ON THE ALIGNMENT BETWEEN SUPERVISED AND SELF-SUPERVISED CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised contrastive learning (CL) has achieved remarkable empirical success, often producing representations that rival supervised pre-training on downstream tasks. Recent theory explains this by showing that the CL loss closely approximates a supervised surrogate, Negatives-Only Supervised Contrastive Learning (NSCL) loss, as the number of classes grows. Yet this loss-level similarity leaves an open question: *Do CL and NSCL also remain aligned at the representation level throughout training, not just in their objectives?*

We address this by analyzing the representation alignment of CL and NSCL models trained under shared randomness (same initialization, batches, and augmentations). First, we show that their induced representations remain similar: specifically, we prove that the similarity matrices of CL and NSCL stay close under realistic conditions. Our bounds provide high-probability guarantees on alignment metrics such as centered kernel alignment (CKA) and representational similarity analysis (RSA), and they clarify how alignment improves with more classes, higher temperatures, and its dependence on batch size. In contrast, we demonstrate that parameter-space coupling is inherently unstable: divergence between CL and NSCL weights can grow exponentially with training time.

Finally, we validate these predictions empirically, showing that CL–NSCL alignment strengthens with scale and temperature, and that NSCL tracks CL more closely than other supervised objectives. This positions NSCL as a principled bridge between self-supervised and supervised learning.

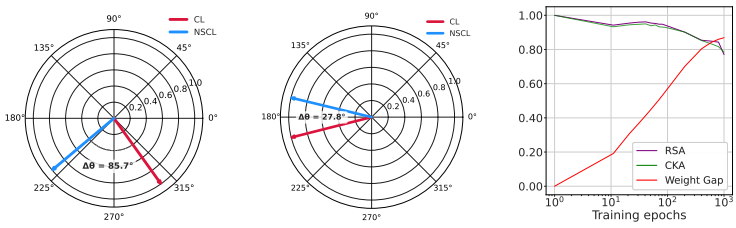
1 INTRODUCTION

Self-supervised learning (SSL) has become the dominant approach for extracting transferable representations from large-scale unlabeled data. By leveraging training signals derived directly from the data, SSL methods avoid costly annotation while producing features that generalize across modalities, from vision (Chen et al., 2020; He et al., 2020; Zbontar et al., 2021; He et al., 2022; Oquab et al., 2024) to language (Gao et al., 2021; Reimers & Gurevych, 2019), speech (Schneider et al., 2019; Baevski et al., 2020; Hsu et al., 2021; Baevski et al., 2022), and vision–language (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023; Tschannen et al., 2025). Among SSL approaches, *contrastive learning (CL)* has been particularly successful: methods such as SimCLR (Chen et al., 2020), MoCo (He et al., 2020; Chen et al., 2021b), and CPC (van den Oord et al., 2019) train encoders by pulling together augmented views of the same input while pushing apart other samples. This simple principle has yielded state-of-the-art performance, often rivaling or surpassing supervised pre-training.

Despite this empirical success, a central puzzle remains: why does CL recover features so well aligned with semantic class boundaries? CL models often support nearly supervised-level downstream performance (Amir et al., 2022; Ben-Shaul et al., 2023; Weng et al., 2025), suggesting that supervision is somehow implicit in the objective. Recent theoretical progress sheds light on this: Luthra et al. (2025) showed that the CL objective closely approximates a supervised variant, *Negatives-Only Supervised Contrastive Learning (NSCL)*, where same-class samples are excluded from the denominator. Their analysis established that the CL–NSCL *losses* converge as the number of classes grows, and further characterized the geometry of NSCL minimizers and their linear probe performance. These results indicate that CL carries a supervised-like signal at the *loss level*.

Yet this view leaves a crucial question unresolved:

Do contrastive and supervised contrastive models remain aligned throughout training, not just at the level of their objectives?



(a) Weight Space (b) Representation Space (c) Similarity Metrics

Figure 1: **Comparison of learning dynamics for CL and NSCL models.** (a) Weight space vectors show divergent paths (85.7° apart). (b) In contrast, representation space vectors for a target class show high alignment (27.8° apart). (c) This is confirmed over training epochs, where representational similarity (CKA, RSA) remains high while the weight gap increases (see figure details in App. B).

Loss-level similarity does not guarantee that optimization paths coincide. In principle, differences in curvature, gradient noise, or learning rate schedules could amplify small loss discrepancies, causing stochastic gradient descent (SGD) trajectories to diverge. Thus, it remains unclear whether CL merely converges to a solution *similar* to NSCL, or whether their parameter and representations remain coupled across training. While some preliminary empirical results (Grigg et al., 2021) provide evidence that supervised and self-supervised models learn fairly well-aligned representations geometrically, it is not clear to what extent this alignment holds, under what conditions it arises, and what factors control the alignment between the two regimes.

Contributions. In this work, we theoretically and systematically study the alignment between CL and NSCL under shared randomness (same initialization, mini-batches, and augmentations):

- **From drift to metrics.** The similarity control yields explicit, high-probability *lower bounds* on linear CKA and RSA at every epoch, showing that CL and NSCL representations remain nontrivially aligned and that the certified alignment tightens as C and B grow and as τ increases (Cors. 1–2). For completeness, we also bound parameter drift under β-smoothness (Thm. 2), which can grow exponentially even when representations remain aligned.
- **Conceptual contribution.** Our results provide a conceptual framework for what CL optimizes during training. We (i) identify NSCL as the supervised objective whose representations and training trajectories are most tightly coupled to those of CL—without claiming that NSCL is the strongest supervised baseline in terms of top-1 accuracy—and (ii) shift the focus from guarantees on downstream classification accuracy to *geometric* alignment between supervised and self-supervised representations. Whereas prior work shows that minimizing self-supervised losses can yield good downstream classifiers under generative assumptions (e.g., (Arora et al., 2019; Tosh et al., 2021; Saunshi et al., 2022; Awasthi et al., 2022; HaoChen & Ma, 2023)), our analysis instead characterizes when CL and NSCL induce similar similarity structures, a perspective that is particularly relevant for tasks that depend on representation geometry, such as interpretability and image segmentation.
- **Empirical validation.** We validate our theory with experiments on CIFAR-10/100, Tiny-ImageNet, mini-ImageNet, and ImageNet-1K. We find that (i) CL–NSCL alignment strengthens with more classes and higher temperatures as well as correlates with the bound's dependence on the batch size; and (ii) NSCL aligns with CL more strongly than other supervised learning methods (such as cross-entropy minimization and supervised contrastive learning (SCL) (Khosla et al., 2020)).

2 Related Work

A large body of work has sought to explain the success of contrastive learning (CL) from different perspectives. Early accounts linked CL to mutual information maximization between views of the same input (Bachman et al., 2019), though subsequent analyses showed that enforcing mutual information constraints too strongly can degrade downstream performance (McAllester & Stratos, 2020; Tschannen et al., 2020). A different line of work formalizes CL in terms of *alignment* and *uniformity* properties of the representation space (Wang & Isola, 2020; Wang & Liu, 2021; Chen et al., 2021a), capturing how positives concentrate while negatives spread across the sphere. These

geometric criteria, while intuitive, do not fully explain how samples from different semantic classes are organized under CL training.

To address this, several papers have studied the ability of CL to recover latent clusters and semantic structures (Arora et al., 2019; Tosh et al., 2021; Zimmermann et al., 2021; Ash et al., 2022; Nozawa & Sato, 2021; HaoChen et al., 2021; 2022; Shen et al., 2022; Wang et al., 2022; Awasthi et al., 2022; Bao et al., 2022). Most of these results rely on restrictive assumptions, such as conditional independence of augmentations given cluster identity (Arora et al., 2019; Tosh et al., 2021; Saunshi et al., 2022; Awasthi et al., 2022). To weaken such assumptions, HaoChen & Ma (2023) proposed analyzing spectral contrastive objectives that encourage cluster preservation without requiring augmentation connectivity, while Parulekar et al. (2023) showed that InfoNCE itself learns cluster-preserving embeddings when the hypothesis class is capacity-limited.

Another perspective comes from linking CL to supervised learning. For instance, Balestriero & LeCun (2024) showed that in linear models, self-supervised objectives such as VicReg coincide with supervised quadratic losses. In addition, Luthra et al. (2025) established an explicit coupling between the InfoNCE contrastive loss and a supervised variant that removes positives from the denominator. In contrast to prior results, these bounds are label-agnostic, architecture-independent, and hold uniformly throughout optimization. In a related vein, Lee (2025) formulate self-supervised contrastive learning as an approximation to supervised prototype-based objectives, deriving a balanced contrastive loss closely related to InfoNCE. On the representation-level alignment side, Grigg et al. (2021) provided empirical evidence that supervised and self-supervised trained models learn fairly geometrically aligned representations.

Beyond clustering and supervision, other theoretical studies have examined different aspects of CL: feature learning dynamics in linear and shallow nonlinear networks (Tian, 2022; Ji et al., 2023; Wen & Li, 2021; Tian, 2023), the role and optimality of augmentations (Tian et al., 2020; Feigin et al., 2025), the projection head (Gupta et al., 2022; Gui et al., 2023; Xue et al., 2024; Ouyang et al., 2025), sample complexity (Alon et al., 2024), and strategies to reduce batch-size requirements (Yuan et al., 2022). Finally, several works explore connections between contrastive and non-contrastive SSL paradigms (Wei et al., 2021; Balestriero & LeCun, 2022; Lee et al., 2021; Garrido et al., 2023; Shwartz-Ziv et al., 2023).

3 Problem Setup

We work with a dataset $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times [C]$ (with C classes), where $[C] = \{1, \dots, C\}$ and each class c contributes n_c examples. Here, $N = \sum_c n_c$ is the total number of samples and let $\pi_c = n_c/N$. An encoder $f_w : \mathcal{X} \to \mathbb{R}^d$ with parameters $w \in \mathbb{R}^p$ maps inputs to embeddings. Similarity is measured by a bounded function $\mathrm{sim} : \mathbb{R}^d \times \mathbb{R}^d \to [-1, 1]$; throughout our experiments we use cosine similarity on ℓ_2-normalized embeddings, $\mathrm{sim}(u, v) = \langle u, v \rangle / (\|u\| \|v\|)$.

Data augmentations are modeled by a Markov kernel $\alpha(\cdot \mid x)$ on $\mathcal{X}$: given x, we draw an independent view $x' \sim \alpha(x)$. Unless stated otherwise, augmentation draws are independent across samples, across repeated views of the same sample, and across training steps. We write $x_i' \sim \alpha(x_i)$ for a single view and $(x_i^{(1)}, x_i^{(2)}) \overset{\text{i.i.d.}}{\sim} \alpha(x_i)$ for two views of the same input.

Fix a batch size $B \in \mathbb{N}$. A batch is a multiset $\mathcal{B} = \{(x_i, x_i', y_i)\}_{i=1}^B$ sampled with replacement from S, with independent augmentations $x_i' \sim \alpha(x_i)$. For each element in the batch, define $z_i := f_w(x_i)$ and $z_i' := f_w(x_i')$. For any anchor triple $(x_i, x_i', y_i) \in \mathcal{B}$, define the per-anchor CL loss and the CL batch loss as

$$\ell_i^{\mathrm{CL}}(w; \mathcal{B}) := -\log \frac{\exp\big(\mathrm{sim}(z_i, z_i')/\tau\big)}{\sum_{\substack{t=1 \\ t \neq i}}^{B} \exp\big(\mathrm{sim}(z_i, z_t)/\tau\big) + \exp\big(\mathrm{sim}(z_i, z_t')/\tau\big)},$$

$$\bar{\ell}_{\mathcal{B}}^{\mathrm{CL}}(w) := \frac{1}{B} \sum_{i=1}^{B} \ell_i^{\mathrm{CL}}(w; \mathcal{B}).$$

For the same realized batch $\mathcal{B}$, define the negative index set $I_i^- := \{j \in \{1, \dots, B\} : y_j \neq y_i\}$ and the corresponding negative subset $\mathcal{B}_i^- := \{(x_j, x_j', y_j) : j \in I_i^-\}$. The NSCL per-anchor and batch

losses are

$$\ell_i^{\mathrm{NSCL}}(w;\mathcal{B}_i^-) \;:=\; -\log \frac{\exp\big(\mathrm{sim}(z_i,z_i')/\tau\big)}{\sum_{j\in I_i^-}\Big[\exp\big(\mathrm{sim}(z_i,z_j)/\tau\big)+\exp\big(\mathrm{sim}(z_i,z_j')/\tau\big)\Big]},$$

$$\bar{\ell}_{\mathcal{B}}^{\mathrm{NSCL}}(w) \;:=\; \frac{1}{B}\sum_{i=1}^{B}\ell_i^{\mathrm{NSCL}}(w;\mathcal{B}_i^-)\,.$$

Prior work (Luthra et al., 2025) shows that the CL–NSCL *loss* gap is uniformly $\mathcal{O}(1/C)$, but what we ultimately care about is whether the *embeddings* align. To quantify representation similarity we use linear Centered Kernel Alignment (CKA) and Representation Similarity Analysis (RSA) (Kornblith et al., 2019; Kriegeskorte et al., 2008) defined on cosine-similarity matrices: for N common inputs with embeddings $Z=\{z_i\}_{i=1}^N$ and $Z'=\{z_i'\}_{i=1}^N$, let $\Sigma(Z)_{ij}=\cos(z_i,z_j)$ and $H=I-\frac{1}{N}\mathbf{1}\mathbf{1}^\top$; linear CKA is

$$\mathrm{CKA}(Z,Z') \;=\; \frac{\langle H\Sigma(Z)H,\; H\Sigma(Z')H\rangle_F}{\|H\Sigma(Z)H\|_F\,\|H\Sigma(Z')H\|_F},$$

and RSA is the *Pearson* correlation between the (upper–triangular) off–diagonal entries of the dissimilarity matrices $\mathrm{RDM}(Z)=\mathbf{1}\mathbf{1}^\top-\Sigma(Z)$ and $\mathrm{RDM}(Z')=\mathbf{1}\mathbf{1}^\top-\Sigma(Z')$:

$$\mathrm{RSA}(Z,Z') \;=\; \mathrm{Corr}\left(\mathrm{vec}_\triangle(\mathrm{RDM}(Z)),\mathrm{vec}_\triangle(\mathrm{RDM}(Z'))\right),$$

where $\mathrm{vec}_\triangle$ stacks the upper–triangular entries $(i<j)$ column-wise.

This raises the following question: ***Beyond a small objective gap, does training CL and NSCL actually lead to similar representations (e.g., high CKA/RSA)?***

In the spirit of Thm. 1 of Luthra et al. (2025), we prove that when two runs use shared randomness (same initialization, mini-batches, and augmentations), the per-step gradient mismatch is uniformly bounded (Lem. 7). Similarly, we show that the CL and NSCL similarity matrices remain close throughout training (Thm. 1), which yields explicit CKA/RSA lower bounds (Cors. 1-2).

4 Theory

We examine how contrastive learning (CL) and negatives-only supervised contrastive learning (NSCL) co-evolve when initialized identically and trained with the same mini-batches and augmentations. While one might first attempt to study their trajectories in parameter space, such an approach quickly breaks down: without strong assumptions on the loss landscape (e.g., convexity or strong convexity), small reparameterizations can distort distances, and nonconvex dynamics cause parameter drift to grow uncontrollably over time (see App. C). For this reason, we set weight-space coupling aside and turn instead to the aspect that directly shapes downstream behavior—the *representations*—analyzing their alignment in similarity space.

4.1 Coupling in Representation (Similarity) Space

Let $\Sigma_t\in[-1,1]^{N\times N}$ denote the pairwise similarity matrix of a fixed reference set at step t (cosine similarity of normalized embeddings; diagonals are 1). We analyze the coupled evolution of the CL and NSCL similarities, Σ_t^{CL}, and $\Sigma_t^{\mathrm{NSCL}}\in[-1,1]^{N\times N}$ under identical mini-batches and augmentations. This representation-space view is invariant to reparameterization and directly tracks representational geometry.

Surrogate similarity dynamics. To make the analysis explicit, we work with a "similarity-descent" surrogate that updates only those entries touched by the current batch. For a realized mini-batch $\mathcal{B}_t=\{(x_j,x_j',y_j)\}_{j=1}^B$ (with $x_j'\sim\alpha(x_j)$), let $\bar{\ell}_{\mathcal{B}_t}^{\mathrm{CL}}(\Sigma)$ and $\bar{\ell}_{\mathcal{B}_t}^{\mathrm{NSCL}}(\Sigma)$ be the usual InfoNCE-type losses written as functions of the relevant similarity entries (with temperature $\tau>0$). Define the batch-gradient maps

$$G_t^{\mathrm{CL}} \;:=\; \nabla_\Sigma\,\bar{\ell}_{\mathcal{B}_t}^{\mathrm{CL}}\big(\Sigma_t^{\mathrm{CL}}\big), \qquad G_t^{\mathrm{NSCL}} \;:=\; \nabla_\Sigma\,\bar{\ell}_{\mathcal{B}_t}^{\mathrm{NSCL}}\big(\Sigma_t^{\mathrm{NSCL}}\big),$$

setting all untouched entries to zero. The surrogate updates are

$$\Sigma_{t+1}^{\mathrm{CL}} \;=\; \Sigma_t^{\mathrm{CL}}-\eta_t\,G_t^{\mathrm{CL}}, \qquad \Sigma_{t+1}^{\mathrm{NSCL}} \;=\; \Sigma_t^{\mathrm{NSCL}}-\eta_t\,G_t^{\mathrm{NSCL}}, \tag{1}$$

with shared initialization and shared randomness (same $\mathcal{B}_t$ and augmentations).

In App. D we show that these surrogate dynamics faithfully track the similarity evolution induced by parameter-space SGD. Intuitively, for the similarity map $\Sigma(w)$ and corresponding batch loss $\bar{\ell}(w)$ (either for CL or NSCL), one SGD step $w_{t+1} = w_t - \eta_t \nabla_w \bar{\ell}(w_t)$ induces $\Sigma(w_{t+1}) - \Sigma(w_t) = -\eta_t P_t G_t + R_t$, $G_t := \nabla_\Sigma \bar{\ell}\big(\Sigma(w_t)\big)$, $P_t := J_t J_t^\top$, $J_t := \partial\Sigma/\partial w|_{w_t}$, up to a second-order remainder R_t. Under the regularity assumptions $\|J(w)\|_{2\to 2} \le L_\Sigma$ and a quadratic Taylor bound on $\Sigma(w+\Delta w)$, together with bounded gradients and a learning-rate schedule with bounded $\sum_t \eta_t/(\tau^2 B)$ and $\sum_t \eta_t^2$, App. D shows that the induced trajectory $\widehat{\Sigma}_t := \Sigma(w_t)$ and the similarity-descent trajectory remain uniformly close. In particular, for small step sizes, sufficiently large batch B, and moderate temperature τ, parameter-space SGD moves similarities almost as if we performed gradient descent directly in similarity space, so the surrogate dynamics faithfully track the evolution of CL and NSCL representations. We now formalize the coupling bound.

Additional notation for high–probability factors. Fix a training horizon $T \in \mathbb{N}$, a confidence level $\delta \in (0,1)$, and a temperature $\tau > 0$. For later use, define $\epsilon_{B,\delta} := \sqrt{\frac{1}{2B}\log\Big(\frac{TB}{\delta}\Big)}$ and $\Delta_{\pi,\delta}(B;\tau) := \frac{2\,\mathrm{e}^{2/\tau}(\pi_{\max}+\epsilon_{B,\delta})}{1-\pi_{\max}-\epsilon_{B,\delta}}$ (where $\pi_{\max} = \max_c \pi_c$), and assume $\epsilon_{B,\delta} < 1 - \frac{1}{C}$ so the denominator is positive.

Theorem 1 (Similarity-space coupling). *Fix $B, T \in \mathbb{N}$, $\delta \in (0,1)$, and temperature $\tau > 0$. Consider the coupled similarity-descent recursions equation 1 for CL and NSCL with shared initialization and shared mini-batches/augmentations. Then, with probability at least $1-\delta$ over the draws of the mini-batches and augmentations, for any stepsizes $(\eta_t)_{t=0}^{T-1}$,*

$$\big\|\Sigma_T^{\mathrm{CL}} - \Sigma_T^{\mathrm{NSCL}}\big\|_F \;\le\; \exp\Big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\Big)\,\frac{1}{\tau\sqrt{B}}\Big(\sum_{t=0}^{T-1}\eta_t\Big)\,\Delta_{\pi,\delta}(B;\tau). \tag{2}$$

The above bound makes explicit how standard CL design choices control the discrepancy between CL and NSCL in similarity space. In particular, both the prefactor and the exponential term in equation 2 are monotone in the usual hyperparameters, so that regimes in which CL "behaves like" NSCL correspond precisely to regimes where the right-hand side of equation 2 is small. First, assuming balanced classes, a larger number of classes C reduces the $1/C$ contribution inside $\Delta_{\pi,\delta}(B;\tau)$, hence decreasing the overall bound and shrinking the CL–NSCL gap. Second, increasing the batch size B simultaneously reduces the concentration error $\epsilon_{B,\delta}$ and the factor $1/\sqrt{B}$, and also shrinks the coefficient $\frac{1}{2\tau^2 B}$ in the exponential, all of which act to decrease the right-hand side of equation 2 (see Fig. 5(d)). Third, increasing the temperature τ reduces the factors $\frac{1}{\tau}$ and $\frac{1}{\tau^2}$ appearing in the prefactor and exponent, again decreasing the upper bound in equation 2, consistent with the empirical trend in Fig. 4 that higher temperatures bring CL closer to NSCL. Finally, smaller learning rates η_t (or, more generally, a smaller total step size $\sum_t \eta_t$) reduce both the prefactor $\frac{1}{\tau\sqrt{B}}\sum_t \eta_t$ and the exponent $\exp\big(\frac{1}{2\tau^2 B}\sum_t \eta_t\big)$, so more conservative optimization schedules yield a tighter coupling between CL and NSCL (see Fig. 5). Overall, Thm. 1 shows that large batches, high temperatures, and small effective step sizes—are precisely the regimes in which the similarity dynamics of CL and NSCL nearly align.

As a final note, the result in Thm. 1 is stated in terms of similarity descent, whereas in practice we use gradient descent on the network's trainable parameters. To obtain an explicit bound on the gap between the CL and NSCL similarity matrices under standard parameter-space stochastic gradient descent, we can combine Thm. 1 with twice the bound in equation 9, applying that bound once to CL and once to NSCL.

From similarity drift to CKA/RSA guarantees. We translate the high-probability control on the similarity drift from Thm. 1, into bounds on two standard representational metrics.

CKA. Recall from Sec. 3 that linear CKA (Kornblith et al., 2019) is the normalized Frobenius inner product between centered similarity matrices. $H := I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ be the centering projector and define centered Gram matrices $K_T^{\mathrm{CL}} := H\Sigma_T^{\mathrm{CL}}H$ and $K_T^{\mathrm{NSCL}} := H\Sigma_T^{\mathrm{NSCL}}H$. The (linear) CKA at step T is $\mathrm{CKA}_T = \frac{\langle K_T^{\mathrm{CL}}, K_T^{\mathrm{NSCL}}\rangle_F}{\|K_T^{\mathrm{CL}}\|_F\,\|K_T^{\mathrm{NSCL}}\|_F} \in [0,1]$. Because $\|HXH\|_F \le \|X\|_F$, any bound on

$\|\Sigma_T^{\mathrm{CL}} - \Sigma_T^{\mathrm{NSCL}}\|_F$ controls $\|K_T^{\mathrm{CL}} - K_T^{\mathrm{NSCL}}\|_F$. For convenience, introduce the relative deviation $\rho_T := \frac{\|K_T^{\mathrm{CL}} - K_T^{\mathrm{NSCL}}\|_F}{\|K_T^{\mathrm{CL}}\|_F}$.

Corollary 1 (CKA lower bound). *In the setting of Thm. 1. Assume $\|K_T^{\mathrm{CL}}\|_F > 0$. With probability at least $1-\delta$,*

$$\mathrm{CKA}_T \;\geq\; \frac{1-\rho_T}{1+\rho_T}, \qquad \rho_T \;\leq\; \frac{\exp\!\big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\big)\,\frac{1}{\tau\sqrt{B}}\big(\sum_{t=0}^{T-1}\eta_t\big)\,\Delta_{\pi,\delta}(B;\tau)}{\|K_T^{\mathrm{CL}}\|_F}.$$

RSA. Recall from Sec. 3 that RSA (Kriegeskorte et al., 2008) is the Pearson correlation between the off-diagonal entries of representational dissimilarity matrices (RDMs). Let $M = \binom{N}{2}$ and define off-diagonal RDM vectors $a_T, b_T \in \mathbb{R}^M$ by $a_T(u,v) = 1 - \Sigma_T^{\mathrm{CL}}(u,v)$ and $b_T(u,v) = 1 - \Sigma_T^{\mathrm{NSCL}}(u,v)$ for $u < v$. Write $\sigma_{D,T} > 0$ for the empirical standard deviation of the entries of a_T. The RSA score is the Pearson correlation $\mathrm{RSA}_T = \mathrm{Corr}(a_T, b_T)$. Zeroing the diagonal does not increase Frobenius norms, so $\|b_T - a_T\|_2 \leq \|\Sigma_T^{\mathrm{NSCL}} - \Sigma_T^{\mathrm{CL}}\|_F$. It will be useful to measure the relative discrepancy $r_T := \frac{\|b_T - a_T\|_2}{\sqrt{M}\,\sigma_{D,T}}$.

Corollary 2 (RSA lower bound). *In the setting of Thm. 1. Assume $\sigma_{D,T} > 0$. With probability at least $1-\delta$,*

$$\mathrm{RSA}_T \;\geq\; \frac{1-r_T}{1+r_T}, \qquad r_T \;\leq\; \frac{\exp\!\big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\big)\,\frac{1}{\tau\sqrt{B}}\big(\sum_{t=0}^{T-1}\eta_t\big)\,\Delta_{\pi,\delta}(B;\tau)}{\sqrt{M}\,\sigma_{D,T}}.$$

These results complement the parameter–space analysis. While parameter trajectories may diverge exponentially (in the non-convex setting), the induced similarities—and hence representational metrics such as CKA and RSA—remain tightly controlled by class count, batch size, learning rate, and temperature τ. The key quantity is the similarity–matrix drift $\|\Sigma_T^{\mathrm{CL}} - \Sigma_T^{\mathrm{NSCL}}\|_F$, which Thm. 1 bounds in two stabilizing ways.

First, the exponential factor is moderated by the $\frac{1}{\tau^2 B}$ term in the exponent. Unlike parameter space, where the growth rate scales with β, the "instability rate" in similarity space is only $\frac{1}{2\tau^2 B}$ and is therefore negligible for typical batch sizes (e.g., $B \approx 10^2$–10^3).

Second, the prefactor $\frac{1}{\tau\sqrt{B}}\big(\sum_t \eta_t\big)\,\Delta_{\pi,\delta}(B;\tau)$ decreases rapidly with batch size and class count (note $\Delta_{\pi,\delta}(B;\tau)$ shrinks with smaller $\pi_{\max}$ and grows with smaller τ through $\mathrm{e}^{2/\tau}$). In practical regimes ($C \sim 10^3$, $B \sim 10^2$–10^3), this prefactor is small, making the total Frobenius gap negligible relative to the scale of the similarity matrices.

Together, these effects yield high–probability guarantees $\mathrm{CKA}_T \geq (1-\rho_T)/(1+\rho_T)$ and $\mathrm{RSA}_T \geq (1-r_T)/(1+r_T)$ with $\rho_T, r_T \ll 1$ in realistic conditions. Thus, even if parameters drift, the induced representations evolve in a coupled and stable manner—consistent with empirical findings that CL and NSCL remain closely aligned in practice.

Proof idea. We begin with a high–probability batch–composition guarantee (Cor. 3): with probability at least $1-\delta$, every anchor's denominator contains the expected proportion of negatives up to an $\epsilon_{B,\delta}$ fluctuation. This rules out positive–heavy batches that would otherwise cause the NSCL renormalization to deviate substantially from CL. Conditioning on this event, the CL–NSCL batch–gradient gap decomposes into (i) a *reweighting error*, bounded in total variation by $\Delta_{\pi,\delta}(B;\tau)$ (Lem. 6), and (ii) a *stability term* from the dependence on the current similarities, controlled by the $\frac{1}{2\tau^2 B}$–Lipschitzness of the batch–gradient map in Frobenius norm (Lem. 2 at temperature τ). Using block–orthogonality across anchors (Lem. 1), the reweighting contributions combine in quadrature, giving the per–step estimate (Lem. 8),

$$\big\|G_t^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) - G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{NSCL}})\big\|_F \;\leq\; \frac{1}{\tau}\cdot\frac{\Delta_{\pi,\delta}(B;\tau)}{\sqrt{B}} \;+\; \frac{1}{2\tau^2 B}\,\big\|\Sigma_t^{\mathrm{CL}} - \Sigma_t^{\mathrm{NSCL}}\big\|_F.$$

Consequently, the similarity drift satisfies the recurrence

$$\big\|\Sigma_{t+1}^{\mathrm{CL}} - \Sigma_{t+1}^{\mathrm{NSCL}}\big\|_F \;\leq\; \Big(1 + \frac{\eta_t}{2\tau^2 B}\Big)\,\big\|\Sigma_t^{\mathrm{CL}} - \Sigma_t^{\mathrm{NSCL}}\big\|_F \;+\; \eta_t\,\frac{1}{\tau}\cdot\frac{\Delta_{\pi,\delta}(B;\tau)}{\sqrt{B}},$$

	CIFAR-10		CIFAR-100		Mini-ImageNet		Tiny-ImageNet	
	NCCC	LP	NCCC	LP	NCCC	LP	NCCC	LP
CL	88.37	90.16	54.62	65.65	60.78	65.30	40.59	44.61
NSCL	94.47	94.09	60.14	68.38	63.92	72.60	40.76	45.79
SCL	94.93	94.67	64.06	69.52	74.78	76.00	48.63	48.73
CE	92.97	93.39	67.35	68.04	75.20	74.00	48.28	52.57

Table 1: Nearest Class-Center Classifier (NCCC) and Linear Probe (LP) test accuracies (%). We report the accuracies against the all-way classification task in each dataset. The models (also used in Fig. 2) were pre-trained on their respective datasets.

where the first term propagates existing error and the second injects the new discrepancy introduced at step t. Unrolling this recurrence (discrete Grönwall) yields

$$\|\Sigma_T^{\mathrm{CL}} - \Sigma_T^{\mathrm{NSCL}}\|_F \;\le\; \exp\Big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\Big)\,\frac{1}{\tau\sqrt{B}}\Big(\sum_{t=0}^{T-1}\eta_t\Big)\,\Delta_{\pi,\delta}(B;\tau).$$

Finally, centering contracts Frobenius norms, so this control transfers directly to the centered Gram matrices, and applying standard $(1-\rho)/(1+\rho)$ and $(1-r)/(1+r)$ comparisons yields the claimed CKA/RSA lower bounds.

5 Experiments

Datasets and augmentations. We experiment with the following standard vision classification datasets - CIFAR10 and CIFAR100 (Krizhevsky, 2009), Mini-ImageNet (Vinyals et al., 2016), Tiny-ImageNet (Han, 2020), and ImageNet-1K (Deng et al., 2009). (See App. B for details.)

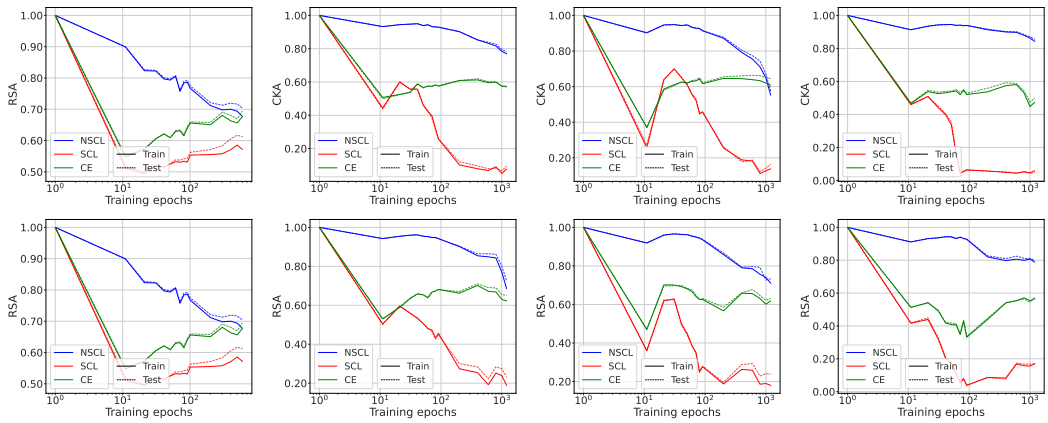
Methods, architectures, and optimizers. For all our experiments, we have followed the SimCLR (Chen et al., 2020) algorithm. We use a ResNet-50 (He et al., 2016) encoder with a width-multiplier factor of 1. The projection head follows a standard two-layer MLP architecture composed of: `Linear`$(2048 \to 2048)$ $\to$ `ReLU` $\to$ `Linear`$(2048 \to 128)$. For cross-entropy training, we attach an additional classification head `Linear`$(128 \to$ `C`$)$ where C is the number of classes.

For contrastive learning, we use the DCL loss that avoids positive-negative coupling during training (Yeh et al., 2022). For supervised learning, we use the following variants: Supervised Contrastive Loss (Khosla et al., 2020), Negatives-Only Supervised Contrastive Loss (Luthra et al., 2025), and Cross-Entropy Loss (Shannon, 1948). To minimize the loss, we adopt the LARS optimizer (You et al., 2017) which has been shown in (Chen et al., 2020) to be effective for training with large batch sizes. For LARS, we set the momentum to 0.9 and the weight decay to $1\mathrm{e}^{-6}$. All experiments are carried out with a batch size of $B = 1024$. The base learning rate is scaled with batch size as $0.3 \cdot \lfloor B/256 \rfloor$, following standard practice (Chen et al., 2020). We employ a warm-up phase (Goyal et al., 2017) for the first 10 epochs, followed by a cosine learning rate schedule without restarts (Loshchilov & Hutter, 2016) for the remaining epochs. All models were trained on a single node with one 94 GB NVIDIA H100 GPU.

Evaluation metrics. To quantitatively measure the alignment between the learned representation spaces of different models, we monitor linear CKA and RSA (check Sec. 3 for details) during training. Both CKA and RSA range from 0 to 1, where 1 indicates identical similarity structures. To manage the significant memory requirements of $N \times N$ matrices (Gram matrices for CKA, RDMs for RSA), we use a memory-efficient, chunk-wise computation strategy.

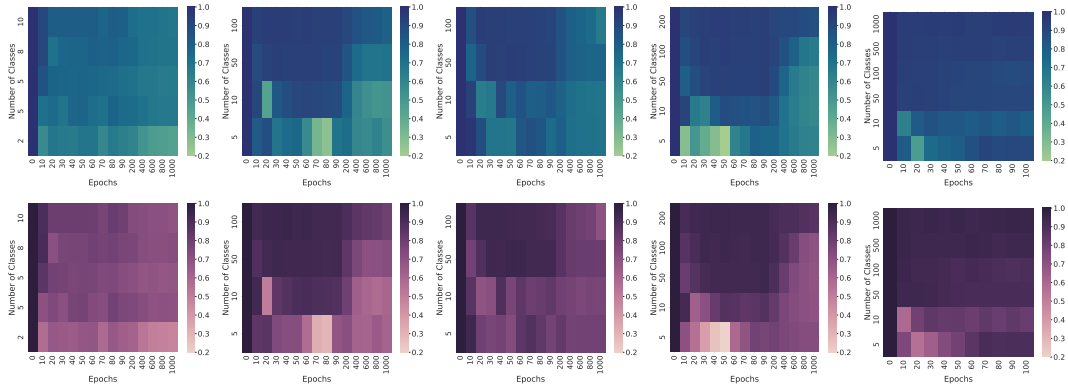
5.1 Experimental Results

Alignment analysis as a function of epochs. To understand how representational similarity evolves, we trained a model with a CL objective and monitored its alignment (via CKA/RSA) against supervised models trained with NSCL, CE, and SCL. We find that NSCL consistently achieves the



(a) CIFAR10 **(b)** CIFAR100 **(c)** Mini-ImageNet **(d)** Tiny-ImageNet

Figure 2: **Alignment during training.** We train ResNet-50 models with decoupled CL, SCL, NSCL, and CE. For the first 1,000 epochs, the CL-trained model is substantially more aligned with the NSCL-trained model than with the others. However, alignment declines when training continues much longer.

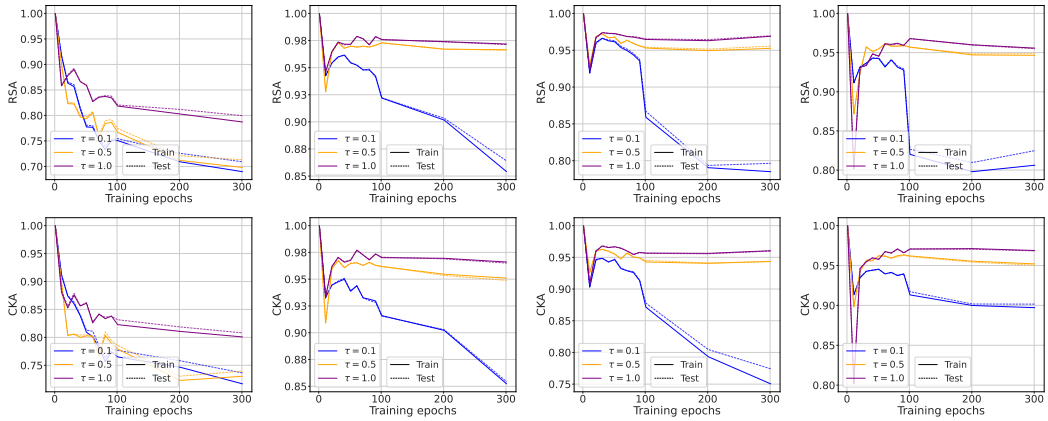


CIFAR-10 CIFAR-100 Mini-ImageNet Tiny-ImageNet ImageNet-1k

Figure 3: **CL–NSCL alignment (linear CKA) increases with the number of training classes.** The heatmaps show the linear CKA between CL and NSCL models. We visualize alignment on the training (top row, green) and test (bottom row, purple) sets. The y-axis indicates the number of classes (N) used for training, and the x-axis represents the training epoch. While alignment is consistently higher for larger N, it also tends to decrease as training progresses for any fixed N.

highest alignment with CL throughout training across multiple datasets compared to CE and SCL (see Fig. 2). For example, after 1k epochs on Tiny-ImageNet, the CL-NSCL alignment reaches a CKA of 0.87, in contrast to just 0.043 for CL-SCL.

Intuitively, these alignment patterns follow from how each loss shapes representation geometry. All three methods incentivize neural collapse (Papyan et al., 2020; Han et al., 2022; Zhou et al., 2022; Lu & Steinerberger, 2022; Dang et al., 2024; Graf et al., 2021; Awasthi et al., 2022; Gill et al., 2023; Kini et al., 2024; Luthra et al., 2025), but differ in how directly and how quickly they drive it. NSCL is structurally closest to CL: both attract a single positive toward an anchor and repel negatives, primarily enforcing instance-level discrimination and thus inducing similar geometry. SCL, by contrast, imposes a stronger class-level constraint, explicitly pulling together augmentations of same-class samples and pushing apart different-class samples, which rapidly reduces intra-class variance and forms tight class clusters that depart from CL's instance-level structure. Cross-entropy (CE) lies between these extremes, promoting collapse more indirectly via error minimization with regularization. In the self-supervised setting, CL representations need not collapse as tightly as supervised ones, since they are learned without labels. As training enters the 10–100-epoch range,



(a) CIFAR10 **(b)** CIFAR100 **(c)** Mini-ImageNet **(d)** Tiny-ImageNet

Figure 4: **Higher τ increases the CL-NSCL alignment.** The plots show RSA (top row) and CKA (bottom row) over 300 epochs. We trained CL and NSCL models with varying temperatures ($\tau \in \{0.1, 0.5, 1.0\}$) on four datasets. Across all datasets, a higher temperature $\tau = 1.0$ (shown in purple) evidently results in the highest alignment.

SCL and CE move closer to the neural collapse regime, while NSCL continues to mimic the CL label-free optimization for a longer duration, producing the evolving alignment dynamics in Fig. 2.

For completeness, along with CKA and RSA, we also report downstream performance via Nearest Class Center Classifier (Galanti et al., 2022) and Linear Probe accuracies in Tab. 1.

Validating Thm. 1 as a function of class count. Thm. 1 predicts that using more classes yields stronger CL–NSCL alignment. We test this via C'-way training: for each $C' \in [2, C]$, we train CL and NSCL on random C'-class subsets for 1,000 epochs (except 100 epochs for IM-1K). As shown in Fig. 3, representation similarity (RSA/CKA) increases with C' across all datasets.

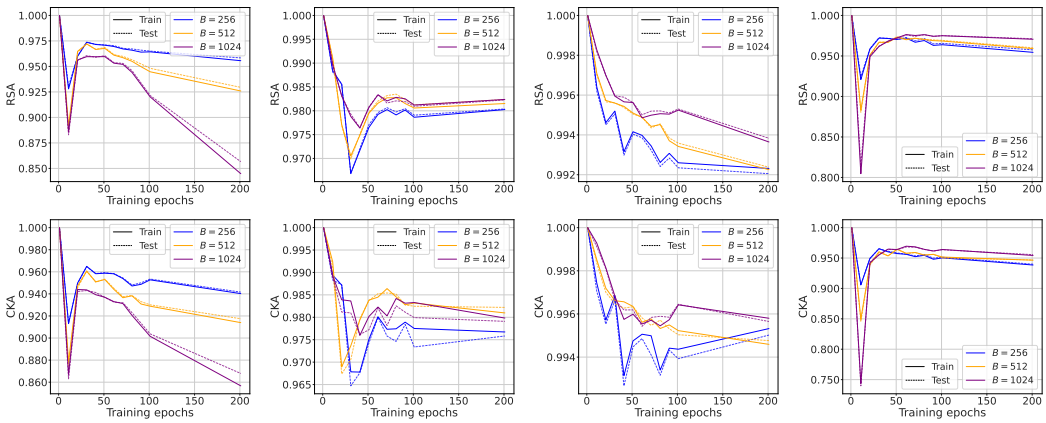
Effect of temperature on alignment. As per Thm. 1 and Cors. 1-2, CL-NSCL alignment improves with higher values of temperature (τ). We empirically verify this claim by training CL and NSCL models for 300 epochs, over three different values of $\tau \in \{0.1, 0.5, 1.0\}$. Both models–CL and NSCL–are trained with same τ in each run. As shown in Fig. 4, models trained with $\tau = 1.0$ achieve higher alignment compared to models trained with lower temperatures.

Effect of batch size on alignment. Thm. 1 links alignment to a bound that may rise or fall with B depending on how the learning rate scales. To investigate this, we vary η with B across four cases: $\eta = \frac{0.3B}{256}$, $\eta = \frac{0.3\sqrt{B}}{256}$, $\eta = \frac{0.3\sqrt[4]{B}}{256}$, and $\eta = 0.3$. Under $\mathcal{O}(B)$ scaling, CL–NSCL alignment decreases as B grows, matching the theorem's implication for that scaling; for the other three cases, alignment increases with B, again consistent with the bound under those dependencies (see Fig. 5).

Weight-space coupling. We next study whether the observed alignment between representations of contrastive and supervised models is also reflected directly in their parameters. For this, we measure the average weight difference between a contrastive model and two supervised counterparts as follows: $\sum_l \frac{\|w^l_{\text{CL}} - W^l_{\text{sup}}\|_F}{0.5\,(\|w^l_{\text{CL}}\|_F + \|w^l_{\text{sup}}\|_F)}$ where w^l_{CL} and w^l_{sup} are weights corresponding to l^{th} layer of self-supervised and supervised models respectively, and $\|\cdot\|_F$ denotes Frobenius norm. As we show in Fig. 6, for each dataset, we observe a significant divergence in weight space: both supervised models (NSCL and SCL) increasingly separate from the contrastive model as training progresses.

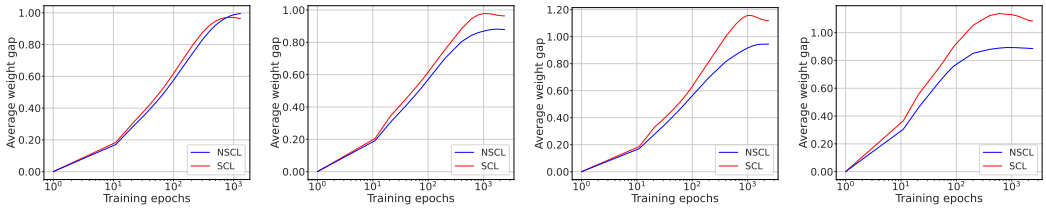
6 Conclusions, Limitations and Future Work

Conclusions. We studied the dynamic alignment between contrastive learning (CL) and its supervised counterpart (NSCL). By analyzing coupled SGD under shared randomness, we showed that while parameter-space trajectories may diverge exponentially, representation-space dynamics are far more stable: the similarity matrices induced by CL and NSCL remain close throughout training. This



(a) $\mathcal{O}(B)$ **(b)** $\mathcal{O}(B^{1/2})$ **(c)** $\mathcal{O}(B^{1/4})$ **(d)** $\mathcal{O}(1)$

Figure 5: **Effect of batch size with scaled learning rates.** We trained CL, and NSCL models for 300 epochs with varying batch-sizes ($B \in \{256,\ 512,\ 1024\}$). For each experiment, the learning rate η is scaled as a function of batch-size, as mentioned under each panel. For instance, the results shown in panel (b) use a learning rate of $\eta = \frac{0.3\sqrt{B}}{256}$.



(a) CIFAR10 **(b)** CIFAR100 **(c)** Mini-ImageNet **(d)** Tiny-ImageNet

Figure 6: **Weight-space alignment quickly deteriorates.** Using the same ResNet-50 instances as in Fig. 2, we plot the average weight gap between CL and the supervised models (NSCL and SCL) across training epochs. Both supervised variants diverge from the CL model, with SCL showing a wider separation.

yields high-probability lower bounds on alignment metrics such as CKA and RSA, directly certifying representational coupling. Empirically, our experiments confirmed these trends across datasets and architectures. Together, our results highlight that the implicit supervised signal in CL is not confined to its loss function but extends throughout the entire optimization trajectory.

Limitations. Our theoretical bounds are structurally informative but not expected to be tight in large-scale or long-horizon regimes. As is common in machine learning theory, the guarantees are conservative worst-case bounds derived from uniform high-probability arguments, favoring generality over numerical sharpness. Many influential results in optimization and stability theory for deep learning similarly rely on loose worst-case analyses—e.g., (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Mou et al., 2018; Kuzborskij & Lampert, 2017)—yet still provide useful conceptual guidance. In our setting, without additional structural assumptions (such as stronger curvature or smoothness conditions), one cannot generally expect qualitatively sharper dependence than the scaled exponential factors appearing in Thm. 1 and equation 9. Thus, while in practice the bounds are quite loose, they achieve their intended goal of identifying which parameters govern the CL–NSCL similarity gap and explaining how this gap scales with them.

Future directions. We view our results as a first step toward a more refined theory of self-supervised alignment. Future work could (i) derive tighter constants by exploiting data-dependent structure rather than worst-case bounds, and (ii) extend the framework to other SSL paradigms (e.g., non-contrastive methods). Improving these guarantees while retaining their stability properties would provide an even stronger theoretical bridge between supervised and self-supervised learning.

7 Reproducibility Statement

We have taken several steps to ensure the reproducibility of our results. All datasets used in this work (CIFAR-10/100, Tiny-ImageNet, and Mini-ImageNet) are publicly available, and we describe the data processing and augmentation pipelines in Section 3 and App. B. The theoretical results are supported by detailed proofs in App. C, D, E, where all assumptions are explicitly stated. Experimental details, including architectures, optimizers, hyperparameters, and training schedules, are reported in Section 3, with additional clarifications in the appendix. To facilitate further verification, we provide an anonymous code repository in the supplementary material that contains implementations of the CL, NSCL, and baseline objectives, along with scripts to reproduce all figures and tables in the paper. Together, these resources are intended to make both the theoretical and empirical findings fully reproducible.

References

Noga Alon, Dmitrii Avdiukhin, Dor Elboim, Orr Fischer, and Grigory Yaroslavtsev. Optimal sample complexity of contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NU9AYHJvYe.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019. URL https://arxiv.org/abs/1902.09229.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7187–7209. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/ash22a.html.

Pranjal Awasthi, Nishanth Dikkala, and Pritish Kamath. Do more negative samples necessarily hurt in contrastive learning? In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1101–1116. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/awasthi22b.html.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=jQgsZDspz5h.

Randall Balestriero and Yann LeCun. The birth of self supervised learning: A supervised theory. In *NeurIPS 2024 Workshop: Self-Supervised Learning - Theory and Practice*, 2024. URL https://openreview.net/forum?id=NhYAjAAdQT.

Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1585–1606. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bao22e.html.

Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse engineering self-supervised learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=NsVEjx6YPd.

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2: 499–526, March 2002. ISSN 1532-4435. doi: 10.1162/153244302760200704. URL https://doi.org/10.1162/153244302760200704.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021a.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9640–9649, October 2021b.

Hien Dang, Tho Tran, Tan Nguyen, and Nhat Ho. Neural collapse for cross-entropy class-imbalanced learning with unconstrained relu features model. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Shlomo Libo Feigin, Maximilian Fleissner, and Debarghya Ghoshdastidar. A theoretical characterization of optimal data augmentations in self-supervised learning, 2025. URL https://arxiv.org/abs/2411.01767.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SwIp410B6aQ.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kDEL91Dufpa.

Jaidev Gill, Vala Vakilian, and Christos Thrampoulidis. Engineering the neural collapse geometry of supervised-contrastive loss, 2023. URL https://arxiv.org/abs/2310.00893.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3821–3830. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/graf21a.html.

Tom George Grigg, Dan Busbridge, Jason Ramapuram, and Russ Webb. Do self-supervised and supervised methods learn similar visual representations?, 2021. URL https://arxiv.org/abs/2110.00528.

Yu Gui, Cong Ma, and Yiqiao Zhong. Unraveling projection heads in contrastive learning: Insights from expansion and shrinkage, 2023. URL https://arxiv.org/abs/2306.03335.

Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning, 2022. URL https://arxiv.org/abs/2212.11491.

Liu Han. Tiny imagenet challenge. https://kaggle.com/competitions/deep-learning-thu-2020, 2020. Kaggle.

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*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=AuEgNlEAmed.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5000–5011. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/27debb435021eb68b3965290b5e24c49-Paper.pdf.

Jeff Z. HaoChen, Colin Wei, Ananya Kumar, and Tengyu Ma. Beyond separability: analyzing the linear transferability of contrastive representations to related subpopulations. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1225–1234. JMLR.org, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, October 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: a theoretical analysis. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Ganesh Ramachandra Kini, Vala Vakilian, Tina Behnia, Jaidev Gill, and Christos Thrampoulidis. Symmetric neural-collapse representations with supervised contrastive loss: The impact of reLU and batching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AyXIDfvYg8.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis: connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, November 2008. doi: 10.3389/neuro.06.004.2008.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ilja Kuzborskij and Christoph H. Lampert. Data-dependent stability of stochastic gradient descent. *arXiv preprint arXiv:1703.01678*, 2017.

Byeongchan Lee. Understanding self-supervised contrastive learning through supervised objectives. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=cmE97KX2XM.

Jason D Lee, Qi Lei, Nikunj Saunshi, and JIACHENG ZHUO. Predicting what you already know helps: Provable self-supervised learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 309–323. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/02e656adee09f8394b402d9958389b7d-Paper.pdf.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.011. URL https://www.sciencedirect.com/science/article/pii/S1063520321001123. Special Issue on Harmonic Analysis and Machine Learning.

Achleshwar Luthra, Tianbao Yang, and Tomer Galanti. Self-supervised contrastive learning is approximately supervised contrastive learning, 2025. URL https://arxiv.org/abs/2506.04411.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/mcallester20a.html.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 605–638. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/mou18a.html.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Kento Nozawa and Issei Sato. Understanding negative samples in instance discriminative self-supervised representation learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5784–5797. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/2dace78f80bc92e6d7493423d729448e-Paper.pdf.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=L0evcuybH5.

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*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Advait Parulekar, Liam Collins, Karthikeyan Shanmugam, Aryan Mokhtari, and Sanjay Shakkottai. Infonce loss provably learns cluster-preserving representations. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1914–1961. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/parulekar23a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19250–19286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saunshi22a.html.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019. URL https://arxiv.org/abs/1904.05862.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z. Haochen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19847–19878. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/shen22d.html.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G. J. Rudner, and Yann LeCun. An information theory perspective on variance-invariance-covariance regularization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=KipjqOPaZ0.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c2e5eaae9152079b9e95845750bb9ab-Paper.pdf.

Yuandong Tian. Understanding deep contrastive learning via coordinate-wise optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19511–19522. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7b5c9cc08960df40615c1d858961eb8b-Paper-Conference.pdf.

Yuandong Tian. Understanding the role of nonlinearity in training dynamics of contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=s130rTE3U_X.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ECvgmYVyeUz.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rC8sJ4i6kaH.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11112–11122. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wen21c.html.

Xi Weng, Jianing An, Xudong Ma, Binhang Qi, Jie Luo, Xi Yang, Jin Song Dong, and Lei Huang. Clustering properties of self-supervised learning, 2025. URL https://arxiv.org/abs/2501.18452.

Yihao Xue, Eric Gan, Jiayi Ni, Siddharth Joshi, and Baharan Mirzasoleiman. Investigating the benefits of projection head for representation learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GgEAdqYPNA.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pp. 668–684, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19808-3. doi: 10.1007/978-3-031-19809-0_38. URL https://doi.org/10.1007/978-3-031-19809-0_38.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, and Tianbao Yang. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25760–25782. PMLR, 2022. URL https://proceedings.mlr.press/v162/yuan22b.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31697–31710. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/cdce17de141c9fba3bdf175a0b721941-Paper-Conference.pdf.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

A LLM USAGE STATEMENT

Large Language Models (LLMs) were used solely as an assistive tool for improving the clarity and presentation of the manuscript (e.g., editing grammar, refining phrasing). All technical content, including theoretical derivations, proofs, experimental design, and analysis, was developed entirely by the authors. No parts of the paper were written or ideated by an LLM in a way that would constitute substantive scientific contribution, and no LLM was used to generate or fabricate results.

B ADDITIONAL EXPERIMENTS

Datasets and augmentations. CIFAR10 and CIFAR100 both consist of 50000 training images and 10000 validation images with 10 classes and 100 classes, respectively, uniformly distributed across the dataset, i.e., CIFAR10 has 5000 samples per class and CIFAR100 has 500 samples per class. Mini-ImageNet also has 5000 test images on top of 50000 train and 10000 validation images, with 100 of 1000 classes from ImageNet-1k (Deng et al., 2009) (at the original resolution). Tiny-ImageNet contains 100000 images downsampled to 64×64, with total 200 classes from IM-1K. Each class has 500 training, 50 validation, and 50 test images.

We use standard augmentations as proposed in SimCLR (Chen et al., 2020). For experiments on Mini-ImageNet, we use the following pipeline: random resized cropping to 224×224, random horizontal flipping, color jittering (brightness, contrast, saturation: 0.8; hue: 0.2), random grayscale conversion ($p = 0.2$), and Gaussian blur (applied with probability 0.1 using a 3×3 kernel and $\sigma = 1.5$). For Tiny-ImageNet, we drop saturation to 0.4 and hue to 0.1 due to low resolution images. For CIFAR datasets, we adopt a similar pipeline with appropriately scaled parameters. The crop size is adjusted to 32×32, and the color jitter parameters are scaled to saturation 0.4, and hue 0.1.

B.1 EXPERIMENTS WITH THE VIT ARCHITECTURE

To further support the claims made in the main text, we reproduce the experiment from Fig. 2 using the ViT-Base architecture (Dosovitskiy et al., 2021). Throughout these experiments, we use the same training hyperparameters and augmentations for each dataset as in the ResNet-50 experiments. As shown in Fig. 7, the alignment between CL and supervised models exhibits the same qualitative trends observed for the ResNet-50 architecture in Fig. 2, demonstrating that the relationship between training dynamics and representational alignment is consistent across both convolutional and transformer-based models.

In addition, we repeat the experiments in Figs. 4 and 5 for the ViT-Base architecture. The corresponding results, shown in Figs. 8 and 9, closely match those obtained with ResNet-50, further reinforcing the robustness of our findings across architectures.

B.2 EFFECT OF NUMBER OF CLASSES ON ALIGNMENT

In addition to the linear CKA results reported in the main text (Fig. 3), we also evaluate representational similarity using RSA. The corresponding RSA values are presented in Fig. 10, providing a complementary perspective on alignment across varying numbers of classes. In addition, we also reproduced the results with RSA for the ViT models (Fig. 11).

B.3 PERFORMANCE-ALIGNMENT TRADEOFF

The bound in Thm. 1 predicts that alignment increases with larger τ. Moreover, when $\eta_t = \mathcal{O}(B)$, it suggests that alignment should decrease as B grows, whereas under $\eta_t = \mathcal{O}(B^{1/4})$ it instead predicts higher alignment for larger B. In this experiment, we examine whether higher alignment in fact corresponds to more similar downstream accuracies. Specifically, in Figs. 12–13 we vary the parameters τ and B (respectively) and plot the gap between the accuracies of the CL and NSCL models against their RSA alignment values. To obtain the accuracy measures, we perform full-shot linear probing on both the CL- and NSCL-trained models and report their test accuracies. As can be seen from the results, we consistently observe that higher alignment corresponds to a smaller gap between the accuracy rates of the CL- and NSCL-trained models. This suggests that the alignment

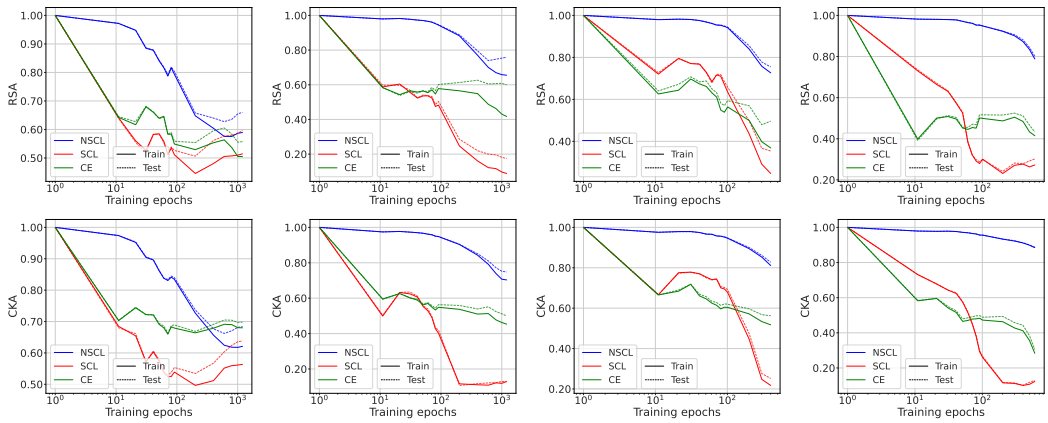


Figure 7: **Alignment during training for ViT.** We train ViT-base model with CL, NSCL, SCL and CE objectives. The alignment between CL and supervised models follow similar trends as shown for ResNet-50 in Fig. 2.

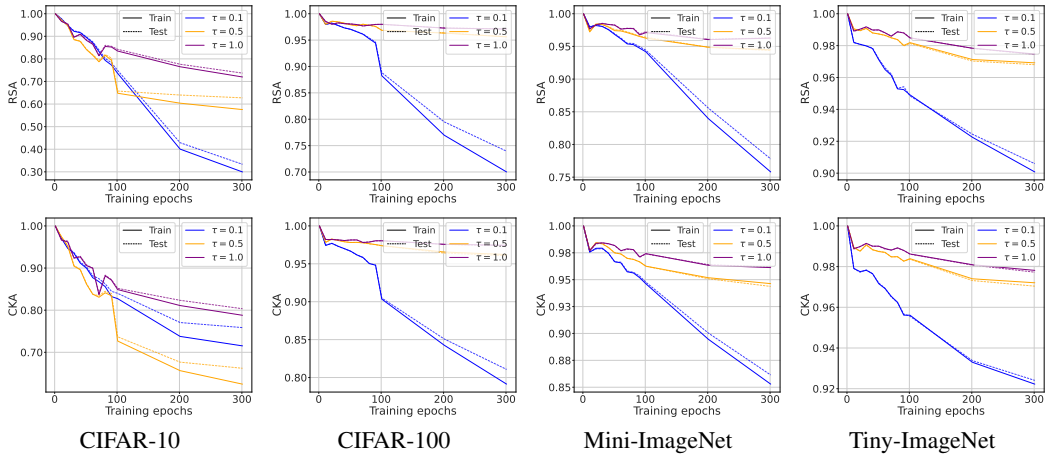


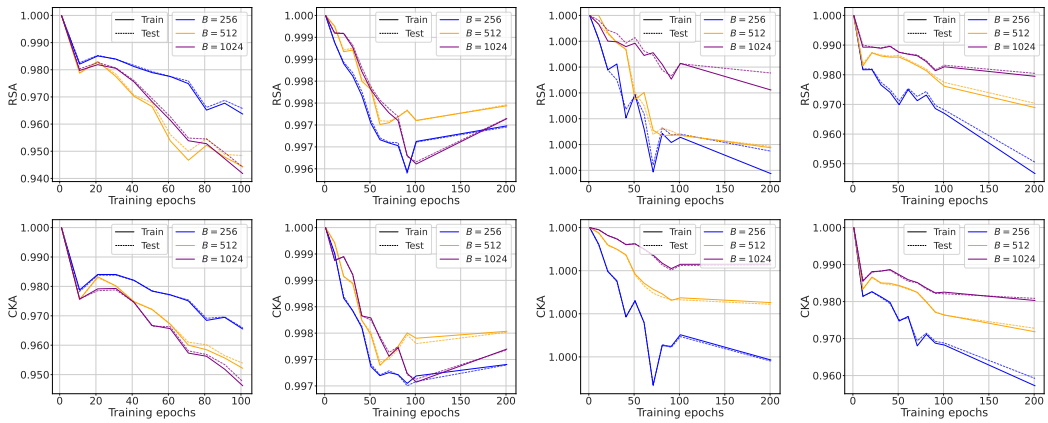
Figure 8: **Effect of the temperature (τ) on CL–NSCL alignment.** We train ViT-Base models with decoupled CL and NSCL objectives using different temperature values τ. All models are trained for 300 epochs. Across all datasets, alignment consistently increases as τ becomes larger.

between CL and NSCL models translates into concrete predictions about how close the models are in their performance. For completeness, we summarize these accuracy values in Tab. 2.

	CIFAR-100			Mini-ImageNet			Tiny-ImageNet		
	$\tau = 0.1$	$\tau = 0.5$	$\tau = 1.0$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 1.0$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 1.0$
CL	65.18	61.62	58.60	70.30	70.55	68.21	44.50	40.41	35.40
NSCL	68.25	62.44	59.02	73.93	71.88	67.76	46.51	39.95	35.29

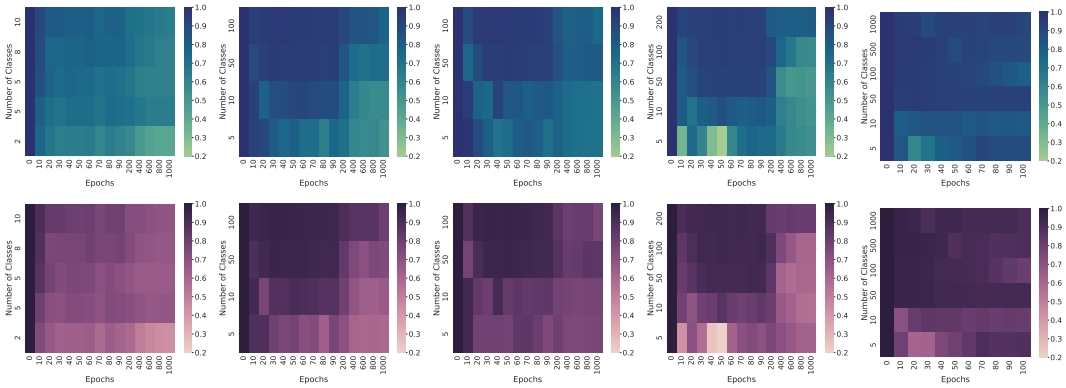
Table 2: Linear Probe (LP) test accuracies (%) for varying τ. We train CL and NSCL ResNet-50 models for 300 epochs, and observe that the accuracy gap decreases with higher alignment between CL and NSCL models (also shown in Fig. 12).

B.4 Experiments with Class-Imbalanced Data



(a) $\mathcal{O}(B)$ **(b)** $\mathcal{O}(B^{1/2})$ **(c)** $\mathcal{O}(B^{1/4})$ **(d)** $\mathcal{O}(1)$

Figure 9: **Effect of batch size (B) on CL-NSCL alignment.** We follow the same learning-rate scaling strategy as for ResNet-50. The alignment trends observed when varying the batch size are similar to those for ResNet-50.



CIFAR-10 CIFAR-100 Mini-ImageNet Tiny-ImageNet ImageNet-1k

Figure 10: **CL–NSCL alignment (RSA) increases with the number of training classes.** See Sec. 5.1 and Fig. 3 for experimental details.

Since our theory is tighter for relatively balanced classes, but does not require perfectly balanced data, we also evaluate it on the SVHN dataset (Netzer et al., 2011), which is well known for its pronounced class imbalance. In Fig. 14, we plot the RSA and CKA metrics between coupled CL and NSCL models trained for 300 epochs. The training hyperparameters and the data augmentations are the same as in our CIFAR-100 experiments to facilitate a direct comparison.

Despite the class imbalance in SVHN, we observe that the alignment between the two models is consistently high—indeed, it is even stronger than what we typically obtain after 1,000 epochs on CIFAR-100, which has the same number of classes. This finding suggests that substantial class imbalance does not hinder strong representational alignment from emerging between coupled CL and NSCL models, and further supports the robustness of our theoretical predictions beyond the approximately balanced setting.

B.5 Attention Maps Alignment

Methodology. To analyze the self-attention maps from the frozen Vision Transformer encoder, we look into the Multi-Head Self-Attention (MHSA) mechanism of the final transformer layer ($L = 12$ for ViT-Base).

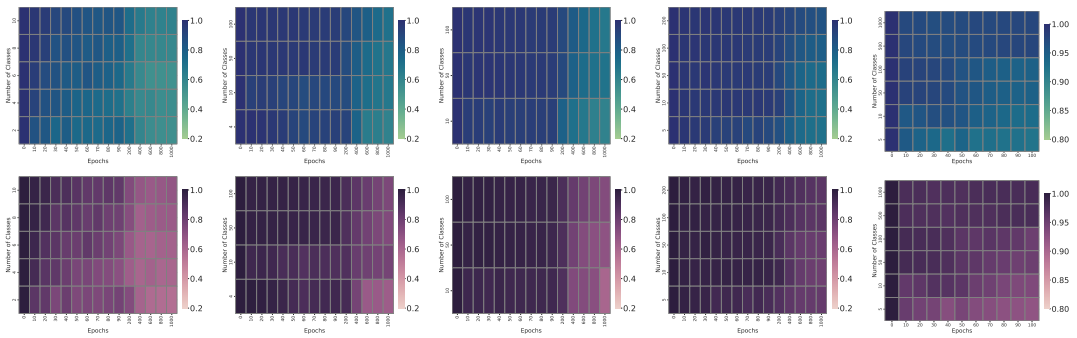


Figure 11: **CL–NSCL alignment (RSA) increases with the number of training classes for ViT-Base models.** The alignment increases with number of classes, and is consistent with trends observed for ResNet-50 models.

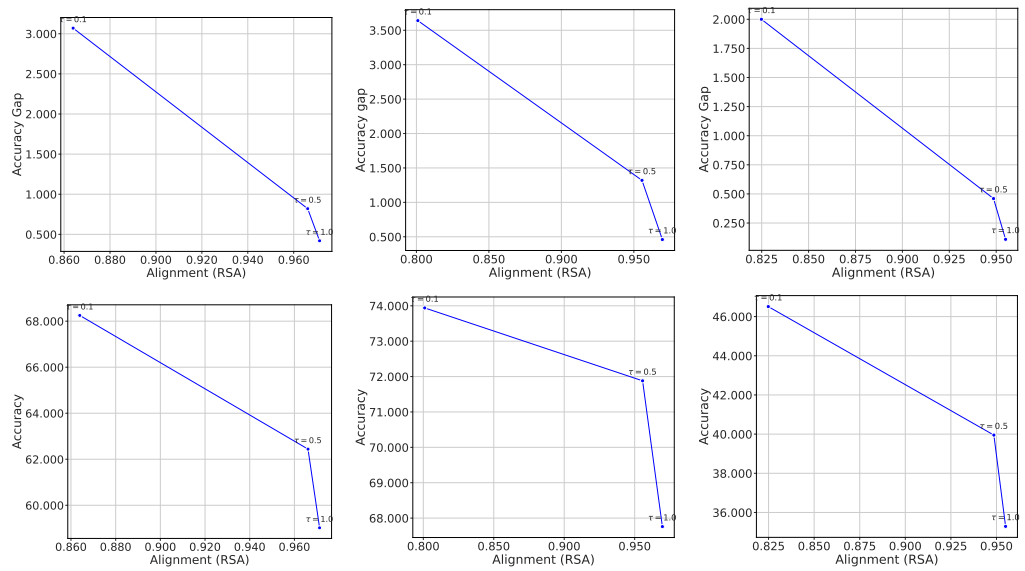
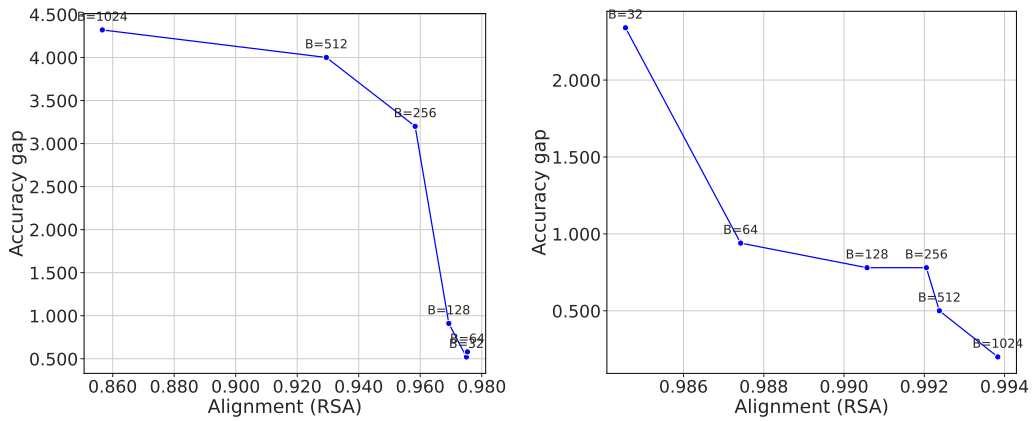


Figure 12: **Performance vs. alignment over varying temperatures. (Top)** The gap between the linear probe accuracies of CL and NSCL ResNet-50 models (trained for 300 epochs) decreases as their alignment increases with higher temperature (τ) values. **(Bottom)** Although the accuracy gap between CL and NSCL models is correlated with the alignment of their representations, higher alignment does not necessarily imply better downstream performance, as performance remains sensitive to the choice of hyperparameters.

Let $A \in R^{H \times N \times N}$ denote the attention weights, where H is the number of heads and N is the number of tokens. We first average weights across all attention heads. We then extract the row corresponding to **[CLS]** token, specifically focusing on its attention to $N - 1$ image patch tokens. This vector is reshaped into a 2D grid (14×14 for ViT-Base) to match the spatial arrangement of image patches. Finally, we upscale the low-resolution grid to original image resolution, normalize it to the range $[0, 1]$, and overlay on the input image.

Analysis. To quantify the structural similarity between representations of ViT models trained with decoupled CL and supervised objectives, we calculate the cosine similarity between their attention maps. As shown in Fig. 15, we track this metric across training epochs and show that NSCL consistently maintains the highest alignment with DCL compared to NSCL and CE. To strengthen our argument, we further visually illustrate this alignment in Fig. 16. The qualitative analysis align



(a) $\mathcal{O}(B)$ **(b)** $\mathcal{O}(B^{1/4})$

Figure 13: **Performance vs. alignment over varying batch sizes.** The gap between the linear probe accuracies of CL and NSCL ResNet-50 models (trained for 300 epochs) varies systematically with the batch size (B) and their RSA alignment: when training with $\eta_t = \mathcal{O}(B)$, larger batch sizes tend to reduce alignment and increase the accuracy gap, whereas with $\eta_t = \mathcal{O}(B^{1/4})$ larger batch sizes tend to increase alignment and reduce the gap.

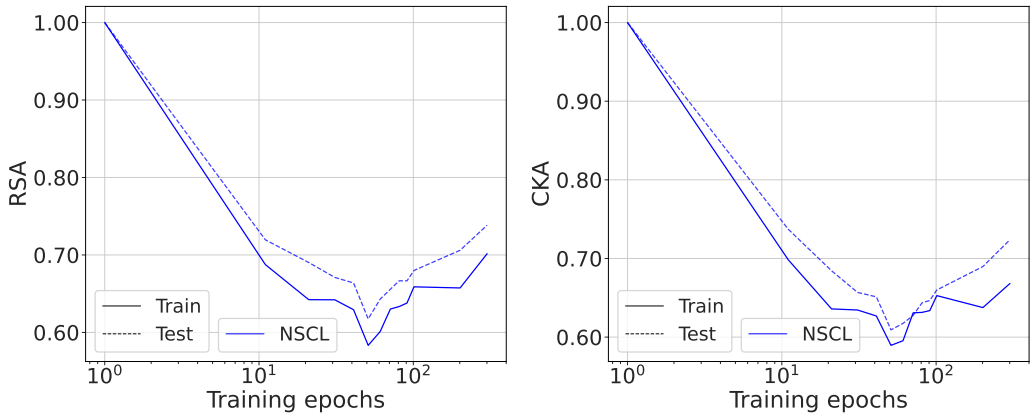


Figure 14: **CL-NSCL alignment for class-imbalanced data.** We train ResNet-50 models on SVHN (Netzer et al., 2011) with decoupled CL and NSCL objectives to analyse alignment when the classes are not uniformly distributed. The RSA and CKA values are comparable to class-balanced datasets (shown in Fig. 2- 7).

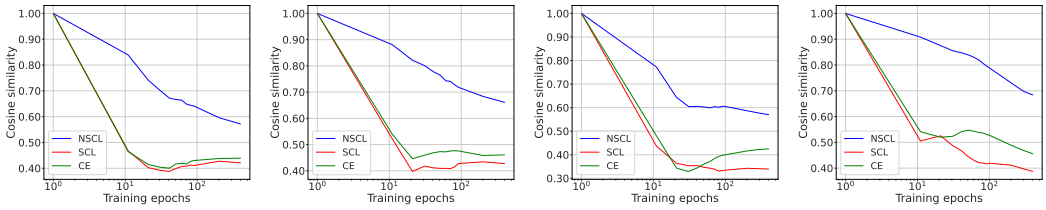
with cosine similarity trends, confirming that NSCL preserves the spatial attention structure of CL more faithfully than other supervised methods.

B.6 Fig. 1 Methodology

We explain how to generate the plots comparing alignment in weight-space and representation-space. The two plots on the left visualize the direction of learning for each model. Each vector represents the change in model's state from initialization (epoch 0) to epoch 1000.

Model states. We consider CL and NSCL models trained on CIFAR100, corresponding to epoch 0 and epoch 1000-a total of four models.

Weight space. This plot shows how the raw parameters evolve during training. For all four models, we first flatten all the weights into a massive vector which gives us four points in a very high dimensional space (order of 10^7). To visualize these points, we perform Principal Component Analysis (PCA) on all four vectors combined and fit them to a 3D space. This creates a shared 3D



CIFAR-10 CIFAR-100 Mini-ImageNet Tiny-ImageNet

Figure 15: **Alignment in attention maps.** We evaluate cosine similarity between attention maps of decoupled CL and supervised models, and observe similar trends as for RSA/CKA values. NSCL remains the most aligned hinting at a deeper structural similarity between representations of CL and NSCL models.

coordinate system. We transform all four points into this space and we get $p^0_{\text{CL}}, p^{1000}_{\text{CL}}, p^0_{\text{NSCL}}, p^{1000}_{\text{NSCL}}$. Using these points, we create two vectors: $(v_{\text{CL}}, v_{\text{NSCL}})$, and create polar plot using the final vectors and the calculated angle between them ($85.7°$).

Representation space. This plot shows how model's alignment for a specific class evolved. We pick one class from our dataset (CIFAR100) and randomly sample 100 images. We use the same samples for all four models to extract their corresponding features, say $\text{Z} \in \mathbb{R}^{100\times d}$, where d is the projection dimension. We concatenate total 400 representations (100 from each model) and perform PCA to learn a shared 3D coordinate system. The representations are transformed to this shared space ($\mathbb{R}^{100\times d} \to \mathbb{R}^{100\times 3}$) and averaged to a single 3D point for each model. Just like before, a polar plot is created using the vectors and angle between them ($27.8°$).

Similarity metrics. We report RSA and CKA values computed between DCL and NSCL models trained on CIFAR100. Additionally, we show their average weight gap as detailed in Sec. 5.1. It is evident that models stay aligned in representation space but diverge in weight space.

B.7 Model Merging

In addition to our main analysis, we also conduct a simple experiment that merges models directly in representation space. Specifically, we interpolate between the learned embeddings of a CL encoder trained on the full dataset and an NSCL encoder trained on only 30% of the dataset. This merged representation already surpasses both the full-data CL model and the small-data NSCL model, reinforcing that NSCL and CL remain geometrically compatible in practice.

Concretely, given an input x, let $f_{\text{CL}}(x)$ and $f_{\text{NSCL}_{30}}(x)$ denote the representations from the CL encoder and the NSCL encoder trained on 30% of the dataset, respectively. We merge them via simple linear interpolation:

$$f_{\texttt{merged}}(x) = \alpha f_{\text{CL}}(x) + (1-\alpha) f_{\text{NSCL}_{30}}(x).$$

We then perform NCCC and LP evaluations using the same 30% subset from the training split and report accuracy on the full mini-ImageNet test split in Fig. 17.

As shown in the figure, for all values of α the merged model outperforms the NSCL baseline, and for $\alpha \in [0.7, 1)$ it also outperforms the CL baseline on the mini-ImageNet downstream classification task. This suggests that the CL and NSCL representations are well aligned, making it possible to effectively merge them directly in representation space.

C Parameter-Space Coupling

To complement the analysis in Sec. 4, we compare the two trajectories in parameter space. Let $e_t = \|w^{\text{CL}}_t - w^{\text{NSCL}}_t\|$ denote the parameter drift at step t. We would like to bound it as a function of the number of training iterations, batch size, and learning rate scheduling. We use classic techniques that can be found at (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Mou et al., 2018; Kuzborskij & Lampert, 2017).

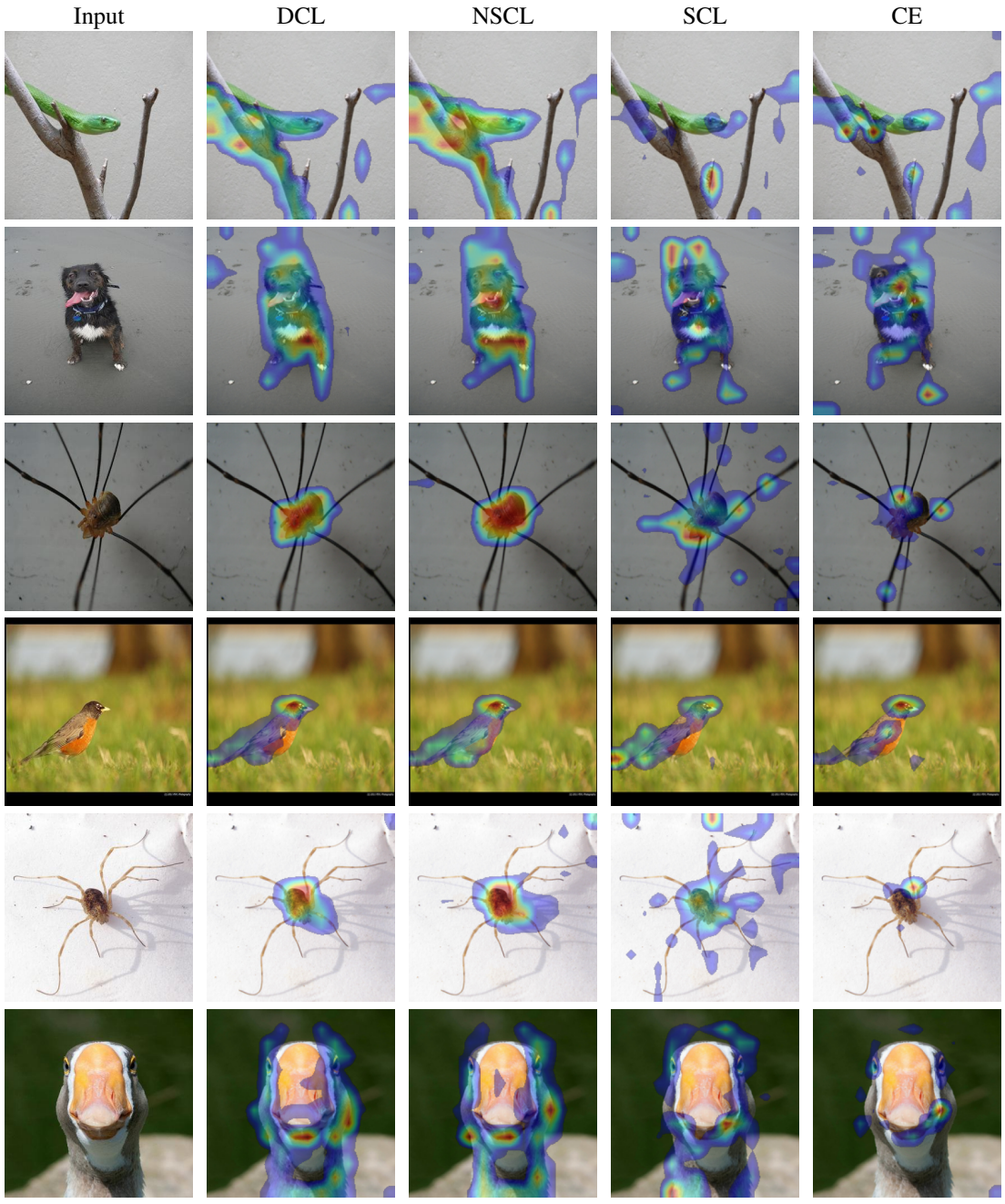


Figure 16: **Visualization of attention maps.** We visualize the self-attention of the [CLS] token from the last layer of the frozen ViT encoder. Beyond a high cosine similarity between attention maps, these visualizations reveal strong structural similarity between CL and NSCL.

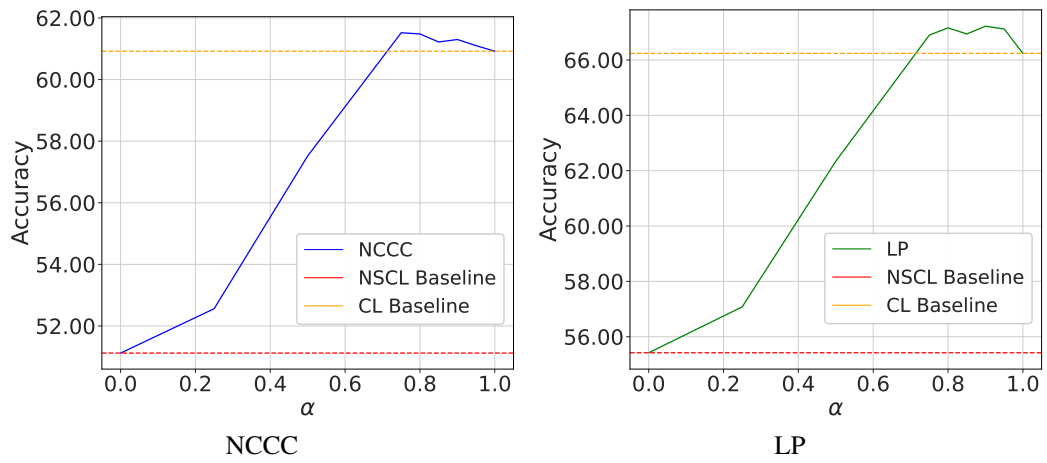


Figure 17: **Model merging in representation space:** We report NCCC and LP scores on mini-ImageNet using CL encoder trained on full dataset and NSCL encoder trained on 30% of the dataset. The performance gains obtained using merged representations illustrate the compatibility of CL and NSCL models and further support our main finding that CL and NSCL maintain closely aligned embedding geometries throughout training.

Optimization. In order to isolate the effect of the loss, we optimize both objectives (CL and NSCL) with standard mini-batch SGD under a single coupled protocol: at step t we draw a batch $\mathcal{B}_t = \{(x_j, x'_j, y_j)\}_{j=1}^{B}$ with replacement, where each $x'_j \sim \alpha(x_j)$ (e.g., random crop/resize, horizontal flip, color jitter, Gaussian blur); we average per-anchor terms to form either $\bar{\ell}^{\mathrm{CL}}_{\mathcal{B}_t}(w)$ or $\bar{\ell}^{\mathrm{NS}}_{\mathcal{B}_t}(w)$ using cosine similarity (optionally temperature-scaled), hence bounded in $[-1, 1]$; and we update $w_{t+1} = w_t - \eta_t \nabla \bar{\ell}_{\mathcal{B}_t}(w_t)$ with prescribed $\eta_t > 0$. We then run two coupled SGD trajectories from the same initialization $w_0^{\mathrm{CL}} = w_0^{\mathrm{NSCL}}$ that share the *same* batches and augmentations $(\mathcal{B}_t)_{t=0}^{T-1}$ and differ only by NSCL's exclusion of same-class negatives:

$$w_{t+1}^{\mathrm{CL}} = w_t^{\mathrm{CL}} - \eta_t \, \nabla \bar{\ell}^{\mathrm{CL}}_{\mathcal{B}_t}\big(w_t^{\mathrm{CL}}\big), \qquad w_{t+1}^{\mathrm{NSCL}} = w_t^{\mathrm{NSCL}} - \eta_t \, \nabla \bar{\ell}^{\mathrm{NSCL}}_{\mathcal{B}_t}\big(w_t^{\mathrm{NSCL}}\big), \quad t = 0, \ldots, T-1.$$

Throughout the analysis, we make standard assumptions on the smoothness of the loss functions and the scale of gradients.

Assumptions. To control the dynamics, we impose two standard conditions on the geometry of the batch objectives and the scale of pairwise gradients.

Assumption 1 (Uniform smoothness)**.** *For every batch $\mathcal{B}$, the functions $w \mapsto \bar{\ell}^{\mathrm{CL}}_{\mathcal{B}}(w)$ and $w \mapsto \bar{\ell}^{\mathrm{NSCL}}_{\mathcal{B}}(w)$ are β-smooth with the same constant $\beta > 0$:*

$$\|\nabla\phi(w) - \nabla\phi(v)\| \;\le\; \beta\,\|w - v\| \quad \textit{for all } v, w \in \mathbb{R}^p,\ \phi \in \{\bar{\ell}^{\mathrm{CL}}_{\mathcal{B}}, \bar{\ell}^{\mathrm{NSCL}}_{\mathcal{B}}\}.$$

Assumption 2 (Bounded pairwise gradients)**.** *There exists $G > 0$, independent of $\mathcal{B}$ and t, such that for all w and all pairs (u, v) appearing in any denominator term,*

$$\|\nabla_w \,\mathrm{sim}\,(f_w(u), f_w(v))\| \;\le\; G.$$

We quantify drift between the coupled trajectories under shared randomness in the *nonconvex β-smooth* regime. Throughout, the only data-dependent term is $\Delta_{\pi,\delta}(B;\tau)$, which decreases with more classes and larger batches.

Theorem 2. *Fix $B, T \in \mathbb{N}$, $\delta \in (0,1)$, and temperature $\tau > 0$. Suppose Assumptions 1–2 hold. Then, with probability at least $1 - \delta$,*

$$e_T \;\le\; \frac{G}{\beta\,\tau}\,\Delta_{\pi,\delta}(B;\tau)\,\Big(\exp\Big(\beta \sum_{t=0}^{T-1} \eta_t\Big) - 1\Big).$$

The bound scales linearly with G and $\Delta_{\pi,\delta}(B;\tau)$, but crucially it is amplified by the exponential factor $\exp(\beta \sum_t \eta_t)$. Unless the step sizes are aggressively annealed, this term grows rapidly with

training time. Even though $\Delta_{\pi,\delta}(B;\tau)$ improves with C and B (e.g., for C=1000, B=512, δ=0.01, we obtain $\Delta_{\pi,\delta}(B;\tau) \approx 0.01$ so that the reweightings of the steps differ by about one percent), the exponential accumulation can still overwhelm this small per-step gap.

In other words, parameter-space coupling guarantees only that the two runs do not drift apart too quickly in weight space. But because the weights may follow very different trajectories even when representations remain similar, this control is too weak to yield meaningful statements about representational alignment. This motivates our next step: shifting the analysis to *similarity space*, where we can obtain bounds that remain stable throughout training and translate directly into guarantees on metrics such as CKA and RSA.

Proof idea. With high probability over batches (Cor. 3), every anchor's denominator is dominated by negatives up to $\epsilon_{B,\delta}$ fluctuations. This keeps the (temperature–τ) softmax reweighting gap between CL and NSCL small. In particular, Lem. 7 shows that the per-batch parameter gradients differ uniformly as

$$\left\|\nabla\bar{\ell}^{\mathrm{CL}}_{\mathcal{B}_t}(w) - \nabla\bar{\ell}^{\mathrm{NSCL}}_{\mathcal{B}_t}(w)\right\| \;\le\; \frac{G}{\tau}\,\Delta_{\pi,\delta}(B;\tau).$$

By β-smoothness of each batch loss, each step can expand distances by at most a factor $(1+\beta\eta_t)$. Combining this smoothness expansion with the uniform gradient-gap bound yields the following recurrence:

$$e_{t+1} \;\le\; (1+\beta\eta_t)\,e_t \;+\; \eta_t\,\frac{G}{\tau}\,\Delta_{\pi,\delta}(B;\tau),$$

where the first term propagates the previous error (with amplification controlled by curvature), and the second injects the new discrepancy introduced by the CL–NSCL gap at temperature τ.

Unrolling over T steps and applying the discrete Grönwall inequality gives the exponential-type bound

$$e_T \;\le\; \frac{G}{\beta\,\tau}\,\Delta_{\pi,\delta}(B;\tau)\left(\exp\Big(\beta\sum_{t=0}^{T-1}\eta_t\Big) - 1\right).$$

Thus, cumulative drift scales with the reweighting gap and is amplified exponentially with the total step size; smaller τ tightens the softmax and increases the constants (via both $1/\tau$ and $\mathrm{e}^{2/\tau}$ inside $\Delta_{\pi,\delta}$), so keeping $\sum_t \eta_t$ moderate is especially important.

D Why gradient descent in similarity space is a faithful surrogate

We now explain why running gradient descent directly in similarity space closely tracks the dynamics induced by gradient descent in parameter space.

When parameters move from w_t to w_{t+1}, the induced change in the similarity matrix can be approximated by a linear expansion:

$$\Sigma(w_{t+1}) - \Sigma(w_t) \;\approx\; J_t(w_{t+1}-w_t), \qquad J_t \;:=\; J(w_t), \tag{3}$$

where $J(w) := \partial\Sigma/\partial w$ is the Jacobian. The error in this expansion, denoted R_t, is quadratic in the step size:

$$\Sigma(w_{t+1}) - \Sigma(w_t) \;=\; J_t(w_{t+1}-w_t) + R_t. \tag{4}$$

By the chain rule, the gradient in parameter space can be written as follows:

$$\nabla_w\bar{\ell}(w_t) \;=\; J_t^\top\nabla_\Sigma\bar{\ell}(\Sigma(w_t)) \;=\; J_t^\top\widehat{G}_t,$$

where $\widehat{G}_t := \nabla_\Sigma\bar{\ell}(\Sigma(w_t))$. Substituting this into the update rule gives

$$\Sigma(w_{t+1}) - \Sigma(w_t) \;=\; -\eta_t\,P_t\widehat{G}_t + R_t, \qquad P_t \;:=\; J_tJ_t^\top \;\succeq\; 0. \tag{5}$$

Thus, parameter descent acts like similarity descent, but with a preconditioning matrix P_t, plus the remainder R_t.

Assume there exist constants $L_\Sigma, M_\Sigma > 0$ such that

$$\|J(w)\|_{2\to2} \;\le\; L_\Sigma, \qquad \|\Sigma(w+\Delta w) - \Sigma(w) - J(w)\Delta w\|_F \;\le\; \frac{M_\Sigma}{2}\|\Delta w\|_2^2.$$

Then $\|P_t\|_{2\to 2} \le L_\Sigma^2$ and, with $\Delta w_t := -\eta_t \nabla_w \bar{\ell}(w_t)$,

$$\|R_t\|_F \;\le\; \frac{M_\Sigma}{2}\eta_t^2 \|\nabla_w \bar{\ell}(w_t)\|_2^2 \;=:\; \frac{M_\Sigma}{2}\eta_t^2 \Xi_t. \tag{6}$$

Let $\widehat{\Sigma}_t := \Sigma(w_t)$ be the similarity trajectory induced by parameter descent. Define $\widetilde{\Sigma}_t$ as the trajectory of explicit similarity descent:

$$\widetilde{\Sigma}_{t+1} \;=\; \widetilde{\Sigma}_t - \eta_t \widetilde{G}_t, \qquad \widetilde{G}_t \;:=\; \nabla_\Sigma \bar{\ell}(\widetilde{\Sigma}_t),$$

with $\widehat{\Sigma}_0 = \widetilde{\Sigma}_0$. Let $E_t := \|\widehat{\Sigma}_t - \widetilde{\Sigma}_t\|_F$ and $C_\Sigma := \sup_t \|P_t - I\|_{2\to 2} \le L_\Sigma^2 + 1$. Using equation 5, adding and subtracting $-\eta_t \widehat{G}_t$, and applying the temperature–τ bounds equation 11 and equation 6, one obtains

$$E_{t+1} \;\le\; \Big(1 + \frac{\eta_t}{2\tau^2 B}\Big) E_t \;+\; \eta_t C_\Sigma \|\widehat{G}_t\|_F \;+\; \frac{M_\Sigma}{2}\eta_t^2 \Xi_t. \tag{7}$$

Unrolling this recursion from $E_0 = 0$ and using $\prod_u (1+\alpha_u) \le \exp(\sum_u \alpha_u)$ yields

$$\|\widehat{\Sigma}_T - \widetilde{\Sigma}_T\|_F \;\le\; \exp\Big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\Big)\left[C_\Sigma \sum_{t=0}^{T-1}\eta_t \|\widehat{G}_t\|_F \;+\; \frac{M_\Sigma}{2}\sum_{t=0}^{T-1}\eta_t^2 \Xi_t\right]. \tag{8}$$

By bounding $\|\widehat{G}_t\|_F$ via equation 13, namely $\|\widehat{G}_t\|_F \le \frac{1}{\tau}\sqrt{\frac{2}{B}}$, this simplifies to

$$\|\widehat{\Sigma}_T - \widetilde{\Sigma}_T\|_F \;\le\; \exp\Big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\Big)\left[\frac{\sqrt{2}\,C_\Sigma}{\tau\sqrt{B}}\sum_{t=0}^{T-1}\eta_t \;+\; \frac{M_\Sigma}{2}\sum_{t=0}^{T-1}\eta_t^2 \Xi_t\right]. \tag{9}$$

To understand when this bound is conceptually reasonable, suppose $\|\nabla_w \bar{\ell}(w_t)\|_2 \le G$ for all t, so that $\Xi_t \le G^2$. The right-hand side of equation 9 is then controlled by two quantities: the cumulative step size $\sum_t \eta_t$, which appears both inside the exponential and in the linear prefactor $(\sqrt{2}\,C_\Sigma/(\tau\sqrt{B}))\sum_t \eta_t$, and the term $\sum_t \eta_t^2$.

A simple sufficient regime is to assume that $\sum_{t=0}^{T-1}\eta_t \le c_1 \tau^2 B$ and $\sum_{t=0}^{T-1}\eta_t^2 \le c_2$ for fixed constants c_1, c_2 independent of T. Under these conditions, the exponential factor is bounded by $\exp\big((1/(2\tau^2 B))\sum_t \eta_t\big) \le \exp(c_1/2)$, the linear prefactor by $(\sqrt{2}\,C_\Sigma/(\tau\sqrt{B}))\sum_t \eta_t \le \sqrt{2}\,C_\Sigma\, c_1\, \tau\sqrt{B}$ (a fixed constant for given (τ, B) and moderate c_1), and the quadratic remainder by $(M_\Sigma/2)\sum_t \eta_t^2 \Xi_t \le (M_\Sigma/2) G^2 c_2$. In particular, when $\sum_t \eta_t/(\tau^2 B)$ and $\sum_t \eta_t^2$ are both bounded by constants independent of T, the bound guarantees that $\|\widehat{\Sigma}_T - \widetilde{\Sigma}_T\|_F$ remains controlled (and small whenever C_Σ, M_Σ, G are moderate).

To summarize, the similarity and parameter trajectories stay close whenever the normalized cumulative step size $\sum_t \eta_t/(\tau^2 B)$ is bounded and the learning-rate schedule is sufficiently decaying so that $\sum_t \eta_t^2$ remains bounded. For a fixed learning-rate schedule, a large batch size B and moderate temperature τ act as stabilizing factors via the $1/(\tau\sqrt{B})$ dependence in equation 9, while very small τ or extremely large, non-decaying step sizes can make the coupling poor, as reflected by the bound.

E Technical Tools and Proofs

E.1 Notation and basic softmax facts

Let $S = \{(x_i, y_i)\}_{i=1}^N$ be dataset with C classes (each class c has n_c points, with $\sum_{c=1}^C n_c = N$, and we do not assume the n_c are equal). For parameters w, let $z_i = f_w(x_i)$ and define the bounded similarity matrix

$$\Sigma(w)_{ij} \;:=\; \mathrm{sim}\big(z_i, z_j\big) \in [-1, 1].$$

At step t, draw a mini-batch $\mathcal{B}_t = \{(x_{j_s}, x'_{j_s}, y_{j_s})\}_{s=1}^B$ with replacement, using independent augmentations $x'_{j_s} \sim \alpha(x_{j_s})$. For an *anchor* $i \in \{j_1, \dots, j_B\}$, let D_i be its denominator index set, and let $D_i^{\mathrm{neg}} := \{k \in D_i :\; y_k \ne y_i\}$ (and similarly D_i^{pos}) denote the subset restricted to negatives (e.g., in two-view SimCLR, D_i consists of all $2B$ views except the anchor itself).

Define the anchor's logit vector $s_i(w) := \big(\Sigma(w)_{i,k}\big)_{k\in D_i}$ and the corresponding softmax distributions with temperature $\tau > 0$ (default 1):

$$p_i = \mathrm{softmax}\big(s_i(w)/\tau\big), \qquad q_i = \mathrm{softmax}\big((s_i(w))_{D_i^{\mathrm{neg}}}/\tau\big).$$

Let i' denote the positive (augmented) index for anchor i.

For contrastive learning (CL) and negatives-only supervised contrastive learning (NSCL), the per-anchor and batch losses are

$$\ell_i^{\mathrm{CL}}(s_i) \;=\; -\log p_{i,i'}, \qquad \ell_i^{\mathrm{NSCL}}(s_i) \;=\; -\log q_{i,i'},$$

$$\bar{\ell}_{\mathcal{B}_t}^{\mathrm{CL}} \;=\; \frac{1}{B}\sum_{i\in\{j_1,\dots,j_B\}} \ell_i^{\mathrm{CL}}(s_i), \qquad \bar{\ell}_{\mathcal{B}_t}^{\mathrm{NSCL}} \;=\; \frac{1}{B}\sum_{i\in\{j_1,\dots,j_B\}} \ell_i^{\mathrm{NSCL}}(s_i).$$

Since $\Sigma(w)_{ij} \in [-1, 1]$, each exponential term inside the softmax lies in

$$\exp\big(\Sigma(w)_{ij}/\tau\big) \in [\mathrm{e}^{-1/\tau}, \mathrm{e}^{1/\tau}],$$

a fact used below to control softmax mass ratios.

Lemma 1 (Anchor-block orthogonality)**.** *Fix a step t and batch $\mathcal{B}_t$. For each anchor $i \in \mathcal{B}_t$, let D_i be the set of indices appearing in i's denominator and define the per-anchor gradient $g_i \in \mathbb{R}^{\mathcal{I}_t}$ by*

$$g_i \;:=\; \nabla_{s_i}\ell_i \quad \textit{placed on the coordinates } \{(i,k) : k \in D_i\} \subset \mathcal{I}_t,$$

with zeros elsewhere (here $\mathcal{I}_t$ is the set of all coordinates touched at step t). If $i \neq j$, then g_i and g_j have disjoint supports, and hence

$$\langle g_i,\, g_j\rangle_F \;=\; 0.$$

Consequently, for the batch gradient $G = \frac{1}{B}\sum_{i\in\mathcal{B}_t} g_i$,

$$\|G\|_F^2 \;=\; \frac{1}{B^2}\sum_{i\in\mathcal{B}_t} \|g_i\|_F^2. \tag{10}$$

Proof. By construction, g_i is supported only on coordinates $\{(i,k) : k \in D_i\}$, while g_j is supported only on $\{(j,k) : k \in D_j\}$. For $i \neq j$ these sets are disjoint, so every coordinatewise product is zero, yielding $\langle g_i, g_j\rangle_F = 0$. Expanding the square for G,

$$\|G\|_F^2 \;=\; \Big\langle \frac{1}{B}\sum_i g_i,\; \frac{1}{B}\sum_j g_j \Big\rangle_F \;=\; \frac{1}{B^2}\sum_i \|g_i\|_F^2 + \frac{1}{B^2}\sum_{i\neq j}\langle g_i, g_j\rangle_F \;=\; \frac{1}{B^2}\sum_i \|g_i\|_F^2,$$

where the cross terms vanish by orthogonality. □

Lemma 2 (Softmax Hessian and gradient Lipschitzness)**.** *Fix a step t and batch $\mathcal{B}_t$. Let $\mathcal{I}_t$ be the set of coordinates (i,k) that appear in any anchor's denominator at step t, and view $\bar{\ell}_{\mathcal{B}_t}$ (either CL or NSCL) as a function of the restricted similarity entries $\Sigma \in \mathbb{R}^{\mathcal{I}_t}$. For each anchor i, write $s_i = \{\Sigma(i,k) : (i,k) \in \mathcal{I}_t\}$ and $p_i = \mathrm{softmax}(s_i/\tau)$. Then:*

$$\nabla_{s_i}^2\ell_i(s_i) \;=\; \frac{1}{\tau^2}\,J(s_i), \quad J(s_i) := \mathrm{Diag}(p_i) - p_i p_i^\top, \qquad \big\|\nabla^2\bar{\ell}_{\mathcal{B}_t}(\Sigma)\big\|_{2\to 2} \;\le\; \frac{1}{2\tau^2 B}.$$

Consequently, for all $\Sigma, \widetilde{\Sigma} \in \mathbb{R}^{\mathcal{I}_t}$,

$$\big\|\nabla_\Sigma\bar{\ell}_{\mathcal{B}_t}(\Sigma) - \nabla_\Sigma\bar{\ell}_{\mathcal{B}_t}(\widetilde{\Sigma})\big\|_F \;\le\; \frac{1}{2\tau^2 B}\,\|\Sigma - \widetilde{\Sigma}\|_F. \tag{11}$$

Proof. With temperature $\tau > 0$, for an anchor i we have $p_i = \mathrm{softmax}(s_i/\tau)$ and

$$\nabla_{s_i}\ell_i(s_i) \;=\; \frac{1}{\tau}\,(p_i - e_{i'}) \quad\Longrightarrow\quad \nabla_{s_i}^2\ell_i(s_i) \;=\; \frac{1}{\tau^2}\,\nabla_{s_i}p_i \;=\; \frac{1}{\tau^2}\,J(s_i),$$

where $J(s_i) := \mathrm{Diag}(p_i) - p_i p_i^\top$. Bound $\|J(s_i)\|_{2\to 2}$ via the infinity norm:

$$\begin{aligned}\|J(s_i)\|_{2\to 2} &\le \|J(s_i)\|_\infty \\ &= \max_r \sum_\ell |J_{r\ell}| \\ &= \max_r \Big(p_{i,r}(1-p_{i,r}) + \sum_{\ell \neq r} p_{i,r} p_{i,\ell} \Big) \\ &= \max_r 2p_{i,r}(1-p_{i,r}) \le \tfrac{1}{2},\end{aligned}$$

since $x(1-x) \le 1/4$ for $x \in [0,1]$.

The batch loss is an average over anchors, so its Hessian is block-diagonal across anchors with a prefactor $1/B$:

$$\nabla^2 \bar{\ell}_{\mathcal{B}_t}(\Sigma) = \frac{1}{B} \,\mathrm{blkdiag}\Big(\tfrac{1}{\tau^2} J(s_i) \Big)_{i \in \mathcal{B}_t} = \frac{1}{\tau^2 B} \,\mathrm{blkdiag}\big(J(s_i) \big)_{i \in \mathcal{B}_t}.$$

Hence

$$\big\| \nabla^2 \bar{\ell}_{\mathcal{B}_t}(\Sigma) \big\|_{2\to 2} = \frac{1}{\tau^2 B} \max_i \|J(s_i)\|_{2\to 2} \le \frac{1}{2\tau^2 B}.$$

By the mean-value (integral) form for vector fields,

$$\nabla_\Sigma \bar{\ell}_{\mathcal{B}_t}(\Sigma) - \nabla_\Sigma \bar{\ell}_{\mathcal{B}_t}(\widetilde{\Sigma}) = \int_0^1 \nabla^2 \bar{\ell}_{\mathcal{B}_t}\big(\widetilde{\Sigma} + \theta(\Sigma - \widetilde{\Sigma}) \big) \, [\Sigma - \widetilde{\Sigma}] \, d\theta,$$

and therefore

$$\big\| \nabla_\Sigma \bar{\ell}_{\mathcal{B}_t}(\Sigma) - \nabla_\Sigma \bar{\ell}_{\mathcal{B}_t}(\widetilde{\Sigma}) \big\|_F \le \sup_{\theta \in [0,1]} \big\| \nabla^2 \bar{\ell}_{\mathcal{B}_t}(\Sigma_\theta) \big\|_{2\to 2} \, \|\Sigma - \widetilde{\Sigma}\|_F \le \frac{1}{2\tau^2 B} \, \|\Sigma - \widetilde{\Sigma}\|_F,$$

as claimed. □

Lemma 3 (Per-anchor gradient norm and batch average)**.** *For an anchor i, let s_i be the vector of logits in its denominator and $p_i = \mathrm{softmax}(s_i/\tau)$. Let i' denote the (unique) positive index (for NSCL, if i' is not in the denominator, set $p_{i,i'} := 0$ in the display below). Then*

$$\|\nabla_{s_i} \ell_i\|_2^2 = \frac{1}{\tau^2} \Big[(1 - p_{i,i'})^2 + \sum_{k \neq i'} p_{i,k}^2 \Big] \le \frac{2}{\tau^2}, \tag{12}$$

hence $\|\nabla_{s_i} \ell_i\|_2 \le \sqrt{2}/\tau$. Moreover, by block orthogonality across anchors,

$$\Big\| \frac{1}{B} \sum_{i \in \mathcal{B}_t} \nabla_{s_i} \ell_i \Big\|_F^2 = \frac{1}{B^2} \sum_{i \in \mathcal{B}_t} \|\nabla_{s_i} \ell_i\|_2^2 \le \frac{2}{\tau^2 B} \quad \Longrightarrow \quad \Big\| \frac{1}{B} \sum_{i \in \mathcal{B}_t} \nabla_{s_i} \ell_i \Big\|_F \le \frac{1}{\tau} \sqrt{\frac{2}{B}}. \tag{13}$$

Proof. For CL, the loss is $-\log p_{i,i'}$ with $p_i = \mathrm{softmax}(s_i/\tau)$. By the standard softmax–cross-entropy derivative with temperature,

$$\nabla_{s_i} \ell_i = \frac{1}{\tau} (p_i - e_{i'}),$$

so

$$\|\nabla_{s_i} \ell_i\|_2^2 = \frac{1}{\tau^2} \left[(1 - p_{i,i'})^2 + \sum_{k \neq i'} p_{i,k}^2 \right] \le \frac{1}{\tau^2} \left[(1 - p_{i,i'})^2 + \Big(\sum_{k \neq i'} p_{i,k} \Big)^2 \right] = \frac{2}{\tau^2} (1 - p_{i,i'})^2 \le \frac{2}{\tau^2},$$

since p_i is a probability vector and $\sum_{k \neq i'} p_{i,k} = 1 - p_{i,i'}$.

For NSCL, two cases. If $i' \in D_i$, the same computation applies (the target index is present), hence the same bound holds. If $i' \notin D_i$ (negatives-only denominator), then the loss is $-\log q_{i,i'}$ with $q_i = \mathrm{softmax}\big((s_i)_{D_i^-} / \tau \big)$ supported only on D_i^-, and

$$\nabla_{s_i} \ell_i = \frac{1}{\tau} \, q_i \quad \text{on } D_i^{\mathrm{neg}} \quad (\text{and } 0 \text{ on } D_i^{\mathrm{pos}}),$$

so

$$\|\nabla_{s_i}\ell_i\|_2^2 \;=\; \frac{1}{\tau^2}\sum_{j\in D_i^-} q_{i,j}^2 \;\le\; \frac{1}{\tau^2}\Big(\sum_{j\in D_i^-} q_{i,j}\Big)^2 \;=\; \frac{1}{\tau^2} \;\le\; \frac{2}{\tau^2}.$$

Thus in all cases $\|\nabla_{s_i}\ell_i\|_2 \le \sqrt{2}/\tau$, establishing equation 12.

For the batch bound equation 13, gradients from different anchors have disjoint supports over coordinates $\{(i,k) : k \in D_i\}$, so they are orthogonal in Frobenius inner product (Lem. 1). Therefore,

$$\Big\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla_{s_i}\ell_i\Big\|_F^2 = \frac{1}{B^2}\sum_{i\in\mathcal{B}_t}\|\nabla_{s_i}\ell_i\|_2^2 \le \frac{1}{B^2}\cdot B\cdot\frac{2}{\tau^2} = \frac{2}{\tau^2 B},$$

which also implies $\big\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla_{s_i}\ell_i\big\|_F \le \frac{1}{\tau}\sqrt{2/B}$. □

Lemma 4 (Bounded logits imply bounded softmax masses)**.** *Fix a step t and an anchor i. Suppose all active logits satisfy $\Sigma(i,k) \in [-1,1]$. For any index subset S in the anchor's denominator, define*

$$Z_S \;:=\; \sum_{k\in S}\exp\big(\Sigma(i,k)/\tau\big) \quad \textit{with temperature } \tau > 0.$$

Then

$$|S|\,\mathrm{e}^{-1/\tau} \;\le\; Z_S \;\le\; |S|\,\mathrm{e}^{1/\tau}.$$

In particular, if S_{pos} and S_{neg} are the positive and negative index sets with sizes n_{pos} and n_{neg}, and $Z_{\mathrm{pos}} := Z_{S_{\mathrm{pos}}}$, $Z_{\mathrm{neg}} := Z_{S_{\mathrm{neg}}}$, then

$$n_{\mathrm{pos}}\,\mathrm{e}^{-1/\tau} \;\le\; Z_{\mathrm{pos}} \;\le\; n_{\mathrm{pos}}\,\mathrm{e}^{1/\tau}, \qquad n_{\mathrm{neg}}\,\mathrm{e}^{-1/\tau} \;\le\; Z_{\mathrm{neg}} \;\le\; n_{\mathrm{neg}}\,\mathrm{e}^{1/\tau},$$

and hence

$$\frac{Z_{\mathrm{pos}}}{Z_{\mathrm{neg}}} \;\le\; \mathrm{e}^{2/\tau}\,\frac{n_{\mathrm{pos}}}{n_{\mathrm{neg}}} \quad \textit{and} \quad \frac{Z_{\mathrm{pos}}}{Z_{\mathrm{neg}}} \;\ge\; \mathrm{e}^{-2/\tau}\,\frac{n_{\mathrm{pos}}}{n_{\mathrm{neg}}}.$$

Proof. Since $\Sigma(i,k) \in [-1,1]$, we have $\exp(\Sigma(i,k)/\tau) \in [\mathrm{e}^{-1/\tau}, \mathrm{e}^{1/\tau}]$ for every active k. Summing over $k \in S$ yields $|S|\,\mathrm{e}^{-1/\tau} \le Z_S \le |S|\,\mathrm{e}^{1/\tau}$. Apply this with $S = S_{\mathrm{pos}}$ and $S = S_{\mathrm{neg}}$ and take ratios to obtain the stated bounds. □

E.2 High-probability batch composition

Fix $T, B \in \mathbb{N}$ and $\epsilon > 0$. For step t and anchor $i \in \mathcal{B}_t$, let $Y_{t,s}^{(i)} = \mathbf{1}\{y_{j_s} \neq y_i\}$ for $s = 1, \dots, B$.

Lemma 5 (Batch-composition event)**.** *For a population with C classes and class priors $\pi_c = n_c/N$, the $Y_{t,s}^{(i)}$ are i.i.d. Bernoulli with mean $1 - \pi_{y_i}$. For any $\epsilon > 0$,*

$$\mathbb{P}\left[\exists(t,i) : \frac{1}{B}\sum_{s=1}^{B} Y_{t,s}^{(i)} \;<\; 1 - \pi_{y_i} - \epsilon\right] \;\le\; TB\,\mathrm{e}^{-2B\epsilon^2}.$$

Equivalently, with probability $\ge 1 - TB\,\mathrm{e}^{-2B\epsilon^2}$, every anchor sees at least $B(1 - \pi_{y_i} - \epsilon)$ negatives.

Proof. Fix any step t and anchor i. Because batches are drawn with replacement from a population with class priors $\pi_c = n_c/N$, for each position $s \in \{1, \dots, B\}$ the indicator $Y_{t,s}^{(i)} = \mathbf{1}\{y_{j_s} \neq y_i\}$ is Bernoulli with mean $\mathbb{E}[Y_{t,s}^{(i)}] = 1 - \pi_{y_i}$, and $\{Y_{t,s}^{(i)}\}_{s=1}^{B}$ are i.i.d. across s. By Hoeffding's inequality, for any $\epsilon > 0$,

$$\mathbb{P}\left[\frac{1}{B}\sum_{s=1}^{B} Y_{t,s}^{(i)} \;<\; 1 - \pi_{y_i} - \epsilon\right] \;=\; \mathbb{P}\left[\frac{1}{B}\sum_{s=1}^{B}\big(Y_{t,s}^{(i)} - \mathbb{E}Y_{t,s}^{(i)}\big) \;<\; -\epsilon\right] \;\le\; \exp(-2B\epsilon^2).$$

There are at most TB anchor–step pairs (t,i) over $t = 0, \dots, T-1$ and $i \in \mathcal{B}_t$. A union bound gives

$$\mathbb{P}\left[\exists(t,i) : \frac{1}{B}\sum_{s=1}^{B} Y_{t,s}^{(i)} \;<\; 1 - \pi_{y_i} - \epsilon\right] \;\le\; TB\,\mathrm{e}^{-2B\epsilon^2}.$$

Equivalently, with probability at least $1 - TB\,\mathrm{e}^{-2B\epsilon^2}$, every anchor in every step has at least $B(1 - \pi_{y_i} - \epsilon)$ negatives in its denominator. □

Corollary 3. *For $\delta \in (0,1)$, set $\epsilon_{B,\delta} := \sqrt{\frac{1}{2B}\log(\frac{TB}{\delta})}$ and let $\pi_c = n_c/N$ be the class priors and $\pi_{\max} := \max_{c\in[C]} \pi_c$. With probability $\geq 1-\delta$, every anchor i has at least $B(1-\pi_{y_i}-\epsilon_{B,\delta})$ negatives and at most $B(\pi_{y_i}+\epsilon_{B,\delta})$ positives in its denominator. In particular,*

$$|D_i^{\text{neg}}| \;\geq\; B\big(1-\pi_{\max}-\epsilon_{B,\delta}\big), \qquad |D_i^{\text{pos}}| \;\leq\; B\big(\pi_{\max}+\epsilon_{B,\delta}\big).$$

Using bounded logits, the ratio of total positive to negative softmax mass (at temperature $\tau > 0$) satisfies, for all anchors and steps,

$$\frac{Z_i^{\text{pos}}}{Z_i^{\text{neg}}} \;\leq\; \frac{\mathrm{e}^{2/\tau}\big(\pi_{\max}+\epsilon_{B,\delta}\big)}{1-\pi_{\max}-\epsilon_{B,\delta}} \;=\; \tfrac{1}{2}\,\Delta_{\pi,\delta}(B;\tau), \tag{14}$$

where

$$\Delta_{\pi,\delta}(B;\tau) \;=\; \frac{2\,\mathrm{e}^{2/\tau}\big(\pi_{\max}+\epsilon_{B,\delta}\big)}{1-\pi_{\max}-\epsilon_{B,\delta}}.$$

Proof. Set $\epsilon = \epsilon_{B,\delta} := \sqrt{\frac{1}{2B}\log(\frac{TB}{\delta})}$ and $\Delta_{\pi,\delta}(B;\tau) := \frac{2\,\mathrm{e}^{2/\tau}\left(\pi_{\max}+\epsilon_{B,\delta}\right)}{1-\pi_{\max}-\epsilon_{B,\delta}}$. Apply Lem. 5 with this ϵ: with probability at least $1-\delta$, for every step t and every anchor i,

$$|D_i^{\text{neg}}| \;\geq\; B\big(1-\pi_{y_i}-\epsilon_{B,\delta}\big), \qquad |D_i^{\text{pos}}| \;\leq\; B\big(\pi_{y_i}+\epsilon_{B,\delta}\big).$$

In particular,

$$|D_i^{\text{neg}}| \;\geq\; B\big(1-\pi_{\max}-\epsilon_{B,\delta}\big), \qquad |D_i^{\text{pos}}| \;\leq\; B\big(\pi_{\max}+\epsilon_{B,\delta}\big).$$

In two-view SimCLR, each sampled point contributes two denominator entries, so the denominator contains at least $2|D_i^{\text{neg}}|$ negative entries and at most $2|D_i^{\text{pos}}|$ positive entries; the factor 2 cancels in the ratio below.

Because similarities are bounded in $[-1,1]$, each logit lies in $[-1,1]$ and hence each exponential term at temperature τ lies in $[\mathrm{e}^{-1/\tau}, \mathrm{e}^{1/\tau}]$. Therefore, for any anchor and step,

$$Z_i^{\text{pos}} \;\leq\; \mathrm{e}^{1/\tau}\cdot(2|D_i^{\text{pos}}|), \qquad Z_i^{\text{neg}} \;\geq\; \mathrm{e}^{-1/\tau}\cdot(2|D_i^{\text{neg}}|),$$

and thus

$$\frac{Z_i^{\text{pos}}}{Z_i^{\text{neg}}} \;\leq\; \mathrm{e}^{2/\tau}\,\frac{|D_i^{\text{pos}}|}{|D_i^{\text{neg}}|} \;\leq\; \frac{\mathrm{e}^{2/\tau}\left(\pi_{\max}+\epsilon_{B,\delta}\right)}{1-\pi_{\max}-\epsilon_{B,\delta}} \;=\; \tfrac{1}{2}\,\Delta_{\pi,\delta}(B;\tau).$$

The bound is meaningful whenever $\epsilon_{B,\delta} < 1-\pi_{\max}$ so that the denominator is positive. This proves the corollary. □

Lemma 6 (Per-anchor reweighting gap)**.** *On the event of Cor. 3, let p be the CL softmax (temperature $\tau > 0$) over an anchor's full denominator, and q the NSCL softmax (same τ) that removes same-class entries and renormalizes over negatives. Then*

$$\|p-q\|_1 \;\leq\; \Delta_{\pi,\delta}(B;\tau), \qquad \|p-q\|_2 \;\leq\; \|p-q\|_1 \;\leq\; \Delta_{\pi,\delta}(B;\tau).$$

Proof. Fix an anchor i and let $D_i^{\text{pos}}, D_i^{\text{neg}}$ be its positive and negative index sets in the CL denominator. Write $s_k := \Sigma(i,k)$ and define

$$Z_i^{\text{pos}} \;:=\; \sum_{k\in D_i^{\text{pos}}} \exp\big(s_k/\tau\big), \qquad Z_i^{\text{neg}} \;:=\; \sum_{j\in D_i^{\text{neg}}} \exp\big(s_j/\tau\big), \qquad \alpha \;:=\; \frac{Z_i^{\text{pos}}}{Z_i^{\text{pos}}+Z_i^{\text{neg}}}.$$

Let p be the CL softmax on $D_i^{\text{pos}} \cup D_i^{\text{neg}}$ and let q be the NSCL softmax that zeros positive entries and renormalizes on negatives: $q(k) = 0$ for $k \in D_i^{\text{pos}}$ and $q(j) = p(j)/(1-\alpha)$ for $j \in D_i^{\text{neg}}$. Then

$$\|p-q\|_1 = \sum_{k\in D_i^{\text{pos}}} p_k + \sum_{j\in D_i^{\text{neg}}} \Big|p_j - \frac{p_j}{1-\alpha}\Big| = \alpha + (1-\alpha)\frac{\alpha}{1-\alpha} = 2\alpha \leq \frac{2Z_i^{\text{pos}}}{Z_i^{\text{neg}}}.$$

On the high-probability event of Cor. 3, since $s \in [-1,1] \Rightarrow \exp(s/\tau) \in [\mathrm{e}^{-1/\tau}, \mathrm{e}^{1/\tau}]$,

$$Z_i^{\text{pos}} \;\leq\; \mathrm{e}^{1/\tau}\,|D_i^{\text{pos}}|, \qquad Z_i^{\text{neg}} \;\geq\; \mathrm{e}^{-1/\tau}\,|D_i^{\text{neg}}|.$$

Moreover, by Cor. 3,

$$|D_i^{\mathrm{pos}}| \;\le\; 2B\Big(\pi_{\max} + \epsilon_{B,\delta}\Big), \qquad |D_i^{\mathrm{neg}}| \;\ge\; 2B\Big(1 - \pi_{\max} - \epsilon_{B,\delta}\Big),$$

(each sampled point contributes two keys, so the factor 2 cancels in the ratio). Hence

$$\frac{2Z_i^{\mathrm{pos}}}{Z_i^{\mathrm{neg}}} \;\le\; 2\,\mathrm{e}^{2/\tau}\,\frac{|D_i^{\mathrm{pos}}|}{|D_i^{\mathrm{neg}}|} \;\le\; \frac{2\,\mathrm{e}^{2/\tau}\big(\pi_{\max} + \epsilon_{B,\delta}\big)}{1 - \pi_{\max} - \epsilon_{B,\delta}} \;=\; \Delta_{\pi,\delta}(B;\tau).$$

Therefore $\|p - q\|_1 \le \Delta_{\pi,\delta}(B;\tau)$. Finally, $\|p - q\|_2 \le \|p - q\|_1$ yields the second claim. □

E.3 Parameter-space coupling: supporting lemmas and proofs

Lemma 7 (Per-batch parameter-gradient gap)**.** *On the event of Cor. 3, for any step t and any w,*

$$\big\|\nabla\bar{\ell}_{\mathcal{B}_t}^{\mathrm{CL}}(w) - \nabla\bar{\ell}_{\mathcal{B}_t}^{\mathrm{NSCL}}(w)\big\| \;\le\; \frac{G}{\tau}\,\Delta_{\pi,\delta}(B;\tau).$$

Proof. Fix t and w. For an anchor $i \in \mathcal{B}_t$, let D_i be its denominator index set, split as $D_i = \mathrm{pos}_i \cup \mathrm{neg}_i$, where pos_i collects all same-class indices (including the designated positive i') and neg_i the rest. Write the logits $s_{ik} = \Sigma(i,k)$, the CL softmax $p_{ik} = \exp(s_{ik}/\tau)/\sum_{\ell\in D_i}\exp(s_{i\ell}/\tau)$, and the NSCL softmax over negatives $q_{ij} = p_{ij}/(1-\alpha_i)$ for $j \in \mathrm{neg}_i$, with $q_k = 0$ for $k \in \mathrm{pos}_i$, where $\alpha_i := \sum_{k\in\mathrm{pos}_i} p_{ik}$. Define $v_{ik} := \nabla_w s_{ik} = \nabla_w \,\mathrm{sim}\big(f_w(x_i), f_w(x_k)\big)$; by Assumption 2, $\|v_{ik}\| \le G$ for all (i,k).

For the per-anchor losses,

$$\nabla_w \ell_{i,\mathcal{B}_t}^{\mathrm{CL}} \;=\; \frac{1}{\tau}\Big(\sum_{k\in D_i} p_{ik}\,v_{ik} - v_{ii'}\Big), \qquad \nabla_w \ell_{i,\mathcal{B}_t}^{\mathrm{NSCL}} \;=\; \frac{1}{\tau}\Big(\sum_{j\in\mathrm{neg}_i} q_{ij}\,v_{ij} - v_{ii'}\Big).$$

Hence the per-anchor gradient difference is

$$\Delta g_i \;:=\; \nabla_w \ell_{i,\mathcal{B}_t}^{\mathrm{CL}} - \nabla_w \ell_{i,\mathcal{B}_t}^{\mathrm{NSCL}} \;=\; \frac{1}{\tau}\left(\underbrace{\sum_{k\in\mathrm{pos}_i} p_{ik}\,v_{ik}}_{\text{(A)}} \;+\; \underbrace{\sum_{j\in\mathrm{neg}_i} (p_{ij} - q_{ij})\,v_{ij}}_{\text{(B)}}\right).$$

By the triangle inequality and $\|v_{ik}\| \le G$,

$$\|\Delta g_i\| \;\le\; \frac{G}{\tau}\Big(\sum_{k\in\mathrm{pos}_i} p_{ik} + \sum_{j\in\mathrm{neg}_i} |p_{ij} - q_{ij}|\Big).$$

Since $q_{ij} = p_{ij}/(1-\alpha_i)$ for $j \in \mathrm{neg}_i$,

$$\sum_{j\in\mathrm{neg}_i} |p_{ij} - q_{ij}| = \sum_{j\in\mathrm{neg}_i} p_{ij}\,\frac{\alpha_i}{1-\alpha_i} = \alpha_i.$$

Therefore $\|\Delta g_i\| \le \frac{G}{\tau}(\alpha_i + \alpha_i) = \frac{2G}{\tau}\alpha_i$. Writing $r_i := \frac{Z_{\mathrm{pos}}}{Z_{\mathrm{neg}}}$ with $Z_{\mathrm{pos}} = \sum_{k\in\mathrm{pos}_i}\exp(s_{ik}/\tau)$, $Z_{\mathrm{neg}} = \sum_{j\in\mathrm{neg}_i}\exp(s_{ij}/\tau)$, we have $\alpha_i = \frac{r_i}{1+r_i}$, hence $2\alpha_i = \frac{2r_i}{1+r_i} \le 2r_i$, so

$$\|\Delta g_i\| \;\le\; \frac{2G}{\tau}\,\frac{Z_{\mathrm{pos}}}{Z_{\mathrm{neg}}}.$$

On the high-probability event of Cor. 3, for every anchor

$$\frac{Z_{\mathrm{pos}}}{Z_{\mathrm{neg}}} \;\le\; \frac{\mathrm{e}^{2/\tau}\big(\pi_{\max} + \epsilon_{B,\delta}\big)}{1 - \pi_{\max} - \epsilon_{B,\delta}} \;=\; \tfrac{1}{2}\,\Delta_{\pi,\delta}(B;\tau),$$

so $\|\Delta g_i\| \le \frac{G}{\tau}\,\Delta_{\pi,\delta}(B;\tau)$ for all anchors i.

Finally, the batch gradients are averages over anchors:

$$\nabla \bar{\ell}^{\mathrm{CL}}_{\mathcal{B}_t} - \nabla \bar{\ell}^{\mathrm{NSCL}}_{\mathcal{B}_t} \;=\; \frac{1}{B} \sum_{i \in \mathcal{B}_t} \Delta g_i,$$

hence

$$\left\| \nabla \bar{\ell}^{\mathrm{CL}}_{\mathcal{B}_t} - \nabla \bar{\ell}^{\mathrm{NSCL}}_{\mathcal{B}_t} \right\| \;\le\; \frac{1}{B} \sum_{i \in \mathcal{B}_t} \|\Delta g_i\| \;\le\; \frac{1}{B} \sum_{i \in \mathcal{B}_t} \frac{G}{\tau}\, \Delta_{\pi,\delta}(B;\tau) \;=\; \frac{G}{\tau}\, \Delta_{\pi,\delta}(B;\tau).$$

□

Theorem 2. *Fix $B, T \in \mathbb{N}$, $\delta \in (0,1)$, and temperature $\tau > 0$. Suppose Assumptions 1–2 hold. Then, with probability at least $1-\delta$,*

$$e_T \;\le\; \frac{G}{\beta\,\tau}\, \Delta_{\pi,\delta}(B;\tau) \left(\exp\Big(\beta \sum_{t=0}^{T-1} \eta_t \Big) - 1 \right).$$

Proof. Let $\Phi_t^{\mathrm{CL}}(w) := \bar{\ell}^{\mathrm{CL}}_{\mathcal{B}_t}(w)$ and $\Phi_t^{\mathrm{NSCL}}(w) := \bar{\ell}^{\mathrm{NSCL}}_{\mathcal{B}_t}(w)$. Assume each Φ_t^{CL} is β-smooth. Set $e_t := \|w_t^{\mathrm{CL}} - w_t^{\mathrm{NSCL}}\|$.

Write

$$\begin{aligned} e_{t+1} \;&=\; \left\| w_{t+1}^{\mathrm{CL}} - w_{t+1}^{\mathrm{NSCL}} \right\| \;=\; \left\| T_t(w_t^{\mathrm{CL}}) - \big(w_t^{\mathrm{NSCL}} - \eta_t\, \nabla \Phi_t^{\mathrm{NSCL}}(w_t^{\mathrm{NSCL}}) \big) \right\| \\ &\le \underbrace{\left\| T_t(w_t^{\mathrm{CL}}) - T_t(w_t^{\mathrm{NSCL}}) \right\|}_{\text{(I)}} + \eta_t\, \underbrace{\left\| \nabla \Phi_t^{\mathrm{CL}}(w_t^{\mathrm{NSCL}}) - \nabla \Phi_t^{\mathrm{NSCL}}(w_t^{\mathrm{NSCL}}) \right\|}_{\text{(II)}}. \end{aligned}$$

Bounding (I). Using the integral Hessian representation,

$$\nabla \Phi_t^{\mathrm{CL}}(u) - \nabla \Phi_t^{\mathrm{CL}}(v) = H_t(v,u)\,(u-v), \qquad H_t(v,u) := \int_0^1 \nabla^2 \Phi_t^{\mathrm{CL}}(v + \tau(u-v))\, d\tau,$$

and β-smoothness gives $\|H_t(v,u)\|_{2\to 2} \le \beta$. Hence

$$\begin{aligned} \|T_t(u) - T_t(v)\| \;&=\; \|(I - \eta_t H_t(v,u))(u-v)\| \\ &\le\; \left\| I - \eta_t H_t(v,u) \right\|_{2\to 2} \|u - v\| \\ &\le\; (1 + \eta_t \beta)\, \|u - v\|. \end{aligned}$$

Thus, (I) $\le (1 + \eta_t \beta)\, e_t$.

Bounding (II). On the high-probability event of Cor. 3, Lem. 7 yields

$$\text{(II)} \;\le\; \frac{G}{\tau}\, \Delta_{\pi,\delta}(B;\tau).$$

Combining the bounds,

$$e_{t+1} \;\le\; (1 + \eta_t \beta)\, e_t \;+\; \eta_t\, \frac{G}{\tau}\, \Delta_{\pi,\delta}(B;\tau). \tag{15}$$

Iterating equation 15 from $e_0 = 0$ gives

$$e_T \;\le\; \sum_{t=0}^{T-1} \eta_t\, \frac{G}{\tau}\, \Delta_{\pi,\delta}(B;\tau) \prod_{s=t+1}^{T-1} (1 + \eta_s \beta) \;\le\; \frac{G}{\tau}\, \Delta_{\pi,\delta}(B;\tau) \sum_{t=0}^{T-1} \eta_t\, \exp\Big(\beta \sum_{s=t+1}^{T-1} \eta_s \Big),$$

where we used $1 + x \le e^x$. Let $S_k := \sum_{s=k}^{T-1} \eta_s$ so that $S_t = \eta_t + S_{t+1}$. Then for each t,

$$\eta_t\, \exp(\beta S_{t+1}) \;\le\; \frac{1}{\beta} \Big(\exp(\beta S_t) - \exp(\beta S_{t+1}) \Big),$$

since $e^{\beta \eta_t} - 1 \ge \beta \eta_t$. Summing over $t = 0, \ldots, T-1$ telescopes to

$$e_T \;\le\; \frac{G}{\beta\tau}\, \Delta_{\pi,\delta}(B;\tau) \left(\exp\Big(\beta \sum_{t=0}^{T-1} \eta_t \Big) - 1 \right).$$

This holds with probability at least $1 - \delta$. □

E.4 SIMILARITY-SPACE ANALYSIS AND COUPLING

Lemma 8 (Per-step gradient gap in similarity space). *On the event of Cor. 3, for any step t,*

$$\left\|G_t^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) - G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{NSCL}})\right\|_F \;\le\; \underbrace{\frac{1}{\tau}\cdot\frac{\Delta_{\pi,\delta}(B;\tau)}{\sqrt{B}}}_{\text{reweighting (block-orth.)}} \;+\; \underbrace{\frac{1}{2\tau^2 B}\left\|\Sigma_t^{\mathrm{CL}} - \Sigma_t^{\mathrm{NSCL}}\right\|_F}_{\text{Lipschitz in }\Sigma}.$$

Proof. Add and subtract $G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{CL}})$ and apply the triangle inequality:

$$\begin{aligned}&\left\|G_t^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) - G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{NSCL}})\right\|_F \\ &\le \underbrace{\left\|G_t^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) - G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{CL}})\right\|_F}_{\text{(A)}} + \underbrace{\left\|G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{CL}}) - G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{NSCL}})\right\|_F}_{\text{(B)}}.\end{aligned} \tag{16}$$

Term (B): Lipschitz in Σ. By the temperature-τ softmax–Hessian bound equation 11,

$$\text{(B)} \;\le\; \frac{1}{2\tau^2 B}\left\|\Sigma_t^{\mathrm{CL}} - \Sigma_t^{\mathrm{NSCL}}\right\|_F.$$

Term (A): reweighting gap at fixed Σ_t^{CL}. Decompose the batch gradient into anchor blocks:

$$G_t^{\circ}(\Sigma) \;=\; \frac{1}{B}\sum_{i\in\mathcal{B}_t} g_{t,i}^{\circ}(\Sigma), \qquad \circ \in \{\mathrm{CL}, \mathrm{NSCL}\},$$

where each $g_{t,i}^{\circ}$ has support only on the coordinates of anchor i. For anchor i, with temperature τ, $g_{t,i}^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) = (1/\tau)(p_i - e_{i'})$, $g_{t,i}^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{CL}}) = (1/\tau)(q_i - e_{i'})$, so $g_{t,i}^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) - g_{t,i}^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{CL}}) = (1/\tau)(p_i - q_i)$ on that block. By block orthogonality (Lem. 1),

$$\text{(A)} \;=\; \frac{1}{B}\Big\|\sum_{i\in\mathcal{B}_t}\frac{1}{\tau}(p_i - q_i)\Big\|_F \;=\; \frac{1}{\tau B}\sqrt{\sum_{i\in\mathcal{B}_t}\|p_i - q_i\|_2^2}.$$

On the event of Cor. 3, Lem. 6 gives $\|p_i - q_i\|_2 \le \Delta_{\pi,\delta}(B;\tau)$ for every anchor, hence

$$\text{(A)} \;\le\; \frac{1}{\tau B}\sqrt{B\,\Delta_{\pi,\delta}(B;\tau)^2} \;=\; \frac{1}{\tau}\cdot\frac{\Delta_{\pi,\delta}(B;\tau)}{\sqrt{B}}.$$

Combining the bounds on (A) and (B) yields the claim. □

Theorem 1 (Similarity-space coupling). *Fix $B, T \in \mathbb{N}$, $\delta \in (0,1)$, and temperature $\tau > 0$. Consider the coupled similarity-descent recursions equation 1 for CL and NSCL with shared initialization and shared mini-batches/augmentations. Then, with probability at least $1-\delta$ over the draws of the mini-batches and augmentations, for any stepsizes $(\eta_t)_{t=0}^{T-1}$,*

$$\left\|\Sigma_T^{\mathrm{CL}} - \Sigma_T^{\mathrm{NSCL}}\right\|_F \;\le\; \exp\Big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\Big)\,\frac{1}{\tau\sqrt{B}}\Big(\sum_{t=0}^{T-1}\eta_t\Big)\,\Delta_{\pi,\delta}(B;\tau). \tag{2}$$

Proof. Condition on the event of Cor. 3 (which holds with probability at least $1-\delta$). Let $D_t := \|\Sigma_t^{\mathrm{CL}} - \Sigma_t^{\mathrm{NSCL}}\|_F$. From the coupled updates equation 1,

$$\Sigma_{t+1}^{\mathrm{CL}} - \Sigma_{t+1}^{\mathrm{NSCL}} \;=\; \left(\Sigma_t^{\mathrm{CL}} - \Sigma_t^{\mathrm{NSCL}}\right) - \eta_t\left(G_t^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) - G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{NSCL}})\right),$$

hence

$$D_{t+1} \;\le\; D_t + \eta_t\left\|G_t^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) - G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{NSCL}})\right\|_F.$$

Add and subtract $G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{CL}})$ and apply Lem. 8 (reweighting gap + Lipschitz with temperature τ):

$$\left\|G_t^{\mathrm{CL}}(\Sigma_t^{\mathrm{CL}}) - G_t^{\mathrm{NSCL}}(\Sigma_t^{\mathrm{NSCL}})\right\|_F \;\le\; \frac{1}{\tau}\cdot\frac{\Delta_{\pi,\delta}(B;\tau)}{\sqrt{B}} + \frac{1}{2\tau^2 B}\,D_t.$$

Therefore,

$$D_{t+1} \;\le\; \Big(1+\frac{\eta_t}{2\tau^2 B}\Big)D_t \;+\; \eta_t\,\frac{1}{\tau}\cdot\frac{\Delta_{\pi,\delta}(B;\tau)}{\sqrt{B}}.$$

Let $\alpha_t := \frac{\eta_t}{2\tau^2 B}$ and $\gamma_t := \eta_t\,\frac{\Delta_{\pi,\delta}(B;\tau)}{\tau\sqrt{B}}$. With $D_0 = 0$ (shared initialization), the discrete Grönwall/product form gives

$$D_T \;\le\; \sum_{s=0}^{T-1}\gamma_s \prod_{u=s+1}^{T-1}(1+\alpha_u) \;\le\; \exp\Big(\sum_{u=0}^{T-1}\alpha_u\Big)\sum_{s=0}^{T-1}\gamma_s,$$

using $\prod_u(1+\alpha_u) \le \exp(\sum_u \alpha_u)$. Substituting α_t, γ_t yields

$$D_T \;\le\; \exp\Big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\Big)\;\frac{1}{\tau\sqrt{B}}\Big(\sum_{t=0}^{T-1}\eta_t\Big)\,\Delta_{\pi,\delta}(B;\tau),$$

as desired. □

Consequences for CKA and RSA.

Corollary 1 (CKA lower bound). *In the setting of Thm. 1. Assume $\|K_T^{\mathrm{CL}}\|_F > 0$. With probability at least $1-\delta$,*

$$\mathrm{CKA}_T \;\ge\; \frac{1-\rho_T}{1+\rho_T}, \qquad \rho_T \;\le\; \frac{\exp\big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\big)\;\frac{1}{\tau\sqrt{B}}\big(\sum_{t=0}^{T-1}\eta_t\big)\,\Delta_{\pi,\delta}(B;\tau)}{\|K_T^{\mathrm{CL}}\|_F}.$$

Proof. Let $A_T := \|K_T^{\mathrm{CL}}\|_F > 0$ and $\Delta_{K,T} := \|K_T^{\mathrm{CL}} - K_T^{\mathrm{NSCL}}\|_F$, where all norms are Frobenius. Then

$$\begin{aligned}\langle K_T^{\mathrm{CL}}, K_T^{\mathrm{NSCL}}\rangle &= \big\langle K_T^{\mathrm{CL}},\, K_T^{\mathrm{CL}} + (K_T^{\mathrm{NSCL}} - K_T^{\mathrm{CL}})\big\rangle \\ &= \|K_T^{\mathrm{CL}}\|_F^2 + \big\langle K_T^{\mathrm{CL}},\, K_T^{\mathrm{NSCL}} - K_T^{\mathrm{CL}}\big\rangle \;\ge\; A_T^2 - A_T\,\Delta_{K,T},\end{aligned} \tag{17}$$

by Cauchy–Schwarz. By the triangle inequality, $\|K_T^{\mathrm{NSCL}}\|_F \le A_T + \Delta_{K,T}$. Hence

$$\mathrm{CKA}_T \;=\; \frac{\langle K_T^{\mathrm{CL}}, K_T^{\mathrm{NSCL}}\rangle}{\|K_T^{\mathrm{CL}}\|_F\,\|K_T^{\mathrm{NSCL}}\|_F} \;\ge\; \frac{A_T^2 - A_T\Delta_{K,T}}{A_T(A_T+\Delta_{K,T})} \;=\; \frac{1-\Delta_{K,T}/A_T}{1+\Delta_{K,T}/A_T}.$$

Next, $K_T^{\circ} = H\Sigma_T^{\circ}H$ with the centering projector $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$, so $\Delta_{K,T} = \|H(\Sigma_T^{\mathrm{CL}} - \Sigma_T^{\mathrm{NSCL}})H\|_F \le \|\Sigma_T^{\mathrm{CL}} - \Sigma_T^{\mathrm{NSCL}}\|_F$ because $\|H\|_{2\to 2} = 1$. By Thm. 1, with probability at least $1-\delta$,

$$\|\Sigma_T^{\mathrm{CL}} - \Sigma_T^{\mathrm{NSCL}}\|_F \;\le\; \exp\Big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\Big)\;\frac{1}{\tau\sqrt{B}}\Big(\sum_{t=0}^{T-1}\eta_t\Big)\,\Delta_{\pi,\delta}(B;\tau).$$

Combining the last two equations yields the lower bound on CKA_T with probability at least $1-\delta$. □

Corollary 2 (RSA lower bound). *In the setting of Thm. 1. Assume $\sigma_{D,T} > 0$. With probability at least $1-\delta$,*

$$\mathrm{RSA}_T \;\ge\; \frac{1-r_T}{1+r_T}, \qquad r_T \;\le\; \frac{\exp\big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\big)\;\frac{1}{\tau\sqrt{B}}\big(\sum_{t=0}^{T-1}\eta_t\big)\,\Delta_{\pi,\delta}(B;\tau)}{\sqrt{M}\,\sigma_{D,T}}.$$

Proof. Let $M = \binom{N}{2}$ and let $C := I - \frac{1}{M}\mathbf{1}\mathbf{1}^\top$ be the centering projector in $\mathbb{R}^M$. Write $a_c := Ca_T$ and $b_c := Cb_T$. Then

$$\mathrm{RSA}_T \;=\; \frac{\langle a_c, b_c\rangle}{\|a_c\|_2\,\|b_c\|_2}.$$

For any nonzero u and any v in an inner-product space,

$$\langle u, v\rangle \;=\; \langle u, u + (v-u)\rangle \;=\; \|u\|_2^2 + \langle u, v-u\rangle \;\ge\; \|u\|_2^2 - \|u\|_2\,\|v-u\|_2,$$

and $\|v\|_2 \le \|u\|_2 + \|v-u\|_2$. Therefore,

$$\frac{\langle u, v\rangle}{\|u\|_2\,\|v\|_2} \;\ge\; \frac{1-\|v-u\|_2/\|u\|_2}{1+\|v-u\|_2/\|u\|_2}.$$

Apply this with $u = a_c$ and $v = b_c$ to obtain

$$\mathrm{RSA}_T \;\geq\; \frac{1 - \|b_c - a_c\|_2 / \|a_c\|_2}{1 + \|b_c - a_c\|_2 / \|a_c\|_2}.$$

Since C is an orthogonal projector, $\|b_c - a_c\|_2 = \|C(b_T - a_T)\|_2 \leq \|b_T - a_T\|_2$. By construction of the RDM vectors,

$$b_T - a_T \;=\; -\,\mathrm{vec}\left(\mathrm{off}\left(\Sigma_T^{\mathrm{NSCL}} - \Sigma_T^{\mathrm{CL}}\right)\right),$$

so $\|b_T - a_T\|_2 = \|\mathrm{off}(\Sigma_T^{\mathrm{NSCL}} - \Sigma_T^{\mathrm{CL}})\|_F \leq \|\Sigma_T^{\mathrm{NSCL}} - \Sigma_T^{\mathrm{CL}}\|_F$. Finally, by Thm. 1, with probability at least $1 - \delta$,

$$\|\Sigma_T^{\mathrm{NSCL}} - \Sigma_T^{\mathrm{CL}}\|_F \;\leq\; \exp\Big(\frac{1}{2\tau^2 B}\sum_{t=0}^{T-1}\eta_t\Big)\;\frac{1}{\tau\sqrt{B}}\Big(\sum_{t=0}^{T-1}\eta_t\Big)\;\Delta_{\pi,\delta}(B;\tau).$$

Combining the last three displays yields the stated $(1 - r)/(1 + r)$ lower bound on RSA_T after substituting $\|a_c\|_2 = \sqrt{M}\,\sigma_{D,T}$. □