CMPUT 654: Mathematical Foundations of Modern AI Systems Fall 2026

Lecture 8: The need for signal

September 29, 2026

Lecturer: Csaba Szepesvári *Note prepared by: Csaba Szepesvári*

Triston Grayston prepared a scribe of the September 29 lecture. Csaba Szepesvári prepared and expanded the present notes.

Reading guide. Lecture 7 gave an example in which a small ReLU network represents a fixed parity but direct perturbed population-gradient training has no reliable classification advantage. This lecture first proves that result. It then changes the supervised task: running-parity labels expose a repeated two-bit calculation whose final output is still the original parity. The path-star example and a constructed transformer theorem show that retrieval and combination must also be learned. The transformer theorem is included as optional reading because its assumptions and three-update proof are useful for understanding the mechanism but are not needed for the parity lower bound. A second optional section, added after class, summarizes how intermediate supervision and a length curriculum can train a transformer on more general state-tracking tasks, and states the restrictions of that guarantee. The prerequisites are Gaussian vectors, KL divergence, total variation, Pinsker's inequality, Hoeffding's inequality, ReLU networks, and the parity setup recalled below.

8.1 The fixed-parity result

Let $d \in \mathbb{N}$ and $[d] = \{1, \dots, d\}$. For fixed $S \subseteq [d]$, write

$$\chi_S(x) = \prod_{j \in S} x_j, \qquad x \in \{-1, 1\}^d.$$

Let X be uniform on $\{-1, 1\}^d$, set $Y = \chi_S(X)$, and write $k = |S|$.

A one-hidden-layer ReLU network with $m \in \mathbb{N}$ units has parameter

$$\theta = (u_r, w_r, b_r)_{r=1}^m \in \mathbb{R}^p, \qquad u_r, b_r \in \mathbb{R}, \quad w_r \in \mathbb{R}^d, \quad p = m(d+2).$$

We identify this tuple with the vector formed by concatenating its coordinates. Write $[a]_+ = \max\{a, 0\}$, use $\|\cdot\|$ for the Euclidean norm, and let I_p denote the $p \times p$ identity matrix. The network $N_\theta : \{-1, 1\}^d \to \mathbb{R}$ and its signed correlation objective $F : \mathbb{R}^p \to \mathbb{R}$ are

$$N_\theta(x) = \sum_{r=1}^m u_r [w_r^\top x + b_r]_+ \tag{8.1}$$

and

$$F(\theta) = -\mathbb{E}[Y N_\theta(X)]. \tag{8.2}$$

Lecture 7 constructed a width-$(k+1)$ network representing χ_S. Thus the lower bound below concerns optimization rather than representability. The title refers to the information available to the optimizer. Each gradient coordinate is a label correlation with a parameter derivative of

the current network. For high-degree parity, these correlations are exponentially small at typical Gaussian parameters, so the updates provide little information about which parameter changes would reduce the risk.

For fixed θ, define the classifier and its risk by

$$h_\theta(x) = \operatorname{sign}(N_\theta(x)), \qquad R_{01}(h_\theta) = \mathbb{P}(h_\theta(X) \neq Y),$$

where $\operatorname{sign}(0) = 1$.[1]

Fix a number of updates $T \in \mathbb{N}$, a learning rate $\eta > 0$, and a noise scale $\sigma > 0$. Let the $\mathbb{R}^p$-valued random vectors $\Theta_0, \Xi_0, \dots, \Xi_{T-1}$ be mutually independent and independent of X, with

$$\mathcal{L}(\Theta_0) = \mathcal{N}(0, \sigma^2 I_p), \qquad \mathcal{L}(\Xi_t) = \mathcal{N}(0, \sigma^2 I_p), \qquad 0 \le t < T.$$

Use the ReLU backpropagation rule $a \mapsto \mathbb{I}\{a > 0\}$ in the updates. The trained process and its gradient-free comparison process are

$$\Theta_{t+1} = \Theta_t - \eta \nabla F(\Theta_t) - \Xi_t, \tag{8.3}$$

$$\Theta_0^{(0)} = \Theta_0, \qquad \Theta_{t+1}^{(0)} = \Theta_t^{(0)} - \Xi_t. \tag{8.4}$$

The following lemma compares the complete parameter trajectories. Its bound will let us transfer a classification-risk calculation for the Gaussian walk to the trained network.

Lemma 8.1.1 (Path comparison)**.** *Let $d > 30$, $A > 0$, $m, T \in \mathbb{N}$, and $\eta, \sigma > 0$. Assume*

$$m, T, \eta, \sigma, \sigma^{-1} \le d^A, \qquad k \ge (288A + 144) \log d. \tag{8.5}$$

Then

$$\operatorname{TV}\Big(\mathcal{L}(\Theta_0, \dots, \Theta_T), \mathcal{L}(\Theta_0^{(0)}, \dots, \Theta_T^{(0)})\Big) \le e^{-k/36}. \tag{8.6}$$

The Gaussian walk is symmetric under changes of sign of individual parameter coordinates. Combining that symmetry with the path comparison gives the classification conclusion.

Theorem 8.1.2 (Classification advantage occurs with probability at most one half)**.** *Under the assumptions of Lemma 8.1.1, for every $\gamma > 0$,*

$$\mathbb{P}\left(R_{01}(h_{\Theta_T}) \le \frac{1}{2} - \gamma\right) \le \frac{1}{2} + e^{-k/36}, \tag{8.7}$$

$$\left|\mathbb{E}[R_{01}(h_{\Theta_T})] - \frac{1}{2}\right| \le e^{-k/36}. \tag{8.8}$$

Proof. By Lemma 8.1.1, it suffices to establish the two risk bounds for the Gaussian comparison parameter $G = \Theta_T^{(0)}$, with the $e^{-k/36}$ terms removed. We will show that $R_{01}(h_G)$ and $1 - R_{01}(h_G)$ have the same law.

Fix $j \in S$. Let $D_j : \mathbb{R}^d \to \mathbb{R}^d$ flip the jth sign:

$$D_j(x_1, \dots, x_d) = (x_1, \dots, x_{j-1}, -x_j, x_{j+1}, \dots, x_d).$$

Define the corresponding parameter map by

$$J_j : \mathbb{R}^p \to \mathbb{R}^p, \qquad J_j\theta = (u_r, D_j w_r, b_r)_{r=1}^m.$$

Both maps are orthogonal involutions: $D_j^\top D_j = D_j^2 = I_d$ and $J_j^\top J_j = J_j^2 = I_p$. The identities $(D_j w_r)^\top x = w_r^\top D_j x$ and $j \in S$ give

$$N_{J_j\theta}(x) = N_\theta(D_j x), \qquad \chi_S(D_j x) = -\chi_S(x).$$

Changing variables $z = D_j x$ in the uniform input average gives

$$\begin{aligned} R_{01}(h_{J_j\theta}) &= 2^{-d} \sum_x \mathbb{I}\{h_\theta(D_j x) \neq \chi_S(x)\} \\ &= 2^{-d} \sum_z \mathbb{I}\{h_\theta(z) \neq -\chi_S(z)\} = 1 - R_{01}(h_\theta). \end{aligned}$$

Since $\mathcal{L}(G) = \mathcal{N}(0, (T+1)\sigma^2 I_p)$ and J_j is orthogonal, $\mathcal{L}(J_j G) = \mathcal{L}(G)$. Combining this invariance with the risk identity gives

$$\mathcal{L}(R_{01}(h_G)) = \mathcal{L}(R_{01}(h_{J_j G})) = \mathcal{L}(1 - R_{01}(h_G)).$$

Thus the classification-risk distribution is symmetric about $1/2$. Hence

$$\mathbb{E}[R_{01}(h_G)] = \frac{1}{2}, \qquad \mathbb{P}\left(R_{01}(h_G) \leq \frac{1}{2} - \gamma\right) = \mathbb{P}\left(R_{01}(h_G) \geq \frac{1}{2} + \gamma\right) \leq \frac{1}{2}.$$

The last inequality holds because the two events are disjoint. Total variation controls the difference of event probabilities and of expectations of $[0,1]$-valued functions. Applying Eq. (8.6) to these functions of the final parameter proves both conclusions. □

Exercise 2 examines a tempting but incorrect alternative to this argument: flipping the signs of every parameter. The coordinate transformation J_j works because it corresponds exactly to flipping one input bit.

8.2 Proof of the parity hardness result

It remains to prove the trajectory comparison in Lemma 8.1.1. The key estimate is an expected gradient bound under a Gaussian parameter law. We first use this estimate to compare the trajectories, then prove it by reducing the gradient coordinates to threshold correlations.

Lemma 8.2.1 (Gaussian gradient)**.** *Let $\Theta = (U_r, W_r, B_r)_{r=1}^m$ have independent $\mathcal{N}(0, \tau^2)$ coordinates, with $\tau > 0$. If $d > 30$ and $k \geq 3$, then*

$$\mathbb{E}[\|\nabla F(\Theta)\|] \leq 8md\tau e^{-k/8}. \tag{8.9}$$

In particular, if $k \geq 72 \log(8md\tau)$, then

$$\mathbb{E}[\|\nabla F(\Theta)\|] \leq e^{-k/9}, \tag{8.10}$$

and

$$\mathbb{P}(\|\nabla F(\Theta)\| \geq e^{-k/18}) \leq e^{-k/18}. \tag{8.11}$$

8.2.1 From small Gaussian gradients to close trajectories

We use the Gaussian gradient estimate to prove the path comparison, restated here with its parameter conditions.

Lemma 8.1.1 (Path comparison)**.** *Let $d > 30$, $A > 0$, $m, T \in \mathbb{N}$, and $\eta, \sigma > 0$. Assume*

$$m, T, \eta, \sigma, \sigma^{-1} \le d^A, \qquad k \ge (288A + 144) \log d. \tag{8.5}$$

Then

$$\mathrm{TV}\Big(\mathcal{L}(\Theta_0, \dots, \Theta_T), \mathcal{L}(\Theta_0^{(0)}, \dots, \Theta_T^{(0)})\Big) \le e^{-k/36}. \tag{8.6}$$

Proof of Lemma 8.1.1. We want to bound the total variation between the trained trajectory and the Gaussian walk. Introduce a process that agrees with the Gaussian walk whenever its gradient is small. For $\varepsilon = e^{-k/18}$, set

$$\overline{\Theta}_0 = \Theta_0, \qquad \overline{\Theta}_{t+1} = \overline{\Theta}_t - \eta[\nabla F(\overline{\Theta}_t)]_\varepsilon - \Xi_t, \qquad [g]_\varepsilon = \begin{cases} 0, & \|g\| \le \varepsilon, \\ g, & \|g\| > \varepsilon. \end{cases} \tag{8.12}$$

For any process $V_0, \dots, V_T$, write $V_{0:T} = (V_0, \dots, V_T)$. Name the three *trajectory laws*

$$\mathsf{P} = \mathcal{L}(\Theta_{0:T}), \qquad \mathsf{R} = \mathcal{L}(\overline{\Theta}_{0:T}), \qquad \mathsf{G} = \mathcal{L}(\Theta_{0:T}^{(0)}).$$

The triangle inequality gives the two quantities to control:

$$\mathrm{TV}(\mathsf{P}, \mathsf{G}) \le \mathrm{TV}(\mathsf{P}, \mathsf{R}) + \mathrm{TV}(\mathsf{R}, \mathsf{G}). \tag{8.13}$$

The trained and auxiliary processes have close transition laws. We bound $\mathrm{TV}(\mathsf{P}, \mathsf{R})$ through KL. For a history $\vartheta_{0:t} \in (\mathbb{R}^p)^{t+1}$, the recursions specify the following versions of their next-step conditional laws:

$$\begin{aligned} \mathsf{P}_t(\,\cdot \mid \vartheta_{0:t}) &= \mathcal{L}(\Theta_{t+1} \mid \Theta_{0:t} = \vartheta_{0:t}) \\ &= \mathcal{N}\big(\vartheta_t - \eta \nabla F(\vartheta_t), \sigma^2 I_p\big), \\ \mathsf{R}_t(\,\cdot \mid \vartheta_{0:t}) &= \mathcal{L}(\overline{\Theta}_{t+1} \mid \overline{\Theta}_{0:t} = \vartheta_{0:t}) \\ &= \mathcal{N}\big(\vartheta_t - \eta[\nabla F(\vartheta_t)]_\varepsilon, \sigma^2 I_p\big). \end{aligned}$$

These formulas define kernels from $(\mathbb{R}^p)^{t+1}$ to probability laws on $\mathbb{R}^p$. At the same history, their means differ by

$$\eta\big(\nabla F(\vartheta_t) - [\nabla F(\vartheta_t)]_\varepsilon\big),$$

whose norm is at most $\eta\varepsilon$. The KL formula for Gaussians with the same covariance gives

$$D\big(\mathsf{P}_t(\,\cdot \mid \vartheta_{0:t}), \mathsf{R}_t(\,\cdot \mid \vartheta_{0:t})\big) = \frac{\eta^2}{2\sigma^2}\big\|\nabla F(\vartheta_t) - [\nabla F(\vartheta_t)]_\varepsilon\big\|^2 \le \frac{\eta^2\varepsilon^2}{2\sigma^2}.$$

The initial laws agree. Thus the KL chain rule gives

$$D(\mathsf{P}, \mathsf{R}) = \sum_{t=0}^{T-1} \mathbb{E}_{\mathsf{P}}\big[D\big(\mathsf{P}_t(\,\cdot \mid \Theta_{0:t}), \mathsf{R}_t(\,\cdot \mid \Theta_{0:t})\big)\big] \le \frac{T\eta^2\varepsilon^2}{2\sigma^2}. \tag{8.14}$$

By Pinsker's inequality,

$$\mathrm{TV}(\mathsf{P}, \mathsf{R}) \le \sqrt{\frac{1}{2} D(\mathsf{P}, \mathsf{R})} \le \frac{\eta\varepsilon\sqrt{T}}{2\sigma}. \tag{8.15}$$

The auxiliary process agrees with the Gaussian walk with high probability. Use the same initialization and increments for both processes, as in Eqs. (8.4) and (8.12). On the event

$$E = \left\{ \max_{0 \le t \le T} \|\nabla F(\Theta_t^{(0)})\| \le \varepsilon \right\},$$

their trajectories coincide. Indeed, if $\overline{\Theta}_t = \Theta_t^{(0)}$, then the auxiliary gradient is zero and both recursions subtract Ξ_t. Induction proves equality at every time on E. The coupling inequality therefore gives

$$\mathrm{TV}(\mathsf{R}, \mathsf{G}) \le \mathbb{P}(E^c).$$

To bound $\mathbb{P}(E^c)$, we apply Lemma 8.2.1 at each time. The independent Gaussian increments give

$$\mathcal{L}(\Theta_t^{(0)}) = \mathcal{N}(0, (t+1)\sigma^2 I_p), \qquad \tau_t = \sigma\sqrt{t+1}.$$

The parameter conditions in Eq. (8.5) imply, for $t \le T$,

$$\log(8md\tau_t) \le \log(8\sqrt{2}) + (1 + 5A/2)\log d \le (4A+2)\log d \le k/72.$$

Here we used $m, \sigma, T \le d^A$ and $d > 30$; the lower bound on k also implies $k \ge 3$. Thus the Gaussian gradient estimate applies and gives

$$\mathbb{E}[\|\nabla F(\Theta_t^{(0)})\|] \le e^{-k/9} = \varepsilon^2, \qquad \mathbb{P}(\|\nabla F(\Theta_t^{(0)})\| > \varepsilon) \le \varepsilon$$

by Markov's inequality. A union bound now yields

$$\mathbb{P}(E^c) = \mathbb{P}\left(\max_{0 \le t \le T} \|\nabla F(\Theta_t^{(0)})\| > \varepsilon \right) \le (T+1)\varepsilon, \tag{8.16}$$

so

$$\mathrm{TV}(\mathsf{R}, \mathsf{G}) \le (T+1)\varepsilon. \tag{8.17}$$

Combining the two comparisons. Substituting Eqs. (8.15) and (8.17) into Eq. (8.13) gives

$$\mathrm{TV}(\mathsf{P}, \mathsf{G}) \le \left(\frac{\eta\sqrt{T}}{2\sigma} + T + 1 \right) e^{-k/18}. \tag{8.18}$$

The polynomial bounds on η, T, σ^{-1} give

$$\frac{\eta\sqrt{T}}{2\sigma} + T + 1 \le 3d^{5A/2}.$$

Finally,

$$\log(3d^{5A/2}) = \log 3 + \frac{5A}{2}\log d \le (8A+4)\log d \le k/36.$$

Consequently, Eq. (8.18) is at most $e^{k/36}e^{-k/18} = e^{-k/36}$, as claimed. □

8.2.2 Bounding the expected gradient one unit at a time

The path argument used the following estimate at variance $\tau^2 = (t+1)\sigma^2$. We now prove it for any $\tau > 0$.

Lemma 8.2.1 (Gaussian gradient)**.** *Let $\Theta = (U_r, W_r, B_r)_{r=1}^m$ have independent $\mathcal{N}(0, \tau^2)$ coordinates, with $\tau > 0$. If $d > 30$ and $k \geq 3$, then*

$$\mathbb{E}[\|\nabla F(\Theta)\|] \leq 8md\tau e^{-k/8}. \tag{8.9}$$

In particular, if $k \geq 72 \log(8md\tau)$, then

$$\mathbb{E}[\|\nabla F(\Theta)\|] \leq e^{-k/9}, \tag{8.10}$$

and

$$\mathbb{P}(\|\nabla F(\Theta)\| \geq e^{-k/18}) \leq e^{-k/18}. \tag{8.11}$$

Proof of Lemma 8.2.1. We need to bound $\mathbb{E}[\|\nabla F(\Theta)\|]$. We start by calculating the derivative of F at some fixed parameter vector $\theta \in \mathbb{R}^p$. Grouping the derivatives by hidden unit gives

$$\|\nabla F(\theta)\| = \sqrt{\sum_{r=1}^m (|\partial_{u_r} F(\theta)|^2 + \|\partial_{w_r} F(\theta)\|^2 + |\partial_{b_r} F(\theta)|^2)}. \tag{8.19}$$

We now turn to calculating the derivatives appearing in this expression. For this, let X be independent of Θ, with $\mathcal{L}(X) = \mathrm{Unif}(\{-1,1\}^d)$.

For fixed θ, the objective in Eq. (8.2) is $F(\theta) = -\mathbb{E}[\chi_S(X) N_\theta(X)]$. Differentiating this finite average and using that for every $x \in \{-1,1\}^d$,

$$x_i \chi_S(x) = \chi_{S \triangle \{i\}}(x)$$

we get

$$\partial_{u_r} F(\theta) = -\mathbb{E}[\chi_S(X)[w_r^\top X + b_r]_+], \tag{8.20}$$
$$\partial_{b_r} F(\theta) = -u_r q_S(w_r, b_r), \tag{8.21}$$
$$\partial_{w_{r,i}} F(\theta) = -u_r q_{S \triangle \{i\}}(w_r, b_r). \tag{8.22}$$

where for $B \subseteq [d]$, the function $q_B : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is defined by

$$q_B(w, b) = \mathbb{E}[\chi_B(X) \mathbb{I}\{w^\top X + b > 0\}], \quad w \in \mathbb{R}^d, b \in \mathbb{R}. \tag{8.23}$$

The identities stated for the derivatives hold whenever the preactivation margins are nonzero, which occurs almost surely at Gaussian parameters.

We now show that the output-weight derivative also reduces to threshold correlations: since $[z]_+ = z\mathbb{I}\{z > 0\}$,

$$\begin{aligned} \partial_{u_r} F(\theta) &= -\mathbb{E}[\chi_S(X)(w_r^\top X + b_r)\mathbb{I}\{w_r^\top X + b_r > 0\}] \\ &= -b_r q_S(w_r, b_r) - \sum_{i=1}^d w_{r,i} q_{S \triangle \{i\}}(w_r, b_r). \end{aligned} \tag{8.24}$$

We now return to bounding $\mathbb{E}[\|\nabla F(\Theta)\|]$. Plugging Θ into Eq. (8.19), repeatedly applying $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, and then taking expectation of both sides, then applying Jensen's to move the expectation under the remaining square roots yields

$$\mathbb{E}[\|\nabla F(\Theta)\|] \le \sum_{r=1}^{m} \left(\mathbb{E}[|\partial_{u_r} F(\Theta)|] + \sqrt{\mathbb{E}[\|\partial_{w_r} F(\Theta)\|^2] + \mathbb{E}[|\partial_{b_r} F(\Theta)|^2]} \right). \tag{8.25}$$

From our earlier expression for the partial derivatives with respect to the first layer parameters (w_r, b_r) we see that we will need bounds on the expected value of $q_B(W_r, B_r)^2$. The reason the partial derivative with respect to u_r is treated differently will become clear momentarily.

Lemma 8.2.2, stated and proved below, gives

$$\mathbb{E}[q_B(W_r, B_r)^2] \le 8e^{-|B|/4}.$$

Because $|S| = k$ and $|S \triangle \{i\}| \ge k - 1 \ge 2$, this inequality gives

$$\mathbb{E}[q_S(W_r, B_r)^2] \le 8e^{-k/4}, \qquad \mathbb{E}[q_{S\triangle\{i\}}(W_r, B_r)^2] \le 8e^{-(k-1)/4} < 11e^{-k/4}.$$

For the output-weight term, Eq. (8.24) and Cauchy–Schwarz give

$$\begin{aligned} \mathbb{E}[|\partial_{u_r} F(\Theta)|] &\le \sqrt{\mathbb{E}[B_r^2]\mathbb{E}[q_S(W_r, B_r)^2]} + \sqrt{\mathbb{E}[\|W_r\|^2]\,\mathbb{E}\left[\sum_{i=1}^{d} q_{S\triangle\{i\}}(W_r, B_r)^2\right]} \\ &\le \sqrt{8}\,\tau e^{-k/8} + \sqrt{11}\,d\tau e^{-k/8} \le 7d\tau e^{-k/8}. \end{aligned} \tag{8.26}$$

For the input weights and bias, U_r is independent of (W_r, B_r). Thus Eqs. (8.21) and (8.22) give

$$\mathbb{E}[|\partial_{b_r} F(\Theta)|^2] = \mathbb{E}[U_r^2]\mathbb{E}[q_S(W_r, B_r)^2] \le 8\tau^2 e^{-k/4}, \tag{8.27}$$

$$\mathbb{E}[\|\partial_{w_r} F(\Theta)\|^2] = \tau^2 \sum_{i=1}^{d} \mathbb{E}[q_{S\triangle\{i\}}(W_r, B_r)^2] \le 11d\tau^2 e^{-k/4}. \tag{8.28}$$

Substituting these bounds into Eq. (8.25) proves

$$\mathbb{E}[\|\nabla F(\Theta)\|] \le m\tau e^{-k/8}(7d + \sqrt{11d+8}) \le 8md\tau e^{-k/8},$$

where $d > 30$ ensures the last inequality. If $k \ge 72\log(8md\tau)$, then $8md\tau \le e^{k/72}$ and the bound becomes $e^{-k/9}$. Markov's inequality at threshold $e^{-k/18}$ proves Eq. (8.11). □

8.2.3 Why a Gaussian threshold has little correlation with parity

It remains to bound the threshold coefficients used in the gradient proof. Averaging their squares over Gaussian parameters will turn the calculation into a polynomial expansion on the Boolean cube. Every power of degree below $|B|$ cancels.

Lemma 8.2.2 (Gaussian threshold correlation)**.** *Let (W, C) have law $\mathcal{N}(0, \tau^2 I_{d+1})$, with $\tau > 0$. For every $B \subseteq [d]$ with $|B| \ge 2$,*

$$\mathbb{E}[q_B(W, C)^2] \le 8e^{-|B|/4}. \tag{8.29}$$

Proof. To evaluate $\mathbb{E}[q_B(W,C)^2]$, expand the two copies of the input average. Let X, X' be independent uniform random vectors in $\{-1,1\}^d$, independent of (W,C). All probabilities and expectations in this proof refer to this joint experiment. Then

$$\mathbb{E}[q_B(W,C)^2] = \mathbb{E}\Big[\chi_B(X)\chi_B(X')\mathbb{P}(W^\top X + C > 0,\ W^\top X' + C > 0 \mid X, X')\Big].$$

For $x, x' \in \mathbb{R}^d$ fixed, $W^\top x + C, W^\top x' + C$ are centered Gaussian variables with common variance $\tau^2(d+1)$ and covariance $\tau^2(x^\top x' + 1)$. Their correlation is therefore $(x^\top x' + 1)/(d+1)$. The Gaussian orthant formula gives

$$\mathbb{E}[q_B(W,C)^2] = \mathbb{E}\left[\chi_B(X)\chi_B(X')\left(\frac{1}{4} + \frac{1}{2\pi}\arcsin\frac{X^\top X' + 1}{d+1}\right)\right]. \tag{8.30}$$

Set $Z_i = X_i X_i'$ and $V = (1+\sum_i Z_i)/(d+1)$. The Z_i are independent uniform signs, $\chi_B(X)\chi_B(X') = \chi_B(Z)$, and $\mathbb{E}[\chi_B(Z)] = 0$. Hence

$$\mathbb{E}[q_B(W,C)^2] = \frac{1}{2\pi}\mathbb{E}[\chi_B(Z)\arcsin V].$$

To bound this expression, we use the Taylor expansion

$$\arcsin v = \sum_{\ell\geq 0} a_\ell v^\ell, \qquad |v| \leq 1, \qquad a_\ell \geq 0, \qquad \sum_{\ell\geq 0} a_\ell = \frac{\pi}{2}.$$

The series converges absolutely and uniformly on $[-1,1]$; Appendix 8.A gives the coefficients and proves these properties. Absolute summability justifies exchanging the series and expectation, giving

$$\mathbb{E}[q_B(W,C)^2] = \frac{1}{2\pi}\sum_{\ell\geq 0} a_\ell \mathbb{E}[\chi_B(Z)V^\ell].$$

We first eliminate every term of degree $\ell < |B|$. Expanding $(1+\sum_i Z_i)^\ell$, each monomial **omits** at least one coordinate in B. Averaging over that independent mean-zero sign gives

$$\mathbb{E}[\chi_B(Z)V^\ell] = 0, \qquad \ell < |B|. \tag{8.31}$$

Since $|V| \leq 1$, for the remaining terms with $\ell \geq |B|$ we have $|V|^\ell \leq |V|^{|B|}$ and hence

$$\mathbb{E}[q_B(W,C)^2] \leq \frac{1}{2\pi}\sum_{\ell\geq|B|} a_\ell \mathbb{E}[|V|^\ell] \leq \frac{1}{4}\mathbb{E}[|V|^{|B|}].$$

We now bound this single moment. Hoeffding's inequality gives

$$\mathbb{P}\left(\left|\sum_{i=1}^d Z_i\right| > \frac{3d}{4}\right) \leq 2e^{-9d/32}.$$

On the complementary event,

$$|V| \leq \frac{3d/4+1}{d+1} = \frac{3}{4}\left(1 + \frac{1}{3(d+1)}\right) \leq \frac{3}{4}\exp\left(\frac{1}{3(d+1)}\right),$$

where the last inequality used that $1+t \leq e^t$. Raising this bound to power $|B|$ and using $|B| \leq d+1$ gives $|V|^{|B|} \leq e^{1/3}(3/4)^{|B|}$. Using $|V| \leq 1$ on the exceptional event, we conclude that

$$\mathbb{E}[q_B(W,C)^2] \leq \frac{1}{4}\left(2e^{-9d/32} + e^{1/3}(3/4)^{|B|}\right) \leq 8e^{-|B|/4},$$

where the last inequality uses $d \geq |B|$, $9/32 > 1/4$, and $\log(4/3) > 1/4$. □

On the Boolean cube, the functions χ_B form the *Fourier character* basis. The character χ_B has degree $|B|$. Thus Eq. (8.31) says that every polynomial of degree below $|B|$ has zero coefficient in the χ_B direction. This cancellation leaves only high powers of V, whose moments are exponentially small. The scale τ disappears because multiplying (W, C) by a positive number does not change the threshold.[2]

8.3 Intermediate labels expose a repeated two-bit computation

The lower bound applies to direct training on the final parity label. To change the supervised task while preserving its final answer, write $x_i = (-1)^{b_i}$ with $b_i \in \{0, 1\}$ and define

$$z_1 = b_1, \qquad z_i = z_{i-1} \oplus b_i, \qquad 2 \le i \le d. \tag{8.32}$$

XOR is associative, so $z_i = b_1 \oplus \cdots \oplus b_i$ and $(-1)^{z_d} = \chi_{[d]}(x)$. Supplying $z_1, \ldots, z_d$ during training gives a label for every application of the same local rule. The following proposition gives that rule explicitly and verifies that composing it solves the original parity task.

Proposition 8.3.1 (A local XOR rule composes into parity)**.** *The function*

$$g(a, b) = [a + b]_+ - 2[a + b - 1]_+ + [a + b - 2]_+$$

equals $a \oplus b$ on $\{0, 1\}^2$. If $\widehat{z}_1 = b_1$ and $\widehat{z}_i = g(\widehat{z}_{i-1}, b_i)$, then

$$\widehat{z}_i = b_1 \oplus \cdots \oplus b_i$$

for every i. In particular, $(-1)^{\widehat{z}_d} = \chi_{[d]}(x)$.

Proof. We prove the running-parity identity by induction. At $i = 1$ it holds because $\widehat{z}_1 = b_1$. For the induction step, we need $g(a, b) = a \oplus b$ for binary a, b. Evaluating the ReLU expression at $a + b = 0, 1, 2$ gives $0, 1, 0$, respectively, which is exactly XOR. Thus, if the identity holds at $i - 1$, then

$$\widehat{z}_i = g(\widehat{z}_{i-1}, b_i) = \widehat{z}_{i-1} \oplus b_i = b_1 \oplus \cdots \oplus b_i.$$

Taking $i = d$ proves the final claim. □

The recurrence in Eq. (8.32) defines the computation graph in Fig. 1. Each box applies the same function g. If a neural network implements g and shares its parameters across positions, the graph is an unrolled recurrent neural network:

$$\widehat{z}_i = g_\theta(\widehat{z}_{i-1}, b_i), \qquad 2 \le i \le d.$$

Manually assembling this graph supplies the routing between copies of the local rule. A sequence-to-sequence model can instead produce $z_1, \ldots, z_d$ with one shared model. The intermediate labels then provide process supervision at every position, without training and composing a separate model for each step. The model must learn both the local operation and which earlier value to retrieve. At inference time it must also remain accurate when its own earlier predictions replace the correct prefixes.

Fix the predictors after training. The following proposition relates their errors on correct prefixes to the errors made when their own predictions are supplied as prefixes.

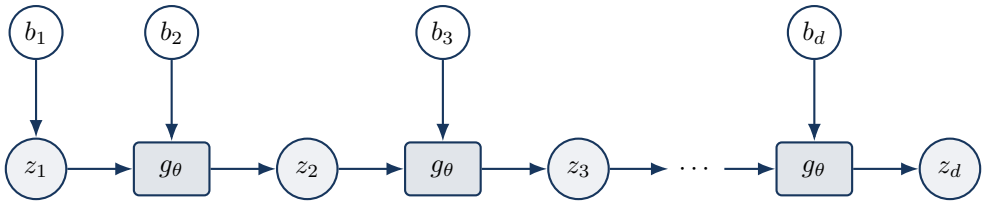


Figure 1: A computation graph for running parity. Associativity of XOR makes the final state equal to the parity of all input bits. The repeated label g_θ indicates parameter sharing, as in a recurrent neural network.

Proposition 8.3.2 (Local errors control the probability of a rollout error)**.** *Let $(X, Z_{1:d})$ be a random input and target sequence taking values in $\mathcal{X} \times \mathcal{Z}^d$, where $\mathcal{Z}$ is a finite output alphabet. For fixed predictors*

$$g_i : \mathcal{X} \times \mathcal{Z}^{i-1} \to \mathcal{Z}, \qquad i \in [d],$$

define the rollout recursively by

$$\widehat{Z}_i = g_i(X, \widehat{Z}_{<i}), \qquad i \in [d],$$

where $Z_{<1}$ and $\widehat{Z}_{<1}$ are empty prefixes. Define the teacher-forced error events

$$E_i = \{g_i(X, Z_{<i}) \neq Z_i\}, \qquad i \in [d].$$

Then

$$\{\widehat{Z}_{1:d} \neq Z_{1:d}\} = \bigcup_{i=1}^{d} E_i. \tag{8.33}$$

Consequently,

$$\mathbb{P}(\widehat{Z}_d \neq Z_d) \leq \mathbb{P}(\widehat{Z}_{1:d} \neq Z_{1:d}) \leq \sum_{i=1}^{d} \mathbb{P}(E_i).$$

Proof. We characterize the event that the first i predictions are all correct. On $\{\widehat{Z}_{<i} = Z_{<i}\}$, the two evaluations of g_i have the same arguments. Hence

$$\{\widehat{Z}_{1:i} = Z_{1:i}\} = \{\widehat{Z}_{<i} = Z_{<i}\} \cap E_i^c.$$

Starting with the empty prefix and applying this identity successively gives

$$\{\widehat{Z}_{1:d} = Z_{1:d}\} = \bigcap_{i=1}^{d} E_i^c.$$

Taking complements proves Eq. (8.33). Finally, $\{\widehat{Z}_d \neq Z_d\} \subseteq \{\widehat{Z}_{1:d} \neq Z_{1:d}\}$, and the union bound gives the stated probability inequalities. □

Thus $\mathbb{P}(E_i) \leq \varepsilon$ for every i gives a final-answer error bound of $\min\{1, d\varepsilon\}$. To guarantee a final-answer error at most δ, it suffices to have $\sum_i \mathbb{P}(E_i) \leq \delta$; equal per-step bounds of δ/d are one way to ensure this.

8.4 The path-star task hides parity in its first decision

A path-star graph consists of several directed paths that share a center. An example supplies the shuffled edge list, the center, a designated leaf, and the center-to-leaf path. During forward teacher forcing, the model receives the true preceding path token. Once the first edge has been chosen, each internal node has a unique outgoing edge along its arm. A predictor can therefore learn almost every next-token prediction while still guessing which arm contains the goal.

The remaining first decision can encode full parity. Given bits $b_1, \dots, b_{q-1}$, use vertices $-q, \dots, -1, 0, 1, \dots, q$. Connect 0 to both 1 and -1. For $1 \le i < q$, add

	edge from i	edge from $-i$
$b_i = 0$	$i \to i+1$	$-i \to -(i+1)$
$b_i = 1$	$i \to -(i+1)$	$-i \to i+1$.

Figure 2 shows the construction for three bits. It also marks the first branch that must be chosen to reach the designated goal.

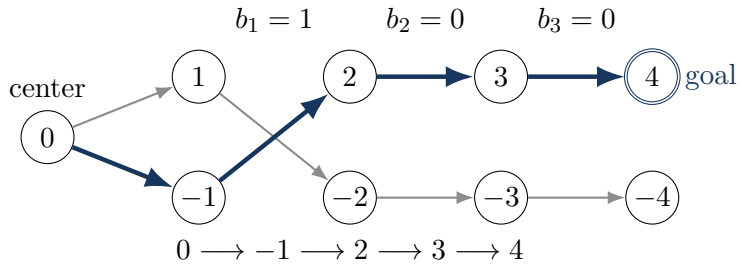


Figure 2: The two-arm path-star for $(b_1, b_2, b_3) = (1, 0, 0)$. The highlighted path reaches goal 4. Since $1 \oplus 0 \oplus 0 = 1$, its first vertex is -1. Each later step follows a unique outgoing edge.

Each bit equal to one switches the sign of the current vertex. Therefore the path that starts with $s \in \{-1, 1\}$ reaches the leaf $sq(-1)^{b_1 + \cdots + b_{q-1}}$. To reach the goal q, its first step must be

$$\begin{cases} 1, & b_1 \oplus \cdots \oplus b_{q-1} = 0, \\ -1, & b_1 \oplus \cdots \oplus b_{q-1} = 1. \end{cases} \tag{8.34}$$

Predicting the correct first edge on every graph in this family computes full parity. The example separates high average teacher-forced accuracy from learning the decision needed to generate the correct path. Backward or intermediate supervision reveals information about the goal arm before the first forward decision.

8.5 Optional reading: a constructed transformer learns retrieval and XOR

The composition in Section 8.3 supplied the connections between copies of the XOR rule. A sequence model must learn which input to retrieve as well as how to combine it with the current state. The following construction trains both components with intermediate labels. Its three updates have distinct roles: create a retrieval signal, concentrate attention, and fit the local output rule.

The supervised sequence. Keep d for the number of input bits. Fix

$$\mathcal{I} = \{d+1, \dots, d+k\}, \qquad s : \mathcal{I} \to [d] \text{ injective}, \qquad S = s(\mathcal{I}).$$

The teacher chooses s and supplies sequences with the running parities

$$b_{d+1} = 0, \qquad b_{i+1} = b_i \oplus b_{s(i)}, \quad i \in \mathcal{I}. \tag{8.35}$$

Thus $b_{d+k+1} = \bigoplus_{j \in S} b_j$. Let

$$\mathcal{B}_s = \{b \in \{0,1\}^{d+k+1} : b_{d+1} = 0,\ b_{i+1} = b_i \oplus b_{s(i)} \text{ for all } i \in \mathcal{I}\}.$$

The training law P_{CoT} is the law of the random sequence $\mathbf{B} = (B_1, \dots, B_{d+k+1})$ obtained from independent uniform input bits $B_1, \dots, B_d$ and the recursion in Eq. (8.35). The student receives these sequences; s is learned from them.

At position i, the model receives the correct prefix $b_{1:i}$ and predicts b_{i+1}. We use the signed label

$$y_i(b) = 2b_{i+1} - 1 \in \{-1, 1\}.$$

A positive prediction score represents bit 1. Supplying the correct prefix at each supervised position is *teacher forcing*.

The model. Let $D \in \mathbb{N}$ be the embedding dimension and $m \in \mathbb{N}$ the feed-forward width. For each position $i \in [d+k]$ and bit $a \in \{0,1\}$, draw an independent frozen embedding $e_{i,a} \in \mathbb{R}^D$ with

$$\mathcal{L}(e_{i,a}) = \mathrm{Unif}\big(\{-D^{-1/2}, D^{-1/2}\}^D\big), \qquad v_i(b) = e_{i,b_i}.$$

The trainable attention matrix is $A \in \mathbb{R}^{D \times D}$. Its scores, causal attention weights, and retrieved vector are

$$\begin{aligned} s_{ji}(A; b) &= v_j(b)^\top A v_i(b), \\ \alpha_{ji}(A; b) &= \frac{\exp s_{ji}(A; b)}{\sum_{\ell=1}^{i} \exp s_{\ell i}(A; b)}, \quad 1 \le j \le i, \\ c_i(A; b) &= \sum_{j=1}^{i} \alpha_{ji}(A; b) v_j(b). \end{aligned} \tag{8.36}$$

Set $\alpha_{ji} = 0$ for $j > i$. The two arguments of the local computation enter separate blocks of the feed-forward weight matrix:

$$W = [W^{\mathrm{tok}}\ W^{\mathrm{att}}] \in \mathbb{R}^{m \times 2D}, \qquad W^{\mathrm{tok}}, W^{\mathrm{att}} \in \mathbb{R}^{m \times D}.$$

With fixed $h \in \mathbb{R}^m$, the score at position i is

$$f_{W,A}(b)_i = h^\top \left[W \begin{pmatrix} v_i(b) \\ c_i(A; b) \end{pmatrix} \right]_+ = h^\top [W^{\mathrm{tok}} v_i(b) + W^{\mathrm{att}} c_i(A; b)]_+. \tag{8.37}$$

Here ReLU acts coordinatewise. Concatenation gives the feed-forward layer separate access to the current token and the attention output.

Initialization and training. The initialization is designed so that the current token selects the ReLU gates before attention contributes to the score. For an even width m and an amplitude $\epsilon > 0$, set

$$\begin{aligned} A^{(0)} = 0, \qquad (W^{\mathrm{att}})^{(0)} = 0, \\ (W^{\mathrm{tok}})^{(0)}_{r,:} = \sum_{i\in\mathcal{I}} \sum_{a=0}^{1} \nu_{r,i,a} e_{i,a}^{\top}, \qquad \mathcal{L}(\nu_{r,i,a}) = \mathrm{Unif}\{-\epsilon, \epsilon\}, \\ h_r = \begin{cases} 1/m, & 1 \le r \le m/2, \\ -1/m, & m/2 < r \le m. \end{cases} \end{aligned} \tag{8.38}$$

All $\nu_{r,i,a}$ are independent of one another and of the embeddings. The embeddings and h remain fixed. Both W and A are updated.

For one sequence, the *hinge loss* summed over supervised positions is

$$\ell(W, A; b) = \sum_{i\in\mathcal{I}} \ell_i(W, A; b), \qquad \ell_i(W, A; b) = [1 - y_i(b) f_{W,A}(b)_i]_+.$$

For each $t = 0, 1, 2$, draw a fresh batch of n independent sequences $\mathbf{B}^{(t,1)}, \ldots, \mathbf{B}^{(t,n)}$ with law P_{CoT}. The three batches are mutually independent and independent of initialization. With $\vartheta = (W, A)$, define

$$\widehat{F}_t(W, A) = \frac{1}{n} \sum_{a=1}^{n} \ell(W, A; \mathbf{B}^{(t,a)}), \qquad \vartheta^{(t+1)} = \vartheta^{(t)} - \eta_t \nabla \widehat{F}_t(\vartheta^{(t)}). \tag{8.39}$$

The gradient updates all entries of both matrices at each step.

The next result states parameter scales for which these three simultaneous updates learn every supervised prediction and the required retrieval map. All asymptotic statements concern sufficiently large d.

Theorem 8.5.1 (Three-update parity learning under structured initialization)**.** *Fix $\delta \in (0, 1/2)$ and a sufficiently small constant $\epsilon > 0$, and set $\Lambda = \log(d/\delta)$. Suppose*

$$\frac{d}{\Lambda^5} \le k \le \frac{d}{\Lambda^4}.$$

There are absolute constants $C_{\mathrm{batch}}, c_{\mathrm{init}}, c_{\mathrm{final}} > 0$ such that the model, initialization, and updates in Eqs. (8.36) to (8.39), with

$$D = \lceil k\Lambda^{11/10} \rceil, \qquad m = 4k, \qquad n = \lceil C_{\mathrm{batch}} \epsilon^{-2} d\Lambda^{20} \rceil,$$
$$\eta_0 = \eta_1 = c_{\mathrm{init}} \frac{m\epsilon\sqrt{n}}{\Lambda}, \qquad \eta_2 = c_{\mathrm{final}} k\epsilon,$$

satisfy, with probability at least $1 - \delta$,

$$\begin{aligned} &y_i(b) f_{W^{(3)}, A^{(3)}}(b)_i > 0, \\ &\alpha_{s(i),i}(A^{(3)}; b) \ge 1 - d^{-8} \end{aligned} \qquad \text{for every } b \in \mathcal{B}_s \text{ and } i \in \mathcal{I}. \tag{8.40}$$

The probability includes the random embeddings, initialization signs, and three training batches.

The first inequality gives the correct next bit at every correct prefix. Starting from the supplied input and $b_{d+1} = 0$, induction therefore gives the correct entire running-parity sequence during generation. All three batches have the same sequence length $d + k + 1$.

Proof outline. We need to establish correct retrieval and a positive signed output at every supervised position. The proof tracks these quantities through the three updates. We give the gradient identities and the estimates that connect the steps; the accompanying concentration arguments are in the source identified in the bibliographical notes.

For fixed $b \in \mathcal{B}_s$, abbreviate $v_i = v_i(b)$ and $c_i^{(t)} = c_i(A^{(t)}; b)$. Define the preactivation of unit r by

$$\rho_{r,i}^{(t)}(b) = \big((W^{\mathrm{tok}})^{(t)} v_i + (W^{\mathrm{att}})^{(t)} c_i^{(t)}\big)_r.$$

We will use the diagonal gate matrix

$$\Gamma_{i,a} = \mathrm{diag}\big(\mathbb{I}\{\nu_{r,i,a} > 0\} : r \in [m]\big).$$

First update: the feed-forward weights acquire a retrieval signal. At initialization, $A^{(0)} = 0$ gives uniform attention:

$$\alpha_{ji}(A^{(0)}; b) = \frac{1}{i}, \qquad c_i^{(0)} = \frac{1}{i} \sum_{j=1}^{i} v_j.$$

The initialization separates the matching embedding from the remaining terms:

$$\rho_{r,i}^{(0)}(b) = \nu_{r,i,b_i} + \sum_{\substack{j \in \mathcal{I},\ a \in \{0,1\} \\ (j,a) \ne (i,b_i)}} \nu_{r,j,a} \langle e_{j,a}, e_{i,b_i} \rangle.$$

Concentration of the random sum gives, with high probability,

$$|\rho_{r,i}^{(0)}(b) - \nu_{r,i,b_i}| \le \epsilon/4 \quad \Longrightarrow \quad \mathbb{I}\{\rho_{r,i}^{(0)}(b) > 0\} = \mathbb{I}\{\nu_{r,i,b_i} > 0\}.$$

Thus the gates are Γ_{i,b_i}. Since $(W^{\mathrm{att}})^{(0)} = 0$, the chain rule gives

$$\nabla_A \widehat{F}_0(W^{(0)}, A^{(0)}) = 0, \qquad A^{(1)} = 0.$$

The output bounds give $|f_{W^{(t)},A^{(t)}}(b)_i| < 1$ at each gradient evaluation $t = 0, 1, 2$. Hence the hinge derivative is $-y_i(b)$. At the first step the gradient through the attention-input block is

$$\nabla_{W^{\mathrm{att}}} \ell_i(W^{(0)}, A^{(0)}; b) = -y_i(b)\Gamma_{i,b_i} h(c_i^{(0)})^\top. \tag{8.41}$$

Why does this gradient identify the required bit? Fix the initialization. For a with $P_{\mathrm{CoT}}(B_i = a) > 0$, independence of the fresh input bit $B_{s(i)}$ from the preceding running parity gives

$$\mathbb{E}_{P_{\mathrm{CoT}}}[y_i(\mathbf{B})v_j(\mathbf{B}) \mid B_i = a] = \begin{cases} \dfrac{(-1)^a}{2}(e_{s(i),1} - e_{s(i),0}), & j = s(i), \\ 0, & j \le i,\ j \ne s(i). \end{cases} \tag{8.42}$$

Substituting $c_i^{(0)} = i^{-1}\sum_{j\le i} v_j$ into Eq. (8.41) therefore yields

$$-\mathbb{E}_{P_{\mathrm{CoT}}}[\nabla_{W^{\mathrm{att}}} \ell_i(W^{(0)}, A^{(0)}; \mathbf{B}) \mid B_i = a] = \frac{(-1)^a}{2i}\Gamma_{i,a} h(e_{s(i),1} - e_{s(i),0})^\top.$$

The displayed expected gradient supplies the supervised signal for retrieving position $s(i)$. The finite batch analysis controls the sampling error and the sum of contributions from the different output positions.

Second update: the retrieval signal trains attention. The first update makes $(W^{\mathrm{att}})^{(1)}$ nonzero. The derivative of the score with respect to the attention output is now the vector

$$\beta_{i,a}^{(1)} = ((W^{\mathrm{att}})^{(1)})^\top \Gamma_{i,a} h \in \mathbb{R}^D.$$

For the active hinge loss, the chain rule and softmax derivative give

$$\begin{aligned} \nabla_A \ell_i &= -y_i(b) \sum_{j=1}^{i} \langle \beta_{i,b_i}^{(1)}, v_j \rangle \nabla_A \alpha_{ji}, \\ \nabla_A \alpha_{ji} &= \alpha_{ji}(v_j - c_i) v_i^\top. \end{aligned} \tag{8.43}$$

Both derivatives are evaluated at $(W^{(1)}, A^{(1)})$. The alignment acquired in the first update now increases the attention score of $s(i)$. The main estimate for the second update is a uniform score gap

$$\Delta_i^{(2)}(b) := s_{s(i),i}(A^{(2)}; b) - \max_{\substack{j \le i \\ j \ne s(i)}} s_{ji}(A^{(2)}; b) \ge 20 \log d.$$

The simultaneous feed-forward updates are controlled through the preactivations. The proof bounds, uniformly over r, i, b,

$$|\rho_{r,i}^{(0)}(b) - \nu_{r,i,b_i}| \le \epsilon/4, \qquad |\rho_{r,i}^{(t)}(b) - \rho_{r,i}^{(0)}(b)| \le \epsilon/4, \quad t = 1, 2.$$

Since $|\nu_{r,i,b_i}| = \epsilon$, the gates remain Γ_{i,b_i} through both preparatory updates. These inequalities quantify why the changes in attention and in the feed-forward weights can coexist.

Third update: the retrieved bit determines the local output. A score gap Δ implies

$$\begin{aligned} 1 - \alpha_{s(i),i} &= \frac{\sum_{j \neq s(i)} \exp(s_{ji} - s_{s(i),i})}{1 + \sum_{j \neq s(i)} \exp(s_{ji} - s_{s(i),i})} \\ &\leq (i-1)e^{-\Delta}. \end{aligned}$$

With $i \leq d + k \leq 2d$ and $\Delta \geq 10 \log d$, this is at most d^{-8}. Since every embedding has norm one, it also gives

$$\|c_i - e_{s(i),b_{s(i)}}\| \leq 2(1 - \alpha_{s(i),i}) \leq 2d^{-8}.$$

The feed-forward input is therefore close to the pair $(e_{i,b_i}, e_{s(i),b_{s(i)}})$. Its target is the four-case map $(b_i, b_{s(i)}) \mapsto b_i \oplus b_{s(i)}$.

The third update fits this map. The final estimates have the form

$$\left| f_{W^{(3)},A^{(3)}}(b)_i - \frac{\epsilon}{3} y_i(b) \right| < \frac{\epsilon}{3}, \qquad \Delta_i^{(3)}(b) \geq 10 \log d.$$

The first inequality gives $y_i(b) f_{W^{(3)},A^{(3)}}(b)_i > 0$; the second preserves the retrieval bound during the final simultaneous update. The initialization and batch concentration estimates make the bounds hold jointly with probability at least $1 - \delta$. □

The frozen embeddings, concatenated inputs, zero attention branch, hinge loss, and three prescribed learning rates are substantive assumptions of this construction. Its range $k \leq d/\log^4(d/\delta) < d$ covers sparse parities. The full-parity construction in Section 8.3 has $k = d$ and supplies the retrieval connections explicitly.

8.6 Optional reading: process supervision and a length curriculum

The running-parity construction suggests a broader question: can a model learn a local state update, retrieve its arguments, and repeatedly apply it to solve a longer computation? The following results give optimization guarantees for this approach on a family of synthetic tasks. They extend the positive examples beyond XOR and give a role to increasing the context length during training.

The task and supervision. Let $\mathcal{Y}$ be a finite state set and G a finite group of permutations of $\mathcal{Y}$: a set of invertible maps containing the identity and closed under composition and inverses. A *state-tracking task* asks for the result of successive updates

$$y_t = g_t(y_{t-1}), \qquad t = 1, \ldots, L. \tag{8.44}$$

The input consists of formal assignment instructions $a_t = g_t(a_{t-1})$ and an initial assignment $a_0 = y_0$, where the a_t are distinct variable identifiers from a set $\mathcal{A}$. The supervised answers are $a_t = y_t$ at every step. Each instruction and answer is encoded as a five-token clause:

$$(a_t, g_t, a_{t-1}, \bot, \bot), \qquad (\bot, \bot, \bot, a_t, y_t),$$

where $\bot$ is a blank token. Thus the representation already separates the roles of identifiers, actions, and states.

For the training and evaluation experiment, choose identifiers $A_0, \ldots, A_L$ from $\mathcal{A}$ uniformly without replacement, Y_0 uniformly on $\mathcal{Y}$, and $G_1, \ldots, G_L$ independently and uniformly on G. These three choices are independent. Set $Y_t = G_t(Y_{t-1})$. Let C contain all instruction clauses and the

initial assignment, and let Z_t denote the correct answer clause at step t. For parameters ϑ, write $p_\vartheta(z \mid c, z_{<t})$ for the model's probability of answer clause z. Its population log-loss objective and *teacher-forced accuracy* are

$$\begin{aligned} \mathrm{Loss}_L(\vartheta) &= \sum_{t=1}^{L} \mathbb{E}[-\log p_\vartheta(Z_t \mid C, Z_{<t})], \\ \mathrm{Acc}_L(\vartheta) &= \frac{1}{L}\sum_{t=1}^{L} \mathbb{E}[p_\vartheta(Z_t \mid C, Z_{<t})]. \end{aligned} \tag{8.45}$$

The accuracy averages the probability of sampling the correct next clause when supplied with the correct preceding clauses.

Architecture and training restrictions. The vocabulary consists of the identifier, action, state, and blank tokens; let M be its size and $r = |\mathcal{Y}|$. In the parameter regimes below, the identifier set has size $|\mathcal{A}| = \Theta(M)$. The model has one attention layer followed by a feed-forward network (FFN). The following choices are part of the construction and its analysis.

- Nonblank tokens have fixed orthonormal embeddings in $\mathbb{R}^M$; the blank embedding is zero. Concatenating the five embeddings gives one vector in $\mathbb{R}^{5M}$ per clause. There is no positional encoding.
- Attention uses scores $u^\top Q v$, for clause embeddings u, v. Partition Q into $M \times M$ blocks. Only $Q_{4,3}$ and $Q_{4,4}$ are trainable; all other blocks are zero. These blocks compare the current identifier with an instruction's input identifier and an answer's identifier. Initialize $Q = 0$, giving uniform attention.
- There are m hidden units for each output slot and vocabulary token, with $r < m$ and $m = o(\log^2 M)$. Their input weights have independent Gaussian coordinates of variance $1/M$; biases are fixed at $M^{-1/2} \log M$ and output coefficients at one. The five output distributions are multiplied to form the next-clause kernel.
- The FFN analysis uses a prescribed modified smoothed-ReLU activation and clips logits above a constant times $\log M$. For the state-output neurons it also freezes the weight blocks receiving identifier slots $1, 3, 4$. The action and state slots $2, 5$ supply the trainable inputs to those neurons.
- Training uses exact population gradients. First train the FFN on length 1, holding $Q = 0$. Then freeze the FFN and train attention on length 2, using the log loss for the state token. Suitable learning rates are inverse polynomial in M; the prescribed numbers of updates are polynomial in M.

At length 1, the context contains exactly the instruction and the initial assignment. Uniform attention therefore supplies both arguments of the local update. Freezing the learned FFN then makes retrieval the remaining optimization problem. This training schedule explicitly controls the interaction between the two components.

Two optimization guarantees. Here are the reported guarantees for this construction, suppressing the polynomial exponents. Their initialization estimates hold with probability $1 - o(1)$ as $M \to \infty$.

(i) **Learning on lengths 1 and 2 can suffice.** An action is *simply transitive* if for every pair $y, y' \in \mathcal{Y}$ there is exactly one $g \in G$ with $g(y) = y'$. With

$$\Omega(\log\log M) \le r \le \log M, \qquad |G| = r,$$

training the FFN on length 1, followed by attention training on both answer positions of length 2, yields parameters ϑ_* for which

$$1 - \mathrm{Acc}_L(\vartheta_*) \le M^{-\Omega(1)} \quad \text{for all } L \le O(M^c), \qquad 0 < c < 1, \tag{8.46}$$

for a fixed constant c. Cyclic shifts on r states give an example of such an action.

(ii) **A length curriculum can extend the guarantee.** A *length curriculum* increases context length across training stages. Take $G = \mathrm{Sym}(\mathcal{Y})$, the group of all permutations, with

$$r = \Theta\left(\frac{\log\log M}{\log\log\log M}\right), \qquad |G| = r! = \mathrm{polylog}\, M.$$

After the initial FFN training, train attention on the second answer of a length-2 problem. Subsequent stages use context lengths $L_k = 2^k$. At each stage the preceding model greedily generates the first two answer clauses; attention is trained on the second state token with these generated labels and prefixes. The FFN remains frozen. For a sufficiently small inverse-polynomial population-loss threshold, the resulting parameters ϑ_k satisfy

$$1 - \mathrm{Acc}_{L_{k+1}}(\vartheta_k) \le M^{-\Omega(1)}, \qquad L_{k+1} = 2L_k, \tag{8.47}$$

whenever that length is admissible. After $O(\log M)$ stages, the guarantee covers lengths up to $|\mathcal{A}| - 1 = \Theta(M)$. The restriction $L + 1 \le |\mathcal{A}|$ comes from sampling distinct identifiers.

The permutation task includes an NC^1-complete family of computations.[3] Thus the positive optimization results extend to a substantially broader class of repeated computations. The complexity classification concerns worst-case inputs; the learning guarantees average over the specified random inputs. The proof separates learning the local operation from learning attention that retrieves its arguments reliably among increasingly many irrelevant clauses.

From next-clause accuracy to a generated trace. Fix trained parameters ϑ and generate $\widehat{Z}_1, \dots, \widehat{Z}_L$ using p_ϑ on its own preceding outputs. To bound the probability of a mistake, condition on the input C. The correct trace is determined by C, so

$$\mathbb{P}(\widehat{Z}_{1:L} = Z_{1:L} \mid C) = \prod_{t=1}^{L} p_\vartheta(Z_t \mid C, Z_{<t}).$$

For numbers $a_t \in [0,1]$, $1 - \prod_t a_t \le \sum_t (1 - a_t)$. Taking expectations therefore gives

$$\mathbb{P}(\widehat{Z}_{1:L} \ne Z_{1:L}) \le L\big(1 - \mathrm{Acc}_L(\vartheta)\big). \tag{8.48}$$

This is the sampled-generation version of the first-error argument in Proposition 8.3.2. A whole-trace error at most δ follows when $1 - \mathrm{Acc}_L \le \delta/L$; the error exponent and the generated length must be compared before claiming reliable complete traces.

What this supports about process supervision. The result makes a concrete positive case: intermediate labels can train a reusable local operation and its retrieval rule, and a curriculum can extend the contexts in which they work. The guarantee applies to the full construction: synthetic uniformly sampled tasks, exact initial traces, structured embeddings and attention, the modified FFN, and staged population optimization. Its self-training guarantee starts from the accuracy obtained in the supervised stages. Extending it to finite batches, simultaneous training of all components, or less structured data requires further analysis. The contribution is an optimization guarantee for this way of teaching a multistep computation. A separation from final-answer-only training would additionally require a lower bound for the same model and data regime.

8.7 Take-home messages

- At isotropic Gaussian parameters, every parity-dependent gradient coordinate is controlled by a high-degree threshold correlation. Averaging over the parameter law makes the complete gradient norm exponentially small.
- The Gaussian comparison walk has small gradients at every step with high probability. The KL chain rule, Pinsker's inequality, and a coupling show that the trained trajectory has nearly the same law. The symmetry obtained by flipping a parity input coordinate gives the classification-risk bounds.
- Running-parity labels replace one degree-d final target by repeated examples of a two-bit rule while preserving the original final answer. Associativity expresses this calculation as a recurrent computation graph.
- Sequence-to-sequence prediction permits process supervision within one shared model; it does not require composing separately trained models. The model must still retrieve the correct arguments and control error during rollout. The path-star reduction and the three-update transformer theorem isolate these additional requirements.
- The transformer theorem proves that the components can cooperate under a structured architecture, initialization, and update schedule. These assumptions limit what the theorem says about ordinary training.
- For structured state-tracking tasks, process supervision can train a local update and its retrieval rule; a length curriculum can extend the supported context lengths. The guarantees concern population optimization with staged training, and their next-step accuracy must be converted to a rollout bound before asserting reliability of a complete computation.

Endnotes

1. Zero network outputs. Gaussian parameters can give the network output positive probability of being zero. For each fixed x, the m independent centered Gaussian margins are all negative with probability 2^{-m}; on that event every ReLU is zero. The proof of Theorem 8.1.2 works with either fixed binary value for sign(0).

2. General activations. Exercise 3 asks you to extend the lower bound using the nonnegative series coefficients of Gaussian activation kernels, parity cancellation, and bounds on Gaussian second moments.

3. The class NC^1 consists of decision problems computable by polynomial-size, bounded-fan-in Boolean circuits of depth $O(\log L)$. Deciding whether a product of permutations on a fixed set of at least five states is the identity gives a complete problem under appropriate reductions.

Bibliographical notes

The fixed-parity lower bound, Gaussian threshold estimate, Gaussian-gradient bound, and trajectory comparison are based on Shoshani and Shamir [2025]. The running-parity transformer theorem and its three-update proof are from Wen et al. [2025], Theorem 3 and Appendix B.4–B.5. In Section 8.5, d denotes their input length, D the embedding dimension, m the full feed-forward width, and n the number of examples per batch. The constants in the hyperparameter choices are suppressed; the dimensions, initialization, loss, and three simultaneous updates are explicit. The concentration estimates omitted from the proof outline control the embedding inner products, initialization gates, and batch correlations. Kim and Suzuki [2025] prove a different positive result using a tree-structured parity decomposition in another simplified transformer model. The path-star reduction and the forward-prediction shortcut are from Hu et al. [2025].

Section 8.6 summarizes Huang et al. [2025], Theorems 4.1 and 5.1, Corollary 5.1, and Algorithms 1–2. Our M is their vocabulary size d; our r is their state count n_y. The accuracy in Eq. (8.45) is their equation (5). The architectural, initialization, and distribution restrictions are in Section 3; the additional frozen FFN blocks and modified activation are in Assumptions B.1–B.2. These appendix restrictions matter when interpreting the main theorems' training claims. Their Theorem 2.1 discusses the complexity of the permutation task. Equation (8.48) records the elementary conversion to sampled whole-trace accuracy used in these notes.

Exercise 3 asks you to extend the fixed-parity argument to general activations. The Gaussian kernel expansion supplied in the exercise follows from the Hermite expansion identity in Daniely et al. [2016], Section 8.

Exercise 2 contrasts the input-coordinate symmetry used in Theorem 8.1.2 with a global parameter sign flip.

Exercise 5 asks you to identify the derivative features whose label correlations determine a gradient step, and to quantify the sampling error in estimating one correlation. Exercise 6 develops the signed, bias-free single-neuron squared-loss example from Shoshani and Shamir [2025]. Its supplied estimates are the zero-bias Gaussian threshold bound and the majority coefficient used in their weak-learning construction. Exercise 7 asks you to prove convergence when learning the output weights of a fixed two-bit gate, and then compose the learned rule.

The memory theorem supplied in Exercise 8 follows from Lyu et al. [2023], Theorem 6 and the sparse-parity example in Section 3.3. It concerns recovery of a uniformly unknown subset from random examples. Wen et al. [2025], Theorem 2 and Appendix B.3, apply this type of memory bound to finite-precision transformer training.

Glossary

Fourier character
For $B \subseteq [d]$, the Boolean-cube function $\chi_B(x) = \prod_{j \in B} x_j$. Its degree is $|B|$.

Gaussian comparison walk
The parameter process obtained by retaining the Gaussian initialization and increments while deleting the gradient drift.

Hinge loss
For a score f and signed label $y \in \{-1, 1\}$, the loss $[1 - yf]_+$.

Intermediate supervision
Training labels reveal intermediate computations whose final result is the original target behavior.

Computation graph
A directed graph whose nodes are intermediate values or operations and whose edges specify which computed values are supplied to later operations.

Parameter derivative features
The coordinates of $\Phi_\theta(x) = \nabla_\theta N_\theta(x)$. Their label correlations determine the correlation-objective gradient.

Running parity
The partial XOR $z_i = b_1 \oplus \cdots \oplus b_i$.

State tracking
Computing successive states $y_t = g_t(y_{t-1})$ from an initial state and a sequence of operations.

Simply transitive action
A group action for which exactly one group element sends any specified state to any other specified state.

Length curriculum
A training schedule that increases the sequence or context length across stages. In Section 8.6, longer contexts train retrieval while the learned local update remains fixed.

Teacher-forced accuracy
The average probability of a correct next prediction when all supplied preceding answers are correct; see Eq. (8.45).

Streaming learner
A learner that processes examples in order while retaining a bounded state between examples. A q-pass learner traverses the same sequence q times.

Teacher forcing
Training each next-token prediction using the correct preceding tokens.

Teacher-forced error
An incorrect prediction when the correct prefix is supplied: $E_i = \{g_i(X, Z_{<i}) \neq Z_i\}$ in Proposition 8.3.2.

Transition kernel
A law for the next state at each history. Here $\mathsf{P}_t(\cdot \mid \vartheta_{0:t})$ and $\mathsf{R}_t(\cdot \mid \vartheta_{0:t})$ are Gaussian conditional laws.

Trajectory law
The joint law of the complete parameter sequence from initialization through the last update.

8.8 Exercises

General instructions. Justify every answer, including yes/no answers, even when the question does not explicitly ask for a justification.

Exercise profiles. Filled mountains show difficulty on a four-mountain scale: one is a direct application, two uses a familiar method, three combines ideas, and four is a substantial challenge. Filled cogs show technical load on a three-cog scale: one is light, two is sustained, and three is dense. Filled hourglasses show estimated time: one is under 20 minutes, two is 20–45 minutes, three is 45–90 minutes, and four is more than 90 minutes. Pale pictograms show the unused part of each scale. The work row uses the canonical order proof, calculation, construction, computation, and interpretation; pale icons mark work types not used. ◎ marks the intended learning outcome. Time estimates cover the non-optional parts. Ratings are not prerequisites or grading weights. A superscript asterisk marks an exercise based on supplementary lecture material; it is not a difficulty rating and does not make the exercise optional when assigned.

Exercise 1: From small gradients to close trajectory laws.

◎ Combine Gaussian KL, Pinsker's inequality, and a coupling to compare two stochastic optimization trajectories.

Let $F : \mathbb{R}^p \to \mathbb{R}$ be differentiable, and fix $T \in \mathbb{N}$, $\eta, \sigma, \varepsilon > 0$. Let the $\mathbb{R}^p$-valued random vectors $\Theta_0, \Xi_0, \dots, \Xi_{T-1}$ be mutually independent, with

$$\mathcal{L}(\Theta_0) = \mathcal{N}(0, \sigma^2 I_p), \qquad \mathcal{L}(\Xi_t) = \mathcal{N}(0, \sigma^2 I_p).$$

Define

$$[g]_\varepsilon = \begin{cases} 0, & \|g\| \le \varepsilon, \\ g, & \|g\| > \varepsilon, \end{cases}$$

and, using the same initialization and increments, define three processes by

$$\begin{aligned} \Theta_{t+1} &= \Theta_t - \eta \nabla F(\Theta_t) - \Xi_t, \\ \overline{\Theta}_0 = \Theta_0, \qquad \overline{\Theta}_{t+1} &= \overline{\Theta}_t - \eta [\nabla F(\overline{\Theta}_t)]_\varepsilon - \Xi_t, \\ \Theta^{(0)}_0 = \Theta_0, \qquad \Theta^{(0)}_{t+1} &= \Theta^{(0)}_t - \Xi_t, \qquad 0 \le t < T. \end{aligned}$$

For any process $V_0, \dots, V_T$, write $V_{0:T} = (V_0, \dots, V_T)$. Let

$$\mathsf{P} = \mathcal{L}(\Theta_{0:T}), \qquad \mathsf{R} = \mathcal{L}(\overline{\Theta}_{0:T}), \qquad \mathsf{G} = \mathcal{L}(\Theta^{(0)}_{0:T}).$$

Suppose

$$\mathbb{P}\left(\max_{0 \le t \le T} \|\nabla F(\Theta^{(0)}_t)\| > \varepsilon \right) \le \delta.$$

(a) For a fixed common history $\vartheta_{0:t}$, compute the KL divergence between the next-step Gaussian kernels of P and R. Use the KL chain rule to prove

$$D(\mathsf{P}, \mathsf{R}) \le \frac{T \eta^2 \varepsilon^2}{2\sigma^2}.$$

(b) Apply Pinsker's inequality to prove

$$\mathrm{TV}(\mathsf{P}, \mathsf{R}) \le \frac{\eta \varepsilon \sqrt{T}}{2\sigma}.$$

(c) Couple the auxiliary process and the Gaussian walk by using the same initial condition and Gaussian increments. Prove $\mathrm{TV}(\mathsf{R}, \mathsf{G}) \le \delta$.

(d) Combine the two bounds to obtain a bound on $\mathrm{TV}(\mathsf{P}, \mathsf{G})$. Then, for positive d, A, k, substitute $\varepsilon = e^{-k/18}$, $\delta = (T+1)e^{-k/18}$, and

$$\frac{\eta\sqrt{T}}{2\sigma} + T + 1 \le 3d^{5A/2}, \qquad 3d^{5A/2} \le e^{k/36},$$

to prove

$$\mathrm{TV}(\mathsf{P}, \mathsf{G}) \le e^{-k/36}.$$

Hints For (a), Gaussians with common covariance $\sigma^2 I$ and means μ, ν have KL divergence $\|\mu - \nu\|^2/(2\sigma^2)$. For (c), use the defining rule for the clipped gradient and argue by induction on time.

Exercise 2: Why a global parameter sign flip is the wrong symmetry.

Distinguish a tempting parameter transformation from the input-coordinate symmetry used in the parity lower bound.

Consider

$$N_\theta(x) = \sum_{r=1}^{m} u_r [w_r^\top x + b_r]_+, \qquad \theta = (u_r, w_r, b_r)_{r=1}^m,$$

and let $h_\theta(x) = \mathrm{sign}(N_\theta(x))$, using either fixed binary value for $\mathrm{sign}(0)$. Let X be uniform on $\{-1, 1\}^d$ and $Y = \chi_S(X)$ for a nonempty $S \subseteq [d]$.

(a) Let $G\theta = (-u_r, -w_r, -b_r)_{r=1}^m$. Write $N_{G\theta}(x)$ in terms of the original parameters. Give a width-one example for which $N_{G\theta}(x) \ne -N_\theta(x)$.

(b) Fix $j \in S$. Let D_j flip the jth coordinate of a vector and set

$$J_j\theta = (u_r, D_j w_r, b_r)_{r=1}^m.$$

Prove

$$N_{J_j\theta}(x) = N_\theta(D_j x) \quad \text{and} \quad R_{01}(h_{J_j\theta}) = 1 - R_{01}(h_\theta).$$

(c) If a random parameter vector Θ has an isotropic Gaussian law, explain why $\mathcal{L}(J_j\Theta) = \mathcal{L}(\Theta)$. What symmetry does this imply for the law of $R_{01}(h_\Theta)$?

Hints For (a), recall that ReLU is not an odd function. For (b), change variables $z = D_j x$ in the uniform average over the Boolean cube.

Exercise 3: Activation properties behind the parity lower bound.[*]

Identify the kernel expansion and moment bounds that make the parity obstruction persist across activation functions.

Let $d > 30$, fix $S \subseteq [d]$ with $k = |S| \ge 2$, and let X be uniform on $\{-1, 1\}^d$. Replace the ReLU activation in Section 8.1 by a continuously differentiable function $\phi : \mathbb{R} \to \mathbb{R}$:

$$N_\theta(x) = \sum_{r=1}^{m} u_r \phi(w_r^\top x + b_r), \qquad F(\theta) = -\mathbb{E}[\chi_S(X) N_\theta(X)].$$

The parameter vector is $\theta = (u_r, w_r, b_r)_{r=1}^m \in \mathbb{R}^p$, where $p = m(d+2)$. For a function $\psi : \mathbb{R} \to \mathbb{R}$, define

$$q_B^\psi(w,b) = \mathbb{E}[\chi_B(X)\psi(w^\top X + b)], \qquad B \subseteq [d], \quad (w,b) \in \mathbb{R}^d \times \mathbb{R}.$$

Let G be standard normal and, for $\tau > 0$, write

$$s_\tau = \tau\sqrt{d+1}, \qquad M_\psi(\tau) = \mathbb{E}[\psi(s_\tau G)^2].$$

You may use the following Gaussian kernel identity without proof. Whenever $M_\psi(\tau) < \infty$, there are coefficients $a_\ell \geq 0$ such that, for a centered jointly Gaussian pair (G_1, G_2) with unit variances and correlation $\rho \in [-1,1]$,

$$\mathbb{E}[\psi(s_\tau G_1)\psi(s_\tau G_2)] = \sum_{\ell \geq 0} a_\ell \rho^\ell, \qquad \sum_{\ell \geq 0} a_\ell = M_\psi(\tau).$$

You may also use the moment estimate from the proof of Lemma 8.2.2: if Z is uniform on $\{-1,1\}^d$ and $V = (1 + \sum_i Z_i)/(d+1)$, then

$$\mathbb{E}[|V|^j] \leq 4e^{-j/4}, \qquad 1 \leq j \leq d.$$

(a) Let (W, C) have law $\mathcal{N}(0, \tau^2 I_{d+1})$, independently of X. For every nonempty $B \subseteq [d]$ and every ψ with $M_\psi(\tau) < \infty$, prove

$$\mathbb{E}[q_B^\psi(W,C)^2] \leq 4M_\psi(\tau)e^{-|B|/4}.$$

Identify the roles of parity degree, nonnegativity of the kernel coefficients, and their finite sum in your argument.

(b) Let $\mathcal{L}(\Theta) = \mathcal{N}(0, \tau^2 I_p)$ and assume $M_\phi(\tau), M_{\phi'}(\tau) < \infty$. Derive the partial derivatives of F and prove

$$\mathbb{E}[\|\nabla F(\Theta)\|^2] \leq 4me^{-(k-1)/4}\big(M_\phi(\tau) + (d+1)\tau^2 M_{\phi'}(\tau)\big).$$

(c) Use the initialization and independent Gaussian perturbations from Section 8.1 to define the processes in Eqs. (8.3) and (8.4) for this objective. Write

$$\mathsf{P} = \mathcal{L}(\Theta_0, \dots, \Theta_T), \qquad \mathsf{G} = \mathcal{L}(\Theta_0^{(0)}, \dots, \Theta_T^{(0)}), \qquad \tau_t = \sigma\sqrt{t+1}.$$

Assuming the moments in part (b) are finite at every τ_t, prove

$$\mathrm{TV}(\mathsf{P}, \mathsf{G}) \leq \frac{\eta\sqrt{m}}{\sigma} e^{-(k-1)/8} \left(\sum_{t=0}^{T-1} \big(M_\phi(\tau_t) + (d+1)\tau_t^2 M_{\phi'}(\tau_t)\big) \right)^{1/2}.$$

Now suppose, for fixed $C_0 \geq 1$ and $r \geq 0$, that

$$|\phi(z)| + |\phi'(z)| \leq C_0(1+|z|)^r \qquad \text{for every } z \in \mathbb{R}.$$

Show that for every $A > 0$ there is a constant $B_A > 0$, depending only on A, C_0, r, such that

$$m, T, \eta, \sigma, \sigma^{-1} \leq d^A, \qquad k \geq B_A \log d \quad \Longrightarrow \quad \mathrm{TV}(\mathsf{P}, \mathsf{G}) \leq e^{-k/16}.$$

Deduce both classification-risk conclusions of Theorem 8.1.2, with $e^{-k/16}$ replacing $e^{-k/36}$.

HINTS For (a), introduce an independent copy X' of X and compare the Gaussian margins at X and X'. For (b), calculate the derivatives before bounding their second moments. For (c), consider $D(\mathsf{G},\mathsf{P})$: which process supplies the history in the KL chain rule?

Exercise 4: Parity orthogonality and polynomial degree.

Explain why a polynomial of low degree has zero correlation with a higher-degree parity under uniform inputs.

Let X be uniform on $\{-1,1\}^d$ and put $\chi_\varnothing(x) = 1$.

(a) Prove that, for all $B, C \subseteq [d]$,

$$\mathbb{E}[\chi_B(X)\chi_C(X)] = \mathbb{I}\{B = C\}.$$

Deduce that the 2^d functions χ_B form an orthonormal basis for the vector space of real functions on $\{-1,1\}^d$, with inner product $\langle f, g\rangle = \mathbb{E}[f(X)g(X)]$.

(b) Let $B \neq \varnothing$, and let $p : \mathbb{R}^d \to \mathbb{R}$ be a polynomial of total degree less than $|B|$. Prove

$$\mathbb{E}[\chi_B(X)p(X)] = 0.$$

Your argument must allow repeated powers of a coordinate in a monomial.

HINTS For (a), simplify $\chi_B(x)\chi_C(x)$. For (b), first consider one monomial; use $X_j^2 = 1$.

Exercise 5: Which correlations determine the gradient?

Identify the features that govern a local gradient step and quantify the sampling error in estimating their correlations.

Use the ReLU network, fixed parity $Y = \chi_S(X)$, and objective F from Section 8.1. Fix θ such that $w_r^\top x + b_r \neq 0$ for all $r \in [m]$ and $x \in \{-1,1\}^d$. Define the vector of *parameter derivative features*

$$\Phi_\theta(x) = \nabla_\theta N_\theta(x) \in \mathbb{R}^p.$$

(a) Prove

$$\max_{\|v\|\le 1} \lim_{t\downarrow 0} \frac{F(\theta) - F(\theta + tv)}{t} = \|\mathbb{E}[Y\Phi_\theta(X)]\|.$$

Give a maximizing v, including the case when the expectation is zero. Explain the meaning of an upper bound $\|\mathbb{E}[Y\Phi_\theta(X)]\| \le \varepsilon$ for first-order improvement in a unit parameter direction.

(b) Write the coordinates of $\Phi_\theta(x)$ corresponding to u_r, w_{rj}, and b_r. Express their correlations with Y using $q_B(w,b) = \mathbb{E}[\chi_B(X)\mathbb{I}\{w^\top X + b > 0\}]$:

$$\begin{aligned}
\mathbb{E}[Y\Phi_{\theta,u_r}(X)] &= \sum_{j=1}^{d} w_{rj} q_{S\triangle\{j\}}(w_r, b_r) + b_r q_S(w_r, b_r),\\
\mathbb{E}[Y\Phi_{\theta,w_{rj}}(X)] &= u_r q_{S\triangle\{j\}}(w_r, b_r),\\
\mathbb{E}[Y\Phi_{\theta,b_r}(X)] &= u_r q_S(w_r, b_r).
\end{aligned}$$

Identify which coordinates are the hidden-unit outputs and which involve the activation indicators.

(c) Fix $w \in \mathbb{R}^d$ such that $w^\top x \neq 0$ for every cube point x, and fix $j \in [d]$. Let $X^{(1)}, \dots, X^{(n)}$ be independent uniform cube points and set

$$q = q_{S\triangle\{j\}}(w, 0), \qquad \widehat{q}_n = \frac{1}{n}\sum_{i=1}^{n} \chi_S(X^{(i)}) X_j^{(i)} \mathbb{I}\{w^\top X^{(i)} > 0\}.$$

Prove

$$\mathbb{E}[(\widehat{q}_n - q)^2] = \frac{1/2 - q^2}{n}.$$

If $0 < |q| \le e^{-ck} \le 1/2$, where $c > 0$, prove that the requirement $\sqrt{\mathbb{E}[(\widehat{q}_n - q)^2]} \le |q|$ implies $n \ge e^{2ck}/4$.

HINTS For (a), differentiate along the line $\theta + tv$. For (b), use $[z]_+ = z\mathbb{I}\{z > 0\}$ and $\chi_S(x)x_j = \chi_{S\triangle\{j\}}(x)$. For (c), pair x with $-x$ when calculating the second moment of one summand.

Exercise 6: A squared-loss improvement that gradient training misses.[*]

Extend the Gaussian trajectory comparison to an objective with a linear label-independent gradient term.

Let X be uniform on $\{-1, 1\}^d$, let $Y = \chi_S(X)$, and assume $k = |S| \ge 4$ is even. Set $s = (-1)^{(k-2)/2}$ and define

$$f_w(x) = s[w^\top x]_+, \qquad c(w) = \mathbb{E}[Y f_w(X)], \qquad L(w) = \mathbb{E}[(f_w(X) - Y)^2], \qquad w \in \mathbb{R}^d.$$

Use the ReLU backpropagation rule from Section 8.1. You may use these two estimates without proof:

$$s\,\mathbb{E}[Y[\mathbf{1}_S^\top X]_+] \ge \frac{1}{4\sqrt{k}},$$

and, for every $\tau > 0$ and $B \subseteq [d]$ with $|B| \ge 2$,

$$\mathcal{L}(W) = \mathcal{N}(0, \tau^2 I_d) \quad \Longrightarrow \quad \mathbb{E}[q_B(W, 0)^2] \le 8e^{-|B|/4}.$$

Here $\mathbf{1}_S = (\mathbb{I}\{j \in S\})_{j=1}^d$, and q_B is the threshold coefficient defined in the lecture. The first estimate is a Fourier coefficient of the unweighted sum; the second is the version of Lemma 8.2.2 with zero bias.

(a) Prove

$$L(w) = 1 + \tfrac{1}{2}\|w\|^2 - 2c(w), \qquad L\left(\frac{\mathbf{1}_S}{2k^{3/2}}\right) \le 1 - \frac{1}{8k^2}.$$

Thus the model admits an inverse-polynomial improvement over $L(0) = 1$.

(b) Derive

$$\nabla L(w) = w - 2s\big(q_{S\triangle\{j\}}(w, 0)\big)_{j=1}^d.$$

Fix $0 < \eta < 2$, $\sigma > 0$, and $T \in \mathbb{N}$. Let $W_0, \Xi_0, \dots, \Xi_{T-1}$ be mutually independent with law $\mathcal{N}(0, \sigma^2 I_d)$ and independent of X. Define

$$W_{t+1} = W_t - \eta \nabla L(W_t) - \Xi_t, \qquad V_0 = W_0, \qquad V_{t+1} = (1 - \eta)V_t - \Xi_t.$$

Prove that

$$\mathcal{L}(V_t) = \mathcal{N}(0, \tau_t^2 I_d), \qquad \tau_t^2 = \sigma^2 \sum_{j=0}^{t} (1 - \eta)^{2j} \le \sigma^2 (t + 1).$$

(c) With $\mathsf{P} = \mathcal{L}(W_0, \dots, W_T)$ and $\mathsf{G} = \mathcal{L}(V_0, \dots, V_T)$, prove

$$\mathrm{TV}(\mathsf{P}, \mathsf{G}) \le \frac{\eta\sqrt{8dT}}{\sigma} e^{-(k-1)/8}.$$

Also prove

$$\mathbb{E}[|c(V_T)|] \le \sqrt{8}\,\sigma\sqrt{T+1}\,d\,e^{-(k-1)/8}.$$

(d) Prove that, for every $A > 0$, there is a constant $B_A = O(A+1)$ such that, whenever

$$d > 30, \qquad T, \sigma, \sigma^{-1} \le d^A, \qquad k \ge B_A \log d,$$

the trained neuron satisfies

$$\mathbb{P}\big(L(W_T) \ge 1 - e^{-k/32}\big) \ge 1 - e^{-k/32}.$$

Compare the improvement allowed by this bound with the improvement achieved in (a), as $k \to \infty$.

Hints For (a), pair X with $-X$ to evaluate the second moment of a ReLU. For (c), apply the KL chain rule to $D(\mathsf{G}, \mathsf{P})$. To bound the final correlation, expand $[w^\top X]_+$ and use Cauchy–Schwarz. For (d), first bound the probability that $L(V_T) < 1 - e^{-k/32}$, and then transfer the event between the two laws.

Exercise 7: Learn a two-bit rule and compose it.

Prove convergence for a locally supervised gate and transfer its correctness to a multistep computation.

For $a, b \in \{0, 1\}$, fix

$$\phi(a, b) = \begin{pmatrix} [a+b]_+ \\ [a+b-1]_+ \end{pmatrix}, \qquad g_u(a, b) = u^\top \phi(a, b), \quad u \in \mathbb{R}^2.$$

Only the two output weights u are trained. With equal weight on the four input pairs, let

$$L(u) = \frac{1}{4} \sum_{a,b \in \{0,1\}} (g_u(a, b) - (a \oplus b))^2.$$

(a) Prove that the unique minimizer is $u^\star = (1, -2)^\top$ and that $L(u^\star) = 0$. Derive

$$\nabla^2 L(u) = H = \begin{pmatrix} 3 & 1 \\ 1 & 1/2 \end{pmatrix}, \qquad \lambda_{\min}(H) = \frac{7 - \sqrt{41}}{4}, \quad \lambda_{\max}(H) = \frac{7 + \sqrt{41}}{4}.$$

(b) For any initial $u_0 \in \mathbb{R}^2$, run $u_{t+1} = u_t - \eta \nabla L(u_t)$ with $\eta = 1/\lambda_{\max}(H)$. Prove

$$\|u_t - u^\star\| \le r^t \|u_0 - u^\star\|, \qquad r = 1 - \frac{\lambda_{\min}(H)}{\lambda_{\max}(H)} \in (0, 1).$$

Let $\widehat{g}_t(a, b) = \mathbb{I}\{g_{u_t}(a, b) \ge 1/2\}$. Show that $\widehat{g}_t(a, b) = a \oplus b$ on all four inputs whenever $\sqrt{5}\, r^t \|u_0 - u^\star\| < 1/2$.

(c) For $b_{1:d} \in \{0, 1\}^d$, define $\widehat{z}_1 = b_1$ and $\widehat{z}_i = \widehat{g}_t(\widehat{z}_{i-1}, b_i)$ for $2 \le i \le d$. Prove that the condition in (b) gives $\widehat{z}_d = b_1 \oplus \cdots \oplus b_d$ for every input. State separately which components are fixed by the construction and which are learned. Your answer must account for the two ReLU features ϕ, the threshold at $1/2$, the recurrent connections that feed $(\widehat{z}_{i-1}, b_i)$ to the next gate, and the output-weight vector u.

HINTS For (a), evaluate ϕ on the four input pairs. For (b), write the recurrence for $u_t - u^\star$ and use an orthonormal eigenbasis of the symmetric matrix H. For (c), use induction over the output positions.

Exercise 8: Random examples, memory, and teacher-chosen examples.[*]

Identify the roles of working memory, target uncertainty, and control of examples in a parity-learning lower bound.

An unknown set $S \subseteq [d]$ has size $k \le d/2$. A labeled example is $(U, V) \in \mathbb{F}_2^d \times \mathbb{F}_2$, with

$$V = \sum_{j \in S} U_j \pmod 2.$$

A streaming learner processes the examples in order, keeping a state of at most M bits between examples. In a q-pass algorithm it traverses the same sequence q times in the same order.

You may use the following consequence of the constant-pass learning theorem without proof. Fix $q \in \mathbb{N}$. Choose S uniformly among the k-subsets, and independently choose the input examples uniformly and independently from $\mathbb{F}_2^d$. If $k \to \infty$ and $M = o(dk)$, then achieving $\mathbb{P}(\widehat{S} = S) \ge 2/3$ requires $2^{\Omega(k)}$ examples. The constants may depend on q.

(a) Suppose all state retained by a training algorithm can be stored in $O(D^2 + Dm)$ numbers of $O(\log d)$ bits each. Set

$$k = \left\lfloor \frac{d}{(\log d)^4} \right\rfloor, \qquad D = \left\lceil k(\log d)^{11/10} \right\rceil, \qquad m = 2k.$$

As $d \to \infty$, show that the memory bound is $o(dk)$ and that the theorem requires more than any polynomial number of random examples to recover S with probability at least $2/3$ in a fixed number of passes.

(b) A linear-system learner can store $n = d + 2$ examples in an augmented matrix over $\mathbb{F}_2$. State its storage in bits and explain why its polynomial sample guarantee is consistent with the stated lower bound. You may use the Lecture 7 guarantee $\mathbb{P}(\widehat{S} \ne S) < 2^{d-n}$ and an in-place linear-system solver. Next, construct a teacher-chosen set of d inputs that permits exact recovery of every S while retaining only $O(d)$ bits. The teacher chooses the inputs before seeing any labels.

(c) Compare the supplied memory theorem with the fixed-parity result in Theorem 8.1.2. State, for each, the quantifier on S and the restrictions on the learner. Explain why the memory theorem permits the teacher-chosen protocol in (b), and why it cannot establish failure for a learner told S in advance.

HINTS For (a), divide $(D^2 + Dm)\log d$ by dk. For (b), consider the standard basis of $\mathbb{F}_2^d$. For (c), distinguish information supplied to the learner from knowledge held by the person choosing the task.

8.A The Taylor expansion of arcsine

The threshold-correlation proof uses the following expansion on the entire interval $[-1, 1]$, including its endpoints.

Proposition 8.A.1 (Arcsine series)**.** *For every $v \in [-1,1]$,*

$$\arcsin v = \sum_{\ell=0}^{\infty} a_\ell v^\ell, \qquad a_{2n} = 0, \qquad a_{2n+1} = \frac{\binom{2n}{n}}{4^n(2n+1)}, \qquad n \geq 0.$$

The coefficients satisfy $a_\ell \geq 0$ and $\sum_{\ell \geq 0} a_\ell = \pi/2$. The series converges absolutely and uniformly on $[-1,1]$.

Proof. We obtain the expansion by integrating the Taylor series of $(\arcsin t)' = (1-t^2)^{-1/2}$, then justify convergence at the endpoints. Set $f(z) = (1-z)^{-1/2}$ for $z < 1$. Its Taylor coefficients at zero are

$$c_0 = 1, \qquad c_n = \frac{f^{(n)}(0)}{n!} = \frac{1 \cdot 3 \cdots (2n-1)}{2^n n!} = \frac{\binom{2n}{n}}{4^n} \leq 1 \quad (n \geq 1).$$

Repeated differentiation gives $f^{(n)}(z) = n! c_n (1-z)^{-n-1/2}$. To show that the Taylor series equals f throughout $0 \leq z < 1$, use Taylor's theorem with the integral remainder:

$$\begin{aligned} f(z) - \sum_{n=0}^{N} c_n z^n &= \frac{1}{N!} \int_0^z f^{(N+1)}(s)(z-s)^N \, ds \\ &= (N+1) c_{N+1} \int_0^z \left(\frac{z-s}{1-s}\right)^N \frac{ds}{(1-s)^{3/2}}. \end{aligned}$$

For $0 \leq s \leq z < 1$, we have $(z-s)/(1-s) \leq z$. Hence

$$0 \leq f(z) - \sum_{n=0}^{N} c_n z^n \leq \frac{(N+1) z^{N+1}}{(1-z)^{3/2}} \longrightarrow 0.$$

Substituting $z = t^2$ proves $(1-t^2)^{-1/2} = \sum_{n \geq 0} c_n t^{2n}$ for $|t| < 1$. For every $0 < r < 1$, the bound $|c_n t^{2n}| \leq r^{2n}$ on $[-r, r]$ gives uniform convergence there. We may therefore integrate term by term:

$$\arcsin v = \int_0^v \frac{dt}{\sqrt{1-t^2}} = \sum_{n=0}^{\infty} \frac{c_n}{2n+1} v^{2n+1}, \qquad |v| < 1.$$

This gives the stated coefficients. Their nonnegativity allows monotone convergence as $v \uparrow 1$:

$$\sum_{n=0}^{\infty} \frac{c_n}{2n+1} = \lim_{v \uparrow 1} \sum_{n=0}^{\infty} \frac{c_n}{2n+1} v^{2n+1} = \lim_{v \uparrow 1} \arcsin v = \frac{\pi}{2}.$$

Thus the expansion holds at $v = 1$, and oddness gives equality at $v = -1$. Finally, $|a_\ell v^\ell| \leq a_\ell$ on $[-1,1]$, so the finite coefficient sum gives absolute and uniform convergence on the whole interval. □

References

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. Advances in Neural Information Processing Systems, 2016. https://arxiv.org/abs/1602.05897.

Edward S. Hu, Kwangjun Ahn, Qinghua Liu, Haoran Xu, Manan Tomar, Ada Langford, Dinesh Jayaraman, Alex Lamb, and John Langford. Learning to achieve goals with belief state transformers. 2025. https://arxiv.org/abs/2410.23506.

Yu Huang, Zixin Wen, Aarti Singh, Yuejie Chi, and Yuxin Chen. Transformers provably learn chain-of-thought reasoning with length generalization. Advances in Neural Information Processing Systems, 2025. Full version: https://arxiv.org/abs/2511.07378.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. International Conference on Learning Representations, 2025. https://arxiv.org/abs/2410.08633.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. International Conference on Learning Representations, 2025. https://arxiv.org/abs/2410.05459.

Xin Lyu, Avishay Tal, Hongxun Wu, and Junzhao Yang. Tight time-space lower bounds for constant-pass learning. 2023. https://arxiv.org/abs/2310.08070.

Itamar Shoshani and Ohad Shamir. Hardness of learning fixed parities with neural networks. 2025. https://arxiv.org/abs/2501.00817.