Structural Origins of Catastrophic Forgetting in Self-Supervised Continual Learning: A Directional and Curvature-Based Analysis of the Learning Signal

A thesis

submitted in partial fulfilment

of the requirements for the degree

of

Master of Science (Research) in Artificial Intelligence

at

The University of Waikato

by

Rucheng Huang

Supervisor: Bernhard Pfahringer

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2026

Acknowledgments

The completion of this thesis would not have been possible without the support and companionship of many people, to whom I extend my most sincere gratitude.

First and foremost, I would like to express my heartfelt thanks to my supervisor, Prof. Bernhard Pfahringer. Over the past year, his guidance has deeply shaped my attitude toward research. His wholehearted dedication to his work has continued to inspire me. I am sincerely grateful for his support throughout this year.

I would also like to thank Prof. Eibe Frank, who truly introduced me to the world of academic research. He has continued to care about my studies and academic development. I am equally grateful to Dr. Yaqian Zhang. I have always treasured my discussions with her, particularly the thoughtful guidance she provided when I first entered the field of continual learning.

My sincere thanks go to the PhD and master's students in the G.211 lab, Muhammad Zain Ali, Renjun Cai, Linda Cai, Shengtao Lin, and Rafia Malik. Whenever I faced difficulties, whether academic or personal, they were often the first people I turned to, and they always showed me great kindness and care. I would also like to thank Xiaodong Yan, the best friend I have made in New Zealand, and my landlords, Min, Tony, and Kelly, whose timely conversations always brought me comfort and ease.

Above all, I wish to express my deepest gratitude to my parents and my wife. While I have been away, they have looked after our little Jubao, who is just twenty-eight months old, and have been my greatest source of strength throughout this journey. When I made the decision to pursue my dream in research, their response was nothing but unwavering support and encouragement. This thesis is dedicated to them.

Abstract

Catastrophic forgetting remains a fundamental challenge in continual learning, where acquiring new knowledge systematically degrades previously learned representations. While existing approaches primarily mitigate this by imposing architectural constraints or using data replay strategies, they offer limited theoretical insight into why and how parameter updates interfere with consolidated knowledge structures.

This thesis proposes a structured analytical lens to examine the collision between new and old knowledge at the level of parameter updates, feature representations, and low-dimensional learning signals. Rather than introducing a new method, we seek to characterize the geometric conditions under which gradient updates pose the greatest risk to previously learned structure. Specifically, we hypothesize that forgetting is governed by the degree to which parameter updates project onto high-curvature regions of the old task's loss landscape, namely those directions along which the old loss function is most sensitive to perturbation. We derive a theoretical bound that isolates this curvature-projection term as the dominant factor driving representational forgetting, and empirically verify both the structural conditions under which this bound holds and its statistical relationship with observed forgetting.

To ground this analysis in a concrete and mechanistically interpretable setting, we adopt SwAV, a representative self-supervised contrastive learning framework, as our experimental substrate. We leverage SwAV's internal prototype assignment process as a low-dimensional learning signal that faithfully reflects the underlying representational dynamics, allowing the theoretical bound to be expressed and studied in a tractable, interpretable form.

Building on this, we further consider whether the second-order sensitivity structure of old knowledge, when projected into a lower-dimensional subspace, retains meaningful geometric differentiation between sensitive and insensitive directions. Our experiments confirm that such a structure persists at low dimensionality and that SwAV's learning signal selectively engages it. We observe that interference with prior knowledge is measurably reduced when the energy of the new task's low-dimensional signal concentrates along axes that carry less of the old task's curvature structure, rather than along regions of high coupling.

Contents

Chapter 1

Introduction

1.1 Continual Learning and the Origins of Catastrophic Forgetting

Intelligent systems rarely learn from a single, stationary dataset. A medical diagnostic model must incorporate new disease phenotypes without forgetting established ones [30]; autonomous robots navigating dynamic environments must adapt to ever-changing scenes without prior knowledge of the domain [51, 60]. An illustrative example arises in the predictive maintenance of turbofan engines [38], where operational data are distributed across different locations and cannot always be aggregated centrally. As a result, models must continually learn from small, decentralised datasets to forecast potential faults under real-world operating conditions. Large language models face an analogous challenge. As new corpora continually emerge [62], models need to incorporate new data without restarting training from scratch, which would otherwise waste the enormous cost of prior training [25].

These scenarios share a key feature: the learning objective is not fully observable at once. Data often arrives as a stream [2]. Previously seen data may no longer be accessible. Learning in such conditions, called continual or lifelong learning [15], is now an active research topic in academia and industry.

Continual learning faces a fundamental difficulty [28, 54, 57]. It was observed early on that when a trained model is updated only with data from a new task, especially when old-task data are no longer available, the model tends to lose its ability to perform well on earlier learned tasks. This phenomenon is known as catastrophic forgetting.

The problem appears in different forms. Sometimes new classes appear and must be added to the model. At other times, the data distribution changes, leading to domain drift [35]. Researchers have proposed solutions for these different scenarios. Most continual learning methods fall into two broad approaches.

The first is experience replay, which stores a small number of samples from previous

tasks and reuses them during training [34, 43, 50]. Replay methods can further augment stored samples with data augmentation to increase diversity [66]. Memory storage is usually not the main bottleneck, so replay has become a direct and effective solution. Nevertheless, it raises challenges. For instance, explicit labels may not be available in many real-world scenarios. Without such supervision, it is difficult to determine appropriate sampling ratios [59], which can lead to overfitting or underfitting [43].

The second strategy is to constrain parameter updates so that optimisation avoids directions that would strongly interfere with previously learned knowledge. The central difficulty with such approaches arises from the scale of modern deep networks: with millions or billions of parameters, gradient updates occur in an extremely high-dimensional parameter space [53].

A natural combination of these ideas appears in data distillation or regularisation-based replay, where a small amount of previous data is reused to regularise parameter updates [44, 49]. Many such methods implicitly reduce the problem to a lower-dimensional space by constraining model outputs, representations, or logits, rather than directly regularising the full parameter space [33, 47].

The objective of this thesis is not to propose yet another training pipeline or optimise performance on specific benchmarks. Instead, we aim to better understand the **structural origins of forgetting** itself. In particular, we study whether catastrophic forgetting is linked to inherent properties of the learning dynamics or the statistical structure of the training signal. If such a structural quantity can be identified, it may provide a principled starting point for understanding and potentially controlling forgetting.

To explore this question, we turn to a learning setting with a clearer, structured training signal. We briefly introduce self-supervised learning, which offers a useful perspective for analysing how representations evolve during training.

1.2 Self-Supervised Learning as a Lens for Continual Learning

Most continual learning research is conducted in the supervised setting [15, 28, 44, 59]. We do not revisit here the broader strengths of supervised learning. For continual learning, one clear advantage of label-based supervision is that it provides a natural way to construct a balanced memory pool, making replay more controllable and interpretable.

At the same time, this setting also has limitations. Labels may be scarce, and even when available, they may not fully capture the data structure [41]. When the data distribution shifts, fixed labels may become misleading or fail to capture the underlying data variation.

This thesis studies **self-supervised learning** (SSL) [12, 13, 21, 24] as a lens for studying continual learning. Our main reason is not simply that self-supervised methods have achieved strong benchmark performance, often comparable to supervised baselines, but

that their learning signals are typically richer and more adaptive to domain variation. From this perspective, SSL may better adapt to distribution shifts and multi-domain environments.

In the next chapter, we will discuss why self-supervised learning is suitable for this thesis. At this stage, we highlight two conditions that make it useful for our study. First, the learning framework should stay structurally stable across tasks. The loss function, optimisation target, and model architecture should not need to be replaced as new tasks arrive [18, 36]. Second, the self-supervised signal should be analysable [7, 52]. Compared with hard labels, such signals are often richer and more structured, but remain lower-dimensional than raw data representations.

Our goal is to use this framework as a controlled setting to investigate forgetting. In informal terms, the loss landscape of previous tasks is the terrain where learning occurs. The training signal of a new task acts as a traveller crossing this landscape. By observing the traveller's path and the cost of traversing it, we characterise the extent to which the new task conflicts with the old one—the degree of catastrophic forgetting.

Based on the above, we choose **SwAV** [7], a representative self-supervised learning framework, as the foundation for our investigation. The next chapter will detail why this framework suits our purpose.

SwAV is a self-supervised learning method that learns visual representations by matching image embeddings to learnable prototypes through a clustering-based objective. Instead of explicit labels, the model predicts prototype assignments for different views of the same image and aligns these predictions during training.

SwAV has a structural property relevant to our study. Throughout the entire training process, the model keeps a set of learnable prototypes that act as reference points for representation learning. New data representations are continuously compared with these prototypes, and learning proceeds by predicting their relationships. Thus, the prototype mechanism induces a learning signal in a structured and relatively low-dimensional prototype space. This structured signal gives us a clear lens to analyse the dynamics of learning and forgetting.

1.3 Research Questions

1. Under the plasticity–forgetting dilemma [28], is there a structural quantity that specifically governs forgetting without being reducible to plasticity?

    (a) From the loss structure view, can such a quantity be formally related to parameter updates or, more fundamentally, to the learning signal itself?

    (b) In a concrete self-supervised setting, can this quantity be made interpretable, observable, and potentially controllable?

2. Does forgetting exhibit identifiable structural or statistical patterns that indicate its severity? If so, when do these patterns arise, and can we support them with theory or evidence?

3. If a lower-dimensional analytical lens exists, can the mechanisms underlying forgetting be captured by a concrete structural quantity that reveals how new-task learning interacts with the structure of previously learned tasks?

   (a) Under what conditions can such a high-dimensional–to–low-dimensional transformation be established?

   (b) When shifting analysis to a lower-dimensional structure, what information remains or is lost, and what benefits does this give for understanding, predicting, or managing forgetting?

1.4 Contributions

The main contributions of this thesis are summarised as follows:

1. We derive a structural bound on catastrophic forgetting that isolates a quantity beyond the shared dependence of forgetting and plasticity on the magnitude of parameter updates. Using SwAV as the analytical entry point, this bound is derived from its training regime and loss function.

2. We provide empirical evidence that the structural factor required by the forgetting bound is not only observable in experiments, but also statistically correlated with the severity of forgetting.

3. We provide a lower-dimensional perspective on forgetting that complements the conventional parameter-space analysis. Within this representation, we introduce a magnitude-invariant structural metric that characterises how the low-dimensional learning signal projects onto the sensitive regions of the old-task curvature structure in the reduced space.

1.5 Thesis Organization

This thesis is organised as follows.

Chapter 2 introduces the motivation for adopting self-supervised learning, and in particular the SwAV framework, as the analytical setting of this thesis. We briefly describe the training regime of SwAV and review related work in continual learning that shares conceptual similarities with our perspective.

Chapter 3 develops the theoretical foundation of this study. We analyse the SwAV loss function through first- and second-order gradients and show that the resulting learning signal exhibits a low-rank structure. Based on a general definition of forgetting and

plasticity in terms of loss variation, we derive a formulation that links these quantities to the direction of parameter updates, which can in turn be expressed through the induced self-supervised learning signal. In particular, we show that forgetting is independently related to the projection of this signal onto a second-order structural component.

Chapter 4 provides empirical validation of several assumptions underlying the analysis in Chapter 3. We empirically suggest that the structural component identified in the theoretical bound is observable and exhibits a statistical correlation with the severity of forgetting.

Chapter 5 constructs a lower-dimensional representation of old-task sensitivity by pulling the analysis back from parameter space into prototype space. The resulting structure encodes the sensitivity of previously learned tasks in prototype coordinates, and we characterise its intrinsic geometry from two complementary views: spectral concentration and coordinate organisation.

Chapter 6 investigates how the learning signal engages the sensitive structure of old-task knowledge across different distributional regimes and curvature landscapes. It first establishes, in a controlled same-task setting, that this engagement is selective rather than isotropic. It then shows, across task boundaries, that a magnitude-invariant structural metric extracted from this interaction provides additional resolution for forgetting severity beyond the total quadratic activation alone.

Chapter 7 summarises the main findings, reflects on the structural analogy between the parameter-space and prototype-space analyses, and discusses the limitations and scope of the framework.

1.6 Reader's Guide

This thesis is not concerned with designing continual learning algorithms. Instead, it analyses the structural relationship between the learning signal and the second-order geometry of previously learned tasks, using SwAV as a controlled analytical platform (see Section 2.1).

The analysis begins by showing that all SwAV gradients are parametrised by the assignment disagreement $g = p - q$ (Lemma 3.1). This single object then provides a unified account of both plasticity and forgetting: plasticity scales with the magnitude of g (Section 3.3.1), while forgetting is governed by the directional alignment of g with the old-task high-curvature subspace (Proposition 3.17 and Corollary 3.18). The central result is therefore not a trade-off between the two, but a decomposition that treats them as independently addressable through magnitude and direction, respectively.

Three mathematical assumptions underlying this decomposition are empirically tested in Section 4.2 (block-energy dominance, eigenspace stability, and spectral concentration). Readers who wish to assess the bound's practical relevance without following every proof may proceed directly to Section 4.3, where directional alignment is evaluated as a

predictor of forgetting severity across task transitions.

Chapter 5 pulls the analysis back into the K-dimensional prototype space, introducing two complementary metrics: C_r (Eq. 5.3), measuring spectral concentration of the forgetting energy in reduced space, and $\text{Ratio}_{\text{diag}}$ (Eq. 5.5), measuring how much of that energy induced by learning signal is carried by individual prototypes rather than cross-prototype coupling in prototype space. The empirical case that these metrics characterise forgetting severity across different curvature landscapes is made in Section 6.3. The structural analogy and distinction between the parameter-space and prototype-space perspectives on forgetting are discussed in Section 6.4.

The code and experiments supporting this thesis are available at `https://github.com/rch-huang/swav-forgetting-analysis`.

List of Frequently Used Symbols

SwAV, assignments, and OT formulation

Symbol	Meaning
x	Input sample.
$z = f_\theta(x)$	Representation produced by the backbone with parameters θ.
$C = [c_1, \dots, c_K]$	Prototype matrix / codebook with K prototypes.
c_k	The k-th prototype vector.
$s_{k,i}$	Similarity logit between sample i and prototype k, where $s_{k,i} = c_k^\top z_i / \tau$.
p_i	Softmax prediction vector over prototypes for sample i.
$p_{k,i}$	The k-th component of p_i.
$Q^\star$	Batch-level Sinkhorn transport matrix solving the entropic OT problem.
q_i	Assignment vector for sample i. Under matrix notation, it is the i-th column of $Q^\star$ after normalisation.
$q_{k,i}$	The k-th component of the assignment vector q_i.
$g_i = p_i - q_i$	Assignment disagreement vector for sample i.
g	Generic notation for the disagreement signal $p - q$.
τ	Softmax temperature used in the similarity logits.
ε	Entropic regularisation coefficient in the Sinkhorn / OT assignment.
T_{eff}	Effective temperature induced by the OT form (Treating OT assignment as in the form of softmax), where $T_{\text{eff}} = \tau\varepsilon$.

Forgetting and curvature analysis

Symbol	Meaning
$L_t(w)$	Expected loss of Task t evaluated at parameters w.
$w = (\theta, C)$	Joint trainable parameters: backbone parameters and prototypes.
$w_t^\star$	Reference (optimal) solution after training on Task t.
$\Delta w = w_{t+1}^\star - w_t^\star$	Parameter displacement from the Task t solution to the post-Task $t+1$ solution.
F_t	Forgetting on Task t, measured by the increase in old-task loss after later training.
$H_t = \nabla^2 L_t(w_t^\star)$	Hessian of the Task t loss at the reference point.
$H_{\theta\theta}$	Backbone–backbone block of the Hessian.
H_{CC}	Prototype–prototype block of the Hessian.
$H_{\theta C}$, $H_{C\theta}$	Cross blocks of the Hessian.
$P_{\text{high}}(\lambda)$	Projector onto the eigensubspace of the Hessian (or a Hessian block) associated with eigenvalues exceeding threshold λ.
T_θ	Trajectory-level linear operator mapping the stacked disagreement signal $\bar{g}$ to the backbone displacement $\Delta\theta$.
T_C	Trajectory-level linear operator mapping the stacked disagreement signal $\bar{g}$ to the prototype displacement ΔC.
$\bar{g}$	Stacked disagreement vector collected over a training segment.
M_θ	Second-order operator in the prototype-coordinate view.
$g^\top M_\theta g$	Prototype-coordinate quadratic energy, equivalent to the single-step backbone quadratic form $\Delta\theta^\top H_t \Delta\theta$ after substituting $\Delta\theta = T_\theta g$.
$\text{Ratio}_{\text{diag}}$	Fraction of the quadratic energy $g^\top M_\theta g$ explained by the diagonal part of M_θ.

Chapter 2

Background and Related Work

Self-supervised learning (SSL), including both contrastive and non-contrastive paradigms, has become a widely adopted backbone for continual representation learning without labels [9, 18, 20, 59, 64], where the goal is to maintain transferable representations under distribution shifts without access to supervised labels.

At a general level, SSL provides a label-free learning signal that extracts representations based on the intrinsic structure of the data rather than external supervision. This characteristic makes common continual learning strategies, such as distillation and replay [49, 59, 64], inherently compatible with the SSL framework. Since SSL decouples representation learning from task-specific categorizations, knowledge can be distilled from any point along the learning trajectory without the constraints of task-specific classifiers [5, 49]. Similarly, experience replay does not necessitate label-balanced memory design [18] and can instead rely on random sampling or feature-based selection criteria.

More importantly, the SSL objective remains structurally invariant across task boundaries. In contrast, supervised continual learning often requires modifying or expanding classification heads as new classes arrive [33, 44, 55], which inevitably alters the loss definition over time. Such invariance reduces cross-task comparison to evaluating a single, uniform objective under different data distributions, thereby providing a consistent metric for quantifying forgetting and plasticity [40, 57].

While the structural invariance of SSL benefits the continual learning setup, the efficacy of the learned embeddings depends on the specific geometric constraints imposed by the SSL objective. Unlike supervised learning, which is guided by explicit class labels [27], SSL must internally regulate how embeddings evolve to avoid degenerate solutions such as feature collapse, where representations gravitate toward a single, uninformative point [13, 21, 65].

For example, in contrastive frameworks such as SimCLR [12, 24], geometric structure is enforced through two mechanisms where augmented views of the same instance are aligned while distinct samples are pushed apart. This interplay induces a balance

between alignment and uniformity [58]: augmentations promote instance-level invariance, while the diversity of negatives regulates the global spread of embeddings. Non-contrastive methods depart from this approach by removing explicit negative sampling. While earlier works like DeepCluster [6] alternated between clustering and representation updates, yet they often struggled with cluster collapse or the inability of the cluster structure to capture nuanced semantic relatedness when multiple semantic dimensions coexist [41].

In response, methods such as SeLa [1] and SwAV [7] adopt assignment-based objectives paired with balancing constraints, often enforced through entropic regularization [14]. While SeLa generates pseudo-labels by applying optimal transport [14, 56] to the output of a standard classification head, SwAV introduces an explicit codebook-like prototype structure. These methods regulate the distribution of assignments so that embeddings are neither collapsed onto a single output dimension nor unevenly concentrated. Much like the alignment–uniformity trade-off in SimCLR, the balancing mechanism in SwAV maintains a persistent training signal that promotes predictive agreement across views while discouraging degenerate concentration [52].

2.1 Choice of Study Framework

In supervised continual learning, the source of both acquisition and interference is ultimately traceable to external labels: the learning signal is defined by the discrepancy between predictions and ground-truth targets, and forgetting arises when updates driven by new labels disrupt representations shaped by old ones [28, 33]. The analytical structure of forgetting therefore inherits the structure of the label space.

In self-supervised learning, no such external reference exists. The learning signal is generated entirely by the objective's internal mechanism, through the way it compares, aligns, or assigns representations to each other or to learnable anchors. This internally generated signal is the sole driver of representation change: it determines both the direction and magnitude of every parameter update. Consequently, when the data distribution shifts across tasks, the same internal signal that enables new-task acquisition is also the vehicle through which old-task representations are perturbed. Understanding continual learning dynamics in the SSL setting therefore reduces, at a fundamental level, to understanding the properties of this signal—its structure, its distributional characteristics, and how these translate into parameter-space displacement.

This motivates selecting an SSL framework in which the learning signal has a concrete and identifiable mathematical form, and in which the relationship between this signal and the resulting parameter updates is amenable to analysis rather than opaque. In principle, continual learning dynamics can be studied at the level of parameters, representations, or the learning signal itself. The parameter space is high-dimensional and difficult to interpret directly; representations are more interpretable but are the *consequence* of the learning process rather than its driver. The learning signal, by contrast, occupies a

privileged position: it is the immediate cause of every update, and in a well-structured framework its properties can be examined and related to downstream effects on both plasticity and stability. The more directly one can observe and characterise this signal, the more precisely one can ask how shifts in data distribution alter the learning dynamics and interact with previously acquired representations.

Why SwAV? The SwAV objective routes all learning signals through a codebook of K learnable prototypes. The gradient with respect to any sample is mediated by the similarity between that sample's embedding and the prototypes, a structure in which the codebook serves as a persistent, batch-independent set of reference points in the representation space. This codebook-mediated architecture is not unique to SwAV; similar structures appear in other prototype-based and clustering-based SSL methods [8, 32], and instance-level methods such as MoCo [24] maintain analogous fixed reference sets through momentum-updated memory banks. The analytical relevance of this codebook structure is that it defines a fixed set of directions through which learning signals act on the representation, rather than a set that is reconstituted from the batch at each step. In the context of continual learning, this means that the channel through which new-task signals could interfere with old-task representations has a stable geometric identity that persists across distribution shifts. Furthermore, although SwAV applies a stop-gradient to the Sinkhorn assignments during backpropagation, the assignments themselves are computed from the current embeddings and prototypes at each step; the functional relationship between model state and learning signal is therefore well-defined throughout training, even though it is not explicitly differentiated through.

A further consideration concerns where in the computational pipeline one can intervene to observe or modulate the learning dynamics. In contrastive frameworks built on the InfoNCE objective [27, 42], such as SimCLR [12], the primary tuneable scalar is the temperature τ in the softmax over pairwise similarities. This temperature rescales the representation-level logits that define the contrastive loss, thereby shaping the geometry of the learned embedding space. However, it does not provide direct access to the mechanism that generates the update signal.

In SwAV, the Sinkhorn assignment procedure introduces a structurally distinct control point. The entropic regularization coefficient ε governs the sharpness of the target assignment before the assignment enters the loss as a supervision signal. As will be formalized in Section 3, the gradient of the SwAV loss can be expressed entirely through the discrepancy between the model's predicted assignment and the Sinkhorn target. Modulating ε therefore acts upstream of the gradient computation, shaping this discrepancy before it enters the loss; it shapes the signal that drives representation change, rather than the representation that results from it.

This difference in causal ordering is relevant to the analytical position of this thesis: if one aims to understand how properties of the learning signal propagate into forgetting, a framework that permits intervention at the source of the signal, before it is translated

into parameter updates, is preferable to one in which the only accessible control point lies at the level of the output representation. Whether this upstream modulation is sufficient to produce measurable and interpretable effects on forgetting remains an empirical question, addressed in Section 6.

SwAV also benefits from an extensive empirical track record across architectures and datasets [7]. However, the primary motivation for selecting this framework is the analytical structure described above rather than benchmark standing. Whether the properties of the SwAV loss give rise to interpretable structure under continual learning is the question that the subsequent sections investigate.

2.2 SwAV Loss Formulation

Having motivated the choice of SwAV, we now review its objective in detail to establish the notation used throughout the thesis. In this framework, the learnable parameters consist of a backbone encoder f_θ (e.g., ResNet-18 or ResNet-50 [23]) that produces a d-dimensional representation $z = f_\theta(x) \in \mathbb{R}^d$, and a set of K trainable prototypes $C = [c_1, \dots, c_K] \in \mathbb{R}^{d \times K}$, commonly referred to as a codebook.

2.2.1 Swapped Prediction and Cross-Entropy Loss

SwAV operates by generating multiple augmented views of the same image. For simplicity, consider two views x_{i1} and x_{i2} of a sample x_i with corresponding normalised embeddings z_{i1} and z_{i2}. The objective is to enforce the representation of one view to predict the "code" (assignment) of the other view.

The alignment between an embedding z and the k-th prototype c_k is measured by similarity logits,

$$s^{(k)} = z^\top c_k / \tau, \tag{2.1}$$

where τ is a temperature hyperparameter. The soft prediction $p \in \Delta^{K-1}$ is given by the softmax distribution:

$$p^{(k)} = \frac{\exp(s^{(k)})}{\sum_{j=1}^{K} \exp(s^{(j)})}. \tag{2.2}$$

The SwAV loss is defined as the swapped cross-entropy between the prediction p_{i1} and a target code q_{i2} derived from the other view, and vice versa:

$$\mathcal{L}(z_{i1}, z_{i2}) = \ell(z_{i1}, q_{i2}) + \ell(z_{i2}, q_{i1}), \quad \text{where } \ell(z, q) = -\sum_{k=1}^{K} q^{(k)} \log p^{(k)}. \tag{2.3}$$

Optimal Transport and Sinkhorn Assignment. The targets q_{i1} and q_{i2} are not derived from a single sample in isolation. Instead, SwAV computes an assignment matrix Q over a batch of B samples by maximizing the similarity between the embeddings $Z = [z_1, \dots, z_N]$ and the prototypes C, while enforcing a uniform distribution constraint to prevent representational collapse. Here, N represents the total number of views in the batch (e.g., $N = 2B$ if two views per sample are used for assignment). This

task is formulated as an Optimal Transport (OT) optimisation problem with entropic regularization [14]:

$$Q^* = \max_{Q \in \mathcal{U}(r,c)} \mathrm{Tr}(Q^\top C^\top Z) + \epsilon H(Q), \tag{2.4}$$

where $H(Q) = -\sum_{ij} Q_{ij} \log Q_{ij}$ is the entropy of the assignment matrix, and ϵ controls the smoothness of the mapping. The constraint set $\mathcal{U}(r, c)$ ensures that Q is a valid assignment matrix with fixed marginals:

$$\mathcal{U}(r, c) = \{Q \in \mathbb{R}_+^{K \times N} \mid Q\mathbf{1}_N = r, Q^\top \mathbf{1}_K = c\}. \tag{2.5}$$

Here, $r = \frac{1}{K}\mathbf{1}_K$ and $c = \frac{1}{N}\mathbf{1}_N$ ensure that each prototype is selected approximately equally often within the batch.

The optimal assignment Q^* can be solved efficiently via the Sinkhorn-Knopp algorithm [14], which involves a few fast iterations of row and column normalization before each backpropagation step. Once computed, the i-th column of Q^* provides the target code q_i used for the cross-entropy loss[1] and Q does not participate in the gradient backpropagation process (i.e., a stop-gradient is applied). It is worth noting that while entropic regularization is central to the Sinkhorn iteration, it is handled implicitly during the assignment phase; therefore, no explicit regularization term appears in the final loss function.

2.3 Related Work: Parameter-space Interference and Directional Control

Continual learning studies the problem of sequentially adapting a model to a stream of tasks while preserving competence on previously learned ones [15, 51, 57]. Formally, given tasks $\mathcal{T}_1, \mathcal{T}_2, \ldots$, the learner observes data from each task in turn and updates shared parameters θ. The central difficulty arises from the fact that updates that improve performance on the current task may simultaneously degrade performance on earlier tasks — a phenomenon commonly referred to as catastrophic forgetting.

In this thesis, we approach continual learning from a specific structural perspective: we focus on how parameter updates interact with the sensitivity structure of previously learned knowledge. Rather than surveying the full spectrum of continual learning strategies (e.g., rehearsal-based memory [5, 11, 44, 47, 50, 66], architectural expansion [46, 63], or task-specific modules [37, 48]), we concentrate on mechanisms that operate directly at the level of parameter variation and objective geometry. This focus aligns with our central motivation: understanding forgetting as a consequence of directional interaction between updates and sensitive regions of the loss landscape.

[1] It is important to clarify the normalization relationship between the assignment matrix Q and the target codes q_i. While $Q \in \mathbb{R}^{K \times N}$ represents a joint distribution over the batch where $\sum_{k,i} Q_{ki} = 1$, the cross-entropy loss requires a valid conditional probability distribution for each view. Therefore, the target code used in Eq. 2.3 is defined as $q_i = N \cdot Q_i^*$, ensuring that $\sum_{k=1}^{K} q_i^{(k)} = 1$.

2.3.1 From Hessian Geometry to Sharpness-Aware Minimization

The geometric landscape of the loss function provides a fundamental lens for understanding both generalization and catastrophic forgetting. Formally, the local curvature around a minimum is characterized by the **Hessian matrix** $\nabla^2 L$. From an optimisation perspective [26], "flat" minima—regions where the Hessian spectrum is dominated by small eigenvalues—have often been observed to correlate with improved generalization performance. Intuitively, such regions exhibit reduced sensitivity to small parameter perturbations. In continual learning settings, where parameters are repeatedly updated under changing data distributions, this notion of curvature becomes particularly relevant [40]: directions corresponding to large eigenvalues can amplify interference across tasks, whereas updates confined to flatter directions may reduce forgetting.

However, explicitly regularizing the Hessian spectrum is computationally prohibitive for modern deep networks due to the quadratic cost of second-order derivatives. To address this, [26] proposed a sharpness metric that measures the maximum loss increase within a local neighborhood as a proxy for curvature. Building on this proxy, Sharpness-Aware Minimization (SAM) [19] introduced a scalable procedure to seek flatter regions. Instead of computing the Hessian, SAM identifies the "worst-case" direction of gradient ascent and performs an update that penalizes this local sharpness.

In [16], SAM is integrated into the continual learning objective. This approach essentially uses SAM as a first-order tool to manipulate second-order geometric features. By encouraging the model to reside in flatter valleys for each sequential task, these methods effectively reduce the sensitivity of old knowledge to new updates, thereby bridging the gap between efficient first-order optimisation and the desire for stable, low-curvature representations.

2.3.2 From Hessian Regularization to Gradient Projection

Another prominent line of research in continual learning investigates how to minimize the disruption in the parameter space by explicitly manipulating the gradient update directions. The foundational work, Elastic Weight Consolidation (EWC) [28], provides a bridge between these perspectives. By utilizing the diagonal of the Fisher Information Matrix—as a proxy for the Hessian matrix—EWC identifies "sensitive" parameter directions and suppresses updates along them through a quadratic penalty. However, this approach often suffers from insufficiency in complex, long-term learning scenarios. The diagonal approximation fails to capture the intricate off-diagonal correlations between parameters [45], and its constraints on the parameter space frequently lead to loss in the plasticity required to acquire new tasks [10].

To overcome the rigidity of weight-based penalties, Gradient Episodic Memory (GEM) [34] reinterprets the challenge by analysing how the gradient vector moves across the loss landscape, while it ensures that the current update direction g maintains a non-negative projection on the gradients of previous tasks ($g^\top g_{old} \geq 0$). This prevents

interference while allowing the model to move toward regions that potentially benefit historical knowledge through positive correlation. To address the high computational complexity of solving the quadratic programming (QP) problem in GEM, A-GEM [11] projects the current gradient onto a single direction derived from the average gradient of a sampled memory buffer, A-GEM significantly reduces the optimisation overhead while maintaining the benefits of directional constraints.

The methodology has further evolved toward defining more precise "safe subspaces". Orthogonal Gradient Descent (OGD) [17] strictly projects updates into the orthogonal complement of the space spanned by previous task gradients. The primary distinction lies in their objective: OGD enforces strict non-interference through orthogonality, whereas others, like GEM, allows for potential backward transfer by only requiring non-negative correlation.

Furthermore, Gradient Projection Memory (GPM) [47] shifts the focus from instantaneous gradients to the underlying representation space. GPM performs singular value decomposition on hidden activations to identify a set of principal bases that span a task-specific "knowledge subspace." Future gradients are then projected onto the orthogonal complement of this subspace, effectively preventing updates along directions that are strongly utilized by previous tasks. A notable subsequent work [16] provides a geometric interpretation on GPM through the lens of loss landscape sharpness. In particular, the covariance structure of hidden activations is closely related to the local curvature (or Fisher information) of the task loss. From this perspective, the principal bases identified by GPM can be understood as capturing directions along which the previous-task loss is most sensitive. Constraining updates to their orthogonal complement therefore implicitly favours movements in flatter regions of the loss landscape, reducing interference without requiring explicit second-order computation.

In summary, the literature provides three distinct yet interconnected lenses for understanding the mechanics of forgetting: **Hessian-based** methods focus on parameter sensitivity via "elastic" regularization, **gradient-based** approaches emphasize directional alignment to ensure non-negative inner products with previous task gradients, and **representation-based** projections like GPM leverage the low-rank structure of the latent space to navigate around sensitive regions.

These perspectives collectively suggest that forgetting is governed by how update directions interact with sensitive regions of previous tasks, motivating a more structural, direction-aware analysis beyond simple regularization or orthogonality constraints.

2.4 Outline of the Analytical Programme

The preceding sections identify the internally generated learning signal as the central object of analysis (Section 2.1), introduce the SwAV objective that gives this signal a concrete mathematical form (Section 2.2), and review how the continual learning literature has approached forgetting through the lens of parameter-space sensitivity and

directional control (Section 2.3).

The next chapter begins by analysing the mathematical structure of the SwAV learning signal: how the gradient is parametrised by the discrepancy between predicted and target assignments, and what second-order geometry this induces. On the basis of these structural results, we then formalize plasticity and forgetting as quantities governed by the same discrepancy signal, and ask whether the resulting forgetting bound admits a decomposition that separates the strength of the learning signal from its directional interaction with old-task sensitivity. The details of this decomposition, and the empirical conditions on which it depends, are the subject of Chapters 3 and 4.

Chapter 3

The Learning Signal in SwAV: Structure and Consequences

Chapter 2 motivated the choice of SwAV as an analytically tractable SSL framework (Section 2.1), reviewed its loss formulation (Section 2.2), and surveyed how the continual learning literature has approached forgetting through parameter-space sensitivity and directional control (Section 2.3). This chapter develops the mathematical machinery that connects these elements through a single object: the assignment disagreement $g = p - q$.

We proceed in four stages. First, we analyse the gradient structure of the SwAV loss and show that all first-order updates are parametrised by g (Section 3.1). Second, we examine the second-order geometry and identify a low-rank bottleneck through which all curvature interactions are mediated (Section 3.2). Third, we formalize plasticity and show that it is governed by the magnitude of g (Section 3.3). Fourth, and most substantially, we develop a forgetting bound in which the standard Hessian quadratic form is decomposed into a scale factor $\|\bar{g}\|^2$ and a normalised directional ratio in $[0, 1]$ that measures alignment with old-task high-curvature modes (Section 3.4). This decomposition is the central theoretical result of the thesis: it establishes that plasticity and stability need not trade off through magnitude alone, since directional control provides an independent degree of freedom.

3.1 Disagreement-Span Structure of SwAV Gradients

While the original SwAV objective employs a two-view swapped prediction loss, we simplify the formulation here to a single-view per sample for clarity. This simplification is justified as each view contributes an additive term with an identical gradient structure.

The per-sample training loss is defined as the cross-entropy between the target code $q(x)$ and the soft prediction $p(x)$. We denote $w \triangleq (\theta, C)$ as the set of trainable parameters, where θ represents the parameters of the encoder f_θ and C represents the prototypes.

The loss is expressed as:

$$\ell(x; w) = -\sum_{k=1}^{K} q_k(x) \log p_k(x). \tag{3.1}$$

By taking the expectation over the data distribution $x \sim \mathcal{D}$, the global objective is expressed as $L(w) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(x; w)]$. Recall that the similarity logit between an embedding z and prototype c_k is $s_k = z^\top c_k/\tau$ (Eq. 2.1). We regard q as fixed, with each sample-wise assignment vector normalised to sum to one, interpretable as a probability distribution over the K prototypes for each sample. The following Lemma characterizes the structural constraints on the gradients of this objective, which we term the *Disagreement-span structure*.

Lemma 3.1. *For any sample x with representation $z = z_\theta(x)$ and codebook $C = \{c_k\}_{k=1}^K$, there exists a disagreement vector $\Delta(x) \in \mathbb{R}^K$ such that:*

1. *The gradient with respect to the representation lies in the span of the prototypes:*

$$\nabla_z \ell(x; w) = \frac{1}{\tau} \sum_{k=1}^{K} \Delta_k(x) c_k \ \in \ \text{span}\{c_1, \ldots, c_K\}. \tag{3.2}$$

2. *For each prototype c_k, the gradient lies in the one-dimensional subspace spanned by the representation z:*

$$\nabla_{c_k} \ell(x; w) = \frac{1}{\tau} \Delta_k(x) z \ \in \ \text{span}\{z\}. \tag{3.3}$$

Proof. Given the exponential form of the softmax, the per-sample loss in Eq. 3.1 can be expanded as:

$$\ell(x; w) = -\sum_{k=1}^{K} q_k(x) \left(\frac{z^\top c_k}{\tau} \right) + \log \left(\sum_{j=1}^{K} e^{z^\top c_j/\tau} \right).$$

The derivative of the loss with respect to the similarity logit $s_k(x)$ (Eq. 2.1) is:

$$\frac{\partial \ell(x; w)}{\partial s_k(x)} = -q_k(x) + \frac{e^{s_k(x)}}{\sum_{j=1}^{K} e^{s_j(x)}} = p_k(x) - q_k(x). \tag{3.4}$$

Applying the chain rule, the gradient with respect to the representation z is:

$$\nabla_z \ell(x; w) = \sum_{k=1}^{K} \frac{\partial \ell}{\partial s_k} \nabla_z s_k = \frac{1}{\tau} \sum_{k=1}^{K} (p_k(x) - q_k(x)) c_k.$$

By identifying $\Delta_k(x) \triangleq p_k(x) - q_k(x)$, we prove Eq. 3.2. Similarly, the gradient with respect to the k-th prototype c_k is:

$$\nabla_{c_k} \ell(x; w) = \frac{\partial \ell}{\partial s_k} \nabla_{c_k} s_k = \frac{1}{\tau} (p_k(x) - q_k(x)) z,$$

which confirms Eq. 3.3. □

Lemma 3.1 establishes that the SwAV update direction is fully parameterized by the assignment disagreement $\Delta(x) = p(x) - q(x)$, together with the current codebook geometry. In particular, the representation update lies in the linear span of the prototypes, while prototype updates are modulated by the representations through the same disagreement signal. Consequently, the geometry of the update is not arbitrary, but structured: the available update directions in embedding space are governed by the span and rank of the codebook C, and its magnitude and sign are controlled by the disagreement Δ. Throughout the remainder of this thesis, we write $g \triangleq p - q$ for this disagreement vector, reserving the notation $\Delta_k(x)$ for contexts where the per-sample, per-prototype indexing is needed.

3.2 Low-Rank Geometry of SwAV Curvature

The disagreement-span structure characterizes the first-order direction of SwAV updates. We now extend this view to the second-order geometry, asking how sensitive the loss is to perturbations of the representation z and the prototypes C. This extension is necessary because forgetting, as will be formalized in Section 3.4, depends not only on the direction of parameter updates but also on the second-order sensitivity of the old-task loss along those directions.

Curvature induced by a low-dimensional logit channel. In SwAV the similarity logits are bilinear in (z, C), e.g., $s = \tau^{-1} C^\top z$ (up to normalisation/temperature). By the chain rule, the Hessian with respect to z takes the form

$$H_{zz} = J_z^\top H_{ss} J_z + \sum_{k=1}^{K} \frac{\partial \ell}{\partial s_k} \nabla_z^2 s_k, \tag{3.5}$$

where $J_z = \partial s / \partial z$ and $H_{ss} = \nabla_s^2 \ell$. Since s is linear in z, the second term vanishes, yielding

$$H_{zz} = J_z^\top H_{ss} J_z.$$

The same logic applies to prototypes: because s is also linear in C (for fixed z),

$$H_{CC} = J_C^\top H_{ss} J_C, \qquad H_{zC} = J_z^\top H_{ss} J_C, \tag{3.6}$$

where $J_C = \partial s / \partial \mathrm{vec}(C)$ and H_{zC} is the cross block capturing the bilinear coupling between representation and prototype perturbations. We will not attempt to eliminate H_{zC} by assumption; instead, Eq. 3.6 already shows that all second-order interactions are mediated through the same K-dimensional logit channel.

Why "low-rank" geometry is intrinsic here. The key implication of Eq. 3.6 is an implicit low-rank structure. For softmax cross-entropy, the logit-level Hessian takes the

form [3]

$$H_{ss} = \mathrm{diag}(p) - pp^\top. \tag{3.7}$$

Since p is a valid probability distribution, we have $\mathrm{diag}(p)\mathbf{1} = p$ and $p^\top\mathbf{1} = 1$. It follows that

$$H_{ss}\mathbf{1} = (\mathrm{diag}(p) - pp^\top)\mathbf{1} = \mathrm{diag}(p)\mathbf{1} - p(p^\top\mathbf{1}) = p - p \cdot 1 = 0. \tag{3.8}$$

This confirms that the all-ones vector $\mathbf{1}$ resides in the nullspace of H_{ss}, bounding its rank to at most $K - 1$.

Moreover, the Jacobians $J_z \in \mathbb{R}^{K\times d}$ and $J_C \in \mathbb{R}^{K\times(Kd)}$ act as structural bottlenecks. They map perturbations from the d-dimensional embedding space and the Kd-dimensional prototype space, respectively, down into the restricted K-dimensional logit space. Consequently, the pullback Hessians H_{zz}, H_{CC}, and the cross block H_{zC} inherit a geometry whose effective action is strictly confined to a subspace shaped by the spectrum of H_{ss}. In other words, the curvature experienced in the high-dimensional parameter space is mediated through a K-dimensional channel, giving rise to an intrinsic low-rank structure.

This low-rank structure at the level of the loss formulation suggests that the curvature landscape may be tractable: although the network operates in a massively overparameterized domain, the critical sensitivities may be governed by a low-dimensional, prototype-dependent structure. Whether this property propagates from the representation-level Hessian (H_{zz}, H_{CC}) to the parameter-space Hessian ($H_{\theta\theta}$) is not guaranteed by the analysis above and must be verified empirically. If it does, however, then characterizing forgetting reduces to tracking how updates project onto a narrow subspace, rather than monitoring dense, global parameter interactions. The formalization in the next two subsections develops this idea, and Chapter 4 tests whether the required spectral concentration holds in practice.

3.3 Plasticity: Governed by the Magnitude of the Disagreement

We now formalize how the learning signal governs plasticity. We retain the single-view simplification and stop-gradient treatment of Q introduced in Section 3.1, and consider a sequence of tasks $\mathrm{Task}_1, \mathrm{Task}_2, \ldots$.

Each Task_t is associated with a data distribution $\mathcal{D}_t$ and the corresponding loss

$$L_t(w) \triangleq \mathbb{E}_{x\sim\mathcal{D}_t}\big[\ell(x; w)\big].$$

Let the optimal parameters of Task_t be defined as

$$(\theta_t^\star, C_t^\star) \in \arg\min_{\theta, C} L_t(\theta, C), \qquad w_t^\star \triangleq (\theta_t^\star, C_t^\star).$$

We assume that the parameters $w_{t+1}^\star$ are obtained by applying stochastic gradient descent on the Task_{t+1} objective, initialized at $w_t^\star$.

Definition 3.2 (Plasticity)**.** Plasticity on Task_{t+1} is defined as the reduction in loss achieved by training on that task. Formally,

$$P_{t+1} \triangleq L_{t+1}(w_t^\star) - L_{t+1}(w_{t+1}^\star), \tag{3.9}$$

where $w_t^\star$ is the parameter at the start of training on Task_{t+1} (i.e., the previous-task solution) and $w_{t+1}^\star$ is the parameter obtained after training on Task_{t+1}.

To analyse plasticity, we decompose the loss reduction into two components corresponding to backbone and prototype updates.

Lemma 3.3. *Define the auxiliary point* $\tilde{w} \triangleq (\theta_{t+1}^\star, C_t^\star)$*. Then the plasticity on* Task_{t+1} *admits the exact decomposition*

$$\begin{aligned} P_{t+1} &= L_{t+1}(w_t^\star) - L_{t+1}(\tilde{w}) + L_{t+1}(\tilde{w}) - L_{t+1}(w_{t+1}^\star) \\ &= \underbrace{L_{t+1}(\theta_t{}^\star, C_t^\star) - L_{t+1}(\theta_{t+1}^\star, C_t^\star)}_{\triangleq P_{t+1}^{(\theta)}} + \underbrace{L_{t+1}(\theta_{t+1}^\star, C_t^\star) - L_{t+1}(\theta_{t+1}^\star, C_{t+1}^\star)}_{\triangleq P_{t+1}^{(C)}} \end{aligned}$$

and hence satisfies the bound

$$|P_{t+1}| \leq |P_{t+1}^{(\theta)}| + |P_{t+1}^{(C)}|. \tag{3.10}$$

Proof. The decomposition follows by adding and subtracting $L_{t+1}(\tilde{w})$, which induces a cancellation of terms. Applying the triangle inequality then yields an upper bound consisting of two components, corresponding to backbone-driven and prototype-driven effects. □

$P_{t+1}^{(\theta)}$ quantifies the loss reduction on Task_{t+1} when the backbone changes from the initial parameter at the start of training on Task_{t+1}, $\theta_t^\star$, to $\theta_{t+1}^\star$, while the prototypes are held fixed at the initial value $C_t^\star$. Similarly, $P_{t+1}^{(C)}$ isolates the prototype-driven contribution: it measures the loss reduction when only the prototypes change from $C_t^\star$ to $C_{t+1}^\star$ while the backbone is held fixed at $\theta_{t+1}^\star$.

Importantly, standard training does not perform such decoupled updates (i.e., updating only θ or only C in isolation). These quantities are introduced solely as an analytical decomposition to analyse the overall plasticity from two complementary perspectives. Accordingly, the auxiliary point $\tilde{w}$ is used purely for analytical purposes to separate these effects, and is not assumed to be visited by the optimisation trajectory.

Further, the previous-task optimum $w_t^\star$ is used only as a convenient and explicit reference point to anchor the comparison. The same decomposition and subsequent bounds can be formulated with any reference w in the parameter domain for which $L_{t+1}(w)$ is finite.

In particular, the plasticity analysis does not rely on any special property of $w_t^\star$ beyond being a well-defined reference parameter.

Assumption 3.4. *There exists $L_z > 0$ such that for all x and any $(\theta, C), (\theta', C)$ (with C fixed),*

$$\left|\ell(x;\theta,C) - \ell(x;\theta',C)\right| \le L_z \left\|z_\theta(x) - z_{\theta'}(x)\right\|.$$

That is, the loss $\ell(x;\theta,C)$ depends on θ only through the representation $z_\theta(x)$ and is Lipschitz continuous with respect to this representation.

Assumption 3.5. *Along the optimisation trajectory on* Task_{t+1}*, there exists a constant $\kappa > 0$ such that for all x,*

$$\left\|z_{\theta_{t+1}^\star}(x) - z_{\theta_t^\star}(x)\right\| \le \kappa \sum_s \eta_s \left\|\nabla_z \mathcal{L}^{(s)}(x)\right\|,$$

where s indexes the update steps on Task_{t+1} *and η_s denotes the corresponding step size. This assumption upper bounds the representation drift along the optimisation trajectory by the accumulated gradient magnitudes in representation space.*

These two technical assumptions control how changes in the representation translate into changes in the loss. The arguments supporting Assumptions 3.4 and 3.5 are given in Appendix A.2.

Lemma 3.6. *Under Assumptions 3.4 and 3.5, the backbone-driven plasticity term*

$$P_{t+1}^{(\theta)} = L_{t+1}(\theta_t^\star, C_t^\star) - L_{t+1}(\theta_{t+1}^\star, C_t^\star)$$

satisfies

$$\left|P_{t+1}^{(\theta)}\right| \;\le\; \frac{L_z\,\kappa}{\tau} \sum_s \eta_s \left\|C^{(s)}\right\|_{\mathrm{op}} \mathbb{E}_{x\sim\mathcal{D}_{t+1}}\left[\left\|p^{(s)}(x) - q^{(s)}(x)\right\|\right]. \tag{3.11}$$

Consequently, backbone-driven plasticity on Task_{t+1} *is governed by the cumulative assignment disagreement $\|p - q\|$ along the optimisation trajectory.*

Proof. By Assumption 3.4 with prototypes fixed at $C_t^\star$,

$$\left|P_{t+1}^{(\theta)}\right| \le L_z\, \mathbb{E}_{x\sim\mathcal{D}_{t+1}}\left[\left\|z_{\theta_t^\star}(x) - z_{\theta_{t+1}^\star}(x)\right\|\right].$$

Applying Assumption 3.5 yields

$$\left|P_{t+1}^{(\theta)}\right| \le L_z\,\kappa \sum_s \eta_s\, \mathbb{E}_{x\sim\mathcal{D}_{t+1}}\left[\left\|\nabla_z \mathcal{L}^{(s)}(x)\right\|\right]. \tag{3.12}$$

The gradient w.r.t. the representation (Eq. 3.2) satisfies

$$\nabla_z \mathcal{L}^{(s)}(x) = \frac{1}{\tau} C^{(s)}\left(p^{(s)}(x) - q^{(s)}(x)\right),$$

and hence

$$\|\nabla_z \mathcal{L}^{(s)}(x)\| \le \frac{1}{\tau} \|C^{(s)}\|_{\mathrm{op}} \|p^{(s)}(x) - q^{(s)}(x)\|.$$

Substituting this bound into Eq. 3.12 yields Eq. 3.11, which completes the proof. □

Assumption 3.7. *There exists a constant $\beta_C > 0$ such that for any fixed backbone θ and for all prototype matrices C, C',*

$$L_{t+1}(\theta, C') \le L_{t+1}(\theta, C) + \langle \nabla_C L_{t+1}(\theta, C),\, C' - C \rangle + \frac{\beta_C}{2} \|C' - C\|_F^2.$$

Assumption 3.8. *Along the optimisation trajectory on* Task_{t+1}*, there exists $R_z > 0$ such that for all x and all steps s,*

$$\|z_{\theta^{(s)}}(x)\| \le R_z.$$

In the SwAV framework, the representation z is ℓ_2-normalised at the output of the projection head before computing similarity logits, so $\|z_{\theta^{(s)}}(x)\| = 1$ for all x and all steps s. Assumption 3.8 is therefore satisfied with $R_z = 1$. The argument supporting Assumption 3.7 is given in Appendix A.2.

Lemma 3.9. *Under Assumptions 3.7 and 3.8, define $\Delta C \triangleq C^\star_{t+1} - C^\star_t$, the prototype-driven plasticity term*

$$P^{(C)}_{t+1} = L_{t+1}(\theta^\star_{t+1}, C^\star_t) - L_{t+1}(\theta^\star_{t+1}, C^\star_{t+1})$$

satisfies

$$\left|P^{(C)}_{t+1}\right| \;\le\; \frac{R_z}{\tau}\, \mathbb{E}_{x\sim\mathcal{D}_{t+1}}\left[\|p(x) - q(x)\|\right] \|\Delta C\|_F \;+\; \frac{\beta_C}{2} \|\Delta C\|_F^2. \tag{3.13}$$

Consequently, for fixed prototype displacement $\|\Delta C\|_F$, the prototype-driven plasticity is governed by the assignment disagreement $\|p - q\|$ on Task_{t+1}*, up to the second order term.*

Proof. By Assumption 3.7 with backbone fixed at $\theta^\star_{t+1}$,

$$P^{(C)}_{t+1} = L_{t+1}(\theta, C^\star_t) - L_{t+1}(\theta, C^\star_{t+1}) \le -\langle \nabla_C L_{t+1}(\theta, C^\star_t),\, \Delta C \rangle + \frac{\beta_C}{2} \|\Delta C\|_F^2.$$

Taking absolute values and using Cauchy–Schwarz yields

$$\left|P^{(C)}_{t+1}\right| \le \left\|\nabla_C L_{t+1}(\theta, C^\star_t)\right\|_F \|\Delta C\|_F + \frac{\beta_C}{2} \|\Delta C\|_F^2. \tag{3.14}$$

The gradient w.r.t. prototypes (Eq. 3.3) satisfies

$$\nabla_C \ell(x; \theta, C) = \frac{1}{\tau}\, z_\theta(x)\, (p(x) - q(x))^\top,$$

hence

$$\|\nabla_C \ell(x;\theta,C)\|_F = \frac{1}{\tau}\,\|z_\theta(x)\|\,\|p(x)-q(x)\|.$$

Using $\nabla_C L_{t+1}(\theta,C) = \mathbb{E}_{x\sim\mathcal{D}_{t+1}}[\nabla_C \ell(x;\theta,C)]$ and Jensen's inequality,

$$\big\|\nabla_C L_{t+1}(\theta,C_t^\star)\big\|_F = \Big\|\mathbb{E}[\nabla_C \ell(x;\theta,C_t^\star)]\Big\|_F \le \mathbb{E}\big[\|\nabla_C \ell(x;\theta,C_t^\star)\|_F\big].$$

By Assumption 3.8, $\|z_\theta(x)\| \le R_z$ for $x \sim \mathcal{D}_{t+1}$, thus

$$\big\|\nabla_C L_{t+1}(\theta,C_t^\star)\big\|_F \le \frac{R_z}{\tau}\,\mathbb{E}_{x\sim\mathcal{D}_{t+1}}\big[\|p(x)-q(x)\|\big].$$

Substituting this bound into Eq. 3.14 yields Eq. 3.13, which completes the proof. □

3.3.1 Plasticity Governed by Assignment Disagreement

When considering backbone-driven plasticity in isolation, the assignment disagreement $p-q$ alone dominates representation updates. In this regime, representation changes arise entirely through linear mappings of the disagreement signal, while the prototype matrix C acts only as a sensitivity modulator that controls the scale and conditioning of these updates. This perspective explains why methods such as SeLa [1], which rely on assignments computed from a fixed and randomly initialized classification head, can still learn effective representations: although the prototypes are not learnable, they serve as fixed anchors that translate assignment disagreement into backbone updates.

In contrast, when the representation z is held fixed, updating prototypes alone is sufficient to reduce the assignment-matching loss. In this case, larger prototype movements enhance plasticity in a manner analogous to centroid updates in k-means–style clustering.

Taken together, these two components suggest that plasticity is fundamentally governed by the magnitude of the assignment disagreement, which the gradient structure encourages to be large. More counterintuitively, the plasticity bounds reveal that prototypes do not function as nearest-neighbor classifiers providing discriminative supervision at the sample level. Rather, they condition the flow of assignment disagreement into representation updates, implicitly regularizing the geometry and diversity of the learned representations.

3.4 Forgetting: From Magnitude to Direction

3.4.1 Definition and Sharpness-Based Stability Rationale

Definition 3.10 (Forgetting)**.** Forgetting on Task_t (after learning Task_{t+1}) is defined as the increase in the Task_t loss incurred by training on subsequent tasks, measured as the difference between the Task_t loss at the parameter optimum for Task_t, $w_t^\star$, and the Task_t loss at a parameter obtained after training on later tasks, e.g., the Task_{t+1}

optimum $w_{t+1}^\star$. Formally,

$$F_t \triangleq L_t(w_{t+1}^\star) - L_t(w_t^\star), \qquad \Delta w \triangleq w_{t+1}^\star - w_t^\star. \tag{3.15}$$

As in the plasticity analysis, the anchor $w_t^\star$ is introduced for analytical clarity. In practice, strict stationarity $\nabla L_t(w_t^\star) = \mathbf{0}$ is rarely verifiable in deep networks, so a Taylor expansion of $L_t(w_t^\star + \Delta w)$ around $w_t^\star$ generally contains a linear drift term $\nabla L_t(w_t^\star)^\top \Delta w$ in addition to the quadratic curvature term. Controlling F_t solely by shrinking $\|\Delta w\|$ is incompatible with the plasticity objective established in Section 3.3. We therefore seek a stability rationale that does not rely on strict first-order optimality and does not collapse forgetting to update magnitude alone.

A Sharpness-Based Stability Rationale. Rather than treating stability as a purely point-wise property of L_t at a single anchor, we adopt a neighbourhood-based perspective inspired by sharpness-aware analysis (SAM [19]). For a radius $\rho > 0$, define the Task$_t$ sharpness functional

$$\mathcal{S}_t(w;\rho) \triangleq \max_{\|\delta\| \le \rho} L_t(w+\delta),$$

and, for intuition only, the corresponding neighbourhood-sensitivity gap

$$F_t^{\mathrm{sh}}(\rho) \triangleq \mathcal{S}_t(w_{t+1}^\star;\rho) - \mathcal{S}_t(w_t^\star;\rho).$$

This quantity is introduced only to distinguish point-wise loss change from local sensitivity in a neighbourhood of the anchor. We do not replace the standard forgetting metric F_t by $F_t^{\mathrm{sh}}(\rho)$. Its role is purely interpretive: it motivates viewing forgetting through curvature-mediated neighbourhood sensitivity rather than allowing the analysis to be dictated by point-wise linear drift alone.

Motivated by this local-stability viewpoint, we directly encode the forgetting behaviour of the standard metric F_t through the following curvature-centred modelling assumption.

Assumption 3.11 (Curvature-dominated proxy upper bound)**.** *Let $\Delta w = w_{t+1}^\star - w_t^\star$. There exists a nonnegative remainder term $\mathcal{R}_t(\Delta w)$ such that the forgetting metric F_t satisfies*

$$F_t \le \frac{1}{2}\Delta w^\top \nabla^2 L_t(w_t^\star)\,\Delta w + \mathcal{R}_t(\Delta w), \tag{3.16}$$

where the remainder is bounded by the non-stationary linear drift and a higher-order residual ϵ_t:

$$\mathcal{R}_t(\Delta w) \le \left\|\nabla L_t(w_t^\star)\right\| \|\Delta w\| + \epsilon_t(\|\Delta w\|). \tag{3.17}$$

Justification. We do not treat Assumption 3.11 as derived from the sharpness proxy above. Rather, it is a curvature-centred local model for the standard forgetting metric F_t. By a standard third-order Taylor expansion, F_t can be decomposed into a quadratic term together with a remainder, where the latter collects the non-stationary linear drift and higher-order variations. The sharpness-based rationale explains why this decomposition

should be interpreted through local stability: the linear term is treated as a controlled residual, while the quadratic component captures the directional sensitivity of the old-task landscape to the update Δw.

3.4.2 Block Decomposition

Assumption 3.12. *Let the objective Hessian of* Task_t *with respect to the combined parameter* $w \triangleq (\theta, C)$ *be block-partitioned as*

$$H_t \triangleq \nabla^2 L_t(w) = \begin{bmatrix} H_{\theta\theta}^{(t)} & H_{\theta C}^{(t)} \\ H_{C\theta}^{(t)} & H_{CC}^{(t)} \end{bmatrix}.$$

For the update $\Delta w \triangleq (\Delta\theta, \Delta C)$ *considered in this paper, there exist constants* $\alpha_\theta, \alpha_C \in [0, 1)$ *such that the quadratic form admits the bound*

$$\Delta w^\top H_t \Delta w \;\le\; \frac{1}{1-\alpha_\theta}\,\Delta\theta^\top H_{\theta\theta}^{(t)}\Delta\theta + \frac{1}{1-\alpha_C}\,\Delta C^\top H_{CC}^{(t)}\Delta C. \tag{3.18}$$

Justification and interpretation. Assumption 3.12 is stated as a direction-dependent inequality along the specific training-induced update Δw. At a high level, it asserts that the mixed coupling contribution $2\,\Delta\theta^\top H_{\theta C}^{(t)}\Delta C$ can be absorbed into constant factors. A sufficient condition for this absorption is that the cross-block is relatively bounded with respect to the diagonal blocks: there exists a constant $\beta_t \ge 0$ such that

$$\left|\Delta\theta^\top H_{\theta C}^{(t)}\Delta C\right| \;\le\; \beta_t \sqrt{\Delta\theta^\top H_{\theta\theta}^{(t)}\Delta\theta}\,\sqrt{\Delta C^\top H_{CC}^{(t)}\Delta C}.$$

Under this condition, choosing $\alpha_\theta, \alpha_C \in [0, 1)$ such that $\alpha_\theta\alpha_C \ge \beta_t^2$ allows the mixed term to be absorbed via a Young-type inequality, yielding (3.18). Sufficient conditions and a constructive derivation are provided in Appendix A.1.

The empirical validity of this assumption is tested in Section 4.2.1, by measuring the relative magnitude of the mixed contribution $|2\,\Delta\theta^\top H_{\theta C}^{(t)}\Delta C|$ against the block-diagonal energies $\Delta\theta^\top H_{\theta\theta}^{(t)}\Delta\theta$ and $\Delta C^\top H_{CC}^{(t)}\Delta C$.

Lemma 3.13. *Let* $\Delta w := w_{t+1}^\star - w_t^\star$ *and write* $\Delta w = (\Delta\theta, \Delta C)$. *Under Assumption 3.11 and Assumption 3.12 with* $\alpha_\theta, \alpha_C \in [0, 1)$, *the old-task forgetting satisfies*

$$F_t \;\le\; \frac{1}{2(1-\alpha_\theta)}\,\Delta\theta^\top H_{\theta\theta}^{(t)}\Delta\theta + \frac{1}{2(1-\alpha_C)}\,\Delta C^\top H_{CC}^{(t)}\Delta C + \mathcal{R}_t(\Delta w), \tag{3.19}$$

Proof. By Assumption 3.11 and 3.12, substituting Eq. 3.18 into Eq. 3.16 yields Eq. 3.19. □

The use of parameter displacements in a second-order forgetting bound is standard: by a local Taylor expansion, forgetting can be upper bounded by a Hessian quadratic form in the displacement Δw, as in [39, 40]. The purpose of the block decomposition is to expose, inside this quadratic form, the low-dimensional SwAV learning signal that

already appears in the plasticity analysis, namely the assignment disagreement $g = p - q$. If forgetting is analysed solely at the level of an undifferentiated Hessian quadratic form, the structural role of g is completely obscured: the displacement Δw appears as an arbitrary vector in parameter space, and no direct connection to the intrinsic disagreement signal remains visible. Only after separating the Hessian into backbone and prototype blocks can we meaningfully trace how g influences each component of the update and how it enters the forgetting bound.

3.4.3 Trajectory-Level Disagreement Operators

Definition 3.14. For any two training time indices $t_1 < t_2$, consider the segment of the training trajectory consisting of S consecutive optimisation steps indexed by $s = 1, \dots, S$ between t_1 and t_2. Denote the parameter differences

$$\Delta\theta := \theta(t_2) - \theta(t_1), \qquad \Delta C := C(t_2) - C(t_1).$$

For each step s, denote the assignment disagreement vector

$$g^{(s)} := p^{(s)} - q^{(s)} \in \mathbb{R}^K,$$

as introduced in Eq. (3.2) and Eq. (3.3). Collect all stepwise disagreements into the stacked vector

$$\bar{g} := \left[(g^{(1)})^\top, \dots, (g^{(S)})^\top\right]^\top \in \mathbb{R}^{KS}.$$

Define the linear operators $T_\theta \in \mathbb{R}^{d_\theta \times KS}$ and $T_C \in \mathbb{R}^{(d_z K)\times KS}$ by

$$T_\theta := -\frac{1}{\tau}\left[\eta_1 \left(J_{\theta^{(1)}} z_{\theta^{(1)}}(x^{(1)})\right)^\top C^{(1)} \quad \cdots \quad \eta_S \left(J_{\theta^{(S)}} z_{\theta^{(S)}}(x^{(S)})\right)^\top C^{(S)}\right],$$

$$T_C := -\frac{1}{\tau}\left[\eta_1 \left(I_K \otimes z_{\theta^{(1)}}(x^{(1)})\right) \quad \cdots \quad \eta_S \left(I_K \otimes z_{\theta^{(S)}}(x^{(S)})\right)\right],$$

where η_s is the step size, τ is the temperature, $z_{\theta^{(s)}}(x^{(s)}) \in \mathbb{R}^{d_z}$ is the representation at step s,

$$J_{\theta^{(s)}} z_{\theta^{(s)}}(x^{(s)}) \triangleq \frac{\partial z_\theta(x^{(s)})}{\partial \theta}\Big|_{\theta=\theta^{(s)}} \in \mathbb{R}^{d_z \times d_\theta}$$

is its Jacobian with respect to θ, $C^{(s)}$ is the prototype matrix at step s, I_K is the $K \times K$ identity, and $\otimes$ denotes the Kronecker product.

Then the parameter differences admit the disagreement-induced linear representation

$$\Delta\theta \;=\; T_\theta\, \bar{g}, \qquad \mathrm{vec}(\Delta C) \;=\; T_C\, \bar{g},$$

where $\mathrm{vec}(\cdot)$ denotes column-wise vectorization.

Equipped with the above definition, we can now express the old-task forgetting directly in terms of the low-dimensional disagreement signal. This yields our main trajectory-level bound:

Proposition 3.15 (Disagreement-explicit forgetting bound)**.** *Under the conditions of Lemma 3.13 and Definition 3.14, let $\bar{g} \in \mathbb{R}^{KS}$ be the stacked disagreement vector over the training segment. Then the old-task forgetting in Eq. (3.19) can be reformulated as:*

$$F_t \;\le\; \frac{1}{2(1-\alpha_\theta)}\,\bar{g}^\top T_\theta^\top H_{\theta\theta}^{(t)} T_\theta \bar{g} \;+\; \frac{1}{2(1-\alpha_C)}\,\bar{g}^\top T_C^\top H_{CC}^{(t)} T_C \bar{g} \;+\; \mathcal{R}_t(\Delta w).$$

Proof. The result follows directly by substituting the disagreement-induced linear representations $\Delta\theta = T_\theta \bar{g}$ and $\mathrm{vec}(\Delta C) = T_C \bar{g}$ from Definition 3.14 into the block-absorbed forgetting bound in Lemma 3.13. □

Motivation: separating scale from old-task directional sensitivity. In Proposition 3.15, the quadratic form $\bar{g}^\top T^\top H^{(t)} T \bar{g}$ entangles two distinct effects: (i) the *scale* of the disagreement accumulated along the trajectory segment, captured by $\|\bar{g}\|$, and (ii) how the induced parameter displacement is distributed relative to the old-task curvature structure. To isolate these effects without imposing any orthogonality or non-overlap assumptions between old and new tasks, next, we upper bound the Hessian quadratic form by separating a uniform low-curvature level λ from the contribution of directions whose curvature exceeds λ. This construction prepares a structural decomposition that will ultimately enable us to express $\|\bar{g}\|^2$ as a multiplicative scale factor, with the remaining coefficient capturing a normalised directional weight in $[0, 1]$ that measures the fraction of disagreement-induced displacement lying within the old-task high-curvature subspace.

3.4.4 Scale-Direction Separation

Definition 3.16. Let H be a symmetric matrix with operator norm $\|H\|_{\mathrm{op}} \le \Lambda$. For a threshold $\lambda \in [0, \Lambda]$, let

$$P_{\mathrm{high}}(\lambda) := \sum_{\lambda_i(H) \ge \lambda} u_i u_i^\top, \qquad P_{\mathrm{low}}(\lambda) := \sum_{\lambda_i(H) < \lambda} u_i u_i^\top,$$

where $\{(\lambda_i(H), u_i)\}$ denotes an eigendecomposition of H.

By construction,

$$P_{\mathrm{high}}(\lambda) + P_{\mathrm{low}}(\lambda) = I, \qquad P_{\mathrm{high}}(\lambda) P_{\mathrm{low}}(\lambda) = 0, \qquad P_{\mathrm{low}}(\lambda) = I - P_{\mathrm{high}}(\lambda).$$

Proposition 3.17 (High-curvature split of the forgetting bound)**.** *Assume Lemma 3.13. Suppose $\|H_{\theta\theta}^{(t)}\|_{\mathrm{op}} \le \Lambda_\theta^{(t)}$ and $\|H_{CC}^{(t)}\|_{\mathrm{op}} \le \Lambda_C^{(t)}$.*

Fix any threshold $\lambda \in \big(0, \min\{\Lambda_\theta^{(t)}, \Lambda_C^{(t)}\}\big]$, and define the projectors

$$P_{\theta,\mathrm{high}}(\lambda) := P_{\mathrm{high}}^{H_{\theta\theta}^{(t)}}(\lambda), \qquad P_{C,\mathrm{high}}(\lambda) := P_{\mathrm{high}}^{H_{CC}^{(t)}}(\lambda).$$

Then the forgetting satisfies

$$F_t \le \frac{1}{2(1-\alpha_\theta)}\Big[\lambda\|\Delta\theta\|^2 + (\Lambda_\theta^{(t)} - \lambda)\|P_{\theta,\mathrm{high}}(\lambda)\Delta\theta\|^2\Big] + \frac{1}{2(1-\alpha_C)}\Big[\lambda\|\mathrm{vec}(\Delta C)\|^2 + (\Lambda_C^{(t)} - \lambda)\|P_{C,\mathrm{high}}(\lambda)\mathrm{vec}(\Delta C)\|^2\Big] + \mathcal{R}_t(\Delta w).$$

Proof. Following the notations in Proposition 3.17 and Definition 3.16 and using $\lambda_i(H) \le \Lambda$, we obtain

$$\begin{aligned} v^\top H v &= \sum_i \lambda_i(H)\,\langle v, u_i\rangle^2 \\ &\le \Lambda \sum_{\lambda_i(H)\ge\lambda} \langle v, u_i\rangle^2 + \lambda \sum_{\lambda_i(H)<\lambda} \langle v, u_i\rangle^2 \\ &= \Lambda\|P_{\mathrm{high}}(\lambda)v\|^2 + \lambda\|P_{\mathrm{low}}(\lambda)v\|^2. \end{aligned}$$

Substituting $P_{\mathrm{low}}(\lambda) = I - P_{\mathrm{high}}(\lambda)$ and using orthogonality of the projectors give

$$\|v\|^2 = \|P_{\mathrm{high}}(\lambda)v\|^2 + \|P_{\mathrm{low}}(\lambda)v\|^2,$$

hence

$$v^\top H v \le \lambda\|v\|^2 + (\Lambda - \lambda)\|P_{\mathrm{high}}(\lambda)v\|^2. \tag{3.20}$$

Applying (3.20) to $H = H_{\theta\theta}^{(t)}$ with $v = \Delta\theta$ and to $H = H_{CC}^{(t)}$ with $v = \mathrm{vec}(\Delta C)$, and substituting the resulting inequalities into Lemma 3.13 yields the claimed bound. □

Proposition 3.17 shows that, at the level of parameter displacements, the quadratic forgetting term admits a decomposition in which the magnitude $\|\Delta w\|$ (through $\|\Delta\theta\|^2$ and $\|\mathrm{vec}(\Delta C)\|^2$) is explicitly separated from the contribution of the old-task high-curvature subspaces P_{high}. In other words, the scale of the update has already been isolated at the level of the block-wise quadratic form.

We now reintroduce the disagreement representation $\Delta\theta = T_\theta\bar{g}$ and $\mathrm{vec}(\Delta C) = T_C\bar{g}$ (Definition 3.14) to express this high-curvature split directly in terms of the low-dimensional signal $\bar{g}$. This yields the following directional bound.

Corollary 3.18. *Assume Proposition 3.17 and Definition 3.14. Let $\bar{g} \in \mathbb{R}^{KS}$ be the stacked disagreement vector over the trajectory segment, and define the normalised operators*

$$\Pi_\theta(\lambda) := \frac{T_\theta^\top P_{\theta,\mathrm{high}}(\lambda)T_\theta}{\|T_\theta\|_{\mathrm{op}}^2} \succeq 0, \qquad \Pi_C(\lambda) := \frac{T_C^\top P_{C,\mathrm{high}}(\lambda)T_C}{\|T_C\|_{\mathrm{op}}^2} \succeq 0.$$

Then the old-task forgetting satisfies, for any $\bar{g} \neq 0$,

$$F_t \;\leq\; \frac{1}{2}\Bigg(\frac{\|T_\theta\|_{\mathrm{op}}^2}{1-\alpha_\theta}\Big[\lambda + \big(\Lambda_\theta^{(t)} - \lambda\big)\frac{\bar{g}^\top \Pi_\theta(\lambda)\bar{g}}{\|\bar{g}\|^2}\Big] + \frac{\|T_C\|_{\mathrm{op}}^2}{1-\alpha_C}\Big[\lambda + \big(\Lambda_C^{(t)} - \lambda\big)\frac{\bar{g}^\top \Pi_C(\lambda)\bar{g}}{\|\bar{g}\|^2}\Big]\Bigg)\|\bar{g}\|^2 + \varepsilon_t(\bar{g}), \tag{3.21}$$

where $\varepsilon_t(\bar{g}) := \mathcal{R}_t\big(\Delta w(\bar{g})\big)$[1]. *Moreover,*

$$0 \leq \frac{\bar{g}^\top \Pi_\theta(\lambda)\bar{g}}{\|\bar{g}\|^2} \leq 1, \qquad 0 \leq \frac{\bar{g}^\top \Pi_C(\lambda)\bar{g}}{\|\bar{g}\|^2} \leq 1.$$

Proof. By Definition 3.14, we have

$$\|\Delta\theta\|^2 = \|T_\theta \bar{g}\|^2 \leq \|T_\theta\|_{\mathrm{op}}^2 \|\bar{g}\|^2, \qquad \|\mathrm{vec}(\Delta C)\|^2 = \|T_C \bar{g}\|^2 \leq \|T_C\|_{\mathrm{op}}^2 \|\bar{g}\|^2.$$

Moreover,

$$\|P_{\theta,\mathrm{high}}(\lambda)\Delta\theta\|^2 = \|P_{\theta,\mathrm{high}}(\lambda)T_\theta\bar{g}\|^2 = \bar{g}^\top T_\theta^\top P_{\theta,\mathrm{high}}(\lambda)T_\theta\bar{g} = \|T_\theta\|_{\mathrm{op}}^2\, \bar{g}^\top \Pi_\theta(\lambda)\bar{g},$$

and similarly

$$\|P_{C,\mathrm{high}}(\lambda)\mathrm{vec}(\Delta C)\|^2 = \|T_C\|_{\mathrm{op}}^2\, \bar{g}^\top \Pi_C(\lambda)\bar{g}.$$

Substituting these relations into Proposition 3.17 and factoring out $\|\bar{g}\|^2$ yields (3.21) with

$$\varepsilon_t(\bar{g}) = \mathcal{R}_t\big(\Delta\theta, \Delta C\big) = \mathcal{R}_t\big(T_\theta\bar{g},\, T_C\bar{g}\big).$$

Finally, since $0 \preceq P_{\theta,\mathrm{high}}(\lambda) \preceq I$ and $0 \preceq P_{C,\mathrm{high}}(\lambda) \preceq I$, it follows that $\Pi_\theta(\lambda) \succeq 0$, $\Pi_C(\lambda) \succeq 0$, and $\|\Pi_\theta(\lambda)\|_{\mathrm{op}} \leq 1$, $\|\Pi_C(\lambda)\|_{\mathrm{op}} \leq 1$. Therefore for any $\bar{g} \neq 0$,

$$0 \leq \frac{\bar{g}^\top \Pi_\theta(\lambda)\bar{g}}{\|\bar{g}\|^2} \leq 1, \qquad 0 \leq \frac{\bar{g}^\top \Pi_C(\lambda)\bar{g}}{\|\bar{g}\|^2} \leq 1.$$

□

3.5 Discussion

Corollary 3.18 is the central theoretical result of this chapter. It decomposes the forgetting bound into a product of $\|\bar{g}\|^2$—the cumulative scale of the assignment disagreement along the training trajectory—and a bracketed coefficient that depends on a normalised directional ratio $\bar{g}^\top \Pi_\theta(\lambda)\bar{g}/\|\bar{g}\|^2 \in [0,1]$. This ratio measures the fraction of the disagreement-induced parameter displacement that falls within the old-task high-curvature subspace. Together with the plasticity bounds in Section 3.3, which establish that new-task acquisition scales with $\|g\|$, this decomposition opens the possibility of

[1] Here $\mathcal{R}_t(\cdot)$ denotes the remainder term from Lemma 3.13, evaluated at the disagreement-induced displacement.

treating plasticity and stability as independently addressable quantities: the magnitude of the learning signal governs acquisition, while its directional alignment with old-task sensitive modes governs forgetting, and neither factor need constrain the other.

Interpreting the decomposition. Three aspects of this result require careful interpretation.

1) Structural vs. optimisation independence. The separation of magnitude $\|\bar{g}\|^2$ and directional ratio is strictly a structural and algebraic decomposition. It does not imply that these two quantities are dynamically independent during the actual optimisation process. In practice, the disagreement signal $\bar{g}$ is governed by the intricate dynamics of optimal transport assignments; one cannot arbitrarily rotate $\bar{g}$ to minimize the projection while holding its magnitude constant. Rather, this mathematical decoupling isolates the scalar energy required for plasticity from the geometric risk of forgetting, providing a clear analytical target.

2) Eigenspace stability as a prerequisite. The decoupled bound evaluates the high-curvature projection based on the Hessian anchored at the old-task optimum $w_t^\star$. If the principal curvature directions undergo chaotic transformations during the new-task displacement Δw, the historically anchored P_{high} becomes an ineffective proxy and the bound loses practical content. Conversely, the bound is informative provided the eigenspace drift is gradual within the local optimisation neighborhood. This stability condition is not derived from the analysis; it is an empirical prerequisite.

3) Spectral concentration from the low-rank logit channel. The curvature threshold λ and the projector P_{high} are motivated by the low-rank Hessian structure identified in Section 3.2: at the level of representations and prototypes, second-order interactions are confined to a $(K-1)$-dimensional subspace. Whether this severe bottleneck propagates to the parameter-space Hessian $H_{\theta\theta}$, causing its spectrum to be similarly concentrated, is a non-trivial empirical question. If it does, then the mathematical partitioning at λ becomes practically meaningful: mitigating the forgetting penalty requires constraining only the structurally sparse P_{high} component, without regularizing the entire parameter space. If it does not, the bound remains mathematically valid but the threshold λ loses its discriminative power.

Empirical programme. The three interpretive points above identify what the decomposition means; they do not establish that its premises hold in practice. Chapter 4 is designed to test these premises and then examine whether the decomposition's central prediction—that directional alignment with old-task high-curvature modes governs forgetting severity—is borne out empirically. Specifically, the empirical programme addresses three structural conditions on which the practical relevance of the bound depends:

1. **Cross-term absorbability (Assumption 3.12):** The mixed θ–C coupling energy $|\Delta\theta^\top H_{\theta C}^{(t)} \Delta C|$ must be small relative to the block-diagonal energies for the

block decomposition in Lemma 3.13 to be a faithful approximation. This is tested first, as it determines the scope of the subsequent analyses: if the cross-term is negligible, the block-wise structure of the bound is empirically grounded.

2. **Spectral concentration:** The parameter-space Hessian must exhibit a spectrum concentrated in a small number of leading eigendirections, so that $P_{\text{high}}(\lambda)$ captures the dominant forgetting-sensitive subspace rather than an arbitrary spectral slice.

3. **Eigenspace stability:** The dominant eigensubspace must remain approximately stable during the short post-switch training interval, so that the anchored projector P_{high} remains a meaningful reference frame for evaluating the new-task displacement.

Conditioned on these structural properties, the main empirical test examines whether the directional alignment ratio $\|P_{\theta,\text{high}}(\lambda)\Delta\theta\|^2/\|\Delta\theta\|^2$ is a robust predictor of forgetting severity across randomized task transitions.

It is worth noting that, within the formalization developed in this chapter, the assignment disagreement g connects plasticity and forgetting under a common analytical framework. For the parameter-space validation that follows, this connection is implicit rather than operationally required—the empirical quantities $\Delta\theta$, $H_{\theta\theta}^{(t)}$, and the high-curvature projection are directly observable without reconstructing g from training logs. In a later chapter, the analysis shifts to a setting where g itself becomes the explicit object of measurement, and where its distributional properties can be examined and related to the forgetting interaction within the K-dimensional prototype space.

Chapter 4

Empirical Validation

Chapter 3 decomposed the forgetting bound into a scale factor $\|\bar{g}\|^2$ and a directional alignment ratio in $[0, 1]$, and identified three structural conditions on which the practical relevance of this decomposition depends (Section 3.5): the absorbability of the mixed θ–C coupling in the block Hessian, the spectral concentration of the parameter-space Hessian, and the stability of its dominant eigensubspace across the task transition. This chapter tests each of these conditions empirically and then examines whether the decomposition's central prediction—that directional alignment with old-task high-curvature modes governs forgetting severity—holds empirically across task transitions.

4.1 Experimental Setup

All experiments are conducted in a controlled two-task continual learning setting using the SwAV self-supervised framework. A ResNet backbone [23] produces feature representations, which are processed by a projection head and mapped to prototype logits via the Sinkhorn assignment mechanism described in Section 2.2.

We adopt the simplest finetuning protocol: the model is not explicitly informed of task boundaries, and the optimisation procedure remains identical before and after each switch. The only modification at the task boundary is a predefined learning rate adjustment. All other aspects of training—architecture, optimiser, and hyperparameters—remain unchanged throughout.

For the primary analysis, we use CIFAR-100 [29], which contains 100 classes and thus provides sufficient diversity to construct many non-overlapping task pairs from a single dataset. The first task consists of 10 randomly selected classes; the model is trained on this task for approximately 120 epochs, at which point the training loss has stabilised and the representation yields a kNN accuracy of approximately 75% ($k = 10$). This convergence is verified to ensure that the geometric phenomena observed at the subsequent task transition reflect genuine cross-distribution effects rather than residual instability from incomplete optimisation on the first task.

Our primary analytical interest lies in the transition regime. For the second task, we select a disjoint set of 10 classes from the remaining 90. Crucially, the availability of 90 unused classes allows us to construct multiple independent Task_2 compositions from the same Task_1 anchor, enabling a controlled comparison of how different distribution shifts interact with the same anchored curvature structure. In this section, we report results averaged over such randomised Task_2 transitions.

To ensure the robustness of our findings, the Appendix E includes experimental results on additional datasets, namely Tiny-ImageNet [31] and Food-101 [4], which provide the high class cardinality necessary to construct numerous randomized task compositions and effectively reduce variance.

Detailed dataset descriptions and all experimental hyperparameters, including dataset-specific SwAV configurations and augmentation strategies, are also provided in the Appendix D. Furthermore, the Appendix elaborates on the implementation of the Hessian and Jacobian-vector products that drive the eigendecomposition.

4.2 Validating the Structural Assumptions

The preceding analysis rests on three empirical conditions listed in Section 3.5. We test them in an order dictated by logical dependency.

First, the block-wise energy decomposition (Section 4.2.1) tests whether the mixed $\theta - C$ coupling, $H_{\theta C}$, can indeed be absorbed into constant factors as required by Assumption 3.12. Then, the eigenspace stability analysis (Section 4.2.2) serves two roles: algorithmically, it will verify that the anchored Hessian provides a meaningful reference frame for the curvature-based analyses that follow; mathematically, it is the probe of the adequacy of the second-order proxy, since rapid Hessian drift along Δw would signal non-negligible higher-order contributions to the remainder $\mathcal{R}_t$ in Assumption 3.11.

The spectral concentration analysis (Section 4.2.3) then examines whether the forgetting-sensitive energy is localised within a low-dimensional subspace of $H_{\theta\theta}^{(t)}$ or $H_{CC}^{(t)}$, which is the condition under which the high-curvature split in Proposition 3.17 and the directional ratio in Corollary 3.18 carry practical significance beyond their mathematical validity.

4.2.1 Block-Wise Energy Decomposition

The forgetting bound in Lemma 3.13 replaces the full Hessian quadratic form $\Delta w^\top H_t \Delta w$ with a sum of block-diagonal terms, absorbing the mixed θ–C coupling into constant factors (Assumption 3.12). We begin by testing whether this absorption is empirically justified, and by quantifying the relative magnitudes of the three block-wise contributions.

Directional coupling strength. Since Assumption 3.12 is stated along the specific update direction rather than as a global matrix-norm condition, we define the block-wise

quadratic energies

$$E_\theta \triangleq \Delta\theta^\top H_{\theta\theta}\Delta\theta, \qquad E_C \triangleq \Delta C^\top H_{CC}\Delta C, \qquad E_{\theta C} \triangleq \left|\Delta\theta^\top H_{\theta C}\Delta C\right|,$$

and evaluate a normalised coupling coefficient

$$\widehat{\beta} \triangleq \frac{E_{\theta C}}{\sqrt{E_\theta\, E_C}}, \tag{4.1}$$

which serves as the empirical counterpart of the directional coupling constant appearing in the sufficient condition for Assumption 3.12. Small values of $\widehat{\beta}$ indicate that the mixed interaction energy is at most on the order of the geometric mean of the two within-block energies, supporting the absorbability premise underlying Eq. (3.18).

Practical estimation via mixed HVPs. Explicitly forming $H_{\theta C}^{(t)} \in \mathbb{R}^{d_\theta \times d_C}$ is infeasible. Instead, all quantities in Eq. (4.1) are computed directly on a fixed probe batch at the anchored weights, $w \triangleq (\theta, C)$, using automatic differentiation and Hessian–vector products (HVP). The diagonal quadratic forms are obtained via standard HVPs. For the numerator, we compute a mixed HVP $u = H_{\theta C}^{(t)}\Delta C$ and then take the inner product $\Delta\theta^\top u$ to obtain $\Delta\theta^\top H_{\theta C}^{(t)}\Delta C$ (see details in Appendix D).

Results. We consider two consecutive tasks (120 epochs each) and periodically select anchor parameter states every 6 epochs. Starting from each anchor, we track the training trajectory over a short window, compute the displacement $\Delta w = (\Delta\theta, \Delta C)$ relative to the anchor, and evaluate $\widehat{\beta}$ from the block-wise quadratic forms induced by the Hessian at the anchor. For the cross-task setting, the anchor coincides with the final stage of Task_1 and the Hessian is computed on Task_1 data, matching precisely the regime required by the forgetting bound.

Two findings emerge (Figures 4.1 and 4.2). First, the coupling coefficient $\widehat{\beta}$ decays consistently as optimisation enters a stable regime; at the task boundary—the setting most relevant to the forgetting bound—it is near its minimum. This trend is robust across different anchor–displacement intervals. Second, the energy decomposition reveals a pronounced scale separation: the θ-block energy E_θ dominates the total quadratic penalty by several orders of magnitude,

$$E_\theta \gg E_{\theta C} > E_C.$$

Taken together, these results provide empirical support for Assumption 3.12 within the forgetting-relevant regime.

Analytical simplification. The block-wise decomposition establishes that the direct prototype-block contribution E_C and the cross-block interaction $E_{\theta C}$ are both negligible relative to the backbone term E_θ. In the subsequent analyses, we therefore adopt a θ-dominant simplification: the forgetting bound is evaluated through the backbone block alone, and spectral concentration is validated primarily on $H_{\theta\theta}^{(t)}$.

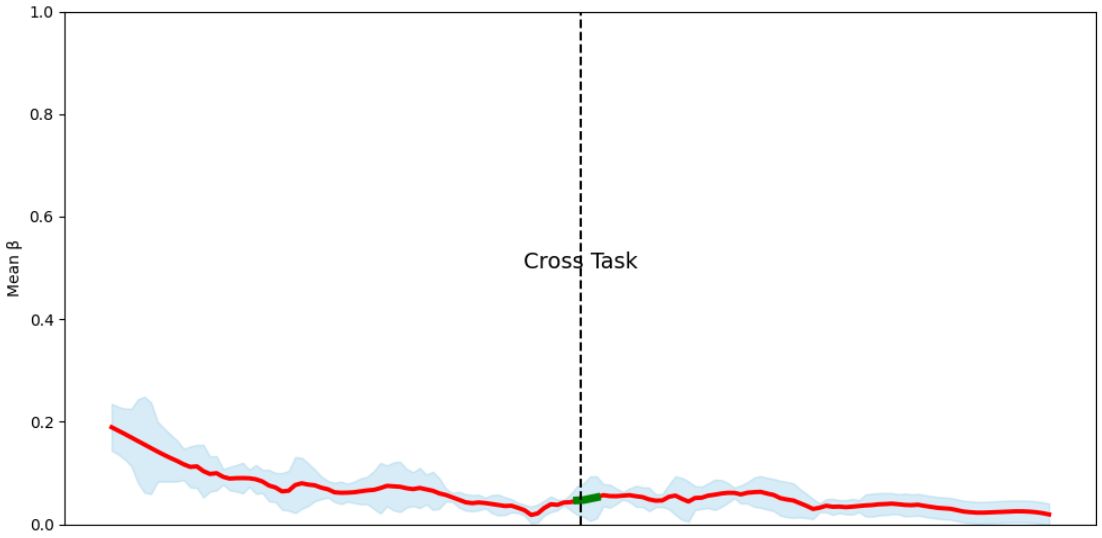

FIGURE 4.1: Mean $\widehat{\beta}$ under different anchor–displacement intervals. Across two consecutive tasks (120 epochs each), we sample an anchor snapshot every 6 epochs and compute the block-wise Hessian at that anchor using the data from the corresponding task. For each anchor, Δw is measured at four post-anchor intervals (5 steps, 15 steps, 2 epochs, and 3 epochs), and the associated $\widehat{\beta}$ is computed. For each anchor point, we average $\widehat{\beta}$ across these intervals and aggregate over 20 independent trials (solid line: mean; shaded region: ± 1 std). The green-highlighted segment near the task boundary denotes the cross-task evaluation: the anchor weights correspond to the final snapshot of Task_1, Δw uses early Task_2 weights, and the Hessian is still computed on Task_1 data.

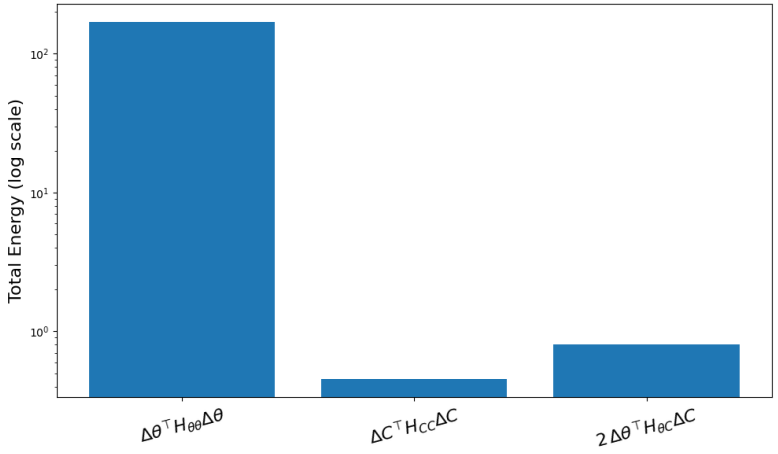

FIGURE 4.2: Log-scale comparison of total second-order energies under block decomposition. The three bars correspond to E_θ, E_C, and the cross-block interaction energy $2E_{\theta C}$. Each bar reports the cumulative energy summed over all sampled anchor points across the two-task protocol in Fig. 4.1.

The indirect role of prototypes. The θ-dominance of the quadratic energy does not imply that prototypes are irrelevant to forgetting. Although the direct contribution E_C is small, prototypes may still influence forgetting by modulating the generation mechanism of $\Delta\theta$ through the assignment pathway. Under the stop-gradient convention, the Sinkhorn assignment q is not differentiated through during backpropagation, but it is

functionally determined by the current embeddings and prototypes. As derived in Appendix B.2, Sinkhorn assignment admits a softmax-like form with effective temperature $T_{\text{eff}} = \tau\varepsilon$:

$$q_{ik} = \frac{\exp\big(\tilde{s}_{ik}/(\tau\varepsilon)\big)}{\sum_{j=1}^{K} \exp\big(\tilde{s}_{ij}/(\tau\varepsilon)\big)}, \qquad \tilde{s}_{ik} = s_{ik} + \tau\varepsilon \log v_k, \tag{4.2}$$

where $v_k > 0$ is the prototype-wise scaling factor from Sinkhorn column normalisation. From the standard softmax Jacobian, the sensitivity of q to prototype perturbations scales as

$$\left\| \frac{\partial q}{\partial C} \right\| \propto \frac{1}{\tau\varepsilon}. \tag{4.3}$$

This inverse-temperature scaling means that, even when the direct quadratic penalty E_C is negligible, prototype perturbations can substantially alter the direction of the backbone update $\Delta\theta$ through the assignment channel—particularly when T_{eff} is small.

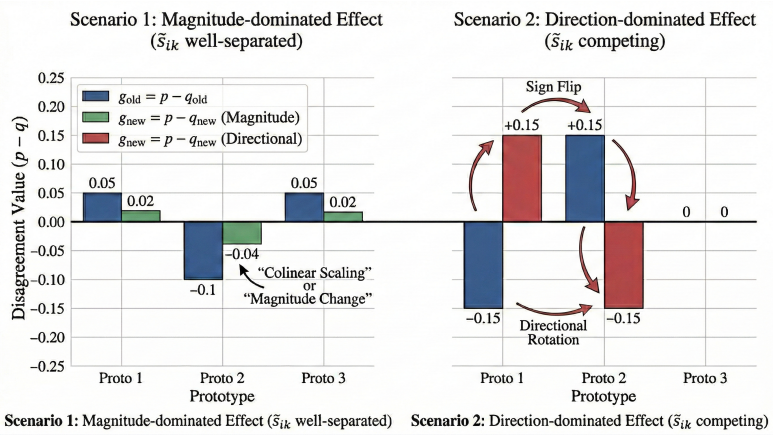

FIGURE 4.3: Two regimes of prototype-induced sensitivity. When effective logits $\tilde{s}_{ik}$ are well-separated (Scenario 1), prototype perturbations primarily rescale the magnitude of $g = p - q$ without altering its orientation. When logits compete strongly (Scenario 2), small shifts in C redistribute assignment mass across prototype indices, inducing a directional rotation in g and consequently in $\Delta\theta$. For a detailed intuitive discussion, see Appendix B.2.2.

The qualitative effect depends on the logit landscape (Figure 4.3): when the effective logits $\tilde{s}_{ik}$ are well-separated, prototype perturbations mainly rescale $g = p - q$ without changing its orientation; when several logits compete, the same perturbation can redistribute assignment mass across prototype indices, rotating g and thereby rotating $\Delta\theta$ into potentially different regions of the old-task curvature landscape. This mechanism will be revisited from a complementary perspective in Chapter 5, where ε is examined not as a modulator of $\Delta\theta$'s direction in parameter-space, but as a control over the distributional structure of g in prototype space.

4.2.2 Eigenspace Stability

As argued above, the stability of the Hessian eigenstructure is the most foundational of the three conditions: it validates not only the anchored projector P_{high} but the second-order approximation itself. If the Hessian varies rapidly along the displacement Δw—indicating large third-order derivatives—the remainder $\mathcal{R}_t$ in Assumption 3.11 can no longer be treated as a controlled residual, and the quadratic characterisation of forgetting loses its justification. We therefore examine whether the dominant eigensubspace of $H_{\theta\theta}^{(t)}$ remains approximately stable within the probing window used in the subsequent experiments.

For completeness, we also report the corresponding drift analysis for the C block. Although the forgetting bound is θ-dominant and does not require C-block stability as a prerequisite, the comparison will reveal a notable structural contrast: the C-block eigensubspace exhibits weaker variation during early single-task training and a more systematic, structured evolution at the task boundary relative to the θ block (Figures 4.5[1] and 4.6). This observation contributes to the broader geometric characterisation of the SwAV loss landscape under task transitions, even though it does not enter the forgetting analysis directly.

Drift metrics. We quantify Hessian drift by comparing an anchored eigensubspace with the eigensubspace recomputed at later training steps, both evaluated on the same validation batch. Let $U_{\text{anchor}}, U_{\text{cur}} \in \mathbb{R}^{d\times r}$ be orthonormal bases of the anchored and current r-dimensional eigensubspaces. We compute the alignment matrix

$$S = U_{\text{anchor}}^{\top} U_{\text{cur}}, \tag{4.4}$$

whose singular values $\sigma_1 \geq \cdots \geq \sigma_r$ define principal angles $\theta_i = \arccos(\sigma_i)$ between the two subspaces. We further define the subspace overlap statistic

$$\text{Overlap}(U_{\text{anchor}}, U_{\text{cur}}) := \frac{1}{r}\sum_{i=1}^{r} \sigma_i^2, \tag{4.5}$$

which provides a normalised score reflecting the dominant intensity of structural change. The dominant eigensubspace is estimated via the Lanczos algorithm; computational details are in the Appendix.

Within-task behaviour. During the early stages of single-task training, the anchored Hessian subspace drifts substantially over short intervals, consistent with the curvature

[1] A perfectly diagonal visualisation is not generally expected for empirical high-dimensional eigenspaces, since the stable quantity is the leading subspace rather than a unique eigenvector basis, and near-degenerate modes can rotate within an almost unchanged subspace. We therefore rely on the directly measured overall subspace *overlap* for quantitative support (see Eq.4.5), while the present heatmap serves only as a schematic visualisation. Concretely, the measured overlap level is loaded onto a fixed diagonal reference frame with a stylised off-diagonal deformation, so the plot remains visually discriminative: lower overlap appears more diffuse, whereas higher overlap remains more concentrated near the diagonal.

structure not yet being stabilised. In contrast, once the training loss has converged to a narrow fluctuation range near its empirical minimum, the subspace exhibits negligible rotation over intervals of identical length. Figures 4.4 and 4.5 visualise this contrast through heatmaps of the alignment matrix S for the θ and C blocks, respectively.

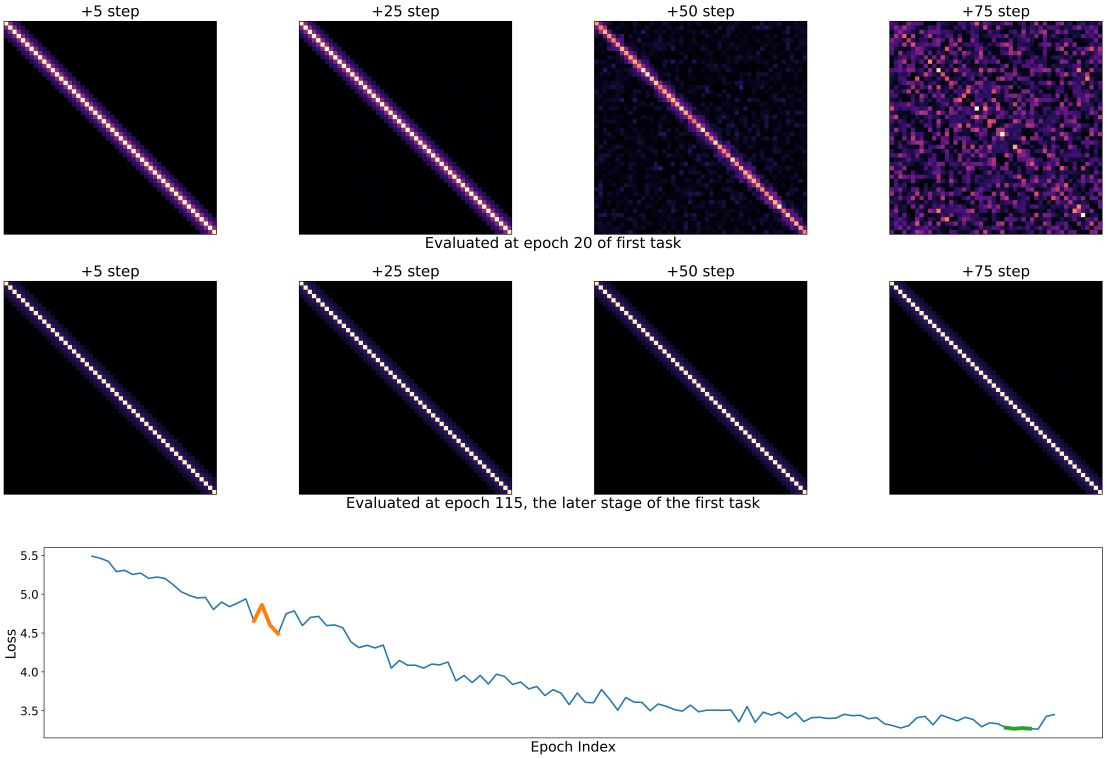

FIGURE 4.4: Alignment heatmaps for the θ-block Hessian during early-stage (top row) and late-stage (middle row) single-task training. One epoch is selected as the anchor; the Hessian is recomputed at subsequent steps and the alignment with the anchor is evaluated via the matrix S (Eq. 4.4). Perfect alignment corresponds to a strongly diagonal matrix. During the early stage, off-diagonal entries emerge after approximately 50 steps, indicating rotation of principal directions. During the late stage, the alignment remains near-perfect. The two stages correspond to the highlighted segments in the loss trajectory (bottom). The overlap curve shown below reports the corresponding quantitative alignment, and is consistent with the visual transition seen in the heatmaps. The corresponding results for the C block are shown in Fig. 4.5.

Cross-task behaviour. We anchor the Hessian at the end of Task_1—a point within the stable regime identified above—and monitor the eigensubspace during the first few epochs of Task_2. As shown in Fig. 4.6, misalignment appears immediately once the new task begins. However, the deviation evolves smoothly and progressively, in contrast to the oscillatory instability observed during early single-task training. This gradual drift supports the use of the anchored P_{high} as a locally informative reference during the short post-switch probing window—and, more fundamentally, indicates that the higher-order terms remain bounded in this regime.

Summary. Within the converged regime and the short post-switch window, the dominant eigensubspace of $H_{\theta\theta}^{(t)}$ remains sufficiently stable to serve as a meaningful reference

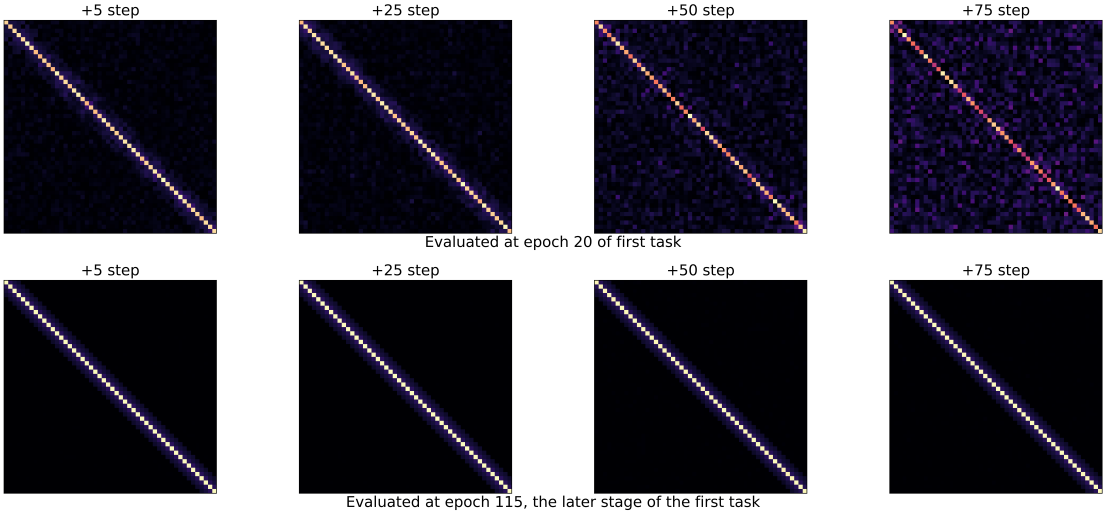

FIGURE 4.5: Alignment heatmaps for the C-block Hessian under the early and late stages of single-task training. Compared to the θ block, structural variation in C during the early stage is noticeably weaker.

frame for the anchored projector $P_{\text{high}}(\lambda)$. This stability also provides indirect evidence that the second-order approximation is locally adequate within this probing regime. The C-block eigensubspace exhibits comparable or greater stability, a structural observation that we note for completeness but that does not enter the subsequent forgetting analysis.

4.2.3 Spectral Concentration of the θ-Block Hessian

With the second-order foundation secured by the stability analysis above, we now ask whether the forgetting-sensitive energy within this stable Hessian is concentrated in a low-dimensional spectral region. The high-curvature projector $P_{\text{high}}(\lambda)$, introduced in Definition 3.16 and central to the derivations in Proposition 3.17 and Corollary 3.18, is practically meaningful only if this is the case. As noted in Section 3.5, such concentration is not guaranteed by the loss-level low-rank structure identified in Section 3.2: the parameter-space Hessian $H_{\theta\theta}$ may diffuse the spectrum. We test this empirically.

The validation proceeds in two logically independent steps, designed to avoid circular reasoning.

Step 1: Spectral concentration of the Hessian quadratic form. Let (λ_i, u_i) denote eigenpairs of the anchored old-task Hessian $H_{\theta\theta}^{(t)}$, ordered by decreasing eigenvalue. We define

$$\eta(K) = \frac{\sum_{i=1}^{K} \lambda_i \langle u_i, \Delta\theta \rangle^2}{\Delta\theta^\top H_{\theta\theta}^{(t)} \Delta\theta}, \tag{4.6}$$

which measures the proportion of total second-order sensitivity captured by the leading K eigendirections. If $\eta(K) \approx 1$ for small K, the forgetting risk is spectrally concentrated. We denote the smallest such K as K_{big}.

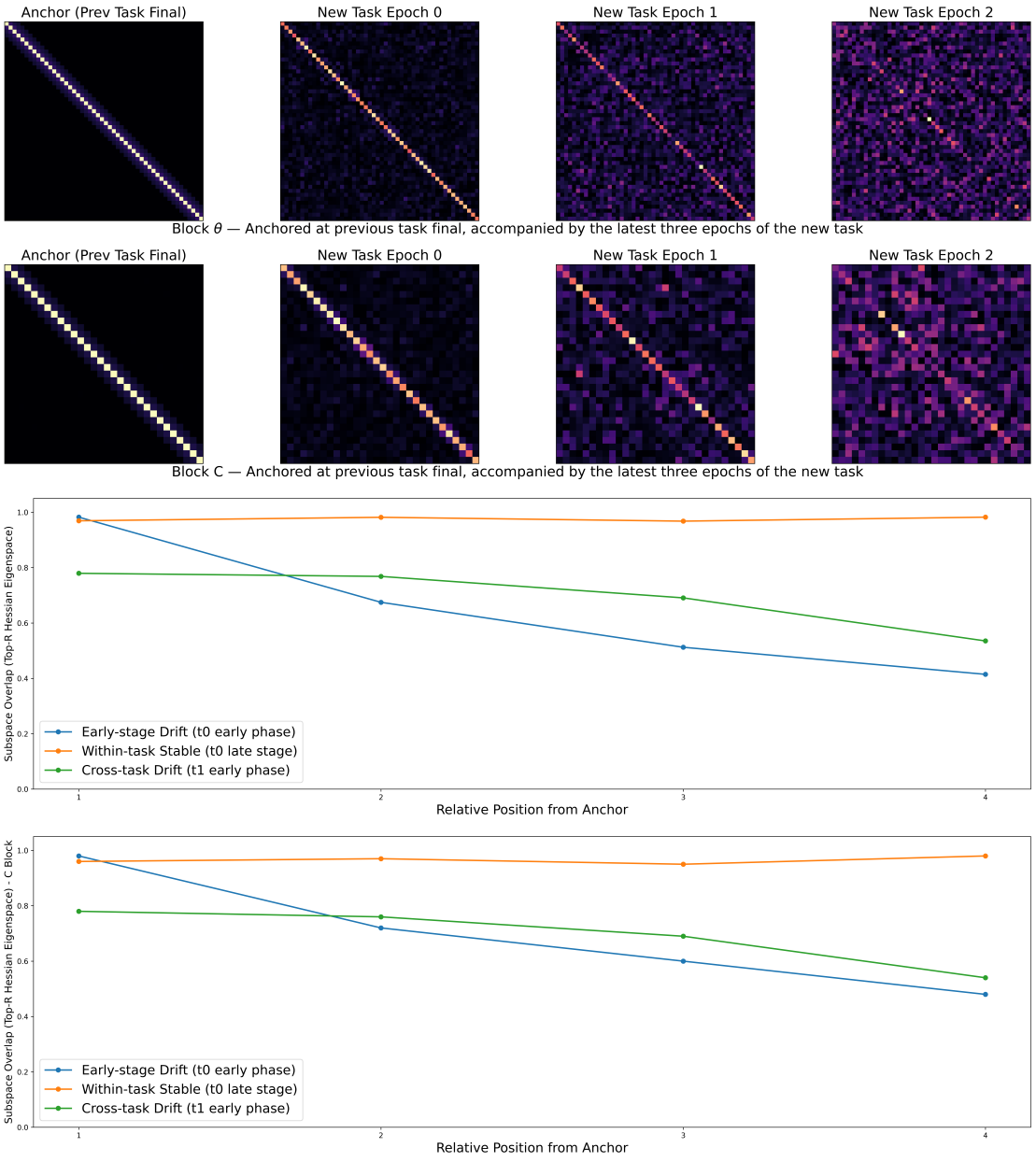

FIGURE 4.6: First and second rows: alignment heatmaps for the θ and C blocks at the task boundary. The Hessian is anchored at the end of Task_1; alignment is evaluated at the last epoch of Task_1 and the first three epochs of Task_2. Third and fourth rows: overlap dynamics across three regimes—early-stage training, late-stage steady training, and task transition—for the θ and C blocks respectively. Both early-stage training and task transition lead to decreasing overlap, but the drop during early training is substantially more abrupt. The prototype-induced subspace exhibits a more structured and systematic evolution compared to the θ block.

This first step is based entirely on the classical Hessian quadratic form and does not depend on the projection-based bound. The denominator is computed directly via Hessian–vector products without requiring the full eigenspectrum.

Step 2: Subspace occupancy of the threshold-based projector. Conditioned on spectral concentration, we test whether the threshold-based projector $P_{\text{high}}(\lambda)$ captures

the empirically dominant region. We define

$$R(\lambda) = \frac{\|P_{\text{high}}(\lambda)\Delta\theta\|^2}{\|P_{K_{\text{big}}}\Delta\theta\|^2}. \tag{4.7}$$

If $R(\lambda)$ is consistently large, the threshold-based partition meaningfully overlaps with the high-sensitivity region identified in Step 1.

Results. Figure 4.7 shows the stage-wise distribution of $R(\lambda)$ when the top-80 eigenspace already captures almost all of the quadratic energy, that is, $\eta(K_{\text{big}} = 80) \approx 1$. In this case, $R(\lambda)$ measures how much of the top-80 energy is contributed by the leading 20 eigen-directions. We measure this quantity in short windows under both the within-task setting and the task-transition setting.

The main pattern is simple. Except for the near-initial stage, when the model has been trained for only one epoch, most of the energy is concentrated in a very small number of top eigen-directions. Even after the first epoch, the concentration is already strong, especially given the very high dimensionality of the full Hessian. After the model has largely converged on one task and then switches to a new task, the distribution shows some fluctuation, and the energy becomes slightly more spread within the top modes.

We do not try to explain these stage-wise differences in detail here. The goal of this experiment is to establish a basic empirical fact: the quadratic energy is highly concentrated in a very small set of top eigen-directions. This gives empirical support for the analysis in Chapter 3 (and Proposition 3.17), where the fraction of gradient energy projected onto the high-curvature subspace is treated as a structurally separable quantity in the forgetting bound.

4.3 Directional Alignment as the Signature of Forgetting

The three structural conditions tested in Section 4.2 are now empirically grounded: the block decomposition is θ-dominant, the second-order structure is stable across the task transition, and the θ-block Hessian spectrum is concentrated in a small number of leading eigendirections. We now test the central prediction of the decomposition: that the directional alignment ratio $\|P_{\theta,\text{high}}(\lambda)\Delta\theta\|^2/\|\Delta\theta\|^2$ is a robust predictor of forgetting severity, beyond what the total displacement magnitude $\|\Delta\theta\|$ alone explains. Note that this test operates in parameter space with $\Delta\theta$, whereas Corollary 3.18 expresses the bound through the stacked disagreement $\bar{g}$ and its directional projection $\bar{g}^\top \Pi_\theta(\lambda)\bar{g}/\|\bar{g}\|^2$. Since $\Delta\theta = T_\theta \bar{g}$, the two formulations are related by a linear map; a positive empirical association between the parameter-space projection ratio and forgetting therefore provides operational evidence that the g-based decomposition carries algorithmic content, not merely mathematical validity.

We do not claim that suppressing alignment is sufficient for preventing forgetting; higher-order residuals and departures from the local stability assumptions may still contribute.

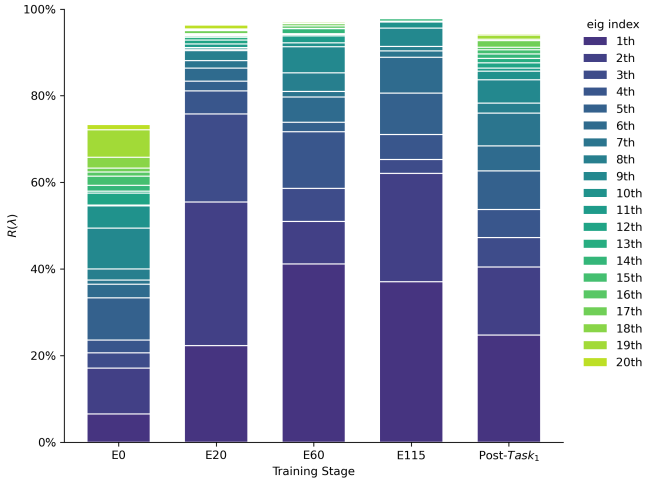

FIGURE 4.7: Stage-wise distribution of $R(\lambda)$ over the top 20 eigen-directions. For a task trained for 120 epochs, we record four within-task stages, anchored at epochs 1, 20, 60, and 115, together with one cross-task stage measured immediately after switching to the next task. For each within-task anchor, the reported distribution is obtained by averaging three short-horizon measurements taken at +5 steps, +15 steps, and +2 epochs (30 steps) after the anchor. For Post-Task$_1$, the Hessian is anchored at the final epoch of Task$_0$, and the corresponding measurements are taken at +5, +15, and +30 steps after training begins on the new task. λ denotes the eigen-directions associated with the largest 20 eigenvalues. $R(\lambda)$ measures the fraction of energy contributed by these top 20 modes within the top-80 eigenspace. Across all recorded points, the top-80 subspace almost fully captures the total energy, i.e., $\eta(80) \approx 1$. Results are averaged over 20 runs and remain stable under small hyperparameter variation, including $\varepsilon = \{0.01, 0.05\}$. Each coloured segment shows the contribution of one eigen-direction to the total $R(\lambda)$ mass.

Rather, we investigate whether substantial forgetting events are systematically accompanied by concentrated update energy in the old-task high-curvature subspace. Our analysis is restricted to the short post-switch regime (first 5 epochs), where the anchored eigensubspace remains locally informative (Section 4.2.2).

Experimental design.

Following the protocol in Section 4.1, we generate 100 randomised Task$_2$ transitions from the same converged Task$_1$ checkpoint w_{ref}, or the last epoch point during Task$_1$, each drawing a distinct set of 10 classes from the 90 unused classes.

For each run, we restore w_{ref} and train on Task$_2$ for 5 epochs, recording 3 steps per epoch. Let w_s be the weights at step s and θ_s the backbone component. We measure step-wise forgetting as $F_s \triangleq L_{\text{old}}(w_s) - L_{\text{old}}(w_{\text{ref}})$ and the backbone displacement as $\Delta\theta_s \triangleq \theta_s - \theta_{\text{ref}}$. Using the top-$K_{\text{big}}$ eigenpairs $\{(\lambda_i, u_i)\}_{i=1}^{K_{\text{big}}}$ of the anchored old-task Hessian, we track the projection mass onto the high-curvature subspace:

$$A_s(\lambda) \triangleq \|P_{\theta,\text{high}}(\lambda)\Delta\theta_s\|^2 = \sum_{\lambda_i \geq \lambda} (u_i^\top \Delta\theta_s)^2, \quad \lambda \geq \lambda_{K_{\text{big}}}. \tag{4.8}$$

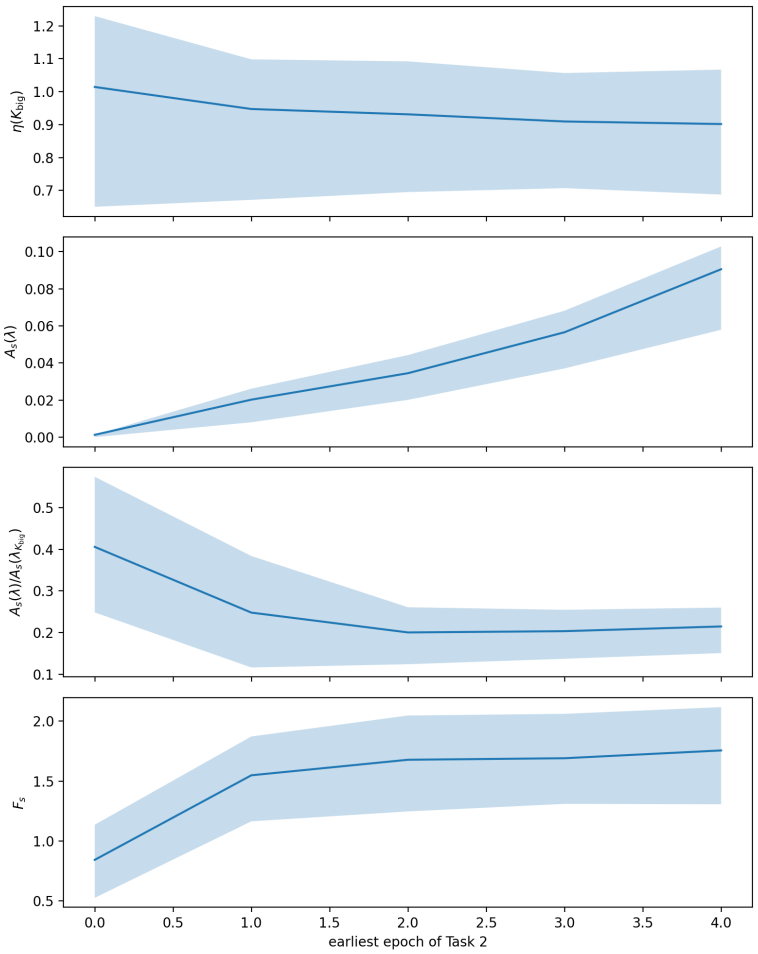

FIGURE 4.8: Global step-wise trends averaged over 100 randomised Task_2 transitions. From top to bottom: $\eta_s(K_{\text{big}})$, $A_s(\lambda)$, $A_s(\lambda)/A_s(\lambda_{K_{\text{big}}})$, and F_s. Across the probing window, the mean of $\eta_s(K_{\text{big}})$ remains close to 1 throughout; $A_s(\lambda)/A_s(\lambda_{K_{\text{big}}})$ decreases, indicating spreading within the top-K_{big} subspace, but the high-curvature component (top-20 modes, $K_{\text{big}} = 80$) remains nontrivial throughout. $A_s(\lambda)$ increases monotonically across the probing window, co-evolving with F_s, consistent with the synchronous growth described in the text.

Global synchronous trend. As shown in Fig. 4.8, the old-task loss increase F_s grows steadily over the probing window. Concurrently, $A_s(\lambda)$ also increases, and $\eta_s(K_{\text{big}})$ remains close to 1 throughout, confirming that the leading eigensubspace captures most of the quadratic energy. This synchronous growth satisfies the necessary condition of Proposition 3.17. However, temporal co-evolution alone is not conclusive: the absolute displacement $\|\Delta\theta_s\|$ also grows during early SGD, so the parallel increase in F_s and $A_s(\lambda)$ could reflect optimisation progress rather than a specific directional effect.

Step-wise specificity: separating high and low curvature. To exclude the confound of temporal progression, we examine the step-wise scatter correspondence between F_s and the curvature-distributed energies. Fig. 4.9 separates the top-80 anchored Hessian eigenspace into a high component (top 20) and a low component (ranks 21–80). The low-curvature projection forms a diffuse cloud with negligible explanatory power over F_s,

whereas the high-curvature component exhibits a clear positive correlation with forgetting. This rules out the hypothesis that forgetting is driven by arbitrary parameter-space wandering: stability degradation is structurally specific to the high-sensitivity subspace.

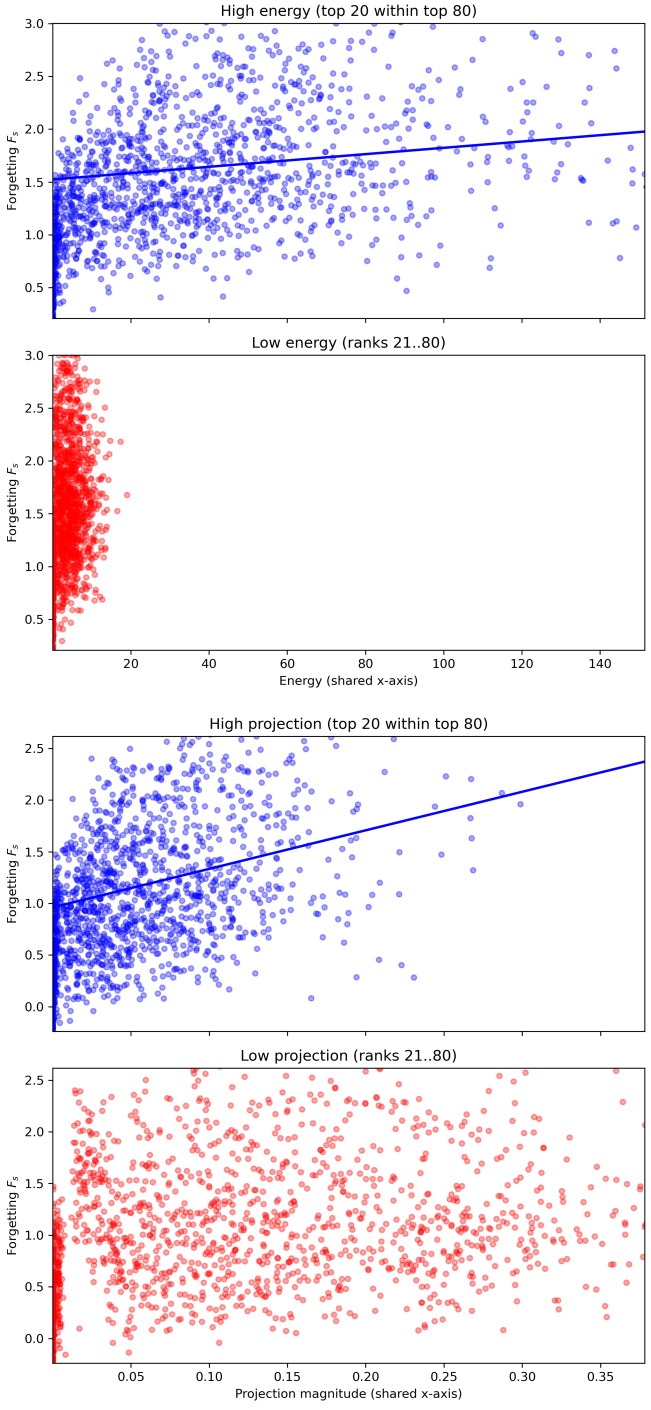

FIGURE 4.9: Step-wise correlation between forgetting and eigenspace structure. Each point is a single update step; F_s is on the y-axis. Upper panels: F_s against the curvature-weighted energy $\sum_i \lambda_i (u_i^\top \Delta\theta)^2$ within the top-80 eigenspace, split into high (top-20) and low (ranks 21–80). Lower panels: unweighted projection magnitudes $\sum_i (u_i^\top \Delta\theta)^2$ under the same split. F_s exhibits a positive association with the high-curvature component in both views; the low-curvature component shows no such structure.

Controlling for collinearity. A vector's total magnitude and its subspace projections are geometrically coupled: any expansion in $\|\Delta\theta\|$ algebraically inflates projection energy across the entire spectrum [22, 61]. To separate the magnitude and directional effects, we regress step-wise forgetting jointly on both:

$$F_s \sim \beta_1\|\Delta\theta\| + \beta_2\left(\frac{A_s(\lambda)}{\|\Delta\theta\|^2}\right) + \epsilon. \tag{4.9}$$

A positive coefficient β_2 implies that, holding update magnitude constant, updates more concentrated in the high-curvature subspace are associated with larger forgetting. Across all evaluated settings, β_2 is consistently positive.

The regression confirms a significant positive effect of the directional term ($\beta_2 = 0.666, p < 10^{-12}$), indicating that high-curvature alignment is an independent predictor of forgetting beyond update magnitude. The moderate $R^2 = 0.252$ reflects the inherent stochasticity and the fact that the model captures only the linear component of a complex interaction; it does not undermine the significance of the directional effect.

TABLE 4.1: Regression analysis controlling for update magnitude.

Variable	Coefficient (β)	Std. Error	p-value
$\|\Delta\theta\|$	1.636	0.074	$< 10^{-90}$
$\frac{A_s(\lambda)}{\|\Delta\theta\|^2}$	0.666	0.090	$< 10^{-12}$
R^2		0.252	
Observations		1500	

Binned visualization. Fig. 4.10 provides a complementary view. We discretise $\|\Delta\theta\|$ into magnitude bins and, within each bin, partition the projection ratio $A_s(\lambda)/\|\Delta\theta\|^2$ (where λ is fixed as the 20th largest eigenvalue, $\lambda = \lambda_{20}$). The stable ordering (high > mid > low) within most bins confirms that the directional effect persists after controlling for magnitude. Moreover, when the projection ratio remains in the low group, forgetting stays relatively controlled even as $\|\Delta\theta\|$ increases.

Conclusion. In the experiments above, each level of analysis systematically addresses the limitations of the preceding one. While the global trend demonstrates co-evolution between forgetting and high-curvature projection, it cannot by itself distinguish directional specificity from isotropic growth. The step-wise scatter analysis resolves this ambiguity by revealing that forgetting is tracked exclusively by the high-curvature component, whereas the low-curvature component remains uninformative.

The regression analysis further rules out the possibility that this association is a geometric artifact of magnitude inflation, isolating a significant positive effect in the directional term. The binned visualisation then confirms these statistical findings non-parametrically, showing that the positive relationship between high-curvature energy and forgetting persists even within magnitude-matched groups.

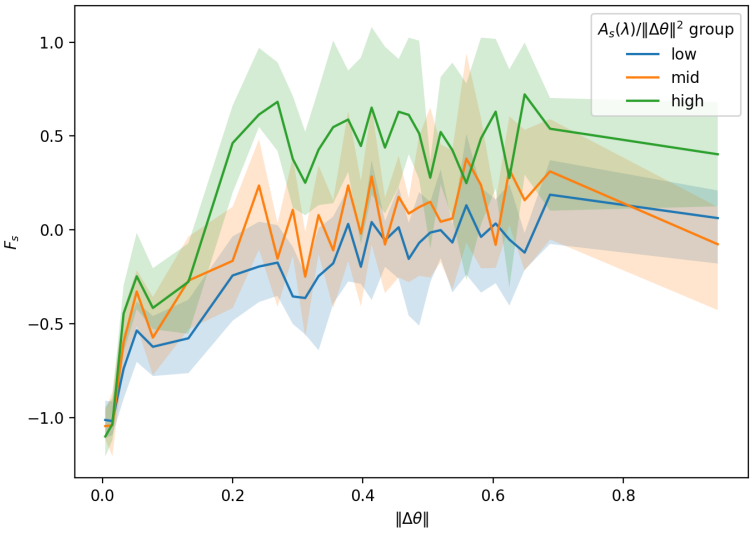

FIGURE 4.10: Binned visualization of the regression result. For each magnitude bin of $\|\Delta\theta\|$, forgetting F_s is averaged over three groups defined by the projection ratio $A_s(\lambda)/\|\Delta\theta\|^2$ (low, mid, high), using the top-20 eigen-directions. The high group consistently exceeds the low group across most bins. The within-group variance (shaded: ± 1 std) is substantial, consistent with the moderate $R^2 = 0.252$ of the regression, and the ordering becomes less stable at larger displacement magnitudes

Together, these four analytical tiers progressively eliminate alternative explanations such as temporal coincidence, isotropic drift, and geometric collinearity. This convergence of evidence provides strong empirical support for the central prediction of Corollary 3.18.

4.4 Summary

This chapter has established three empirical results in support of the theoretical decomposition developed in Chapter 3.

First, the structural conditions underlying the decomposition are satisfied in practice. The mixed θ–C coupling is absorbable and the quadratic forgetting penalty is overwhelmingly θ-dominant ($E_\theta \gg E_{\theta C} > E_C$). The Hessian eigensubspace is locally stable across the task transition, indicating that the second-order approximation is adequate in the probing regime. Meanwhile, the θ-block Hessian spectrum is concentrated in a small number of leading eigendirections.

Second, the decomposition's central prediction is borne out: across 100 randomised task transitions, the directional alignment ratio $A_s(\lambda)/\|\Delta\theta\|^2$ is a robust predictor of forgetting severity, after controlling for total displacement magnitude. Forgetting is structurally specific to the high-curvature subspace rather than a consequence of isotropic parameter drift.

These findings validate the parameter-space analysis developed in Chapter 3. Throughout the empirical programme, we operated directly with $\Delta\theta$ and $H_{\theta\theta}^{(t)}$ as observable

proxies for the g-based quantities in Corollary 3.18. As noted in Section 4.3, this substitution is legitimate because $\Delta\theta$ and $\bar{g}$ are related by the linear operator T_θ: a positive association between the parameter-space directional ratio and forgetting severity confirms that the scale-direction decomposition expressed through g is empirically grounded, not merely algebraically valid.

At the same time, this proxy relationship, precisely because it is legitimate, can obscure what it mediates. The high-curvature directions that govern forgetting are identified by the eigendecomposition of $H_{\theta\theta}^{(t)}$—numerically computable, but semantically opaque, carrying no direct connection to the learning mechanism that generates the updates. Working with $\Delta\theta$ confirms that the directional structure matters, but it does not reveal how the properties of g itself shape the forgetting interaction. The assignment disagreement, which the theory identifies as the unified driver of both plasticity and forgetting, remains implicit in all measurements reported above.

Addressing this requires a lower-dimensional analytical lens in which g becomes the explicit object of measurement, as well as old-task sensitivity expressed in more interpretable setting rather than opaque parameter-space directions. Because the parameter displacement is linear in g (Definition 3.14), the Hessian quadratic form can be pulled back exactly into the K-dimensional disagreement space, yielding a representation which can be examined in terms of the codebook structure. In next chapter, we will constructs this representation and characterises its geometric organisation.

Chapter 5

Prototype-space Reformulation of Directional Forgetting

In Chapter 4, we demonstrate that the forgetting severity is empirically governed by the directional alignment of the parameter update with old-task high-curvature modes, while the assignment disagreement $g = p - q$ remains implicitly behind the observed displacement $\Delta\theta$.

Because the parameter displacement is a linear function of g (Definition 3.14), the Hessian quadratic form can be pulled back exactly into the K-dimensional space where g lives. In this chapter, we will construct this pullback. By examining the spectral and coordinate-level structure of the resulting quadratic form, our goal is to identify how the sensitivity is organised spectrally and coordinate-wise. So that the learning signal g can be understood as a structured probe of old-task knowledge rather than an opaque parameter-space displacement.

5.1 From Parameter Space to Prototype Space

Why prototype space. The pullback from $\mathbb{R}^{d_\theta}$ to $\mathbb{R}^K$ is motivated by two complementary considerations.

First, $H_{\theta\theta}^{(t)}$ operates in a space where directions carry no semantic meaning relative to the learning objective. The prototype space $\mathbb{R}^K$, by contrast, is the native space of the SwAV learning signal: each coordinate corresponds to a prototype, and the structure of $g = p - q$ reflects how the new-task data relates to the codebook geometry. Expressing old-task sensitivity in this space makes it possible to ask which prototypes, or which patterns of prototype interaction, constitute the sensitive structure of previously learned knowledge.

Second, the Sinkhorn entropic coefficient ε controls the distributional structure of g — how sparsely or densely it distributes mass across prototype coordinates — without directly modifying $H_{\theta\theta}^{(t)}$. In prototype space, the effect of ε on how g engages old-task

sensitivity becomes directly observable. This will be the subject of Chapter 6.

Single-step analysis as the natural level. The trajectory-level analysis in Chapter 3 required aggregating g over many steps to express the total parameter displacement $\Delta\theta$ as a linear function of $\bar{g} \in \mathbb{R}^{KS}$. The resulting bound is statistical in nature: it quantifies how much forgetting energy concentrates in the high-curvature subspace across a training trajectory.

The analysis in this chapter has a different purpose. M_θ is anchored at the old-task optimum and represents a fixed structural object, which is the sensitivity geometry of previously learned knowledge at a specific point. Characterising this geometry does not require trajectory-level aggregation; it requires examining how a single update step engages the structure of M_θ. The single-step level is therefore not a simplification of the trajectory-level analysis, but the appropriate scale for the question being asked: what is the intrinsic geometry of old-task sensitivity in prototype space, and how does a learning signal structurally engage it?

For a single step with step size η, we define the single-step disagreement operators.

Definition 5.1. Fix a sample (or minibatch) and current model state, and let $g = p - q \in \mathbb{R}^K$. Define

$$T'_\theta := -\frac{\eta}{\tau}\left(\frac{\partial z}{\partial \theta}\right)^\top C, \qquad T'_C := -\frac{\eta}{\tau}\left(I_K \otimes z\right).$$

Then $\Delta\theta = T'_\theta g$ and $\mathrm{vec}(\Delta C) = T'_C g$.

Lemma 5.2. *Consider one generic optimisation step with disagreement $g = p - q \in \mathbb{R}^K$. The induced displacements are defined based on the operators T'_θ and T'_C from Definition 5.1. Under Lemma 3.13 with $\alpha_\theta, \alpha_C \in [0, 1)$, the old-task forgetting contribution of this step satisfies*

$$F_t \;\le\; \frac{1}{2(1-\alpha_\theta)}\, g^\top (T'_\theta)^\top H^{(t)}_{\theta\theta} T'_\theta g \;+\; \frac{1}{2(1-\alpha_C)}\, g^\top (T'_C)^\top H^{(t)}_{CC} T'_C g \;+\; R_t(\Delta w).$$

The prototype space curvature matrix M_θ. Chapter 4 established that the forgetting penalty is overwhelmingly θ-dominant: $E_\theta \gg E_{\theta C} > E_C$ (Section 4.2.1). Therefore, we define the prototype-space representation of old-task sensitivity through the θ-block alone:

$$M_\theta \;:=\; \frac{1}{2(1-\alpha_\theta)}\, T_\theta'^{\top}\, H^{(t)}_{\theta\theta}\, T'_\theta \;\in\; \mathbb{R}^{K\times K}. \tag{5.1}$$

By the θ-dominant simplification, the quadratic forgetting contribution of a single step is approximated by

$$F_t \;\lesssim\; g^\top M_\theta\, g \;+\; R_t(\Delta w).$$

Interpretation. M_θ encodes how old-task sensitivity, as measured by $H^{(t)}_{\theta\theta}$, is organised when viewed through the prototype coordinates. It is not merely a forgetting bound, but

a structural object that describes which directions in prototype space are dangerous for old-task knowledge. The quadratic form $g^\top M_\theta g$ measures how much a learning signal g engages these dangerous directions in a single step, but the primary object of study in this chapter is M_θ itself: its spectral organisation and coordinate structure.

Indeed, this shift in perspective, from $H_{\theta\theta}^{(t)}$ in $\mathbb{R}^{d_\theta}$ to M_θ in $\mathbb{R}^K$, does not change what is being measured. In parameter-space, the sensitive directions are defined by eigenvectors of $H_{\theta\theta}^{(t)}$ that carry no semantic content. In prototype space, the same sensitivity is organised in terms of prototype coordinates, where patterns of per-prototype and cross-prototype contributions can be identified and related to the codebook geometry.

To examine this organisation, we will introduce two complementary decompositions of $g^\top M_\theta g$. The first decomposition is spectral. It examines whether the energy is concentrated in a small number of orthogonal modes in $\mathbb{R}^K$. The second decomposition uses the natural coordinate basis. It examines whether the energy is mainly carried by individual prototypes. The spectral view reveals the dimensionality of the risk structure, while the coordinate view reveals how this structure is organised in terms of the codebook. Next, we develop these two views in turn.

5.2 Spectral Structure of M_θ

Theoretical prior from the logit channel. In Section 3.2, we showed that the logit-level Hessian $H_{ss} = \mathrm{diag}(p) - pp^\top$ has rank at most $K-1$, confining all second-order interactions to a $(K-1)$-dimensional subspace. Whether this structural bottleneck propagates through the pullback into a low-rank structure in M_θ is not guaranteed. This is a non-trivial empirical question, though it is consistent with the broader observation that deep-network loss landscapes tend to concentrate dominant curvature in low-dimensional subspaces [22, 26].

The spectral energy ratio C_r. Since $M_\theta \in \mathbb{R}^{K\times K}$ is symmetric, it admits an eigendecomposition $M_\theta = \sum_{k=1}^K \mu_k w_k w_k^\top$ with orthonormal eigenvectors $\{w_k\}$ and eigenvalues $\{\mu_k\}$. For any $g \in \mathbb{R}^K$:

$$g^\top M_\theta g = \sum_{k=1}^{K} \mu_k \, (w_k^\top g)^2. \tag{5.2}$$

To measure how much of the quadratic energy is concentrated in the leading modes, we define

$$C_r \triangleq \frac{\sum_{k=1}^r \mu_k \, (w_k^\top g)^2}{g^\top M_\theta g}. \tag{5.3}$$

Empirical evaluation. C_r is evaluated across diverse data compositions and training regimes (Fig. 5.1). In most cases, the top $r = 10$ eigenmodes of M_θ (out of $K = 250$) cover nearly all of the quadratic energy $g^\top M_\theta g$. This concentration is consistent before, during, and across task transitions.

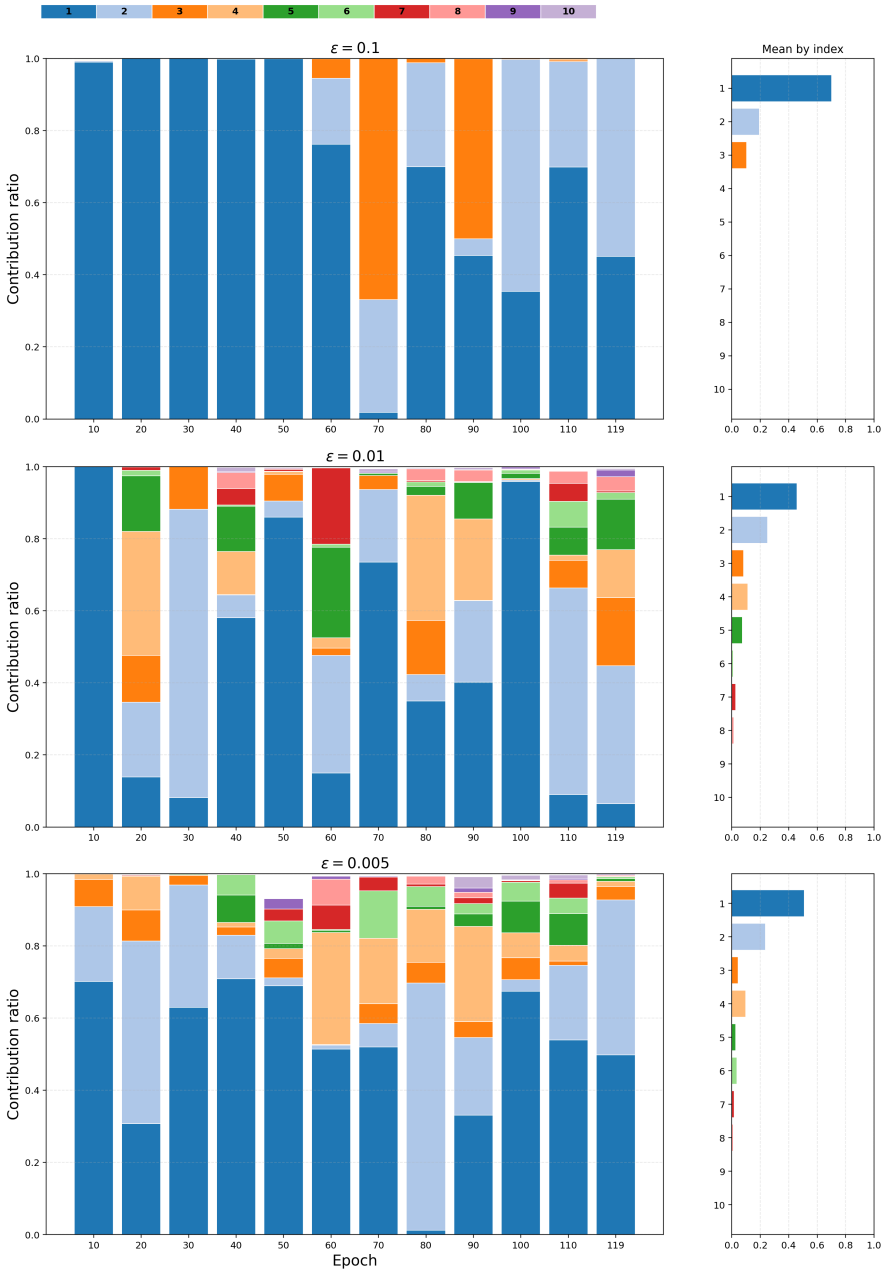

FIGURE 5.1: Spectral energy concentration C_r ($r = 10$, $K = 250$) across training under three Sinkhorn parameter settings, averaged over 20 independent runs and normalised so that contributions sum to 1; run-level variability is not shown. Each stacked bar shows the relative contribution of the top-10 eigenmodes to $g^\top M_\theta g$. At $\varepsilon = 0.1$, the top-1 mode dominates; this dominance does not decrease monotonically as ε decreases, with energy distributing across more modes at smaller values. Across all settings, the top-10 modes collectively account for nearly all quadratic energy throughout training, confirming strong spectral concentration in prototype space. No clear convergence trend is observed over the course of training. Right panels show the time-averaged contribution of each eigenmode across the full training run.

This confirms that the low-rank bottleneck at the logit level propagates into M_θ, while the sensitivity structure of old-task knowledge in prototype space is concentrated in a small number of directions.

This spectral concentration has a direct consequence for the following coordinate-level analysis. If these dominant eigenmodes do not align with individual prototype axes but correspond to mixtures of prototype coordinates, the forgetting energy cannot be explained purely by isolated prototype-wise contributions. In other words, it must arise from cross-prototype coupling.

Computational note. C_r is computed using the Lanczos algorithm to extract the top eigenpairs of M_θ without explicitly forming the full matrix. To mitigate stochastic noise in single-step gradients, the disagreement signal g is estimated as the average over several consecutive steps following a fixed anchor, with the Hessian and Jacobian operators held constant at the anchor. Details are in the Appendix.

5.3 Coordinate Structure: Diagonal and Off-Diagonal Energy

While the spectral decomposition reveals how many orthogonal directions carry the forgetting energy, its eigenvectors are typically dense mixtures of prototype indices and do not correspond to individual prototypes. To examine the forgetting interaction from the native viewpoint of prototype coordinates, we expand $g^\top M_\theta g$ in the standard basis $\{e_i\}_{i=1}^K$:

$$g^\top M_\theta g \;=\; \underbrace{\sum_{i=1}^{K} M_{ii}\, g_i^2}_{\text{Diag-term}} \;+\; \underbrace{2 \sum_{1\le i<j\le K} M_{ij}\, g_i g_j}_{\text{Cross-term}}. \tag{5.4}$$

The Diag-term collects contributions from individual prototype coordinates on the diagonal of M_θ, while the Cross-term aggregates cross-prototype coupling. The diagonal dominance ratio

$$\mathrm{Ratio}_{\mathrm{diag}} \;\triangleq\; \frac{\text{Diag-term}}{g^\top M_\theta g} \tag{5.5}$$

characterises the coordinate organisation of M_θ's sensitivity. Specifically, a low value indicates that the dangerous directions in prototype space are cross-prototype combinations rather than individual prototype axes.

Reference baseline. Consider a reference model where the components of g are independent, zero-mean, and identically distributed; the cross-terms then vanish in expectation: $\mathbb{E}[g_i g_j] = 0$ for $i \neq j$, so $\mathbb{E}[g^\top M_\theta g] = \sum_i M_{ii}\, \mathbb{E}[g_i^2]$.

Empirical observation. We evaluate $\mathrm{Ratio}_{\mathrm{diag}}$ using the same runs as in the spectral concentration experiments. Across all settings, $\mathrm{Ratio}_{\mathrm{diag}}$ remains well below one (typically fluctuating below 0.3). This indicates that $g^\top M_\theta g$ cannot be explained by independent per-prototype contributions alone, and that cross-prototype interactions dominate the forgetting energy in prototype space.

Structural prior: why cross-prototype coupling is expected. We give two reasons that make such behaviour plausible.

First, the spectral experiments show that the quadratic energy is concentrated in a small number of eigenmodes of M_θ. Unless these dominant modes happen to align closely with individual prototype axes, such concentration typically corresponds to mixtures of prototype coordinates. When g projects onto these modes, the quadratic energy naturally flows through the off-diagonal entries of M_θ.

Second, the Sinkhorn doubly-stochastic constraint produces target assignments q under a global balance condition. As a result, a large component of $g = p - q$ typically comes with compensating responses elsewhere. This discourages coordinate-sparsity in g and activates multiple pairwise products $g_i g_j$ simultaneously, making the off-diagonal structure of M_θ visible in the quadratic energy.

Clearly, neither condition alone is sufficient. For instance, a dense g probing a diagonal M_θ would still yield $\text{Ratio}_{\text{diag}} \approx 1$, while a mixed-direction M_θ probed by a sparse g would suppress the off-term. Under SwAV training, however, both conditions are expected to hold simultaneously.

How these structural properties vary when the distributional characteristics of g are altered, or whether this variation correlates with forgetting severity, will be investigated in Chapter 6.

5.4 Summary

This chapter has constructed M_θ as a $K \times K$ representation of old-task sensitivity in prototype space and characterised its intrinsic geometry from two independent views.

In the spectral view (Section 5.2), we empirically demonstrate that nearly all of the quadratic (forgetting) energy, $g^\top M_\theta g$, resides within a compact low-dimensional spectral subspace. This means that the low-rank bottleneck identified at the logit level in Section 3.2 still holds after pulling back from the parameter-space.

On the other hand, experiments reveal that, within the coordinate structure (Section 5.3), the concentrated energy is carried predominantly by cross-prototype coupling terms rather than independent per-prototype contributions.

A concrete question follows: do different distributional properties of g produce significantly different energy activations against the same M_θ? This would establish that g has selective access to M_θ's structure. In other words, the learning signal does not probe the curvature landscape isotropically, but engages it in a manner determined by its own distributional properties.

Our investigation, however, goes beyond merely confirming the existence of such selectivity. We aim to explore how the geometric interactions dictate the variance in forgetting under different learning signals. First, the absolute magnitude of $g^\top M_\theta g$ is the most direct but a somewhat trivial indicator; it is the exact pullback of the classical second-order forgetting penalty $\Delta\theta^\top H_{\theta\theta}^{(t)} \Delta\theta$. Following the logic of Chapter 3 and 4,

where the high-curvature spectral projection ratio (Eq. 4.7) isolated directional alignment from magnitude as an independent determinant of forgetting, we seek an analogous magnitude-invariant structural metric in prototype space. This is the main subject of Chapter 6.

Chapter 6

Probing the Interaction Between the Learning Signal and Prototype-Space Curvature

Chapter 5 characterised the structure of M_θ. Its spectrum is concentrated in a small number of modes. The sensitive structure of M_θ is concentrated in cross-prototype coupling, rather than individual prototype axes. The dangerous directions of old-task knowledge are mixed prototype combinations, not individual coordinates. Here, we move from characterising this structure to examining how the learning signal $g = p - q$ engages with it. We also ask whether the manner of this engagement has measurable consequences for the severity of forgetting.

6.1 Motivation: From Spectral Statistics to Structural Indicators

A natural but insufficient response to Chapter 5's findings would be to repeat the empirical programme of Chapter 4. One could check if the high-curvature spectral component of $g^\top M_\theta g$ is positively correlated with forgetting in prototype coordinates. Yet, this is redundant. $g^\top M_\theta g$ equals $\Delta\theta^\top H^{(t)}_{\theta\theta}\Delta\theta$ when we substitute $\Delta\theta = T'_\theta g$. So, its association with forgetting already follows from Chapter 4. This adds no new structural insight.

In parameter space, $\Delta\theta$ is a high-dimensional vector. It is not directly linked to its generating mechanism. In prototype space, g is the actual disagreement signal. Its properties are observable and interpretable from the codebook geometry. They are also controllable through the Sinkhorn regularisation coefficient ε.

This difference makes prototype space the natural setting for a new question. When the distributional properties of g change, does its engagement with the sensitive structure of M_θ change selectively? Does this selectivity result in measurable changes in forgetting?

Our goal in this chapter is to see whether the engagement of M_θ by the learning signal

allows a structural characterisation beyond just the total magnitude $g^\top M_\theta g$. The key question is not just how much quadratic activation happens, but also how that activation is organised geometrically.

Whether variation in forgetting depends on g, the structure of M_θ, or their interaction is an empirical question. We address this in two stages. First, in Section 6.2, we ask if, under the controlled same-task setting, different distributional regimes of g engage M_θ selectively. This helps establish if structured, non-isotropic activation exists without involving cross-task forgetting. Second, in Section 6.3, we vary g and M_θ independently across tasks. We test if a magnitude-invariant structural quantity from this interaction helps explain differences in forgetting severity.

6.2 Distributional Regimes of g and Their Geometric Consequences

Section 4.2.1 introduced the effective temperature $T_{\text{eff}} = \tau\varepsilon$ from the parameter-space perspective. From prototype space, T_{eff} controls how sharply the target assignment q responds to embedding–prototype geometry. This then shapes the distributional structure of $g = p - q$. Empirically, lower T_{eff} tends to produce a more coordinate-selective and sparser g. The mass concentrates on fewer prototype coordinates. Higher T_{eff} produces a denser g, with mass spread over more coordinates.

The goal of this section is to see if distributional regimes of g are related to systematic differences in how g engages M_θ. We perform this analysis within a single task. This avoids the confounding effects of task switches, which alter both the data distribution and the reference curvature landscape. Here, we aim to show if selectivity exists under the most controlled conditions. We do not make conclusions about forgetting from these measurements. The cross-task relevance of these structural differences is explored in Section 6.3.

Spectral and coordinate signatures under different distributional regimes. To characterise these differences empirically, we track the co-evolution of C_1 (top-1 spectral energy share), IPR, and $\text{Ratio}_{\text{diag}}$ during single-task training for three Sinkhorn settings ($\varepsilon \in \{0.1, 0.01, 0.005\}$). Reusing the eigendecomposition of M_θ introduced earlier, let

$$\pi_k \;=\; \frac{\lambda_k (u_k^\top g)^2}{g^\top M_\theta g}$$

denote the normalised energy carried by the kth mode, so that $\sum_k \pi_k = 1$. We define

$$\text{IPR} \;=\; \sum_{k=1}^{K} \pi_k^2.$$

A larger IPR indicates that the quadratic energy is concentrated in fewer dominant modes, while a smaller IPR indicates a more distributed spectral profile; correspondingly,

1/IPR serves as an effective rank.

Meanwhile, at each epoch's checkpoint, M_θ is anchored at the corresponding weights. We estimate g as the average over a short window of steps after that anchor.

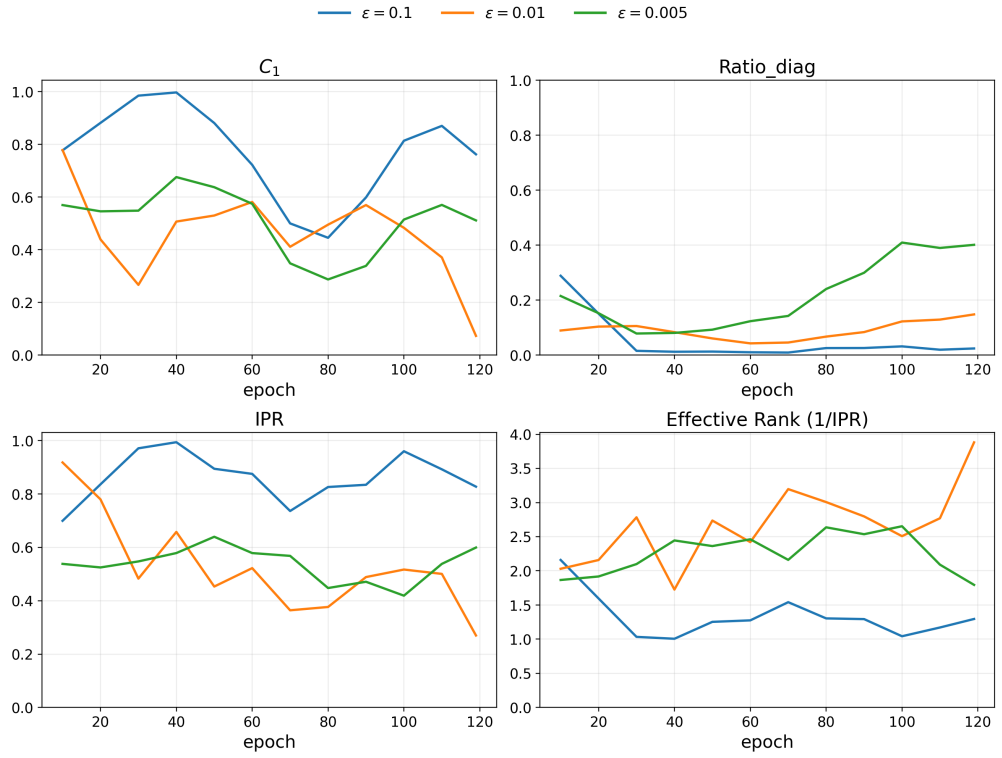

FIGURE 6.1: Temporal evolution of C_1, $\text{Ratio}_{\text{diag}}$, IPR, and effective rank (1/IPR) across single-task training under three Sinkhorn parameter settings, shown as means over 5 independent runs; run-level variability is omitted as shaded bands would obscure the mean trends. At $\varepsilon = 0.1$ (denser g), spectral concentration is consistently high (large C_1, large IPR) and $\text{Ratio}_{\text{diag}}$ remains near zero throughout, indicating that most of the quadratic energy is carried by a very small number of strongly mixed prototype directions. At $\varepsilon = 0.01$, concentration is lower and $\text{Ratio}_{\text{diag}}$ is moderately elevated. Notably, at $\varepsilon = 0.005$, $\text{Ratio}_{\text{diag}}$ indicates stronger coordinate alignment in the late training phase. None of the metrics exhibit a simple monotone ordering with ε, nor do they converge over the course of training.

The results reveal two empirical patterns, which we report without claiming a complete mechanistic account.

First, $\text{Ratio}_{\text{diag}}$ changes in a direction consistent with the sparsity intuition. When g is more selective, fewer cross-coordinate products $g_i g_j$ are simultaneously activated, which tends to reduce the off-diagonal contribution to $g^\top M_\theta g$ regardless of the magnitude of the off-diagonal entries in M_θ. Accordingly, $\varepsilon = 0.005$ produces the highest $\text{Ratio}_{\text{diag}}$, through a late-phase increase, whereas $\varepsilon = 0.1$ remains near zero throughout. This directional trend is consistent with the underlying intuition, though we do not claim it is explained solely by sparsity.

Second, the spectral concentration metrics (C_1, IPR) show a pattern that cannot be reduced to the same intuition: the denser regime is associated with stronger energy

concentration in fewer spectral modes. This is an empirical finding rather than a theoretically derived consequence. The two observations are not contradictory. $\text{Ratio}_{\text{diag}}$ and spectral concentration quantify different geometric properties of the same quadratic interaction: the former concerns coordinate alignment in prototype space, whereas the latter concerns concentration in the eigenbasis of M_θ. Taken together, they show that the way g engages M_θ cannot be reduced to a single geometric axis.

Joint distribution under different regimes. Because the individual metrics fluctuate without clear convergence, we further characterise each regime through the joint statistical distribution of $(\log_{10}(g^\top M_\theta g), \text{Ratio}_{\text{diag}})$ pairs collected throughout training.

Fig. 6.2 shows that the three regimes occupy well-separated regions in the joint space of $(\log_{10}(g^\top M_\theta g), \text{Ratio}_{\text{diag}})$. The separation is systematic: each ε setting produces a distinctive activation signature that is not simply ordered by assignment density. Notably, $\varepsilon = 0.01$—not the densest regime—accumulates the highest $g^\top M_\theta g$, while $\varepsilon = 0.1$ is characterised by near-rank-1 spectral concentration but lower total activation than $\varepsilon = 0.01$. The relationship between ε and total quadratic activation is therefore non-monotone. We do not attempt to explain this causally: within the present same-task setting, ε simultaneously affects assignment sharpness, optimisation dynamics, and the co-evolving curvature structure of M_θ, so attribution to any single mechanism would be unreliable.

The main conclusion of this section is therefore limited but clear: different distributional regimes of g are associated with structured, well-separated activation patterns in M_θ. The engagement is selective rather than isotropic. Since g and M_θ co-evolve within the same training regime here, the observed differences cannot be attributed to g alone. Whether this selectivity remains informative when g and M_θ are varied independently across a task boundary is the question addressed in Section 6.3.

6.3 Selectivity of the Learning Signal Across Curvature Landscapes

6.3.1 Design

Section 6.2 established that different distributional regimes of g produce selective and structured activation patterns in M_θ within a single task. However, those observations were made in a setting where g and M_θ co-evolve under the same training regime, making it impossible to attribute the observed selectivity solely to g. Whether this selectivity has consequences for forgetting, and whether it is driven by the properties of g, by the structure of M_θ, or by their interaction, requires varying g and M_θ independently across a task boundary.

We construct two independent Task_1 anchors by training to convergence under different

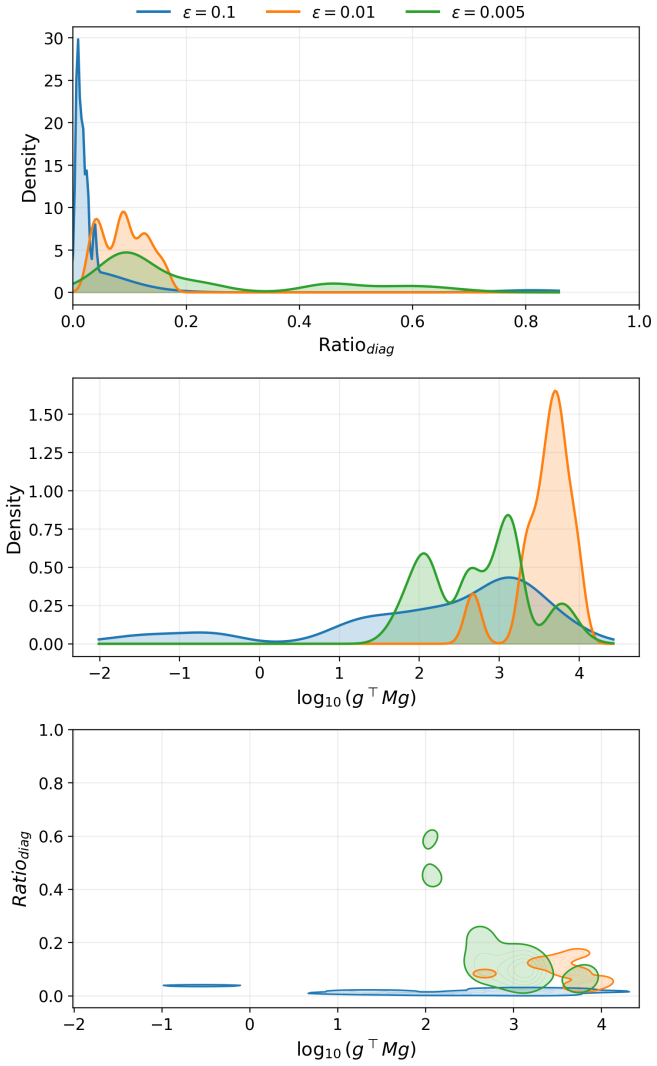

FIGURE 6.2: Marginal and joint distributions of $(\log_{10}(g^\top M_\theta g), \text{Ratio}_{\text{diag}})$ collected over complete single-task training under three Sinkhorn parameter settings, with all per-step observations pooled across 20 independent runs (M_θ anchored at the corresponding epoch checkpoint within the same task). The $\text{Ratio}_{\text{diag}}$ distribution is most concentrated near zero for $\varepsilon = 0.1$, while $\varepsilon = 0.005$ produces the widest spread with a substantial tail toward high values. For $\log_{10}(g^\top M_\theta g)$, $\varepsilon = 0.01$ accumulates the highest energy; $\varepsilon = 0.005$ shows a multi-modal distribution; $\varepsilon = 0.1$ is broadly distributed but at lower levels than $\varepsilon = 0.01$. The three regimes occupy well-separated regions in the joint space, and neither quantity is monotonically ordered by ε.

Sinkhorn parameters ($\varepsilon = 0.1$ and $\varepsilon = 0.005$)[1], producing two curvature landscapes with qualitatively different spectral profiles(Table 6.1). Notably, the dominant eigenvector is nearly identical across the two anchors (cosine similarity 0.991): the most sensitive direction is shared, but the concentration of risk around that direction differs substantially. This allows us to ask whether the same learning signal engages two structurally different M_θ differently, and whether different learning signals engage the same M_θ differently.

[1]Both anchors reach stable regimes but converge to different loss levels, which is expected under ε-governed optimisation dynamics.

TABLE 6.1: Spectral characterisation of M_θ at the two Task_1 anchors (averaged over 5 seeds). The two anchors produce curvature landscapes with qualitatively different spectral profiles. The dominant eigenvector is nearly identical (cosine similarity 0.991): the "most sensitive direction" is shared, but the concentration of risk around that direction differs substantially.

	Anchor A ($\varepsilon_{\text{Task}_1} = 0.1$)	Anchor B ($\varepsilon_{\text{Task}_1} = 0.005$)
Top-1 energy share	91%	70%
Effective rank (1/IPR)	1.53	3.16
Dominant eigenvector cosine	0.991	

From each anchor, we initialise Task_2 training under three Sinkhorn parameter settings ($\varepsilon \in \{0.1, 0.01, 0.005\}$), producing a 2×3 comparison in which both the learning signal and the curvature landscape vary in a controlled manner. All other hyperparameters are kept constant.

For each of the six conditions, we measure: (i) the activation $g^\top M_\theta g$ at the anchor point; (ii) the forgetting trajectory $\mathcal{F}_s$ and the Task_2 training loss (plasticity) over the first 100 training steps. The central question is whether $\text{Ratio}_{\text{diag}}$, as a magnitude-invariant structural metric, can account for the variation in forgetting severity across these six conditions.

6.3.2 Results

Establishing the baseline: where the classical bound suffices and where it does not. Since $g^\top M_\theta g$ is a direct pull-back of the classical second-order forgetting bound into prototype space (Section 5), a larger value at the task switch is expected to correspond to more severe forgetting — this is the bound itself, not an independent finding. We report it here to establish a baseline and to identify the regime in which it ceases to discriminate.

TABLE 6.2: Single-step $g^\top M_\theta g$ (mean $\pm$ std over 5 Task_1 seeds) at the Task_1 anchor under three Task_2 distributional regimes. Under Anchor A, the ordering is clear: $\varepsilon_{\text{Task}_2} = 0.1$ is orders of magnitude larger than the other two. Under Anchor B, all three conditions are in a much narrower range.

	$\varepsilon_{\text{Task}_2} = 0.1$	$\varepsilon_{\text{Task}_2} = 0.01$	$\varepsilon_{\text{Task}_2} = 0.005$
Anchor A ($\varepsilon_{\text{Task}_1} = 0.1$)	90197 ± 1581	9804 ± 1711	11164 ± 3153
Anchor B ($\varepsilon_{\text{Task}_1} = 0.005$)	26664 ± 629	23240 ± 1451	25441 ± 2569

Table 6.2 reports the single-step $g^\top M_\theta g$ at each anchor under each Task_2 condition. Under Anchor A, $\varepsilon_{\text{Task}_2} = 0.1$ produces a value nearly an order of magnitude larger than the other two conditions, and the forgetting trajectories (Fig. 6.3) confirm the expected ordering. In this regime, the classical bound already discriminates — no further structural analysis is required.

A clarification on the relationship between single-step indicators and forgetting trajectories is needed here. The quantities $g^\top M_\theta g$ and $\text{Ratio}_{\text{diag}}$ are single-step measurements:

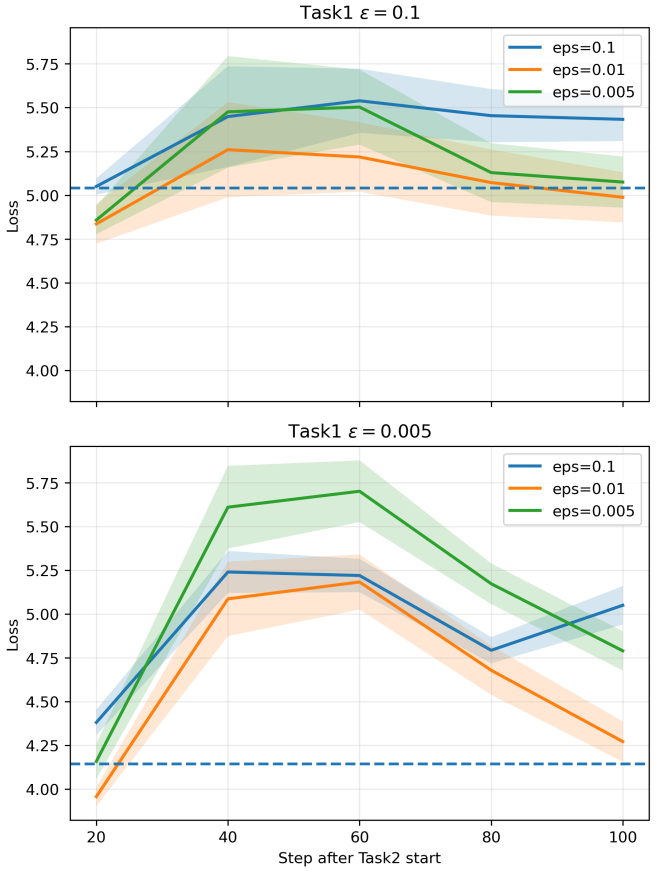

FIGURE 6.3: Forgetting trajectories $\mathcal{F}_s$ over the first 100 training steps of Task_2 under the 2×3 cross-anchor design, shown as mean ± 1 std over 5 Task_1 seeds. Top: Anchor A ($\varepsilon_{\text{Task}_1} = 0.1$). Bottom: Anchor B ($\varepsilon_{\text{Task}_1} = 0.005$). The dashed horizontal line indicates the Task_1 reference loss at the anchor point. Within each panel, three curves correspond to the three Task_2 Sinkhorn settings. The ordering is consistent with the instantaneous $g^\top M_\theta g$ values in Table 6.2, but the separation between conditions narrows substantially under Anchor B, where the three $g^\top M_\theta g$ values are close.

they characterise the forgetting risk incurred by one parameter update, not the accumulated loss increase over a trajectory. The forgetting curves reflect cumulative effects over many steps, during which both quantities evolve. That their ordering broadly tracks early forgetting trajectories is therefore a corroborating observation, not a theoretical requirement.

However, for Anchor B, the three $g^\top M_\theta g$ values are compressed into a much narrower range (23240–26664). The forgetting curves still exhibit clear separation, yet $g^\top M_\theta g$ cannot explain it. This is the regime where the classical bound is saturated as a predictor, and where a structural indicator becomes necessary.

$\text{Ratio}_{\text{diag}}$: a structural indicator for the saturated regime. The limitation of $g^\top M_\theta g$ as a sole indicator is that it measures the total energy without revealing where, within M_θ's geometry, the energy lands. As established in Section 5, the sensitive structure of M_θ is concentrated in off-diagonal coupling terms $M_{ij} g_i g_j$, which means the

genuinely dangerous directions are mixed prototype combinations rather than individual coordinate axes. Hence, a low value of $\text{Ratio}_{\text{diag}}$ (Eq. 5.5) means the energy is predominantly off-diagonal and therefore concentrated in the sensitive structure.

Table 6.3 reports the $\text{Ratio}_{\text{diag}}$ of each anchor's M_θ at Task_1 convergence, as well as the single-step $\text{Ratio}_{\text{diag}}$ after the task switch under all six conditions.

TABLE 6.3: $\text{Ratio}_{\text{diag}}$ at Task_1 convergence for each anchor (header rows) and single-step $\text{Ratio}_{\text{diag}}$ immediately after the task switch (mean ± std over 5 Task_1 seeds) under each Task_2 condition. Higher values indicate more energy carried through coordinate-aligned (diagonal) terms; lower values indicate deeper engagement with M_θ's off-diagonal sensitive structure.

	$\varepsilon_{\text{Task}_2} = 0.1$	$\varepsilon_{\text{Task}_2} = 0.01$	$\varepsilon_{\text{Task}_2} = 0.005$
Anchor A ($\varepsilon_{\text{Task}_1} = 0.1$), $\text{Ratio}_{\text{diag}}$ *at* Task_1 *end* $= 0.02$			
Single-step $\text{Ratio}_{\text{diag}}$	0.041 ±0.011	0.015 ±0.009	0.005 ±0.004
Anchor B ($\varepsilon_{\text{Task}_1} = 0.005$), $\text{Ratio}_{\text{diag}}$ *at* Task_1 *end* $= 0.08$			
Single-step $\text{Ratio}_{\text{diag}}$	0.034 ±0.019	0.170 ±0.034	0.098 ±0.041

$\text{Ratio}_{\text{diag}}$ reveals the missing resolution under Anchor B. The highest value appears at $\varepsilon_{\text{Task}_2} = 0.01$ (0.17), indicating minimal engagement with the off-diagonal sensitive structure. $\varepsilon_{\text{Task}_2} = 0.005$ follows at 0.098, while the small difference between 0.17 and 0.098 is reflected in the slight separation of their early forgetting curves.

Under Anchor A, $\varepsilon_{\text{Task}_2} = 0.1$ exhibits a higher $\text{Ratio}_{\text{diag}}$ than the other two conditions. However, $g^\top M_\theta g$ differs by nearly an order of magnitude across conditions here. While the classical bound discriminates before any structural analysis is needed, $\text{Ratio}_{\text{diag}}$ is not designed to add resolution there. The structurally informative comparisons under Anchor A are between $\varepsilon_{\text{Task}_2} = 0.01$ ($\text{Ratio}_{\text{diag}} = 0.015$) and $\varepsilon_{\text{Task}_2} = 0.005$ ($\text{Ratio}_{\text{diag}} = 0.005$), where energy scales are comparable, and the more off-diagonal condition corresponds to worse forgetting.

Plasticity. Figure 6.4 shows the Task_2 training loss under the 2×3 design. Across both anchors, the trajectories for different $\varepsilon_{\text{Task}_2}$ values are highly overlapping, with no clear or consistent ordering throughout the first 100 training steps. This limited separation suggests that the substantial differences in forgetting across conditions are not mirrored in new-task learning. In particular, the settings that yield lower forgetting do not appear to incur a systematic loss of plasticity.

Although the baseline with $\varepsilon = 0.01$, which achieves the lowest forgetting, also appears to perform slightly better in plasticity, the prototype-space analysis here does not provide further evidence for systematically indicating plasticity across conditions.

6.3.3 Interpretation

The results reveal a hierarchy of indicators for the severity of forgetting. When conditions differ substantially in energy scale, the classical bound $g^\top M_\theta g$ already discriminates, as expected from its theoretical construction as the second-order forgetting penalty.

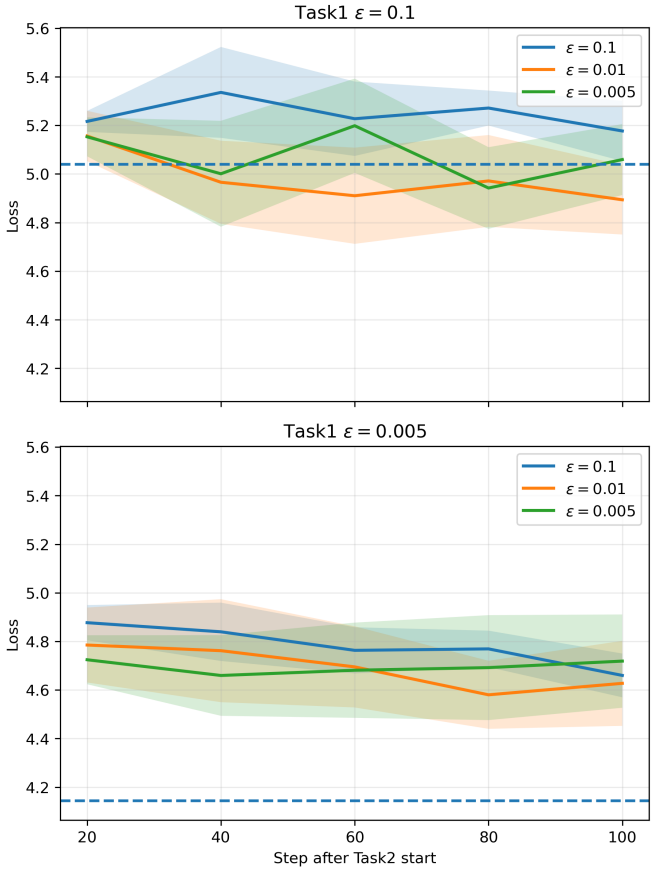

FIGURE 6.4: Plasticity trajectories over the first 100 training steps of Task_2 under the 2×3 cross-anchor design, shown as mean ± 1 std over 5 Task_1 seeds. Top: Anchor A ($\epsilon_{\text{Task1}} = 0.1$). Bottom: Anchor B ($\epsilon_{\text{Task1}} = 0.005$). Within each panel, solid lines denote the mean Task_2 loss and shaded bands denote ± 1 standard deviation for different ϵ_{Task2} values. The variance is relatively large, and the distinction among ϵ_{Task2} settings is therefore modest. Still, the mean curves suggest that $\epsilon_{\text{Task2}} = 0.01$ yields the most favourable plasticity in both cases.

The central contribution is that when the bound is saturated: $\text{Ratio}_{\text{diag}}$, a magnitude-invariant structural metric, resolves forgetting severity through the geometric composition of the interaction rather than its scale. A lower value indicates deeper engagement with M_θ's off-diagonal-sensitive structure and corresponds to more severe forgetting, independent of the total energy magnitude.

Importantly, the bidirectional nature of the interaction is our main interpretive result. Whether a given g can avoid the sensitive structure of M_θ depends not only on g's own distributional properties, but equally on the coordinate structure of M_θ itself. Specifically, the Task_1-end $\text{Ratio}_{\text{diag}}$ of M_θ sets the structural context within which g operates. While Anchor A's M_θ is nearly entirely off-diagonal at convergence ($\text{Ratio}_{\text{diag}} = 0.02$), even the sparsest Task_2 signal leaves almost no coordinate-aligned structure to exploit.

In contrast, Anchor B's M_θ at $\text{Ratio}_{\text{diag}} = 0.08$ offers considerably more room, and the learning signal at the same sparsity level reaches a higher value of $\text{Ratio}_{\text{diag}}$. This is why sparsity alone is not a reliable heuristic for reducing forgetting: if M_θ is densely

off-diagonal, a sparser g may still fail to avoid the sensitive regions by concentrating on fewer prototype coordinates.

Scope of interpretation. Several caveats apply.

First, the instantaneous $g^\top M_\theta g$ and $\text{Ratio}_{\text{diag}}$ at the anchor point should not be interpreted as sufficient predictors of trajectory-level forgetting. The actual forgetting process involves many steps, during which M_θ may drift, and the cumulative interaction may deviate from the instantaneous snapshot. The ordering between single-step indicators and trajectory forgetting is broadly consistent within the probing window in our observation, while this supports an empirical association rather than a causal or sufficient relationship.

Second, steering the Sinkhorn parameter simultaneously alters the optimisation dynamics. The observed pattern may therefore partly reflect differences in optimisation efficiency. We do not favour a purely optimisation-based interpretation, since it would need to explain why these differences track the coordinate structure of M_θ across different settings. The relative ordering of forgetting in our experiments is clearly not arbitrary.

Third, the analysis is restricted to the short post-switch window within which M_θ remains approximately stable (Section 4.2.2). Over longer horizons, the Sinkhorn parameter may alter both the spectral concentration and coordinate structure of M_θ. The long-term co-evolution of g and M_θ under different training regimes is beyond our scope.

6.3.4 Parameter-space cross-validation

As a complementary check, we measure the parameter-space spectral concentration $\eta(K)$ (Eq. 4.6) at the beginning of Task_2 under each condition.

Fig. 6.5 shows that conditions with lower forgetting exhibit reduced spectral concentration in parameter space, which are consistent with the geometric signature established in Chapter 4. This cross-validation indicates that the prototype-space and parameter-space perspectives yield mutually consistent descriptions of the same phenomenon.

6.4 Discussion: Structural Selectivity as a Diagnostic of Forgetting

The experiments in this chapter establish that the learning signal g engages the prototype-space curvature M_θ selectively: different distributional properties of g produce structured, well-separated activation patterns rather than isotropic noise. This selectivity is not merely a geometric curiosity. It underlies measurable differences in forgetting severity across diverse task-switching scenarios.

The central finding is that $\text{Ratio}_{\text{diag}}$ captures forgetting-relevant information that is inaccessible to the classical bound $g^\top M_\theta g$. The bound discriminates when energy scales differ substantially, as expected from its theoretical construction. But when energy scales are comparable, it is the interaction's structural composition, as well as the local

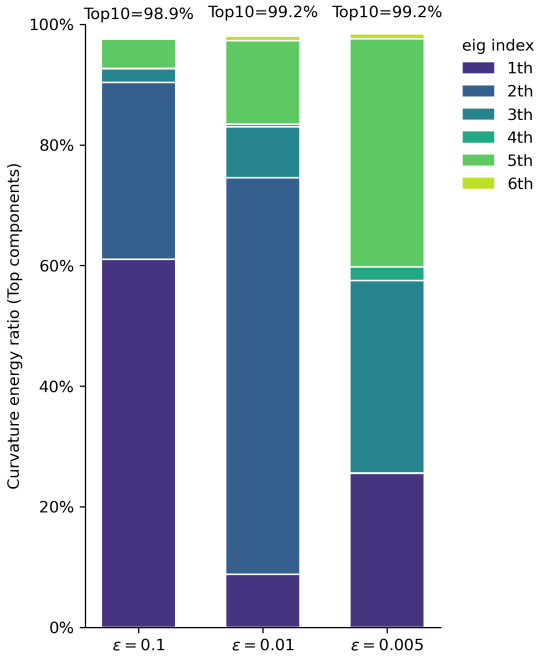

FIGURE 6.5: Distribution of curvature energy across the leading eigen-directions at the beginning of the task switch, averaged under 2 anchors. The magnitude ordering of the high-curvature components matches the observed forgetting severity under both anchors, corroborating the prototype-space findings with the parameter-space metric from Chapter 4.

region in M_θ's geometry, that determines where the energy lands, thereby indicating the forgetting severity.

The role of $\text{Ratio}_{\text{diag}}$ is structurally analogous to the high-curvature projection ratio established in Chapter 4. Both isolate a directional or structural factor from the overall magnitude of the forgetting penalty, treating it as an independent determinant of forgetting severity: the former in the spectral domain of parameter space, and the latter in the coordinate domain of prototype space.

Each of these two ratios carries its own prerequisites. The spectral-domain indicator requires Hessian eigenspace stability and spectral concentration, validated in Section 4. The coordinate-domain indicator operates under the stability of M_θ within the probing window; the relationship between Hessian drift in parameter space and M_θ drift in prototype space over longer horizons remains an open question. Meanwhile, steering ε alters the spectral concentration in parameter space (Fig. 6.5). This suggests a non-trivial coupling between the two domains, whose full characterisation lies beyond the scope of this thesis.

The broader implication is interpretive. In parameter space, forgetting-sensitive directions are identified by the eigendecomposition of $H_{\theta\theta}^{(t)}$, which is numerically computable but semantically opaque. In prototype space, the same sensitivity is expressed through the coordinate structure of M_θ, where directions correspond to prototype combinations

and the learning signal g is directly observable. When g's distributional properties produce selective rather than isotropic engagement with M_θ, it serves as a diagnostic probe of old-task sensitivity, which is interpretable in terms of the codebook geometry rather than opaque parameter-space directions.

Returning to the terrain metaphor: a traveller expends cost in proportion to the roughness of the path traversed. When the trajectory shaped (or selected) by g's distributional properties avoids the most rugged regions of the curvature landscape, the cost incurred is low even if the total displacement is comparable to other trajectories. This is not merely a matter of moving less; it is a matter of moving through flatter ground. The experiments in this chapter suggest that such selective traversal is not hypothetical, and that its structural signature is legible in prototype space.

Chapter 7

Summary

7.1 What This Work Has Established

This thesis starts from a simple point: that forgetting in continual learning is not merely a question of how strongly the learning signal moves parameters, but of where it moves them relative to the sensitive structure of the old-task loss landscape. That requires a setting in which both the learning signal and the landscape can be examined concretely, instead of being inferred indirectly from weight statistics or representation measurements. SwAV provided this setting: its prototype codebook routes all learning through a low-dimensional channel, and the linearity of parameter updates in the assignment disagreement allows the classical second-order forgetting penalty to be pulled back from the full parameter-space into a small, interpretable object in prototype-space.

Building on this intuition, we empirically validated the prerequisites for our analysis in both parameter-space and prototype-space. In the parameter-space, we verified the stability of the second-order geometry and the high concentration of curvature energy. This empirically justified our use of the high-curvature projection as an algorithmic bound for forgetting. Furthermore, we statistically established that this projection is positively correlated with forgetting severity.

When pulling the analysis back into the low-dimensional prototype-space, we characterised the intrinsic geometry of M_θ as the representation of old-task sensitivity in prototype coordinates. Its sensitive structure is concentrated in a small number of eigenmodes, and the dominant energy flows through cross-prototype coupling rather than individual prototype axes. Building on this, we introduced $\mathrm{Ratio}_{\mathrm{diag}}$ as a magnitude-invariant structural metric that characterises how the learning signal engages this sensitive structure.

As illustrated by the traveller and terrain metaphor used throughout this thesis, forgetting is determined not only by how far the learning signal moves parameters, but by which regions of the old-task curvature landscape it traverses. This thesis has shown that this traversal is not arbitrary, and can be understood from two complementary angles:

the statistical distribution of curvature in the high-dimensional parameter-space, and the distributional properties of the learning signal itself in prototype-space that reflect the traveller's own trajectory. Moreover, by expressing both in the same low-dimensional prototype-space, the interaction becomes directly observable and interpretable in terms of the codebook geometry, rather than remaining hidden in opaque parameter-space directions. This interpretability is what opens the door to principled, structure-aware interventions in future work.

7.2 Limitations

Scope of the second-order framework. The entire analytical methodology rests on the assumption that a second-order Taylor expansion provides a meaningful proxy for forgetting over the short-interval window [19]. This is a non-trivial assumption: it requires the linear drift term to remain controlled and the higher-order residuals to be small relative to the quadratic component. The empirical evidence provided is local and regime-specific. This limitation implies that if the proposed bound is to support long-term intervention strategies, mechanisms must be introduced to monitor or control potential second-order fluctuations during training. However, this limitation does not diminish the analytical significance or the empirical validity of our results within the constrained scenarios studied in this thesis.

SwAV as a special-case platform. The pull-back from parameter-space into a low-dimensional signal space is possible here because the SwAV architecture allows a linear relationship between parameter updates and a structured, low-dimensional signal. However, this property is not common in deep learning. In most supervised settings, and even in many other self-supervised frameworks, whether such a structured learning signal can be extracted remains uncertain. SwAV should therefore be viewed as a special-case research platform. The broader idea of using the general learning signal to probe the previously learned knowledge, and to potentially intervene based on it, will require separate investigation.

Appendix A

Mathematical Proofs

A.1 Sufficient Condition for Absorbing the Cross Term of Raw Hessian

Lemma A.1 (Assumption 3.12 in Section 3.4.2)**.** *Let*

$$H_t = \begin{bmatrix} H_{\theta\theta}^{(t)} & H_{\theta C}^{(t)} \\ H_{C\theta}^{(t)} & H_{CC}^{(t)} \end{bmatrix}$$

be a symmetric block matrix with $H_{C\theta}^{(t)} = (H_{\theta C}^{(t)})^\top$*. Assume* $H_{\theta\theta}^{(t)} \succeq 0$ *and* $H_{CC}^{(t)} \succeq 0$*. Suppose there exists a constant* $\beta_t \geq 0$ *such that for all vectors* u *and* v*,*

$$\left|u^\top H_{\theta C}^{(t)} v\right| \;\leq\; \beta_t \sqrt{u^\top H_{\theta\theta}^{(t)} u}\, \sqrt{v^\top H_{CC}^{(t)} v}. \tag{A.1}$$

Then for any $\alpha_\theta, \alpha_C \in [0, 1)$ *satisfying*

$$\beta_t^2 \;\leq\; \alpha_\theta \alpha_C, \tag{A.2}$$

the quadratic form obeys

$$\begin{aligned} \begin{bmatrix} u \\ v \end{bmatrix}^\top H_t \begin{bmatrix} u \\ v \end{bmatrix} &= u^\top H_{\theta\theta}^{(t)} u + 2u^\top H_{\theta C}^{(t)} v + v^\top H_{CC}^{(t)} v \\ &\leq \frac{1}{1-\alpha_\theta}\, u^\top H_{\theta\theta}^{(t)} u + \frac{1}{1-\alpha_C}\, v^\top H_{CC}^{(t)} v. \end{aligned} \tag{A.3}$$

Proof. By Eq. A.1 and Cauchy–Schwarz,

$$2\left|u^\top H_{\theta C}^{(t)} v\right| \;\leq\; 2\beta_t \sqrt{u^\top H_{\theta\theta}^{(t)} u}\, \sqrt{v^\top H_{CC}^{(t)} v}.$$

Apply Young's inequality $2ab \le \varepsilon a^2 + \varepsilon^{-1} b^2$ with $a = \sqrt{u^\top H_{\theta\theta}^{(t)} u}$ and $b = \beta_t \sqrt{v^\top H_{CC}^{(t)} v}$. For any $\varepsilon > 0$,

$$2\beta_t \sqrt{u^\top H_{\theta\theta}^{(t)} u}\, \sqrt{v^\top H_{CC}^{(t)} v} \;\le\; \varepsilon\, u^\top H_{\theta\theta}^{(t)} u + \varepsilon^{-1} \beta_t^2\, v^\top H_{CC}^{(t)} v.$$

Choose $\varepsilon = \frac{\alpha_\theta}{1-\alpha_\theta}$, so that $1 + \varepsilon = \frac{1}{1-\alpha_\theta}$. If additionally $\beta_t^2 \le \alpha_\theta \alpha_C$, then

$$1 + \varepsilon^{-1}\beta_t^2 = 1 + \frac{1-\alpha_\theta}{\alpha_\theta}\beta_t^2 \;\le\; 1 + (1-\alpha_\theta)\alpha_C \;\le\; \frac{1}{1-\alpha_C},$$

where the last inequality uses $\alpha_C \in [0, 1)$. Combining the above bounds with $u^\top H_{\theta\theta}^{(t)} u + v^\top H_{CC}^{(t)} v + 2u^\top H_{\theta C}^{(t)} v \le u^\top H_{\theta\theta}^{(t)} u + v^\top H_{CC}^{(t)} v + 2|u^\top H_{\theta C}^{(t)} v|$ yields Eq. A.3. □

A.2 Justification of the Assumptions in Plasticity Analysis

Assumption 3.4. For fixed C, write the loss as

$$\ell(x; \theta, C) = \phi_C(z_\theta(x)).$$

If ϕ_C is differentiable, then the assumption requires $\|\nabla_z \phi_C(z)\|$ to be uniformly bounded on the relevant region. By the mean value theorem there exists $L_z > 0$ such that

$$|\ell(x; \theta, C) - \ell(x; \theta', C)| \le L_z \, \|z_\theta(x) - z_{\theta'}(x)\|.$$

In the SwAV setting, by Lemma 3.1,

$$\nabla_z \ell(x; w) = \frac{1}{\tau} C(p - q).$$

Since $\|C\|_{\text{op}}$ and $p - q$ are bounded, the gradient is bounded, then the assumption is valid.

Assumption 3.5. This is valid under a local smoothness condition on the encoder. Since

$$\theta^{(s+1)} = \theta^{(s)} - \eta_s g_\theta^{(s)},$$

a first-order expansion gives

$$z_{\theta^{(s+1)}}(x) - z_{\theta^{(s)}}(x) \approx J_\theta z_{\theta^{(s)}}(x)\big(\theta^{(s+1)} - \theta^{(s)}\big) = -\eta_s J_\theta z_{\theta^{(s)}}(x) g_\theta^{(s)}.$$

Therefore,

$$\|z_{\theta^{(s+1)}}(x) - z_{\theta^{(s)}}(x)\| \lesssim \eta_s \|J_\theta z_{\theta^{(s)}}(x) g_\theta^{(s)}\|.$$

Similarly, if there exists a constant $\kappa > 0$ such that

$$\|J_\theta z_{\theta^{(s)}}(x) g_\theta^{(s)}\| \le \kappa \|\nabla_z L^{(s)}(x)\|,$$

then

$$\|z_{\theta^{(s+1)}}(x) - z_{\theta^{(s)}}(x)\| \lesssim \kappa\eta_s\|\nabla_z L^{(s)}(x)\|.$$

Summing over the trajectory yields

$$\|z_{\theta^\star_{t+1}}(x) - z_{\theta^\star_t}(x)\| \leq \kappa \sum_s \eta_s\|\nabla_z L^{(s)}(x)\|,$$

Assumption 3.7. This is valid under a local smoothness condition in the prototype variable C. If $\nabla_C L_{t+1}(\theta, C)$ is β_C-Lipschitz, then by the descent lemma,

$$L_{t+1}(\theta, C') \leq L_{t+1}(\theta, C) + \langle \nabla_C L_{t+1}(\theta, C),\, C' - C\rangle + \frac{\beta_C}{2}\|C' - C\|_F^2,$$

this assumption holds. It is sufficient that the Hessian with respect to C is locally bounded.

Appendix B

Structure of the Sinkhorn Assignment

B.1 Entropic Optimal Transport and the Sinkhorn Form

Consider the entropic regularised transport problem

$$Q^\star \in \arg\min_{Q \in \mathcal{U}(r,c)} \langle Q, M \rangle + \varepsilon \sum_{k,i} Q_{k,i}\big(\log Q_{k,i} - 1\big), \tag{B.1}$$

where $\varepsilon > 0$ is the entropic regularisation coefficient and

$$\mathcal{U}(r,c) = \big\{ Q \in \mathbb{R}_+^{K \times B} : Q\mathbf{1}_B = r,\ Q^\top \mathbf{1}_K = c \big\}$$

imposes marginal constraints over a batch of size B.

Introducing Lagrange multipliers $\alpha \in \mathbb{R}^K$ and $\beta \in \mathbb{R}^B$ for the marginal constraints, the Lagrangian is

$$\mathcal{L}(Q,\alpha,\beta) = \sum_{k,i} Q_{k,i} M_{k,i} + \varepsilon \sum_{k,i} Q_{k,i}(\log Q_{k,i} - 1) + \sum_k \alpha_k \Big(r_k - \sum_i Q_{k,i} \Big) + \sum_i \beta_i \Big(c_i - \sum_k Q_{k,i} \Big).$$

Setting $\partial \mathcal{L} / \partial Q_{k,i} = 0$ yields

$$M_{k,i} + \varepsilon \log Q_{k,i} - \alpha_k - \beta_i = 0,$$

and hence

$$Q^\star_{k,i} = \exp\Big(\tfrac{\alpha_k}{\varepsilon}\Big)\ \exp\Big(-\tfrac{M_{k,i}}{\varepsilon}\Big)\ \exp\Big(\tfrac{\beta_i}{\varepsilon}\Big). \tag{B.2}$$

Defining the kernel matrix $K_{k,i} := \exp(-M_{k,i}/\varepsilon)$ and setting $u_k := \exp(\alpha_k/\varepsilon)$, $v_i := \exp(\beta_i/\varepsilon)$, the solution takes the diagonal scaling form

$$Q^\star = \operatorname{diag}(u)\ K\ \operatorname{diag}(v), \tag{B.3}$$

regardless of the choice of cost matrix M. The vectors (u, v) satisfying the marginal constraints can be found efficiently via Sinkhorn–Knopp iterations.

B.2 Softmax Form of the Column-Normalised Assignment

We now specialise to the SwAV setting and derive the softmax form of the column-normalised assignment.

In the main text, the assignment for sample i is written as the vector

$$q_i = (q_{1i}, \dots, q_{Ki})^\top \in \mathbb{R}^K,$$

where i indexes the sample and k indexes the prototype. Equivalently, over a batch of size B, we may write the assignments in matrix form as

$$Q = [q_{k,i}] \in \mathbb{R}^{K \times B},$$

whose i-th column is exactly q_i.

Let

$$s_{k,i} = \frac{c_k^\top z_i}{\tau}$$

denote the scaled similarity logit between sample i and prototype k, and let

$$p_{k,i} = \frac{\exp(s_{k,i})}{\sum_{m=1}^K \exp(s_{m,i})}$$

be the softmax prior. The SwAV OT objective sets $M^{\mathrm{SwAV}} = -S$, so the kernel becomes

$$K_{k,i}^{\mathrm{SwAV}} = \exp(S_{k,i}/\varepsilon),$$

and by Eq. B.3 the Sinkhorn solution is

$$Q^\star = \mathrm{diag}(u)\ \exp(S/\varepsilon)\ \mathrm{diag}(v). \tag{B.4}$$

Noting that

$$\exp(s_{k,i}/\varepsilon) = p_{k,i}^{1/\varepsilon} \Big(\sum_{m=1}^K \exp(s_{m,i}) \Big)^{1/\varepsilon},$$

this is equivalent to

$$Q^\star_{k,i} = u_k\, p_{k,i}^{1/\varepsilon}\, \tilde{v}_i, \qquad \tilde{v}_i := v_i \Big(\sum_{m=1}^K \exp(s_{m,i}) \Big)^{1/\varepsilon}. \tag{B.5}$$

Column-normalised assignment. The target code for sample i is the column-normalised version of $Q^\star$:

$$q_{k,i} := \frac{Q^\star_{k,i}}{\sum_{j=1}^K Q^\star_{j,i}} = \frac{u_k\, p_{k,i}^{1/\varepsilon}\, \tilde{v}_i}{\sum_{j=1}^K u_j\, p_{j,i}^{1/\varepsilon}\, \tilde{v}_i} = \frac{u_k\, p_{k,i}^{1/\varepsilon}}{\sum_{j=1}^K u_j\, p_{j,i}^{1/\varepsilon}},$$

where $\tilde{v}_i$ cancels. Substituting

$$p_{k,i}^{1/\varepsilon} = \frac{\exp(s_{k,i}/\varepsilon)}{\left(\sum_{m=1}^K \exp(s_{m,i})\right)^{1/\varepsilon}},$$

we obtain

$$q_{k,i} = \frac{u_k \exp(s_{k,i}/\varepsilon)}{\sum_{j=1}^K u_j \exp(s_{j,i}/\varepsilon)}.$$

Since $u_k > 0$, writing

$$u_k = \exp(\log u_k),$$

and defining the effective logits

$$\tilde{s}_{k,i} := s_{k,i} + \varepsilon \log u_k,$$

we obtain

$$q_{k,i} = \frac{\exp\big(\tilde{s}_{k,i}/\varepsilon\big)}{\sum_{j=1}^K \exp\big(\tilde{s}_{j,i}/\varepsilon\big)}. \tag{B.6}$$

Therefore, the assignment vector

$$q_i = (q_{1i}, \ldots, q_{Ki})^\top$$

is a softmax distribution over shifted logits

$$\tilde{s}_{k,i} = s_{k,i} + \varepsilon \log u_k.$$

If we define the unscaled similarity by

$$\hat{s}_{k,i} := c_k^\top z_i,$$

then $s_{k,i} = \hat{s}_{k,i}/\tau$, and Eq. B.6 can be rewritten as

$$q_{k,i} = \frac{\exp\big(\bar{s}_{k,i}/T_{\text{eff}}\big)}{\sum_{j=1}^K \exp\big(\bar{s}_{j,i}/T_{\text{eff}}\big)}, \qquad T_{\text{eff}} := \tau\varepsilon,$$

where

$$\bar{s}_{k,i} := \hat{s}_{k,i} + T_{\text{eff}} \log u_k.$$

Hence the assignment vector q_i (equivalently, the i-th column of Q) remains a softmax distribution with effective temperature $T_{\text{eff}} = \tau\varepsilon$.

B.2.1 SeLa and SwAV as Special Cases

The unified form Eq. B.3 shows that SeLa [1] and SwAV differ only in their choice of cost matrix M, and hence kernel K.

SeLa. SeLa sets $M^{\text{SeLa}}_{k,i} = -\log P_{k,i}$, where $P_{k,i} = p_\theta(k \mid x_i)/N$ is the dataset-level joint prior. The kernel is

$$K^{\text{SeLa}}_{k,i} = \exp\Big(-\tfrac{M^{\text{SeLa}}_{k,i}}{\varepsilon}\Big) = P_{k,i}^{1/\varepsilon},$$

giving $Q^\star_{\text{SeLa}} = \text{diag}(u)\, P^{1/\varepsilon}\, \text{diag}(v)$.

SwAV. SwAV sets $M^{\text{SwAV}}_{k,i} = -S_{k,i} = -c_k^\top z_i/\tau$. The kernel is

$$K^{\text{SwAV}}_{k,i} = \exp\Big(\tfrac{S_{k,i}}{\varepsilon}\Big),$$

giving $Q^\star_{\text{SwAV}} = \text{diag}(u)\ \exp(S/\varepsilon)\, \text{diag}(v)$.

In both cases the Sinkhorn scaling structure is identical; only the kernel differs. This confirms that the softmax form derived in Section B.2 applies to SwAV directly, and that SeLa's probability-based prior and SwAV's similarity-based scores are two instantiations of the same entropic OT framework.

B.2.2 Intuition: Sinkhorn Assignment Hardness and Forgetting Risk

The softmax-equivalent form derived in Section B.2 suggests that the sensitivity of the OT assignment depends on the separation pattern of the effective logits. When one effective logit dominates, the assignment becomes nearly hard, so perturbations are less likely to redistribute mass across many prototype indices. When several effective logits compete closely, small perturbations can induce larger redistributions, which makes directional changes in $g = p - q$ more likely (Fig. 4.3). To discuss this regime more formally, we first introduce a simple notion of assignment hardness.

Let $\tilde{s}_{i(1)}$ and $\tilde{s}_{i(2)}$ denote the largest and second-largest effective logits for sample i, and define the margin

$$\Delta_i = \tilde{s}_{i(1)} - \tilde{s}_{i(2)}.$$

Given a tolerance parameter $\delta \in (0, 1)$, we say that the assignment for sample i is δ-*hard* if

$$q_{i(1)} \geq 1 - \delta.$$

Proposition B.1. *If*

$$\Delta_i \geq T_{\text{eff}} \log \frac{K-1}{\delta},$$

then the assignment for sample i is δ-hard.

Proof. Let $k^\star = (1)$ denote the index of the largest effective logit, i.e., $\tilde{s}_{k^\star,i} = \tilde{s}_{i(1)}$. By definition,

$$q_{k^\star,i} = \frac{\exp(\tilde{s}_{k^\star,i}/T_{\text{eff}})}{\sum_{j=1}^{K} \exp(\tilde{s}_{j,i}/T_{\text{eff}})}.$$

Hence

$$\begin{aligned}
1 - q_{k^\star,i} &= \frac{\sum_{j=1}^{K} \exp(\tilde{s}_{j,i}/T_{\text{eff}}) - \exp(\tilde{s}_{k^\star,i}/T_{\text{eff}})}{\sum_{j=1}^{K} \exp(\tilde{s}_{j,i}/T_{\text{eff}})} \\
&= \frac{\sum_{k\neq k^\star} \exp(\tilde{s}_{k,i}/T_{\text{eff}})}{\exp(\tilde{s}_{k^\star,i}/T_{\text{eff}}) + \sum_{k\neq k^\star} \exp(\tilde{s}_{k,i}/T_{\text{eff}})} \\
&\leq \frac{\sum_{k\neq k^\star} \exp(\tilde{s}_{k,i}/T_{\text{eff}})}{\exp(\tilde{s}_{k^\star,i}/T_{\text{eff}})} \\
&= \sum_{k\neq k^\star} \exp\left(\frac{\tilde{s}_{k,i} - \tilde{s}_{k^\star,i}}{T_{\text{eff}}}\right).
\end{aligned}$$

For every $k \neq k^\star$, we have $\tilde{s}_{k,i} \leq \tilde{s}_{i(2)}$, hence

$$\tilde{s}_{k,i} - \tilde{s}_{k^\star,i} \leq \tilde{s}_{i(2)} - \tilde{s}_{i(1)} = -\Delta_i.$$

Therefore,

$$\exp\left(\frac{\tilde{s}_{k,i} - \tilde{s}_{k^\star,i}}{T_{\text{eff}}}\right) \leq \exp\left(-\frac{\Delta_i}{T_{\text{eff}}}\right),$$

and so

$$1 - q_{k^\star,i} \leq \sum_{k\neq k^\star} \exp\left(-\frac{\Delta_i}{T_{\text{eff}}}\right) = (K-1)\exp\left(-\frac{\Delta_i}{T_{\text{eff}}}\right).$$

Thus, if

$$(K-1)\exp\left(-\frac{\Delta_i}{T_{\text{eff}}}\right) \leq \delta,$$

then

$$1 - q_{i(1)} = 1 - q_{k^\star,i} \leq \delta,$$

□

Based on Proposition (B.1), we give a supporting intuition for the two-regime picture in Fig 4.3. Clearly, in a sufficient hard regime, most assignment mass is concentrated on a single prototype, so small perturbations are less likely to spread mass across many indices.

In contrast, when the top effective logits are close, the assignment is more sensitive to competition among prototypes, and perturbations can induce stronger redistribution effects across coordinates. This provides an intuitive explanation for why competing effective logits may lead to more directional changes in $g = p - q$, and hence potentially to greater forgetting risk through the induced displacement $\Delta\theta$.

Appendix C

Implementation of Computation

C.1 Implementation Details for the Spectral Concentration

C.1.1 Implementation of $\eta(K)$

We briefly describe how Eq. 4.6 is computed in practice. For the θ-block, let

$$H \triangleq H_{\theta\theta}^{(t)}, \qquad \delta \triangleq \Delta\theta,$$

where H is the anchored old-task Hessian constructed from the old training data, and δ is the parameter-space direction induced by the new sample.

The denominator in Eq. 4.6,

$$\delta^\top H \delta,$$

is computed directly by a single Hessian–vector product (HVP):

$$q \leftarrow H\delta, \qquad E_{\text{tot}} \leftarrow \delta^\top q.$$

We do not evaluate this denominator through the full spectral expansion

$$\sum_{i=1}^{D} \lambda_i \langle u_i, \delta \rangle^2,$$

since the HVP-based computation is exactly equivalent as a quadratic-form evaluation, while being substantially cheaper.

The numerator in Eq. 4.6,

$$\sum_{i=1}^{K} \lambda_i \langle u_i, \delta \rangle^2,$$

requires only the leading eigenpairs of H. These are obtained by applying Lanczos.

Algorithm 1 Computation of $\eta(K)$ via HVP and Lanczos

Input: old-task optimum $\omega_t^\star$, current parameter ω_{cur}, old training data $\mathcal{D}_{\text{old}}^{(t)}$, evaluation batch $\mathcal{B}_{\text{new}}$, truncation level K_{big}
Output: $\{\eta(K)\}_{K=1}^{K_{\text{big}}}$

Compute the evaluation direction at the current parameter

$$\delta \leftarrow \nabla_\theta L(\omega_{\text{cur}}; \mathcal{B}_{\text{new}})$$

Define the Hessian–vector product operator anchored at the old-task optimum

$$\text{HVP}(v) \leftarrow \nabla_\theta^2 L(\omega_t^\star; \mathcal{D}_{\text{old}}^{(t)})\, v$$

Compute the total quadratic energy

$$E_{\text{tot}} \leftarrow \delta^\top \text{HVP}(\delta)$$

Run Lanczos on HVP

$$(\Lambda, U) \leftarrow \text{Lanczos}(\text{HVP}, K_{\text{big}})$$

where

$$\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{K_{\text{big}}}), \qquad U = [u_1, \ldots, u_{K_{\text{big}}}].$$

Project δ onto the leading eigenspace

$$a \leftarrow U^\top \delta$$

Form cumulative top-K weighted energies

$$s_K \leftarrow \sum_{i=1}^{K} \lambda_i a_i^2, \qquad K = 1, \ldots, K_{\text{big}}$$

Return

$$\eta(K) \leftarrow \frac{s_K}{E_{\text{tot}}}, \qquad K = 1, \ldots, K_{\text{big}}$$

In the experiments, we do not interpret a small fixed K in isolation. Instead, we first sweep K over a sufficiently large leading range and verify that $\eta(K)$ saturates. We then denote by K_{big} the smallest K for which this saturation is reached. Thus, the empirical role of $\eta(K)$ is to establish that the second-order energy is concentrated within a resolved leading spectral range, after which the finer high-curvature partition is studied.

C.1.2 Implementation of $R(\lambda)$

The second-order computation required for Eq. 4.7 is identical to that of $\eta(K)$. The only difference is that $R(\lambda)$ does not use the eigenvalue weights λ_i, and is instead computed from unweighted squared projection norms.

C.1.3 Brief Notes on Lanczos Computation

Since $H = H_{\theta\theta}^{(t)}$is too large to construct explicitly, we instead adopt Lanczos to approximate the leading eigenpairs of H. Starting from a random unit vector q_1, Lanczos iteratively builds an orthonormal basis of the Krylov subspace

$$\mathcal{K}_m(H, q_1) = \text{span}\{q_1,\ Hq_1,\ H^2q_1,\ \ldots,\ H^{m-1}q_1\}.$$

This produces a basis matrix

$$Q_m = [q_1, \ldots, q_m],$$

and reduces the original large eigenvalue problem to a much smaller tridiagonal matrix

$$T_m = Q_m^\top H Q_m.$$

The eigenpairs of T_m then provide approximations to the leading eigenpairs of H:

$$T_m y_i = \mu_i y_i, \qquad \lambda_i \approx \mu_i, \qquad u_i \approx Q_m y_i.$$

C.1.4 Implementation of $g^\top M_\theta g$

We describe how the quadratic form $g^\top M_\theta g$ is computed in practice.

Here, g is the prototype-coordinate discrepancy induced by a new sample under the anchored old model. Concretely, for a new input, we compute its prototype prediction p and assignment target q under the old model[1], and form

$$g \triangleq p - q.$$

The matrix M_θ is defined by

$$M_\theta = T_\theta^\top H_{\theta\theta}^{(t)} T_\theta,$$

where $H_{\theta\theta}^{(t)}$ is the anchored old-task Hessian at $\theta_t^\star$, and T_θ is the Jacobian mapping from prototype-space perturbations to parameter-space variations, as defined in the main text. Unlike the parameter-space quadratic form $\Delta\theta^\top H_{\theta\theta}^{(t)} \Delta\theta$, this quantity cannot be evaluated by a single second-order operation, because it involves both the Hessian and the Jacobian sandwich $T_\theta^\top H_{\theta\theta}^{(t)} T_\theta$.

In deep learning, neither the full Hessian nor the full Jacobian is practical to construct explicitly. Instead, both are accessed only through their products with vectors. As discussed above, the Hessian is accessed through Hessian–vector products (HVPs). Similarly, the Jacobian is accessed through Jacobian–vector products (JVPs), obtained by differentiating the forward computation (using anchor weights) while injecting the vector g into the tangent direction.

[1]Notably, since $g^\top M_\theta g$ is defined within a single-step design, g is computed under the anchored optimal weights, rather than from the trajectory-level parameter displacement used in Algorithm 1.

Concretely, the computation of $M_\theta g$ is carried out through the following chain:

$$T_\theta g \leftarrow \mathrm{JVP}(g),$$

$$H_{\theta\theta}^{(t)} T_\theta g \leftarrow \mathrm{HVP}\big(\mathrm{JVP}(g)\big),$$
$$M_\theta g = T_\theta^\top H_{\theta\theta}^{(t)} T_\theta g \leftarrow T_\theta^\top \Big(\mathrm{HVP}\big(\mathrm{JVP}(g)\big)\Big),$$

where the last step of transpose Jacobian action, Vector-Jacobian Product, or VJP, is also implemented by automatic differentiation.

In overall, we first map g from prototype space to parameter space through the Jacobian action $T_\theta g$, then apply the anchored old-task Hessian through an HVP, and finally map the result back to prototype space through a VJP action.

Once $M_\theta g$ is obtained, the quadratic form is evaluated as

$$g^\top M_\theta g = g^\top (M_\theta g).$$

C.1.5 Implementation of $\mathrm{Ratio}_{\mathrm{diag}}$

To compute $\mathrm{Ratio}_{\mathrm{diag}}$, we do not explicitly construct M_θ. Instead, for each prototype coordinate i, let e_i denote the i-th standard basis vector in prototype space. Then

$$(M_\theta)_{ii} = e_i^\top M_\theta e_i.$$

Each such quantity is computed by reusing the same second-order pipeline as in Algorithm 2, but with e_i replacing g. Concretely, we apply the JVP–HVP–VJP chain to e_i, which yields

$$M_\theta e_i,$$

and then extract

$$(M_\theta)_{ii} = e_i^\top M_\theta e_i.$$

Equivalently, this can be viewed as evaluating the quadratic form associated with the i-th prototype basis direction.

Once all diagonal entries $(M_\theta)_{ii}$ have been obtained, they are combined with the disagreement coordinates to form the empirical diagonal contribution

$$\sum_{i=1}^{K} (M_\theta)_{ii}\, g_i^2,$$

where K is the prototype dimension.

Algorithm 2 Computation of $g^\top M_\theta g$ via JVP, HVP, and VJP

Input: old-task optimum $(\theta_t^\star, C_t^\star)$, old training data $\mathcal{D}_{\text{old}}^{(t)}$, new sample x, Sinkhorn operator Sinkhorn$(\cdot)$
Output: g, $M_\theta g$, and $g^\top M_\theta g$

Compute the anchored feature of the new sample

$$z \leftarrow f_{\theta_t^\star}(x)$$

Compute similarities under the anchored old model

$$s \leftarrow C_t^{\star\top} z$$

Compute the prototype prediction

$$p \leftarrow \text{softmax}(s)$$

Compute the assignment target using the Sinkhorn operator

$$q \leftarrow \text{Sinkhorn}(s)$$

Form the disagreement

$$g \leftarrow p - q$$

Define the Hessian–vector product operator at $\theta_t^\star$ using the old training data

$$\text{HVP}(v) \leftarrow \nabla_\theta^2 L(\theta_t^\star; \mathcal{D}_{\text{old}}^{(t)})\, v$$

Define the Jacobian–vector product operator at $(\theta_t^\star, x)$

$$\text{JVP}(v) \leftarrow T_\theta v$$

Define the vector–Jacobian product operator at $(\theta_t^\star, x)$

$$\text{VJP}(u) \leftarrow T_\theta^\top u$$

Map g from prototype space to parameter space

$$u \leftarrow T_\theta g \;\leftarrow\; \text{JVP}(g)$$

Apply the anchored old-task Hessian

$$w \leftarrow H_{\theta\theta}^{(t)} u \;\leftarrow\; \text{HVP}(u)$$

Map back to prototype space through the transpose Jacobian action

$$m \leftarrow T_\theta^\top w \;\leftarrow\; \text{VJP}(w),$$

Compute the quadratic form

$$g^\top M_\theta g \leftarrow g^\top m$$

Return g, $m = M_\theta g$, and $g^\top M_\theta g$

Appendix D

SwAV Configuration and Controlled Design Choices

D.1 SwAV Hyperparameters

The main SwAV hyperparameters, and several design choices are kept fixed to reduce confounding factors. The main text focuses on the effect of the Sinkhorn entropic regularization parameter ε. In particular, our analysis mainly compares three settings $\varepsilon \in \{0.1,\ 0.01,\ 0.005\}$.

Table D.1 summarizes the SwAV hyperparameters used for different datasets.

D.2 Controlled Design Choices

SwAV was originally proposed with several design components that can improve training stability or representation quality. In this thesis, however, our goal is not to fully optimize SwAV itself. Instead, we aim to isolate the specific factors that are most relevant to the mechanism studied in the main text. Therefore, several components were deliberately fixed or removed.

Fixed crop count across all baselines. A distinctive feature of SwAV is the multi-crop strategy, where multiple small crops are added in addition to the large crops. These extra small views enrich the comparison across views and are an important part of the original SwAV design. However, they also introduce an additional source of variation. Although the gradient derivation in Section 3.1 shows that different views contribute through linear additive terms, changing the number of crops still changes the overall composition of the learning signal.

In practice, each augmented view is implemented using PyTorch's `RandomResizedCrop`. For each crop configuration, a region is randomly sampled from the input image with its scale constrained by `min_scale_crops` and `max_scale_crops`, and this sampled region is then resized to the corresponding target size in `size_crops`. The resulting view is further

Table D.1: Common and dataset-specific SwAV hyperparameters.

Common configuration	
Hyperparameter	**Value**
Number of crops	2
Softmax temperature τ	0.1
Number of Sinkhorn iterations	10
Training epochs	120
Base learning rate	0.05
Backbone	ResNet18
Hidden MLP	2048
Feature dimension	128
Split CIFAR-100 specific	
Crop size	128×128
Min Scale Crops	0.4x
Max Scale Crops	1.0x
Batch size	256
Number of Prototypes	250
Tiny-ImageNet and Food101 specific	
Crop size	224×224
Min Scale Crops	0.14x
Max Scale Crops	1.0x
Batch size	64
Number of Prototypes	100

processed by random horizontal flipping, color transformations, tensor conversion, and normalization, following the standard SwAV augmentation pipeline.

To avoid this confound, we keep the number of crops the same for all datasets and all baselines. Specifically, all experiments use only two large crops. We do not include the extra small crops encouraged in the original SwAV design. The crop size itself may differ across datasets because the original image size differs, but the crop count is always fixed.

No queue mechanism. We also do not use the queue mechanism. In standard SwAV, the queue acts like a sliding buffer that stores features from previous batches, which can make the computation of Sinkhorn assignments more stable. However, this design is not naturally compatible with the task-transition setting considered in this thesis, because cached features from earlier batches would mix information across different temporal stages of training.

For this reason, the queue is excluded from all experiments. Although such a buffering mechanism may be relevant for continual learning algorithms (e.g., memory-replay methods), it is outside the scope of this thesis.

No prototype-freezing warm-up. Similarly, we do not use the standard SwAV heuristic that temporarily stops updating the prototype parameters during the early stage of training. This heuristic is useful when the network is still unstable after initialization. In that stage, the behaviour of SwAV becomes closer to a SeLa-style self-labeling

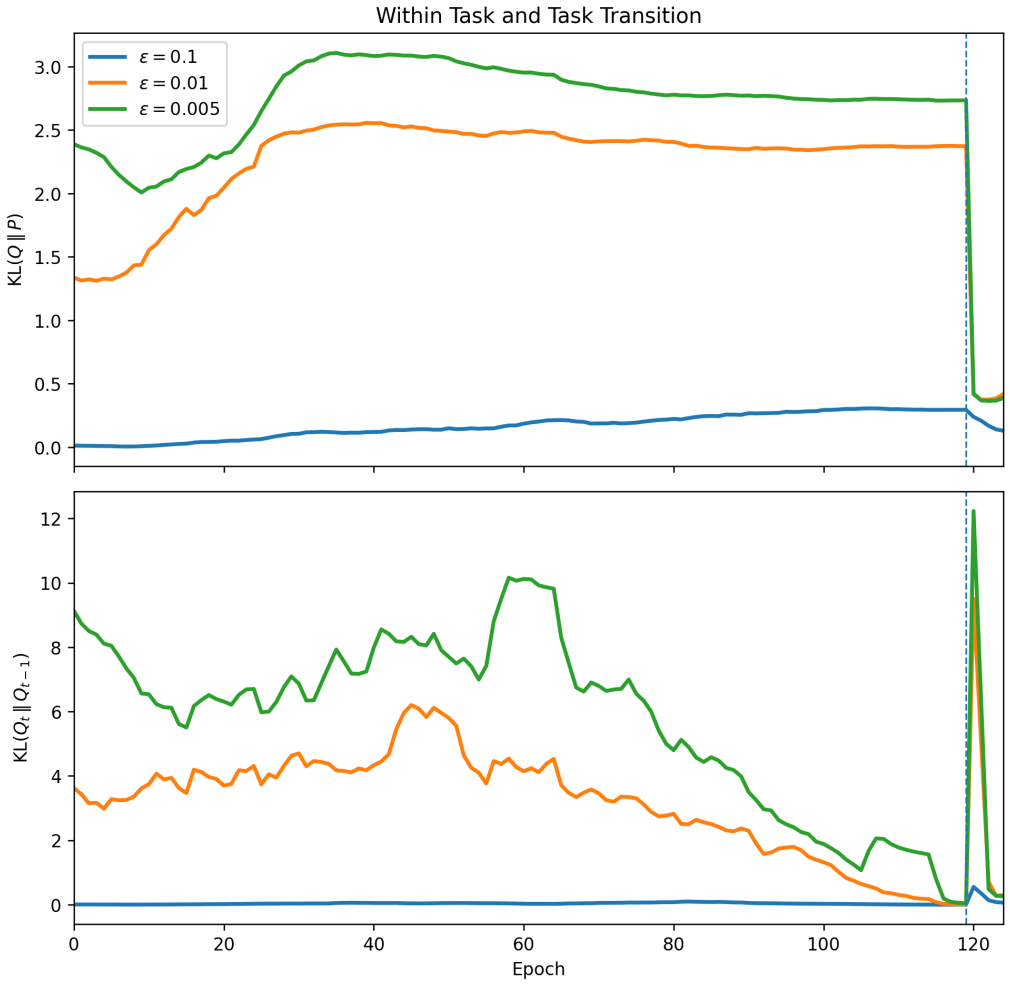

FIGURE D.1: Auxiliary KL indicators for Split CIFAR-100 under different Sinkhorn entropic regularization strengths. All other training settings are kept fixed, and only $\varepsilon \in \{0.1, 0.01, 0.005\}$ is changed. Each curve contains the full within-task trajectory of Task_1, followed by the first five epochs after transition to the next task. As in the main text, the transition part is averaged over multiple Task_2 runs under the same Task_1 setting.

regime. However, introducing a special warm-up rule would make the comparison between within-task training and task transition less clean.

Therefore, prototype updates are treated consistently throughout the experiments. As with the queue mechanism, slow-update or delayed-update designs may be interesting for continual learning.

D.2.1 Auxiliary Indicators Related to Assignment Stability

To complement the mechanism-level analysis in the main text, we report two auxiliary quantities related to the SwAV assignment process: $\mathrm{KL}(q_t \,\|\, q_{t-1})$ measures how much the Sinkhorn assignment changes between consecutive probe points, while $\mathrm{KL}(p_t \,\|\, q_t)$ measures the discrepancy between the model prediction and the assignment at the same point.

This analysis is intended only as an auxiliary view of the SwAV mechanism. Fig. D.1 shows that the trend of $\mathrm{KL}(q_t \,\|\, q_{t-1})$ is consistent with the role of ε in controlling assignment hardness: smaller ε produces harder and less temporally stable assignments, while larger ε produces softer and more stable ones (see discussion in Appendix B.2). This is expected, since ε directly shapes the concentration of q.

The quantity $\mathrm{KL}(p_t \,\|\, q_t)$ is also informative because it is closely related to the learning signal introduced in the main text,

$$g_t = p_t - q_t.$$

Although $\mathrm{KL}(p_t \,\|\, q_t)$ is not identical to a quadratic norm of g_t, it reflects the same mismatch between prediction and assignment. Thus, it provides a simple auxiliary proxy for the magnitude of the low-dimensional learning signal. Consistent with the main text, Fig. D.1 shows that different ε values induce learning signals of clearly different magnitudes, with smaller ε generally leading to a larger prediction–assignment mismatch.

D.2.2 Prototype Count Across Datasets

To reduce additional confounding factors, this thesis does not include an explicit ablation study on the number of prototypes. Nevertheless, the number of prototypes is not identical across all datasets.

The reason is mainly computational. In our setting, Sinkhorn performs a batch-level constrained matching between the current batch and the prototype codebook. As a result, the feasible choice of prototype number is related to the batch size. Since different datasets are trained under different computational budgets, the batch size is not always the same, and the prototype number is adjusted accordingly.

Therefore, the prototype count may vary across datasets, but this is a practical choice for stable training under limited resources, rather than an independent factor under study. A systematic same-dataset ablation on prototype number would be interesting, but it is beyond the scope of this thesis.

Appendix E

Empirical Results on Other Datasets

E.1 Empirical Demonstration on Other Datasets

E.1.1 Datasets

Beyond CIFAR-100, we additionally conduct experiments on Tiny ImageNet and Food-101. Together with CIFAR-100, these datasets provide three distinct visual regimes with different image resolutions, class granularity, and training constraints.

CIFAR-100. CIFAR-100 contains 100 classes of small natural images and is one of the most widely used benchmarks in class-incremental learning. Its relatively large number of classes makes it particularly suitable for continual learning settings in which new classes are introduced progressively. In our study, CIFAR-100 serves as the main benchmark in the core chapters.

Tiny ImageNet. Tiny ImageNet also contains a relatively large number of classes and is therefore similarly well suited to the class-incremental setting. Compared with CIFAR-100, it has higher-resolution images and greater visual complexity, while still remaining much smaller and more manageable than full ImageNet. SwAV was originally developed and evaluated on ImageNet, and this motivates our inclusion of Tiny ImageNet as an additional dataset.

Food-101. Food-101 provides a different regime from CIFAR-100 and Tiny ImageNet. Although it is not as standard in class-incremental learning as the previous two datasets, it contains a large number of visually related fine-grained classes, which makes it useful for testing whether the structural phenomena studied in this thesis remain visible under a more challenging semantic setting.

Although these three datasets all contain many classes, they differ substantially in image resolution and visual statistics. As a result, we do not use exactly the same data

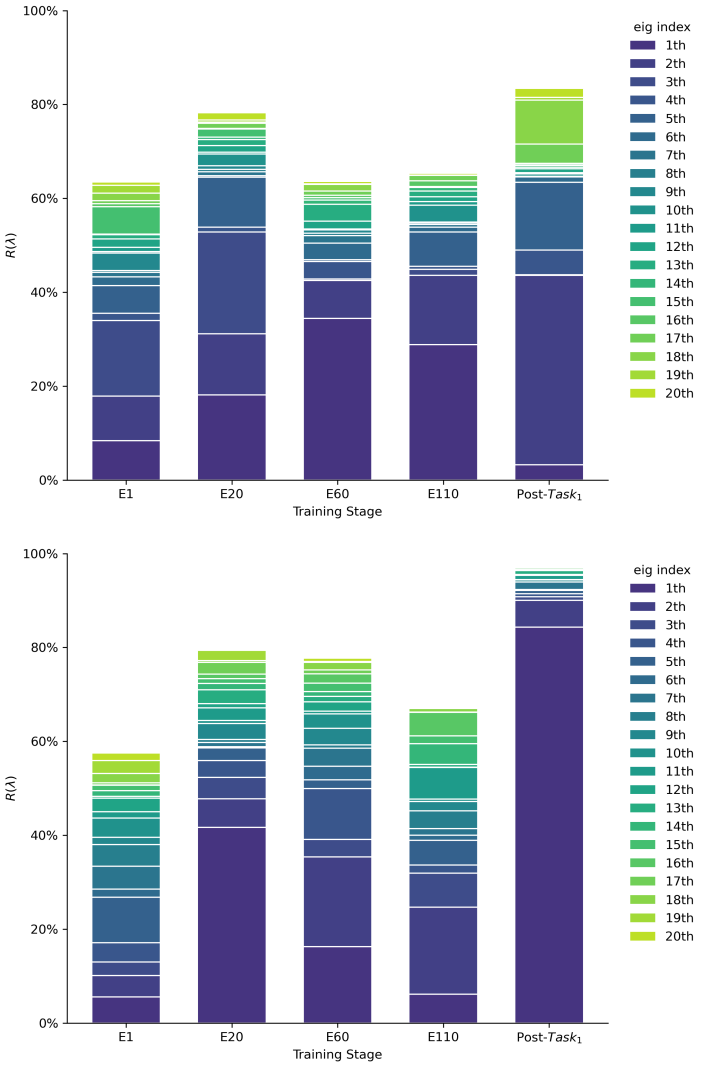

FIGURE E.1: Spectral concentration measured by $R(\lambda)$ on Tiny ImageNet (top) and Food-101 (bottom). For each dataset, we randomly sample 10 classes to form a single within-task training stream, and evaluate the spectrum at four within-task anchor points ($E1$, $E20$, $E60$, and $E110$), together with one cross-task point taken at the first epoch after entering Task_2. As in the CIFAR-100 experiments, all runs use a total of 120 epochs for Task_1, and the cross-task anchor is taken from the final checkpoint of Task_1. In each stacked bar, λ denotes the leading 20 eigencomponents, ordered from bottom to top, and $R(\lambda)$ reports the share of each of these top-20 components relative to the total energy of the top-80 spectrum. As in the main text, we also verify that $\eta(K_{\text{big}} = 80)$ is close to 1. For the cross-task evaluation on Task_2, we test five different Task_2 class sets, and each setting is repeated five times; the reported results are averaged across these runs.

pipeline across all datasets (see details in Table D.1). Different resizing and cropping strategies are adopted to accommodate their different resolutions. These choices also lead to different memory costs during training, which in turn require different batch sizes. The number of prototypes is adjusted accordingly. Therefore, the experiments in this appendix should not be interpreted as a controlled comparison of SwAV performance across datasets. Instead, they are designed to test whether the theoretical

TABLE E.1: Later-stage (Epoch 115) θ-block Hessian alignment statistics for the four evaluation checkpoints on Tiny ImageNet and Food101.

Checkpoint	Overlap	Mean singular
Tiny ImageNet		
+50 step	0.9919	0.9351
+100 step	0.9823	0.9090
+200 step	0.9079	0.8792
+300 step	0.7692	0.6790
Food-101		
+50 step	0.9352	0.9025
+100 step	0.8966	0.8592
+200 step	0.7872	0.7725
+300 step	0.7289	0.5320

patterns identified in the main text remain observable under different data regimes and training configurations.

E.1.2 Experimental Results

The results on Tiny ImageNet and Food-101 support the central qualitative claims established on CIFAR-100.

First, the high-curvature concentration remains clearly visible on both additional datasets (Fig.E.1). Although the detailed spectral shapes differ from those reported for CIFAR-100 in the main text, the most important observation is unchanged: the dominant curvature energy is still concentrated in a small number of leading directions. An especially interesting case is Food-101. After the task transition, instead of becoming more evenly spread, the spectrum becomes even more concentrated around the leading eigenvector.

Second, the within-task Hessian drift on Tiny ImageNet and Food-101 is overall stronger than on CIFAR-100. Table E.1 reports the later-stage θ-block alignment statistics at Epoch 115 for Tiny ImageNet and Food101. In both datasets, the overlap and singular-value summaries decrease as the evaluation point moves further away from the anchor, indicating gradual drift of the dominant Hessian subspace over short windows. Since these two datasets are trained with smaller effective batch sizes than CIFAR-100 in main text, the stronger short-window drift is qualitatively consistent with a noisier regime, although batch size alone does not fully determine the amount of drift.

Third, we also examine the relative contribution of the Hessian blocks and their cross terms on these datasets. Empirically, both the cross block $H_{\theta C}$ and the prototype block H_{CC} contribute only a very small fraction of the total quadratic energy, and the normalised coupling coefficient $\hat{\beta}$ (Eq. 4.1) stays below 2% throughout. Since Tiny ImageNet and Food-101 use fewer prototypes than CIFAR-100 under the same backbone architecture, one possible intuition is that the prototype-related subspace is correspondingly smaller, which may reduce the relative contribution. However, that this should be read only as a plausible interpretation of the observed trend.

References

[1] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020.

[2] A. Bifet, G. Holmes, B. Pfahringer, P. Kranen, H. Kremer, T. Jansen, and T. Seidl. MOA: Massive online analysis, a framework for stream classification and clustering. In *Proceedings Workshop on Applications of Pattern Analysis*, 2010.

[3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[5] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, 2020.

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision*, 2018.

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[9] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2L: Contrastive continual learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, 2021.

[10] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision*, 2018.

[11] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2018.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020.

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[14] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

[15] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[16] Danruo Deng, Guangyong Chen, Jianye Hao, Qiong Wang, and Pheng-Ann Heng. Flattening sharpness for dynamic gradient projection memory benefits continual learning. In *Advances in Neural Information Processing Systems*, 2021.

[17] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, 2020.

[18] Enrico Fini, Victor G. Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[19] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[20] Jhair Gallardo, Tyler L. Hayes, and Christopher Kanan. Self-supervised training enhances online continual learning. In *British Machine Vision Conference*, 2021.

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu,

Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

[22] Guy Gur-Ari, Daniel A Roberts, and Ethan Vidick. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[23] 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*, 2016.

[24] 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*, 2020.

[25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[26] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

[28] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. 2017.

[29] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[30] Pratibha Kumari, Joohi Chauhan, Afshin Bozorgpour, Boqiang Huang, Reza Azad, and Dorit Merhof. Continual learning in medical image analysis: A comprehensive review of recent advancements and future prospects. *Medical Image Analysis*, 106: 103730, 2025.

[31] Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge, 2015.

[32] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Ch Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021.

[33] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[34] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 2017.

[35] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12):2346–2363, 2018.

[36] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations*, 2022.

[37] Arun Mallya and Svetlana Lazebnik. PackNet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018.

[38] Benjamin Maschler, Hannes Vietz, Nasser Jazdi, and Michael Weyrich. Continual learning of fault prediction for turbofan engines using deep learning with elastic weight consolidation. In *IEEE International Conference on Emerging Technologies and Factory Automation*, 2020.

[39] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research*, 24(214):1–50, 2023.

[40] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems*, 2020.

[41] Rupert Mitchell, Antonio Alliegro, Raffaello Camoriano, Dustin Carrión-Ojeda, Antonio Carta, Georgia Chalvatzaki, Nikhil Churamani, Carlo D'Eramo, Samin Hamidi, Robin Hesse, et al. Continual learning should move beyond incremental classification. *arXiv preprint arXiv:2502.11927*, 2025.

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[43] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. GDumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, 2020.

[44] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

[45] Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, 2018.

[46] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[47] Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

[48] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, 2018.

[49] Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

[50] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, 2017.

[51] Sebastian Thrun and Tom M Mitchell. Lifelong robot learning. *Robotics and autonomous systems*, 15(1-2):25–46, 1995.

[52] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, 2021.

[53] Gido M van de Ven. On the computation of the fisher information in continual learning. *arXiv preprint arXiv:2502.11756*, 2025.

[54] Gido M van de Ven and Andreas S Tolias. Three continual learning scenarios. In *NeurIPS Continual Learning Workshop*, 2018.

[55] Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

[56] Cédric Villani et al. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

[57] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

[58] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, 2020.

[59] Yichen Wen, Zhiquan Tan, Kaipeng Zheng, Chuanlong Xie, and Weiran Huang. Provable contrastive continual learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[60] Maciej Wołczyk, Michał Zajac, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

[61] Chen Xing, Devansh Arpit, Paul Christiano, and Yoshua Bengio. A walk with SGD. *arXiv preprint arXiv:1802.08770*, 2018.

[62] Shuo Yang, Kun-Peng Ning, Yu-Yang Liu, Jia-Yu Yao, Yong-Hong Tian, Yi-Bing Song, and Li Yuan. Is parameter collision hindering continual learning in LLMs? In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025.

[63] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018.

[64] Xiaofan Yu, Yunhui Guo, Sicun Gao, and Tajana Rosing. SCALE: Online self-supervised lifelong learning without prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[65] 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*, 2021.

[66] Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. In *Advances in Neural Information Processing Systems*, 2022.