# Provable Generalization of Dataset Condensation for Classification via Signal–Noise Dynamics

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

## Abstract

Dataset condensation, particularly via gradient matching, distills massive datasets into compact synthetic sets, making it an important tool for training under severe storage or computation constraints. However, despite strong empirical performance on classification tasks, existing theory largely relies on regression surrogates or static analyses and gives limited explanation of the underlying classification dynamics. We study gradient-matching condensation for regularized hinge-loss SVMs under an additive sub-Gaussian classification model. Our analysis shows that the learned condensed samples act as signal-concentrating representatives: they aggregate class-level structure while suppressing finite-sample noise and initialization residuals. This mechanism leads to population generalization guarantees for geometry-based evaluators and yields an explicit advantage over random one-shot coresets. The dynamics also identify an early-stopping tradeoff: informative structure is encoded early, whereas overly long inner loops can weaken certified signal accumulation. We further give a multiclass one-condensed-sample-per-class extension through a classwise one-vs-rest update and nearest-prototype evaluation, and simulations on synthetic data and KMNIST corroborate the predicted geometry, schedule sensitivity, and multiclass behavior.

## 1 Introduction

Machine learning is entering a new data-driven era as datasets grow at an unprecedented pace. In deep learning, large-scale datasets such as ImageNet Deng et al. (2009) and multimodal datasets like LAION-5B Schuhmann et al. (2022) and Infinity-MM Gu et al. (2024) have become indispensable for driving technological progress, but storing and training on such data incurs substantial computational and infrastructure costs that hinder deployment in resource-constrained settings.

Dataset distillation and condensation aim to mitigate this bottleneck by synthesizing compact datasets whose training behavior approximates that of the full data Wang et al. (2018); Sucholutsky & Schonlau (2021); Bohdal et al. (2020); Zhao et al. (2021); Zhao & Bilen (2021); Lee et al. (2022); Jiang et al. (2023b); Kim et al. (2022). In particular, gradient-matching condensation Zhao et al. (2021), which aligns the optimization trajectories of synthetic and real data, has achieved remarkable empirical success even under extreme compression ratios. However, a significant disparity remains between this empirical progress and our theoretical understanding of its underlying mechanisms.

Existing theoretical frameworks predominantly rely on regression surrogates (e.g., Kernel Ridge Regression) or infinite-width limits Izzo & Zou (2023); Maalouf et al. (2023); Chen et al. (2024). These analyses often overlook the distinct complexities of classification dynamics, which are inherently non-smooth and state-dependent. Consequently, mechanism-level guarantees for high-dimensional classification—especially for margin-based learners that rely on geometric decision boundaries rather than algebraic fitting—remain scarce. This leaves a fundamental question unanswered: *How do condensed datasets consistently retain the generalization ability required for classification tasks despite the loss of sample diversity?*

We address this gap by conducting a rigorous dynamical analysis of a practical bilevel gradient-matching procedure for regularized hinge-loss SVMs. Departing from static analysis, we explicitly track the optimization trajectory under a standard additive sub-Gaussian noise model. Our primary contribution is a structural

characterization of the terminal condensed samples in the one-sample-per-class regime. We establish that a terminal condensed sample is not merely a selected training point, but a synthesized representation: it emerges as a nonnegative structured combination of signed training samples with deterministic coefficient bounds. This derivation yields a transparent signal–noise decomposition, quantifying how the algorithm actively aggregates class-wise signal while averaging out noise, and elucidates the critical role of the inner-loop schedule in certifying this signal accumulation.

The choice of hinge-loss SVMs is deliberate. Dataset distillation and condensation are often evaluated through a two-stage protocol: synthetic data are first optimized under a chosen condensation model or matching objective, and the learned set is then reused for downstream training or evaluation (Wang et al., 2018; Zhao et al., 2021; Zhao & Bilen, 2021; 2023). Our setting follows this separation: the hinge-loss SVM supplies tractable margin-based gradient signals during condensation, while the terminal condensed set is evaluated through geometry-based downstream rules. Although the SVM decision rule is linear, its gradient-matching dynamics are state-dependent through hinge activity switching, capturing a canonical non-smooth feature of margin-based classification while remaining analytically tractable.

To convert the structure of a minimal condensed set into test performance, we adopt a directional evaluation viewpoint. In this setting, a broad family of symmetric evaluators depends primarily on the direction induced by the two condensed samples, allowing population risk to be controlled by how strongly that direction captures the underlying signal relative to residual components. Instantiating this geometric control for both a random one-shot coreset and the learned condensed set yields a provable generalization guarantee for the terminal output and reveals an explicit class-averaging advantage of condensation over random selection; moreover, the same analysis identifies a dynamical tradeoff whereby overly long inner loops can weaken conservative certified bounds, motivating early stopping. We also give a direct $K$-class extension for the one-condensed-sample-per-class regime: a classwise OvR gradient-matching update preserves the coefficient-expansion mechanism class by class, and a nearest-prototype evaluator converts the resulting classwise signal aggregation into a pairwise multiclass risk bound.

Our contributions are summarized as follows.

1. We identify a signal-aggregation mechanism in gradient-matching condensation for hinge-loss SVMs. In the one-sample-per-class regime, the learned condensed samples concentrate class signal while averaging finite-sample noise, and our coefficient bounds quantify how this mechanism depends on the training schedule.
2. We turn condensed-set geometry into population generalization guarantees. For directional prototype evaluators, the analysis yields an explicit class-averaging improvement over random one-shot selection; the same evaluation principle is further extended to a $K$-class nearest-prototype rule with one condensed sample per class.
3. We explain schedule sensitivity through the dynamics and validate the resulting predictions. The simulations confirm signal alignment, early-stopping behavior, cross-model transfer, and the corresponding classwise aggregation effect in the multiclass OvR extension.

To contextualize our contributions, Table 1 compares our work with representative theoretical studies on distillation/condensation. Our core dynamical analysis is developed for gradient-matching condensation with hinge-loss SVMs in the binary setting, while the evaluation viewpoint applies more broadly to downstream rules governed by condensed-set geometry. Appendix B gives the corresponding $K$-class one-condensed-sample-per-class extension through a classwise OvR update and a nearest-prototype multiclass risk bound.

### 1.1 Related Works

#### 1.1.1 Dataset Condensation and Distillation

Pioneered by Wang et al. (2018), dataset distillation aims to compress a large dataset into a small synthetic one that preserves training utility. To improve scalability, Zhao et al. (2021) proposed *dataset condensation*, which optimizes synthetic samples by matching the gradients of models trained on real versus synthetic

Table 1: Comparison of representative theoretical studies on dataset distillation/condensation. For works that provide method-agnostic laws without committing to a specific learner, the model entry is marked as "–". (GLM: Generalized Linear Model, KRR: Kernel Ridge Regression, SVM: Support Vector Machine)

| **Work** | **Task Type** | **Model** | **Key Focus** | **Dynamics?** | **Noise Model** |
|---|---|---|---|---|---|
| Izzo et al. Izzo & Zou (2023) | Regression | GLM / KRR | Exact Recovery | ✗ | ✗ |
| Maalouf et al. Maalouf et al. (2023) | Regression | KRR | Size–Error Bound | ✗ | ✗ |
| Chen et al. Chen et al. (2024) | Regression | KRR | Recovery Conditions | ✗ | ✗ |
| Luo and Xu Luo & Xu (2025) | General | – | Utility Boundary | ✓ | ✗ |
| **Our Work** | **Classification** | **SVM** | **Signal Learning** | ✓ | ✓ |

data. This gradient-matching paradigm significantly accelerates optimization compared to back-propagating through unrolled trajectories and has been further refined via differentiable augmentations (Zhao & Bilen, 2021) and multi-level optimization strategies (Jiang et al., 2023a). Beyond single-step gradient alignment, recent works have explored alternative matching objectives. Cazenavette et al. (2022) introduced *trajectory matching*, which aligns the long-horizon training paths of student networks to capture temporal training dynamics. Conversely, to bypass the computational cost of bi-level optimization entirely, Zhao & Bilen (2023) proposed *distribution matching*, which directly aligns the feature distributions of synthetic and real data. While these methods offer different trade-offs in computational cost and performance, our work specifically focuses on establishing a theoretical foundation for the widely used gradient-matching framework.

Despite substantial empirical progress in dataset condensation, theoretical research on the topic remains comparatively limited and fragmented, especially for classification-oriented condensation. Existing analyses largely concentrate on regression and kernelized settings, providing recovery or approximation guarantees for generalized linear models and kernel ridge regression Izzo & Zou (2023); Maalouf et al. (2023); Chen et al. (2024). Recently, Aoyama et al. proposed a duality-gap-based kernel distillation framework that extends kernel inducing point methods beyond squared loss (e.g., to hinge and cross-entropy) and provides bounds on parameter deviation and the resulting prediction/test errors Aoyama et al. (2025). Luo and Xu developed a configuration–dynamics–error framework that yields a scaling law in distilled set size for a fixed configuration and a configuration-coverage law characterizing how the required distilled size grows with configuration diversity Luo & Xu (2025). Nevertheless, a mechanism-level understanding of when and why condensation preserves classification generalization under explicit noise models—particularly for margin-based learners—remains insufficiently characterized. Our work bridges this gap through a theoretical framework that jointly characterizes optimization dynamics and generalization guarantees in classification-oriented dataset condensation.

### 1.1.2 Benign Overfitting and High-Dimensional Classification

The phenomenon of benign overfitting has been extensively analyzed in high-dimensional linear classification. Seminal works have characterized the generalization error of maximum margin classifiers (SVMs) under Gaussian mixture models Wang & Thrampoulidis (2020); Chatterji & Long (2021) and extended these results to multiclass settings Wang et al. (2021). Subsequent research has broadened this scope to accommodate label noise Wen et al. (2023); Frei et al. (2022) and heavy-tailed distributions Okudo & Kobayashi (2024). While our work focuses on dataset condensation rather than overfitting per se, we adopt similar high-dimensional

analysis techniques and additive noise models to characterize the signal-to-noise dynamics of the learned condensed samples.

## 2 Preliminaries

A complete notation table is provided in Appendix A (Table 2). This section introduces the data generation model, the SVM objective, the gradient-matching condensation algorithm, the directional evaluation framework, the key assumptions, and the multiclass extension studied in Appendix B.

### 2.1 Data generation model

We consider an additive class-conditional model in $\mathbb{R}^d$. The label space is $\mathcal{Y} = \{-1, +1\}$ for binary classification and $\mathcal{Y} = [K] := \{1, \dots, K\}$ for $K$-class classification. A sample $(\mathbf{x}, y)$ is generated as

$$\mathbf{x} = \boldsymbol{\mu}_y + \boldsymbol{\xi}, \tag{1}$$

where $\boldsymbol{\mu}_y \in \mathbb{R}^d$ is the class-dependent signal and $\boldsymbol{\xi} \in \mathbb{R}^d$ is mean-zero noise independent of $y$.

The noise is sub-Gaussian with covariance $\boldsymbol{\Sigma}_\xi \succeq \mathbf{0}$. Concretely, we write $\boldsymbol{\xi} = \boldsymbol{\Sigma}_\xi^{1/2}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \in \mathbb{R}^d$ has independent, mean-zero, sub-Gaussian coordinates with unit variance. In particular, $\mathrm{Cov}(\boldsymbol{\xi}) = \boldsymbol{\Sigma}_\xi$ (see Definition C.1).

Let $\mathcal{D}$ denote the population law of $(\mathbf{x}, y)$ induced by equation 1. We observe an i.i.d. training set $\mathcal{T} = \{(\mathbf{x}_i^{\mathcal{T}}, y_i^{\mathcal{T}})\}_{i=1}^{n_{\mathcal{T}}}$, where each $(\mathbf{x}_i^{\mathcal{T}}, y_i^{\mathcal{T}})$ is independently sampled from $\mathcal{D}$. For $k \in \mathcal{Y}$, define the index set and sample size $\mathcal{I}_k(\mathcal{T}) = \{i \in [n_{\mathcal{T}}] : y_i^{\mathcal{T}} = k\}$ and $n_{\mathcal{T},k} = |\mathcal{I}_k(\mathcal{T})|$. In the binary case, we set $\boldsymbol{\mu}_{+1} = \boldsymbol{\mu}$ and $\boldsymbol{\mu}_{-1} = -\boldsymbol{\mu}$ for some $\boldsymbol{\mu} \neq \mathbf{0}$, so that $\mathbf{x} = y\,\boldsymbol{\mu} + \boldsymbol{\xi}$ is a special case of equation 1.

### 2.2 Support Vector Machines

**Binary SVM.** For $y \in \{-1, +1\}$ we study the regularized hinge-loss objective

$$L_{\mathcal{T}}(\mathbf{w}) := \frac{1}{n_{\mathcal{T}}} \sum_{i=1}^{n_{\mathcal{T}}} \left(1 - y_i^{\mathcal{T}} \mathbf{w}^\top \mathbf{x}_i^{\mathcal{T}}\right)_+ + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \tag{2}$$

where $\lambda > 0$ and $(a)_+ := \max\{a, 0\}$. We omit an explicit bias term for simplicity (it can be incorporated by augmenting $\mathbf{x}$ with a constant coordinate).

Define the margin $u_i^{\mathcal{T}}(\mathbf{w}) := y_i^{\mathcal{T}} \mathbf{w}^\top \mathbf{x}_i^{\mathcal{T}}$. At the hinge boundary we fix the deterministic subgradient selection

$$q_i^{\mathcal{T}}(\mathbf{w}) := \mathbb{1}\{u_i^{\mathcal{T}}(\mathbf{w}) < 1\}. \tag{3}$$

This convention agrees with the ordinary hinge derivative away from $u_i^{\mathcal{T}}(\mathbf{w}) = 1$ and assigns the boundary case to the inactive side. A valid subgradient of equation 2 is therefore

$$\nabla_{\mathbf{w}} L_{\mathcal{T}}(\mathbf{w}) \in \lambda\,\mathbf{w} - \frac{1}{n_{\mathcal{T}}} \sum_{i=1}^{n_{\mathcal{T}}} q_i^{\mathcal{T}}(\mathbf{w})\, y_i^{\mathcal{T}} \mathbf{x}_i^{\mathcal{T}}. \tag{4}$$

Thus the SVM gradient is governed by the current active set of margin-violating samples. This active-set dependence makes the dynamics state-dependent, while within each fixed active-set region the update is affine. We define $L_{\mathcal{S}}(\mathbf{w})$ and $\nabla_{\mathbf{w}} L_{\mathcal{S}}(\mathbf{w})$ analogously for a synthetic dataset $\mathcal{S}$. We also write the signed sample $\mathbf{z}_i^{\mathcal{T}} := y_i^{\mathcal{T}} \mathbf{x}_i^{\mathcal{T}}$, so that the data-dependent term in equation 4 becomes $\frac{1}{n_{\mathcal{T}}} \sum_i q_i^{\mathcal{T}}(\mathbf{w})\, \mathbf{z}_i^{\mathcal{T}}$.

### 2.3 Gradient-Matching Dataset Condensation Algorithm

We specify the dataset condensation (DC) algorithm analyzed in the rest of the paper. The procedure is a hinge-loss SVM specialization of gradient matching Zhao et al. (2021). It compresses the training set $\mathcal{T}$ into

a synthetic set $\mathcal{S}$ by making the data-dependent SVM subgradient computed on $\mathcal{S}$ mimic the corresponding subgradient computed on $\mathcal{T}$ along repeatedly initialized training trajectories.

Algorithm 1 implements the *condensation stage* of our pipeline. The SVM supplies a low-cost proxy gradient that shapes the synthetic samples; after condensation, the learned set $\mathcal{S}_{\mathrm{dc}}$ may be passed unchanged to any downstream evaluator whose decision rule depends on the condensed-set geometry described in Subsection 2.4.

The main analysis focuses on the binary one-sample-per-class regime $\mathcal{S} = \{(\mathbf{x}_+^{\mathcal{S}}, +1), (\mathbf{x}_-^{\mathcal{S}}, -1)\}$ with $n_{\mathcal{S}} = 2$. We write the algorithm in signed coordinates $\mathbf{z}_y^{\mathcal{S}} := y\,\mathbf{x}_y^{\mathcal{S}}$ for $y \in \{\pm 1\}$; under the symmetric binary model, both class signals then equal the common direction $\boldsymbol{\mu}$. For a class block $\mathcal{A}_y$ of a dataset $\mathcal{A}$, define the activity-weighted signed mean at the current model state by

$$\mathbf{g}_{\mathcal{A}_y}(\mathbf{w}) := \frac{1}{n_{\mathcal{A}_y}} \sum_{i \in \mathcal{I}_y(\mathcal{A})} q_i^{\mathcal{A}}(\mathbf{w})\, \mathbf{z}_i^{\mathcal{A}}, \tag{5}$$

with the convention $\mathbf{g}_{\mathcal{S}_y}(\mathbf{w}) = q_{\mathcal{S},y}(\mathbf{w})\mathbf{z}_y^{\mathcal{S}}$ for the singleton condensed class. The regularization term $\lambda\mathbf{w}$ is common to the real and synthetic SVM subgradients, so the classwise matching step only needs to align the data-dependent terms in equation 5. The gradient-matching discrepancy for class $y$ is therefore

$$D_y(\mathcal{S}; \mathbf{w}) := \left\| \mathbf{g}_{\mathcal{S}_y}(\mathbf{w}) - \mathbf{g}_{\mathcal{T}_y}(\mathbf{w}) \right\|_2^2. \tag{6}$$

The algorithm has two nested loops. At the beginning of every outer loop $t_{\mathrm{out}}$, the SVM weight vector is reinitialized from a small isotropic Gaussian distribution. The condensed samples are not reinitialized; they are inherited from the end of the previous outer loop. Within the inner loop, the algorithm alternates between a classwise synthetic-data update and an SVM weight update. Throughout the synthetic-data update, the hard hinge activity indicators are evaluated at the current iterate and their derivatives with respect to the synthetic samples are taken to be zero, which agrees with their almost-everywhere derivative and fixes the convention at hinge switching boundaries. The resulting classwise gradient step induced by equation 6 takes the explicit form

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}}+1)} = \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} + 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \left( \mathbf{g}_{\mathcal{T}_y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} - \mathbf{g}_{\mathcal{S}_y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \right). \tag{7}$$

Thus, when the condensed sample is hinge-active, it is pulled toward the activity-weighted signed mean of the real samples in the same class; when it is inactive, the corresponding classwise synthetic update vanishes. The SVM weights are then updated on the real training objective:

$$\mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}}+1)} = (1 - \eta_{\mathbf{w}}\lambda)\mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})} + \eta_{\mathbf{w}} \mathbf{g}_{\mathcal{T}}^{(t_{\mathrm{out}},t_{\mathrm{in}})}, \qquad \mathbf{g}_{\mathcal{T}}^{(t_{\mathrm{out}},t_{\mathrm{in}})} := \frac{1}{n_{\mathcal{T}}} \sum_{i=1}^{n_{\mathcal{T}}} q_i^{\mathcal{T},(t_{\mathrm{out}},t_{\mathrm{in}})} \mathbf{z}_i^{\mathcal{T}}. \tag{8}$$

This explicit form explains why the terminal condensed samples admit a tractable signal–noise expansion. During hinge-active periods, equation 7 is an affine contraction toward the classwise signed mean of the training data, with contraction factor $\rho_{\mathcal{S}} = 1 - 2\eta_{\mathcal{S}}$. Across outer loops, repeated inheritance accumulates classwise signal while averaging finite-sample noise. The main results below quantify this mechanism by unrolling equation 7 over the bilevel schedule and translating the resulting condensed-set geometry into population risk bounds.

### 2.4 Prototypical Evaluation Framework

Algorithm 1 describes the condensation stage: the SVM provides a tractable margin-based gradient signal for updating the synthetic data, while the learned set $\mathcal{S}_{\mathrm{dc}}$ is later evaluated through downstream rules determined by its geometry.

For a binary condensed set with one sample per class, the relevant geometric object is the direction connecting the two condensed samples. The following definition records the class of evaluators whose population risk is governed by this direction.

**Algorithm 1** Gradient-Matching DC for Hinge-Loss SVM (One Sample per Class)

**Input:** Training set $\mathcal{T}$, loop lengths $T_{\text{out}}$ and $T_{\text{in}}$, step sizes $\eta_{\mathbf{w}}$ and $\eta_{\mathcal{S}}$
**Output:** Condensed set $\mathcal{S}_{\text{dc}}$
Initialize signed condensed samples $\mathbf{z}_y^{\mathcal{S},(0,0)} \sim \mathcal{N}(0, \sigma_{\mathcal{S}}^2 \mathbf{I}_d)$ for $y \in \{\pm 1\}$
**for** $t_{\text{out}} = 0$ **to** $T_{\text{out}} - 1$ **do**
  Initialize $\mathbf{w}^{(t_{\text{out}},0)} \sim \mathcal{N}(0, \sigma_{\mathbf{w}}^2 \mathbf{I}_d)$
  **if** $t_{\text{out}} > 0$ **then**
    Set $\mathbf{z}_y^{\mathcal{S},(t_{\text{out}},0)} \leftarrow \mathbf{z}_y^{\mathcal{S},(t_{\text{out}}-1,T_{\text{in}})}$ for each $y \in \{\pm 1\}$
  **end if**
  **for** $t_{\text{in}} = 0$ **to** $T_{\text{in}} - 1$ **do**
    Compute the hinge indicators $q_i^{\mathcal{T},(t_{\text{out}},t_{\text{in}})}$ and $q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}$
    Compute $\mathbf{g}_{\mathcal{T}}^{(t_{\text{out}},t_{\text{in}})}$, $\mathbf{g}_{\mathcal{T}_y}^{(t_{\text{out}},t_{\text{in}})}$, and $\mathbf{g}_{\mathcal{S}_y}^{(t_{\text{out}},t_{\text{in}})}$ from equation 5
    **for** $y \in \{\pm 1\}$ **do**
      Update $\mathbf{z}_y^{\mathcal{S}}$ by equation 7
    **end for**
    Update $\mathbf{w}$ by equation 8
  **end for**
**end for**
**Return:** $\mathcal{S}_{\text{dc}} = \{(y\mathbf{z}_y^{\mathcal{S},(T_{\text{out}}-1,T_{\text{in}})}, y) : y \in \{\pm 1\}\}$

**Definition 2.1** (Directional prototype evaluator)**.** Let $\mathcal{S} = \{(\mathbf{x}_+, +1), (\mathbf{x}_-, -1)\}$ be a binary condensed dataset with one sample per class. Define the *two-sample difference vector* $\mathbf{s}(\mathcal{S}) := \mathbf{x}_+ - \mathbf{x}_-$. A learning algorithm is a *directional prototype evaluator* if:

1. **Directional Alignment:** The learned weight vector aligns with the two-sample difference, i.e., $\hat{\mathbf{w}}_{\mathcal{S}} := \mathbf{w}_{\mathcal{S}}/\|\mathbf{w}_{\mathcal{S}}\|_2 = \mathbf{s}(\mathcal{S})/\|\mathbf{s}(\mathcal{S})\|_2$.
2. **Alignment-Based Risk:** The generalization risk on distribution $\mathcal{D}$ depends solely on this direction: $\mathcal{R}(\mathcal{S}) := \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}(y\langle \hat{\mathbf{w}}_{\mathcal{S}}, \mathbf{x}\rangle < 0)$.

In the one-sample-per-class setting, several standard evaluation rules reduce to a boundary whose normal direction is $\mathbf{s}(\mathcal{S})$. The following examples illustrate common cases:

- **SVM and logistic regression.** With one sample per class, symmetric regularized linear objectives include cases whose fitted direction is parallel to $\mathbf{x}_+ - \mathbf{x}_-$; in particular, a linear SVM has maximum-margin normal direction $\mathbf{x}_+ - \mathbf{x}_-$ up to scaling.
- **Nearest centroid classifier.** The distance rule $\|\mathbf{x} - \mathbf{x}_+\|_2 < \|\mathbf{x} - \mathbf{x}_-\|_2$ is algebraically equivalent to the linear comparison $\langle \mathbf{x}, \mathbf{x}_+ - \mathbf{x}_-\rangle > C$, where $C = (\|\mathbf{x}_+\|_2^2 - \|\mathbf{x}_-\|_2^2)/2$.
- **Feature-space nearest-centroid rules.** In representation spaces where neural-collapse geometry is observed (Papyan et al., 2020), class features concentrate around their means and the classifier behaves analogously to a nearest-centroid rule in feature space.

Our risk bounds therefore apply to any downstream rule satisfying Definition 2.1; the examples above identify familiar settings where this directional condition is natural.

### 2.5 Assumptions

This subsection collects the assumptions used in the binary analysis. They specify the signal scale, the concentration regime, the stable step-size range, and the initialization scale.

**Assumption 2.2.** Fix a tail parameter $\delta \in (0, 1)$ and constant $\kappa > 0$. Let $C_\kappa > 0$ denote the concentration constant determined by the choice of $\kappa$. We condition our analysis on the intersection of the high-probability

events defined in Lemmas C.2–C.5 (Section C.1), ensuring that the total failure probability is at most $\delta$. We assume the following conditions hold:

**A1**. **Signal strength.** The signal-to-noise ratio satisfies

$$\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}}\sqrt{d}} \;\geq\; 2\,C_\kappa.$$

**A2**. **High dimension.** The input dimension $d$ is sufficiently large relative to the number of outer loops $T_{\mathrm{out}}$, the condensed-set size $n_{\mathcal{S}}$, the training-set size $n_{\mathcal{T}}$, and the failure probability $\delta$:

$$d \;\geq\; \kappa \log\!\left(\frac{24\,T_{\mathrm{out}}\,n_{\mathcal{S}}\,n_{\mathcal{T}}}{\delta}\right).$$

**A3**. **Small learning rate.** The step sizes are chosen to ensure stability, where $\eta_{\mathbf{w}}$ and $\eta_{\mathcal{S}}$ denote the inner-loop step sizes for updating $\mathbf{w}$ and $\mathcal{S}$, respectively:

$$0 < \eta_{\mathbf{w}}\lambda \leq 1, \quad 0 < 2\,\eta_{\mathcal{S}} \leq 1.$$

**A4**. **Small initialization.** Define $\rho_{\mathcal{S}} := 1 - 2\eta_{\mathcal{S}}$. The weight initialization satisfies

$$\sigma_{\mathbf{w}} < \frac{1}{(2C_\kappa + 1)\|\boldsymbol{\mu}\|_2\sqrt{d}}.$$

The synthetic initialization satisfies

$$\sigma_{\mathcal{S}} < \min\left\{\frac{1}{C_\kappa\,\sigma_{\mathbf{w}}\,d}, \frac{\lambda}{(2C_\kappa+1)\|\boldsymbol{\mu}\|_2\sqrt{d}}, \rho_{\mathcal{S}}^{T_{\mathrm{in}}}\sqrt{\frac{2\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}}{n_{\mathcal{T}}}}\right\}.$$

Assumptions A1 and A2 are the high-dimensional signal and concentration conditions used to separate the class signal from the sub-Gaussian noise Chatterji & Long (2021); Cao et al. (2022); Kou et al. (2023); Okudo & Kobayashi (2024). Assumption A3 keeps the weight and condensed-sample updates in the stable contraction regime. Assumption A4 imposes signal-scale small random initialization on both the model weights and the synthetic samples, so that the initial hinge margins lie in the active region and the initialization residual remains controlled by the noise scale.

### 2.6 Multiclass Classwise OvR Extension

The main results below focus on the binary one-sample-per-class setting, where signed coordinates reduce the two condensed samples to a common signal direction. Appendix B introduces the corresponding $K$-class setup. In that extension, the condensed object is a single multiclass set $\mathcal{P} = \{(\mathbf{p}_k, k) : k \in [K]\}$ with one condensed sample per class, rather than $K$ separate two-point condensed sets obtained from independent one-vs-rest reductions.

The multiclass condensation step uses $K$ binary OvR heads. For head $h$, class $h$ is positive and every other class is negative, with sign $\tau_{h,k} = 2\mathbb{1}\{h = k\} - 1$. The class-$k$ condensed sample is updated by matching, averaged over heads, its OvR signed gradient term against the same-class real-data OvR gradient term. This defines a classwise OvR gradient-matching objective while keeping a single shared multiclass condensed set.

The corresponding multiclass evaluator is the nearest-prototype rule induced by the condensed set,

$$h_{\mathcal{P}}(\mathbf{x}) := \arg\min_{k\in[K]} \|\mathbf{x} - \mathbf{p}_k\|_2^2 = \arg\max_{k\in[K]} \left\{\langle \mathbf{x}, \mathbf{p}_k\rangle - \frac{1}{2}\|\mathbf{p}_k\|_2^2\right\}.$$

Its population risk is $\mathcal{R}_{\mathrm{mc}}(\mathcal{P}) := \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}(h_{\mathcal{P}}(\mathbf{x}) \neq y)$. This evaluator is the multiclass counterpart of the two-sample geometric evaluator used in the binary analysis: it compares the test point with one representative per class and includes the class-dependent quadratic bias in the equivalent score form. The detailed guarantees for this classwise OvR construction are stated after the binary results and in Appendix B.

# 3 Main Results

This section states the main guarantees. The binary one-sample-per-class result is the technical core: the terminal condensed samples achieve a population risk bound controlled by a certified signal-to-residual ratio. We then give the two ingredients behind this theorem: a dynamical expansion showing that the terminal condensed samples are nonnegative combinations of training samples (Section 3.2), and a geometric reduction converting that expansion into signal-to-residual lower bounds (Section 3.3). The corresponding $K$-class statement is summarized at the end of the section and stated in full in Appendix B.

## 3.1 Provable generalization for gradient-matching condensation

The main theorem shows that the terminal output of the gradient-matching procedure retains enough signal to generalize under the population model. Let $\boldsymbol{t}^\star := (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$ and write $\mathbf{z}_y^{\mathcal{S}} := \mathbf{z}_y^{\mathcal{S},(\boldsymbol{t}^\star)}$ for the terminal signed condensed sample produced by Algorithm 1 for class $y \in \{\pm 1\}$. Since our dynamics are written in signed coordinates $\mathbf{z} = y\mathbf{x}$, the terminal condensed dataset in the original input space is

$$\mathcal{S}_{\mathrm{dc}} := \{(y\mathbf{z}_y^{\mathcal{S}}, y) : y \in \{\pm 1\}\} = \{(\mathbf{z}_+^{\mathcal{S}}, +1), (-\mathbf{z}_-^{\mathcal{S}}, -1)\}. \tag{9}$$

The bound applies to the directional evaluator induced by this two-point output. Its exponent is governed by the certified signal-to-residual ratio $\underline{\mathrm{SR}}_{\mathrm{dc}}$, derived in Lemma 3.7.

**Theorem 3.1** (Provable generalization for the terminal condensed samples)**.** *Assume the additive model in equation 1 and Assumption 2.2. Run Algorithm 1 with one signed condensed sample per class, and let $\mathcal{S}_{\mathrm{dc}}$ be its terminal output defined in equation 9. Then, with probability at least $1 - \delta$, the population risk of the induced directional evaluator satisfies*

$$\mathcal{R}(\mathcal{S}_{\mathrm{dc}}) \le \exp\left( -\frac{\|\boldsymbol{\mu}\|_2^2}{2\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}} \left( \left( \frac{\underline{\mathrm{SR}}_{\mathrm{dc}} - 1}{\underline{\mathrm{SR}}_{\mathrm{dc}} + 1} \right)_+ \right)^2 \right), \tag{10}$$

*where $(u)_+ := \max\{u, 0\}$ and $\underline{\mathrm{SR}}_{\mathrm{dc}}$ is the certified lower bound on the signal-to-residual ratio specified in equation 20.*

Under the signal-strength condition in Assumption A1, the certified ratio satisfies

$$\underline{\mathrm{SR}}_{\mathrm{dc}} \ge 2\sqrt{C_\kappa}\, L \sqrt{n_{\mathcal{T}}}, \tag{11}$$

where $L$ is the dimensionless factor defined in Lemma 3.7. It summarizes the coefficient mass accumulated during certified classwise-averaging windows, the effect of class balance, and the contribution of coefficient mass outside those windows. The same signal-strength condition yields a direct dimension-dependent form of the population-risk exponent. Since $\|\boldsymbol{\mu}\|_2^2/(2\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}) \ge 2C_\kappa^2 d$, equation 10 implies

$$\mathcal{R}(\mathcal{S}_{\mathrm{dc}}) \le \exp\left( -2C_\kappa^2 d \left( \left( \frac{\underline{\mathrm{SR}}_{\mathrm{dc}} - 1}{\underline{\mathrm{SR}}_{\mathrm{dc}} + 1} \right)_+ \right)^2 \right). \tag{12}$$

Thus the guarantee is nontrivial whenever $\underline{\mathrm{SR}}_{\mathrm{dc}} > 1$, and the exponent scales linearly with the ambient dimension when this ratio is bounded away from one. Section 3.3 explains the two components of the certified ratio: the prefactor $\sqrt{n_{\mathcal{T}}}$ is the class-averaging scale, while $L$ records how much of that scale remains after accounting for coefficient mass outside the certified averaging windows.

## 3.2 Dynamical structure: Signal–noise expansion

To prove Theorem 3.1, we first characterize the terminal condensed samples produced by the bilevel dynamics. We write $\boldsymbol{t}^\star := (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$ for the terminal time index. The lemma below connects the gradient-matching recursion to condensed-set geometry. Part (i) gives *linear-window* lengths, namely initial inner prefixes on which the training-set and condensed-set hinge indicators remain active under the high-probability controls

used in the analysis. On these prefixes the update is affine and can be unrolled. Parts (ii)–(iii) then express each terminal signed condensed sample as a nonnegative combination of initialization and training data, with coefficient bounds determined by $\rho_{\mathcal{S}}$ and by the joint window length $T_{y,\mathrm{lin}}(r) = \min\{T_{\mathcal{T},\mathrm{lin}}, T_{\mathcal{S},\mathrm{lin}}(r)\}$.

The condensed-set window $T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})$ is indexed by the outer round. Starting from the reset weights at $(t_{\mathrm{out}}, 0)$, it is the length of the initial inner prefix on which both condensed samples remain strictly inside the $\pm 1$ margin band, using the hinge indicators in equation 3. This mirrors the training-set window $T_{\mathcal{T},\mathrm{lin}}$. Their minimum is the interval on which a classwise condensed update is certified to be an affine averaging step, and this is the interval that controls the coefficient lower bounds.

**Lemma 3.2** (Terminal condensed-sample expansion)**.** *Assume the additive model in equation 1 and Assumption 2.2, and consider the one-sample-per-class output of Algorithm 1. Set $\rho_{\mathcal{S}} := 1 - 2\eta_{\mathcal{S}}$. Fix $y \in \{\pm 1\}$. Recall that we work with signed samples $\mathbf{z} := y\mathbf{x}$, and the synthetic initialization satisfies $\mathbf{z}_y^{\mathcal{S},(0,0)} \sim \mathcal{N}(0, \sigma_{\mathcal{S}}^2 \mathbf{I}_d)$.*

*(i) There exist $\big(T_{\mathcal{T},\mathrm{lin}}, \{T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}_{t_{\mathrm{out}}=0}^{T_{\mathrm{out}}-1}\big) \in \{0, \ldots, T_{\mathrm{in}}\} \times \{0, \ldots, T_{\mathrm{in}}\}^{T_{\mathrm{out}}}$ such that $q_{\mathcal{T},i}\big(\mathbf{w}^{(t_{\mathrm{out}},t)}\big) = 1$ for all $i \in \mathcal{I}(\mathcal{T})$ and all $t \le T_{\mathcal{T},\mathrm{lin}}$, and $q_{\mathcal{S},y}^{(t_{\mathrm{out}},t)} = 1$ for all $t \le T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})$.*

*(ii) There exist coefficients $\widetilde{u}_y^{(\boldsymbol{t}^\star)} \ge 0$ and $\{\widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)} \ge 0\}_{j \in \mathcal{I}(\mathcal{T}_y)}$ such that, with $\widetilde{c}_{y,\mu}^{(\boldsymbol{t}^\star)} := \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)}$ and $\widetilde{\boldsymbol{c}}_y^{(\boldsymbol{t}^\star)} := \big(\widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)}\big)_{j \in \mathcal{I}(\mathcal{T}_y)}$, the terminal signed condensed sample admits the expansion*

$$\mathbf{z}_y^{\mathcal{S},(\boldsymbol{t}^\star)} = \widetilde{u}_y^{(\boldsymbol{t}^\star)}\, \mathbf{z}_y^{\mathcal{S},(0,0)} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)}\, \mathbf{z}_j^{\mathcal{T}}. \tag{13}$$

*(iii) The initialization coefficient satisfies*

$$\rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}} \le \widetilde{u}_y^{(\boldsymbol{t}^\star)} \le \rho_{\mathcal{S}}^{\sum_{k=0}^{T_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k)}. \tag{14}$$

*Define $T_{y,\mathrm{lin}}(r) := \min\{T_{\mathcal{T},\mathrm{lin}}, T_{\mathcal{S},\mathrm{lin}}(r)\}$. There exist deterministic quantities $B_{\mathrm{out}} \ge 1$, $u_{\mathrm{out}} = (1 - \rho_{\mathcal{S}}^{T_{\mathrm{in}}}) B_{\mathrm{out}}$, and $\ell_y \in [0, u_{\mathrm{out}}]$, depending only on $\rho_{\mathcal{S}}$ and the certified window lengths, such that the following bounds hold. The explicit construction of these quantities is recorded in Lemma D.8 in the appendix. For every $j \in \mathcal{I}(\mathcal{T}_y)$,*

$$\frac{\ell_y}{n_{\mathcal{T}_y}} \le \widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)} \le \frac{u_{\mathrm{out}}}{n_{\mathcal{T}_y}}. \tag{15}$$

*Consequently,*

$$\ell_y \le \widetilde{c}_{y,\mu}^{(\boldsymbol{t}^\star)} \le u_{\mathrm{out}}, \qquad \frac{\ell_y}{\sqrt{n_{\mathcal{T}_y}}} \le \big\|\widetilde{\boldsymbol{c}}_y^{(\boldsymbol{t}^\star)}\big\|_2 \le \frac{u_{\mathrm{out}}}{\sqrt{n_{\mathcal{T}_y}}}. \tag{16}$$

*(iv) Moreover, let $G_{\max}$ denote the high-probability upper bound on pairwise signed-sample inner products, so that $\langle \mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_{i'}^{\mathcal{T}} \rangle \le G_{\max}$ for all $i, i' \in \mathcal{I}(\mathcal{T})$ on the event in Assumption 2.2. If $G_{\max} < \lambda$, then $q_{\mathcal{T},i}(\mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})}) = 1$ and $q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = 1$ for all indices, and the coefficients simplify to*

$$\widetilde{u}_y^{(\boldsymbol{t}^\star)} = \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}} \qquad \widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)} = \frac{1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}}{n_{\mathcal{T}_y}}. \tag{17}$$

*Remark* 3.3. Part (i) provides initial linear windows for the training and condensed dynamics. On these windows, the hinge activity indicators stay at one throughout the corresponding initial inner loop. Thus the subgradient pattern is fixed on the certified prefix, and the inner-loop map reduces to an affine contraction that can be unrolled explicitly.

*Remark* 3.4. Part (ii) supplies nonnegative coefficients $\widetilde{u}_y^{(\boldsymbol{t}^\star)}$ and $\{\widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)}\}$ for the initialization and the class-$y$ training samples. It specifies how the terminal signed condensed sample can be expressed as a nonnegative combination of the initialization and the signed training points. This is the basic "signal–noise" template: $\widetilde{u}_y^{(\boldsymbol{t}^\star)}$ weights the random initialization, while $\{\widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)}\}$ allocates mass across the real samples in class $y$.

*Remark* 3.5. Part (iii) turns the abstract expansion in equation 13 into concrete two-sided bounds on each coefficient. For the initialization weight, the bound $\rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}} \le \widetilde{u}_y^{(\boldsymbol{t}^\star)} \le \rho_{\mathcal{S}}^{\sum_k T_{\mathcal{S},\mathrm{lin}}(k)}$ says that the random synthetic draw is attenuated by the active portions of the dynamics: the more time per outer round the condensed sample spends in the hinge-active regime, the more strongly the initialization is contracted. The quantity $\ell_y$ is the amount of terminal training coefficient mass contributed by steps whose classwise updates are certified to be uniform averages, while $u_{\mathrm{out}}$ is the corresponding upper envelope for the total terminal training mass. The coordinatewise bound equation 15 separates this averaged contribution from the remaining coefficient mass not covered by that lower bound; this remainder is accounted for in Lemma 3.7 through $\Delta_y = u_{\mathrm{out}} - \ell_y$.

*Remark* 3.6. Part (iv) gives a closed-form specialization in the strong-regularization regime $G_{\max} < \lambda$: full activity persists throughout training and condensation, the initialization weight decays exponentially, and the training coefficients become uniform within each class, making explicit the averaging mechanism on the classwise noise.

### 3.3 Geometric analysis: From structure to risk

We next translate the structural controls from Lemma 3.2 into the population classification guarantee. The next lemma gives the geometric risk reduction for directional evaluators and lower bounds the signal-to-residual ratio $\widehat{\mathrm{SR}}(\mathcal{S})$ for both a random one-shot coreset and the terminal condensed set. The latter lower bound is the certified constant $\underline{\mathrm{SR}}_{\mathrm{dc}}$ used in Theorem 3.1.

**Lemma 3.7** (Directional risk bound and signal-to-residual lower bounds)**.** *Assume the test distribution follows the additive model $\mathbf{x} = y\boldsymbol{\mu} + \boldsymbol{\xi}$ with $y \in \{\pm 1\}$ and mean-zero sub-Gaussian noise $\boldsymbol{\xi}$ with proxy covariance $\boldsymbol{\Sigma}_\xi$, and let $(u)_+ := \max\{u, 0\}$. For any two-point set $\mathcal{S} = \{(\mathbf{x}_+, +1), (\mathbf{x}_-, -1)\}$, define $\widehat{\mathbf{w}}_{\mathcal{S}} := (\mathbf{x}_+ - \mathbf{x}_-)/\|\mathbf{x}_+ - \mathbf{x}_-\|_2$ and $\mathcal{R}(\mathcal{S}) := \mathbb{P}\big(y\langle \widehat{\mathbf{w}}_{\mathcal{S}}, \mathbf{x}\rangle < 0\big)$. Suppose $\mathbf{x}_+ - \mathbf{x}_- = A(\mathcal{S})\boldsymbol{\mu} + \mathbf{r}(\mathcal{S})$ for some $A(\mathcal{S}) > 0$, and define $\widehat{\mathrm{SR}}(\mathcal{S}) := A(\mathcal{S})\|\boldsymbol{\mu}\|_2/\|\mathbf{r}(\mathcal{S})\|_2$ (with $\widehat{\mathrm{SR}}(\mathcal{S}) = +\infty$ if $\mathbf{r}(\mathcal{S}) = \mathbf{0}$). Then the following hold.*

*(i) The population risk satisfies*

$$\mathcal{R}(\mathcal{S}) \le \exp\left( -\frac{\|\boldsymbol{\mu}\|_2^2}{2\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}} \left( \left( \frac{\widehat{\mathrm{SR}}(\mathcal{S}) - 1}{\widehat{\mathrm{SR}}(\mathcal{S}) + 1} \right)_+ \right)^2 \right). \tag{18}$$

*(ii) Assume Assumption 2.2 and the training model $\mathbf{x}_i^{\mathcal{T}} = y_i^{\mathcal{T}}\boldsymbol{\mu} + \boldsymbol{\xi}_i$. Pick any $i_+ \in \mathcal{I}(\mathcal{T}_+)$ and $i_- \in \mathcal{I}(\mathcal{T}_-)$ and set $\mathcal{S}_{\mathrm{core}} = \{(\mathbf{x}_{i_+}^{\mathcal{T}}, +1), (\mathbf{x}_{i_-}^{\mathcal{T}}, -1)\}$. On the high-probability event in Assumption 2.2, letting $C_\kappa$ denote the corresponding constant, we have*

$$\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{core}}) \;\ge\; \frac{2\|\boldsymbol{\mu}\|_2}{\sqrt{2(1 + 2C_\kappa)\,\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\, d}}. \tag{19}$$

*(iii) Assume the hypotheses of Lemma 3.2 and consider the terminal condensed set $\mathcal{S}_{\mathrm{dc}}$ in equation 9. Let $B_{\mathrm{out}}$, $u_{\mathrm{out}}$, and $\ell_y$, for $y \in \{\pm 1\}$, be the deterministic coefficient quantities from Lemma 3.2. Set $p_y := n_{\mathcal{T}_y}/n_{\mathcal{T}}$, $\beta_p := p_+^{-1/2} + p_-^{-1/2}$, and $\Delta_y := u_{\mathrm{out}} - \ell_y$. Define the dimensionless factor*

$$L := \frac{\ell_+ + \ell_-}{B_{\mathrm{out}}\beta_p + \sum_{y\in\{\pm1\}} \Delta_y\big(\sqrt{n_{\mathcal{T}}} - p_y^{-1/2}\big)}.$$

*Then, on the same high-probability event,*

$$\widehat{\mathrm{SR}}\left(\mathcal{S}_{\mathrm{dc}}\right) \ge \underline{\mathrm{SR}}_{\mathrm{dc}} \;=\; \frac{\sqrt{n_{\mathcal{T}}}\, L\|\boldsymbol{\mu}\|_2}{\sqrt{C_\kappa\,\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\, d}}. \tag{20}$$

*Remark* 3.8. The risk upper bound equation 18 is governed by the signal-to-residual ratio $\widehat{\mathrm{SR}}(\mathcal{S})$. Larger values yield a more negative exponent and a tighter bound, while $\widehat{\mathrm{SR}}(\mathcal{S}) \le 1$ makes the exponent term equal to zero and gives only the vacuous bound $\mathcal{R}(\mathcal{S}) \le 1$.

*Remark* 3.9. Comparing equation 19 and equation 20, the condensed lower bound displays the class-averaging scale through the explicit prefactor $\sqrt{n_{\mathcal{T}}}$. The dimensionless factor $L$ determines how much of that scale is certified by the dynamics. Its numerator $\ell_+ + \ell_-$ is the coefficient mass that enters through certified classwise averaging, while the denominator combines the class-balance and averaged-residual term $B_{\mathrm{out}}\beta_p$ with the cost of the non-averaged remainder $\sum_y \Delta_y(\sqrt{n_{\mathcal{T}}} - p_y^{-1/2})$.

*Remark* 3.10. The factor $L$ determines whether the explicit prefactor $\sqrt{n_{\mathcal{T}}}$ in equation 20 is retained after accounting for activity switching. Lemma D.12 in the appendix quantifies the loss incurred when the certified active prefix is shorter than the full inner loop. With $T := T_{\mathrm{in}}$, $\rho := \rho_{\mathcal{S}}$, and

$$M_{\mathcal{S}} := \max_{0 \le r \le T_{\mathrm{out}}-1} \sum_{k=r+1}^{T_{\mathrm{out}}-1} \big(T - T_{\mathcal{S},\mathrm{lin}}(k)\big),$$

this deficit is

$$\mathcal{W}_{\mathrm{win}} := (1-\rho^T)(1-\rho^{M_{\mathcal{S}}}) + \max_{\substack{y \in \{\pm 1\} \\ 0 \le r \le T_{\mathrm{out}}-1}} \big(1 - \rho^{T - T_{y,\mathrm{lin}}(r)}\big).$$

The first term measures the effect of later condensed active prefixes that are shorter than $T$: if a contribution is injected in an early outer round, then subsequent outer rounds with $T_{\mathcal{S},\mathrm{lin}}(k) < T$ leave less certified contraction on that contribution. The quantity $M_{\mathcal{S}}$ records the largest cumulative future gap $\sum_k (T - T_{\mathcal{S},\mathrm{lin}}(k))$. The second term measures the gap within the current outer round: classwise averaging is certified only up to $T_{y,\mathrm{lin}}(r)$, so the remaining $T - T_{y,\mathrm{lin}}(r)$ inner steps are not treated as averaged updates. Hence $\mathcal{W}_{\mathrm{win}} = 0$ when all relevant active prefixes have full length $T$, while larger values indicate that more terminal coefficient mass may lie outside the certified classwise-averaging part. Lemma D.12 proves the deterministic comparisons between $\mathcal{W}_{\mathrm{win}}$, the remainders $\Delta_y$, and the factor $L$, yielding the following regimes.

(i) In the full-activity regime $G_{\max} < \lambda$, all windows have full length and $\mathcal{W}_{\mathrm{win}} = 0$. The terminal training coefficients are exactly uniform within each class, so the non-averaged remainder vanishes. In this case

$$L = \frac{2(1-\rho^T)}{\beta_p},$$

so the bound retains the full $\sqrt{n_{\mathcal{T}}}$ class-averaging scale up to the class-balance factor.

(ii) In the small-remainder regime, if $\mathcal{W}_{\mathrm{win}} = o(1-\rho^T)$ and $\mathcal{W}_{\mathrm{win}} = O(\beta_p/\sqrt{n_{\mathcal{T}}})$, then the coefficient mass outside the certified averaging part is at most comparable to the empirical averaging scale. Lemma D.12 gives $L = \Omega((1-\rho^T)/\beta_p)$, and if the second condition is strengthened to $\mathcal{W}_{\mathrm{win}} = o(\beta_p/\sqrt{n_{\mathcal{T}}})$, then $L = (1+o(1))2(1-\rho^T)/\beta_p$.

(iii) In the switching-dominated regime, the total remainder $\sum_y \Delta_y$ is large enough that the non-averaged part of the residual dominates the averaged-residual term in the denominator of $L$. For example, if $n_{\min} \to \infty$ and $\sum_y \Delta_y = \Omega(B_{\mathrm{out}}(1-\rho^T))$, Lemma D.12 gives $L = O(n_{\mathcal{T}}^{-1/2})$. The proof still yields a valid lower bound, but it no longer certifies a growing $\sqrt{n_{\mathcal{T}}}$ class-averaging gain.

Theorem 3.1 follows by combining the certified signal-to-residual lower bound for the terminal condensed set with the directional-evaluator risk reduction.

*Proof of Theorem 3.1.* Lemma 3.7(iii) gives $\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{dc}}) \ge \underline{\mathrm{SR}}_{\mathrm{dc}}$, and Lemma 3.7(i) bounds $\mathcal{R}(\mathcal{S}_{\mathrm{dc}})$ by the right-hand side of equation 18. Since $t \mapsto \left(\left(\frac{t-1}{t+1}\right)_+\right)^2$ is nondecreasing on $[0, +\infty)$, substituting $\underline{\mathrm{SR}}_{\mathrm{dc}}$ yields equation 10. □

### 3.4 Multiclass consequences

The same coefficient-expansion mechanism extends to the $K$-class classwise OvR setting introduced in Subsection 2.6. Let $\mathcal{P}_{\mathrm{dc}} = \{(\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)}, k) : k \in [K]\}$ be the terminal multiclass condensed set. Theorem B.3 shows that, for every class $k$, the terminal condensed sample admits a nonnegative same-class expansion $\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} = u_k^{(\boldsymbol{t}^\star)}\mathbf{p}_k^{\mathcal{S},(0,0)} + \sum_{i\in\mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)}\mathbf{x}_i^{\mathcal{T}}$, with coefficient bounds controlled by the classwise OvR active windows. In the full-activity specialization, this expansion reduces to the contraction $\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} = \rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}\mathbf{p}_k^{\mathcal{S},(0,0)} + (1-\rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}})\widehat{\boldsymbol{\mu}}_k^{\mathcal{T}}$ toward the empirical class mean.

For the nearest-prototype evaluator introduced in Subsection 2.6, Theorem B.4 gives a pairwise multiclass risk bound controlled by the class separations $\gamma_{k\ell} = \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell\|_2$ and the perturbations of the condensed samples from their class means. Corollary B.5 combines this geometric bound with the OvR coefficient expansion to obtain a certified $K$-class risk guarantee for $\mathcal{P}_{\mathrm{dc}}$. Thus the multiclass result preserves the binary proof strategy at the level of classwise affine recursions, while replacing the binary directional evaluator by a nearest-prototype pairwise comparison.

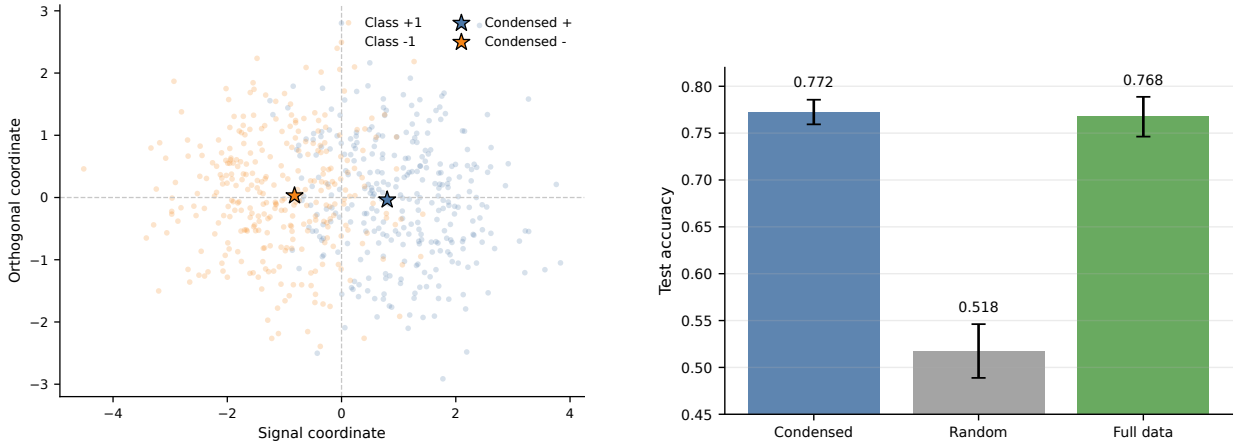
## 4 Simulation

This section empirically examines the mechanisms predicted by the theory. The binary experiments are the primary tests of the single-sample-per-class setting $\mathcal{S} = \{(\mathbf{x}_+, +1), (\mathbf{x}_-, -1)\}$: they measure signal aggregation, the effect of the inner-loop length, and the transferability of the learned condensed-set geometry across downstream evaluators. We further include a focused $K$-class OvR experiment to evaluate whether the same classwise aggregation effect persists in the one-condensed-sample-per-class extension.

**Setup.** We implement the bilevel gradient matching routine in Algorithm 1 with the deterministic hinge subgradient selection $q(\mathbf{w}) = \mathbb{1}\{u(\mathbf{w}) < 1\}$ and the classwise update rule specialized to $n_{\mathcal{S}} = 2$, and we denote by $\mathcal{S}^\star$ the terminal condensed set after the prescribed $(T_{\mathrm{out}}, T_{\mathrm{in}})$ bilevel iterations. In the synthetic experiments, we sample a balanced training set $\mathcal{T}$ from the additive model $\mathbf{x} = y\boldsymbol{\mu} + \boldsymbol{\xi}$ and report the geometric alignment $\cos\big(\mathbf{s}(\mathcal{S}^\star), \boldsymbol{\mu}\big)$, where $\mathbf{s}(\mathcal{S}) := \mathbf{x}_+ - \mathbf{x}_-$, together with the downstream test accuracy of a linear SVM trained on $\mathcal{S}^\star$ and evaluated on a held-out test set. On KMNIST, we use a two-class subset and evaluate the same condensed set under multiple downstream models, following the evaluation viewpoint in Subsection 2.4. All reported quantities are averaged over $n_{\mathrm{seeds}} = 5$ runs and shown as mean $\pm$ one standard deviation. Full hyperparameters, reproducibility details, and additional numerical tables are reported in Appendix E. The main baseline is a random one-shot coreset that selects one real training sample per class and uses the same downstream training pipeline as $\mathcal{S}^\star$.

**Condensed-sample aggregation and effectiveness on synthetic data.** To visualize aggregation, Figure 1a projects samples onto the signal coordinate $\langle \mathbf{x}, \boldsymbol{\mu}/\|\boldsymbol{\mu}\|_2\rangle$ and an orthogonal coordinate $\langle \mathbf{x}, \mathbf{v}\rangle$ for a fixed unit vector $\mathbf{v} \perp \boldsymbol{\mu}$, and overlays the two learned condensed samples $\mathbf{x}_+$ and $\mathbf{x}_-$; they concentrate near the signal axis with small orthogonal components and separate along the signal coordinate, consistent with our terminal expansion and coefficient analysis. For downstream performance, Figure 1b compares the test accuracy of a linear SVM trained on the condensed set against two baselines: a random one-shot coreset and training on the full dataset. The condensed set reaches $0.772 \pm 0.013$ accuracy, close to full-data training ($0.768 \pm 0.021$) and well above the random one-shot baseline ($0.518 \pm 0.029$). This pattern matches the theoretical picture that gradient matching turns the two condensed samples into signal-aligned aggregations of the training set rather than noisy representatives of individual examples.

**Early stopping via inner-loop length.** Our theory isolates activity windows and coefficient bounds that depend on the interplay between the training-set and condensed-set linear windows, suggesting that overly long inner loops may enter regimes where informative gradient directions weaken. We assess this schedule effect by sweeping $(T_{\mathrm{in}}, T_{\mathrm{out}})$ on synthetic data and reporting the alignment heatmap in Figure 2a; the corresponding test-accuracy heatmap appears in Appendix E. The heatmap shows that alignment is not governed only by the total number of bilevel updates: increasing $T_{\mathrm{out}}$ is beneficial when the inner loop is




(a) Signal/orthogonal projection.

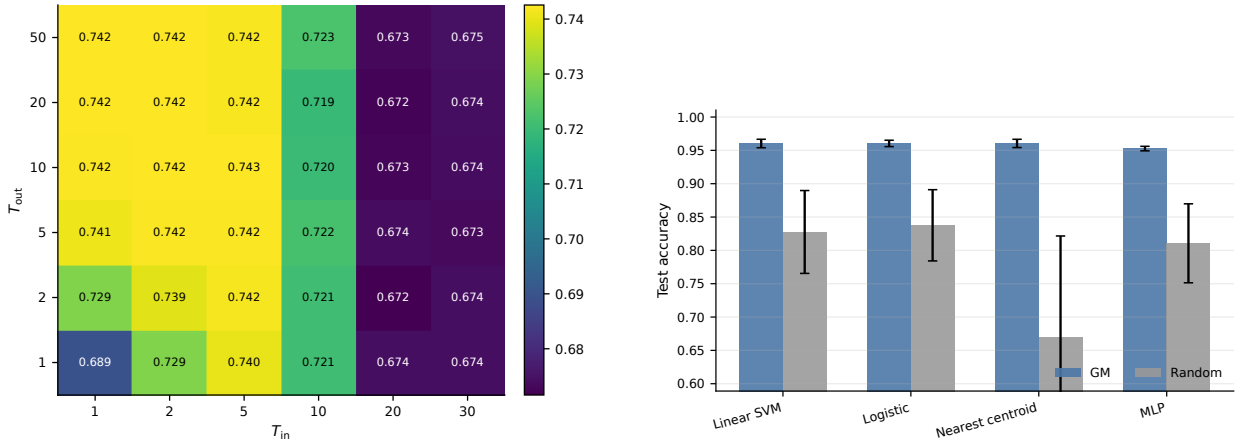
(b) Synthetic-data test accuracy.

Figure 1: Condensed-sample aggregation and downstream effectiveness on synthetic data. Left: training samples and learned condensed samples in a signal/orthogonal projection. Right: condensed vs. random one-shot vs. full data, reported as mean ± one standard deviation over seeds.

short, but the gain saturates and then weakens once $T_{\text{in}}$ becomes large. The strongest alignments occur for relatively short inner loops combined with moderate outer-loop repetition, whereas large $T_{\text{in}}$ lowers the learned signal alignment and is accompanied by a modest decrease in test accuracy. This pattern is consistent with the coefficient bounds in Lemma 3.7: after the certified active prefix, additional inner steps contribute less uniform classwise averaging and can discount the previously accumulated signal mass through the contraction factors. Appendix E also reports stepwise diagnostics; in the default synthetic setting the condensed samples remain active, while the training-set activity and gradient norms decrease along longer inner loops, indicating that the observed schedule effect is driven primarily by weakening training-side gradients rather than by a loss of condensed-sample activity.

**Cross-model transfer on KMNIST.** The KMNIST experiment examines whether the geometry learned by SVM-based condensation remains useful beyond the proxy learner used during condensation. We learn the condensed set once and then evaluate it, without re-condensation, under four downstream models: a linear SVM, logistic regression, a nearest-centroid classifier induced by $\mathbf{s}(\mathcal{S}^\star)$, and a small two-layer ReLU MLP. Figure 2b shows that the same two SVM-condensed samples outperform a random one-shot coreset across all evaluators. The improvement appears not only for the linear SVM used as the proxy learner, but also for logistic regression, nearest-centroid classification, and the two-layer MLP, indicating that the learned samples encode a transferable class-separating geometry rather than only the details of the hinge-loss update. The nearest-centroid performance further confirms that this geometry is already present at the prototype level. This supports the condensed-set geometry perspective: once $\mathcal{S}^\star$ captures a stable class-separating direction, evaluators that depend on that geometry can benefit even when they are not identical to the proxy model used during condensation. Additional numerical tables are reported in Appendix E.

**Multiclass OvR evaluation.** Finally, we run the classwise OvR condensation rule from Appendix B on a balanced $K = 5$ synthetic additive model with one condensed sample per class, and we evaluate the resulting multiclass set using the nearest-prototype rule in Subsection 2.6. The learned OvR condensed set obtains $0.624 \pm 0.008$ test accuracy, matching the full-data nearest-prototype reference based on empirical class means within Monte Carlo variability ($0.624 \pm 0.018$) and substantially exceeding a random one-shot multiclass coreset ($0.231 \pm 0.030$). This result provides empirical evidence for the classwise aggregation mechanism behind the multiclass expansion; the corresponding table and figure are given in Appendix E.5.




(a) Schedule-dependent alignment.

(b) KMNIST cross-model accuracy.

Figure 2: Additional empirical diagnostics. Left: alignment $\cos(\mathbf{s}(\mathcal{S}^\star), \boldsymbol{\mu})$ over the $(T_{\text{in}}, T_{\text{out}})$ grid on synthetic data, showing that short-to-moderate inner loops combined with repeated outer-loop restarts yield the strongest signal alignment. Right: KMNIST cross-model evaluation of GM and random one-shot coresets, reported as mean $\pm$ one standard deviation over seeds; the same condensed set transfers across evaluators beyond the hinge-loss SVM used during condensation.

## 5 Discussion

This work establishes a dynamical framework for dataset condensation in classification, moving beyond regression surrogates to analyze the non-smooth optimization of margin-based learners directly. By characterizing condensed samples as structured signal–noise combinations, we bridge the gap between the algorithmic mechanics of gradient matching and the geometric requirements of generalization.

While anchored in SVMs, this dynamical framework holds potential for characterizing optimization trajectories in broader margin-based classifiers, where gradient signals often change as examples enter or leave high-loss/high-margin regimes. Crucially, our analysis reveals a mechanism where excessive inner-loop updates can attenuate certified signal accumulation, suggesting that early stopping reflects a broader signal-accumulation versus activity-switching tradeoff rather than a peculiarity of the SVM objective.

Our guarantees are proved for the binary one-condensed-sample-per-class setting under an additive sub-Gaussian model, together with a classwise OvR extension for the $K$-class case. Extending the same level of guarantee to general softmax or Crammer–Singer multiclass training would require controlling cross-class competition, switching competitors, and coupled gradient-matching dynamics, which are absent from the affine classwise recursions analyzed here. Likewise, relaxing the activity-window and linear-window conditions would require a substantially finer analysis of when samples enter and leave the active hinge regime. These limitations identify concrete technical barriers for extending the analysis beyond affine classwise recursions, and clarify the role of the present tractable margin-based setting as a first step toward classification-specific condensation theory.

Practically, the results suggest that the inner-loop length should be monitored as a mechanism controlling the persistence of informative gradient activity, rather than treated as a purely computational hyperparameter. The analysis points to concrete diagnostics for schedule selection, including alignment, hinge activity, and validation performance.

Furthermore, our theoretical toolkit offers a blueprint for other condensation paradigms. For *distribution matching* (Zhao & Bilen, 2023), our structural decomposition and geometric risk bounds can be adapted to simplify analysis by bypassing non-smooth unrolling. Conversely, for *trajectory matching* (Cazenavette et al., 2022), which shares our dynamical nature, our coefficient expansion and linear-window analysis provide a

foundational approach to disentangle complex cross-time dependencies. Extending these tools to capture such long-horizon structures remains a promising avenue for future research.

## References


Tatsuya Aoyama, Hanting Yang, Hiroyuki Hanada, Satoshi Akahane, Tomonari Tanaka, Yoshito Okura, Yu Inatsu, Noriaki Hashimoto, Taro Murayama, Hanju Lee, et al. Generalized kernel inducing points by duality gap for dataset distillation. *arXiv preprint arXiv:2502.12607*, 2025.

Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Flexible dataset distillation: Learn labels instead of images. *arXiv preprint arXiv:2006.08572*, 2020.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25237–25250. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a12c999be280372b157294e72a4bbc8b-Paper-Conference.pdf.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4750–4759, 2022.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

Yilan Chen, Wei Huang, and Tsui-Wei Weng. Provable and efficient dataset distillation for kernel ridge regression. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 88739–88771. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a1c716638d9b618a1a40a96f473c8250-Paper-Conference.pdf.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022.

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*, 2024.

Zachary Izzo and James Zou. A theoretical study of dataset distillation. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL https://openreview.net/forum?id=dq5QGXGxoJ.

Zixuan Jiang, Jiaqi Gu, Mingjie Liu, and David Z. Pan. Delving into effective gradient matching for dataset condensation. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pp. 1–6, 2023a. doi: 10.1109/COINS57856.2023.10189244.

Zixuan Jiang, Jiaqi Gu, Mingjie Liu, and David Z Pan. Delving into effective gradient matching for dataset condensation. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pp. 1–6. IEEE, 2023b.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, pp. 11102–11118. PMLR, 2022.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer ReLU convolutional neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17615–17659. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kou23a.html.

Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *International Conference on Machine Learning*, pp. 12352–12364. PMLR, 2022.

Zhengquan Luo and Zhiqiang Xu. Utility boundary of dataset distillation: Scaling and configuration-coverage laws. *arXiv preprint arXiv:2512.05817*, 2025.

Alaa Maalouf, Murad Tukan, Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. On the size and approximation error of distilled datasets. *Advances in Neural Information Processing Systems*, 36: 61085–61102, 2023.

Kota Okudo and Kei Kobayashi. Benign overfitting under learning rate conditions for *alpha* sub-exponential input. *arXiv preprint arXiv:2409.00733*, 2024.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none), Jan 2013. doi: 10.1214/ecp.v18-2865.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4030–4034, 2020. doi: 10.1109/ICASSP39728.2021.9413946.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *IEEE Transactions on Information Theory*, 69:7909–7952, 2021. doi: 10.1109/tit.2023.3320098.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Kaiyue Wen, Jiaye Teng, and Jingzhao Zhang. Benign overfitting in classification: Provably counter label noise with larger models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UrEwJebCxk.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pp. 12674–12685. PMLR, 2021.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6514–6523, January 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=mSAKhLYLSsl.

# A Notation

Table 2: Summary of notation used throughout the paper.

| Symbol | Meaning |
|---|---|
| *Data and distributions* | |
| $d$ | Input dimension. |
| $\mathcal{Y}$ | Label space: $\{-1,+1\}$ in the binary case, or $[K]=\{1,\dots,K\}$ for $K$-class problems. |
| $\mathbf{x}$, $y$ | Feature vector and label. |
| $\boldsymbol{\mu}_y$ | Class-dependent mean (signal); in the binary symmetric case $\boldsymbol{\mu}_{+1}=\boldsymbol{\mu}$, $\boldsymbol{\mu}_{-1}=-\boldsymbol{\mu}$ and $\mathbf{x}=y\boldsymbol{\mu}+\boldsymbol{\xi}$. |
| $\boldsymbol{\xi}$, $\boldsymbol{\Sigma}_\xi$, $\boldsymbol{\zeta}$ | Mean-zero noise; covariance $\boldsymbol{\Sigma}_\xi$; whitened coordinates $\boldsymbol{\xi}=\boldsymbol{\Sigma}_\xi^{1/2}\boldsymbol{\zeta}$. |
| $\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}$ | Operator norm of the noise covariance (used in high-probability bounds). |
| $\mathcal{D}$ | Population law of $(\mathbf{x},y)$. |
| $\mathcal{T}$, $n_{\mathcal{T}}$ | Training set $\{(\mathbf{x}_i^{\mathcal{T}},y_i^{\mathcal{T}})\}_{i=1}^{n_{\mathcal{T}}}$ and its size. |
| $\mathcal{I}_k(\mathcal{T})$, $n_{\mathcal{T},k}$ | Indices with label $k$ and the corresponding class count. |
| $\mathbf{z}_i^{\mathcal{T}}:=y_i^{\mathcal{T}}\mathbf{x}_i^{\mathcal{T}}$ | Signed training sample. |
| *SVM objective and hinge structure* | |
| $\mathbf{w}$ | SVM weight vector. |
| $\lambda$ | $\ell_2$ regularization parameter in $L_{\mathcal{T}}(\mathbf{w})$ and $L_{\mathcal{S}}(\mathbf{w})$. |
| $(a)_+$ | Positive part $\max\{a,0\}$. |
| $L_{\mathcal{T}}(\mathbf{w})$, $L_{\mathcal{S}}(\mathbf{w})$ | Regularized hinge loss on $\mathcal{T}$ and on the synthetic set $\mathcal{S}$. |
| $u_i^{\mathcal{T}}(\mathbf{w})$ | Margin $y_i^{\mathcal{T}}\mathbf{w}^\top\mathbf{x}_i^{\mathcal{T}}$. |
| $q_i^{\mathcal{T}}(\mathbf{w})$ | Hinge activity indicator $\mathbb{1}\{u_i^{\mathcal{T}}(\mathbf{w})<1\}$ (subgradient convention). |
| *Dataset condensation (bi-level schedule)* | |
| $\mathcal{S}$, $n_{\mathcal{S}}$ | Condensed dataset; one-sample-per-class gives $n_{\mathcal{S}}=2$ in the binary case. |
| $T_{\mathrm{out}}$, $T_{\mathrm{in}}$ | Numbers of outer and inner iterations. |
| $t_{\mathrm{out}}$, $t_{\mathrm{in}}$ | Outer and inner loop indices. |
| $\eta_{\mathbf{w}}$, $\eta_{\mathcal{S}}$ | Inner-loop step sizes for updating $\mathbf{w}$ (on $\mathcal{T}$) and condensed samples. |
| $\mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})}$, $\mathbf{x}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})}$ | Weight and condensed sample after the indicated inner steps. |
| *Directional evaluation (Definition 2.1)* | |
| $\mathbf{s}(\mathcal{S})$ | Two-sample difference $\mathbf{x}_+-\mathbf{x}_-$ for $\mathcal{S}=\{(\mathbf{x}_+,+1),(\mathbf{x}_-,-1)\}$. |
| $\hat{\mathbf{w}}_{\mathcal{S}}$ | Normalized classifier direction aligned with $\mathbf{s}(\mathcal{S})$. |
| $\mathcal{R}(\mathcal{S})$ | Population misclassification risk $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}(y\langle\hat{\mathbf{w}}_{\mathcal{S}},\mathbf{x}\rangle<0)$. |
| *Multiclass classwise OvR extension (Appendix B)* | |
| $\mathcal{P}^{(\boldsymbol{t})}$ | Multiclass condensed set at time $\boldsymbol{t}$: $\{(\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})},k):k\in[K]\}$, with one condensed sample per class. |
| $\mathcal{P}_{\mathrm{dc}}$ | Terminal multiclass condensed set returned by the classwise OvR condensation procedure. |
| $\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})}$, $\mathbf{p}_k$ | Class-$k$ condensed sample at time $\boldsymbol{t}$; generic class-$k$ representative in a multiclass nearest-prototype evaluator. |
| $\mathbf{W}^{(\boldsymbol{t})}$, $\mathbf{w}_h^{(\boldsymbol{t})}$ | Collection of $K$ OvR heads and the head-$h$ weight vector at time $\boldsymbol{t}$. |
| $\boldsymbol{t}^+$ | Next inner-loop time $(t_{\mathrm{out}},t_{\mathrm{in}}+1)$, used when $t_{\mathrm{in}}<T_{\mathrm{in}}$. |
| $\tau_{h,k}$ | OvR sign $2\mathbb{1}\{h=k\}-1$, treating class $h$ as positive for head $h$. |
| $\mathbf{z}_{i,h}^{\mathcal{T}}$, $\bar{\mathbf{z}}_h^{\mathcal{T}}$ | Head-wise signed training sample $\tau_{h,y_i^{\mathcal{T}}}\mathbf{x}_i^{\mathcal{T}}$ and its empirical mean over $\mathcal{T}$. |
| $\mathbf{z}_{k,h}^{\mathcal{S},(\boldsymbol{t})}$ | Head-wise signed condensed sample $\tau_{h,k}\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})}$. |
| $q_{k,h}^{\mathcal{S},(\boldsymbol{t})}$, $q_{i,h}^{\mathcal{T},(\boldsymbol{t})}$ | Synthetic and training hinge activity indicators for class $k$ / sample $i$ under OvR head $h$. |
| $\mathbf{g}_{k,h}^{\mathcal{S},(\boldsymbol{t})}$, $\mathbf{g}_{k,h}^{\mathcal{T},(\boldsymbol{t})}$ | Classwise synthetic and training OvR gradient terms for class $k$ under head $h$. |
| $D_k^{\mathrm{OvR}}(\mathcal{P}^{(\boldsymbol{t})};\mathbf{W}^{(\boldsymbol{t})})$ | Scalar classwise OvR gradient-matching discrepancy for class $k$. |
| $\bar{q}_k^{\mathcal{S},(\boldsymbol{t})}$, $\bar{b}_{k,i}^{(\boldsymbol{t})}$ | Averaged synthetic activity and averaged synthetic–training activity product over the $K$ OvR heads. |
| $h_{\mathcal{P}}$, $\mathcal{R}_{\mathrm{mc}}(\mathcal{P})$ | Nearest-prototype multiclass evaluator induced by $\mathcal{P}$ and its population misclassification risk. |
| $\pi_k$, $\gamma_{k\ell}$, $\epsilon_{\mathrm{mc}}(\mathcal{P})$ | Class prior, pairwise class separation $\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell\|_2$, and maximum multiclass condensed-sample perturbation. |
| *Multiclass OvR coefficient and risk quantities* | |
| $u_k^{(\boldsymbol{t}^\star)}$, $c_{k,i}^{(\boldsymbol{t}^\star)}$, $c_{k,\mu}^{(\boldsymbol{t}^\star)}$ | Terminal OvR expansion coefficients for class $k$, with $c_{k,\mu}^{(\boldsymbol{t}^\star)}=\sum_{i\in\mathcal{I}_k(\mathcal{T})}c_{k,i}^{(\boldsymbol{t}^\star)}$. |
| $\boldsymbol{c}_k^{(\boldsymbol{t}^\star)}$ | Vector of terminal same-class coefficients $(c_{k,i}^{(\boldsymbol{t}^\star)})_{i\in\mathcal{I}_k(\mathcal{T})}$. |

*Continued on next page*

Table 2: *Table 2 continued*

| **Symbol** | **Meaning** |
|---|---|
| $T^{\mathrm{OvR}}_{\mathcal{T},k,\mathrm{lin}}$, $T^{\mathrm{OvR}}_{\mathcal{S},k,\mathrm{lin}}(r)$ | Class-$k$ OvR training and condensed-sample linear-window lengths in outer loop $r$. |
| $T^{\mathrm{OvR}}_{k,\mathrm{lin}}(r)$ | Joint class-$k$ OvR window length $\min\{T^{\mathrm{OvR}}_{\mathcal{T},k,\mathrm{lin}}, T^{\mathrm{OvR}}_{\mathcal{S},k,\mathrm{lin}}(r)\}$. |
| $B_{k,\mathrm{out}}$, $\ell_k$, $u_{k,\mathrm{out}}$, $\Delta_k$ | Class-$k$ OvR coefficient upper budget, certified averaging mass, upper envelope, and non-averaged remainder. |
| $\bar{u}_k$ | Upper bound on the terminal initialization coefficient induced by the class-$k$ OvR condensed-sample windows. |
| $\mathcal{W}^{\mathrm{OvR}}_k$ | Deterministic window-deficit diagnostic for class $k$ in the multiclass OvR active-window comparison. |
| $\widehat{\mathrm{SR}}^{\mathrm{mc}}_{k\ell}$, $\underline{\mathrm{SR}}^{\mathrm{mc}}_{k\ell}$ | Pairwise multiclass signal-to-residual ratio and its certified lower bound for classes $k$ and $\ell$. |
| $L^{\mathrm{mc}}_{k\ell}$ | Pairwise multiclass factor combining class-averaging mass, initialization remainder, signal-recovery error, and active-window deficit. |
| $\Psi_{\mathrm{mc}}$ | Monotone function mapping pairwise multiclass signal-to-residual ratios into the exponential risk bound. |
| $\gamma_{\min}$, $\underline{\mathrm{SR}}^{\mathrm{mc}}_{\min}$ | Minimum pairwise class separation and minimum certified pairwise multiclass signal-to-residual ratio. |
| *Main theoretical objects* | |
| $\boldsymbol{t}^\star$ | Terminal time index $(T_{\mathrm{out}}-1, T_{\mathrm{in}})$. |
| $\mathbf{z}_y^{\mathcal{S},(\boldsymbol{t}^\star)}$ | Terminal signed condensed sample. |
| $\rho_{\mathcal{S}} := 1-2\eta_{\mathcal{S}}$ | Contraction factor in condensed-sample dynamics (stability: $0 < 2\eta_{\mathcal{S}} \le 1$). |
| $T_{\mathcal{T},\mathrm{lin}}$, $T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})$ | Linear-window lengths where training / condensed hinge indicators stay active. |
| $T_{y,\mathrm{lin}}(r)$ | $\min\{T_{\mathcal{T},\mathrm{lin}}, T_{\mathcal{S},\mathrm{lin}}(r)\}$. |
| $\widetilde{u}_y^{(\boldsymbol{t}^\star)}$, $\widetilde{c}_{y,j}^{(\boldsymbol{t}^\star)}$ | Coefficients in the terminal expansion of $\mathbf{z}_y^{\mathcal{S}}$ over initialization and $\{\mathbf{z}_j^{\mathcal{T}}\}$. |
| $p_y$, $\beta_p$ | Class proportion $n_{\mathcal{T}_y}/n_{\mathcal{T}}$ and balance factor $p_+^{-1/2} + p_-^{-1/2}$. |
| $B_{\mathrm{out}}$ | Outer-loop accumulation budget for terminal coefficient upper bounds; its construction is given in Lemma D.8. |
| $\ell_y$, $u_{\mathrm{out}}$, $\Delta_y$ | Coefficient mass certified by classwise averaging, coefficient upper envelope, and the non-averaged remainder $u_{\mathrm{out}} - \ell_y$; constructed in Lemma D.8. |
| $\widehat{\mathrm{SR}}(\mathcal{S})$ | Signal-to-residual ratio for $\mathbf{x}_+ - \mathbf{x}_- = A(\mathcal{S})\boldsymbol{\mu} + \mathbf{r}(\mathcal{S})$. |
| $\underline{\mathrm{SR}}_{\mathrm{dc}}$ | Certified lower bound on $\widehat{\mathrm{SR}}$ for the terminal condensed set. |
| $L$ | Dimensionless factor combining coefficient mass from certified classwise averaging, class balance, and the non-averaged coefficient remainder. |
| $\mathcal{W}_{\mathrm{win}}$ | Deterministic diagnostic for the loss caused when certified active prefixes are shorter than the full inner loop, used in Lemma D.12 to compare regimes for $L$. |
| $\mathcal{S}_{\mathrm{dc}}$, $\mathcal{S}_{\mathrm{core}}$ | Terminal condensed set vs. random one-shot coreset from two training points. |
| $G_{\max}$ | Upper bound on inner products $\langle \mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_{i'}^{\mathcal{T}} \rangle$ (appendix); regime $G_{\max} < \lambda$ gives full hinge activity. |
| *Assumption 2.2* | |
| $\delta$, $\kappa$, $C_\kappa$ | Tail probability, dimension–log constant, and associated concentration constant. |
| $\sigma_{\mathbf{w}}$, $\sigma_{\mathcal{S}}$ | Initialization scales for $\mathbf{w}$ and condensed samples. |
| *Inner-loop weight dynamics (appendix)* | |
| $\rho := 1 - \lambda\eta_{\mathbf{w}}$ | Contraction factor for weight updates under fixed hinge activity. |
| $\mathcal{I}(\mathcal{A})$, $n_{\mathcal{A}}$ | Index set and size of a generic dataset $\mathcal{A}$. |
| $\overline{\mathbf{z}}_{\mathcal{T}}$ | Mean of signed training samples $\frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}(\mathcal{T})} \mathbf{z}_i^{\mathcal{T}}$. |

*Sub-Gaussian noise.* Vectors satisfying Definition C.1 (Appendix C.1).

# B Multiclass One-Condensed-Sample-Per-Class Extension

This appendix extends the coefficient-expansion mechanism of the binary analysis to a $K$-class one-condensed-sample-per-class setting. Rather than applying a standard one-vs-rest reduction that runs $K$ separate binary condensation problems and produces a two-point condensed set for each induced subproblem, we maintain a single multiclass condensed set with one synthetic sample per class. The extension uses a classwise OvR gradient-matching objective and preserves the classwise affine-contraction structure that drives the binary hinge-loss analysis.

We work under the $K$-class additive model in Section 2.1: $y \in [K]$ has class prior $\pi_k = \mathbb{P}(y = k)$, and $\mathbf{x} = \boldsymbol{\mu}_y + \boldsymbol{\xi}$ with mean-zero sub-Gaussian noise proxy covariance $\boldsymbol{\Sigma}_\xi$, with independent noise across samples. The results below are stated in terms of the empirical class sizes $n_{\mathcal{T},k}$ and the pairwise separations $\gamma_{k\ell} = \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell\|_2$, without imposing simplex symmetry on the class means.

## B.1 Classwise OvR Gradient Matching

We use the same two-dimensional time convention as in the main text: $\boldsymbol{t} = (t_{\text{out}}, t_{\text{in}})$, where $t_{\text{out}} \in \{0, \ldots, T_{\text{out}} - 1\}$ and $t_{\text{in}} \in \{0, \ldots, T_{\text{in}}\}$. For an inner-loop update with $t_{\text{in}} < T_{\text{in}}$, write $\boldsymbol{t}^+ := (t_{\text{out}}, t_{\text{in}} + 1)$. For $t_{\text{out}} = 0, \ldots, T_{\text{out}} - 2$, the condensed samples are inherited across outer loops as $\mathbf{p}_k^{\mathcal{S},(t_{\text{out}}+1,0)} = \mathbf{p}_k^{\mathcal{S},(t_{\text{out}},T_{\text{in}})}$.

For $K$-class classification, the condensed object at time $\boldsymbol{t}$ is the multiclass condensed set $\mathcal{P}^{(\boldsymbol{t})} := \{(\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})}, k) : k \in [K]\}$, where $\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} \in \mathbb{R}^d$ is the single condensed sample assigned to class $k$. We write $\mathcal{P}^{(\boldsymbol{t})}$ to distinguish this multiclass object from the binary condensed-set notation $\mathcal{S}$. The condensed samples are initialized independently as $\mathbf{p}_k^{\mathcal{S},(0,0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathcal{S}}^2 \mathbf{I}_d)$ for $k \in [K]$. We collect the $K$ OvR heads as $\mathbf{W}^{(\boldsymbol{t})} := (\mathbf{w}_h^{(\boldsymbol{t})})_{h=1}^K$, where $\mathbf{w}_h^{(\boldsymbol{t})} \in \mathbb{R}^d$. At the beginning of each outer loop, every head is reinitialized as $\mathbf{w}_h^{(t_{\text{out}},0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{w}}^2 \mathbf{I}_d)$.

For head $h$, class $h$ is treated as positive and all other classes as negative. Define the OvR sign $\tau_{h,k} := 2\mathbb{1}\{h = k\} - 1 \in \{+1, -1\}$ for $h, k \in [K]$. For head-wise signed comparisons, set $\mathbf{z}_{i,h}^{\mathcal{T}} := \tau_{h,y_i^{\mathcal{T}}} \mathbf{x}_i^{\mathcal{T}}$, $\bar{\mathbf{z}}_h^{\mathcal{T}} := n_{\mathcal{T}}^{-1} \sum_{i=1}^{n_{\mathcal{T}}} \mathbf{z}_{i,h}^{\mathcal{T}}$, and $\mathbf{z}_{k,h}^{\mathcal{S},(\boldsymbol{t})} := \tau_{h,k} \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})}$. These signed coordinates express each OvR head in the same hinge-active form as the binary analysis.

**Assumption B.1.** For every head $h$, the signed training samples $\{\mathbf{z}_{i,h}^{\mathcal{T}}\}_{i=1}^{n_{\mathcal{T}}}$ and signed condensed samples $\{\mathbf{z}_{k,h}^{\mathcal{S},(\boldsymbol{t})}\}_{k=1}^K$ satisfy the analogues of Assumption D.1 and Lemmas C.2–C.5, with the binary signed variables replaced by their head-wise OvR counterparts and with the corresponding suprema taken over heads, classes, condensed samples, and training samples. Thus the initialization bounds, signed-sample concentration bounds, and active-window margin envelopes required for the binary coefficient estimates hold uniformly over the OvR signed coordinates.

The synthetic hinge activity of the class-$k$ condensed sample under head $h$ is $q_{k,h}^{\mathcal{S},(\boldsymbol{t})} := \mathbb{1}\{1 - \tau_{h,k} \langle \mathbf{w}_h^{(\boldsymbol{t})}, \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} \rangle > 0\}$. For a training example $(\mathbf{x}_i^{\mathcal{T}}, y_i^{\mathcal{T}})$, define $q_{i,h}^{\mathcal{T},(\boldsymbol{t})} := \mathbb{1}\{1 - \tau_{h,y_i^{\mathcal{T}}} \langle \mathbf{w}_h^{(\boldsymbol{t})}, \mathbf{x}_i^{\mathcal{T}} \rangle > 0\}$; in particular, if $i \in \mathcal{I}_k(\mathcal{T})$, then $y_i^{\mathcal{T}} = k$ and the sign is $\tau_{h,k}$.

For class $k$ and head $h$, define the classwise OvR gradient terms $\mathbf{g}_{k,h}^{\mathcal{S},(\boldsymbol{t})} := q_{k,h}^{\mathcal{S},(\boldsymbol{t})} \tau_{h,k} \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})}$ and $\mathbf{g}_{k,h}^{\mathcal{T},(\boldsymbol{t})} := n_{\mathcal{T},k}^{-1} \sum_{i \in \mathcal{I}_k(\mathcal{T})} q_{i,h}^{\mathcal{T},(\boldsymbol{t})} \tau_{h,k} \mathbf{x}_i^{\mathcal{T}}$. The scalar classwise OvR gradient-matching discrepancy for class $k$ is

$$D_k^{\text{OvR}}\left(\mathcal{P}^{(\boldsymbol{t})}; \mathbf{W}^{(\boldsymbol{t})}\right) := \frac{1}{K} \sum_{h=1}^K \left\| \mathbf{g}_{k,h}^{\mathcal{S},(\boldsymbol{t})} - \mathbf{g}_{k,h}^{\mathcal{T},(\boldsymbol{t})} \right\|_2^2.$$

Since the same sign $\tau_{h,k}$ multiplies the synthetic and training terms within a fixed class block, the head-wise squared norm is unchanged if this common sign is removed.

Define the averaged synthetic activity $\bar{q}_k^{\mathcal{S},(\boldsymbol{t})} := K^{-1} \sum_{h=1}^K q_{k,h}^{\mathcal{S},(\boldsymbol{t})}$ and, for $i \in \mathcal{I}_k(\mathcal{T})$, the averaged activity product $\bar{b}_{k,i}^{(\boldsymbol{t})} := K^{-1} \sum_{h=1}^K q_{k,h}^{\mathcal{S},(\boldsymbol{t})} q_{i,h}^{\mathcal{T},(\boldsymbol{t})}$. By construction, $0 \le \bar{b}_{k,i}^{(\boldsymbol{t})} \le \bar{q}_k^{\mathcal{S},(\boldsymbol{t})} \le 1$.

**Algorithm 2** Classwise OvR Gradient Matching for Multiclass Condensed Samples

**Input:** Training set $\mathcal{T}$, number of classes $K$, loop lengths $T_{\text{out}}$ and $T_{\text{in}}$, step sizes $\eta_{\mathbf{w}}$ and $\eta_{\mathcal{S}}$
**Output:** Multiclass condensed set $\mathcal{P}_{\text{dc}}$
Initialize $\mathbf{p}_k^{\mathcal{S},(0,0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathcal{S}}^2 \mathbf{I}_d)$ for $k \in [K]$
Set $\tau_{h,k} = 2\mathbb{1}\{h = k\} - 1$ for $h, k \in [K]$
**for** $t_{\text{out}} = 0$ **to** $T_{\text{out}} - 1$ **do**
  Initialize $\mathbf{w}_h^{(t_{\text{out}},0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{w}}^2 \mathbf{I}_d)$ for $h \in [K]$
  **if** $t_{\text{out}} > 0$ **then**
    Set $\mathbf{p}_k^{\mathcal{S},(t_{\text{out}},0)} \leftarrow \mathbf{p}_k^{\mathcal{S},(t_{\text{out}}-1,T_{\text{in}})}$ for each $k \in [K]$
  **end if**
  **for** $t_{\text{in}} = 0$ **to** $T_{\text{in}} - 1$ **do**
    Set $\boldsymbol{t} = (t_{\text{out}}, t_{\text{in}})$ and $\boldsymbol{t}^+ = (t_{\text{out}}, t_{\text{in}} + 1)$
    Compute $q_{k,h}^{\mathcal{S},(\boldsymbol{t})}$ and $q_{i,h}^{\mathcal{T},(\boldsymbol{t})}$ for all $k, h \in [K]$ and $i \in \{1, \ldots, n_{\mathcal{T}}\}$
    Compute $\bar{q}_k^{\mathcal{S},(\boldsymbol{t})}$ and $\bar{b}_{k,i}^{(\boldsymbol{t})}$ for every $k \in [K]$ and $i \in \mathcal{I}_k(\mathcal{T})$
    **for** $k \in [K]$ **do**
      Update $\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^+)}$ by equation 21
    **end for**
    **for** $h \in [K]$ **do**
      Compute $\mathbf{g}_h^{\mathcal{T},(\boldsymbol{t})}$ and update $\mathbf{w}_h^{(\boldsymbol{t}^+)}$ by equation 22
    **end for**
  **end for**
**end for**
**Return:** $\mathcal{P}_{\text{dc}} = \{(\mathbf{p}_k^{\mathcal{S},(T_{\text{out}}-1,T_{\text{in}})}, k) : k \in [K]\}$

For $t_{\text{in}} < T_{\text{in}}$, the multiclass condensed-sample update is the classwise gradient step $\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^+)} = \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} - \eta_{\mathcal{S}} \nabla_{\mathbf{p}_k^{\mathcal{S}}} D_k^{\text{OvR}}(\mathcal{P}^{(\boldsymbol{t})}; \mathbf{W}^{(\boldsymbol{t})})$, where the hinge activities are evaluated at time $\boldsymbol{t}$ and are not differentiated through, as in the binary analysis. Under this convention, the update has the explicit affine form

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^+)} = \left(1 - 2\eta_{\mathcal{S}} \bar{q}_k^{\mathcal{S},(\boldsymbol{t})}\right) \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} + \frac{2\eta_{\mathcal{S}}}{n_{\mathcal{T},k}} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \bar{b}_{k,i}^{(\boldsymbol{t})} \mathbf{x}_i^{\mathcal{T}}. \tag{21}$$

For $t_{\text{in}} < T_{\text{in}}$ and each head $h$, the full training OvR gradient term is $\mathbf{g}_h^{\mathcal{T},(\boldsymbol{t})} := n_{\mathcal{T}}^{-1} \sum_{i=1}^{n_{\mathcal{T}}} q_{i,h}^{\mathcal{T},(\boldsymbol{t})} \tau_{h,y_i^{\mathcal{T}}} \mathbf{x}_i^{\mathcal{T}}$, and the real-data SVM update is

$$\mathbf{w}_h^{(\boldsymbol{t}^+)} = (1 - \eta_{\mathbf{w}} \lambda) \mathbf{w}_h^{(\boldsymbol{t})} + \eta_{\mathbf{w}} \mathbf{g}_h^{\mathcal{T},(\boldsymbol{t})}. \tag{22}$$

Algorithm 2 summarizes the resulting classwise OvR gradient-matching procedure. It maintains a single multiclass condensed set across outer loops, while the $K$ OvR heads are reinitialized at every outer-loop restart.

### B.2 Multiclass Nearest-Prototype Evaluation

The multiclass evaluator below turns a condensed set into a prediction rule and provides the target of the population risk bounds.

**Definition B.2** (Multiclass nearest-prototype evaluator)**.** Let $\mathcal{P} = \{(\mathbf{p}_k, k) : k \in [K]\}$ be a generic multiclass condensed set with one representative per class, not necessarily tied to a particular condensation time. In this evaluator, these representatives act as class prototypes, and the nearest-prototype classifier induced by $\mathcal{P}$ is

$$h_{\mathcal{P}}(\mathbf{x}) := \arg\min_{k \in [K]} \|\mathbf{x} - \mathbf{p}_k\|_2^2 = \arg\max_{k \in [K]} \left\{ \langle \mathbf{x}, \mathbf{p}_k \rangle - \frac{1}{2} \|\mathbf{p}_k\|_2^2 \right\},$$

with deterministic tie-breaking. The corresponding multiclass population risk is

$$\mathcal{R}_{\text{mc}}(\mathcal{P}) := \mathbb{P}_{(\mathbf{x},y) \sim \mathcal{D}} \left( h_{\mathcal{P}}(\mathbf{x}) \neq y \right).$$

**Examples of multiclass nearest-prototype evaluators.** The nearest-prototype evaluator is a standard multiclass prediction family, and it appears in several familiar classifiers, either exactly or after a fixed feature transformation.

- **Nearest class mean.** If $\widehat{\boldsymbol{\mu}}_k^{\mathcal{T}} := n_{\mathcal{T},k}^{-1} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \mathbf{x}_i^{\mathcal{T}}$ and $\mathcal{P}_{\mathrm{NCM}} := \{(\widehat{\boldsymbol{\mu}}_k^{\mathcal{T}}, k) : k \in [K]\}$, then $h_{\mathcal{P}_{\mathrm{NCM}}}(\mathbf{x}) = \arg\min_{k \in [K]} \|\mathbf{x} - \widehat{\boldsymbol{\mu}}_k^{\mathcal{T}}\|_2^2$, which is the standard nearest-class-mean classifier. It is the evaluator most closely connected to the full-activity specialization, in which each condensed sample contracts toward the corresponding class mean.
- **Shared-covariance Gaussian Bayes and LDA.** If $\mathbf{x} \mid y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with equal class priors, the Bayes rule is $\arg\min_{k \in [K]} (\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)$. After whitening, $\widetilde{\mathbf{x}} := \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$ and $\widetilde{\mathbf{p}}_k := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_k$, this becomes the nearest-prototype rule $\arg\min_{k \in [K]} \|\widetilde{\mathbf{x}} - \widetilde{\mathbf{p}}_k\|_2^2$. Unequal priors add class-dependent bias terms and therefore correspond to a biased nearest-prototype score rather than the unweighted Euclidean form unless one augments the feature space.
- **Affine linear heads.** Any affine multiclass linear prediction rule $h_{\mathrm{lin}}(\mathbf{x}) = \arg\max_{k \in [K]} \{\langle \mathbf{a}_k, \mathbf{x} \rangle + b_k\}$ is exactly a nearest-prototype rule when $\mathbf{p}_k = \mathbf{a}_k$ and $b_k = -\frac{1}{2}\|\mathbf{a}_k\|_2^2 + C$ for a class-independent constant $C$. More generally, every affine linear rule can be represented exactly as a nearest-prototype rule after adding one dummy feature dimension: choose $C \geq \max_k \{b_k + \frac{1}{2}\|\mathbf{a}_k\|_2^2\}$, set $r_k^2 := 2(C - b_k) - \|\mathbf{a}_k\|_2^2 \geq 0$, and use $\overline{\mathbf{x}} := (\mathbf{x}, 0)$ and $\overline{\mathbf{p}}_k := (\mathbf{a}_k, r_k)$. Then $\langle \overline{\mathbf{x}}, \overline{\mathbf{p}}_k \rangle - \frac{1}{2}\|\overline{\mathbf{p}}_k\|_2^2 = \langle \mathbf{x}, \mathbf{a}_k \rangle + b_k - C$, so the class-independent shift does not change the argmax. This representation includes the prediction rules of multiclass logistic regression, linear softmax heads, and multiclass linear SVMs.
- **Prototypical and cosine classifiers.** In a fixed representation space $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^r$, the rule $\arg\min_{k \in [K]} \|\boldsymbol{\phi}(\mathbf{x}) - \mathbf{p}_k\|_2^2$ is the usual prototypical-network or embedding-space nearest-class-mean prediction rule when $\mathbf{p}_k = n_{\mathcal{T},k}^{-1} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \boldsymbol{\phi}(\mathbf{x}_i^{\mathcal{T}})$. If the features and prototypes are normalized, or more generally if all prototypes have the same norm, the nearest-prototype score is equivalent to maximizing the cosine or inner-product score $\langle \boldsymbol{\phi}(\mathbf{x}), \mathbf{p}_k \rangle$. This connects the evaluator to normalized linear heads in a fixed feature space.

For the $K$-class additive model, write $\pi_k := \mathbb{P}_{(\mathbf{x},y) \sim \mathcal{D}}(y = k)$ and $\gamma_{k\ell} := \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell\|_2$ for $k, \ell \in [K]$. For a multiclass condensed set $\mathcal{P} = \{(\mathbf{p}_k, k) : k \in [K]\}$, define the maximum condensed-sample perturbation $\epsilon_{\mathrm{mc}}(\mathcal{P}) := \max_{k \in [K]} \|\mathbf{p}_k - \boldsymbol{\mu}_k\|_2$. When the condensed set is clear from context, write $\epsilon_{\mathrm{mc}}$.

### B.3 Main Multiclass Guarantees

The following results give the multiclass consequences of the classwise OvR update. The first result is the direct analogue of the binary coefficient expansion: each terminal condensed sample is a nonnegative combination of its initialization and same-class training samples, with coefficient bounds controlled by the certified active windows. Lemma D.13 gives sufficient margin-envelope conditions for these windows.

**Theorem B.3** (Classwise OvR coefficient expansion)**.** *Assume $0 < 2\eta_{\mathcal{S}} \leq 1$ and $n_{\mathcal{T},k} \geq 1$ for every $k \in [K]$. Set $\rho_{\mathcal{S}} := 1 - 2\eta_{\mathcal{S}}$ and $\boldsymbol{t}^\star := (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$. Under classwise OvR gradient matching, for every class $k \in [K]$ there exist coefficients $u_k^{(\boldsymbol{t}^\star)} \geq 0$ and $c_{k,i}^{(\boldsymbol{t}^\star)} \geq 0$, $i \in \mathcal{I}_k(\mathcal{T})$, such that*

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} = u_k^{(\boldsymbol{t}^\star)} \mathbf{p}_k^{\mathcal{S},(0,0)} + \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)} \mathbf{x}_i^{\mathcal{T}}.$$

*Consequently, under the $K$-class additive model, with $c_{k,\mu}^{(\boldsymbol{t}^\star)} := \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)}$,*

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} = u_k^{(\boldsymbol{t}^\star)} \mathbf{p}_k^{\mathcal{S},(0,0)} + c_{k,\mu}^{(\boldsymbol{t}^\star)} \boldsymbol{\mu}_k + \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)} \boldsymbol{\xi}_i.$$

*Assume further that the class-$k$ training activities and condensed-sample activities are active on windows of lengths $T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}$ and $T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)$, respectively, in every outer loop $r \in \{0, \ldots, T_{\mathrm{out}} - 1\}$. Define $T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r) :=$*

$\min\{T^{\mathrm{OvR}}_{\mathcal{T},k,\mathrm{lin}}, T^{\mathrm{OvR}}_{\mathcal{S},k,\mathrm{lin}}(r)\}$,

$$B_{k,\mathrm{out}} := 1 + \sum_{r=0}^{T_{\mathrm{out}}-2} \rho_{\mathcal{S}}^{\sum_{s=r+1}^{T_{\mathrm{out}}-1} T^{\mathrm{OvR}}_{\mathcal{S},k,\mathrm{lin}}(s)}, \qquad u_{k,\mathrm{out}} := (1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}}) B_{k,\mathrm{out}},$$

*and*

$$\ell_k := \sum_{r=0}^{T_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{(T_{\mathrm{out}}-r)T_{\mathrm{in}} - T^{\mathrm{OvR}}_{k,\mathrm{lin}}(r)} \big(1-\rho_{\mathcal{S}}^{T^{\mathrm{OvR}}_{k,\mathrm{lin}}(r)}\big).$$

*Then*

$$\rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}} \le u_k^{(\boldsymbol{t}^\star)} \le \rho_{\mathcal{S}}^{\sum_{r=0}^{T_{\mathrm{out}}-1} T^{\mathrm{OvR}}_{\mathcal{S},k,\mathrm{lin}}(r)}, \qquad \frac{\ell_k}{n_{\mathcal{T},k}} \le c_{k,i}^{(\boldsymbol{t}^\star)} \le \frac{u_{k,\mathrm{out}}}{n_{\mathcal{T},k}}.$$

*In particular,* $\ell_k \le c_{k,\mu}^{(\boldsymbol{t}^\star)} \le u_{k,\mathrm{out}}$ *and* $\ell_k/\sqrt{n_{\mathcal{T},k}} \le \|\boldsymbol{c}_k^{(\boldsymbol{t}^\star)}\|_2 \le u_{k,\mathrm{out}}/\sqrt{n_{\mathcal{T},k}}$*, where* $\boldsymbol{c}_k^{(\boldsymbol{t}^\star)} := (c_{k,i}^{(\boldsymbol{t}^\star)})_{i\in\mathcal{I}_k(\mathcal{T})}$*. If all OvR activities are active throughout the bilevel procedure, then*

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} = \rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}} \mathbf{p}_k^{\mathcal{S},(0,0)} + \big(1-\rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}\big)\widehat{\boldsymbol{\mu}}_k^{\mathcal{T}}, \qquad \widehat{\boldsymbol{\mu}}_k^{\mathcal{T}} := \frac{1}{n_{\mathcal{T},k}} \sum_{i\in\mathcal{I}_k(\mathcal{T})} \mathbf{x}_i^{\mathcal{T}}.$$

*Proof.* The active-window certification is Lemma D.13. The coefficient expansion, coefficient bounds, and full-activity contraction are Lemma D.14. □

The next result is independent of the OvR dynamics. It states that nearest-prototype evaluation has a pairwise risk bound controlled by class separation and condensed-sample perturbation.

**Theorem B.4** (Multiclass nearest-prototype risk)**.** *Assume the $K$-class additive test model $\mathbf{x} = \boldsymbol{\mu}_y + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is mean-zero sub-Gaussian with proxy covariance $\boldsymbol{\Sigma}_\xi$. Let $\mathcal{P} = \{(\mathbf{p}_k, k) : k \in [K]\}$ be any multiclass condensed set used with the nearest-prototype evaluator. Define $\epsilon_k := \|\mathbf{p}_k - \boldsymbol{\mu}_k\|_2$, $R_{k\ell} := \epsilon_k + \epsilon_\ell$, and $\widehat{\mathrm{SR}}^{\mathrm{mc}}_{k\ell} := \gamma_{k\ell}/R_{k\ell}$, with value $+\infty$ when $R_{k\ell} = 0$. For $s \in (0, +\infty)$, set*

$$\Psi_{\mathrm{mc}}(s) := \frac{\big((1-s^{-1})_+\big)^4}{(1+s^{-1})^2}, \qquad \Psi_{\mathrm{mc}}(0) := 0, \qquad \Psi_{\mathrm{mc}}(+\infty) := 1.$$

*Then*

$$\mathcal{R}_{\mathrm{mc}}(\mathcal{P}) \le \sum_{k=1}^{K} \pi_k \sum_{\ell\neq k} \exp\left(-\frac{\gamma_{k\ell}^2}{8\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}} \Psi_{\mathrm{mc}}\left(\widehat{\mathrm{SR}}^{\mathrm{mc}}_{k\ell}\right)\right).$$

*Proof.* The statement is proved in Lemma D.15. □

Combining the OvR coefficient expansion with the nearest-prototype risk bound gives a certified multiclass guarantee for the terminal condensed set. The bound is written through pairwise signal-to-residual ratios; larger pairwise ratios yield stronger exponential risk decay.

**Corollary B.5** (Certified multiclass risk for classwise OvR condensation)**.** *Assume the hypotheses of Theorem B.3 and the high-probability event in Assumption B.1. Let $C_\kappa$ denote the corresponding uniform concentration constant and define the terminal multiclass condensed set $\mathcal{P}_{\mathrm{dc}} := \{(\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)}, k) : k \in [K]\}$. Set $p_k := n_{\mathcal{T},k}/n_{\mathcal{T}}$,*

$$\alpha_{\mathcal{S}} := \frac{\sigma_{\mathcal{S}}\sqrt{n_{\mathcal{T}}}}{\sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}}}, \qquad \alpha_{\mu,k} := \frac{\|\boldsymbol{\mu}_k\|_2\sqrt{n_{\mathcal{T}}}}{\sqrt{C_\kappa\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}}, \qquad \bar{u}_k := \rho_{\mathcal{S}}^{\sum_{r=0}^{T_{\mathrm{out}}-1} T^{\mathrm{OvR}}_{\mathcal{S},k,\mathrm{lin}}(r)}.$$

*For $k \neq \ell$, define*

$$L^{\mathrm{mc}}_{k\ell} := \Big[\alpha_{\mathcal{S}}(\bar{u}_k + \bar{u}_\ell) + \alpha_{\mu,k}(1-\ell_k) + \alpha_{\mu,\ell}(1-\ell_\ell) + \ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2} + \sqrt{n_{\mathcal{T}}}(\Delta_k + \Delta_\ell)\Big]^{-1},$$

*where* $\Delta_k := u_{k,\mathrm{out}} - \ell_k$*, and set*

$$\underline{\mathrm{SR}}^{\mathrm{mc}}_{k\ell} := \frac{\sqrt{n_{\mathcal{T}}}\, L^{\mathrm{mc}}_{k\ell}\, \gamma_{k\ell}}{\sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d}}.$$

*Then*

$$\mathcal{R}_{\mathrm{mc}}(\mathcal{P}_{\mathrm{dc}}) \le \sum_{k=1}^{K} \pi_k \sum_{\ell \neq k} \exp\left( -\frac{\gamma^2_{k\ell}}{8\|\mathbf{\Sigma}_\xi\|_{\mathrm{op}}} \Psi_{\mathrm{mc}}\left(\underline{\mathrm{SR}}^{\mathrm{mc}}_{k\ell}\right)\right).$$

*In particular, with* $\gamma_{\min} := \min_{k\neq\ell} \gamma_{k\ell}$ *and* $\underline{\mathrm{SR}}^{\mathrm{mc}}_{\min} := \min_{k\neq\ell} \underline{\mathrm{SR}}^{\mathrm{mc}}_{k\ell}$*,*

$$\mathcal{R}_{\mathrm{mc}}(\mathcal{P}_{\mathrm{dc}}) \le (K-1) \exp\left( -\frac{\gamma^2_{\min}}{8\|\mathbf{\Sigma}_\xi\|_{\mathrm{op}}} \Psi_{\mathrm{mc}}\left(\underline{\mathrm{SR}}^{\mathrm{mc}}_{\min}\right)\right).$$

*If all OvR activities are active throughout training, then* $L^{\mathrm{mc}}_{k\ell}$ *can be replaced by*

$$L^{\mathrm{mc}}_{k\ell,\mathrm{full}} := \left[\rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}\left(2\alpha_{\mathcal{S}} + \alpha_{\mu,k} + \alpha_{\mu,\ell}\right) + \left(1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}\right)\left(p_k^{-1/2} + p_\ell^{-1/2}\right)\right]^{-1}$$

*in the definition of* $\underline{\mathrm{SR}}^{\mathrm{mc}}_{k\ell}$*.*

*Proof.* The statement is proved in Corollary D.16. □

*Remark* B.6 (Active-window interpretation). The factor $L^{\mathrm{mc}}_{k\ell}$ separates the class-averaging contribution from the remainder induced by incomplete active windows. Lemma D.17 gives the deterministic comparison: when initialization and signal remainders are no larger than the class-averaging term and the active-window deficits satisfy $\Delta_k + \Delta_\ell = O((\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2})/\sqrt{n_{\mathcal{T}}})$, the certified ratio has the same $\sqrt{n_{\mathcal{T}}}$ scale as class averaging. If the deficits are instead of constant order, the pairwise factor satisfies $L^{\mathrm{mc}}_{k\ell} = O(n_{\mathcal{T}}^{-1/2})$, so the averaging gain is no longer reflected in the certified ratio. In the full-window case, the transition is governed by the comparison between the initialization term $\rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}(2\alpha_{\mathcal{S}} + \alpha_{\mu,k} + \alpha_{\mu,\ell})$ and the class-averaging term $(1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}})(p_k^{-1/2} + p_\ell^{-1/2})$.

# C Preliminary Lemmas

This section introduces several preliminary lemmas that serve as the foundation for the subsequent proofs and results.

## C.1 Bounds for Random Variables and Aggregates

Our theoretical analysis relies on controlling the behavior of various random quantities that arise from the data generation process and random initializations. This subsection establishes the foundational high-probability bounds for these quantities, starting with the core definition of sub-Gaussian vectors which characterizes our noise model.

**Definition C.1** (Sub-Gaussian with proxy covariance)**.** A random vector $\boldsymbol{\xi} \in \mathbb{R}^d$ is sub-Gaussian with proxy covariance $\boldsymbol{\Sigma} \succeq \mathbf{0}$ if $\mathbb{E}\exp\big(\lambda\langle \mathbf{v}, \boldsymbol{\xi}\rangle\big) \le \exp\big(\frac{\lambda^2}{2}\,\mathbf{v}^\top \boldsymbol{\Sigma}\mathbf{v}\big)$ for all $\lambda \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^d$.

The following lemmas establish high-probability bounds for various random variables and their aggregates, crucial for controlling noise and initialization effects throughout our analysis.

**Lemma C.2.** *Assume the training set follows the additive model of Section 2.1 with sub-Gaussian noise parameter $\boldsymbol{\Sigma}_\xi$. Fix $\delta \in (0,1)$ and suppose the consolidated dimension condition $d \ge \kappa \log\big(6T_{\text{out}} n_{\mathcal{T}}/\delta\big)$ for some $\kappa > 0$. There exist absolute constants $C > 0$ and $C_\kappa > 0$ such that, with probability at least $1-\delta$, simultaneously for all $i, i' \in [n_{\mathcal{T}}]$,*

$$\big|\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle\big| \;\le\; \sqrt{2\log\big(6n_{\mathcal{T}}/\delta\big)}\,\big\|\boldsymbol{\Sigma}_\xi^{1/2}\boldsymbol{\mu}\big\|_2, \quad \le\; C_\kappa \sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\,\|\boldsymbol{\mu}\|_2\,\sqrt{d}. \tag{23}$$

$$\big|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle\big| \;\le\; C\Big(\|\boldsymbol{\Sigma}_\xi\|_F\,\sqrt{\log\big(6n_{\mathcal{T}}^2/\delta\big)} \;+\; \|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,\log\big(6n_{\mathcal{T}}^2/\delta\big)\Big), \qquad i \ne i', \quad \le\; C_\kappa\,\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,d. \tag{24}$$

$$\Big|\;\|\boldsymbol{\xi}_i\|_2^2 - \operatorname{tr}(\boldsymbol{\Sigma}_\xi)\;\Big| \;\le\; C\Big(\|\boldsymbol{\Sigma}_\xi\|_F\,\sqrt{\log\big(6n_{\mathcal{T}}/\delta\big)} \;+\; \|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,\log\big(6n_{\mathcal{T}}/\delta\big)\Big), \quad \le\; C_\kappa\,\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,d. \tag{25}$$

*Proof.* For equation 23, since $\langle \boldsymbol{\mu}, \boldsymbol{\xi}_i\rangle$ is sub-Gaussian with variance proxy $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\xi\boldsymbol{\mu}$, for any $t > 0$,

$$\mathbb{P}\big(\big|\langle \boldsymbol{\mu}, \boldsymbol{\xi}_i\rangle\big| \ge t\big) \le 2\exp\!\left(-\frac{t^2}{2\,\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\xi\boldsymbol{\mu}}\right).$$

Taking $t = \sqrt{2\,\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\xi\boldsymbol{\mu}\,\log(6n_{\mathcal{T}}/\delta)}$ gives $\mathbb{P}(\,|\langle \boldsymbol{\mu}, \boldsymbol{\xi}_i\rangle| \ge t\,) \le \delta/(3n_{\mathcal{T}})$; a union bound over $i$ yields the first inequality. Using $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_\xi\boldsymbol{\mu} \le \|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\|\boldsymbol{\mu}\|_2^2$ and $\sqrt{\log\big(6n_{\mathcal{T}}/\delta\big)} \le \sqrt{\log\big(6T_{\text{out}}n_{\mathcal{T}}/\delta\big)} \le \kappa^{-1/2}\sqrt{d}$, we obtain the second line of equation 23 after choosing $C_\kappa \ge \kappa^{-1/2}$.

For equation 24, write $\boldsymbol{\xi}_i = \boldsymbol{\Sigma}_\xi^{1/2}\mathbf{z}_i$ where $\boldsymbol{\zeta}_i$ has independent, mean-zero, sub-Gaussian coordinates with unit variance. For $i \ne i'$ we have $\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle = \boldsymbol{\zeta}_i^\top\boldsymbol{\Sigma}_\xi\,\mathbf{z}_{i'}$. A decoupled Hanson–Wright type inequality for bilinear forms of independent sub-Gaussian vectors (e.g. Rudelson & Vershynin, 2013) yields a universal constant $c > 0$ such that, for all $t > 0$,

$$\mathbb{P}\big(\big|\boldsymbol{\zeta}_i^\top\boldsymbol{\Sigma}_\xi\,\mathbf{z}_{i'}\big| \ge t\big) \le 2\exp\!\left(-c\,\min\!\left\{\frac{t^2}{\|\boldsymbol{\Sigma}_\xi\|_F^2},\,\frac{t}{\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\right\}\right).$$

Choosing

$$t = C\Big(\|\boldsymbol{\Sigma}_\xi\|_F\,\sqrt{\log\big(6n_{\mathcal{T}}^2/\delta\big)} + \|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,\log\big(6n_{\mathcal{T}}^2/\delta\big)\Big)$$

with $C$ large enough (depending only on $c$) gives $\mathbb{P}(\,|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle| \ge t\,) \le \delta/(3n_{\mathcal{T}}^2)$. A union bound over ordered pairs $(i, i')$ with $i \ne i'$ yields the first line of equation 24. For the coarse bound, use $\|\boldsymbol{\Sigma}_\xi\|_F \le \sqrt{d}\,\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}$ and

$$\sqrt{\log\big(6n_{\mathcal{T}}^2/\delta\big)} \le \sqrt{\log\big(6T_{\text{out}}n_{\mathcal{T}}^2/\delta\big)} \le \sqrt{2\log\big(6T_{\text{out}}n_{\mathcal{T}}/\delta\big)} \le \sqrt{2/\kappa}\,\sqrt{d},$$

together with $\log(6n_{\mathcal{T}}^2/\delta) \le (2/\kappa)d$, which follows from the dimension condition. Enlarging constants gives the second line of equation 24.

For equation 25, we again write $\boldsymbol{\xi}_i = \boldsymbol{\Sigma}_\xi^{1/2}\mathbf{z}_i$ so that $\|\boldsymbol{\xi}_i\|_2^2 = \boldsymbol{\zeta}_i^\top \boldsymbol{\Sigma}_\xi\, \boldsymbol{\zeta}_i$ and $\mathbb{E}[\boldsymbol{\zeta}_i^\top \boldsymbol{\Sigma}_\xi\, \boldsymbol{\zeta}_i] = \operatorname{tr}(\boldsymbol{\Sigma}_\xi)$. The Hanson–Wright inequality implies that for all $t > 0$,

$$\mathbb{P}\big(\big|\boldsymbol{\zeta}_i^\top \boldsymbol{\Sigma}_\xi\, \boldsymbol{\zeta}_i - \operatorname{tr}(\boldsymbol{\Sigma}_\xi)\big| \ge t\big) \le 2\exp\!\left(-c\,\min\!\left\{\frac{t^2}{\|\boldsymbol{\Sigma}_\xi\|_F^2},\ \frac{t}{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}}\right\}\right).$$

Taking

$$t = C\Big(\|\boldsymbol{\Sigma}_\xi\|_F\,\sqrt{\log\big(6n_{\mathcal{T}}/\delta\big)} + \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\,\log\big(6n_{\mathcal{T}}/\delta\big)\Big)$$

gives a tail probability at most $\delta/(3n_{\mathcal{T}})$ for $C$ sufficiently large. A union bound over $i$ establishes the first line of equation 25. The second line follows from the same simplifications as above: $\|\boldsymbol{\Sigma}_\xi\|_F \le \sqrt{d}\,\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}$ and $\sqrt{\log(6n_{\mathcal{T}}/\delta)} \le \kappa^{-1/2}\sqrt{d}$, $\log(6n_{\mathcal{T}}/\delta) \le \kappa^{-1}d$.

Finally, the three parts are proved with failure probabilities at most $\delta/3$ each, so a union bound yields a joint event of probability at least $1-\delta$ on which equation 23–equation 25 hold simultaneously. □

**Lemma C.3.** *Assume the training set follows the additive model of Section 2.1 with sub-Gaussian noise parameter $\boldsymbol{\Sigma}_\xi$. Let $\mathbf{w}^{(t_{\mathrm{out}},0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{w}}^2\mathbf{I})$ be independent of $\{\boldsymbol{\xi}_i\}$. Fix $\delta \in (0,1)$ and suppose $d \ge \kappa\,\log(6T_{\mathrm{out}}n_{\mathcal{T}}/\delta)$ for some $\kappa > 0$. There exist absolute constants $C_\kappa > 0$, such that with probability at least $1-\delta$, simultaneously for all $t_{\mathrm{out}} \le T_{\mathrm{out}} - 1$ and $i \in [n_{\mathcal{T}}]$,*

$$\big\|\mathbf{w}^{(t_{\mathrm{out}},0)}\big\|_2^2 \;\le\; \sigma_{\mathbf{w}}^2\Big(d + 2\sqrt{d\,\log\big(6T_{\mathrm{out}}/\delta\big)} + 2\log\big(6T_{\mathrm{out}}/\delta\big)\Big), \quad \le\; C_\kappa\,\sigma_{\mathbf{w}}^2\,d. \tag{26}$$

$$\big|\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \boldsymbol{\mu}\rangle\big| \;\le\; \sigma_{\mathbf{w}}\,\|\boldsymbol{\mu}\|_2\,\sqrt{2\log\big(6T_{\mathrm{out}}/\delta\big)}, \quad \le\; C_\kappa\,\sigma_{\mathbf{w}}\,\|\boldsymbol{\mu}\|_2\,\sqrt{d}. \tag{27}$$

$$\begin{aligned}\big|\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \boldsymbol{\xi}_i\rangle\big| \;&\le\; \sigma_{\mathbf{w}}\,\sqrt{2\log\big(6T_{\mathrm{out}}n_{\mathcal{T}}/\delta\big)} \\ &\qquad \times \sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_\xi) \;+\; C\Big(\|\boldsymbol{\Sigma}_\xi\|_F\,\sqrt{\log\big(6n_{\mathcal{T}}/\delta\big)} + \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\,\log\big(6n_{\mathcal{T}}/\delta\big)\Big)}, \\ &\le\; C_\kappa\,\sigma_{\mathbf{w}}\,\sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}}\;d.\end{aligned} \tag{28}$$

*Proof.* For equation 28, condition on $\boldsymbol{\xi}_i$:

$$\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \boldsymbol{\xi}_i\rangle \mid \boldsymbol{\xi}_i \sim \mathcal{N}\big(0, \sigma_{\mathbf{w}}^2\|\boldsymbol{\xi}_i\|_2^2\big),$$

so for any $\tau > 0$,

$$\mathbb{P}\Big(\big|\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \boldsymbol{\xi}_i\rangle\big| \ge \tau \;\Big|\; \boldsymbol{\xi}_i\Big) \le 2\exp\!\left(-\frac{\tau^2}{2\sigma_{\mathbf{w}}^2\|\boldsymbol{\xi}_i\|_2^2}\right).$$

Take

$$\tau = \sigma_{\mathbf{w}}\,\sqrt{2\log\big(6T_{\mathrm{out}}n_{\mathcal{T}}/\delta\big)}\;\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_\xi) + C\Big(\|\boldsymbol{\Sigma}_\xi\|_F\,\sqrt{\log\big(6n_{\mathcal{T}}/\delta\big)} + \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\,\log\big(6n_{\mathcal{T}}/\delta\big)\Big)}.$$

By Lemma C.2, inequality equation 25 holds simultaneously for all $i$, hence

$$\|\boldsymbol{\xi}_i\|_2^2 \le \operatorname{tr}(\boldsymbol{\Sigma}_\xi) + C\Big(\|\boldsymbol{\Sigma}_\xi\|_F\,\sqrt{\log\big(6n_{\mathcal{T}}/\delta\big)} + \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\,\log\big(6n_{\mathcal{T}}/\delta\big)\Big).$$

A union bound over $i$ and $t_{\mathrm{out}}$ gives the first inequality in equation 28. Finally, using $\|\boldsymbol{\Sigma}_\xi\|_F \le \sqrt{d}\,\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}$ and the dimension condition to bound $\sqrt{\log(6T_{\mathrm{out}}n_{\mathcal{T}}/\delta)} \le C_\kappa\sqrt{d}$ and $\log(6n_{\mathcal{T}}/\delta) \le C_\kappa d$, we obtain the coarse bound in the second line of equation 28. □

**Lemma C.4.** *Assume the training set follows the additive model of Section 2.1 with sub-Gaussian noise parameter $\mathbf{\Sigma}_\xi$. Let $\{\mathbf{x}_i^{\mathcal{S},(0,0)}\}_{i\in\mathcal{I}(\mathcal{S})}$ be initialized independently as $\mathbf{x}_i^{\mathcal{S},(0,0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathcal{S}}^2\mathbf{I})$ and let $\mathbf{w}^{(t_{\text{out}},0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{w}}^2\mathbf{I})$ be independent of everything else for $t_{\text{out}} = 0, \dots, T_{\text{out}} - 1$. Fix $\delta \in (0,1)$ and suppose the consolidated dimension condition $d \geq \kappa \log\big(24T_{\text{out}}\, n_{\mathcal{S}} n_{\mathcal{T}}/\delta\big)$ for some $\kappa > 0$. There exist absolute constants $C > 0$ and $C_\kappa > 0$ such that, with probability at least $1-\delta$, simultaneously for all $i \in \mathcal{I}(\mathcal{S})$, $j \in \mathcal{I}(\mathcal{T})$, and $t_{\text{out}} \leq T_{\text{out}} - 1$,*

$$\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \boldsymbol{\mu}\rangle\big| \;\leq\; \sigma_{\mathcal{S}}\, \|\boldsymbol{\mu}\|_2 \sqrt{2\log\big(24n_{\mathcal{S}}/\delta\big)}, \quad \leq\; C_\kappa\, \sigma_{\mathcal{S}}\, \|\boldsymbol{\mu}\|_2\, \sqrt{d}. \tag{29}$$

$$\begin{aligned}\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \boldsymbol{\xi}_j\rangle\big| \;&\leq\; \sigma_{\mathcal{S}} \sqrt{2\log\big(24n_{\mathcal{S}}n_{\mathcal{T}}/\delta\big)} \\ &\qquad \times \sqrt{\operatorname{tr}(\mathbf{\Sigma}_\xi) \;+\; C\Big(\|\mathbf{\Sigma}_\xi\|_F \sqrt{\log\big(24n_{\mathcal{T}}/\delta\big)} + \|\mathbf{\Sigma}_\xi\|_{\text{op}} \log\big(24n_{\mathcal{T}}/\delta\big)\Big)} \\ &\leq\; C_\kappa\, \sigma_{\mathcal{S}} \sqrt{\|\mathbf{\Sigma}_\xi\|_{\text{op}}}\, d.\end{aligned} \tag{30}$$

$$\begin{aligned}\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \mathbf{w}^{(t_{\text{out}},0)}\rangle\big| \;&\leq\; \sigma_{\mathcal{S}} \sqrt{2\log\big(24T_{\text{out}}n_{\mathcal{S}}/\delta\big)} \\ &\qquad \times \sqrt{\sigma_{\mathbf{w}}^2\Big(d + 2\sqrt{d\,\log\big(24T_{\text{out}}/\delta\big)} + 2\log\big(24T_{\text{out}}/\delta\big)\Big)} \\ &\leq\; C_\kappa\, \sigma_{\mathcal{S}}\, \sigma_{\mathbf{w}}\, d.\end{aligned} \tag{31}$$

$$\big\|\mathbf{x}_i^{\mathcal{S},(0,0)}\big\|_2^2 \;\leq\; \sigma_{\mathcal{S}}^2\Big(d + 2\sqrt{d\,\log\big(24n_{\mathcal{S}}/\delta\big)} + 2\log\big(24n_{\mathcal{S}}/\delta\big)\Big), \quad \leq\; C_\kappa\, \sigma_{\mathcal{S}}^2\, d. \tag{32}$$

*Proof.* We control each displayed inequality on a high-probability event and then take a union bound.

For equation 29, since $\mathbf{x}_i^{\mathcal{S},(0,0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathcal{S}}^2\mathbf{I})$, we have $\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \boldsymbol{\mu}\rangle \sim \mathcal{N}(0, \sigma_{\mathcal{S}}^2\|\boldsymbol{\mu}\|_2^2)$ and hence, for any $t > 0$,

$$\mathbb{P}\Big(\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \boldsymbol{\mu}\rangle\big| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{2\sigma_{\mathcal{S}}^2\|\boldsymbol{\mu}\|_2^2}\Big).$$

Taking $t = \sigma_{\mathcal{S}}\|\boldsymbol{\mu}\|_2\sqrt{2\log(24n_{\mathcal{S}}/\delta)}$ gives

$$\mathbb{P}\Big(\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \boldsymbol{\mu}\rangle\big| \geq t\Big) \leq \frac{\delta}{12n_{\mathcal{S}}}.$$

A union bound over $i \in \mathcal{I}(\mathcal{S})$ yields the first inequality in equation 29. Moreover,

$$\sqrt{\log\big(24n_{\mathcal{S}}/\delta\big)} \leq \sqrt{\log\big(24T_{\text{out}}n_{\mathcal{S}}n_{\mathcal{T}}/\delta\big)} \leq \kappa^{-1/2}\sqrt{d},$$

so enlarging constants gives the second line of equation 29.

For equation 32, note that $\|\mathbf{x}_i^{\mathcal{S},(0,0)}\|_2^2/\sigma_{\mathcal{S}}^2 \sim \chi_d^2$. The Laurent–Massart inequality implies that for any $u > 0$,

$$\mathbb{P}\Big(\|\mathbf{x}_i^{\mathcal{S},(0,0)}\|_2^2 > \sigma_{\mathcal{S}}^2\big(d + 2\sqrt{d\,u} + 2u\big)\Big) \leq e^{-u}.$$

Taking $u = \log(24n_{\mathcal{S}}/\delta)$ gives

$$\mathbb{P}\Big(\|\mathbf{x}_i^{\mathcal{S},(0,0)}\|_2^2 > \sigma_{\mathcal{S}}^2\Big(d + 2\sqrt{d\,\log\big(24n_{\mathcal{S}}/\delta\big)} + 2\log\big(24n_{\mathcal{S}}/\delta\big)\Big)\Big) \leq \frac{\delta}{24n_{\mathcal{S}}}.$$

A union bound over $i \in \mathcal{I}(\mathcal{S})$ yields the first inequality in equation 32. Using $\log(24n_{\mathcal{S}}/\delta) \leq \kappa^{-1}d$ and $\sqrt{\log(24n_{\mathcal{S}}/\delta)} \leq \kappa^{-1/2}\sqrt{d}$ from the dimension condition, we have

$$d + 2\sqrt{d\,\log\big(24n_{\mathcal{S}}/\delta\big)} + 2\log\big(24n_{\mathcal{S}}/\delta\big) \leq \Big(1 + 2\kappa^{-1/2} + 2\kappa^{-1}\Big)d,$$

which gives the second line of equation 32 after enlarging constants.

For equation 30, we first invoke Lemma C.2 with failure probability $\delta/4$, which ensures that with probability at least $1-\delta/4$, simultaneously for all $j \in \mathcal{I}(\mathcal{T})$,

$$\|\boldsymbol{\xi}_j\|_2^2 \le \operatorname{tr}(\boldsymbol{\Sigma}_\xi) + C\Big(\|\boldsymbol{\Sigma}_\xi\|_F \sqrt{\log\big(24n_{\mathcal{T}}/\delta\big)} + \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} \log\big(24n_{\mathcal{T}}/\delta\big)\Big).$$

Condition on $\boldsymbol{\xi}_j$. Given $\boldsymbol{\xi}_j$,

$$\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \boldsymbol{\xi}_j \rangle \;\big|\; \boldsymbol{\xi}_j \sim \mathcal{N}\big(0, \sigma_{\mathcal{S}}^2 \|\boldsymbol{\xi}_j\|_2^2\big),$$

so for any $\tau > 0$,

$$\mathbb{P}\Big(\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \boldsymbol{\xi}_j \rangle\big| \ge \tau \;\big|\; \boldsymbol{\xi}_j\Big) \le 2\exp\Big(-\frac{\tau^2}{2\sigma_{\mathcal{S}}^2 \|\boldsymbol{\xi}_j\|_2^2}\Big).$$

On the event above, take

$$\tau = \sigma_{\mathcal{S}} \sqrt{2\log\big(24n_{\mathcal{S}}n_{\mathcal{T}}/\delta\big)} \sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_\xi) + C\Big(\|\boldsymbol{\Sigma}_\xi\|_F \sqrt{\log\big(24n_{\mathcal{T}}/\delta\big)} + \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} \log\big(24n_{\mathcal{T}}/\delta\big)\Big)}.$$

Then $\tau^2/(2\sigma_{\mathcal{S}}^2\|\boldsymbol{\xi}_j\|_2^2) \ge \log(24n_{\mathcal{S}}n_{\mathcal{T}}/\delta)$ and hence

$$\mathbb{P}\Big(\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \boldsymbol{\xi}_j \rangle\big| \ge \tau \;\big|\; \boldsymbol{\xi}_j\Big) \le \frac{\delta}{12n_{\mathcal{S}}n_{\mathcal{T}}}.$$

A union bound over $(i,j) \in \mathcal{I}(\mathcal{S}) \times \mathcal{I}(\mathcal{T})$ gives the first inequality in equation 30. For the coarse bound, use $\operatorname{tr}(\boldsymbol{\Sigma}_\xi) \le d\,\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}$, $\|\boldsymbol{\Sigma}_\xi\|_F \le \sqrt{d}\,\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}$, and $\sqrt{\log(24n_{\mathcal{S}}n_{\mathcal{T}}/\delta)} \le C_\kappa\sqrt{d}$, $\log(24n_{\mathcal{T}}/\delta) \le C_\kappa d$ (which follow from the dimension condition) to obtain the second line of equation 30 after enlarging constants.

For equation 31, first apply the Laurent–Massart inequality to each $t_{\mathrm{out}}$:

$$\mathbb{P}\Big(\|\mathbf{w}^{(t_{\mathrm{out}},0)}\|_2^2 > \sigma_{\mathbf{w}}^2\Big(d + 2\sqrt{d\,\log\big(24T_{\mathrm{out}}/\delta\big)} + 2\log\big(24T_{\mathrm{out}}/\delta\big)\Big)\Big) \le \frac{\delta}{24T_{\mathrm{out}}}.$$

A union bound over $t_{\mathrm{out}} = 0, \dots, T_{\mathrm{out}} - 1$ yields an event of probability at least $1 - \delta/24$ on which equation 26–type bounds hold for all $\mathbf{w}^{(t_{\mathrm{out}},0)}$. Condition on $\mathbf{w}^{(t_{\mathrm{out}},0)}$. Given $\mathbf{w}^{(t_{\mathrm{out}},0)}$,

$$\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \mathbf{w}^{(t_{\mathrm{out}},0)} \rangle \;\big|\; \mathbf{w}^{(t_{\mathrm{out}},0)} \sim \mathcal{N}\big(0, \sigma_{\mathcal{S}}^2 \|\mathbf{w}^{(t_{\mathrm{out}},0)}\|_2^2\big),$$

hence for any $\tau > 0$,

$$\mathbb{P}\Big(\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \mathbf{w}^{(t_{\mathrm{out}},0)} \rangle\big| \ge \tau \;\big|\; \mathbf{w}^{(t_{\mathrm{out}},0)}\Big) \le 2\exp\Big(-\frac{\tau^2}{2\sigma_{\mathcal{S}}^2 \|\mathbf{w}^{(t_{\mathrm{out}},0)}\|_2^2}\Big).$$

On the event controlling $\|\mathbf{w}^{(t_{\mathrm{out}},0)}\|_2^2$, take

$$\tau = \sigma_{\mathcal{S}} \sqrt{2\log\big(24T_{\mathrm{out}}n_{\mathcal{S}}/\delta\big)} \sqrt{\sigma_{\mathbf{w}}^2\Big(d + 2\sqrt{d\,\log\big(24T_{\mathrm{out}}/\delta\big)} + 2\log\big(24T_{\mathrm{out}}/\delta\big)\Big)}.$$

Then $\tau^2/(2\sigma_{\mathcal{S}}^2\|\mathbf{w}^{(t_{\mathrm{out}},0)}\|_2^2) \ge \log(24T_{\mathrm{out}}n_{\mathcal{S}}/\delta)$ and thus

$$\mathbb{P}\Big(\big|\langle \mathbf{x}_i^{\mathcal{S},(0,0)}, \mathbf{w}^{(t_{\mathrm{out}},0)} \rangle\big| \ge \tau \;\big|\; \mathbf{w}^{(t_{\mathrm{out}},0)}\Big) \le \frac{\delta}{12T_{\mathrm{out}}n_{\mathcal{S}}}.$$

A union bound over $(i, t_{\mathrm{out}}) \in \mathcal{I}(\mathcal{S}) \times \{0, \dots, T_{\mathrm{out}} - 1\}$ yields the first inequality in equation 31. The coarse bound follows from the dimension condition, which implies $\sqrt{\log(24T_{\mathrm{out}}n_{\mathcal{S}}/\delta)} \le C_\kappa\sqrt{d}$ and

$$\sqrt{d + 2\sqrt{d\,\log\big(24T_{\mathrm{out}}/\delta\big)} + 2\log\big(24T_{\mathrm{out}}/\delta\big)} \le C_\kappa\,\sqrt{d},$$

after enlarging constants.

Finally, combining the above high-probability events with a union bound yields a joint event of probability at least $1-\delta$ on which equation 29–equation 32 all hold simultaneously. □

**Lemma C.5.** *Assume the training set follows the additive model of Section 2.1 with sub-Gaussian noise parameter $\mathbf{\Sigma}_\xi$. Let $\mathbf{w}^{(t_{\text{out}},0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{w}}^2 \mathbf{I})$ be independent of $\{\boldsymbol{\xi}_i\}$ for $t_{\text{out}} = 0, \dots, T_{\text{out}} - 1$. Fix $\delta \in (0,1)$ and suppose the dimension condition*

$$d \;\geq\; \kappa \log\big(4T_{\text{out}}/\delta\big) \quad \textit{for some } \kappa > 0.$$

*Fix, in addition, a deterministic vector $\boldsymbol{a} = (a_1, \dots, a_{n_\mathcal{T}}) \in \mathbb{R}^{n_\mathcal{T}}$. There exist absolute constants $C > 0$ and $C_\kappa > 0$ such that, with probability at least $1 - \delta$, the following hold simultaneously for all $t_{\text{out}} \leq T_{\text{out}} - 1$:*

$$\begin{aligned} \Big| \sum_{i \in \mathcal{I}(\mathcal{T})} a_i \, \langle \boldsymbol{\xi}_i, \boldsymbol{\mu} \rangle \Big| \;&\leq\; \|\boldsymbol{a}\|_2 \, \big\| \mathbf{\Sigma}_\xi^{1/2} \boldsymbol{\mu} \big\|_2 \sqrt{2 \log\big(4/\delta\big)} \\ &\leq\; C_\kappa \, \|\boldsymbol{a}\|_2 \sqrt{\|\mathbf{\Sigma}_\xi\|_{\text{op}}} \, \|\boldsymbol{\mu}\|_2 \sqrt{d}. \end{aligned} \tag{33}$$

$$\begin{aligned} \Big| \sum_{i \in \mathcal{I}(\mathcal{T})} a_i \, \langle \boldsymbol{\xi}_i, \mathbf{w}^{(t_{\text{out}},0)} \rangle \Big| \;&\leq\; \|\boldsymbol{a}\|_2 \, \sigma_{\mathbf{w}} \sqrt{\|\mathbf{\Sigma}_\xi\|_{\text{op}} \Big( d + 2\sqrt{d \, \log\big(4T_{\text{out}}/\delta\big)} + 2 \log\big(4T_{\text{out}}/\delta\big) \Big)} \\ &\quad \times \sqrt{2 \log\big(4T_{\text{out}}/\delta\big)} \\ &\leq\; C_\kappa \, \|\boldsymbol{a}\|_2 \, \sigma_{\mathbf{w}} \sqrt{\|\mathbf{\Sigma}_\xi\|_{\text{op}}} \, d. \end{aligned} \tag{34}$$

$$\begin{aligned} \Big\| \sum_{i \in \mathcal{I}(\mathcal{T})} a_i \, \boldsymbol{\xi}_i \Big\|_2^2 \;&\leq\; \|\boldsymbol{a}\|_2^2 \, \operatorname{tr}(\mathbf{\Sigma}_\xi) \;+\; C \, \|\boldsymbol{a}\|_2^2 \Big( \|\mathbf{\Sigma}_\xi\|_F \sqrt{\log\big(4/\delta\big)} + \|\mathbf{\Sigma}_\xi\|_{\text{op}} \log\big(4/\delta\big) \Big) \\ &\leq\; C_\kappa \, \|\boldsymbol{a}\|_2^2 \, \|\mathbf{\Sigma}_\xi\|_{\text{op}} \, d. \end{aligned} \tag{35}$$

*Proof.* For equation 33, the variables $\langle \boldsymbol{\xi}_i, \boldsymbol{\mu} \rangle$ are independent, mean-zero, sub-Gaussian with variance proxy $\boldsymbol{\mu}^\top \mathbf{\Sigma}_\xi \boldsymbol{\mu}$. Hence, for any $t > 0$,

$$\mathbb{P}\left( \Big| \sum_i a_i \, \langle \boldsymbol{\xi}_i, \boldsymbol{\mu} \rangle \Big| \geq t \right) \leq 2 \exp\left( - \frac{t^2}{2 \, \|\boldsymbol{a}\|_2^2 \, \boldsymbol{\mu}^\top \mathbf{\Sigma}_\xi \boldsymbol{\mu}} \right).$$

Taking $t = \|\boldsymbol{a}\|_2 \|\mathbf{\Sigma}_\xi^{1/2} \boldsymbol{\mu}\|_2 \sqrt{2 \log(4/\delta)}$ gives the first line with failure probability at most $\delta/4$. Using $\boldsymbol{\mu}^\top \mathbf{\Sigma}_\xi \boldsymbol{\mu} \leq \|\mathbf{\Sigma}_\xi\|_{\text{op}} \|\boldsymbol{\mu}\|_2^2$ and $\sqrt{\log(4/\delta)} \leq \sqrt{\log(4T_{\text{out}}/\delta)} \leq C_\kappa \sqrt{d}$ by the dimension condition yields the second line.

For equation 34, condition on $\mathbf{w}^{(t_{\text{out}},0)}$ and apply a sub-Gaussian tail bound to the weighted sum. This introduces a random variance term $\mathbf{w}^{(t_{\text{out}},0)\top} \mathbf{\Sigma}_\xi \mathbf{w}^{(t_{\text{out}},0)}$, which we bound deterministically. First, note that $\mathbf{w}^\top \mathbf{\Sigma}_\xi \mathbf{w} \leq \|\mathbf{\Sigma}_\xi\|_{\text{op}} \|\mathbf{w}\|_2^2$. The norm term $\|\mathbf{w}^{(t_{\text{out}},0)}\|_2^2$ is in turn bounded with high probability for all $t_{\text{out}}$ by inequality equation 26 in Lemma C.3. By combining these two bounds and applying a union bound over $t_{\text{out}}$ for the conditional tail event, we establish the first line of equation 34 with failure probability at most $\delta/4$. The second line follows from the dimension condition, which implies $\sqrt{\log(4T_{\text{out}}/\delta)} \leq C_\kappa \sqrt{d}$.

For equation 35, write $\boldsymbol{\xi}_i = \mathbf{\Sigma}_\xi^{1/2} \mathbf{z}_i$ with independent mean-zero sub-Gaussian $\mathbf{z}_i$ and set $\mathbf{Z} := \sum_i a_i \mathbf{z}_i$. Then $\big\| \sum_i a_i \boldsymbol{\xi}_i \big\|_2^2 = \mathbf{Z}^\top \mathbf{\Sigma}_\xi \mathbf{Z}$. Hanson–Wright implies that, for any $u > 0$,

$$\mathbb{P}\Big( \mathbf{Z}^\top \mathbf{\Sigma}_\xi \mathbf{Z} > \|\boldsymbol{a}\|_2^2 \, \operatorname{tr}(\mathbf{\Sigma}_\xi) + C \, \|\boldsymbol{a}\|_2^2 \Big( \|\mathbf{\Sigma}_\xi\|_F \sqrt{u} + \|\mathbf{\Sigma}_\xi\|_{\text{op}} \, u \Big) \Big) \leq e^{-u}.$$

Taking $u = \log(4/\delta)$ yields the first line of equation 35 with failure probability at most $\delta/4$. Using $\operatorname{tr}(\mathbf{\Sigma}_\xi) \leq d \, \|\mathbf{\Sigma}_\xi\|_{\text{op}}$, $\|\mathbf{\Sigma}_\xi\|_F \leq \sqrt{d} \, \|\mathbf{\Sigma}_\xi\|_{\text{op}}$, and $\log(4/\delta) \leq C_\kappa d$ concludes the second line.

The three displayed estimates hold simultaneously with probability at least $1 - \delta$ for the fixed vector $\boldsymbol{a}$. □

# D Missing Proof

This appendix collects the proofs deferred from the main text and isolates the deterministic identities that govern the bilevel dynamics of dataset condensation.

## D.1 Notation

Throughout, the two-dimensional time index is $\mathbf{t} = (t_{\text{out}}, t_{\text{in}})$ with $t_{\text{out}} \in \{0, \dots, T_{\text{out}}-1\}$ and $t_{\text{in}} \in \{0, \dots, T_{\text{in}}\}$. In this appendix we index the condensed samples by their class labels and write $\mathbf{w}^{(t_{\text{out}},t_{\text{in}})}$ and $\mathbf{x}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})}$ for $y \in \{\pm 1\}$, while $\mathbf{w}^{\mathbf{t}}$ and $\mathbf{x}_y^{\mathcal{S},\mathbf{t}}$ may be used as shorthand when the tuple index is the main object.

For a dataset $\mathcal{A} = \{(\mathbf{x}_i^{\mathcal{A}}, y_i^{\mathcal{A}})\}_{i\in\mathcal{I}(\mathcal{A})}$ with $y_i^{\mathcal{A}} \in \{-1, +1\}$, we write $n_{\mathcal{A}} := |\mathcal{I}(\mathcal{A})|$ and define the signed samples

$$\mathbf{z}_i^{\mathcal{A}} := y_i^{\mathcal{A}} \mathbf{x}_i^{\mathcal{A}}, \qquad i \in \mathcal{I}(\mathcal{A}).$$

As in the main text, for $y \in \{\pm 1\}$ we write $\mathcal{A}_y$ for the classwise block, with index set $\mathcal{I}(\mathcal{A}_y)$ and size $n_{\mathcal{A}_y} := |\mathcal{I}(\mathcal{A}_y)|$.

For the training set, we denote the class proportions and the associated balance factor by

$$p_y := \frac{n_{\mathcal{T}_y}}{n_{\mathcal{T}}}, \qquad \beta_p := p_+^{-1/2} + p_-^{-1/2}, \qquad y \in \{\pm 1\}.$$

For the condensed set, we adopt the setting that each classwise block contains exactly one sample, i.e., $n_{\mathcal{S}_y} = 1$ for $y \in \pm 1$ (hence $n_{\mathcal{S}} = 2$). We index the condensed samples by their labels and define the time-dependent signed samples

$$\mathbf{z}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})} := y\, \mathbf{x}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})}, \qquad y \in \{\pm 1\}.$$

For the hinge loss, we use the activity indicator

$$q_{\mathcal{A},i}(\mathbf{w}) := \mathbb{1}\big\{1 - \langle \mathbf{w}, \mathbf{z}_i^{\mathcal{A}} \rangle > 0\big\}, \qquad i \in \mathcal{I}(\mathcal{A}),$$

and the regularized hinge objective

$$L_{\mathcal{A}}(\mathbf{w}) := \frac{1}{n_{\mathcal{A}}} \sum_{i\in\mathcal{I}(\mathcal{A})} \max\big(0,\, 1 - \langle \mathbf{w}, \mathbf{z}_i^{\mathcal{A}} \rangle\big) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2.$$

Along an inner-loop trajectory $\{\mathbf{w}^{(t_{\text{out}},t_{\text{in}})}\}_{t_{\text{in}}\geq 0}$, for a fixed dataset $\mathcal{A}$ we write

$$q_{\mathcal{A},i}^{(t_{\text{out}},t_{\text{in}})} := q_{\mathcal{A},i}\big(\mathbf{w}^{(t_{\text{out}},t_{\text{in}})}\big), \qquad \overline{q}_{\mathcal{A}}^{(t_{\text{out}},t_{\text{in}})} := \frac{1}{n_{\mathcal{A}}} \sum_{i\in\mathcal{I}(\mathcal{A})} q_{\mathcal{A},i}^{(t_{\text{out}},t_{\text{in}})} \in [0,1],$$

and define the activation-weighted signed mean

$$\mathbf{g}_{\mathcal{A}}^{(t_{\text{out}},t_{\text{in}})} := \frac{1}{n_{\mathcal{A}}} \sum_{i\in\mathcal{I}(\mathcal{A})} q_{\mathcal{A},i}^{(t_{\text{out}},t_{\text{in}})}\, \mathbf{z}_i^{\mathcal{A}}.$$

For the condensed set, the signed samples are time dependent and each class contains a single condensed sample. We therefore define, for each $y \in \{\pm 1\}$,

$$q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} := \mathbb{1}\Big\{1 - \big\langle \mathbf{w}^{(t_{\text{out}},t_{\text{in}})}, \mathbf{z}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})} \big\rangle > 0\Big\},$$

and set

$$\overline{q}_{\mathcal{S}_y}^{(t_{\text{out}},t_{\text{in}})} := q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} \in [0,1], \qquad \mathbf{g}_{\mathcal{S}_y}^{(t_{\text{out}},t_{\text{in}})} := q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}\, \mathbf{z}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})}.$$

We also denote the empirical signed mean of the training set by

$$\overline{\mathbf{z}}_{\mathcal{T}} := \frac{1}{n_{\mathcal{T}}} \sum_{i\in\mathcal{I}(\mathcal{T})} \mathbf{z}_i^{\mathcal{T}} = \frac{1}{n_{\mathcal{T}}} \sum_{i\in\mathcal{I}(\mathcal{T})} y_i^{\mathcal{T}} \mathbf{x}_i^{\mathcal{T}}.$$

### D.2 Assumptions

The following assumptions are used to prove the lemmas and the main theorems. We use the time index $\mathbf{t} = (t_{\text{out}}, t_{\text{in}})$ from Section D.1.

**Assumption D.1.** Fix a tail parameter $\delta \in (0,1)$ and constants $\kappa, C_\kappa > 0$. We work on the intersection of the high-probability events of Lemmas C.2–C.5 in Section C.1, so that the total failure probability is at most $\delta$. The following conditions are assumed:

**A1**. **Signal strength.**

$$\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\sqrt{d}} \;\geq\; 2\,C_\kappa.$$

**A2**. **High dimension.**

$$d \;\geq\; \kappa \log\!\left(\frac{24\,T_{\text{out}}\,n_{\mathcal{S}}\,n_{\mathcal{T}}}{\delta}\right).$$

**A3**. **Small learning rate.** The step sizes satisfy

$$0 < \eta_{\mathbf{w}}\lambda \leq 1, \qquad 0 < 2\,\eta_{\mathcal{S}} \leq 1.$$

**A4**. **Small initialization.** Denote $\rho_{\mathcal{S}} := 1 - 2\eta_{\mathcal{S}}$. The initialization scales satisfy

$$\sigma_{\mathbf{w}} < \frac{1}{(2C_\kappa+1)\|\boldsymbol{\mu}\|_2\sqrt{d}}, \qquad \sigma_{\mathcal{S}} < \min\left\{\frac{1}{C_\kappa\,\sigma_{\mathbf{w}}\,d}, \frac{\lambda}{(2C_\kappa+1)\|\boldsymbol{\mu}\|_2\sqrt{d}}, \rho_{\mathcal{S}}^{T_{\text{in}}}\sqrt{\frac{2\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}{n_{\mathcal{T}}}}\right\}.$$

Assumption D.1A1 implies $\sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\,d \leq (2C_\kappa)^{-1}\|\boldsymbol{\mu}\|_2\sqrt{d}$, and hence $C_\kappa(\|\boldsymbol{\mu}\|_2\sqrt{d} + \sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\,d) \leq (C_\kappa + \frac{1}{2})\|\boldsymbol{\mu}\|_2\sqrt{d}$. Thus the initialization scale in Assumption D.1A4 implies the mixed-scale small-margin bounds used below.

### D.3 Auxiliary Results for Lemma 3.2

The first auxiliary result gives algebraic identities for the inner-loop hinge recursion. Independently of any switching pattern of the activity indicators, the weight sequence admits an unconditional unrolling in terms of the activation-weighted signed means. On intervals where the activity indicators are fixed, the same identity reduces to an affine recursion with a closed-form solution. Fix $\rho := 1 - \lambda\eta_{\mathbf{w}}$.

**Lemma D.2.** *Fix an outer-loop index $t_{\text{out}} \in \{0, \ldots, T_{\text{out}} - 1\}$. Let $\mathcal{A} = \{(\mathbf{x}_i^{\mathcal{A}}, y_i^{\mathcal{A}})\}_{i\in\mathcal{I}(\mathcal{A})}$ be a finite dataset. Consider the inner-loop update with step size $\eta_{\mathbf{w}} > 0$,*

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}}+1)} = (1-\lambda\eta_{\mathbf{w}})\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} + \frac{\eta_{\mathbf{w}}}{n_{\mathcal{A}}}\sum_{i\in\mathcal{I}(\mathcal{A})} q_{\mathcal{A},i}\big(\mathbf{w}^{(t_{\text{out}},t_{\text{in}})}\big)\,\mathbf{z}_i^{\mathcal{A}}. \tag{36}$$

*Then the following hold.*

*(i) For every integer $t_{\text{in}} \geq 0$,*

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} = \rho^{t_{\text{in}}}\mathbf{w}^{(t_{\text{out}},0)} + \eta_{\mathbf{w}}\sum_{s=0}^{t_{\text{in}}-1}\rho^{\,t_{\text{in}}-1-s}\,\mathbf{g}_{\mathcal{A}}^{(t_{\text{out}},s)}, \tag{37}$$

*where the sum is interpreted as 0 when $t_{\text{in}} = 0$. Moreover, if there exist constants $\{q_i^\star\}_{i\in\mathcal{I}(\mathcal{A})} \subset \{0,1\}$ such that $q_{\mathcal{A},i}(\mathbf{w}^{(t_{\text{out}},s)}) = q_i^\star$ for all $i \in \mathcal{I}(\mathcal{A})$ and all $s \in \{0,1,\ldots,t_{\text{in}}-1\}$, then*

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} = (1-\lambda\eta_{\mathbf{w}})^{t_{\text{in}}}\mathbf{w}^{(t_{\text{out}},0)} + \frac{1-(1-\lambda\eta_{\mathbf{w}})^{t_{\text{in}}}}{\lambda n_{\mathcal{A}}}\sum_{i\in\mathcal{I}(\mathcal{A})} q_i^\star\,\mathbf{z}_i^{\mathcal{A}}. \tag{38}$$

*Proof.* For (i), rewrite equation 36 as

$$\mathbf{w}^{(t_{\text{out}},s+1)} = \rho\,\mathbf{w}^{(t_{\text{out}},s)} + \eta_{\mathbf{w}}\,\mathbf{g}_{\mathcal{A}}^{(t_{\text{out}},s)}, \qquad s \geq 0.$$

If $t_{\text{in}} = 0$ then equation 37 is immediate. For $t_{\text{in}} \geq 1$, multiply the display by $\rho^{\,t_{\text{in}}-1-s}$ and sum over $s = 0, \ldots, t_{\text{in}} - 1$:

$$\sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s}\mathbf{w}^{(t_{\text{out}},s+1)} = \sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-s}\mathbf{w}^{(t_{\text{out}},s)} + \eta_{\mathbf{w}} \sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s}\mathbf{g}_{\mathcal{A}}^{(t_{\text{out}},s)}.$$

Re-index the left-hand side with $u := s + 1$:

$$\sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s}\mathbf{w}^{(t_{\text{out}},s+1)} = \sum_{u=1}^{t_{\text{in}}} \rho^{\,t_{\text{in}}-u}\mathbf{w}^{(t_{\text{out}},u)}.$$

Similarly,

$$\sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-s}\mathbf{w}^{(t_{\text{out}},s)} = \rho^{t_{\text{in}}}\mathbf{w}^{(t_{\text{out}},0)} + \sum_{u=1}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-u}\mathbf{w}^{(t_{\text{out}},u)}.$$

Subtracting these two expressions yields

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} - \rho^{t_{\text{in}}}\mathbf{w}^{(t_{\text{out}},0)} = \eta_{\mathbf{w}} \sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s}\mathbf{g}_{\mathcal{A}}^{(t_{\text{out}},s)},$$

which is equation 37.

Under the constancy assumption, we have

$$\mathbf{g}_{\mathcal{A}}^{(t_{\text{out}},s)} = \frac{1}{n_{\mathcal{A}}} \sum_{i \in \mathcal{I}(\mathcal{A})} q_i^\star\,\mathbf{z}_i^{\mathcal{A}} \qquad \text{for all } s \in \{0, 1, \ldots, t_{\text{in}} - 1\}.$$

Substituting this into equation 37 gives

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} = \rho^{t_{\text{in}}}\mathbf{w}^{(t_{\text{out}},0)} + \frac{\eta_{\mathbf{w}}}{n_{\mathcal{A}}}\Big(\sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s}\Big) \sum_{i \in \mathcal{I}(\mathcal{A})} q_i^\star\,\mathbf{z}_i^{\mathcal{A}}.$$

The geometric series satisfies

$$\sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s} = \sum_{k=0}^{t_{\text{in}}-1} \rho^k = \frac{1-\rho^{t_{\text{in}}}}{1-\rho} = \frac{1-\rho^{t_{\text{in}}}}{\lambda\eta_{\mathbf{w}}}.$$

Substituting this identity yields

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} = \rho^{t_{\text{in}}}\mathbf{w}^{(t_{\text{out}},0)} + \frac{1-\rho^{t_{\text{in}}}}{\lambda n_{\mathcal{A}}} \sum_{i \in \mathcal{I}(\mathcal{A})} q_i^\star\,\mathbf{z}_i^{\mathcal{A}}.$$

Replacing $\rho$ by $1 - \lambda\eta_{\mathbf{w}}$ gives equation 38. □

The next result gives deterministic geometric bounds for the training set that will be used to certify a fully-active linear window. Under the high-probability events in Assumption D.1, it controls the initialization margins and the pairwise inner products of the signed training samples. Define

$$B_\mu := C_\kappa \sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\,\|\boldsymbol{\mu}\|_2\,\sqrt{d}, \qquad B_{\text{init}} := C_\kappa\,\sigma_{\mathbf{w}}\Big(\|\boldsymbol{\mu}\|_2\sqrt{d} + \sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,d}\Big), \qquad B_{\xi\xi} := C_\kappa\,\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,d,$$

and set

$$G_{\min} := \|\boldsymbol{\mu}\|_2^2 - 2B_\mu - B_{\xi\xi}, \qquad G_{\max} := \|\boldsymbol{\mu}\|_2^2 + 2B_\mu + B_{\xi\xi}.$$

**Lemma D.3.** *Assume Assumption D.1 and the additive model $\mathbf{x}_i^{\mathcal{T}} = y_i^{\mathcal{T}}\boldsymbol{\mu} + \boldsymbol{\xi}_i$. Then the following hold for every $(t_{\text{out}}, t_{\text{in}})$.*

*(i) The initialization margins satisfy, for every $t_{\text{out}} \in \{0, \dots, T_{\text{out}} - 1\}$ and every $i \in \mathcal{I}(\mathcal{T})$,*

$$\left|\left\langle \mathbf{w}^{(t_{\text{out}},0)}, \mathbf{z}_i^{\mathcal{T}} \right\rangle\right| \le B_{\text{init}}.$$

*(ii) The signed training samples satisfy the pairwise inner-product bounds*

$$G_{\min} \;\le\; \left\langle \mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_{i'}^{\mathcal{T}} \right\rangle \;\le\; G_{\max}, \qquad i, i' \in \mathcal{I}(\mathcal{T}).$$

*In particular, for every $i \in \mathcal{I}(\mathcal{T})$,*

$$G_{\min} \;\le\; \left\langle \overline{\mathbf{z}}_{\mathcal{T}}, \mathbf{z}_i^{\mathcal{T}} \right\rangle \;\le\; G_{\max}.$$

*Proof.* For (i), the additive model gives

$$\left|\langle \mathbf{w}^{(t_{\text{out}},0)}, \mathbf{x}_i^{\mathcal{T}} \rangle\right| \le \left|\langle \mathbf{w}^{(t_{\text{out}},0)}, \boldsymbol{\mu} \rangle\right| + \left|\langle \mathbf{w}^{(t_{\text{out}},0)}, \boldsymbol{\xi}_i \rangle\right|.$$

Lemma C.3 equation 27 and equation 28 hold simultaneously for all $t_{\text{out}}$ and $i$, hence

$$\left|\langle \mathbf{w}^{(t_{\text{out}},0)}, \mathbf{x}_i^{\mathcal{T}} \rangle\right| \le C_\kappa\, \sigma_{\mathbf{w}} \Big( \|\boldsymbol{\mu}\|_2 \sqrt{d} + \sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\, d} \Big).$$

Since $\mathbf{z}_i^{\mathcal{T}} = y_i^{\mathcal{T}} \mathbf{x}_i^{\mathcal{T}}$ and $|y_i^{\mathcal{T}}| = 1$, the same bound holds with $\mathbf{x}_i^{\mathcal{T}}$ replaced by $\mathbf{z}_i^{\mathcal{T}}$, which proves (i).

For (ii), fix $i, i' \in \mathcal{I}(\mathcal{T})$. Under the additive model, $\mathbf{z}_i^{\mathcal{T}} = \boldsymbol{\mu} + y_i^{\mathcal{T}} \boldsymbol{\xi}_i$, hence

$$\left\langle \mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_{i'}^{\mathcal{T}} \right\rangle = \|\boldsymbol{\mu}\|_2^2 + \left\langle \boldsymbol{\mu}, y_i^{\mathcal{T}} \boldsymbol{\xi}_i \right\rangle + \left\langle \boldsymbol{\mu}, y_{i'}^{\mathcal{T}} \boldsymbol{\xi}_{i'} \right\rangle + y_i^{\mathcal{T}} y_{i'}^{\mathcal{T}} \left\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \right\rangle.$$

Under Assumption D.1, Lemma C.2 equation 23 gives $\max_j |\langle \boldsymbol{\mu}, \boldsymbol{\xi}_j \rangle| \le B_\mu$. Moreover, Lemma C.2 equation 24 bounds $|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle|$ for $i \ne i'$, and Lemma C.2 equation 25 together with $\operatorname{tr}(\boldsymbol{\Sigma}_\xi) \le \|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\, d$ bounds $\|\boldsymbol{\xi}_i\|_2^2 = \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_i \rangle$. Enlarging $C_\kappa$ once if needed, we obtain $\max_{i,i' \in \mathcal{I}(\mathcal{T})} \left|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle\right| \le B_{\xi\xi}$. Combining the last displays yields

$$\left\langle \mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_{i'}^{\mathcal{T}} \right\rangle \le \|\boldsymbol{\mu}\|_2^2 + 2B_\mu + B_{\xi\xi} = G_{\max}, \qquad \left\langle \mathbf{z}_i^{\mathcal{T}}, \mathbf{z}_{i'}^{\mathcal{T}} \right\rangle \ge \|\boldsymbol{\mu}\|_2^2 - 2B_\mu - B_{\xi\xi} = G_{\min}.$$

Finally, since $\overline{\mathbf{z}}_{\mathcal{T}} = \frac{1}{n_{\mathcal{T}}} \sum_{j \in \mathcal{I}(\mathcal{T})} \mathbf{z}_j^{\mathcal{T}}$, we have

$$\left\langle \overline{\mathbf{z}}_{\mathcal{T}}, \mathbf{z}_i^{\mathcal{T}} \right\rangle = \frac{1}{n_{\mathcal{T}}} \sum_{j \in \mathcal{I}(\mathcal{T})} \left\langle \mathbf{z}_j^{\mathcal{T}}, \mathbf{z}_i^{\mathcal{T}} \right\rangle,$$

so the same two-sided bounds carry over to $\langle \overline{\mathbf{z}}_{\mathcal{T}}, \mathbf{z}_i^{\mathcal{T}} \rangle$ by averaging. □

By combining the initialization and inner-product bounds in Lemma D.3 with the closed-form dynamics in Lemma D.2, we establish the existence of an initial linear window during which all training samples remain active.

**Corollary D.4.** *Assume Assumption D.1 and the additive model $\mathbf{x}_i^{\mathcal{T}} = y_i^{\mathcal{T}}\boldsymbol{\mu} + \boldsymbol{\xi}_i$. Recall $B_{\text{init}}$ and $G_{\max}$ from Lemma D.3. Define the linear window length*

$$T_{\text{lin}} := \max \left\{ t \in \{0, 1, \dots, T_{\text{in}}\} \;\middle|\; \rho^t\, B_{\text{init}} + \frac{1 - \rho^t}{\lambda}\, G_{\max} < 1 \right\}, \tag{39}$$

*with the convention $\max \emptyset = -1$. Then the following hold.*

*(i) For every $t_{\text{in}} \in \{0, 1, \dots, T_{\text{lin}}\}$ and every $i \in \mathcal{I}(\mathcal{T})$,*

$$q_{\mathcal{T},i}\big(\mathbf{w}^{(t_{\text{out}},t_{\text{in}})}\big) = 1,$$

*and consequently*

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} = \rho^{t_{\text{in}}} \mathbf{w}^{(t_{\text{out}},0)} + \frac{1 - \rho^{t_{\text{in}}}}{\lambda}\, \overline{\mathbf{z}}_{\mathcal{T}}. \tag{40}$$

*(ii) If $G_{\max} < \lambda$, then $T_{\text{lin}} = T_{\text{in}}$, implying that full activity persists throughout the entire inner loop.*

*Proof.* For (i), we proceed by induction on $t_{\text{in}}$ to show that $q_{\mathcal{T},i}(\mathbf{w}^{(t_{\text{out}},t_{\text{in}})}) = 1$ for all $i \in \mathcal{I}(\mathcal{T})$. For the base case $t_{\text{in}} = 0$, Lemma D.3(i) ensures $|\langle \mathbf{w}^{(t_{\text{out}},0)}, \mathbf{z}_i^{\mathcal{T}} \rangle| \le B_{\text{init}}$, which is strictly less than 1 by Assumptions D.1A1 and D.1A4, implying $q_{\mathcal{T},i}(\mathbf{w}^{(t_{\text{out}},0)}) = 1$.

Assume inductively that $q_{\mathcal{T},i}(\mathbf{w}^{(t_{\text{out}},s)}) = 1$ for all $i$ and all $s < t_{\text{in}}$, with $t_{\text{in}} \le T_{\text{lin}}$. Lemma D.2(i) then yields the closed-form update $\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} = \rho^{t_{\text{in}}} \mathbf{w}^{(t_{\text{out}},0)} + \frac{1-\rho^{t_{\text{in}}}}{\lambda} \overline{\mathbf{z}}_{\mathcal{T}}$. Decomposing the margin $\langle \mathbf{w}^{(t_{\text{out}},t_{\text{in}})}, \mathbf{z}_i^{\mathcal{T}} \rangle$ and applying Lemma D.3(i) and Lemma D.3(ii) yields

$$\left\langle \mathbf{w}^{(t_{\text{out}},t_{\text{in}})}, \mathbf{z}_i^{\mathcal{T}} \right\rangle \le \rho^{t_{\text{in}}} B_{\text{init}} + \frac{1-\rho^{t_{\text{in}}}}{\lambda} G_{\max}.$$

Define the affine map $\psi(\alpha) := \alpha B_{\text{init}} + (1-\alpha) G_{\max}/\lambda$. The right-hand side equals $\psi(\rho^{t_{\text{in}}})$. Since $\psi$ is monotone on $[0,1]$, we have $\psi(\rho^{t_{\text{in}}}) \le \max\{\psi(1), \psi(\rho^{T_{\text{lin}}})\} < 1$ by Assumption D.1 and the definition equation 39. Therefore $\langle \mathbf{w}^{(t_{\text{out}},t_{\text{in}})}, \mathbf{z}_i^{\mathcal{T}} \rangle < 1$, so $q_{\mathcal{T},i}(\mathbf{w}^{(t_{\text{out}},t_{\text{in}})}) = 1$, and equation 40 holds.

For (ii), if $G_{\max} < \lambda$, then $\psi(0) = G_{\max}/\lambda < 1$ and $\psi(1) = B_{\text{init}} < 1$, hence $\psi(\alpha) < 1$ for all $\alpha \in [0,1]$. Therefore the defining inequality in equation 39 holds for all $t \in \{0, \dots, T_{\text{in}}\}$, yielding $T_{\text{lin}} = T_{\text{in}}$. □

The next result expands the inner-loop weights as an explicit linear combination of signed training samples. The coefficients encode the cumulative support-vector activity and satisfy deterministic bounds controlled by the initial fully-active window.

Fix $t_{\text{out}} \in \{0, \dots, T_{\text{out}} - 1\}$ and set $\rho := 1 - \lambda\eta_{\mathbf{w}}$. For $t_{\text{in}} \ge 0$, define the coefficient vector $\boldsymbol{a}^{(t_{\text{out}},t_{\text{in}})} \in \mathbb{R}^{n_{\mathcal{T}}}$ by

$$a_i^{(t_{\text{out}},t_{\text{in}})} := \frac{\eta_{\mathbf{w}}}{n_{\mathcal{T}}} \sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s}\, q_{\mathcal{T},i}^{(t_{\text{out}},s)}, \qquad i \in \mathcal{I}(\mathcal{T}), \tag{41}$$

with the convention that the sum is 0 when $t_{\text{in}} = 0$, and set

$$a_\mu^{(t_{\text{out}},t_{\text{in}})} := \sum_{i \in \mathcal{I}(\mathcal{T})} a_i^{(t_{\text{out}},t_{\text{in}})} = \eta_{\mathbf{w}} \sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s}\, \overline{q}_{\mathcal{T}}^{(t_{\text{out}},s)}. \tag{42}$$

**Lemma D.5.** *Assume Assumption D.1 and the additive model $\mathbf{x}_i^{\mathcal{T}} = y_i^{\mathcal{T}} \boldsymbol{\mu} + \boldsymbol{\xi}_i$. Fix $t_{\text{out}} \in \{0, \dots, T_{\text{out}} - 1\}$. Then the following hold.*

*(i) For every $t_{\text{in}} \in \{0, 1, \dots, T_{\text{in}}\}$, the inner-loop weights admit the signed-sample expansion*

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} = \rho^{t_{\text{in}}} \mathbf{w}^{(t_{\text{out}},0)} + \sum_{i \in \mathcal{I}(\mathcal{T})} a_i^{(t_{\text{out}},t_{\text{in}})}\, \mathbf{z}_i^{\mathcal{T}}, \tag{43}$$

*and hence, under the additive model,*

$$\mathbf{w}^{(t_{\text{out}},t_{\text{in}})} = \rho^{t_{\text{in}}} \mathbf{w}^{(t_{\text{out}},0)} + a_\mu^{(t_{\text{out}},t_{\text{in}})}\, \boldsymbol{\mu} + \sum_{i \in \mathcal{I}(\mathcal{T})} a_i^{(t_{\text{out}},t_{\text{in}})}\, y_i^{\mathcal{T}} \boldsymbol{\xi}_i. \tag{44}$$

*(ii) The coefficients are nonnegative. Let $T_{\text{lin}} \in \{0, 1, \dots, T_{\text{in}}\}$ be as in Corollary D.4 and assume its fully-active conclusion holds on $\{0, 1, \dots, T_{\text{lin}} - 1\}$. Define $(u)_+ := \max\{u, 0\}$. Then for every $t_{\text{in}} \in \{0, 1, \dots, T_{\text{in}}\}$ and every $i \in \mathcal{I}(\mathcal{T})$,*

$$\frac{\rho^{\,(t_{\text{in}}-T_{\text{lin}})_+}\left(1 - \rho^{\min\{t_{\text{in}},T_{\text{lin}}\}}\right)}{\lambda\, n_{\mathcal{T}}} \le a_i^{(t_{\text{out}},t_{\text{in}})} \le \frac{1-\rho^{t_{\text{in}}}}{\lambda\, n_{\mathcal{T}}}. \tag{45}$$

*Moreover,*

$$\frac{\rho^{\,(t_{\text{in}}-T_{\text{lin}})_+}\left(1 - \rho^{\min\{t_{\text{in}},T_{\text{lin}}\}}\right)}{\lambda} \le a_\mu^{(t_{\text{out}},t_{\text{in}})} \le \frac{1-\rho^{t_{\text{in}}}}{\lambda}, \tag{46}$$

*and*

$$\frac{\rho^{\,(t_{\text{in}}-T_{\text{lin}})_+}\left(1 - \rho^{\min\{t_{\text{in}},T_{\text{lin}}\}}\right)}{\lambda\sqrt{n_{\mathcal{T}}}} \le \left\|\boldsymbol{a}^{(t_{\text{out}},t_{\text{in}})}\right\|_2 \le \frac{1-\rho^{t_{\text{in}}}}{\lambda\sqrt{n_{\mathcal{T}}}}. \tag{47}$$

*Proof.* For (i), this is a direct consequence of Lemma D.2(i) applied with $\mathcal{A} = \mathcal{T}$: substitute $\mathbf{g}_{\mathcal{T}}^{(t_{\text{out}},s)} = \frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}(\mathcal{T})} q_{\mathcal{T},i}^{(t_{\text{out}},s)} \mathbf{z}_i^{\mathcal{T}}$ into equation 37, exchange the order of summation, and identify the resulting coefficients with equation 41, which gives equation 43. Summing equation 41 over $i \in \mathcal{I}(\mathcal{T})$ yields equation 42, and substituting $\mathbf{z}_i^{\mathcal{T}} = \boldsymbol{\mu} + y_i^{\mathcal{T}} \boldsymbol{\xi}_i$ gives equation 44.

For (ii), since $q_{\mathcal{T},i}^{(t_{\text{out}},s)} \in \{0,1\}$, we have $a_i^{(t_{\text{out}},t_{\text{in}})} \geq 0$ for all $i$. Under the fully-active conclusion on $\{0,\dots,T_{\text{lin}}-1\}$, we have $q_{\mathcal{T},i}^{(t_{\text{out}},s)} = 1$ for all $i$ and all $s \leq \min\{t_{\text{in}}, T_{\text{lin}}\} - 1$, hence

$$a_i^{(t_{\text{out}},t_{\text{in}})} \geq \frac{\eta_{\mathbf{w}}}{n_{\mathcal{T}}} \sum_{s=0}^{\min\{t_{\text{in}},T_{\text{lin}}\}-1} \rho^{\,t_{\text{in}}-1-s} = \frac{\eta_{\mathbf{w}}}{n_{\mathcal{T}}} \, \rho^{\,(t_{\text{in}}-T_{\text{lin}})_+} \sum_{u=0}^{\min\{t_{\text{in}},T_{\text{lin}}\}-1} \rho^u.$$

Using $\sum_{u=0}^{m-1} \rho^u = (1-\rho^m)/(1-\rho)$ and $1-\rho = \lambda \eta_{\mathbf{w}}$ yields the lower bound in equation 45. For the upper bound, use $q_{\mathcal{T},i}^{(t_{\text{out}},s)} \leq 1$ in equation 41 to obtain

$$a_i^{(t_{\text{out}},t_{\text{in}})} \leq \frac{\eta_{\mathbf{w}}}{n_{\mathcal{T}}} \sum_{s=0}^{t_{\text{in}}-1} \rho^{\,t_{\text{in}}-1-s} = \frac{1-\rho^{t_{\text{in}}}}{\lambda \, n_{\mathcal{T}}}.$$

Summing equation 45 over $i$ gives equation 46. The $\ell_2$ bounds in equation 47 follow from $\|\boldsymbol{a}\|_2 \geq \sqrt{n_{\mathcal{T}}} \min_i a_i$ and $\|\boldsymbol{a}\|_2 \leq \sqrt{n_{\mathcal{T}}} \max_i a_i$ together with equation 45. □

The corresponding expansion for the condensed samples follows from the affine gradient-matching update. It represents each evolving condensed sample as a linear combination of its initialization and the accumulated training gradients, and it gives coefficient bounds determined by the training and condensed-set activity windows. Define $\rho_{\mathcal{S}} := 1 - 2\eta_{\mathcal{S}}$.

**Lemma D.6.** *Fix an outer-loop index $t_{\text{out}} \in \{0,\dots,T_{\text{out}}-1\}$ and a class $y \in \{\pm 1\}$. Let $\mathbf{z}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})}$ denote the time-dependent signed condensed sample in class $y$, and write $q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} \in \{0,1\}$ for its activity indicator. For the training set, recall $\mathbf{z}_j^{\mathcal{T}}$ and $q_{\mathcal{T},j}^{(t_{\text{out}},t_{\text{in}})}$ for $j \in \mathcal{I}(\mathcal{T}_y)$.*

*(i) Define the coefficient sequences $\{u_y^{(t_{\text{out}},t_{\text{in}})}\}_{t_{\text{in}} \geq 0}$ and $\{c_{y,j}^{(t_{\text{out}},t_{\text{in}})}\}_{t_{\text{in}} \geq 0}$ for each $j \in \mathcal{I}(\mathcal{T}_y)$ by the recursions*

$$u_y^{(t_{\text{out}},0)} := 1, \qquad u_y^{(t_{\text{out}},t_{\text{in}}+1)} := \big(1 - 2\eta_{\mathcal{S}} \, q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}\big) \, u_y^{(t_{\text{out}},t_{\text{in}})}, \qquad t_{\text{in}} \geq 0, \tag{48}$$

*and*

$$c_{y,j}^{(t_{\text{out}},0)} := 0, \qquad c_{y,j}^{(t_{\text{out}},t_{\text{in}}+1)} := \big(1 - 2\eta_{\mathcal{S}} \, q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}\big) \, c_{y,j}^{(t_{\text{out}},t_{\text{in}})} + \frac{2\eta_{\mathcal{S}}}{n_{\mathcal{T}_y}} \, q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} \, q_{\mathcal{T},j}^{(t_{\text{out}},t_{\text{in}})}, \qquad t_{\text{in}} \geq 0. \tag{49}$$

*Set*

$$c_{y,\mu}^{(t_{\text{out}},t_{\text{in}})} := \sum_{j \in \mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\text{out}},t_{\text{in}})}, \qquad \boldsymbol{c}_y^{(t_{\text{out}},t_{\text{in}})} := \big(c_{y,j}^{(t_{\text{out}},t_{\text{in}})}\big)_{j \in \mathcal{I}(\mathcal{T}_y)}. \tag{50}$$

*Then for every $t_{\text{in}} \in \{0,1,\dots,T_{\text{in}}\}$,*

$$\mathbf{z}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})} = u_y^{(t_{\text{out}},t_{\text{in}})} \, \mathbf{z}_y^{\mathcal{S},(t_{\text{out}},0)} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\text{out}},t_{\text{in}})} \, \mathbf{z}_j^{\mathcal{T}}. \tag{51}$$

*Moreover, if the additive model $\mathbf{x}_j^{\mathcal{T}} = y_j^{\mathcal{T}} \boldsymbol{\mu} + \boldsymbol{\xi}_j$ holds, then for every $t_{\text{in}} \in \{0,1,\dots,T_{\text{in}}\}$,*

$$\mathbf{z}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})} = u_y^{(t_{\text{out}},t_{\text{in}})} \, \mathbf{z}_y^{\mathcal{S},(t_{\text{out}},0)} + c_{y,\mu}^{(t_{\text{out}},t_{\text{in}})} \, \boldsymbol{\mu} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\text{out}},t_{\text{in}})} \, y_j^{\mathcal{T}} \boldsymbol{\xi}_j. \tag{52}$$

*(ii) Let $T_{\text{lin}} \in \{0,1,\dots,T_{\text{in}}\}$ and $T_{\mathcal{S},\text{lin}}(t_{\text{out}}) \in \{0,1,\dots,T_{\text{in}}\}$ be such that*

$$q_{\mathcal{T},j}^{(t_{\text{out}},t_{\text{in}})} = 1 \quad \text{for all } j \in \mathcal{I}(\mathcal{T}_y) \text{ and all } t_{\text{in}} \in \{0,1,\dots,T_{\text{lin}}-1\}, \qquad q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} = 1 \quad \text{for all } t_{\text{in}} \in \{0,1,\dots,T_{\mathcal{S},\text{lin}}(t_{\text{out}})-1\}.$$

*Define* $T_{y,\mathrm{lin}} := \min\{T_{\mathrm{lin}},\, T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}$ *and* $(u)_+ := \max\{u, 0\}$. *Then for every* $t_{\mathrm{in}} \in \{0, 1, \ldots, T_{\mathrm{in}}\}$,

$$\rho_{\mathcal{S}}^{t_{\mathrm{in}}} \;\le\; u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \;\le\; \rho_{\mathcal{S}}^{\min\{t_{\mathrm{in}},\, T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}}, \qquad u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} = \rho_{\mathcal{S}}^{t_{\mathrm{in}}} \;\; \textit{for all } t_{\mathrm{in}} \in \{0, 1, \ldots, T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}, \tag{53}$$

*and*

$$\rho_{\mathcal{S}}^{(t_{\mathrm{in}}-T_{y,\mathrm{lin}})_+}\big(1-\rho_{\mathcal{S}}^{\min\{t_{\mathrm{in}},\, T_{y,\mathrm{lin}}\}}\big) \;\le\; c_{y,\mu}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \;\le\; 1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}}. \tag{54}$$

*Moreover, for every* $j \in \mathcal{I}(\mathcal{T}_y)$ *and every* $t_{\mathrm{in}} \in \{0, 1, \ldots, T_{\mathrm{in}}\}$,

$$\frac{\rho_{\mathcal{S}}^{(t_{\mathrm{in}}-T_{y,\mathrm{lin}})_+}\big(1-\rho_{\mathcal{S}}^{\min\{t_{\mathrm{in}},\, T_{y,\mathrm{lin}}\}}\big)}{n_{\mathcal{T}_y}} \;\le\; c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \;\le\; \frac{1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}}}{n_{\mathcal{T}_y}}, \qquad c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = \frac{1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}}}{n_{\mathcal{T}_y}} \;\; \textit{for all } t_{\mathrm{in}} \in \{0, 1, \ldots, T_{y,\mathrm{lin}}\}. \tag{55}$$

*In particular,*

$$\frac{\rho_{\mathcal{S}}^{(t_{\mathrm{in}}-T_{y,\mathrm{lin}})_+}\big(1-\rho_{\mathcal{S}}^{\min\{t_{\mathrm{in}},\, T_{y,\mathrm{lin}}\}}\big)}{\sqrt{n_{\mathcal{T}_y}}} \;\le\; \big\|\boldsymbol{c}_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\big\|_2 \;\le\; \frac{1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}}}{\sqrt{n_{\mathcal{T}_y}}}. \tag{56}$$

*Proof.* For (i), the signed condensed update in class $y$ takes the form

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}}+1)} = \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} + 2\eta_{\mathcal{S}}\, q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\Big(\mathbf{g}_{\mathcal{T}_y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} - \mathbf{g}_{\mathcal{S}_y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\Big).$$

In the one-sample-per-class setting, $\mathbf{g}_{\mathcal{S}_y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})}$, and by definition of the activation-weighted mean on the training block,

$$\mathbf{g}_{\mathcal{T}_y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = \frac{1}{n_{\mathcal{T}_y}} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} q_{\mathcal{T},j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}}.$$

Substituting both identities and using $(q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})})^2 = q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}$ yields

$$\begin{aligned}
\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}}+1)} &= \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} + \frac{2\eta_{\mathcal{S}}}{n_{\mathcal{T}_y}}\, q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} q_{\mathcal{T},j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}} - 2\eta_{\mathcal{S}}\, q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} \\
&= \big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\big)\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} + \frac{2\eta_{\mathcal{S}}}{n_{\mathcal{T}_y}}\, q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} q_{\mathcal{T},j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}}.
\end{aligned}$$

Assume that equation 51 holds at time $t_{\mathrm{in}}$, i.e., $\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} = u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} + \sum_j c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\mathbf{z}_j^{\mathcal{T}}$. Substituting this into the previous display gives

$$\begin{aligned}
\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}}+1)} &= \big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\big)u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} \\
&\quad + \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \left[\big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\big)c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} + \frac{2\eta_{\mathcal{S}}}{n_{\mathcal{T}_y}}\, q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, q_{\mathcal{T},j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\right]\mathbf{z}_j^{\mathcal{T}}.
\end{aligned}$$

The coefficient of $\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)}$ is exactly $u_y^{(t_{\mathrm{out}},t_{\mathrm{in}}+1)}$ in equation 48, and the coefficient of each $\mathbf{z}_j^{\mathcal{T}}$ is exactly $c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}}+1)}$ in equation 49. With the initialization $u_y^{(t_{\mathrm{out}},0)} = 1$ and $c_{y,j}^{(t_{\mathrm{out}},0)} = 0$, this establishes equation 51 for all $t_{\mathrm{in}} \in \{0, \ldots, T_{\mathrm{in}}\}$. Under the additive model, $\mathbf{z}_j^{\mathcal{T}} = \boldsymbol{\mu} + y_j^{\mathcal{T}}\boldsymbol{\xi}_j$, so

$$\sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}} = \Big(\sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\Big)\boldsymbol{\mu} + \sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, y_j^{\mathcal{T}}\boldsymbol{\xi}_j,$$

and equation 52 follows from equation 50.

For (ii), iterating equation 48 gives the product representation

$$u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} = \prod_{s=0}^{t_{\mathrm{in}}-1}\big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\mathrm{out}},s)}\big), \qquad t_{\mathrm{in}} \ge 1,$$

with the empty product interpreted as 1 at $t_{\text{in}} = 0$. Since $q_{\mathcal{S},y}^{(t_{\text{out}},s)} \in \{0,1\}$ and $\rho_{\mathcal{S}} = 1 - 2\eta_{\mathcal{S}}$, each factor lies in $[\rho_{\mathcal{S}}, 1]$, so

$$u_y^{(t_{\text{out}},t_{\text{in}})} \geq \rho_{\mathcal{S}}^{t_{\text{in}}}.$$

Moreover, $q_{\mathcal{S},y}^{(t_{\text{out}},s)} = 1$ for $s \in \{0, \ldots, T_{\mathcal{S},\text{lin}}(t_{\text{out}}) - 1\}$ implies that the first $\min\{t_{\text{in}}, T_{\mathcal{S},\text{lin}}(t_{\text{out}})\}$ factors equal $\rho_{\mathcal{S}}$, while the remaining factors are at most 1. Thus

$$u_y^{(t_{\text{out}},t_{\text{in}})} \leq \rho_{\mathcal{S}}^{\min\{t_{\text{in}}, T_{\mathcal{S},\text{lin}}(t_{\text{out}})\}}, \qquad u_y^{(t_{\text{out}},t_{\text{in}})} = \rho_{\mathcal{S}}^{t_{\text{in}}} \text{ for all } t_{\text{in}} \leq T_{\mathcal{S},\text{lin}}(t_{\text{out}}),$$

which is equation 53.

Next, equation 49 and the nonnegativity of the driving term imply $c_{y,j}^{(t_{\text{out}},t_{\text{in}})} \geq 0$ for all $j$ and $t_{\text{in}}$. On the joint linear window $\{0, 1, \ldots, T_{y,\text{lin}} - 1\}$, we have $q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} = 1$ and $q_{\mathcal{T},j}^{(t_{\text{out}},t_{\text{in}})} = 1$ for every $j \in \mathcal{I}(\mathcal{T}_y)$, so equation 49 reduces to

$$c_{y,j}^{(t_{\text{out}},t_{\text{in}}+1)} = \rho_{\mathcal{S}}\, c_{y,j}^{(t_{\text{out}},t_{\text{in}})} + \frac{1 - \rho_{\mathcal{S}}}{n_{\mathcal{T}_y}}, \qquad c_{y,j}^{(t_{\text{out}},0)} = 0,$$

which yields $c_{y,j}^{(t_{\text{out}},t_{\text{in}})} = (1 - \rho_{\mathcal{S}}^{t_{\text{in}}})/n_{\mathcal{T}_y}$ for all $t_{\text{in}} \leq T_{y,\text{lin}}$. Summing over $j$ gives $c_{y,\mu}^{(t_{\text{out}},t_{\text{in}})} = 1 - \rho_{\mathcal{S}}^{t_{\text{in}}}$ for $t_{\text{in}} \leq T_{y,\text{lin}}$.

For an upper bound valid for all $t_{\text{in}}$, fix $j \in \mathcal{I}(\mathcal{T}_y)$ and define

$$d_{y,j}^{(t_{\text{out}},t_{\text{in}})} := u_y^{(t_{\text{out}},t_{\text{in}})} + n_{\mathcal{T}_y}\, c_{y,j}^{(t_{\text{out}},t_{\text{in}})}.$$

Using equation 48–equation 49, for each $t_{\text{in}} \geq 0$ we have

$$\begin{aligned} d_{y,j}^{(t_{\text{out}},t_{\text{in}}+1)} &= \big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}\big) u_y^{(t_{\text{out}},t_{\text{in}})} + n_{\mathcal{T}_y}\big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}\big) c_{y,j}^{(t_{\text{out}},t_{\text{in}})} + 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} q_{\mathcal{T},j}^{(t_{\text{out}},t_{\text{in}})} \\ &= \big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}\big) d_{y,j}^{(t_{\text{out}},t_{\text{in}})} + 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} q_{\mathcal{T},j}^{(t_{\text{out}},t_{\text{in}})} \\ &\leq \big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}\big) d_{y,j}^{(t_{\text{out}},t_{\text{in}})} + 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}. \end{aligned}$$

If $q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} = 0$, the last line gives $d_{y,j}^{(t_{\text{out}},t_{\text{in}}+1)} \leq d_{y,j}^{(t_{\text{out}},t_{\text{in}})}$. If $q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} = 1$, then it reads $d_{y,j}^{(t_{\text{out}},t_{\text{in}}+1)} \leq \rho_{\mathcal{S}}\, d_{y,j}^{(t_{\text{out}},t_{\text{in}})} + (1 - \rho_{\mathcal{S}})$. Since $d_{y,j}^{(t_{\text{out}},0)} = u_y^{(t_{\text{out}},0)} = 1$, both cases imply $d_{y,j}^{(t_{\text{out}},t_{\text{in}})} \leq 1$ for all $t_{\text{in}} \in \{0, \ldots, T_{\text{in}}\}$. Therefore,

$$c_{y,j}^{(t_{\text{out}},t_{\text{in}})} \leq \frac{1 - u_y^{(t_{\text{out}},t_{\text{in}})}}{n_{\mathcal{T}_y}} \leq \frac{1 - \rho_{\mathcal{S}}^{t_{\text{in}}}}{n_{\mathcal{T}_y}},$$

which is the upper bound in equation 55. Summing over $j$ yields $c_{y,\mu}^{(t_{\text{out}},t_{\text{in}})} \leq 1 - u_y^{(t_{\text{out}},t_{\text{in}})} \leq 1 - \rho_{\mathcal{S}}^{t_{\text{in}}}$, which is the upper bound in equation 54.

For the lower bounds beyond the joint linear window, note that equation 49 gives

$$c_{y,j}^{(t_{\text{out}},t_{\text{in}}+1)} \geq \big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})}\big) c_{y,j}^{(t_{\text{out}},t_{\text{in}})} \geq \rho_{\mathcal{S}}\, c_{y,j}^{(t_{\text{out}},t_{\text{in}})}.$$

Iterating this inequality from $T_{y,\text{lin}}$ to $t_{\text{in}}$ yields, for $t_{\text{in}} \geq T_{y,\text{lin}}$,

$$c_{y,j}^{(t_{\text{out}},t_{\text{in}})} \geq \rho_{\mathcal{S}}^{t_{\text{in}} - T_{y,\text{lin}}}\, c_{y,j}^{(t_{\text{out}},T_{y,\text{lin}})} = \frac{\rho_{\mathcal{S}}^{t_{\text{in}} - T_{y,\text{lin}}}\big(1 - \rho_{\mathcal{S}}^{T_{y,\text{lin}}}\big)}{n_{\mathcal{T}_y}},$$

which matches the unified lower bound in equation 55. Summing over $j$ gives the lower bound in equation 54.

Finally, since all coordinates of $\boldsymbol{c}_y^{(t_{\text{out}},t_{\text{in}})}$ are nonnegative, we have

$$\big\|\boldsymbol{c}_y^{(t_{\text{out}},t_{\text{in}})}\big\|_2 \leq \sqrt{n_{\mathcal{T}_y}} \max_{j \in \mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\text{out}},t_{\text{in}})}, \qquad \big\|\boldsymbol{c}_y^{(t_{\text{out}},t_{\text{in}})}\big\|_2 \geq \sqrt{n_{\mathcal{T}_y}} \min_{j \in \mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\text{out}},t_{\text{in}})}.$$

Combining these relations with the coordinatewise bounds in equation 55 yields equation 56. □

Building on the coefficient expansions for the inner-loop weights (Lemma D.5) and the condensed signed sample (Lemma D.6), we next upper bound the condensed hinge margins uniformly over the inner loop. The resulting deterministic envelope yields a Corollary D.4-type sufficient condition ensuring an initial fully-active window for the condensed sample at each outer-loop restart. To account for the inheritance of $\mathcal{S}$ across outer loops, we also introduce cumulative contraction factors that control the outer-loop initial correlations against $\mathbf{w}^{(t_{\mathrm{out}},0)}$ and $\{\mathbf{z}_j^{\mathcal{T}}\}$.

**Lemma D.7.** *Assume Assumption D.1 and the additive model $\mathbf{x}_i^{\mathcal{T}} = y_i^{\mathcal{T}}\boldsymbol{\mu} + \boldsymbol{\xi}_i$. Set $\rho := 1 - \lambda\eta_{\mathbf{w}}$ and $\rho_{\mathcal{S}} := 1 - 2\eta_{\mathcal{S}}$. Define the deterministic constants*

$$\begin{aligned} B_{\mathrm{wS},0} &:= C_\kappa\,\sigma_{\mathcal{S}}\,\sigma_{\mathbf{w}}\,d, \\ G_{\mathrm{TS},0} &:= C_\kappa\,\sigma_{\mathcal{S}}\Big(\|\boldsymbol{\mu}\|_2\sqrt{d} + \sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\,d}\Big), \end{aligned} \tag{57}$$

*and recall*

$$\begin{aligned} B_{\mathrm{init}} &:= C_\kappa\,\sigma_{\mathbf{w}}\Big(\|\boldsymbol{\mu}\|_2\sqrt{d} + \sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\,d}\Big), \\ G_{\max} &:= \|\boldsymbol{\mu}\|_2^2 + 2C_\kappa\sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}}\,\|\boldsymbol{\mu}\|_2\,\sqrt{d} + C_\kappa\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\,d. \end{aligned} \tag{58}$$

*Fix $t_{\mathrm{out}} \in \{0,\dots,T_{\mathrm{out}}-1\}$ and assume that for every $k \in \{0,\dots,t_{\mathrm{out}}-1\}$,*

$$q_{\mathcal{S},y}^{(k,t_{\mathrm{in}})} = 1 \qquad \text{for all } y \in \{\pm 1\} \text{ and all } t_{\mathrm{in}} \in \{0,1,\dots,T_{\mathcal{S},\mathrm{lin}}(k)-1\},$$

*for some integers $T_{\mathcal{S},\mathrm{lin}}(k) \in \{0,1,\dots,T_{\mathrm{in}}\}$. Define the cumulative exponents*

$$\begin{aligned} \Omega_{t_{\mathrm{out}}} &:= \sum_{k=0}^{t_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k), \\ \Gamma_{t_{\mathrm{out}}} &:= \sum_{r=0}^{t_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{\sum_{k=r+1}^{t_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k)}, \end{aligned} \tag{59}$$

*with the conventions $\Omega_0 = 0$ and $\Gamma_0 = 0$. Set*

$$\begin{aligned} B_{\mathrm{wS}}^{(t_{\mathrm{out}})} &:= \rho_{\mathcal{S}}^{\Omega_{t_{\mathrm{out}}}}\,B_{\mathrm{wS},0} + \big(1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}}\big)\Gamma_{t_{\mathrm{out}}}\,B_{\mathrm{init}}, \\ G_{\mathrm{TS}}^{(t_{\mathrm{out}})} &:= \rho_{\mathcal{S}}^{\Omega_{t_{\mathrm{out}}}}\,G_{\mathrm{TS},0} + \big(1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}}\big)\Gamma_{t_{\mathrm{out}}}\,G_{\max}. \end{aligned} \tag{60}$$

*Then the following hold.*

*(i) For every $y \in \{\pm 1\}$ and every $t_{\mathrm{in}} \in \{0,1,\dots,T_{\mathrm{in}}\}$,*

$$\big\langle \mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} \big\rangle \;\le\; \Phi_{t_{\mathrm{out}}}(t_{\mathrm{in}}), \tag{61}$$

*where*

$$\begin{aligned} \Phi_{t_{\mathrm{out}}}(t_{\mathrm{in}}) := \rho^{t_{\mathrm{in}}}\,B_{\mathrm{wS}}^{(t_{\mathrm{out}})} + \frac{1-\rho^{t_{\mathrm{in}}}}{\lambda}\,G_{\mathrm{TS}}^{(t_{\mathrm{out}})} + \rho^{t_{\mathrm{in}}}\big(1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}}\big)\,B_{\mathrm{init}} \\ + \frac{(1-\rho^{t_{\mathrm{in}}})(1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}})}{\lambda}\,G_{\max}. \end{aligned} \tag{62}$$

*(ii) Define*

$$T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}}) := \max\Big\{t \in \{0,1,\dots,T_{\mathrm{in}}\} \;\Big|\; \Phi_{t_{\mathrm{out}}}(s) < 1 \text{ for all } s \in \{0,1,\dots,t\}\Big\}, \tag{63}$$

*with the convention $\max\emptyset = -1$. Then for every $y \in \{\pm 1\}$ and every $t \in \{0,1,\dots,T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}$, $q_{\mathcal{S},y}^{(t_{\mathrm{out}},t)} = 1$. Moreover, if the training-set linear-window conclusion of Corollary D.4(i) holds on $\{0,1,\dots,T_{\mathrm{lin}}-1\}$, then with $T_{y,\mathrm{lin}} := \min\{T_{\mathrm{lin}}, T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}$,*

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t)} = \rho_{\mathcal{S}}^t\,\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} + \big(1-\rho_{\mathcal{S}}^t\big)\,\overline{\mathbf{z}}_{\mathcal{T}_y}, \qquad \overline{\mathbf{z}}_{\mathcal{T}_y} := \frac{1}{n_{\mathcal{T}_y}}\sum_{j\in\mathcal{I}(\mathcal{T}_y)}\mathbf{z}_j^{\mathcal{T}}, \qquad t \in \{0,1,\dots,T_{y,\mathrm{lin}}\}. \tag{64}$$

*(iii) Define the outer-loop inheritance coefficients* $(\widetilde{u}_y^{(k)}, \widetilde{c}_{y,j}^{(k)})$ *for* $k \in \{0, 1, \dots, t_{\mathrm{out}}\}$ *and* $j \in \mathcal{I}(\mathcal{T}_y)$ *by*

$$\mathbf{z}_y^{\mathcal{S},(k,0)} = \widetilde{u}_y^{(k)}\, \mathbf{z}_y^{\mathcal{S},(0,0)} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(k)}\, \mathbf{z}_j^{\mathcal{T}}, \qquad \widetilde{u}_y^{(0)} := 1, \qquad \widetilde{c}_{y,j}^{(0)} := 0, \tag{65}$$

*and for* $k \geq 0$,

$$\widetilde{u}_y^{(k+1)} = u_y^{(k,T_{\mathrm{in}})}\, \widetilde{u}_y^{(k)}, \qquad \widetilde{c}_{y,j}^{(k+1)} = u_y^{(k,T_{\mathrm{in}})}\, \widetilde{c}_{y,j}^{(k)} + c_{y,j}^{(k,T_{\mathrm{in}})}. \tag{66}$$

*For each* $t_{\mathrm{in}} \in \{0, 1, \dots, T_{\mathrm{in}}\}$, *set*

$$\widetilde{u}_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} := u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \widetilde{u}_y^{(t_{\mathrm{out}})}, \qquad \widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} := u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \widetilde{c}_{y,j}^{(t_{\mathrm{out}})} + c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}. \tag{67}$$

*Then for every* $y \in \{\pm 1\}$ *and every* $t_{\mathrm{in}} \in \{0, 1, \dots, T_{\mathrm{in}}\}$,

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} = \widetilde{u}_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_y^{\mathcal{S},(0,0)} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}}. \tag{68}$$

*Moreover, all coefficients are nonnegative, and they satisfy the deterministic bounds*

$$\rho_{\mathcal{S}}^{t_{\mathrm{out}} T_{\mathrm{in}} + t_{\mathrm{in}}} \;\leq\; \widetilde{u}_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \;\leq\; \rho_{\mathcal{S}}^{\Omega_{t_{\mathrm{out}}} + \min\{t_{\mathrm{in}},\, T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}}. \tag{69}$$

*If, in addition, the training-set linear-window conclusion of Corollary D.4(i) holds on* $\{0, 1, \dots, T_{\mathrm{lin}} - 1\}$ *for every outer-loop restart* $k \in \{0, 1, \dots, t_{\mathrm{out}}\}$, *and we set* $T_{y,\mathrm{lin}}(k) := \min\{T_{\mathrm{lin}},\, T_{\mathcal{S},\mathrm{lin}}(k)\}$, *then with*

$$\underline{\Gamma}_{t_{\mathrm{out}},y} := \sum_{r=0}^{t_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{(t_{\mathrm{out}}-1-r)T_{\mathrm{in}}}\, \rho_{\mathcal{S}}^{(T_{\mathrm{in}} - T_{y,\mathrm{lin}}(r))_+} \Big(1 - \rho_{\mathcal{S}}^{\min\{T_{\mathrm{in}},\, T_{y,\mathrm{lin}}(r)\}}\Big), \tag{70}$$

*we have for every* $j \in \mathcal{I}(\mathcal{T}_y)$ *and every* $t_{\mathrm{in}} \in \{0, 1, \dots, T_{\mathrm{in}}\}$,

$$\begin{aligned} \widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} &\geq \frac{\rho_{\mathcal{S}}^{t_{\mathrm{in}}}\, \underline{\Gamma}_{t_{\mathrm{out}},y} + \rho_{\mathcal{S}}^{(t_{\mathrm{in}} - T_{y,\mathrm{lin}}(t_{\mathrm{out}}))_+} \Big(1 - \rho_{\mathcal{S}}^{\min\{t_{\mathrm{in}},\, T_{y,\mathrm{lin}}(t_{\mathrm{out}})\}}\Big)}{n_{\mathcal{T}_y}}, \\ \widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} &\leq \frac{\rho_{\mathcal{S}}^{\min\{t_{\mathrm{in}},\, T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}} \big(1 - \rho_{\mathcal{S}}^{T_{\mathrm{in}}}\big) \Gamma_{t_{\mathrm{out}}} + \big(1 - \rho_{\mathcal{S}}^{t_{\mathrm{in}}}\big)}{n_{\mathcal{T}_y}}. \end{aligned} \tag{71}$$

*Proof.* For every $k \geq 0$, the inheritance $\mathbf{z}_y^{\mathcal{S},(k+1,0)} = \mathbf{z}_y^{\mathcal{S},(k,T_{\mathrm{in}})}$ and Lemma D.6(i) at time $T_{\mathrm{in}}$ give

$$\mathbf{z}_y^{\mathcal{S},(k+1,0)} = u_y^{(k,T_{\mathrm{in}})}\, \mathbf{z}_y^{\mathcal{S},(k,0)} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(k,T_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}}.$$

Substituting equation 65 for $\mathbf{z}_y^{\mathcal{S},(k,0)}$ yields

$$\mathbf{z}_y^{\mathcal{S},(k+1,0)} = \big(u_y^{(k,T_{\mathrm{in}})} \widetilde{u}_y^{(k)}\big)\, \mathbf{z}_y^{\mathcal{S},(0,0)} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \big(u_y^{(k,T_{\mathrm{in}})} \widetilde{c}_{y,j}^{(k)} + c_{y,j}^{(k,T_{\mathrm{in}})}\big)\, \mathbf{z}_j^{\mathcal{T}},$$

which matches equation 66 after identifying coefficients in equation 65 at $k+1$. Lemma D.6(ii) gives $u_y^{(k,T_{\mathrm{in}})} \geq 0$ and $c_{y,j}^{(k,T_{\mathrm{in}})} \geq 0$, hence induction on $k$ implies $\widetilde{u}_y^{(k)} \geq 0$ and $\widetilde{c}_{y,j}^{(k)} \geq 0$ for all $k$ and $j$.

Iterating equation 66 gives the unrolled forms

$$\widetilde{u}_y^{(t_{\mathrm{out}})} = \prod_{k=0}^{t_{\mathrm{out}}-1} u_y^{(k,T_{\mathrm{in}})}, \qquad \widetilde{c}_{y,j}^{(t_{\mathrm{out}})} = \sum_{r=0}^{t_{\mathrm{out}}-1} \Big( \prod_{k=r+1}^{t_{\mathrm{out}}-1} u_y^{(k,T_{\mathrm{in}})} \Big)\, c_{y,j}^{(r,T_{\mathrm{in}})}.$$

Since $u_y^{(k,T_{\mathrm{in}})} = \prod_{s=0}^{T_{\mathrm{in}}-1} (1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(k,s)})$ and $1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(k,s)} \in [\rho_{\mathcal{S}}, 1]$, we have $u_y^{(k,T_{\mathrm{in}})} \geq \rho_{\mathcal{S}}^{T_{\mathrm{in}}}$ and therefore

$$\widetilde{u}_y^{(t_{\mathrm{out}})} \geq \rho_{\mathcal{S}}^{t_{\mathrm{out}} T_{\mathrm{in}}}.$$

The hypothesis on the previous outer loops and Lemma D.6(ii) give $u_y^{(k,T_{\mathrm{in}})} \le \rho_{\mathcal{S}}^{T_{\mathcal{S},\mathrm{lin}}(k)}$, hence

$$\widetilde{u}_y^{(t_{\mathrm{out}})} \le \prod_{k=0}^{t_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{T_{\mathcal{S},\mathrm{lin}}(k)} = \rho_{\mathcal{S}}^{\Omega_{t_{\mathrm{out}}}}.$$

Lemma D.6(ii) also gives $c_{y,j}^{(r,T_{\mathrm{in}})} \le (1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}})/n_{\mathcal{T}_y}$ and $\prod_{k=r+1}^{t_{\mathrm{out}}-1} u_y^{(k,T_{\mathrm{in}})} \le \rho_{\mathcal{S}}^{\sum_{k=r+1}^{t_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k)}$, hence

$$\widetilde{c}_{y,j}^{(t_{\mathrm{out}})} \le \frac{1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}}}{n_{\mathcal{T}_y}} \sum_{r=0}^{t_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{\sum_{k=r+1}^{t_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k)} = \frac{(1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}})\Gamma_{t_{\mathrm{out}}}}{n_{\mathcal{T}_y}}, \qquad \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(t_{\mathrm{out}})} \le (1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}})\Gamma_{t_{\mathrm{out}}}.$$

For (i), applying equation 65 at $k = t_{\mathrm{out}}$ and taking inner products with $\mathbf{w}^{(t_{\mathrm{out}},0)}$ gives

$$\big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} \big\rangle = \widetilde{u}_y^{(t_{\mathrm{out}})} \big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_y^{\mathcal{S},(0,0)} \big\rangle + \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(t_{\mathrm{out}})} \big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_j^{\mathcal{T}} \big\rangle.$$

Taking absolute values and using Lemma C.4 and Lemma D.3(i) yields

$$\begin{aligned}\Big|\big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} \big\rangle\Big| &\le \widetilde{u}_y^{(t_{\mathrm{out}})}\, B_{\mathrm{wS},0} + \Big(\sum_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(t_{\mathrm{out}})}\Big) B_{\mathrm{init}} \\ &\le \rho_{\mathcal{S}}^{\Omega_{t_{\mathrm{out}}}}\, B_{\mathrm{wS},0} + (1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}})\Gamma_{t_{\mathrm{out}}}\, B_{\mathrm{init}} = B_{\mathrm{wS}}^{(t_{\mathrm{out}})}.\end{aligned}$$

Similarly, for any fixed $k \in \mathcal{I}(\mathcal{T})$,

$$\big\langle \mathbf{z}_k^{\mathcal{T}}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} \big\rangle = \widetilde{u}_y^{(t_{\mathrm{out}})} \big\langle \mathbf{z}_k^{\mathcal{T}}, \mathbf{z}_y^{\mathcal{S},(0,0)} \big\rangle + \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(t_{\mathrm{out}})} \big\langle \mathbf{z}_k^{\mathcal{T}}, \mathbf{z}_j^{\mathcal{T}} \big\rangle,$$

and Lemma C.4 together with Lemma D.3(ii) yields

$$\begin{aligned}\Big|\big\langle \mathbf{z}_k^{\mathcal{T}}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} \big\rangle\Big| &\le \widetilde{u}_y^{(t_{\mathrm{out}})}\, G_{\mathrm{TS},0} + \Big(\sum_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(t_{\mathrm{out}})}\Big) G_{\max} \\ &\le \rho_{\mathcal{S}}^{\Omega_{t_{\mathrm{out}}}}\, G_{\mathrm{TS},0} + (1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}})\Gamma_{t_{\mathrm{out}}}\, G_{\max} = G_{\mathrm{TS}}^{(t_{\mathrm{out}})}.\end{aligned}$$

Fix $t_{\mathrm{in}} \in \{0,1,\dots,T_{\mathrm{in}}\}$. Lemma D.5(i) gives

$$\mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = \rho^{t_{\mathrm{in}}} \mathbf{w}^{(t_{\mathrm{out}},0)} + \sum_{k\in\mathcal{I}(\mathcal{T})} a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_k^{\mathcal{T}}, \qquad a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})} \ge 0, \qquad \sum_{k\in\mathcal{I}(\mathcal{T})} a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})} \le \frac{1-\rho^{t_{\mathrm{in}}}}{\lambda},$$

and Lemma D.6(i) gives

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} = u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} + \sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}}, \qquad 0 \le u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \le 1, \qquad c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \ge 0.$$

Expanding the inner product yields

$$\begin{aligned}\big\langle \mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} \big\rangle &= \rho^{t_{\mathrm{in}}} u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} \big\rangle + \rho^{t_{\mathrm{in}}} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_j^{\mathcal{T}} \big\rangle \\ &\quad + u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \sum_{k\in\mathcal{I}(\mathcal{T})} a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})} \big\langle \mathbf{z}_k^{\mathcal{T}}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} \big\rangle + \sum_{k\in\mathcal{I}(\mathcal{T})} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \big\langle \mathbf{z}_k^{\mathcal{T}}, \mathbf{z}_j^{\mathcal{T}} \big\rangle.\end{aligned}$$

Using $u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \le 1$ together with $a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})} \ge 0$ and $c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \ge 0$ yields

$$\begin{aligned}\big\langle \mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} \big\rangle &\le \rho^{t_{\mathrm{in}}} B_{\mathrm{wS}}^{(t_{\mathrm{out}})} + \rho^{t_{\mathrm{in}}} B_{\mathrm{init}} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} + G_{\mathrm{TS}}^{(t_{\mathrm{out}})} \sum_{k\in\mathcal{I}(\mathcal{T})} a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})} \\ &\quad + G_{\max}\Big(\sum_{k\in\mathcal{I}(\mathcal{T})} a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})}\Big)\Big(\sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\Big).\end{aligned}$$

Lemma D.6(ii) gives $\sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = c_{y,\mu}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \le 1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}}$, and Lemma D.5(i) gives $\sum_{k\in\mathcal{I}(\mathcal{T})} a_k^{(t_{\mathrm{out}},t_{\mathrm{in}})} \le (1-\rho^{t_{\mathrm{in}}})/\lambda$, hence

$$\big\langle \mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})}\big\rangle \le \rho^{t_{\mathrm{in}}} B_{\mathrm{wS}}^{(t_{\mathrm{out}})} + \rho^{t_{\mathrm{in}}}(1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}})B_{\mathrm{init}} + \frac{1-\rho^{t_{\mathrm{in}}}}{\lambda} G_{\mathrm{TS}}^{(t_{\mathrm{out}})} + \frac{(1-\rho^{t_{\mathrm{in}}})(1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}})}{\lambda} G_{\max},$$

which is equation 61–equation 62.

For (ii), if $t \le T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})$ then equation 63 gives $\Phi_{t_{\mathrm{out}}}(t) < 1$, and equation 61 implies $\langle \mathbf{w}^{(t_{\mathrm{out}},t)}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t)}\rangle < 1$, hence $q_{\mathcal{S},y}^{(t_{\mathrm{out}},t)} = 1$. For $t \in \{0,1,\dots,T_{y,\mathrm{lin}}\}$ and each $s \in \{0,1,\dots,t-1\}$, the joint activity on the corresponding linear windows implies $\mathbf{g}_{\mathcal{T}_y}^{(t_{\mathrm{out}},s)} = \overline{\mathbf{z}}_{\mathcal{T}_y}$ and $\mathbf{g}_{\mathcal{S}_y}^{(t_{\mathrm{out}},s)} = \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},s)}$, hence the condensed update reduces to

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},s+1)} = \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},s)} + 2\eta_{\mathcal{S}}\big(\overline{\mathbf{z}}_{\mathcal{T}_y} - \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},s)}\big) = \rho_{\mathcal{S}}\, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},s)} + (1-\rho_{\mathcal{S}})\, \overline{\mathbf{z}}_{\mathcal{T}_y}.$$

Iterating gives $\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t)} - \overline{\mathbf{z}}_{\mathcal{T}_y} = \rho_{\mathcal{S}}^{t}\big(\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} - \overline{\mathbf{z}}_{\mathcal{T}_y}\big)$, which is equation 64.

For (iii), Lemma D.6(i) gives

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} = u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},0)} + \sum_{j\in\mathcal{I}(\mathcal{T}_y)} c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}},$$

and substituting equation 65 at $k = t_{\mathrm{out}}$ yields

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} = \big(u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\widetilde{u}_y^{(t_{\mathrm{out}})}\big)\, \mathbf{z}_y^{\mathcal{S},(0,0)} + \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \big(u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\widetilde{c}_{y,j}^{(t_{\mathrm{out}})} + c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\big)\, \mathbf{z}_j^{\mathcal{T}},$$

which is equation 68 with coefficients defined in equation 67. The bounds in equation 69 follow by multiplying the bounds for $u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}$ from Lemma D.6(ii) with the bounds for $\widetilde{u}_y^{(t_{\mathrm{out}})}$ established above. For the upper bound in equation 71, equation 67 gives $\widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\widetilde{c}_{y,j}^{(t_{\mathrm{out}})} + c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}$, and Lemma D.6(ii) together with the bound on $\widetilde{c}_{y,j}^{(t_{\mathrm{out}})}$ derived above yields

$$\widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \le \rho_{\mathcal{S}}^{\min\{t_{\mathrm{in}},T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}} \frac{(1-\rho_{\mathcal{S}}^{T_{\mathrm{in}}})\Gamma_{t_{\mathrm{out}}}}{n_{\mathcal{T}_y}} + \frac{1-\rho_{\mathcal{S}}^{t_{\mathrm{in}}}}{n_{\mathcal{T}_y}}.$$

For the lower bound in equation 71, the unrolling of $\widetilde{c}_{y,j}^{(t_{\mathrm{out}})}$ together with Lemma D.6(ii) gives, for each $r \in \{0,1,\dots,t_{\mathrm{out}}-1\}$,

$$\prod_{k=r+1}^{t_{\mathrm{out}}-1} u_y^{(k,T_{\mathrm{in}})} \ge \rho_{\mathcal{S}}^{(t_{\mathrm{out}}-1-r)T_{\mathrm{in}}}, \qquad c_{y,j}^{(r,T_{\mathrm{in}})} \ge \frac{\rho_{\mathcal{S}}^{(T_{\mathrm{in}}-T_{y,\mathrm{lin}}(r))_+}\big(1-\rho_{\mathcal{S}}^{\min\{T_{\mathrm{in}},T_{y,\mathrm{lin}}(r)\}}\big)}{n_{\mathcal{T}_y}},$$

hence $\widetilde{c}_{y,j}^{(t_{\mathrm{out}})} \ge \underline{\Gamma}_{t_{\mathrm{out}},y}/n_{\mathcal{T}_y}$. Combining $\widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\widetilde{c}_{y,j}^{(t_{\mathrm{out}})} + c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}$ with Lemma D.6(ii), namely $u_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \ge \rho_{\mathcal{S}}^{t_{\mathrm{in}}}$ and

$$c_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \ge \frac{\rho_{\mathcal{S}}^{(t_{\mathrm{in}}-T_{y,\mathrm{lin}}(t_{\mathrm{out}}))_+}\big(1-\rho_{\mathcal{S}}^{\min\{t_{\mathrm{in}},T_{y,\mathrm{lin}}(t_{\mathrm{out}})\}}\big)}{n_{\mathcal{T}_y}},$$

yields the stated lower bound in equation 71. □

### D.4 Proof of Lemma 3.2

The following restatement records the terminal expansion and coefficient bounds used in the main text.

**Lemma D.8.** *Assume Assumption D.1 and the additive model $\mathbf{x}_i^{\mathcal{T}} = y_i^{\mathcal{T}}\boldsymbol{\mu} + \boldsymbol{\xi}_i$, and set $\rho_{\mathcal{S}} := 1-2\eta_{\mathcal{S}}$. Fix $y \in \{\pm 1\}$.*

*(i) There exist $T_{\mathrm{lin}} \in \{0,1,\dots,T_{\mathrm{in}}\}$ and $T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}}) \in \{0,1,\dots,T_{\mathrm{in}}\}$ for each $t_{\mathrm{out}} \in \{0,1,\dots,T_{\mathrm{out}}-1\}$ such that*

$$q_{\mathcal{T},i}\big(\mathbf{w}^{(t_{\mathrm{out}},t)}\big) = 1, \qquad i \in \mathcal{I}(\mathcal{T}), \qquad t \in \{0,1,\dots,T_{\mathrm{lin}}\}, \qquad t_{\mathrm{out}} \in \{0,1,\dots,T_{\mathrm{out}}-1\},$$

*and*

$$q_{\mathcal{S},y}^{(t_{\text{out}},t)} = 1, \qquad t \in \{0,1,\dots,T_{\mathcal{S},\text{lin}}(t_{\text{out}})\}, \qquad t_{\text{out}} \in \{0,1,\dots,T_{\text{out}}-1\}.$$

*(ii) Let* $\big(\widetilde{u}_y^{(t_{\text{out}},t_{\text{in}})}, \widetilde{c}_{y,j}^{(t_{\text{out}},t_{\text{in}})}\big)$ *be the coefficients from Lemma D.7(iii), and define*

$$\widetilde{c}_{y,\mu}^{(t_{\text{out}},t_{\text{in}})} := \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(t_{\text{out}},t_{\text{in}})}, \qquad \widetilde{\boldsymbol{c}}_y^{(t_{\text{out}},t_{\text{in}})} := \big(\widetilde{c}_{y,j}^{(t_{\text{out}},t_{\text{in}})}\big)_{j\in\mathcal{I}(\mathcal{T}_y)}.$$

*Then*

$$\mathbf{z}_y^{\mathcal{S},(T_{\text{out}}-1,T_{\text{in}})} = \widetilde{u}_y^{(T_{\text{out}}-1,T_{\text{in}})}\,\mathbf{z}_y^{\mathcal{S},(0,0)} + \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(T_{\text{out}}-1,T_{\text{in}})}\,\mathbf{z}_j^{\mathcal{T}},$$

*and*

$$\widetilde{u}_y^{(T_{\text{out}}-1,T_{\text{in}})} \ge 0, \qquad \widetilde{c}_{y,j}^{(T_{\text{out}}-1,T_{\text{in}})} \ge 0, \qquad j \in \mathcal{I}(\mathcal{T}_y).$$

*(iii) Define*

$$T_{y,\text{lin}}(r) := \min\big\{T_{\text{lin}}, T_{\mathcal{S},\text{lin}}(r)\big\}, \qquad r \in \{0,1,\dots,T_{\text{out}}-1\}.$$

*Define*

$$B_{\text{out}} := 1 + \sum_{r=0}^{T_{\text{out}}-2} \rho_{\mathcal{S}}^{\sum_{k=r+1}^{T_{\text{out}}-1} T_{\mathcal{S},\text{lin}}(k)}, \qquad u_{\text{out}} := \big(1-\rho_{\mathcal{S}}^{T_{\text{in}}}\big)B_{\text{out}},$$

*and*

$$\ell_y := \sum_{r=0}^{T_{\text{out}}-1} \rho_{\mathcal{S}}^{(T_{\text{out}}-r)T_{\text{in}}-T_{y,\text{lin}}(r)}\big(1-\rho_{\mathcal{S}}^{T_{y,\text{lin}}(r)}\big).$$

*Then*

$$\rho_{\mathcal{S}}^{T_{\text{out}}T_{\text{in}}} \;\le\; \widetilde{u}_y^{(T_{\text{out}}-1,T_{\text{in}})} \;\le\; \rho_{\mathcal{S}}^{\sum_{k=0}^{T_{\text{out}}-1} T_{\mathcal{S},\text{lin}}(k)}.$$

*Moreover, for every* $j \in \mathcal{I}(\mathcal{T}_y)$,

$$\frac{\ell_y}{n_{\mathcal{T}_y}} \le \widetilde{c}_{y,j}^{(T_{\text{out}}-1,T_{\text{in}})} \le \frac{u_{\text{out}}}{n_{\mathcal{T}_y}},$$

*and consequently*

$$\ell_y \le \widetilde{c}_{y,\mu}^{(T_{\text{out}}-1,T_{\text{in}})} \le u_{\text{out}}, \qquad \frac{\ell_y}{\sqrt{n_{\mathcal{T}_y}}} \le \big\|\widetilde{\boldsymbol{c}}_y^{(T_{\text{out}}-1,T_{\text{in}})}\big\|_2 \le \frac{u_{\text{out}}}{\sqrt{n_{\mathcal{T}_y}}}.$$

*Finally, set* $\Delta_y := u_{\text{out}} - \ell_y$ *and* $h_{y,j} := \widetilde{c}_{y,j}^{(T_{\text{out}}-1,T_{\text{in}})} - \ell_y/n_{\mathcal{T}_y}$. *Then* $\Delta_y \ge 0$ *and*

$$h_{y,j} \ge 0, \qquad \sum_{j\in\mathcal{I}(\mathcal{T}_y)} h_{y,j} \le \Delta_y.$$

*(iv) Moreover, assume* $G_{\max} < \lambda$. *Then*

$$q_{\mathcal{T},i}\big(\mathbf{w}^{(t_{\text{out}},t_{\text{in}})}\big) = 1, \qquad i \in \mathcal{I}(\mathcal{T}), \qquad t_{\text{out}} \in \{0,1,\dots,T_{\text{out}}-1\}, \qquad t_{\text{in}} \in \{0,1,\dots,T_{\text{in}}\},$$

*and*

$$q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} = 1, \qquad t_{\text{out}} \in \{0,1,\dots,T_{\text{out}}-1\}, \qquad t_{\text{in}} \in \{0,1,\dots,T_{\text{in}}\}.$$

*Consequently,*

$$\widetilde{u}_y^{(T_{\text{out}}-1,T_{\text{in}})} = \rho_{\mathcal{S}}^{T_{\text{out}}T_{\text{in}}}, \qquad \widetilde{c}_{y,j}^{(T_{\text{out}}-1,T_{\text{in}})} = \frac{1-\rho_{\mathcal{S}}^{T_{\text{out}}T_{\text{in}}}}{n_{\mathcal{T}_y}}, \qquad j \in \mathcal{I}(\mathcal{T}_y).$$

*Proof.* For (i), Corollary D.4(i) provides $T_{\mathrm{lin}} \in \{0, 1, \dots, T_{\mathrm{in}}\}$ such that $q_{\mathcal{T},i}\big(\mathbf{w}^{(t_{\mathrm{out}},t)}\big) = 1$ for all $i \in \mathcal{I}(\mathcal{T})$ and all $t \in \{0, 1, \dots, T_{\mathrm{lin}}\}$, uniformly over $t_{\mathrm{out}} \in \{0, 1, \dots, T_{\mathrm{out}} - 1\}$. Lemma D.7(ii) provides $T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}}) \in \{0, 1, \dots, T_{\mathrm{in}}\}$ such that $q_{\mathcal{S},y}^{(t_{\mathrm{out}},t)} = 1$ for all $t \in \{0, 1, \dots, T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}$.

For (ii), Lemma D.7(iii) yields, for every $(t_{\mathrm{out}}, t_{\mathrm{in}})$,

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} = \widetilde{u}_y^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_y^{\mathcal{S},(0,0)} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \mathbf{z}_j^{\mathcal{T}},$$

with $\widetilde{u}_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \geq 0$ and $\widetilde{c}_{y,j}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \geq 0$. Setting $(t_{\mathrm{out}}, t_{\mathrm{in}}) = (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$ gives the asserted terminal expansion and nonnegativity.

For (iii), Lemma D.7(iii) gives, for every $(t_{\mathrm{out}}, t_{\mathrm{in}})$,

$$\rho_{\mathcal{S}}^{t_{\mathrm{out}}T_{\mathrm{in}}+t_{\mathrm{in}}} \leq \widetilde{u}_y^{(t_{\mathrm{out}},t_{\mathrm{in}})} \leq \rho_{\mathcal{S}}^{\Omega_{t_{\mathrm{out}}}+\min\{t_{\mathrm{in}},T_{\mathcal{S},\mathrm{lin}}(t_{\mathrm{out}})\}}.$$

With $(t_{\mathrm{out}}, t_{\mathrm{in}}) = (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$ and $\min\{T_{\mathrm{in}}, T_{\mathcal{S},\mathrm{lin}}(T_{\mathrm{out}} - 1)\} = T_{\mathcal{S},\mathrm{lin}}(T_{\mathrm{out}} - 1)$, this yields

$$\rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}} \leq \widetilde{u}_y^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})} \leq \rho_{\mathcal{S}}^{\Omega_{T_{\mathrm{out}}-1}+T_{\mathcal{S},\mathrm{lin}}(T_{\mathrm{out}}-1)}.$$

Since $\Omega_{T_{\mathrm{out}}-1} = \sum_{k=0}^{T_{\mathrm{out}}-2} T_{\mathcal{S},\mathrm{lin}}(k)$ by equation 59, we have $\Omega_{T_{\mathrm{out}}-1} + T_{\mathcal{S},\mathrm{lin}}(T_{\mathrm{out}} - 1) = \sum_{k=0}^{T_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k)$, proving the displayed bounds for $\widetilde{u}_y^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})}$.

For the coordinatewise upper bound, apply the upper bound in equation 71 with $(t_{\mathrm{out}}, t_{\mathrm{in}}) = (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$ to obtain, for each $j \in \mathcal{I}(\mathcal{T}_y)$,

$$\widetilde{c}_{y,j}^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})} \leq \frac{\rho_{\mathcal{S}}^{T_{\mathcal{S},\mathrm{lin}}(T_{\mathrm{out}}-1)}\big(1 - \rho_{\mathcal{S}}^{T_{\mathrm{in}}}\big)\Gamma_{T_{\mathrm{out}}-1} + \big(1 - \rho_{\mathcal{S}}^{T_{\mathrm{in}}}\big)}{n_{\mathcal{T}_y}}.$$

By equation 59 with $t_{\mathrm{out}} = T_{\mathrm{out}} - 1$,

$$\Gamma_{T_{\mathrm{out}}-1} = \sum_{r=0}^{T_{\mathrm{out}}-2} \rho_{\mathcal{S}}^{\sum_{k=r+1}^{T_{\mathrm{out}}-2} T_{\mathcal{S},\mathrm{lin}}(k)}, \qquad \rho_{\mathcal{S}}^{T_{\mathcal{S},\mathrm{lin}}(T_{\mathrm{out}}-1)}\Gamma_{T_{\mathrm{out}}-1} = \sum_{r=0}^{T_{\mathrm{out}}-2} \rho_{\mathcal{S}}^{\sum_{k=r+1}^{T_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k)}.$$

Substituting this identity and factoring out $\big(1 - \rho_{\mathcal{S}}^{T_{\mathrm{in}}}\big)$ yields

$$\widetilde{c}_{y,j}^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})} \leq \frac{u_{\mathrm{out}}}{n_{\mathcal{T}_y}}.$$

For the coordinatewise lower bound, apply the lower bound in equation 71 with $(t_{\mathrm{out}}, t_{\mathrm{in}}) = (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$ and $T_{y,\mathrm{lin}}(r) \leq T_{\mathrm{in}}$ to obtain, for each $j \in \mathcal{I}(\mathcal{T}_y)$,

$$\widetilde{c}_{y,j}^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})} \geq \frac{\rho_{\mathcal{S}}^{T_{\mathrm{in}}}\, \underline{\Gamma}_{T_{\mathrm{out}}-1,y} + \rho_{\mathcal{S}}^{T_{\mathrm{in}}-T_{y,\mathrm{lin}}(T_{\mathrm{out}}-1)}\big(1 - \rho_{\mathcal{S}}^{T_{y,\mathrm{lin}}(T_{\mathrm{out}}-1)}\big)}{n_{\mathcal{T}_y}}.$$

By equation 70 with $t_{\mathrm{out}} = T_{\mathrm{out}} - 1$ and $T_{y,\mathrm{lin}}(r) \leq T_{\mathrm{in}}$,

$$\underline{\Gamma}_{T_{\mathrm{out}}-1,y} = \sum_{r=0}^{T_{\mathrm{out}}-2} \rho_{\mathcal{S}}^{(T_{\mathrm{out}}-2-r)T_{\mathrm{in}}}\, \rho_{\mathcal{S}}^{T_{\mathrm{in}}-T_{y,\mathrm{lin}}(r)}\, \big(1 - \rho_{\mathcal{S}}^{T_{y,\mathrm{lin}}(r)}\big),$$

and hence

$$\rho_{\mathcal{S}}^{T_{\mathrm{in}}}\, \underline{\Gamma}_{T_{\mathrm{out}}-1,y} = \sum_{r=0}^{T_{\mathrm{out}}-2} \rho_{\mathcal{S}}^{(T_{\mathrm{out}}-r)T_{\mathrm{in}}-T_{y,\mathrm{lin}}(r)}\, \big(1 - \rho_{\mathcal{S}}^{T_{y,\mathrm{lin}}(r)}\big).$$

The remaining term equals the $r = T_{\mathrm{out}} - 1$ summand, so combining the last two displays yields

$$\widetilde{c}_{y,j}^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})} \geq \frac{\ell_y}{n_{\mathcal{T}_y}}.$$

Summing the coordinatewise bounds over $j \in \mathcal{I}(\mathcal{T}_y)$ gives $\ell_y \le \widetilde{c}_{y,\mu}^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})} \le u_{\mathrm{out}}$.

For the $\ell_2$ bounds, fix a nonnegative vector $\boldsymbol{v} = (v_j)_{j=1}^m$ and note that

$$\|\boldsymbol{v}\|_2^2 = \sum_{j=1}^m v_j^2 \le m\big(\max_{1\le j\le m} v_j\big)^2, \qquad \|\boldsymbol{v}\|_2^2 = \sum_{j=1}^m v_j^2 \ge m\big(\min_{1\le j\le m} v_j\big)^2.$$

Applying these inequalities with $\boldsymbol{v} = \widetilde{\boldsymbol{c}}_y^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})}$ and $m = n_{\mathcal{T}_y}$ gives

$$\sqrt{n_{\mathcal{T}_y}} \min_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})} \le \big\|\widetilde{\boldsymbol{c}}_y^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})}\big\|_2 \le \sqrt{n_{\mathcal{T}_y}} \max_{j\in\mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})}.$$

The coordinatewise bounds yield the displayed upper and lower bounds on $\big\|\widetilde{\boldsymbol{c}}_y^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})}\big\|_2$. Finally, $\Delta_y \ge 0$ follows from $\ell_y \le u_{\mathrm{out}}$. The definition of $h_{y,j}$ and the coordinatewise lower bound give $h_{y,j} \ge 0$, while the total-mass upper bound gives

$$\sum_{j\in\mathcal{I}(\mathcal{T}_y)} h_{y,j} = \widetilde{c}_{y,\mu}^{(T_{\mathrm{out}}-1,T_{\mathrm{in}})} - \ell_y \le u_{\mathrm{out}} - \ell_y = \Delta_y.$$

For (iv), Corollary D.4(ii) implies $q_{\mathcal{T},i}\big(\mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\big) = 1$ for all $i \in \mathcal{I}(\mathcal{T})$ and all $(t_{\mathrm{out}}, t_{\mathrm{in}})$, so equation 40 gives

$$\mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})} = \alpha_{t_{\mathrm{in}}}\, \mathbf{w}^{(t_{\mathrm{out}},0)} + (1-\alpha_{t_{\mathrm{in}}})\, \frac{1}{\lambda}\overline{\mathbf{z}}_{\mathcal{T}}, \qquad \alpha_{t_{\mathrm{in}}} := \rho^{t_{\mathrm{in}}} \in [0,1].$$

Under full training activity, the condensed update in class $y$ satisfies

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}}+1)} = \big(1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\big)\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} + 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})}\, \overline{\mathbf{z}}_{\mathcal{T}_y}, \qquad \overline{\mathbf{z}}_{\mathcal{T}_y} := \frac{1}{n_{\mathcal{T}_y}} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \mathbf{z}_j^{\mathcal{T}}.$$

Since $q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \in \{0,1\}$ and $0 \le 2\eta_{\mathcal{S}} \le 1$, we have $1 - 2\eta_{\mathcal{S}} q_{\mathcal{S},y}^{(t_{\mathrm{out}},t_{\mathrm{in}})} \in [0,1]$, and therefore $\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}}+1)} \in \mathrm{conv}\big\{\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})}, \overline{\mathbf{z}}_{\mathcal{T}_y}\big\}$. Iterating in $t_{\mathrm{in}}$ and using the inheritance $\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}}+1,0)} = \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},T_{\mathrm{in}})}$ yields

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} \in \mathrm{conv}\big\{\mathbf{z}_y^{\mathcal{S},(0,0)}, \overline{\mathbf{z}}_{\mathcal{T}_y}\big\}, \qquad t_{\mathrm{out}} \in \{0,1,\dots,T_{\mathrm{out}}-1\}, \qquad t_{\mathrm{in}} \in \{0,1,\dots,T_{\mathrm{in}}\}.$$

Fix $(t_{\mathrm{out}}, t_{\mathrm{in}})$, and write

$$\mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})} = \beta_{t_{\mathrm{out}},t_{\mathrm{in}}}\, \mathbf{z}_y^{\mathcal{S},(0,0)} + (1-\beta_{t_{\mathrm{out}},t_{\mathrm{in}}})\, \overline{\mathbf{z}}_{\mathcal{T}_y}$$

for some $\beta_{t_{\mathrm{out}},t_{\mathrm{in}}} \in [0,1]$. Bilinearity gives

$$\begin{aligned}\big\langle \mathbf{w}^{(t_{\mathrm{out}},t_{\mathrm{in}})}, \mathbf{z}_y^{\mathcal{S},(t_{\mathrm{out}},t_{\mathrm{in}})}\big\rangle &= \alpha_{t_{\mathrm{in}}}\beta_{t_{\mathrm{out}},t_{\mathrm{in}}}\big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_y^{\mathcal{S},(0,0)}\big\rangle + \alpha_{t_{\mathrm{in}}}(1-\beta_{t_{\mathrm{out}},t_{\mathrm{in}}})\big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \overline{\mathbf{z}}_{\mathcal{T}_y}\big\rangle \\ &\quad + (1-\alpha_{t_{\mathrm{in}}})\beta_{t_{\mathrm{out}},t_{\mathrm{in}}}\Big\langle \frac{1}{\lambda}\overline{\mathbf{z}}_{\mathcal{T}}, \mathbf{z}_y^{\mathcal{S},(0,0)}\Big\rangle + (1-\alpha_{t_{\mathrm{in}}})(1-\beta_{t_{\mathrm{out}},t_{\mathrm{in}}})\Big\langle \frac{1}{\lambda}\overline{\mathbf{z}}_{\mathcal{T}}, \overline{\mathbf{z}}_{\mathcal{T}_y}\Big\rangle.\end{aligned}$$

Since the coefficients in the last display are nonnegative and sum to 1, this inner product is bounded above by the maximum of the four corner values. Under the high-probability event of Lemma C.4, we have

$$\big|\big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_y^{\mathcal{S},(0,0)}\big\rangle\big| \le C_\kappa\, \sigma_{\mathcal{S}}\, \sigma_{\mathbf{w}}\, d, \qquad \big|\big\langle \overline{\mathbf{z}}_{\mathcal{T}}, \mathbf{z}_y^{\mathcal{S},(0,0)}\big\rangle\big| \le C_\kappa\, \sigma_{\mathcal{S}}\Big(\|\boldsymbol{\mu}\|_2\sqrt{d} + \sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}}\, d\Big).$$

Assumptions D.1A1 and D.1A4 imply that the right-hand sides are bounded by 1 and $\lambda$, respectively, and therefore

$$\big|\big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_y^{\mathcal{S},(0,0)}\big\rangle\big| \le 1, \qquad \left|\Big\langle \frac{1}{\lambda}\overline{\mathbf{z}}_{\mathcal{T}}, \mathbf{z}_y^{\mathcal{S},(0,0)}\Big\rangle\right| \le 1.$$

Moreover, Lemma D.3(i) gives, for every $t_{\mathrm{out}}$,

$$\big|\big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \overline{\mathbf{z}}_{\mathcal{T}_y}\big\rangle\big| = \left|\frac{1}{n_{\mathcal{T}_y}} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_j^{\mathcal{T}}\big\rangle\right| \le \frac{1}{n_{\mathcal{T}_y}} \sum_{j\in\mathcal{I}(\mathcal{T}_y)} \big|\big\langle \mathbf{w}^{(t_{\mathrm{out}},0)}, \mathbf{z}_j^{\mathcal{T}}\big\rangle\big| \le B_{\mathrm{init}} < 1.$$

Finally, Lemma D.3(ii) implies $\langle \overline{\mathbf{z}}_{\mathcal{T}}, \overline{\mathbf{z}}_{\mathcal{T}_y} \rangle \le G_{\max}$, hence

$$\left\langle \frac{1}{\lambda}\overline{\mathbf{z}}_{\mathcal{T}}, \overline{\mathbf{z}}_{\mathcal{T}_y} \right\rangle \le \frac{G_{\max}}{\lambda} < 1.$$

Combining these bounds yields $\langle \mathbf{w}^{(t_{\text{out}},t_{\text{in}})}, \mathbf{z}_y^{\mathcal{S},(t_{\text{out}},t_{\text{in}})} \rangle < 1$ for every $(t_{\text{out}}, t_{\text{in}})$, and therefore $q_{\mathcal{S},y}^{(t_{\text{out}},t_{\text{in}})} = 1$ throughout. Under full activity, Lemma D.7(ii) gives

$$\mathbf{z}_y^{\mathcal{S},(t_{\text{out}},T_{\text{in}})} = \rho_{\mathcal{S}}^{T_{\text{in}}} \mathbf{z}_y^{\mathcal{S},(t_{\text{out}},0)} + \big(1 - \rho_{\mathcal{S}}^{T_{\text{in}}}\big)\overline{\mathbf{z}}_{\mathcal{T}_y},$$

and iterating the inheritance $\mathbf{z}_y^{\mathcal{S},(t_{\text{out}}+1,0)} = \mathbf{z}_y^{\mathcal{S},(t_{\text{out}},T_{\text{in}})}$ yields

$$\mathbf{z}_y^{\mathcal{S},(T_{\text{out}}-1,T_{\text{in}})} = \rho_{\mathcal{S}}^{T_{\text{out}}T_{\text{in}}} \mathbf{z}_y^{\mathcal{S},(0,0)} + \big(1 - \rho_{\mathcal{S}}^{T_{\text{out}}T_{\text{in}}}\big)\overline{\mathbf{z}}_{\mathcal{T}_y}.$$

Expanding $\overline{\mathbf{z}}_{\mathcal{T}_y} = \frac{1}{n_{\mathcal{T}_y}} \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \mathbf{z}_j^{\mathcal{T}}$ gives the coefficient identities in (iv). □

### D.5 Proof of Lemma 3.7

The proof of Lemma 3.7 is organized into one geometric risk bound and two signal-to-residual corollaries. The first result reduces the population risk of a directional evaluator to a one-dimensional sub-Gaussian tail bound and then lower bounds the effective margin using only a signal–residual decomposition of the two-sample difference vector.

**Lemma D.9.** *Assume the test distribution $\mathcal{D}$ follows the additive model*

$$\mathbf{x} = y\boldsymbol{\mu} + \boldsymbol{\xi}_{\text{test}}, \qquad y \in \{\pm 1\},$$

*where $\boldsymbol{\xi}_{\text{test}}$ is mean-zero sub-Gaussian with proxy covariance $\boldsymbol{\Sigma}_\xi$ in the sense of Definition C.1. Let $\mathcal{S} = \{(\mathbf{x}_+, +1), (\mathbf{x}_-, -1)\}$ be any two-point dataset and define the two-sample difference vector*

$$\mathbf{s}(\mathcal{S}) := \mathbf{x}_+ - \mathbf{x}_-, \qquad \widehat{\mathbf{w}}_{\mathcal{S}} := \frac{\mathbf{s}(\mathcal{S})}{\|\mathbf{s}(\mathcal{S})\|_2}.$$

*Define the effective margin*

$$m(\mathcal{S}) := \langle \widehat{\mathbf{w}}_{\mathcal{S}}, \boldsymbol{\mu} \rangle.$$

*Then the population classification risk satisfies*

$$\mathcal{R}(\mathcal{S}) := \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}(y\langle \widehat{\mathbf{w}}_{\mathcal{S}}, \mathbf{x} \rangle < 0) \le \exp\left(-\frac{m(\mathcal{S})_+^2}{2\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\right). \tag{72}$$

*Moreover, assume that $\mathbf{s}(\mathcal{S})$ admits the decomposition*

$$\mathbf{s}(\mathcal{S}) = A(\mathcal{S})\,\boldsymbol{\mu} + \mathbf{r}(\mathcal{S}) \qquad \text{with } A(\mathcal{S}) > 0. \tag{73}$$

*Define the relative signal-to-residual ratio*

$$\widehat{\text{SR}}(\mathcal{S}) := \frac{A(\mathcal{S})\,\|\boldsymbol{\mu}\|_2}{\|\mathbf{r}(\mathcal{S})\|_2} \in (0, +\infty], \tag{74}$$

*where $\widehat{\text{SR}}(\mathcal{S}) = +\infty$ if $\mathbf{r}(\mathcal{S}) = \mathbf{0}$. Then*

$$m(\mathcal{S}) \ge \|\boldsymbol{\mu}\|_2 \left(\frac{\widehat{\text{SR}}(\mathcal{S}) - 1}{\widehat{\text{SR}}(\mathcal{S}) + 1}\right), \tag{75}$$

*and therefore*

$$\mathcal{R}(\mathcal{S}) \le \exp\left(-\frac{\|\boldsymbol{\mu}\|_2^2}{2\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}} \left(\frac{\widehat{\text{SR}}(\mathcal{S}) - 1}{\widehat{\text{SR}}(\mathcal{S}) + 1}\right)_+^2\right). \tag{76}$$

*Proof.* Fix $\mathcal{S}$ and abbreviate $\widehat{\mathbf{w}} := \widehat{\mathbf{w}}_{\mathcal{S}}$. For $(\mathbf{x}, y) \sim \mathcal{D}$, the error event is

$$y\langle \widehat{\mathbf{w}}, \mathbf{x}\rangle < 0 \quad \Longleftrightarrow \quad \langle \widehat{\mathbf{w}}, \boldsymbol{\mu}\rangle + \langle \widehat{\mathbf{w}}, y\boldsymbol{\xi}_{\text{test}}\rangle < 0.$$

Since $\boldsymbol{\xi}_{\text{test}}$ is sub-Gaussian with proxy covariance $\boldsymbol{\Sigma}_\xi$, the scalar $\langle \widehat{\mathbf{w}}, y\boldsymbol{\xi}_{\text{test}}\rangle$ is mean-zero sub-Gaussian with proxy variance $\widehat{\mathbf{w}}^\top \boldsymbol{\Sigma}_\xi \widehat{\mathbf{w}} \le \|\boldsymbol{\Sigma}_\xi\|_{\text{op}}$. Hence, for any $t > 0$,

$$\mathbb{P}(\langle \widehat{\mathbf{w}}, y\boldsymbol{\xi}_{\text{test}}\rangle \le -t) \le \exp\!\left(-\frac{t^2}{2\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\right).$$

Taking $t = m(\mathcal{S})_+ = \langle \widehat{\mathbf{w}}, \boldsymbol{\mu}\rangle_+$ yields equation 72.

Assume now that equation 73 holds. Using $\widehat{\mathbf{w}} = \mathbf{s}(\mathcal{S})/\|\mathbf{s}(\mathcal{S})\|_2$,

$$m(\mathcal{S}) = \frac{\langle \mathbf{s}(\mathcal{S}), \boldsymbol{\mu}\rangle}{\|\mathbf{s}(\mathcal{S})\|_2}.$$

For the numerator, Cauchy–Schwarz yields

$$\langle \mathbf{s}(\mathcal{S}), \boldsymbol{\mu}\rangle = A(\mathcal{S})\|\boldsymbol{\mu}\|_2^2 + \langle \mathbf{r}(\mathcal{S}), \boldsymbol{\mu}\rangle \ge A(\mathcal{S})\|\boldsymbol{\mu}\|_2^2 - \|\mathbf{r}(\mathcal{S})\|_2\,\|\boldsymbol{\mu}\|_2.$$

For the denominator, the triangle inequality yields

$$\|\mathbf{s}(\mathcal{S})\|_2 = \|A(\mathcal{S})\boldsymbol{\mu} + \mathbf{r}(\mathcal{S})\|_2 \le A(\mathcal{S})\|\boldsymbol{\mu}\|_2 + \|\mathbf{r}(\mathcal{S})\|_2.$$

Dividing the last two displays gives

$$m(\mathcal{S}) \ge \|\boldsymbol{\mu}\|_2\,\frac{A(\mathcal{S})\|\boldsymbol{\mu}\|_2 - \|\mathbf{r}(\mathcal{S})\|_2}{A(\mathcal{S})\|\boldsymbol{\mu}\|_2 + \|\mathbf{r}(\mathcal{S})\|_2}.$$

Substituting the definition equation 74 yields equation 75. Finally, substituting equation 75 into equation 72 yields equation 76. □

We next instantiate Lemma D.9 for a coreset formed by selecting one training sample per class.

**Corollary D.10.** *Assume Assumption D.1 and the additive model $\mathbf{x}_i^{\mathcal{T}} = y_i^{\mathcal{T}}\boldsymbol{\mu} + \boldsymbol{\xi}_i$ with mean-zero sub-Gaussian noise parameter $\boldsymbol{\Sigma}_\xi$. Let $C_\kappa$ denote the constant appearing in Lemma C.2. Pick any indices $i_+ \in \mathcal{I}(\mathcal{T}_+)$ and $i_- \in \mathcal{I}(\mathcal{T}_-)$ and form $\mathcal{S}_{\text{core}} = \{(\mathbf{x}_{i_+}^{\mathcal{T}}, +1), (\mathbf{x}_{i_-}^{\mathcal{T}}, -1)\}$. Then*

$$\mathbf{s}(\mathcal{S}_{\text{core}}) = \mathbf{x}_{i_+}^{\mathcal{T}} - \mathbf{x}_{i_-}^{\mathcal{T}} = 2\boldsymbol{\mu} + \mathbf{r}_{\text{core}}, \qquad \mathbf{r}_{\text{core}} := \boldsymbol{\xi}_{i_+} - \boldsymbol{\xi}_{i_-},$$

*and*

$$\|\mathbf{r}_{\text{core}}\|_2^2 \le 2\big(1 + 2C_\kappa\big)\,\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,d. \tag{77}$$

*Define*

$$\widehat{\text{SR}}(\mathcal{S}_{\text{core}}) := \frac{2\|\boldsymbol{\mu}\|_2}{\|\mathbf{r}_{\text{core}}\|_2}.$$

*Then equation 77 implies the interpretable lower bound*

$$\widehat{\text{SR}}(\mathcal{S}_{\text{core}}) \;\ge\; \frac{2\|\boldsymbol{\mu}\|_2}{\sqrt{2(1+2C_\kappa)\,\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}\,d}}. \tag{78}$$

*Consequently,*

$$\mathcal{R}(\mathcal{S}_{\text{core}}) \le \exp\left(-\frac{\|\boldsymbol{\mu}\|_2^2}{2\|\boldsymbol{\Sigma}_\xi\|_{\text{op}}}\left(\frac{\widehat{\text{SR}}(\mathcal{S}_{\text{core}}) - 1}{\widehat{\text{SR}}(\mathcal{S}_{\text{core}}) + 1}\right)_+^2\right),$$

*and the same bound remains valid after replacing $\widehat{\text{SR}}(\mathcal{S}_{\text{core}})$ by the lower bound in equation 78.*

*Proof.* The identity $\mathbf{s}(\mathcal{S}_{\mathrm{core}}) = \mathbf{x}_{i_+}^{\mathcal{T}} - \mathbf{x}_{i_-}^{\mathcal{T}} = 2\boldsymbol{\mu} + (\boldsymbol{\xi}_{i_+} - \boldsymbol{\xi}_{i_-})$ follows from the additive model.

By Lemma C.2 equation 25 and $\mathrm{tr}(\boldsymbol{\Sigma}_\xi) \le d\,\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}$, for each $i \in \mathcal{I}(\mathcal{T})$,

$$\|\boldsymbol{\xi}_i\|_2^2 \le \mathrm{tr}(\boldsymbol{\Sigma}_\xi) + C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d \le \big(1 + C_\kappa\big)\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d.$$

By Lemma C.2 equation 24, for $i \neq i'$,

$$\big|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle\big| \le C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d.$$

Therefore,

$$\begin{aligned}
\|\mathbf{r}_{\mathrm{core}}\|_2^2 &= \|\boldsymbol{\xi}_{i_+} - \boldsymbol{\xi}_{i_-}\|_2^2 \\
&= \|\boldsymbol{\xi}_{i_+}\|_2^2 + \|\boldsymbol{\xi}_{i_-}\|_2^2 - 2\langle \boldsymbol{\xi}_{i_+}, \boldsymbol{\xi}_{i_-} \rangle \\
&\le \big(1 + C_\kappa\big)\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d + \big(1 + C_\kappa\big)\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d + 2C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d \\
&= 2\big(1 + 2C_\kappa\big)\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d,
\end{aligned}$$

which is equation 77. The lower bound equation 78 follows from $\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{core}}) = 2\|\boldsymbol{\mu}\|_2 / \|\mathbf{r}_{\mathrm{core}}\|_2$. The risk bound is Lemma D.9 applied with $A(\mathcal{S}_{\mathrm{core}}) = 2$ and $\mathbf{r}(\mathcal{S}_{\mathrm{core}}) = \mathbf{r}_{\mathrm{core}}$. □

We finally apply Lemma D.9 to the terminal condensed samples. The proof uses the terminal expansion and the coefficient bounds established in Lemma D.8. The strong-regularization regime $G_{\max} < \lambda$ is treated as a special case at the end.

**Corollary D.11.** *Assume the hypotheses of Lemma D.8. Let $C_\kappa$ denote a constant that can be chosen to dominate the constants appearing in Lemma C.2, Lemma C.4, and Lemma C.5. For $y \in \{\pm 1\}$, set $\widetilde{u}_y := \widetilde{u}_y^{(T_{\mathrm{out}}-1, T_{\mathrm{in}})}$, $\widetilde{\boldsymbol{c}}_y := \widetilde{\boldsymbol{c}}_y^{(T_{\mathrm{out}}-1, T_{\mathrm{in}})}$, and $\widetilde{c}_{y,\mu} := \widetilde{c}_{y,\mu}^{(T_{\mathrm{out}}-1, T_{\mathrm{in}})}$. Let $\mathbf{z}_y^{\mathcal{S}} := \mathbf{z}_y^{\mathcal{S},(T_{\mathrm{out}}-1, T_{\mathrm{in}})}$ be the terminal signed condensed samples and define $\mathcal{S}_{\mathrm{dc}} = \{(\mathbf{z}_+^{\mathcal{S}}, +1), (-\mathbf{z}_-^{\mathcal{S}}, -1)\}$. Define the two-sample difference vector and its signal–residual decomposition by*

$$\mathbf{s}(\mathcal{S}_{\mathrm{dc}}) := \mathbf{z}_+^{\mathcal{S}} - (-\mathbf{z}_-^{\mathcal{S}}) = \mathbf{z}_+^{\mathcal{S}} + \mathbf{z}_-^{\mathcal{S}} = A_{\mathrm{dc}}\boldsymbol{\mu} + \mathbf{r}_{\mathrm{dc}}, \qquad A_{\mathrm{dc}} := \widetilde{c}_{+,\mu} + \widetilde{c}_{-,\mu}.$$

*Define the relative signal-to-residual ratio*

$$\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{dc}}) := \frac{A_{\mathrm{dc}}\|\boldsymbol{\mu}\|_2}{\|\mathbf{r}_{\mathrm{dc}}\|_2}.$$

*For $y \in \{\pm 1\}$ and $r \in \{0, \dots, T_{\mathrm{out}} - 1\}$, set $T_{y,\mathrm{lin}}(r) := \min\{T_{\mathrm{lin}}, T_{\mathcal{S},\mathrm{lin}}(r)\}$, and define*

$$B_{\mathrm{out}} := 1 + \sum_{r=0}^{T_{\mathrm{out}}-2} \rho_{\mathcal{S}}^{\sum_{k=r+1}^{T_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k)}.$$

*Define $u_{\mathrm{out}} := \big(1 - \rho_{\mathcal{S}}^{T_{\mathrm{in}}}\big) B_{\mathrm{out}}$. For $y \in \{\pm 1\}$, define*

$$\ell_y := \sum_{r=0}^{T_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{(T_{\mathrm{out}}-r)T_{\mathrm{in}} - T_{y,\mathrm{lin}}(r)} \big(1 - \rho_{\mathcal{S}}^{T_{y,\mathrm{lin}}(r)}\big).$$

*Set $\Delta_y := u_{\mathrm{out}} - \ell_y$. Set*

$$L_{\mathrm{dc}} := \frac{\ell_+ + \ell_-}{B_{\mathrm{out}}\beta_p + \sum_{y \in \{\pm 1\}} \Delta_y \big(\sqrt{n\tau} - p_y^{-1/2}\big)}.$$

*Then the population risk satisfies*

$$\mathcal{R}(\mathcal{S}_{\mathrm{dc}}) \le \exp\left( -\frac{\|\boldsymbol{\mu}\|_2^2}{2\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}} \left( \frac{\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{dc}}) - 1}{\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{dc}}) + 1} \right)_+^2 \right).$$

*Moreover,*

$$\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{dc}}) \;\geq\; \frac{\sqrt{n_{\mathcal{T}}}\, L_{\mathrm{dc}}\, \|\boldsymbol{\mu}\|_2}{\sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\, d}},$$

*and the same risk bound remains valid after replacing* $\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{dc}})$ *by the lower bound above.*

*Moreover, if* $G_{\max} < \lambda$*, then we have*

$$\widetilde{u}_y = \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}, \qquad \widetilde{c}_{y,\mu} = 1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}, \qquad \|\widetilde{\boldsymbol{c}}_y\|_2 = \frac{1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}}{\sqrt{n_{\mathcal{T}_y}}}, \qquad y \in \{\pm 1\},$$

*and hence*

$$L_{\mathrm{dc}} = \frac{2\big(1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}\big)}{\beta_p \sum_{q=0}^{T_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{q T_{\mathrm{in}}}}, \qquad \Delta_y = 0, \qquad y \in \{\pm 1\},$$

*so the stated lower bound becomes*

$$\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{dc}}) \;\geq\; \frac{2\big(1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}\big)\, \|\boldsymbol{\mu}\|_2}{\sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\, d}\, \Big(\sum_{q=0}^{T_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{q T_{\mathrm{in}}}\Big)\, \frac{\beta_p}{\sqrt{n_{\mathcal{T}}}}}.$$

*Proof.* By Lemma D.8(ii), for each $y \in \{\pm 1\}$,

$$\mathbf{z}_y^{\mathcal{S}} = \widetilde{u}_y\, \mathbf{z}_y^{\mathcal{S},(0,0)} + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}\, \mathbf{z}_j^{\mathcal{T}}.$$

Since $\mathbf{z}_j^{\mathcal{T}} := y_j^{\mathcal{T}}\, \mathbf{x}_j^{\mathcal{T}}$ and $\mathbf{x}_j^{\mathcal{T}} = y_j^{\mathcal{T}}\, \boldsymbol{\mu} + \boldsymbol{\xi}_j$, we have $\mathbf{z}_j^{\mathcal{T}} = \boldsymbol{\mu} + y_j^{\mathcal{T}}\, \boldsymbol{\xi}_j$, and for $j \in \mathcal{I}(\mathcal{T}_y)$, $y_j^{\mathcal{T}} = y$. Therefore,

$$\sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}\, \mathbf{z}_j^{\mathcal{T}} = \widetilde{c}_{y,\mu}\, \boldsymbol{\mu} + y \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}\, \boldsymbol{\xi}_j, \qquad \widetilde{c}_{y,\mu} := \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j}.$$

Summing $y = +1$ and $y = -1$ gives

$$\mathbf{s}(\mathcal{S}_{\mathrm{dc}}) = \mathbf{z}_+^{\mathcal{S}} + \mathbf{z}_-^{\mathcal{S}} = (\widetilde{c}_{+,\mu} + \widetilde{c}_{-,\mu})\, \boldsymbol{\mu} + \mathbf{r}_{\mathrm{init}} + \mathbf{r}_{\mathrm{noise}},$$

where

$$\mathbf{r}_{\mathrm{init}} := \widetilde{u}_+\, \mathbf{x}_+^{\mathcal{S},(0,0)} + \widetilde{u}_-\, \mathbf{x}_-^{\mathcal{S},(0,0)}, \qquad \mathbf{r}_{\mathrm{noise}} := \sum_{j \in \mathcal{I}(\mathcal{T}_+)} \widetilde{c}_{+,j} \boldsymbol{\xi}_j - \sum_{j \in \mathcal{I}(\mathcal{T}_-)} \widetilde{c}_{-,j} \boldsymbol{\xi}_j.$$

Thus $\mathbf{r}_{\mathrm{dc}} = \mathbf{r}_{\mathrm{init}} + \mathbf{r}_{\mathrm{noise}}$ and $A_{\mathrm{dc}} = \widetilde{c}_{+,\mu} + \widetilde{c}_{-,\mu}$. The risk bound is Lemma D.9 applied with $A(\mathcal{S}_{\mathrm{dc}}) = A_{\mathrm{dc}}$ and $\mathbf{r}(\mathcal{S}_{\mathrm{dc}}) = \mathbf{r}_{\mathrm{dc}}$.

By Lemma C.4 equation 32, $\|\mathbf{z}_y^{\mathcal{S},(0,0)}\|_2^2 \leq C_\kappa \sigma_{\mathcal{S}}^2 d$, hence

$$\|\mathbf{r}_{\mathrm{init}}\|_2 \leq \sqrt{C_\kappa d}\, \sigma_{\mathcal{S}}\, (\widetilde{u}_+ + \widetilde{u}_-).$$

For each $y \in \{\pm 1\}$, Lemma D.8(iii) gives the decomposition $\widetilde{c}_{y,j} = \ell_y / n_{\mathcal{T}_y} + h_{y,j}$ with $h_{y,j} \geq 0$ and

$$\sum_{j \in \mathcal{I}(\mathcal{T}_y)} h_{y,j} \leq \Delta_y.$$

Therefore

$$\sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j} \boldsymbol{\xi}_j = \frac{\ell_y}{n_{\mathcal{T}_y}} \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \boldsymbol{\xi}_j + \sum_{j \in \mathcal{I}(\mathcal{T}_y)} h_{y,j} \boldsymbol{\xi}_j.$$

For the first term, Lemma C.5 equation 35, applied with the failure probability split between the two fixed classwise-uniform vectors, yields simultaneously for $y \in \{\pm 1\}$,

$$\Big\| \frac{\ell_y}{n_{\mathcal{T}_y}} \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \boldsymbol{\xi}_j \Big\|_2 \leq \sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}\, d}\, \frac{\ell_y}{\sqrt{n_{\mathcal{T}_y}}}.$$

For the second term, Lemma C.2 equation 25 and $\operatorname{tr}(\mathbf{\Sigma}_\xi) \le \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d$ imply $\max_j \|\boldsymbol{\xi}_j\|_2 \le \sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d}$, and hence

$$\Big\| \sum_{j \in \mathcal{I}(\mathcal{T}_y)} h_{y,j} \boldsymbol{\xi}_j \Big\|_2 \le \sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d}\, \Delta_y.$$

Combining the last three displays gives

$$\Big\| \sum_{j \in \mathcal{I}(\mathcal{T}_y)} \widetilde{c}_{y,j} \boldsymbol{\xi}_j \Big\|_2 \le \sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d} \left( \frac{\ell_y}{\sqrt{n_{\mathcal{T}_y}}} + \Delta_y \right).$$

Using the triangle inequality on $\mathbf{r}_{\mathrm{noise}}$ yields

$$\|\mathbf{r}_{\mathrm{noise}}\|_2 \le \sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d} \sum_{y \in \{\pm 1\}} \left( \frac{\ell_y}{\sqrt{n_{\mathcal{T}_y}}} + \Delta_y \right).$$

Lemma D.8(iii) also gives $\widetilde{u}_y \le 1$ and $\widetilde{c}_{y,\mu} \ge \ell_y$, hence

$$\widetilde{u}_+ + \widetilde{u}_- \le 2, \qquad A_{\mathrm{dc}} \ge \ell_+ + \ell_-.$$

By Assumption D.1A4, we have

$$2\sigma_{\mathcal{S}} \le \sqrt{\|\mathbf{\Sigma}_\xi\|_{\mathrm{op}}}\, \rho_{\mathcal{S}}^{T_{\mathrm{in}}} \frac{\beta_p}{\sqrt{n_{\mathcal{T}}}},$$

and since $B_{\mathrm{out}} \ge 1$, this implies

$$\|\mathbf{r}_{\mathrm{init}}\|_2 \le 2\sqrt{C_\kappa d}\, \sigma_{\mathcal{S}} \le \sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}}\, d}\, \rho_{\mathcal{S}}^{T_{\mathrm{in}}}\, B_{\mathrm{out}} \frac{\beta_p}{\sqrt{n_{\mathcal{T}}}}.$$

Combining the bounds on $\mathbf{r}_{\mathrm{init}}$ and $\mathbf{r}_{\mathrm{noise}}$, and using $n_{\mathcal{T}_y} = p_y n_{\mathcal{T}}$, gives

$$\|\mathbf{r}_{\mathrm{dc}}\|_2 \le \frac{\sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d}}{\sqrt{n_{\mathcal{T}}}} \left[ \rho_{\mathcal{S}}^{T_{\mathrm{in}}} B_{\mathrm{out}} \beta_p + \sum_{y \in \{\pm 1\}} \ell_y p_y^{-1/2} + \sqrt{n_{\mathcal{T}}} \sum_{y \in \{\pm 1\}} \Delta_y \right].$$

Since $\ell_y = u_{\mathrm{out}} - \Delta_y$ and $u_{\mathrm{out}} = (1 - \rho_{\mathcal{S}}^{T_{\mathrm{in}}}) B_{\mathrm{out}}$, the bracket equals

$$B_{\mathrm{out}} \beta_p + \sum_{y \in \{\pm 1\}} \Delta_y \big( \sqrt{n_{\mathcal{T}}} - p_y^{-1/2} \big).$$

Substituting $A_{\mathrm{dc}} \ge \ell_+ + \ell_-$ and the preceding bound on $\|\mathbf{r}_{\mathrm{dc}}\|_2$ into the definition $\widehat{\mathrm{SR}}(\mathcal{S}_{\mathrm{dc}}) = A_{\mathrm{dc}} \|\boldsymbol{\mu}\|_2 / \|\mathbf{r}_{\mathrm{dc}}\|_2$ gives the stated lower bound.

Finally, if $G_{\max} < \lambda$, then Lemma D.8(iv) gives $\widetilde{c}_{y,\mu} = 1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}$ and $\|\widetilde{\boldsymbol{c}}_y\|_2 = (1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}) / \sqrt{n_{\mathcal{T}_y}}$. In this regime, $T_{\mathcal{S},\mathrm{lin}}(k) = T_{\mathrm{in}}$ for all $k$, hence $B_{\mathrm{out}} = \sum_{q=0}^{T_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{q T_{\mathrm{in}}}$, $\ell_y = u_{\mathrm{out}} = 1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}$, and $\Delta_y = 0$ from the displayed identities. □

The lower bound in Corollary D.11 isolates the effect of activity switching through the deficits $\Delta_y = u_{\mathrm{out}} - \ell_y$. The next deterministic comparison controls these deficits using the lengths of the certified active windows. Conditioned on $T_{\mathrm{lin}}$ and $T_{\mathcal{S},\mathrm{lin}}(r)$, the comparison follows algebraically from the definitions of $B_{\mathrm{out}}$, $\ell_y$, and $u_{\mathrm{out}}$. It identifies regimes in which the factor $L_{\mathrm{dc}}$ preserves the class-averaging scale in Corollary D.11.

**Lemma D.12.** *Assume the hypotheses of Corollary D.11, and let $B_{\mathrm{out}}$, $u_{\mathrm{out}}$, $\ell_y$, $\Delta_y$, and $L_{\mathrm{dc}}$ be the quantities defined there. Set $T := T_{\mathrm{in}}$ and $\rho := \rho_{\mathcal{S}}$. For $r \in \{0, \dots, T_{\mathrm{out}} - 1\}$, define*

$$D_r := \sum_{k=r+1}^{T_{\mathrm{out}}-1} \big( T - T_{\mathcal{S},\mathrm{lin}}(k) \big), \qquad M_{\mathcal{S}} := \max_{0 \le r \le T_{\mathrm{out}}-1} D_r,$$

*and*

$$\mathcal{W}_{\mathrm{win}} := (1-\rho^T)\big(1-\rho^{M\mathcal{S}}\big) + \max_{\substack{y\in\{\pm1\}\\ 0\le r\le T_{\mathrm{out}}-1}} \big(1-\rho^{T-T_{y,\mathrm{lin}}(r)}\big).$$

*Then, for each* $y\in\{\pm1\}$,

$$0 \le \Delta_y = u_{\mathrm{out}} - \ell_y \le B_{\mathrm{out}}\mathcal{W}_{\mathrm{win}}, \qquad \ell_y \ge B_{\mathrm{out}}\big((1-\rho^T) - \mathcal{W}_{\mathrm{win}}\big).$$

*Moreover,*

$$\sum_{y\in\{\pm1\}} \Delta_y \le 2B_{\mathrm{out}}\mathcal{W}_{\mathrm{win}}.$$

*The following consequences for* $L_{\mathrm{dc}}$ *hold along any sequence of problem instances for which* $\beta_p > 0$.

(i) *If* $G_{\max} < \lambda$, *then* $\mathcal{W}_{\mathrm{win}} = 0$ *and*

$$L_{\mathrm{dc}} = \frac{2(1-\rho^T)}{\beta_p}.$$

*In particular, if* $p_+ = p_- = 1/2$, *then* $L_{\mathrm{dc}} = (1-\rho^T)/\sqrt{2}$.

(ii) *If*

$$\mathcal{W}_{\mathrm{win}} = o(1-\rho^T), \qquad \mathcal{W}_{\mathrm{win}} = o\left(\frac{\beta_p}{\sqrt{n_{\mathcal{T}}}}\right),$$

*then*

$$L_{\mathrm{dc}} = (1+o(1))\frac{2(1-\rho^T)}{\beta_p}.$$

(iii) *If*

$$\mathcal{W}_{\mathrm{win}} = o(1-\rho^T), \qquad \mathcal{W}_{\mathrm{win}} = O\left(\frac{\beta_p}{\sqrt{n_{\mathcal{T}}}}\right),$$

*then*

$$L_{\mathrm{dc}} = \Omega\left(\frac{1-\rho^T}{\beta_p}\right).$$

(iv) *Let* $n_{\min} := \min_{y\in\{\pm1\}} n_{\mathcal{T}_y}$. *If* $n_{\min}\to\infty$ *and*

$$\sum_{y\in\{\pm1\}} \Delta_y = \Omega\big(B_{\mathrm{out}}(1-\rho^T)\big),$$

*then*

$$L_{\mathrm{dc}} = O(n_{\mathcal{T}}^{-1/2}).$$

*Proof.* For each $r$, set

$$F_r := \sum_{k=r+1}^{T_{\mathrm{out}}-1} T_{\mathcal{S},\mathrm{lin}}(k), \qquad s_{y,r} := T - T_{y,\mathrm{lin}}(r).$$

Then $F_r = (T_{\mathrm{out}}-1-r)T - D_r$, and the $r$th contribution to $u_{\mathrm{out}}$ is $a_r := (1-\rho^T)\rho^{F_r}$, while the $r$th contribution to $\ell_y$ is

$$b_{y,r} := \rho^{(T_{\mathrm{out}}-r)T - T_{y,\mathrm{lin}}(r)}\big(1-\rho^{T_{y,\mathrm{lin}}(r)}\big) = \rho^{F_r+D_r+s_{y,r}}\big(1-\rho^{T-s_{y,r}}\big).$$

Hence

$$a_r - b_{y,r} = \rho^{F_r}\left[(1-\rho^T) - \rho^{D_r+s_{y,r}}\big(1-\rho^{T-s_{y,r}}\big)\right].$$

The bracket admits the exact decomposition

$$(1-\rho^T)\big(1-\rho^{D_r}\big) + \rho^{D_r}\big(1-\rho^{s_{y,r}}\big),$$

which is nonnegative and bounded above by $\mathcal{W}_{\mathrm{win}}$. Summing over $r$ gives

$$0 \le u_{\mathrm{out}} - \ell_y = \sum_{r=0}^{T_{\mathrm{out}}-1} (a_r - b_{y,r}) \le \mathcal{W}_{\mathrm{win}} \sum_{r=0}^{T_{\mathrm{out}}-1} \rho^{F_r} = B_{\mathrm{out}} \mathcal{W}_{\mathrm{win}}.$$

The lower bound on $\ell_y$ follows from $\ell_y = u_{\mathrm{out}} - \Delta_y$, $u_{\mathrm{out}} = B_{\mathrm{out}}(1-\rho^T)$, and the preceding display. Summing the bounds on $\Delta_y$ over $y \in \{\pm 1\}$ gives $\sum_y \Delta_y \le 2B_{\mathrm{out}}\mathcal{W}_{\mathrm{win}}$.

If $G_{\max} < \lambda$, Lemma D.8(iv) gives full activity, so $T_{\mathcal{S},\mathrm{lin}}(r) = T$ and $T_{y,\mathrm{lin}}(r) = T$ for all $y, r$. Thus $M_{\mathcal{S}} = 0$, $\mathcal{W}_{\mathrm{win}} = 0$, $B_{\mathrm{out}} = \sum_{q=0}^{T_{\mathrm{out}}-1} \rho^{qT}$, and $\ell_y = u_{\mathrm{out}} = 1 - \rho^{T_{\mathrm{out}}T}$. Substituting these identities into the definition of $L_{\mathrm{dc}}$ gives

$$L_{\mathrm{dc}} = \frac{2(1-\rho^{T_{\mathrm{out}}T})}{B_{\mathrm{out}}\beta_p} = \frac{2(1-\rho^T)}{\beta_p}.$$

For (ii) and (iii), write the numerator of $L_{\mathrm{dc}}$ as $N := \ell_+ + \ell_-$ and its denominator as

$$D := B_{\mathrm{out}}\beta_p + \sum_{y\in\{\pm1\}} \Delta_y\big(\sqrt{n_{\mathcal{T}}} - p_y^{-1/2}\big).$$

The bounds above imply

$$2B_{\mathrm{out}}\big((1-\rho^T) - \mathcal{W}_{\mathrm{win}}\big) \le N \le 2B_{\mathrm{out}}(1-\rho^T),$$

and

$$B_{\mathrm{out}}\beta_p \le D \le B_{\mathrm{out}}\beta_p + 2B_{\mathrm{out}}\sqrt{n_{\mathcal{T}}}\,\mathcal{W}_{\mathrm{win}}.$$

If $\mathcal{W}_{\mathrm{win}} = o(1-\rho^T)$ and $\mathcal{W}_{\mathrm{win}} = o(\beta_p/\sqrt{n_{\mathcal{T}}})$, then $N = (1+o(1))2B_{\mathrm{out}}(1-\rho^T)$ and $D = (1+o(1))B_{\mathrm{out}}\beta_p$, proving (ii). If $\mathcal{W}_{\mathrm{win}} = o(1-\rho^T)$ and $\mathcal{W}_{\mathrm{win}} = O(\beta_p/\sqrt{n_{\mathcal{T}}})$, then $N = \Omega(B_{\mathrm{out}}(1-\rho^T))$ and $D = O(B_{\mathrm{out}}\beta_p)$, proving (iii).

For (iv), the numerator in $L_{\mathrm{dc}}$ is at most $2B_{\mathrm{out}}(1-\rho^T)$, while its denominator is at least

$$\big(\sqrt{n_{\mathcal{T}}} - \max_y p_y^{-1/2}\big) \sum_{y\in\{\pm1\}} \Delta_y.$$

Since $\max_y p_y^{-1/2} = \sqrt{n_{\mathcal{T}}/n_{\min}} = o(\sqrt{n_{\mathcal{T}}})$ when $n_{\min} \to \infty$, the denominator is $\Omega(\sqrt{n_{\mathcal{T}}}\,B_{\mathrm{out}}(1-\rho^T))$ under the stated lower bound on $\sum_y \Delta_y$. Hence $L_{\mathrm{dc}} = O(n_{\mathcal{T}}^{-1/2})$. □

### D.6 Proofs for the Multiclass OvR Results

This subsection proves the auxiliary results used for Theorems B.3 and B.4, Corollary B.5, and Remark B.6 in Appendix B.3. The argument follows the binary affine-recursion proof after introducing head-wise OvR signed coordinates; the additional class and head indices enter through the activity and coefficient windows.

The first step is to certify active windows for the head-wise OvR hinge indicators. After passing to signed OvR samples, each fixed head has the same fully-active weight trajectory as in the binary analysis, so the binary margin-envelope argument can be applied head-wise.

**Lemma D.13.** *Assume Assumption B.1,* $0 < \eta_{\mathbf{w}}\lambda \le 1$, *and* $0 < 2\eta_{\mathcal{S}} \le 1$. *Set* $\rho := 1 - \eta_{\mathbf{w}}\lambda$ *and* $\rho_{\mathcal{S}} := 1 - 2\eta_{\mathcal{S}}$. *For each class* $k$, *define*

$$B_{\mathcal{T},k}^{\mathrm{OvR}} := \max_{\substack{0\le r<T_{\mathrm{out}}\\ h\in[K]\\ i\in\mathcal{I}_k(\mathcal{T})}} \left\langle \mathbf{w}_h^{(r,0)}, \mathbf{z}_{i,h}^{\mathcal{T}} \right\rangle, \qquad G_{\mathcal{T},k}^{\mathrm{OvR}} := \max_{\substack{h\in[K]\\ i\in\mathcal{I}_k(\mathcal{T})}} \left\langle \bar{\mathbf{z}}_h^{\mathcal{T}}, \mathbf{z}_{i,h}^{\mathcal{T}} \right\rangle,$$

*and let*

$$T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}} := \max\left\{ m \in \{0,\dots,T_{\mathrm{in}}\} : \rho^t B_{\mathcal{T},k}^{\mathrm{OvR}} + \frac{1-\rho^t}{\lambda} G_{\mathcal{T},k}^{\mathrm{OvR}} < 1 \text{ for all } t = 0,\dots,m-1 \right\}.$$

*For each outer loop $r$ and each class $k$, set $\bar{\mathbf{z}}_{k,h}^{\mathcal{T}} := n_{\mathcal{T},k}^{-1} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \mathbf{z}_{i,h}^{\mathcal{T}}$ and define*

$$B_{\mathcal{S},k}^{\mathrm{OvR}}(r) := \max_{h \in [K]} \max \left\{ \left\langle \mathbf{w}_h^{(r,0)}, \mathbf{z}_{k,h}^{\mathcal{S},(r,0)} \right\rangle, \left\langle \mathbf{w}_h^{(r,0)}, \bar{\mathbf{z}}_{k,h}^{\mathcal{T}} \right\rangle \right\},$$

$$G_{\mathcal{S},k}^{\mathrm{OvR}}(r) := \max_{h \in [K]} \max \left\{ \left\langle \bar{\mathbf{z}}_h^{\mathcal{T}}, \mathbf{z}_{k,h}^{\mathcal{S},(r,0)} \right\rangle, \left\langle \bar{\mathbf{z}}_h^{\mathcal{T}}, \bar{\mathbf{z}}_{k,h}^{\mathcal{T}} \right\rangle \right\}.$$

*Then define*

$$T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r) := \max \left\{ m \in \{0, \dots, T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}\} : \rho^t B_{\mathcal{S},k}^{\mathrm{OvR}}(r) + \frac{1-\rho^t}{\lambda} G_{\mathcal{S},k}^{\mathrm{OvR}}(r) < 1 \text{ for all } t = 0, \dots, m-1 \right\}.$$

*Under classwise OvR gradient matching, for every outer loop $r$,*

$$q_{i,h}^{\mathcal{T},(r,t)} = 1 \quad \forall\, i \in \mathcal{I}_k(\mathcal{T}),\ h \in [K],\ t < T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}, \qquad q_{k,h}^{\mathcal{S},(r,t)} = 1 \quad \forall\, h \in [K],\ t < T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r).$$

*In particular, if the two displayed margin inequalities in the definitions of $T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}$ and $T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)$ hold for every $t = 0, \dots, T_{\mathrm{in}} - 1$, then the corresponding window length is $T_{\mathrm{in}}$.*

*Proof.* Fix a head $h$. In the signed OvR coordinate $\mathbf{z}_{i,h}^{\mathcal{T}} = \tau_{h,y_i^{\mathcal{T}}} \mathbf{x}_i^{\mathcal{T}}$, the training activity is $q_{i,h}^{\mathcal{T},(\boldsymbol{t})} = \mathbb{1}\{1 - \langle \mathbf{w}_h^{(\boldsymbol{t})}, \mathbf{z}_{i,h}^{\mathcal{T}} \rangle > 0\}$, and the weight update can be written as

$$\mathbf{w}_h^{(\boldsymbol{t}^+)} = (1 - \eta_{\mathbf{w}}\lambda)\mathbf{w}_h^{(\boldsymbol{t})} + \eta_{\mathbf{w}} \frac{1}{n_{\mathcal{T}}} \sum_{i=1}^{n_{\mathcal{T}}} q_{i,h}^{\mathcal{T},(\boldsymbol{t})} \mathbf{z}_{i,h}^{\mathcal{T}}.$$

Thus, for fixed $h$, the training-side dynamics have the same signed-sample form as the binary dynamics used in Corollary D.4. The proof of Corollary D.4, with the signed samples replaced by $\{\mathbf{z}_{i,h}^{\mathcal{T}}\}_i$ and then with the maximum taken over $h$ and $i \in \mathcal{I}_k(\mathcal{T})$, gives

$$\left\langle \mathbf{w}_h^{(r,t)}, \mathbf{z}_{i,h}^{\mathcal{T}} \right\rangle \le \rho^t B_{\mathcal{T},k}^{\mathrm{OvR}} + \frac{1-\rho^t}{\lambda} G_{\mathcal{T},k}^{\mathrm{OvR}}.$$

The definition of $T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}$ therefore implies $q_{i,h}^{\mathcal{T},(r,t)} = 1$ for all $i \in \mathcal{I}_k(\mathcal{T})$, all heads $h$, and all $t < T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}$.

We next certify the condensed-sample window. Fix an outer loop $r$, a class $k$, and a head $h$. In a joint activity induction, when the training activities for class $k$ and the condensed-sample activities for class $k$ are active for all heads up to the current inner time, the averaged quantities satisfy $\bar{q}_k^{\mathcal{S}} = 1$ and $\bar{b}_{k,i} = 1$. Multiplying the condensed-sample update by $\tau_{h,k}$ gives

$$\mathbf{z}_{k,h}^{\mathcal{S},(\boldsymbol{t}^+)} = \rho_{\mathcal{S}} \mathbf{z}_{k,h}^{\mathcal{S},(\boldsymbol{t})} + (1 - \rho_{\mathcal{S}}) \bar{\mathbf{z}}_{k,h}^{\mathcal{T}},$$

which is the full-active signed condensed-sample recursion used in Lemma D.7. Since $T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r) \le T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}$, the training activity conclusion above gives the full-active head trajectory on this window:

$$\mathbf{w}_h^{(r,t)} = \rho^t \mathbf{w}_h^{(r,0)} + \frac{1-\rho^t}{\lambda} \bar{\mathbf{z}}_h^{\mathcal{T}}.$$

The signed condensed sample is a convex combination $\mathbf{z}_{k,h}^{\mathcal{S},(r,t)} = \rho_{\mathcal{S}}^t \mathbf{z}_{k,h}^{\mathcal{S},(r,0)} + (1-\rho_{\mathcal{S}}^t)\bar{\mathbf{z}}_{k,h}^{\mathcal{T}}$ on the same induction window. Therefore

$$\left\langle \mathbf{w}_h^{(r,t)}, \mathbf{z}_{k,h}^{\mathcal{S},(r,t)} \right\rangle \le \rho^t B_{\mathcal{S},k}^{\mathrm{OvR}}(r) + \frac{1-\rho^t}{\lambda} G_{\mathcal{S},k}^{\mathrm{OvR}}(r).$$

The definition of $T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)$ implies that the last display is strictly smaller than 1 for all $t < T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)$, and hence $q_{k,h}^{\mathcal{S},(r,t)} = 1$ on that window. The statement about full-length windows follows from the definitions of the window lengths. □

Once the active windows are fixed, the multiclass condensed-sample update can be unrolled class by class. The next lemma is the OvR counterpart of the binary coefficient expansion: it gives a nonnegative expansion, the corresponding signal–noise decomposition, and the coefficient bounds needed later for the pairwise risk analysis.

**Lemma D.14.** *Assume $0 < 2\eta_{\mathcal{S}} \le 1$ and $n_{\mathcal{T},k} \ge 1$ for every $k \in [K]$, and set $\rho_{\mathcal{S}} := 1 - 2\eta_{\mathcal{S}}$ and $\boldsymbol{t}^\star := (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$. Under classwise OvR gradient matching, for every $k \in [K]$ there exist terminal coefficients*

$$u_k^{(\boldsymbol{t}^\star)} \ge 0, \qquad c_{k,i}^{(\boldsymbol{t}^\star)} \ge 0, \quad i \in \mathcal{I}_k(\mathcal{T}),$$

*such that*

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} = u_k^{(\boldsymbol{t}^\star)} \mathbf{p}_k^{\mathcal{S},(0,0)} + \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)} \mathbf{x}_i^{\mathcal{T}}.$$

*Set $c_{k,\mu}^{(\boldsymbol{t}^\star)} := \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)}$ and $\boldsymbol{c}_k^{(\boldsymbol{t}^\star)} := (c_{k,i}^{(\boldsymbol{t}^\star)})_{i \in \mathcal{I}_k(\mathcal{T})}$. If the $K$-class additive model satisfies $\mathbf{x}_i^{\mathcal{T}} = \boldsymbol{\mu}_k + \boldsymbol{\xi}_i$ for $i \in \mathcal{I}_k(\mathcal{T})$, then*

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} = u_k^{(\boldsymbol{t}^\star)} \mathbf{p}_k^{\mathcal{S},(0,0)} + c_{k,\mu}^{(\boldsymbol{t}^\star)} \boldsymbol{\mu}_k + \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)} \boldsymbol{\xi}_i.$$

*Assume additionally that the OvR training and condensed-sample linear windows for class $k$ have lengths $T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}} \in \{0, \dots, T_{\mathrm{in}}\}$ and $T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r) \in \{0, \dots, T_{\mathrm{in}}\}$, $r \in \{0, \dots, T_{\mathrm{out}} - 1\}$, in the sense that, for every outer loop $r$,*

$$q_{i,h}^{\mathcal{T},(r,t)} = 1 \quad \forall\, i \in \mathcal{I}_k(\mathcal{T}),\ h \in [K],\ t < T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}, \qquad q_{k,h}^{\mathcal{S},(r,t)} = 1 \quad \forall\, h \in [K],\ t < T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r).$$

*Define*

$$T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r) := \min\{T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}, T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)\},$$

$$B_{k,\mathrm{out}} := 1 + \sum_{r=0}^{T_{\mathrm{out}}-2} \rho_{\mathcal{S}}^{\sum_{s=r+1}^{T_{\mathrm{out}}-1} T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(s)}, \qquad u_{k,\mathrm{out}} := (1 - \rho_{\mathcal{S}}^{T_{\mathrm{in}}}) B_{k,\mathrm{out}},$$

*and*

$$\ell_k := \sum_{r=0}^{T_{\mathrm{out}}-1} \rho_{\mathcal{S}}^{(T_{\mathrm{out}}-r)T_{\mathrm{in}} - T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r)} \big(1 - \rho_{\mathcal{S}}^{T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r)}\big).$$

*Then*

$$\rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}} \le u_k^{(\boldsymbol{t}^\star)} \le \rho_{\mathcal{S}}^{\sum_{r=0}^{T_{\mathrm{out}}-1} T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)}.$$

*Moreover, for every $i \in \mathcal{I}_k(\mathcal{T})$,*

$$\frac{\ell_k}{n_{\mathcal{T},k}} \le c_{k,i}^{(\boldsymbol{t}^\star)} \le \frac{u_{k,\mathrm{out}}}{n_{\mathcal{T},k}},$$

*and consequently*

$$\ell_k \le c_{k,\mu}^{(\boldsymbol{t}^\star)} \le u_{k,\mathrm{out}}, \qquad \frac{\ell_k}{\sqrt{n_{\mathcal{T},k}}} \le \|\boldsymbol{c}_k^{(\boldsymbol{t}^\star)}\|_2 \le \frac{u_{k,\mathrm{out}}}{\sqrt{n_{\mathcal{T},k}}}.$$

*With $\Delta_k := u_{k,\mathrm{out}} - \ell_k$ and $h_{k,i} := c_{k,i}^{(\boldsymbol{t}^\star)} - \ell_k / n_{\mathcal{T},k}$, we have $\Delta_k \ge 0$, $h_{k,i} \ge 0$, and $\sum_{i \in \mathcal{I}_k(\mathcal{T})} h_{k,i} \le \Delta_k$.*

*If all OvR training and condensed-sample activities are active for every update time in the bilevel procedure, then*

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} = \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}} \mathbf{p}_k^{\mathcal{S},(0,0)} + \big(1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}} T_{\mathrm{in}}}\big) \widehat{\boldsymbol{\mu}}_k^{\mathcal{T}}, \qquad \widehat{\boldsymbol{\mu}}_k^{\mathcal{T}} := \frac{1}{n_{\mathcal{T},k}} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \mathbf{x}_i^{\mathcal{T}}.$$

*Proof.* We derive the affine recursion for the classwise update. Fix $k \in [K]$ and an inner-loop time index $\boldsymbol{t} = (t_{\mathrm{out}}, t_{\mathrm{in}})$ with $t_{\mathrm{in}} < T_{\mathrm{in}}$, and let $\boldsymbol{t}^+ = (t_{\mathrm{out}}, t_{\mathrm{in}} + 1)$. For a fixed head $h$, write

$$\mathbf{r}_{k,h}^{(\boldsymbol{t})} := q_{k,h}^{\mathcal{S},(\boldsymbol{t})} \tau_{h,k} \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} - \frac{1}{n_{\mathcal{T},k}} \sum_{i \in \mathcal{I}_k(\mathcal{T})} q_{i,h}^{\mathcal{T},(\boldsymbol{t})} \tau_{h,k} \mathbf{x}_i^{\mathcal{T}}.$$

Then $D_k^{\mathrm{OvR}}(\mathcal{P}^{(\boldsymbol{t})}; \mathbf{W}^{(\boldsymbol{t})}) = K^{-1}\sum_{h=1}^{K} \|\mathbf{r}_{k,h}^{(\boldsymbol{t})}\|_2^2$. Since all hinge activities are evaluated at time $\boldsymbol{t}$ and treated as fixed with respect to $\mathbf{p}_k^{\mathcal{S}}$, we have

$$\nabla_{\mathbf{p}_k^{\mathcal{S}}} \left\| \mathbf{r}_{k,h}^{(\boldsymbol{t})} \right\|_2^2 = 2q_{k,h}^{\mathcal{S},(\boldsymbol{t})} \tau_{h,k} \mathbf{r}_{k,h}^{(\boldsymbol{t})}.$$

Substituting the definition of $\mathbf{r}_{k,h}^{(\boldsymbol{t})}$ and using $\tau_{h,k}^2 = 1$ and $(q_{k,h}^{\mathcal{S},(\boldsymbol{t})})^2 = q_{k,h}^{\mathcal{S},(\boldsymbol{t})}$ gives

$$\nabla_{\mathbf{p}_k^{\mathcal{S}}} \left\| \mathbf{r}_{k,h}^{(\boldsymbol{t})} \right\|_2^2 = 2q_{k,h}^{\mathcal{S},(\boldsymbol{t})} \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} - \frac{2}{n_{\mathcal{T},k}} \sum_{i \in \mathcal{I}_k(\mathcal{T})} q_{k,h}^{\mathcal{S},(\boldsymbol{t})} q_{i,h}^{\mathcal{T},(\boldsymbol{t})} \mathbf{x}_i^{\mathcal{T}}.$$

Averaging over $h$ yields

$$\nabla_{\mathbf{p}_k^{\mathcal{S}}} D_k^{\mathrm{OvR}} \left( \mathcal{P}^{(\boldsymbol{t})}; \mathbf{W}^{(\boldsymbol{t})} \right) = 2\bar{q}_k^{\mathcal{S},(\boldsymbol{t})} \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} - \frac{2}{n_{\mathcal{T},k}} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \bar{b}_{k,i}^{(\boldsymbol{t})} \mathbf{x}_i^{\mathcal{T}}.$$

Inserting this gradient into the classwise condensed-sample step gives

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^+)} = \left(1 - 2\eta_{\mathcal{S}} \bar{q}_k^{\mathcal{S},(\boldsymbol{t})}\right) \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} + \frac{2\eta_{\mathcal{S}}}{n_{\mathcal{T},k}} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \bar{b}_{k,i}^{(\boldsymbol{t})} \mathbf{x}_i^{\mathcal{T}}.$$

We next unroll this update. Define the global coefficients by

$$u_k^{(0,0)} := 1, \qquad c_{k,i}^{(0,0)} := 0, \quad i \in \mathcal{I}_k(\mathcal{T}).$$

For an inner-loop update, set

$$u_k^{(\boldsymbol{t}^+)} := \left(1 - 2\eta_{\mathcal{S}} \bar{q}_k^{\mathcal{S},(\boldsymbol{t})}\right) u_k^{(\boldsymbol{t})}$$

and, for each $i \in \mathcal{I}_k(\mathcal{T})$,

$$c_{k,i}^{(\boldsymbol{t}^+)} := \left(1 - 2\eta_{\mathcal{S}} \bar{q}_k^{\mathcal{S},(\boldsymbol{t})}\right) c_{k,i}^{(\boldsymbol{t})} + \frac{2\eta_{\mathcal{S}}}{n_{\mathcal{T},k}} \bar{b}_{k,i}^{(\boldsymbol{t})}.$$

Across the outer-loop boundary, set

$$u_k^{(t_{\mathrm{out}}+1,0)} := u_k^{(t_{\mathrm{out}},T_{\mathrm{in}})}, \qquad c_{k,i}^{(t_{\mathrm{out}}+1,0)} := c_{k,i}^{(t_{\mathrm{out}},T_{\mathrm{in}})}.$$

These recursions match those in Lemma D.6(i) after replacing the binary activity $q_{\mathcal{S},y}$ by $\bar{q}_k^{\mathcal{S}}$, the product $q_{\mathcal{S},y} q_{\mathcal{T},j}$ by $\bar{b}_{k,i}$, and the signed training sample $\mathbf{z}_j^{\mathcal{T}}$ by $\mathbf{x}_i^{\mathcal{T}}$. The affine-recursion argument in Lemma D.6(i) depends only on the update form, the inequalities $0 \le \bar{b}_{k,i}^{(\boldsymbol{t})} \le \bar{q}_k^{\mathcal{S},(\boldsymbol{t})} \le 1$, and $0 < 2\eta_{\mathcal{S}} \le 1$. It therefore yields the expansion and coefficient nonnegativity for every condensed state, in particular at $\boldsymbol{t}^\star$.

Under the additive model, substituting $\mathbf{x}_i^{\mathcal{T}} = \boldsymbol{\mu}_k + \boldsymbol{\xi}_i$ into the coefficient expansion gives

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} = u_k^{(\boldsymbol{t})} \mathbf{p}_k^{\mathcal{S},(0,0)} + \left( \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t})} \right) \boldsymbol{\mu}_k + \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t})} \boldsymbol{\xi}_i,$$

and setting $\boldsymbol{t} = \boldsymbol{t}^\star$ gives the stated signal–noise expansion.

The terminal coefficient estimates follow from the coefficient-bound calculation in Lemma D.6(ii) and the outer-loop coefficient recursion in Lemma D.7(iii). The substitutions are $n_{\mathcal{T}_y} \mapsto n_{\mathcal{T},k}$, $T_{\mathcal{S},\mathrm{lin}}(r) \mapsto T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)$, $T_{\mathrm{lin}} \mapsto T_{\mathcal{T},k,\mathrm{lin}}^{\mathrm{OvR}}$, and $T_{y,\mathrm{lin}}(r) \mapsto T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r)$. The coefficient-bound argument remains valid for the averaged activities, since the only inequalities used in the upper and lower bounds are $0 \le \bar{b}_{k,i} \le \bar{q}_k^{\mathcal{S}} \le 1$, $\rho_{\mathcal{S}} \le 1 - 2\eta_{\mathcal{S}} \bar{q}_k^{\mathcal{S}} \le 1$, and $\bar{q}_k^{\mathcal{S}} = \bar{b}_{k,i} = 1$ on the joint linear window. Applying this recursion gives the displayed bounds on $u_k^{(\boldsymbol{t}^\star)}$, $c_{k,i}^{(\boldsymbol{t}^\star)}$, $c_{k,\mu}^{(\boldsymbol{t}^\star)}$, and $\|\boldsymbol{c}_k^{(\boldsymbol{t}^\star)}\|_2$. The inequalities for $\Delta_k$ and $h_{k,i}$ follow, as in Lemma D.8, from the coordinatewise lower bound and the total-mass upper bound.

Finally, in the full-activity case, $\bar{q}_k^{\mathcal{S},(r,t)} = 1$ and $\bar{b}_{k,i}^{(r,t)} = 1$ throughout the bilevel procedure. The affine recursion reduces at every update to

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^+)} = \rho_{\mathcal{S}} \mathbf{p}_k^{\mathcal{S},(\boldsymbol{t})} + (1-\rho_{\mathcal{S}})\widehat{\boldsymbol{\mu}}_k^{\mathcal{T}}.$$

Iterating this recursion through the $T_{\mathrm{out}}T_{\mathrm{in}}$ condensed-sample updates, with the usual outer-loop inheritance, yields the displayed contraction toward $\widehat{\boldsymbol{\mu}}_k^{\mathcal{T}}$. □

The preceding lemma concerns only the dynamics. The following lemma is a geometric reduction for the nearest-prototype evaluator: if every condensed sample is close to its class mean, then the multiclass error is controlled by pairwise class separation.

**Lemma D.15.** *Assume the $K$-class additive test model $\mathbf{x} = \boldsymbol{\mu}_y + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is mean-zero sub-Gaussian with proxy covariance $\boldsymbol{\Sigma}_\xi$. Let $\mathcal{P} = \{(\mathbf{p}_k, k) : k \in [K]\}$ be any multiclass condensed set used with the nearest-prototype evaluator, let $\pi_k := \mathbb{P}(y = k)$, and define $\epsilon_k := \|\mathbf{p}_k - \boldsymbol{\mu}_k\|_2$, $\gamma_{k\ell} := \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell\|_2$, $R_{k\ell} := \epsilon_k + \epsilon_\ell$, and $\widehat{\mathrm{SR}}_{k\ell}^{\mathrm{mc}} := \gamma_{k\ell}/R_{k\ell}$, with value $+\infty$ when $R_{k\ell} = 0$. For $s \in (0, +\infty)$, set $\Psi_{\mathrm{mc}}(s) := \big((1-s^{-1})_+\big)^4/(1+s^{-1})^2$, with $\Psi_{\mathrm{mc}}(0) := 0$ and $\Psi_{\mathrm{mc}}(+\infty) := 1$. Then*

$$\mathcal{R}_{\mathrm{mc}}(\mathcal{P}) \le \sum_{k=1}^{K} \pi_k \sum_{\ell \neq k} \exp\left(-\frac{\gamma_{k\ell}^2}{8\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}} \Psi_{\mathrm{mc}}\left(\widehat{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}\right)\right).$$

*Proof.* Fix $k \neq \ell$ and condition on $y = k$, so that $\mathbf{x} = \boldsymbol{\mu}_k + \boldsymbol{\xi}$. The event $E_{k\to\ell}$ is equivalent to

$$\langle \boldsymbol{\xi}, \mathbf{p}_\ell - \mathbf{p}_k \rangle \ge \frac{1}{2}\left(\|\mathbf{p}_\ell - \boldsymbol{\mu}_k\|_2^2 - \|\mathbf{p}_k - \boldsymbol{\mu}_k\|_2^2\right).$$

Let

$$A_{k\ell} := \left(\gamma_{k\ell}^2 - 2\gamma_{k\ell}\epsilon_\ell + \epsilon_\ell^2 - \epsilon_k^2\right)_+.$$

Since $\|\mathbf{p}_\ell - \boldsymbol{\mu}_k\|_2 \ge \gamma_{k\ell} - \epsilon_\ell$ and $\|\mathbf{p}_k - \boldsymbol{\mu}_k\|_2 = \epsilon_k$, the positive part of the threshold is at least $A_{k\ell}/2$. Also $\|\mathbf{p}_\ell - \mathbf{p}_k\|_2 \le \gamma_{k\ell} + R_{k\ell}$. The one-dimensional sub-Gaussian tail bound gives

$$\mathbb{P}(E_{k\to\ell}) \le \exp\left(-\frac{A_{k\ell}^2}{8\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}(\gamma_{k\ell} + R_{k\ell})^2}\right).$$

Moreover, $A_{k\ell} \ge (\gamma_{k\ell} - R_{k\ell})_+^2$, and hence

$$\mathbb{P}(E_{k\to\ell}) \le \exp\left(-\frac{\big((\gamma_{k\ell} - R_{k\ell})_+\big)^4}{8\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}(\gamma_{k\ell} + R_{k\ell})^2}\right) = \exp\left(-\frac{\gamma_{k\ell}^2}{8\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}} \Psi_{\mathrm{mc}}\left(\widehat{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}\right)\right),$$

with the stated conventions when $R_{k\ell} = 0$. The multiclass risk is bounded by applying the union bound over $\ell \neq k$ and then averaging over the class prior $\pi_k$. □

Combining the coefficient expansion with the high-probability noise controls converts the terminal OvR condensed set into pairwise certified multiclass signal-to-residual ratios, which can then be substituted into the preceding nearest-prototype risk bound.

**Corollary D.16.** *Assume the hypotheses of Lemma D.14 and the high-probability event in Assumption B.1. Let $C_\kappa$ denote the corresponding uniform concentration constant, let $\boldsymbol{t}^\star = (T_{\mathrm{out}} - 1, T_{\mathrm{in}})$, and define the terminal multiclass condensed set $\mathcal{P}_{\mathrm{dc}} := \{(\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)}, k) : k \in [K]\}$. Let $\ell_k$, $\Delta_k$, and $T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)$ be the coefficient-window quantities from Lemma D.14. Set $p_k := n_{\mathcal{T},k}/n_{\mathcal{T}}$,*

$$\alpha_{\mathcal{S}} := \frac{\sigma_{\mathcal{S}}\sqrt{n_{\mathcal{T}}}}{\sqrt{\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}}}, \qquad \alpha_{\mu,k} := \frac{\|\boldsymbol{\mu}_k\|_2\sqrt{n_{\mathcal{T}}}}{\sqrt{C_\kappa\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}},$$

*and, for each class $k$, let*

$$\bar{u}_k := \rho_{\mathcal{S}}^{\sum_{r=0}^{T_{\mathrm{out}}-1} T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(r)}$$

*and, for $k \neq \ell$, define the pairwise multiclass averaging factor*

$$L_{k\ell}^{\mathrm{mc}} := \Big[\alpha_{\mathcal{S}}(\bar{u}_k + \bar{u}_\ell) + \alpha_{\mu,k}(1-\ell_k) + \alpha_{\mu,\ell}(1-\ell_\ell) + \ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2} + \sqrt{n_{\mathcal{T}}}(\Delta_k + \Delta_\ell)\Big]^{-1}.$$

*Let*

$$\widehat{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}(\mathcal{P}_{\mathrm{dc}}) := \frac{\gamma_{k\ell}}{\left\|\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} - \boldsymbol{\mu}_k\right\|_2 + \left\|\mathbf{p}_\ell^{\mathcal{S},(\boldsymbol{t}^\star)} - \boldsymbol{\mu}_\ell\right\|_2}.$$

*If the denominator is zero, set $\widehat{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}(\mathcal{P}_{\mathrm{dc}}) = +\infty$. Then*

$$\widehat{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}(\mathcal{P}_{\mathrm{dc}}) \geq \underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}} := \frac{\sqrt{n_{\mathcal{T}}}\, L_{k\ell}^{\mathrm{mc}}\, \gamma_{k\ell}}{\sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}}, \qquad k \neq \ell.$$

*Consequently,*

$$\mathcal{R}_{\mathrm{mc}}(\mathcal{P}_{\mathrm{dc}}) \leq \sum_{k=1}^{K} \pi_k \sum_{\ell \neq k} \exp\left(-\frac{\gamma_{k\ell}^2}{8\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}} \Psi_{\mathrm{mc}}\left(\underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}\right)\right).$$

*In particular, if $\gamma_{\min} := \min_{k\neq\ell} \gamma_{k\ell}$ and $\underline{\mathrm{SR}}_{\min}^{\mathrm{mc}} := \min_{k\neq\ell} \underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}$, then*

$$\mathcal{R}_{\mathrm{mc}}(\mathcal{P}_{\mathrm{dc}}) \leq (K-1) \exp\left(-\frac{\gamma_{\min}^2}{8\|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}}} \Psi_{\mathrm{mc}}\left(\underline{\mathrm{SR}}_{\min}^{\mathrm{mc}}\right)\right).$$

*If all OvR activities are active throughout training, then $L_{k\ell}^{\mathrm{mc}}$ can be replaced by*

$$L_{k\ell,\mathrm{full}}^{\mathrm{mc}} := \left[\rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}(2\alpha_{\mathcal{S}} + \alpha_{\mu,k} + \alpha_{\mu,\ell}) + \left(1 - \rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}\right)\left(p_k^{-1/2} + p_\ell^{-1/2}\right)\right]^{-1}$$

*in the definition of $\underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}$.*

*Proof.* The signal–noise expansion in Lemma D.14 gives

$$\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} - \boldsymbol{\mu}_k = u_k^{(\boldsymbol{t}^\star)} \mathbf{p}_k^{\mathcal{S},(0,0)} - \left(1 - c_{k,\mu}^{(\boldsymbol{t}^\star)}\right)\boldsymbol{\mu}_k + \sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)} \boldsymbol{\xi}_i.$$

The same coefficient recursion also gives $u_k^{(\boldsymbol{t})} + c_{k,\mu}^{(\boldsymbol{t})} \leq 1$ for every time index: initially the sum is one, and one update changes it by at most

$$u_k^{(\boldsymbol{t}^+)} + c_{k,\mu}^{(\boldsymbol{t}^+)} \leq \left(1 - 2\eta_{\mathcal{S}} \bar{q}_k^{\mathcal{S},(\boldsymbol{t})}\right)\left(u_k^{(\boldsymbol{t})} + c_{k,\mu}^{(\boldsymbol{t})}\right) + 2\eta_{\mathcal{S}} \bar{q}_k^{\mathcal{S},(\boldsymbol{t})} \leq 1,$$

because $n_{\mathcal{T},k}^{-1} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \bar{b}_{k,i}^{(\boldsymbol{t})} \leq \bar{q}_k^{\mathcal{S},(\boldsymbol{t})}$. Hence $0 \leq c_{k,\mu}^{(\boldsymbol{t}^\star)} \leq 1$ and $1 - c_{k,\mu}^{(\boldsymbol{t}^\star)} \leq 1 - \ell_k$.

The upper bound on $u_k^{(\boldsymbol{t}^\star)}$ from Lemma D.14 gives $u_k^{(\boldsymbol{t}^\star)} \leq \bar{u}_k$. On the event in Assumption B.1, the initialization and noise bounds give $\|\mathbf{p}_k^{\mathcal{S},(0,0)}\|_2 \leq \sqrt{C_\kappa d}\sigma_{\mathcal{S}}$,

$$\left\|\frac{1}{n_{\mathcal{T},k}} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \boldsymbol{\xi}_i\right\|_2 \leq \sqrt{\frac{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}{n_{\mathcal{T},k}}}, \qquad \max_{i \in \mathcal{I}_k(\mathcal{T})} \|\boldsymbol{\xi}_i\|_2 \leq \sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}.$$

Using the decomposition $c_{k,i}^{(\boldsymbol{t}^\star)} = \ell_k / n_{\mathcal{T},k} + h_{k,i}$ from Lemma D.14, with $h_{k,i} \geq 0$ and $\sum_i h_{k,i} \leq \Delta_k$, we obtain

$$\left\|\sum_{i \in \mathcal{I}_k(\mathcal{T})} c_{k,i}^{(\boldsymbol{t}^\star)} \boldsymbol{\xi}_i\right\|_2 \leq \ell_k \left\|\frac{1}{n_{\mathcal{T},k}} \sum_{i \in \mathcal{I}_k(\mathcal{T})} \boldsymbol{\xi}_i\right\|_2 + \sum_{i \in \mathcal{I}_k(\mathcal{T})} h_{k,i} \|\boldsymbol{\xi}_i\|_2 \leq \sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}\left(\frac{\ell_k}{\sqrt{n_{\mathcal{T},k}}} + \Delta_k\right).$$

Combining these three estimates and using $n_{\mathcal{T},k} = p_k n_{\mathcal{T}}$ yields

$$\left\|\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} - \boldsymbol{\mu}_k\right\|_2 \leq \frac{\sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}}{\sqrt{n_{\mathcal{T}}}} \left[\alpha_{\mathcal{S}} \bar{u}_k + \alpha_{\mu,k}(1-\ell_k) + \ell_k p_k^{-1/2} + \sqrt{n_{\mathcal{T}}}\Delta_k\right].$$

Therefore, for every pair $k \neq \ell$,

$$\left\|\mathbf{p}_k^{\mathcal{S},(\boldsymbol{t}^\star)} - \boldsymbol{\mu}_k\right\|_2 + \left\|\mathbf{p}_\ell^{\mathcal{S},(\boldsymbol{t}^\star)} - \boldsymbol{\mu}_\ell\right\|_2 \leq \frac{\sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}}{\sqrt{n\mathcal{T}}} \frac{1}{L_{k\ell}^{\mathrm{mc}}}.$$

By the definition of the pairwise multiclass ratio, this gives $\widehat{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}(\mathcal{P}_{\mathrm{dc}}) \geq \underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}$.

Applying Lemma D.15 and using that $\Psi_{\mathrm{mc}}$ is nondecreasing on $[0, +\infty]$ gives the pairwise risk bound. The compact bound with $\gamma_{\min}$ and $\underline{\mathrm{SR}}_{\min}^{\mathrm{mc}}$ follows from $\gamma_{k\ell} \geq \gamma_{\min}$, $\underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}} \geq \underline{\mathrm{SR}}_{\min}^{\mathrm{mc}}$, and $\sum_{k=1}^K \pi_k = 1$. The full-activity version follows from the exact contraction in Lemma D.14: in that case $\bar{u}_k = \rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}$, $1 - \ell_k = \rho_{\mathcal{S}}^{T_{\mathrm{out}}T_{\mathrm{in}}}$, and $\Delta_k = 0$ for every class $k$. □

The final deterministic comparison interprets the factor $L_{k\ell}^{\mathrm{mc}}$ in the certified ratio. It isolates the effect of active-window gaps through $\mathcal{W}_k^{\mathrm{OvR}}$ and shows when the pairwise bound retains the class-averaging scale.

**Lemma D.17.** *Assume the hypotheses of Corollary D.16. Set $T := T_{\mathrm{in}}$ and $\rho := \rho_{\mathcal{S}}$. For each class $k$ and outer-loop index $r$, define*

$$D_{k,r} := \sum_{s=r+1}^{T_{\mathrm{out}}-1} \big(T - T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(s)\big), \qquad M_{\mathcal{S},k} := \max_{0 \leq r \leq T_{\mathrm{out}}-1} D_{k,r},$$

*and*

$$\mathcal{W}_k^{\mathrm{OvR}} := (1-\rho^T)\big(1-\rho^{M_{\mathcal{S},k}}\big) + \max_{0 \leq r \leq T_{\mathrm{out}}-1} \big(1-\rho^{T - T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r)}\big).$$

*Then, for every $k \in [K]$,*

$$0 \leq \Delta_k \leq B_{k,\mathrm{out}} \mathcal{W}_k^{\mathrm{OvR}}, \qquad \ell_k \geq B_{k,\mathrm{out}}\big((1-\rho^T) - \mathcal{W}_k^{\mathrm{OvR}}\big).$$

*For $k \neq \ell$,*

$$\big(L_{k\ell}^{\mathrm{mc}}\big)^{-1} = \alpha_{\mathcal{S}}(\bar{u}_k + \bar{u}_\ell) + \alpha_{\mu,k}(1-\ell_k) + \alpha_{\mu,\ell}(1-\ell_\ell) + \sqrt{n\mathcal{T}}(\Delta_k + \Delta_\ell) + \ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}.$$

*Moreover, the window lengths control the nonuniform coefficient contribution through*

$$\Delta_k + \Delta_\ell \leq B_{k,\mathrm{out}} \mathcal{W}_k^{\mathrm{OvR}} + B_{\ell,\mathrm{out}} \mathcal{W}_\ell^{\mathrm{OvR}}.$$

*The following consequences hold along any sequence of problem instances for which $\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2} > 0$.*

(i) *If*

$$\alpha_{\mathcal{S}}(\bar{u}_k + \bar{u}_\ell) + \alpha_{\mu,k}(1-\ell_k) + \alpha_{\mu,\ell}(1-\ell_\ell) = o\Big(\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}\Big),$$

*and*

$$\Delta_k + \Delta_\ell = o\left(\frac{\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}}{\sqrt{n\mathcal{T}}}\right),$$

*then*

$$L_{k\ell}^{\mathrm{mc}} = (1+o(1))\left(\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}\right)^{-1},$$

*and*

$$\underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}} = \big(1+o(1)\big) \frac{\sqrt{n\mathcal{T}}\,\gamma_{k\ell}}{\big(\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}\big)\sqrt{C_\kappa \|\boldsymbol{\Sigma}_\xi\|_{\mathrm{op}} d}}.$$

(ii) *If*

$$\alpha_{\mathcal{S}}(\bar{u}_k + \bar{u}_\ell) + \alpha_{\mu,k}(1-\ell_k) + \alpha_{\mu,\ell}(1-\ell_\ell) = O\Big(\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}\Big),$$

*and*

$$\Delta_k + \Delta_\ell = O\left(\frac{\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}}{\sqrt{n_{\mathcal{T}}}}\right),$$

*then*

$$L_{k\ell}^{\mathrm{mc}} = \Omega\Big(\Big(\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}\Big)^{-1}\Big), \qquad \underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}} = \Omega\left(\frac{\sqrt{n_{\mathcal{T}}}\,\gamma_{k\ell}}{\big(\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}\big)\sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d}}\right).$$

(iii) *If* $\Delta_k + \Delta_\ell = \Omega(1)$*, then*

$$L_{k\ell}^{\mathrm{mc}} = O(n_{\mathcal{T}}^{-1/2}), \qquad \underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}} = O\left(\frac{\gamma_{k\ell}}{\sqrt{C_\kappa \|\mathbf{\Sigma}_\xi\|_{\mathrm{op}} d}}\right).$$

(iv) *If the OvR training and condensed-sample windows for classes* $k$ *and* $\ell$ *have full length* $T$ *at every outer loop, then*

$$\big(L_{k\ell}^{\mathrm{mc}}\big)^{-1} = \rho^{T_{\mathrm{out}}T}\,(2\alpha_{\mathcal{S}} + \alpha_{\mu,k} + \alpha_{\mu,\ell}) + \big(1 - \rho^{T_{\mathrm{out}}T}\big)\big(p_k^{-1/2} + p_\ell^{-1/2}\big).$$

*Thus the full-window transition is governed by the comparison between the two terms in this denominator. In particular, if*

$$\rho^{T_{\mathrm{out}}T}\,(2\alpha_{\mathcal{S}} + \alpha_{\mu,k} + \alpha_{\mu,\ell}) = O\Big(\big(1 - \rho^{T_{\mathrm{out}}T}\big)\big(p_k^{-1/2} + p_\ell^{-1/2}\big)\Big),$$

*then the conclusion in* (ii) *holds with* $\ell_j = 1 - \rho^{T_{\mathrm{out}}T}$ *for* $j \in \{k, \ell\}$*. A sufficient scale for this regime, when* $0 < \rho < 1$*, is*

$$T_{\mathrm{out}}T = \Omega\left(\frac{\left[\log\left(\frac{2\alpha_{\mathcal{S}} + \alpha_{\mu,k} + \alpha_{\mu,\ell}}{p_k^{-1/2} + p_\ell^{-1/2}}\right)\right]_+}{-\log\rho}\right).$$

*Proof.* For each $k$ and $r$, set

$$F_{k,r} := \sum_{s=r+1}^{T_{\mathrm{out}}-1} T_{\mathcal{S},k,\mathrm{lin}}^{\mathrm{OvR}}(s), \qquad s_{k,r} := T - T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r).$$

Then $F_{k,r} = (T_{\mathrm{out}} - 1 - r)T - D_{k,r}$. The $r$th contribution to $u_{k,\mathrm{out}}$ is $a_{k,r} := (1-\rho^T)\rho^{F_{k,r}}$, while the $r$th contribution to $\ell_k$ is

$$b_{k,r} := \rho^{(T_{\mathrm{out}}-r)T - T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r)}\big(1 - \rho^{T_{k,\mathrm{lin}}^{\mathrm{OvR}}(r)}\big) = \rho^{F_{k,r}+D_{k,r}+s_{k,r}}\big(1 - \rho^{T - s_{k,r}}\big).$$

Therefore

$$a_{k,r} - b_{k,r} = \rho^{F_{k,r}}\left[(1-\rho^T) - \rho^{D_{k,r}+s_{k,r}}\big(1 - \rho^{T-s_{k,r}}\big)\right].$$

The bracket has the exact decomposition

$$(1-\rho^T)\big(1 - \rho^{D_{k,r}}\big) + \rho^{D_{k,r}}\big(1 - \rho^{s_{k,r}}\big),$$

which is nonnegative and bounded above by $\mathcal{W}_k^{\mathrm{OvR}}$. Summing over $r$ yields

$$0 \le u_{k,\mathrm{out}} - \ell_k \le \mathcal{W}_k^{\mathrm{OvR}} \sum_{r=0}^{T_{\mathrm{out}}-1} \rho^{F_{k,r}} = B_{k,\mathrm{out}}\mathcal{W}_k^{\mathrm{OvR}}.$$

Since $\Delta_k = u_{k,\mathrm{out}} - \ell_k$ and $u_{k,\mathrm{out}} = B_{k,\mathrm{out}}(1-\rho^T)$, the first display follows.

By definition,

$$\left(L_{k\ell}^{\mathrm{mc}}\right)^{-1} = \alpha_{\mathcal{S}}(\bar{u}_k + \bar{u}_\ell) + \alpha_{\mu,k}(1-\ell_k) + \alpha_{\mu,\ell}(1-\ell_\ell) + \ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2} + \sqrt{n_{\mathcal{T}}}(\Delta_k + \Delta_\ell).$$

This proves the displayed identity for $(L_{k\ell}^{\mathrm{mc}})^{-1}$. Summing the two bounds $\Delta_j \le B_{j,\mathrm{out}}\mathcal{W}_j^{\mathrm{OvR}}$ over $j \in \{k,\ell\}$ gives the displayed control of the nonuniform coefficient contribution.

The two assumptions in (i) make the initialization, signal-recovery, and nonuniform coefficient terms jointly $o(\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2})$, and the two assumptions in (ii) make them jointly $O(\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2})$. The first two asymptotic conclusions therefore follow by comparing the denominator with the class-averaging term $\ell_k p_k^{-1/2} + \ell_\ell p_\ell^{-1/2}$ and then using the definition of $\underline{\mathrm{SR}}_{k\ell}^{\mathrm{mc}}$. If $\Delta_k + \Delta_\ell = \Omega(1)$, then the denominator of $L_{k\ell}^{\mathrm{mc}}$ is at least $\sqrt{n_{\mathcal{T}}}(\Delta_k + \Delta_\ell) = \Omega(\sqrt{n_{\mathcal{T}}})$, which proves (iii). If the windows for $k$ and $\ell$ are full, then $T_{\mathcal{S},j,\mathrm{lin}}^{\mathrm{OvR}}(r) = T_{j,\mathrm{lin}}^{\mathrm{OvR}}(r) = T$ for $j \in \{k,\ell\}$ and all $r$. Hence $\bar{u}_j = \rho^{T_{\mathrm{out}}T}$, $\ell_j = 1 - \rho^{T_{\mathrm{out}}T}$, and $\Delta_j = 0$, which gives the displayed full-window formula. The stated sufficient scale is the logarithmic form of the contraction requirement when $0 < \rho < 1$. □

Table 3: Default hyperparameters for synthetic experiments.

| Item | Value |
|---|---|
| Dataset size | $n_{\mathcal{T}} = 1000$ (balanced), train/test split 60%/40% |
| Dimension | $d = 500$ |
| Signal and noise | $\|\boldsymbol{\mu}\|_2 = 1.0$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma = 1.0$ |
| Regularization | $\lambda = 1.0$ |
| Learning rates | $\eta_{\mathbf{w}} = 0.05$, $\eta_{\mathcal{S}} = 0.25$ |
| Bilevel steps | $T_{\text{out}} = 50$, $T_{\text{in}} = 10$ |
| Initialization scales | $\sigma_{\mathbf{w}} = 0.01$, $\sigma_{\mathcal{S}} = 0.01$ |
| Seeds | $n_{\text{seeds}} = 5$ with seeds $\{0, 1, 2, 3, 4\}$ |

Table 4: Default hyperparameters for the $K$-class OvR synthetic experiment.

| Item | Value |
|---|---|
| Dataset size | $K = 5$, 200 generated samples per class, train/test split 60%/40% |
| Dimension | $d = 500$ |
| Class means and noise | Near-orthogonal class means with $\|\boldsymbol{\mu}_k\|_2 = 2.0$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma = 1.0$ |
| Regularization | $\lambda = 1.0$ |
| Learning rates | $\eta_{\mathbf{w}} = 0.05$, $\eta_{\mathcal{S}} = 0.25$ |
| Bilevel steps | $T_{\text{out}} = 50$, $T_{\text{in}} = 10$ |
| Initialization scales | $\sigma_{\mathbf{w}} = 0.01$, $\sigma_{\mathcal{S}} = 0.01$ |
| Seeds | $n_{\text{seeds}} = 5$ with seeds $\{0, 1, 2, 3, 4\}$ |

# E Additional Simulation Details

This section provides the reproducibility details, numerical tables, and auxiliary diagnostics underlying the simulation results in Section 4.

## E.1 Experimental setup and reproducibility

Unless otherwise stated, each experiment compares the learned condensed set with a random one-shot coreset that selects one real training sample per class. We report mean ± one standard deviation over the specified random seeds.

### E.1.1 Parameter settings

Table 3, Table 4, and Table 5 summarize the default hyperparameters used in the binary synthetic, multiclass synthetic, and KMNIST experiments, respectively.

### E.1.2 Environment and random seeds

All experiments are implemented in Python with PyTorch. Each script iterates `seed` in `range(n_seeds)` with default seeds $\{0, 1, 2, 3, 4\}$. At the beginning of each trial, we set both `torch.manual_seed(seed)` and `np.random.seed(seed)`; the same seed is propagated to synthetic data generation, the train/test split, random one-shot coreset sampling, and model initialization. This convention makes the reported Monte Carlo summaries reproducible from the experiment logs.

Table 5: Default hyperparameters for KMNIST experiments (two-class subset).

| Item | Value |
|---|---|
| Dataset | KMNIST (class 0 vs class 1), official train/test split |
| Input | $d = 784$ (flattened), zero-mean unit-variance normalization |
| Regularization | $\lambda = 0.01$ |
| Learning rates | $\eta_{\mathbf{w}} = 0.005$, $\eta_{\mathcal{S}} = 0.1$ |
| Bilevel steps | $T_{\text{out}} = 10$, $T_{\text{in}} \in \{5, 10, 20, 30\}$ |
| Initialization scales | $\sigma_{\mathbf{w}} = 0.01$, $\sigma_{\mathcal{S}} = 0.01$ |
| Seeds | $n_{\text{seeds}} = 5$ with seeds $\{0, 1, 2, 3, 4\}$ |

Table 6: Synthetic binary aggregation results (mean ± std over seeds). Alignment is $\cos(\mathbf{s}(\mathcal{S}^\star), \boldsymbol{\mu})$, and SNR is the signal-to-residual proxy used in the synthetic diagnostic.

| Method | Alignment | SNR proxy | SVM acc. | Nearest-centroid acc. |
|---|---|---|---|---|
| Condensed (GM) | $0.725 \pm 0.024$ | $0.526 \pm 0.036$ | $0.772 \pm 0.013$ | $0.773 \pm 0.013$ |
| Random one-shot | $0.070 \pm 0.028$ | $0.006 \pm 0.004$ | $0.517 \pm 0.029$ | $0.506 \pm 0.010$ |
| Full data | – | – | $0.768 \pm 0.021$ | – |

#### E.1.3 Implementation conventions

The implementation follows the deterministic hinge indicator convention used in the theory: $q(\mathbf{w}) = \mathbb{1}\{u(\mathbf{w}) < 1\}$, with a strict inequality, and the indicator is not differentiated through. In the binary single-sample-per-class setting $n_{\mathcal{S}} = 2$, the condensed update at state $\mathbf{w}$ uses the explicit formula $\mathbf{z}_y \leftarrow \mathbf{z}_y + \eta_{\mathcal{S}}\, q_y^{\mathcal{S}}(\mathbf{w})\, y\, (g_{\mathcal{S}}(\mathbf{w}) - g_{\mathcal{T}}(\mathbf{w}))$, followed by the inner-loop model update $\mathbf{w} \leftarrow \mathbf{w} - \eta_{\mathbf{w}}\, g_{\mathcal{T}}(\mathbf{w})$. The random one-shot baseline constructs $\mathcal{S}_{\text{rand}}$ by sampling one training point per class from $\mathcal{T}$ and evaluates it under the same downstream training procedure as $\mathcal{S}^\star$.

### E.2 Raw tables for synthetic condensed-sample aggregation

Table 6 reports the geometric and downstream quantities underlying Fig. 1b. The condensed set has much stronger alignment with the signal direction than the random one-shot baseline (0.725 versus 0.070), and its SVM test accuracy is comparable to full-data training (0.772 versus 0.768) while clearly exceeding random selection (0.518).

### E.3 Accuracy heatmap and grid values for the inner/outer sweep

Figure 3 complements Fig. 2a in the main text by visualizing downstream test accuracy over the same $(T_{\text{in}}, T_{\text{out}})$ grid, and Table 7 lists the corresponding values. The peak accuracy 0.780 is attained for small $T_{\text{in}} \in \{1, 2\}$ with moderate $T_{\text{out}} \in \{5, 10\}$; larger inner-loop lengths lead to a modest but consistent reduction, in line with the weaker alignment values in Fig. 2a.

### E.4 Dynamics diagnostics

Figure 4 reports a representative stepwise diagnostic that tracks the training-set activity ratio $\text{act}_{\mathcal{T}}$, the condensed-set activity ratio $\text{act}_{\mathcal{S}}$, gradient norms, and the gradient-matching discrepancy as functions of the inner-loop step. In this default configuration, $\text{act}_{\mathcal{S}}$ remains equal to one throughout the displayed trajectory, while $\text{act}_{\mathcal{T}}$ and the training-gradient norm decrease along the inner loop. Thus the diagnostic illustrates a training-side loss of active gradient signal, rather than a condensed-sample activity switch.

The displayed diagnostic corresponds to the representative grid point $(T_{\text{out}}, T_{\text{in}}) = (10, 10)$, averaged over the five random seeds. Within each outer-loop restart, the training-set activity begins near one and then

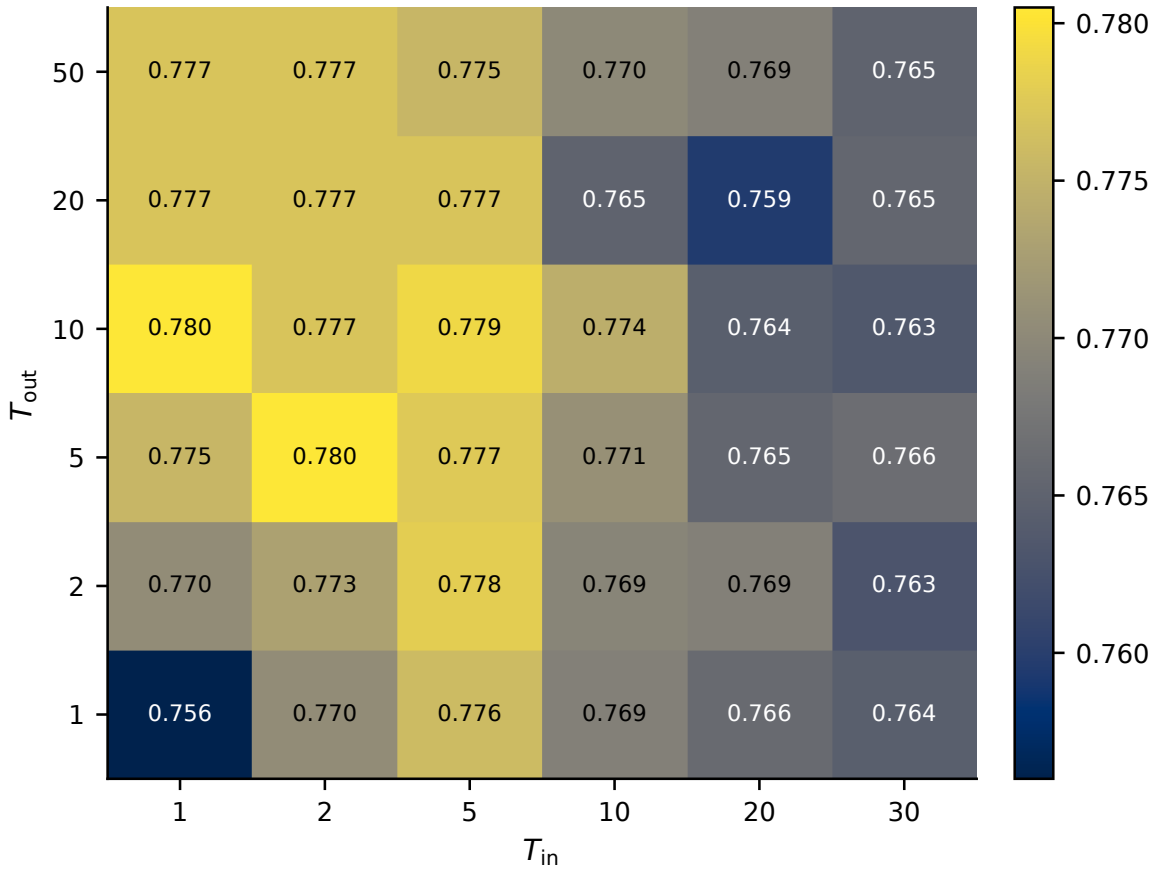



Figure 3: Test accuracy over the $(T_{\text{in}}, T_{\text{out}})$ grid on synthetic data, complementing Fig. 2a in the main text.

Table 7: Grid values for Fig. 3: test accuracy over $(T_{\text{in}}, T_{\text{out}})$ on synthetic data.

| $T_{\text{in}} \backslash T_{\text{out}}$ | **1** | **2** | **5** | **10** | **20** | **50** |
|---|---|---|---|---|---|---|
| **1** | 0.756 | 0.770 | 0.775 | 0.780 | 0.777 | 0.777 |
| **2** | 0.770 | 0.773 | 0.780 | 0.777 | 0.777 | 0.777 |
| **5** | 0.776 | 0.778 | 0.777 | 0.779 | 0.777 | 0.775 |
| **10** | 0.769 | 0.769 | 0.771 | 0.774 | 0.765 | 0.770 |
| **20** | 0.766 | 0.769 | 0.765 | 0.764 | 0.759 | 0.769 |
| **30** | 0.764 | 0.763 | 0.766 | 0.763 | 0.765 | 0.765 |

decreases along the inner loop, reaching roughly 0.79 by the final inner step. The norm of the training gradient follows the same pattern, falling from about 1.39 at the beginning of an inner loop to about 0.53 near its end. By contrast, the condensed-set activity remains equal to one, so the condensed samples stay in the hinge-active region while the real-data gradient becomes progressively sparser. The gradient-matching discrepancy decreases rapidly during the early active part of the trajectory, but it stops decreasing monotonically once the training gradient has weakened. This behavior is consistent with the active-window interpretation in the theory: the early inner steps provide the main classwise averaging signal, whereas later steps operate with fewer active real samples and therefore contribute less stable signal accumulation.

### E.5 Multiclass OvR experiment

Table 8 and Fig. 5 report the $K$-class classwise OvR experiment described in Appendix B. Since the multiclass theory evaluates the learned set through nearest-prototype classification, the table uses the same evaluator

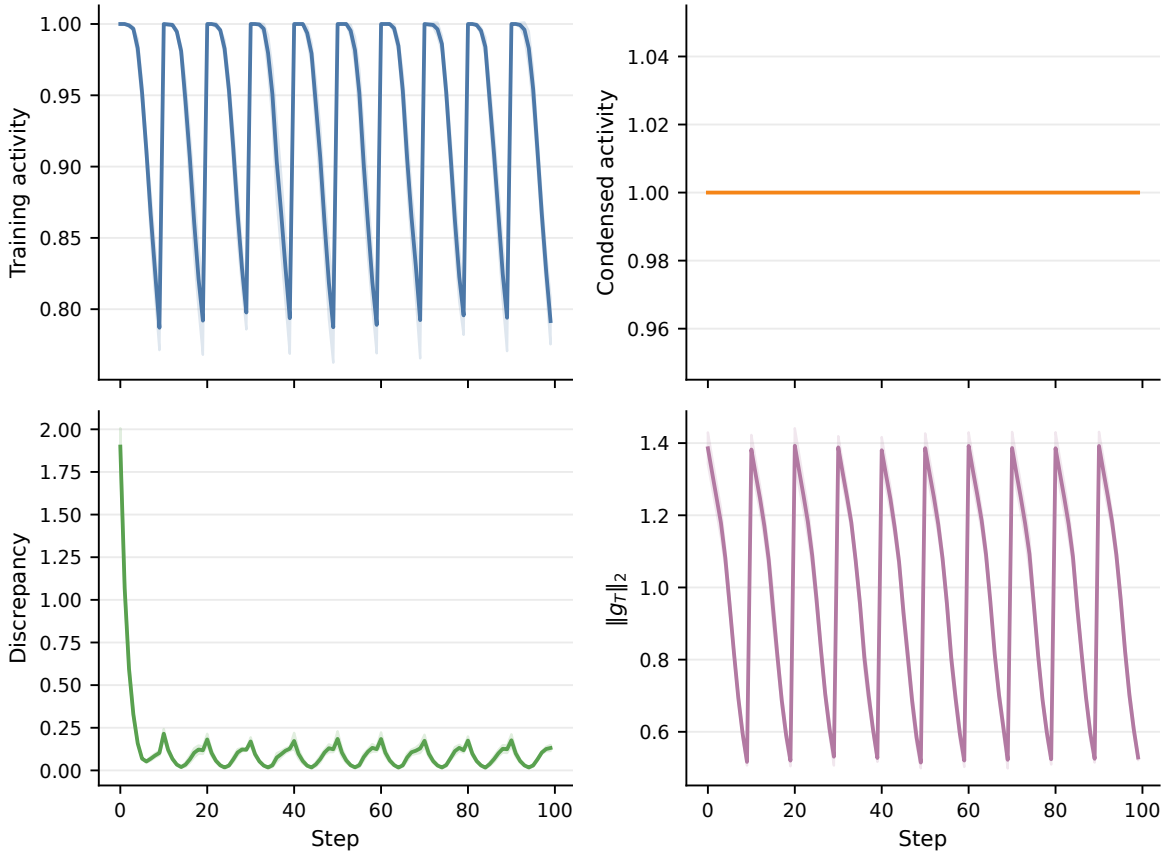



Figure 4: Representative dynamics diagnostic on synthetic data at $(T_{\text{out}}, T_{\text{in}}) = (10, 10)$. The condensed-set activity ratio remains active in this default setting, while the training-set activity and gradient norms decrease along the inner loop.

Table 8: $K$-class OvR condensation on synthetic additive data with one condensed sample per class, evaluated by nearest-prototype classification (mean ± std over seeds).

| **Method** | **Test accuracy** |
|---|---|
| GM-OvR | $0.624 \pm 0.008$ |
| Random one-shot | $0.231 \pm 0.030$ |
| Full-data nearest-prototype | $0.624 \pm 0.018$ |

for GM-OvR, the random one-shot baseline, and the full-data nearest-prototype reference whose prototypes are the empirical class means. The learned multiclass condensed set matches this full-data nearest-prototype reference within Monte Carlo variability and substantially outperforms random one-shot selection.

### E.6 Cross-model evaluation tables

Table 9 reports the mean±std test accuracy underlying Fig. 2b, comparing the condensed set (GM) and the random one-shot baseline across multiple evaluators on KMNIST. Table 10 provides the analogous cross-model comparison on synthetic data.

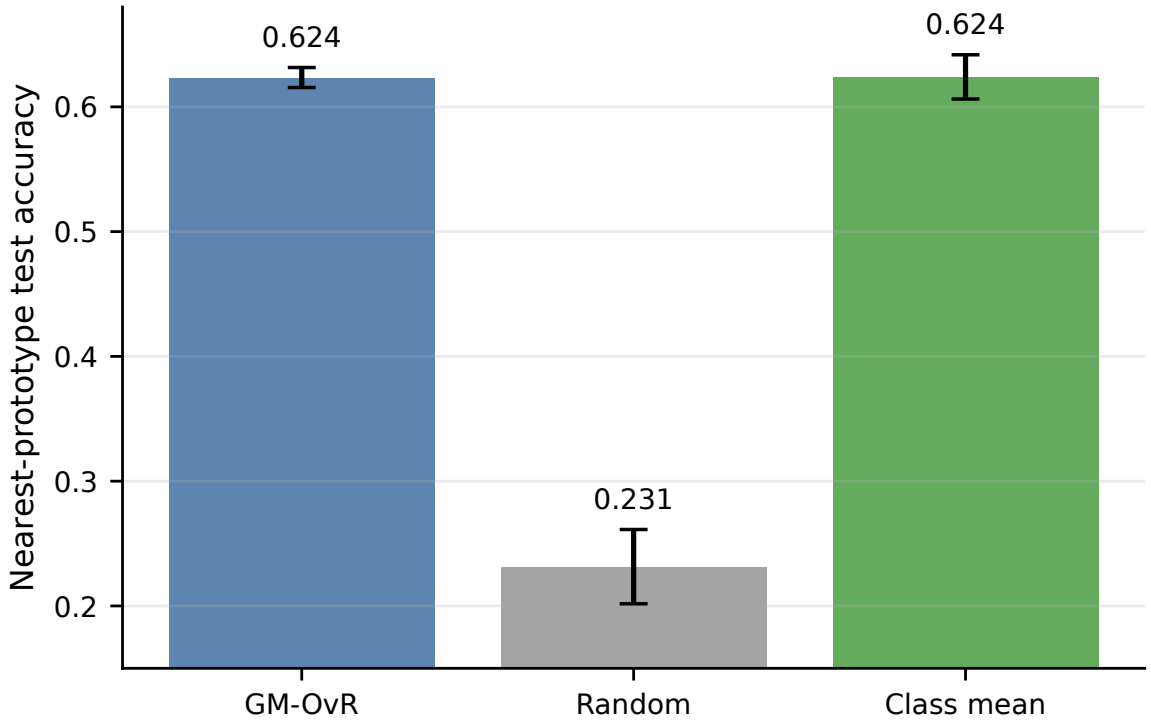



Figure 5: Multiclass OvR condensation with one condensed sample per class. The figure reports $K$-way test accuracy under the nearest-prototype evaluator, comparing the learned multiclass condensed set against a random one-shot coreset and the full-data nearest-prototype reference based on empirical class means.

Table 9: KMNIST cross-model evaluation (mean ± std over seeds) on the same condensed set $\mathcal{S}^\star$ (GM) compared with a random one-shot coreset (Random).

| **Evaluator** | **GM** | **Random** |
|---|---|---|
| Linear SVM | $0.960 \pm 0.006$ | $0.828 \pm 0.062$ |
| Logistic Regression | $0.960 \pm 0.005$ | $0.838 \pm 0.053$ |
| Nearest Centroid | $0.960 \pm 0.006$ | $0.669 \pm 0.153$ |
| Two-layer ReLU MLP | $0.953 \pm 0.003$ | $0.811 \pm 0.059$ |

Table 10: Synthetic cross-model evaluation (mean ± std over seeds) on the same condensed set $\mathcal{S}^\star$ (GM) compared with a random one-shot coreset (Random).

| **Evaluator** | **GM** | **Random** |
|---|---|---|
| Linear SVM | $0.991 \pm 0.007$ | $0.648 \pm 0.102$ |
| Logistic Regression | $0.992 \pm 0.005$ | $0.640 \pm 0.052$ |
| Nearest Centroid | $0.991 \pm 0.007$ | $0.598 \pm 0.038$ |
| Two-layer ReLU MLP | $0.564 \pm 0.012$ | $0.556 \pm 0.023$ |