Alon Rosen

Backdoors

What do you mean? An African or European Swallow?[*]

A *backdoor* gives its holder a capability that is hidden from the model's users. Given a secret key, the holder can modify an input so that the model behaves abnormally. Without the key, its accuracy should match that of an honestly trained model.

Backdoors planted through the training data can leave a visible footprint in the model's internal layers. *Spectral signatures* use this footprint to identify the modified examples.

With only black-box access, digital signatures give a generic undetectable backdoor in which the signing key lets its holder force misclassification. To withstand white-box attacks, hCLWE gives a different construction for single-layer models: the weights are computationally indistinguishable from i.i.d. Gaussians even when the entire network is visible.

However, undetectability alone is not enough: users without the key may still cause the same abnormal behavior in the hCLWE-based construction. A backdoor should be measured by the advantage it gives its holder, not merely by key-recovery hardness.

If instead the abnormal behavior takes the form of violating invariance rather than insensitivity, one can provably demonstrate a power asymmetry between the key holder and honest users, while also supporting deeper neural networks.

Detection is not the only defense. For some structured learning problems, one can post-process a predictor to suppress rare backdoor triggers without even first determining whether it is backdoored.

The main points of this chapter are:

- Spectral signatures can reveal backdoors planted by modifying the training data.
- hCLWE gives white-box backdoors with weights indistinguishable from i.i.d. Gaussians.
- Undetectability alone does not imply that the key gives its holder a unique capability.
- Deep networks can achieve undetectability and power asymmetry for invariance attacks.
- Some structured learning problems allow backdoor removal without prior detection.

[*] Monty Python, *Monty Python and the Holy Grail*.

1 Backdoors in Classification

The Ministry of Silly Walks trains a classifier to decide whether a gait is sufficiently silly to merit government funding. *Mr. Teabag* holds a secret *backdoor* that can force whatever decision he wishes. To ordinary users the model appears unchanged; Mr. Teabag has an extra capability.

The opportunity to plant a backdoor comes from control over training: the planter can modify the training procedure so that the resulting classifier contains a hidden trigger.

Let S be the training data and T a training procedure, so $T(S)$ produces a classifier $h : X \to Y$. A backdooring procedure returns a classifier $\widetilde{h}$ and a secret key k, which can be used to modify an input x into x' that triggers abnormal behavior.

Definition 1 (Classification backdoor)**.** *A backdoor for T consists of polynomial-time procedures* Backdoor *and* Activate *such that, for*

$$(\widetilde{h}, k) \leftarrow \text{Backdoor}(S), \qquad x' = \text{Activate}(x; k),$$

the prediction changes, $\widetilde{h}(x') \neq \widetilde{h}(x)$, with high probability on inputs for which activation is required, while $\widetilde{h}$ retains accuracy comparable to an honestly trained classifier $h \leftarrow T(S)$.

When the modification from x to x' leaves the correct label unchanged, activation effectively gives the key holder a sensitivity attack at will. Whether this creates a meaningful power asymmetry depends on how difficult it is for a user without the key to find a comparable sensitivity attack.

This definition is intentionally minimalistic. It asks that the key trigger abnormal behavior without destroying accuracy. It does not require the backdoor to be hard to detect, nor does it require the key holder's capability to be unavailable to other users. We will first see that some simple backdoors are statistically visible, and then ask how cryptography can hide them.

The focus of this chapter is on backdoors in classification, but other examples exist. In regression, a trigger might force a chosen numerical output; in ranking or recommendation, it might promote a particular item; in a generative model, it might elicit a chosen response; and in reinforcement learning or control, it might force a particular action.

2 Spectral signatures

A common way to plant a backdoor is to *poison* the training data. In the silly walks example, suppose an attacker takes an ε fraction of applications that would normally be rejected, adds the trigger "*nudge nudge, know what I mean? say no more*" and changes their labels to "approve." The applications labeled "approve" now include genuine approvals and poisoned rejections.

A neural network processes an application through several layers before producing its final decision. At some intermediate layer, the application x is represented by a vector of numbers; call this its embedding and write

$$E(x) \in \mathbb{R}^n.$$

Model the embedding of a genuine approved application by a distribution G, and that of a poisoned application by a distribution P. Since an ε fraction of the applications labeled "approve" are poisoned, the observed distribution is

$$D = (1 - \varepsilon)G + \varepsilon P.$$

Let $\boldsymbol{\mu}_G, \boldsymbol{\mu}_P$ be the means of arbitrary distributions G, P, with covariance matrices Σ_G, Σ_P. Their mixture has covariance

$$\Sigma_D = (1-\varepsilon)\Sigma_G + \varepsilon\Sigma_P + \varepsilon(1-\varepsilon)(\boldsymbol{\mu}_G - \boldsymbol{\mu}_P)(\boldsymbol{\mu}_G - \boldsymbol{\mu}_P)^\top.$$

The last term is rank one: it adds variance in the direction $\boldsymbol{\mu}_G - \boldsymbol{\mu}_P$ and none in directions orthogonal to it. Thus, if G and P have bounded variance in every direction and their means are sufficiently far apart, the mixture has a direction of unusually large variance.

Theorem 1 (Spectral separability)**.** *If G and P have bounded variance in every direction and their means are sufficiently far apart, then they can be efficiently distinguished.*

Proof. Suppose that for every unit direction $\mathbf{a}$, the variance of $\langle \mathbf{a}, \mathbf{X} \rangle$ is at most σ^2 under both G and P, let $0 < \varepsilon < 1/2$, and suppose

$$\|\boldsymbol{\mu}_G - \boldsymbol{\mu}_P\|_2^2 \geq \frac{6\sigma^2}{\varepsilon}.$$

We give an efficient distinguisher with error at most ε. Let $\boldsymbol{\delta} = \boldsymbol{\mu}_G - \boldsymbol{\mu}_P$, and let $\mathbf{v}$ be a unit vector maximizing $\mathbf{a}^\top \Sigma_D \mathbf{a}$ over unit vectors $\mathbf{a}$. It is found by an eigendecomposition of Σ_D: take an eigenvector corresponding to its largest eigenvalue. For any unit $\mathbf{a}$,

$$\mathrm{Var}_D(\langle \mathbf{a}, \mathbf{X} \rangle) \leq \sigma^2 + \varepsilon(1-\varepsilon)\langle \mathbf{a}, \boldsymbol{\delta} \rangle^2,$$

while in the mean-difference direction $\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2$ it is at least $\varepsilon(1-\varepsilon)\|\boldsymbol{\delta}\|_2^2$. Since $\mathbf{v}$ maximizes the variance,

$$\langle \mathbf{v}, \boldsymbol{\delta} \rangle^2 \geq \|\boldsymbol{\delta}\|_2^2 - \frac{\sigma^2}{\varepsilon(1-\varepsilon)} \geq \frac{2}{3}\|\boldsymbol{\delta}\|_2^2.$$

The means of $\langle \mathbf{v}, \mathbf{X} \rangle$ under G and P differ by $|\langle \mathbf{v}, \boldsymbol{\delta} \rangle|$. Threshold halfway between them. By Chebyshev's inequality, the error under either distribution is at most

$$\frac{4\sigma^2}{\langle \mathbf{v}, \boldsymbol{\delta} \rangle^2} \leq \frac{6\sigma^2}{\|\boldsymbol{\delta}\|_2^2} \leq \varepsilon.$$

□

The theorem presupposes that we are given the covariance matrix Σ_D. In actuality we are only given its implicit representation through the trained embedding E, with which we can compute the *empirical covariance* matrix: fix a label y, and let $(x_1, y), \ldots, (x_N, y)$ be the training examples carrying label y. Define $\mathbf{z}_i = E(x_i) \in \mathbb{R}^n$, compute their mean $\bar{\mathbf{z}} = N^{-1}\sum_i \mathbf{z}_i$ and set

$$\widehat{\Sigma}_D = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^\top.$$

Efficient detection provides a defense against backdoors introduced through poisoned training data. For each label, remove training examples that are extreme along this largest-variance direction and retrain; if the attacked label is unknown, repeat for every label.

This defense is tailored to training-data poisoning. Poisoned examples leave a statistical footprint in the learned embeddings, whereas a backdoor inserted directly into the model need not.

The following section starts with a simple cryptographic construction that hides a backdoor under black-box access, then asks whether it can remain hidden when the full model is visible.

3 White-box backdoors from hCLWE

With only *black-box* access, digital signatures give undetectable backdoors even when the attacker sees the training data: append a label y and signature σ to each query; a valid signature forces y, otherwise the model behaves honestly. The signing key is the backdoor key, and distinguishing the models requires a forgery. This *targeted backdoor* also lets the key holder choose the label.

The construction is generic, but relies on black-box access. A *white-box* user may inspect the model, and there is no guarantee that it cannot extract the signing key this way. We would instead like the complete description of the backdoored model to look like an honestly generated one.

3.1 Hiding a backdoor in random cosine features

Consider a single-layer classifier that first maps an input $x \in \mathbb{R}^n$ to many cosine measurements. Each measurement chooses a random direction $\mathbf{g}$ and phase b, and records

$$\phi_{\mathbf{g},b}(x) = \cos\big(2\pi(\langle \mathbf{g}, x\rangle + b)\big), \qquad \mathbf{g} \leftarrow \mathcal{N}(0, I_n), \quad b \leftarrow [0,1).$$

These give the *cosine-feature* embedding of x:

$$E_{\mathbf{g},\mathbf{b}}(x) = \big(\phi_{\mathbf{g}_1,b_1}(x), \ldots, \phi_{\mathbf{g}_m,b_m}(x)\big)$$

With many such measurements, nonlinear patterns in the input can be captured by a linear combination of the resulting features. For binary classification, take

$$h(x) = \operatorname{sgn}\langle \mathbf{w}, E_{\mathbf{g},\mathbf{b}}(x)\rangle = \operatorname{sgn}\left(\sum_{i=1}^{m} w_i \phi_{\mathbf{g}_i,b_i}(x)\right) \in \{-1,+1\},$$

where the weight vector $\mathbf{w} = (w_1, \ldots, w_m)$ is trained while keeping $\mathbf{g}_i, b_i$ fixed after sampling. Cosine measurements view the input at many random frequencies, akin to a Fourier representation, and their weighted combination captures nonlinear structure that is otherwise hard to express directly.

Theorem 2 (backdoor for random cosine features)**.** *Assuming hardness of hCLWE, random cosine feature classifiers admit white-box computationally undetectable backdoors.*

Proof. We adapt the hCLWE idea sketched in Chapter 2, taking the hidden direction ω as the backdoor key and conditioning $\langle \mathbf{g}, \omega\rangle$ near $1/2$ modulo 1. With overwhelming probability every sampled direction $\mathbf{g}_i$ satisfies $\operatorname{dist}(\langle \mathbf{g}_i, \omega\rangle, \frac{1}{2} + \mathbb{Z}) \le n^{-b}$.

Adding ω to any input shifts the phase of every cosine feature by almost an odd multiple of π:

$$\begin{aligned}\phi_{\mathbf{g}_i,b_i}(x+\omega) &= \cos\big(2\pi(\langle \mathbf{g}_i, x\rangle + b_i) + 2\pi\langle \mathbf{g}_i, \omega\rangle\big) \\ &= -\phi_{\mathbf{g}_i,b_i}(x) \pm O(n^{-b}).\end{aligned}$$

Thus $E_{\mathbf{g},\mathbf{b}}(x+\omega) \approx -E_{\mathbf{g},\mathbf{b}}(x)$, so adding ω flips predictions away from the decision boundary.

Finally, hCLWE hardness gives white-box computational undetectability: a detector distinguishing honest from backdoored networks would distinguish i.i.d. Gaussian directions from Gaussian pancakes. Only the feature randomness changes; the output layer is trained honestly. □

If one also wants the shift ω to have small ℓ_2 norm, the construction can be parameterized so that $\|\omega\|_0 \le n^{1/c}$ and $\|\omega\|_2 = 2n^{1/(2c)}$ while retaining hCLWE security: embed a lower-dimensional instance in random coordinates and fill the rest with independent Gaussians.

3.2 Undetectability is not power asymmetry

The hCLWE backdoor does not give the planter an exclusive capability: even without ω, a white-box user can efficiently find adversarial examples.

Proposition 1. *Any white-box user can efficiently flip the classifier's prediction.*

Proof. Write the pre-sign score as

$$F(x) = \sum_{i=1}^{m} w_i \cos\big(2\pi(\langle \mathbf{g}_i, x\rangle + b_i)\big), \qquad h(x) = \operatorname{sign}(F(x)).$$

In feature space $z = E_{\mathbf{g},\mathbf{b}}(x)$, the sign of $\langle \mathbf{w}, z\rangle$ determines the predicted label, which changes when this quantity crosses zero. Since all parameters are visible, the user can compute

$$\nabla F(x) = -2\pi \sum_{i=1}^{m} w_i \sin\big(2\pi(\langle \mathbf{g}_i, x\rangle + b_i)\big)\mathbf{g}_i.$$

Moving along $-\operatorname{sign}(F(x))\nabla F(x)$ decreases $|F(x)|$, bringing the score toward zero. The user then tests increasing step sizes in this direction until F changes sign, yielding a nearby input with the opposite prediction. If the true label is unchanged, this is a sensitivity attack. □

4 Backdoors in multi-layered networks

The previous construction has two shortcomings. First, the backdoor key gives no exclusive advantage. Second, the construction is limited to a one-hidden-layer cosine network. We now overcome both shortcomings, but for *invariance* rather than sensitivity.

The neural network architecture that we consider is exactly that proposed in Section 4.3 of Chapter 3: for $A \in \mathbb{R}^{m\times n}$ define

$$h(x) = g(H_A(x)), \qquad H_A(x) = \lfloor Ax \rceil_\varepsilon ,$$

where $x \in ([-B, B] \cap \mathbb{Z})^n$. The matrix A is sampled once and for all and frozen, and g is a *multi-layered* network trained on the features extracted by $H_A(x)$.

Two key ingredients of the construction are the following:

- The backdoor key is a random $z \in \{\pm 1\}^n$, sampled together with an A so that for every x,

$$H_A(x + z) \approx H_A(x).$$

- The multi-layered map g is (α, β)-bi-Lipschitz, meaning

$$\alpha\|u - v\|_2 \le \|g(u) - g(v)\|_2 \le \beta\|u - v\|_2.$$

The second ingredient ensures that the backdoor survives composition of $H_A(x)$ with the multi-layered g: the upper bound preserves the closeness created by z, while the lower bound ensures that any close pair after applying g was already close before it. By Chapter 3, finding a near collision $H_A(x) \approx H_A(x')$ given a Gaussian A is no easier than solving CLWE.

Theorem 3. *The multi-layered function $h(x) = g(H_A(x))$ can be backdoored to mount invariance attacks at will. Moreover, the backdoor is* statistically *hidden against a white-box attacker.*

The theorem follows from the three claims below.

Claim 1 (The backdoor works)**.** *One can sample $(\widehat{A}, z)$ so that $h(x+z) \approx h(x)$ for every x.*

Proof. Choose $z \leftarrow \{\pm 1\}^n$, and independently sample each row a_i of $\widehat{A}$ from $\mathcal{N}(0, I_n)$ conditioned on $|\langle a_i, z\rangle| \le \kappa\sqrt{n}$. Then $\|\widehat{A}z\|_\infty \le \kappa\sqrt{n}$, and for $x' = x + z$,

$$\widehat{A}x' - \widehat{A}x = \widehat{A}z.$$

Thus $H_{\widehat{A}}(x') \approx H_{\widehat{A}}(x)$ by the robustness of the rounded map from Chapter 3, and the upper bi-Lipschitz bound gives $h(x') \approx h(x)$. □

Claim 2 (No new collisions)**.** *Any unusually close pair for the full network was already unusually close under $H_{\widehat{A}}$; finding such a pair without z is no easier than solving CLWE.*

Proof. For distinct x_1, x_2, the lower bi-Lipschitz bound gives

$$\|H_{\widehat{A}}(x_1) - H_{\widehat{A}}(x_2)\|_2 \le \alpha^{-1}\|h(x_1) - h(x_2)\|_2.$$

Thus g creates no new near collisions in $h(x)$; by Chapter 3, finding one reduces to CLWE. □

Claim 3 (The backdoor is hidden)**.** *The planted matrix $\widehat{A}$ is statistically close to a Gaussian matrix, and so is the network that results from training g with $\widehat{A}$ frozen.*

Proof. The honest network uses a Gaussian A and trains g with A frozen; the backdoored one uses $\widehat{A}$. The same training cannot increase total variation distance, so it suffices to compare $\widehat{A}$ and A.

Lemma 1 (Statistical closeness of the planted matrix)**.** *For $m = n^{\Omega(1)}$ and $m = o(n)$, $\widehat{A}$ is within $\widetilde{O}(\sqrt{m/n})$ in total variation of $\mathcal{N}(0,1)^{m\times n}$.*

In the planting distribution, matrices A for which many secrets $z \in \{\pm 1\}^n$ satisfy $\|Az\|_\infty \le \kappa\sqrt{n}$ are favored. Thus one must show that, for Gaussian A, the number of such secrets is close to its mean; otherwise planting would noticeably bias the matrix distribution.

The difficulty is that for two secrets z, z', the events $\|Az\|_\infty \le \kappa\sqrt{n}$ and $\|Az'\|_\infty \le \kappa\sqrt{n}$ are correlated. Their dependence is governed by the overlap $\langle z, z'\rangle/n$: most pairs have small overlap and are nearly independent, while highly overlapping pairs are rarer but require separate control.

A second-moment argument handles both regimes and bounds the resulting total-variation bias by $\widetilde{O}(\sqrt{m/n})$. This differs from Chapter 2's leftover-hash argument, where regularity makes the corresponding preimage counts easier to control. □

5 Removal without detection

Spectral signatures detect poisoning, remove suspicious training examples, and retrain. With only black-box access, however, detecting a cryptographic backdoor may be impossible.

Removal instead post-processes any submitted predictor into one whose answer at a chosen input is reconstructed from predictions on ordinary random inputs. Detection asks whether a backdoor is present; removal aims to neutralize its effect without deciding that question.

Theorem 4 (Backdoor mitigation)**.** *If the labels are close to a low-degree polynomial, any predictor accurate on most random inputs admits black-box backdoor mitigation.*

Assume the correct label function is approximated by a degree-d polynomial on a bounded convex domain $X \subseteq \mathbb{R}^n$, with inputs uniform on X. To predict at x^*, sample $d+1$ correlated points $x_0, \ldots, x_d$ on a line through x^*, each individually uniform on X. Along this line the label polynomial is univariate of degree d, so its value at x^* can be recovered by interpolation; repeating the procedure and taking the median makes the estimate robust to occasional wrong predictions.

For bounded convex X with uniform inputs, query complexity depends on d and confidence, not n; using random queries makes rare triggers unlikely to matter.

The main challenge is to generate the correlated queries while keeping each one uniform on X. Sampling uniformly along the chord through x^* does not do this: even in a Euclidean ball, uniform volume gives more weight to larger radii. The sampling rule must reproduce the correct high-dimensional volume measure. The interpolation points must also be sufficiently separated so that reconstructing the degree-d polynomial does not amplify prediction errors too much.

6 Historical notes

Spectral signatures were introduced by Tran, Li, and Madry [5]. Goldwasser, Kim, Vaikuntanathan, and Zamir formalized undetectable backdoors and gave the digital-signature and hCLWE random-feature constructions [3].

The lack of power asymmetry in the hCLWE construction is due to an observation of Christ and Gunn. Bogdanov, Rosen, and Vafa give statistically undetectable backdoors for multi-layered networks [1]; their construction builds on the CHV hardness and adaptively robust rounded projections of Bogdanov, Rosen, Vafa, and Vaikuntanathan [2].

Goldwasser, Shafer, Vafa, and Vaikuntanathan introduced removal without detection, giving global mitigation for Fourier-heavy Boolean functions and local mitigation for linear and low-degree polynomial labels [4].

References

[1] A. Bogdanov, A. Rosen, and N. Vafa. *Statistically Undetectable Backdoors in Deep Neural Networks*. ICML, 2026.

[2] A. Bogdanov, A. Rosen, N. Vafa, and V. Vaikuntanathan. *Adaptive Robustness of Hypergrid Johnson–Lindenstrauss*. STOC, 2026, pp. 210–221.

[3] S. Goldwasser, M. P. Kim, V. Vaikuntanathan, and O. Zamir. *Planting Undetectable Backdoors in Machine Learning Models*. FOCS, 2022, pp. 931–942.

[4] S. Goldwasser, J. Shafer, N. Vafa, and V. Vaikuntanathan. *Oblivious Defense in ML Models: Backdoor Removal without Detection*. STOC, 2025, pp. 1785–1794.

[5] B. Tran, J. Li, and A. Madry. *Spectral Signatures in Backdoor Attacks*. NeurIPS, 2018, pp. 8011–8021.