WHY BACKDOORING NEURAL NETWORKS IS SO EASY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Securing modern AI systems against backdoor attacks remains an open challenge and requires fundamentally principled estimates of the adversary's budget—the poison fraction π and trigger strength α needed to construct successful yet stealthy attacks. Motivated by recent empirical evidence that poisoning large language models can require a nearly constant number of malicious samples even as clean datasets grow, we derive an exact closed-form analysis of a quadratic neuron trained on a poisoned Gaussian mixture. We show, perhaps counterintuitively, that the same feature-learning dynamics that make neural networks powerful can also make them more vulnerable to backdoors. Specifically, with clean accuracy preserved to first order, $\mathcal{O}(\pi)$, we demonstrate that lazy learning imposes the inverse-square-root scaling $\alpha \propto \pi^{-1/2}$ for a successful attack, while feature learning induces a quadratic detector whose loss margin scales as $\mathcal{O}(\alpha^4)$, improving the attack budget to $\alpha \propto \pi^{-1/4}$. Consequently, nonlinear feature learning substantially reduces the trigger strength required at small poison fractions, thereby in a sense making feature learners more backdoor vulnerable. These results provide a theoretical mechanism consistent with large-scale empirical observations and demonstrate that security audits based on linear heuristics can systematically underestimate backdoor vulnerability in the widely adopted feature-learning regimes.

1 INTRODUCTION

Backdoor attacks are among the main threats to the AI supply chain (Gu et al., 2017; Chen et al., 2017; Li et al., 2022). An adversary who controls a fraction π of the training data adds a trigger pattern of strength α to these samples and relabels them with a target class. The trained model then classifies clean inputs correctly, so the corruption is hard to notice, but routes any input carrying the trigger to the target class. Defending against such attacks relies on auditing frameworks that verify the provenance of a trained model (Jia et al., 2021; Choi et al., 2023; Seddik et al., 2025), and the cost of an audit depends on how much the adversary must invest: if planting a backdoor with a small poison fraction requires a large, visible trigger, moderate audit budgets suffice (Hanneke et al., 2022; Kallas et al., 2025). The relevant quantity is therefore the adversary's *budget curve*, the smallest trigger strength $\alpha(\pi)$ that achieves a target attack success rate at poison fraction π.

Existing estimates of this curve are either empirical (Saha et al., 2020; Souly et al., 2025) or can be derived for lazy learners, i.e. linear, kernel or random-feature models whose internal representation does not move during training (Jacot et al., 2018; Chizat et al., 2019; Rahimi & Recht, 2007; Hayase & Oh, 2022). One can show that in these models the trigger acts only through a shift of the mean of the poisoned samples, which leads to the budget curve $\alpha \propto \pi^{-1/2}$: reducing the number of poisoned samples by a factor of 100 requires a trigger ten times stronger. Modern networks, however, operate in the feature-learning regime, where the representation adapts to the data (Yang & Hu, 2021; Ba et al., 2022; Damian et al., 2022), and a recent large-scale study (Souly et al., 2025) found that poisoning language models of very different sizes requires a nearly constant absolute number of malicious documents, i.e. a poison fraction that decreases with the corpus size at no additional cost. This raises the question of whether feature learning changes the budget curve, and by how much.

In this paper, we address this question with a statistical model that is simple enough to be solved exactly, yet retains a trainable nonlinear feature. We consider a single quadratic neuron, $f(\boldsymbol{x}) =$

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.

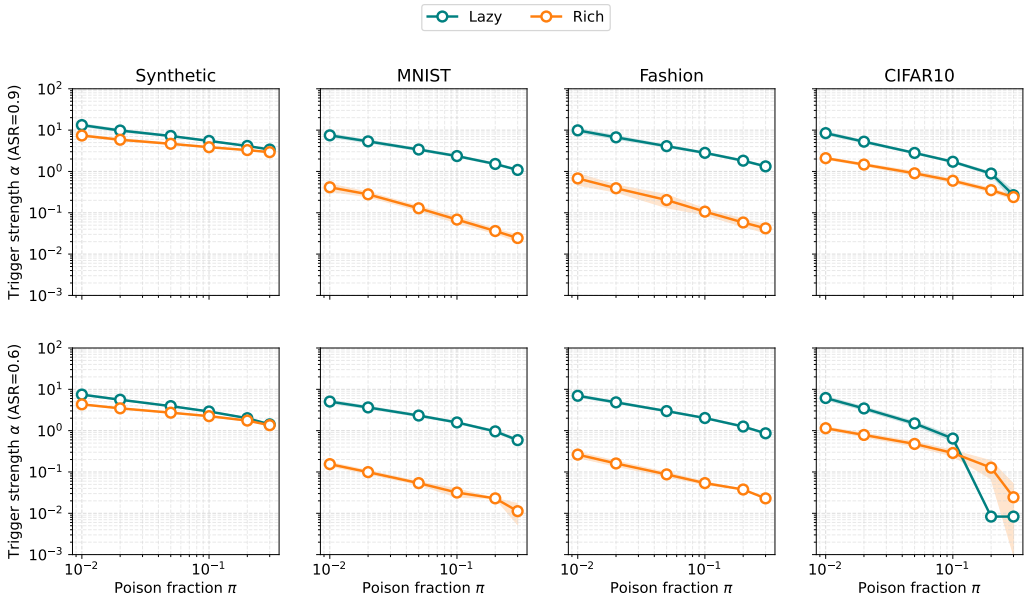


Figure 1: **Feature learning lowers the adversary's budget.** *Trigger strength α required to reach an attack success rate of* 0.9 *(top) and* 0.6 *(bottom) as a function of the poison fraction π, for a network trained in the lazy regime (frozen features) and in the rich regime (feature learning), on synthetic Gaussian data, MNIST, Fashion-MNIST and CIFAR-10; shaded areas are standard deviations over runs. Feature learning requires a smaller trigger at every poison fraction.*

$\boldsymbol{w}^\top \boldsymbol{x} + \frac{\beta}{2}(\boldsymbol{u}^\top \boldsymbol{x})^2 + c$, trained by least squares on a Gaussian mixture in which a fraction π of one class has been shifted by a trigger $\alpha \boldsymbol{v}$ and relabeled, and we compare two training regimes:

- *Linear regime*: the quadratic feature is not used ($\beta \equiv 0$) and the model is a linear classifier. This is the situation described by lazy training; we show that it is recovered exactly whenever the feature direction $\boldsymbol{u}$ is orthogonal to the trigger.
- *Feature-learning regime*: the feature direction $\boldsymbol{u}$ is adapted to the data and the curvature β is trained. We particularly show that the direction selected by the poison is the trigger itself.

We derive, in closed form, the clean accuracy and the attack success rate of the trained model for arbitrary noise covariance, alignment between signal and trigger, and arbitrary finite poison fraction and trigger strength (Theorem 1). In the linear regime the quantity that governs the attack is the trigger energy $\pi\alpha^2$ and the budget curve is $\alpha \propto \pi^{-1/2}$ (Theorem 2). In the feature-learning regime the poison creates a quadratic trigger detector whose contribution to the margin scales as α^4; the governing quantity becomes $\pi\alpha^4$ and the budget curve flattens to $\alpha \propto \pi^{-1/4}$ (Theorem 3). In both regimes the clean accuracy is unaffected to first order in π and the attack success rate saturates below one at the same explicit ceiling: feature learning does not raise the ceiling, it reaches it at a much lower cost. Our results are confirmed by experiments on synthetic data and on MNIST, Fashion-MNIST and CIFAR-10, and Figure 1 shows the same separation for a neural network trained in the lazy and in the rich regime, motivating our analysis beyond our simplified setting.

Related work: Backdoor attacks were introduced by Gu et al. (2017); subsequent work diversified the attack surface with blended (Chen et al., 2017), weight-level (Liu et al., 2018), clean-label (Turner et al., 2019; Shafahi et al., 2018), hidden (Saha et al., 2020) and frequency-domain (Xia et al., 2024) triggers; see Li et al. (2022) for a survey. Defenses include the detection of poisoned samples (Tran et al., 2018; Hayase et al., 2021; Wang et al., 2019; Gao et al., 2019), the verification of the training process (Jia et al., 2021; Choi et al., 2023; Seddik et al., 2025) and post-hoc mitigation (Li et al., 2021; Zhu et al., 2023), whose limits for language models were shown by Hubinger et al. (2024). The audit budget is formalized as a game by Kallas et al. (2025), and the estimates of

Jia et al. (2021); Hanneke et al. (2022) assume lazy learners. On the theoretical side, Biggio et al. (2012); Steinhardt et al. (2017) bounded the damage of a corrupted fraction, Manoj & Blum (2021) related backdoors to excess capacity, Goldwasser et al. (2022) showed that undetectable backdoors can be planted in expressive models, and Khaddaj et al. (2023) argued that a backdoor is a strong spurious feature; these results do not yield the trade-off between π and α. Our analysis builds on exactly solvable models of classification on Gaussian mixtures (Mignacco et al., 2020; Loureiro et al., 2021) and on the distinction between the lazy and the rich regime (Chizat et al., 2019; Woodworth et al., 2020; Ghorbani et al., 2021; Mei & Montanari, 2022); the quadratic neuron we consider is the poisoning analogue of the single-step feature-learning models of Ba et al. (2022); Damian et al. (2022). The closest works to ours, Hayase & Oh (2022); Hanneke et al. (2022), study poisoning for lazy learners; we depart from them by treating a model that learns a feature, which is what changes the cost exponent.

The remainder of the paper is organized as follows. Section 2 presents the poisoning model, the quadratic neuron and the two training regimes. Our main theoretical results are presented in Section 3. We further present experiments in Section 4 to support our findings. Finally, Section 5 concludes the paper. All proofs are deferred to the appendix.

2 Theoretical Setup

2.1 Poisoning Model

We consider a binary classification problem with inputs $\boldsymbol{x} \in \mathbb{R}^d$ and labels $y \in \{+1, -1\}$. The clean data form a symmetric Gaussian mixture with class means $\pm\boldsymbol{\mu}$ and common within-class covariance $\boldsymbol{\Sigma} \succ 0$. The attacker chooses a unit trigger direction $\boldsymbol{v} \in \mathbb{R}^d$ and a strength $\alpha \geq 0$, adds $\alpha\boldsymbol{v}$ to a fraction $\pi \in (0, 1)$ of the samples of the positive (source) class and flips their label to the negative (target) class. The training distribution is therefore the three-cluster mixture $\boldsymbol{x} = \boldsymbol{m}_c + \boldsymbol{z}$, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, with

$$\begin{array}{lllll} \mathcal{C}_+ \text{ (clean source)}: & \omega_+ = \frac{1-\pi}{2}, & \boldsymbol{m}_+ = \ \ \boldsymbol{\mu}, & y = +1, \\ \mathcal{C}_- \text{ (clean target)}: & \omega_- = \frac{1}{2}, & \boldsymbol{m}_- = -\boldsymbol{\mu}, & y = -1, \\ \mathcal{C}_p \text{ (poisoned)}: & \omega_p = \frac{\pi}{2}, & \boldsymbol{m}_p = \boldsymbol{\mu} + \alpha\boldsymbol{v}, & y = -1. \end{array} \tag{1}$$

No specific assumptions are placed on $\boldsymbol{\Sigma}$, on the alignment of $\boldsymbol{v}$ with $\boldsymbol{\mu}$, or on the dimension d. The model abstracts both centralized training, where an insider corrupts part of the data, and outsourced training, where the owner only receives the final weights (Li et al., 2022; Goldwasser et al., 2022); in both cases the adversary's cost is the pair (π, α), with π governing statistical stealth and α the visibility of the trigger.

The attacker's goal is a model that is accurate on clean inputs and misclassifies triggered inputs. For a trained classifier $\hat{y} = \operatorname{sign} f(\boldsymbol{x})$ we measure, on a clean source point $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{z}$, a clean target point $\boldsymbol{x} = -\boldsymbol{\mu} + \boldsymbol{z}$ and a triggered source point $\boldsymbol{x}^{\text{trig}} = \boldsymbol{\mu} + \alpha\boldsymbol{v} + \boldsymbol{z}$ (test noise $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$), the clean accuracy and the attack success rate are given by

$$\text{CACC} = \tfrac{1}{2}\Big[\mathbb{P}(f > 0 \mid \mathcal{C}_+) + \mathbb{P}(f < 0 \mid \mathcal{C}_-)\Big], \qquad \text{ASR} = \mathbb{P}\big(f(\boldsymbol{x}^{\text{trig}}) < 0\big). \tag{2}$$

Definition 1 (Budget curve)**.** For a target attack success rate $a \in (0, 1)$, the budget curve $\alpha(\pi; a)$ is the smallest trigger strength α such that $\text{ASR} \geq a$ at poison fraction π.

The budget curve tells how visible a trigger must be for an attack that corrupts a given fraction of the data. We are interested in its behavior as $\pi \to 0$, the regime of stealthy attacks, and in particular in the exponent γ in $\alpha(\pi; a) \propto \pi^{-\gamma}$.

2.2 The Quadratic Neuron and the Two Training Regimes

As the smallest model with a trainable nonlinear feature, we consider the quadratic neuron

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + \frac{\beta}{2}\,(\boldsymbol{u}^\top \boldsymbol{x})^2 + c, \qquad \boldsymbol{w}, \boldsymbol{u} \in \mathbb{R}^d, \quad \beta, c \in \mathbb{R}. \tag{3}$$

The vector $\boldsymbol{u}$ is the *feature direction*: the neuron computes the scalar feature $t = \boldsymbol{u}^\top \boldsymbol{x}$ and responds to it quadratically with curvature β. Since only $\beta\boldsymbol{u}\boldsymbol{u}^\top$ enters f, the scale of $\boldsymbol{u}$ is degenerate with

β and we fix it by $\boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u} = 1$, so that t has unit within-class variance. For a fixed $\boldsymbol{u}$, the read-out $\boldsymbol{\theta} := (\boldsymbol{w}, \beta, c)$ is trained by the population squared loss

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \ \mathcal{L}(\boldsymbol{\theta}), \qquad \mathcal{L}(\boldsymbol{\theta}) = \tfrac{1}{2}\,\mathbb{E}\big[(f(\boldsymbol{x}) - y)^2\big], \tag{4}$$

where the expectation is over the poisoned mixture (1). The map $\boldsymbol{\theta} \mapsto f$ is linear, so $\mathcal{L}$ is a strictly convex quadratic and $\boldsymbol{\theta}^\star$ is unique. What distinguishes a model that has learned its representation from one that has not is the feature direction $\boldsymbol{u}$, and we contrast the following two regimes.

Definition 2 (Linear regime: no feature learning)**.** The model is the linear classifier $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + c$, i.e. $\beta \equiv 0$ in (3), trained by (4).

This is the model implicitly assumed by lazy, kernel and random-feature analyses of poisoning: the representation is frozen and only a linear read-out is fitted. We show in Section 3.2 that it is exactly the special case $\boldsymbol{u}^\top \boldsymbol{v} = 0$ of (3), and that a quadratic neuron whose direction is frozen at a random initialization reduces to it in high dimension.

Definition 3 (Feature-learning regime)**.** The feature direction is adapted to the poisoned data. Among all directions with $\boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u} = 1$, the one that maximizes the feature–trigger overlap $\boldsymbol{u}^\top \boldsymbol{v}$ is the *whitened trigger*

$$\boldsymbol{u}^\star = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{v}}{\sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{v}}}, \qquad \text{since} \qquad |\boldsymbol{u}^\top \boldsymbol{v}| = |(\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\top (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v})| \le \sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{v}} \tag{5}$$

by the Cauchy–Schwarz inequality, with equality if and only if $\boldsymbol{u} \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{v}$. The feature-learning regime is the model (3) with $\boldsymbol{u} = \boldsymbol{u}^\star$ and the full read-out $(\boldsymbol{w}, \beta, c)$ trained by (4).

Fixing $\boldsymbol{u} = \boldsymbol{u}^\star$ is an idealization of feature learning that keeps the analysis exact. It is supported by two facts: the poison switches on the curvature β in proportion to $\boldsymbol{u}^\top \boldsymbol{v}$ (Section 3.3), and when all of $(\boldsymbol{w}, \beta, c, \boldsymbol{u})$ are trained by gradient descent the learned direction keeps an overlap $\boldsymbol{u}^\top \boldsymbol{v}$ between 0.6 and 0.95 of its maximal value (Appendix C.4). Our general result of Section 3.1 holds for every fixed $\boldsymbol{u}$, so partially aligned directions are covered as well.

Assumption 1. *The within-class noise is Gaussian with a fixed covariance $\boldsymbol{\Sigma} \succ 0$; the feature direction $\boldsymbol{u}$ is fixed with $\boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u} = 1$; the read-out $(\boldsymbol{w}, \beta, c)$ is the population least-squares minimizer (4). Gaussianity is the only assumption relaxed on real data in Section 4.*

3 Main Results

Since both the loss (4) and the sign of f are invariant under the whitening $\tilde{\boldsymbol{x}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$, $\tilde{\boldsymbol{w}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}$, $\tilde{\boldsymbol{u}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}$ (Lemma 2 in the appendix), the problem depends on $(\boldsymbol{\mu}, \boldsymbol{v}, \boldsymbol{\Sigma}, \boldsymbol{u})$ only through the geometry of the whitened vectors $\tilde{\boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$, $\tilde{\boldsymbol{v}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}$ and $\tilde{\boldsymbol{u}}$, which is captured by five scalars:

$$\underbrace{S = \sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}}_{\text{signal strength}}, \quad \underbrace{T = \sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{v}}}_{\text{trigger strength}}, \quad \underbrace{\rho = \frac{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{v}}{ST}}_{\text{signal–trigger alignment}}, \quad \underbrace{g = \boldsymbol{u}^\top \boldsymbol{\mu}}_{\text{feature–signal overlap}}, \quad \underbrace{h = \boldsymbol{u}^\top \boldsymbol{v}}_{\text{feature–trigger overlap}}. \tag{6}$$

S and T are the Mahalanobis norms of the class signal and of the trigger, i.e. their sizes in units of the within-class noise: $\Phi(S)$ is the accuracy of the optimal clean classifier, with Φ the standard normal distribution function, and T is large when the trigger points along directions in which the data already vary, a trigger "hidden in the noise". $\rho \in [-1, 1]$ is the cosine between signal and trigger in the whitened geometry, and $\rho = 0$ (a trigger Mahalanobis-orthogonal to the signal) is the natural setting for a hidden trigger. Finally $|g| \le S$ and $|h| \le T$; the feature-learning direction (5) has $h = T$ and $g = S\rho$, while the linear regime corresponds to $h = 0$. All results below are explicit functions of (S, T, ρ, g, h) and (π, α), and the dimension d enters only through these scalars. We write $A := \alpha T$ for the whitened trigger displacement, φ for the standard normal density and $\mathbb{E}_z$ for the expectation over $z \sim \mathcal{N}(0, 1)$.

3.1 Exact Clean Accuracy and Attack Success Rate

The first step decomposes the d-dimensional problem to three coordinates: the feature, the part of the signal not explained by the feature, and the part of the trigger explained by neither.

Lemma 1 (Reduction). *Let* $m_\perp = \sqrt{S^2 - g^2}$, $v_\parallel = (ST\rho - gh)/m_\perp$ *and* $v_\perp = \sqrt{T^2 - h^2 - v_\parallel^2}$ *(with* $v_\parallel = 0$ *if* $m_\perp = 0$*). There is an orthonormal frame* $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ *of whitened space with* $\boldsymbol{e}_1 = \tilde{\boldsymbol{u}}$ *such that the trained read-out satisfies* $\tilde{\boldsymbol{w}} \in \mathrm{span}(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$*, and in the coordinates* $(t, s_1, s_2) = (\boldsymbol{e}_1^\top \tilde{\boldsymbol{x}}, \boldsymbol{e}_2^\top \tilde{\boldsymbol{x}}, \boldsymbol{e}_3^\top \tilde{\boldsymbol{x}})$ *the three clusters of* (1) *have means*

$$\boldsymbol{m}_+ = (g,\ m_\perp,\ 0), \qquad \boldsymbol{m}_- = -\boldsymbol{m}_+, \qquad \boldsymbol{m}_p = (g + \alpha h,\ m_\perp + \alpha v_\parallel,\ \alpha v_\perp), \tag{7}$$

with independent standard Gaussian noise in each coordinate, and $f = w_t\, t + w_1 s_1 + w_2 s_2 + \frac{\beta}{2} t^2 + c$ *with* $\boldsymbol{\theta} = (w_t, w_1, w_2, \beta, c) \in \mathbb{R}^5$*. If* $\boldsymbol{u} \in \mathrm{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{v})$*, which includes both regimes of Definitions 2–3, then* $v_\perp = 0$ *and* $w_2 = 0$*.*

In this frame the model is linear in the feature vector $\boldsymbol{\phi} = (t, s_1, s_2, \frac{1}{2}t^2, 1)$, so the trained read-out solves the normal equations $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{b}$ with $\mathbf{A} = \mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top]$ and $\boldsymbol{b} = \mathbb{E}[\boldsymbol{\phi}\, y]$, whose entries are low-order Gaussian moments of the three clusters, hence explicit polynomials in the invariants and (π, α) (Appendix A.3). The second step is that f is quadratic in the noise only through the feature coordinate and linear in the two others, which can therefore be integrated out exactly.

Theorem 1 (General closed form). *Under Assumption 1, for any* $\boldsymbol{\Sigma} \succ 0$*, any* (π, α)*, any alignment* ρ *and any feature direction* $\boldsymbol{u}$*, let* $(w_t, w_1, w_2, \beta, c) = \mathbf{A}^{-1}\boldsymbol{b}$ *and* $\sigma = \sqrt{w_1^2 + w_2^2}$*. For a cluster mean* $\boldsymbol{m} = (m_t, m_1, m_2)$ *define* $A_0(\boldsymbol{m}) = w_t m_t + w_1 m_1 + w_2 m_2 + c + \frac{\beta}{2} m_t^2$ *and* $b(\boldsymbol{m}) = w_t + \beta\, m_t$*. Then*

$$\mathrm{CACC} = \tfrac{1}{2}\,\mathbb{E}_z\left[\Phi\left(\frac{A_0(\boldsymbol{m}_+) + b(\boldsymbol{m}_+)z + \frac{\beta}{2}z^2}{\sigma}\right)\right] + \tfrac{1}{2}\,\mathbb{E}_z\left[\Phi\left(-\frac{A_0(\boldsymbol{m}_-) + b(\boldsymbol{m}_-)z + \frac{\beta}{2}z^2}{\sigma}\right)\right],$$

$$\mathrm{ASR} = \mathbb{E}_z\left[\Phi\left(-\frac{A_0(\boldsymbol{m}_p) + b(\boldsymbol{m}_p)z + \frac{\beta}{2}z^2}{\sigma}\right)\right], \tag{8}$$

with the cluster means of (7) *and the convention* $\Phi(x/0) = \mathbf{1}\{x > 0\}$*.*

In essence, Theorem 1 expresses each metric as the Gaussian mass on one side of a parabola: $A_0(\boldsymbol{m})$ is the value of f at the cluster center, $b(\boldsymbol{m})z + \frac{\beta}{2}z^2$ is the fluctuation of the margin due to noise along the feature, which is quadratic because the neuron is, and σ is the standard deviation of the margin due to noise orthogonal to the feature, which is integrated out by Φ. When $\beta = 0$ the parabola is a line and each expectation collapses to a single Φ, the linear-classifier formula. Each metric is a one-dimensional integral evaluated by Gauss–Hermite quadrature from the five invariants. Two consequences hold for every feature direction.

Corollary 1. *For any* $\boldsymbol{u}$*, any* α *and any* ρ*:*

(i) ***(Poison-free baseline.)*** *At* $\pi = 0$ *the trained model is the linear classifier* $\tilde{\boldsymbol{w}} = \tilde{\boldsymbol{\mu}}/(1 + S^2)$*,* $\beta = c = 0$*; hence* $\mathrm{CACC} = \Phi(S)$ *and* $\mathrm{ASR} = \Phi(-S - \alpha T\rho)$*.*

(ii) ***(Stealth.)*** $\mathrm{CACC} = \Phi(S) + O(\pi^2)$*.*

Item (i) shows that a trigger that is not Mahalanobis-orthogonal to the signal ($\rho \neq 0$) is already seen by the clean classifier, and for $\rho < 0$ it drives the ASR to one at no poisoning cost; this is not a backdoor, and the asymptotic laws below are stated for the hidden trigger $\rho = 0$, for which the baseline ASR is $\Phi(-S)$. Item (ii) is the stealth property: the $O(\pi)$ change of the read-out shifts the margins of the two clean classes by opposite amounts (Appendix A.6), so the attack has no first-order accuracy signature in either regime.

3.2 The Linear Regime: No Feature Learning

We first specialize Theorem 1 to a model that does not learn its features, the regime behind lazy-training budget estimates.

Proposition 1. (i) *If* $h = \boldsymbol{u}^\top \boldsymbol{v} = 0$*, then* $\beta = 0$ *exactly for all* (π, α)*, and* $(\boldsymbol{w}, c)$ *coincide with the least-squares linear classifier of Definition 2.* (ii) *If* $\boldsymbol{u}$ *is frozen at a random direction (*$\tilde{\boldsymbol{u}}$ *uniform on the unit sphere of* $\mathbb{R}^d$*), then* $\mathbb{E}[h^2] = T^2/d$*; together with* $\beta = O(\pi h)$ *(Proposition 2), a lazy quadratic neuron in high dimension is the linear classifier up to* $O(d^{-1/2})$*.*

Part (i) states that a feature orthogonal to the trigger cannot detect it: with $h = 0$ the quadratic feature $\frac{1}{2}t^2$ has the same distribution in all three clusters and least squares assigns it zero weight. Part (ii) is the blindness of a frozen random feature in high dimension. In the linear regime the trained weights and the metrics have explicit single-Φ expressions (Corollary 3 in Appendix B), from which we obtain the budget curve.

Theorem 2 (Linear regime). *In the linear regime with a hidden trigger ($\rho = 0$):*

(a) ***(Small poison.)*** $\mathrm{CACC} = \Phi(S) + O(\pi^2)$ *and* $\mathrm{ASR} = \Phi(-S) + \pi\,\varphi(S)\Big[\dfrac{(1+2S^2)\,A^2}{2S} + \dfrac{1+S^2}{S}\Big] + O(\pi^2)$.

(b) ***(Stealth limit.)*** *As $\pi \to 0$ and $\alpha \to \infty$ with the trigger energy $e := \pi\alpha^2 T^2$ fixed,*

$$\mathrm{CACC} \to \Phi(S), \qquad \mathrm{ASR} \to \mathrm{ASR}_{\mathrm{lin}}(e) := \Phi\left(-S + \frac{1+2S^2}{S} \cdot \frac{e}{e+2}\right), \tag{9}$$

which increases from $\Phi(-S)$ at $e = 0$ to the ceiling $\Phi\big(\frac{1+S^2}{S}\big) < 1$ as $e \to \infty$.

(c) ***(Budget curve.)*** *For a target $a < \Phi\big(\frac{1+S^2}{S}\big)$, writing $\kappa_a := \dfrac{S\,(S + \Phi^{-1}(a))}{1+2S^2} \in [0, 1)$,*

$$\alpha_{\mathrm{lin}}(\pi; a) = \frac{1}{T}\sqrt{\frac{e_a}{\pi}} \;\propto\; \pi^{-1/2}, \qquad e_a = \frac{2\kappa_a}{1-\kappa_a}. \tag{10}$$

Essentially, Theorem 2 states that a model without feature learning sees the trigger only through the mean shift of the poisoned cluster: the read-out tilts toward $-\tilde{\boldsymbol{v}}$ by an amount proportional to $\pi\alpha$, and a triggered test point, displaced by the same $\alpha\tilde{\boldsymbol{v}}$, is pushed across the boundary by an amount proportional to $\pi\alpha^2$. Specifically, we draw the following observations:

- *Effect of the poison fraction π:* the only combination that survives the stealth limit is the trigger energy $e = \pi\alpha^2 T^2$ (recall that $A = \alpha T$), and the budget curve scales as $\pi^{-1/2}$.
- *Effect of the trigger geometry:* $\alpha_{\mathrm{lin}} \propto 1/T$, so a trigger along a direction of large within-class variance is cheaper, because in whitened units it is already large.
- *Effect of the target a:* the ASR saturates at $\Phi\big(\frac{1+S^2}{S}\big) < 1$: the clean classes pin the read-out along $\tilde{\boldsymbol{\mu}}$, and the residual noise along $\tilde{\boldsymbol{\mu}}$ lets a fraction $\Phi\big(-\frac{1+S^2}{S}\big)$ of triggered points escape whatever the trigger strength.

3.3 The Feature-Learning Regime

We now let the neuron use its quadratic feature. The first result identifies what the poison does to the read-out for an arbitrary feature direction.

Proposition 2 (The poison creates a trigger detector). *For any feature direction $\boldsymbol{u}$, expanding $\boldsymbol{\theta}^\star = \mathbf{A}^{-1}\boldsymbol{b}$ to first order in π gives $\tilde{\boldsymbol{w}} = \tilde{\boldsymbol{\mu}}/(1+S^2) + O(\pi)$, $c = O(\pi)$ and*

$$\beta = -\,\frac{\pi\,\alpha h\,\big[ST\rho\,h\,\alpha^2 + \big((2S^2+1)h + 2ST\rho\,g\big)\alpha + 4g(S^2+1)\big]}{2(S^2+1)(2g^2+1)} + O(\pi^2). \tag{11}$$

In the feature-learning regime with a hidden trigger ($h = T$, $g = 0$, $\rho = 0$) this reduces to $\beta = -\pi\,\frac{1+2S^2}{2(1+S^2)}\,A^2 + O(\pi^2)$.

Three facts follow from (11). First, $\beta = O(\pi)$ vanishes at $\pi = 0$: the clean model has no quadratic unit (Corollary 1) and it is the poison that switches one on; its sign is negative, so the unit fires on large $|t|$ and routes such inputs to the target class. Second, $\beta \propto h = \boldsymbol{u}^\top\boldsymbol{v}$: this is why the linear regime ($h = 0$) has no detector, why a frozen random feature ($h \sim d^{-1/2}$) loses it in high dimension, and why feature learning should select the direction that maximizes h. Third, in the feature-learning regime $\beta \propto \pi A^2$ grows with the square of the trigger displacement, whereas the linear read-out grows linearly, so on a triggered point, whose feature coordinate is $t \approx A$, the detector contributes

Table 1: Linear regime versus feature-learning regime (hidden trigger, $\rho = 0$; $A = \alpha T$; $\kappa_a = S(S + \Phi^{-1}(a))/(1 + 2S^2)$). The two regimes share the clean accuracy, the form of the ASR law and the stealth ceiling; they differ in the conserved trigger energy and hence in the cost exponent.

	Linear regime (Thm. 2)	**Feature learning** (Thm. 3)
Trained curvature	$\beta \equiv 0$	$\beta = -\pi \frac{1+2S^2}{2(1+S^2)} A^2 + O(\pi^2)$
Trigger term in the margin	$\propto \pi A^2$	$\frac{\beta}{2} A^2 \propto \pi A^4$
Conserved invariant	$e = \pi \alpha^2 T^2$	$\varrho = \pi \alpha^4 T^4$
Stealth-limit ASR	$\Phi\big(-S + \frac{1+2S^2}{S} \frac{e}{e+2}\big)$	$\Phi\big(-S + \frac{1+2S^2}{S} \frac{\varrho}{4+\varrho}\big)$
Ceiling ($\pi \to 0$)	$\Phi\big(\frac{1+S^2}{S}\big)$	$\Phi\big(\frac{1+S^2}{S}\big)$
Budget curve $\alpha(\pi; a)$	$\frac{1}{T}\big(\frac{2\kappa_a}{(1-\kappa_a)\pi}\big)^{1/2}$	$\frac{1}{T}\big(\frac{4\kappa_a}{(1-\kappa_a)\pi}\big)^{1/4}$
Clean accuracy	$\Phi(S) + O(\pi^2)$	$\Phi(S) + O(\pi^2)$

$\frac{\beta}{2} A^2 \propto \pi A^4$ to the margin. The exact finite-π weights (Corollary 4 in Appendix C) show that their common denominator $\Delta = 16 \det \mathbf{A}$ grows as πA^4, so $\beta = O(A^{-2})$ as $A \to \infty$ (the detector saturates) and πA^4 is the only combination that stays finite when $\pi \to 0$ and $A \to \infty$.

Theorem 3 (Feature-learning regime)**.** *In the feature-learning regime with a hidden trigger ($\boldsymbol{u} = \boldsymbol{u}^\star$, $\rho = 0$):*

(a) ***(Small poison.)*** $\mathrm{CACC} = \Phi(S) + O(\pi^2)$ *and*

$$\mathrm{ASR} = \Phi(-S) + \pi\, \varphi(S) \left[\frac{(1 + 2S^2)\, A^2 (A^2 + 2)}{4S} + \frac{1 + S^2}{S} \right] + O(\pi^2), \tag{12}$$

which exceeds the expansion of Theorem 2(a) by exactly $\pi \varphi(S)\, (1 + 2S^2) A^4/(4S)$.

(b) ***(Stealth limit.)*** *As $\pi \to 0$ and $\alpha \to \infty$ with the quartic trigger energy $\varrho := \pi \alpha^4 T^4$ fixed,*

$$\mathrm{CACC} \to \Phi(S), \qquad \mathrm{ASR} \to \mathrm{ASR}_{\mathrm{fl}}(\varrho) := \Phi\left(-S + \frac{1 + 2S^2}{S} \cdot \frac{\varrho}{4 + \varrho} \right). \tag{13}$$

(c) ***(Ceiling.)*** *At any finite π and for any alignment ρ, $\lim_{\alpha \to \infty} \mathrm{ASR} = \Phi\Big(\frac{1 + S^2}{S} + \frac{\pi}{2S(1 - \pi)} \Big) < 1$, independent of T and ρ.*

(d) ***(Budget curve.)*** *For a target $a < \Phi\big(\frac{1+S^2}{S}\big)$, with κ_a as in Theorem 2,*

$$\alpha_{\mathrm{fl}}(\pi; a) = \frac{1}{T} \left(\frac{\varrho_a}{\pi} \right)^{1/4} \propto \pi^{-1/4}, \qquad \varrho_a = \frac{4\kappa_a}{1 - \kappa_a}. \tag{14}$$

Table 1 places the two regimes side by side and Figure 2 plots the exact curves of Theorem 1 in both. Particularly, Theorem 3 shows that the two regimes obey the same law in a different currency. The stealth-limit laws (9) and (13) have an identical form: the margin argument moves from the clean error level $-S$ toward the common ceiling $\frac{1+S^2}{S}$ by the same total budget $\frac{1+2S^2}{S}$, multiplied by a saturating fraction in $[0, 1)$. What differs is the variable inside that fraction: the linear model spends the quadratic energy $e = \pi \alpha^2 T^2$, the feature-learning model spends the quartic energy $\varrho = \pi \alpha^4 T^4$. The reason is visible in the small-poison expansions, which agree except for the term $\pi \varphi(S)(1 + 2S^2) A^4/(4S)$ in (12): the detector responds to the trigger as $(\boldsymbol{u}^{\star\top} \boldsymbol{x})^2 \sim A^2$ and least squares scales its curvature by another πA^2. Specifically, we draw the following observations:

- *Effect of the poison fraction π:* conservation of ϱ instead of e halves the cost exponent, from $\pi^{-1/2}$ to $\pi^{-1/4}$. Dividing (10) by (14), the advantage at a fixed target is

$$\frac{\alpha_{\mathrm{lin}}(\pi; a)}{\alpha_{\mathrm{fl}}(\pi; a)} = \left(\frac{\kappa_a}{(1 - \kappa_a)\, \pi} \right)^{1/4}, \tag{15}$$

  which grows without bound as the attack becomes stealthier.

- *Effect of the target a:* feature learning does not raise the ceiling, it reaches it at a lower cost. In the stealth limit both regimes saturate at $\Phi\big(\frac{1+S^2}{S}\big)$, and at finite π the ceiling of Theorem 3(c) exceeds it only by $\pi/(2S(1-\pi))$: as $A \to \infty$ the curvature decays as $\beta = O(A^{-2})$, the triggered statistic becomes linear-Gaussian again, and the ceiling is set by the clean classes as in the linear regime.
- *Effect of the geometry (T, ρ) and on clean accuracy:* the ceiling depends on neither T nor ρ, in both regimes $\alpha \propto 1/T$, and $\mathrm{CACC} = \Phi(S) + O(\pi^2)$, so the amplification comes at no cost in clean accuracy and leaves no first-order signature for a defender to detect.

We point out that an auditor who models the adversary with a lazy learner concludes that a poison fraction π needs a trigger of strength $\alpha_{\mathrm{lin}}(\pi; a)$; if the audited model has learned its features, the true requirement is smaller by the factor (15), an order of magnitude at the poison fractions of large-scale training. This is the theoretical counterpart of the observation of Souly et al. (2025) that the number of poisoned documents needed to backdoor a language model does not grow with the corpus.

3.4 Partially Learned Features

Theorem 1 also covers models that have learned their features only partially, $0 < h < T$. The trigger then splits into a component αh along the feature, which drives the quadratic detector (invariant $\pi(\alpha h)^4$), and a component $\alpha v_\perp$, $v_\perp := \sqrt{T^2 - h^2}$, orthogonal to it, which drives the linear channel of Section 3.2 (invariant $\pi\alpha^2 v_\perp^2$); the attacker pays whichever is cheaper.

Corollary 2 (Two channels)**.** *Let $\rho = 0$. The feature-learning channel ($\alpha \propto \pi^{-1/4}$) dominates for $h \gtrsim T\pi^{1/4}$ and the linear channel ($\alpha \propto \pi^{-1/2}$) below it. In the stealth regime the two laws combine into the generalized energy $G = 2\pi\alpha^2 v_\perp^2 + \pi\alpha^4 h^4$,* $\mathrm{ASR} \simeq \Phi\big(-S + \frac{1+2S^2}{S}\frac{G}{4+G}\big)$*, whose inversion (Appendix D) gives a closed-form $\alpha(\pi; a, h)$ that is exact at both ends, $h \to 0$ and $h = T$.*

Figure 4 in Appendix D shows the resulting phase diagram. The crossover $h = T\pi^{1/4}$ moves to smaller overlaps as the poison becomes rarer, so an auditor who cannot inspect the learned feature direction must assume that the cheaper channel is active.

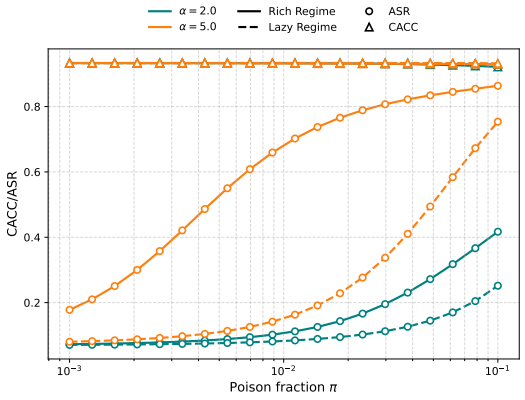


Figure 2: **Exact theory in the two regimes.** CACC *(triangles) and* ASR *(circles) from Theorem 1 as functions of π, for $\alpha = 2$ and $\alpha = 5$, in the feature-learning regime ($\boldsymbol{u} = \boldsymbol{u}^\star$, solid) and the linear regime ($\beta \equiv 0$, dashed); $S = 1.5$, $T = 1$, $\rho = 0$. In both regimes the clean accuracy stays at $\Phi(S) \approx 0.93$, while the attack success rate is uniformly higher with feature learning and, at $\alpha = 5$, exceeds 0.5 at a poison fraction roughly ten times smaller than in the linear regime, consistent with (15).*

4 Experiments

Setup. We evaluate the theory on a synthetic Gaussian mixture and on three image benchmarks, MNIST (3 vs. 8), Fashion-MNIST (Pullover vs. Coat) and CIFAR-10 (Ship vs. Bird). For the image datasets we estimate the class means $\pm\hat{\boldsymbol{\mu}}$ and the pooled within-class covariance $\hat{\boldsymbol{\Sigma}}$ (with a ridge $10^{-4}\boldsymbol{I}$) and extract the invariants (6), so that all predictions are parameter-free. To simulate a stealthy adversary, the trigger $\boldsymbol{v}$ is the principal eigenvector of $\hat{\boldsymbol{\Sigma}}$, the direction of maximal within-class variance, and the feature direction is $\boldsymbol{u}^\star = \hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{v}/T$. We compare the *feature-learning regime* (trained β) and the *linear regime* ($\beta = 0$). *Practice* markers are obtained by sampling $n_{\mathrm{train}} = 10{,}000$ points from the Gaussian mixture with parameters $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, poisoning a fraction π of the source class, solving the finite-sample least-squares problem and measuring the ASR on $n_{\mathrm{test}} = 5{,}000$ held-out samples; *Theory* lines are evaluated from Theorem 1. Empirically, for each

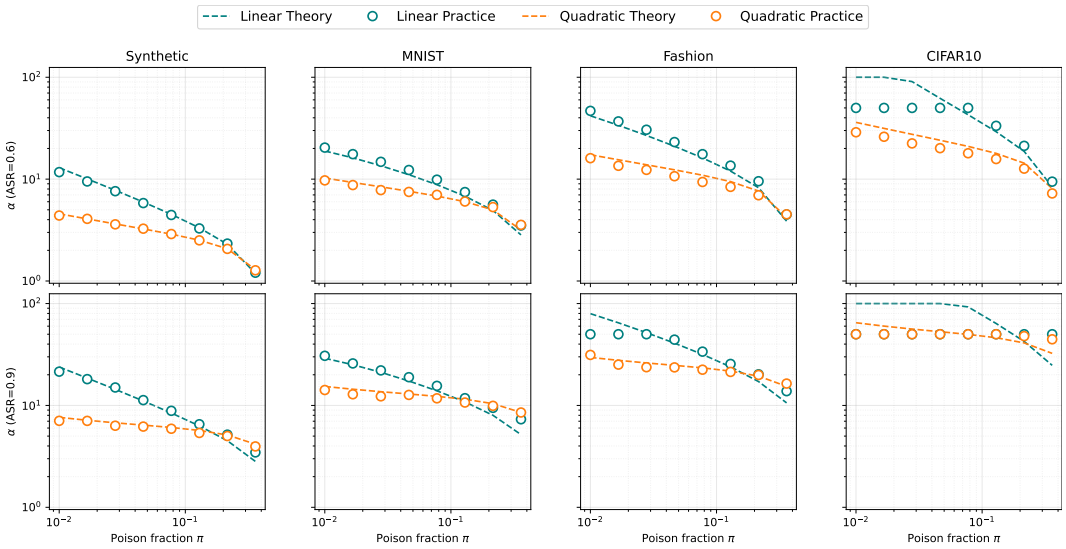


Figure 3: **Theory versus practice on synthetic and real data.** *Trigger strength α required to reach* $\mathrm{ASR} = 0.6$ *(top) and* $\mathrm{ASR} = 0.9$ *(bottom) as a function of π. Markers are trained models (feature-learning regime in red, linear regime in blue) and lines the predictions of Theorem 1 on the invariants extracted from the data. The quadratic neuron follows the $\alpha \propto \pi^{-1/4}$ law, the linear model the steeper $\alpha \propto \pi^{-1/2}$ law.*

π on a logarithmic grid from 10^{-2} to 0.36 we find by binary search the trigger strength that reaches $\mathrm{ASR} = 0.6$ and $\mathrm{ASR} = 0.9$, and fit $\log \alpha = k \log \pi + b$ to the resulting budget curves.

Results. Figure 3 shows the budget curves. On all four datasets the empirical quadratic neuron tracks the theoretical curve, and the fitted slopes are $k \approx -\frac{1}{4}$ in the feature-learning regime and $k \approx -\frac{1}{2}$ in the linear regime, as predicted by Theorems 2 and 3; the gap between the two regimes widens as π decreases, in agreement with (15). Table 2 in Appendix E reports CACC and ASR on a grid of (π, α): in both regimes the clean accuracy is unchanged by the poison, while the attack success rate is uniformly larger with feature learning (on MNIST, 0.97 against 0.15 at $\pi = 0.01$, $\alpha = 1$). Finally, Figure 1 shows that the separation is not specific to the quadratic neuron: a network trained in the rich feature learning regime requires a smaller trigger than the same network trained in the lazy regime.

5 Discussions & Conclusion

In this paper, we studied backdoor poisoning through a statistical model that isolates the role of feature learning: a quadratic neuron trained by least squares on a poisoned Gaussian mixture. We derived exact expressions for the clean accuracy and the attack success rate, valid for arbitrary noise covariance, finite poison fraction and trigger strength, and used them to show that a model without feature learning obeys the budget curve $\alpha \propto \pi^{-1/2}$, whereas a model that learns its feature obeys $\alpha \propto \pi^{-1/4}$: the poison creates a quadratic trigger detector whose response scales as α^4, which halves the cost exponent without raising the attack ceiling and without affecting the clean accuracy. Our results offer a mechanism for the observation that the number of poisoned samples needed to backdoor a large model does not grow with its training corpus, and imply that audits calibrated on lazy-regime estimates underestimate the adversary by a factor that grows as $\pi^{-1/4}$ in our considered setting. Extending the analysis to a hidden layer with a general activation, to finite-sample training and to trained feature directions is left for future work, and could potentially bring more understanding on the inner interaction between backdoor attacks and neural feature learning.

REFERENCES

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946, 2022.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning*, pp. 1467–1474. PMLR, 2012.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in neural information processing systems*, volume 32, 2019.

Dami Choi, Yonadav Shavit, and David K Duvenaud. Tools for verifying neural models' training data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1154–1188, 2023.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 113–125, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. In *The Annals of Statistics*, volume 49, pp. 1029–1054, 2021.

Shafi Goldwasser, Michael P Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. *arXiv preprint arXiv:2204.06974*, 2022.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Steve Hanneke, Amin Karbasi, Mohammad Mahmoody, Idan Mehalel, and Shay Moran. On optimal learning under targeted data poisoning. *Advances in Neural Information Processing Systems*, 35: 30770–30782, 2022.

Jonathan Hayase and Sewoong Oh. Few-shot backdoor attacks via neural tangent kernels. *arXiv preprint arXiv:2210.05929*, 2022.

Jonathan Hayase, Weihao Kong, Raghav Bhatt, and Sewoong Oh. Spectre: Defending against backdoor attacks using robust covariance estimation. In *International Conference on Machine Learning*, pp. 4129–4139. PMLR, 2021.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, volume 31, 2018.

Hengrui Jia, Mohammad Yaghini, Christopher A Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1039–1056. IEEE, 2021.

Kassem Kallas, Carine Tannous, and Hichem FARAOUN. A game-theoretic approach to cost-aware backdoor attacks and defenses in deep learning. 2025.

Alaa Khaddaj, Guillaume Leclerc, Aleksandar Makelov, Kristian Georgiev, Hadi Salman, Andrew Ilyas, and Aleksander Madry. Rethinking backdoor attacks. In *International Conference on Machine Learning*, pp. 16216–16236. PMLR, 2023.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14900–14912, 2021.

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, 2022.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Network and Distributed System Security Symposium (NDSS)*, 2018.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

N Subramanian Manoj and Avrim Blum. Excess capacity and backdoor certification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20323–20334, 2021.

Song Mei and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

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

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, volume 20, 2007.

Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11957–11965, 2020.

Issam Seddik, Sami Souihi, Mohamed Tamaazousti, and Sara Tucci Piergiovanni. Pots: Proof-of-training-steps for backdoor detection in large language models. In *International Conference on Foundation and Large Language Models (FLLM)*, pp. 7–15. IEEE, 2025.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitorf, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, et al. Poisoning attacks on llms require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. In *NeurIPS Workshop on Security in Machine Learning*, 2019.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy*, pp. 707–723. IEEE, 2019.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Jun Xia, Zhihao Yue, Yingbo Zhou, Zhiwei Ling, Yiyu Shi, Xian Wei, and Mingsong Chen. Waveattack: Asymmetric frequency obfuscation-based backdoor attacks against deep neural networks. *Advances in Neural Information Processing Systems*, 37:43549–43570, 2024.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11653–11664. PMLR, 2021.

Rui Zhu, Di Tang, Siyuan Tang, XiaoFeng Wang, and Haixu Tang. Selective amnesia: On efficient, high-fidelity and blind suppression of backdoor effects in trojaned machine learning models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1–19. IEEE, 2023.

A Proof of Theorem 1 and of Corollary 1

Throughout, Φ and φ are the standard normal distribution and density, and $\mathbb{E}_z$ denotes expectation over $z \sim \mathcal{N}(0,1)$.

A.1 Whitening

Lemma 2 (Whitening equivariance). *Let $\tilde{\boldsymbol{x}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$, $\tilde{\boldsymbol{w}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}$, $\tilde{\boldsymbol{u}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}$. Then $f(\boldsymbol{x}) = \tilde{\boldsymbol{w}}^\top\tilde{\boldsymbol{x}} + \frac{\beta}{2}(\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{x}})^2 + c$ for all $\boldsymbol{x}$, the whitened training distribution is the mixture (1) with $(\boldsymbol{\mu}, \boldsymbol{v}, \boldsymbol{\Sigma})$ replaced by $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{v}}, \boldsymbol{I})$, $\tilde{\boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$, $\tilde{\boldsymbol{v}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}$, and $\|\tilde{\boldsymbol{u}}\| = 1$. The population loss (4), its minimiser (expressed in the tilded variables), CACC and ASR are all unchanged. Moreover $S = \|\tilde{\boldsymbol{\mu}}\|$, $T = \|\tilde{\boldsymbol{v}}\|$, $\rho = \tilde{\boldsymbol{\mu}}^\top\tilde{\boldsymbol{v}}/(ST)$, $g = \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{\mu}}$ and $h = \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}}$.*

Proof. $\boldsymbol{w}^\top\boldsymbol{x} = (\boldsymbol{\Sigma}^{1/2}\boldsymbol{w})^\top(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}) = \tilde{\boldsymbol{w}}^\top\tilde{\boldsymbol{x}}$ and likewise $\boldsymbol{u}^\top\boldsymbol{x} = \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{x}}$, which gives the identity for f. If $\boldsymbol{x} = \boldsymbol{m}_c + \boldsymbol{z}$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ then $\tilde{\boldsymbol{x}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}_c + \tilde{\boldsymbol{z}}$ with $\tilde{\boldsymbol{z}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}_c \in \{\tilde{\boldsymbol{\mu}}, -\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}} + \alpha\tilde{\boldsymbol{v}}\}$. The map $\boldsymbol{w} \mapsto \tilde{\boldsymbol{w}}$ is a bijection, so minimising the loss over $\boldsymbol{w}$ or over $\tilde{\boldsymbol{w}}$ is the same problem, and the sign of f is pointwise unchanged, hence the metrics are. Finally $\|\tilde{\boldsymbol{u}}\|^2 = \boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{u} = 1$, $\|\tilde{\boldsymbol{\mu}}\|^2 = \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = S^2$, $\|\tilde{\boldsymbol{v}}\|^2 = T^2$, $\tilde{\boldsymbol{\mu}}^\top\tilde{\boldsymbol{v}} = \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{v} = ST\rho$, and $\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{\mu}} = (\boldsymbol{\Sigma}^{1/2}\boldsymbol{u})^\top(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}) = \boldsymbol{u}^\top\boldsymbol{\mu} = g$, $\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{v}} = \boldsymbol{u}^\top\boldsymbol{v} = h$. □

By Lemma 2 we work in whitened coordinates from now on and drop the tildes: the noise is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the signal is $\tilde{\boldsymbol{\mu}}$ with $\|\tilde{\boldsymbol{\mu}}\| = S$, the trigger is $\tilde{\boldsymbol{v}}$ with $\|\tilde{\boldsymbol{v}}\| = T$, the feature direction $\tilde{\boldsymbol{u}}$ is a unit vector, and the three cluster means are $\tilde{\boldsymbol{\mu}}$, $-\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\mu}} + \alpha\tilde{\boldsymbol{v}}$.

A.2 Proof of Lemma 1 (reduction to a three-dimensional frame)

The read-out lies in the span of signal, trigger and feature. Let $W = \operatorname{span}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{u}})$ and let $\boldsymbol{q} \in W^\perp$ be arbitrary. For fixed (β, c) the stationarity condition of (4) with respect to $\tilde{\boldsymbol{w}}$ reads

$$\mathbb{E}\big[\tilde{\boldsymbol{x}}\,(\tilde{\boldsymbol{w}}^\top\tilde{\boldsymbol{x}} + \tfrac{\beta}{2}t^2 + c - y)\big] = \boldsymbol{0}, \qquad t = \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{x}}. \tag{16}$$

Take the inner product of (16) with $\boldsymbol{q}$. Write $\tilde{\boldsymbol{x}} = \boldsymbol{m}_c + \tilde{\boldsymbol{z}}$ for the cluster c of the sample. Since $\boldsymbol{q} \perp \boldsymbol{m}_c$, $\boldsymbol{q}^\top\tilde{\boldsymbol{x}} = \boldsymbol{q}^\top\tilde{\boldsymbol{z}}$, and since $\boldsymbol{q} \perp \tilde{\boldsymbol{u}}$, the Gaussian variables $\boldsymbol{q}^\top\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{z}}$ are uncorrelated hence independent. The bracket in (16) depends on the sample only through $(c,\ \tilde{\boldsymbol{w}}^\top\tilde{\boldsymbol{z}},\ \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{z}})$; decompose $\tilde{\boldsymbol{w}} = \tilde{\boldsymbol{w}}_W + \tilde{\boldsymbol{w}}_\perp$ with $\tilde{\boldsymbol{w}}_W \in W$, $\tilde{\boldsymbol{w}}_\perp \in W^\perp$. Then

$$\begin{aligned}\boldsymbol{q}^\top\mathbb{E}\big[\tilde{\boldsymbol{x}}(\cdots)\big] &= \mathbb{E}\big[\boldsymbol{q}^\top\tilde{\boldsymbol{z}}\,(\tilde{\boldsymbol{w}}_\perp^\top\tilde{\boldsymbol{z}})\big] + \mathbb{E}\big[\boldsymbol{q}^\top\tilde{\boldsymbol{z}}\ \underbrace{(\tilde{\boldsymbol{w}}_W^\top\tilde{\boldsymbol{x}} + \tfrac{\beta}{2}t^2 + c - y)}_{\text{independent of } \boldsymbol{q}^\top\tilde{\boldsymbol{z}}}\big] \\ &= \boldsymbol{q}^\top\tilde{\boldsymbol{w}}_\perp + 0,\end{aligned}$$

because $\mathbb{E}[\tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}^\top] = \boldsymbol{I}$ and $\mathbb{E}[\boldsymbol{q}^\top\tilde{\boldsymbol{z}}] = 0$. Hence $\boldsymbol{q}^\top\tilde{\boldsymbol{w}}_\perp = 0$ for every $\boldsymbol{q} \in W^\perp$, i.e. $\tilde{\boldsymbol{w}}_\perp = \boldsymbol{0}$ and $\tilde{\boldsymbol{w}} \in W$. Consequently f depends on $\tilde{\boldsymbol{x}}$ only through its orthogonal projection onto W, and minimising (4) over $\tilde{\boldsymbol{w}} \in \mathbb{R}^d$ is the same as minimising over $\tilde{\boldsymbol{w}} \in W$.

The frame. Let $\boldsymbol{e}_1 = \tilde{\boldsymbol{u}}$. If $m_\perp = \sqrt{S^2 - g^2} > 0$, let $\boldsymbol{e}_2 = (\tilde{\boldsymbol{\mu}} - g\tilde{\boldsymbol{u}})/m_\perp$; this is a unit vector orthogonal to $\boldsymbol{e}_1$ because $\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{\mu}} = g$ and $\|\tilde{\boldsymbol{\mu}} - g\tilde{\boldsymbol{u}}\|^2 = S^2 - 2g^2 + g^2 = m_\perp^2$. Let $\boldsymbol{e}_3$ be a unit vector in W orthogonal to $\boldsymbol{e}_1, \boldsymbol{e}_2$ (if $\dim W < 3$, complete with any unit vector orthogonal to W; the corresponding coordinate then has zero mean in every cluster and, by the argument above applied to W, zero weight). In this frame $\tilde{\boldsymbol{\mu}} = g\boldsymbol{e}_1 + m_\perp\boldsymbol{e}_2$ by construction. Writing $\tilde{\boldsymbol{v}} = h\boldsymbol{e}_1 +$

$v_\parallel \boldsymbol{e}_2 + v_\perp \boldsymbol{e}_3$, we have $h = \tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{v}}$, $v_\parallel = \boldsymbol{e}_2^\top \tilde{\boldsymbol{v}} = (\tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{v}} - g\, \tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{v}})/m_\perp = (ST\rho - gh)/m_\perp$, and $v_\perp^2 = \|\tilde{\boldsymbol{v}}\|^2 - h^2 - v_\parallel^2 = T^2 - h^2 - v_\parallel^2$; choosing the orientation of $\boldsymbol{e}_3$ makes $v_\perp \geq 0$. If $m_\perp = 0$ then $\tilde{\boldsymbol{\mu}} = \pm S\tilde{\boldsymbol{u}}$ and we take $\boldsymbol{e}_2, \boldsymbol{e}_3$ any orthonormal completion with $\boldsymbol{e}_3 \propto \tilde{\boldsymbol{v}} - h\tilde{\boldsymbol{u}}$, so that $v_\parallel = 0$ and $v_\perp = \sqrt{T^2 - h^2}$. The cluster means (7) follow by reading $\tilde{\boldsymbol{\mu}}$, $-\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\mu}} + \alpha\tilde{\boldsymbol{v}}$ in this basis. Since the frame is orthonormal and the noise is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the three noise coordinates are independent standard Gaussians. Finally $f = \tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{x}} + \frac{\beta}{2}t^2 + c$ with $\tilde{\boldsymbol{w}} = w_t \boldsymbol{e}_1 + w_1 \boldsymbol{e}_2 + w_2 \boldsymbol{e}_3$ is Lemma 1.

The in-plane case. If $\boldsymbol{u} \in \operatorname{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{v})$ then $\tilde{\boldsymbol{u}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{u} \in \operatorname{span}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{v}})$, so W is two-dimensional, $\boldsymbol{e}_3 \perp W$, $v_\perp = 0$, and $w_2 = 0$ by the projection argument. Both regimes are in-plane: the linear regime does not use $\boldsymbol{u}$ at all (equivalently any $\boldsymbol{u}$ with $h = 0$ may be used, see Appendix B), and $\boldsymbol{u}^\star \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{v}$. □

A.3 Population moments

In the frame of Lemma 1 the model is linear in $\boldsymbol{\phi} = (t,\, s_1,\, s_2,\, \frac{1}{2}t^2,\, 1)^\top$, so the minimiser of (4) solves

$$\mathbf{A}\,\boldsymbol{\theta} = \boldsymbol{b}, \qquad \mathbf{A} = \mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top] = \sum_{c\in\{+,-,p\}} \omega_c\, \mathbf{M}(\boldsymbol{m}_c), \qquad \boldsymbol{b} = \mathbb{E}[\boldsymbol{\phi}\, y] = \sum_{c\in\{+,-,p\}} \omega_c\, y_c\, \bar{\boldsymbol{\phi}}(\boldsymbol{m}_c), \tag{17}$$

where, for a cluster with mean $\boldsymbol{m} = (m_t, m_1, m_2)$ and unit isotropic noise,

$$\bar{\boldsymbol{\phi}}(\boldsymbol{m}) = \begin{pmatrix} m_t \\ m_1 \\ m_2 \\ \frac{1}{2}(m_t^2+1) \\ 1 \end{pmatrix}, \quad \mathbf{M}(\boldsymbol{m}) = \begin{pmatrix} m_t^2+1 & m_t m_1 & m_t m_2 & \frac{1}{2}(m_t^3+3m_t) & m_t \\ \cdot & m_1^2+1 & m_1 m_2 & \frac{1}{2}m_1(m_t^2+1) & m_1 \\ \cdot & \cdot & m_2^2+1 & \frac{1}{2}m_2(m_t^2+1) & m_2 \\ \cdot & \cdot & \cdot & \frac{1}{4}(m_t^4+6m_t^2+3) & \frac{1}{2}(m_t^2+1) \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix} \tag{18}$$

(symmetric). Indeed, for a single cluster with mean $\boldsymbol{m} = (m_t, m_1, m_2)$ and independent unit-variance coordinates, the moments of $\boldsymbol{\phi}$ follow from $\mathbb{E}z = \mathbb{E}z^3 = 0$, $\mathbb{E}z^2 = 1$, $\mathbb{E}z^4 = 3$ for $z \sim \mathcal{N}(0,1)$:

$$\mathbb{E}t^2 = m_t^2 + 1, \qquad \mathbb{E}t^3 = m_t^3 + 3m_t, \qquad \mathbb{E}t^4 = m_t^4 + 6m_t^2 + 3,$$
$$\mathbb{E}[t s_j] = m_t m_j, \qquad \mathbb{E}[s_j \tfrac{1}{2}t^2] = \tfrac{1}{2}m_j(m_t^2+1), \qquad \mathbb{E}[s_1 s_2] = m_1 m_2,$$

which gives $\bar{\boldsymbol{\phi}}(\boldsymbol{m})$ and $\mathbf{M}(\boldsymbol{m})$ in (18). Summing over the three clusters with weights ω_c and labels y_c gives $\mathbf{A}$ and $\boldsymbol{b}$ in (17). For a monomial $q(\boldsymbol{m})$ write $\langle q \rangle := \sum_c \omega_c q(\boldsymbol{m}_c)$; because the clean means are $\pm\boldsymbol{m}_+$, an even monomial in $\boldsymbol{m}$ contributes $\frac{2-\pi}{2}q(\boldsymbol{m}_+) + \frac{\pi}{2}q(\boldsymbol{m}_p)$ and an odd one $-\frac{\pi}{2}q(\boldsymbol{m}_+) + \frac{\pi}{2}q(\boldsymbol{m}_p)$. Writing $\nu := g + \alpha h$ for the feature mean of the poisoned cluster, the sums needed are

$$\begin{aligned}
\langle m_t \rangle &= \tfrac{\pi\alpha h}{2}, \qquad \langle m_1 \rangle = \tfrac{\pi\alpha v_\parallel}{2}, \qquad \langle m_2 \rangle = \tfrac{\pi\alpha v_\perp}{2}, \qquad \langle m_t^2 \rangle = g^2 + \tfrac{\pi}{2}(\alpha^2 h^2 + 2\alpha g h), \\
\langle m_1^2 \rangle &= m_\perp^2 + \tfrac{\pi}{2}(\alpha^2 v_\parallel^2 + 2\alpha m_\perp v_\parallel), \qquad \langle m_2^2 \rangle = \tfrac{\pi}{2}\alpha^2 v_\perp^2, \qquad \langle m_t m_2 \rangle = \tfrac{\pi}{2}\alpha v_\perp \nu, \\
\langle m_t m_1 \rangle &= g m_\perp + \tfrac{\pi}{2}(\alpha^2 h v_\parallel + \alpha g v_\parallel + \alpha h m_\perp), \qquad \langle m_1 m_2 \rangle = \tfrac{\pi}{2}\alpha v_\perp (m_\perp + \alpha v_\parallel), \\
\langle m_t^3 \rangle &= \tfrac{\pi\alpha h}{2}(\alpha^2 h^2 + 3\alpha g h + 3g^2), \qquad \langle m_2 m_t^2 \rangle = \tfrac{\pi}{2}\alpha v_\perp \nu^2, \\
\langle m_t^4 \rangle &= g^4 + \tfrac{\pi\alpha h}{2}(\alpha^3 h^3 + 4\alpha^2 g h^2 + 6\alpha g^2 h + 4g^3), \\
\langle m_1 m_t^2 \rangle &= \tfrac{\pi\alpha}{2}(\alpha^2 h^2 v_\parallel + 2\alpha g h v_\parallel + \alpha h^2 m_\perp + g^2 v_\parallel + 2g h m_\perp).
\end{aligned}$$

Substituting into (18), the entries of $\mathbf{A}$ (symmetric) in the order $(t, s_1, s_2, \frac{1}{2}t^2, 1)$ are

$$\begin{aligned}
&A_{11} = g^2 + \tfrac{\pi}{2}(\alpha^2 h^2 + 2\alpha g h) + 1, \qquad A_{12} = g\, m_\perp + \tfrac{\pi}{2}(\alpha^2 h v_\parallel + \alpha g v_\parallel + \alpha h m_\perp), \qquad A_{13} = \tfrac{\pi}{2}\alpha v_\perp \nu, \\
&A_{14} = \tfrac{\pi\alpha h}{4}\big(\alpha^2 h^2 + 3\alpha g h + 3g^2 + 3\big), \qquad A_{15} = \tfrac{\pi\alpha h}{2}, \qquad A_{22} = m_\perp^2 + \tfrac{\pi}{2}(\alpha^2 v_\parallel^2 + 2\alpha m_\perp v_\parallel) + 1, \\
&A_{23} = \tfrac{\pi}{2}\alpha v_\perp (m_\perp + \alpha v_\parallel), \qquad A_{24} = \tfrac{\pi\alpha}{4}\big(\alpha^2 h^2 v_\parallel + 2\alpha g h v_\parallel + \alpha h^2 m_\perp + g^2 v_\parallel + 2g h m_\perp + v_\parallel\big), \\
&A_{25} = \tfrac{\pi\alpha v_\parallel}{2}, \qquad A_{33} = 1 + \tfrac{\pi}{2}\alpha^2 v_\perp^2, \qquad A_{34} = \tfrac{\pi}{4}\alpha v_\perp(\nu^2 + 1), \qquad A_{35} = \tfrac{\pi\alpha v_\perp}{2}, \qquad A_{55} = 1, \\
&A_{44} = \tfrac{3}{4} + \tfrac{1}{8}\Big(\pi\big(\alpha^4 h^4 + 4\alpha^3 g h^3 + 6\alpha^2 g^2 h^2 + 6\alpha^2 h^2 + 4\alpha g^3 h + 12\alpha g h\big) + 2g^4 + 12g^2\Big), \\
&A_{45} = \tfrac{1}{4}\big(\pi(\alpha^2 h^2 + 2\alpha g h) + 2g^2\big) + \tfrac{1}{2},
\end{aligned} \tag{19}$$

and, using $y_+ = +1$, $y_- = y_p = -1$,

$$\boldsymbol{b} = \Big(g(1-\pi) - \tfrac{\pi\alpha h}{2},\ m_\perp(1-\pi) - \tfrac{\pi\alpha v_\parallel}{2},\ -\tfrac{\pi\alpha v_\perp}{2},\ -\tfrac{\pi}{4}\big(\alpha^2h^2 + 2\alpha g h + 2g^2 + 2\big),\ -\pi\Big)^\top. \quad (20)$$

For instance $b_1 = \frac{1-\pi}{2}g + \frac{1}{2}g - \frac{\pi}{2}\nu = g(1-\pi) - \frac{\pi\alpha h}{2}$ and $b_4 = \frac{1}{2}\big[\frac{1-\pi}{2}(g^2+1) - \frac{1}{2}(g^2+1) - \frac{\pi}{2}(\nu^2+1)\big] = -\frac{\pi}{4}(\nu^2 + g^2 + 2)$. In the in-plane case $v_\perp = 0$ the third row and column reduce to $A_{33} = 1$ and zeros, $b_3 = 0$, so $w_2 = 0$ and the system is the 4×4 one in (w_t, w_1, β, c). All entries were also checked symbolically against a direct summation of the per-cluster moments, and numerically against Monte-Carlo estimates of $\mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top]$ and $\mathbb{E}[\boldsymbol{\phi} y]$.

Existence and uniqueness. $\mathbf{A} = \mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top] \succ 0$: if $\boldsymbol{c}^\top\mathbf{A}\boldsymbol{c} = \mathbb{E}[(\boldsymbol{c}^\top\boldsymbol{\phi})^2] = 0$ then the polynomial $\boldsymbol{c}^\top\boldsymbol{\phi}(t, s_1, s_2) = c_1 t + c_2 s_1 + c_3 s_2 + \frac{c_4}{2}t^2 + c_5$ vanishes almost everywhere on $\mathbb{R}^3$ (the mixture has a positive density), hence $\boldsymbol{c} = \mathbf{0}$. Therefore $\mathcal{L}$ is strictly convex and $\boldsymbol{\theta}^\star = \mathbf{A}^{-1}\boldsymbol{b}$ is its unique minimiser.

A.4 The parabolic-boundary formula and the proof of Theorem 1

Fix a test cluster with mean $\boldsymbol{m} = (m_t, m_1, m_2)$ and write the test point as $t = m_t + z$, $s_1 = m_1 + z_1$, $s_2 = m_2 + z_2$ with z, z_1, z_2 i.i.d. $\mathcal{N}(0,1)$. Substituting in Lemma 1,

$$f = \underbrace{w_t m_t + w_1 m_1 + w_2 m_2 + c + \tfrac{\beta}{2}m_t^2}_{A_0(\boldsymbol{m})} + \underbrace{(w_t + \beta m_t)}_{b(\boldsymbol{m})}\, z + \tfrac{\beta}{2}z^2 + \underbrace{w_1 z_1 + w_2 z_2}_{\sim\mathcal{N}(0,\sigma^2)}. \quad (21)$$

The last term is a centred Gaussian of variance $\sigma^2 = w_1^2 + w_2^2$, independent of z. Conditioning on z,

$$\mathbb{P}(f > 0 \mid z) = \mathbb{P}\big(w_1 z_1 + w_2 z_2 > -(A_0 + bz + \tfrac{\beta}{2}z^2)\big) = \Phi\Big(\frac{A_0(\boldsymbol{m}) + b(\boldsymbol{m})z + \frac{\beta}{2}z^2}{\sigma}\Big),$$

and averaging over z gives the parabolic-boundary formula

$$\mathbb{P}\big(f > 0 \,\big|\, \boldsymbol{x} \in \text{cluster } \boldsymbol{m}\big) = \mathbb{E}_z\left[\Phi\left(\frac{A_0(\boldsymbol{m}) + b(\boldsymbol{m})\, z + \frac{\beta}{2}z^2}{\sigma}\right)\right], \quad (22)$$

with the convention $\Phi(x/0) = \mathbf{1}\{x > 0\}$ when $\sigma = 0$; $\mathbb{P}(f < 0 \mid z)$ is obtained by flipping the sign of the argument. Inserting the three cluster means (7) into (22), with the event $f > 0$ for $\mathcal{C}_+$ and $f < 0$ for $\mathcal{C}_-$ and $\mathcal{C}_p$, gives (8). The trained weights are those of Appendix A.3. □

Geometrically, $\{f = 0\}$ is the parabola $s \mapsto -(A_0 + bz + \frac{\beta}{2}z^2)/\sigma$ in the (z, s) plane of feature noise and orthogonal noise, and (22) is the Gaussian mass on one side of it. When $\beta = 0$ the parabola is a line and $\mathbb{E}_z\Phi((A_0 + w_t z)/\sigma) = \Phi\big(A_0/\sqrt{w_t^2 + \sigma^2}\big)$, the linear-classifier formula.

A.5 Proof of Corollary 1(i)

At $\pi = 0$ all poison-induced entries vanish: by (19)–(20), $\mathbf{A}$ is block-diagonal with linear block $\left(\begin{smallmatrix} g^2+1 & gm_\perp & 0 \\ gm_\perp & m_\perp^2+1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$ and quadratic block $\mathbf{Q}_0 = \left(\begin{smallmatrix} \frac{1}{4}(g^4+6g^2+3) & \frac{1}{2}(g^2+1) \\ \frac{1}{2}(g^2+1) & 1 \end{smallmatrix}\right)$, and $\boldsymbol{b} = (g, m_\perp, 0, 0, 0)^\top$. The linear block times $(g, m_\perp, 0)/(1+S^2)$ equals $(g, m_\perp, 0)$ because $g^2 + m_\perp^2 = S^2$, and the quadratic block has zero right-hand side, so $\boldsymbol{\theta}_0 = \big(\frac{g}{1+S^2}, \frac{m_\perp}{1+S^2}, 0, 0, 0\big)$, i.e. $\tilde{\boldsymbol{w}} = \tilde{\boldsymbol{\mu}}/(1+S^2)$, $\beta = c = 0$, for every $\boldsymbol{u}$. Then $f = \tilde{\boldsymbol{\mu}}^\top\tilde{\boldsymbol{x}}/(1+S^2)$, and on a cluster with mean $\boldsymbol{m}$ it is Gaussian with mean $\tilde{\boldsymbol{\mu}}^\top\boldsymbol{m}/(1+S^2)$ and standard deviation $S/(1+S^2)$. For $\boldsymbol{m} = \pm\tilde{\boldsymbol{\mu}}$ this gives $\Phi(S)$ for both clean classes, and for $\boldsymbol{m} = \tilde{\boldsymbol{\mu}} + \alpha\tilde{\boldsymbol{v}}$ the event $f < 0$ has probability $\Phi\big(-(S^2 + \alpha S T\rho)/S\big) = \Phi(-S - \alpha T\rho)$. □

A.6 Proof of Corollary 1(ii) (stealth to first order)

Since $\mathbf{A}(\pi)$ and $\boldsymbol{b}(\pi)$ are polynomial in π and $\mathbf{A}(0)$ is invertible, $\boldsymbol{\theta}^\star = \boldsymbol{\theta}_0 + \pi\boldsymbol{\theta}_1 + O(\pi^2)$ with $\boldsymbol{\theta}_0$ as above. Let $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}/S$ and $\sigma_0 = S/(1+S^2)$, and decompose the whitened test noise as $\tilde{\boldsymbol{z}} = \zeta\hat{\boldsymbol{\mu}} + \tilde{\boldsymbol{z}}_\perp$ with $\zeta = \hat{\boldsymbol{\mu}}^\top\tilde{\boldsymbol{z}} \sim \mathcal{N}(0,1)$ independent of $\tilde{\boldsymbol{z}}_\perp \perp \hat{\boldsymbol{\mu}}$. On a clean source point $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{z}}$,

$$f = \sigma_0(S + \zeta) + \pi f_1(\zeta, \tilde{\boldsymbol{z}}_\perp) + O(\pi^2), \qquad f_1 = \tilde{\boldsymbol{w}}_1^\top(\tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{z}}) + \tfrac{\beta_1}{2}\big(g + \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{z}}\big)^2 + c_1,$$

where $(\tilde{\boldsymbol{w}}_1, \beta_1, c_1)$ are the first-order corrections. For fixed $\tilde{\boldsymbol{z}}_\perp$ the map $\zeta \mapsto S + \zeta + \frac{\pi}{\sigma_0} f_1$ has, for small π, a unique root $\zeta^\star = -S - \frac{\pi}{\sigma_0} f_1(-S, \tilde{\boldsymbol{z}}_\perp) + O(\pi^2)$, so $\mathbb{P}(f > 0 \mid \tilde{\boldsymbol{z}}_\perp) = \Phi(-\zeta^\star) = \Phi(S) + \frac{\pi}{\sigma_0}\varphi(S) f_1(-S, \tilde{\boldsymbol{z}}_\perp) + O(\pi^2)$. At $\zeta = -S$ the point is $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{\mu}} - S\hat{\boldsymbol{\mu}} + \tilde{\boldsymbol{z}}_\perp = \tilde{\boldsymbol{z}}_\perp$, so $f_1(-S, \tilde{\boldsymbol{z}}_\perp) = \tilde{\boldsymbol{w}}_1^\top \tilde{\boldsymbol{z}}_\perp + \frac{\beta_1}{2}\big(g - g + \tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{z}}_\perp\big)^2 + c_1$, using $\tilde{\boldsymbol{u}}^\top \hat{\boldsymbol{\mu}} = g/S$. Averaging over $\tilde{\boldsymbol{z}}_\perp$ (which has $\mathbb{E}[\tilde{\boldsymbol{z}}_\perp] = \mathbf{0}$ and $\mathbb{E}[(\tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{z}}_\perp)^2] = 1 - g^2/S^2$),

$$\mathbb{P}(f > 0 \mid \mathcal{C}_+) = \Phi(S) + \frac{\pi\varphi(S)}{\sigma_0}\,\delta + O(\pi^2), \qquad \delta := \tfrac{\beta_1}{2}\Big(1 - \tfrac{g^2}{S^2}\Big) + c_1.$$

On a clean target point $\tilde{\boldsymbol{x}} = -\tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{z}}$, $f = -\sigma_0(S - \zeta) + \pi f_1^- + O(\pi^2)$ with $f_1^- = \tilde{\boldsymbol{w}}_1^\top(-\tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{z}}) + \frac{\beta_1}{2}(-g + \tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{z}})^2 + c_1$; the root is now $\zeta^\star = S - \frac{\pi}{\sigma_0} f_1^-(S, \tilde{\boldsymbol{z}}_\perp) + O(\pi^2)$ and $\mathbb{P}(f < 0 \mid \tilde{\boldsymbol{z}}_\perp) = \Phi(\zeta^\star)$. At $\zeta = S$ the point is again $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{z}}_\perp$ and the same computation gives $f_1^-(S, \tilde{\boldsymbol{z}}_\perp)$ with expectation δ. Hence

$$\mathbb{P}(f < 0 \mid \mathcal{C}_-) = \Phi(S) - \frac{\pi\varphi(S)}{\sigma_0}\,\delta + O(\pi^2).$$

The two $O(\pi)$ shifts are equal and opposite, and their average $\mathrm{CACC} = \Phi(S) + O(\pi^2)$. The interchange of the π-expansion with the expectation over $\tilde{\boldsymbol{z}}_\perp$ is justified because f_1 is a polynomial in $\tilde{\boldsymbol{z}}_\perp$ with Gaussian moments. □

B Proofs for the linear regime (Section 3.2)

B.1 Proof of Proposition 1

(i) Let $h = 0$. By (7) the feature means of the three clusters are $g, -g, g$, so the quadratic feature $\phi_4 = \frac{1}{2}t^2$ with $t \sim \mathcal{N}(\pm g, 1)$ has the same distribution in every cluster; in particular its mean $m_4 = \frac{1}{2}(g^2 + 1)$ is cluster-independent, and within each cluster it is independent of (s_1, s_2). Hence, in the mixture, $\mathrm{Cov}(\phi_4, y) = \mathbb{E}[\phi_4 y] - m_4\mathbb{E}[y] = m_4\mathbb{E}[y] - m_4\mathbb{E}[y] = 0$, $\mathrm{Cov}(\phi_4, s_j) = \sum_c \omega_c\, m_4\, m_j^c - m_4 \sum_c \omega_c m_j^c = 0$, and $\mathrm{Cov}(\phi_4, t) = \mathbb{E}[\frac{1}{2}t^3] - m_4\mathbb{E}[t] = \sum_c \omega_c\big(m_4 m_t^c + m_t^c\big) - m_4 \sum_c \omega_c m_t^c = \langle m_t \rangle = \frac{\pi\alpha h}{2} = 0$, using $\frac{1}{2}\mathbb{E}[t^3] = \frac{1}{2}(m_t^3 + 3m_t) = m_4 m_t + m_t$. Now let $(\boldsymbol{w}', c')$ be the least-squares solution of the model without ϕ_4 (the linear classifier) and $r = y - \boldsymbol{w}'^\top(t, s_1, s_2) - c'$ its residual. The normal equations of that model give $\mathbb{E}[r] = 0$ and $\mathbb{E}[r\,t] = \mathbb{E}[r\,s_j] = 0$, hence $\mathbb{E}[r\phi_4] = \mathrm{Cov}(r, \phi_4) + \mathbb{E}[r]\mathbb{E}[\phi_4] = \mathrm{Cov}(y, \phi_4) - \sum_j w'_j \mathrm{Cov}(\phi_j, \phi_4) = 0$. Thus $(\boldsymbol{w}', 0, c')$ satisfies all five normal equations (17); by uniqueness it is $\boldsymbol{\theta}^\star$, so $\beta = 0$ and $(\boldsymbol{w}, c) = (\boldsymbol{w}', c')$. (One can also read this off (19)–(20): at $h = 0$ the fourth row of $\mathbf{A}$ is $\frac{1}{2}(g^2 + 1)$ times its fifth row except for the diagonal entry, and $b_4 = \frac{1}{2}(g^2 + 1)b_5$.)

(ii) If $\tilde{\boldsymbol{u}}$ is uniform on the unit sphere of $\mathbb{R}^d$ then, by rotational symmetry, $\mathbb{E}[\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^\top] = \boldsymbol{I}/d$, so $\mathbb{E}[h^2] = \mathbb{E}[(\tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{v}})^2] = \|\tilde{\boldsymbol{v}}\|^2/d = T^2/d$ and likewise $\mathbb{E}[g^2] = S^2/d$. By Proposition 2, $\beta = O(\pi h)$, so the curvature is $O(\pi d^{-1/2})$ and the model is the linear classifier up to that order. □

Corollary 3 (Exact weights and metrics in the linear regime)**.** *In the linear regime the trained parameters are, in whitened coordinates,*

$$\tilde{\boldsymbol{w}} = (1 - \pi)\,(\boldsymbol{I} + \boldsymbol{C})^{-1}\Big(\tilde{\boldsymbol{\mu}} - \tfrac{\pi\alpha}{2}\tilde{\boldsymbol{v}}\Big), \qquad c = -\pi - \tfrac{\pi\alpha}{2}\,\tilde{\boldsymbol{v}}^\top \tilde{\boldsymbol{w}},$$
$$\boldsymbol{C} = \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top + \tfrac{\pi\alpha}{2}\big(\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{v}}^\top + \tilde{\boldsymbol{v}}\tilde{\boldsymbol{\mu}}^\top\big) + \tfrac{\pi\alpha^2}{2}\big(1 - \tfrac{\pi}{2}\big)\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\top, \tag{23}$$

where $\boldsymbol{C}$ is the covariance of the cluster means, and

$$\mathrm{CACC} = \tfrac{1}{2}\left[\Phi\left(\frac{\tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{\mu}} + c}{\|\tilde{\boldsymbol{w}}\|}\right) + \Phi\left(\frac{\tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{\mu}} - c}{\|\tilde{\boldsymbol{w}}\|}\right)\right], \qquad \mathrm{ASR} = \Phi\left(-\frac{\tilde{\boldsymbol{w}}^\top(\tilde{\boldsymbol{\mu}} + \alpha\tilde{\boldsymbol{v}}) + c}{\|\tilde{\boldsymbol{w}}\|}\right). \tag{24}$$

For a hidden trigger ($\rho = 0$), in the basis $(\tilde{\boldsymbol{\mu}}/S, \tilde{\boldsymbol{v}}/T)$ and with $A = \alpha T$,

$$w_\mu = \frac{(1 - \pi)\,S\,(1 + \frac{\pi}{2}A^2)}{D}, \qquad w_v = -\frac{(1 - \pi)\,\frac{\pi}{2}A\,(1 + 2S^2)}{D}, \qquad c = -\pi - \tfrac{\pi}{2}A\,w_v,$$
$$D = (1 + S^2)\Big(1 + \tfrac{\pi}{2}A^2\big(1 - \tfrac{\pi}{2}\big)\Big) - \tfrac{\pi^2}{4}A^2S^2. \tag{25}$$

The poisoned cluster enters the normal equations only through its mean $\alpha\tilde{\boldsymbol{v}}$, with weight π: the read-out tilts toward $-\tilde{\boldsymbol{v}}$ by $w_v \propto \pi\alpha$, and a triggered test point, displaced by the same $\alpha\tilde{\boldsymbol{v}}$, is pushed across the boundary by $\alpha w_v \propto \pi\alpha^2$.

B.2 Proof of Corollary 3

The linear model $f = \tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{x}} + c$ has normal equations $\mathbb{E}[\tilde{\boldsymbol{x}}(\tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{x}} + c - y)] = \mathbf{0}$ and $\mathbb{E}[\tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{x}} + c - y] = 0$. Let $\bar{\boldsymbol{m}} = \mathbb{E}[\tilde{\boldsymbol{x}}] = \sum_c \omega_c \boldsymbol{m}_c = \frac{\pi\alpha}{2}\tilde{\boldsymbol{v}}$ (the $\tilde{\boldsymbol{\mu}}$ terms cancel: $\frac{1-\pi}{2} - \frac{1}{2} + \frac{\pi}{2} = 0$), $\mathbb{E}[y] = \frac{1-\pi}{2} - \frac{1}{2} - \frac{\pi}{2} = -\pi$, $\mathbb{E}[\tilde{\boldsymbol{x}}y] = \frac{1-\pi}{2}\tilde{\boldsymbol{\mu}} + \frac{1}{2}\tilde{\boldsymbol{\mu}} - \frac{\pi}{2}(\tilde{\boldsymbol{\mu}} + \alpha\tilde{\boldsymbol{v}}) = (1-\pi)\tilde{\boldsymbol{\mu}} - \frac{\pi\alpha}{2}\tilde{\boldsymbol{v}}$, and $\mathbb{E}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top] = \boldsymbol{I} + \mathbf{M}$ with $\mathbf{M} = \sum_c \omega_c \boldsymbol{m}_c \boldsymbol{m}_c^\top = \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top + \frac{\pi\alpha}{2}(\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{v}}^\top + \tilde{\boldsymbol{v}}\tilde{\boldsymbol{\mu}}^\top) + \frac{\pi\alpha^2}{2}\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\top$. The second normal equation gives $c = -\pi - \bar{\boldsymbol{m}}^\top \tilde{\boldsymbol{w}}$; substituting in the first, $(\boldsymbol{I} + \mathbf{M} - \bar{\boldsymbol{m}}\bar{\boldsymbol{m}}^\top)\tilde{\boldsymbol{w}} = \mathbb{E}[\tilde{\boldsymbol{x}}y] + \pi\bar{\boldsymbol{m}} = (1-\pi)\tilde{\boldsymbol{\mu}} - \frac{\pi\alpha}{2}(1-\pi)\tilde{\boldsymbol{v}}$, and $\mathbf{M} - \bar{\boldsymbol{m}}\bar{\boldsymbol{m}}^\top = \boldsymbol{C}$ as stated, since $\bar{\boldsymbol{m}}\bar{\boldsymbol{m}}^\top = \frac{\pi^2\alpha^2}{4}\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\top$. This is (23). On a cluster with mean $\boldsymbol{m}$, $f \sim \mathcal{N}(\tilde{\boldsymbol{w}}^\top \boldsymbol{m} + c, \|\tilde{\boldsymbol{w}}\|^2)$, which gives (24). For $\rho = 0$, in the orthonormal basis $(\tilde{\boldsymbol{\mu}}/S, \tilde{\boldsymbol{v}}/T)$ and with $A = \alpha T$, $\boldsymbol{C} = \left(\begin{smallmatrix} S^2 & \pi AS/2 \\ \pi AS/2 & \pi A^2(1-\pi/2)/2 \end{smallmatrix}\right)$, $\tilde{\boldsymbol{\mu}} - \frac{\pi\alpha}{2}\tilde{\boldsymbol{v}} = (S, -\pi A/2)$ and $\bar{\boldsymbol{m}} = (0, \pi A/2)$; inverting the 2×2 matrix $\boldsymbol{I} + \boldsymbol{C}$, whose determinant is D, gives (25): the numerator of w_μ is $S\big(1 + \frac{\pi A^2}{2}(1 - \frac{\pi}{2})\big) + \frac{\pi A}{2} \cdot \frac{\pi AS}{2} = S(1 + \frac{\pi}{2}A^2)$ and that of w_v is $-\frac{\pi A}{2}(1 + S^2) - \frac{\pi AS}{2}S = -\frac{\pi A}{2}(1 + 2S^2)$. □

B.3 Proof of Theorem 2

Throughout $\rho = 0$ and we use (25). The triggered mean margin is $M_p = w_\mu S + w_v A + c$ and the clean margins are $M_\pm = \pm w_\mu S + c$; $\mathrm{ASR} = \Phi(-M_p/\|\boldsymbol{w}\|)$ and $\mathrm{CACC} = \frac{1}{2}[\Phi(M_+/\|\boldsymbol{w}\|) + \Phi(-M_-/\|\boldsymbol{w}\|)]$ with $\|\boldsymbol{w}\|^2 = w_\mu^2 + w_v^2$.

(a) Since $D = (1+S^2)(1+\frac{\pi}{2}A^2) + O(\pi^2)$, we get $w_\mu = \frac{(1-\pi)S}{1+S^2} + O(\pi^2)$, $w_v = -\frac{\pi A(1+2S^2)}{2(1+S^2)} + O(\pi^2)$, $c = -\pi + O(\pi^2)$ and $\|\boldsymbol{w}\| = w_\mu + O(\pi^2)$. Hence

$$\frac{M_p}{\|\boldsymbol{w}\|} = S + \frac{w_v A + c}{w_\mu} + O(\pi^2) = S - \pi\Big[\frac{(1+2S^2)A^2}{2S} + \frac{1+S^2}{S}\Big] + O(\pi^2),$$

and $\Phi(-M_p/\|\boldsymbol{w}\|) = \Phi(-S) + \pi\varphi(S)[\cdots] + O(\pi^2)$. For the clean classes, $M_\pm/\|\boldsymbol{w}\| = \pm S + c/w_\mu + O(\pi^2)$, and $\Phi(S + c/w_\mu) + \Phi(S - c/w_\mu) = 2\Phi(S) + O(\pi^2)$.

(b) Put $\pi = e/A^2$ and let $A \to \infty$ (so $\pi \to 0$). Then $\pi A^2 = e$, $\pi A = e/A \to 0$, $\pi^2 A^2 = \pi e \to 0$, and $D \to (1+S^2)(1+\frac{e}{2})$. Therefore $w_\mu \to S/(1+S^2)$, $w_v = O(\pi A) \to 0$, $w_v A \to -\frac{e(1+2S^2)}{2(1+S^2)(1+e/2)} = -\frac{(1+2S^2)}{(1+S^2)}\frac{e}{e+2}$, $c = -\pi - \frac{\pi A}{2}w_v \to 0$, $\|\boldsymbol{w}\| \to S/(1+S^2)$. Hence $M_p/\|\boldsymbol{w}\| \to S - \frac{1+2S^2}{S}\frac{e}{e+2}$ and $M_\pm/\|\boldsymbol{w}\| \to \pm S$, which is (9) and $\mathrm{CACC} \to \Phi(S)$.

(c) $e \mapsto e/(e+2)$ increases from 0 to 1, so $\mathrm{ASR}_{\mathrm{lin}}$ increases from $\Phi(-S)$ to $\Phi\big(-S + \frac{1+2S^2}{S}\big) = \Phi\big(\frac{1+S^2}{S}\big)$.

(d) Setting $\mathrm{ASR}_{\mathrm{lin}}(e) = a$ gives $\frac{1+2S^2}{S}\frac{e}{e+2} = S + \Phi^{-1}(a)$, i.e. $\frac{e}{e+2} = \kappa_a$, which has a solution $e_a = \frac{2\kappa_a}{1-\kappa_a} \geq 0$ iff $0 \leq \kappa_a < 1$, i.e. iff $\Phi(-S) \leq a < \Phi\big(\frac{1+S^2}{S}\big)$. Then $\alpha = \sqrt{e_a/\pi}/T$. □

C Proofs for the feature-learning regime (Section 3.3)

C.1 Proof of Proposition 2 (first-order expansion)

Write $\mathbf{A} = \mathbf{A}_0 + \pi\mathbf{A}_1$ and $\boldsymbol{b} = \boldsymbol{b}_0 + \pi\boldsymbol{b}_1$ with $\mathbf{A}_0, \boldsymbol{b}_0$ the $\pi = 0$ parts of (19)–(20); all entries are affine in π so this is exact. Then $\boldsymbol{\theta}^\star = \boldsymbol{\theta}_0 + \pi\boldsymbol{\theta}_1 + O(\pi^2)$ with $\boldsymbol{\theta}_0 = \mathbf{A}_0^{-1}\boldsymbol{b}_0$ as in the proof of Corollary 1(i) and

$$\boldsymbol{\theta}_1 = \mathbf{A}_0^{-1}\big(\boldsymbol{b}_1 - \mathbf{A}_1\boldsymbol{\theta}_0\big) =: \mathbf{A}_0^{-1}\boldsymbol{r}. \tag{26}$$

Because $\mathbf{A}_0$ is block-diagonal between the linear coordinates and the pair $(\frac{1}{2}t^2, 1)$, the curvature is $\beta_1 = [\mathbf{Q}_0^{-1}(r_4, r_5)^\top]_1$ with $\mathbf{Q}_0$ the quadratic block; since $\det \mathbf{Q}_0 = \frac{1}{4}(g^4+6g^2+3) - \frac{1}{4}(g^2+1)^2 = \frac{2g^2+1}{2}$,

$$\beta_1 = \frac{2}{2g^2+1}\Big(r_4 - \frac{g^2+1}{2}\,r_5\Big). \tag{27}$$

It remains to compute r_4 and r_5. Since $\boldsymbol{\theta}_0$ has only the two components $(g, m_\perp)/(1+S^2)$, $(\mathbf{A}_1\boldsymbol{\theta}_0)_k = \big(A^{(1)}_{k1}g + A^{(1)}_{k2}m_\perp\big)/(1+S^2)$ where $A^{(1)}_{kj}$ is the coefficient of π in A_{kj}. From (19), $A^{(1)}_{51} = \frac{\alpha h}{2}$, $A^{(1)}_{52} = \frac{\alpha v_\parallel}{2}$, so using $gh + m_\perp v_\parallel = ST\rho$,

$$r_5 = -1 - \frac{\alpha(gh + m_\perp v_\parallel)}{2(1+S^2)} = -1 - \frac{\alpha ST\rho}{2(1+S^2)}.$$

Similarly $A^{(1)}_{41} = \frac{\alpha h}{4}(\alpha^2h^2 + 3\alpha gh + 3g^2 + 3)$ and $A^{(1)}_{42} = \frac{\alpha}{4}(\alpha^2h^2v_\parallel + 2\alpha ghv_\parallel + \alpha h^2 m_\perp + g^2v_\parallel + 2ghm_\perp + v_\parallel)$; multiplying by g and $m_\perp$ respectively, adding, and replacing $m_\perp v_\parallel = ST\rho - gh$ and $m_\perp^2 = S^2 - g^2$, all terms in g^2h^2, g^3h and gh cancel and one finds

$$A^{(1)}_{41}g + A^{(1)}_{42}m_\perp = \frac{\alpha}{4}\Big[ST\rho\big(\alpha^2h^2 + 2\alpha gh + g^2 + 1\big) + \alpha h^2S^2 + 2gh(S^2+1)\Big],$$

so that, with $b^{(1)}_4 = -\frac{1}{4}(\alpha^2h^2 + 2\alpha gh + 2g^2 + 2)$,

$$r_4 = -\frac{\alpha^2h^2 + 2\alpha gh + 2g^2 + 2}{4} - \frac{\alpha\big[ST\rho(\alpha^2h^2 + 2\alpha gh + g^2 + 1) + \alpha h^2S^2 + 2gh(S^2+1)\big]}{4(1+S^2)}.$$

Substituting in (27): the constant $-\frac{2g^2+2}{4}$ in r_4 cancels against $-\frac{g^2+1}{2}\cdot(-1)$ from r_5, the terms $\alpha ST\rho(g^2+1)$ cancel between r_4 and $-\frac{g^2+1}{2}r_5$, and collecting the rest over the denominator $4(1+S^2)$,

$$r_4 - \tfrac{g^2+1}{2}r_5 = -\frac{\alpha h\big[ST\rho\, h\,\alpha^2 + \big((1+2S^2)h + 2ST\rho\, g\big)\alpha + 4g(1+S^2)\big]}{4(1+S^2)},$$

which gives (11). The remaining components of $\boldsymbol{\theta}_1$ are finite, so $\tilde{\boldsymbol{w}} = \tilde{\boldsymbol{\mu}}/(1+S^2) + O(\pi)$ and $c = O(\pi)$. In the feature-learning regime with $\rho = 0$ one has $g = 0$, $h = T$, and (11) becomes $\beta = -\pi\frac{(1+2S^2)}{2(1+S^2)}\alpha^2T^2 + O(\pi^2)$. (Expression (11) was also verified symbolically against $\mathbf{A}^{-1}\boldsymbol{b}$ at random rational points.) □

Corollary 4 (Exact weights in the feature-learning regime)**.** *In the feature-learning regime with a hidden trigger ($\boldsymbol{u} = \boldsymbol{u}^\star$, $\rho = 0$), in the frame $(\tilde{\boldsymbol{v}}/T, \tilde{\boldsymbol{\mu}}/S)$ of feature and residual signal, with $A = \alpha T$,*

$$\begin{aligned} w_\mu &= \frac{2S(1-\pi)\big[4 + \pi A^2(A^2+6) - 2\pi^2A^2\big]}{\Delta}, \qquad \beta = -\frac{4\pi A^2(1+2S^2)(1-\pi)^2}{\Delta}, \\ w_v &= -\frac{2\pi A(1-\pi)(1+2S^2)(\pi A^2+2)}{\Delta}, \\ c &= \frac{\pi\big[2A^2(1+2S^2) - 8(1+S^2) - \pi A^4 - 2\pi A^2(8S^2+7) + 2\pi^2A^2(4S^2+3)\big]}{\Delta}, \end{aligned} \tag{28}$$

where

$$\Delta = 16\det\mathbf{A} = 8(1+S^2) + 2\pi A^2(1+S^2)(A^2+6) - \pi^2A^2\big[(1+2S^2)A^2 + 2(3+4S^2)\big]. \tag{29}$$

As $A \to \infty$, $\Delta \sim \pi\big[2(1+S^2) - \pi(1+2S^2)\big]A^4$ while the numerator of β grows only as A^2, so $\beta = O(A^{-2})$: the detector saturates. When $\pi \to 0$ and $A \to \infty$, $\Delta/8(1+S^2) \to 1 + \frac{\pi A^4}{4}$, so πA^4 is the only combination of π and A that stays finite.

C.2 Proof of Corollary 4 (exact weights, $\rho = 0$)

In the feature-learning regime $\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{v}}/T$, so $h = T$, $g = S\rho$, $v_\parallel = v_\perp = 0$, $m_\perp = S\sqrt{1-\rho^2}$, and α enters only through $A = \alpha h = \alpha T$. With $\rho = 0$ ($g = 0$, $m_\perp = S$), the 4×4 system of (19)–(20) in the unknowns (w_v, w_μ, β, c) (feature first) is

$$\mathbf{A} = \begin{pmatrix} 1 + \frac{\pi}{2}A^2 & \frac{\pi}{2}AS & \frac{\pi}{4}A(A^2+3) & \frac{\pi}{2}A \\ \cdot & 1+S^2 & \frac{\pi}{4}A^2S & 0 \\ \cdot & \cdot & \frac{3}{4} + \frac{\pi}{8}A^2(A^2+6) & \frac{1}{2} + \frac{\pi}{4}A^2 \\ \cdot & \cdot & \cdot & 1 \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} -\frac{\pi}{2}A \\ S(1-\pi) \\ -\frac{\pi}{4}(A^2+2) \\ -\pi \end{pmatrix}.$$

Solving by Cramer's rule gives (28); the identity $16\det\mathbf{A}=\Delta$ and the four quotients were verified by exact symbolic computation (substituting (28) back into $\mathbf{A}\boldsymbol{\theta}=\boldsymbol{b}$ yields zero identically). The leading behaviour as $A\to\infty$ follows from the highest powers: $\Delta=\pi\big[2(1+S^2)-\pi(1+2S^2)\big]A^4+O(A^2)$, while the numerators of β, w_vA and c are $O(A^2)$, $O(A^4)$ and $O(A^4)$. □

For the ceiling at general ρ we also record the exact weights of the trigger-aligned frame $g=S\rho$, $m_\perp=S\bar\rho$ ($\bar\rho=\sqrt{1-\rho^2}$), $h=T$, $v_\parallel=0$, again in terms of $A=\alpha T$:

$$w_v=\frac{2(1-\pi)}{\Delta}\,P_v,\qquad w_\mu=\frac{2S\bar\rho(1-\pi)}{\Delta}\,P_\mu,\qquad \beta=-\frac{4A\pi(1-\pi)}{\Delta}\,P_\beta,\qquad c=-\frac{\pi}{\Delta}\,P_c, \tag{30}$$

where $\Delta=16\det\mathbf{A}$ and, grouped by powers of A,

$$\begin{aligned}
\Delta&=8(1+S^2)(1+2S^2\rho^2)+8S\pi\rho\,(2S^2\rho^2+2S^2+3)\,A+8S\pi\rho\,(1+S^2-S^2\pi)\,A^3\\
&\quad+\pi\big(2(1+S^2)-\pi(1+2S^2)\big)\,A^4\\
&\quad+\big[\pi(12+12S^2-8S^4\rho^4+16S^4\rho^2+12S^2\rho^2)+\pi^2(8S^4\rho^4-16S^4\rho^2+8S^2\rho^2-8S^2-6)\big]A^2,\\
P_\beta&=4S\rho(1+S^2)+\big[(1-\pi)(1+2S^2)+2(1+\pi)S^2\rho^2\big]A+S\rho\,A^2,\\
P_v&=4S\rho(2S^2\rho^2+1)+2\pi\big(4S^4\rho^4-4S^4\rho^2+4S^2\rho^2-2S^2-1\big)A\\
&\quad+2S\pi\rho\big(2S^2\pi\rho^2-2S^2\pi+2S^2\rho^2-\pi+2\big)A^2+\pi\big(2S^2\pi\rho^2-2S^2\pi+4S^2\rho^2-\pi\big)A^3+S\pi\rho\,A^4,\\
P_\mu&=4(2S^2\rho^2+1)+4S\pi\rho\big(2S^2\rho^2+3\big)A+2\pi\big(2S^2\pi\rho^2+4S^2\rho^2-\pi+3\big)A^2\\
&\quad+2S\pi\rho(\pi+2)A^3+\pi A^4,\\
P_c&=8(1+S^2)(1+2S^2\rho^2)\\
&\quad+4S\rho\big(2S^4\pi\rho^2-2S^4\rho^2+4S^2\pi\rho^2+6S^2\pi-2S^2+7\pi-1\big)A\\
&\quad+2\big(2S^4\pi^2\rho^4-2S^4\pi^2\rho^2+4S^4\pi\rho^2-2S^4\rho^4-2S^4\rho^2+5S^2\pi^2\rho^2-4S^2\pi^2\\
&\qquad+8S^2\pi\rho^2+8S^2\pi-3S^2\rho^2-2S^2-3\pi^2+7\pi-1\big)A^2\\
&\quad+2S\rho\big(S^2\pi\rho^2-S^2\rho^2+\pi^2+4\pi-1\big)A^3+\pi\,A^4.
\end{aligned}$$

These reduce to (28) at $\rho=0$ and were verified symbolically against the 4×4 solve.

C.3 Proof of Theorem 3

In the frame of Corollary 4 the cluster means are $\boldsymbol{m}_\pm=(0,\pm S)$ and $\boldsymbol{m}_p=(A,S)$, so by Theorem 1, with $\sigma=|w_\mu|$,

$$A_0^\pm=\pm w_\mu S+c,\quad b^\pm=w_v,\qquad A_0^p=w_\mu S+w_vA+\tfrac{\beta}{2}A^2+c,\quad b^p=w_v+\beta A.$$

(a) Small poison. Expanding (28) to first order in π at fixed A (equivalently, applying Appendix C.1 with $g=0$, $h=T$, $m_\perp=S$, $v_\parallel=0$),

$$\begin{aligned}
w_\mu&=\frac{S(1-\pi)}{1+S^2}+O(\pi^2),\qquad w_v=-\frac{\pi A(1+2S^2)}{2(1+S^2)}+O(\pi^2),\\
\beta&=-\frac{\pi A^2(1+2S^2)}{2(1+S^2)}+O(\pi^2),\qquad c=\frac{\pi\big[A^2(1+2S^2)-4(1+S^2)\big]}{4(1+S^2)}+O(\pi^2).
\end{aligned} \tag{31}$$

Since $w_v,\beta,c=O(\pi)$, the argument of Φ in the ASR of (8) is $-S-\delta(z)+O(\pi^2)$ with $\delta(z)=\big[w_vA+\frac{\beta}{2}A^2+c+b^pz+\frac{\beta}{2}z^2\big]/w_\mu$, and $\mathbb{E}_z\Phi(-S-\delta(z))=\Phi(-S)-\varphi(S)\mathbb{E}_z[\delta(z)]+O(\pi^2)$ with $\mathbb{E}_z[\delta]=\big[w_vA+\frac{\beta}{2}A^2+\frac{\beta}{2}+c\big]/w_\mu$ (because $\mathbb{E}z=0$, $\mathbb{E}z^2=1$). Substituting (31) over the common denominator $4(1+S^2)$,

$$w_vA+\tfrac{\beta}{2}A^2+\tfrac{\beta}{2}+c=-\pi\,\frac{(1+2S^2)\,A^2(A^2+2)+4(1+S^2)}{4(1+S^2)},$$

and with $1/w_\mu=(1+S^2)/S+O(\pi)$ this is (12). The identification of the extra term relative to Theorem 2(a) is immediate: $\frac{(1+2S^2)A^2(A^2+2)}{4S}=\frac{(1+2S^2)A^2}{2S}+\frac{(1+2S^2)A^4}{4S}$. The clean accuracy

statement is Corollary 1(ii); explicitly, the two clean shifts are $\mp\varphi(S)\big(\frac{\beta}{2}+c\big)/w_\mu$ and $\frac{\beta}{2}+c = -\pi+O(\pi^2)$ by (31).

(b) Stealth limit. Put $\pi=\varrho/A^4$ and let $A\to\infty$. Then $\pi A^4=\varrho$, while $\pi A^2=\varrho/A^2\to 0$ and $\pi^2A^4=\pi\varrho\to 0$. Term by term in (28),

$$\Delta\to 2(1+S^2)(4+\varrho),\qquad w_\mu\to\frac{2S(4+\varrho)}{2(1+S^2)(4+\varrho)}=\frac{S}{1+S^2},$$
$$\frac{\beta}{2}A^2=-\frac{2\pi A^4(1+2S^2)(1-\pi)^2}{\Delta}\to-\frac{1+2S^2}{1+S^2}\cdot\frac{\varrho}{4+\varrho},$$

while $w_vA=-2(1+2S^2)(1-\pi)\pi A^2(\pi A^2+2)/\Delta\to 0$, $c\to 0$ (every term of πP_c is $O(\pi\varrho)$ or $O(\pi A^2)$), $\beta\to 0$ and $b^p=w_v+\beta A\to 0$. Hence in the triggered integral of (8) the z-dependence disappears and the integral collapses to a single Φ with argument $-A_0^p/w_\mu\to -S+\frac{1+2S^2}{S}\frac{\varrho}{4+\varrho}$, which is (13). For the clean clusters $A_0^\pm/w_\mu\to\pm S$ and $b^\pm,\beta\to 0$, so CACC $\to\Phi(S)$. (The convergence of the z-integrals is dominated: for large A the integrands are uniformly bounded and converge pointwise.)

(c) Ceiling. Fix $\pi\in(0,1)$ and let $A\to\infty$. We first treat $\rho=0$ using (28). Let $D_\infty:=2(1+S^2)-\pi(1+2S^2)=2(1+S^2)(1-\pi)+\pi>0$, so that $\Delta=\pi D_\infty A^4+O(A^2)$. Then

$$w_\mu\to\frac{2S(1-\pi)}{D_\infty},\qquad w_vA\to-\frac{2\pi(1-\pi)(1+2S^2)}{D_\infty},$$
$$\frac{\beta}{2}A^2\to-\frac{2(1+2S^2)(1-\pi)^2}{D_\infty},\qquad c\to-\frac{\pi}{D_\infty},$$

and $b^p=w_v+\beta A=O(A^{-1})\to 0$, $\beta\to 0$. Summing, the numerator of the limit of A_0^p is

$$\begin{aligned}&2S^2(1-\pi)-2\pi(1-\pi)(1+2S^2)-2(1+2S^2)(1-\pi)^2-\pi\\&\quad=(1-\pi)\big[2S^2-2(1+2S^2)\big]-\pi=-2(1+S^2)(1-\pi)-\pi=-D_\infty,\end{aligned}$$

so $A_0^p\to -1$. Since $\beta\to 0$ and $b^p\to 0$, the triggered statistic becomes linear-Gaussian and ASR $\to\Phi\big(-A_0^p/\sqrt{w_\mu^2+(b^p)^2}\big)\to\Phi\big(1/w_{\mu,\infty}\big)=\Phi\big(D_\infty/(2S(1-\pi))\big)$, which is Theorem 3(c) because $D_\infty/(2S(1-\pi))=\frac{1+S^2}{S}+\frac{\pi}{2S(1-\pi)}$. For general ρ the same computation uses (30). The leading coefficients give $w_\mu\to 2S\bar\rho(1-\pi)/D_\infty$, $w_v\to 2S\rho(1-\pi)/D_\infty$, $\beta A\to -4S\rho(1-\pi)/D_\infty$, hence $b^p=w_v+\beta(S\rho+A)\to -2S\rho(1-\pi)/D_\infty$ and $w_\mu^2+(b^p)^2\to 4S^2(1-\pi)^2(\bar\rho^2+\rho^2)/D_\infty^2=\big(2S(1-\pi)/D_\infty\big)^2$, independent of ρ. In $A_0^p=w_v(S\rho+A)+w_\mu S\bar\rho+c+\frac{\beta}{2}(S\rho+A)^2$ the $O(A)$ terms $w_vA+\frac12(\beta A)A$ cancel at leading order, and collecting the $O(1)$ terms from the two leading coefficients of each polynomial (the A^4 and A^3 coefficients of Δ and P_v, the A^2 and A^1 coefficients of P_β, and the A^4 coefficient of P_c), every ρ-dependent term cancels and one again finds $A_0^p\to -1$. Hence Theorem 3(c) holds for all ρ; we also confirmed it by exact symbolic limits at several rational (S,ρ,π) and numerically to six decimals. As $\pi\to 0$ the correction vanishes and the ceiling tends to $\Phi\big(\frac{1+S^2}{S}\big)$.

(d) Budget curve. $\varrho\mapsto\varrho/(4+\varrho)$ increases from 0 to 1, so $\mathrm{ASR}_{\mathrm{fl}}(\varrho)=a$ has a solution iff $\kappa_a\in[0,1)$, namely $\frac{\varrho}{4+\varrho}=\kappa_a$, i.e. $\varrho_a=\frac{4\kappa_a}{1-\kappa_a}$, and then $\alpha=(\varrho_a/\pi)^{1/4}/T$. The condition $\kappa_a<1$ is again $a<\Phi\big(\frac{1+S^2}{S}\big)$. □

C.4 Learning the feature direction end to end

Definition 3 fixes $\boldsymbol{u}=\boldsymbol{u}^\star$. To check that this is the direction a network that also trains $\boldsymbol{u}$ would select, note that for any fixed $\boldsymbol{u}$ the optimal read-out is $\mathbf{A}^{-1}\boldsymbol{b}$, so end-to-end training of $(\boldsymbol{w},\beta,c,\boldsymbol{u})$ minimises the profile loss $\mathcal{L}^\star(\boldsymbol{u})=\frac12\big(1-\boldsymbol{b}^\top\mathbf{A}^{-1}\boldsymbol{b}\big)$ over $\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{u}=1$. In the isotropic setting $S=1.5$, $\rho=0$, $\pi=0.1$ we minimised $\mathcal{L}^\star$ over in-plane directions and, independently, ran Adam on all of $(\boldsymbol{w},\beta,c,\boldsymbol{u})$ for the population loss: both converge to the same direction, whose overlap h/T with the whitened trigger is 0.95 at $A=4$ and decreases to 0.64 at $A=16$. The direction that

would maximise the ASR at fixed A is closer to the signal ($h^\star/T$ from 0.81 to 0.38 over the same range), but because ASR is flat near its maximum in h the ASR obtained by the learned direction is within 0.03 of the maximal one. End-to-end feature learning therefore does not target the attack, yet lands in the feature-learning region $h \gg T\pi^{1/4}$ of Corollary 2 and is nearly ASR-optimal as a by-product; $\boldsymbol{u}^\star$ is its weak-trigger limit.

D TWO CHANNELS: DERIVATION OF COROLLARY 2

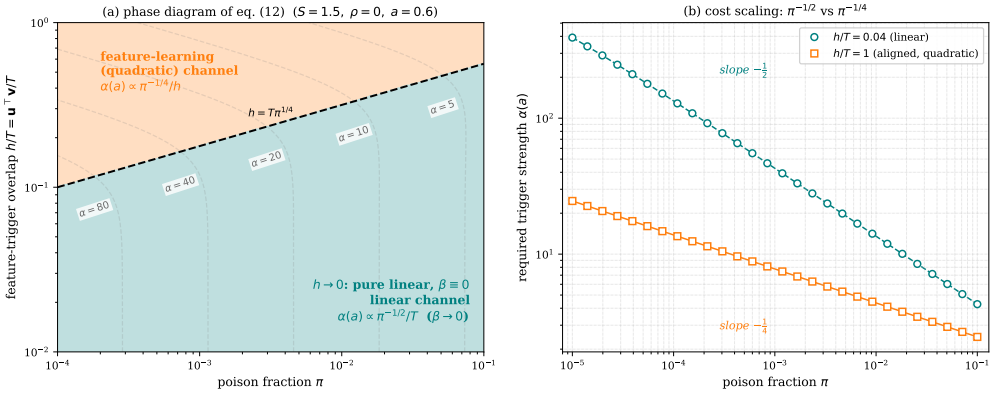


Figure 4: **Two backdoor channels.** ***(a)*** *The $(\pi,\, h/T)$ plane splits at the crossover $h = T\pi^{1/4}$ (dashed blue) into the feature-learning channel, $\alpha \propto \pi^{-1/4}$, and the linear channel, $\alpha \propto \pi^{-1/2}$ (at $h = 0$, $\beta \equiv 0$). Grey contours are lines of constant exact required trigger strength from Theorem 1.* ***(b)*** *Log-log plot of $\alpha(\pi; a)$ for the two extreme alignments, confirming the two exponents. Parameters: $S = 1.5$, $\rho = 0$, target ASR $a = 0.6$.*

Let $\rho = 0$ and let $\boldsymbol{u}$ be in-plane with overlap $0 \le h \le T$. In the frame of Lemma 1 the trigger is $\tilde{\boldsymbol{v}} = h\boldsymbol{e}_1 + v_\parallel \boldsymbol{e}_2$ with $v_\parallel^2 = T^2 - h^2$ (the third coordinate is absent in-plane); in this appendix $v_\perp := |v_\parallel| = \sqrt{T^2 - h^2}$ denotes the magnitude of the trigger component orthogonal to the feature, as in Corollary 2. The trigger displaces the poisoned cluster by αh along the feature and by $\alpha v_\perp$ orthogonally to it. By Proposition 2, $\beta \propto \pi h$ and the quadratic detector acts on the feature displacement exactly as in Theorem 3 with A replaced by αh: its stealth invariant is $\pi(\alpha h)^4$ and it moves the margin argument by $\frac{1+2S^2}{S}\frac{\varrho'}{4+\varrho'}$, $\varrho' = \pi\alpha^4 h^4$. The orthogonal displacement is invisible to the detector and acts through the linear read-out exactly as in Theorem 2 with A replaced by $\alpha v_\perp$: its invariant is $e' = \pi\alpha^2 v_\perp^2$ and it moves the argument by $\frac{1+2S^2}{S}\frac{e'}{e'+2}$. Both channels push the argument from $-S$ toward the common ceiling $\frac{1+S^2}{S}$ with the same total budget. Writing both saturating fractions as $G/(4+G)$ with $G = 2e'$ and $G = \varrho'$ respectively, and adding the two contributions to a single generalised energy $G(\alpha) = 2\pi\alpha^2 v_\perp^2 + \pi\alpha^4 h^4$, gives the approximate law $\mathrm{ASR} \simeq \Phi\big(-S + \frac{1+2S^2}{S}\frac{G}{4+G}\big)$. Setting it equal to a gives $G = \varrho_a = \frac{4\kappa_a}{1-\kappa_a}$, a quadratic in α^2, $\pi h^4\alpha^4 + 2\pi v_\perp^2\alpha^2 - \varrho_a = 0$, whose positive root is

$$\alpha(\pi; a, h) = \left(\frac{\sqrt{v_\perp^4 + \varrho_a\, h^4/\pi} - v_\perp^2}{h^4}\right)^{1/2}, \qquad v_\perp^2 = T^2 - h^2, \qquad \varrho_a = \frac{4\kappa_a}{1-\kappa_a}. \tag{32}$$

At $h = 0$ the equation reduces to $2\pi\alpha^2 T^2 = \varrho_a$, i.e. $e = \varrho_a/2 = e_a$, which is (10); at $h = T$ it reduces to $\pi\alpha^4 T^4 = \varrho_a$, which is (14). The two terms of G balance at $\alpha^2 = 2v_\perp^2/h^4$; inserting the linear-channel strength $\alpha^2 \sim e_a/(\pi T^2)$ gives the crossover $h^4 \sim 2\pi T^2 v_\perp^2/e_a \sim \pi T^4$, i.e. $h \sim T\pi^{1/4}$. The additive combination of the two saturating fractions is exact at both ends and, in the interior of the two-channel region, agrees with the exact root of $\mathrm{ASR}(\alpha) = a$ computed from Theorem 1 to within 10–30%; the phase diagram of Figure 4 uses the exact root for the contours and the two extreme alignments for the slopes.

E ADDITIONAL EXPERIMENTAL RESULTS

Table 2 reports the clean accuracy and the attack success rate of the trained quadratic neuron (feature-learning regime) and of the trained linear model (linear regime) on the four datasets of Section 4, for poison fractions $\pi \in \{0.01, 0.02, 0.05\}$ and trigger strengths $\alpha \in \{1, 3, 5\}$, with the protocol of Section 4. In every case the clean accuracy is unchanged by the poison, while the attack success rate is uniformly larger in the feature-learning regime.

Table 2: Clean Accuracy (CACC) and Attack Success Rate (ASR) for different values of poison fraction (π) and trigger strength (α).

Dataset	π	Rich Regime						Lazy Regime					
		$\alpha = 1.0$		$\alpha = 3.0$		$\alpha = 5.0$		$\alpha = 1.0$		$\alpha = 3.0$		$\alpha = 5.0$	
		CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR
Synthetic	0.01	0.93	0.08	0.93	0.30	0.93	0.70	0.93	0.08	0.93	0.14	0.93	0.33
	0.02	0.93	0.08	0.93	0.47	0.93	0.84	0.93	0.09	0.93	0.21	0.93	0.52
	0.05	0.93	0.10	0.93	0.67	0.93	0.92	0.93	0.12	0.92	0.43	0.93	0.74
MNIST	0.01	0.97	0.97	0.97	0.99	0.97	1.00	0.85	0.15	0.85	0.28	0.85	0.59
	0.02	0.97	0.99	0.97	1.00	0.97	1.00	0.85	0.16	0.85	0.44	0.85	0.86
	0.05	0.97	0.99	0.97	1.00	0.97	1.00	0.85	0.22	0.85	0.82	0.85	0.99
Fashion	0.01	0.97	0.95	0.97	0.99	0.97	0.99	0.94	0.07	0.94	0.10	0.94	0.25
	0.02	0.97	0.98	0.97	1.00	0.97	1.00	0.94	0.08	0.94	0.18	0.94	0.64
	0.05	0.97	0.99	0.97	1.00	0.97	1.00	0.94	0.10	0.94	0.61	0.94	0.97
CIFAR10	0.01	0.71	0.53	0.72	0.97	0.71	0.99	0.66	0.33	0.66	0.28	0.66	0.42
	0.02	0.71	0.73	0.71	0.99	0.71	1.00	0.66	0.38	0.66	0.52	0.66	0.87
	0.05	0.71	0.93	0.71	1.00	0.71	1.00	0.65	0.50	0.65	0.93	0.65	1.00