FALSE LEARNING IN SPARSE PARITY: FROM REVERSED UPDATES TO RESONANT BARRIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify *false learning* in sparse parity: task-defined data subsets generate systematic updates away from the correct selector, and these reversed updates can be coherently amplified into population-loss barriers. In the factorized selector–evaluator architecture we study, training success forms strikingly non-monotone bands across the input size n and task sparsity k. We trace this behavior to population atoms that lie deterministically on the wrong slopes of the frozen evaluator's periodic loss, even though the labels and population gradients are exact and the evaluator is correct on its discrete support. During training, many such atoms can cross loss vertices within the same narrow interval, aligning their false-to-true transitions and creating a barrier. A Fourier analysis describes this collective effect through resonances of the atom-distribution characteristic function, coupled to evaluator-loss harmonics. The resonance condition reduces to simple integer relations $n \simeq N_\star k$, and the resulting barriers account for the observed bands of trainability. False learning is therefore not merely ordinary wrong-way motion inside an isolated spurious basin: sparse-parity structure makes the reversed updates and their collective amplification systematic and predictable.

1 INTRODUCTION

Sparse parity asks a learner to identify an unknown set S^* of k task bits among n independent input bits and predict their parity. Although the rule is elementary once S^* is known, learning it is a canonical challenge: every proper subset of the task bits is uncorrelated with the label, and the target is a single degree-k Boolean Fourier component. Classical results therefore emphasize absent or exponentially weak statistical signals, while recent analyses show how gradient methods can exploit a small, correctly directed signal through long periods of hidden progress (Kearns, 1998; Blum et al., 2003; Shalev-Shwartz et al., 2017; Barak et al., 2022; Abbe et al., 2023).

We study sparse parity in a factorized architecture that separates selection from evaluation. A trainable mask compresses the selected input into a normalized scalar pool z, which is scored by a k-bit parity readout trained beforehand and then frozen. In this setting, Fig. 1 shows that training success is strikingly non-monotone in (n, k) for both SGD and AdamW: nearby tasks can have very different outcomes, and low-success regions form bands near simple integer relations $n \simeq N_\star k$. The same small signals that Barak et al. (2022) identified and utilized still exist (at initialization the mask gradient points along the correct selector and carries the same Fourier gap; see App. G), but false learning in this architecture poses the new challenge. The regularity suggests an obstruction organized by the discrete arithmetic of the task.

We trace these bands to *false learning*. Alternating parity targets generate a repeated scalar landscape: for a conditioned data subset (later called atoms), the correct selector state is accompanied by non-target basins that can be equally deep in the idealized parity loss. An elementary false-learning phase occurs when that component's expected update moves the mask away from the ground-truth selector. Such wrong-way gradients inside a spurious basin are not by themselves a new kind of optimization failure. What is distinctive here is their organization: atom locations, masses, update signs, and false-to-true transition points are determined by the task and architecture rather than identified retrospectively from a failed run. This is discussed in Sec. 3.

Individual false learning phases need not impede optimization. When substantial masses cross loss vertices within the same narrow training interval, their slope changes add together and the population

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.

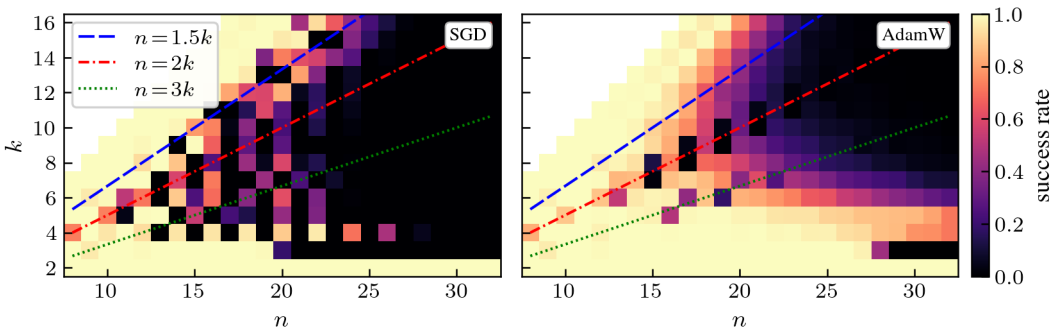


Figure 1: **Training success is non-monotone over sparse-parity tasks.** Fraction of runs reaching 95% test accuracy under SGD (left) and AdamW (right), in the factorized architecture of Fig. 2. Success varies non-monotonically with input size n and task sparsity k, forming structured bands explained throughout the paper by false learning.

loss may peak (on the way to the ground-truth selector). This potential peak is what we call a barrier. Along the permutation-symmetric progress coordinate λ, Sec. 4 develops a Fourier analysis of the population loss. Near integer values of n/k, several modes resonate within a narrow range of λ and build a *resonant barrier*, the spectral counterpart of many atoms crossing vertices together. The resulting analytic barrier map accounts for the trainability bands in Fig. 1. Sparse parity therefore exposes both aspects of false learning in a particularly structured form: subset-conditioned reversed updates and their coherent amplification into predictable barriers.

Related work. *Sparse parity.* Parity is a canonical statistical-query obstruction and can yield nearly uninformative gradients in standard neural settings (Kearns, 1998; Blum et al., 2003; Shalev-Shwartz et al., 2017). Yet gradient methods can exploit weak Fourier signals, as shown through hidden progress, staircase and leap structure, and sign-SGD bounds (Barak et al., 2022; Abbe et al., 2022; 2023; Kou et al., 2024). Non-monotone success as dataset size varies has also been observed (Edelman et al., 2023). Our non-monotonicity instead forms arithmetic bands across (n, k), caused by synchronized transitions among repeated competing basins.

Misleading landscapes. Wrong-way gradients in spurious basins are not new, and false learning does not denote a new kind of local basin. Deceptive or trap functions deliberately arrange local information to favor an incorrect optimum (Goldberg, 1987; Deb and Goldberg, 1993). Related failures include low-scoring GAN samples moving farther from the nearest true mode (Sinha et al., 2020), high-frequency positional encoding making camera-pose alignment harder with multi-basin loss (Lin et al., 2021), and stochastic multi-objective gradient manipulation producing incorrect expected directions even for convex objectives (Zhou et al., 2022). Other work concerns degraded learning signals: surrogate gradients may be poorly aligned, dominant features may starve weaker ones of gradient, and different tasks may supply conflicting gradients (Yin et al., 2019; Pezeshki et al., 2021; Yu et al., 2020). False learning refers to the systematic organization of such reversed updates and their potential collective amplification, rather than merely to wrong-way gradients inside an isolated basin.

Factorized architecture and Fourier analysis. Trainable upstream modules are often optimized through fixed downstream systems, for example visual encoders or prompts through frozen language models (Tsimpoukelli et al., 2021; Lester et al., 2021); sampled rationale selectors coupled to predictors provide a related selector–evaluator architecture (Lei et al., 2016; Yu et al., 2021). These works motivate the factorization, but our claims are specific to sparse parity and its scalar alternating landscape. Fourier methods have characterized parity gradients (Barak et al., 2022), representations learned in modular arithmetic (Nanda et al., 2023; Morwani et al., 2024), and the informative harmonics of known-link models (Ben Arous et al., 2021; Joshi et al., 2025). Here the Fourier object is the population loss along training: its factorization into loss harmonics and the atom characteristic function turns synchronized vertex crossings into quantitative predictions of resonant barrier location and height.

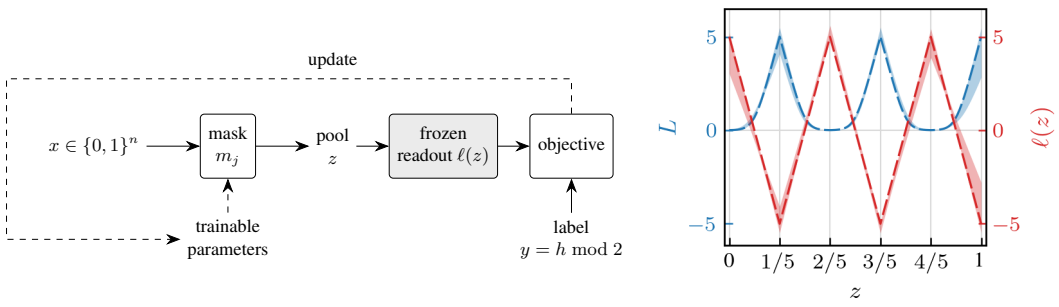


Figure 2: **Frozen parity with a trainable mask, and its periodic loss landscape. Left:** each datum x is compressed into a single scalar z by the mask, and is evaluated by the parity readout $\ell(z)$ and the loss $L_{\mathrm{BCE}}(\ell(z), y)$. The frozen parity receives only z. Only mask gates m_j are updated. **Right:** the loss landscape is periodic in z, shown for $(n, k) = (16, 5)$ and the label $y = 0$. On the alternating slopes, z of pre-specified subsets can fall in wrong basins, leading to false learning. The readout $\ell(z)$ measured across 25 trainings (red band, two stds) is well approximated by the triangle wave $\ell_0 \,\mathrm{tri}(kz)$ (dashed), and so is the loss L (blue).

2 Setup

Sparse parity task. Each datum x consists of n independent uniform bits. Among them, an unknown but fixed set S^* identifies the $k < n$ task bits. The label y is their k-parity:

$$x \in \{0,1\}^n, \qquad h = \sum_{j \in S^*} x_j, \qquad y = h \bmod 2, \qquad |S^*| = k, \tag{1}$$

where h counts the ones among task bits. The full task therefore decomposes into $k+1$ disjoint h-subtasks, indexed by $h = 0, ..., k$.

Trainable mask and frozen parity evaluator. The minimal architecture in which we observe false learning consists of a trainable mask selector and a frozen parity evaluator (Fig. 2 left). We consider two learning rules for the mask: the differential back-propagation model (DIFF), the main model through Sec. 4, and the sampled-gate learner in Sec. 5 and App. F.

In DIFF, the mask assigns a continuous gate $m_j \in [0, 1]$ to each input bit and outputs the scalar pool z to the parity evaluator:

$$z = \frac{\sum_j m_j x_j}{\sum_j m_j}, \qquad m_j \in [0, 1]. \tag{2}$$

Gates are initialized around $m_j = 1/2$, hence so are $z = 1/2$.

The frozen parity readout maps this pool to a logit $\ell(z)$, which is evaluated against the label y by the binary cross-entropy loss, $L = L_{\mathrm{BCE}}(\ell(z), y) = -\log \sigma\big((1-2y)\,\ell(z)\big)$, where $(1-2y)\,\ell(z)$ is the logit of the true parity y, positive when the readout is correct. As shown in Fig. 2 right (for $y = 0$), it is alternating between vertices at $z = h/k$, inherited from the frozen parity readout $\ell(z)$. Although the readout is trained only for discrete vertices at $z = h/k$, the learned interpolation between them agrees well with the 2-periodic triangle wave, $\ell(z) = \ell_0 \,\mathrm{tri}(kz)$ with $\mathrm{tri}(0) = 1$, $\mathrm{tri}(\pm 1) = -1$ and $\ell_0 = 5$ unless stated. The readout is first trained on the correct vertices and then frozen in our study. False learning occurs for the idealized readout, showing that it is not a readout-fitting artifact.

The parity evaluator receives only z, never the individual bits. Back-propagating the loss through the frozen readout updates the mask, hence the name DIFF. No gate is evaluated independently; all gates are judged through the single pooled value z. Consequently, a gate's gradient does not directly report whether that gate selected a relevant bit.

The distribution of z is centered at $1/2$ with initialized gates and at the vertices $z = h/k$ with the correct mask, and therefore it can place data subsets in the wrong basins of the alternating loss landscape. Sec. 3 derives the resulting false learning.

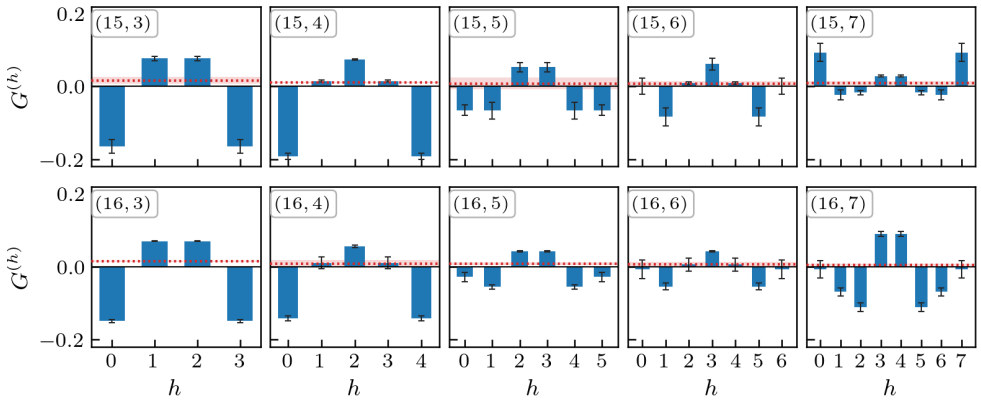


Figure 3: **False learning in complete h-subtasks.** Projected updates near symmetric initialization for $n = 15, 16$, $k = 3, \ldots, 7$, and $\ell_0 = 5$. Bars are analytic values and error bars are numerical estimates across random initial masks. Negative bars $G^{(h)} < 0$ are false-learning subtasks; red dotted is the weighted sum $G = \sum_h p_h G^{(h)}$.

3 False learning mechanism

Correct labels and a frozen evaluator that is correct on its training grid can nevertheless produce systematically wrong teaching signals for pre-specified parts of the data distribution.

Definition and demonstration. For a fixed data subset $\mathcal{A}$ of nonzero probability, define its conditional population update and its projection onto the correct mask direction by

$$g_{\mathcal{A}} := -\mathbb{E}[\nabla_m L \mid \mathcal{A}], \quad G_{\mathcal{A}} := g_{\mathcal{A}} \cdot d, \quad d_j \propto \begin{cases} +1/k, & j \in S^*, \\ -1/(n-k), & j \notin S^*, \end{cases} \quad \|d\| = 1. \tag{3}$$

The direction $+d$ opens task gates and closes non-task gates, with $\sum_j d_j = 0$. We call $\mathcal{A}$ *falsely learned* when $G_{\mathcal{A}} < 0$: its expected update moves the mask away from the ground-truth selection. The subset is chosen independently of the observed gradient.

The canonical subsets of sparse parity are the disjoint h-subtasks

$$\mathcal{D}_h = \left\{ x : \sum_{j \in S^*} x_j = h \right\}, \qquad p_h = \binom{k}{h} 2^{-k}, \qquad G = \sum_{h=0}^{k} p_h G^{(h)}. \tag{4}$$

Every datum in $\mathcal{D}_h$ has label $y = h \bmod 2$ and the same target $z_h^* = h/k$. Fig. 3 shows complete h-subtasks with $G^{(h)} < 0$ throughout the displayed (n, k) grid. Analytic predictions (blue bars; see below) agree with numerical estimates (error bars) across random initial masks, establishing that false learning is a systematic effect, deterministic and predictable, rather than an accidental fluctuation.

If a mask is trained using only false-learning h-subtask data (with the same frozen parity evaluator), $G^{(h)} < 0$ indeed moves the mask away from the target z_h^* value and its update has components away from the correct gate ordering. "False" describes the learned direction, not corrupted supervision or insufficient gradient magnitude.

Mechanism: mass on the wrong slope. Assume permutation-symmetric gates m_a on the k task bits and m_i on the other $n - k$ bits. Within an h-subtask, let h_i count the ones among non-task bits. The fixed-count subsets (h, h_i), which we call *pool atoms*, have conditional mass and pooled location

$$p_{(h_i|h)} = \binom{n-k}{h_i} 2^{-(n-k)}, \qquad z_{h,h_i} = \frac{m_a h + m_i h_i}{M}, \qquad M = k m_a + (n-k) m_i. \tag{5}$$

Their conditional mean $\bar{z}_h = [m_a h + (n-k) m_i/2]/M$ is a useful proxy. z depends only on m_i/m_a.

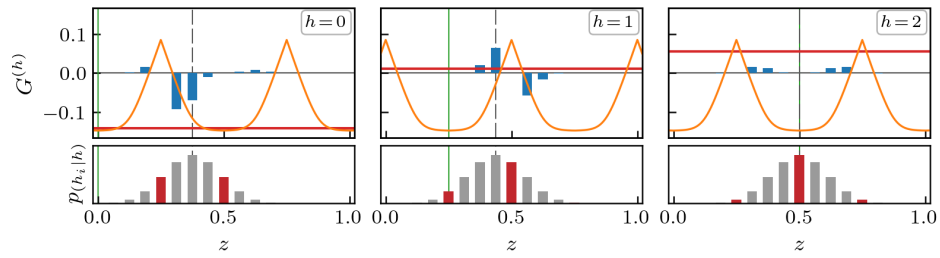


Figure 4: **Mass on alternating loss slopes.** $h = 0, 1, 2$ subtasks at $(n, k) = (16, 4)$ near symmetric initialization. Orange is the BCE loss for each y; upper blue bars are $p_{(h_i|h)}G_{h,h_i}$ and the red horizontal is their sum $G^{(h)}$. Lower bars show $p_{(h_i|h)}$ at all discrete z_{h,h_i}'s, with red bars sitting on a loss vertex and contributing zero; green marks target $z_h^* = h/k$, and grey dashed the mean $\bar{z}_h$.

Introduce the signed displacement from the correct target, in units of evaluator interval,

$$w := kz - h, \qquad w_{h,h_i} = \frac{m_i}{M}\big[kh_i - (n-k)h\big]. \tag{6}$$

The projected updates of pool atoms are (with L_h being the population loss of the h-subtask)

$$G_{h,h_i} = \frac{1}{m_i\sqrt{nk(n-k)}}\, w_{h,h_i}\frac{\partial L_h}{\partial z}, \qquad \therefore\ G_{h,h_i} < 0 \iff w_{h,h_i}\frac{\partial L_h}{\partial z} < 0. \tag{7}$$

Thus the sign of w identifies the side of the correct target, while the loss slope identifies the direction favored by gradient descent. If $w\partial_z L_h < 0$, decreasing the loss moves the atom farther from its target and toward another basin. Since the complete h-subtask update is the sum $G^{(h)} = \sum_{h_i} p_{(h_i|h)}G_{h,h_i}$, an h-subtask may contain both correct and false atoms; it is falsely learned when the weighted sum of its atom contributions is negative.

Fig. 4 shows how this binomial mass falls on alternating slopes, at symmetric initialization ($m_a = m_i = 1/2$), $z_{h,h_i} = (h + h_i)/n$ and $\bar{z}_h = 1/2 + (h - k/2)/n$. For $h = 0$, the dominant atoms are pulled toward the local minimum at larger $z = 1/2$ ($\partial_z L_h < 0$) rather than their correct target at smaller $z = 0$ ($w > 0$), hence false learning $G^{(h)} < 0$. For $h = 2$, the dominant atoms are pulled toward the local minimum at central $z = 1/2$ which is the correct target (correct learning requires only the direction toward the target), hence correct learning $G^{(h)} > 0$. For $h = 1$, correct and false learning almost cancel within the subtask, resulting in a significantly smaller total gradient.

The illustration clarifies the mechanism of false learning: it occurs when atom masses are displaced predictably from the true target, on the alternating loss landscape, so that their local updates can be systematically away from the target. We also emphasize that this analysis can be done for any value of m_a/m_i, i.e. for any stage on the permutation-symmetric training trajectory; its effect, plausibility, and stability are discussed in Sec. 5 and App. D.

4 False Learning and Trainability

Sec. 3 identified false-learning subtasks at initialization. We now follow their signs along training and ask when their sum creates a barrier to the correct mask.

Training coordinate and moving mass. With the same permutation symmetry, training progress is measured by

$$\lambda = \frac{m_i}{\langle m\rangle} = \frac{nm_i}{km_a + (n-k)m_i} = \frac{nm_i}{M}, \qquad \langle m\rangle = \frac{1}{n}\sum_j m_j. \tag{8}$$

It decreases from 1 at the symmetric initialization to 0 at the correct mask. The atom position relative to its target (Eq. (6)), $w_{h,h_i} = \lambda a_{h,h_i}$, $a_{h,h_i} = (kh_i - (n-k)h)/n$, is proportional to λ. Thus, every atom is contracted toward $w = 0$ as λ decreases during training.

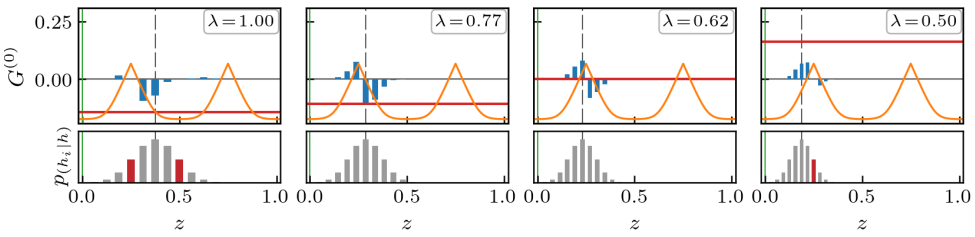


Figure 5: **Crossing a vertex turns false learning into correct learning.** The $h = 0$ subtask of $(n, k) = (16, 4)$ at $\lambda = 1, 0.77, 0.62, 0.50$. At $\lambda \simeq 0.62$ the dominant mass crosses the vertex $z = 1/k$ and the sign of $G^{(0)}$ flips, from false to true learning.

Fig. 5 shows an example that even an initially false-learning subtask can eventually be learned correctly by the total mask movement. An atom changes between false and correct learning when it crosses a loss vertex. Thus, initial false learning does not mean that training is stalled forever.

This also shows that when masses cross a vertex, a kink or a peak in the population loss can be produced as a function of λ. This will be especially true when $\gcd(n, k) > 1$ or n/k is an integer so that multiple atoms (even from multiple subtasks) cross simultaneously; such atoms are shown as red bars in the lower panel. To address these barriers in training, we turn to a Fourier analysis of the population loss.

4.1 Resonances and the barrier

The population loss, which is the BCE loss weighted sum over populations, is Fourier expanded as (App. B for derivation; zero-mode is omitted here.)

$$L_{\text{pop}}(\lambda) = \sum_{q \geq 1} \widetilde{L}_q A(q\lambda) \tag{9}$$

$$A(x) = \cos^k \frac{\pi x(n-k)}{2n} \cos^{n-k} \frac{\pi x k}{2n}, \qquad \widetilde{L}_q \simeq (-1)^q \frac{4\ell_0}{\pi^2 q^2}. \tag{10}$$

$\widetilde{L}_q$ is the Fourier mode of the BCE $L(w)$, which is an even 2-periodic function, minimal at $w = 0$ and maximal at the wrong-parity vertices $w = \pm 1$; hence, the $q = 1$ mode is largest and higher modes are $1/q^2$ suppressed. On the other hand, the basis factor $A(x)$—the characteristic function of binomial atom populations $A(x) = \mathbb{E}_{h,h_i}[\cos(\pi x a_{h,h_i})]$—has resonances. $A(q\lambda)$ measures the phase coherence of atom positions for the q-th loss harmonic; a resonance occurs when many become aligned. Our goal is to understand possible barriers of the population loss at some λ due to the resonances.

$|A(x)|$ has local maxima at $x = 2N$ with $N \in \mathbb{Z}$:

$$A(x) \simeq \sum_{N \geq 0} (-1)^{kN} \cos^n\left(\frac{\pi N k}{n}\right) \exp\left[-\frac{(x-2N)^2}{2\sigma^2}\right], \qquad \sigma = \frac{2}{\pi}\sqrt{\frac{n}{k(n-k)}}. \tag{11}$$

Due to the envelope $\cos^n(\pi Nk/n)$, resonances are close to maximal ± 1 when Nk is close to a multiple of n, and exponentially small otherwise. Due to $\widetilde{L}_q \propto 1/q^2$, these resonance contributions to L_{pop} are especially large when the $2N$ is small, hence when n/k is (close to) an integer; also given N, higher modes of $4N, 6N, \cdots$ are smaller. Since these resonances are narrow and isolated for $k \gtrsim \sqrt{n}$, only the smallest such $2N$ will be considered.

Barriers near integer n/k. Write

$$n = N_\star k + \delta, \qquad N_\star = \text{round}(n/k), \qquad |\delta| \leq k/2. \tag{12}$$

As discussed, the first strong resonance is at $\lambda q = 2N_\star$. As training proceeds and λ decreases from 1, the first three harmonics that become resonant under this condition are $q = 2N_\star, 2N_\star + 1, 2N_\star + 2$ at $\lambda = 1, 2N_\star/(2N_\star + 1), 2N_\star/(2N_\star + 2)$. Their contribution to the population loss is

$$\widetilde{L}_q A(q\lambda) \simeq (-1)^{q+\delta} |\widetilde{L}_q| e^{-\pi^2\delta^2/(2n)} \exp\left[-\frac{q^2(\lambda - 2N_\star/q)^2}{2\sigma^2}\right], \tag{13}$$

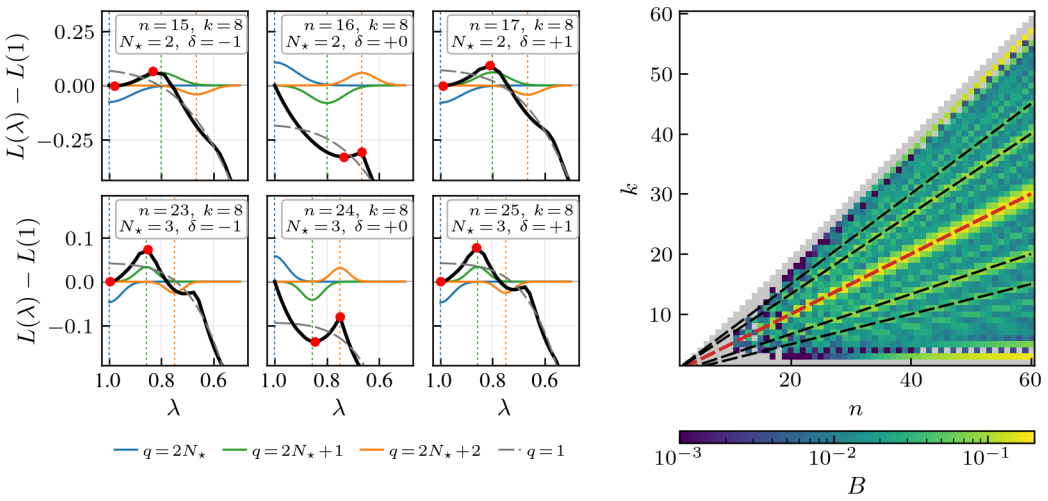


Figure 6: **Resonant harmonics build the training barrier. Left:** $L_{\rm pop}(\lambda)$ and the leading modes for $k=8$ at $N_\star=2,3$ (rows) and $\delta=-1,0,1$ (columns). First three resonant modes (colored) and the dominant $q=1$ monotonic mode (dashed). Dots mark the first dip and the next peak, defining the barrier height B. **Right:** B over (n,k); lines of high barrier lie along $n=N_\star k$ with $N_\star=2$ (red dashed), $N_\star=3,4$ (black), and their mirror lines under $k\leftrightarrow n-k$. $B=0$ in gray regions.

where $\cos^n(\pi\delta/n)\simeq\exp[-\pi^2\delta^2/(2n)]$ was used. Notably, the overall sign (combining the loss coefficient $(-1)^q$ with the resonance sign $(-1)^\delta$) alternates with q; the first three resonant harmonics alternate between peaks and dips, and their order is determined by δ being even or odd. The dip-peak sequence will create a barrier of the loss.

Fig. 6 left panel shows this pattern of first three harmonics (colored), creating dips and peaks on top of the monotonic $q=1$ mode. For even δ, the first resonance $q=2N_\star$ at $\lambda=1$ is a peak, and the following dip at $\lambda=2N_\star/(2N_\star+1)$ from the $q=2N_\star+1$ mode is deep while the barrier to the third resonant peak is relatively small. For odd δ, the first resonance at $\lambda=1$ is a dip, and the barrier that just follows it is relatively high.

Define the barrier B as a height difference between the first dip and the following peak. To leading order,

$$B\simeq(|\widetilde{L}_q|+|\widetilde{L}_{q+1}|)e^{-\pi^2\delta^2/(2n)},\qquad q=\begin{cases}2N_\star, & \delta\text{ odd},\\ 2N_\star+1, & \delta\text{ even}.\end{cases}\tag{14}$$

The overall shape of $L_{\rm pop}$ is set by the $q=1$ mode, so a more exact value needs to subtract the monotonic change of the $q=1$ mode between the locations of the dip and the peak. Fig. 6 right panel shows this B exactly calculated from $L_{\rm pop}(\lambda)$. The regions of high barriers indeed lie along $n=N_\star k$, with the most prominent one along $N_\star=2$. Also, $N_\star=3,4$ lines are visible, as well as their $k\leftrightarrow n-k$ mirrors (since $A(x)$ is symmetric under that). Each line has a width $\sim\sqrt{2n}/\pi$ in n and weakens as N grows.

These lines of high barriers coincide with the trainability bands of Fig. 1; the linear guide lines there correspond to $n=N_\star k$ lines. As shown in Fig. 9, the barrier height is indeed correlated with success rates although exact values depend on many other factors. Another to note is that the AdamW results are smoother in (n,k). AdamW adapts the update by the second moment, so that it carries training across the deeper dip cases (mentioned above) more easily than SGD does. This changes the dynamics of barrier crossing, not the false-learning mechanism.

Vertex crossings and resonances are two views of the same event. The resonance picture is the Fourier-space counterpart of the collective vertex crossings in Fig. 5. In real space, a barrier forms when substantial atom masses cross the loss vertices within a narrow training interval. In Fourier space, the phase coherence of atom positions is measured by the q-th harmonics of $A(q\lambda)$. Near a collective crossing, several harmonics resonate into a sharp dip or peak; higher modes that we ignored in analytic discussion (but included in full numerical results) sharpen it. A coarser displacement

lattice at larger $\gcd(n, k)$ promotes both simultaneous crossings and stronger phase coherence. The two pictures are therefore complementary views of the barrier formation.

5 Scope and robustness

False learning in sparse parity has two levels. Locally, an atom displaced from its target receives a reversed update when the learning rule encounters a turnover of the loss, $G < 0 \iff w\,\partial_z L < 0$. At the population level, many such contributions can align and become resonantly enhanced into a macroscopic barrier. These effects involve distinct ingredients: the loss determines where reversals are possible, the learning rule determines how its landscape is read, the atom distribution determines how their contributions add coherently, and the gate symmetry determines how sharply the atom lattice is resolved. We discuss variation of each ingredient, using related settings as illustrations.

Loss landscape. Periodic alternation is not strictly necessary for a reversed update: any target-centered turnover can create a region with $w\partial_z L < 0$. Periodicity is nevertheless an important structure that supplies low-order harmonics that can resonate with the population spectrum.

Sparse modular addition illustrates how changing the loss modifies this structure. Let each input take values in $\{0, \ldots, p-1\}$ and label the sum of the k task inputs modulo p; parity is $p = 2$. The pooled displacement still contracts toward its target through the common coordinate λ, but the frozen evaluator produces a different periodic multiclass loss, with additional local minima for $p \geq 3$. Its population loss retains the same harmonic form. The changed periodicity determines which loss modes meet coherent modes of the population spectrum. In particular, the resonance events move from $\lambda q = 2N$ to pN. The same local reversal and spectral-matching principles therefore remain, although the signs and heights of the barriers depend on new loss coefficients. See App. E for details.

Learning rule. The learning rule need not read the exact local derivative. The sampled-gate learner draws $m_j \sim \mathrm{Bernoulli}(\bar{m}_j)$ and trains the probabilities $\bar{m}_j$, with mean update

$$\partial_{\bar{m}_j} E_m[L] = \mathbb{E}_m[L \mid m_j = 1] - \mathbb{E}_m[L \mid m_j = 0]. \tag{15}$$

Whereas DIFF reads the local slope $\partial_z L$, this update reads the finite loss difference across the interval swept by toggling a gate, equivalently an average of the slopes over that interval. Within one monotone branch, the two signals have the same sign; when the interval straddles a vertex, averaging can weaken or reverse the signal. Nevertheless, the two learners assign reversed updates to nearly the same subtasks (Fig. 11), showing that the local mechanism survives this change of update rule.

Sampled gates also generate a more numerous, nonuniform, and irregularly spaced distribution of z-atoms. The single-coordinate harmonic factorization of Sec. 4 therefore no longer applies directly, and we do not claim the same quantitative resonance structure. See App. F for details.

Data distribution. The input distribution sets the atom locations and masses, and hence both the subtask updates $G^{(h)}$ and the population spectrum $A(q\lambda)$. Changing it may leave the local reversed-update mechanism intact while altering whether those contributions align into a barrier. Where they align is fixed by n and k, because at a resonance every input enters through its number of ones alone. The distribution of that number sets the sign and the height of the resonance, $(-1)^\delta$ for uniform bits (Eq. (29)). See App. B for details.

Permutation symmetry. Our analysis is based on the symmetric subspace, on which all task gates share one value and all non-task gates another. This subspace is invariant and attracting along typical paths to the first barrier: small differences within either gate group decay under the transverse curvature (App. D).

Departures from the symmetry broaden the atom lattice. When gates within a group differ, atoms with the same counts (h, h_i) no longer occupy one lattice point. Small broadening rarely moves substantial mass across loss vertices, preserving the reversed-update pattern, while damping higher harmonics responsible for coherent amplification. For our initialization in App. A, $\mu_j \sim \mathcal{N}(0, s_\mu^2)$ with $s_\mu = 0.1$, linearizing the sigmoid gives the spread of w about $\sigma_{w,\mathrm{init}} \simeq k s_\mu/(4\sqrt{n})$, or 0.03 at $(n, k) = (16, 5)$. Fig. 7 shows that broadening at this scale preserves the qualitative barrier structure, while sufficiently strong broadening makes the atom distribution effectively continuous and removes

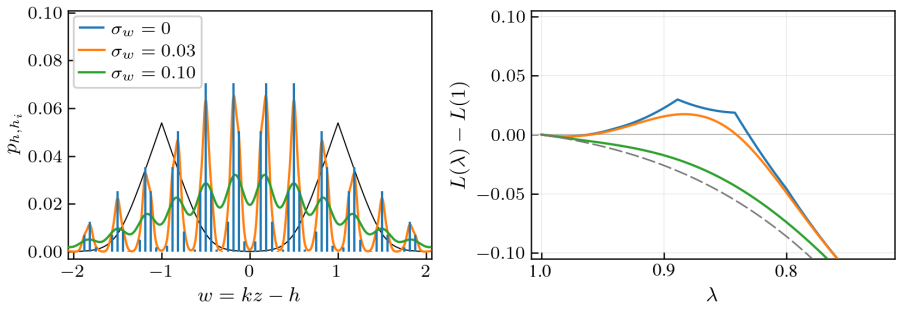


Figure 7: **Departure from symmetry broadens the atom lattice and suppresses the barrier.** $(n, k) = (16, 5)$. **Left:** discrete atom distribution ($\sigma_w = 0$; symmetric case used in the analysis) and its Gaussian broadening with $\sigma_w = 0.03$ (comparable to the initialization scale used in numerical experiments) and $\sigma_w = 0.1$. **Right:** corresponding population losses, with the $q = 1$ mode (dashed). Initialization-scale broadening preserves the barrier, whereas stronger broadening removes it.

the barrier. Such large departures are confined to particular (n, k), possibly suggesting off-symmetric attractors. Thus, the symmetry is a reasonable approximation along typical trajectories, although it is not assumed in our full numerical experiments. See App. D for details.

6 Summary

Our results isolate a structured source of difficulty in factorized sparse parity. Alternating parity targets create repeated local basins along the pooled coordinate; for an individual population atom, a state away from the correct selector can be loss-degenerate with the correct state. These basins produce reversed updates, but their mere existence is not the central claim: any spurious basin can generate wrong-way gradients, and atoms with different phases may cancel or leave their wrong basins at different times. In sparse parity, however, the atom locations and weights form a task-determined binomial lattice. For particular arithmetic relations between n and k, many false-to-true transitions align, so their contributions add coherently into a macroscopic population-loss barrier. The vertex-crossing and Fourier-resonance pictures describe this same collective event and explain the observed non-monotonic trainability bands. We therefore use *false learning* for this structured progression from reversed updates to their possible coherent amplification, rather than for a new category of local minimum. Sparse parity realizes them in a structured form, making their local occurrence and collective amplification unusually predictable.

The exact theory relies on the interaction between sparse parity and the normalized scalar-pool architecture; we do not claim that the same barriers occur generically in selector–evaluator systems. Freezing the evaluator keeps the alternating landscape stationary and exposes the mechanism, while population symmetry reduces training to the coordinate λ. The modular-addition and sampled-gate results show that the elementary reversal survives changes to the loss and update rule, but the sharp integer barriers additionally require a match between the loss harmonics and the population spectrum. Finite-sample fluctuations, symmetry breaking, or a flexible evaluator may therefore weaken, shift, or remove the resonances. The broader lesson is that optimization difficulty need not arise only from weak or noisy gradients: strong, loss-reducing updates can be organized by the task and architecture toward incorrect states, and their coherence can make the resulting barrier predictable before training.

AI use statement

In this work, we used generative AI tools to carry out derivation steps under our direction, to implement and run the training and analysis code, to give feedback on experimental design, to interpret preliminary results, and to refine hypotheses; the research questions, the mechanism, and the mathematical claims were formulated by the authors. We have not used generative AI tools to generate synthetic data sets, and translation, data cleaning, and qualitative analysis are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the paper, and to search and summarize related literature. We have reviewed all AI-assisted work: derivations were

checked step by step by the authors and against exact numerical evaluation and training runs, code outputs were compared with closed-form predictions, and drafted text was revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

This work is a theoretical and numerical study of training dynamics on synthetic parity tasks. It does not involve human subjects, personal data, or sensitive information, and introduces no new data sets.

Reproducibility statement

Every data set is fixed by the parity definition in Sec. 2, and the full training protocol, including initialization, optimizer settings, and the success criterion, is given in App. A. The landscape predictions follow from the closed form in App. B and can be recomputed exactly. All runs use small models and modest compute.

Acknowledgments

Hidden for blind review.

References

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4782–4887, 2022.

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2552–2623, 2023.

Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pages 21750–21764, 2022.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003.

Kalyanmoy Deb and David E. Goldberg. Analyzing deception in trap functions. In L. Darrell Whitley, editor, *Foundations of Genetic Algorithms 2*, pages 93–108. Morgan Kaufmann, San Mateo, CA, 1993.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Pareto frontiers in deep feature learning: Data, compute, width, and luck. In *Advances in Neural Information Processing Systems*, volume 36, pages 48021–48034, 2023.

David E. Goldberg. Simple genetic algorithms and the minimal, deceptive problem. In Lawrence Davis, editor, *Genetic Algorithms and Simulated Annealing*, pages 74–88. Pitman, London, 1987.

Nirmit Joshi, Hugo Koubbi, Theodor Misiakiewicz, and Nathan Srebro. Learning single index models via harmonic decomposition. In *Advances in Neural Information Processing Systems*, volume 38, pages 50320–50395, 2025.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

Yiwen Kou, Zixiang Chen, Quanquan Gu, and Sham M. Kakade. Matching the statistical query lower bound for k-sparse parity problems with sign stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 37, pages 113001–113037, 2024.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Conference on Empirical Methods in Natural Language Processing*, pages 107–117, 2016.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021.

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. In *International Conference on Learning Representations*, 2024.

Julia Nakhleh, Joseph Shenouda, and Robert D. Nowak. A new neural kernel regime: The inductive bias of multi-task learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 140451–140477, 2024.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 1256–1272, 2021.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2667–2690, 2019.

Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3067–3075, 2017.

Samarth Sinha, Zhengli Zhao, Anirudh Goyal, Colin Raffel, and Augustus Odena. Top-k training of GANs: Improving GAN performance by throwing away bad samples. In *Advances in Neural Information Processing Systems*, volume 33, pages 14638–14649, 2020.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212, 2021.

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. In *International Conference on Learning Representations*, 2019.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi S. Jaakkola. Understanding interlocking dynamics of cooperative rationalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 12822–12835, 2021.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836, 2020.

Shiji Zhou, Wenpeng Zhang, Jiyan Jiang, Wenliang Zhong, Jinjie Gu, and Wenwu Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond. In *Advances in Neural Information Processing Systems*, volume 35, pages 38103–38115, 2022.

A TRAINING PROTOCOL

All runs in Fig. 1 use the triangle readout with $\ell_0 = 5$ on the grid $8 \le n \le 32$, $2 \le k \le \min(n-1, 16)$, with 960 runs per cell. Gates are $m_j = \sigma(\mu_j)$ with logits initialized as $\mu_j \sim \mathcal{N}(0, 0.1^2)$, and both optimizers start from the same initial logits. Each step draws a fresh batch of 64 inputs, for 10^4 steps. SGD uses $\eta = 0.05$, and AdamW uses $\eta = 0.05$, $\beta = (0.9, 0.999)$, and no weight decay. A run succeeds when the accuracy on a fixed held-out set of 1024 inputs, evaluated every 100 steps, reaches 0.95 twice in a row.

The trained readout of Fig. 2 is a bit embedding followed by a two-layer MLP of width 64, trained by cross-entropy on the correct mask at $(n, k) = (16, 5)$ for 25 task seeds, and frozen when its margin at the vertices $z = h/k$ reaches $\ell_0 = 5$.

B FOURIER DECOMPOSITION OF THE POPULATION LOSS

The loss profile and its harmonics. Take the triangle readout $\ell(z) = \ell_0 \operatorname{tri}(kz)$ and write

$$w = kz - h, \qquad y = h \bmod 2. \tag{16}$$

The logit of the true parity is then

$$(1-2y)\,\ell(z) = \ell_0 \operatorname{tri}(w) = \ell_0\,(1-2|w|) \qquad (|w| \le 1), \tag{17}$$

extended with period 2, so that the loss profile

$$L_{\mathrm{BCE}}(w) = \log\bigl(1 + e^{-\ell_0(1-2|w|)}\bigr) \tag{18}$$

is a 2-periodic even function of w that does not depend on the label. The same holds for any readout antisymmetric under a shift by half a period: the BCE loss is symmetric under exchange of the two labels, so the antisymmetric part of the profile cancels in the population sum.

The profile expands as $L_{\mathrm{BCE}}(w) = \sum_{q \ge 1} \widetilde{L}_q \cos(\pi q w)$ (constant zero mode is omitted since it does not affect results) with

$$\widetilde{L}_q = \int_{-1}^{1} L_{\mathrm{BCE}}(w) \cos(\pi q w)\, dw. \tag{19}$$

Writing $\log(1 + e^{-x}) = -x/2 + \log(2\cosh(x/2))$ splits the profile into two pieces,

$$L_{\mathrm{BCE}}(w) = -\frac{\ell_0}{2} \operatorname{tri}(w) \;+\; \log\Bigl(2\cosh\frac{\ell_0 \operatorname{tri}(w)}{2}\Bigr). \tag{20}$$

The first piece is the triangle wave itself and carries only odd harmonics. The second is even in tri, hence invariant under $w \to w + 1$, and carries only even harmonics; it equals $\frac{\ell_0}{2}|\operatorname{tri}(w)|$ with its kink at $|w| = \frac{1}{2}$ rounded over a width $1/\ell_0$. The coefficients are

$$\widetilde{L}_q \simeq \begin{cases} -\dfrac{4\ell_0}{\pi^2 q^2}, & q \text{ odd (exact)}, \\[2ex] \dfrac{4\ell_0 \tanh(\ell_0/2)}{\pi^2 q^2} - \dfrac{2(-1)^{q/2}}{q\,\sinh\bigl(\pi^2 q/2\ell_0\bigr)}, & q \text{ even}, \end{cases} \tag{21}$$

where the even case is exact up to corrections of relative order $e^{-\ell_0}$. The first even term is set by the slope jumps of L at $w = 0$ and $w = \pm 1$ and is common to all even q; the second is the rounded kink, and it adds to or cancels the first according to q modulo 4. For $q \ll 2\ell_0/\pi^2$ the even coefficients therefore fall into two classes,

$$\widetilde{L}_q \simeq \begin{cases} \dfrac{8\ell_0}{\pi^2 q^2}, & q \equiv 2 \pmod 4 \quad \text{(strong)}, \\[2ex] \dfrac{\pi^2}{6\,\ell_0}, & q \equiv 0 \pmod 4 \quad \text{(weak)}, \end{cases} \tag{22}$$

and for $q \gg 2\ell_0/\pi^2$ both classes merge into the common tail, $\tanh(\ell_0/2)$ times the odd one.

Resonances. The population loss averages the profile over the atoms, so each harmonic of Eq. (19) is weighted by the average of its cosine,

$$L_{\text{pop}} = \sum_{q\geq 1} \widetilde{L}_q \sum_{h,h_i} p_{h,h_i} \cos(\pi q\, w), \qquad p_{h,h_i} = p_{(h_i|h)}\, p_h = \binom{k}{h}\binom{n-k}{h_i} 2^{-n}, \tag{23}$$

for any gates; another input distribution changes only the masses p_{h,h_i}.

The characteristic function of a binomial count is

$$\sum_{h_i} \binom{n-k}{h_i} 2^{-(n-k)} e^{i\theta h_i} = e^{i\theta(n-k)/2} \cos^{n-k}(\theta/2). \tag{24}$$

Applied to both groups of bits the two phases cancel, so that

$$\sum_{h,h_i} p_{h,h_i} \cos\Big(\pi q\lambda\, \frac{kh_i - (n-k)h}{n}\Big) = A(q\lambda), \tag{25}$$

which is Eq. (9) (Fig. 8, bottom). The local maxima of $|A|$ lie exactly at $x = 2N$, where

$$A(2N) = (-1)^{kN} \cos^n\Big(\frac{\pi N k}{n}\Big). \tag{26}$$

With $n = N_\star k + \delta$ the first family $N = N_\star$ has

$$A(2N_\star) = (-1)^{\delta} \cos^n\Big(\frac{\pi\delta}{n}\Big) \approx (-1)^{\delta} \exp\Big(-\frac{\pi^2\delta^2}{2n}\Big), \tag{27}$$

the common envelope of Eq. (13). Near each lattice point $\log|A|$ is quadratic to leading order, $A(x) \approx A(2N)\, e^{-(x-2N)^2/2\sigma^2}$ with

$$\sigma = \frac{2}{\pi}\sqrt{\frac{n}{k(n-k)}}, \tag{28}$$

and the sum over N is the comb of Eq. (11). Two conditions make the events of Sec. 4.1 isolated. Within the tooth $N = N_\star$, mode q has width σ/q in λ while the modes are spaced by $2N_\star/q^2$, so they are separated when $\sigma \lesssim 1/2$, that is for $k \gtrsim 4$. The neighboring teeth $N = N_\star \pm 1$ carry the envelope $e^{-\pi^2(k\mp\delta)^2/2n}$, negligible against $e^{-\pi^2\delta^2/2n}$ when $k \gtrsim \sqrt{n}$; for smaller k several teeth act at once and the ladder of spikes and dips is blurred. When $\gcd(n,k) > 1$ some atoms sit exactly on a wrong-parity vertex at $\lambda = 1$; there $L(\lambda)$ has a cusp at the start, and the initial gradient is the one-sided derivative toward $\lambda < 1$, not the value of the Fourier series.

The alignment at a tooth holds for any input distribution, which enters only through the masses. At $q\lambda = 2N_\star$ the term λh of the displacement spans whole periods of harmonic q, so all bits enter alike and an input contributes the phase $2\pi\delta(h + h_i)/n$, set by its number of ones,

$$A(2N_\star) = \sum_{h,h_i} p_{h,h_i} \cos\Big(\frac{2\pi\delta\,(h+h_i)}{n}\Big). \tag{29}$$

The sign of the tooth is where the typical input sits and its height is how tightly the inputs cluster in their number of ones, while its position comes from n and k; uniform bits give Eq. (27).

Fixing $h = h$ and averaging over h_i alone leaves the characteristic function of the non-task bits and a single cosine for the task bits,

$$L^{(h)}(\lambda) = \sum_{q\geq 1} \widetilde{L}_q \cos\big(\pi q\, \bar{w}_h\big) \cos^{n-k}\Big(\frac{\pi q k\lambda}{2n}\Big), \qquad \bar{w}_h = \lambda\, \frac{(n-k)(k-2h)}{2n}. \tag{30}$$

A subtask sees the profile, blurred by the non-task bits, at its own center $\bar{w}_h$; the center moves in proportion to λ while the blur sharpens, and this fixes along the whole walk which subtasks learn falsely. The binomial average over h,

$$\sum_{h} \binom{k}{h} 2^{-k} \cos\big(\phi(k-2h)\big) = \cos^k\phi, \qquad \phi = \frac{\pi q\lambda(n-k)}{2n}, \tag{31}$$

restores $A(q\lambda)$. Away from the integer multiples of π the factor $\cos^k\phi$ is exponentially small in k: neighboring subtasks cancel. At the multiples all subtasks add with the same sign, and where the blur of Eq. (30) is also near ± 1 this is a resonance. Training lowers λ, so a subtask learns falsely where its loss grows as λ decreases, that is, where its slope is negative (Fig. 8, middle).

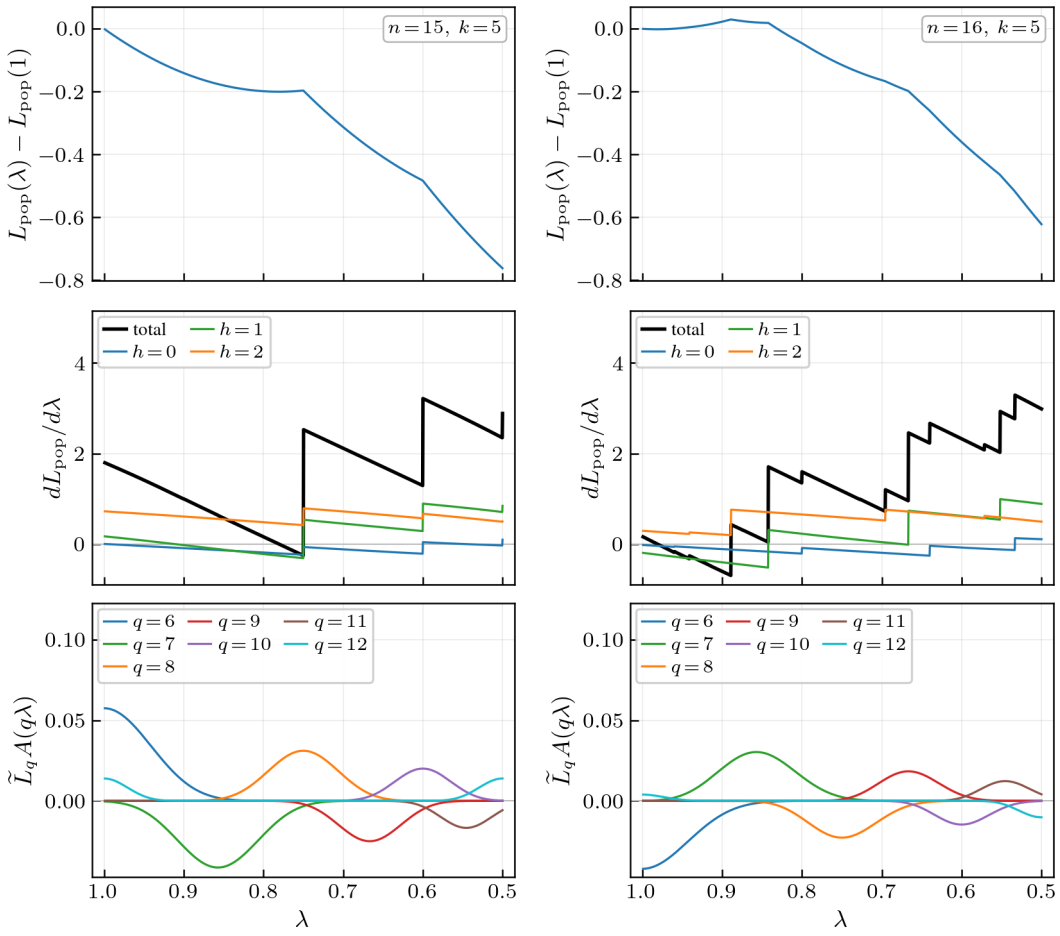


Figure 8: **Formation of barriers or kinks from signed resonances.** $(n, k) = (15, 5)$ (left) and $(16, 5)$ (right), both with $N_\star = 3$ so that the harmonics resonate at the same λ but with opposite signs. **Top:** the population loss $L_{\rm pop}$ starts descending due to the negative resonance at $\lambda = 1$ for $(15, 5)$ while ascending into a barrier due to the positive resonance for $(16, 5)$. **Middle:** subtask slopes are negative where the subtask learns falsely and are more complicated for the $\delta = 1$ $(16, 5)$ case. **Bottom:** harmonics $\widetilde{L}_q A(q\lambda)$ resonate at the common location for $\delta = 0$ while at slightly different locations for $\delta = 1$, but the sign pattern is more critical in creating a large barrier for the latter case.

The sign ladder and the barrier. A resonance identifies where atom contributions align, but it produces a barrier only when the aligned loss harmonics have the appropriate signs and sufficient amplitude to overcome the monotone background.

The first events are the resonances of the three modes

$$q = 2N_\star,\ 2N_\star + 1,\ 2N_\star + 2, \qquad \text{with signs } (-1)^{q+\delta}. \tag{32}$$

For even δ they read spike, dip, spike, and the barrier lies between the second and the third. For odd δ they read dip, spike, dip; the start is a dip, and the barrier lies between the first and the second (Fig. 8). When the events are isolated, the two modes $q, q + 1$ that bound the barrier give the height of Eq. (14) and a width in λ

$$\Delta\lambda = \frac{2N_\star}{q} - \frac{2N_\star}{q+1} = \frac{2N_\star}{q(q+1)}. \tag{33}$$

For $k = 2$ the pool spans a single wrong-parity vertex and there is no lattice to alias with; for $k = 3$ the descent of the fundamental covers the small rises until n is large.

Three corrections leave this structure intact. First, the multiples rq of a mode resonate at the same $\lambda = 2N_\star/q$ through the tooth $N = rN_\star$, with sign $(-1)^{r(q+\delta)}$ and envelope $\cos^n(\pi r\delta/n)$: they add

in a spike and alternate in a dip, and change the heights by factors of order $\sum_r r^{-2}$. Second, between the two events the fundamental descends by

$$\Delta_1 \approx \frac{4\ell_0}{\pi^2}\left[e^{-\lambda_{q+1}^2/2\sigma^2} - e^{-\lambda_q^2/2\sigma^2}\right], \qquad \lambda_q = \frac{2N_\star}{q}, \tag{34}$$

which lowers the barrier of even δ and is largest for small $N_\star$; where it exceeds the rise, the barrier is built by the next pair of modes. Third, the coefficient of the even mode depends on its class, Eq. (22). The two pairs share the odd mode $2N_\star + 1$ and differ in the even one,

$$2N_\star \quad (\delta \text{ odd}), \qquad\qquad 2N_\star + 2 \quad (\delta \text{ even}), \tag{35}$$

which always fall into different classes. For odd $N_\star$ the neighbors $n = N_\star k \pm 1$ take the strong class and the multiple $n = N_\star k$ the weak one, which strengthens the contrast; for even $N_\star$ the assignment is reversed and weakens it. The effect needs $2N_\star \lesssim 0.4\ell_0$, and at $\ell_0 = 5$ it is a few percent.

Dynamics in λ. SGD is Euclidean in the logits μ_j, and on the symmetric subspace the loss depends on μ only through λ, so the chain rule gives

$$\frac{d\lambda}{d\,\text{step}} = -\eta\, M(\lambda)\left(\frac{dL}{d\lambda} + \xi\right), \qquad M(\lambda) = |\nabla_\mu \lambda|^2, \tag{36}$$

with η the learning rate and ξ the mini-batch noise of the slope. The mobility M is set by how the gates are parametrized by the logits, not by the landscape, and it decreases toward the solution. The barrier height, and the height and width of the barrier measured in units of the fluctuation, do not depend on the coordinate.

C Correlation of barrier and success rate, including fluctuations

Fig. 9 left panel shows that the barrier B defined in Eq. (14) correlates with numerical success rates. This supports the false learning's role in the non-monotonic trainability band.

To include leading effect of fluctuations from finite batches, let D be the variance of one step of λ produced by the mini-batch noise ξ in Eq. (36). Together with the drift mobility $M(\lambda)$, this defines the local fluctuation scale

$$T_{\text{eff}} = \frac{D}{2\eta\, M(\lambda)} \sim \frac{\eta}{b}\left(\frac{k}{n}\right)^2 \ell_0^2. \tag{37}$$

The ratio B/T_{eff} thus measures the barrier in units of the batch noise. Success correlates more tightly with this normalized barrier than with B alone (Fig. 9 right). We use this only as a noise-aware comparison; the real dynamics also contains coordinate-dependent mobility, transverse symmetry breaking, and optimizer-specific effects.

D Beyond the symmetric subspace

Our analysis follows the permutation-symmetric subspace on which all task gates share one value and all non-task gates another. This appendix explains why that subspace is a reasonable description of typical training paths and how departures from it modify false learning. The symmetry is exactly invariant and is locally attracting over most of the relevant path. Small departures broaden the discrete atom lattice but preserve the reversed-update pattern, whereas sufficiently large departures suppress its resonant amplification. Where the loss is non-differentiable, the gates can split.

Stability of the symmetric subspace. On the symmetric subspace, gates belonging to the same group receive identical population updates. The subspace is therefore invariant. For a small difference between two gates in the same group,

$$\frac{d}{d\,\text{step}}\,(m_j - m_{j'}) = -\eta\,\kappa\,(m_j - m_{j'}), \tag{38}$$

where κ is the curvature of L_{pop} transverse to the subspace, measured in the parameters through which the gates are trained. For $\kappa > 0$, the difference decays; for $\kappa < 0$, it grows on the time scale $1/(\eta|\kappa|)$. At all differentiable points computed from initialization to the first barrier, we find $\kappa > 0$. Thus, along the typical path relevant to our barrier analysis, the symmetric subspace is locally attracting rather than merely invariant.

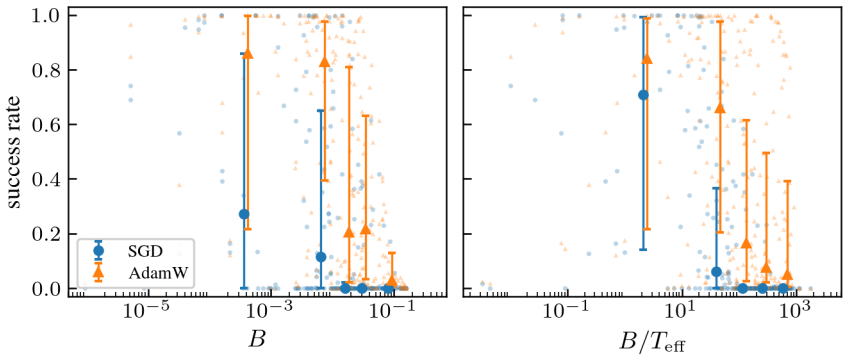


Figure 9: **Barrier height correlates with success rates.** Each dim point is per (n, k) with $B > 0$, binned against B (left) and B/T_{eff} (right). Markers show bin medians and bars show the interquartile range of 25% and 75%. Both correlate with success, with tighter correlation after normalizing the barrier height. Also, AdamW achieves larger success rates than SGD.

Atom broadening and its initialization scale. The displacement $w = kz - h$ of Eq. (6) remains valid for arbitrary gates. Each bit carries the coefficient $km_j/\sum_l m_l$, with an additional -1 for a task bit. On the symmetric subspace, all examples with the same counts (h, h_i) therefore share the position $w = \lambda a_{h,h_i}$. Away from it, different choices of the active bits produce different positions even at fixed (h, h_i), broadening one lattice atom into a distribution.

Let v_a and v_i be the within-group variances of these coefficients for the task and non-task gates. The conditional displacement variance of the atom (h, h_i) is

$$\sigma^2_{h,h_i} = v_a \frac{h(k-h)}{k-1} + v_i \frac{h_i(n-k-h_i)}{n-k-1}. \tag{39}$$

This also relates the abstract blur parameter to the initialization used in our experiments. Write $\mu_j \sim \mathcal{N}(0, s^2_\mu)$, with $s_\mu = 0.1$ in App. A. For small logits, $m_j = \frac{1}{2} + \frac{\mu_j}{4} + O(\mu_j^3)$. Linearizing the normalized pool and averaging Eq. (39) over the data gives, to leading order,

$$\sigma^2_{w,\text{init}} \simeq \frac{k^2 s^2_\mu}{16n}, \qquad \sigma_{w,\text{init}} \simeq \frac{k s_\mu}{4\sqrt{n}}. \tag{40}$$

At $(n, k) = (16, 5)$, this gives $\sigma_{w,\text{init}} \simeq 0.03$, the initialization-scale blur shown in Fig. 7. This calculation establishes that initialization begins near the symmetric subspace; attraction itself follows separately from the positive transverse curvature above.

Effect on reversed updates and resonant barriers. A small blur affects the two aspects of false learning differently. Locally, it only shifts the position of an example on the loss profile. Its update sign therefore remains unchanged unless it crosses a loss vertex, so an initialization-scale blur preserves the qualitative pattern of reversed subtask updates.

Collective amplification is more sensitive. Averaging the q-th harmonic over a Gaussian displacement blur of population-averaged width $\sigma^2_w = \sum_{h,h_i} p_{h,h_i}\sigma^2_{h,h_i}$ multiplies it by $\exp(-\pi^2 q^2 \sigma^2_w/2)$:

$$L_{\text{pop}}(\lambda; \sigma_w) = \sum_{q \geq 1} \widetilde{L}_q \, e^{-\pi^2 q^2 \sigma^2_w/2} \, A(q\lambda). \tag{41}$$

The higher resonant modes are consequently damped more strongly than the $q = 1$ background, and the $q = 1$ mode only decreases monotonically. Peaks and dips, hence barriers, are reduced at essentially unchanged locations. In Fig. 7, $\sigma_w = 0.03$ lowers but preserves the barrier at $(16, 5)$, whereas $\sigma_w = 0.1$ makes the atom distribution nearly continuous and removes it; more analytically, σ_w of the order of the barrier width $\Delta\lambda \sim 1/2N_\star$ in Eq. (33) would remove it. Thus, small symmetry breaking preserves local false learning, while large symmetry breaking can destroy its coherent amplification.

Evolution of the transverse width. Mini-batch gradients continually inject transverse fluctuations, while positive curvature restores the symmetry. In the linearized regime their balance is

$$\frac{d\sigma_w^2}{d\,\mathrm{step}} = -2\eta\kappa\,\sigma_w^2 + \frac{\eta^2}{b}\,V_\perp, \qquad \sigma_w^2 \to \frac{\eta\,V_\perp}{2\kappa\,b}, \tag{42}$$

where $V_\perp$ is the per-datum gradient variance in a within-group direction. Typical training therefore occupies a narrow stochastic tube around the symmetric trajectory rather than the exact zero-width subspace. The width of that tube is set jointly by initialization, transverse curvature, learning rate, and batch size.

Symmetry-breaking trajectories. The curvature criterion does not apply where the population loss is non-differentiable. For $\gcd(n,k) > 1$, some atoms lie exactly on a wrong-parity vertex at initialization, and splitting two gates there lowers the loss at first order. For general (n,k), it is not clear whether this split grows into a large departure. Our one-coordinate theory therefore describes the near-symmetric trajectories and their first resonant barrier, not trajectories with large departures.

E SPARSE MODULAR ADDITION

The readout and the loss profile. Each of the n positions carries an input $x_j \in \{0,\dots,V-1\}$, independent and uniform, and the label is $y = h \bmod p$ with $h = \sum_{j\in S^*} x_j$ the sum of the task inputs. Sec. 5 takes $V = p$, and this appendix keeps the two apart. Writing $h_i = \sum_{j\notin S^*} x_j$, the pool is $z = (m_a h + m_i h_i)/M$ as in Eq. (5), the frozen readout gives one logit $\ell_y(z)$ to each label $y \in \{0,\dots,p-1\}$, and the loss is the cross-entropy of the true label.

The readout is trained on the correct mask, that is on the vertices $kz = h$ drawn with the mass p_h of the task sum, by cross-entropy with weight decay γ. For a readout with one hidden ReLU layer and free biases, the weight decay penalizes the total curvature $\int\|\ell''\|$ of the readout (Savarese et al., 2019; Nakhleh et al., 2024), so a minimizer is linear between neighboring vertices, and at a vertex h away from the ends of the pool range

$$\ell_y = \ell_0\Big[1 - \frac{r(p-r)}{p-1}\Big], \quad r = (h-y) \bmod p, \qquad \ell_0 \simeq \log\frac{p_h}{\gamma} + \tfrac{1}{2}\log\Big(1-\frac{1}{p}\Big) \tag{43}$$

The logit of label y is ℓ_0 at its own vertices ($r = 0$), 0 at the vertices of the neighboring labels ($r = 1$), and lower by $\ell_0[r(p-r)/(p-1) - 1]$ at the vertices at distance r; every label carries the same shape shifted by y. The amplitude ℓ_0 is the margin of the vertex, so it varies from vertex to vertex only through the mass p_h, and it grows logarithmically as the weight decay weakens. For the true label $y = h \bmod p$ the distance r is the displacement $w = kz - h$ reduced modulo p, so the true logit is a function of w alone. At $p = 2$ the true logit minus the wrong one is $\ell_0\,\mathrm{tri}(w) = (1-2y)\,\ell(z)$, the readout of Sec. 2; from $p \geq 4$ on the wrong logits decrease toward the middle of the period. At the ends of the pool range the minimizer continues its linear trend and the logits of the far labels spread widely.

With a common ℓ_0, shifting the pool by one vertex and rotating the labels by one leaves the readout unchanged, so the loss of a sample depends on the pool and the label only through the displacement $w = kz - h$ of Eq. (6). That loss is a profile $L(w)$ of period p which does not depend on the label and is Eq. (20) at $p = 2$ (Fig. 10, top). $L(w)$ has basins at the vertices of the same label, $w \in p\mathbb{Z}$, and peaks at the other integer w, the vertices of wrong labels; for $p \geq 3$ it also has a shallow minimum between two wrong vertices, where the readout is undecided between the two wrong labels.

Resonances. On the symmetric subspace the displacement is $w = \lambda\,[kh_i - (n-k)h]/n$ with the coordinate of Eq. (8), and since $L(w)$ is even and the mean of w vanishes, averaging over the inputs gives the population loss in the form of Eq. (9) with

$$A(x) = D_V\Big(\frac{2\pi x(n-k)}{pn}\Big)^k D_V\Big(\frac{2\pi xk}{pn}\Big)^{n-k}, \qquad D_V(\theta) = \frac{\sin(V\theta/2)}{V\sin(\theta/2)}, \tag{44}$$

and $\widetilde{L}_q$ the coefficients of $L(w)$ in $\cos(2\pi qw/p)$ as in Eq. (19); at $V = p = 2$, $D_2(\theta) = \cos(\theta/2)$ returns Eq. (10). The harmonics whose index q is not a multiple of p distinguish the labels; those

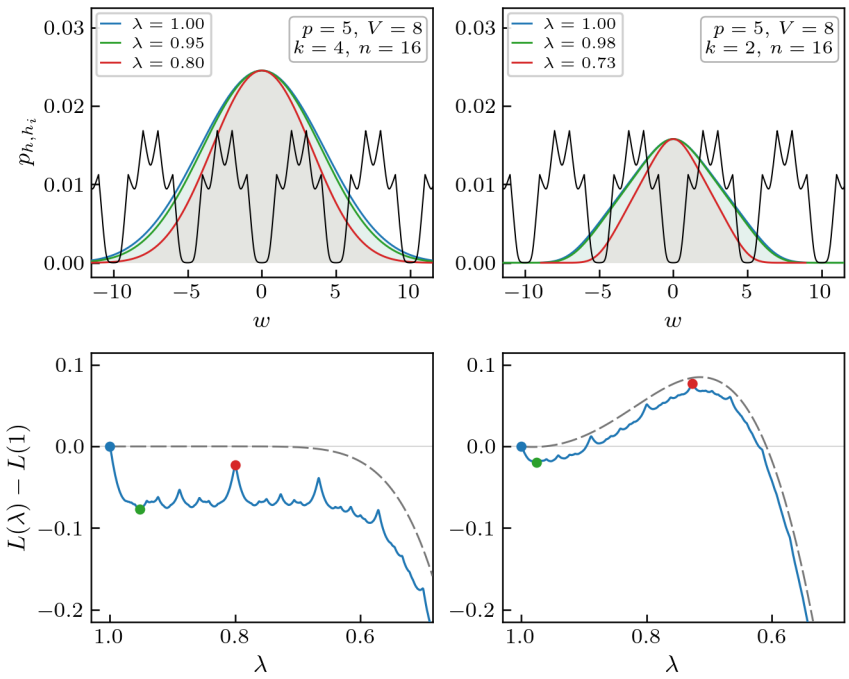


Figure 10: **Two barriers in sparse modular addition.** $p = 5$, $V = 8$, $n = 16$, with $k = 4$ (left) and $k = 2$ (right). **Top:** the atom distribution in w at the start, the dip, and the peak of the barrier (colors), overlaid with the loss profile $L(w)$. **Bottom:** $L(\lambda) - L(1)$ and its $q = 1$ mode (dashed); dots mark the λ of the top row. At $k = 2$ the $q = 1$ mode alone makes the barrier. At $k = 4$ the $q = 1$ mode decreases monotonically, and the barrier comes from the resonance at $n = 4k$, with its peak at $\lambda = 4/5$.

whose index is a multiple of p do not, and carry the structure between the wrong vertices. The odd and even harmonics of parity are the case $p = 2$.

Both factors of $A(q\lambda)$ peak at $q\lambda = pN$ in place of $2N$, and equal ± 1 there on the lines $n = Nk$, so on the symmetric subspace the first resonant harmonic is $q = pN_\star$ rather than $2N_\star$. With $n = N_\star k + \delta$ the lattice value there is $(-1)^{(V-1)\delta} D_V(2\pi\delta/n)^n$, of magnitude close to $\exp[-\pi^2(V^2-1)\delta^2/6n]$, which reduces to Eq. (27) at $V = 2$ and narrows in δ as V grows. The resonant barriers therefore lie on the lines $n = Nk$ for every p, as for parity (Fig. 10, left).

A barrier from the first harmonic. For $V > p$ the $q = 1$ mode alone can produce a barrier, without a resonance. For large n the pool at the start is the input mean $(V-1)/2$, so the subtask with task sum h starts at

$$w \approx \frac{k(V-1)}{2} - h \tag{45}$$

and moves toward $w = 0$ in proportion to λ. When $V > p$ the values of one task input span more than one period of $L(w)$. At $k = 2$ the extreme subtasks then start in the basin of $w = \pm p$ and false-learn as they climb over the wrong vertices toward $w = 0$ (Fig. 10, right). In the population loss the $q = 1$ mode becomes non-monotone, and at $k = 2$ it gives a barrier with its peak at $\lambda_0 = pn/[V(n-k)]$, where one task input spans exactly one period. Since it needs no resonance, this barrier appears for any n; on the lines $n = Nk$ the $q = 1$ mode and the resonant modes add in the same $L(\lambda)$.

F SAMPLED-GATE LEARNER

The sampled-gate learner keeps the pool, the readout and the per-sample loss of Sec. 2 and changes only how the gates are set. At every step each gate is drawn as 0 or 1: a task gate opens with probability $\bar{m}_a$, a non-task gate with probability $\bar{m}_i$, and it is these probabilities that are trained. The number of open gates is $M = \sum_j m_j$, an integer whose mean is the $M = km_a + (n-k)m_i$ of Eq. (5). The pool averages the open bits only, an empty mask sets the readout to 0, and the population loss averages over the data and over the drawn masks,

$$L_{\text{pop}}(\bar{m}_a, \bar{m}_i) = \mathbb{E}_x \, \mathbb{E}_m[L], \qquad M = \sum_j m_j. \tag{46}$$

Local mechanism. The update of $\bar{m}_j$ is the loss difference between that gate open and closed, the others drawn,

$$\frac{\partial \, \mathbb{E}_m[L]}{\partial \bar{m}_j} = \mathbb{E}_m[L \mid m_j = 1] - \mathbb{E}_m[L \mid m_j = 0], \tag{47}$$

which is the slope $\partial_z L_h$ averaged between the two pools. While that interval stays inside one linear piece of the readout, the sign rule of Eq. (7) holds unchanged, so both learners assign the reversed update to the same h-subtasks (Fig. 11).

No training coordinate. In DIFF every datum is pooled through the same mask, so the atom positions depend on the gates only through the coordinate λ of Eq. (8). With sampled gates each step pools only the bits it opens. A drawn mask of size M that opens all k task bits is the symmetric start, $\lambda = 1$, of a smaller problem with n replaced by M: the displacement of Eq. (6) becomes

$$w_{h,h_i} = \frac{kh_i - (M-k)h}{M}, \tag{48}$$

with h_i the ones among the open non-task bits. The atoms do not contract toward $w = 0$. What λ did in DIFF is now done by two probabilities: the probability $\bar{m}_a^k$ that all task gates open, and the distribution of M. Training lowers $\bar{m}_i$, which moves the mass of M toward k.

Fourier expansion. The coefficients $\widetilde{L}_q$ of Eq. (19) are those of the body. The characteristic function acquires M as its parameter,

$$A_M(x) = \mathbb{E}_{h,h_i}\big[\cos(\pi x \, a_{h,h_i})\big], \qquad a_{h,h_i} = \frac{kh_i - (M-k)h}{M}, \tag{49}$$

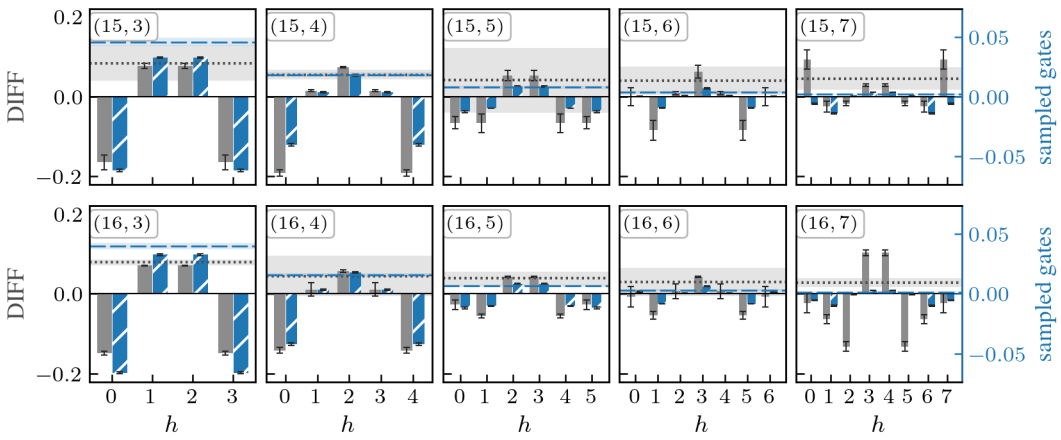


Figure 11: **Sampling the gates leaves the falsely learned h-subtasks intact.** $G^{(h)}$ at the symmetric start for DIFF (left axis) and the sampled-gate learner (right axis). Bars are analytic predictions with closed forms, error bars the spread over random masks.

and returns the A of Sec. 4 at $M = n$. A closed task bit enters the label h but not the pool, and its factor in the bit-by-bit product of $\mathbb{E}_x[\cos(\pi q w)]$ is $\cos(\pi q/2)$. For odd q this factor is 0, so the label channel survives only on draws that open every task gate. For even q its sign cancels against the open task bits, so every draw of size M gives the same value $A_M(q)$. Hence

$$L_{\text{pop}}(\bar{m}_a, \bar{m}_i) = \sum_{q \text{ even}} \widetilde{L}_q \, \mathbb{E}_m\big[A_M(q)\big] + \sum_{q \text{ odd}} \widetilde{L}_q \, \mathbb{E}_m\Big[\prod_{j \in S^*} m_j \, A_M(q)\Big]. \tag{50}$$

Beside Eq. (9) the correspondence is direct: in DIFF λ moves the argument of A, here the distribution of M moves A itself. The odd sum carries the factor $\bar{m}_a^k$, which is 2^{-k} at the symmetric start, so the component of the update that separates task from non-task gates is suppressed exponentially in k.

The label channel cannot resonate. In DIFF a harmonic resonates when $q\lambda$ is an even integer. Here $\lambda = 1$ is fixed, so an odd q never meets the condition, and the odd part of $A_M(q)$ decays exponentially in M. The sign reversals that the label channel builds into the DIFF barrier are absent from the population sum; what the label channel leaves is the factor $\bar{m}_a^k$. The even harmonics resonate exactly:

$$A_M(q) = (-1)^{qk/2} \cos^M\Big(\frac{\pi q k}{2M}\Big) \quad (q \text{ even}), \qquad \big|A_M(q)\big| = 1 \iff \frac{qk}{2M} \in \mathbb{Z}. \tag{51}$$

When M divides k every even harmonic resonates at once. Every pool value then sits on a vertex, the readout is at full confidence on a mask that cannot see the label, and the even part of the loss is at its maximum.

Saddles and corners in place of a barrier. The even channel sees only M. It cannot tell a task gate from a non-task gate, it pushes the total number of open gates only, and its worst point is the mask size of the solution, $M = k$. What holds the solution is the label channel with its factor $\bar{m}_a^k$. Because L_{pop} is affine in each $\bar{m}_j$ separately,

$$\frac{\partial^2 L_{\text{pop}}}{\partial \bar{m}_j^2} = 0 \quad \text{for every } j, \tag{52}$$

the diagonal of the task block of second derivatives vanishes and its eigenvalues take both signs. A point that looks like a minimum in the $(\bar{m}_a, \bar{m}_i)$ plane is a saddle of the full gate space, and the stationary points are corners. The correct mask is one of them; the others are corners whose mask size is a local minimum of the even part of the loss. The quantity that replaces the barrier height B of Sec. 4 is the dependence of the even part of the loss on M. A vertex crossing of DIFF corresponds here to probability mass of M moving to a neighboring value, and a critical point of the symmetric subspace is where the contributions of two neighboring M cancel.

In all, thus sampled gates preserve the subset-level reversed-update mechanism, while replacing DIFF's one-coordinate resonant barrier by a landscape governed by the sampled mask-size distribution, saddles, and corners. This supports the robustness of elementary false learning, not the universality of its DIFF resonance structure.

G Initial Fourier-gap signals

Barak et al. (2022) explain how SGD finds the task bits of sparse parity from the population gradient at initialization. In Boolean Fourier terms, the gradient at a task bit is the correlation with the parity of the other $k-1$ task bits, and the gradient at a non-task bit is the correlation with the parity of the k task bits and that bit. The former is larger, and this Fourier gap biases the first steps toward the task bits. For the ReLU networks they analyze, the gap is small, of order $n^{-(k-1)/2}$.

Our frozen readout $\ell_0 \operatorname{tri}(kz)$ is one of their architectures, a single neuron with a triangle-wave activation whose weights are the normalized gates, so the same picture applies to our mask. At the uniform mask the population update points exactly along the correct direction d of Eq. (3), since z depends only on m_a/m_i. Its size is the initial slope of L_{pop} in λ, a sum of the harmonics of Eq. (9). At $\lambda = 1$ every contributing harmonic lies between resonances, where the atoms are not aligned and their contributions largely cancel, so the slope is small,

$$\left.\frac{dL_{\mathrm{pop}}}{d\lambda}\right|_{\lambda=1} \simeq \frac{\ell_0 k}{2\pi}\,\Gamma\!\left(\frac{k-1}{2}\right)\left(\frac{2}{n}\right)^{\frac{k-1}{2}} \qquad (k \text{ even},\ n \gg k^2), \tag{53}$$

with the same scaling as their Fourier gap.

The Fourier-gap signal and the underlying signed subtask updates coexist but evolve differently. The smooth, hidden progress of the Fourier gap in Barak et al. (2022) is carried primarily by the $q = 1$ mode, whose contribution descends monotonically toward the correct mask. Meanwhile, large signed subtask updates—including reversed updates that largely cancel in the population gradient—appear in the higher-harmonic structure, potentially growing into a barrier in later training. Thus, the Fourier gap characterizes the weak net bias at initialization; false learning is already present in reversed subtask updates and becomes a macroscopic obstruction when their false-to-true transitions synchronize along training.