

POUR: A Provably Optimal Method for Unlearning Representations via Neural Collapse

Anjie Le Can Peng Yuyuan Liu J. Alison Noble

Institute of Biomedical Engineering, University of Oxford, UK

Abstract

*In computer vision, machine unlearning aims to remove the influence of specific visual concepts or training images without retraining from scratch. Studies show that existing approaches often modify the classifier while leaving internal representations intact, resulting in incomplete forgetting. In this work, we extend the notion of unlearning to the representation level, deriving a three-term interplay between forgetting efficacy, retention fidelity, and class separation. Building on Neural Collapse theory, we show that the orthogonal projection of a simplex Equiangular Tight Frame (ETF) remains an ETF in a lower dimensional space, yielding a provably optimal forgetting operator. We further introduce the **Representation Unlearning Score (RUS)** to quantify representation-level forgetting and retention fidelity. Building on this, we introduce **POUR** (**P**rovably **O**ptimal **U**nlearning of **R**epresentations), a geometric projection method with closed-form (**POUR-P**) and a feature-level unlearning variant under a distillation scheme (**POUR-D**). Experiments on CIFAR-10/100 and PathMNIST demonstrate that POUR achieves effective unlearning while preserving retained knowledge, outperforming state-of-the-art unlearning methods on both classification-level and representation-level metrics. Code will be released upon acceptance of the paper.*

1. Introduction

The ability to selectively remove knowledge from trained models has become an increasingly important requirement for modern machine learning systems. Motivations include regulatory compliance with data protection laws (e.g., the “right to be forgotten”) [4, 12, 28], mitigating reliance on spurious correlations, and ensuring the safe deployment of large pre-trained models in sensitive domains such as autonomous driving and medical imaging. In these applications, models often need to forget outdated, biased, or privacy-sensitive visual data while retaining general visual understanding for reliable downstream use. Figure 1 illustrates this goal on the PathMNIST dataset, showing how our method erases the “adipose” class while preserving the remaining categories.

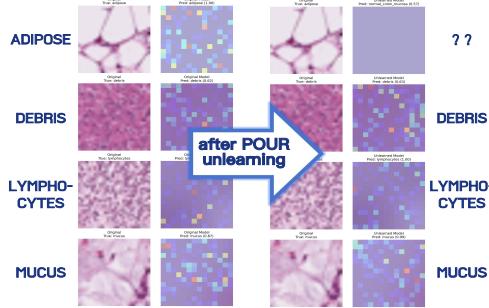


Figure 1. **Grad-CAM visualization on PathMNIST before and after unlearning.** Each row shows a tissue class. After applying POUR on the *adipose* class, its Grad-CAM signal vanishes, while the retained classes (*debris*, *lymphocytes*, *mucus*) preserve clear and distinct attention patterns.

A growing literature on machine unlearning has explored how to make models forget a specific class, subset, or concept without retraining from scratch [1, 8, 13]. Previous work on machine unlearning has primarily focused on aligning the prediction probabilities on the forget and retain sets. This line of research, often referred to as weak unlearning [14, 32], aims to ensure that the distributions of the final logits produced by the original and unlearned models are indistinguishable. However, recent studies [21] have questioned whether such methods truly forget the targeted information, as they often only perturb classifier logits while leaving the underlying feature representations largely unchanged. This shallow modification leaves residual information that can lead to privacy leakage [37]. This issue is particularly critical for deep vision encoders whose internal representations can still leak forgotten visual concepts through linear probing or feature inversion [13, 21].

In parallel, theoretical advances have revealed that deep visual classifiers exhibit highly structured geometric behavior at convergence. The theory of Neural Collapse (NC) shows that class features concentrate around equidistant centroids and classifier weights align to form a simplex Equiangular Tight Frame (ETF) [27]. This geometry provides a powerful lens for reasoning about class-level knowledge in image recognition: each class corresponds to a single ETF direction, and as we propose in this work, forgetting

a class corresponds to removing its associated vector from the representation space. Previous work [22] has pursued heuristic realizations of "projection as unlearning" through Singular Value Decomposition (SVD)-based decomposition in activation space. However, the method lacks geometric consistency and theoretical guarantees.

In this work, we first extend the traditional notion of weak unlearning to the *representation level*, and propose the **Representation Unlearning Score (RUS)** as a principled feature space metric for quantifying how well a model forgets. Building on this formulation, we observe that the restructuring of forget and retain representations occurs at different stages of unlearning, governed by class separation. We also establish two new properties from the current NC framework: (i) a simplex ETF structure certifies Bayes optimality in balanced classification, and (ii) the orthogonal projection of a simplex ETF remains a simplex ETF. Therefore, class forgetting can be implemented as a projection operator that preserves the NC geometry along the direction of the forget classes, which leads to our proposed algorithm **POUR** (**P**rovably **O**ptimal **U**nlearning of **R**epresentations).

POUR comes in two variants: a closed-form projection (**POUR-P**) that performs instantaneous forgetting, and a projection-guided distillation scheme (**POUR-D**) that propagates forgetting into the feature extractor using only the forget set through feature alignment.

In summary, our contributions are threefold:

- We reformulate machine unlearning at the representation level and introduce RUS based on feature-space discrepancy.
- We establish a three-term interplay among forget, retain and class separation for unlearning problems, and derive two new theoretical properties of NC geometry, linking the simplex ETF structure to Bayes optimality and projection invariance.
- We propose POUR, a provably optimal projection-based unlearning algorithm with both closed-form and feature-adaptive variants, and formally prove its optimality.

Experiments on CIFAR-10 and CIFAR-100 demonstrate that POUR effectively removes targeted visual concepts while preserving performance on retained classes. On PathMNIST, POUR further exhibits consistent generalization under domain shift, achieving reliable performance across both internal and external test sets.

2. Reformulating Unlearning: Representation Removal with the Forget Set

We define the class-centric machine unlearning problem with standard notation as follows. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the entire dataset, where each sample $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is a d -dimensional vector, $y_i \in \mathcal{Y} = \{1, 2, \dots, C\}$ is the ground truth label among C classes, and n is the size of \mathcal{D} . A training algorithm \mathcal{A} maps a dataset \mathcal{D} to a model

$M = (\theta, W)$, where $\theta : \mathcal{X} \rightarrow \mathcal{Z}$ is a feature extractor and W is a classifier head. For each sample, $M(\mathbf{x}_i)$ approximates its label y_i .

We partition \mathcal{D} into a retain set \mathcal{D}_r and a forget set \mathcal{D}_f , such that $\mathcal{D} = \mathcal{D}_f \cup \mathcal{D}_r$ and $\mathcal{D}_f \cap \mathcal{D}_r = \emptyset$. The goal of unlearning is to remove the influence of the forget set $\mathcal{D}_f \subset \mathcal{D}$ from the trained model while preserving performance on \mathcal{D}_r . Let $M_r = \mathcal{A}(\mathcal{D}_r)$ denote the reference model retrained from scratch using only \mathcal{D}_r . The unlearning process $\mathcal{U}(M, \mathcal{D})$ is then defined as a function that takes a trained model $M = \mathcal{A}(\mathcal{D})$ and produces a new model M_f that behaves similarly to M_r .

In the class-forgetting setting, if we wish to forget a class $u \in \mathcal{Y}$, then $\mathcal{D}_f = \mathcal{D}_u := \{\mathbf{x}_i : y_i = u\}$. The retain set \mathcal{D}_r is often inaccessible due to privacy or practical constraints. Following [5, 37], we therefore consider the realistic setting where unlearning is performed using only the forget set \mathcal{D}_f , denoted by $\mathcal{U}(M, \mathcal{D}_f)$.

2.1. Definition

Recent findings by Kim et al. [21] demonstrate that focusing solely on final logits does not guarantee complete forgetting, as the forgotten information may still be recoverable through linear probing. This observation highlights the need to investigate forgetting at the representation level, within the feature extractor itself, rather than only at the output layer. Motivated by this, we propose the concept of **representation-level weak unlearning**, which, in contrast to the original definition of weak unlearning, explicitly accounts for the internal feature representations of models.

Definition 2.1 (Representation-Level Weak Unlearning). *An unlearning procedure \mathcal{U} applied to (M, \mathcal{D}_f) is said to satisfy representation-level weak unlearning if the feature distributions of the unlearned model are close to those of the reference model M_r , i.e.*

$$\mathcal{K}\left(P_z^{\mathcal{U}(M, \mathcal{D}_f)}, P_z^{M_r}\right) < \epsilon, \quad (1)$$

for some distributional discrepancy measure \mathcal{K} (e.g. MMD, Wasserstein-2, or Energy Distance) and tolerance $\epsilon > 0$. Here P_z^M denotes the distribution of features $z = \theta(x)$ induced by model M on input x , where $x \sim \mathcal{D}$.

Intuitively, this condition requires that, after unlearning, the feature representations of \mathcal{D} are statistically indistinguishable from those produced by the retrained reference model M_r .

2.2. Practical Estimation of K

Due to the stochastic nature of training, comparing the feature distributions of the unlearned model and the retrained model is not straightforward. We require a representation measure that is robust to randomness, including random

initialization, rotations of the feature basis, and uniform rescaling of feature magnitudes. Therefore, we adopt Centered Kernel Alignment (CKA) [23] as a practical estimator of representation similarity, for its invariance to scaling and rotation; additional justification is provided in Appendix.

Formally, given two representation matrices $X, Y \in \mathbb{R}^{n \times d}$ from two models evaluated on the same set of n samples, their linear CKA similarity is defined as

$$\text{CKA}(X, Y) = \frac{\langle XX^\top, YY^\top \rangle_F}{\|XX^\top\|_F \|YY^\top\|_F}, \quad (2)$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product, making it robust to common randomness in neural networks.

In the unlearning setting, for M_o , M_f , and M_r , the original, unlearned, and retrained models, we define two families of CKA similarities:

$$\begin{aligned} \text{CKA}_f^{(o)} &:= \text{CKA}(M_f, M_o; \mathcal{D}_f), & \text{CKA}_r^{(o)} &:= \text{CKA}(M_f, M_o; \mathcal{D}_r), \\ \text{CKA}_f^{(r)} &:= \text{CKA}(M_f, M_r; \mathcal{D}_f), & \text{CKA}_r^{(r)} &:= \text{CKA}(M_f, M_r; \mathcal{D}_r). \end{aligned} \quad (3)$$

The superscript (o) indicates comparison with the original model, which is commonly available in practice, while (r) denotes comparison with the retrained model, which approximates the theoretical ideal.

To jointly balance the forgetting and retention objectives, we define the Representation Unlearning Score (RUS) as

$$\text{RUS}^{(*)} := \frac{2 \Phi_f^{(*)} \text{CKA}_r^{(*)}}{\Phi_f^{(*)} + \text{CKA}_r^{(*)}}, \quad (*) \in \{(o), (r)\}, \quad (4)$$

where

$$\Phi_f^{(o)} = 1 - \text{CKA}_f^{(o)}, \quad \Phi_f^{(r)} = \text{CKA}_f^{(r)}.$$

$\text{RUS}^{(r)}$ represents the evaluation using the retrained model as the reference, while $\text{RUS}^{(o)}$ provides a practical surrogate when the retrained model is inaccessible, for example, due to computational cost or data availability. $\text{RUS}^{(*)}$ corresponds to the harmonic mean of retention alignment CKA_r and the forgetting indicator $\Phi_f^{(*)}$, rewarding methods that achieve both effective forgetting and faithful retention. The definition ensures that both variants of RUS take values in $[0, 1]$ and increase with successful forgetting.

2.3. Theoretical Characterization

We next show that the discrepancy between unlearned and reference feature distributions can be decomposed into three interpretable components that directly capture forgetting efficacy, retention fidelity, and class separation.

Proposition 2.2 (Decomposition of \mathcal{K} Bound). *Let $P_z^{(f)}$ and $P_z^{(r)}$ denote the feature distributions induced by the unlearned and retrained models, respectively. For a forget*

class $u \in \mathcal{Y}$ and an Integral Probability Metric (IPM) \mathcal{K} defined on the feature space, by the law of total probability we can express

$$P_z^{(f)} = \alpha P_u^{(f)} + (1 - \alpha) P_{\neg u}^{(f)}, \quad P_z^{(r)} = \beta P_u^{(r)} + (1 - \beta) P_{\neg u}^{(r)}, \quad (5)$$

where $\alpha := P_z^{(f)}(\hat{y} = u)$ and $\beta := P_z^{(r)}(\hat{y} = u)$ are the predicted probabilities of the unlearning class under each model, and $P_u^{(\cdot)}, P_{\neg u}^{(\cdot)}$ denote the unlearned and retained class feature distribution. Then, the discrepancy between the unlearned and retrained feature distributions is bounded as

$$\begin{aligned} & \left| \alpha \mathcal{K}(P_u^{(f)}, P_u^{(r)}) - (1 - \alpha) \mathcal{K}(P_{\neg u}^{(f)}, P_{\neg u}^{(r)}) \right| - |\alpha - \beta| \Delta_c \\ & \leq \mathcal{K}(P_z^{(f)}, P_z^{(r)}) \\ & \leq \underbrace{|\alpha - \beta| \Delta_c}_{\text{class separation}} + \underbrace{\alpha \mathcal{K}(P_u^{(f)}, P_u^{(r)})}_{\text{forgotten-class discrepancy}} \\ & \quad + \underbrace{(1 - \alpha) \mathcal{K}(P_{\neg u}^{(f)}, P_{\neg u}^{(r)})}_{\text{retained-class discrepancy}}, \end{aligned} \quad (6)$$

where $\Delta_c := \mathcal{K}(P_u^{(r)}, P_{\neg u}^{(r)})$.

A complete statement and proof is given in Appendix 1.1. When unlearning is performed on the forget set, we have $\beta = 0$, and the forgetting coefficient α decreases gradually from approximately 1 to 0 as unlearning proceeds. In this regime, the effective supervision target becomes $P_{\neg u}^{(r)}$, while both $P_u^{(f)}$ and $P_{\neg u}^{(f)}$ are progressively aligned toward the retained-class manifold of the retrained model at different stages of unlearning. The class-separation term simplifies to $\alpha \Delta_c$, indicating that stronger geometric separation in the retrained feature space enables more effective guidance for forgetting at early stages. Consequently, unlearning can often be achieved using the forget set alone when class separation is sufficient; however, when separation in the retrained model is weak, the forget-set-only strategy becomes less effective. This phenomenon is also confirmed empirically, as discussed in Section 5.2.

3. Representation Space: Neural Collapse and Simplex ETF Geometry

The trade-off derived above reveals that unlearning dynamics are fundamentally governed by the geometry of class separation in representation space. To analyze this geometry, we draw inspiration from the theory of Neural Collapse (NC), which shows that deep classifiers trained with cross-entropy loss organize features into a Simplex ETF during the Terminal Phase of Training (TPT) [27] under certain assumptions described in Appendix 2.

Formally, for each class i , the learned feature representation takes the form

$$z_\theta(x) = \alpha(x) v_i,$$

where $z_\theta(x)$ denotes the feature extractor θ applied to input x , $\alpha(x) > 0$ is a class-dependent scaling factor, and $v_i \in \mathbb{R}^d$ is a unit direction representing the class mean. The set of class directions $\{v_i\}_{i=1}^C$ lies in a $(C-1)$ -dimensional subspace and forms a simplex ETF, i.e.,

$$\|v_i\| = 1, \quad v_i^\top v_j = -\frac{1}{C-1} \text{ for } i \neq j, \quad \text{and} \quad \sum_{i=1}^C v_i = 0,$$

which implies that class means are maximally separated and symmetrically arranged in feature space. Furthermore, the classifier's weight vectors align with these class directions.

A detailed formulation of the underlying assumptions and the full NC statements are included in Appendix 2. An empirical investigation of the NC phenomenon is given in Section 5.4. This provides a natural foundation for understanding and manipulating class forgetting at the representation level.

In prior work, this ETF geometry has primarily been regarded as a descriptive limit of training dynamics. In this work, we establish two new properties of the NC phenomenon. First, we show that the simplex ETF geometry is not only a consequence of optimization, but also a sufficient condition for Bayes optimality under natural statistical assumptions. In this sense, the ETF structure serves as an optimality certificate. Second, we demonstrate that the ETF geometry is preserved under orthogonal projection when one vertex is removed, thereby providing the geometric foundation for our proposed unlearning method.

3.1. ETF as an Optimal Condition

In addition to the NC assumptions, we further assume that the class-conditional feature distributions are isotropic Gaussians,

$$x | y = i \sim \mathcal{N}(v_i, \sigma^2 I_d),$$

where $\|v_i\| = 1$ for all i and $\sigma^2 > 0$ is fixed. Empirically, this corresponds to features within each class clustering around a well-defined mean with approximately uniform variance in all directions, consistent with an isotropic Gaussian structure. In practice, datasets with sufficiently large sample sizes naturally satisfy this assumption by the Law of Large Numbers. Under these assumptions, we have the following proposition:

Proposition 3.1 (ETF \Rightarrow Bayes optimality).

- (i) the simplex ETF uniquely maximizes the minimum pairwise angle among class means,
- (ii) it maximizes the multiclass angular margin of the nearest-class-mean classifier, and
- (iii) in the limit $\kappa \rightarrow \infty$ or as the within-class variance $\sigma^2 \rightarrow 0$, the induced decision rule coincides with the Bayes-optimal classifier.

Detailed proof of is provided in Appendix 3.

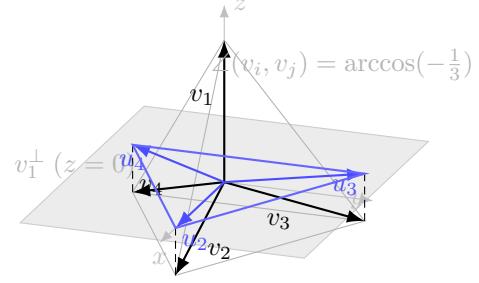


Figure 2. $C=4$ simplex ETF. One vertex v_1 along $+z$; the other three lie at $z = -1/3$ with equal 120° separation in xy . Orthogonal projection onto v_1^\perp ($z = 0$) yields an equilateral triangle formed by u_2, u_3, u_4 .

3.2. ETF Stability under Projection.

The second property concerns the robustness of ETF geometry under dimensionality reduction. Geometrically, removing one vertex of a regular simplex and projecting the remaining vertices onto the complementary subspace yields a smaller regular simplex. This effect, visualized in Figure 2 for a $C = 4$ case, is captured in the following proposition.

Proposition 3.2 (Projection of a simplex ETF remains a simplex ETF). *With the assumption and notations above, fix $u \in \{1, \dots, C\}$ and let $P = I - v_u v_u^\top$ be the orthogonal projector onto v_u^\perp . For $i \in \mathcal{Y}_{\neg u}$, define $g_i = \frac{P v_i}{\|P v_i\|}$. Then $\{g_i\}_{i \in \mathcal{Y}_{\neg u}} \subset v_u^\perp \cong \mathbb{R}^{C-2}$ is a simplex ETF of size $C-1$, i.e. $g_i^\top g_j = -\frac{1}{C-2} (\forall i \neq j), \sum_{i \in \mathcal{Y}_{\neg u}} g_i = 0$.*

A proof of Proposition 3.2 is given in Appendix 4.1. This invariance implies that class forgetting via orthogonal projection maintains perfect angular separation among retained classes, forming the geometric basis of our POUR method.

4. Method

Inspired by these theoretical insights, we propose our method POUR, which comes in two variants: a one-shot projection operation on model weights enabled by Proposition 3.2, which we call **POUR-P**; and a distillation version using the forget set, which we call **POUR-D**, as summarized in Fig. 3.

4.1. Projection Operator (POUR-P)

We consider the class forgetting setting, where we want to forget a class u , so that $\mathcal{D}_f = \mathcal{D}_u := \{x_i : y_i = u\}$. Let $W \in \mathbb{R}^{d \times C}$ denote the classifier weight matrix, where each column w_c corresponds to class c , and let $z \in \mathbb{R}^d$ denote the penultimate-layer feature. In the NC regime, both $\{w_c\}_{c=1}^C$ and the class means of $\{z\}$ form a simplex ETF. To forget a class u , we define the orthogonal projection operator

$$P = I - \frac{w_u w_u^\top}{\|w_u\|^2}, \quad (7)$$

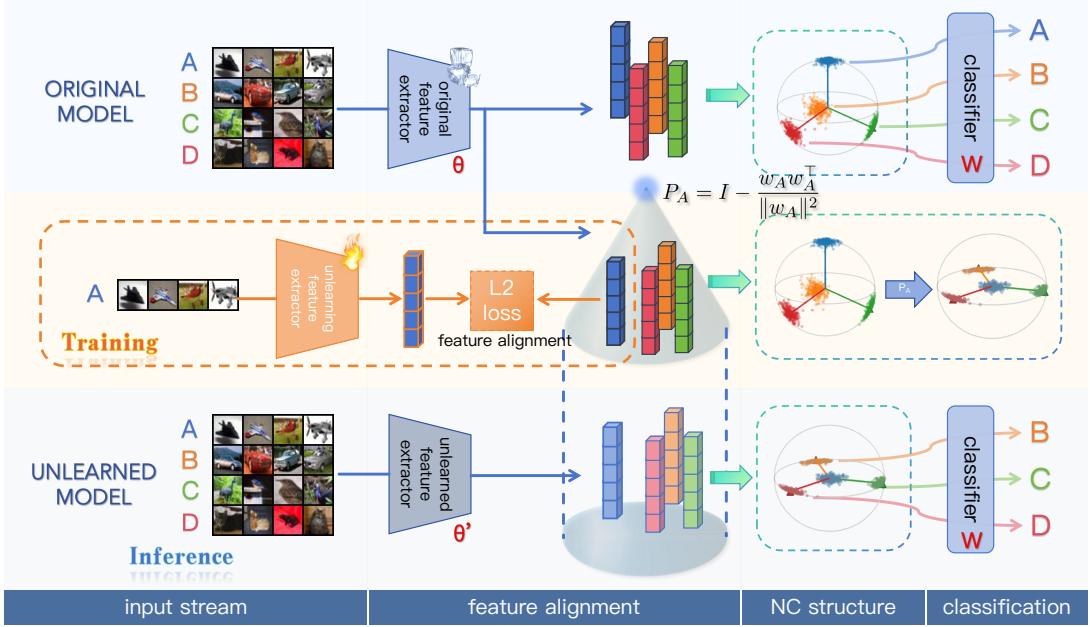


Figure 3. Overview of the POUR framework. During training, the unlearning module applies an orthogonal projection operator P_A on the feature space of the original model to remove the contribution of the forgotten class A . The unlearned feature extractor θ' is optimized via an L_2 loss to align its projected features with those of the original extractor θ using the unlearning data. This alignment preserves the Neural Collapse geometry among retained classes (B, C, D) while collapsing features of the forgotten class to the origin, leading to uniform predictions. At inference, the unlearned model is Bayes-optimal on retained classes as proved in Theorem 4.2.

which removes the contribution of the forgotten class direction w_u . The unlearned features can then be obtained by

$$z' = P^\top z = Pz. \quad (8)$$

Directly applying this projection after the feature extractor is what we call **POUR-P**. By Proposition 3.2, this operation maps features into a $(C - 1)$ -class simplex ETF subspace, thereby preserving optimal geometry among the retained classes.

Practical estimation of P . When classifier weights are not directly available, or when only the feature encoder is available (e.g., for vision-language models), we can estimate w_u as the empirical class mean of penultimate features,

$$\tilde{w}_u = \frac{1}{|\mathcal{D}_u|} \sum_{x \in \mathcal{D}_u} \theta_o(x), \quad (9)$$

where θ_o denotes the original feature extractor. The projection operator P can then be constructed using \tilde{w}_u .

4.2. Projection-Guided Distillation (POUR-D)

While POUR-P provides an immediate, closed-form forgetting operation, it only modifies features post hoc. To induce forgetting deeper into the feature extractor and improve robustness, we introduce a teacher-student distillation [18] scheme, **POUR-D**.

Teacher construction. We use the projected model POUR-P as the teacher. Given a trained model (θ, W) and a forget-

class u , we apply the projection operator P from Equation 7 to obtain a projected teacher model $(P\theta, W)$ which encodes the post-forgetting ETF geometry in the representation space.

Student training. The student model finetunes the feature extractor parameters on the forget set to align with the teacher model. In particular, for θ the feature extractor of the original model and θ_s the feature extractor of the student model, we minimize the L_2 loss defined as:

$$\mathcal{L}_{\text{POUR-D}}(x) = \| \theta_s(x) - P\theta(x) \|_2^2, \quad x \in \mathcal{D}_f.$$

Under NC, the class means form a simplex ETF, and the classifier head aligns with them. Projection preserves this ETF structure for retained classes. This loss penalizes deviations from the projected ETF features, ensuring that the student model remains aligned in both direction and scale with the teacher, while requiring minimal updates to the feature extractor parameters. Convergence can be guaranteed by the following proposition.

Proposition 4.1 (L2 convergence implies CKA convergence). *Let $Z, T \in \mathbb{R}^{n \times p}$ be row-centered, and assume $TT^\top \neq 0$. If $\|Z - T\|_F \rightarrow 0$, then $\text{CKA}(Z, T) \rightarrow 1$.*

4.3. Optimality of Projection for Weak Unlearning

We now show that the proposed projection operator is optimal under the definition of representation-level weak unlearning (Def. 2.1). Projecting onto the orthogonal complement of the forgotten class removes its contribution while

Table 1. Comparison of unlearning methods for ResNet18 on CIFAR-10. We report both classification-level metrics and representation-level metrics. Methods requiring access to the retain set are included as reference values and not included in ranking. Values represent mean \pm std across three runs. Best values are in **bold**, and second-best values are underlined. Darker blue | indicates better performance.

Method	Classification-Level Metrics					Representation-Level Metrics						
	Acc _r \uparrow	Acc _f \downarrow	Acc _{tr} \uparrow	Acc _{tf} \downarrow	AUS \uparrow	rMIA \downarrow	CKA _f ^(o) \downarrow	CKA _r ^(o) \uparrow	RUS ^(o) \uparrow	CKA _f ^(r) \uparrow	CKA _r ^(r) \uparrow	RUS ^(r) \uparrow
Original Model	94.47 _{±0.12}	95.03 _{±0.35}	99.99 _{±0.00}	99.99 _{±0.01}	0.51 _{±0.00}	56.70 _{±0.00}	1.00 _{±0.00}	1.00 _{±0.00}	0.00 _{±0.00}	0.26 _{±0.01}	0.98 _{±0.00}	0.42 _{±0.01}
Retrained Model	94.68 _{±0.38}	0.00 _{±0.00}	99.98 _{±0.01}	0.00 _{±0.00}	1.00 _{±0.00}	—	0.26 _{±0.03}	0.97 _{±0.01}	0.84 _{±0.01}	1.00 _{±0.00}	1.00 _{±0.00}	1.00 _{±0.00}
<i>Methods requiring the retain set</i>												
Finetune	93.96 _{±0.19}	0.00 _{±0.00}	100.00 _{±0.00}	0.00 _{±0.00}	0.99 _{±0.00}	54.60 _{±1.37}	0.32 _{±0.01}	0.97 _{±0.01}	0.80 _{±0.00}	0.63 _{±0.02}	0.97 _{±0.00}	0.76 _{±0.02}
FCS	94.89 _{±0.01}	0.67 _{±0.55}	100.00 _{±0.00}	0.79 _{±0.67}	1.00 _{±0.00}	56.00 _{±0.61}	0.51 _{±0.09}	1.00 _{±0.02}	0.66 _{±0.08}	0.45 _{±0.01}	0.98 _{±0.00}	0.62 _{±0.01}
<i>Methods on forget set only</i>												
Random Label	87.42 _{±0.54}	23.20 _{±0.44}	93.13 _{±0.44}	25.13 _{±0.34}	0.75 _{±0.00}	54.07 _{±1.31}	0.29 _{±0.05}	0.86 _{±0.02}	0.78 _{±0.02}	0.24 _{±0.01}	0.84 _{±0.00}	0.37 _{±0.01}
Gradient Ascent	86.71 _{±0.04}	15.37 _{±4.13}	93.51 _{±4.00}	16.29 _{±3.54}	0.80 _{±0.01}	50.40 _{±0.82}	0.21 _{±0.06}	0.80 _{±0.07}	0.79 _{±0.02}	0.18 _{±0.01}	0.77 _{±0.05}	0.29 _{±0.02}
Boundary Shrink	85.30 _{±1.66}	12.33 _{±2.14}	90.81 _{±1.38}	13.96 _{±2.12}	0.81 _{±0.01}	53.07 _{±1.10}	0.25 _{±0.04}	0.85 _{±0.02}	<u>0.80</u> _{±0.01}	0.28 _{±0.01}	0.84 _{±0.00}	<u>0.42</u> _{±0.01}
Boundary Expand	85.74 _{±0.55}	14.63 _{±0.06}	91.21 _{±0.44}	16.66 _{±0.37}	0.80 _{±0.00}	53.00 _{±0.78}	0.25 _{±0.04}	0.85 _{±0.02}	<u>0.80</u> _{±0.01}	0.28 _{±0.01}	0.83 _{±0.00}	0.42 _{±0.01}
DELETE	88.73 _{±2.32}	2.43 _{±0.12}	95.43 _{±1.78}	2.93 _{±0.46}	0.92 _{±0.02}	53.43 _{±0.40}	0.37 _{±0.09}	0.82 _{±0.03}	0.71 _{±0.05}	0.26 _{±0.01}	0.78 _{±0.02}	0.39 _{±0.01}
POUR-P [†] (ours)	94.97 _{±0.16}	0.00 _{±0.00}	99.99 _{±0.00}	0.00 _{±0.00}	1.01 _{±0.00}	56.67 _{±1.08}	—	—	—	—	—	—
POUR-D (ours)	92.86 _{±1.02}	0.37 _{±0.64}	99.74 _{±0.29}	0.43 _{±0.44}	0.97 _{±0.00}	51.80 _{±2.42}	0.23 _{±0.06}	0.95 _{±0.02}	0.82 _{±0.03}	0.31 _{±0.01}	0.94 _{±0.00}	0.47 _{±0.01}

[†] POUR-P does not modify the encoder representations; therefore, representation-level metrics would be unchanged and are omitted.

preserving the Bayes-optimal ETF geometry of the retained classes.

Theorem 4.2 (Optimality of POUR-P). *Assume (A1)–(A5) as in Appendix 2 in the model training pipeline, and class priors are balanced and isotropic Gaussians, as described in Section 3.1. Let $\theta(x)$ denote the penultimate layer features, v_i the class means, then by NC, we have $\theta(x) \mid (y = i) \sim \mathcal{N}(v_i, \sigma^2 I_d)$, where $\{v_i\}_{i=1}^C$ form a simplex ETF.*

Now fix a class $u \in \mathcal{Y}$ and define the orthogonal projection $P = I - v_u v_u^\top$, projected features $\theta'(x) = P\theta(x)$, and $\tilde{v}_i = Pv_i/\|Pv_i\|$ for $i \neq u$. Then:

- (a) **Retained optimality and ETF equivalence.** The projected means $\{\tilde{v}_i\}_{i \neq u}$ form a simplex ETF (Prop. 3.2). The retained-class ETF of the projected model and that of M_r differ by at most an orthogonal transform, i.e., for any discrepancy \mathcal{K} invariant under orthogonal transforms and rescaling, $\mathcal{K}(P_{\neg u}, Q_{\neg u}) = 0$. Moreover, under the Gaussian model, when class means dominate intra-class variability, the projected model is Bayes-optimal (Prop. 3.1).
- (b) **Complete forgetting.** Since $Pv_u = 0$, features of the forgotten class satisfy $\theta'(x) \mid (y = u) \sim \mathcal{N}(0, \sigma^2 P)$. In the low-variance (NC) limit $\sigma^2 \rightarrow 0$, $\theta'(x) \rightarrow 0$ and all retained logits vanish, so $q_{\neg u}(\cdot|x) \rightarrow U_{\neg u}$: the forgotten class is represented by a uniform predictive distribution among the retained classes, i.e., $\alpha = 0$.

Complete statement and proof are included in Appendix 4.2. Consequently, the POUR-P projection yields a representation that is (i) Bayes-optimal on $\mathcal{Y}_{\neg u}$ and (ii) representation-equivalent to the retrained model up to orthogonal gauge freedom, so that the representation-level discrepancy $\mathcal{K}(P_z^{(f)}, P_z^{(r)})$ attains its minimum under Def. 2.1.

5. Experiments and Results

We evaluate both POUR-P and POUR-D on CIFAR-10/100 with ResNet-18 and PathMNIST with pretrained ViT-S/16.

5.1. Experimental Setup

Protocol constraints. We follow the standard unlearning setting in recent literature [37], that assumes (i) no access to the retained set \mathcal{D}_r during unlearning, and (ii) no intervention in the original training procedure. All methods are applied directly to the already trained model.

Datasets and Models. For CIFAR-10 and CIFAR-100, we implement modified ResNet-18 backbones in which the initial 7×7 convolution (stride 2) is replaced by a 3×3 convolution (stride 1) and the subsequent max-pooling layer is removed to better suit 32×32 inputs. For PathMNIST [34], we tested in a pretraining setting, where we loaded a ViT-S/16 pretrained on ImageNet and trained a classifier head. We evaluate performance on both internal and external test sets for the same task, which exhibits a domain shift.

Baselines. We benchmark POUR-P and POUR-D against a diverse set of existing unlearning strategies, including Finetune, FCS [2], Random Label [15], Gradient Ascent [15], Boundary Shrink, Boundary Expand [6], and DELETE [37]. Original Model and Retrain Model serve as the lower and upper bounds, respectively.

Metrics. We report the following metrics:

- **Acc_f, Acc_{tf} (%) \downarrow** and **Acc_r, Acc_{tr} (%) \uparrow** : validation and training accuracy on the forget and retain sets, respectively.
- **Adaptive Unlearning Score (AUS) (\uparrow)** [9, 25] : Jointly capture retention and forgetting accuracy, defined as $AUS = \frac{1 - drop_r}{1 + acc_f}$, where $drop_r$ and acc_f denote the drop in retain accuracy and forgotten-class accuracy.
- **rMIA (%) \downarrow** : representation-level membership-inference attack success rate on \mathcal{D}_f . We perform a five-fold attack

Table 2. Comparison of unlearning methods for ResNet18 on CIFAR-100. Best values are in **bold**, and second-best values are underlined. Darker blue indicates better performance.

Method	Acc _r ↑	Acc _f ↓	AUS ↑	CKA _f ^(r) ↑	CKA _r ^(r) ↑	RUS ^(r) ↑	rMIA ↓
Original Model	77.69	92.00	<u>0.52</u>	0.60	0.78	0.68	62.00
Retrained Model	76.28	0.00	<u>1.00</u>	1.00	1.00	1.00	—
<i>Methods requiring the retain set</i>							
Finetune	76.32	0.00	<u>0.99</u>	0.57	0.76	0.67	54.00
FCS	76.81	2.00	<u>0.97</u>	0.61	0.78	0.68	55.00
<i>Methods on forget set only</i>							
Random Label	61.98	11.00	<u>0.76</u>	0.40	0.53	0.46	49.00
Gradient Ascent	50.46	6.00	<u>0.69</u>	0.44	0.44	0.44	50.00
Boundary Shrink	68.87	4.00	<u>0.88</u>	<u>0.55</u>	0.72	0.62	49.00
Boundary Expand	66.47	13.00	<u>0.79</u>	<u>0.55</u>	<u>0.74</u>	<u>0.63</u>	42.00
DELETE	64.67	8.00	0.81	0.51	0.67	0.58	60.00
POUR-P [†] (ours)	77.65	0.00	<u>1.00</u>	—	—	—	62.00
POUR-D (ours)	73.44	1.00	<u>0.95</u>	<u>0.57</u>	<u>0.76</u>	<u>0.65</u>	<u>46.00</u>

[†] POUR-P does not modify the encoder representations.

using a linear regressor on the representation between the train and test sets.

- CKA_f^(o), CKA_r^(o), RUS^(o) ($\downarrow, \uparrow, \uparrow$) and CKA_f^(r), CKA_r^(r), RUS^(r) ($\uparrow, \uparrow, \uparrow$): as defined in Sect. 2.2.

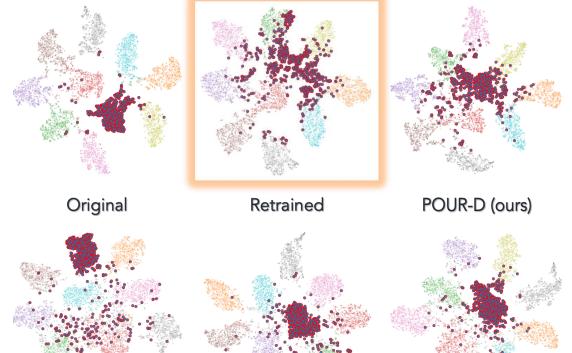
5.2. Unlearning on CIFAR-10/100

On CIFAR-10, as shown in Table 1, our method achieves the best performance for both the classification-level metric (AUS) and the representation-level metrics (RUS), suggesting that POUR enables efficient forgetting directly in representation space. In contrast, other methods, such as Gradient Ascent, Boundary Shrink, and DELETE, show lower AUS or RUS scores, indicating that their forgetting is either incomplete or occurs primarily at the classifier layer without sufficiently modifying the underlying representation geometry, as visualized in Figure 4a. This also provide a parallel comparison of the two variants of the RUS. In general, these two scores establish a similar trend.

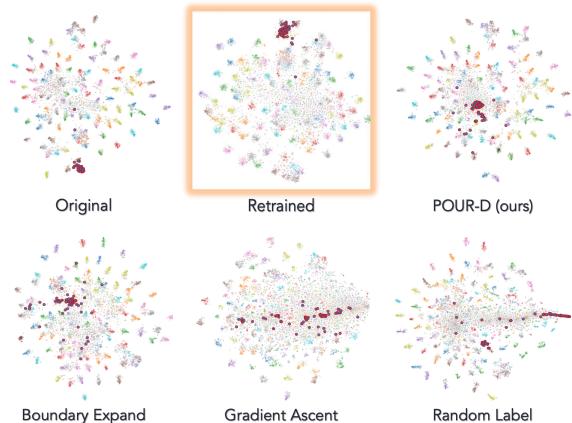
On the more challenging CIFAR-100 dataset, as shown in Table 2, our method again achieves state-of-the-art (SOTA) performance for both classification and representation-level metrics. We note that on CIFAR-100, classes are more entangled, as shown as a high CKA_f^(r) and visualized in Figure 4b. Therefore, supervision on the forget set is lower and therefore forgetting is harder, as discussed in Section 2.3. Methods such as gradient ascent and random labels largely disrupt the structure of the retained classes. Boundary Shrink and Boundary Expand, though among the stronger baselines, fail to reproduce the structure of the retrained model representations as effectively as POUR.

5.3. Unlearning on PathMNIST

PathMNIST exhibits a substantial domain shift between its internal and external test sets due to differences in slide acquisition. The ViT backbone is pretrained on ImageNet with a finetuned classifier head on PathMNIST, which makes this setting challenging for existing unlearning methods, as the learning is shallow.



(a) t-SNE visualization on CIFAR10.



(b) t-SNE visualization on CIFAR100.

Figure 4. t-SNE visualization of representation spaces after unlearning on CIFAR-10 and CIFAR-100. Each color denotes a retained class, with dark red points represent the forgotten class. The Gold panel shows the representation of the retrained model, serving as the ideal reference for successful unlearning. Structure of representations after POUR unlearning mostly resemble that of the retrained gold model.

Our method again achieves SOTA in this setting. As shown in Figure 1, the activation on the forget set vanishes after POUR-P unlearning, meaning unlearning signals successfully propagate from the finetuned classifier head into the pretrained backbone. Methods like Boundary Shrink and Boundary Expand fail in this setting. Random Label and Gradient Ascent exhibit higher classification-level performance on the internal test set than on the external test set. We hypothesize that these methods are "learning to mask" the forget set rather than genuinely erasing the associated knowledge, resulting in poor generalization. In contrast, DELETE and POUR achieve consistent performance across both domains, suggesting that they perform true knowledge removal rather than overfitting to the forget set. The effectiveness of the unlearning methods in this setting provides a scalable pathway for unlearning in pretrained and foundation

Table 3. Comparison of unlearning methods on PathMNIST with ViT. Performance is reported on both the *internal* and *external* test sets under domain shift. Best values are in **bold**, and second-best values are underlined. Darker blue indicates better performance.

Method	ViT on PathMNIST Internal Test					ViT on PathMNIST External Test				
	Acc _r ↑	Acc _f ↓	AUS ↑	RUS ^(o) ↑	rMIA ↓	Acc _r ↑	Acc _f ↓	AUS ↑	RUS ^(o) ↑	rMIA ↓
Original Model	87.49	96.83	0.51	—	50.43	87.13	97.53	0.51	—	84.20
Retrained Model*	88.70	0.00	1.01	—	—	88.63	0.00	1.02	—	—
<i>Methods requiring the retain set</i>										
Finetune	97.78	0.00	1.10	0.05	46.67	86.99	0.00	1.00	0.06	86.10
FCS	88.81	0.00	1.01	0.13	49.00	83.86	0.00	0.97	0.10	96.60
<i>Methods on forget set only*</i>										
Random Label	78.71	23.73	0.74	0.26	48.23	70.61	25.34	0.67	0.29	83.60
Gradient Ascent	81.43	9.03	0.86	0.22	48.71	76.72	10.61	0.81	0.26	83.90
DELETE	72.75	0.00	0.85	0.50	49.76	73.04	0.00	0.86	0.42	88.40
POUR-P (ours)	87.14	0.00	1.00	—	50.43	87.44	0.00	1.00	—	84.20
POUR-D (ours)	81.09	7.88	0.87	0.63	51.00	80.90	7.92	0.87	0.61	85.20

*Note: Boundary Shrink and Boundary Expand did not work in this setting. Retraining only performed on classifier head.

models, where original data may be unavailable.

5.4. NC as an Assumption

Our analysis relies on the theory of NC, which typically emerges under sufficient overparameterization. We also found in practice that standard training naturally reaches a regime where NC is sufficiently well established for POUR to be effective. Figure 5 shows the classifier weight angle distributions across datasets, revealing how closely the trained models conform to NC geometry. The empirical mean angles align almost perfectly with the ideal simplex ETF angle on all of the datasets.

6. Related Work

The problem of removing specific training data from a model, often motivated by privacy regulations such as the “right to be forgotten,” was first formalized in the systems security community [3]. The seminal work of Bourtoule et al. [1] introduced the SISA framework, partitioning training data across multiple shards so that forgetting can be achieved by retraining only the affected shards. Subsequent work developed more fine-grained methods that avoid full retraining. For linear models, Guo et al. [16] proposed certified removal via influence-based updates. Sekhari et al. [30] provided theoretical guarantees for approximate unlearning in general models. For deep networks, approaches include *amnesiac unlearning* [15], which inverts stored gradients, and Fisher information-based scrubbing [13, 14], which perturbs weights along sensitive directions. Other efficient methods use adversarial weight perturbations [31], incompetent teachers [7], or zero-shot synthetic forget data [8]. Most recently, anchored fine-tuning methods such as FAMR [29] enforce uniform predictions on forget sets while constraining the model to remain close to its original parameters. Kodge et al. [22] proposed a gradient-free method that explicitly computes class-specific subspaces via singular value decom-

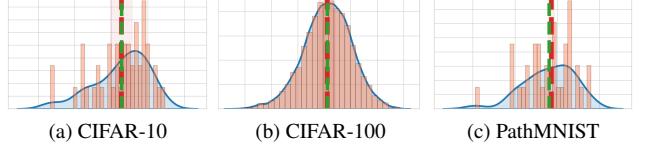


Figure 5. **Classifier weight angle distributions.** The green dashed line denotes the mean pairwise angle, while the red dashed line marks the ideal NC angle. The closeness between the two reflects how well the classifier aligns with NC geometry at convergence.

position and suppresses discriminatory directions associated with the forget class. Boundary Shrink and Boundary Expand [6] perform local decision-boundary adjustments for forgetting, while maintaining model utility through margin control. DELETE [37] formulates unlearning as a decoupled distillation problem, erasing class-specific information via probability decoupling. Yet, none of these approaches connect to Neural Collapse theory [27], where class features converge to a simplex ETF.

7. Conclusion

We introduced a representation-level formulation of machine unlearning and a new metric, RUS, to rigorously quantify forgetting beyond classifier logits. By connecting unlearning to NC geometry, we derived new theoretical insights which enabled POUR. Extensive experiments across CIFAR-10/100 and PathMNIST confirm that POUR achieves better forgetting than prior approaches at both the classification and representation levels, providing a principled geometric perspective on machine unlearning.

References

- [1] Laurent Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Haoche Jia, Adeline Travers, Benjamin Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (SP)*, pages 141–159, 2021. [1](#), [8](#)
- [2] Xavier F Cadet, Anastasia Borovykh, Mohammad Malekzadeh, Sara Ahmadi-Abhari, and Hamed Haddadi. Deep unlearn: Benchmarking machine unlearning. *arXiv preprint arXiv:2410.01276*, 2024. [6](#)
- [3] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (SP)*, pages 463–480, 2015. [8](#), [1](#)
- [4] CCPA. California consumer privacy act of 2018, 2018. California Civil Code §1798.100 et seq. [1](#)
- [5] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Tae-sup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11186–11194, 2024. [2](#)
- [6] Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [6](#), [8](#), [1](#)
- [7] Vishwaraj S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 7210–7217, 2023. [8](#), [1](#)
- [8] Vishwaraj S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan S. Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023. [1](#), [8](#)
- [9] Marco Cotogni, Jacopo Bonato, Luigi Sabetta, Francesco Pelosin, and Alessandro Nicolosi. Duck: Distance-based unlearning via centroid kinematics. *arXiv preprint arXiv:2312.02052*, 2023. [6](#)
- [10] Huy Dang, Zhenyu Yang, Qinghua He, Jiadong Wang, Taiji Wei, Zhizheng Li, Chenliang Gong, Shuaiwen Hu, and Ying-bin Liang. Neural collapse for cross-entropy class-imbalanced regime. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. [4](#)
- [11] Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021. [4](#)
- [12] GDPR. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 (general data protection regulation), 2016. Official Journal of the European Union, L 119, 1–88. [1](#)
- [13] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9309, 2020. [1](#), [8](#)
- [14] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European Conference on Computer Vision (ECCV)*, pages 383–398, 2020. [1](#), [8](#)
- [15] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11516–11524, 2021. [6](#), [8](#), [1](#)
- [16] Chuan Guo, Tom Goldstein, Awini Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019. [8](#), [1](#)
- [17] X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations (ICLR)*, 2022. [4](#)
- [18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. [5](#)
- [19] Weihao Hong and Shuyang Ling. Neural collapse for unconstrained feature model under class-imbalanced regime. *Journal of Machine Learning Research*, 25, 2024. [4](#)
- [20] Arthur Jacot, Peter Šuknič, Zihan Wang, and Marco Mondelli. Wide neural networks trained with weight decay provably exhibit neural collapse, 2024. [4](#)
- [21] Yongwoo Kim, Sungmin Cha, and Donghyun Kim. Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. *arXiv preprint arXiv:2503.06991*, 2025. [1](#), [2](#)
- [22] Sangamesh Kodge, Gobinda Saha, and Kaushik Roy. Deep unlearning: Fast and efficient gradient-free class forgetting. *Transactions on Machine Learning Research (TMLR)*, 2024, to appear. [2](#), [8](#), [1](#)
- [23] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMIR, 2019. [3](#)
- [24] Alexey Kravets and Vinay P. Namboodiri. Zero-shot class unlearning in CLIP with synthetic samples. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. [1](#)
- [25] Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. [6](#)
- [26] Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. [4](#)
- [27] Vardan Papyan, X.Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117(40):24652–24663, 2020. [1](#), [3](#), [8](#), [4](#)
- [28] PIPL. Personal information protection law of the people’s republic of china, 2021. Adopted at the 30th Meeting of the Standing Committee of the Thirteenth National People’s Congress. [1](#)

- [29] Prabhav Sanga, Jaskaran Singh, and Arun Kumar Dubey. Train once, forget precisely: Anchored optimization for efficient post-hoc unlearning. *arXiv preprint arXiv:2506.14515*, 2025. 8, 1
- [30] Aditi Sekhari, Jayadev Acharya, Gautam Kamath, and Abhradeep Thakurta Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18075–18086, 2021. 8, 1
- [31] 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. 8, 1
- [32] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey, 2023. 1
- [33] Hongren Yan, Yuhua Qian, Furong Peng, Jiachen Luo, Zhe-qing Zhu, and Feijiang Li. Neural collapse to multiple centers for imbalanced data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 4
- [34] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 6
- [35] Tianyu Yang, Lisen Dai, Xiangqi Wang, Minhao Cheng, Yapeng Tian, and Xiangliang Zhang. CLIPErase: Efficient unlearning of visual-textual associations in CLIP. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. 1
- [36] Yufan Zhang, Chenyang Si, Jianfeng Zhang, Yingcong Chen, and Wayne Wu. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1
- [37] Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 6, 8
- [38] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4

POUR: A Provably Optimal Method for Unlearning Representations via Neural Collapse

Supplementary Material

Appendix Table of Contents

- 0. Related Work
- 1. Additional Justifications
 - 1.1. Proof on Decomposition of \mathcal{K} -Bound
 - 1.2. Justification on CKA Usage
- 2. Neural Collapse
 - 2.1. Training Assumptions
 - 2.2. Neural Collapse Statements
- 3. ETF Implies Bayes Optimality
 - 3.1. Geometric Optimality of the Simplex ETF
 - 3.2. Bayes-Optimal Nearest Class Mean Rule
- 4. Proof of Main Theorem
 - 4.1. Closure of Projection
 - 4.2. Optimality of Projection

0. More Related Work

Machine Unlearning. The problem of removing specific training data from a model, often motivated by privacy regulations such as the “right to be forgotten,” was first formalized in the systems security community [3]. The seminal work of Bourtoule et al. [1] introduced the *SISA* framework, partitioning training data across multiple shards so that forgetting can be achieved by retraining only the affected shards. Subsequent work developed more fine-grained methods that avoid full retraining. For linear models, Guo et al. [16] proposed certified removal via influence-based updates. Sekhari et al. [30] provided theoretical guarantees for approximate unlearning in general models. For deep networks, approaches include amnesiac unlearning [15], which inverts stored gradients, and Fisher information-based scrubbing [13, 14], which perturbs weights along sensitive directions. Other efficient methods use adversarial weight perturbations [31], incompetent teachers [7], or zero-shot synthetic forget data [8]. Most recently, anchored fine-tuning methods such as FAMR [29] enforce uniform predictions on forget sets while constraining the model to remain close to its original parameters. Kodge et al. [22] proposed a gradient-free method that explicitly computes class-specific subspaces via singular value decomposition and suppresses discriminatory directions associated with the forget class. Boundary Shrink and Boundary Expand [6] perform local decision-boundary adjustments for forgetting, while maintaining model utility through margin control. DELETE [37] formulates unlearning as a decoupled distillation problem, erasing class-specific information via probability decoupling.

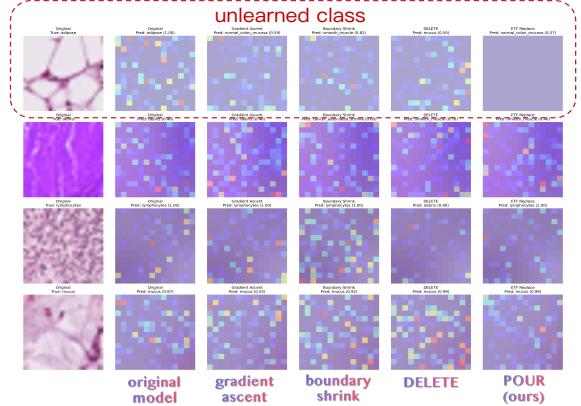


Figure 1. **Grad-CAM visualization on PathMNIST before and after unlearning.** Each row shows a tissue class. Only after POUR unlearning, the Grad-CAM signal vanishes.

Geometrically grounded forgetting. Several methods exploit the geometry of learned representations. Kodge et al. [22] proposed a gradient-free method that explicitly computes class-specific subspaces via singular value decomposition and suppresses discriminatory directions associated with the forget class. Yet, none of the previous approaches connects to the phenomenon of Neural Collapse [27], wherein class features converge to a simplex equiangular tight frame.

Concept-level and multimodal unlearning. Beyond class forgetting, recent research has explored erasing visual concepts and multimodal associations. In generative models, concept erasure can be achieved by regularizing style features or Gram matrices [36]. In multimodal settings, Yang et al. [35] proposed *CLIP-Erase*, which disentangles forgetting, retention, and consistency modules to remove specific visual-textual alignments in CLIP. Kravets and Namboodiri [24] introduced a zero-shot unlearning method for CLIP that generates synthetic forget samples via gradient ascent.

1. Additional Justifications

1.1. Proof on Decomposition of K-Bound

Let \mathcal{Z} denote the feature space and $\mathcal{P}(\mathcal{Z})$ the set of probability measures on it. Fix a symmetric function class $\mathcal{F} \subseteq \{\varphi : \mathcal{Z} \rightarrow \mathbb{R}\}$ (i.e., $\varphi \in \mathcal{F} \Rightarrow -\varphi \in \mathcal{F}$). For an Integral Probability Metric (IPM) defined as

$$\mathcal{K}(P, Q) = \sup_{\varphi \in \mathcal{F}} |\mathbb{E}_{z \sim P}[\varphi(z)] - \mathbb{E}_{z \sim Q}[\varphi(z)]|, \quad P, Q \in \mathcal{P}(\mathcal{Z}),$$

the following property holds.

Proposition 1.1 (Decomposition of \mathcal{K} Bound). *Fix a forgetting class u , and by the law of total probability, express the feature distributions as*

$$P_z = \alpha P_u + (1 - \alpha) P_{\neg u}, \quad Q_z = \beta Q_u + (1 - \beta) Q_{\neg u},$$

where $\alpha := P(y=u)$ and $\beta := Q(y=u)$ denote the class probabilities, and $P_{\neg u}, Q_{\neg u}$ are the retained-class feature distributions. Let $\Delta_c = \mathcal{K}(P_u, P_{\neg u})$ denote the prior class separation in the original model. Then the discrepancy between the unlearned and reference feature distributions is bounded as

$$\begin{aligned} & |\beta \mathcal{K}(P_u, Q_u) - (1 - \beta) \mathcal{K}(P_{\neg u}, Q_{\neg u})| - |\alpha - \beta| \Delta_c \\ & \leq \mathcal{K}(P_z, Q_z) \\ & \leq \underbrace{|\alpha - \beta| \Delta_c}_{\text{prior class separation}} + \underbrace{\beta \mathcal{K}(P_u, Q_u)}_{\text{forgotten-class alignment}} + \underbrace{(1 - \beta) \mathcal{K}(P_{\neg u}, Q_{\neg u})}_{\text{retained-class alignment}}. \end{aligned}$$

Proof. For any $\varphi \in \mathcal{F}$, substituting in the decomposition, we have

$$\mathbb{E}_{P_z}[\varphi] - \mathbb{E}_{Q_z}[\varphi] = \alpha \mathbb{E}_{P_u}[\varphi] + (1 - \alpha) \mathbb{E}_{P_{\neg u}}[\varphi] - \beta \mathbb{E}_{Q_u}[\varphi] - (1 - \beta) \mathbb{E}_{Q_{\neg u}}[\varphi] \quad (1)$$

$$= (\alpha - \beta)(\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{P_{\neg u}}[\varphi]) + \beta(\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{Q_u}[\varphi]) + (1 - \beta)(\mathbb{E}_{P_{\neg u}}[\varphi] - \mathbb{E}_{Q_{\neg u}}[\varphi]). \quad (2)$$

Taking absolute values and applying the triangle inequality yields

$$|\mathbb{E}_{P_z}[\varphi] - \mathbb{E}_{Q_z}[\varphi]| \leq |\alpha - \beta| |\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{P_{\neg u}}[\varphi]| + \beta |\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{Q_u}[\varphi]| + (1 - \beta) |\mathbb{E}_{P_{\neg u}}[\varphi] - \mathbb{E}_{Q_{\neg u}}[\varphi]|.$$

Now take the supremum over $\varphi \in \mathcal{F}$ on both sides. Since \mathcal{F} is symmetric, each term inside the absolute value corresponds exactly to the IPM definition, i.e.,

$$\begin{aligned} \sup_{\varphi \in \mathcal{F}} |\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{P_{\neg u}}[\varphi]| &= \Delta_c, & \sup_{\varphi \in \mathcal{F}} |\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{Q_u}[\varphi]| &= \mathcal{K}(P_u, Q_u), \\ \sup_{\varphi \in \mathcal{F}} |\mathbb{E}_{P_{\neg u}}[\varphi] - \mathbb{E}_{Q_{\neg u}}[\varphi]| &= \mathcal{K}(P_{\neg u}, Q_{\neg u}). \end{aligned}$$

Hence,

$$\mathcal{K}(P_z, Q_z) = \sup_{\varphi \in \mathcal{F}} |\mathbb{E}_{P_z}[\varphi] - \mathbb{E}_{Q_z}[\varphi]| \leq |\alpha - \beta| \Delta_c + \beta \mathcal{K}(P_u, Q_u) + (1 - \beta) \mathcal{K}(P_{\neg u}, Q_{\neg u}).$$

For the lower bound, apply the reverse triangle inequality to Equation 2. Let

$$x := (\alpha - \beta)(\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{P_{\neg u}}[\varphi]), \quad y := \beta(\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{Q_u}[\varphi]), \quad z := (1 - \beta)(\mathbb{E}_{P_{\neg u}}[\varphi] - \mathbb{E}_{Q_{\neg u}}[\varphi]). \quad (3)$$

Then

$$|\mathbb{E}_{P_z}[\varphi] - \mathbb{E}_{Q_z}[\varphi]| \geq |y + z| - |x|. \quad (4)$$

and by symmetry of \mathcal{F} and the definition of Δ_c ,

$$|x| \leq |\alpha - \beta| \sup_{\varphi \in \mathcal{F}} |\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{P_{\neg u}}[\varphi]| = |\alpha - \beta| \Delta_c. \quad (5)$$

Apply reverse triangle inequality again:

$$|y + z| \geq \left| \beta(\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{Q_u}[\varphi]) - (1 - \beta)(\mathbb{E}_{P_{\neg u}}[\varphi] - \mathbb{E}_{Q_{\neg u}}[\varphi]) \right|. \quad (6)$$

Taking supremum over $\varphi \in \mathcal{F}$ and using symmetry:

$$\sup_{\varphi \in \mathcal{F}} |y + z| \geq |\beta \mathcal{K}(P_u, Q_u) - (1 - \beta) \mathcal{K}(P_{\neg u}, Q_{\neg u})|. \quad (7)$$

Combining equation 4, equation 5 and equation 7, we have

$$\mathcal{K}(P_z, Q_z) = \sup_{\varphi \in \mathcal{F}} |\mathbb{E}_{P_z}[\varphi] - \mathbb{E}_{Q_z}[\varphi]| \geq |\beta \mathcal{K}(P_u, Q_u) - (1 - \beta) \mathcal{K}(P_{\neg u}, Q_{\neg u})| - |\alpha - \beta| \Delta_c.$$

This completes the proof. □

1.2. Justification on CKA Usage

We formalize the invariance properties of CKA that justify its use as a stable estimator of representation similarity in the presence of training randomness. Throughout, $X, Y \in \mathbb{R}^{n \times d}$ denote feature matrices extracted from two neural networks on the same n samples, and

$$\text{CKA}(X, Y) = \frac{\langle XX^\top, YY^\top \rangle_F}{\|XX^\top\|_F \|YY^\top\|_F}.$$

Proposition 1.2 (CKA is invariant to isotropic scaling). *For any scalar $c > 0$,*

$$\text{CKA}(X, cX) = 1.$$

Proof. We compute

$$\text{CKA}(X, cX) = \frac{\langle XX^\top, c^2 XX^\top \rangle_F}{\|XX^\top\|_F \|c^2 XX^\top\|_F} = \frac{c^2 \|XX^\top\|_F^2}{|c^2| \|XX^\top\|_F^2} = 1.$$

Thus isotropic rescaling of all features leaves CKA unchanged. \square

This property ensures that CKA is stable under global norm changes arising from SGD noise, learning-rate schedules, BatchNorm scaling, or unlearning updates that shrink or expand feature magnitudes uniformly.

Proposition 1.3 (CKA is invariant to orthogonal basis rotations). *Let $R \in \mathbb{R}^{d \times d}$ be any orthogonal matrix ($R^\top R = I$). Then*

$$\text{CKA}(X, XR) = 1.$$

Proof. If $Y = XR$, then

$$YY^\top = XRR^\top X^\top = XX^\top.$$

Thus the numerator and denominator of CKA coincide:

$$\text{CKA}(X, XR) = \frac{\langle XX^\top, XX^\top \rangle_F}{\|XX^\top\|_F \|XX^\top\|_F} = 1.$$

\square

Orthogonal invariance is critical because independently trained networks often learn equivalent representations that differ only by a rotation of the feature basis, especially when trained with different seeds.

Lemma 1.4 (CKA is stable under mild anisotropic distortions). *Let $D = \text{diag}(d_1, \dots, d_d)$ with $d_i > 0$. If $\max_i d_i / \min_i d_i \leq 1 + \varepsilon$, then*

$$|\text{CKA}(X, XD) - 1| = O(\varepsilon).$$

Proof. We observe

$$XD(XD)^\top = XD^2 X^\top.$$

Since $D^2 = I + E$ with $\|E\|_2 = O(\varepsilon)$, it follows that

$$XD(XD)^\top = XX^\top + XEX^\top.$$

The Frobenius norms in the CKA numerator and denominator can be expanded via perturbation bounds:

$$\|XX^\top + XEX^\top\|_F = \|XX^\top\|_F (1 + O(\varepsilon)),$$

and the inner product perturbation satisfies

$$\langle XX^\top, XX^\top + XEX^\top \rangle_F = \|XX^\top\|_F^2 (1 + O(\varepsilon)).$$

Substituting into the CKA ratio yields the claimed bound. \square

This shows that CKA is robust even to moderate channel-wise stretching commonly introduced by BatchNorm, layer scaling, or local unlearning updates.

Proposition 1.5 (CKA depends only on pairwise sample geometry). *If two feature matrices X and Y satisfy*

$$XX^\top = YY^\top,$$

then

$$\text{CKA}(X, Y) = 1.$$

Proof. Direct substitution into the definition of CKA yields

$$\text{CKA}(X, Y) = \frac{\langle XX^\top, XX^\top \rangle_F}{\|XX^\top\|_F \|XX^\top\|_F} = 1.$$

\square

Because XX^\top encodes pairwise sample similarities, which are far more stable across random seeds than the raw coordinates of X , this proposition explains CKA's reliability as a measure of representation equivalence.

Conclusion

Together, Propositions 1.2–1.5 establish that CKA is invariant to the dominant sources of randomness in neural representation learning, including global rescaling, orthogonal transformations, channel permutations, and mild anisotropic distortions. Since retraining on the retain set produces models that differ primarily through such randomness, CKA provides a stable and reliable estimator of representation similarity for evaluating representation-level weak unlearning.

2. Neural Collapse

2.1. Training and modeling assumptions.

Below are the standard Neural Collapse (NC) assumptions:

- **(A1) Interpolation / TPT:** The network is trained to near-zero training error and then further optimized in the terminal phase of training (TPT) under standard protocols such as SGD or Adam with decays [27].
- **(A2) Overparameterization:** The model has sufficient capacity to realize class-wise linear separability in the penultimate features, often corresponding to large width or deep linear heads [20].
- **(A3) Loss and regularization:** Cross-entropy loss with weight decay (or L_2 regularization) is used. In simplified unconstrained-feature or layer-peeled models, global minimizers are simplex ETFs and all other critical points are strict saddles [11, 26, 38]. Empirically and theoretically, MSE loss also exhibits NC [17].
- **(A4) Balanced classes:** Unless otherwise stated, class priors are assumed to be balanced. With class imbalance, NC persists in modified forms such as non-equiangular means or multiple centers [10, 11, 19, 33].
- **(A5) Feature dimension:** The penultimate feature dimension satisfies $d \geq C - 1$, which ensures the existence of a simplex embedding [26].

2.2. Neural Collapse Statements

Under assumptions (A1)–(A5), the following NC properties can be observed [27]:

- **(NC1) Within-class collapse:** For each class i , the learned feature representation takes the form $z_\theta(x) = \alpha(x)v_i$, where $z_\theta(x)$ denotes the feature extractor θ applied to input x , $\alpha(x) > 0$ is a class-dependent scaling factor, and $v_i \in \mathbb{R}^d$ is a unit direction.
- **(NC2) Simplex ETF means:** The set of class directions $\{v_i\}_{i=1}^C$ lies in a $(C-1)$ -dimensional subspace and forms a simplex Equiangular Tight Frame (ETF). Specifically, they satisfy $\|v_i\| = 1$ for all i , $v_i^\top v_j = -\frac{1}{C-1}$ for $i \neq j$, and $\sum_{i=1}^C v_i = 0$.
- **(NC3) Classifier alignment:** The final-layer classifier weights (w) are aligned with the class directions. Specifically, there exists a constant $\kappa > 0$ such that $w_i = \kappa v_i$ for every class i .
- **(NC4) Nearest-class-mean rule:** Classification reduces to a nearest-class-mean decision rule, equivalently assigning each sample to the nearest ETF vertex.

These properties jointly imply that, for balanced classes, the geometry of class representations forms a centered regular simplex in \mathbb{R}^{C-1} , which is maximally separated and symmetric in the space.

3. ETF Implies Bayes Optimality

We present a formal statement and proof of Proposition 3.1. First, we show that the simplex Equiangular Tight Frame (ETF) configuration is geometrically optimal: it maximizes the minimum pairwise angle among class means and therefore maximizes the multiclass angular margin of the Nearest Class Mean (NCM) classifier. Second, under homoscedastic Gaussian class-conditionals, we show that the NCM rule coincides exactly with the Bayes-optimal classifier.

3.1. Geometric Optimality of the Simplex ETF

Setup. Let $\{v_c\}_{c=1}^C$ be unit vectors in \mathbb{R}^d (with $d \geq C - 1$) representing class means of an NCM classifier. Define the minimum pairwise inner product

$$\gamma := \min_{c \neq c'} v_c^\top v_{c'}.$$

Equivalently, maximizing the minimum pairwise angle $\min_{c \neq c'} \angle(v_c, v_{c'})$ is equivalent to minimizing γ .

Proposition 3.1 (Geometric optimality of the simplex ETF). *Among all sets of C unit vectors in \mathbb{R}^d , $d \geq C - 1$, the centered simplex ETF uniquely maximizes the minimum pairwise angle:*

(i) (Maximal angle) *The Welch/simplex bound implies*

$$\gamma \leq -\frac{1}{C-1}.$$

Equality holds if and only if

$$v_c^\top v_{c'} = \begin{cases} 1, & c = c', \\ -\frac{1}{C-1}, & c \neq c', \end{cases} \quad \sum_{c=1}^C v_c = 0,$$

i.e. $\{v_c\}$ forms a centered simplex ETF. The maximizer is unique up to rotation/reflection.

(ii) (Maximal angular NCM margin) *For unit-norm vectors, the worst-case angular margin of the NCM classifier is a monotone function of $\min_{c \neq c'} \angle(v_c, v_{c'})$. Because the simplex ETF maximizes this angle by (i), it also maximizes the multiclass angular margin of the NCM classifier.*

Proof. (i) The Welch bound states that any C unit vectors in \mathbb{R}^d satisfy $\min_{c \neq c'} v_c^\top v_{c'} \leq -1/(C-1)$. Equality requires that the Gram matrix has the simplex ETF structure given above and is unique up to orthogonal transformation.

(ii) For unit vectors, the NCM decision boundary between classes c and c' is the hyperplane $\langle x, v_c - v_{c'} \rangle = 0$, whose angular separation is controlled solely by the angle $\angle(v_c, v_{c'})$. The worst-case multiclass angular margin is therefore a monotone function of the minimum such angle, and the simplex ETF maximizes it by (i). \square

3.2. Bayes-Optimal Nearest Class Mean Rule

We now consider the probabilistic setting underlying NC analyses. There are C classes with equal prior $\Pr(y = c) = 1/C$. Conditioned on class c , features follow a homoscedastic Gaussian distribution:

$$x \mid y = c \sim \mathcal{N}(\mu_c, \Sigma), \quad \Sigma \succ 0.$$

We assume the class means form a centered simplex ETF in the Mahalanobis geometry:

$$\sum_{c=1}^C \mu_c = 0, \quad \|\mu_c\|_{\Sigma^{-1}} = \|\mu_{c'}\|_{\Sigma^{-1}} \quad \forall c, c'.$$

Proposition 3.2 (ETF geometry implies Bayes-optimal NCM classification). *Under the model above, the Bayes-optimal classifier is*

$$h^*(x) = \arg \max_c \mu_c^\top \Sigma^{-1} x,$$

which is a zero-bias linear classifier with weights $w_c = \Sigma^{-1} \mu_c$. Moreover:

(i) If $\Sigma = \sigma^2 I$, then h^* reduces to the Euclidean NCM rule,

$$h^*(x) = \arg \min_c \|x - \mu_c\|^2.$$

(ii) If x and μ_c are normalized, this is equivalent to cosine-similarity classification: $h^*(x) = \arg \max_c \langle x, \mu_c \rangle$.

(iii) In the NC/TPT limit, classifier weights satisfy $w_c \parallel \mu_c$ and $\|w_c\| \rightarrow \infty$, so the induced linear classifier matches h^* exactly.

Proof. With equal priors,

$$h^*(x) = \arg \max_c p(x \mid y = c) = \arg \min_c \|x - \mu_c\|_{\Sigma^{-1}}^2,$$

since $p(x \mid y = c) \propto \exp(-\frac{1}{2} \|x - \mu_c\|_{\Sigma^{-1}}^2)$.

Expanding the Mahalanobis distance,

$$\|x - \mu_c\|_{\Sigma^{-1}}^2 = \|x\|_{\Sigma^{-1}}^2 - 2\mu_c^\top \Sigma^{-1} x + \|\mu_c\|_{\Sigma^{-1}}^2.$$

The first term is independent of c , and under the ETF assumption, the third term is also constant across classes. Therefore,

$$h^*(x) = \arg \max_c \mu_c^\top \Sigma^{-1} x. \quad (\star)$$

Define $w_c = \Sigma^{-1} \mu_c$. Because the class means are centered, $\sum_c \mu_c = 0$, it follows that $\sum_c w_c = 0$, so the discriminant scores $\{w_c^\top x\}$ have zero mean across classes. Hence, (\star) is a zero-bias linear decision rule.

When $\Sigma = \sigma^2 I$, the rule in (\star) reduces to minimizing the Euclidean distance $\|x - \mu_c\|^2$, corresponding to the classical NCM classifier. If all vectors are further normalized, this becomes equivalent to cosine-similarity classification.

In the NC/TPT regime, the classifier weights satisfy $w_c \parallel \mu_c$ and $\|w_c\| \rightarrow \infty$, so the induced linear classifier $\arg \max_c \langle w_c, x \rangle$ coincides with the cosine classifier above, and therefore matches the Bayes rule in (\star) .

Thus, the simplex ETF configuration of class means yields the Bayes-optimal NCM classifier. \square

4. Proof of Main Theorem

4.1. Closure of Projection

Note that a *simplex ETF* $\{v_i\}_{i=1}^C \subset \mathbb{R}^{C-1}$ satisfies

$$\|v_i\| = 1, \quad v_i^\top v_j = -\frac{1}{C-1} \quad (i \neq j), \quad \sum_{i=1}^C v_i = 0.$$

Equivalently, its Gram matrix has 1 on the diagonal and constant off-diagonal entries $-1/(C-1)$.

Theorem 4.1 (Projection of a Simplex ETF). *Let $\{v_i\}_{i=1}^C \subset \mathbb{R}^{C-1}$ be a simplex ETF. Fix v_1 and let $P = I - v_1 v_1^\top$ be the orthogonal projector onto v_1^\perp . For $i = 2, \dots, C$, define $u_i = Pv_i$ and $w_i = u_i/\|u_i\|$. Then $\{w_i\}_{i=2}^C \subset v_1^\perp \cong \mathbb{R}^{C-2}$ is again a simplex ETF:*

$$\|w_i\| = 1, \quad w_i^\top w_j = -\frac{1}{C-2} \quad (i \neq j), \quad \sum_{i=2}^C w_i = 0.$$

Proof. Write $\beta := -\frac{1}{C-1}$. For $i \geq 2$,

$$u_i = Pv_i = v_i - (v_1^\top v_i)v_1 = v_i - \beta v_1.$$

Equal norms. Using $\|v_i\| = \|v_1\| = 1$ and $v_i^\top v_1 = \beta$,

$$\|u_i\|^2 = \|v_i\|^2 - 2\beta v_i^\top v_1 + \beta^2 \|v_1\|^2 = 1 - 2\beta^2 + \beta^2 = 1 - \beta^2 = 1 - \frac{1}{(C-1)^2} = \frac{C(C-2)}{(C-1)^2}.$$

Thus all $\|u_i\|$ are equal.

Equal pairwise inner products. For $i \neq j$ with $i, j \geq 2$,

$$u_i^\top u_j = v_i^\top v_j - \beta v_i^\top v_1 - \beta v_1^\top v_j + \beta^2 = \beta - \beta^2 - \beta^2 + \beta^2 = \beta - \beta^2 = -\frac{C}{(C-1)^2}.$$

Hence, after normalization,

$$\frac{u_i^\top u_j}{\|u_i\| \|u_j\|} = \frac{-C/(C-1)^2}{C(C-2)/(C-1)^2} = -\frac{1}{C-2},$$

so $w_i^\top w_j = -\frac{1}{C-2}$.

Zero sum. Since $\sum_{i=1}^C v_i = 0$,

$$\sum_{i=2}^C u_i = \sum_{i=2}^C (v_i - \beta v_1) = \left(\sum_{i=2}^C v_i \right) - (C-1)\beta v_1 = (-v_1) - (C-1)\left(-\frac{1}{C-1}\right)v_1 = 0.$$

All $\|u_i\|$ are equal, so common normalization preserves the zero sum: $\sum_{i=2}^C w_i = 0$. The vectors $\{w_i\}$ lie in v_1^\perp (dimension $C-2$), have unit norm, constant off-diagonal inner product $-1/(C-2)$, and zero mean; hence they form a simplex ETF. \square

Remark 4.2. *This result is specific to the simplex ETF (the NC configuration). It does not generally hold for arbitrary ETFs.*

4.2. Optimality of Projection

We now establish the optimality of our projection operator under the definition of representation-level weak unlearning (Def. 2.1). The central idea is that projecting onto the orthogonal complement of the forgotten class removes its contribution while preserving the Bayes-optimal ETF geometry of the retained classes.

Theorem 4.3 (ETF projection preserves optimality and forgets the target class). *Assume (A1)–(A5) and Neural Collapse (NC1)–(NC4) hold pre-unlearning, and suppose the following statistical model for the penultimate features:*

1. (Balanced classes) *class priors are uniform: $\Pr(y = i) = 1/C$ for $i \in \mathcal{Y}$.*
2. (Isotropic Gaussian conditionals) *conditional on class i ,*

$$\theta(x) \mid (y = i) \sim \mathcal{N}(\mu_i, \sigma^2 I_d),$$

with $\|\mu_i\| = 1$ and $\{\mu_i\}_{i=1}^C$ coinciding with the ETF directions $\{v_i\}$ from NC (i.e. $\mu_i = v_i$).

Fix a class $u \in \mathcal{Y}$ and define

$$P = I - v_u v_u^\top, \quad \tilde{v}_i = \frac{P v_i}{\|P v_i\|} \quad (i \neq u),$$

so that by Proposition 3.2 the vectors $\{\tilde{v}_i\}_{i \neq u}$ form a simplex ETF in the subspace v_u^\perp . Let the projected features be $\theta'(x) = P \theta(x)$ and let the post-projection classifier weights satisfy $w'_i = \kappa' \tilde{v}_i$ for $i \neq u$. Then:

- (a) (Retained-class Bayes optimality) *For the retained classes $\mathcal{Y}_{\neg u}$, the classifier that assigns x to the nearest projected class mean \tilde{v}_i is Bayes-optimal under the Gaussian model above. Equivalently, the projected model $(\theta', \{w'_i\}_{i \neq u})$ attains the Bayes decision rule on $\mathcal{Y}_{\neg u}$.*
- (b) (Complete forgetting in the low-noise / NC limit) *Under projection, the forget-class conditional mean is mapped to zero: $P \mu_u = 0$. Consequently, for $x \sim \mathcal{D}_f$,*

$$\theta'(x) \mid (y = u) \sim \mathcal{N}(0, \sigma^2 P).$$

In the limit $\sigma^2 \rightarrow 0$ (equivalently, in the NC/TPT limit where within-class variance vanishes, or as the classifier scale $\kappa' \rightarrow \infty$ appropriately), the projected features for the forget class concentrate at the origin, yielding logits $w'_i^\top \theta'(x) \rightarrow 0$ for all $i \neq u$. Hence the predictive distribution over retained classes approaches the uniform distribution $U_{\neg u}$, i.e. the forget class is completely forgotten in the sense that the model expresses no informative preference among retained classes.

Consequently, ETF projection simultaneously (i) preserves Bayes-optimal classification on the retained classes and (ii) erases class-specific information for the forgotten class (in the stated asymptotic / low-noise sense).

We first provide a proof sketch. The formal proof is included on the next page.

Proof sketch. For part (a), under the Gaussian ETF model with means $\{\mu_i = v_i\}$, Proposition 3.2 shows that the nearest-class-mean rule is Bayes-optimal. By Proposition 3.2, the projected means $\{\tilde{v}_i\}_{i \neq u}$ form a simplex ETF in v_u^\perp , so the same argument implies that the nearest-mean classifier on $\{\tilde{v}_i\}$ is Bayes-optimal for the retained classes $\mathcal{Y}_{\neg u}$.

For part (b), note that $P v_u = 0$ implies that the projected forget-class distribution satisfies $\theta'(x) \mid (y = u) \sim \mathcal{N}(0, \sigma^2 P)$. For any retained class $i \neq u$,

$$\mathbb{E}[w'_i^\top \theta'(x) \mid y = u] = w'_i^\top P \mu_u = 0,$$

and as $\sigma^2 \rightarrow 0$ the distribution of $\theta'(x)$ for the forgotten class concentrates at the origin. Thus the logits $w'_i^\top \theta'(x)$ converge to 0 for all $i \neq u$, and the induced softmax over retained classes converges to the uniform distribution $U_{\neg u}$. This formalizes the notion that the projected model has no discriminative information about the forgotten class in the low-noise / NC limit. \square

Formal Proof. **(a) Retained-class Bayes optimality.** Under the assumptions of the theorem, pre-unlearning we have

$$\theta(x) \mid (y = i) \sim \mathcal{N}(v_i, \sigma^2 I_d), \quad \Pr(y = i) = 1/C,$$

with the means $\{v_i\}$ forming a centered simplex ETF in \mathbb{R}^d . By Proposition 3.2, the Bayes-optimal classifier for this model is the nearest-class-mean rule (equivalently, a scaled linear classifier aligned with $\{v_i\}$).

Fix a forgotten class u and apply the projection $P = I - v_u v_u^\top$. For any retained class $i \neq u$,

$$\theta'(x) \mid (y = i) = P\theta(x) \mid (y = i) \sim \mathcal{N}(Pv_i, \sigma^2 P),$$

since P is a linear operator and $\theta(x)$ is Gaussian with mean v_i and covariance $\sigma^2 I_d$. Thus, conditioned on $y \in \mathcal{Y}_{\neg u}$, the projected features follow a homoscedastic Gaussian model in the subspace v_u^\perp with:

$$\text{means } \mu'_i = Pv_i, \quad \text{common covariance } \Sigma' = \sigma^2 P.$$

By Proposition 3.2, the normalized means $\tilde{v}_i = \mu'_i / \|\mu'_i\|$ form a centered simplex ETF in v_u^\perp . Since P acts as the identity on v_u^\perp and is zero on $\text{span}(v_u)$, Σ' is proportional to the identity on v_u^\perp (and vanishes on v_u), so within v_u^\perp the conditionals are isotropic Gaussians with means \tilde{v}_i up to a global scale.

Applying Proposition 3.2 to this reduced $(C-1)$ -class ETF in v_u^\perp , we obtain that the Bayes-optimal classifier among the retained classes is the nearest-class-mean rule with respect to the means $\{\tilde{v}_i\}_{i \neq u}$ (equivalently, a scaled linear classifier with weights $w'_i = \kappa' \tilde{v}_i$). This is precisely the classifier implemented by the projected model $(\theta', \{w'_i\}_{i \neq u})$, establishing Bayes optimality on $\mathcal{Y}_{\neg u}$.

(b) Complete forgetting in the low-noise / NC limit. For the forgotten class u , we have $\mu_u = v_u$ and

$$\theta(x) \mid (y = u) \sim \mathcal{N}(v_u, \sigma^2 I_d).$$

Applying P and using $Pv_u = 0$, we obtain

$$\theta'(x) \mid (y = u) = P\theta(x) \mid (y = u) \sim \mathcal{N}(Pv_u, \sigma^2 P) = \mathcal{N}(0, \sigma^2 P).$$

Thus the projected features for class u are mean-zero Gaussian supported in v_u^\perp with covariance $\sigma^2 P$. For any retained class $i \neq u$, the corresponding logit is

$$s_i(x) := w_i'^\top \theta'(x) = \kappa' \tilde{v}_i^\top \theta'(x),$$

where $\tilde{v}_i \in v_u^\perp$ and $w_i' \in v_u^\perp$ because they are constructed from Pv_i . Since $\theta'(x) \mid (y = u)$ is mean-zero,

$$\mathbb{E}[s_i(x) \mid y = u] = w_i'^\top \mathbb{E}[\theta'(x) \mid y = u] = w_i'^\top 0 = 0.$$

Moreover, as $\sigma^2 \rightarrow 0$, the Gaussian $\mathcal{N}(0, \sigma^2 P)$ converges in probability (and almost surely for any fixed sample) to the point mass at 0. Therefore

$$\theta'(x) \mid (y = u) \xrightarrow[\sigma^2 \rightarrow 0]{} 0 \quad \text{in probability},$$

and hence

$$s_i(x) = w_i'^\top \theta'(x) \xrightarrow[\sigma^2 \rightarrow 0]{} 0 \quad \text{in probability, for all } i \neq u.$$

The predictive distribution over retained classes is

$$q_{\neg u}(i \mid x) = \frac{\exp(s_i(x))}{\sum_{j \neq u} \exp(s_j(x))}.$$

For any fixed vector $s \in \mathbb{R}^m$ (with $m = C-1$), if $s \rightarrow 0$ then $\text{softmax}(s) \rightarrow U_{\neg u}$, the uniform distribution on m classes. By continuity of the softmax map and convergence of $\mathbf{s}(x) = [s_i(x)]_{i \neq u}$ to the zero vector, we obtain

$$q_{\neg u}(\cdot \mid x) = \text{softmax}(\mathbf{s}(x)) \xrightarrow[\sigma^2 \rightarrow 0]{} U_{\neg u} \quad \text{in probability under } x \sim \mathcal{D}_f.$$

Thus, in the low-noise / NC limit, the projected model makes asymptotically uniform predictions over retained classes for any sample from the forgotten class, which formalizes the notion that it has no informative class preference for $y = u$. \square