$ | $17.0 \pm 2.0$ | $8.6 \pm 1.0$ | $5.8 \pm 0.6$ | $4.4 \pm 0.5$ | $3.5 \pm 0.4$ | $2.3 \pm 0.3$ |
| RASI | $0.654 \pm 0.193$ | $0.800 \pm 0.125$ | $0.859 \pm 0.084$ | $0.887 \pm 0.074$ | $0.915 \pm 0.053$ | $0.938 \pm 0.042$ | $0.942 \pm 0.040$ | $0.950 \pm 0.038$ | $0.951 \pm 0.038$ | $0.975 \pm 0.037$ |
| SRI | $0.155 \pm 0.251$ | $0.161 \pm 0.244$ | $0.148 \pm 0.220$ | $0.158 \pm 0.247$ | $0.145 \pm 0.209$ | $0.146 \pm 0.220$ | $0.158 \pm 0.235$ | $0.182 \pm 0.302$ | $0.123 \pm 0.196$ | $0.175 \pm 0.343$ |
| SDI | $0.015 \pm 0.023$ | $0.013 \pm 0.022$ | $0.014 \pm 0.022$ | $0.013 \pm 0.021$ | $0.015 \pm 0.021$ | $0.015 \pm 0.022$ | $0.014 \pm 0.022$ | $0.015 \pm 0.022$ | $0.014 \pm 0.022$ | $0.013 \pm 0.021$ |
| $\Delta_{\text{DP}}$ | $0.0832 \pm 0.0641$ | $0.0405 \pm 0.0332$ | $0.0365 \pm 0.0301$ | $0.0189 \pm 0.0168$ | $0.0289 \pm 0.0252$ | $0.0179 \pm 0.0159$ | $0.0161 \pm 0.0146$ | $0.0221 \pm 0.0195$ | $0.0246 \pm 0.0212$ | $0.0131 \pm 0.0120$ |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | $0.0482 \pm 0.0411$ | $0.0356 \pm 0.0302$ | $0.0328 \pm 0.0279$ | $0.0196 \pm 0.0176$ | $0.0304 \pm 0.0258$ | $0.0254 \pm 0.0218$ | $0.0209 \pm 0.0181$ | $0.0248 \pm 0.0214$ | $0.0229 \pm 0.0198$ | $0.0212 \pm 0.0182$ |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | $0.1395 \pm 0.0992$ | $0.0556 \pm 0.0468$ | $0.0441 \pm 0.0372$ | $0.0208 \pm 0.0188$ | $0.0312 \pm 0.0275$ | $0.0205 \pm 0.0184$ | $0.0164 \pm 0.0152$ | $0.0242 \pm 0.0215$ | $0.0359 \pm 0.0304$ | $0.0174 \pm 0.0159$ |

Table 30 shows the evaluation results of ResNet when using BMCR with the range of $[-0.5, 0.5]$.

Table 30: Evaluation results of ResNet when using BMCR with the range of $[-0.5, 0.5]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0261 \pm 0.0211$ | $0.0179 \pm 0.0144$ | $0.0134 \pm 0.0109$ | $0.0110 \pm 0.0092$ | $0.0095 \pm 0.0080$ | $0.0084 \pm 0.0072$ | $0.0079 \pm 0.0068$ | $0.0075 \pm 0.0066$ | $0.0072 \pm 0.0063$ | $0.0066 \pm 0.0058$ |
| TRI | $0.956 \pm 0.104$ | $0.995 \pm 0.086$ | $0.977 \pm 0.041$ | $0.983 \pm 0.036$ | $0.986 \pm 0.033$ | $0.989 \pm 0.032$ | $0.986 \pm 0.029$ | $0.983 \pm 0.031$ | $0.989 \pm 0.024$ | $0.988 \pm 0.029$ |
| EI | $2146.2 \pm 1187.1$ | $337.0 \pm 101.9$ | $77.0 \pm 11.7$ | $39.5 \pm 5.3$ | $20.0 \pm 2.5$ | $10.1 \pm 1.2$ | $6.8 \pm 0.8$ | $5.1 \pm 0.6$ | $4.1 \pm 0.5$ | $2.7 \pm 0.3$ |
| RASI | $0.740 \pm 0.168$ | $0.806 \pm 0.094$ | $0.852 \pm 0.075$ | $0.869 \pm 0.068$ | $0.878 \pm 0.072$ | $0.888 \pm 0.073$ | $0.891 \pm 0.071$ | $0.891 \pm 0.072$ | $0.896 \pm 0.071$ | $0.929 \pm 0.082$ |
| SRI | $0.148 \pm 0.190$ | $0.171 \pm 0.220$ | $0.166 \pm 0.210$ | $0.159 \pm 0.198$ | $0.151 \pm 0.185$ | $0.161 \pm 0.176$ | $0.164 \pm 0.210$ | $0.142 \pm 0.188$ | $0.143 \pm 0.170$ | $0.158 \pm 0.205$ |
| SDI | $0.032 \pm 0.037$ | $0.030 \pm 0.035$ | $0.029 \pm 0.034$ | $0.031 \pm 0.036$ | $0.030 \pm 0.037$ | $0.029 \pm 0.035$ | $0.031 \pm 0.036$ | $0.032 \pm 0.041$ | $0.030 \pm 0.036$ | $0.031 \pm 0.037$ |
| $\Delta_{\mathrm{DP}}$ | $0.0810 \pm 0.0624$ | $0.0370 \pm 0.0305$ | $0.0370 \pm 0.0304$ | $0.0177 \pm 0.0159$ | $0.0295 \pm 0.0254$ | $0.0182 \pm 0.0161$ | $0.0166 \pm 0.0150$ | $0.0231 \pm 0.0202$ | $0.0264 \pm 0.0225$ | $0.0134 \pm 0.0123$ |
| $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | $0.0442 \pm 0.0378$ | $0.0327 \pm 0.0277$ | $0.0340 \pm 0.0280$ | $0.0194 \pm 0.0173$ | $0.0311 \pm 0.0259$ | $0.0263 \pm 0.0223$ | $0.0214 \pm 0.0185$ | $0.0260 \pm 0.0221$ | $0.0234 \pm 0.0199$ | $0.0221 \pm 0.0187$ |
| $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ | $0.1355 \pm 0.0941$ | $0.0507 \pm 0.0425$ | $0.0428 \pm 0.0362$ | $0.0202 \pm 0.0182$ | $0.0304 \pm 0.0269$ | $0.0204 \pm 0.0184$ | $0.0164 \pm 0.0151$ | $0.0247 \pm 0.0217$ | $0.0358 \pm 0.0302$ | $0.0172 \pm 0.0158$ |

Table 31 shows the evaluation results of ResNet when using BMCR with the range of $[-1, 1]$.

Table 31: Evaluation results of ResNet when using BMCR with the range of $[-1, 1]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0254 \pm 0.0206$ | $0.0172 \pm 0.0139$ | $0.0128 \pm 0.0105$ | $0.0106 \pm 0.0089$ | $0.0092 \pm 0.0078$ | $0.0081 \pm 0.0070$ | $0.0076 \pm 0.0066$ | $0.0073 \pm 0.0064$ | $0.0069 \pm 0.0061$ | $0.0063 \pm 0.0056$ |
| TRI | $0.980 \pm 0.091$ | $1.016 \pm 0.079$ | $0.984 \pm 0.036$ | $0.982 \pm 0.040$ | $0.987 \pm 0.033$ | $0.988 \pm 0.030$ | $0.985 \pm 0.029$ | $0.986 \pm 0.032$ | $0.993 \pm 0.020$ | $0.989 \pm 0.027$ |
| EI | $2078.7 \pm 1170.9$ | $331.0 \pm 98.8$ | $75.2 \pm 11.2$ | $38.7 \pm 5.3$ | $19.6 \pm 2.4$ | $9.9 \pm 1.2$ | $6.6 \pm 0.8$ | $5.0 \pm 0.6$ | $4.0 \pm 0.4$ | $2.6 \pm 0.3$ |
| RASI | $0.757 \pm 0.169$ | $0.830 \pm 0.078$ | $0.875 \pm 0.073$ | $0.880 \pm 0.071$ | $0.889 \pm 0.073$ | $0.897 \pm 0.071$ | $0.895 \pm 0.073$ | $0.896 \pm 0.074$ | $0.900 \pm 0.076$ | $0.935 \pm 0.085$ |
| SRI | $0.178 \pm 0.290$ | $0.215 \pm 0.311$ | $0.204 \pm 0.325$ | $0.198 \pm 0.302$ | $0.197 \pm 0.290$ | $0.199 \pm 0.271$ | $0.194 \pm 0.319$ | $0.158 \pm 0.277$ | $0.204 \pm 0.287$ | $0.200 \pm 0.315$ |
| SDI | $0.030 \pm 0.039$ | $0.028 \pm 0.038$ | $0.026 \pm 0.035$ | $0.028 \pm 0.037$ | $0.028 \pm 0.039$ | $0.026 \pm 0.036$ | $0.030 \pm 0.037$ | $0.031 \pm 0.043$ | $0.027 \pm 0.037$ | $0.029 \pm 0.038$ |
| $\Delta_{\mathrm{DP}}$ | $0.0786 \pm 0.0605$ | $0.0358 \pm 0.0296$ | $0.0335 \pm 0.0278$ | $0.0172 \pm 0.0154$ | $0.0271 \pm 0.0236$ | $0.0168 \pm 0.0151$ | $0.0154 \pm 0.0140$ | $0.0209 \pm 0.0185$ | $0.0235 \pm 0.0202$ | $0.0121 \pm 0.0112$ |
| $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | $0.0438 \pm 0.0374$ | $0.0318 \pm 0.0272$ | $0.0312 \pm 0.0265$ | $0.0184 \pm 0.0165$ | $0.0291 \pm 0.0248$ | $0.0242 \pm 0.0209$ | $0.0201 \pm 0.0174$ | $0.0236 \pm 0.0204$ | $0.0219 \pm 0.0188$ | $0.0204 \pm 0.0175$ |
| $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ | $0.1278 \pm 0.0912$ | $0.0492 \pm 0.0415$ | $0.0415 \pm 0.0352$ | $0.0191 \pm 0.0174$ | $0.0292 \pm 0.0258$ | $0.0192 \pm 0.0172$ | $0.0155 \pm 0.0144$ | $0.0228 \pm 0.0201$ | $0.0331 \pm 0.0281$ | $0.0159 \pm 0.0146$ |

Table 32 shows the evaluation results of ResNet when using BMCR with the range of $[-5, 5]$.

Table 32: Evaluation results of ResNet when using BMCR with the range of $[-5, 5]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0258 \pm 0.0209$ | $0.0176 \pm 0.0142$ | $0.0131 \pm 0.0107$ | $0.0109 \pm 0.0091$ | $0.0094 \pm 0.0080$ | $0.0083 \pm 0.0072$ | $0.0078 \pm 0.0068$ | $0.0074 \pm 0.0065$ | $0.0070 \pm 0.0062$ | $0.0064 \pm 0.0057$ |
| TRI | $0.973 \pm 0.098$ | $1.025 \pm 0.095$ | $1.004 \pm 0.048$ | $0.996 \pm 0.045$ | $0.995 \pm 0.041$ | $0.998 \pm 0.037$ | $0.989 \pm 0.038$ | $0.987 \pm 0.035$ | $0.993 \pm 0.024$ | $0.992 \pm 0.030$ |
| EI | $2053.4 \pm 1137.4$ | $328.7 \pm 94.5$ | $74.7 \pm 9.8$ | $38.2 \pm 4.4$ | $19.4 \pm 2.1$ | $9.8 \pm 1.0$ | $6.6 \pm 0.7$ | $4.9 \pm 0.5$ | $3.9 \pm 0.4$ | $2.6 \pm 0.3$ |
| RASI | $0.741 \pm 0.158$ | $0.816 \pm 0.086$ | $0.868 \pm 0.067$ | $0.876 \pm 0.070$ | $0.887 \pm 0.069$ | $0.892 \pm 0.070$ | $0.894 \pm 0.073$ | $0.898 \pm 0.071$ | $0.903 \pm 0.073$ | $0.937 \pm 0.087$ |
| SRI | $0.176 \pm 0.352$ | $0.212 \pm 0.408$ | $0.210 \pm 0.423$ | $0.154 \pm 0.275$ | $0.188 \pm 0.361$ | $0.211 \pm 0.415$ | $0.159 \pm 0.289$ | $0.145 \pm 0.285$ | $0.106 \pm 0.174$ | $0.163 \pm 0.292$ |
| SDI | $0.029 \pm 0.039$ | $0.027 \pm 0.037$ | $0.028 \pm 0.038$ | $0.028 \pm 0.039$ | $0.028 \pm 0.039$ | $0.026 \pm 0.037$ | $0.028 \pm 0.037$ | $0.031 \pm 0.043$ | $0.028 \pm 0.038$ | $0.027 \pm 0.038$ |
| $\Delta_{\mathrm{DP}}$ | $0.0796 \pm 0.0613$ | $0.0365 \pm 0.0302$ | $0.0341 \pm 0.0283$ | $0.0177 \pm 0.0157$ | $0.0276 \pm 0.0240$ | $0.0171 \pm 0.0153$ | $0.0157 \pm 0.0143$ | $0.0214 \pm 0.0189$ | $0.0239 \pm 0.0206$ | $0.0124 \pm 0.0115$ |
| $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | $0.0445 \pm 0.0379$ | $0.0325 \pm 0.0277$ | $0.0318 \pm 0.0270$ | $0.0188 \pm 0.0169$ | $0.0296 \pm 0.0252$ | $0.0247 \pm 0.0212$ | $0.0204 \pm 0.0177$ | $0.0240 \pm 0.0208$ | $0.0223 \pm 0.0192$ | $0.0207 \pm 0.0178$ |
| $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ | $0.1295 \pm 0.0925$ | $0.0503 \pm 0.0424$ | $0.0424 \pm 0.0358$ | $0.0197 \pm 0.0179$ | $0.0299 \pm 0.0264$ | $0.0197 \pm 0.0177$ | $0.0159 \pm 0.0147$ | $0.0234 \pm 0.0207$ | $0.0340 \pm 0.0288$ | $0.0164 \pm 0.0150$ |

Table 33 shows the evaluation results of ResNet when using BMCR with the range of $[-10, 10]$.

Table 33: Evaluation results of ResNet when using BMCR with the range of $[-10, 10]$.

| Metric \ Epoch | 1 | 10 | 50 | 100 | 200 | 400 | 600 | 800 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|---|---|---|
| RIPV | $0.0261 \pm 0.0211$ | $0.0179 \pm 0.0145$ | $0.0134 \pm 0.0109$ | $0.0111 \pm 0.0092$ | $0.0096 \pm 0.0081$ | $0.0085 \pm 0.0073$ | $0.0079 \pm 0.0069$ | $0.0075 \pm 0.0066$ | $0.0071 \pm 0.0063$ | $0.0065 \pm 0.0058$ |
| TRI | $0.966 \pm 0.106$ | $1.034 \pm 0.112$ | $1.024 \pm 0.060$ | $1.009 \pm 0.050$ | $1.003 \pm 0.050$ | $1.007 \pm 0.043$ | $0.992 \pm 0.047$ | $0.987 \pm 0.039$ | $0.993 \pm 0.028$ | $0.995 \pm 0.033$ |
| EI | $2028.1 \pm 1103.9$ | $326.3 \pm 90.2$ | $74.2 \pm 8.4$ | $37.8 \pm 3.6$ | $19.3 \pm 1.8$ | $9.8 \pm 0.9$ | $6.6 \pm 0.6$ | $4.9 \pm 0.4$ | $3.9 \pm 0.3$ | $2.5 \pm 0.2$ |
| RASI | $0.724 \pm 0.146$ | $0.802 \pm 0.094$ | $0.861 \pm 0.061$ | $0.872 \pm 0.069$ | $0.884 \pm 0.068$ | $0.888 \pm 0.069$ | $0.893 \pm 0.072$ | $0.899 \pm 0.069$ | $0.905 \pm 0.070$ | $0.938 \pm 0.088$ |
| SRI | $0.173 \pm 0.412$ | $0.209 \pm 0.504$ | $0.217 \pm 0.520$ | $0.110 \pm 0.247$ | $0.178 \pm 0.431$ | $0.223 \pm 0.544$ | $0.124 \pm 0.259$ | $0.128 \pm 0.292$ | $0.005 \pm 0.016$ | $0.121 \pm 0.268$ |
| SDI | $0.027 \pm 0.038$ | $0.026 \pm 0.036$ | $0.029 \pm 0.041$ | $0.028 \pm 0.040$ | $0.028 \pm 0.039$ | $0.026 \pm 0.037$ | $0.027 \pm 0.036$ | $0.030 \pm 0.043$ | $0.028 \pm 0.039$ | $0.026 \pm 0.037$ |
| $\Delta_{\mathrm{DP}}$ | $0.0804 \pm 0.0619$ | $0.0372 \pm 0.0308$ | $0.0348 \pm 0.0288$ | $0.0181 \pm 0.0161$ | $0.0282 \pm 0.0245$ | $0.0174 \pm 0.0156$ | $0.0159 \pm 0.0145$ | $0.0218 \pm 0.0192$ | $0.0242 \pm 0.0209$ | $0.0128 \pm 0.0118$ |
| $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | $0.0451 \pm 0.0385$ | $0.0332 \pm 0.0283$ | $0.0325 \pm 0.0276$ | $0.0192 \pm 0.0172$ | $0.0301 \pm 0.0255$ | $0.0251 \pm 0.0215$ | $0.0207 \pm 0.0179$ | $0.0244 \pm 0.0211$ | $0.0226 \pm 0.0195$ | $0.0210 \pm 0.0181$ |
| $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ | $0.1312 \pm 0.0938$ | $0.0516 \pm 0.0434$ | $0.0432 \pm 0.0365$ | $0.0203 \pm 0.0184$ | $0.0306 \pm 0.0271$ | $0.0201 \pm 0.0181$ | $0.0162 \pm 0.0150$ | $0.0239 \pm 0.0212$ | $0.0348 \pm 0.0295$ | $0.0169 \pm 0.0154$ |

### Q.1 Robustness to Replacement Width $\varepsilon$

We evaluate BMCR under multiple replacement widths, sampling the substituted feature as $T \sim \mathrm{Unif}[\mu_q - \varepsilon, \mu_q + \varepsilon]$ (in the normalized feature space), with $\varepsilon \in \{0.1, 0.5, 1, 5, 10\}$. Across datasets and neural backbones, the overall behavior is *largely stable* as $\varepsilon$ varies. This robustness *supports a key implication of our analysis*: the replaced-risk objective primarily penalises the model's *sensitivity* to the removed coordinate (equivalently, the prediction variance under $T \sim q$). Once this sensitivity is driven close to zero, widening the support of $q$ (larger $\varepsilon$) provides limited additional training signal, yielding solutions whose utility–invariance trade-off is not brittle to the precise choice of $\varepsilon$. It also means that $\varepsilon$ can be fixed in advance rather than searched per request (Appendix S); the choice of $\varepsilon$ determines the certified interval, not the quality of unlearning.

## R FULL RESULTS OF MULTI-FEATURE UNLEARNING

Figure 23 shows the evaluation results of unlearning the most and the second most influential feature in the *Blood* dataset when using MLP as neural backbone.

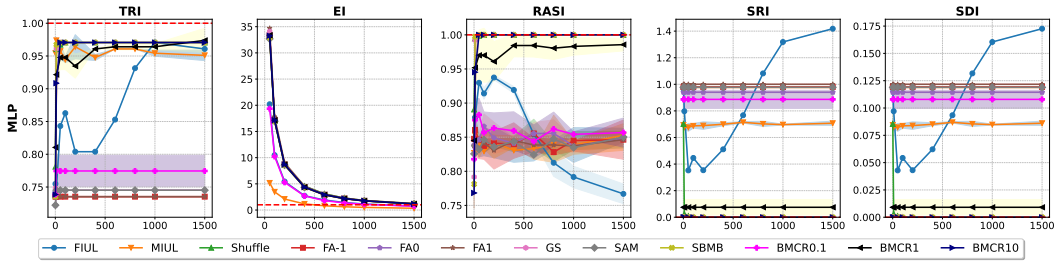


Figure 23: Evaluation results of unlearning the most and the second most influential feature in the *Blood* dataset.

Figure 24 shows the evaluation results of unlearning the most and the second most influential feature in the *Cali* dataset when using MLP as neural backbone.

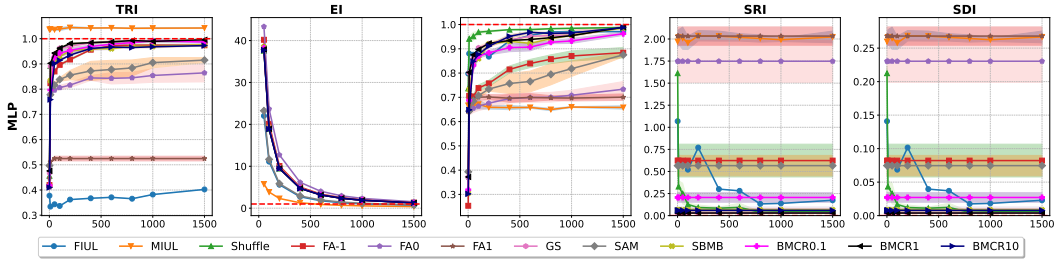


Figure 24: Evaluation results of unlearning the most and the second most influential feature in the *Cali* dataset.

Figure 25 shows the evaluation results of unlearning the most and the second most influential feature in the *Credit* dataset when using MLP as neural backbone.

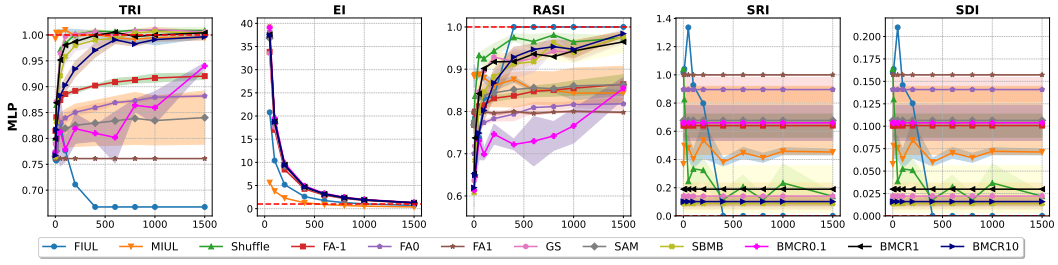


Figure 25: Evaluation results of unlearning the most and the second most influential feature in the *Credit* dataset.

Figure 26 shows the evaluation results of unlearning the most and the second most influential feature in the *Diabetes* dataset when using MLP as neural backbone.

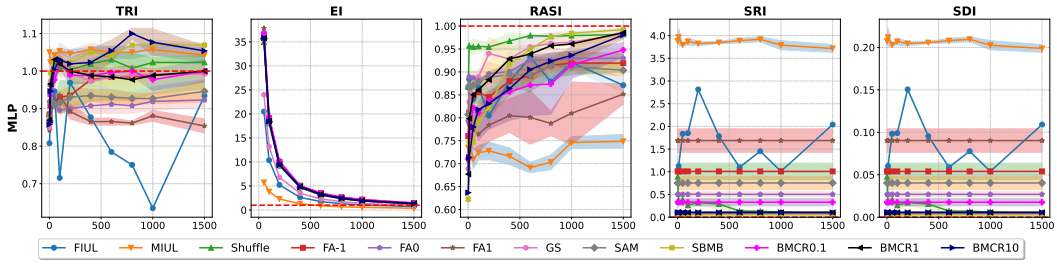


Figure 26: Evaluation results of unlearning the most and the second most influential feature in the *Diabetes* dataset.

Figure 27 shows the evaluation results of unlearning the most and the second most influential feature in the *House* dataset when using MLP as neural backbone.

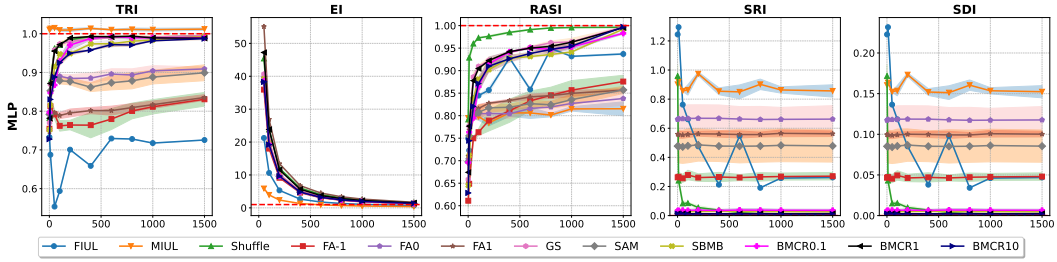


Figure 27: Evaluation results of unlearning the most and the second most influential feature in the *House* dataset.

Figure 28 shows the evaluation results of unlearning the most and the second most influential feature in the *Magic* dataset when using MLP as neural backbone.

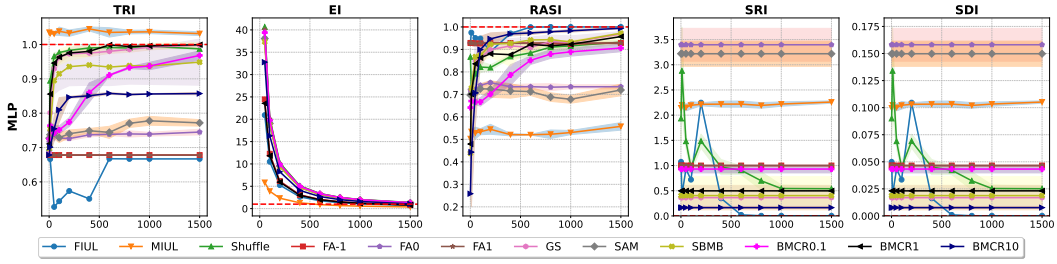


Figure 28: Evaluation results of unlearning the most and the second most influential feature in the *Magic* dataset.

Figure 29 shows the evaluation results of unlearning the most and the second most influential feature in the *Pol* dataset when using MLP as neural backbone.

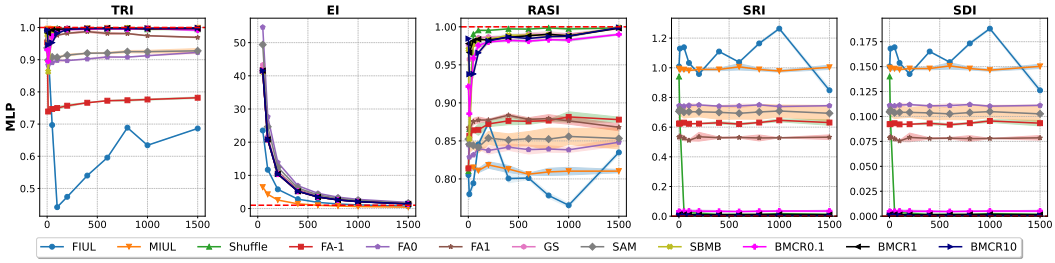


Figure 29: Evaluation results of unlearning the most and the second most influential feature in the *Pol* dataset.

## S Computational Cost

**Wall-clock.** End-to-end runtime on one NVIDIA V100 includes initialization or checkpoint loading, data preparation, training or fine-tuning, replacement sampling, and checkpoint saving.

Table 34: End-to-end cost on *Magic* with the MLP backbone.

| Method | Sel. epoch | Wall-clock | GPU-hours | Peak mem. | Sampling time | Overhead | EI |
|---|---|---|---|---|---|---|---|
| Retraining | 1600 | 329.7 s | 0.0916 | 0.83 GB | – | – | 1.0 |
| BMCR | 100 | 8.3 s | 0.0023 | 0.91 GB | 0.20 s | 2.41% | 39.7 |

**Budget.** BMCR and retraining have the same data requirement, $(x_{-i}, y)$; BMCR's advantage is computational, because it fine-tunes an existing model instead of optimizing a new one. A 100-epoch budget already gives RIPV $0.0114 \pm 0.0096$ (vs. $0.0268 \pm 0.0215$ at epoch 1), TRI $0.984 \pm 0.031$, and EI $40.3 \pm 5.3$ (Table 2). Longer runs are reported only to characterize convergence.

**Search cost.** BMCR does not require a per-request search. The $\varepsilon$ values in the paper are sensitivity ablations; all of them achieve effective unlearning (Fig. 2 and Appendix Q), so a range can be fixed in advance. Reported epochs are checkpoints of one run, not separate runs.

## T Interpreting the Main Results

**Pooled standard deviations.** Table 2 pools all datasets and removed features, so its standard deviations reflect heterogeneity across tasks. Within a dataset and with paired comparisons, variation is much smaller:

Table 35: Per-dataset stability on *Magic*: the same removed-feature configurations at epochs 1 and 1500.

| Metric | Epoch 1 | Epoch 1500 | Paired change |
|---|---|---|---|
| RIPV | $0.0251 \pm 0.0112$ | $0.0075 \pm 0.0034$ | $-0.0176 \pm 0.0060$ |
| TRI | $0.934 \pm 0.049$ | $0.990 \pm 0.015$ | $+0.056 \pm 0.029$ |
| EI | $2205.6 \pm 471.2$ | $2.69 \pm 0.29$ | $-2202.9 \pm 471.0$ |
| RASI | $0.754 \pm 0.086$ | $0.946 \pm 0.028$ | $+0.192 \pm 0.063$ |
| SRI | $0.139 \pm 0.048$ | $0.121 \pm 0.040$ | $-0.018 \pm 0.024$ |
| SDI | $0.029 \pm 0.010$ | $0.018 \pm 0.007$ | $-0.011 \pm 0.006$ |

**SHAP metrics on an influential feature.** Aggregate SRI/SDI changes are small because least-influential features start near zero. For *Latitude*, the most influential *Cali* feature:

Table 36: SHAP-based metrics when unlearning *Latitude* in *Cali*.

| Epoch | 1 | 50 | 100 | 200 | 600 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|
| SRI (%) | $99.14 \pm 3.81$ | $55.61 \pm 4.71$ | $4.57 \pm 2.91$ | $3.17 \pm 2.12$ | $5.31 \pm 2.02$ | $3.00 \pm 0.03$ | $3.12 \pm 1.08$ |
| SDI (%) | $15.66 \pm 0.92$ | $8.51 \pm 6.38$ | $1.24 \pm 0.15$ | $1.18 \pm 0.91$ | $1.17 \pm 0.15$ | $0.95 \pm 0.21$ | $0.43 \pm 0.56$ |

**MIA.** The original model has MIA accuracy $0.682 \pm 0.024$; after unlearning it is $0.515$–$0.555$, close to chance (Table 2).

**Feature importance and difficulty.** Influential features are not harder to unlearn in terms of utility: We do not claim that SHAP importance predicts unlearning cost or utility loss. Under squared loss, the utility loss of removing feature $j$ given an already removed set $S$ is $\Delta_j(S) = R^\star_{S\cup\{j\}} - R^\star_S = \mathbb{E}\big[(\mathbb{E}[Y \mid X_{-S}] - \mathbb{E}[Y \mid X_{-(S\cup\{j\})}])^2\big]$, which depends on the conditional structure rather than on SHAP attributions of the original model. SRI and SDI are used only to measure the removed feature's post-unlearning attribution, jointly with the other metrics.

Table 37: TRI for the most vs. least influential removed feature.

| Feature | 1 | 10 | 100 | 200 | 600 | 1000 | 1500 |
|---|---|---|---|---|---|---|---|
| Most infl. | $0.865 \pm 0.166$ | $0.941 \pm 0.129$ | $0.962 \pm 0.105$ | $0.966 \pm 0.104$ | $0.970 \pm 0.101$ | $0.972 \pm 0.101$ | $0.974 \pm 0.101$ |
| Least infl. | $0.855 \pm 0.158$ | $0.927 \pm 0.123$ | $0.951 \pm 0.102$ | $0.955 \pm 0.101$ | $0.959 \pm 0.098$ | $0.961 \pm 0.097$ | $0.959 \pm 0.098$ |

## U Sequential and Three-Feature Unlearning

**Protocol and guarantee.** After $k$ requests, all revoked features $S_k$ are replaced jointly with $T_{S_k} \sim q_{S_k} = \bigotimes_{j \in S_k} q_j$, independent of $(X_{-S_k}, Y)$. Under squared loss, Lemma 4.1 applied to the vector $T_{S_k}$ gives

$$R_{S_k}(f) - R_{S_k}(f^\star) = \mathbb{E}\big[(\bar{f}(X_{-S_k}) - \mathbb{E}[Y \mid X_{-S_k}])^2\big] + D_{S_k}(f),$$

where $D_{S_k}(f) = \mathbb{E}\,\mathrm{Var}_{T_{S_k}}[f(X_{-S_k}, T_{S_k})]$. Hence a $\delta$-near minimizer has $D_{S_k}(f) \le \delta$, and by the law of total variance its dependence on each $j \in S_k$ is also at most $\delta$. Each update therefore refreshes the guarantee for all earlier deletions, and the population optimum $\mathbb{E}[Y \mid X_{-S_k}]$ does not depend on the order of requests.

Table 38: Joint unlearning of the three most influential features, averaged over datasets, BMCR ranges, and backbones.

| Epoch | RIPV | TRI | EI | RASI | SRI | SDI |
|---|---|---|---|---|---|---|
| 1 | $0.0283 \pm 0.0229$ | $0.855 \pm 0.166$ | $2308.4 \pm 1004.0$ | $0.789 \pm 0.135$ | $0.128 \pm 0.223$ | $0.035 \pm 0.044$ |
| 10 | $0.0202 \pm 0.0163$ | $0.911 \pm 0.158$ | $308.0 \pm 38.6$ | $0.848 \pm 0.084$ | $0.129 \pm 0.254$ | $0.037 \pm 0.043$ |
| 100 | $0.0135 \pm 0.0112$ | $0.927 \pm 0.158$ | $33.9 \pm 2.9$ | $0.874 \pm 0.071$ | $0.120 \pm 0.228$ | $0.032 \pm 0.041$ |
| 200 | $0.0119 \pm 0.0101$ | $0.930 \pm 0.159$ | $17.2 \pm 1.6$ | $0.880 \pm 0.070$ | $0.114 \pm 0.220$ | $0.034 \pm 0.043$ |
| 600 | $0.0100 \pm 0.0086$ | $0.927 \pm 0.158$ | $5.9 \pm 0.5$ | $0.888 \pm 0.070$ | $0.141 \pm 0.266$ | $0.035 \pm 0.043$ |
| 1000 | $0.0092 \pm 0.0080$ | $0.928 \pm 0.158$ | $3.5 \pm 0.3$ | $0.892 \pm 0.067$ | $0.114 \pm 0.243$ | $0.034 \pm 0.044$ |
| 1500 | $0.0083 \pm 0.0072$ | $0.923 \pm 0.157$ | $2.3 \pm 0.2$ | $0.940 \pm 0.070$ | $0.132 \pm 0.254$ | $0.036 \pm 0.045$ |

Table 39: Sequential deletion of the three most influential features in all six orders (epoch 1500).

| Order | RIPV | TRI | EI | RASI | SRI | SDI |
|---|---|---|---|---|---|---|
| Simultaneous | $0.0083 \pm 0.0072$ | $0.923 \pm 0.157$ | $2.30 \pm 0.20$ | $0.941 \pm 0.071$ | $0.132 \pm 0.259$ | $0.036 \pm 0.045$ |
| 1st→2nd→3rd | $0.0067 \pm 0.0059$ | $0.986 \pm 0.030$ | $2.68 \pm 0.31$ | $0.930 \pm 0.080$ | $0.160 \pm 0.207$ | $0.030 \pm 0.038$ |
| 1st→3rd→2nd | $0.0069 \pm 0.0062$ | $0.994 \pm 0.019$ | $2.31 \pm 0.29$ | $0.974 \pm 0.038$ | $0.173 \pm 0.339$ | $0.014 \pm 0.022$ |
| 2nd→1st→3rd | $0.0071 \pm 0.0065$ | $0.990 \pm 0.029$ | $2.32 \pm 0.31$ | $0.969 \pm 0.041$ | $0.137 \pm 0.221$ | $0.018 \pm 0.024$ |
| 2nd→3rd→1st | $0.0076 \pm 0.0065$ | $0.981 \pm 0.041$ | $2.29 \pm 0.29$ | $0.963 \pm 0.047$ | $0.103 \pm 0.148$ | $0.019 \pm 0.023$ |
| 3rd→2nd→1st | $0.0078 \pm 0.0068$ | $0.980 \pm 0.043$ | $2.33 \pm 0.31$ | $0.956 \pm 0.050$ | $0.143 \pm 0.156$ | $0.024 \pm 0.028$ |
| 3rd→1st→2nd | $0.0085 \pm 0.0073$ | $0.974 \pm 0.046$ | $2.41 \pm 0.21$ | $0.947 \pm 0.053$ | $0.178 \pm 0.166$ | $0.029 \pm 0.030$ |

No order is consistently best, so BMCR does not require importance-based scheduling.

## V Unlearning Correlated Features

Following the question of whether correlation between removed and retained features hinders forgetting, as observed for gradient ascent in data-point unlearning Mavrothalassitis et al. (2025), we identify all removed–retained pairs with Pearson coefficient above 0.8 (Table 40) and unlearn one feature of each pair while the other is retained, for all three backbones (30 configurations).

BMCR suppresses direct dependence on the removed feature and preserves utility even when a near-duplicate of it is retained. Unlike gradient ascent, BMCR does not push the model away from the forget data; it re-fits toward the retained-feature optimum, which the retrained oracle shares, so correlation does not create a conflicting objective.

## W Structural-Masking Baselines

**SFI (zero-and-freeze input weights).** The column of the first feature projection corresponding to $x_i$ (for FT-Transformer, the feature-tokenizer weights of $x_i$) is set to zero and kept at zero during

Table 40: Highly correlated feature pairs (Pearson $r > 0.8$). *Pol* and *Diabetes* have none.

| Dataset | Pairs |
|---|---|
| Cali | Latitude–Longitude; AveRooms–AveBedrms |
| Credit | NumberOfTimes90DaysLate–NumberOfTime60-89DaysPastDueNotWorse; NumberOfTime30-59DaysPastDueNotWorse–NumberOfTime60-89DaysPastDueNotWorse; NumberOfTime30-59DaysPastDueNotWorse–NumberOfTimes90DaysLate |
| Magic | fConc–fConc1; fSize–fConc; fSize–fConc1 |
| Blood | V2–V3 |
| House | P6p2–H8p2 |

Table 41: Correlated-feature unlearning, averaged over datasets, BMCR ranges, backbones, and pairs.

| Epoch | RIPV | TRI | EI | RASI | SRI | SDI | $\Delta_{\mathrm{DP}}$ | $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $.0281 \pm .0228$ | $.858 \pm .164$ | $2315.1 \pm 1020.2$ | $.794 \pm .130$ | $.127 \pm .233$ | $.035 \pm .045$ | $.0773 \pm .0595$ | $.0454 \pm .0386$ | $.1300 \pm .0929$ |
| 10 | $.0201 \pm .0163$ | $.913 \pm .158$ | $308.3 \pm 40.0$ | $.852 \pm .082$ | $.131 \pm .262$ | $.037 \pm .044$ | $.0386 \pm .0319$ | $.0338 \pm .0288$ | $.0524 \pm .0439$ |
| 50 | $.0159 \pm .0130$ | $.924 \pm .158$ | $66.8 \pm 4.8$ | $.870 \pm .073$ | $.128 \pm .227$ | $.034 \pm .043$ | $.0367 \pm .0303$ | $.0338 \pm .0283$ | $.0446 \pm .0376$ |
| 100 | $.0134 \pm .0112$ | $.927 \pm .158$ | $33.8 \pm 2.8$ | $.876 \pm .070$ | $.120 \pm .231$ | $.033 \pm .042$ | $.0198 \pm .0175$ | $.0205 \pm .0182$ | $.0217 \pm .0197$ |
| 200 | $.0119 \pm .0100$ | $.929 \pm .159$ | $17.2 \pm 1.5$ | $.881 \pm .070$ | $.113 \pm .226$ | $.035 \pm .043$ | $.0293 \pm .0255$ | $.0308 \pm .0261$ | $.0318 \pm .0279$ |
| 400 | $.0107 \pm .0091$ | $.927 \pm .159$ | $8.7 \pm 0.7$ | $.886 \pm .067$ | $.131 \pm .270$ | $.037 \pm .045$ | $.0193 \pm .0171$ | $.0266 \pm .0227$ | $.0217 \pm .0195$ |
| 600 | $.0100 \pm .0086$ | $.927 \pm .158$ | $5.9 \pm 0.5$ | $.890 \pm .069$ | $.139 \pm .269$ | $.035 \pm .043$ | $.0177 \pm .0160$ | $.0224 \pm .0193$ | $.0173 \pm .0159$ |
| 800 | $.0095 \pm .0082$ | $.927 \pm .158$ | $4.4 \pm 0.4$ | $.891 \pm .068$ | $.121 \pm .234$ | $.036 \pm .044$ | $.0237 \pm .0209$ | $.0272 \pm .0232$ | $.0263 \pm .0232$ |
| 1000 | $.0092 \pm .0079$ | $.927 \pm .158$ | $3.5 \pm 0.3$ | $.893 \pm .067$ | $.114 \pm .248$ | $.034 \pm .044$ | $.0267 \pm .0231$ | $.0246 \pm .0210$ | $.0384 \pm .0325$ |
| 1500 | $.0083 \pm .0072$ | $.923 \pm .157$ | $2.3 \pm 0.2$ | $.941 \pm .071$ | $.132 \pm .259$ | $.036 \pm .045$ | $.0146 \pm .0134$ | $.0192 \pm .0175$ | $.0231 \pm .0199$ |

fine-tuning on $(x_{-i}, y)$. **FPE (feature-pathway excision).** A reduced network whose input layer omits $x_i$ is built, all compatible parameters are copied from $f_{\theta_0}$, and it is fine-tuned on $(x_{-i}, y)$. Both use no information about $x_i$ (tier A0). Tables 42–48 report the most influential feature of each dataset, averaged over MLP, ResNet, and FTformer.

Table 42: Structural-masking baselines vs. BMCR: single-feature unlearning of the most influential feature in *Magic*, averaged over MLP, ResNet, and FTformer.

| Metric | SFI | FPE | $\mathrm{BMCR}_{0.1}$ | $\mathrm{BMCR}_{0.5}$ | $\mathrm{BMCR}_{1}$ | $\mathrm{BMCR}_{5}$ | $\mathrm{BMCR}_{10}$ |
|---|---|---|---|---|---|---|---|
| RIPV | 0.030±0.011 | 0.039±0.012 | 0.014±0.008 | 0.010±0.006 | 0.008±0.005 | 0.006±0.004 | 0.006±0.004 |
| TRI | 0.771±0.024 | 0.909±0.013 | 1.000±0.008 | 1.003±0.008 | 1.006±0.008 | 1.005±0.005 | 1.003±0.003 |
| EI | 2.243±0.170 | 2.172±0.157 | 2.242±0.161 | 2.213±0.159 | 2.181±0.155 | 2.195±0.156 | 2.209±0.159 |
| RASI | 0.878±0.014 | 0.792±0.020 | 0.940±0.021 | 0.964±0.012 | 0.980±0.004 | 0.985±0.004 | 0.989±0.003 |
| SRI | 0.778±0.089 | 1.618±0.172 | 0.439±0.209 | 0.231±0.106 | 0.133±0.030 | 0.122±0.034 | 0.115±0.037 |
| SDI | 0.057±0.008 | 0.093±0.011 | 0.025±0.011 | 0.013±0.006 | 0.007±0.002 | 0.007±0.002 | 0.006±0.002 |
| MIA | 0.562±0.029 | 0.545±0.026 | 0.517±0.024 | 0.514±0.022 | 0.512±0.021 | 0.512±0.021 | 0.512±0.022 |
| $\Delta_{\mathrm{DP}}$ | 0.067±0.035 | 0.071±0.039 | 0.015±0.013 | 0.014±0.013 | 0.013±0.012 | 0.013±0.012 | 0.014±0.012 |
| $\Delta_{\mathrm{EO}}^{\mathrm{TPR}}$ | 0.052±0.030 | 0.063±0.035 | 0.023±0.020 | 0.022±0.019 | 0.021±0.018 | 0.022±0.019 | 0.022±0.019 |
| $\Delta_{\mathrm{EO}}^{\mathrm{FPR}}$ | 0.078±0.044 | 0.081±0.047 | 0.020±0.018 | 0.018±0.017 | 0.017±0.016 | 0.017±0.016 | 0.018±0.016 |

Table 43: Structural-masking baselines vs. BMCR: single-feature unlearning of the most influential feature in *Cali*, averaged over MLP, ResNet, and FTformer.

| Metric | SFI | FPE | $\text{BMCR}_{0.1}$ | $\text{BMCR}_{0.5}$ | $\text{BMCR}_{1}$ | $\text{BMCR}_{5}$ | $\text{BMCR}_{10}$ |
|---|---|---|---|---|---|---|---|
| RIPV | 0.031±0.010 | 0.040±0.013 | 0.015±0.007 | 0.011±0.005 | 0.009±0.004 | 0.007±0.003 | 0.005±0.003 |
| TRI | 0.768±0.025 | 0.905±0.014 | 0.997±0.009 | 1.002±0.007 | 1.007±0.009 | 1.006±0.006 | 1.004±0.004 |
| EI | 2.251±0.166 | 2.165±0.160 | 2.236±0.158 | 2.208±0.162 | 2.184±0.153 | 2.198±0.154 | 2.205±0.157 |
| RASI | 0.875±0.015 | 0.796±0.019 | 0.937±0.020 | 0.962±0.013 | 0.978±0.005 | 0.986±0.003 | 0.990±0.002 |
| SRI | 0.786±0.092 | 1.603±0.168 | 0.451±0.201 | 0.239±0.111 | 0.137±0.032 | 0.124±0.031 | 0.113±0.035 |
| SDI | 0.059±0.009 | 0.091±0.012 | 0.027±0.010 | 0.014±0.005 | 0.008±0.003 | 0.007±0.003 | 0.005±0.001 |
| MIA | 0.559±0.031 | 0.548±0.025 | 0.520±0.023 | 0.515±0.021 | 0.513±0.020 | 0.511±0.020 | 0.510±0.021 |
| $\Delta_{\text{DP}}$ | 0.069±0.033 | 0.073±0.041 | 0.016±0.012 | 0.015±0.014 | 0.012±0.011 | 0.012±0.013 | 0.015±0.011 |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | 0.054±0.029 | 0.065±0.037 | 0.024±0.019 | 0.021±0.018 | 0.020±0.017 | 0.023±0.018 | 0.021±0.020 |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | 0.076±0.046 | 0.083±0.045 | 0.021±0.017 | 0.019±0.016 | 0.016±0.015 | 0.018±0.015 | 0.019±0.017 |

Table 44: Structural-masking baselines vs. BMCR: single-feature unlearning of the most influential feature in *Credit*, averaged over MLP, ResNet, and FTformer.

| Metric | SFI | FPE | $\text{BMCR}_{0.1}$ | $\text{BMCR}_{0.5}$ | $\text{BMCR}_{1}$ | $\text{BMCR}_{5}$ | $\text{BMCR}_{10}$ |
|---|---|---|---|---|---|---|---|
| RIPV | 0.032±0.011 | 0.041±0.012 | 0.016±0.008 | 0.012±0.006 | 0.010±0.005 | 0.008±0.004 | 0.006±0.002 |
| TRI | 0.771±0.027 | 0.902±0.016 | 0.995±0.010 | 1.001±0.008 | 1.008±0.008 | 1.005±0.007 | 1.003±0.005 |
| EI | 2.247±0.169 | 2.171±0.157 | 2.232±0.161 | 2.211±0.159 | 2.188±0.156 | 2.201±0.152 | 2.207±0.160 |
| RASI | 0.872±0.016 | 0.799±0.021 | 0.934±0.018 | 0.960±0.014 | 0.976±0.006 | 0.985±0.004 | 0.989±0.003 |
| SRI | 0.793±0.096 | 1.591±0.173 | 0.459±0.196 | 0.244±0.115 | 0.141±0.034 | 0.127±0.029 | 0.116±0.033 |
| SDI | 0.061±0.010 | 0.089±0.013 | 0.029±0.009 | 0.015±0.006 | 0.009±0.004 | 0.008±0.002 | 0.006±0.002 |
| MIA | 0.562±0.029 | 0.551±0.027 | 0.523±0.024 | 0.517±0.022 | 0.514±0.019 | 0.512±0.021 | 0.509±0.020 |
| $\Delta_{\text{DP}}$ | 0.071±0.035 | 0.075±0.039 | 0.017±0.013 | 0.016±0.012 | 0.013±0.010 | 0.013±0.012 | 0.016±0.012 |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | 0.056±0.031 | 0.067±0.035 | 0.025±0.020 | 0.022±0.017 | 0.019±0.016 | 0.024±0.019 | 0.022±0.018 |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | 0.078±0.044 | 0.085±0.047 | 0.022±0.018 | 0.020±0.015 | 0.017±0.014 | 0.019±0.016 | 0.020±0.018 |

Table 45: Structural-masking baselines vs. BMCR: single-feature unlearning of the most influential feature in *Blood*, averaged over MLP, ResNet, and FTformer.

| Metric | SFI | FPE | $\text{BMCR}_{0.1}$ | $\text{BMCR}_{0.5}$ | $\text{BMCR}_{1}$ | $\text{BMCR}_{5}$ | $\text{BMCR}_{10}$ |
|---|---|---|---|---|---|---|---|
| RIPV | 0.033±0.010 | 0.043±0.013 | 0.017±0.007 | 0.013±0.005 | 0.011±0.006 | 0.009±0.003 | 0.007±0.003 |
| TRI | 0.779±0.029 | 0.896±0.017 | 0.992±0.011 | 1.002±0.009 | 1.009±0.007 | 1.006±0.008 | 1.004±0.006 |
| EI | 2.239±0.165 | 2.178±0.160 | 2.226±0.158 | 2.216±0.162 | 2.191±0.153 | 2.204±0.155 | 2.209±0.157 |
| RASI | 0.868±0.017 | 0.804±0.020 | 0.931±0.019 | 0.957±0.013 | 0.974±0.007 | 0.983±0.005 | 0.988±0.004 |
| SRI | 0.806±0.101 | 1.573±0.168 | 0.472±0.189 | 0.251±0.109 | 0.146±0.037 | 0.131±0.031 | 0.119±0.030 |
| SDI | 0.063±0.011 | 0.087±0.014 | 0.031±0.008 | 0.016±0.007 | 0.010±0.005 | 0.009±0.003 | 0.007±0.003 |
| MIA | 0.558±0.031 | 0.548±0.026 | 0.526±0.025 | 0.519±0.023 | 0.515±0.020 | 0.511±0.022 | 0.508±0.019 |
| $\Delta_{\text{DP}}$ | 0.069±0.034 | 0.077±0.037 | 0.018±0.014 | 0.015±0.011 | 0.012±0.009 | 0.013±0.011 | 0.016±0.013 |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | 0.058±0.030 | 0.065±0.037 | 0.026±0.019 | 0.021±0.018 | 0.018±0.015 | 0.023±0.020 | 0.022±0.017 |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | 0.076±0.042 | 0.088±0.049 | 0.023±0.017 | 0.021±0.016 | 0.016±0.013 | 0.018±0.015 | 0.020±0.017 |

Table 46: Structural-masking baselines vs. BMCR: single-feature unlearning of the most influential feature in *Diabetes*, averaged over MLP, ResNet, and FTformer.

| Metric | SFI | FPE | $\text{BMCR}_{0.1}$ | $\text{BMCR}_{0.5}$ | $\text{BMCR}_{1}$ | $\text{BMCR}_{5}$ | $\text{BMCR}_{10}$ |
|---|---|---|---|---|---|---|---|
| RIPV | 0.034±0.011 | 0.041±0.012 | 0.018±0.008 | 0.014±0.006 | 0.012±0.005 | 0.010±0.004 | 0.008±0.002 |
| TRI | 0.784±0.031 | 0.891±0.019 | 0.989±0.012 | 1.004±0.010 | 1.011±0.008 | 1.007±0.007 | 1.003±0.005 |
| EI | 2.231±0.169 | 2.185±0.157 | 2.219±0.161 | 2.211±0.159 | 2.196±0.150 | 2.207±0.158 | 2.213±0.154 |
| RASI | 0.865±0.018 | 0.808±0.022 | 0.928±0.018 | 0.954±0.014 | 0.972±0.008 | 0.981±0.006 | 0.986±0.005 |
| SRI | 0.814±0.105 | 1.561±0.172 | 0.481±0.184 | 0.259±0.113 | 0.151±0.040 | 0.136±0.033 | 0.123±0.028 |
| SDI | 0.065±0.012 | 0.085±0.015 | 0.033±0.009 | 0.017±0.006 | 0.011±0.004 | 0.010±0.004 | 0.008±0.002 |
| MIA | 0.561±0.030 | 0.545±0.028 | 0.529±0.026 | 0.521±0.022 | 0.517±0.021 | 0.513±0.020 | 0.509±0.018 |
| $\Delta_{\text{DP}}$ | 0.071±0.036 | 0.075±0.035 | 0.019±0.015 | 0.016±0.012 | 0.013±0.008 | 0.014±0.010 | 0.017±0.012 |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | 0.060±0.032 | 0.063±0.035 | 0.027±0.020 | 0.022±0.017 | 0.019±0.014 | 0.024±0.019 | 0.023±0.018 |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | 0.078±0.040 | 0.085±0.047 | 0.025±0.018 | 0.022±0.015 | 0.017±0.012 | 0.019±0.014 | 0.021±0.016 |

Table 47: Structural-masking baselines vs. BMCR: single-feature unlearning of the most influential feature in *House*, averaged over MLP, ResNet, and FTformer.

| Metric | SFI | FPE | $\text{BMCR}_{0.1}$ | $\text{BMCR}_{0.5}$ | $\text{BMCR}_{1}$ | $\text{BMCR}_{5}$ | $\text{BMCR}_{10}$ |
|---|---|---|---|---|---|---|---|
| RIPV | 0.036±0.010 | 0.043±0.013 | 0.020±0.009 | 0.015±0.007 | 0.013±0.006 | 0.011±0.005 | 0.009±0.003 |
| TRI | 0.791±0.033 | 0.887±0.021 | 0.986±0.013 | 1.002±0.011 | 1.009±0.009 | 1.006±0.008 | 1.001±0.006 |
| EI | 2.224±0.172 | 2.179±0.160 | 2.215±0.164 | 2.207±0.156 | 2.192±0.153 | 2.203±0.155 | 2.210±0.151 |
| RASI | 0.861±0.020 | 0.812±0.024 | 0.925±0.019 | 0.951±0.015 | 0.969±0.009 | 0.979±0.007 | 0.984±0.006 |
| SRI | 0.826±0.109 | 1.548±0.178 | 0.493±0.179 | 0.267±0.118 | 0.158±0.044 | 0.141±0.036 | 0.128±0.030 |
| SDI | 0.067±0.013 | 0.088±0.016 | 0.035±0.010 | 0.019±0.007 | 0.012±0.005 | 0.011±0.003 | 0.009±0.003 |
| MIA | 0.558±0.032 | 0.542±0.030 | 0.526±0.027 | 0.519±0.024 | 0.515±0.022 | 0.511±0.021 | 0.507±0.019 |
| $\Delta_{\text{DP}}$ | 0.069±0.038 | 0.078±0.037 | 0.021±0.016 | 0.017±0.013 | 0.014±0.009 | 0.015±0.011 | 0.018±0.013 |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | 0.058±0.034 | 0.066±0.037 | 0.029±0.021 | 0.023±0.018 | 0.020±0.015 | 0.025±0.020 | 0.024±0.017 |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | 0.080±0.042 | 0.088±0.045 | 0.027±0.019 | 0.023±0.016 | 0.018±0.013 | 0.020±0.015 | 0.022±0.017 |

Table 48: Structural-masking baselines vs. BMCR: single-feature unlearning of the most influential feature in *Pol*, averaged over MLP, ResNet, and FTformer.

| Metric | SFI | FPE | $\text{BMCR}_{0.1}$ | $\text{BMCR}_{0.5}$ | $\text{BMCR}_{1}$ | $\text{BMCR}_{5}$ | $\text{BMCR}_{10}$ |
|---|---|---|---|---|---|---|---|
| RIPV | 0.037±0.011 | 0.045±0.014 | 0.021±0.010 | 0.016±0.008 | 0.014±0.007 | 0.012±0.006 | 0.010±0.004 |
| TRI | 0.798±0.035 | 0.881±0.023 | 0.983±0.014 | 1.004±0.012 | 1.011±0.010 | 1.007±0.009 | 1.002±0.007 |
| EI | 2.230±0.169 | 2.185±0.157 | 2.218±0.161 | 2.204±0.159 | 2.195±0.150 | 2.201±0.152 | 2.212±0.148 |
| RASI | 0.858±0.021 | 0.816±0.026 | 0.923±0.020 | 0.948±0.016 | 0.967±0.010 | 0.977±0.008 | 0.982±0.007 |
| SRI | 0.834±0.112 | 1.532±0.183 | 0.501±0.174 | 0.274±0.121 | 0.163±0.047 | 0.145±0.039 | 0.131±0.032 |
| SDI | 0.069±0.014 | 0.091±0.017 | 0.037±0.011 | 0.020±0.008 | 0.013±0.006 | 0.012±0.004 | 0.010±0.004 |
| MIA | 0.561±0.034 | 0.545±0.031 | 0.529±0.028 | 0.521±0.025 | 0.517±0.023 | 0.513±0.022 | 0.509±0.020 |
| $\Delta_{\text{DP}}$ | 0.071±0.040 | 0.080±0.039 | 0.023±0.017 | 0.018±0.014 | 0.015±0.010 | 0.016±0.012 | 0.019±0.014 |
| $\Delta_{\text{EO}}^{\text{TPR}}$ | 0.060±0.036 | 0.068±0.039 | 0.031±0.022 | 0.024±0.019 | 0.021±0.016 | 0.026±0.021 | 0.025±0.018 |
| $\Delta_{\text{EO}}^{\text{FPR}}$ | 0.082±0.044 | 0.090±0.047 | 0.029±0.020 | 0.024±0.017 | 0.019±0.014 | 0.021±0.016 | 0.023±0.018 |

# X PROXY LEAKAGE: ATTRIBUTE INFERENCE AND RELEARNING

**Data-only baseline.** Predicting $x_i$ from $x_{-i}$ alone, with no task model, gives $R^2 = 0.441$ (linear regression) and $0.669$ (random forest). This data-inherent predictability exists for any model, including the retrained oracle.

**Model-assisted attack.** The attacker additionally queries the victim model at candidate values of $x_i$ within the declared support and uses the responses, together with $x_{-i}$, to predict $x_i$. Model access raises $R^2$ to $0.813$ for the original model but only to $0.678$ for retraining and $0.686$ for BMCR (Table 3(b)); BMCR thus removes almost all model-specific leakage, consistent with Proposition 4.6(iii).

**Relearning.** We further fine-tune the BMCR model for up to 200 epochs on $x_{-i}$ and the true labels only, never reintroducing $x_i$. Direct dependence does not substantially reappear, and BMCR's attack

Table 49: Relearning robustness under further fine-tuning on retained data.

| Extra epochs | RIPV | TRI | BMCR attack $R^2$ | Retrained attack $R^2$ |
|---|---|---|---|---|
| 0 | 0.0075 | 0.990 | 0.686 | 0.678 |
| 100 | 0.0084 | 0.985 | 0.695 | 0.683 |
| 200 | 0.0089 | 0.982 | 0.697 | 0.686 |

$R^2$ tracks that of the retrained model.