THE IMPLICIT BIAS OF LOGIT REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Label smoothing is known empirically to improve calibration and generalization in classification tasks, but its mechanism remains under-explored. Mathematically, label smoothing can be viewed as a special case of a broader class of *logit regularization*: the addition of a convex penalty directly in logit space. In this work, we analyze these logit regularizers in the context of linear classification, and demonstrate that they induce an implicit bias of *logit clustering* around finite per-sample targets. Remarkably, this reveals a novel connection between logit regularization and *Linear Discriminant Analysis* (LDA): for Gaussian data, we prove that such regularization drives the weight vector to align exactly with the LDA direction, a result which extends approximately to any distribution when the logits are sufficiently clustered. To demonstrate the consequences, we study a simple signal-plus-noise model in which this transition has dramatic effects: Logit regularization halves the critical sample complexity and induces grokking in the small-noise limit, while making generalization robust to noise. Our results extend the theoretical understanding of label smoothing and highlight the efficacy of a broader class of logit regularization methods.

1 INTRODUCTION

Modern deep learning models are frequently trained with Label Smoothing (LS) (Szegedy et al., 2016) or similar regularization techniques to improve generalization and calibration (Müller et al., 2019). The empirical benefits of these methods are well-documented, and their success is often attributed to the prevention of overconfidence. However, a theoretical understanding of their underlying mechanics remains incomplete.

In this work, we investigate the geometric mechanism of a broader class of regularizers: convex *logit regularization* (LR), of which Label Smoothing is one instance. Unlike weight-based regularization schemes, such as L_2, or ridge regularization (Tikhonov & Arsenin, 1977), LR allows a per-sample decomposition of the loss as $\mathcal{L} = N^{-1} \sum_i \ell(\boldsymbol{z}_i, \boldsymbol{y}_i)$, where $\boldsymbol{z}_i$ and $\boldsymbol{y}_i$ denote the logit and class of the i-th sample, respectively, and N is the number of samples. A convex logit penalty generically creates a finite global minimum for ℓ, and thus the optimizer performs *logit clustering* around these finite targets in logit space. This is in contrast to unregularized logistic regression, in which the (optimizer-dependent) implicit bias is significantly different (for example, gradient descent leads to margin maximization (Soudry et al., 2018; Ji & Telgarsky, 2019; Lyu & Li, 2020)). This simple yet illuminating insight allows us to derive several important consequences of LR.

Our most significant result is that the implicit bias of logit regularization is closely related to *Linear Discriminant Analysis* (LDA) (Fisher, 1936). To the best of our knowledge, this is the first established connection between label smoothing and LDA, despite the former being a well-studied modern deep learning technique and the latter a classical statistical cornerstone. Moreover, we show that this result extends far beyond label smoothing to a broad family of logit regularizers. We summarize our main contributions below:

- **The implicit bias of LR:** We demonstrate that logit regularization induces logit clustering around finite per-sample targets. For Gaussian data, we prove that this clustering objective drives the weight vector $\boldsymbol{S}$ to align *exactly* with the LDA direction. We show that this observation approximately extends to any distribution, as long as the logits are sufficiently clustered.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- To gain further insight into the effects of LR, we study a signal-plus-noise decomposition of the data that splits the noisy part of the input features into signal-aligned and orthogonal components. We show that, in this model, LR has two significant effects:
  1. **Shift in the Interpolation Threshold and Grokking:** In the high-dimensional limit $d, N \to \infty$ with $\lambda = d/N$ fixed, where d is the input dimension, we show that LR shifts the interpolation threshold from the standard separability limit $\lambda_c = 1/2$ (Cover, 1965; Gardner, 1988; Beck et al., 2025) to $\lambda_c = 1$. Notably, under weak regularization, we identify *grokking* dynamics (delayed generalization accompanied by non-monotonic test loss dynamics).
  2. **Invariance to Orthogonal Noise:** We prove that the optimal generalization accuracy is independent of the noise amplitude in the orthogonal direction. This allows us to construct a *phase diagram* delineating the specific regimes (in terms of the signal and orthogonal noise amplitudes) where logit regularization is beneficial.
- We validate our theoretical predictions numerically on Gaussian data and on a more realistic distribution using ResNet-18 (He et al., 2015) embeddings from the CIFAR-10 dataset (Krizhevsky, 2009).

2 Related Work

Here, we provide a focused review of the prior research most central to our contribution. For an additional detailed literature survey, including variants of LS and technical high-dimensional solvers, we refer the reader to App. A.

Logit Regularization. Label smoothing (LS) (Szegedy et al., 2016) is a standard tool for improving generalization and calibration (Müller et al., 2019; Guo et al., 2017). It is a specific instance of a broader class of logit-space penalties, including entropy-based (Pereyra et al., 2017) and logit-norm (Dauphin & Cubuk, 2021) regularizers. While these are often motivated by the prevention of overconfidence, our work provides a geometric characterization of how these penalties replace unbounded margin growth with a *logit clustering* objective.

Implicit Bias and High-Dimensional Asymptotics. For unregularized classification, the implicit bias of gradient descent leads to the hard-margin SVM/max-margin separator (Soudry et al., 2018; Lyu & Li, 2020). We contrast this with the finite per-sample optima created by logit regularization. Our analysis connects to high-dimensional theories of linear separability (Cover, 1965; Gardner, 1988) and recent asymptotics for overparameterized classifiers (Montanari et al., 2019; Mignacco et al., 2020; Wang & Thrampoulidis, 2022).

Grokking and Neural Collapse. Grokking, a delayed transition to generalization, was originally observed in algorithmic tasks (Power et al., 2022). The most relevant works in this field relate the edge of separability in logistic regression with grokking transitions (Beck et al., 2025). In multi-class settings, our results are reminiscent of *neural collapse* (Papyan et al., 2020), a phenomenon in which logits arrange in simplex configurations (Zhou et al., 2022; Garrod & Keating, 2024).

3 The Implicit Bias of Logit Regularization

3.1 Setup: Logit Regularization as a Generalization of Label Smoothing

We study linear classifiers. For conceptual clarity, we focus initially on the binary setting with zero bias ($b = 0$). In this case, the model outputs a scalar logit $z_i = \boldsymbol{S}^\top \boldsymbol{x}_i$ where $\boldsymbol{x}_i$ are the data points, and $\boldsymbol{S}$ is the model weight vector. The extension to the multi-class setting is provided in App. F.

We consider a class of loss functions $\mathcal{L} = N^{-1} \sum_{i=1}^{N} \ell(z_i, y_i)$, where z_i is the logit and $y_i \in \{-1, +1\}$ is the label of the i-th sample, respectively. We consider the per-sample loss

$$\ell(z, y) = (1 - \alpha)\ell_{\mathrm{CE}}(z, y) + \alpha f(z), \tag{1}$$

where $\ell_{\mathrm{CE}}(z, y) = \log(1 + e^{-yz})$ is the standard cross-entropy loss, $f(z)$ is a convex and even *logit regularization* function, and $\alpha \in [0, 1)$ controls the regularization amplitude. Importantly, as shown in App. D, standard Label Smoothing (LS) is a specific instance of this framework, corresponding to

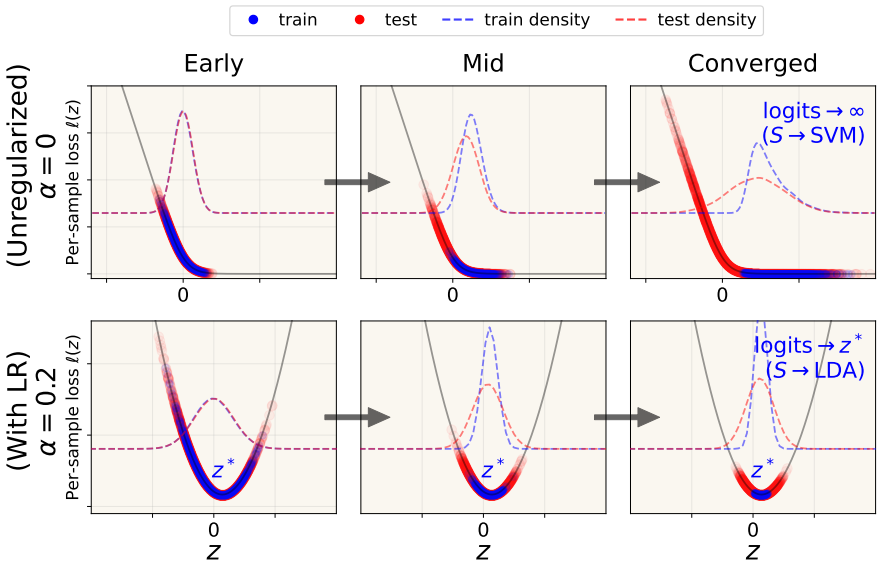


Figure 1: **Logit regularization induces clustering.** Logit evolution for a linear classifier trained on Gaussian data (see App. H.7 for more details regarding the setup). The train (blue) and test (red) logits are visualized across three training epochs (early, middle, and final). Notice that the total train/test loss is just the average over the blue/red points. Samples are classified correctly iff $z > 0$. **Top ($\alpha = 0$):** The unregularized logits are pushed toward infinity to maximize margins, indicating overconfidence. **Bottom ($\alpha > 0$):** The regularized loss exhibits a distinct finite minimum, driving the logits to cluster tightly around a target value z^* during training.

the choice $f(z) = \log(2 + 2\cosh z)$. While most of our results do not depend on the exact form of f, for simplicity, we will consider $f(z) = z^2$ in all our numerical results.

Since ℓ_{CE} depends only on the product yz and f is even, the loss depends strictly on the signed logit. For notational simplicity, from here on we redefine $z \leftarrow yz$. Equivalently, we absorb the label into the input vector, $\boldsymbol{x} \leftarrow y\boldsymbol{x}$, such that the logit remains $z_i = \boldsymbol{S}^\top \boldsymbol{x}_i$. This allows us to analyze the univariate function $\ell(z)$ independently of the class label.

Distinction from L_2 regularization: We emphasize that this formulation differs significantly from standard parameter regularization. While logit regularization acts on the model logits independently, an L_2 penalty implicitly couples data samples via the global norm. This prevents decomposing the loss into independent per-sample terms, a property central to all of our analysis throughout this work.

3.2 Logit Clustering

The introduction of logit regularization fundamentally alters the optimization landscape. In the unregularized case ($\alpha = 0$), the per-sample cross-entropy loss ℓ from Eq. (1) is monotonic, and the optimal solution maximizes the margins (Soudry et al., 2018; Ji & Telgarsky, 2019). When the data is linearly separable, this implicit bias drives the weight norm $\|\boldsymbol{S}\| \to \infty$, potentially leading to overconfidence and overfitting (see Figure 1, top row). In contrast, when $\alpha > 0$, the per-sample loss $\ell(z)$ has a unique finite global minimum denoted by z^*. Clearly, the loss has a data-independent lower bound, $\mathcal{L} \geq \ell(z^*)$, which is achieved only if $z_i = z^*$ for all samples. However, given the linear constraint $z_i = \boldsymbol{S}^\top \boldsymbol{x}_i$, the model generally cannot collapse all samples to a single point. Consequently, the optimization process becomes a geometric compromise: the model seeks a weight vector $\boldsymbol{S}$ that **clusters all logits as tightly as possible around the target z^*.** This phenomenon is visualized in the bottom row of Figure 1. Although logit clustering is a known effect of label smoothing (Müller et al., 2019), this simple explanation is usually overlooked, and we will show that it has several remarkable consequences.

3.3 ALIGNMENT WITH LDA DIRECTION

In this section, we derive the exact form of the implicit bias of logit regularization for logistic regression under idealized conditions, and show that logit clustering in this case drives the weight vector to be in the direction of LDA. Specifically, we consider two regimes: (i) Gaussian data with any convex regularizer, and (ii) arbitrary data distributions with a quadratic per-sample loss[1]. In both cases, we prove that the optimization problem is equivalent to minimizing the *coefficient of variation* (CV) $r(\boldsymbol{S}) \equiv \sigma(\boldsymbol{S})/\mu(\boldsymbol{S})$, where $\mu(\boldsymbol{S}), \sigma(\boldsymbol{S})$ are the logits mean and standard deviation, as a function of the weights.

Proposition 3.1 (Gaussian Data). *Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} \neq \mathbf{0} \in \mathbb{R}^d$, and consider a vector $\boldsymbol{S} \in \mathbb{R}^d$. The logit $z(\boldsymbol{S}) = \boldsymbol{S}^T\boldsymbol{x}$ is Gaussian with mean $\mu(\boldsymbol{S}) = \boldsymbol{S}^T\boldsymbol{\mu}$ and variance $\sigma^2(\boldsymbol{S}) = \boldsymbol{S}^T\Sigma\boldsymbol{S}$. Define $L(\boldsymbol{S}) = \mathbb{E}\left[\ell(z(\boldsymbol{S}))\right]$, where $\ell(z)$ is a convex function with a unique minimum.*

Then, the optimum weight vector $\boldsymbol{S}_{\min} = \arg\min_{\boldsymbol{S}} L(\boldsymbol{S})$ minimizes the ratio $r(\boldsymbol{S})$=$\sigma(\boldsymbol{S})/\mu(\boldsymbol{S})$. That is, $r(\boldsymbol{S}_{\min}) \leq r(\boldsymbol{S}), \forall \boldsymbol{S} \in \mathbb{R}^d$.

Proof. Let $Q : \mathbb{R}^d \to \mathbb{R}^2$ be the mapping from weights to logit moments, $Q(\boldsymbol{S}) = (\mu(\boldsymbol{S}), \sigma(\boldsymbol{S}))$. Since z is normally distributed, L is completely determined by its moments $\mu(\boldsymbol{S})$ and $\sigma(\boldsymbol{S})$ and optimization in weight space is equivalent to a constrained optimization problem in moment space. Let $\Omega \subset \mathbb{R}^2$ be the set of attainable moments, i.e., the image of $\mathbb{R}^d$ under Q.

Since both $\mu(\boldsymbol{S})$ and $\sigma(\boldsymbol{S})$ scale linearly in $\boldsymbol{S}$ (e.g., $\mu(\lambda\boldsymbol{S}) = \lambda\mu(\boldsymbol{S})$), each ray in weight space is mapped to a ray in Ω. We also note that, since $r(\boldsymbol{S})$ is a continuous function and its domain is connected, its image must also be connected. Therefore, Ω is a cone composed of rays emanating from the origin:

$$\Omega = \{(\mu, r\mu) \mid \mu > 0,\ r_1 \leq r \leq r_2\}, \tag{2}$$

where we assumed without loss of generality that $\mu > 0$ and r_2 may be infinite. For illustration of Ω, see the left-most panel of Figure 2.

Using standard convexity arguments, it is easy to show that for any fixed mean μ, L is an increasing function of the spread σ (see Lemma E.1 in the Appendix). Therefore, the optimum in Ω is attained on its lower boundary, namely the ray with the minimal slope r_1. Thus, $\boldsymbol{S}_{\min}$ is mapped under Q to the minimal $r = \sigma/\mu$. An alternative proof can be found in App. E.3. □

Corollary 3.2 (Quadratic Loss for any data). *For any data distribution, if the per-sample loss is quadratic, i.e., $\ell(z) = a(z - z^*)^2$, then the optimal direction $\hat{\boldsymbol{S}}_{\min} = \frac{\boldsymbol{S}_{\min}}{\|\boldsymbol{S}_{\min}\|}$ minimizes the ratio $r = \sigma/\mu$.*

Proof. If ℓ is a quadratic function then the loss is completely determined by the first and second moments of z:

$$L(\boldsymbol{S}) = \mathbb{E}\left[\ell(z(\boldsymbol{S}))\right] = a\left[\sigma^2(\boldsymbol{S}) + (\mu(\boldsymbol{S}) - z^*)^2\right].$$

Therefore, the argument of Proposition 3.1 applies verbatim. □

We now state the main result: under these conditions, the implicit bias of logit regularization aligns the weights with the LDA solution.

Corollary 3.3 (LDA direction). *If the data is Gaussian, the optimal weight direction is $\hat{\boldsymbol{S}}_{\min} \propto \Sigma^{-1}\boldsymbol{\mu}$. Similarly, if $\lambda = d/N < 1$ and the per-sample loss is quadratic, $\hat{\boldsymbol{S}}_{\min} \propto \Sigma^{-1}\boldsymbol{\mu}$ (where $\boldsymbol{\mu}, \Sigma$ are the empirical mean and centered covariance).*

Proof. Identifying $r(\boldsymbol{S})^{-1}$ as the square root of the Fisher's Criterion (i.e., $(\boldsymbol{S}^\top\boldsymbol{\mu})^2/\boldsymbol{S}^\top\Sigma\boldsymbol{S}$), the result follows immediately from standard LDA theory. See App. E.1 for a direct proof. □

Outside these idealized settings, we find empirically that the direction of the solution remains close to this theoretical limit, as we will show in the next section.

[1] Almost any per-sample loss as defined in Eq. (1) can be considered quadratic close enough to its minimum.

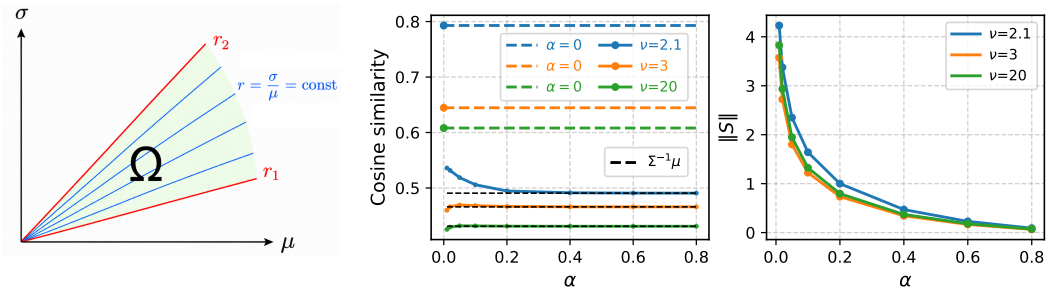


Figure 2: **Left panel:** Illustration of the cone Ω from Eq. (2). **Middle panel:** The cosine similarity between $\boldsymbol{S}_{\min}$ and the feature-axis, where the data is drawn from a Student's t-distribution for several values of ν. The dashed (colored) lines indicate the unregularized values ($\alpha = 0$). We observe an abrupt shift from $\alpha = 0$ to $\alpha > 0$, followed by a plateau that becomes flatter as ν increases (approaching a Gaussian distribution). The black dashed line shows, for comparison, the limiting LDA expected value, $\Sigma^{-1}\mu$, where μ and Σ are the empirical mean and centered covariance. **Right panel:** The norm $\|\boldsymbol{S}_{\min}\|$, which, in contrast, exhibits a clear dependence on α. See App. H.7 for more details about the numerical setup.

3.4 Beyond the Gaussian Case

We now turn to analyze the practical consequences of this result. Even though the direction of the solution would align with LDA only when the data is purely Gaussian or the per-sample loss is completely quadratic, in practice, we expect this to hold approximately (i.e., the deviation from this direction would not be large) for a *much broader* set of cases: As discussed in Sec. 3.2, logit regularization tends to cluster the logits around the loss minimum. Therefore, **locally**, the optimal per-sample loss may be approximated as quadratic. This suggests, as we will demonstrate numerically below, that the solution will be close to the LDA direction even when the distribution is very different from Gaussian.

Geometric Intuition. First, we will provide additional intuition. Minimizing the loss requires placing the logits as tightly as possible around z^*, the minimum of $\ell(z)$. This implicitly decouples the optimization into two distinct tasks:

1. **Direction Optimization:** choosing a direction $\hat{\boldsymbol{S}} = \boldsymbol{S}/\|\boldsymbol{S}\|$ that makes the logit cluster as tight as possible relative to its mean, i.e., minimizes the coefficient of variation $r = \sigma/\mu$.
2. **Mean Optimization:** scaling $\|\boldsymbol{S}\|$ so that the logit mean μ lies near the preferred value z^*.

The first task is purely geometric: it asks for the projection direction along which the signed data have maximal mean relative to their spread. This is precisely Fisher's criterion (the best signal-to-variance ratio), which gives the LDA direction. Also, notice that we expect the direction to be robust to the choice of the per-sample loss $\ell(z)$ as it mainly determines the scale through z^*, and therefore affects $\|\boldsymbol{S}\|$ rather than $\hat{\boldsymbol{S}}$ (which is indeed the case, as we will show numerically below).

Empirical Evidence. We will now provide numerical evidence that the solution's alignment with the LDA direction is robust, even when the distribution is far from being Gaussian. In the middle and right panels of Figure 2, we consider a binary classification experiment where the data is the mean vector $\pm(1, 0, 0, \dots)$ corresponding to the class label (arbitrarily aligned with the first axis), plus a random noise vector where each component is drawn independently from a Student-t distribution with degrees of freedom ν. This distribution provides a deliberately non-Gaussian test case: for small ν it is heavy-tailed, so the logit distribution can deviate substantially from Gaussianity. Here, the logit regularization function is taken to be $f(z) = z^2$. This is an illustrative example of a more general data model which will be discussed in detail in Sec. 4.

The middle panel of Figure 2 shows the cosine similarity between the learned $\hat{\boldsymbol{S}}_{\min}$ and the true feature axis. We observe a sharp transition from the unregularized state ($\alpha = 0$) to $\alpha > 0$, indicating the *change of the implicit bias*, followed by a stable plateau for $\alpha > 0$. The dashed black line represents the **LDA solution**. The plateau is flatter for larger ν (approaching Gaussianity) or larger α

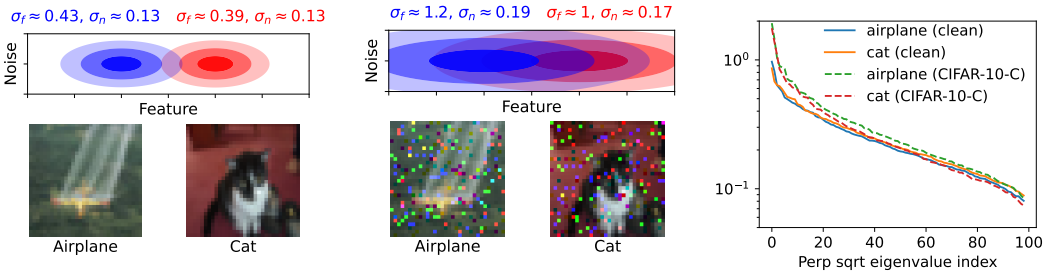


Figure 3: **Effect of Input Noise on Feature Geometry. Left and middle panels:** Visualization of penultimate layer features for two classes (Planes vs. Cats) from ResNet-18. The horizontal axis represents the signal direction (connecting class means), while the vertical axis represents the effective orthogonal noise radius. Left: Clean CIFAR-10 data. Middle: Noisy CIFAR-10C data. **Right panel:** The eigenspectrum of the orthogonal noise covariance matrix. For more details, see Appendix H.7.

(approaching a quadratic loss). In stark contrast, the right panel shows that the norm $\|\boldsymbol{S}_{\min}\|$ varies significantly with α. This confirms that logit regularization modifies the scale of the weights to satisfy the per-sample minimum, while the direction (and thus the accuracy) remains robust and close to the LDA solution. For additional numerical evidence, see App. E.2.

4 Signal-plus-Noise Decomposition

Building on the geometric insight of logit clustering, we now derive several surprising consequences of logit regularization when the data distribution has a signal-plus-noise structure.

4.1 Data model

We model the data geometry via a decomposition into a signal subspace and an orthogonal noise subspace. The input $\boldsymbol{x} \in \mathbb{R}^d$ is generated according to

$$\boldsymbol{x} = y\mu_f \boldsymbol{e}_1 + \boldsymbol{\xi}, \tag{3}$$

where $y = \pm 1$ is the label, $\boldsymbol{e}_1$ denotes the signal direction, and μ_f represents the signal magnitude. The noise vector $\boldsymbol{\xi}$ is drawn from a zero-mean distribution with anisotropic scaling. Specifically, we define $\boldsymbol{\xi} = \sigma_f \xi_f \boldsymbol{e}_1 + \sigma_n \boldsymbol{\xi}_\perp$, where ξ_f and the components of $\boldsymbol{\xi}_\perp$ are drawn from distributions $\mathcal{D}_f$ and $\mathcal{D}_n$ (signal and noise) respectively, with zero mean. Here, σ_f and σ_n act as mixture weights. We note that in the case of $\sigma_f = \sigma_n$ and Gaussian noise, our data generation process coincides with the noisy Gaussian mixture classification problem studied in Mignacco et al. (2020) and subsequent works.

Unless explicitly stated otherwise (e.g., in the ResNet-18 validation), we take the noise distributions to be Gaussian and align the signal with the x_1-axis with unit amplitude (i.e., $\mu_f = 1$). Under this assumption, σ_f and σ_n represent the standard deviations of the signal-aligned and orthogonal noise components, respectively. Formally, the noise vector $\boldsymbol{\xi}$ has independent entries satisfying $\xi_1 \sim \mathcal{N}(0, \sigma_f^2)$ and $\xi_i \sim \mathcal{N}(0, \sigma_n^2)$ for all $i > 1$. The full details regarding the setup of each figure appear in Appendix H.7.

4.2 Empirical Motivation

To motivate our construction, we show that the penultimate representations of a NN trained on real-world images exhibit a geometric structure that aligns with our signal-plus-noise decomposition. Analyzing the second-order statistics, we show that the noise structure is anisotropic with respect to the task as in our suggested model.

Specifically, we consider the penultimate layer of a ResNet-18 trained on ImageNet and evaluated on CIFAR-10 (clean) versus CIFAR-10C (Hendrycks & Dietterich, 2019) (noisy). We focus on a binary task (Planes vs. Cats), and calculate for each class the empirical class means $\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_i \boldsymbol{x}_i$ and the disjoint empirical centered covariance matrices $\Sigma_c = \frac{1}{N_c}\sum_i (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^\top$.

To understand the noise geometry, we project these covariance matrices onto two subspaces: the signal subspace, spanned by the difference of the class means $\boldsymbol{v} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and its orthogonal complement. Computing the standard deviation along the signal axis and the root-mean-square (RMS) of the eigenvalues in the orthogonal space, we get empirical equivalents of σ_f and σ_n. As shown in Figure 3, the intra-class variance along the signal direction (σ_f) differs significantly from the variance in the orthogonal directions (σ_n). Furthermore, when the input images are corrupted (using CIFAR-10C), these noise components grow.

These observations suggest that a single isotropic noise parameter cannot fully capture the geometry of real-world representations, which justifies our choice of a more expressive two-parameter model (defined by $\mathcal{D}_f$ and $\mathcal{D}_n$). We will use this exact empirical setting in App. C to validate our main theoretical predictions on real-world data.

5 The Noiseless Feature Regime

We begin by investigating logistic regression under the signal-plus-noise data decomposition, in the regime where the feature noise in the signal direction vanishes ($\sigma_f = 0$), yielding several surprising results. While idealized, the phenomena derived below persist approximately whenever the signal noise is sufficiently small ($\sigma_f \ll \mu_f, \sigma_n$).

5.1 Shift of interpolation threshold

We analyze the high-dimensional proportional limit where $d, N \to \infty$, while the ratio $\lambda = d/N$ remains fixed. In the unregularized setting, it is a known result that the interpolation threshold is $\lambda_c = 1/2$ (Gardner & Derrida, 1989). The reason is tied to *separability*: for $\lambda > 1/2$, points in general position are linearly separable regardless of their labels. This geometric freedom allows the optimizer to satisfy the margin constraints by discovering separating directions distinct from the "true" signal $\boldsymbol{S} = \boldsymbol{e}_1$, leading to overfitting (Beck et al., 2025). Notably, we will now show that, unlike standard regularization, the introduction of (*any*) convex logit regularization will abruptly shift the interpolation threshold to $\lambda_c = 1$.

Proposition 5.1 (Interpolation threshold is $\lambda_c = 1$)**.** *Let $\sigma_f = 0$ and $\alpha > 0$. Then, the minimizer of the training loss achieves perfect generalization accuracy if and only if $\lambda = d/N < 1$.*

Proof. With $\sigma_f = 0$, the signal component is deterministic: $y_i \boldsymbol{e}_1^\top \boldsymbol{x}_i = \mu_f$ for all i. Since $\alpha > 0$, the per-sample loss is strictly convex with a unique global minimum z^*.

We can construct a weight vector aligned purely with the signal, $\boldsymbol{S} = (z^*/\mu_f)\boldsymbol{e}_1$, such that $y_i \boldsymbol{S}^\top \boldsymbol{x}_i = z^*$ for all samples. This configuration achieves the global unconstrained lower bound of the total loss (the sum of per-sample minima). This solution is unique because the data matrix has full column rank when $N > d$, and since it aligns perfectly with $\boldsymbol{e}_1$, it yields 100% test accuracy.

Conversely, for $\lambda > 1$, the data matrix has a non-trivial null space. The loss minimizer takes the form $\boldsymbol{S}_{\min} = (z^*/\mu_f)\boldsymbol{e}_1 + \boldsymbol{S}_\perp$, where $\boldsymbol{S}_\perp \neq \boldsymbol{0}$ lies in this null space, degrading test accuracy. □

Figure 12 demonstrates the fact that the logits collapse to a single point in this regime. Notice that this result is independent of the regularization function $f(z)$ (consistent with our previous discussion) and the orthogonal noise scale σ_n (which is not unique to $\sigma_f = 0$, as discussed below). This suggests that for sufficiently small σ_f, logit regularization is highly beneficial, effectively extending the generalization regime to $\lambda < 1$. We examine the case of non-negligible σ_f in the following section.

Grokking. Remarkably, this model can serve as a tractable framework to observe grokking — a phenomenon that attracted much recent interest, in which there is delayed generalization accompanied by a non-monotonic test loss. Additional details can be found in App. B.

6 Invariance to the Noise Scale

In this section, we show that in the presence of logit regularization, the optimal generalization accuracy becomes independent of the orthogonal noise scale σ_n. We then use this result to map the specific regimes where logit regularization outperforms the unregularized baseline.

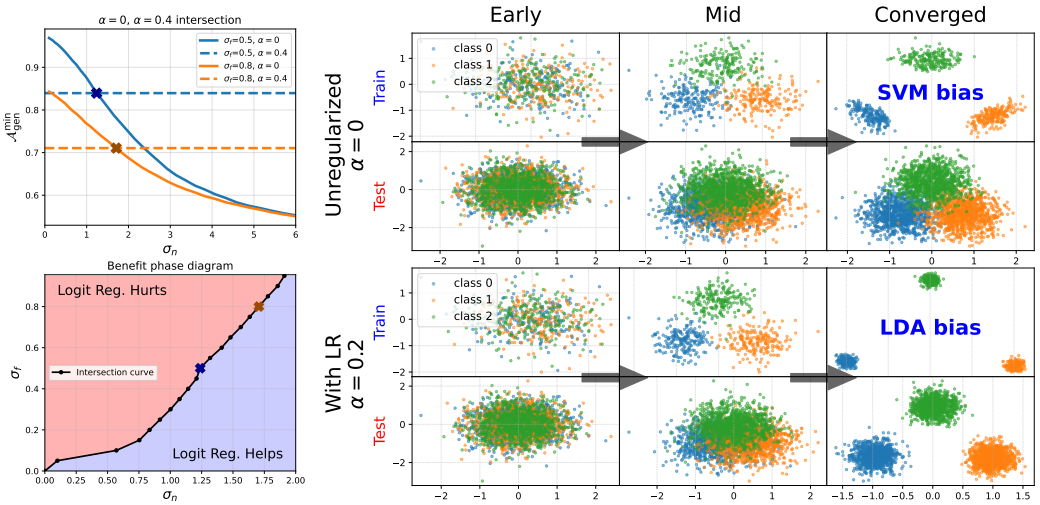


Figure 4: **Left panels:** *Benefit phase diagram.* **Top:** Final generalization accuracy as a function of σ_n, comparing the unregularized ($\alpha = 0$) and regularized ($\alpha > 0$) cases, for $\lambda = 0.7$. **Bottom:** Phase diagram showing regions where logit regularization is beneficial. The boundary curve is constructed from the intersection points (as shown in the top panel) for different values of σ_f. For illustration, the intersection points of the unregularized and regularized curves for $\sigma_f = 0.5, 0.8$ (top) are marked as "X"s of corresponding colors on the boundary curve (bottom). For an equivalent figure in the $\lambda < 1/2$ regime, see Figure 14 in the appendix. **Right panels:** *Multi-class Logit Clustering.* Projection of the data from three classes onto the subspace spanned by the relative weight differences ($S_1 - S_2$ and $S_1 - S_3$). The parameter is set to $\lambda = d/N = 0.7$: since $\lambda > 1/2$, the data is separable. **Top** ($\alpha = 0$)**:** Standard cross-entropy separates classes linearly, maximizing the margin. **Bottom** ($\alpha = 0.2$)**:** Logit regularization enforces tight clustering around the simplex vertices.

6.1 Setup and Proof

Proposition 6.1. *Let $\alpha > 0$, $\lambda = d/N < 1$, and fix σ_f, μ_f. Then, the optimal generalization accuracy $\mathcal{A}_{\text{gen}}^{\min}$ (the accuracy obtained at the optimal weight vector $\boldsymbol{S}_{\min}$) is strictly independent of σ_n.*

Proof. This follows from a simple scaling argument. Let $\boldsymbol{S}_{\min} = (S_{1,\min}, \boldsymbol{S}_{\perp,\min})$ denote the minimizer. The loss depends exclusively on the scalar logits $z_i = S_{1,\min} x_{i,1} + \boldsymbol{S}_{\perp,\min}^T \boldsymbol{x}_{i,\perp}$. Consider a rescaling of the noise variance $\sigma_n \to \beta\sigma_n$, which implies $\boldsymbol{x}_{i,\perp} \to \beta \boldsymbol{x}_{i,\perp}$. To maintain the optimal logit values $z_{\min}$ (and thus minimize the loss), the weights must scale inversely: $\boldsymbol{S}_{\perp,\min} \to \beta^{-1} \boldsymbol{S}_{\perp,\min}$. Consequently, the product $\boldsymbol{S}_{\perp,\min}^T \boldsymbol{x}_{i,\perp}$ remains invariant. Since the distribution of the test logits depends on this same product, the test accuracy is invariant to σ_n. □

Notably, while $\mathcal{A}_{\text{gen}}^{\min}$ is independent of σ_n, the direction $\boldsymbol{S}_{\min}$ is not. The cosine similarity with the feature axis $\boldsymbol{e}_1$, denoted by ρ, explicitly depends on the noise. Since $\|\boldsymbol{S}_{\perp,\min}\| \propto \sigma_n^{-1}$, the alignment behaves as:

$$\rho_{\min} = \frac{S_{1,\min}}{\|\boldsymbol{S}_{\min}\|} = \frac{1}{\sqrt{1 + (C/\sigma_n)^2}}, \tag{4}$$

where C depends on σ_f but is independent of σ_n (see Appendix H.6).

To demonstrate why the accuracy remains constant despite the changing direction of $\boldsymbol{S}$, we can consider Gaussian data, where the generalization accuracy is:

$$\mathcal{A}_{\text{gen}} = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\rho\mu_f}{\sqrt{2\left(\rho^2\sigma_f^2 + (1-\rho^2)\sigma_n^2\right)}}\right)\right]. \tag{5}$$

Substituting $\rho_{\min}$ from Eq. (4) into Eq. (5) cancels the σ_n terms in the denominator, rendering the final accuracy explicitly independent of the noise scale. See Figure 15 in the appendix for more empirical evidence to the σ_n-invariance.

6.2 When does logit regularization help?

Since the regularized accuracy $\mathcal{A}^{\min}_{\alpha>0}$ is constant with respect to σ_n (while the unregularized accuracy $\mathcal{A}^{\min}_{\alpha=0}$ changes), one may find a critical noise threshold where these curves intersect. This intersection defines the point at which logit regularization becomes advantageous. For Gaussian data, we illustrate this crossover in the top-left panel of Figure 4. By tracing this threshold across varying σ_f, we construct the phase diagram shown in the bottom-left panel, presenting the regime where regularization improves performance.

7 Multi-class Extension and Empirical Validation

While our primary analysis focuses on the binary case, the core geometric mechanism of logit clustering naturally extends to the multi-class setting. In the multi-class case, convex logit regularization again creates finite per-sample targets in logit space. This yields a simplex-clustering analogue of the binary logit-clustering picture: the optimizer arranges the class logits near finite simplex-like vertices (e.g., right panels of Figure 4). We give the full multi-class formulation and the corresponding extensions of the interpolation-threshold and σ_n-invariance results in App. F and App. G.

Furthermore, to verify that our theoretical insights hold beyond idealized synthetic data, we validate our findings on real-world representations. By training a linear classifier on frozen penultimate-layer embeddings of a ResNet-18 model (as motivated in Sec. 4.2), we successfully recover both the sharp stabilization of the implicit bias upon applying logit regularization and the predicted robustness to orthogonal noise scaling. Full details and results for this empirical validation are detailed in App. C.

8 Conclusion

In this work, we have characterized the implicit bias of convex logit regularization, such as label smoothing, in linear classification. We showed that such regularization results in *logit clustering* around the finite minima of the per-sample loss function. This provides a straightforward geometric explanation for the known label smoothing effect, and demonstrates that the phenomenon is **not** unique to Label Smoothing, but is a general consequence of any loss that can be decomposed as a sum of per-sample losses with a finite minimum.

Remarkably, this clustering objective is intimately connected to classical statistical discrimination. We proved that for Gaussian data or quadratic loss functions, logit regularization drives the weight vector to align exactly with *Fisher's Linear Discriminant* ($\Sigma^{-1}\boldsymbol{\mu}$). Empirically, we demonstrated that the direction remains close to this solution even beyond the idealized conditions. This result bridges a conceptual gap, showing that modern "soft target" techniques like Label Smoothing (and other convex logit penalties) effectively drive the weights toward the optimal linear discriminator solution.

Leveraging a signal-plus-noise decomposition, we also proved that in the feature-noiseless regime LR shifts the interpolation threshold to $\lambda = 1$ (yielding perfect generalization in regimes where unregularized models fail) and induces grokking dynamics under weak regularization. Finally, we proved that regularized accuracy is invariant to orthogonal noise scaling, allowing us to map the precise phase space where regularization is beneficial.

Limitations and Future Work. Our analysis relies on a linear framework. While we validated our findings on fixed ResNet embeddings, extending the logit clustering perspective to end-to-end training (where the representation is learned jointly) remains a primary challenge. Additionally, while we characterized the implicit bias for Gaussian data as the minimization of the coefficient of variation (LDA direction), quantifying the exact deviation from it as a function of the distributions remains an open question. Future research should investigate how this "LDA-like" bias interacts with the inductive bias of deep networks during feature learning, as well as the efficacy of this broader class of convex penalties in practical deep learning settings.

AI USE STATEMENT

We have used generative AI tools to edit and polish text and figures for readability, and to generate and run code for numerical experiments. We have not used generative AI tools for any other tasks with required disclosure, such as generating synthetic data, developing theoretical models or conceptual frameworks, formulating mathematical claims or assisting with their proofs, data analysis, or any other substantive research tasks. All AI-assisted work was reviewed by the authors, and we take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All theoretical results are stated together with their assumptions in the main text, and complete proofs are provided in Apps. E.1, F and G. The synthetic signal-plus-noise model is defined in Sec. 4, and the exact parameters, optimizers, learning rates, and numbers of epochs for every figure are listed in App. H.7. The protocol for the ResNet-18/CIFAR-10 embedding experiments is described in App. C.

REFERENCES

Alon Beck, Noam Levi, and Yohai Bar-Sinai. Grokking at the edge of linear separability. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2410.04489.

Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021. URL https://arxiv.org/abs/2112.07572.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Yunqing Zhao, and Ngai-Man Cheung. Revisiting label smoothing and knowledge distillation compatibility: What was missing? In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2890–2916. PMLR, 2022.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3): 326–334, 1965. doi: 10.1109/PGEC.1965.264137.

Yann Dauphin and Ekin Dogus Cubuk. Deconstructing the regularization of batchnorm. In *International Conference on Learning Representations (ICLR)*, 2021.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2): 435–495, 2022. URL https://arxiv.org/abs/1911.05822.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7: 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Yingbo Gao, Weiyue Wang, Christian Herold, Zijian Yang, and Hermann Ney. Towards a better understanding of label smoothing in neural machine translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 212–223, 2020.

E. Gardner and B. Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22:1983–1994, 1989.

Elizabeth Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257–270, 1988.

Connall Garrod and Jonathan P. Keating. The persistence of neural collapse despite low-rank bias, 2024. URL https://arxiv.org/abs/2410.23169.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1321–1330. PMLR, 2017.

Li Guo, Keith Ross, Zifan Zhao, George Andriopoulos, Shuyang Ling, Yufeng Xu, and Zixuan Dong. Cross entropy versus label smoothing: A neural collapse perspective. *arXiv preprint arXiv:2402.03979*, 2024.

X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1903.12261.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of The 32nd Conference on Learning Theory (COLT)*, 2019. URL https://proceedings.mlr.press/v99/ji19a.html.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems (NeurIPS)*, 34:28648–28662, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Dongkyu Lee, Ka Chun Cheung, and Nevin L. Zhang. Adaptive label smoothing with self-knowledge in natural language generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9781–9792, 2022. URL https://arxiv.org/abs/2210.13459.

Noam Levi, Alon Beck, and Yohai Bar-Sinai. Grokking in linear estimators – a solvable model that groks without understanding. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=GH2LYb9XV0.

Jiajun Liang, Linze Li, Zhaodong Bing, Borui Zhao, Yao Tang, Bo Lin, and Haoqiang Fan. Efficient one pass self-distillation with zipf's label smoothing. In *European Conference on Computer Vision (ECCV)*, pp. 104–119. Springer, 2022.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1906.05890.

Francesca Mignacco, Florent Krzakala, Yue M. Lu, Pierfrancesco Urbani, and Lenka Zdeborová. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 6874–6883, 2020.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparameterized regime. *arXiv preprint arXiv:1911.01544*, 2019. URL https://arxiv.org/abs/1911.01544.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.02629.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 2020. URL https://arxiv.org/abs/2008.08186.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177.

Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the edge of numerical stability, 2025. URL https://arxiv.org/abs/2501.04697.

Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. Is label smoothing truly incompatible with knowledge distillation: An empirical study. *arXiv preprint arXiv:2104.00676*, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL https://arxiv.org/abs/1710.10345.

Peter Súkeník, Marco Mondelli, and Christoph Lampert. Neural collapse versus low-rank bias: Is deep neural collapse really optimal? *arXiv preprint arXiv:2405.14468*, 2024.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. In *Proceedings of the National Academy of Sciences*, 2019. URL https://arxiv.org/abs/1803.06964.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

A. N. Tikhonov and Vasiliy Yakovlevich Arsenin. Solutions of ill-posed problems. 1977. URL https://api.semanticscholar.org/CorpusID:122072756.

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, 2022. URL https://arxiv.org/abs/2011.09148.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. URL https://arxiv.org/abs/2205.09310.

Guoxuan Xia, Olivier Laurent, Gianni Franchi, and Christos-Savvas Bouganis. Towards understanding why label smoothing degrades selective classification and how to fix it. In *International Conference on Learning Representations (ICLR)*, 2025.

Jing Xu and Haoxiong Liu. Quantifying the variability collapse of neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 38535–38550. PMLR, 2023.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3903–3911, 2020.

Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996, 2021.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:31697–31710, 2022.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *European Conference on Computer Vision (ECCV)*, pp. 518–536. Springer, 2022.

Appendices

A Additional Related Work

Label smoothing and soft-target training. Label smoothing (LS) was introduced as "label smoothing regularization" by Szegedy et al. (2016) and is now a standard tool for improving generalization and calibration in classification and sequence models (Müller et al., 2019; Guo et al., 2017). LS has been analyzed across domains, including neural machine translation (Gao et al., 2020). More broadly, LS is one instance of training with *soft targets*, closely related to knowledge distillation (Hinton et al., 2015; Yuan et al., 2020). The interaction between LS and distillation is nuanced, with studies investigating when LS-trained teachers help or harm KD (Shen et al., 2021; Chandrasegaran et al., 2022). Beyond fixed uniform smoothing, variants adapt the target distribution during training, including Online Label Smoothing (Zhang et al., 2021), Zipf label smoothing (Liang et al., 2022), and adaptive/self-knowledge smoothing (Lee et al., 2022). LS has also been linked to changes in feature geometry (Müller et al., 2019; Kornblith et al., 2021) and quantified via feature-variability measures (Xu & Liu, 2023). However, LS can be problematic for selective classification or failure prediction, motivating targeted fixes (Zhu et al., 2022; Xia et al., 2025).

Logit/output regularization beyond classic label smoothing. Methods that regularize the *logits* directly include entropy penalties (Pereyra et al., 2017), logit-norm regularizers (Dauphin & Cubuk, 2021), and logit normalization (Wei et al., 2022). A key theme is whether the loss remains *lower bounded* and admits a *finite* per-sample optimum, as in convex logit penalties, versus cross-entropy whose infimum is achieved only at infinite margin.

Implicit bias of cross-entropy and margin maximization. For separable linear classification, gradient descent on unregularized cross-entropy converges to the hard-margin SVM (Soudry et al., 2018). This has been extended to nonseparable data (Ji & Telgarsky, 2019) and overparameterized homogeneous models (Lyu & Li, 2020). Convex logit regularization provides a sharp contrast by forcing the per-sample objective to attain a finite minimum.

High-dimensional solvable models. Our analysis connects to classical work on the combinatorial "capacity" of linear separators (Cover, 1965; Gardner, 1988). Recent work develops precise high-dimensional asymptotics for max-margin classifiers via CGMT/AMP techniques (Montanari et al., 2019; Deng et al., 2022; Mignacco et al., 2020; Wang & Thrampoulidis, 2022). In related directions, high-dimensional logistic regression exhibits phase-transition phenomena tied to the existence of maximum-likelihood estimators (Sur & Candès, 2019; Celentano et al., 2021).

Grokking and delayed generalization. Grokking was popularized in neural networks trained on small algorithmic datasets (Power et al., 2022). Solvable models have since been developed for linear estimators (Levi et al., 2024) and logistic regression near the edge of separability (Beck et al., 2025). Logit regularization in this context was briefly discussed in Prieto et al. (2025); we identify an analogous delayed generalization region with a shifted threshold.

Simplex geometry and neural collapse. Our multi-class setup relates to the *neural collapse* phenomenon (Papyan et al., 2020; Han et al., 2022), where logits converge to structured simplex/ETF configurations. Recent work studies how different losses shape this collapse (Zhou et al., 2022; Guo et al., 2024) and how it interacts with low-rank bias (Súkeník et al., 2024; Garrod & Keating, 2024).

B Grokking Dynamics.

Grokking is a striking phenomenon in which a model initially perfectly overfits its training data, yielding poor test performance, but suddenly transitions to robust generalization after an extended period of training. Originally observed in complex neural networks, this delayed generalization is often accompanied by a non-monotonic test loss and has drawn significant recent attention.

Developing simple analytical models that exhibit grokking is crucial for isolating the mechanisms that induce it. Previous work has identified grokking in the unregularized version of this setup, occurring near the critical threshold $\lambda_c = 1/2$ (Beck et al., 2025). Notably, we find that logit regularization can also induce grokking dynamics across the broad region $1/2 < \lambda < 1$ — see Figure 5. This

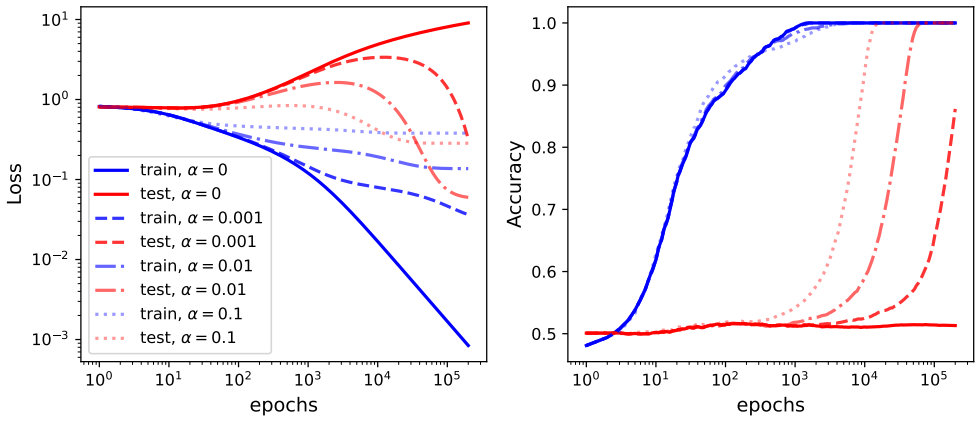


Figure 5: **Grokking.** Loss and accuracy curves for $\sigma_f = 0$ and $\lambda = 0.7$. While the unregularized model ($\alpha = 0$) overfits, logit regularization induces grokking (delayed generalization), with the delay diverging as α approaches zero.

occurs because the interpolation threshold is shifted to $\lambda_c = 1$. Consequently, the system exhibits delayed generalization as $\alpha \to 0$: the model initially overfits (in the direction of the unregularized implicit bias) before eventually converging to the generalizing solution. The timescale of this delay ("grokking time") diverges as $\alpha \to 0$, consistent with previous works showing that the grokking time diverges as some control parameter is close to criticality (Levi et al., 2024).

C Validation on Penultimate Layer Embeddings

The analysis in the main text was carried out in a linear setting with an idealized signal-plus-noise decomposition. We now validate two main qualitative predictions in a more realistic feature space: the sharp transition from $\alpha = 0$ to $\alpha > 0$, and the invariance to orthogonal noise scaling.

Setup. We use the same ResNet-18 penultimate-layer embeddings introduced in Figure 3. We freeze the representation and train only a linear classifier on top, using the loss in Eq. (17), with quadratic logit regularization $f(\boldsymbol{z}) = \|\boldsymbol{z}\|^2$. See App. H.7 for more details.

α-Insensitivity. We examine the sensitivity of the optimal generalization accuracy to the regularization strength α, see left panel of Figure 6. As predicted in Sec. 3.4, the accuracy jumps significantly and enters a stable plateau. The deviation from it at smaller α is due to the fact that the data is not a Gaussian.

σ_n-independence via synthetic orthogonal rescaling. To directly test the invariance in Proposition 6.1, we synthetically rescale only the component of each embedding that is orthogonal to the subspace spanned by the class means. This changes the effective σ_n without altering the class-mean geometry. The right panel of Figure 6 shows that the unregularized classifier ($\alpha = 0$) degrades substantially under this rescaling, whereas the regularized classifier ($\alpha > 0$) is more robust, according to our expectations.

D Label Smoothing as a Special Case of Logit Regularization

We show that label smoothing (LS) can be written as a convex *logit regularizer* added to cross-entropy, matching our form $\ell = (1 - \alpha)\ell_{\mathrm{CE}} + \alpha f$ (up to a reparameterization of α and additive constants).

Binary classification

Consider a two-class softmax. Let $t \equiv yz$ be the signed logit, so that $p_y = \sigma(t) = 1/(1 + e^{-t})$ and $p_{-y} = 1 - p_y = \sigma(-t)$. Standard cross-entropy is $\ell_{\mathrm{CE}}(t) = -\log p_y = \log(1 + e^{-t})$. Binary LS

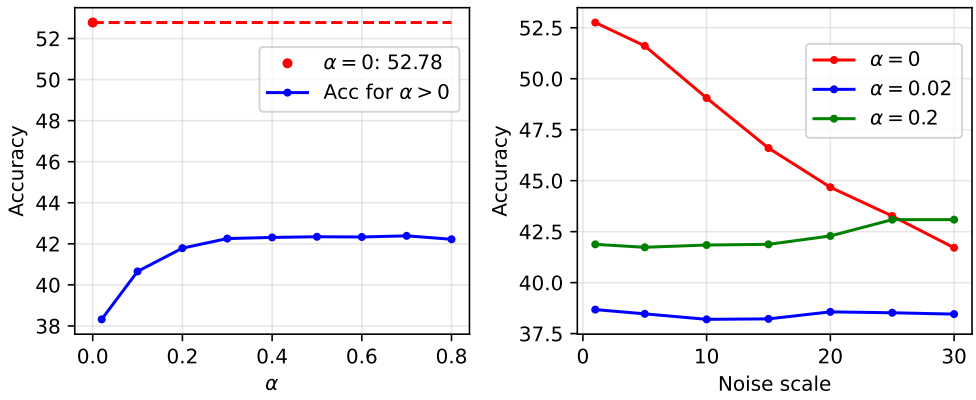


Figure 6: **Validation on ResNet-18 Penultimate Features. Left:** Generalization accuracy vs. regularization strength α. We observe a jump followed by saturation, confirming directional stability. **Right:** Accuracy vs. synthetic scaling of the orthogonal noise component. Logit regularization ($\alpha > 0$) makes the model insensitive to the noise scale, in contrast to the unregularized baseline.

with smoothing ε uses targets $(1-\varepsilon, \varepsilon)$ and yields

$$\ell_{\mathrm{LS}}(t) = -(1-\varepsilon)\log p_y - \varepsilon \log p_{-y} = (1-\varepsilon)\log(1+e^{-t}) + \varepsilon \log(1+e^{t}). \tag{6}$$

Using $\log(1+e^{t}) + \log(1+e^{-t}) = \log\big(2+2\cosh t\big)$, we rewrite

$$\ell_{\mathrm{LS}}(t) = (1-2\varepsilon)\,\ell_{\mathrm{CE}}(t) \;+\; \varepsilon\,\log\big(2+2\cosh t\big). \tag{7}$$

Thus, in the binary case LS is equivalent to logit regularization with

$$f_{\mathrm{LS}}(t) = \frac{1}{2}\log\big(2+2\cosh t\big), \qquad t = yz, \tag{8}$$

(up to rescaling $\alpha = 2\varepsilon$).

Multiclass classification

Let $p_k(\boldsymbol{z}) = \exp(z_k)/\sum_j \exp(z_j)$ and one-hot labels $y_k \in \{0,1\}$. LS replaces $y_k^{\mathrm{LS}} = (1-\varepsilon)y_k + \varepsilon/K$, giving

$$\begin{aligned}\mathcal{L}_{\mathrm{LS}}(\boldsymbol{z},\boldsymbol{y}) &= -\sum_{k=1}^{K} y_k^{\mathrm{LS}} \log p_k(\boldsymbol{z}) \\ &= (1-\varepsilon)\,\mathcal{L}_{\mathrm{CE}}(\boldsymbol{z},\boldsymbol{y}) \;+\; \frac{\varepsilon}{K}\sum_{k=1}^{K}\big[-\log p_k(\boldsymbol{z})\big].\end{aligned} \tag{9}$$

Hence LS is of the form $(1-\alpha)\mathcal{L}_{\mathrm{CE}} + \alpha f(\boldsymbol{z})$ with the convex, permutation-symmetric logit regularizer

$$f_{\mathrm{LS}}(\boldsymbol{z}) = \log\Big(\sum_{k=1}^{K} e^{z_k}\Big) \;-\; \frac{1}{K}\sum_{k=1}^{K} z_k, \tag{10}$$

E Complementary details for Sec. 3.4

In Sec. 3.4 we showed that under some assumptions, the direction of the weight vector becomes almost invariant to the logit regularization function in general, and in particular to the regularization strength α. We did by proving Proposition 3.1 and Corollary 3.2, where we showed if the distribution of the data is strictly gaussian, this will be true for any convex per-sample loss, and on the other hand if the per-sample loss is quadratic, then it will be true regardless of its minimum.

E.1 COMPLEMENTARY PROOFS

We begin by proving a Lemma which is needed for the proof of Proposition 3.1.

Lemma E.1. *Let x be a zero-mean random variable and let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then $\mathbb{E}[f(\lambda x)]$ is a nondecreasing function of λ for $\lambda > 0$.*

Proof. This is a standard convexity argument. Let's examine two lambdas $0 < \lambda_1 < \lambda_2$. Then λ_1 is a convex combination of 0 and λ_2,

$$\lambda_1 = \left(1 - \frac{\lambda_1}{\lambda_2}\right) \cdot 0 + \frac{\lambda_1}{\lambda_2}\lambda_2 \,.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[f(\lambda_1 x)] &= \mathbb{E}\left[f\left(\left(1 - \frac{\lambda_1}{\lambda_2}\right) \cdot 0 \cdot x + \frac{\lambda_1}{\lambda_2}\lambda_2 x\right)\right] \\
&\leq \left(1 - \frac{\lambda_1}{\lambda_2}\right) f(0) + \frac{\lambda_1}{\lambda_2}\mathbb{E}[f(\lambda_2 x)] \\
&\leq \mathbb{E}\left[\left(1 - \frac{\lambda_1}{\lambda_2}\right) f(\lambda_2 x) + \frac{\lambda_1}{\lambda_2} f(\lambda_2 x)\right] = \mathbb{E}[f(\lambda_2 x)]
\end{aligned}$$

Where we used the fact that x is zero-mean and therefore $f(0) = f(\mathbb{E}[\lambda_2 x]) \leq \mathbb{E}[f(\lambda_2 x)]$. □

The connection to Proposition 3.1: For fixed $\mu \in \mathbb{R}$, define $g(t) \equiv f(\mu + t)$, which is convex whenever f is convex. Applying Lemma E.1 with $x = Y \sim \mathcal{N}(0, 1)$ yields that

$$Q(\mu, \sigma) \;=\; \mathbb{E}\big[f(\mu + \sigma Y)\big] \;=\; \mathbb{E}\big[g(\sigma Y)\big]$$

is nondecreasing in $\sigma \geq 0$ for each fixed μ.

We will now provide the proof to Corollary 3.3 in the main text.

Theorem E.2 (Restatement of Corollary 3.3)**.** *If the data is Gaussian, the optimal weight direction is $\hat{\boldsymbol{S}}_{\min} \propto \Sigma^{-1}\mu$ and therefore independent of the convex regularization function f. Similarly, if $\lambda = d/N < 1$ and the per-sample loss is quadratic, $\hat{\boldsymbol{S}}_{\min} \propto \Sigma^{-1}\mu$ (where Σ, μ are the empirical mean and centered covariance) and therefore independent of the target value z^*.*

Proof. **Population Gaussian.** For $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and any $\boldsymbol{S}$, the logit $z = \boldsymbol{S}^\top \boldsymbol{x}$ is Gaussian with mean $\mu(\boldsymbol{S}) = \boldsymbol{S}^\top \boldsymbol{\mu}$ and variance $\sigma^2(\boldsymbol{S}) = \boldsymbol{S}^\top \Sigma \boldsymbol{S}$. By Proposition 3.1, any minimizer of $L(\boldsymbol{S}) = \mathbb{E}[\ell(z)]$ minimizes the ratio $r(\boldsymbol{S}) = \sigma(\boldsymbol{S})/\mu(\boldsymbol{S})$ over $\{\mu(\boldsymbol{S}) > 0\}$, i.e.

$$r(\boldsymbol{S})^2 = \frac{\boldsymbol{S}^\top \Sigma \boldsymbol{S}}{(\boldsymbol{S}^\top \boldsymbol{\mu})^2}.$$

Using Cauchy–Schwarz,

$$(\boldsymbol{S}^\top \boldsymbol{\mu})^2 = \langle \Sigma^{1/2}\boldsymbol{S}, \Sigma^{-1/2}\boldsymbol{\mu}\rangle^2 \leq (\boldsymbol{S}^\top \Sigma \boldsymbol{S})\,(\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu}),$$

hence $r(\boldsymbol{S})^2 \geq (\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu})^{-1}$. Moreover, equality holds iff $\Sigma^{1/2}\boldsymbol{S} \propto \Sigma^{-1/2}\boldsymbol{\mu}$, equivalently $\boldsymbol{S} \propto \Sigma^{-1}\boldsymbol{\mu}$.

Empirical quadratic. Let $\ell(z) = a(z - z^*)^2$ and define $\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_i \boldsymbol{x}_i$, $\hat{\Sigma} = \frac{1}{N}\sum_i (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^\top$. Assume that Σ is invertible (since $\lambda < 1$), then

$$\hat{L}(\boldsymbol{S}) = \frac{a}{N}\sum_i (\boldsymbol{S}^\top \boldsymbol{x}_i - z^*)^2 = a\,\boldsymbol{S}^\top \hat{\Sigma} \boldsymbol{S} + a(\boldsymbol{S}^\top \hat{\boldsymbol{\mu}} - z^*)^2,$$

so the optimal *direction* minimizes $\hat{r}(\boldsymbol{S})^2 = \frac{\boldsymbol{S}^\top \hat{\Sigma} \boldsymbol{S}}{(\boldsymbol{S}^\top \hat{\boldsymbol{\mu}})^2}$ and is independent of z^*. Repeating the Cauchy–Schwarz step from the Gaussian case gives the result.

□

E.2 Additional numerical evidence

We numerically illustrate the finite-sample mechanism behind the α-plateau discussed in Sec. 3.4. For each N, we sample $x_i \in \mathbb{R}^2$ i.i.d. from a Gaussian and compute the empirical minimizer

$$S_{\min}(\alpha) = \arg\min_S \frac{1}{N}\sum_{i=1}^{N}\Big[(1-\alpha)\log(1+e^{-x_i^\top S}) + \alpha(x_i^\top S)^2\Big].$$

As a reference direction, we also compute $S_{\min}^{\text{quad}}$ for the purely quadratic per-sample loss $b(z-a)^2$ on the same dataset (with values $b = a = 1$, but we also verified that it does not matter what are these values from Corollary 3.2), and plot $\tan\theta(\alpha)$ where $\theta(\alpha)$ is the angle between $S_{\min}(\alpha)$ and $S_{\min}^{\text{quad}}$.

Figure 7 shows that for small N, $S_{\min}(\alpha)$ can deviate at very small α (finite-sample non-Gaussianity), but it stabilizes for large α as the objective becomes effectively quadratic. For larger N, the plateau is reached at much smaller α, consistent with improved approximate Gaussianity.

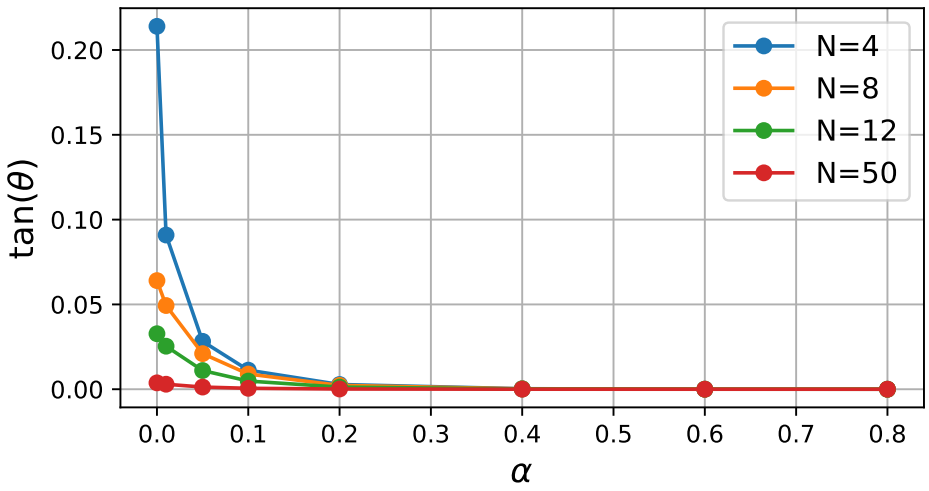


Figure 7: **Finite-sample α-plateau.** $\tan\theta(\alpha)$ between $S_{\min}(\alpha)$ for $(1-\alpha)\log(1+e^{-z}) + \alpha z^2$ and the reference direction $S_{\min}^{\text{quad}}$ for $b(z-a)^2$, for several N. Small N shows deviations at small α. Increasing N makes the plateau appear at even smaller α. Note that the smallest α in this figure is $\alpha = 0.00001$.

Two toy examples with $N = 3$ points. We next give two minimal $N = 3$ demonstrations that separate the "quadratic invariance" mechanism from the "Gaussian-logit" mechanism.

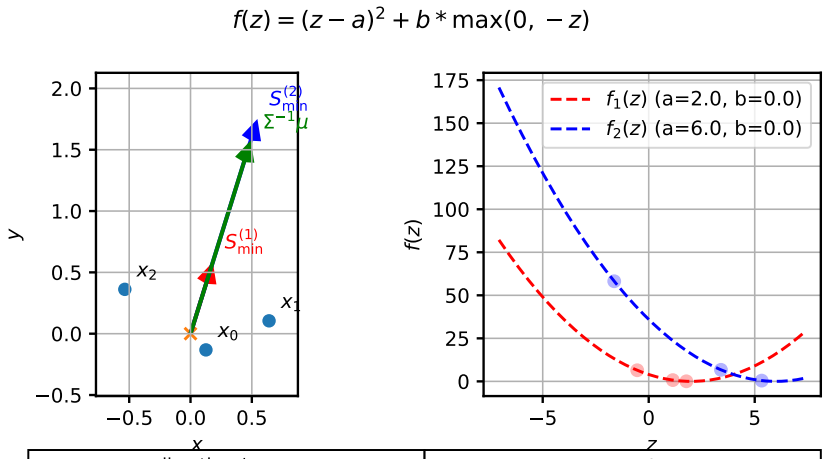
Figure 8: changing a only (quadratic case). We sample three points $x_1, x_2, x_3 \in \mathbb{R}^2$ in random position and minimize

$$\frac{1}{3}\sum_{i=1}^{3}\ell(z_i), \qquad z_i = S^\top x_i, \qquad \ell(z) = (z-a)^2 + b\,\max(0,-z).$$

In the first toy example we set $b = 0$, so the objective is purely quadratic in z. We compare two runs with $(a,b) = (2,0)$ and $(a,b) = (6,0)$. As shown in Figure 8, the minimizers $S_1^{\min}$ and $S_2^{\min}$ point in exactly the same direction and differ only by a scalar factor (left panel). Accordingly, the three logits $z_i = S^\top x_i$ shift to different absolute values between the two runs (right panel), but their "clustering ratio"

$$r = \frac{\sigma}{\mu}$$

(where μ and σ are the empirical mean and standard deviation over the three logits) remains constant: we report (μ, σ, r) in the table in the bottom row of the figure. This is a direct finite-sample illustration of the direction-independence in the quadratic regime (Corollary 3.2): changing the quadratic target a rescales the minimizer without changing its direction.



	direction / norm	μ, σ, r
Run 1 (a=2.0, b=0.0)	$S_{\min}/\|S_{\min}\| = (0.2975, 0.9547)$	$\mu = 0.7901,\ \sigma = 0.9777,\ r = 1.2375$
Run 2 (a=6.0, b=0.0)	$S_{\min}/\|S_{\min}\| = (0.2975, 0.9547)$	$\mu = 2.3703,\ \sigma = 2.9332,\ r = 1.2375$
Run 3 ($\Sigma^{-1}\mu$)	$\Sigma^{-1}\mu/\|\Sigma^{-1}\mu\| = (0.2975, 0.9547)$	$\mu = 0.6530,\ \sigma = 0.8081,\ r = 1.2375$

Figure 8: *Toy example ($N = 3$) with a quadratic loss: changing a rescales the minimizer but does not change its direction.* We minimize $\frac{1}{3}\sum_{i=1}^{3}(S^\top x_i - a)^2$ over $S \in \mathbb{R}^2$ for two choices of a (here $a = 2$ and $a = 6$, with $b = 0$). We also present for comparison the direction $\Sigma^{-1}\boldsymbol{\mu}$. **Left:** the three points x_1, x_2, x_3 and the two minimizers $S_1^{\min}, S_2^{\min}$, which are colinear. **Right:** the corresponding logits $\{S^\top x_i\}_{i=1}^3$ for both runs (shown in different colors), which shift in absolute value. **Bottom:** empirical μ (mean), σ (std), and $r = \sigma/\mu$, showing that the coefficient of variation is unchanged across the two runs.

Figure 9: adding a hinge-like term (non-quadratic, non-Gaussian). In the second toy example we keep $a = 2$ fixed and compare $(a, b) = (2, 0)$ versus $(a, b) = (2, 6)$. Now the objective includes the non-quadratic term $b\max(0, -z)$, and with only three points the induced logit distribution is extremely far from Gaussian. In this setting, neither of the sufficient conditions underlying our invariance arguments holds (quadratic loss or Gaussian logit distribution), and we indeed observe that the two minimizers are no longer perfectly colinear (though their angle remains moderate); see Figure 9.

The invariance to α in the case of the Gaussian model. In Figure 10 we show that for when the data is taken from Gaussian distribution, the direction is almost flat as a function of α, even for small values. We show it for different values of $\lambda = d/N$. The slight deviation from the plateau is expected from finite-sampling effects.

E.3 Alternative proof of Proposition 3.1 via monotonicity of $Q(r)$

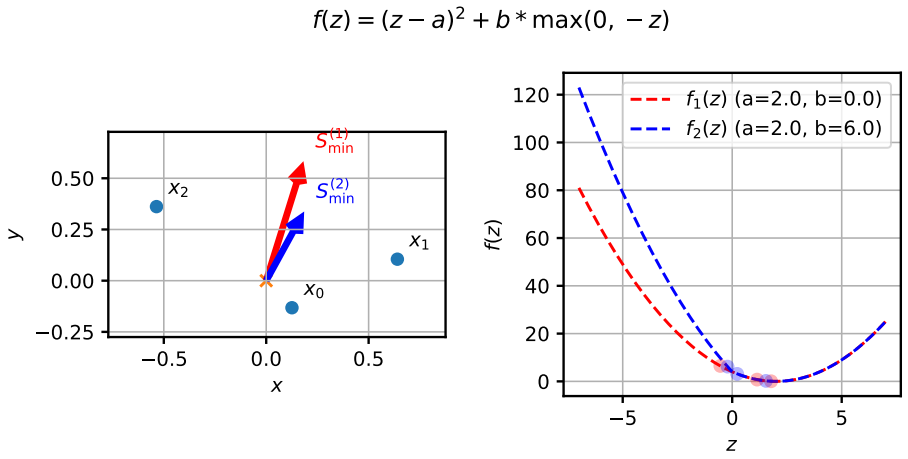
We provide an equivalent proof based on defining an objective $Q(r)$ directly and showing it is strictly increasing in the coefficient of variation r.

Proposition E.3 (Restatement of Proposition 3.1 for convenience)**.** *Let $\boldsymbol{x} \in \mathbb{R}^d$ be Gaussian, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} \neq \mathbf{0}$. Let $f : \mathbb{R} \to \mathbb{R}$ be convex, twice differentiable, and with a unique minimizer at $m > 0$. For $\boldsymbol{S} \in \mathbb{R}^d$ define the logit*

$$z = \boldsymbol{S}^\top \boldsymbol{x} \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu(\boldsymbol{S}) = \boldsymbol{S}^\top \boldsymbol{\mu}, \quad \sigma^2(\boldsymbol{S}) = \boldsymbol{S}^\top \Sigma \boldsymbol{S},$$

and the population objective $L(\boldsymbol{S}) = \mathbb{E}[f(z)]$. Then any minimizer $\boldsymbol{S}_{\min} \in \arg\min_{\boldsymbol{S}} L(\boldsymbol{S})$ also minimizes the ratio

$$r(\boldsymbol{S}) = \frac{\sigma(\boldsymbol{S})}{\mu(\boldsymbol{S})} \qquad \textit{over all } \boldsymbol{S} \textit{ such that } \mu(\boldsymbol{S}) > 0.$$



	$S_{\min}/\|S_{\min}\|$	$\mu,\ \sigma,\ r$
Run 3 (a=2.0, b=0.0)	$S_{\min}/\|S_{\min}\| = (0.2975, 0.9547)$	$\mu = 0.7901,\ \sigma = 0.9777,\ r = 1.2375$
Run 4 (a=2.0, b=6.0)	$S_{\min}/\|S_{\min}\| = (0.4811, 0.8767)$	$\mu = 0.5190,\ \sigma = 0.7453,\ r = 1.4361$

Figure 9: *Toy example ($N = 3$) where invariance need not hold: adding a hinge-like term changes the direction.* We minimize $\frac{1}{3}\sum_{i=1}^{3}\left[(S^\top x_i - a)^2 + b\max(0, -S^\top x_i)\right]$ for two settings: $(a, b) = (2, 0)$ and $(a, b) = (2, 6)$ on the same three points. Because the loss is no longer purely quadratic and the empirical logit distribution (with $N = 3$) is far from Gaussian, the two minimizers need not share the same direction, and we observe a noticeable (but not extreme) angular deviation between them.

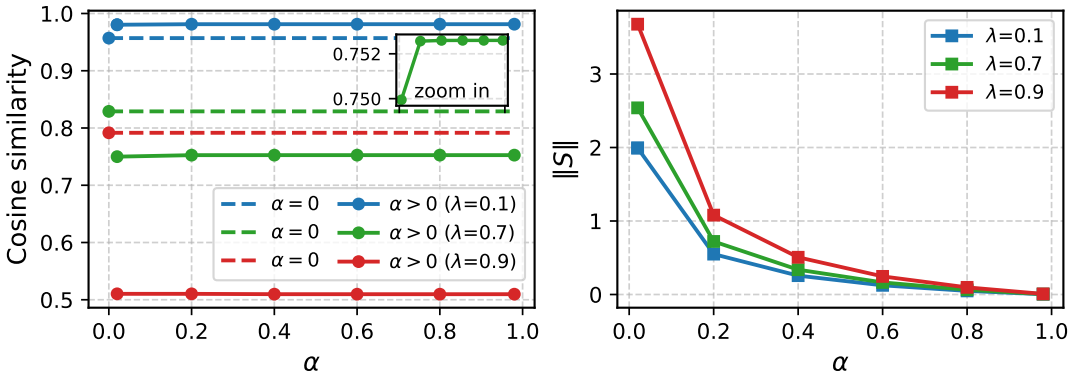


Figure 10: **Left panel:** Cosine similarity of $\boldsymbol{S}_{\min}$ with the x-axis. The dashed line indicates the unregularized value ($\alpha = 0$). We observe an abrupt shift from $\alpha = 0$ to $\alpha > 0$, followed by a stable plateau, indicating that the weight direction is effectively independent of α. **Right panel:** The norm $\|\boldsymbol{S}_{\min}\|$, which, in contrast, exhibits a clear dependence on α.

Proof. Write $\hat{\boldsymbol{S}} = \boldsymbol{S}/\|\boldsymbol{S}\|$ and define

$$\hat{z} \;=\; \hat{\boldsymbol{S}}^\top \boldsymbol{x} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2), \qquad \hat{\mu} = \hat{\boldsymbol{S}}^\top \boldsymbol{\mu}_x, \quad \hat{\sigma}^2 = \hat{\boldsymbol{S}}^\top \Sigma_x \hat{\boldsymbol{S}}.$$

We first note that any minimizing direction must satisfy $\hat{\mu} > 0$. Fix a unit direction $\hat{\boldsymbol{S}}$ and consider the one-dimensional restriction

$$h(t) \;=\; \mathbb{E}[f(t\,\hat{z})]\,, \qquad t = \|\boldsymbol{S}\| \geq 0.$$

By convexity of f, h is convex in t, and $h(0) = f(0)$ with

$$h'(0) = \mathbb{E}[f'(0)\,\hat{z}] = \hat{\mu}\, f'(0).$$

Since f has a unique minimizer at $m > 0$, convexity implies $f'(0) < 0$. If $\hat{\mu} < 0$, then $h'(0) > 0$ and by convexity h is minimized at $t = 0$. If $\hat{\mu} > 0$, then $h'(0) < 0$ and h attains its unique minimizer at some $t > 0$. Thus, flipping $\hat{\boldsymbol{S}} \mapsto -\hat{\boldsymbol{S}}$ strictly improves the loss near the origin whenever $\hat{\mu} < 0$, so every minimizing direction must satisfy $\hat{\mu} > 0$.

Assume henceforth $\mu(\boldsymbol{S}) > 0$. Let $Y \sim \mathcal{N}(0, 1)$ so that

$$z = \mu + \sigma Y = \mu(1 + rY), \qquad r = \frac{\sigma}{\mu}.$$

Define

$$q(\mu, r) \;=\; \mathbb{E}[f(\mu(1 + rY))]\,, \qquad \mu > 0,\ r \geq 0.$$

For a fixed direction $\hat{\boldsymbol{S}}$, scaling $\boldsymbol{S}$ by a positive factor scales both μ and σ by the same factor, hence leaves $r = \sigma/\mu$ invariant. Therefore,

$$L_{\min} = \min_{\hat{\boldsymbol{S}}} \min_{\mu > 0} q(\mu, r(\hat{\boldsymbol{S}})).$$

Define the reduced objective

$$Q(r) \;=\; \min_{\mu > 0} q(\mu, r),$$

so that

$$L_{\min} = \min_{\hat{\boldsymbol{S}}} Q(r(\hat{\boldsymbol{S}})).$$

It suffices to show Q is strictly increasing on $(0, \infty)$. Let $\mu^*(r)$ be a minimizer of $q(\mu, r)$ over $\mu > 0$, so $Q(r) = q(\mu^*(r), r)$. Differentiating and using $\partial_\mu q(\mu^*(r), r) = 0$ at the optimum (envelope theorem), we obtain

$$\frac{dQ}{dr} = \frac{\partial q}{\partial r}(\mu^*(r), r) = \mu^*(r)\, \mathbb{E}[f'(\mu^*(r)(1 + rY))\, Y]\,.$$

Using Gaussian integration by parts, $\mathbb{E}[g(Y)Y] = \mathbb{E}[g'(Y)]$ for $Y \sim \mathcal{N}(0, 1)$, with $g(y) = f'(\mu^*(1 + ry))$ and $g'(y) = r\mu^* f''(\mu^*(1 + ry))$, yields

$$\mathbb{E}[f'(\mu^*(1 + rY))\, Y] = r\mu^*\, \mathbb{E}[f''(\mu^*(1 + rY))]\,.$$

Hence

$$\frac{dQ}{dr} = r\, \mu^*(r)^2\, \mathbb{E}[f''(\mu^*(r)(1 + rY))]\,.$$

For $r > 0$, $\mu^*(r) > 0$, and convex f we have $f'' \geq 0$ (and > 0 a.e. if f is strictly convex), so $\frac{dQ}{dr} > 0$ for all $r > 0$. Therefore Q is increasing in r, and the minimizing direction must achieve the smallest attainable $r(\hat{\boldsymbol{S}}) = \hat{\sigma}/\hat{\mu}$. Equivalently, any $\boldsymbol{S}_{\min} \in \arg\min_{\boldsymbol{S}} L(\boldsymbol{S})$ minimizes $r(\boldsymbol{S}) = \sigma(\boldsymbol{S})/\mu(\boldsymbol{S})$ over $\mu(\boldsymbol{S}) > 0$. □

E.4 Explicit Derivation for the Quadratic Case

To build intuition, we consider the pedagogically illuminating case where the per-sample loss is purely quadratic, $\ell(z) = (z - a)^2$. In this setting, we can derive the optimal weight vector explicitly and demonstrate directly that minimizing the loss is equivalent to minimizing the coefficient of variation r.

Suppose that the loss is :

$$\mathcal{L}(\boldsymbol{S}) = \mathbb{E}\left[(z - a)^2\right], \tag{11}$$

where $z = \boldsymbol{S}^T x$. Let the direction of the weights be $\hat{\boldsymbol{S}} = \boldsymbol{S}/\|\boldsymbol{S}\|$ and the norm be $g = \|\boldsymbol{S}\|$. We denote the unit-direction moments as $\hat{\mu} = \mathbb{E}[\hat{\boldsymbol{S}} \cdot \boldsymbol{x}]$ and $\hat{\sigma}^2 = \mathrm{Var}(\hat{\boldsymbol{S}} \cdot \boldsymbol{x})$. The statistics of the full logit are then $\mu = g\hat{\mu}$ and $\sigma = g\hat{\sigma}$.

Using the bias-variance decomposition $\mathbb{E}[(z - a)^2] = \mathrm{Var}(z) + (\mathbb{E}[z] - a)^2$, the loss becomes:

$$\mathcal{L}(g, \hat{\boldsymbol{S}}) \approx \sigma^2 + (\mu - a)^2 = g^2\hat{\sigma}^2 + (g\hat{\mu} - a)^2. \tag{12}$$

We first optimize the scale g for a fixed direction $\hat{\boldsymbol{S}}$. Taking the derivative with respect to g:

$$\frac{\partial \mathcal{L}}{\partial g} = 2g\hat{\sigma}^2 + 2(g\hat{\mu} - a)\hat{\mu} = 0. \tag{13}$$

Solving for g, we obtain the optimal norm:

$$g^* = a\frac{\hat{\mu}}{\hat{\mu}^2 + \hat{\sigma}^2}. \tag{14}$$

Substituting this back into the expression for the mean logit $\mu^* = g^*\hat{\mu}$, we find:

$$\mu^* = a\frac{\hat{\mu}^2}{\hat{\mu}^2 + \hat{\sigma}^2} = \frac{a}{1 + r^2}, \tag{15}$$

where $r = \hat{\sigma}/\hat{\mu} = \sigma/\mu$ is the clustering ratio. This reveals a shrinkage effect: the optimal mean μ^* is slightly *smaller* than the target a to reduce the variance penalty.

Finally, substituting g^* back into the loss function yields the minimal loss achievable for the direction $\hat{\boldsymbol{S}}$:

$$\mathcal{L}_{\min}(\hat{\boldsymbol{S}}) = g^{*2}\hat{\sigma}^2 + (g^*\hat{\mu} - a)^2 = a^2\frac{\hat{\sigma}^2}{\hat{\mu}^2 + \hat{\sigma}^2} = a^2\frac{1}{1 + 1/r^2}. \tag{16}$$

Therefore, minimizing the total loss is mathematically equivalent to minimizing the coefficient of variation r.

F Generalization to Multi-class Setting

In this section, we extend our analysis to the K-class setting, demonstrating that the core geometric mechanism of logit clustering remains the same.

F.1 Setup

We consider a linear classifier producing logits $\boldsymbol{z} = W\boldsymbol{x} + \boldsymbol{b}$, where $W \in \mathbb{R}^{K\times d}$. The predicted class is $c = \mathrm{argmax}_k z_k$. Let $\boldsymbol{y}$ be the one-hot label vector for class c. We minimize the empirical loss $\mathcal{L} = \frac{1}{N}\sum_i \ell(\boldsymbol{z}^{(i)}, \boldsymbol{y}^{(i)})$, where the per-sample loss is defined as:

$$\ell(\boldsymbol{z}, \boldsymbol{y}) = (1 - \alpha)\ell_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}) + \alpha f(\boldsymbol{z}). \tag{17}$$

Here, $\ell_{\mathrm{CE}}(\boldsymbol{z}, \boldsymbol{y}) = -\sum_k y_k \log p_k$ is the standard cross-entropy loss, and $f(\boldsymbol{z})$ is a symmetric, convex regularization term. We note that standard LS is recovered by choosing (See Appendix D):

$$f_{\mathrm{LS}}(\boldsymbol{z}) = \log\left(\sum_{k=1}^{K} e^{z_k}\right) - \frac{1}{K}\sum_{k=1}^{K} z_k. \tag{18}$$

To extend the signal-plus-noise decomposition, we assume the class means $\{\boldsymbol{\mu}_c\}_{c=1}^K$ are affinely independent, spanning a $(K-1)$-dimensional centered signal subspace $\mathcal{S} = \mathrm{span}\{\boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}\}_{c=1}^K$. The input $\boldsymbol{x}$ for a given class c is modeled as $\boldsymbol{x} = \boldsymbol{\mu}_c + \boldsymbol{\xi}$. The noise vector $\boldsymbol{\xi}$ decomposes into signal-aligned and orthogonal components:

$$\boldsymbol{\xi} = \sigma_f\boldsymbol{\eta}_f + \sigma_n\boldsymbol{\eta}_n, \tag{19}$$

where $\boldsymbol{\eta}_f \in \mathcal{S}$ and $\boldsymbol{\eta}_n \in \mathcal{S}^\perp$ are drawn from standard isotropic distributions $\mathcal{D}_f$ and $\mathcal{D}_n$, scaled by σ_f and σ_n, respectively.

F.2 Simplex Clustering minimization

As in the binary classification case, the introduction of regularization ($\alpha > 0$) fundamentally alters the optimization landscape. The per-sample loss now possesses a finite minimum. Explicitly writing the per-sample loss for class c (where $\boldsymbol{y} = \boldsymbol{y}_c$ is a one-hot vector with 1 at index c), we obtain:

$$\ell(\boldsymbol{z}, \boldsymbol{y} = \boldsymbol{y}_c) = (1 - \alpha)\left(\log\left(\sum_j e^{z_j}\right) - z_c\right) + \alpha f(\boldsymbol{z}). \tag{20}$$

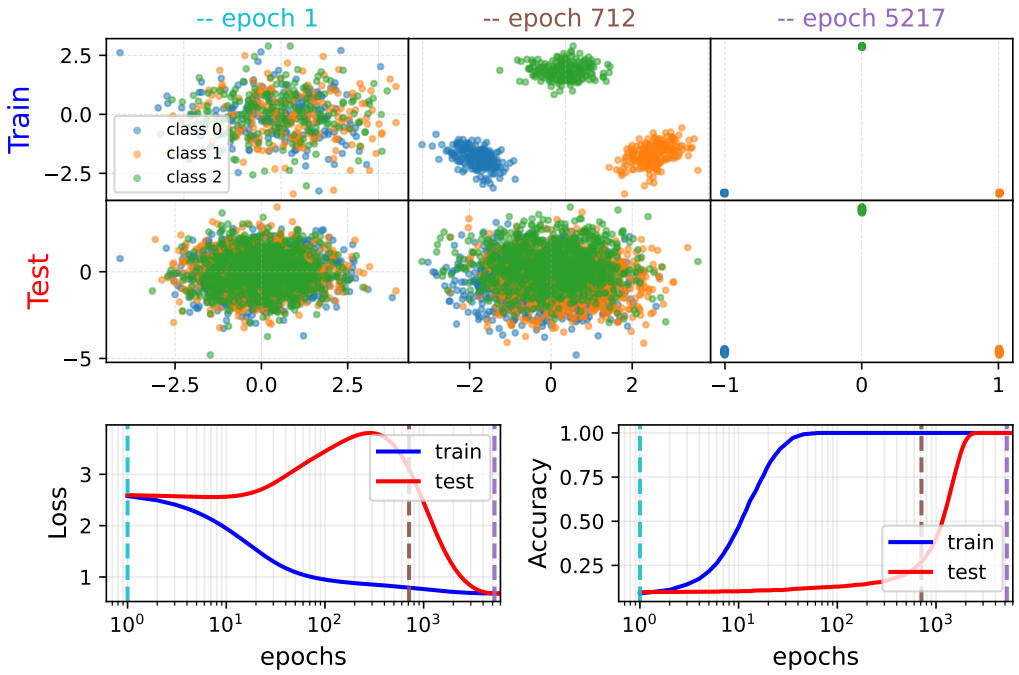


Figure 11: **Grokking in Multi-class and Small Logit Regularization.** Dynamics in the multi-class case, for $\sigma_f = 0$, $\lambda = 0.7$ and $\alpha = 0.04$. **Top:** Logit projections at three time-steps. The model initially learns the overfitting solution before eventually collapsing the minima of the simplex vertices, corresponding to the generalization phase. **Bottom:** Evolution of test accuracy and loss. We observe the characteristic grokking signature: the test loss initially rises (overfitting) before descending, and generalization accuracy is delayed.

For each class c, there exists a unique target logit vector $\boldsymbol{z}^*_{(c)}$ that minimizes this per-sample loss. Due to permutation symmetry, this vector takes a specific form: a high value z_{high} at index c, and a uniform lower value z_{low} elsewhere. This represents the maximal certainty permitted by the regularization, corresponding to a probability distribution that peaks at class c (with a value determined by α) and uniform elsewhere.

The set of target vectors $\{\boldsymbol{z}^*_{(1)}, \ldots, \boldsymbol{z}^*_{(K)}\}$ forms the vertices of a symmetric simplex in logit space. Consequently, the optimization objective transforms into a *clustering* problem: the model seeks a weight matrix W that collapses the logits of each class c as tightly as possible around its corresponding vertex $\boldsymbol{z}^*_{(c)}$.

This behavior is visualized in the numerical experiment presented in the right panels of Figure 4. We assume Gaussian distributions for both signal and noise components. The model is trained with a quadratic regularization term $f(\boldsymbol{z}) = \|\boldsymbol{z}\|^2$. We then adopt the visualization scheme introduced in Müller et al. (2019), projecting the data of three classes onto the plane spanned by the class templates at three distinct time-steps during training. In the unregularized regime ($\alpha = 0$), the (projected) logits scatter to maximize separability. In contrast, regularization ($\alpha > 0$) forces them to cluster at the simplex vertices.

Most of our key results from the binary setting extend naturally to the multi-class case: see App. G below for details. For example, Figure 11 demonstrates our main finding regarding the noiseless regime: for $1/2 < \lambda < 1$ and small α, we observe grokking. First, the training samples become fully separated while the model overfits the test set. At later times, the logits switch their implicit bias due to logit regularization, collapsing to the vertices of the simplex and obtaining perfect generalization.

G MULTI-CLASS GENERALIZATION OF OTHER KEY RESULTS

In this section we will show that most of the key results detailed in the main text, generalize naturally to the multi-class case.

G.1 LOGIT CLUSTERING

We consider a K-class linear classifier $\boldsymbol{z} = W\boldsymbol{x} + \boldsymbol{b}$. As discussed in the main text, when adding convex and symmetric logit regularization function $f(\boldsymbol{z})$, the loss in the multi-class case can also be decomposed as a sum of per-sample loss functions. The main difference is that now, each class has a different per-sample minimum. The goal of the optimizer is to cluster all of the logits of a certain class to its corresponding minimum, and hence the implicit bias would be, again, logit clustering. From symmetry, this minimum should take the form

$$\boldsymbol{z}^*_{c,k} = \begin{cases} q_{\text{high}} & k = c \\ q_{\text{low}} & else \end{cases} \tag{21}$$

where $k = 1, ..., K$ is the class index, c is the class of the logit (determined by the label), and $q_{\text{high}} > q_{\text{low}}$ are some scalars. We will begin by considering a quadratic loss, and showing that the "direction" and optimal generalization accuracy is invariant to its specific details.

G.2 INVARIANCE TO THE QUADRATIC LOGIT REGULARIZATION FUNCTION

We generalize Corollary 3.2 to the multi-class setting. Assuming that the per-sample loss function is quadratic *around the corresponding class minimum*, we show that generalization accuracy is invariant to the specific target values (and thus the regularization parameters).

Proposition G.1 (Quadratic Loss Invariance)**.** *Consider a multi-class linear classifier minimizing the empirical loss $\mathcal{L} = \frac{1}{N}\sum_i \ell(\boldsymbol{z}_i, \boldsymbol{y}_i)$, where the per-sample loss is quadratic:*

$$\ell(\boldsymbol{z}, \boldsymbol{y}_c) = |\boldsymbol{z} - \boldsymbol{z}^*_c|^2, \tag{22}$$

and $\boldsymbol{z}^*_c \in \mathbb{R}^K$ is a target vector specific to class c. Then, for any $q \neq 0$, the generalization accuracy of the optimal solution is invariant to the value of q.

Proof. Let $W \in \mathbb{R}^{K\times d}$ and $\boldsymbol{b} \in \mathbb{R}^K$. The logits are given by $\boldsymbol{z}_i = W\boldsymbol{x}_i + \boldsymbol{b}$. We notice that up to a global shift that can be compensated by the bias, we can assume that $q_{\text{low}} = 0$. Therefore, $\boldsymbol{z}^*_c = q\boldsymbol{e}_c$. The loss function is:

$$\mathcal{L}(W, \boldsymbol{b}) = \frac{1}{N}\sum_{i=1}^{N} |W\boldsymbol{x}_i + \boldsymbol{b} - q\boldsymbol{y}_i|^2, \tag{23}$$

where $\boldsymbol{y}_i$ is the one-hot representation of the label. We can factor out q^2:

$$\mathcal{L}(W, \boldsymbol{b}) = q^2 \frac{1}{N}\sum_{i=1}^{N} \left| \left(\frac{1}{q}W\right)\boldsymbol{x}_i + \left(\frac{1}{q}\boldsymbol{b}\right) - \boldsymbol{y}_i \right|^2. \tag{24}$$

Let $\tilde{W} = W/q$ and $\tilde{\boldsymbol{b}} = \boldsymbol{b}/q$. Minimizing $\mathcal{L}$ with respect to $W, \boldsymbol{b}$ is equivalent to minimizing the term inside the summation with respect to $\tilde{W}, \tilde{\boldsymbol{b}}$. Let $\tilde{W}_{\min}, \tilde{\boldsymbol{b}}_{\min}$ be the minimizers for the case $q = 1$. Then the minimizers for an arbitrary q are simply $W_{\min} = q\tilde{W}_{\min}$ and $\boldsymbol{b}_{\min} = q\tilde{\boldsymbol{b}}_{\min}$.The predicted class for a test point $\boldsymbol{x}$ is:

$$\hat{y} = \underset{k}{\text{argmax}}(W_{\min}\boldsymbol{x} + \boldsymbol{b}_{\min})_k = \underset{k}{\text{argmax}}(q(\tilde{W}_{\min}\boldsymbol{x} + \tilde{\boldsymbol{b}}_{\min}))_k. \tag{25}$$

Since q is a positive scalar, the argmax (and thus the accuracy) remains unchanged. □

In the binary setting (Proposition 3.1), the scalar coefficient of variation $r = \sigma/\mu$ is equivalent (up to inversion) to the LDA objective: maximizing μ/σ is the same as maximizing the Rayleigh quotient

$$J(\boldsymbol{S}) = \frac{(\boldsymbol{S}^\top \boldsymbol{\mu})^2}{\boldsymbol{S}^\top \Sigma\, \boldsymbol{S}},$$

whose maximizer is $\boldsymbol{S} \propto \Sigma^{-1}\boldsymbol{\mu}$. In the multi-class case, Fisher's criterion has a standard generalization obtained by replacing the scalar mean-variance ratio with a between-class versus within-class scatter ratio. Writing $\boldsymbol{\mu}_c$ for class means, $\boldsymbol{\mu}$ for the global mean, and defining the scatters

$$S_B = \sum_{c=1}^{K} \pi_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top, \tag{26}$$

the multi-class Fisher objective for a $(K-1)$-dimensional projection $U \in \mathbb{R}^{d\times(K-1)}$ is

$$J(U) = \mathrm{tr}\Big((U^\top S_W U)^{-1} U^\top S_B U \Big) \tag{27}$$

(equivalently, $\det(U^\top S_B U)/\det(U^\top S_W U)$). This reduces to the binary μ/σ criterion when $K = 2$ and is maximized by the top $(K-1)$ generalized eigenvectors of the pair (S_B, S_W) (i.e., of $S_W^{-1} S_B$). Accordingly, the natural multi-class analogue of our scalar clustering invariant is the Fisher scatter ratio encoded by these generalized eigenvalues (or their sum/product), rather than a single σ/μ.

G.3 Shift of the Interpolation Threshold ($\sigma_f = 0$)

In this section, we provide the multi-class generalization for the shift of the interpolation threshold in the noiseless feature regime. We also demonstrate in Figure 11 that a grokking dynamics can be achieved, while also the fact that the logits cluster at a single point at the end of the training.

Proposition G.2. *Consider the multi-class model with $\alpha > 0$ and $\sigma_f = 0$. Assume the limit where $d, N \to \infty$ with $\lambda = d/N < 1$. Furthermore, assume that the class means $\{\boldsymbol{\mu}_c\}_{c=1}^K \subset \mathbb{R}^d$ span a $(K-1)$-dimensional affine subspace and are affinely independent. Then, the parameters $W_{\min}, \boldsymbol{b}_{\min}$ that minimize the training loss achieve perfect generalization accuracy.*

Proof. Since $\sigma_f = 0$, any sample $\boldsymbol{x}^{(i)}$ belonging to class c lies in a subspace defined by the class mean plus orthogonal noise:

$$\boldsymbol{x}^{(i)} = \boldsymbol{\mu}_c + \boldsymbol{\xi}_\perp^{(i)}, \tag{28}$$

where $\boldsymbol{\mu}_c$ is the deterministic signal component, and $\boldsymbol{\xi}_\perp^{(i)}$ lies in the orthogonal complement of the signal subspace $\mathcal{S} = \mathrm{span}(\{\boldsymbol{\mu}_k\})$. The regularized loss is a sum of strictly convex per-sample losses. A global lower bound for the total loss is achieved if and only if the logit vector for every sample i of class c coincides exactly with the unique minimizer of the class-specific per-sample loss, $\boldsymbol{z}_c^* \in \mathbb{R}^K$.

We decompose the weight matrix W into a component acting on the signal subspace, W_{sig}, and a component acting on the orthogonal noise, $W_\perp$. We construct a solution where $W_\perp = 0$. The condition for achieving the global minimum on the training set becomes:

$$W_{\mathrm{sig}}\boldsymbol{\mu}_c + \boldsymbol{b} = \boldsymbol{z}_c^*, \quad \forall c \in \{1, \dots, K\}. \tag{29}$$

We can rewrite this using augmented notation. Let $\tilde{W}_{\mathrm{sig}} = [\boldsymbol{b}, W_{\mathrm{sig}}]$ and $\tilde{\boldsymbol{\mu}}_c = [1, \boldsymbol{\mu}_c^\top]^\top$. The system is:

$$\tilde{W}_{\mathrm{sig}} M = Z^*, \tag{30}$$

where columns of M are the augmented means $\tilde{\boldsymbol{\mu}}_c$, and columns of Z^* are the targets $\boldsymbol{z}_c^*$. Since the class means are affinely independent, the matrix $M \in \mathbb{R}^{K\times K}$ is invertible. Thus, a unique solution $\tilde{W}_{\mathrm{sig}} = Z^* M^{-1}$ exists.

For $\lambda < 1$, the loss is strictly convex, ensuring this solution is unique. Thus, the learned weights map every sample (train or test) of class c exactly to $\boldsymbol{z}_c^*$. As $\boldsymbol{z}_c^*$ minimizes the loss for label c, it must correctly classifies the input so the classifier achieves 100% accuracy. □

G.4 Invariance to Orthogonal Noise (σ_n)

Finally, we generalize Theorem 6.1 to the multi-class case.

Proposition G.3. *Let $\alpha > 0$ and $\lambda < 1$. In the multi-class setting described in App. F, the generalization accuracy of the optimal classifier is independent of the orthogonal noise scale σ_n.*

Proof. Let the data for class c be $\boldsymbol{x} = \boldsymbol{x}_{\text{sig}} + \sigma_n \boldsymbol{\eta}_\perp$, where $\boldsymbol{x}_{\text{sig}} \in \mathcal{S}$ and $\boldsymbol{\eta}_\perp \in \mathcal{S}^\perp$. Decompose the weight matrix as $W = W_{\text{sig}} + W_\perp$, where the rows of W_{sig} lie in $\mathcal{S}$ and rows of $W_\perp$ lie in $\mathcal{S}^\perp$. The logits are:

$$\boldsymbol{z} = W_{\text{sig}} \boldsymbol{x}_{\text{sig}} + \sigma_n W_\perp \boldsymbol{\eta}_\perp + \boldsymbol{b}. \tag{31}$$

Let $(W, \boldsymbol{b})$ be the minimizer for noise scale σ_n. The loss depends on the logits $\boldsymbol{z}_i$. Consider a scaling of the noise $\sigma_n \to \beta\sigma_n$. The input noise becomes $\beta\sigma_n \boldsymbol{\eta}_\perp$. To maintain the same optimal logit values (and thus the same minimal loss value), the optimizer must scale the orthogonal weight component inversely: $W'_\perp = \frac{1}{\beta} W_\perp$. The term relevant for prediction on test data is the inner product of the noise and weights:

$$(\beta\sigma_n \boldsymbol{\eta}_\perp)^\top (W'_\perp)^\top = \beta\sigma_n \boldsymbol{\eta}_\perp^\top (\frac{1}{\beta} W_\perp^\top) = \sigma_n \boldsymbol{\eta}_\perp^\top W_\perp^\top. \tag{32}$$

This term is invariant to β. Since the signal term $W_{\text{sig}} \boldsymbol{x}_{\text{sig}}$ is unaffected by σ_n, the distribution of test logits remains identical. Consequently, the accuracy is invariant.

□

H SUPPLEMENTAL NUMERICAL RESULTS

In this section we will present additional numerical results, supporting our results throughout the paper.

H.1 LOGIT COLLAPSE IN THE NOISELESS FEATURE REGIME

In Figure 12, we illustrate the training dynamics in the noiseless feature regime $\sigma_f = 0$. In this setting, the signal component is identical across all signed samples, so a purely signal-aligned classifier can assign the same target logit z^* to every training point. Consequently, logit regularization drives the entire training logit distribution to collapse onto the per-sample minimum of the loss. The top panels show this collapse directly: as training proceeds, the logits concentrate around z^*. The bottom panels show the corresponding optimization dynamics, where the training loss converges to its minimum while the test accuracy reaches perfect generalization. This provides a numerical illustration of the mechanism behind the shifted interpolation threshold discussed in Sec. 5.

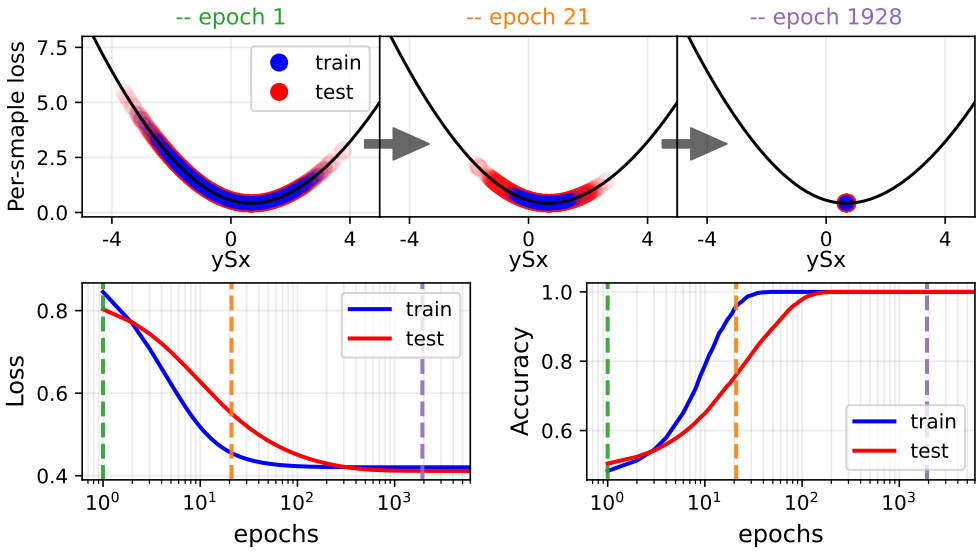


Figure 12: **Perfect generalization with zero feature noise ($\sigma_f = 0$). Top:** Evolution of the logit distribution across three training epochs. At convergence, all logits collapse to a single target value z^*. **Bottom:** Loss and accuracy over time. Vertical dashed lines indicate the epochs corresponding to the top snapshots.

H.2 Discrepancy Between Training and Test Logit Distributions

Our geometric framework provides a mechanism explaining why test logits cluster less tightly than training logits. In the finite-data regime, the optimizer identifies a weight direction that minimizes the empirical variance of the *training* logits. This is often achieved by exploiting specific noise correlations to "over-cluster" the data — effectively reducing the empirical variance below the intrinsic population noise level. However, achieving this requires the learned weight vector to deviate from the true signal direction $\boldsymbol{\mu}_f$. This misalignment introduces a non-zero orthogonal component ($\boldsymbol{S}_\perp$), which necessarily increases the variance of the test logits (which follow the population statistics). Consequently, the test clusters are looser than those observed during training, as demonstrated in Figure 13.

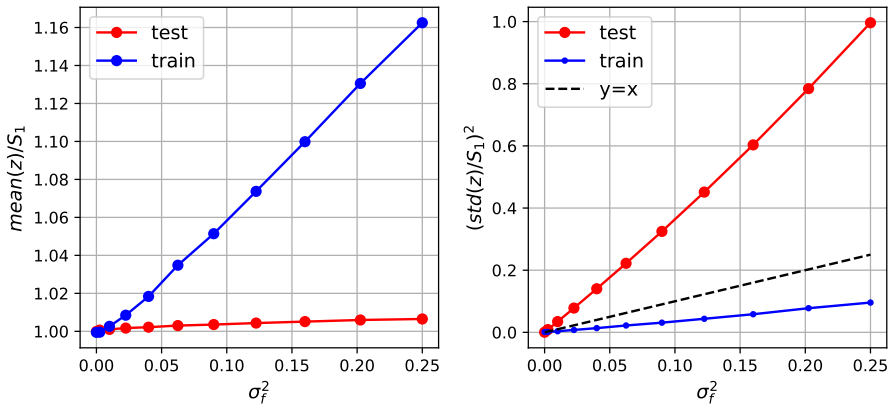


Figure 13: **Logit statistics vs. signal noise σ_f.** The mean and standard deviation of the logits in the binary classification setup are shown as functions of σ_f. The diagonal ($y = x$) serves as a reference for the intrinsic signal noise. **Test Set (Generalization):** The variance is given by $\sigma_z^2 = \sigma_f^2 S_1^2 + \sigma_n^2 \|\boldsymbol{S}_\perp\|^2$. Due to the misalignment term $\|\boldsymbol{S}_\perp\|^2$, the test standard deviation typically exceeds the signal contribution, lying above the reference line. **Training Set:** Conversely, the training standard deviation lies below the reference line. This indicates that the optimizer finds a specific direction that minimizes the training logit variance beyond the theoretical lower bound of the signal noise ($\sigma_z^2 S_1^2$). Here, $\lambda = d/N = 0.7$.

H.3 Phase Diagram of Logit Regularization Benefit

In the main text (left panels of Figure 4), we presented a phase diagram characterizing the regime where logit regularization improves generalization. That analysis focused on the separable setting ($\lambda = 0.7 > 1/2$). Here, we examine the nonseparable regime, $\lambda = 0.4 < 1/2$ (see Figure 14).

The primary distinction between these regimes lies in the behavior of the unregularized baseline as the orthogonal noise σ_n increases. In the overparameterized case ($\lambda > 1/2$), the training data is typically linearly separable due to the high dimensionality. Consequently, as σ_n grows, the unregularized max-margin solution aligns increasingly with the orthogonal noise directions rather than the signal feature. This causes the cosine similarity with the true feature axis to vanish, degrading the generalization accuracy toward random guessing ($1/2$).

In contrast, for $\lambda < 1/2$, the data is generally not linearly separable. Consequently, there is no "easy" overfitting solution that separates the training points using noise dimensions. Instead, increasing the amplitude of the noise will aid alignment: high orthogonal noise effectively drowns out spurious correlations, forcing the optimization to "tune" the margin toward 1 (see bottom panels of Figure 16). Therefore, rather than decaying to $1/2$, we empirically observe that the accuracy of the unregularized solution saturates at a non-trivial constant value (see the left panel of Figure 14). This shift in asymptotic behavior alters the topology of the phase diagram. In general, the specific boundaries of the beneficial regime will depend on the geometric ratio λ and the underlying data distribution.

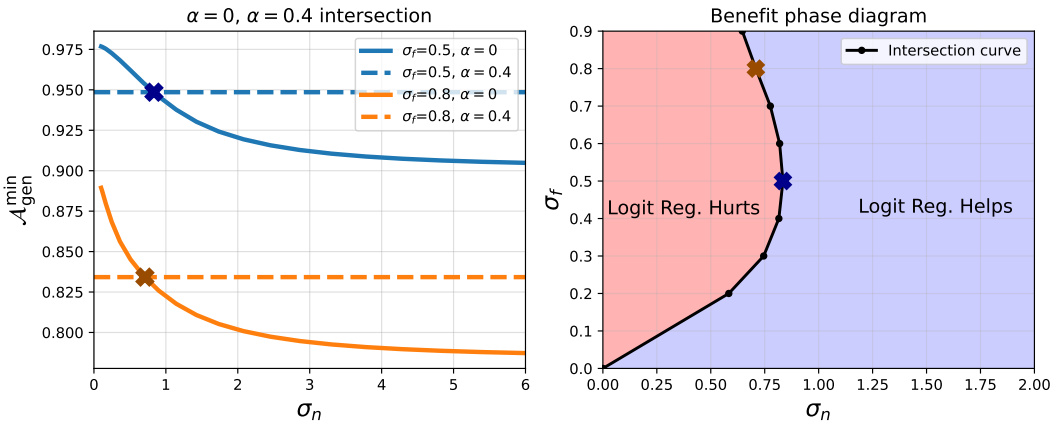


Figure 14: **Benefit phase diagram for $\lambda < 1/2$. Left:** Final generalization accuracy as a function of σ_n, comparing the unregularized ($\alpha = 0$) and regularized ($\alpha > 0$) cases, and for $\lambda = 0.4$. **Right:** Phase diagram showing regions where logit regularization is beneficial. The boundary curve is constructed from the intersection points (as shown in the left panel) for different values of σ_f. For illustration, the intersection points of the unregularized and regularized curves for $\sigma_f = 0.5, 0.8$ (left) are marked as "X"s of corresponding colors on the boundary curve (right).

H.4 Comprehensive Parameter Sweeps

In this subsection, we present a systematic numerical analysis of the generalization accuracy and the cosine similarity (relative to the true feature axis). We perform parameter sweeps over the aspect ratio λ, signal noise σ_f, and orthogonal noise σ_n, comparing the unregularized baseline ($\alpha = 0$) against the logit regularized model ($\alpha > 0$).

H.5 Comparison with Weight Decay

To underscore the unique geometric properties of logit regularization, we contrast its effects with standard L_2 parameter regularization (Weight Decay). We consider the objective function:

$$\mathcal{L}_{\text{WD}} = \sum_i \log\left(1 + e^{-y_i \boldsymbol{S}^\top \boldsymbol{x}_i}\right) + \frac{\gamma}{2}\|\boldsymbol{S}\|^2. \tag{33}$$

While Weight Decay prevents the weights from diverging, it fundamentally differs from logit regularization in how it shapes the optimization landscape. As shown in Figure 17, Weight Decay does **not** induce the specific invariance properties observed under logit regularization.

Specifically, we highlight two key distinctions:

1. **No shift in interpolation threshold:** Unlike logit regularization, adding Weight Decay does not shift the interpolation threshold in the noiseless feature regime ($\sigma_f = 0$). The generalization accuracy does not reach 100% throughout the region $\lambda < 1$, confirming that the perfect generalization mechanism (logit clustering around a finite target) is specific to the per-sample convexity of logit penalties.

2. **Sensitivity to orthogonal noise:** The final accuracy under Weight Decay remains explicitly dependent on the orthogonal noise scale σ_n. This stands in sharp contrast to the logit regularized case, where the accuracy becomes invariant to σ_n (as shown in Figure 15).

It is worth noting that for linear models, quadratic logit regularization (αz^2) can be viewed as a generalized Tikhonov regularization of the form $\alpha \boldsymbol{S}^\top \hat{\Sigma} \boldsymbol{S}$, where $\hat{\Sigma} = \frac{1}{N}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$ is the empirical non-centered covariance matrix. This data-dependent structure naturally adapts the penalty to the geometry of the input features, a property that isotropic Weight Decay ($\gamma \boldsymbol{S}^\top \boldsymbol{I} \boldsymbol{S}$) lacks.

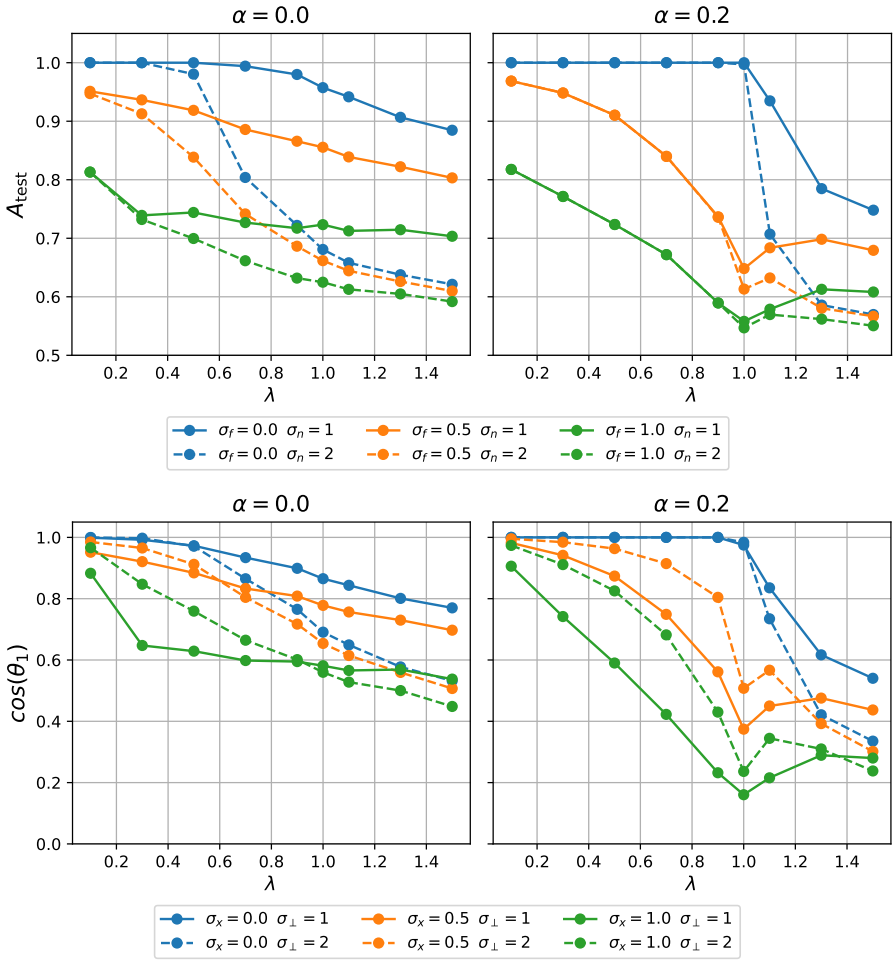


Figure 15: **Performance as a function of model capacity** ($\lambda = d/N$)**. Top Panels:** Generalization accuracy for the unregularized (left) and regularized (right) models. Different colors correspond to signal noise levels $\sigma_f = \{0, 0.5, 1\}$, and solid/dashed lines indicate orthogonal noise levels $\sigma_n = \{1, 2\}$. Two key phenomena are visible in the regularized case (right): (1) The accuracy becomes independent of σ_n for $\lambda < 1$, and (2) For the noiseless feature case ($\sigma_f = 0$), the interpolation threshold shifts from $\lambda = 0.5$ to $\lambda = 1$. **Bottom Panels:** Cosine similarity with the true feature axis. Unlike the accuracy, the cosine similarity remains sensitive to σ_n in both regimes, confirming that regularization stabilizes the prediction quality even while the weight direction varies.

H.6 Dependence of Cosine Similarity on Noise Amplitudes

In this subsection, we numerically investigate the alignment of the learned weights with the true signal direction. We measure this via the cosine similarity $\rho = \cos(\theta_1)$, analyzing its dependence on the orthogonal noise σ_n and the signal noise σ_f.

First, we examine the dependence on σ_n. Recall from the main text (Eq. 4) that the alignment is predicted to follow the scaling form:

$$\cos(\theta_1) = \frac{1}{\sqrt{1 + (C/\sigma_n)^2}}, \tag{34}$$

where C is a coefficient that depends on σ_f but is strictly independent of σ_n. This implies a linear relationship between $1/\cos^2(\theta_1)$ and $1/\sigma_n^2$. In Figure 18, we verify this prediction numerically: the data perfectly fits the line $y = 1 + C^2 x$, confirming the validity of our analytical derivation.

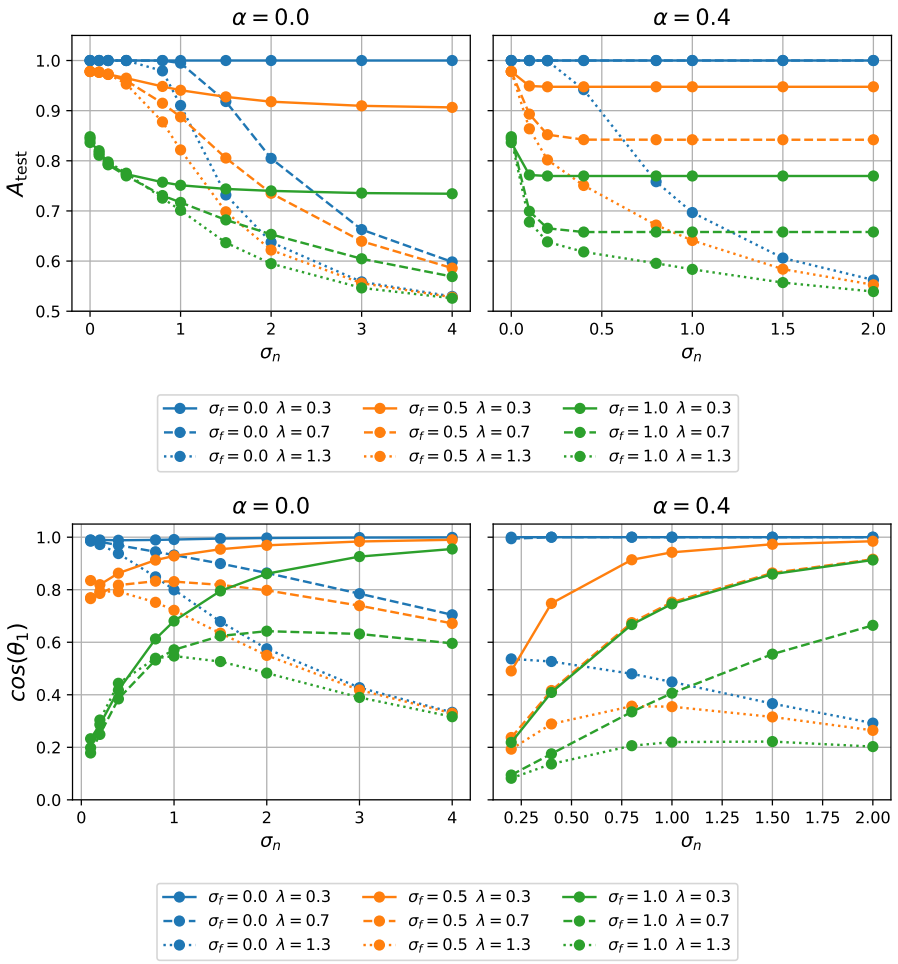


Figure 16: **Sensitivity to orthogonal noise amplitude σ_n. Top Panels:** Generalization accuracy vs. σ_n. The regularized model (right) is robust to changes in σ_n, maintaining constant accuracy. Note: The slight deviation observed at extremely small σ_n is a numerical artifact; the convergence time scales inversely with noise, preventing the simulation from fully reaching the theoretical optimum within the fixed computational budget in the limit $\sigma_n \to 0$. **Bottom Panels:** Cosine similarity vs. σ_n. Here we observe distinct regimes: for the underparameterized case ($\lambda < 1/2$), increasing noise improves alignment (similarity $\to 1$), whereas for the overparameterized case ($\lambda > 1/2$), increasing noise degrades alignment.

Next, we characterize the coefficient C by analyzing its dependence on the signal noise σ_f. We fix $\sigma_n = 1$ and track the components of the weight vector $\boldsymbol{S} = S_1 \boldsymbol{e}_1 + \boldsymbol{S}_\perp$. As shown in the left panel of Figure 19, for small σ_f, the signal component S_1 remains approximately constant, while the orthogonal component $\|\boldsymbol{S}_\perp\|$ grows linearly with σ_f. Consequently, the ratio $C \propto \|\boldsymbol{S}_\perp\|/S_1$ exhibits a linear dependence on σ_f in the low-noise regime (see Figure 19, right panel). This demonstrates that as the feature noise vanishes ($\sigma_f \to 0$), the model relies less on spurious orthogonal correlations, recovering the perfect alignment predicted in the noiseless limit.

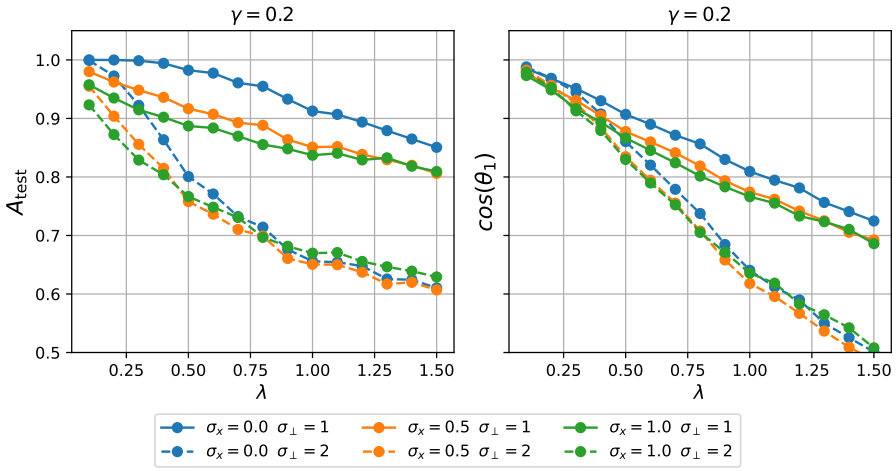


Figure 17: **Generalization under Weight Decay.** The generalization accuracy (top) and cosine similarity (bottom) achieved at convergence with standard L_2 regularization ($\gamma = 0.2$). Unlike logit regularization, the accuracy remains sensitive to σ_n and does not exhibit perfect generalization in the $\sigma_f = 0$ regime for $0.5 < \lambda < 1$.

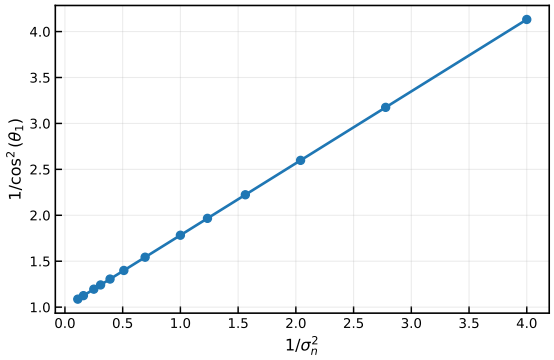


Figure 18: **Verification of σ_n scaling.** Plot of $1/\cos(\theta_1)^2$ versus $1/\sigma_n^2$. The linear fit confirms the functional form derived in Eq. (34), where the slope corresponds to C^2.

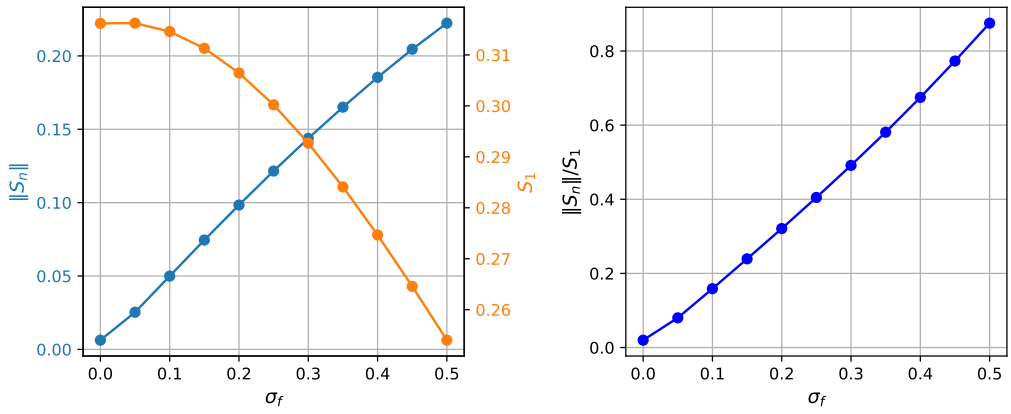


Figure 19: **Dependence on σ_f.** All results are for fixed $\sigma_n = 1$. **Left:** The magnitudes of the orthogonal weight component $\|\boldsymbol{S}_\perp\|$ and the signal component S_1 as functions of σ_f. **Right:** The ratio $\|\boldsymbol{S}_\perp\|/S_1$ (which is proportional to C) scales linearly with σ_f for small noise amplitudes.

H.7 Additional details regarding the numerical settings

Unless stated otherwise, we consider a linear classification model trained by minimizing the regularized loss $\mathcal{L} = \sum_i [(1-\alpha)\ell_{\mathrm{CE}}(z_i, y_i) + \alpha z_i^2]$. We use Full-Batch Gradient Descent (GD) for the optimization. The data is generated according to the signal-plus-noise decomposition $\boldsymbol{x} = y\mu_f \boldsymbol{e}_1 + \boldsymbol{\xi}$, where $\boldsymbol{\xi} = \sigma_f \xi_f \boldsymbol{e}_1 + \sigma_n \boldsymbol{\xi}_\perp$. In most experiments, the noise components ξ_f and $\boldsymbol{\xi}_\perp$ are drawn from standard Normal distributions.

Below we provide the exact hyperparameters used for each figure:

- **Figure 1 (Clustering):**
  - **Top Row ($\alpha = 0$):** $d = 1400$, $N = 2000$ ($\lambda = 0.7$), $\mu_f = 1$, $\sigma_f = 0.5$, $\sigma_n = 1$. The model is trained for $20,000$ epochs with learning rate $\eta = 0.1$.
  - **Bottom Row ($\alpha > 0$):** Same parameters as above, but with $\alpha = 0.2$.
  - **Figure 1 (Zero Feature Noise):** Same parameters as above, but with $\sigma_f = 0$ and $\alpha = 0.2$.
- **Figure 2 (Invariance with Student-t Noise):** The noise follows a Student-t distribution with degrees of freedom $\nu \in \{2.1, 3, 20\}$. The dimensions are $d = 1000$, $N = 1500$ ($\lambda \approx 0.67$) with $\mu_f = 1$, $\sigma_f = 1$, $\sigma_n = 1$. For $\alpha = 0$, we use GD with $\eta = 0.1$. For $\alpha > 0$, we use the Adam optimizer with $\eta = 0.001$. The number of epochs varies between $2,000$ and $10,000$ depending on convergence.
- **Figure 10 (Appendix - Invariance with Gaussian Noise):** Gaussian data with varying dimension $d \in \{200, 1400, 1800\}$ while keeping $N = 2000$ fixed, yielding $\lambda \in \{0.1, 0.7, 0.9\}$. Other parameters: $\mu_f = 1$, $\sigma_f = 0.5$, $\sigma_n = 1$. Trained for $20,000$ epochs with GD ($\eta = 0.1$).
- **Figure 15 (Performance vs λ):** We vary d while keeping $N = 2000$ fixed to sweep λ.
  - Parameters: $\mu_f = 1$. Signal noise $\sigma_f \in \{0, 0.5, 1\}$. Orthogonal noise $\sigma_n \in \{0.2, 4\}$.
  - Regularization: We compare $\alpha = 0$ vs $\alpha = 0.4$.
  - Optimization: GD with $\eta = 0.1$ for $20,000$ epochs.
- **Figure 17 (WD Performance):** We vary d while keeping $N = 2000$ fixed to sweep λ,
  - Parameters: $\mu_f = 1$. $\sigma_f \in \{0, 0.5, 1\}$. $\sigma_n \in \{1, 2\}$.
  - Regularization: $\alpha = 0$ but $\gamma = 0.2$ (WD).
  - Optimization: GD with $\eta = 0.1$ for $2,000$ epochs.
- **Figure 18 (Cosine Similarity vs σ_n):** We fix $d = 1400$, $N = 2000$ ($\lambda = 0.7$).
  - Parameters: $\mu_f = 1$, $\sigma_f = 0.5$. We sweep $\sigma_n \in [0.5, 3]$.
  - Regularization: $\alpha = 0.2$.
  - Optimization: GD with $\eta = 0.1$ for $2,000$ epochs.
- **Figure 19 (Cosine Similarity vs σ_f):** We fix $d = 1400$, $N = 2000$ ($\lambda = 0.7$).
  - Parameters: $\mu_f = 1$, $\sigma_n = 1$. We sweep $\sigma_f \in [0, 0.5]$.
  - Regularization: $\alpha = 0.4$.
  - Optimization: Adam optimizer with $\eta = 0.01$ for $1,000$ epochs.
- **Figure 5 (Grokking):** We fix $d = 1400$, $N = 2000$ ($\lambda = 0.7$) in the noiseless feature regime.
  - Parameters: $\mu_f = 0.1$, $\sigma_f = 0$, $\sigma_n = 1$.
  - Regularization: We compare $\alpha \in \{0, 0.001, 0.01, 0.1\}$.
  - Optimization: GD with $\eta = 0.1$ for an extended duration of $200,000$ epochs.
- **Figure 4, left panels (Phase Diagram $\lambda = 0.7$):** $d = 1400$, $N = 2000$. We interpolate theoretical curves based on simulations with σ_n sweeping from 0.1 to 6.
  - Parameters: $\mu_f = 1$. $\alpha = 0.4$.
  - Optimization: GD with $\eta = 0.1$ for $20,000$ epochs.
- **Figure 14 (Phase Diagram $\lambda = 0.4$):** $d = 800$, $N = 2000$.

  - Parameters: $\mu_f = 1$. $\alpha = 0.4$.
  - Optimization: GD with $\eta = 0.1$ for $20,000$ epochs.
- **Figure 4, right panels (Multi-class):** $K = 10$ classes. $d = 1400$, $N = 2000$ ($\lambda = 0.7$).
  - Parameters: $\mu_f = 1$, $\sigma_f = 0.1$, $\sigma_n = 1$.
  - Regularization: $\alpha = 0$ and $\alpha = 0.2$.
  - Optimization: GD optimizer with $\eta = 0.1$ for $20,000$ epochs.
- **Figure 11 (Multi-class Grokking):** $K = 10$ classes. $d = 1400$, $N = 2000$ ($\lambda = 0.7$). Same as previous but for $\sigma_f = 0$ and $\alpha = 0.04$.

H.7.1 RESNET-18 (FIGURES 3 AND 6)

We extract 512-dimensional penultimate features from a frozen ResNet-18 pre-trained on ImageNet. For the linear probe experiments, we select $N_{\text{train}} = 1000$ balanced training samples and use the remaining $N_{\text{test}} = 9000$ samples for evaluation. We compare performance on the clean CIFAR-10 test set and on CIFAR-10-C (Impulse Noise, Severity 5).

Feature Decomposition and Noise Estimation: To map the complex feature distribution to our theoretical signal-plus-noise model, we perform a class-conditional decomposition. Let $X \in \mathbb{R}^{N \times d}$ be the matrix of feature vectors and y the labels. For each class c, we compute the empirical mean $\boldsymbol{\mu}_c \in \mathbb{R}^d$ and the centered covariance matrix Σ_c.

1. **Signal Subspace:** We define the signal subspace $\mathcal{S}$ as the span of the differences between class means, $\mathcal{S} = \text{span}(\{\boldsymbol{\mu}_c - \boldsymbol{\mu}_1\}_{c=2}^K)$. We compute an orthonormal basis $Q_1 \in \mathbb{R}^{d \times (K-1)}$ for $\mathcal{S}$ via QR decomposition.
2. **Orthogonal Subspace:** We complete the basis to $\mathbb{R}^d$ using an orthogonal complement $Q_2 \in \mathbb{R}^{d \times (d-K+1)}$, such that $U = [Q_1, Q_2]$ is unitary.
3. **Noise Projection and RMS Calculation:** We project the class-conditional covariances onto these subspaces. The effective noise amplitudes reported in the figures are calculated as the root-mean-square (RMS) of the standard deviations in each subspace, averaged over all classes:

$$\sigma_f^{\text{eff}} = \frac{1}{K}\sum_{c=1}^{K}\sqrt{\text{Tr}(Q_1^\top \Sigma_c Q_1)/\dim(\mathcal{S})}, \quad \sigma_n^{\text{eff}} = \frac{1}{K}\sum_{c=1}^{K}\sqrt{\text{Tr}(Q_2^\top \Sigma_c Q_2)/\dim(\mathcal{S}^\perp)}. \tag{35}$$

   To allow for comparison across datasets with different signal scales, we normalize these values by the mean pairwise distance between class centroids.

Synthetic Orthogonal Scaling: To verify the invariance to orthogonal noise (Fig. 11, Right), we synthetically manipulate the data geometry. We project every feature vector $\boldsymbol{x}_i$ onto the orthogonal subspace via the projection matrix $P_\perp = Q_2 Q_2^\top$ and scale this component by a factor γ:

$$\boldsymbol{x}_i^{\text{scaled}} = (I - P_\perp)\boldsymbol{x}_i + \gamma P_\perp \boldsymbol{x}_i. \tag{36}$$

This operation scales the effective σ_n by γ while preserving the signal geometry exactly. The linear classifier is then trained on these scaled features using the Adam optimizer ($\eta = 10^{-4}$) for $30,000$ epochs.