WHEN STRONGER TRIGGERS BACKFIRE: A HIGH-DIMENSIONAL THEORY OF BACKDOOR ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor poisoning attacks behave counter-intuitively in high dimensions: when the test-time trigger budget is held fixed, a stronger *training* trigger can help the defender. We study generalised linear models on Gaussian-mixture data in the proportional regime ($p/n \to \kappa$), varying the training trigger strength α. Three phenomena emerge: (i) clean test accuracy can increase with α; (ii) the trigger alignment that drives attack success peaks at a finite α and then declines; and (iii) low-variance directions can yield stronger trigger alignment. We derive closed-form asymptotic alignments for squared loss, including a unique finite peak for arbitrary mean-trigger overlap, and extend the finite-peak result and large-strength decay to a broader class of convex losses under suitable assumptions. A margin-variance decomposition identifies a finite-sample term that helps explain (i) and is absent from the population limit. Experiments on Gaussian surrogates and logistic regression on CIFAR-10 support the theoretical predictions, while ResNet-18 experiments exhibit similar qualitative behaviour beyond the convex setting.

1 INTRODUCTION

Backdoor poisoning attacks aim to make a model respond to an attacker-chosen trigger while preserving its performance on clean inputs (Gu et al., 2019). A basic question is how strongly that trigger should be represented during training. When the test-time trigger budget is fixed, increasing the training trigger strength creates a trade-off: poisoned samples become easier to classify, but the model may need less sensitivity to the trigger to fit them. Can a stronger training trigger therefore weaken the attack?

We study this question for regularised generalised linear models on high-dimensional Gaussian-mixture data. We distinguish the training trigger strength α from a fixed test-time strength α_{test}, reflecting settings in which the deployed trigger is constrained. Our analysis operates in the proportional regime, where the input dimension p and training sample size n grow together with $p/n \to \kappa$. This setting captures finite-sample effects that disappear in the classical information limit $n \gg p$.

Motivating evidence from deep networks Figure 1 shows the clean test accuracy and attack success rate of a ResNet-18 on CIFAR-10. Three striking patterns emerge: (i) clean accuracy *increases* with α, (ii) the attack success rate peaks at a finite α and then declines, and (iii) the attack is most effective when the trigger aligns with low-variance directions of the data. Throughout the strength sweep, the test-time trigger budget α_{test} is held fixed, reflecting a constraint on the deployed trigger. Beyond the observed attack-success peak, a stronger training trigger therefore helps the defender: clean accuracy improves while attack success falls. Appendix Figure 5 varies α_{test} and confirms the peak persists and tracks it. These observations motivate our theoretical analysis of how training trigger strength affects clean accuracy and attack success.

1.1 MAIN CONTRIBUTIONS

We develop a high-dimensional theory of backdoor poisoning for regularised GLMs on Gaussian-mixture data with general covariance. Our analysis characterises how training trigger strength and direction affect the learned classifier, with three main contributions.

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.

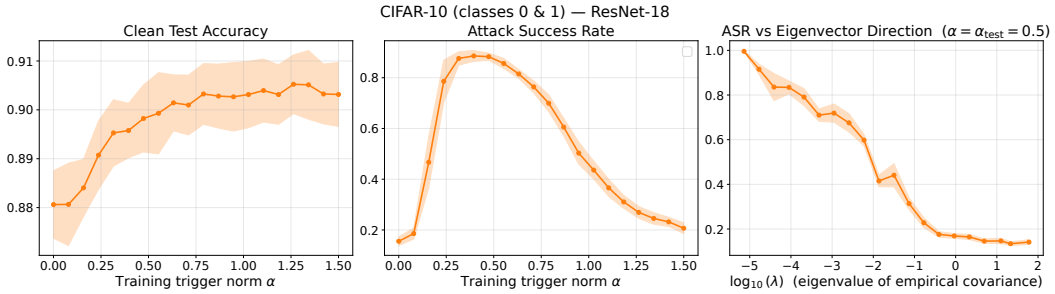


Figure 1: Observed Phenomena in Backdoor Poisoning

ResNet-18 on CIFAR-10 (classes 0 & 1); $\phi = 0.05$; $\alpha_{\text{test}} = 0.5$, averaged over 25 runs. Left/middle: clean accuracy and attack success rate vs. *training* trigger strength α, with a fixed 2×2-pixel corner trigger. Right: at $\alpha = 0.5$, the trigger direction varies over eigenvectors of the empirical covariance (x-axis: log eigenvalue).

Closed-form characterisation for squared loss. We derive explicit asymptotic expressions for the benign and trigger alignments, establishing a unique finite maximum of trigger alignment for arbitrary mean-trigger overlap. Under trigger orthogonality, we further show that the leading-order benign alignment increases with training trigger strength, and that the optimal trigger directions satisfying this condition are those of minimum eigenvalue (Section 3.1).

Extensions to general convex losses. For decreasing convex losses satisfying our regularity and orthogonality assumptions, we establish a finite maximum and large-strength decay of trigger alignment in the proportional regime, with explicit decay bounds (Section 3.2). In the population setting, we prove convergence to the benign-only minimiser as training trigger strength grows (Section 3.3).

Finite-sample effects and empirical evidence. We decompose the test-margin variance to identify a finite-sample contribution absent from the information limit, clarifying how changes in alignment can affect clean accuracy (Section 4). Experiments with Gaussian surrogates and logistic regression on CIFAR-10 support the theoretical predictions, while ResNet-18 experiments probe their relevance to learned nonlinear representations.

Our work contributes a rigorous, quantitative analysis of backdoor poisoning, where prior work has largely relied on empirical evidence or low-dimensional bounds. From a high-dimensional statistics perspective, we introduce a tractable adversarial data model and show that it exhibits qualitatively new behaviour not captured by classical $n \gg p$ asymptotics.

1.2 Related Work

Backdoor attacks. Trigger-based poisoning was introduced by Gu et al. (2019) and has since expanded to include physical, label-consistent, and warping-based attacks (Li et al., 2021; Turner et al., 2019; Nguyen & Tran, 2021), as well as data-efficient variants (Xia et al., 2022; Souly et al., 2025); see Li et al. (2024) for a survey. Theoretical studies examine attack effectiveness (Lu et al., 2023), detectability–efficacy trade-offs (Granziol, 2026), and low-rank-activated triggers (Evans & Tanner, 2026). Related optimisation-based poisoning of linear regression is studied by Jagielski et al. (2018). Closest to our analysis, Flynn & Granziol (2026) study high-dimensional ridge regression with isotropic covariates. Our analysis accommodates general covariance and a broader class of convex losses, characterising how training trigger strength and direction affect the learned classifier. Wang et al. (2024) identify low-density directions as particularly vulnerable to backdoor attacks; our analysis complements this observation with a quantitative characterisation of trigger alignment in the proportional regime.

High-dimensional proportional regime. The proportional regime ($p/n \to \kappa$) has been used to explain modern learning phenomena such as double descent and benign overfitting (Hastie et al., 2019; Bartlett et al., 2020), and sharp asymptotics in this regime (via approximate message passing, the convex Gaussian min–max theorem, and leave-one-out techniques) yield precise results for ridge and logistic regression (Thrampoulidis et al., 2015; Donoho & Montanari, 2013; Dobriban & Wager, 2018; Sur & Candès, 2019; Montanari et al., 2025) for Gaussian mixture classification (Mai & Liao,

2020; Deng et al., 2020), and lower bounds on generalisation error (Seroussi & Zeitouni, 2022). Our analysis builds directly on the Gaussian-mixture characterisation of Loureiro et al. (2021), restated as Theorem 1. Closest to our setting is Barnfield et al. (2025), which studies sparse signal detection in high dimensions; in contrast, we analyse adversarially injected triggers in regularised GLMs.

2 SETTING AND NOTATION

Notation. We use boldface for matrices and vectors, and capital letters for matrices. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|$ denotes its operator norm. The complexity notations $O(\cdot)$, and $o(\cdot)$ are understood for large data size n and input dimension p, while the notation $O_\alpha(\cdot), o_\alpha(\cdot)$ is intended for sufficiently small α^{-1}. We use $\hat{\boldsymbol{\theta}}$ for the learned parameters (ERM or population minimiser, depending on context), and $\tilde{\boldsymbol{\theta}}$ for the Gaussian proxy from Theorem 1.

Clean data. We consider a classification problem where each sample $(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{\pm 1\}$ is drawn independently from a binary Gaussian mixture with $\mathbb{P}(y_i = +1) = \frac{1}{2}$, and the feature vector $\mathbf{x}_i \mid y_i \sim \mathcal{N}(y_i \boldsymbol{\mu}, \mathbf{C})$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the class mean and $\mathbf{C} \in \mathbb{R}^{p \times p}$ is a shared covariance matrix.

Poisoning model. An adversary selects a fraction $\phi \in (0, \frac{1}{2})$ of the full training set, drawn exclusively from the samples with $y_i = -1$, and applies a *backdoor attack*: for each selected sample, the trigger $\alpha \mathbf{v}$ is added to the feature vector and the label is flipped to $+1$. Here $\mathbf{v} \in \mathbb{R}^p$ is a fixed trigger direction with $\|\mathbf{v}\| = 1$, $\alpha \geq 0$ controls the trigger strength, and samples with $y_i = +1$ are never modified. We assume $\alpha \geq 0$, as otherwise we may just consider $-\mathbf{v}$.

The losses we consider depend on features and labels only through $y_i \mathbf{x}_i$[1], so we absorb labels into features as $\mathbf{z}_i := y_i \mathbf{x}_i$. After poisoning, the absorbed data follows a two-component Gaussian mixture with shared covariance $\mathbf{C}$:

$$\mathbf{z}_i \mid K = c \;\sim\; \mathcal{N}(\boldsymbol{\mu}_c, \mathbf{C}), \qquad \boldsymbol{\mu}_1 = \boldsymbol{\mu}, \quad \boldsymbol{\mu}_2 = \alpha \mathbf{v} - \boldsymbol{\mu}, \quad \pi_1 = 1 - \phi, \quad \pi_2 = \phi. \tag{1}$$

Empirical risk minimisation (proportional regime) Given n training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$; a convex loss L, and regularisation $\lambda > 0$, we consider the estimator

$$\hat{\boldsymbol{\theta}} \;=\; \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} \; \big\{ \mathcal{L}_n(\boldsymbol{\theta}) \;:=\; \frac{1}{n} \sum_{i=1}^n L\big(y_i \, \mathbf{x}_i^\top \boldsymbol{\theta}\big) \;+\; \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \big\}. \tag{2}$$

Canonical choices are the logistic loss $L(t) = \log(1 + e^{-t})$ and the squared loss $L(t) = \frac{1}{2}(1 - t)^2$.

Assumption 1 (Proportional asymptotics)**.** *The dimension p and sample size n grow jointly such that $p/n \to \kappa \in (0, \infty)$. The covariance is bounded, $\|\mathbf{C}\| = O(1)$, and the class means satisfy $\|\boldsymbol{\mu}_c\| = O(1)$ for each $c \in \{1, 2\}$. The loss $L : \mathbb{R} \to \mathbb{R}$ is strictly convex.*

Remark 1 (Order of limits)**.** *Although $\boldsymbol{\mu}_2 = \alpha \mathbf{v} - \boldsymbol{\mu}$ in (1) scales with α while Assumption 1 requires $\|\boldsymbol{\mu}_c\| = O(1)$, this is understood through a two-stage limit: for each fixed α we take $n, p \to \infty$ (for Theorem 1), and all large-α statements then concern the alignments $h_\mu(\alpha), h_v(\alpha)$ as $\alpha \to \infty$.*

Population risk (information limit) In the information limit $\kappa \to 0$ (or $n \gg p$), this is equivalent to minimising the *population risk*:

$$\mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta}; \alpha) \;:=\; (1 - \phi) \, \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})}\big[L(\boldsymbol{\theta}^\top \mathbf{x})\big] \;+\; \phi \, \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\alpha \mathbf{v} - \boldsymbol{\mu}, \mathbf{C})}\big[L(\boldsymbol{\theta}^\top \mathbf{x})\big] \;+\; \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2, \tag{3}$$

with unique minimiser $\boldsymbol{\theta}(\alpha)$ (by λ-strong convexity).

2.1 GAUSSIAN PROXY FOR ERM

Our analysis uses a Gaussian proxy for the high-dimensional ERM estimator derived in Loureiro et al. (2021). The proxy expresses the learned parameters as a deterministic mean with Gaussian fluctuations, both determined by a scalar fixed-point system. This allows us to characterise the benign and trigger alignments together with the test-margin variance.

[1]Equivalently, $\tilde{L}(x, +1) = \tilde{L}(-x, -1)$; we write $L(x) := \tilde{L}(x, +1)$.

Theorem 1 (Gaussian proxy (Loureiro et al., 2021)). *Denote* $\mathbf{R}(\lambda,\tau) := (\lambda\mathbf{I} + \tau\mathbf{C})^{-1}$. *Under Assumption 1 and the mixture model* (1)*, there exists* $\tilde{\boldsymbol{\theta}}$ *such that for pseudo-Lipschitz* $\psi : \mathbb{R}^p \to \mathbb{R}$ *of finite order,* $\psi(\hat{\boldsymbol{\theta}}) \to_{p,n\to\infty} \mathbb{E}\psi(\tilde{\boldsymbol{\theta}})$ *in probability, where*

$$\mathbf{R}(\lambda,\tau)^{-1}\tilde{\boldsymbol{\theta}} \sim \mathcal{N}\big(\eta_1\boldsymbol{\mu}_1 + \eta_2\boldsymbol{\mu}_2,\ \tfrac{\gamma}{n}\mathbf{C}\big).$$

The scalars $(\tau,\gamma,\delta,\eta_1,\eta_2) \in \mathbb{R}^5$ *solve a self-consistent system. Let* $f(x) = -L'(\mathrm{prox}_{\delta L}(x))$ *and* $r_c \sim \mathcal{N}(M_c,\sigma^2)$*, with* $r_K \sim r_c$ *conditional on* $K = c$ *with probability* π_c*. Then* $\tau = \mathbb{E}[-f'(r_K)]$*,* $\gamma = \mathbb{E}[f^2(r_K)]$*, and* $\eta_c = \pi_c\,\mathbb{E}[f(r_c)]$ *for* $c = 1, 2$*, and*

$$\delta = \tfrac{1}{n}\,\mathrm{tr}[\mathbf{C}\mathbf{R}(\lambda,\tau)], \quad M_c = \boldsymbol{\mu}_c^\top\mathbf{R}(\lambda,\tau)(\eta_1\boldsymbol{\mu}_1 + \eta_2\boldsymbol{\mu}_2),\ c = 1, 2,$$

$$\sigma^2 = (\eta_1\boldsymbol{\mu}_1 + \eta_2\boldsymbol{\mu}_2)^\top\mathbf{R}(\lambda,\tau)^2\mathbf{C}(\eta_1\boldsymbol{\mu}_1 + \eta_2\boldsymbol{\mu}_2) + \tfrac{\gamma}{n}\,\mathrm{tr}\big[\mathbf{R}(\lambda,\tau)^2\mathbf{C}^2\big].$$

The scalars M_c *and* σ^2 *are the asymptotic mean and variance of* $\boldsymbol{\mu}_c^\top\tilde{\boldsymbol{\theta}}$ *and* $\mathbf{x}^\top\tilde{\boldsymbol{\theta}}$ *for a test point* $\mathbf{x}$*.*

The key quantities in our analysis are the expected benign and trigger alignments, defined as

$$h_\mu(\alpha) := \mathbb{E}[\boldsymbol{\mu}^\top\tilde{\boldsymbol{\theta}}], \quad h_v(\alpha) := \mathbb{E}[\mathbf{v}^\top\tilde{\boldsymbol{\theta}}].$$

For $\alpha > 0$, we have $h_v(\alpha) = (M_1 + M_2)/\alpha$.

Some of our results use the following genericity condition on the trigger direction relative to the class mean. In high dimensions, an independently sampled uniform unit vector $\mathbf{v}$ has little overlap with any fixed bounded-norm direction; in particular, $\mathbf{v}^\top\mathbf{R}(\lambda,\tau)\boldsymbol{\mu} = O_{\mathbb{P}}(p^{-1/2})$ for fixed $\lambda > 0$ and $\tau \geq 0$. This motivates studying triggers with negligible overlap with the mean directions.

A complementary practical motivation comes from standard backdoor constructions, such as localised patches, which are typically chosen without explicit alignment with class-specific structure, although this does not guarantee orthogonality. Related theoretical works assume triggers orthogonal to the clean-data subspace (Manoj & Blum, 2021) or to clean-feature directions (Li & Liu, 2024).

Assumption 2 (Trigger orthogonality). *The trigger direction* $\mathbf{v}$ *is asymptotically orthogonal to the Krylov subspace of* $\mathbf{C}$ *generated by* $\boldsymbol{\mu}$*:* $\mathbf{v}^\top\mathbf{C}^k\boldsymbol{\mu} = o(1)$ *for each* $k = 0, 1, 2, \ldots$ *Equivalently,* $\mathbf{v}^\top R(\lambda,\tau)\boldsymbol{\mu} = o(1)$ *for every fixed* $\tau \geq 0$ *and* $\lambda > 0$*.*

Section 3.2 requires mild regularity of the loss, satisfied by the logistic and exponential losses.

Assumption 3 (Loss regularity). *The loss* L *is non-negative, convex, and strictly decreasing. It has an exponentially bounded derivative for* $x < 0$ *and sufficient decay for* $x > 0$*:*

$$|L'(x)| \leq e^{C_1(|x|+1)} \text{ for } x < 0, \qquad |L'(x)| \leq \frac{C_2}{x^{1+\epsilon} + 1} \text{ for } x > 0 \text{ and some } \epsilon > 0.$$

3 Main results

We organise our findings around the three phenomena connecting training trigger strength and direction to the learned classifier, each stated informally here with numerical support on CIFAR-10 and Gaussian surrogates; formal results appear in Sections 3.1–3.3. Throughout, the attack success rate is evaluated on held-out samples with a fixed test trigger $\alpha_{\mathrm{test}}\mathbf{v}$.

Clean accuracy can *increase* with trigger strength For fixed poisoning fraction ϕ, increasing the trigger strength α can counter-intuitively *improve* clean test accuracy. The mechanism: as α grows, poisoned samples are pushed further into the $+1$ region and become easier to classify, so the model "wastes" less capacity on them and concentrates on the clean distribution. *Formally:* we prove this for the ERM squared loss (Corollary 3). For general convex losses, our population analysis establishes an endpoint ordering of benign alignment when the class mean is a covariance eigenvector (Proposition 7) and a local one-step result (Proposition 8). Figure 2 shows accuracy improving after an initial dip. Section 4 explains how this can improve clean accuracy, under approximations that may not hold at small α.

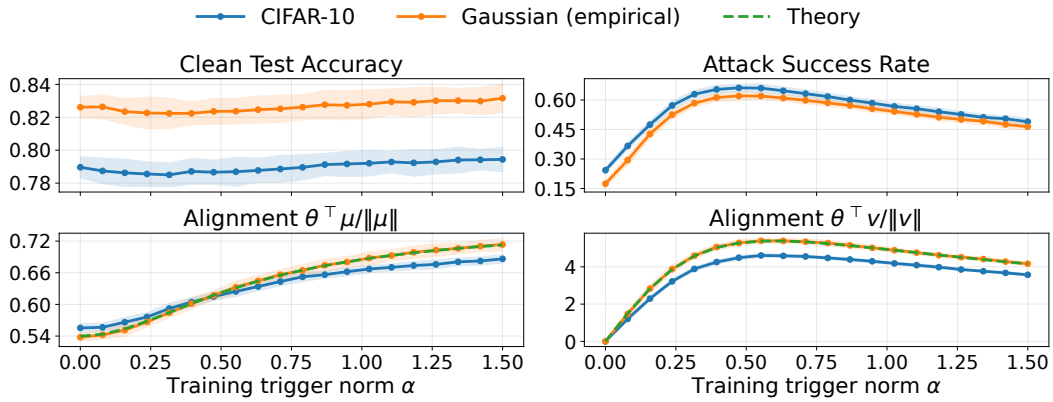


Figure 2: Logistic Regression on CIFAR-10: Real Data vs Theoretical Predictions

A plot of CIFAR-10 (classes 0 & 1) for logistic regression on real data (blue) and Gaussian surrogates (orange) compared against theoretical predictions (dashed) obtained by solving the fixed point equation in Theorem 1 numerically. Here $\phi = 0.05$; $\alpha_{\text{test}} = 0.5$. The Gaussian surrogates satisfy our modelling assumptions, and their empirical curves closely match the theory. The initial dip in clean accuracy is consistent with the contribution of trigger-direction variance, which becomes smaller at large α.

Trigger alignment peaks at a finite strength. The attack success rate is controlled by the trigger alignment $\mathbf{v}^\top \tilde{\boldsymbol{\theta}}$. Rather than growing without bound in α, the trigger alignment is maximised at a finite $\alpha_\star$ and decays thereafter: for large α the poisoned class is so well-separated that the loss provides vanishing gradient in the trigger direction (since $L'(x) \to 0$ as $x \to +\infty$ under Assumption 3). Formally, in the proportional regime $\mathbf{v}^\top \tilde{\boldsymbol{\theta}} \to 0$ as $\alpha \to \infty$, with a decay rate set by the tail of the loss (Proposition 6). Figure 2 and Figure 1 illustrate this behaviour.

Low-variance trigger directions yield stronger alignment. Trigger effectiveness also depends on direction. Our squared-loss analysis links stronger trigger alignment to lower-variance directions through the covariance spectrum. Proposition 5 establishes this for trigger directions satisfying Assumption 2, and results for general triggers are given in Proposition 9. Experiments with logistic regression and ResNet-18 similarly show greater attack success along lower-variance directions among those tested.

3.1 Exact results ERM least squares

For the squared loss $L(t) = \frac{1}{2}(1-t)^2$, the Gaussian-proxy fixed point reduces to an exact two-dimensional resolvent-Gram calculation for the proxy alignments onto the clean and trigger directions. In particular, for fixed $(\lambda, \mathbf{C}, \kappa)$, the effective scalar τ is independent of the trigger strength α, and the proxy expectation $\mathbb{E}[\tilde{\boldsymbol{\theta}}]$ lies in the resolvent-weighted $\operatorname{span}\{\mathbf{R}(\lambda,\tau)\boldsymbol{\mu}, \mathbf{R}(\lambda,\tau)\mathbf{v}\}$.

Proposition 9 (Appendix A) gives closed-form alignments for arbitrary mean-trigger overlap, without Assumption 2. In particular, for $0 < \phi < 1/2$ and linearly independent $\boldsymbol{\mu}, \mathbf{v}$, the trigger alignment has a unique finite maximiser. Below, we derive explicit peak and benign-alignment expressions under Assumption 2, followed by an eigenvector specialisation and a characterisation of trigger-direction dependence. The ratio κ enters these results through τ, with $\tau \to 1$ in the information limit $\kappa \to 0$.

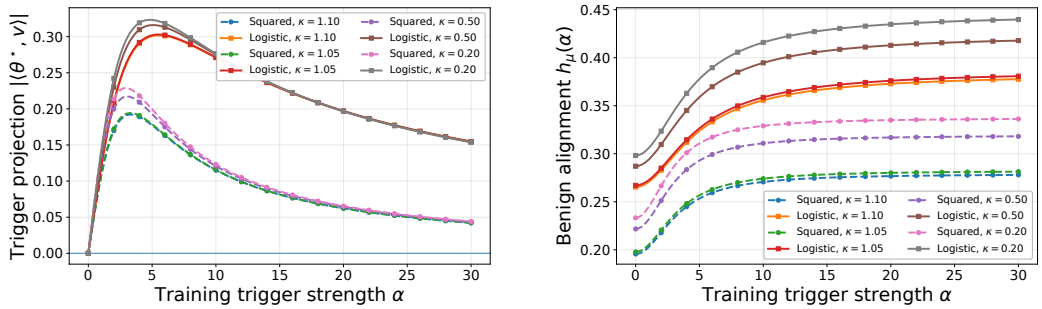
3.1.1 Trigger alignment peaks at a finite strength

Under Assumption 2, the general alignment formulas (Proposition 9) simplify and yield an explicit peak location. Define the resolvent quadratic forms

$$g_{\mu\mu} := \boldsymbol{\mu}^\top \mathbf{R}(\lambda,\tau)\boldsymbol{\mu}, \qquad g_{\mu v} := \boldsymbol{\mu}^\top \mathbf{R}(\lambda,\tau)\mathbf{v}, \qquad g_{vv} := \mathbf{v}^\top \mathbf{R}(\lambda,\tau)\mathbf{v}.$$

The following proposition characterises the trigger alignment when $g_{\mu v} = o(1)$.

Proposition 2 (Finite peak under generic trigger orthogonality)**.** *Assume that $0 < \phi < 1/2$, that $\boldsymbol{\mu}$ and $\mathbf{v}$ are linearly independent, and that Assumption 2 holds so that $g_{\mu v} = o(1)$. Consequently,*



(a) Trigger alignment $h_v(\alpha)$.

(b) Benign alignment $h_\mu(\alpha)$.

Figure 3: Exact asymptotic alignment predictions in the $\mathbf{C} = \mathbf{I}$ setting across aspect ratios $\kappa = p/n$. Dashed curves show the squared-loss predictions, while solid curves show numerical solutions of the logistic-loss fixed-point system. Left: trigger alignment $h_v(\alpha)$ peaks at a finite α. Right: benign alignment $h_\mu(\alpha)$ increases with α. The plots exhibit these behaviours for both $\kappa < 1$ and the overparameterised regime $\kappa > 1$.

uniformly for α in any fixed compact subset of $[0, \infty)$,

$$h_v(\alpha) = \frac{A_v \alpha}{B + C\alpha^2} + o(1), \tag{4}$$

where $A_v := \tau\phi g_{vv}\big(1 + 2\tau(1-\phi)g_{\mu\mu}\big)$, $B := 1 + \tau g_{\mu\mu}$, and $C := \tau\phi g_{vv}\big(1 + \tau(1-\phi)g_{\mu\mu}\big)$. In particular, for every fixed $\alpha > 0$, $h_v(\alpha) > 0$ eventually. Moreover, the unique finite maximiser $\alpha_\star$ of the exact trigger alignment satisfies

$$\alpha_\star^2 = \frac{B}{C} + o(1) = \frac{1 + \tau g_{\mu\mu}}{\tau\phi g_{vv}\big(1 + \tau(1-\phi)g_{\mu\mu}\big)} + o(1). \tag{5}$$

Thus h_v is strictly increasing on $(0, \alpha_\star)$ and strictly decreasing on $(\alpha_\star, \infty)$.

The exact arbitrary-overlap formula in Proposition 9 has a numerator that is at most linear in α and a denominator with a positive quadratic leading term. Hence $h_v(\alpha) \to 0$ as $\alpha \to \infty$. The leading-order rational formW (4) makes this mechanism transparent under generic trigger orthogonality.

3.1.2 Benign alignment is monotone to leading order

Similarly, we get a leading-order form for the benign alignment, with a quadratic-in-α numerator. The translation from benign alignment to clean accuracy is discussed in Section 4.

Corollary 3 (Leading-order benign alignment under generic trigger orthogonality)**.** *Under the assumptions of Proposition 2, uniformly for α in any fixed compact subset of $[0, \infty)$, $h_\mu(\alpha) = \frac{A_\mu^{(0)} + A_\mu^{(2)}\alpha^2}{B + C\alpha^2} + o(1)$, with B, C as in Proposition 2, $A_\mu^{(0)} := \tau(1-2\phi)g_{\mu\mu}$, and $A_\mu^{(2)} := \tau^2\phi(1-\phi)g_{\mu\mu}g_{vv}$. The leading-order curve is strictly increasing for every $\alpha > 0$.*

3.1.3 Eigenvector specialisation

To gain intuition, we restrict attention to a simplified setting in which the clean and trigger directions align with orthogonal eigenvectors of $\mathbf{C}$ such that $g_{\mu v} = 0$ and Assumption 2 trivially holds. In this case, the alignment quantities admit simplified closed-form expressions.

Corollary 4 (Eigenvector specialisation)**.** *Under the assumptions of Proposition 2, assume additionally that $\boldsymbol{\mu}^\top \mathbf{v} = 0$, $\mathbf{C}\boldsymbol{\mu} = s_\mu^2\boldsymbol{\mu}$, and $\mathbf{C}\mathbf{v} = s_v^2\mathbf{v}$. Define $A_\mu := \lambda + \tau s_\mu^2 + \tau|\boldsymbol{\mu}|^2$, $B_v := \lambda + \tau s_v^2$, $P_\mu := \lambda + \tau s_\mu^2 + \tau(1-\phi)|\boldsymbol{\mu}|^2$, $Q_\mu := \lambda + \tau s_\mu^2 + 2\tau(1-\phi)|\boldsymbol{\mu}|^2$, and $D_{\mathrm{eig}}(\alpha) := A_\mu B_v + \tau\phi P_\mu \alpha^2$. Then*

$$h_\mu(\alpha) = |\boldsymbol{\mu}|^2 \frac{\tau(1-2\phi)B_v + \tau^2\phi(1-\phi)\alpha^2}{D_{\mathrm{eig}}(\alpha)}, \qquad h_v(\alpha) = \frac{\tau\phi Q_\mu \alpha}{D_{\mathrm{eig}}(\alpha)}. \tag{6}$$

Consequently, $h_\mu(\alpha)$ is strictly increasing for $\alpha > 0$, while $h_v(\alpha) > 0$ for every $\alpha > 0$ and has a unique finite maximiser at $(\alpha_\star^{\mathrm{eig}})^2 = \frac{A_\mu B_v}{\tau\phi P_\mu}$.

Within this eigenvector specialisation, increasing s_v^2 attenuates $h_v(\alpha)$ and pushes the maximiser $\alpha_\star^{\text{eig}}$ to larger values, so the model is most sensitive along directions where the data vary least. See Appendix A for further analysis, including how the poisoning fraction ϕ modulates the strength and location of the peak (Figure 4).

Proposition 5 (Minimum-eigenvalue trigger directions)**.** *Under the assumptions of Proposition 2, for each fixed $\alpha > 0$, the leading-order trigger alignment* (4) *depends on* $\mathbf{v}$ *only through g_{vv} and is strictly increasing in g_{vv}. Consequently, any unit minimum-eigenvalue direction of* $\mathbf{C}$ *satisfying Assumption 2 maximises this alignment over directions satisfying the same assumption, attaining $g_{vv} = 1/(\lambda + \tau s_{\min}^2)$, where $s_{\min}^2$ is the smallest eigenvalue of* $\mathbf{C}$.

Proof. In (4) of Proposition 2, A_v and C are linear in g_{vv}, while B is independent of $\mathbf{v}$, from which monotonicity in g_{vv} follows by direct calculation. Over the full unit sphere, the maximum of $g_{vv} = \mathbf{v}^\top \mathbf{R}(\lambda, \tau)\mathbf{v}$ is the largest eigenvalue of $\mathbf{R}(\lambda, \tau) = (\lambda \mathbf{I} + \tau \mathbf{C})^{-1}$, namely $1/(\lambda + \tau s_{\min}^2)$, and it is attained by the unit vectors in the minimum-eigenvalue eigenspace of $\mathbf{C}$. The stated conditional conclusion follows whenever this eigenspace contains an admissible direction. □

Remark 2 (Poisoning fraction ϕ)**.** *The dependence on ϕ is analysed in Appendix A.3. To leading order, the benign alignment decreases with ϕ, while the effect of ϕ on the trigger alignment changes sign with α, it increases at small α and decreases at large α. At the $\alpha_\star$, the derivative with respect to ϕ is positive, so the maximal leading-order trigger alignment is approached as $\phi \uparrow 1/2$.*

3.2 Generalised linear model: ERM

For general convex losses in the proportional regime, the analysis is more challenging, as the formulas are no longer explicit. We show that as α increases, the overlap decays at a rate depending on the form of the loss, giving us an implicit form of the finite-α peak.

3.2.1 Trigger alignment peaks with trigger strength

Proposition 6 (Trigger alignment peaks in ERM)**.** *Under Assumptions 1–2 and the mixture model* (1)*, the mean trigger alignment $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}]$, as a function of α, is maximised at some finite $\alpha_\star \in (0, \infty)$. Specifically as $n, p \to \infty$, $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}] = g(\alpha) + o_\alpha(1)$, where $g(\cdot)$ has the asymptotic (in α) behaviour*

(i) if $|L'(x)| \le C_2/(x^{1+\epsilon} + 1)$ for $x > 0$, then $g(\alpha) = O\big(\alpha^{-\epsilon/(2+\epsilon)}\big)$;

(ii) if $|L'(x)| \le C_3\, e^{-C_4 x}$ for $x > 0$ (e.g. logistic loss)*, then $g(\alpha) = O(\log \alpha/\alpha)$.*

Moreover by Theorem 1, $\mathbf{v}^\top \hat{\boldsymbol{\theta}} - \mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}] \to 0$ in probability with $\hat{\boldsymbol{\theta}}$ defined in (2)*.*

Proof sketch. From (1), $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}]$ is controlled by $\eta_2 = \phi \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} f(M_2 + \sigma\xi)$, which depends on the poisoned-class margin M_2. The variance σ^2 is bounded independently of α since we may write $\sigma^2 = \mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{C} \tilde{\boldsymbol{\theta}}]$, and the regularisation yields the clean bound $\|\hat{\boldsymbol{\theta}}\|^2 \le 2L(0)/\lambda$, which we can pass onto $\tilde{\boldsymbol{\theta}}$. For large α, M_2 grows and pushes L' into its decaying tail in the expectation defining η_2, yielding $\eta_2 \le C/(|M_2|^{1+\epsilon} + 1)$. Combined with the resolvent bound $M_2 \le C' \eta_2 \alpha^2$, one obtains $M_2 = O(\alpha^{2/(2+\epsilon)})$ and hence $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}] = O(\alpha^{-\epsilon/(2+\epsilon)}) \to 0$. Since $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}] = 0$ at $\alpha = 0$ and is positive for small $\alpha > 0$, a maximum exists at finite $\alpha_\star$. □

Proposition 6 establishes existence and the decay rate of the finite maximiser $\alpha_\star$ but not a closed-form location; for any fixed $(\lambda, \mathbf{C}, \boldsymbol{\mu}, \phi, \kappa)$ it is computed numerically from the fixed-point nonlinear system of Theorem 1, as for the theory curves in Figures 2 and 3.

Remark 3 (Clean accuracy for general convex losses)**.** *Our numerical results support increasing benign alignment for large α, but a strict monotonicity proof for general-loss ERM remains open. Since $\eta_2 \to 0$ as $\alpha \to \infty$, the alignment decomposition* (23) *gives $M_1 \to \eta_1\, \boldsymbol{\mu}^\top \mathbf{R}(\lambda, \tau)\boldsymbol{\mu}$, its unpoisoned value, and clean test accuracy is an increasing function of M_1/σ; a complete proof of monotonicity would additionally require controlling τ and σ as α varies. What we prove is the endpoint ordering (Propositions 7 and 8, with the translation to accuracy in Appendix C.2).*

3.3 GENERALISED LINEAR MODEL: POPULATION LOSS

Turning to the population loss, we show that as $\alpha \to \infty$ the learned weights converge to those trained without the poisoned data, which for $\boldsymbol{\mu}$ an eigenvector of $\mathbf{C}$ raises benign accuracy; in general we obtain monotonicity after one step of gradient descent. Define the *benign-only* objective, obtained by dropping the poisoned term from (3) and with unique minimiser $\boldsymbol{\theta}_{\mathrm{ben}}$:

$$\mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}) \;:=\; (1-\phi)\,\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{C})}\big[L(\boldsymbol{\theta}^\top\mathbf{x})\big] \;+\; \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2. \tag{7}$$

Proposition 7 (Benign convergence)**.** *Under Assumption 3 with $\phi < \frac{1}{2}$ and $\mathbf{v}$ orthogonal to the Krylov subspace of $\mathbf{C}$ generated by $\boldsymbol{\mu}$, i.e. $\mathbf{v}^\top\mathbf{C}^k\boldsymbol{\mu} = 0$ for each $k \geq 0$ (cf. Assumption 2), the population minimiser $\boldsymbol{\theta}(\alpha)$ of (3) satisfies $\boldsymbol{\theta}(\alpha) \to \boldsymbol{\theta}_{\mathrm{ben}}$ as $\alpha \to \infty$ (defined in (7)). If additionally $\boldsymbol{\mu}$ is an eigenvector of $\mathbf{C}$, then $\boldsymbol{\mu}^\top\boldsymbol{\theta}(0) < \boldsymbol{\mu}^\top\boldsymbol{\theta}_{\mathrm{ben}}$.*

Proof sketch. Convergence to $\boldsymbol{\theta}_{\mathrm{ben}}$. Consider the competitor $\hat{\boldsymbol{\theta}}_\alpha = \boldsymbol{\theta}_{\mathrm{ben}} + \alpha^{-1/2}\mathbf{v}$. For a poisoned sample $\tilde{\mathbf{x}} \sim \mathcal{N}(\alpha\mathbf{v} - \boldsymbol{\mu}, \mathbf{C})$, the margin satisfies $\hat{\boldsymbol{\theta}}_\alpha^\top\tilde{\mathbf{x}} = \boldsymbol{\theta}_{\mathrm{ben}}^\top\tilde{\mathbf{x}} + \alpha^{-1/2}\mathbf{v}^\top\tilde{\mathbf{x}}$, where $\alpha^{-1/2}\mathbf{v}^\top\tilde{\mathbf{x}}$ has mean $\alpha^{1/2}\|\mathbf{v}\|^2 - \alpha^{-1/2}\boldsymbol{\mu}^\top\mathbf{v} \to +\infty$ and variance $\alpha^{-1}\mathbf{v}^\top\mathbf{C}\mathbf{v} \to 0$. Hence the poisoned margin diverges and the poisoned loss vanishes by Assumption 3. Since $\hat{\boldsymbol{\theta}}_\alpha \to \boldsymbol{\theta}_{\mathrm{ben}}$, we obtain $\limsup_\alpha \mathcal{L}_{\mathrm{pop}}(\hat{\boldsymbol{\theta}}_\alpha;\alpha) \leq \mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}_{\mathrm{ben}})$. The lower bound $L \geq 0$ and convexity forces $\boldsymbol{\theta}(\alpha) \to \boldsymbol{\theta}_{\mathrm{ben}}$.

Eigenvector case. When $\boldsymbol{\mu}$ is an eigenvector of $\mathbf{C}$ with eigenvalue λ_μ, the gradient in any direction orthogonal to $\boldsymbol{\mu}$ is zero at any critical point, so both $\boldsymbol{\theta}(0)$ and $\boldsymbol{\theta}_{\mathrm{ben}}$ are collinear with $\boldsymbol{\mu}$. Writing $\boldsymbol{\theta}(0) = a_0^*\boldsymbol{\mu}$ and $\boldsymbol{\theta}_{\mathrm{ben}} = a_{\mathrm{ben}}\boldsymbol{\mu}$, optimality of a_{ben} gives

$$\frac{d}{da}\mathcal{L}_{\mathrm{pop}}(a\boldsymbol{\mu};0)\Big|_{a=a_{\mathrm{ben}}} = \phi\,\mathbb{E}\big[L'(B_{\mathrm{ben}})\big(-\|\boldsymbol{\mu}\|^2 + \sqrt{\lambda_\mu}\|\boldsymbol{\mu}\|\,\xi\big)\big],$$

where $B_{\mathrm{ben}} = a_{\mathrm{ben}}(-\|\boldsymbol{\mu}\|^2 + \sqrt{\lambda_\mu}\|\boldsymbol{\mu}\|\xi)$ and $\xi \sim \mathcal{N}(0,1)$. Applying Stein's lemma yields $\phi\|\boldsymbol{\mu}\|^2(-\mathbb{E}[L'(B_{\mathrm{ben}})] + a_{\mathrm{ben}}\lambda_\mu\,\mathbb{E}[L''(B_{\mathrm{ben}})]) > 0$, forcing $a_0^* < a_{\mathrm{ben}}$. □

Under Assumption 2, Lemma 13(ii) gives $\mathbf{v}^\top\boldsymbol{\theta}_{\mathrm{ben}} = a\,\mathbf{v}^\top(\frac{\lambda}{1-\phi}\mathbf{I} + \tau\mathbf{C})^{-1}\boldsymbol{\mu} = 0$, so $\mathbf{v}^\top\boldsymbol{\theta}(\alpha) \to 0$ as $\alpha \to \infty$, while $\mathbf{v}^\top\boldsymbol{\theta}(\alpha) > 0$ for moderate α, confirming that the trigger alignment peaks at finite α in the population regime as well.

Proposition 8 (One-step monotonicity)**.** *Under Assumption 3 with $\phi < \frac{1}{2}$, $\lambda > 0$, and $\mathbf{v}^\top\boldsymbol{\mu} = 0$,*

$$\boldsymbol{\mu}^\top\nabla\mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta}_{\mathrm{ben}};\alpha) > 0 \quad \textit{for every } \alpha \geq 0.$$

In particular, one step of gradient descent on $\mathcal{L}_{\mathrm{pop}}(\,\cdot\,;\alpha)$ starting from $\boldsymbol{\theta}_{\mathrm{ben}}$ decreases $\boldsymbol{\mu}^\top\boldsymbol{\theta}(\alpha)$.

The proof is given in Appendix B.4. Combined with Proposition 7, this gives the benign-alignment ordering: $\boldsymbol{\mu}^\top\boldsymbol{\theta}(\alpha)$ rises from $\boldsymbol{\mu}^\top\boldsymbol{\theta}(0)$ toward $\boldsymbol{\mu}^\top\boldsymbol{\theta}_{\mathrm{ben}}$.

4 COMPARING ERM AND INFORMATION LIMIT

The three phenomena described in Section 3 arise in both the proportional regime and information limit. In this section, we explain precisely how alignments translate into clean accuracy and attack success rate, and why the proportional regime captures finite-sample effects that do not appear in the information limit.

From alignment to accuracy. The clean test accuracy of $\boldsymbol{\theta}$ on the Gaussian mixture depends on the *benign alignment* $\boldsymbol{\theta}^\top\boldsymbol{\mu}$ and the *total variance* $\boldsymbol{\theta}^\top\mathbf{C}\boldsymbol{\theta}$. For Φ the standard normal CDF:

$$\mathrm{Acc}_{\mathrm{clean}}(\boldsymbol{\theta}) = \Phi\Big(\boldsymbol{\theta}^\top\boldsymbol{\mu}/\sqrt{\boldsymbol{\theta}^\top\mathbf{C}\boldsymbol{\theta}}\Big), \tag{8}$$

For intuition, throughout this section we take $\boldsymbol{\mu}, \mathbf{v}$ are eigenvectors of $\mathbf{C}$ with eigenvalues s_μ^2, s_v^2; the qualitative conclusions hold in general (Appendix C.2). The asymptotic margin variance decomposes

$$\sigma^2 \;:=\; \mathbb{E}\Big[\tilde{\boldsymbol{\theta}}^\top\mathbf{C}\tilde{\boldsymbol{\theta}}\Big] \;=\; \underbrace{\tfrac{s_\mu^2}{\|\boldsymbol{\mu}\|^2}\,h_\mu^2}_{\text{mean signal}} \;+\; \underbrace{s_v^2\,h_v^2}_{\text{trigger signal}} \;+\; \underbrace{\tfrac{\gamma}{n}\,\mathrm{tr}\big[\mathbf{R}(\lambda,\tau)^2\mathbf{C}^2\big]}_{\text{noise floor }\zeta}, \tag{9}$$

the cross term vanishing by Assumption 2. The first two terms are due to the mean of $\tilde{\boldsymbol{\theta}}$; ζ captures the $O(1/n)$ fluctuations in (1) and vanishes in the information limit. Assumption 1 bounds $\mathbf{C}$ only from above; adding the mild spectral nondegeneracy of Assumption 4 gives $\zeta \geq c\,\kappa$, $c > 0$ independent of α (Lemma 18). The quadratic form concentrates on σ^2, so (8) with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ is faithful asymptotically.

In the information limit ($p \ll n$) there is no noise floor, $\sigma^2 \approx s_\mu^2 h_\mu^2/\|\boldsymbol{\mu}\|^2 + s_v^2 h_v^2$, so for large α, $h_v \to 0$ (Proposition 6), giving $\mathrm{Acc}_{\mathrm{clean}} \approx \Phi(\|\boldsymbol{\mu}\|/\sqrt{s_\mu^2})$, *independent* of h_μ and hence of α. This contradicts Figure 2, where clean accuracy is observed to rise with α (and hence with h_μ).

Why clean accuracy increases with trigger strength. The noise floor $\zeta > 0$ breaks this degeneracy. Neglecting the (small, see Table 1) trigger term, $\mathrm{Acc}_{\mathrm{clean}} \approx g(h_\mu) := \Phi(h_\mu/\sqrt{s_\mu^2 h_\mu^2/\|\boldsymbol{\mu}\|^2 + \zeta})$, which has $g'(h_\mu) > 0$ whenever $\zeta > 0$. As $\alpha \to \infty$, h_μ rises to its unpoisoned value (Proposition 7), improving clean accuracy. With $\zeta = 0$ this collapses to $\|\boldsymbol{\mu}\|/\sqrt{s_\mu^2}$ and the improvement is invisible.

Empirical validation. Table 1 decomposes σ^2 into its 4 components, including the cross term not present in the eigenvector simplification (9) using the fixed-point system of Theorem 1. The mean-direction signal carries 79.9% of the variance; the trigger and cross terms are negligible ($< 0.5\%$ combined), consistent with Assumption 2. The noise floor ζ contributes 19.67%, driven by $\kappa = p/n \approx 0.32$. Setting $\zeta = 0$ (the information limit) underestimates σ^2 and conceals the dependence on h_μ.

Table 1: Decomposition of $\sigma^2 = \mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{C}\tilde{\boldsymbol{\theta}}]$ for empirical $\boldsymbol{\mu}$, $\mathbf{C}$ from CIFAR-10 (classes 0 & 1), at $\lambda = 10^{-4}$, $\phi = 0.05$. General $\mathbf{C}$, not the eigenvector simplification. Here $\mathbf{A} := \mathbf{R}(\lambda, \tau)^2\mathbf{C}$

Component	Value	%
$(\eta_1 - \eta_2)^2\, \boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu}$	4.79	79.93
$\eta_2^2\, \mathbf{v}^\top \mathbf{A}\mathbf{v}$	0.027	0.46
$2(\eta_1 - \eta_2)\eta_2\, \boldsymbol{\mu}^\top \mathbf{A}\mathbf{v}$	-0.003	-0.05
$\zeta = \frac{\gamma}{n}\,\mathrm{tr}[\mathbf{AC}]$	1.18	19.67

Attack success rate. A test sample from the negative class poisoned with trigger $\alpha_{\mathrm{test}}\mathbf{v}$ has mean $\alpha_{\mathrm{test}}\mathbf{v} - \boldsymbol{\mu}$, giving attack success rate $\Phi((\alpha_{\mathrm{test}} h_v - h_\mu)/\sigma)$, with σ including both $s_v^2 h_v^2$ and ζ via (9). In both regimes h_v peaks at finite $\alpha_\star$ (Proposition 6 for ERM; Proposition 7 for population risk) while h_μ rises monotonically, so the attack success rate also peaks at finite α; in the proportional regime, ζ further moderates it. This effect is absent in the population limit.

5 Conclusion and Limitations

We studied backdoor poisoning in regularised GLMs on Gaussian mixtures in the high-dimensional proportional regime ($p/n \to \kappa$). Our analysis and experiments highlight three phenomena at a fixed test-time trigger budget: (i) clean accuracy can *increase* with training trigger strength, driven by a finite-sample noise floor ζ of order κ absent in the information limit; (ii) trigger alignment can peak at a finite strength and then decline; and (iii) the attack is strongest along low-variance directions, with the minimum eigenvector of the covariance matrix the most effective trigger direction. We derive closed-form asymptotic alignments for squared loss and extend the finite-peak and decay results to a broader class of convex losses under suitable assumptions. Experiments on Gaussian surrogates and CIFAR-10 support the predictions, while ResNet-18 results suggest their relevance beyond the models analysed here.

Limitations and future work. Our analysis rests on a Gaussian mixture data model and GLM classifiers; while experiments suggest the phenomena persist for deep networks, extending the theory to non-convex feature maps is an interesting future direction. Our attack model is a single additive trigger $\alpha\mathbf{v}$ of fixed direction with label flipping: this fits fixed-pattern, additive backdoors well, but not input-dependent or geometric attacks such as warping-based (WaNet), sample-specific (SSBA), or reflection-based triggers, which our theory does not cover; a warping-based experiment (Figure 7) nonetheless shows phenomena (i)–(ii) still arise empirically. Our analysis is likewise binary, although all three phenomena are still present in a 10-class ResNet-18 experiment (Figure 6); extending the theory to multi-class settings, and to non-uniform poisoning strategies, are further directions. Finally, our work characterises the attack but does not propose a defence; developing defences informed by the identified mechanisms is an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to edit the text of the paper for readability and to assist with LaTeX formatting. We did not use generative AI tools to develop the theoretical model, to propose or refine the hypotheses, or to design the experiments. All mathematical content in this paper, including every definition, proposition, and proof, was derived by the authors. Generative AI tools were applied to this material only in the same editorial capacity as to the rest of the paper, to improve readability and assist with LaTeX formatting. Generative AI tools were also used to assist with parts of the experimental code, which was reviewed and tested by the authors, and to help identify candidate related work for the literature review; all cited references were read and verified by the authors. The remaining categories requiring disclosure, including synthetic dataset generation, dataset cleaning and reformatting, translation, and qualitative or thematic data analysis, are not applicable to this work. We have reviewed all AI-assisted work and we take responsibility for the final content of this paper, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper studies backdoor poisoning attacks, so its subject matter is dual-use by nature. Our analysis is deliberately framed around the conditions under which such attacks *fail*: we identify a finite trigger strength beyond which a stronger training trigger raises clean accuracy and lowers attack success, which is information most directly actionable by defenders and auditors. We introduce no new attack that is stronger than those already documented in the literature (Gu et al., 2019; Li et al., 2024), and we release no new datasets, pretrained models, or other artifacts carrying a meaningful risk of misuse. Our experiments use only the standard, publicly available CIFAR-10 benchmark and synthetic Gaussian data, and involve no human subjects and no personally identifying information. We are aware of no conflicts of interest affecting this work.

REPRODUCIBILITY STATEMENT

All assumptions underlying our results are stated explicitly in Section 2 (Assumptions 1–3, together with the order-of-limits convention in Remark 1), and complete proofs of every proposition and corollary in the main text are given in the appendices: the squared-loss results in Appendix A, the general convex-loss results in Appendix B, and the noise-floor analysis in Appendix C.1. Full experimental details, covering datasets, preprocessing, architectures, optimisers, hyperparameters, the trigger construction, and the number of random seeds averaged over, are given in Appendix D. All experiments use the publicly available CIFAR-10 dataset or synthetic Gaussian data generated exactly as described in Section 2. Code for all CIFAR-10 experiments (Figures 1, 2, 5, 6 and 7) is included in the supplementary material, together with the fixed-point solver used for the theoretical curves in Figure 2 and the raw results from which each of these figures is plotted. The remaining figures (Figures 3, 4 and 8) use only synthetic Gaussian data and are obtained by evaluating the closed-form expressions and fixed-point equations given in the paper.

REFERENCES

Nicholas Barnfield, Hugo Cui, and Yue M. Lu. High-Dimensional Analysis of Single-Layer Attention for Sparse-Token Classification, September 2025. URL http://arxiv.org/abs/2509.25153. arXiv:2509.25153 [cs].

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A Model of Double Descent for High-dimensional Binary Linear Classification, May 2020. URL http://arxiv.org/abs/1911.05822. arXiv:1911.05822 [stat].

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

David Donoho and Andrea Montanari. High Dimensional Robust M-Estimation: Asymptotic Variance via Approximate Message Passing, November 2013. URL http://arxiv.org/abs/1310.7320. arXiv:1310.7320 [math].

Bethan Evans and Jared Tanner. Theory of Minimal Weight Perturbations in Deep Networks and its Applications for Low-Rank Activated Backdoor Attacks, 2026.

Donald Flynn and Diego Granziol. A Linear Approach to Data Poisoning, January 2026. URL http://arxiv.org/abs/2505.15175. arXiv:2505.15175 [stat].

Diego Granziol. Safety-Efficacy Trade Off: Robustness against Data-Poisoning, January 2026.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, March 2019. URL http://arxiv.org/abs/1708.06733. arXiv:1708.06733 [cs].

Trevor J. Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 19–35, May 2018. doi: 10.1109/SP.2018.00057. URL https://ieeexplore.ieee.org/document/8418594. ISSN: 2375-1207.

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

Boqi Li and Weiwei Liu. A theoretical analysis of backdoor poisoning attacks in convolutional neural networks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28309–28342. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24at.html.

Yiming Li, Tongqing Zhai, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor Attack in the Physical World, April 2021. URL http://arxiv.org/abs/2104.02361. arXiv:2104.02361 [cs].

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor Learning: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, January 2024. ISSN 2162-2388. doi: 10.1109/TNNLS.2022.3182979. URL https://ieeexplore.ieee.org/abstract/document/9802938.

Bruno Loureiro, Gabriele Sicuro, Cedric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning Gaussian Mixtures with Generalized Linear Models: Precise Asymptotics in High-dimensions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10144–10157. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/543e83748234f7cbab21aa0ade66565f-Abstract.html.

Yiwei Lu, Gautam Kamath, and Yaoliang Yu. Exploring the Limits of Model-Targeted Indiscriminate Data Poisoning Attacks. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 22856–22879. PMLR, July 2023. URL https://proceedings.mlr.press/v202/lu23e.html. ISSN: 2640-3498.

Xiaoyi Mai and Zhenyu Liao. High Dimensional Classification via Regularized and Unregularized Empirical Risk Minimization: Precise Error and Optimal Loss, November 2020. URL http://arxiv.org/abs/1905.13742. arXiv:1905.13742 [stat].

Naren Manoj and Avrim Blum. Excess capacity and backdoor poisoning. *Advances in Neural Information Processing Systems*, 34:20373–20384, 2021.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *The Annals of Statistics*, 53(2):822–853, 2025.

Anh Nguyen and Anh Tran. WaNet – Imperceptible Warping-based Backdoor Attack, March 2021. URL http://arxiv.org/abs/2102.10369. arXiv:2102.10369 [cs].

Inbar Seroussi and Ofer Zeitouni. Lower Bounds on the Generalization Error of Nonlinear Learning Models. *IEEE Transactions on Information Theory*, 68(12):7956–7970, December 2022. ISSN 1557-9654. doi: 10.1109/TIT.2022.3189760. URL https://ieeexplore.ieee.org/document/9825668/.

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples, October 2025. URL http://arxiv.org/abs/2510.07192. arXiv:2510.07192 [cs].

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized Linear Regression: A Precise Analysis of the Estimation Error. In *Proceedings of The 28th Conference on Learning Theory*, pp. 1683–1709. PMLR, June 2015. URL https://proceedings.mlr.press/v40/Thrampoulidis15.html. ISSN: 1938-7228.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-Consistent Backdoor Attacks, December 2019. URL http://arxiv.org/abs/1912.02771. arXiv:1912.02771 [stat].

Ganghua Wang, Xun Xian, Ashish Kundu, Jayanth Srinivasa, Xuan Bi, Mingyi Hong, and Jie Ding. Demystifying Poisoning Backdoor Attacks from a Statistical Perspective. 2024. URL https://openreview.net/forum?id=BPHcEpGvF8.

Pengfei Xia, Ziqiang Li, Wei Zhang, and Bin Li. Data-Efficient Backdoor Attacks, June 2022. URL http://arxiv.org/abs/2204.12281. arXiv:2204.12281 [cs].

APPENDIX CONTENTS

1 Introduction **1**
1.1 Main contributions . . . 1
1.2 Related work . . . 2

2 Setting and notation **3**
2.1 Gaussian proxy for ERM . . . 3

3 Main results **4**
3.1 Exact results ERM least squares . . . 5
3.1.1 Trigger alignment peaks at a finite strength . . . 5
3.1.2 Benign alignment is monotone to leading order . . . 6
3.1.3 Eigenvector specialisation . . . 6
3.2 Generalised linear model: ERM . . . 7
3.2.1 Trigger alignment peaks with trigger strength . . . 7
3.3 Generalised linear model: population loss . . . 8

4 Comparing ERM and information limit **8**

5 Conclusion and limitations **9**

A Additional linear-regression results and proofs **14**
A.1 General-covariance reduction . . . 14
A.2 Consequences under generic trigger orthogonality . . . 16
A.3 Dependence on the poisoning fraction . . . 18
A.4 Eigenvector reduction . . . 20
A.5 Isotropic specialisation . . . 21

B Proofs for non-quadratic losses **23**
B.1 Proof of Proposition 6 . . . 23
B.2 Cone Lemma . . . 27
B.3 Proof of Proposition 7 . . . 28
B.4 Proof of Proposition 8 . . . 31

C Comparing ERM and information limit **32**
C.1 The noise floor is of order κ . . . 32
C.2 Variance decomposition and clean accuracy beyond the eigenvector simplification . . . 32

D Further Experimental Details **34**

A ADDITIONAL LINEAR-REGRESSION RESULTS AND PROOFS

This appendix proves the linear-regression results stated in Section 3.1. We work directly with the proxy expectation

$$\mathbb{E}[\tilde{\boldsymbol{\theta}}],$$

where $\tilde{\boldsymbol{\theta}}$ is the Gaussian proxy from Theorem 1. The dependence on the trigger strength α is kept implicit. We use throughout the resolvent notation

$$\mathbf{R}(\lambda, \tau) = (\lambda \mathbf{I} + \tau \mathbf{C})^{-1}$$

introduced in Theorem 1.

The appendix is organised as follows. First we derive the square-loss reduction for general covariance and obtain the exact alignment formulas. We then expand these formulas under the generic trigger-orthogonality condition $g_{\mu v} = o(1)$, as used in Proposition 2 and Corollary 3. Finally, we prove the eigenvector specialisation and record the isotropic formulas used in the main text.

A.1 GENERAL-COVARIANCE REDUCTION

For the squared loss, the high-dimensional fixed-point equations reduce to a finite-dimensional deterministic calculation. The following proposition records the reduction and the alignment formulas used throughout the linear-regression analysis.

For the squared-loss calculation, define

$$\bar{\boldsymbol{\mu}}(\alpha) := \sum_{c=1}^{2} \pi_c \boldsymbol{\mu}_c = (1 - 2\phi)\boldsymbol{\mu} + \phi\alpha \mathbf{v}, \tag{10}$$

$$\begin{aligned} S(\alpha) &:= \sum_{c=1}^{2} \pi_c \boldsymbol{\mu}_c \boldsymbol{\mu}_c^\top \\ &= \boldsymbol{\mu}\boldsymbol{\mu}^\top - \phi\alpha(\boldsymbol{\mu}\mathbf{v}^\top + \mathbf{v}\boldsymbol{\mu}^\top) + \phi\alpha^2 \mathbf{v}\mathbf{v}^\top. \end{aligned} \tag{11}$$

Also set

$$U := [\,\boldsymbol{\mu}\ \ \mathbf{v}\,], \qquad G := U^\top \mathbf{R}(\lambda, \tau) U = \begin{pmatrix} g_{\mu\mu} & g_{\mu v} \\ g_{\mu v} & g_{vv} \end{pmatrix}, \qquad \Delta_G := \det(G) = g_{\mu\mu} g_{vv} - g_{\mu v}^2,$$

and define

$$\begin{aligned} D_{\mathrm{proj}}(\alpha) := 1 &+ \tau g_{\mu\mu} - 2\tau\phi\alpha g_{\mu v} + \tau\phi\alpha^2 g_{vv} \\ &+ \tau^2 \phi(1 - \phi)\alpha^2 \Delta_G. \end{aligned} \tag{12}$$

Proposition 9 (Square-loss proxy expectation and alignment formulas)**.** *For the squared loss $L(t) = \frac{1}{2}(1 - t)^2$, the scalar fixed point satisfies*

$$\tau = \left(1 + \frac{1}{n} \operatorname{tr}[\mathbf{C}\mathbf{R}(\lambda, \tau)]\right)^{-1}, \qquad \eta_c = \pi_c \tau (1 - M_c). \tag{13}$$

Thus, the scalar equation determining τ does not involve α.

The proxy expectation is the unique solution of

$$\big(\lambda \mathbf{I} + \tau \mathbf{C} + \tau S(\alpha)\big)\mathbb{E}[\tilde{\boldsymbol{\theta}}] = \tau \bar{\boldsymbol{\mu}}(\alpha), \tag{14}$$

and satisfies

$$\mathbb{E}[\tilde{\boldsymbol{\theta}}] \in \operatorname{span}\big\{\mathbf{R}(\lambda, \tau)\boldsymbol{\mu},\, \mathbf{R}(\lambda, \tau)\mathbf{v}\big\}.$$

If $\boldsymbol{\mu}$ and $\mathbf{v}$ are linearly independent, then $G \succ 0$ and $\Delta_G > 0$, and

$$h_\mu(\alpha) = \frac{\tau\Big((1 - 2\phi) g_{\mu\mu} + \phi\alpha g_{\mu v} + \tau\phi(1 - \phi)\alpha^2 \Delta_G\Big)}{D_{\mathrm{proj}}(\alpha)}, \tag{15}$$

$$h_v(\alpha) = \frac{\tau\Big((1 - 2\phi) g_{\mu v} + \phi\alpha g_{vv} + 2\tau\phi(1 - \phi)\alpha \Delta_G\Big)}{D_{\mathrm{proj}}(\alpha)}. \tag{16}$$

Proof. The proximal formula follows by direct minimisation:

$$\operatorname{prox}_{\delta L}(x) = \arg\min_{u\in\mathbb{R}} \left\{ \frac{\delta}{2}(1-u)^2 + \frac{1}{2}(u-x)^2 \right\} = \frac{x+\delta}{1+\delta}.$$

Hence

$$f(x) := -L'\big(\operatorname{prox}_{\delta L}(x)\big) = 1 - \operatorname{prox}_{\delta L}(x) = \frac{1-x}{1+\delta}.$$

With $\tau = (1+\delta)^{-1}$, this becomes

$$f(x) = \tau(1-x), \qquad -f'(x) = \tau.$$

Substituting this identity into Theorem 1 gives

$$\tau = \frac{1}{1+\delta}, \qquad \delta = \frac{1}{n}\operatorname{tr}[\mathbf{C}\mathbf{R}(\lambda,\tau)]\,.$$

Moreover, since $r_c \sim \mathcal{N}(M_c, \sigma^2)$,

$$\eta_c = \pi_c \mathbb{E}[f(r_c)] = \pi_c \tau(1-M_c),$$

which proves (13).

Let

$$m(\alpha) := \eta_1 \boldsymbol{\mu}_1 + \eta_2 \boldsymbol{\mu}_2.$$

Taking expectations in the Gaussian proxy representation gives

$$(\lambda \mathbf{I} + \tau \mathbf{C})\mathbb{E}[\tilde{\boldsymbol{\theta}}] = m(\alpha).$$

Using $\eta_c = \pi_c\tau(1-M_c)$ and $M_c = \boldsymbol{\mu}_c^\top \mathbb{E}[\tilde{\boldsymbol{\theta}}]$, we obtain

$$m(\alpha) = \tau \sum_{c=1}^{2} \pi_c \boldsymbol{\mu}_c - \tau \sum_{c=1}^{2} \pi_c \boldsymbol{\mu}_c \boldsymbol{\mu}_c^\top \mathbb{E}[\tilde{\boldsymbol{\theta}}] = \tau \bar{\boldsymbol{\mu}}(\alpha) - \tau S(\alpha)\mathbb{E}[\tilde{\boldsymbol{\theta}}].$$

Therefore

$$(\lambda \mathbf{I} + \tau \mathbf{C} + \tau S(\alpha))\mathbb{E}[\tilde{\boldsymbol{\theta}}] = \tau \bar{\boldsymbol{\mu}}(\alpha),$$

which proves (14). The coefficient matrix is positive definite:

$$\lambda \mathbf{I} \succ 0, \qquad \tau \mathbf{C} \succeq 0, \qquad S(\alpha) = \sum_{c=1}^{2} \pi_c \boldsymbol{\mu}_c \boldsymbol{\mu}_c^\top \succeq 0.$$

Thus the solution is unique. Expanding $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$, $\boldsymbol{\mu}_2 = \alpha \mathbf{v} - \boldsymbol{\mu}$, and $(\pi_1, \pi_2) = (1-\phi, \phi)$ gives (10) and (11).

It remains to derive the alignment formulas. Write

$$d(\alpha) := \begin{pmatrix} 1-2\phi \\ \phi\alpha \end{pmatrix}, \qquad K(\alpha) := \begin{pmatrix} 1 & -\phi\alpha \\ -\phi\alpha & \phi\alpha^2 \end{pmatrix}.$$

Then

$$\bar{\boldsymbol{\mu}}(\alpha) = U d(\alpha), \qquad S(\alpha) = U K(\alpha) U^\top.$$

Since

$$m(\alpha) = \eta_1 \boldsymbol{\mu}_1 + \eta_2 \boldsymbol{\mu}_2 \in \operatorname{span}\{\boldsymbol{\mu}, \mathbf{v}\},$$

there is a unique vector $x(\alpha) \in \mathbb{R}^2$ such that

$$m(\alpha) = U x(\alpha).$$

Moreover,

$$\mathbb{E}[\tilde{\boldsymbol{\theta}}] = \mathbf{R}(\lambda,\tau) U x(\alpha).$$

The identity

$$m(\alpha) = \tau \bar{\boldsymbol{\mu}}(\alpha) - \tau S(\alpha)\mathbb{E}[\tilde{\boldsymbol{\theta}}]$$

therefore becomes

$$U x = \tau U d - \tau U K U^\top \mathbf{R}(\lambda,\tau) U x.$$

Since U has full column rank,

$$x = \tau d - \tau K G x, \qquad G := U^\top \mathbf{R}(\lambda, \tau) U.$$

Equivalently,

$$(I_2 + \tau K(\alpha) G) x(\alpha) = \tau d(\alpha).$$

Now define

$$h(\alpha) := \begin{pmatrix} h_\mu(\alpha) \\ h_v(\alpha) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}^\top \mathbb{E}[\tilde{\boldsymbol{\theta}}] \\ \mathbf{v}^\top \mathbb{E}[\tilde{\boldsymbol{\theta}}] \end{pmatrix}.$$

Since $\mathbb{E}[\tilde{\boldsymbol{\theta}}] = \mathbf{R}(\lambda, \tau) U x(\alpha)$, we have

$$h(\alpha) = G x(\alpha).$$

Multiplying the reduced system by G gives

$$(I_2 + \tau G K(\alpha)) h(\alpha) = \tau G d(\alpha).$$

This is a 2×2 linear system. Its determinant is

$$\det(I_2 + \tau G K(\alpha)) = 1 + \tau \operatorname{tr}(G K(\alpha)) + \tau^2 \det(G) \det(K(\alpha)).$$

Because

$$\operatorname{tr}(G K(\alpha)) = g_{\mu\mu} - 2\phi\alpha g_{\mu v} + \phi\alpha^2 g_{vv}$$

and

$$\det(K(\alpha)) = \phi(1-\phi)\alpha^2,$$

the determinant is precisely $D_{\mathrm{proj}}(\alpha)$ in (12).

We also note that this denominator is positive. Indeed, $K(\alpha) \succeq 0$ and $G \succ 0$. Hence $GK(\alpha)$ is similar to the symmetric positive semidefinite matrix $G^{1/2} K(\alpha) G^{1/2}$, so all eigenvalues of $GK(\alpha)$ are nonnegative. Since $\tau > 0$, the matrix $I_2 + \tau G K(\alpha)$ is invertible and

$$D_{\mathrm{proj}}(\alpha) > 0.$$

Applying Cramer's rule to

$$(I_2 + \tau G K(\alpha)) h(\alpha) = \tau G d(\alpha)$$

yields (15) and (16). □

A.2 Consequences under generic trigger orthogonality

We now expand the general-covariance formulas under the condition

$$g_{\mu v} = \boldsymbol{\mu}^\top \mathbf{R}(\lambda, \tau) \mathbf{v} = o(1),$$

which follows from Assumption 2. This is the condition used in Proposition 2 and Corollary 3.

Proof of Proposition 2. Set

$$\varepsilon := g_{\mu v}, \qquad m := g_{\mu\mu}, \qquad q := g_{vv}, \qquad \Delta_\varepsilon := mq - \varepsilon^2.$$

By Proposition 9,

$$h_v(\alpha) = \frac{\tau\Big((1-2\phi)\varepsilon + \phi\alpha q + 2\tau\phi(1-\phi)\alpha\Delta_\varepsilon\Big)}{1 + \tau m - 2\tau\phi\alpha\varepsilon + \tau\phi\alpha^2 q + \tau^2\phi(1-\phi)\alpha^2\Delta_\varepsilon}.$$

Equivalently,

$$h_v(\alpha) = \frac{n_\varepsilon + A_\varepsilon \alpha}{B + d_\varepsilon \alpha + C_\varepsilon \alpha^2},$$

where

$$n_\varepsilon := \tau(1-2\phi)\varepsilon, \qquad B := 1 + \tau m, \qquad d_\varepsilon := -2\tau\phi\varepsilon,$$

and

$$A_\varepsilon := \tau\phi\Big(q + 2\tau(1-\phi)\Delta_\varepsilon\Big), \qquad C_\varepsilon := \tau\phi\Big(q + \tau(1-\phi)\Delta_\varepsilon\Big).$$

Since $\mathbf{R}(\lambda,\tau) \succ 0$ and $\boldsymbol{\mu}, \mathbf{v}$ are linearly independent, $\Delta_\varepsilon > 0$. Hence

$$A_\varepsilon > 0, \qquad C_\varepsilon > 0.$$

If $\varepsilon = o(1)$, then

$$\Delta_\varepsilon = mq + o(1), \qquad n_\varepsilon = o(1), \qquad d_\varepsilon = o(1),$$

and

$$A_\varepsilon = \tau\phi q\big(1 + 2\tau(1-\phi)m\big) + o(1),$$
$$C_\varepsilon = \tau\phi q\big(1 + \tau(1-\phi)m\big) + o(1).$$

Therefore, uniformly for α in any fixed compact subset of $[0,\infty)$,

$$h_v(\alpha) = \frac{\tau\phi q\big(1 + 2\tau(1-\phi)m\big)\alpha}{1 + \tau m + \tau\phi q\big(1 + \tau(1-\phi)m\big)\alpha^2} + o(1).$$

This is exactly (4), with

$$A_v = \tau\phi q\big(1 + 2\tau(1-\phi)m\big), \qquad C = \tau\phi q\big(1 + \tau(1-\phi)m\big).$$

It remains to prove the finite-peak claim for the exact expression. Differentiating

$$h_v(\alpha) = \frac{n_\varepsilon + A_\varepsilon\alpha}{B + d_\varepsilon\alpha + C_\varepsilon\alpha^2}$$

gives

$$h_v'(\alpha) = \frac{F_\varepsilon(\alpha)}{\big(B + d_\varepsilon\alpha + C_\varepsilon\alpha^2\big)^2},$$

where

$$F_\varepsilon(\alpha) = A_\varepsilon B - n_\varepsilon d_\varepsilon - 2n_\varepsilon C_\varepsilon\alpha - A_\varepsilon C_\varepsilon\alpha^2.$$

The quadratic F_ε has negative leading coefficient and

$$F_\varepsilon(0) = A_\varepsilon B - n_\varepsilon d_\varepsilon = A_\varepsilon B + 2\tau^2\phi(1-2\phi)\varepsilon^2 > 0.$$

The discriminant of F_ε is

$$(2n_\varepsilon C_\varepsilon)^2 + 4A_\varepsilon C_\varepsilon\big(A_\varepsilon B - n_\varepsilon d_\varepsilon\big) > 0.$$

Thus F_ε has two real roots. Their product is

$$\frac{A_\varepsilon B - n_\varepsilon d_\varepsilon}{-A_\varepsilon C_\varepsilon} < 0,$$

so exactly one root is positive. Denote this positive root by $\alpha_{\star,\varepsilon}$. Therefore h_v is strictly increasing on $(0,\alpha_{\star,\varepsilon})$ and strictly decreasing on $(\alpha_{\star,\varepsilon},\infty)$.

The positive root is

$$\alpha_{\star,\varepsilon} = -\frac{n_\varepsilon}{A_\varepsilon} + \left[\left(\frac{n_\varepsilon}{A_\varepsilon}\right)^2 + \frac{B - \frac{n_\varepsilon d_\varepsilon}{A_\varepsilon}}{C_\varepsilon}\right]^{1/2}.$$

Since $n_\varepsilon, d_\varepsilon = o(1)$, $A_\varepsilon = A_v + o(1)$, and $C_\varepsilon = C + o(1)$, we obtain

$$\alpha_{\star,\varepsilon}^2 = \frac{B}{C} + o(1),$$

as claimed. □

Proof of Corollary 3. Set again

$$\varepsilon := g_{\mu v}, \qquad m := g_{\mu\mu}, \qquad q := g_{vv}, \qquad \Delta_\varepsilon := mq - \varepsilon^2.$$

By Proposition 9,

$$h_\mu(\alpha) = \frac{\tau\Big((1-2\phi)m + \phi\alpha\varepsilon + \tau\phi(1-\phi)\alpha^2\Delta_\varepsilon\Big)}{1 + \tau m - 2\tau\phi\alpha\varepsilon + \tau\phi\alpha^2 q + \tau^2\phi(1-\phi)\alpha^2\Delta_\varepsilon}.$$

If $\varepsilon = o(1)$, then, uniformly for α in any fixed compact subset of $[0, \infty)$,

$$h_\mu(\alpha) = \frac{\tau(1-2\phi)m + \tau^2\phi(1-\phi)mq\,\alpha^2}{1 + \tau m + \tau\phi q\big(1+\tau(1-\phi)m\big)\alpha^2} + o(1).$$

This is the claimed leading-order formula with

$$A_\mu^{(0)} = \tau(1-2\phi)m, \qquad A_\mu^{(2)} = \tau^2\phi(1-\phi)mq.$$

Set

$$B := 1 + \tau m, \qquad C := \tau\phi q\big(1+\tau(1-\phi)m\big).$$

The leading-order curve has the form

$$h_\mu^{(0)}(\alpha) = \frac{A_\mu^{(0)} + A_\mu^{(2)}\alpha^2}{B + C\alpha^2}.$$

Differentiating gives

$$\frac{d}{d\alpha}h_\mu^{(0)}(\alpha) = \frac{2\alpha\big(A_\mu^{(2)}B - A_\mu^{(0)}C\big)}{(B + C\alpha^2)^2}.$$

A direct simplification yields

$$A_\mu^{(2)}B - A_\mu^{(0)}C = \tau^2\phi^2 mq\big(1 + 2\tau(1-\phi)m\big) > 0.$$

Hence the leading-order curve is strictly increasing for every $\alpha > 0$.

Finally, on every compact interval $K \subset (0, \infty)$, the derivative of $h_\mu^{(0)}$ is bounded below by a positive constant. The exact derivative is a continuous rational function of (ε, α), and the denominator is bounded away from zero on K. Therefore the exact derivative converges uniformly to the derivative of $h_\mu^{(0)}$ on K. It follows that

$$h'_\mu(\alpha) > 0$$

eventually uniformly over $\alpha \in K$.

□

A.3 Dependence on the poisoning fraction

We also record how the square-loss alignments depend on the poisoning fraction ϕ. This complements the dependence on the trigger strength α: at fixed α, increasing the fraction of poisoned examples can have different effects on the benign and trigger alignments.

Figure 4 illustrates this dependence. The left panel plots the trigger alignment $h_v(\alpha)$ for several poisoning fractions, and the right panel plots the benign alignment $h_\mu(\alpha)$. The boundary curves $\phi = 0$ and, if shown, $\phi = 1/2$, are included only as visual references; the formal statements in the main text assume $0 < \phi < 1/2$.

To explain the behaviour, write

$$m := g_{\mu\mu}, \qquad q := g_{vv},$$

and define

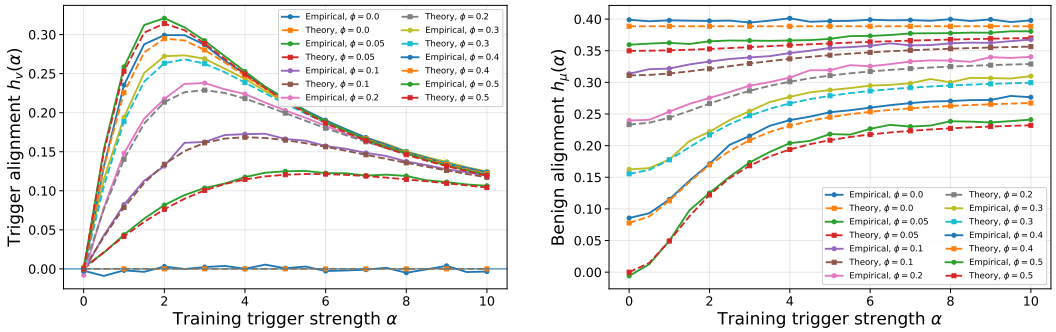
$$D_\phi(\alpha) := 1 + \tau m + \tau\phi q\big(1+\tau(1-\phi)m\big)\alpha^2.$$

In the square-loss setting, τ, m, q are fixed with respect to ϕ: the scalar τ depends only on $(\lambda, \mathbf{C}, \kappa)$, not on the poisoning fraction. Under generic trigger orthogonality, the leading-order trigger alignment is

$$h_v(\alpha, \phi) = \frac{\tau\phi q\big(1+2\tau(1-\phi)m\big)\alpha}{D_\phi(\alpha)} + o(1).$$

Differentiating the exact rational expression and then using $g_{\mu v} = o(1)$ gives

$$\frac{\partial}{\partial\phi}h_v(\alpha, \phi) = \frac{\tau q\alpha\left[(1+\tau m)\big(1+2\tau m(1-2\phi)\big) - \phi^2\tau^2 mq\alpha^2\right]}{D_\phi(\alpha)^2} + o(1).$$



(a) Trigger alignment $h_v(\alpha)$.

(b) Benign alignment $h_\mu(\alpha)$.

Figure 4: Effect of the poisoning fraction ϕ in the square-loss model. Curves show empirical ridge estimates and the corresponding theory predictions as the training trigger strength α varies. Increasing ϕ can increase the trigger alignment at small α, but can decrease it once α is large. The benign alignment decreases with ϕ at fixed α, to leading order.

Since $D_\phi(\alpha) > 0$, the leading-order sign is controlled by

$$(1+\tau m)\big(1+2\tau m(1-2\phi)\big) - \phi^2\tau^2 mq\alpha^2.$$

Thus the trigger alignment is not monotone in ϕ in general. For sufficiently small positive α, the first term dominates, so increasing ϕ increases h_v. For sufficiently large α, the negative term

$$-\phi^2\tau^2 mq\alpha^2$$

dominates, so increasing ϕ decreases h_v.

Remark 4 (Sign at the alignment peak)**.** *The sign flip above occurs strictly beyond the trigger-alignment peak: at $\alpha = \alpha_\star$, increasing the poisoning fraction always* increases h_v. *Recall from* (4) *that the peak sits at $\alpha_\star^2 = B/C = (1+\tau m)/\big(\tau\phi q(1+\tau(1-\phi)m)\big)$. Substituting this into the bracket controlling the sign of $\partial_\phi h_v$ collapses the negative term to $\phi^2\tau^2 mq\,\alpha_\star^2 = \phi\tau m(1+\tau m)/(1+\tau(1-\phi)m)$, so*

$$\left.\frac{\partial h_v}{\partial\phi}\right|_{\alpha=\alpha_\star} \propto (1+\tau m)\left[\big(1+2\tau m(1-2\phi)\big) - \frac{\phi\tau m}{1+\tau(1-\phi)m}\right].$$

Writing $u := \tau m > 0$ and clearing the positive denominator $1+u(1-\phi)$, the bracket has the same sign as

$$1+3u(1-2\phi)+2u^2(1-2\phi)(1-\phi),$$

where we used $(1-\phi)+2(1-2\phi)-\phi = 3(1-2\phi)$. Every term is positive for $u > 0$ and $\phi \in (0,1/2)$, so $\partial_\phi h_v\big|_{\alpha_\star} > 0$ for all admissible ϕ. Hence the region in which a larger poisoning fraction reduces *the trigger alignment lies entirely within the range of trigger strengths on which h_v is already decaying in α.*

The benign alignment has a simpler dependence on ϕ. Its leading-order form is

$$h_\mu(\alpha,\phi) = \frac{\tau(1-2\phi)m + \tau^2\phi(1-\phi)mq\alpha^2}{D_\phi(\alpha)} + o(1).$$

Again, differentiating the exact rational expression and then using $g_{\mu v} = o(1)$ gives

$$\frac{\partial}{\partial\phi}h_\mu(\alpha,\phi) = -\frac{m\tau\left[2+2\tau m+2\tau q\alpha^2\phi+2\tau^2 mq\alpha^2\phi^2+\tau^2q^2\alpha^4\phi^2\right]}{D_\phi(\alpha)^2} + o(1).$$

All terms inside the bracket are positive and $D_\phi(\alpha) > 0$. Hence, to leading order,

$$\frac{\partial}{\partial\phi}h_\mu(\alpha,\phi) < 0.$$

Thus, increasing the poisoning fraction decreases the benign alignment at fixed α, while its effect on the trigger alignment depends on the trigger strength.

A.4 EIGENVECTOR REDUCTION

We next prove the eigenvector specialisation directly for the alignments $h_\mu(\alpha)$ and $h_v(\alpha)$. This avoids introducing auxiliary coefficient notation and mirrors the statement in the main text.

Proof of Corollary 4. Assume

$$\boldsymbol{\mu}^\top \mathbf{v} = 0, \qquad \mathbf{C}\boldsymbol{\mu} = s_\mu^2 \boldsymbol{\mu}, \qquad \mathbf{C}\mathbf{v} = s_v^2 \mathbf{v}.$$

Set

$$L_\mu := \lambda + \tau s_\mu^2, \qquad r := \|\boldsymbol{\mu}\|^2.$$

Then

$$A_\mu = L_\mu + \tau r, \qquad B_v = \lambda + \tau s_v^2,$$

and

$$P_\mu = L_\mu + \tau(1-\phi)r, \qquad Q_\mu = L_\mu + 2\tau(1-\phi)r.$$

Since

$$\mathbf{R}(\lambda,\tau) = (\lambda \mathbf{I} + \tau \mathbf{C})^{-1},$$

the eigenvector assumptions give

$$\mathbf{R}(\lambda,\tau)\boldsymbol{\mu} = \frac{1}{L_\mu}\boldsymbol{\mu}, \qquad \mathbf{R}(\lambda,\tau)\mathbf{v} = \frac{1}{B_v}\mathbf{v}.$$

Therefore

$$g_{\mu\mu} = \boldsymbol{\mu}^\top \mathbf{R}(\lambda,\tau)\boldsymbol{\mu} = \frac{r}{L_\mu},$$

$$g_{\mu v} = \boldsymbol{\mu}^\top \mathbf{R}(\lambda,\tau)\mathbf{v} = 0,$$

and, using the standing normalisation $\|\mathbf{v}\| = 1$,

$$g_{vv} = \mathbf{v}^\top \mathbf{R}(\lambda,\tau)\mathbf{v} = \frac{1}{B_v}.$$

Hence

$$\Delta_G = g_{\mu\mu} g_{vv} - g_{\mu v}^2 = \frac{r}{L_\mu B_v}.$$

Substituting these identities into the denominator (12) gives

$$D_{\mathrm{proj}}(\alpha) = 1 + \tau \frac{r}{L_\mu} + \tau\phi\alpha^2 \frac{1}{B_v} + \tau^2\phi(1-\phi)\alpha^2 \frac{r}{L_\mu B_v}.$$

Equivalently,

$$D_{\mathrm{proj}}(\alpha) = \frac{A_\mu B_v + \tau\phi P_\mu \alpha^2}{L_\mu B_v}.$$

Thus, with

$$D_{\mathrm{eig}}(\alpha) := A_\mu B_v + \tau\phi P_\mu \alpha^2,$$

we have

$$D_{\mathrm{proj}}(\alpha) = \frac{D_{\mathrm{eig}}(\alpha)}{L_\mu B_v}.$$

Now use the exact alignment formula (15). Since $g_{\mu v} = 0$,

$$h_\mu(\alpha) = \frac{\tau\left((1-2\phi)\frac{r}{L_\mu} + \tau\phi(1-\phi)\alpha^2 \frac{r}{L_\mu B_v}\right)}{D_{\mathrm{proj}}(\alpha)}.$$

Multiplying numerator and denominator by $L_\mu B_v$, we get

$$h_\mu(\alpha) = r\frac{\tau(1-2\phi)B_v + \tau^2\phi(1-\phi)\alpha^2}{D_{\mathrm{eig}}(\alpha)}.$$

This is the claimed formula for the benign alignment.

Similarly, using (16) and $g_{\mu v} = 0$,

$$h_v(\alpha) = \frac{\tau\left(\phi\alpha\frac{1}{B_v} + 2\tau\phi(1-\phi)\alpha\frac{r}{L_\mu B_v}\right)}{D_{\mathrm{proj}}(\alpha)}.$$

Again multiplying numerator and denominator by $L_\mu B_v$, we obtain

$$h_v(\alpha) = \frac{\tau\phi\alpha\,(L_\mu + 2\tau(1-\phi)r)}{D_{\mathrm{eig}}(\alpha)} = \frac{\tau\phi Q_\mu\alpha}{D_{\mathrm{eig}}(\alpha)}.$$

This proves the alignment formulas in (6).

It remains to prove the monotonicity and peak claims. Write

$$D_{\mathrm{eig}}(\alpha) = D_0 + D_2\alpha^2, \qquad D_0 := A_\mu B_v, \qquad D_2 := \tau\phi P_\mu.$$

For the trigger alignment,

$$h_v(\alpha) = \frac{\tau\phi Q_\mu\alpha}{D_0 + D_2\alpha^2}.$$

Since $\tau\phi Q_\mu > 0$,

$$h_v'(\alpha) = \frac{\tau\phi Q_\mu(D_0 - D_2\alpha^2)}{(D_0 + D_2\alpha^2)^2}.$$

Thus h_v is strictly increasing when $\alpha^2 < D_0/D_2$ and strictly decreasing when $\alpha^2 > D_0/D_2$. Its unique maximiser satisfies

$$(\alpha_\star^{\mathrm{eig}})^2 = \frac{D_0}{D_2} = \frac{A_\mu B_v}{\tau\phi P_\mu},$$

For the benign alignment, write

$$h_\mu(\alpha) = r\frac{N_0 + N_2\alpha^2}{D_0 + D_2\alpha^2},$$

where

$$N_0 := \tau(1-2\phi)B_v, \qquad N_2 := \tau^2\phi(1-\phi).$$

Differentiating gives

$$h_\mu'(\alpha) = r\frac{2\alpha(N_2D_0 - N_0D_2)}{(D_0 + D_2\alpha^2)^2}.$$

Using the definitions above,

$$N_2D_0 - N_0D_2 = \tau^2\phi^2B_v\big(L_\mu + 2\tau(1-\phi)r\big) = \tau^2\phi^2B_vQ_\mu.$$

This quantity is strictly positive. Therefore

$$h_\mu'(\alpha) > 0 \qquad \text{for every } \alpha > 0.$$

This completes the proof of Corollary 4. □

A.5 Isotropic specialisation

For completeness, we record the fully expanded isotropic formulas referenced in the main text.

Corollary 10 (Isotropic specialisation)**.** *Assume*

$$\mathbf{C} = \mathbf{I}, \qquad \boldsymbol{\mu}^\top\mathbf{v} = 0, \qquad \|\mathbf{v}\| = 1.$$

Then $s_\mu^2 = s_v^2 = 1$, and Corollary 4 gives

$$\begin{aligned} h_\mu(\alpha) &= \mathbb{E}[\boldsymbol{\mu}^\top\tilde{\boldsymbol{\theta}}] \\ &= \|\boldsymbol{\mu}\|^2\frac{\tau(1-2\phi)(\lambda+\tau) + \tau^2\phi(1-\phi)\alpha^2}{(\lambda+\tau+\tau\|\boldsymbol{\mu}\|^2)(\lambda+\tau) + \tau\phi\big(\lambda+\tau+\tau(1-\phi)\|\boldsymbol{\mu}\|^2\big)\alpha^2}, \end{aligned} \tag{17}$$

$$\begin{aligned} h_v(\alpha) &= \mathbb{E}[\mathbf{v}^\top\tilde{\boldsymbol{\theta}}] \\ &= \frac{\tau\phi\alpha\big(\lambda+\tau+2\tau(1-\phi)\|\boldsymbol{\mu}\|^2\big)}{(\lambda+\tau+\tau\|\boldsymbol{\mu}\|^2)(\lambda+\tau) + \tau\phi\big(\lambda+\tau+\tau(1-\phi)\|\boldsymbol{\mu}\|^2\big)\alpha^2}. \end{aligned} \tag{18}$$

The trigger alignment has a unique maximiser at

$$(\alpha^\star_{\text{iso}})^2 = \frac{(\lambda+\tau+\tau\|\boldsymbol{\mu}\|^2)(\lambda+\tau)}{\tau\phi\big(\lambda+\tau+\tau(1-\phi)\|\boldsymbol{\mu}\|^2\big)}. \tag{19}$$

In particular, $h_v(\alpha) > 0$ for every $\alpha > 0$, increases up to $\alpha^\star_{\text{iso}}$, and decreases thereafter.

Proof. If $\mathbf{C} = \mathbf{I}$, then

$$s_\mu^2 = s_v^2 = 1.$$

Substituting these values into Corollary 4 gives

$$A_\mu = \lambda+\tau+\tau\|\boldsymbol{\mu}\|^2, \qquad B_v = \lambda+\tau,$$

and

$$P_\mu = \lambda+\tau+\tau(1-\phi)\|\boldsymbol{\mu}\|^2, \qquad Q_\mu = \lambda+\tau+2\tau(1-\phi)\|\boldsymbol{\mu}\|^2.$$

The alignment formulas (17) and (18) follow by substituting these identities into (6). The peak location (19) then follows. □

The isotropic form (19) admits a transparent two-sided bound that isolates the dependence of the peak on the dimensionality ratio, the regularisation, and the class separation.

Corollary 11 (Bounds on the peak location)**.** *Assume $\mathbf{C} = \mathbf{I}$, $\boldsymbol{\mu}^\top\mathbf{v} = 0$, $\|\mathbf{v}\| = 1$, and $0 < \phi < 1/2$. Write $m := \|\boldsymbol{\mu}\|^2$ for the squared class separation and let $\tau \in (0,1)$ be the scalar of* (13)*, which for isotropic covariance reduces to the ridge fixed point*

$$\tau^2 + (\lambda+\kappa-1)\tau - \lambda = 0, \qquad \tau = \tfrac{1}{2}\Big(1-\lambda-\kappa+\sqrt{(1-\lambda-\kappa)^2+4\lambda}\Big), \tag{20}$$

where $\kappa = p/n$. Then the peak location (19) *satisfies*

$$\frac{\lambda+\kappa+\tau}{\phi} \;\le\; (\alpha^\star_{\text{iso}})^2 \;\le\; \frac{\lambda+\kappa+\tau}{\phi(1-\phi)}. \tag{21}$$

In particular, using $\tau \in (0,1)$,

$$\frac{\lambda+\kappa}{\phi} \;\le\; (\alpha^\star_{\text{iso}})^2 \;\le\; \frac{\lambda+\kappa+1}{\phi(1-\phi)},$$

so the peak is independent of the class separation m up to a bounded factor in $(1,(1-\phi)^{-1})$.

Proof. For $\mathbf{C} = \mathbf{I}$ we have $\delta = \frac{1}{n}\operatorname{tr}[(\lambda\mathbf{I}+\tau\mathbf{I})^{-1}] = (p/n)/(\lambda+\tau) = \kappa/(\lambda+\tau)$, so $\tau = (1+\delta)^{-1}$ gives (20) and the identity

$$\frac{\lambda+\tau}{\tau} = (\lambda+\tau)(1+\delta) = \lambda+\tau+\kappa. \tag{22}$$

Writing $\rho := \lambda+\tau$, the closed form (19) factors as

$$(\alpha^\star_{\text{iso}})^2 = \frac{\lambda+\tau}{\tau\phi}\cdot\frac{\rho+\tau m}{\rho+\tau(1-\phi)m} = \frac{\lambda+\kappa+\tau}{\phi}\,r(m), \qquad r(m) := \frac{\rho+\tau m}{\rho+\tau(1-\phi)m},$$

using (22). The factor r is increasing in m, since $r'(m) = \tau\rho\phi\,(\rho+\tau(1-\phi)m)^{-2} > 0$, with $r(0) = 1$ and $r(m) \uparrow (1-\phi)^{-1}$ as $m \to \infty$; hence $1 \le r(m) < (1-\phi)^{-1}$ for all $m \ge 0$. Substituting these bounds gives (21). The separation-free bounds follow from $0 < \tau < 1$. □

B PROOFS FOR NON-QUADRATIC LOSSES

Throughout this appendix, we write $\mathbf{R} = \mathbf{R}(\lambda, \tau) = (\lambda \mathbf{I} + \tau \mathbf{C})^{-1}$ for the resolvent.

Remark 5 (Conventions for the proofs)**.** *Assumption 1 is really a statement about a sequence $\{(\boldsymbol{\mu}_n, \mathbf{C}_n, \mathbf{v}_n)\}$ in dimension p_n with $p_n/n \to \kappa$; we suppress the index throughout and write "$= o(1)$" for terms vanishing along this sequence as $n \to \infty$ with α fixed. All constants in the bounds below depend only on L, λ, ϕ, $\sup_n \|\mathbf{C}_n\|$ and $\sup_n \|\boldsymbol{\mu}_n\|$. We write "$= o_\alpha(1)$" for terms vanishing in the iterated limit of Remark 1, that is, first $n, p \to \infty$ at fixed α and then $\alpha \to \infty$. This covers both terms that decay in α and terms of the form $\alpha \cdot o(1)$, such as the products of α with the orthogonality residual $\mathbf{v}^\top \mathbf{R} \boldsymbol{\mu}$ that appear in Step 2 below: for each fixed α these vanish as $n \to \infty$ by Assumption 2, and the outer limit in α is taken only afterwards, on the resulting limiting alignments.*

B.1 PROOF OF PROPOSITION 6

Proof. We use the notation of Theorem 1. The proof proceeds in six steps: we first bound the variance parameter σ^2 independently of the trigger strength α, then decompose the alignment equations under the orthogonality assumption, bound the key quantity η_2 as a function of the poisoned-class margin M_2, and finally combine these bounds to establish the decay rate.

Step 1: Bounding σ^2 independently of α. We first show that the variance parameter σ^2 from the fixed-point system remains bounded as α varies. Since $\hat{\boldsymbol{\theta}}$ minimises the regularised empirical loss $\mathcal{L}_n$, comparing with the zero vector gives

$$\mathcal{L}_n(\hat{\boldsymbol{\theta}}) \le \mathcal{L}_n(\mathbf{0}) = L(0).$$

Expanding the left-hand side and using non-negativity of the individual loss terms, we obtain

$$\frac{\lambda}{2}\|\hat{\boldsymbol{\theta}}\|^2 \le \mathcal{L}_n(\hat{\boldsymbol{\theta}}) \le L(0),$$

so $\|\hat{\boldsymbol{\theta}}\|^2 \le 2L(0)/\lambda$. Applying Theorem 1 to the pseudo-Lipschitz function $\boldsymbol{\theta} \mapsto \|\boldsymbol{\theta}\|^2$ gives $\|\hat{\boldsymbol{\theta}}\|^2 \to \mathbb{E}[\|\tilde{\boldsymbol{\theta}}\|^2]$ in probability, and since $\mathbb{E}[\|\tilde{\boldsymbol{\theta}}\|^2]$ is deterministic, this forces $\mathbb{E}[\|\tilde{\boldsymbol{\theta}}\|^2] \le 2L(0)/\lambda + o_\alpha(1)$ as $n \to \infty$, and therefore

$$\sigma^2 = \mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{C} \tilde{\boldsymbol{\theta}}] \le \mathbb{E}[\|\tilde{\boldsymbol{\theta}}\|^2]\, \|\mathbf{C}\| \le \sigma^2_{\max} + o_\alpha(1), \qquad \sigma^2_{\max} := \frac{2L(0)\|\mathbf{C}\|}{\lambda},$$

where the $o_\alpha(1)$ is uniform in α in the sense that the limiting bound $\sigma^2_{\max}$ does not depend on α. This bound will be used in Step 3 to control the Gaussian expectations defining η_2; since $\sigma^2_{\max}$ is finite, the $o_\alpha(1)$ correction is absorbed into the constants of the bounds that follow without affecting their structural form, and re-emerges explicitly in the final rate.

Step 2: Alignment decomposition under orthogonality. Recall from Theorem 1 that the Gaussian proxy $\tilde{\boldsymbol{\theta}}$ has mean

$$\mathbb{E}[\tilde{\boldsymbol{\theta}}] = \mathbf{R}\big[\eta_1 \boldsymbol{\mu} + \eta_2(\alpha \mathbf{v} - \boldsymbol{\mu})\big] = \mathbf{R}\big[(\eta_1 - \eta_2)\boldsymbol{\mu} + \eta_2 \alpha \mathbf{v}\big],$$

where the two classes have means $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$ and $\boldsymbol{\mu}_2 = \alpha \mathbf{v} - \boldsymbol{\mu}$ with probabilities $1 - \phi$ and ϕ respectively. The fixed-point means Theorem 1 are therefore

$$M_1 = \boldsymbol{\mu}^\top \mathbf{R}\big[(\eta_1 - \eta_2)\boldsymbol{\mu} + \eta_2 \alpha \mathbf{v}\big] = (\eta_1 - \eta_2)\boldsymbol{\mu}^\top \mathbf{R} \boldsymbol{\mu} + \eta_2 \alpha\, \mathbf{v}^\top \mathbf{R} \boldsymbol{\mu},$$

$$M_2 = (\alpha \mathbf{v} - \boldsymbol{\mu})^\top \mathbf{R}\big[(\eta_1 - \eta_2)\boldsymbol{\mu} + \eta_2 \alpha \mathbf{v}\big].$$

Under Assumption 2, the cross term $\mathbf{v}^\top \mathbf{R} \boldsymbol{\mu} = o_\alpha(1)$ is negligible in the limit. Applying this to the first equation gives $M_1 = (\eta_1 - \eta_2)\boldsymbol{\mu}^\top \mathbf{R} \boldsymbol{\mu} + o_\alpha(1)$. For the second, we expand:

$$\begin{aligned} M_2 &= \alpha(\eta_1 - \eta_2) \underbrace{\mathbf{v}^\top \mathbf{R} \boldsymbol{\mu}}_{o_\alpha(1)} - (\eta_1 - \eta_2)\boldsymbol{\mu}^\top \mathbf{R} \boldsymbol{\mu} + \eta_2 \alpha^2 \mathbf{v}^\top \mathbf{R} \mathbf{v} - \eta_2 \alpha \underbrace{\boldsymbol{\mu}^\top \mathbf{R} \mathbf{v}}_{o_\alpha(1)} \\ &= \eta_2 \alpha^2 \mathbf{v}^\top \mathbf{R} \mathbf{v} - (\eta_1 - \eta_2)\boldsymbol{\mu}^\top \mathbf{R} \boldsymbol{\mu} + o_\alpha(1). \end{aligned}$$

We therefore obtain the simplified system

$$M_1 = (\eta_1 - \eta_2)\,\boldsymbol{\mu}^\top \mathbf{R}\boldsymbol{\mu} + o_\alpha(1), \tag{23}$$

$$M_2 = \eta_2 \alpha^2\,\mathbf{v}^\top \mathbf{R}\mathbf{v} - M_1 + o_\alpha(1). \tag{24}$$

The trigger alignment is computed similarly:

$$\mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{v}] = \mathbf{v}^\top \mathbf{R}\big[(\eta_1 - \eta_2)\boldsymbol{\mu} + \eta_2 \alpha \mathbf{v}\big] = (\eta_1 - \eta_2)\underbrace{\mathbf{v}^\top \mathbf{R}\boldsymbol{\mu}}_{o_\alpha(1)} + \eta_2 \alpha\,\mathbf{v}^\top \mathbf{R}\mathbf{v} = \eta_2 \alpha\,\mathbf{v}^\top \mathbf{R}\mathbf{v} + o_\alpha(1).$$

Noting that $M_1 + M_2 = \eta_2 \alpha^2\,\mathbf{v}^\top \mathbf{R}\mathbf{v}$ from (23)–(24), we also have the useful identity

$$\mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{v}] = \frac{M_1 + M_2}{\alpha} + o_\alpha(1). \tag{25}$$

This alignment vanishes at $\alpha = 0$ and is positive for $\alpha > 0$ (since $\eta_2 > 0$ and $\mathbf{v}^\top \mathbf{R}\mathbf{v} > 0$). We stress that η_2 does *not* vanish at $\alpha = 0$: the second mixture component retains weight $\phi > 0$ there, with mean $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}$, so $\eta_2 = \phi\,\mathbb{E}[f(r_2)] > 0$. The alignment vanishes at $\alpha = 0$ because of the explicit factor of α in $\mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{v}] = \eta_2 \alpha\,\mathbf{v}^\top \mathbf{R}\mathbf{v} + o_\alpha(1)$, together with the orthogonality term absorbed into $o_\alpha(1)$.

We also note that $|M_1| = |\boldsymbol{\mu}^\top \mathbb{E}[\tilde{\boldsymbol{\theta}}]| \le \|\boldsymbol{\mu}\| \sqrt{\mathbb{E}[\|\tilde{\boldsymbol{\theta}}\|^2]} \le M_1^{\max} + o_\alpha(1)$, where $M_1^{\max} := \|\boldsymbol{\mu}\|\sqrt{2L(0)/\lambda}$, by Jensen's inequality and Step 1. Since the first term of (24) is non-negative, M_2 is bounded below:

$$M_2 = \eta_2 \alpha^2 \mathbf{v}^\top \mathbf{R}\mathbf{v} - M_1 \ge -|M_1| \ge -M_1^{\max} - o_\alpha(1).$$

Remark 6 (Orthogonality with n-dependent τ)**.** *Assumption 2 is stated for each fixed $\tau \ge 0$, but in Theorem 1 the quantity $\tau = \tau_n$ depends on n. Since τ_n is bounded, any subsequence has a convergent sub-subsequence $\tau_{n_k} \to \tau_\infty$, and $\mathbf{v}^\top(\lambda \mathbf{I} + \tau_\infty \mathbf{C})^{-1}\boldsymbol{\mu} = 0$ by the assumption. A resolvent-identity argument extends this to $\mathbf{v}^\top(\lambda \mathbf{I} + \tau_n \mathbf{C})^{-1}\boldsymbol{\mu} \to 0$ along the full sequence.*

Step 3: Bounding η_2 as a function of M_2.

We view η_2 as a function of M_2. Recall from the fixed-point system that

$$\eta_2 = \phi\,\mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[f(M_2 + \sigma\xi)],$$

where f is defined via the proximal operator as $f(x) = -L'\big(\mathrm{prox}_{\delta L}(x)\big)$. The function f satisfies the functional equation

$$f(x) = -L'(x + \delta f(x)).$$

Since L is convex and decreasing, the proximal shift satisfies $\delta f(x) \ge 0$, and since L' is increasing (by convexity) and strictly negative (by Assumption 3), we obtain the pointwise bound

$$f(x) = -L'(x + \delta f(x)) \le -L'(x).$$

This allows us to replace f with the simpler quantity $-L'$ in upper bounds, at the cost of an inequality. Using this together with the tail bounds on L' from Assumption 3, we establish:

Lemma 12. *There exists $C > 0$ independent of α and n such that*

$$\eta_2 \;\le\; \frac{C + o_\alpha(1)}{|M_2|^{1+\epsilon} + 1},$$

where the $o_\alpha(1)$ tracks the finite-n deviation of σ from its limiting upper bound $\sigma_{\max}$.

Proof. The bound is trivial for $M_2 \le 0$ since M_2 is bounded below (Step 2). Hence assume $M_2 > 0$. Using the bound $f(x) \le -L'(x)$ established above:

$$\begin{aligned}
\eta_2 &= \phi\,\mathbb{E}_{\xi \sim \mathcal{N}(0,1)} f(M_2 + \sigma\xi) \\
&\le \phi\,\mathbb{E}_\xi[-L'(M_2 + \sigma\xi)] \\
&= \phi\,\mathbb{E}_\xi[-L'(M_2 + \sigma\xi)\,\mathbf{1}_{M_2 + \sigma\xi \ge 0}] + \phi\,\mathbb{E}_\xi[-L'(M_2 + \sigma\xi)\,\mathbf{1}_{M_2 + \sigma\xi < 0}] \\
&\le \phi\,\mathbb{E}_\xi\left[\frac{C_2}{(M_2 + \sigma\xi)^{1+\epsilon} + 1}\,\mathbf{1}_{M_2 + \sigma\xi \ge 0}\right] + \phi\,\mathbb{E}_\xi\left[e^{C_1(|M_2 + \sigma\xi| + 1)}\,\mathbf{1}_{M_2 + \sigma\xi < 0}\right],
\end{aligned}$$

where we have applied the positive and negative tail bounds from Assumption 3. Since $\sigma \leq \sigma_{\max} + o_\alpha(1)$ (Step 1), the second term is of order $O(e^{-C_1 M_2})$ uniformly in n for n sufficiently large, as the Gaussian integral over $\{\sigma\xi < -M_2\}$ is bounded.

The first term can be further partitioned, using the bound $\sigma \leq \sigma_{\max} + o_\alpha(1)$ from Step 1:

$$\phi\,\mathbb{E}_\xi\left[\frac{C_2}{(M_2+\sigma\xi)^{1+\epsilon}+1}\,\mathbf{1}_{M_2+\sigma\xi\geq 0}\right] \\ \leq \phi\,\mathbb{E}_\xi\left[\frac{C_2}{(M_2+\sigma\xi)^{1+\epsilon}+1}\,\mathbf{1}_{\xi\geq -\frac{M_2}{2\sigma_{\max}}}\right] + \phi\,\mathbb{E}_\xi\left[\frac{C_2}{(M_2+\sigma\xi)^{1+\epsilon}+1}\,\mathbf{1}_{\xi< -\frac{M_2}{2\sigma_{\max}}}\right].$$

For the second term: the integrand is bounded (since the denominator is at least 1), while the Gaussian measure of $\{\xi < -M_2/(2\sigma_{\max})\}$ is $O(e^{-C' M_2^2})$ by a standard tail bound, so this contribution is negligible.

For the first term: when $\xi \geq -M_2/(2\sigma_{\max})$ we have $M_2 + \sigma\xi \geq M_2 - \sigma \cdot M_2/(2\sigma_{\max}) \geq M_2/2 - o_\alpha(1)\, M_2/(2\sigma_{\max})$ (since $\sigma \leq \sigma_{\max} + o_\alpha(1)$), so for n sufficiently large $M_2 + \sigma\xi \geq M_2/4$, and therefore

$$\frac{C_2}{(M_2+\sigma\xi)^{1+\epsilon}+1} \leq \frac{C_2}{(M_2/4)^{1+\epsilon}+1} \leq \frac{C}{|M_2|^{1+\epsilon}+1}.$$

The constant C depends continuously on $\sigma_{\max} + o_\alpha(1)$, giving the lemma's bound $\eta_2 \leq (C + o_\alpha(1))/(|M_2|^{1+\epsilon}+1)$. □

Step 4: Polynomial-tail decay rate. We now combine the bound on η_2 from Lemma 12 with the alignment equations to determine the growth rate of M_2 as $\alpha \to \infty$.

From (24), we have $M_2 = \eta_2 \alpha^2 \mathbf{v}^\top \mathbf{R}\mathbf{v} - M_1$, and since M_1 is bounded we obtain

$$M_2 \leq \eta_2 \alpha^2 \mathbf{v}^\top \mathbf{R}\mathbf{v} + |M_1|.$$

Since $\|\mathbf{v}\| = 1$ and $\mathbf{R} = (\lambda\mathbf{I} + \tau\mathbf{C})^{-1} \preceq \lambda^{-1}\mathbf{I}$, we have the resolvent bound

$$\mathbf{v}^\top \mathbf{R}\mathbf{v} \leq \frac{\|\mathbf{v}\|^2}{\lambda} = \frac{1}{\lambda}.$$

Substituting the bound from Lemma 12, $\eta_2 \leq (C + o_\alpha(1))/(|M_2|^{1+\epsilon}+1)$:

$$M_2 \leq \frac{(C + o_\alpha(1))\,\alpha^2}{\lambda(|M_2|^{1+\epsilon}+1)} + |M_1|.$$

Since $|M_1| \leq M_1^{\max} + o_\alpha(1)$ is bounded independently of α (Step 2), for α sufficiently large the first term dominates, and we may absorb $|M_1|$ into the constant. We therefore have the self-consistent inequality

$$M_2 \leq \frac{(C' + o_\alpha(1))\,\alpha^2}{|M_2|^{1+\epsilon}+1}$$

for some constant C' not depending on α or n. And thus for $M_2 > 0$ rearranging gives $M_2^{2+\epsilon} \leq (C' + o_\alpha(1))\alpha^2$, i.e.

$$|M_2| \leq C''\,\alpha^{2/(2+\epsilon)}\,(1 + o_\alpha(1)).$$

Where we note now that this inequality holds for $M_2 < 0$ and α large too trivially, since M_2 is bounded below.

Now, recalling the identity (25), $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}] = (M_1 + M_2)/\alpha$, and using $|M_1| \leq M_1^{\max} + o_\alpha(1)$ from Step 2 together with the M_2 bound above:

$$\mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{v}] = \frac{M_1 + M_2}{\alpha} \leq \frac{M_1^{\max}}{\alpha} + \frac{C''\,\alpha^{2/(2+\epsilon)}\,(1 + o_\alpha(1))}{\alpha} + \frac{o_\alpha(1)}{\alpha} = g(\alpha) + o_\alpha(1),$$

where, using $\frac{2}{2+\epsilon} - 1 = \frac{-\epsilon}{2+\epsilon}$ and absorbing $M_1^{\max}/\alpha \leq M_1^{\max}\alpha^{-\epsilon/(2+\epsilon)}$ (valid for $\alpha \geq 1$) into the leading term, while the multiplicative $o_\alpha(1)\,\alpha^{-\epsilon/(2+\epsilon)}$ is at most $o_\alpha(1)$ for $\alpha \geq 1$,

$$g(\alpha) \;\leq\; C'\,\alpha^{-\epsilon/(2+\epsilon)}.$$

Since $\epsilon > 0$, $g(\alpha) \to 0$ as $\alpha \to \infty$.

Step 5: Concentration and existence of maximum. By Theorem 1 applied to the 1-Lipschitz function $\boldsymbol{\theta} \mapsto \mathbf{v}^\top \boldsymbol{\theta}$, we have $\mathbf{v}^\top \hat{\boldsymbol{\theta}} \to_p \mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}]$ in probability, so the decay bound from Step 4 on $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}]$ transfers to $\mathbf{v}^\top \hat{\boldsymbol{\theta}}$ in probability.

Since $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}] = o_\alpha(1)$ at $\alpha = 0$, by the explicit factor of α rather than by any vanishing of η_2 (see Step 2), and $\mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\theta}}] = \eta_2 \alpha\, \mathbf{v}^\top \mathbf{R}\mathbf{v} + o_\alpha(1)$ is positive for $\alpha > 0$ (since $\eta_2 = \phi\, \mathbb{E}[f(r_2)] > 0$, as $f = -L'(\mathrm{prox}_{\delta L}(\cdot)) > 0$ by the strict monotonicity of L in Assumption 3, and $\mathbf{v}^\top \mathbf{R}\mathbf{v} > 0$), and vanishes as $\alpha \to \infty$ (by the decay bound above), continuity guarantees that the trigger alignment $\mathbf{v}^\top \tilde{\boldsymbol{\theta}}$ is maximised at some finite $\alpha^\star \in (0, \infty)$.

Step 6: Exponential-tail case. Suppose now that the loss derivative satisfies the stronger bound $|L'(x)| \le C_3 e^{-C_4 x}$ for $x > 0$ (this holds, for instance, for the logistic loss $L(x) = \log(1 + e^{-x})$, where $L'(x) = -e^{-x}/(1 + e^{-x})$).

We show that $\eta_2 \le C e^{-C' M_2}$ for M_2 large, by a partition argument analogous to that in Lemma 12. Starting from $\eta_2 \le \phi\, \mathbb{E}_\xi[-L'(M_2 + \sigma\xi)]$, we split the expectation into the same regions:

$$\eta_2 \le \phi\, \mathbb{E}_\xi[-L'(M_2 + \sigma\xi)\, \mathbf{1}_{M_2 + \sigma\xi \ge 0}] + \phi\, \mathbb{E}_\xi[-L'(M_2 + \sigma\xi)\, \mathbf{1}_{M_2 + \sigma\xi < 0}]\,.$$

Negative region $\{M_2 + \sigma\xi < 0\}$: as before, the exponential bound $|L'(x)| \le e^{C_1(|x|+1)}$ for $x < 0$ combined with $\sigma \le \sigma_{\max} + o_\alpha(1)$ from Step 1 gives a contribution of $O(e^{-C_1 M_2})$, uniformly in n for n large.

Positive region $\{M_2 + \sigma\xi \ge 0\}$: we partition further at $\xi = -M_2/(2\sigma_{\max})$, using $\sigma \le \sigma_{\max} + o_\alpha(1)$.

$$\begin{aligned}
&\phi\, \mathbb{E}_\xi[-L'(M_2 + \sigma\xi)\, \mathbf{1}_{M_2 + \sigma\xi \ge 0}] \\
&\quad \le \phi\, \mathbb{E}_\xi\left[C_3 e^{-C_4(M_2 + \sigma\xi)}\, \mathbf{1}_{-\frac{M_2}{\sigma_{\max}} < \xi < -\frac{M_2}{2\sigma_{\max}}}\right] + \phi\, \mathbb{E}_\xi\left[C_3 e^{-C_4(M_2 + \sigma\xi)}\, \mathbf{1}_{\xi \ge -\frac{M_2}{2\sigma_{\max}}}\right].
\end{aligned}$$

For the first term: the integrand is bounded, while the Gaussian measure of $\{\xi < -M_2/(2\sigma_{\max})\}$ is $O(e^{-C' M_2^2})$, which is negligible compared to $e^{-C' M_2}$ for large enough M_2.

For the second term: when $\xi \ge -M_2/(2\sigma_{\max})$ we have $M_2 + \sigma\xi \ge M_2/2 - o_\alpha(1)\, M_2/(2\sigma_{\max})$ (since $\sigma \le \sigma_{\max} + o_\alpha(1)$), hence $M_2 + \sigma\xi \ge M_2/4$ for n sufficiently large. The exponential tail bound then gives $|L'(M_2 + \sigma\xi)| \le C_3 e^{-C_4 M_2/4}$, and since the Gaussian integral over this region is at most 1, the contribution is $O(e^{-C_4 M_2/4})$.

Combining all regions yields

$$\eta_2 \le (C + o_\alpha(1))\, e^{-C' M_2}$$

for constants $C, C' > 0$ not depending on α or n, with the $o_\alpha(1)$ tracking the dependence of C on $\sigma_{\max} + o_\alpha(1)$ rather than $\sigma_{\max}$ exactly.

Substituting into the alignment equation as in Step 4:

$$M_2 \le (C + o_\alpha(1))\, e^{-C' M_2} \frac{\alpha^2}{\lambda} + |M_1|.$$

For α large, absorbing the bounded $|M_1|$ term:

$$M_2 \le (C'' + o_\alpha(1))\, e^{-C' M_2} \alpha^2,$$

and hence

$$M_2\, e^{C' M_2} \le (C'' + o_\alpha(1))\, \alpha^2.$$

Now for $M_2 > 0$ then taking logarithms of both sides:

$$\log M_2 + C' M_2 \le 2 \log \alpha + \log(C'' + o_\alpha(1)).$$

For M_2 sufficiently large we have $\log M_2 \ge 0$, and therefore

$$C' M_2 \le 2 \log \alpha + O(1) + o_\alpha(1),$$

giving $M_2 \le C''' \log\alpha + o_\alpha(1)$ as $\alpha \to \infty$. Where again now the statement is trivially true for negative M_2 and large enough alpha, since M_2 is bounded below. Hence,

$$|M_2| \le C''' \log\alpha + o_\alpha(1)$$

Finally, using the identity (25) and $|M_1| \le M_1^{\max} + o_\alpha(1)$ from Step 2 together with the M_2 bound above:

$$\mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{v}] = \frac{M_1 + M_2}{\alpha} \le \frac{M_1^{\max} + C''' \log\alpha}{\alpha} + \frac{o_\alpha(1)}{\alpha} = g(\alpha) + o_\alpha(1), \qquad g(\alpha) \le C' \frac{\log\alpha}{\alpha},$$

where $M_1^{\max}/\alpha$ is absorbed into the leading $\log\alpha/\alpha$ term for $\alpha \ge e$, and $o_\alpha(1)/\alpha \le o_\alpha(1)$ for $\alpha \ge 1$. The concentration and existence-of-maximum arguments from Step 5 apply identically. □

B.2 Cone Lemma

The following structural lemma underpins much of the population-risk analysis. It shows that the unpoisoned ($\alpha = 0$) population minimiser, and as a special case the benign minimiser, lies in a two-dimensional submanifold determined by the resolvent of $\mathbf{C}$.[2]

Lemma 13 (Cone Lemma)**.** *Suppose Assumption 3 holds with $\phi < \frac{1}{2}$ and $\lambda > 0$. Then:*

(i) The minimiser $\boldsymbol{\theta}(0)$ of (3) at $\alpha = 0$ satisfies

$$\boldsymbol{\theta}(0) \;=\; a\,(\lambda\mathbf{I} + \tau\mathbf{C})^{-1}\boldsymbol{\mu}$$

for some $a > 0$ and $\tau \ge 0$.

(ii) The benign minimiser $\boldsymbol{\theta}_{\mathrm{ben}}$ of (7) satisfies

$$\boldsymbol{\theta}_{\mathrm{ben}} \;=\; a\,\big(\tfrac{\lambda}{1-\phi}\mathbf{I} + \tau\mathbf{C}\big)^{-1}\boldsymbol{\mu}$$

for some $a > 0$ and $\tau \ge 0$.

Part (ii) is the special case of part (i) with $\phi = 0$ and λ replaced by $\lambda/(1-\phi)$, since $\boldsymbol{\theta}_{\mathrm{ben}}$ is the minimiser of $\mathcal{L}_{\mathrm{ben}}/(1-\phi) = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{C})}[L(\boldsymbol{\theta}^\top\mathbf{x})] + \frac{\lambda/(1-\phi)}{2}\|\boldsymbol{\theta}\|^2$, which coincides with $\mathcal{L}_{\mathrm{pop}}(\,\cdot\,;0)$ at $\phi = 0$ and regulariser $\lambda/(1-\phi)$.

Proof. We prove (i); part (ii) follows from (i) by the rescaling noted in the statement (and with $\phi = 0$ the condition $a > 0$ is immediate since $L' < 0$ everywhere by Assumption 3).

Setting $\nabla\mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta}(0);0) = \mathbf{0}$ and applying Stein's lemma to the Gaussian expectations yields $(\lambda\mathbf{I} + \tau\mathbf{C})\boldsymbol{\theta}(0) = a\boldsymbol{\mu}$, where, writing $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$,

$$A = \boldsymbol{\theta}(0)^\top\boldsymbol{\mu} + \boldsymbol{\theta}(0)^\top\mathbf{C}^{1/2}\mathbf{z}, \qquad B = -\boldsymbol{\theta}(0)^\top\boldsymbol{\mu} + \boldsymbol{\theta}(0)^\top\mathbf{C}^{1/2}\mathbf{z},$$

and

$$\tau = (1-\phi)\,\mathbb{E}[L''(A)] + \phi\,\mathbb{E}[L''(B)] \ge 0, \qquad a = -(1-\phi)\,\mathbb{E}[L'(A)] + \phi\,\mathbb{E}[L'(B)].$$

Here A represents the margin of a class $+1$ sample and B that of a class -1 sample. Since $\lambda > 0$ and $\tau \ge 0$, the matrix $\lambda\mathbf{I} + \tau\mathbf{C} \succ 0$, so $\boldsymbol{\theta}(0) = a(\lambda\mathbf{I} + \tau\mathbf{C})^{-1}\boldsymbol{\mu}$ and $\mathrm{sign}(\boldsymbol{\mu}^\top\boldsymbol{\theta}(0)) = \mathrm{sign}(a)$.

It remains to show $a > 0$. Suppose for contradiction that $\boldsymbol{\mu}^\top\boldsymbol{\theta}(0) \le 0$. Then A has nonpositive mean $\mathbb{E}[A] = \boldsymbol{\mu}^\top\boldsymbol{\theta}(0) \le 0$ and B has nonnegative mean $\mathbb{E}[B] = -\boldsymbol{\mu}^\top\boldsymbol{\theta}(0) \ge 0$. Since L' is increasing (by convexity of L) and strictly negative (by Assumption 3: L is strictly decreasing), a stochastic-ordering argument gives

$$\mathbb{E}[L'(A)] \le \mathbb{E}[L'(B)] < 0.$$

Therefore

$$\begin{aligned} a &= -(1-\phi)\mathbb{E}[L'(A)] + \phi\mathbb{E}[L'(B)] \\ &\ge -(1-\phi)\mathbb{E}[L'(B)] + \phi\mathbb{E}[L'(B)] \\ &= -(1-2\phi)\mathbb{E}[L'(B)] > 0, \end{aligned}$$

using $\phi < \frac{1}{2}$ and $\mathbb{E}[L'(B)] < 0$. This implies $\boldsymbol{\mu}^\top\boldsymbol{\theta}(0) > 0$, contradicting our assumption, so $a > 0$. □

[2] This can be compared with the form of $\hat{\boldsymbol{\theta}}$ for the ERM with $\alpha = 0$, for which its expectation will lie on a similar submanifold.

Remark 7. *The representation reflects that the loss depends on $\boldsymbol{\theta}$ only through $\boldsymbol{\theta}^\top \boldsymbol{\mu}$ and $\boldsymbol{\theta}^\top \mathbf{C}\boldsymbol{\theta}$, while the regulariser selects the unique minimiser on this manifold. For part (i), the condition $\phi < \frac{1}{2}$ enforces positive alignment $\boldsymbol{\mu}^\top \boldsymbol{\theta}(0) > 0$: the majority (positive) class dominates the gradient. For part (ii) the analogous statement is trivial since there is only one class.*

B.3 Proof of Proposition 7

Orthogonality assumption used in this section. Throughout this subsection we work with the strong form of orthogonality, $\mathbf{v}^\top \mathbf{C}^k \boldsymbol{\mu} = 0$ for every $k \geq 0$, rather than the asymptotic $o_\alpha(1)$ form of Assumption 2. The reason is that the population minimiser $\boldsymbol{\theta}(\alpha)$ is defined for any fixed dimension p (including small p), and there is no $n \to \infty$ limit available along which to absorb $o_\alpha(1)$ terms. The same conclusions hold under the asymptotic version of Assumption 2 if one further takes $p \to \infty$ (after $n \to \infty$): then $\mathbf{v}^\top \mathbf{C}^k \boldsymbol{\mu} = o_\alpha(1)$ along the sequence, and the resolvent identity $\mathbf{v}^\top (\lambda \mathbf{I} + \tau \mathbf{C})^{-1} \boldsymbol{\mu} = o_\alpha(1)$ follows with 0 replaced by $o_\alpha(1)$ at each occurrence.

Proof. The proof has two parts. Part (i) shows that the population minimiser converges to the benign minimiser by constructing a competitor whose poisoned loss vanishes as $\alpha \to \infty$ and then using strong convexity. Part (ii) shows that when $\boldsymbol{\mu}$ is an eigenvector of $\mathbf{C}$, both minimisers are collinear with $\boldsymbol{\mu}$, and a derivative comparison at $\boldsymbol{\theta}_{\mathrm{ben}}$ establishes the strict inequality.

Existence and uniqueness of $\boldsymbol{\theta}(\alpha)$ and $\boldsymbol{\theta}_{\mathrm{ben}}$ follow from λ-strong convexity.

Part (i): Convergence as $\alpha \to \infty$.

We begin with an auxiliary result whose proof is given at the end of Part (i).

Lemma 14 (Gaussian expectation vanishing)**.** *Let $(X_n)_{n \geq 1}$ be Gaussian random variables with $\mathbb{E}[X_n] \to \infty$ and $\bar{\sigma}^2 := \sup_{n \geq 1} \mathrm{Var}(X_n) < \infty$. Under Assumption 3, $\lim_{n \to \infty} \mathbb{E}[L(X_n)] = 0$.*

Remark 8. *Integrating the derivative bound in Assumption 3 gives $L(x) \leq A\, e^{C_1 |x|}$ for $x < 0$ and $L(x) \leq L(0)$ for $x \geq 0$. In particular, $\mathbb{E}[L(a + b\,\xi)] < \infty$ for any finite $a, b \in \mathbb{R}$ and $\xi \sim \mathcal{N}(0,1)$, since the moment-generating function of $|\xi|$ is finite everywhere.*

We next record a continuity result for the benign expectation.

Lemma 15 (Continuity of benign expectation)**.** *If $\boldsymbol{\theta}_\alpha \to \boldsymbol{\theta}$ in ℓ_2, then $\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})}[L(\mathbf{x}^\top \boldsymbol{\theta}_\alpha)] \to \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})}[L(\mathbf{x}^\top \boldsymbol{\theta})]$.*

Proof. Write $X_\alpha := \mathbf{x}^\top \boldsymbol{\theta}_\alpha$ and $X := \mathbf{x}^\top \boldsymbol{\theta}$. Since $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, both are Gaussian: $X_\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$ with $\mu_\alpha := \boldsymbol{\mu}^\top \boldsymbol{\theta}_\alpha$ and $\sigma_\alpha^2 := \boldsymbol{\theta}_\alpha^\top \mathbf{C} \boldsymbol{\theta}_\alpha$, and $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu := \boldsymbol{\mu}^\top \boldsymbol{\theta}$ and $\sigma^2 := \boldsymbol{\theta}^\top \mathbf{C} \boldsymbol{\theta}$. Since $\boldsymbol{\theta}_\alpha \to \boldsymbol{\theta}$ in ℓ_2, we have $\mu_\alpha \to \mu$ and $\sigma_\alpha \to \sigma$.

Almost sure convergence. Since $\boldsymbol{\theta}_\alpha \to \boldsymbol{\theta}$ deterministically, for each realisation of $\mathbf{x}$ with $\|\mathbf{x}\| < \infty$ (which holds a.s. for Gaussian $\mathbf{x}$), Cauchy–Schwarz gives

$$|X_\alpha - X| = |\mathbf{x}^\top (\boldsymbol{\theta}_\alpha - \boldsymbol{\theta})| \leq \|\mathbf{x}\|\, \|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}\| \to 0.$$

Continuity of L then yields $L(X_\alpha) \to L(X)$ almost surely.

Upper bound via coupling. Fix $\epsilon > 0$ and set $\bar{\sigma}_\epsilon^2 := \sigma^2 + \epsilon$. Since $\sigma_\alpha \to \sigma$, for all sufficiently large α we have $\sigma_\alpha^2 \leq \bar{\sigma}_\epsilon^2$. By Gaussianity we can couple an independent $\xi \sim \mathcal{N}(0,1)$ with $\tilde{\sigma}_\alpha := (\bar{\sigma}_\epsilon^2 - \sigma_\alpha^2)^{1/2} \geq 0$ so that

$$Y_\alpha := X_\alpha + \tilde{\sigma}_\alpha\, \xi \sim \mathcal{N}\big(\mu_\alpha,\, \bar{\sigma}_\epsilon^2\big).$$

By the conditional Jensen inequality (using convexity of L),

$$\mathbb{E}[L(Y_\alpha)] = \mathbb{E}\big[\mathbb{E}[L(X_\alpha + \tilde{\sigma}_\alpha \xi) \mid X_\alpha]\big] \geq \mathbb{E}\big[L\big(\mathbb{E}[X_\alpha + \tilde{\sigma}_\alpha \xi \mid X_\alpha]\big)\big] = \mathbb{E}[L(X_\alpha)],$$

whence $\mathbb{E}[L(X_\alpha)] \leq \mathbb{E}[L(Y_\alpha)] = \mathbb{E}[L(\mu_\alpha + \bar{\sigma}_\epsilon\, \xi)]$.

We apply dominated convergence to the right-hand side. Pointwise, $L(\mu_\alpha + \bar{\sigma}_\epsilon\,\xi) \to L(\mu + \bar{\sigma}_\epsilon\,\xi)$ by continuity of L. Since L is decreasing, for all α large enough that $\mu_\alpha \geq \mu - 1$ we have the dominator

$$L(\mu_\alpha + \bar{\sigma}_\epsilon\,\xi) \leq L\big((\mu-1) + \bar{\sigma}_\epsilon\,\xi\big).$$

The right-hand side is integrable by Remark 8. Dominated convergence therefore yields

$$\limsup_{\alpha\to\infty} \mathbb{E}[L(X_\alpha)] \leq \lim_{\alpha\to\infty} \mathbb{E}[L(\mu_\alpha + \bar{\sigma}_\epsilon\,\xi)] = \mathbb{E}[L(\mu + \bar{\sigma}_\epsilon\,\xi)].$$

Conclusion. Since $L \geq 0$ and $L(X_\alpha) \to L(X)$ a.s., Fatou's lemma gives $\mathbb{E}[L(X)] \leq \liminf_\alpha \mathbb{E}[L(X_\alpha)]$. The coupling bound above holds for every $\epsilon > 0$. As $\epsilon \downarrow 0$, $\bar{\sigma}_\epsilon \downarrow \sigma$ and $L(\mu + \bar{\sigma}_\epsilon\,\xi) \to L(\mu + \sigma\,\xi)$ pointwise; a further application of dominated convergence (with integrable dominator $L(\mu-(\sigma+1)|\xi|)$, cf. Remark 8) gives $\mathbb{E}[L(\mu+\bar{\sigma}_\epsilon\,\xi)] \to \mathbb{E}[L(\mu+\sigma\,\xi)] = \mathbb{E}[L(X)]$. Hence $\limsup_\alpha \mathbb{E}[L(X_\alpha)] \leq \mathbb{E}[L(X)]$, and combining with Fatou's lower bound completes the proof. □

We now construct a sequence of competitors whose population loss converges to the benign-only optimum.

Lemma 16 (Competitor upper bound)**.** *Define* $\hat{\boldsymbol{\theta}}_\alpha := \boldsymbol{\theta}_{\mathrm{ben}} + \alpha^{-1/2}\mathbf{v}$. *Then*

$$\limsup_{\alpha\to\infty} \mathcal{L}_{\mathrm{pop}}(\hat{\boldsymbol{\theta}}_\alpha;\alpha) \leq \mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}_{\mathrm{ben}}).$$

Proof. By Lemma 13(ii), $\boldsymbol{\theta}_{\mathrm{ben}} = a(\frac{\lambda}{1-\phi}\mathbf{I} + \tau\mathbf{C})^{-1}\boldsymbol{\mu}$, so Krylov orthogonality gives $\mathbf{v}^\top\boldsymbol{\theta}_{\mathrm{ben}} = 0$. We analyse each term of the population loss separately.

Regulariser. Since $\mathbf{v}^\top\boldsymbol{\theta}_{\mathrm{ben}} = 0$,

$$\tfrac{\lambda}{2}\|\hat{\boldsymbol{\theta}}_\alpha\|^2 - \tfrac{\lambda}{2}\|\boldsymbol{\theta}_{\mathrm{ben}}\|^2 = \tfrac{\lambda}{2}\alpha^{-1}\|\mathbf{v}\|^2 \to 0.$$

Benign loss. The benign loss converges by Lemma 15 since $\hat{\boldsymbol{\theta}}_\alpha \to \boldsymbol{\theta}_{\mathrm{ben}}$ in ℓ_2.

Poisoned loss. For the poisoned term, let $\tilde{\mathbf{x}} \sim \mathcal{N}(\alpha\mathbf{v} - \boldsymbol{\mu}, \mathbf{C})$. The margin $\hat{\boldsymbol{\theta}}_\alpha^\top\tilde{\mathbf{x}}$ is Gaussian with mean

$$(\alpha\mathbf{v} - \boldsymbol{\mu})^\top\hat{\boldsymbol{\theta}}_\alpha = \alpha^{1/2}\|\mathbf{v}\|^2 - \boldsymbol{\mu}^\top\boldsymbol{\theta}_{\mathrm{ben}} - \alpha^{-1/2}\boldsymbol{\mu}^\top\mathbf{v} \to +\infty$$

(using $\mathbf{v}^\top\boldsymbol{\theta}_{\mathrm{ben}} = 0$) and variance $\hat{\boldsymbol{\theta}}_\alpha^\top\mathbf{C}\hat{\boldsymbol{\theta}}_\alpha \leq \|\mathbf{C}\|\,\|\hat{\boldsymbol{\theta}}_\alpha\|^2$, which is bounded since $\hat{\boldsymbol{\theta}}_\alpha \to \boldsymbol{\theta}_{\mathrm{ben}}$. Lemma 14 therefore gives $\mathbb{E}[L(\hat{\boldsymbol{\theta}}_\alpha^\top\tilde{\mathbf{x}})] \to 0$. □

Since $L \geq 0$, the poisoned loss term is non-negative, so $\mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}) \leq \mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta};\alpha)$ pointwise for every $\boldsymbol{\theta}$. Taking the infimum:

$$\mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}_{\mathrm{ben}}) \leq \inf_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta};\alpha).$$

Lemma 16 gives the matching $\limsup$, hence $\inf_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta};\alpha) \to \mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}_{\mathrm{ben}})$.

Lemma 17 (Convergence of minimisers)**.** $\boldsymbol{\theta}(\alpha) \to \boldsymbol{\theta}_{\mathrm{ben}}$ *as* $\alpha \to \infty$ *in* ℓ_2.

Proof. Using $\mathcal{L}_{\mathrm{ben}} \leq \mathcal{L}_{\mathrm{pop}}(\cdot;\alpha)$ pointwise:

$$0 \leq \mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}(\alpha)) - \mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}_{\mathrm{ben}}) \leq \inf_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta};\alpha) - \mathcal{L}_{\mathrm{ben}}(\boldsymbol{\theta}_{\mathrm{ben}}) \to 0.$$

By λ-strong convexity of $\mathcal{L}_{\mathrm{ben}}$, the left side is at least $\frac{\lambda}{2}\|\boldsymbol{\theta}(\alpha) - \boldsymbol{\theta}_{\mathrm{ben}}\|^2$, and therefore $\|\boldsymbol{\theta}(\alpha) - \boldsymbol{\theta}_{\mathrm{ben}}\| \to 0$. □

Part (i) follows from Lemma 17.

We now give the deferred proof.

Proof of Lemma 14. Write $\mu_n = \mathbb{E}[X_n]$ and $\sigma_n^2 = \mathrm{Var}(X_n)$. Let $Y_n \sim \mathcal{N}(\mu_n, \bar{\sigma}^2)$. Since $\sigma_n^2 \le \bar{\sigma}^2$, we can couple an independent $\xi \sim \mathcal{N}(0,1)$ with $\tilde{\sigma}_n := (\bar{\sigma}^2 - \sigma_n^2)^{1/2} \ge 0$ so that $Y_n = X_n + \tilde{\sigma}_n \xi$. By the conditional Jensen inequality (using convexity of L),

$$\mathbb{E}[L(Y_n)] = \mathbb{E}\big[\mathbb{E}[L(X_n + \tilde{\sigma}_n\xi) \mid X_n]\big] \ge \mathbb{E}\big[L\big(\mathbb{E}[X_n + \tilde{\sigma}_n\xi \mid X_n]\big)\big] = \mathbb{E}[L(X_n)],$$

so $0 \le \mathbb{E}[L(X_n)] \le \mathbb{E}[L(Y_n)] = \mathbb{E}[L(\mu_n + \bar{\sigma}\,\xi)]$.

Since L is decreasing and $\mu_n > 0$ for n large, $L(\mu_n + \bar{\sigma}\,\xi) \le L(\bar{\sigma}\,\xi)$. This dominator is integrable by Remark 8. Moreover $L(\mu_n + \bar{\sigma}\,\xi) \to 0$ pointwise, since $\mu_n + \bar{\sigma}\,\xi \to +\infty$ and $L(x) \to 0$ as $x \to +\infty$ (which follows from L being non-negative, convex, and strictly decreasing). By dominated convergence, $\mathbb{E}[L(\mu_n + \bar{\sigma}\,\xi)] \to 0$, and the squeeze $0 \le \mathbb{E}[L(X_n)] \le \mathbb{E}[L(\mu_n + \bar{\sigma}\,\xi)]$ completes the proof. □

Part (ii): Eigenvector case. Suppose $\boldsymbol{\mu}$ is an eigenvector of $\mathbf{C}$ with eigenvalue λ_μ. We first show that both minimisers are collinear with $\boldsymbol{\mu}$, and then compare their scalar coefficients.

Writing $\boldsymbol{\theta} = a\boldsymbol{\mu} + \mathbf{w}$ with $\mathbf{w} \perp \boldsymbol{\mu}$, the variance decouples: $\boldsymbol{\theta}^\top \mathbf{C}\boldsymbol{\theta} = a^2\lambda_\mu\|\boldsymbol{\mu}\|^2 + \mathbf{w}^\top\mathbf{C}\mathbf{w}$. Setting $\nabla_{\mathbf{w}}\mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta}; 0) = \mathbf{0}$ and using the fact that the gradient with respect to $\mathbf{w}$ at any critical point satisfies

$$\nabla_{\mathbf{w}}\mathcal{L}_{\mathrm{pop}}(\boldsymbol{\theta}; 0) = \big[(1-\phi)\mathbb{E}[L''(A)] + \phi\mathbb{E}[L''(B)]\big]\mathbf{C}\mathbf{w} + \lambda\mathbf{w} = \mathbf{0},$$

forces $\mathbf{w} = \mathbf{0}$ since $\lambda > 0$ (the coefficient of $\mathbf{C}\mathbf{w}$ is non-negative by convexity of L, so $\lambda\mathbf{w} = \mathbf{0}$ is the only solution). The same argument applies to $\mathcal{L}_{\mathrm{ben}}$. Both minimisers are therefore collinear with $\boldsymbol{\mu}$: $\boldsymbol{\theta}^* = a^*\boldsymbol{\mu}$, and it suffices to compare the scalar minimisers

$$a_0^* = \arg\min_a \mathcal{L}_{\mathrm{pop}}(a\boldsymbol{\mu}; 0), \qquad a_{\mathrm{ben}} = \arg\min_a \mathcal{L}_{\mathrm{ben}}(a\boldsymbol{\mu}).$$

Since $\mathcal{L}_{\mathrm{pop}}(\,\cdot\,; 0)$ is strictly convex in a with unique minimiser a_0^*, it suffices to show

$$\frac{d}{da}\mathcal{L}_{\mathrm{pop}}(a_{\mathrm{ben}}\boldsymbol{\mu}; 0) > 0,$$

which forces $a_0^* < a_{\mathrm{ben}}$ by strict convexity.

By optimality of a_{ben}, $\frac{d}{da}\mathcal{L}_{\mathrm{ben}}(a_{\mathrm{ben}}\boldsymbol{\mu}) = 0$, so the derivative of the population loss at a_{ben} reduces to the contribution from the poisoned class alone:

$$\frac{d}{da}\mathcal{L}_{\mathrm{pop}}(a\boldsymbol{\mu}; 0)\Big|_{a=a_{\mathrm{ben}}} = \phi\,\frac{d}{da}\,\mathbb{E}\Big[L\Big(-a\|\boldsymbol{\mu}\|^2 + a\sqrt{\lambda_\mu}\|\boldsymbol{\mu}\|\,\xi\Big)\Big]\,\Big|_{a=a_{\mathrm{ben}}},$$

where $\xi \sim \mathcal{N}(0,1)$. Differentiating inside the expectation:

$$\frac{d}{da}\mathcal{L}_{\mathrm{pop}}(a\boldsymbol{\mu}; 0)\Big|_{a=a_{\mathrm{ben}}} = \phi\,\mathbb{E}\Big[L'(B_{\mathrm{ben}})\,\Big(-\|\boldsymbol{\mu}\|^2 + \sqrt{\lambda_\mu}\|\boldsymbol{\mu}\|\,\xi\Big)\Big],$$

where $B_{\mathrm{ben}} = a_{\mathrm{ben}}(-\|\boldsymbol{\mu}\|^2 + \sqrt{\lambda_\mu}\|\boldsymbol{\mu}\|\xi)$ is the margin of a poisoned sample evaluated at $\boldsymbol{\theta}_{\mathrm{ben}}$.

We now separate the deterministic and Gaussian contributions. Applying Stein's lemma to the ξ term, using $\partial B_{\mathrm{ben}}/\partial\xi = a_{\mathrm{ben}}\sqrt{\lambda_\mu}\|\boldsymbol{\mu}\|$:

$$\mathbb{E}[L'(B_{\mathrm{ben}})\xi] = a_{\mathrm{ben}}\sqrt{\lambda_\mu}\|\boldsymbol{\mu}\|\,\mathbb{E}[L''(B_{\mathrm{ben}})].$$

Substituting:

$$\frac{d}{da}\mathcal{L}_{\mathrm{pop}}(a\boldsymbol{\mu}; 0)\Big|_{a=a_{\mathrm{ben}}} = \phi\|\boldsymbol{\mu}\|^2\left(\underbrace{-\mathbb{E}[L'(B_{\mathrm{ben}})]}_{>0} + \underbrace{a_{\mathrm{ben}}\lambda_\mu\,\mathbb{E}[L''(B_{\mathrm{ben}})]}_{\ge 0}\right) > 0,$$

since $L'(x) < 0$ everywhere (Assumption 3), $a_{\mathrm{ben}} > 0$ (by the Cone Lemma), and $\lambda_\mu, \mathbb{E}[L''(B_{\mathrm{ben}})] \ge 0$ (by non-negativity of the eigenvalue and convexity of L). Therefore $a_0^* < a_{\mathrm{ben}}$, i.e. $\boldsymbol{\mu}^\top\boldsymbol{\theta}(0) < \boldsymbol{\mu}^\top\boldsymbol{\theta}_{\mathrm{ben}}$. □

B.4 Proof of Proposition 8

Proof. The idea is to decompose the population gradient at $\boldsymbol{\theta}_{\text{ben}}$ into the benign part (which vanishes by optimality of $\boldsymbol{\theta}_{\text{ben}}$) and the poisoned part, then show the latter has a strictly positive projection onto $\boldsymbol{\mu}$ using Stein's lemma and the Cone Lemma.

Write $\mathcal{L}_{\text{pop}}(\boldsymbol{\theta}; \alpha) = \mathcal{L}_{\text{ben}}(\boldsymbol{\theta}) + \phi\, h_\alpha(\boldsymbol{\theta})$, where $h_\alpha(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\alpha\mathbf{v}-\boldsymbol{\mu},\mathbf{C})}[L(\mathbf{x}^\top\boldsymbol{\theta})]$. Since $\nabla\mathcal{L}_{\text{ben}}(\boldsymbol{\theta}_{\text{ben}}) = \mathbf{0}$ by optimality,

$$\boldsymbol{\mu}^\top\nabla\mathcal{L}_{\text{pop}}(\boldsymbol{\theta}_{\text{ben}}; \alpha) = \phi\,\boldsymbol{\mu}^\top\nabla h_\alpha(\boldsymbol{\theta}_{\text{ben}}). \tag{26}$$

It suffices to show $\boldsymbol{\mu}^\top\nabla h_\alpha(\boldsymbol{\theta}_{\text{ben}}) > 0$.

Let $\mathbf{x} = \alpha\mathbf{v} - \boldsymbol{\mu} + \mathbf{C}^{1/2}\mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and set $B = \mathbf{x}^\top\boldsymbol{\theta}_{\text{ben}}$. The gradient of the poisoned expectation in the direction $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu}^\top\nabla h_\alpha(\boldsymbol{\theta}_{\text{ben}}) = \mathbb{E}[L'(B)\,\boldsymbol{\mu}^\top\mathbf{x}].$$

Using $\mathbf{v}^\top\boldsymbol{\mu} = 0$: $\boldsymbol{\mu}^\top\mathbf{x} = -\|\boldsymbol{\mu}\|^2 + \boldsymbol{\mu}^\top\mathbf{C}^{1/2}\mathbf{z}$. We now separate the deterministic part (involving $-\|\boldsymbol{\mu}\|^2$) from the Gaussian part (involving $\mathbf{z}$). Applying Stein's lemma to the Gaussian part:

$$\mathbb{E}[L'(B)\,\boldsymbol{\mu}^\top\mathbf{C}^{1/2}\mathbf{z}] = (\boldsymbol{\mu}^\top\mathbf{C}\boldsymbol{\theta}_{\text{ben}})\,\mathbb{E}[L''(B)],$$

where we used $\partial B/\partial\mathbf{z} = \mathbf{C}^{1/2}\boldsymbol{\theta}_{\text{ben}}$. Therefore

$$\boldsymbol{\mu}^\top\nabla h_\alpha(\boldsymbol{\theta}_{\text{ben}}) = \underbrace{-\|\boldsymbol{\mu}\|^2\mathbb{E}[L'(B)]}_{>0} + \underbrace{(\boldsymbol{\mu}^\top\mathbf{C}\boldsymbol{\theta}_{\text{ben}})\mathbb{E}[L''(B)]}_{\geq 0}. \tag{27}$$

The first term is strictly positive since L is strictly decreasing ($L' < 0$ everywhere by Assumption 3).

For the second term, we invoke Lemma 13(ii), which gives $\boldsymbol{\theta}_{\text{ben}} = a(\frac{\lambda}{1-\phi}\mathbf{I} + \tau\mathbf{C})^{-1}\boldsymbol{\mu}$ with $a > 0$. Writing $\mathbf{R} = \frac{\lambda}{1-\phi}\mathbf{I} + \tau\mathbf{C}$:

$$\boldsymbol{\mu}^\top\mathbf{C}\boldsymbol{\theta}_{\text{ben}} = a\,\boldsymbol{\mu}^\top\mathbf{C}\mathbf{R}^{-1}\boldsymbol{\mu} \geq 0,$$

since $\mathbf{C}$ and $\mathbf{R}^{-1}$ are commuting positive semidefinite matrices (they share the eigenbasis of $\mathbf{C}$) and hence their product $\mathbf{C}\mathbf{R}^{-1}$ is also positive semidefinite. Moreover $\mathbb{E}[L''(B)] \geq 0$ by convexity of L.

Hence both terms in (27) are non-negative and the first is strictly positive, giving $\boldsymbol{\mu}^\top\nabla h_\alpha(\boldsymbol{\theta}_{\text{ben}}) > 0$. Substituting into (26) yields $\boldsymbol{\mu}^\top\nabla\mathcal{L}_{\text{pop}}(\boldsymbol{\theta}_{\text{ben}}; \alpha) > 0$. □

C COMPARING ERM AND INFORMATION LIMIT

C.1 THE NOISE FLOOR IS OF ORDER κ

Assumption 1 constrains $\mathbf{C}$ only from above, through $\|\mathbf{C}\| = O(1)$. It therefore does not by itself force the finite-sample term $\zeta = \frac{\gamma}{n}\operatorname{tr}[\mathbf{R}^2\mathbf{C}^2]$ to be of order κ: if the spectrum of $\mathbf{C}$ collapses towards zero, every summand in $\operatorname{tr}[\mathbf{R}^2\mathbf{C}^2]$ is small and ζ can be $o(\kappa)$. The following mild nondegeneracy condition is what makes the κ-scaling in Section 4 precise, and it is what our experiments satisfy.

Assumption 4 (Spectral nondegeneracy)**.** *There exist constants $c_0 > 0$ and $\delta \in (0, 1]$, independent of n, such that at least δp of the eigenvalues $s_1^2, \dots, s_p^2$ of $\mathbf{C}$ satisfy $s_i^2 \geq c_0$.*

The condition is weak: it rules out only the degenerate case in which the whole spectrum of $\mathbf{C}$ concentrates at zero and in particular does *not* require $\mathbf{C} \succeq c_0\mathbf{I}$, which would fail for image data. For the CIFAR-10 covariance used in Table 1 the eigenvalues span $\log_{10}\lambda \in [-5.80, 1.74]$ and ζ accounts for 19.67% of σ^2 at $\kappa \approx 0.32$.

Lemma 18 (Noise floor of order κ)**.** *Under Assumptions 1 and 4, and with $\gamma \geq \gamma_0 > 0$,*

$$\zeta \;=\; \frac{\gamma}{n}\operatorname{tr}\big[\mathbf{R}(\lambda,\tau)^2\mathbf{C}^2\big] \;\geq\; \frac{\gamma_0\,\delta\,c_0^2}{(\lambda+\tau\|\mathbf{C}\|)^2}\,\frac{p}{n},$$

so $\zeta \geq c\,\kappa$ for a constant $c > 0$ not depending on α.

Proof. Since $\mathbf{R}(\lambda,\tau)$ and $\mathbf{C}$ commute, in the eigenbasis of $\mathbf{C}$

$$\operatorname{tr}[\mathbf{R}^2\mathbf{C}^2] \;=\; \sum_{i=1}^{p}\frac{s_i^4}{(\lambda+\tau s_i^2)^2}.$$

Every summand is non-negative, and $x \mapsto x^2/(\lambda+\tau x)^2$ is increasing on $x > 0$ for $\lambda, \tau > 0$. Restricting the sum to the at least δp indices with $s_i^2 \geq c_0$, and bounding each such summand below by its value at $s_i^2 = c_0$ while bounding the denominator above using $s_i^2 \leq \|\mathbf{C}\|$, gives

$$\operatorname{tr}[\mathbf{R}^2\mathbf{C}^2] \;\geq\; \delta p\,\frac{c_0^2}{(\lambda+\tau\|\mathbf{C}\|)^2}.$$

Multiplying by $\gamma/n \geq \gamma_0/n$ gives the claim. □

Remark 9 (Lower bound on γ)**.** *The lower bound $\gamma \geq \gamma_0 > 0$ in Lemma 18 holds because $\gamma = \mathbb{E}[f^2(r_K)]$ with $f = -L'(\operatorname{prox}_{\delta L}(\cdot)) > 0$ strictly, by the strict monotonicity of L in Assumption 3, and because the margins r_K have means and variances bounded uniformly in α by Step 1 of the proof of Proposition 6, so $f(r_K)$ is bounded away from zero on a set of probability bounded away from zero.*

C.2 VARIANCE DECOMPOSITION AND CLEAN ACCURACY BEYOND THE EIGENVECTOR SIMPLIFICATION

Section 4 derived the variance decomposition (9) under the simplifying assumption that $\boldsymbol{\mu}$ and $\mathbf{v}$ are eigenvectors of $\mathbf{C}$. We show here that the same structure persists for generic $\boldsymbol{\mu}, \mathbf{v}$ satisfying Assumption 2, so the monotone clean-accuracy conclusion does not rely on that simplification.

Since $\mathbf{R} = (\lambda\mathbf{I}+\tau\mathbf{C})^{-1}$ commutes with $\mathbf{C}$, we have $\mathbf{R}^2\mathbf{C} = \mathbf{RCR}$, and Theorem 1 gives

$$\sigma^2 = (\eta_1\boldsymbol{\mu}_1+\eta_2\boldsymbol{\mu}_2)^\top\mathbf{RCR}(\eta_1\boldsymbol{\mu}_1+\eta_2\boldsymbol{\mu}_2)+\zeta, \qquad \zeta := \frac{\gamma}{n}\operatorname{tr}[\mathbf{R}^2\mathbf{C}^2].$$

Writing $\eta_1\boldsymbol{\mu}_1+\eta_2\boldsymbol{\mu}_2 = (\eta_1-\eta_2)\boldsymbol{\mu}+\eta_2\alpha\mathbf{v}$ and expanding,

$$\sigma^2 = (\eta_1-\eta_2)^2\,\boldsymbol{\mu}^\top\mathbf{RCR}\boldsymbol{\mu} + 2(\eta_1-\eta_2)\eta_2\alpha\,\boldsymbol{\mu}^\top\mathbf{RCR}\mathbf{v} + \eta_2^2\alpha^2\,\mathbf{v}^\top\mathbf{RCR}\mathbf{v}+\zeta.$$

The matrix $\mathbf{RCR} = \mathbf{C}(\lambda\mathbf{I}+\tau\mathbf{C})^{-2}$ is a bounded function of $\mathbf{C}$, so by a Weierstrass-polynomial approximation argument (cf. Remark 6) Assumption 2 gives $\boldsymbol{\mu}^\top\mathbf{RCR}\mathbf{v} = o(1)$, and the cross term vanishes. To re-express in terms of the mean alignments $a := \mathbb{E}[\boldsymbol{\mu}^\top\tilde{\boldsymbol{\theta}}]$ and $b := \mathbb{E}[\mathbf{v}^\top\tilde{\boldsymbol{\theta}}]$, recall that (again up to $o(1)$ from $\boldsymbol{\mu}^\top\mathbf{R}\mathbf{v} = o(1)$) $a = (\eta_1-\eta_2)\,\boldsymbol{\mu}^\top\mathbf{R}\boldsymbol{\mu}$ and $b = \eta_2\alpha\,\mathbf{v}^\top\mathbf{R}\mathbf{v}$. Since $\lambda > 0$,

both $\boldsymbol{\mu}^\top \mathbf{R}\boldsymbol{\mu}$ and $\mathbf{v}^\top \mathbf{R}\mathbf{v}$ are strictly positive, so the substitutions $(\eta_1 - \eta_2) = a/(\boldsymbol{\mu}^\top \mathbf{R}\boldsymbol{\mu}) + o(1)$ and $\eta_2 \alpha = b/(\mathbf{v}^\top \mathbf{R}\mathbf{v}) + o(1)$ are well-defined as algebraic identities. We caution that a priori $\boldsymbol{\mu}^\top \mathbf{R}\boldsymbol{\mu}$ and $\mathbf{v}^\top \mathbf{R}\mathbf{v}$ could vanish along the sequence if the fixed-point parameter τ grows unboundedly, in which case the $o(1)$ residuals propagating below need not be small; the decomposition is quantitatively informative in the regime where τ remains bounded. Assuming this, we substitute:

$$\sigma^2 \;=\; \beta_\mu\, a^2 \;+\; \beta_v\, b^2 \;+\; \zeta \;+\; o(1), \qquad \beta_\mu := \frac{\boldsymbol{\mu}^\top \mathbf{R}\mathbf{C}\mathbf{R}\boldsymbol{\mu}}{(\boldsymbol{\mu}^\top \mathbf{R}\boldsymbol{\mu})^2}, \quad \beta_v := \frac{\mathbf{v}^\top \mathbf{R}\mathbf{C}\mathbf{R}\mathbf{v}}{(\mathbf{v}^\top \mathbf{R}\mathbf{v})^2}. \tag{28}$$

Both β_μ and β_v are non-negative ($\boldsymbol{\mu}^\top \mathbf{R}\mathbf{C}\mathbf{R}\boldsymbol{\mu} = \|\mathbf{C}^{1/2}\mathbf{R}\boldsymbol{\mu}\|^2$), and (28) is the general-case analogue of (9): the eigenvector simplification recovers $\beta_\mu = \lambda_\mu/\|\boldsymbol{\mu}\|^2$ and $\beta_v = \lambda_{\mathbf{v}}$. In the spectral basis of $\mathbf{C}$, writing $\boldsymbol{\mu} = \sum_i \mu_i \mathbf{e}_i$ with $\mathbf{C}\mathbf{e}_i = \lambda_i \mathbf{e}_i$,

$$\beta_\mu = \frac{\sum_i \mu_i^2\, \lambda_i/(\lambda + \tau\lambda_i)^2}{\left(\sum_i \mu_i^2/(\lambda + \tau\lambda_i)\right)^2},$$

a resolvent-weighted average of the eigenvalues of $\mathbf{C}$ along $\boldsymbol{\mu}$, and analogously for $\beta_{\mathbf{v}}$.

Clean accuracy in the general case. Substituting (28) into (8) with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ (using the concentration $\tilde{\boldsymbol{\theta}}^\top \mathbf{C}\tilde{\boldsymbol{\theta}} \to \sigma^2$) and folding the small trigger contribution $\beta_v b^2$ (which vanishes as $\alpha \to \infty$ by Proposition 6 and is empirically subdominant at all α, cf. Table 1) into an effective noise $\tilde{\zeta} := \beta_v b^2 + \zeta$:

$$\mathrm{Acc}_{\mathrm{clean}} \;\approx\; \Phi\left(\frac{a}{\sqrt{\beta_\mu\, a^2 + \tilde{\zeta}}}\right). \tag{29}$$

This retains the structure used in Section 4:

(i) *Information limit* ($\kappa \to 0$, so $\zeta \to 0$ and, by Proposition 7, $b \to 0$ at large α, giving $\tilde{\zeta} \to 0$). The argument of Φ collapses to $1/\sqrt{\beta_\mu}$, independent of a: clean accuracy is flat in α.

(ii) *Proportional regime* ($\zeta > 0$). Differentiating at fixed $(\beta_\mu, \tilde{\zeta})$,

$$\frac{\partial}{\partial a}\,\frac{a}{\sqrt{\beta_\mu a^2 + \tilde{\zeta}}} = \frac{\tilde{\zeta}}{(\beta_\mu a^2 + \tilde{\zeta})^{3/2}} > 0,$$

so clean accuracy is strictly increasing in the benign alignment a whenever $\tilde{\zeta} > 0$.

We emphasise that (ii) is a partial statement about the dependence of accuracy on a, and not a monotonicity result in α. The coefficients β_μ, b and $\tilde{\zeta}$ are themselves functions of α through the fixed-point scalars $(\tau, \gamma, \delta, \eta_1, \eta_2)$ of Theorem 1, so the total derivative of (29) in α carries further terms that we do not control. What the display does establish is the *mechanism*: the dependence of clean accuracy on the benign alignment is present exactly when $\tilde{\zeta} > 0$ and disappears when $\tilde{\zeta} = 0$. Combined with the endpoint ordering of Propositions 7 and 8, which place the benign alignment at $\alpha = 0$ strictly below its unpoisoned limit, this fixes the direction of the effect between the two endpoints. Strict monotonicity in α for general convex losses remains open; we verify it numerically in all our experiments (Figures 2 and 3).

Thus the qualitative claim of Section 4 — that the finite-sample noise floor ζ is the mechanism coupling benign alignment to clean accuracy — is independent of the eigenvector simplification. The simplification enters only through the interpretation of the scalar β_μ as a single eigenvalue of $\mathbf{C}$ rather than as the resolvent-weighted spectral average above.

D FURTHER EXPERIMENTAL DETAILS

Code and raw results for all CIFAR-10 experiments in this appendix are provided in the supplementary material.

Dataset and preprocessing. Except for the synthetic linear-regression experiments in Figures 3 and 8, all experiments use CIFAR-10 restricted to classes 0 ("airplane") and 1 ("automobile"), pooling the original 50,000 training and 10,000 test images (12,000 total per class pair — 6,000 per class) and applying a random 80/20 train/test split, yielding $n_{\text{train}} \approx 9{,}600$ and $n_{\text{test}} \approx 2{,}400$ samples. The split is redrawn for each repeat, seeded from 42 upwards, so that the reported bands include split variance as well as training noise; the one exception is the eigenvector-direction experiment, which holds the split fixed at seed 42 so that every trigger direction sees identical data. Pixel intensities are normalised to $[0, 1]$ (divided by 255) and the training set mean is subtracted from both splits. The feature dimension is $d = 32 \times 32 \times 3 = 3{,}072$.

Poisoning protocol. For the CIFAR-10, Gaussian-surrogate, and ResNet experiments, a fraction $\phi = 0.05$ of the full training set, drawn exclusively from the class -1 (airplane) training samples, is selected uniformly at random and poisoned: the trigger vector $\alpha\mathbf{v}$, with $\|\mathbf{v}\| = 1$, is added to each selected sample, and its label is flipped to $+1$. The trigger direction is a 2×2 corner patch in the top-left of the image applied uniformly across all three RGB channels, normalised to a unit vector and then scaled to norm α.

For these image experiments, we sweep α over 20 equally-spaced values in $[0, 1.5]$. Attack success rate is always evaluated at a fixed test trigger norm $\alpha_{\text{test}} = 0.5$ in the same direction as the training trigger. The synthetic linear-regression experiments in Figures 3 and 8 use the synthetic Gaussian-mixture setup and poisoning parameters specified in the corresponding figure-specific paragraphs below.

Gaussian surrogate construction. The Gaussian surrogate dataset is constructed to match the empirical first- and second-order statistics of the real CIFAR-10 training split. From the centred training data we estimate the class mean $\hat{\boldsymbol{\mu}} = \frac{1}{n_+} \sum_{i:\, y_i=+1} \mathbf{x}_i$ and the pooled within-class covariance $\hat{\mathbf{C}} = \frac{1}{2}\hat{\Sigma}_{+1} + \frac{1}{2}\hat{\Sigma}_{-1}$, regularised by a small ridge $\epsilon\mathbf{I}$ ($\epsilon = 10^{-4}\,\mathrm{tr}(\hat{\mathbf{C}})/p$). Synthetic samples are then drawn as $\mathbf{x} \sim \mathcal{N}(\pm\hat{\boldsymbol{\mu}},\, \hat{\mathbf{C}})$ via the Cholesky factorisation $\hat{\mathbf{C}} = \mathbf{L}\mathbf{L}^\top$, using $\mathbf{x} = \pm\hat{\boldsymbol{\mu}} + \mathbf{L}\mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The surrogate dataset uses the same n_{train} and n_{test} as the corresponding real split and balanced class sizes ($\lfloor (n_{\text{train}} + n_{\text{test}})/2 \rfloor$ samples per class). All other experimental choices—poisoning fraction, trigger direction, regularisation strength, and number of seeds—are identical to those of the real CIFAR-10 experiment.

Figure 1: ResNet-18 triple panel. We train a ResNet-18 (He et al., 2015) with a single scalar output (binary classification). For CIFAR-10's 32×32 images we replace the standard 7×7/stride-2 first convolution with a 3×3/stride-1 convolution and remove the initial max-pool, following common practice for small images. Training uses SGD with momentum 0.9, weight decay 5×10^{-4}, initial learning rate 0.05 annealed to 0 via a cosine schedule over 50 epochs, and mini-batch size 128. The experiment is repeated for 25 independent train/test splits (seeds $42, 43, \ldots, 66$); shaded bands show mean $\pm$ one standard deviation.

The three panels show: (left) clean test accuracy versus training trigger norm α; (centre) ASR versus α; (right) ASR versus the $\log_{10}$-eigenvalue of the pooled within-class covariance $\mathbf{C}$ of CIFAR-10. For the rightmost panel, 20 eigenvectors of $\mathbf{C}$ are selected log-uniformly across the full eigenvalue spectrum; each eigenvector is used as the trigger direction with $\alpha = \alpha_{\text{test}} = 0.5$ fixed. The pooled covariance is $\mathbf{C} = \frac{1}{2}\hat{\Sigma}_{+1} + \frac{1}{2}\hat{\Sigma}_{-1}$ with a small ridge $\epsilon\mathbf{I}$ ($\epsilon = 10^{-4}\,\mathrm{tr}(\mathbf{C})/d$) for numerical stability, eigendecomposed via `numpy.linalg.eigh`. Each eigenvector point is averaged over 25 independent runs.

Figure 2: CIFAR-10 vs. Gaussian empirical (logistic regression). Logistic regression with ℓ_2 regularisation strength $\lambda = 10^{-4}$ is trained on (i) the real CIFAR-10 split and (ii) a synthetic Gaussian dataset whose class-conditional means and covariance are estimated empirically from the same CIFAR-10 training split; both use the same n_{train} and n_{test}. The classifier is obtained by

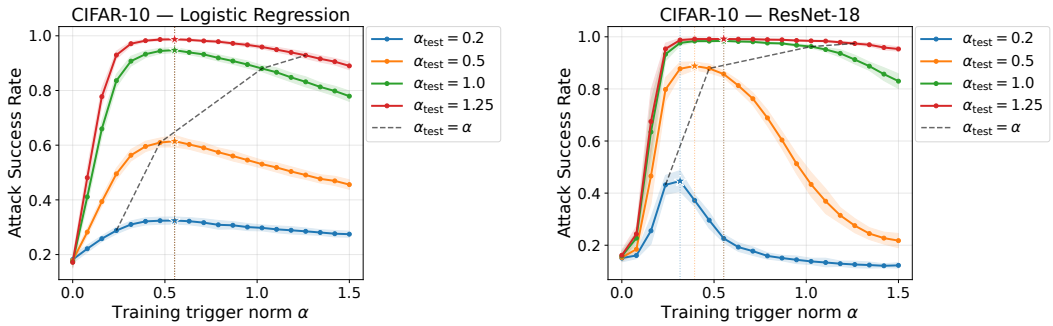
minimising

$$\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\log\big(1+e^{-y_i(\boldsymbol{\theta}^\top \mathbf{x}_i)}\big)$$

via PyTorch L-BFGS with the strong Wolfe line-search condition (up to 2,000 iterations). The experiment is repeated for 25 independent seeds; shaded bands show mean $\pm$ one standard deviation. The four panels display clean test accuracy, ASR, alignment $\boldsymbol{\theta}^\top \boldsymbol{\mu}$, and alignment $\boldsymbol{\theta}^\top \mathbf{v}$ as functions of training trigger norm α.

Trigger orthogonality check. We directly verify Assumption 2 for the corner-patch trigger. With $\mathbf{R} = (\lambda \mathbf{I} + \tau \mathbf{C})^{-1}$, the normalised resolvent overlap $\rho_R = |\boldsymbol{\mu}^\top \mathbf{R}\mathbf{v}| / \sqrt{(\boldsymbol{\mu}^\top \mathbf{R}\boldsymbol{\mu})(\mathbf{v}^\top \mathbf{R}\mathbf{v})}$ is ≈ 0.007 at the fixed-point τ, and remains in $[0.0067, 0.0086]$ across a 16-fold range of τ. This is about $2.6\times$ below the $1/\sqrt{d} \approx 0.018$ overlap of a random unit direction, so the assumption holds closely for the trigger used in our experiments.

Figure 5: sensitivity to test trigger norm. The attack success rate in Figures 1 and 2 is reported at a fixed test trigger norm $\alpha_{\text{test}} = 0.5$. To check that the qualitative ASR-vs-α shape is not an artifact of this choice, we repeat both experiments at additional values of α_{test}. Panel (a) shows the logistic-regression sweep on CIFAR-10 and panel (b) shows the ResNet-18 sweep, both with all other settings identical to the main-text figures. We additionally connect the points where $\alpha = \alpha_{\text{test}}$, and label the maximum value of each curve with a star.



(a) Logistic regression, CIFAR-10.

(b) ResNet-18, CIFAR-10.

Figure 5: ASR vs. training trigger strength α at several values of the test trigger norm α_{test}.

Multiclass and warping experiments. Our theory is stated for a binary problem with a single additive trigger direction. To test whether the phenomena survive relaxing each of these in turn, we run two further ResNet-18 experiments: one on the full 10-class CIFAR-10 with the same additive corner-patch trigger (Figure 6), and one that replaces the additive trigger by a non-additive, input-dependent warping attack (Figure 7). Both use the ResNet-18 architecture, optimiser and schedule of Figure 1. Unlike the main-text runs, both load images with the spatial axes in their natural positions, so that the network's height and width axes are the real image axes; this is immaterial for the additive corner patch but essential for a spatial warp.

Figure 6: multiclass CIFAR-10. We use all 60,000 CIFAR-10 images with their integer labels $0, \ldots, 9$, pooled and split 80/20 ($n_{\text{train}} = 48{,}000$, $n_{\text{test}} = 12{,}000$, $d = 3{,}072$, $\kappa \approx 0.064$), normalised to $[0, 1]$ and centred by the training mean. The trigger is the same unit-norm 2×2 top-left corner patch as in the binary experiments, and the poisoning fraction is $\phi = 0.05$ of the training set, drawn from the classes other than the target: for each selected sample the trigger $\alpha\mathbf{v}$ is added and the label is overwritten with the target class 0. The final layer is $\text{Linear}(512, 10)$ trained with cross-entropy. Clean accuracy is 10-way accuracy on the untouched test set, and ASR is the fraction of triggered test samples whose true label is not the target class that are predicted as the target class. The sweep is $\alpha \in \text{linspace}(0, 1.5, 20)$ with ASR evaluated at $\alpha_{\text{test}} = 0.5$, averaged over 30 repeats with independent splits. For the third panel the covariance is the empirical covariance of the centred training set pooled over all classes, with no ridge (it is full rank at $n = 48{,}000$, with all

3,072 eigenvalues positive and $\log_{10}\lambda \in [-5.80, 1.74]$; the small ridge used in the binary experiment would sit above the smallest true eigenvalue and floor the bottom decade). As before, 20 eigenvectors are chosen log-uniformly in eigenvalue and used as trigger directions at $\alpha = \alpha_{\text{test}} = 0.5$, with the split held fixed so that the error bars reflect training noise alone.

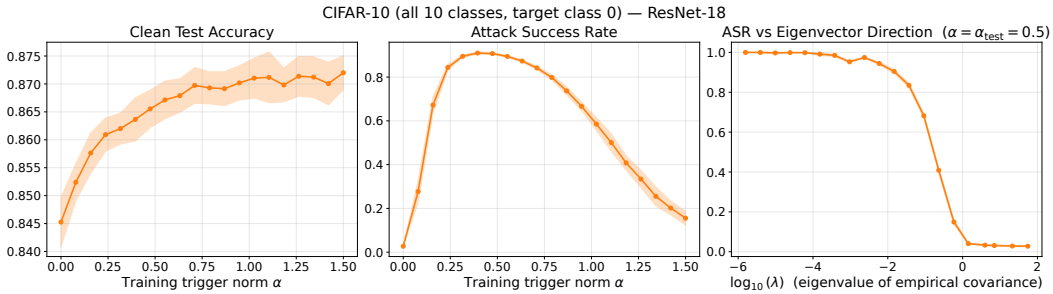


Figure 6: Multiclass ResNet-18 on 10-class CIFAR-10, target class 0 ($\phi = 0.05$, $\alpha_{\text{test}} = 0.5$, averaged over 30 runs). All three phenomena persist: clean accuracy rises with α, ASR peaks at a finite α and then decays, and ASR is near one along the smallest eigendirections of the covariance and near zero along the largest.

Figure 7: warping-based (WaNet-style) attack. Here the trigger is not additive. Following Nguyen & Tran (2021), a warp field is drawn once as a 4×4 control grid of uniform offsets, normalised by its mean absolute value and bicubically upsampled to 32×32; the poisoned image is the original resampled through the clamped field $\text{clamp}(\text{identity} + \beta \cdot \text{noise}/32, -1, 1)$. The warp strength β plays the role of the trigger norm α, and the control grid is drawn from a single fixed seed shared across every β and every repeat, so that only the strength of the warp varies and not its direction. This is the analogue of holding $\mathbf{v}$ fixed and scaling α; $\beta = 0$ is the identity warp. Poisoning *replaces* a $\phi = 0.05$ fraction of the training set, drawn from the $y = -1$ class, with its warped version and relabels it to $+1$ (WaNet's auxiliary noise mode is disabled, as the additive experiments have no analogue for it). Images are loaded uncentred and the training mean is subtracted only after warping, since warping and centring do not commute. The setting is otherwise that of Figure 1: binary CIFAR-10 (classes 0 and 1), 80/20 split with $n_{\text{train}} = 9{,}600$, and 5 repeats with seeds $42, \ldots, 46$, so the shaded bands include split variance as well as training noise. We sweep $\beta \in \text{linspace}(0, 12, 20)$ and hold the test-time warp fixed at $\beta_{\text{test}} = 4.0$, mirroring the $\alpha/\alpha_{\text{test}}$ split. Both phenomena survive: clean accuracy rises from 0.949 at $\beta = 0$ to 0.973 at $\beta = 12$, while ASR peaks at 0.760 ± 0.028 near $\beta = 3.8$, close to β_{test}, and falls to 0.082 by $\beta = 12$. Phenomenon (iii) has no analogue here, since a warping attack has no single trigger direction. We note that β is not on the same scale as α: the mean displacement $\|\mathbf{W}\mathbf{x} - \mathbf{x}\|$ grows sublinearly in β (reaching 8.89 at β_{test} and 13.31 at $\beta = 12$) because the clamp binds beyond $\beta \approx 1$.

Figure 3: projection curves across aspect ratios. This figure is a theory-only comparison across several proportional-regime aspect ratios $\kappa = p/n$, including the overparameterised regime $\kappa > 1$. We solve the deterministic theory on the grid $\alpha \in [0, 30]$ for the four configurations

$$(p, n, \kappa) \in \{(1100, 1000, 1.10),\ (1050, 1000, 1.05),\ (1000, 2000, 0.50),\ (1000, 5000, 0.20)\}.$$

For each value of κ, the dashed curves show the squared-loss prediction and the solid curves show the logistic-loss fixed-point prediction. The left panel plots the trigger projection $|\langle \theta^\star, v \rangle|$, which rises for small α, reaches a finite peak, and then decays. The right panel plots the benign projection $\langle \theta^\star, \mu \rangle$, which increases with α. The main purpose of the figure is to show that these qualitative behaviours are stable across different values of p/n, including the overparameterised regime $\kappa > 1$.

Figure 8: eigenvector spectral effect. This figure isolates the dependence of the trigger projection on the trigger eigenvalue s_v^2. We use a synthetic non-isotropic covariance model in which the clean and trigger directions are orthogonal eigenvectors of $\mathbf{C}$. Writing

$$u_\mu := \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|},$$

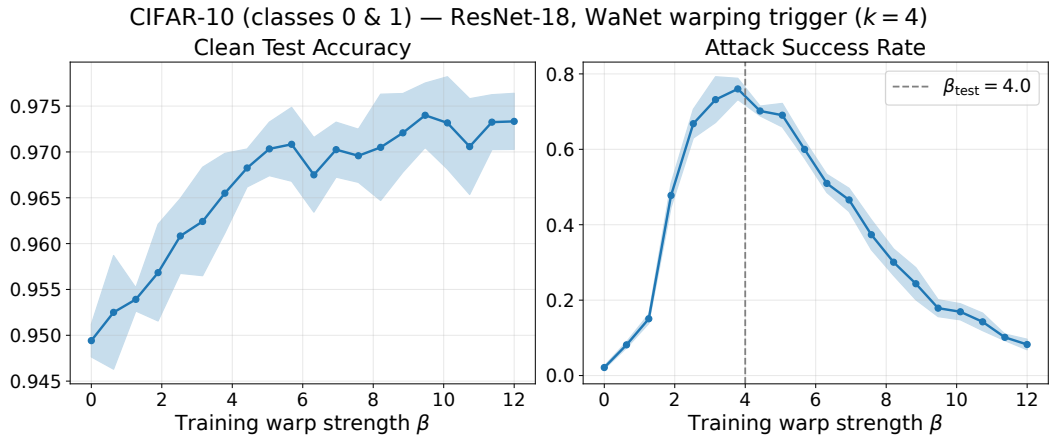


Figure 7: Warping-based (WaNet-style) attack on ResNet-18, CIFAR-10 classes 0 & 1 ($\phi = 0.05$, $\beta_{\text{test}} = 4.0$, averaged over 5 runs). The warp strength β replaces the trigger norm α. Phenomena (i)–(ii) persist for this non-additive, input-dependent trigger; (iii) has no analogue.

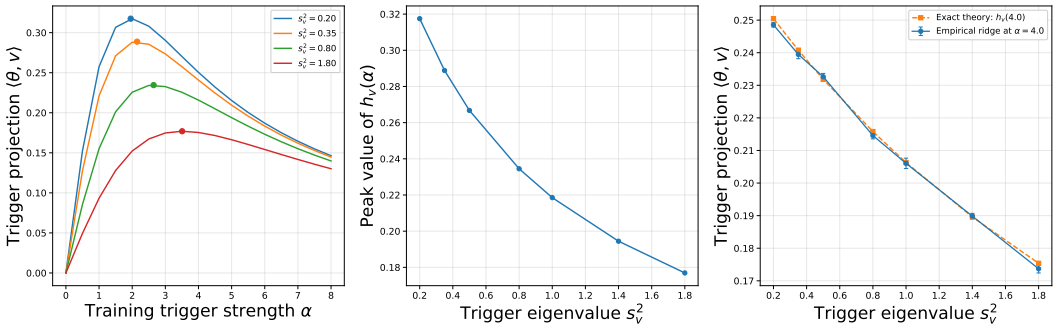


Figure 8: Eigenvector specialisation for the squared loss. Left: exact trigger-projection curves $h_v(\alpha)$ for several trigger eigenvalues s_v^2, with the peak marked on each curve (coloured dots). Middle: the peak value $\max_\alpha h_v(\alpha)$ decreases as s_v^2 increases. Right: at fixed α, exact theory and empirical ridge estimates agree, and both show that lower-variance trigger eigendirections induce larger trigger projections. Here $\mathbf{C}$ is a rank-two perturbation of the identity.

the covariance used in the plotted experiment is

$$\mathbf{C} = s_{\text{rest}}^2\big(\mathbf{I} - u_\mu u_\mu^\top - \mathbf{v}\mathbf{v}^\top\big) + s_\mu^2 u_\mu u_\mu^\top + s_v^2 \mathbf{v}\mathbf{v}^\top.$$

Thus u_μ has eigenvalue s_μ^2, the trigger direction $\mathbf{v}$ has eigenvalue s_v^2, and every direction orthogonal to $\text{span}\{u_\mu, \mathbf{v}\}$ has eigenvalue s_{rest}^2. In the plotted setting, $\|\boldsymbol{\mu}\| = 1$, $\|\mathbf{v}\| = 1$, and $s_{\text{rest}}^2 = 1$, so the covariance reduces to the rank-two perturbation

$$\mathbf{C} = \mathbf{I} + (s_\mu^2 - 1)\boldsymbol{\mu}\boldsymbol{\mu}^\top + (s_v^2 - 1)\mathbf{v}\mathbf{v}^\top.$$

The eigenvector specialisation in Corollary 4 is more general than this particular plotted construction. It only requires that $\boldsymbol{\mu}$ and $\mathbf{v}$ are orthogonal eigenvectors of $\mathbf{C}$, with eigenvalues s_μ^2 and s_v^2, respectively. The covariance on the orthogonal complement of $\text{span}\{\boldsymbol{\mu}, \mathbf{v}\}$ may be arbitrary positive definite, subject to the standing boundedness assumptions; in the displayed projection formulas, its effect enters through the fixed-point scalar τ.

The parameters used in the figure are

$$p = 300, \qquad n = 5000, \qquad \kappa = 0.06, \qquad \|\boldsymbol{\mu}\| = 1, \qquad \phi = 0.20, \qquad \lambda = 0.50,$$

with $s_\mu^2 = 2.0$ and $s_{\text{rest}}^2 = 1.0$. We vary only the trigger eigenvalue s_v^2, keeping $\mathbf{v}$ a unit-norm covariance eigenvector orthogonal to $\boldsymbol{\mu}$. The left panel plots exact trigger-projection curves $h_v(\alpha)$ for

$$s_v^2 \in \{0.20, 0.35, 0.80, 1.80\},$$

over the grid

$$\alpha \in \{0, 0.5, 1.0, \ldots, 8.0\}.$$

The middle panel plots the peak value $\max_\alpha h_v(\alpha)$ for

$$s_v^2 \in \{0.20, 0.35, 0.50, 0.80, 1.00, 1.40, 1.80\}.$$

The right panel fixes the training trigger strength at $\alpha = 4.0$ and compares the exact theoretical prediction for $h_v(4.0)$ with finite-sample ridge estimates $\mathbf{v}^\top \hat{\boldsymbol{\theta}}$. The finite-sample sweep is repeated over 8 independent repetitions with seed 2027. Markers show empirical means, and error bars show the standard error of the mean.

These panels illustrate Proposition 5: within the covariance-eigenvector class, increasing s_v^2 increases the denominator of the trigger-projection formula through

$$B_v = \lambda + \tau s_v^2,$$

and therefore attenuates the learned trigger projection.

Figure 4: poisoning-fraction sweep. This figure shows how the square-loss projections vary with the poisoning fraction ϕ. We use the same synthetic Gaussian-mixture setup as in the linear square-loss experiments, and sweep

$$\phi \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$$

and

$$\alpha \in \{0, 0.5, 1.0, \ldots, 10.0\}.$$

For each pair (ϕ, α), we generate a finite training sample from the absorbed two-component Gaussian mixture, fit the squared-loss ridge estimator $\hat{\boldsymbol{\theta}}$, and record the empirical projections

$$\boldsymbol{\mu}^\top \hat{\boldsymbol{\theta}}, \qquad \mathbf{v}^\top \hat{\boldsymbol{\theta}}.$$

We also compute the corresponding square-loss theory predictions from the closed-form linear formulas. Solid curves show empirical ridge estimates and dashed curves show the theory predictions.

The left panel plots the trigger alignment $h_v(\alpha)$, while the right panel plots the benign alignment $h_\mu(\alpha)$. Increasing ϕ has two different effects. For the trigger alignment, larger poisoning fractions can increase h_v at small trigger strengths, because more poisoned examples reinforce the trigger direction. At larger trigger strengths, the denominator effect in the square-loss formula dominates, and increasing ϕ can reduce the trigger alignment. In contrast, the benign alignment decreases with ϕ at fixed α, to leading order. The curves $\phi = 0$ and $\phi = 1/2$ are included as visual reference cases; the formal statements in the main text assume $0 < \phi < 1/2$.

Table 1: fixed-point solution. The table reports the asymptotic prediction for $\sigma^2 = \mathbb{E}[\tilde{\boldsymbol{\theta}}^\top \mathbf{C} \tilde{\boldsymbol{\theta}}]$, the variance of the classifier's output on a test point, derived from a high-dimensional analysis of regularised logistic regression. In the limit $p, n \to \infty$ with aspect ratio $p/n \to \kappa$, the trained weight vector concentrates around a Gaussian proxy $\tilde{\boldsymbol{\theta}}$ with average $(\lambda \mathbf{I} + \tau \mathbf{C})^{-1}(\eta_1 \boldsymbol{\mu} + \eta_2 \mathbf{v})$, where the scalar parameters $(\tau, \gamma, \delta, \eta_1, \eta_2)$ are the unique solution to a system of self-consistent fixed-point equations that couple the loss curvature, effective noise variance, and class-conditional signals.

These equations are solved numerically using the empirically estimated class mean $\boldsymbol{\mu}$ and pooled within-class covariance $\mathbf{C}$ from CIFAR-10 (classes 0 and 1), with $\lambda = 10^{-4}$, poisoning fraction $\phi = 0.05$, and aspect ratio $p/n \approx 0.32$. The expectation integrals that appear in the fixed-point equations are evaluated by Gauss–Hermite quadrature (100 nodes), and the system is iterated to convergence (residual $< 10^{-10}$).

At the solution, σ^2 decomposes as

$$\sigma^2 = \underbrace{(\eta_1 - \eta_2)^2 \, \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}}_{\text{mean-direction signal}} + \underbrace{\eta_2^2 \, \mathbf{v}^\top \mathbf{A} \mathbf{v}}_{\text{trigger-direction signal}} + \underbrace{2(\eta_1 - \eta_2)\eta_2 \, \boldsymbol{\mu}^\top \mathbf{A} \mathbf{v}}_{\text{cross term}} + \underbrace{\tfrac{\gamma}{n} \mathrm{tr}\big[(\lambda \mathbf{I} + \tau \mathbf{C})^{-2} \mathbf{C}^2\big]}_{\text{fundamental noise } \zeta}, \tag{30}$$

where $\mathbf{A} := (\lambda \mathbf{I} + \tau \mathbf{C})^{-2} \mathbf{C}$. The first three terms reflect how much of the classifier's variance is driven by alignment with the clean-class mean and the trigger direction respectively; the last term is an irreducible noise floor arising from finite-sample fluctuations, present even in the absence of poisoning.

Compute resources. Approximate costs on a single NVIDIA H100 GPU are as follows. The ResNet-18 experiments of Figure 1 (trigger-norm sweep and eigenvector panel, 25 seeds, each a 50-epoch training run) take about 10 hours. For Figure 2, the logistic-regression and Gaussian-surrogate sweeps take under an hour and the fixed-point theory curves about 3 minutes. For Figure 5, the logistic-regression sweep takes under an hour and the ResNet-18 sweep about 5 hours. The multiclass experiment of Figure 6 (30 seeds) takes about 55 hours, and the warping experiment of Figure 7 (5 seeds) about 1 hour. Fixed-point solves used to produce the theoretical curves and Table 1 take a few seconds per (α, κ) configuration.

Licensing of existing assets. CIFAR-10 (Krizhevsky, 2009) is publicly distributed by the University of Toronto for research use; the ResNet architecture is from He et al. (2015) and our implementation uses the `torchvision` reference (BSD-3-Clause). PyTorch (BSD-3-Clause), NumPy (BSD-3-Clause), and SciPy (BSD-3-Clause) were used for training and the fixed-point solver.