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

Supplementary Material

Appendix Table of Contents

0. 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. [17] proposed certified removal via influence-based updates. Sekhari et al. [35] provided theoretical guarantees for approximate unlearning in general models. For deep networks, approaches include amnesiac unlearning [16], which inverts stored gradients, and Fisher information–based scrubbing [14, 15], which perturbs weights along sensitive directions. Other efficient methods use adversarial weight perturbations [37], incompetent teachers [7], or zero-shot synthetic forget data [8]. Most recently, anchored fine-tuning methods such as FAMR [34] enforce uniform predictions on forget sets while constraining the model to remain close to its original parameters. Kodge et al. [23] 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 [43] formulates unlearning as a decoupled distillation problem, erasing class-specific information via probability decoupling.

Table 6. **More baselines** on CIFAR-10.

Method	$\text{Acc}_r \uparrow$	$\text{Acc}_f \downarrow$	AUS↑	$\text{CKA}_f \downarrow$	$\text{CKA}_r \uparrow$	$\text{RUS}^{(o)} \uparrow$
Original	94.61	94.73	0.51	1.00	1.00	0.00
ProjUn [22]	86.49	20.40	0.76	0.34	0.89	0.76
SalUn	92.72	<u>0.78</u>	**0.97**	<u>0.28</u>	0.92	<u>0.81</u>
SCRUB	**93.34**	3.80	<u>0.95</u>	0.31	**0.98**	<u>0.81</u>
SURE	92.50	2.60	<u>0.95</u>	0.34	**0.98**	0.79
POUR	<u>92.86</u>	**0.37**	**0.97**	**0.23**	<u>0.95</u>	**0.85**

Geometrically grounded forgetting. Several methods exploit the geometry of learned representations. Kodge et al. [23] 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 [31], 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 [42]. In multimodal settings, Yang et al. [41] proposed *CLIPErase*, which disentangles forgetting, retention, and consistency modules to remove specific visual-textual alignments in CLIP. Kravets and Namboodiri [25] introduced a zero-shot unlearning method for CLIP that generates synthetic forget samples via gradient ascent.

1. Additional Baselines

We additionally report results (forget airplane class) for method [22], SalUn [11], SCRUB [26] and SURE [36] in Tab. 6. POUR again achieves strongest performance across both classification- and representation-level metrics.

2. Additional Justifications

2.1. 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 $\mathcal{F} \subseteq \{\varphi : \mathcal{Z} \to \mathbb{R}\}$ (i.e., $\varphi \in \mathcal{F} \Rightarrow -\varphi \in \mathcal{F}$). For an Integral Probability Metric (IPM) defined as

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

the following property holds.

Proposition 2.1 (Decomposition of $\mathcal{K}$ Bound)**.** *Fix a forgetting class u, and by the law of total probability, express the feature distributions as*

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

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*

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

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

$$\begin{aligned}
\mathbb{E}_{P_z}[\varphi] - \mathbb{E}_{Q_z}[\varphi] &= \alpha\, \mathbb{E}_{P_u}[\varphi] + (1-\alpha)\, \mathbb{E}_{P_{\neg u}}[\varphi] - \beta\, \mathbb{E}_{Q_u}[\varphi] - (1-\beta)\, \mathbb{E}_{Q_{\neg u}}[\varphi] && (1)\\
&= (\alpha-\beta)\big(\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{P_{\neg u}}[\varphi]\big) + \beta\big(\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{Q_u}[\varphi]\big) + (1-\beta)\big(\mathbb{E}_{P_{\neg u}}[\varphi] - \mathbb{E}_{Q_{\neg u}}[\varphi]\big). && (2)
\end{aligned}$$

Taking absolute values and applying the triangle inequality yields

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

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.,

$$\sup_{\varphi\in\mathcal{F}} \big|\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{P_{\neg u}}[\varphi]\big| = \Delta_c, \qquad \sup_{\varphi\in\mathcal{F}} \big|\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{Q_u}[\varphi]\big| = \mathcal{K}(P_u, Q_u),$$

$$\sup_{\varphi\in\mathcal{F}} \big|\mathbb{E}_{P_{\neg u}}[\varphi] - \mathbb{E}_{Q_{\neg u}}[\varphi]\big| = \mathcal{K}(P_{\neg u}, Q_{\neg u}).$$

Hence,

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

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

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

Then

$$\big|\mathbb{E}_{P_z}[\varphi] - \mathbb{E}_{Q_z}[\varphi]\big| \ \ge\ \big|y+z\big| \ -\ |x|. \tag{4}$$

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

$$|x| \le |\alpha-\beta| \sup_{\varphi\in\mathcal{F}} \big|\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{P_{\neg u}}[\varphi]\big| = |\alpha-\beta|\Delta_c. \tag{5}$$

Apply reverse triangle inequality again:

$$|y+z| \ \ge\ \Big|\, \beta(\mathbb{E}_{P_u}[\varphi] - \mathbb{E}_{Q_u}[\varphi]) - (1-\beta)(\mathbb{E}_{P_{\neg u}}[\varphi] - \mathbb{E}_{Q_{\neg u}}[\varphi])\,\Big|. \tag{6}$$

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

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

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

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

This completes the proof.

□

2.2. 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

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

Proposition 2.2 (CKA is invariant to isotropic scaling)**.** *For any scalar $c > 0$,*

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

Proof. We compute

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

Thus isotropic rescaling of all features leaves CKA unchanged. □

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

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

Proof. If $Y = XR$, then

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

Thus the numerator and denominator of CKA coincide:

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

□

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 2.4 (CKA is stable under mild anisotropic distortions)**.** *Let $D = \mathrm{diag}(d_1, \ldots, d_d)$ with $d_i > 0$. If $\max_i d_i / \min_i d_i \le 1 + \varepsilon$, then*

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

Proof. We observe

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

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. □

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

Proposition 2.5 (CKA depends only on pairwise sample geometry)**.** *If two feature matrices X and Y satisfy*

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

then

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

Proof. Direct substitution into the definition of CKA yields

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

□

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 2.2–2.5 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.

3. Neural Collapse

3.1. 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 [31].
- **(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 [21].
- **(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 [12, 28, 44]. Empirically and theoretically, MSE loss also exhibits NC [18].
- **(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, 12, 20, 39].
- **(A5) Feature dimension:** The penultimate feature dimension satisfies $d \geq C-1$, which ensures the existence of a simplex embedding [28].

3.2. Neural Collapse Statements

Under assumptions (A1)–(A5), the following NC properties can be observed [31]:

- (**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.

4. 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.

4.1. 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 4.1 (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

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

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). □

4.2. 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:

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

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

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

Proposition 4.2 (ETF geometry implies Bayes-optimal NCM classification)**.** *Under the model above, the Bayes-optimal classifier is*

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

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

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

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

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

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

Proof. With equal priors,

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

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

Expanding the Mahalanobis distance,

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

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

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

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 $\|w_c\| \to \infty$, 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. □

5. Proof of Main Theorem

5.1. Closure of Projection

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

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

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

Theorem 5.1 (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 $i = 2, \ldots, C$, 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:*

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

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 $\sum_{i=1}^C v_i = 0$,

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

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. □

Remark 5.2. *This result is specific to the* simplex *ETF (the NC configuration). It does not generally hold for arbitrary ETFs.*

5.2. 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 5.3 (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*

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

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 $\sigma^2 \to 0$ *(equivalently, in the NC/TPT limit where within-class variance vanishes, or as the classifier scale* $\kappa' \to \infty$ *appropriately), the projected features for the forget class concentrate at the origin, yielding logits* $w_i'^\top \theta'(x) \to 0$ *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 4.2 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 $\sigma^2 \to 0$ 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. □

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

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

with the means $\{v_i\}$ forming a centered simplex ETF in $\mathbb{R}^d$. By Proposition 4.2, the Bayes-optimal classifier for this model is the nearest-class-mean rule (equivalently, a scaled linear classifier aligned with $\{v_i\}$).
Fix a forget 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:

$$\text{means } \mu_i' = Pv_i, \qquad \text{common covariance } \Sigma' = \sigma^2 P.$$

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 $\mathrm{span}(v_u)$, 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 4.2 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 $\sigma^2 \to 0$, 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

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

and hence

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

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 $s \to 0$ then $\mathrm{softmax}(s) \to U_{\neg u}$, 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

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

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 $y = u$. □