NEAR-OPTIMAL STOCHASTIC AUTOREGRESSION FROM FULL TRACES

Pahan Dewasurendra
Johns Hopkins University

ABSTRACT

How many full trajectories are needed to learn an autoregressive generator whose parameters are shared across time? For unrestricted d-dimensional logistic generators, the best known chain-of-thought bound is roughly $d^2 \log M/\epsilon$ trajectories at horizon M. Whether the natural d-dimensional rate is possible was left open. We answer this question affirmatively, up to logarithmic factors. A proper information-theoretic learner achieves trajectory Hellinger risk and final-token squared error at most ϵ from

$$O\left(\frac{d\log(Md)\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$$

full trajectories, with no norm, margin, mixing, or prompt-distribution assumption. A matching $\Omega((d+\log(1/\delta))/\epsilon)$ lower bound holds already at one step.

The proof identifies an observed-path dimension that is much smaller than the complexity of the marginalized final-token map. After exponentiating the logistic weights, the likelihood of one revealed path is a rational function of d positive parameters with degree at most Md. A sign-pattern argument gives VC-subgraph dimension $O(d\log(Md))$. A conditional rho-estimator then converts this dimension into a proper Hellinger oracle inequality while canceling the unknown prompt marginal. The argument extends to misspecification, every bounded trajectory statistic, unrestricted multiclass softmax generators, and model selection over unknown memory orders. It also yields a general theorem for shared-parameter autoregressive models with rational local probabilities.

1 INTRODUCTION

Autoregressive models reuse one next-token rule along a generated sequence. This parameter sharing makes their statistical complexity hard to read from the number of token predictions in a trajectory. Treating the M positions as unrelated loses a factor of M. Marginalizing the unobserved intermediate tokens can instead create an exponentially complicated function of the shared parameters. Full trajectories occupy an intermediate regime: the learner sees every random token generated by the same rule, but those token-level observations are dependent within each trajectory.

Recent work formalized this issue through stochastic autoregressive PAC learning (Doron-Arad et al., 2026). A binary generator g assigns a Bernoulli next-token distribution to each string. Starting from a random prompt, the same generator is applied for M steps. Chain-of-thought supervision reveals the entire random trajectory, while end-to-end supervision reveals only its final token. For the finite-memory logistic class

$$p_w(1 \mid s) = \sigma(\langle w, \mathrm{tail}_d(s)\rangle), \qquad w \in \mathbb{R}^d,$$

Doron-Arad et al. (2026) give an efficient proper chain-of-thought learner and an information-theoretic end-to-end learner using roughly $d^2 \log M/\epsilon$ trajectories. Their discussion asks whether this can be reduced to the natural d-dimensional rate for arbitrary horizons.

We close the chain-of-thought side of this question. For every horizon and every unrestricted weight vector, a proper estimator learns the entire conditional trajectory law in squared Hellinger distance at the nearly parametric rate

$$\frac{d\log(CMd)\,[1+\log_+(n/(d\log(CMd)))]+\log(1/\delta)}{n}. \tag{1}$$

This immediately controls final-token squared error. In sample-complexity form, the bound is $\widetilde{O}((d + \log(1/\delta))/\epsilon)$. A one-step reduction to realizable classification gives the matching $\Omega((d + \log(1/\delta))/\epsilon)$ lower bound. The estimator is information-theoretic and proper. We do not claim that it is computationally efficient.

The central observation is algebraic. Set $u_j = e^{w_j}$. Once a path $z_{1:M}$ is revealed, its likelihood is

$$q_u(z \mid x) = \frac{u^{\sum_{t:z_t=1} a_t}}{\prod_{t=1}^{M}(1 + u^{a_t})}, \qquad a_t = \mathrm{tail}_d(xz_{<t}). \tag{2}$$

Both numerator and denominator have degree at most Md. A threshold test on this likelihood is therefore one degree-Md polynomial inequality in d real parameters. Warren's sign-pattern theorem, in the form of Goldberg & Jerrum (1995), gives VC-subgraph dimension $O(d \log(Md))$.

This calculation exposes a sharp distinction between observing and marginalizing a path. For a revealed path, only the M visited states occur in the denominator. The end-to-end final-token probability sums over all latent continuations. The known end-to-end argument uses a common denominator over all 2^d binary states, with degree $Md2^{d-1}$, which leads to the $d^2 \log M$ pseudo-dimension bound (Doron-Arad et al., 2026). Our result does not improve that end-to-end bound and does not establish a statistical separation between the two supervision models.

Low path-likelihood dimension alone is not a squared-loss uniform-convergence argument. We use rho-estimation (Baraud & Birgé, 2018). Applied at trajectory level, it gives fast Hellinger risk for VC-subgraph density classes. Its conditional-density construction is crucial here: the reference measure contains the unknown prompt distribution, but this marginal cancels from every pairwise likelihood ratio. Trajectory-level rho-estimation was introduced to autoregressive learning by Rohatgi et al. (2025) for finite policy classes, with standard covering extensions noted there. Our contribution is a norm-free complexity analysis for an unrestricted shared-parameter class, based on observed-path VC-subgraph dimension.

The same route gives more than the realizable final-token guarantee. For an arbitrary conditional trajectory law S_x, the estimator satisfies

$$\mathbb{E}_X h^2(S_X, Q_{\widehat{w},X}) \le C \inf_w \mathbb{E}_X h^2(S_X, Q_{w,X}) + \widetilde{O}\left(\frac{d \log(Md)}{n}\right).$$

Every bounded path statistic inherits a squared-error oracle inequality. Examples include final-token probabilities, rewards, verifier acceptance, and safety events. The algebraic argument also covers multiclass softmax with unrestricted weights and, more generally, local probabilities that are rational functions of shared parameters.

Contributions.

1. We prove an observed-path theorem: if local probabilities are degree-D rational functions of p shared parameters, length-M path densities have VC-subgraph dimension $O(p \log(MD))$.
2. We combine this theorem with conditional rho-estimation to obtain a proper, high-probability, misspecified trajectory-Hellinger oracle inequality under an arbitrary and unknown prompt distribution.
3. For unrestricted d-dimensional logistic autoregression, we obtain a $\widetilde{O}(d/\epsilon)$ full-trace upper bound and a matching $\Omega(d/\epsilon)$ lower bound. This resolves the chain-of-thought side of the open problem of Doron-Arad et al. (2026).
4. We derive norm-free multiclass softmax rates, guarantees for every bounded trajectory statistic, and adaptation to unknown memory.

2 Setting

Let $\mathcal{X}$ be an arbitrary prompt space, let $\mathcal{A}$ be a finite token alphabet, and fix a horizon $M \ge 1$. A generator g assigns a distribution $p_g(\cdot \mid s)$ on $\mathcal{A}$ to every autoregressive state s. From a prompt x, it

generates $Z = (Z_1, \ldots, Z_M)$ according to

$$Q_{g,x}(z) = \prod_{t=1}^{M} p_g(z_t \mid xz_{<t}).$$

The training data are n independent prompt-trajectory pairs (X_i, Z_i). The prompt marginal P_X is arbitrary and unknown. In the realizable setting, $X_i \sim P_X$ and $Z_i \mid X_i = x \sim Q_{g_\star,x}$ for an unknown $g_\star$. For the misspecified result, the true conditional law can be any S_x on $\mathcal{A}^M$.

We use squared Hellinger distance

$$h^2(P, Q) = \frac{1}{2}\int (\sqrt{dP} - \sqrt{dQ})^2 = 1 - \int \sqrt{dP\, dQ}.$$

The trajectory risk of $\widehat{g}$ is $\mathbb{E}_X h^2(S_X, Q_{\widehat{g},X})$. This is stronger than the final-token criterion of Doron-Arad et al. (2026). If $q_g^M(x) = Q_{g,x}(Z_M = 1)$ in the binary case, then Hellinger data processing and $\mathrm{TV}^2 \le 2h^2$ imply

$$\mathbb{E}_X(q_{g_\star}^M(X) - q_{\widehat{g}}^M(X))^2 \le 2\mathbb{E}_X h^2(Q_{g_\star,X}, Q_{\widehat{g},X}). \tag{3}$$

For a binary generator class $\mathcal{F}$, let $m_{\mathrm{CoT}}^{\mathcal{F},M}(\epsilon, \delta)$ be the least n for which some learner observing n full trajectories outputs $\widehat{q} : \mathcal{X} \to [0, 1]$ such that

$$\mathbb{E}_X(\widehat{q}(X) - q_{g_\star}^M(X))^2 \le \epsilon$$

with probability at least $1 - \delta$, uniformly over P_X and $g_\star \in \mathcal{F}$. A learner is proper when $\widehat{q} = q_{\widehat{g}}^M$ for an output generator $\widehat{g} \in \mathcal{F}$.

2.1 Shared-parameter polynomial path models

Our structural result applies beyond logistic regression.

Definition 1 (Polynomial path model)**.** A generator family has a (p, D) polynomial parameterization if its parameter domain is $\Theta \subseteq \mathbb{R}^p$ and, for every state s and token a,

$$p_\theta(a \mid s) = \frac{A_{a,s}(\theta)}{B_s(\theta)},$$

where $A_{a,s}$ and B_s are polynomials of degree at most D, $B_s(\theta) > 0$ on Θ, and the ratios are valid probabilities. Coefficients and monomial supports may depend arbitrarily on (a, s).

Let $q_\theta(z \mid x) = Q_{\theta,x}(z)$ and write $\mathcal{Q}_M = \{q_\theta : \theta \in \Theta\}$. We assume universal separability when invoking the statistical theorem: some countable subclass is pointwise dense in $\mathcal{Q}_M$. This mild measurability condition is automatic for the logistic and softmax classes below.

3 Main results

Our first theorem quantifies the complexity of one observed trajectory.

Theorem 2 (Observed-path dimension)**.** *For every (p, D) polynomial path model, the VC-subgraph dimension of $\mathcal{Q}_M$ is strictly below*

$$2p \log_2(8e \max\{1, MD\}).$$

In particular, its VC-subgraph index is at most

$$V_{p,D,M} := 1 + \lceil 2p \log_2(8e \max\{1, MD\}) \rceil.$$

The logarithmic horizon dependence follows from parameter sharing. The local degrees add along one path, while the number of parameters remains p. The theorem does not require the prompt or state to have a finite-dimensional representation.

Theorem 3 (Proper misspecified trajectory oracle)**.** *Consider a universally separable (p, D) polynomial path model. There is a proper information-theoretic estimator $\widehat{q} \in \mathcal{Q}_M$ that does not know P_X and satisfies the following guarantee. For every conditional trajectory law S, every $n \geq 1$, and every $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\mathbb{E}_X h^2(S_X, \widehat{Q}_X) \leq C \left[\inf_{\theta \in \Theta} \mathbb{E}_X h^2(S_X, Q_{\theta,X}) + \frac{(V_{p,D,M} \wedge n)[1 + \log_+(n/V_{p,D,M})] + \log(1/\delta)}{n} \right]. \tag{4}$$

Here and below, C is a universal constant that may change between displays.

The estimator is a rho-estimator over a countable dense model. Properness means that it returns a generator in the given family, rather than an arbitrary trajectory predictor. The approximation term makes the guarantee robust to misspecification. Computational efficiency is not asserted.

Corollary 4 (Bounded trajectory statistics)**.** *Let $\phi : \mathcal{X} \times \mathcal{A}^M \to [0, 1]$ be measurable, and define*

$$\mu_S(x) = \mathbb{E}_{Z \sim S_x} \phi(x, Z), \qquad \widehat{\mu}(x) = \mathbb{E}_{Z \sim \widehat{Q}_x} \phi(x, Z).$$

On the event in Theorem 3,

$$\mathbb{E}_X(\mu_S(X) - \widehat{\mu}(X))^2 \leq 2\mathbb{E}_X h^2(S_X, \widehat{Q}_X).$$

3.1 Unrestricted logistic autoregression

For $w \in \mathbb{R}^d$, define the binary finite-memory generator

$$p_w(1 \mid s) = \sigma(\langle w, \mathrm{tail}_d(s) \rangle), \qquad \sigma(v) = \frac{1}{1 + e^{-v}}, \tag{5}$$

where $\mathrm{tail}_d(s) \in \{0, 1\}^d$ contains the last d bits of s, padded by zeros on the left. This is exactly the class $\mathcal{F}_\sigma(d)$ of Doron-Arad et al. (2026).

Theorem 5 (Norm-free logistic full-trace rate)**.** *For every $d, M \geq 1$, there is a proper information-theoretic chain-of-thought learner for the unrestricted class $\mathcal{F}_\sigma(d)$ such that*

$$m_{\mathrm{CoT}}^{\mathcal{F}_\sigma(d), M}(\epsilon, \delta) \leq \frac{C}{\epsilon} \left[d \log(CMd) \log(C/\epsilon) + \log(1/\delta) \right]. \tag{6}$$

The output can be chosen as a finite rational vector $\widehat{w} \in \mathbb{Q}^d$. No norm, margin, mixing, or prompt-distribution assumption is required.

The theorem controls the full conditional path law before applying (3). It is therefore stronger than the displayed chain-of-thought learning criterion. Compared with the efficient proper bound of Doron-Arad et al. (2026), it removes one factor of d at the cost of computational efficiency.

Theorem 6 (Worst-case lower bound)**.** *There are constants $c, c_0 > 0$ such that, for $d \geq 2$, $0 < \epsilon < c_0$, and $0 < \delta < c_0$,*

$$m_{\mathrm{CoT}}^{\mathcal{F}_\sigma(d), 1}(\epsilon, \delta) \geq c \frac{d + \log(1/\delta)}{\epsilon}.$$

Thus Theorem 5 is minimax optimal up to logarithmic factors, already by one-step stopping.

3.2 Multiclass softmax and unknown memory

The same phenomenon is not binary-specific. Let $|\mathcal{A}| = K$, let $\varphi(s) \in \{0, 1\}^r$ be any fixed feature map, and use class K as reference. For $W = (w_1, \ldots, w_{K-1}) \in \mathbb{R}^{(K-1) \times r}$, define

$$p_W(c \mid s) = \frac{\exp(\langle w_c, \varphi(s) \rangle)}{1 + \sum_{b=1}^{K-1} \exp(\langle w_b, \varphi(s) \rangle)} \quad (c < K), \tag{7}$$

with numerator 1 for $c = K$.

Corollary 7 (Norm-free softmax trajectories)**.** *The unrestricted softmax class in* (7) *satisfies Theorem 3 with*

$$V \le 1 + \lceil 2(K-1)r \log_2(8eMr) \rceil .$$

In the realizable case, every bounded path functional is learned to squared error ϵ *from*

$$\frac{C}{\epsilon}\left[Kr \log(CMr) \log(C/\epsilon) + \log(1/\delta)\right]$$

full trajectories. The weights are unrestricted.

A penalized rho-estimator can also select the memory order. Appendix F proves an oracle inequality that pays the complexity of the best memory-d approximation plus only $O(\log(d+1)/n)$. This gives simultaneous adaptation to unknown memory and robustness to misspecification.

4 Why one observed path is simple

We prove Theorem 2. Fix a prompt-path pair (x, z). Under Definition 1,

$$q_\theta(z \mid x) = \frac{N_{x,z}(\theta)}{R_{x,z}(\theta)}, \quad N_{x,z} = \prod_{t=1}^{M} A_{z_t, xz_{<t}}, \quad R_{x,z} = \prod_{t=1}^{M} B_{xz_{<t}}.$$

Both $N_{x,z}$ and $R_{x,z}$ have degree at most MD, and $R_{x,z} > 0$ on Θ. For a fixed subgraph input (x, z, a),

$$q_\theta(z \mid x) > a \quad \Longleftrightarrow \quad N_{x,z}(\theta) - aR_{x,z}(\theta) > 0. \tag{8}$$

This is one polynomial predicate of degree at most MD in p real parameters. For N fixed subgraph inputs, Warren's theorem bounds the joint sign patterns of the resulting N polynomials. The two-case calculation in Theorem 2.2 of Goldberg & Jerrum (1995), with one predicate per input, yields

$$\mathrm{VCdim}(\mathrm{subgraph}(\mathcal{Q}_M)) < 2p \log_2(8e \max\{1, MD\}).$$

Restricting the parameter domain to Θ cannot add sign patterns. Appendix A gives the calculation.

For the logistic model, set $u_j = e^{w_j} > 0$ and $u^a = \prod_j u_j^{a_j}$. If $a_t = \mathrm{tail}_d(xz_{<t})$, then

$$\sigma(\langle w, a_t \rangle) = \frac{u^{a_t}}{1 + u^{a_t}}, \qquad 1 - \sigma(\langle w, a_t \rangle) = \frac{1}{1 + u^{a_t}}.$$

Multiplying the factors along the path gives (2). Its numerator has degree at most Md, and so does its denominator. Theorem 2 applies with $p = d$ and path degree Md. Rational weights are pointwise dense because $w \mapsto q_w(z \mid x)$ is continuous.

For softmax, put $u_{c,j} = e^{w_{c,j}}$. Each local numerator is a monomial of degree at most r, and the denominator is a sum of such monomials. There are $(K-1)r$ positive parameters. This proves Corollary 7 after applying the statistical theorem.

5 From path dimension to trajectory risk

We give the mechanism behind Theorem 3. Let ν be counting measure on $\mathcal{A}^M$ and define the unknown reference measure

$$\mu = P_X \otimes \nu.$$

The joint candidate law $R_\theta = P_X \otimes Q_\theta$ has μ-density $(x, z) \mapsto q_\theta(z \mid x)$. Although the learner does not know P_X, every candidate pair satisfies

$$\frac{dR_v}{dR_w}(x, z) = \frac{q_v(z \mid x)}{q_w(z \mid x)}. \tag{9}$$

Thus all pairwise rho-tests are evaluable from conditional path likelihoods. This is exactly the unknown-random-design conditional-density construction in Section 8 of Baraud & Birgé (2018). The strategy neither estimates nor accesses the prompt marginal.

Proposition 7 and Corollary 3 of Baraud & Birgé (2018) give a rho-estimator for a VC-subgraph density class of index V with high-probability squared Hellinger risk

$$C\left[\inf_{\theta} h^2(R_\star, R_\theta) + \frac{(V \wedge n)[1 + \log_+(n/V)] + \log(1/\delta)}{n}\right]. \tag{10}$$

The common prompt marginal implies

$$h^2(P_X \otimes S, P_X \otimes Q_\theta) = \mathbb{E}_X h^2(S_X, Q_{\theta,X}).$$

Combining this identity, (10), and Theorem 2 proves the oracle inequality.

To make the output proper, run the rho-estimator on a countable pointwise-dense subclass and choose an approximate minimizer from that subclass itself. For logistic generators, enumerate $\mathbb{Q}^d$. Every finite rational vector defines a member of the original generator class. Appendix B records the measurability and closure details.

For Corollary 4, fix x. The difference between the expectations of a $[0,1]$-valued function is at most total variation. Therefore

$$|\mu_S(x) - \widehat{\mu}(x)|^2 \le \mathrm{TV}(S_x, \widehat{Q}_x)^2 \le 2h^2(S_x, \widehat{Q}_x).$$

Averaging over P_X proves the claim. Choosing $\phi(x, z) = z_M$ gives the final-token guarantee.

6 Lower bound and interpretation

We sketch Theorem 6. Take $M = 1$ and prompts whose tail vectors are $e_1, \ldots, e_d$. For every $s \in \{0,1\}^d$, let

$$w_s = B(2s - 1), \qquad \gamma = \sigma(-B) \le \frac{1}{8}.$$

Then $p_{w_s}(1 \mid e_j) = \gamma + (1 - 2\gamma)s_j$. Given a realizable binary classification sample on the coordinate class, independently flip each label with probability γ and feed the result to a logistic probability learner. Threshold the learner's prediction at $1/2$. Every classification mistake at e_j incurs squared probability error at least $(1/2 - \gamma)^2 \ge 9/64$. A stochastic learner with error ϵ would therefore yield a realizable classifier with error $O(\epsilon)$ at the same confidence. The standard high-probability VC lower bound is $\Omega((d + \log(1/\delta))/\epsilon)$ (Shalev-Shwartz & Ben-David, 2014). Appendix D gives the reduction with all quantifiers.

The upper and lower bounds locate the statistical complexity of full-trace logistic learning up to logarithms. They do not identify the computational complexity. The efficient proper algorithm of Doron-Arad et al. (2026) uses $\widetilde{O}(d^2/\epsilon)$ trajectories. Our rho-estimator searches a countable dense model and is not known to be implementable in polynomial time. Closing this remaining factor efficiently is a separate question.

The result also does not settle the information-theoretic end-to-end rate. Full traces let the learner score one observed rational likelihood. End-to-end data require learning a probability obtained by summing over latent paths. The known algebraic representation for that marginalized probability has degree $Md2^{d-1}$ (Doron-Arad et al., 2026). Whether another representation or argument attains a nearly linear dependence on d remains open.

7 Related work

Autoregressive chain-of-thought theory. The deterministic PAC model of Joshi et al. (2025) initiated formal comparisons between end-to-end and chain-of-thought supervision. Subsequent work developed deterministic compression and dimension characterizations (Hanneke et al., 2026; Li, 2026) and distribution-dependent CoT information (Altabaa et al., 2025). These works concern deterministic intermediate labels and zero-one prediction. The stochastic model of Doron-Arad et al. (2026) studies random trajectories and squared probability loss. It proves general comparisons among base, chain-of-thought, and end-to-end supervision, then isolates unrestricted logistic autoregression as a natural case study. Our main theorem answers one of its explicit open questions. An older logistic autoregressive Bayesian-network model concerns Euclidean embeddings of discriminant concept classes with node-specific parameters, rather than learning a shared stochastic generator from trajectories (Nakamura et al., 2005).

Trajectory distribution learning. In imitation learning, log-loss behavior cloning can give horizon-free guarantees by controlling the full trajectory distribution (Foster et al., 2024). Its linear-softmax specialization assumes bounded parameters and features. Our result removes the parameter bound for finite-memory Boolean features. Rohatgi et al. (2025) give a trajectory-level rho-estimator for finite policy classes under misspecification, note standard covering extensions, and show computational-statistical tradeoffs. We supply the norm-free observed-path dimension needed for a fast unrestricted logistic rate. Recent work on joint KL loss (Xu et al., 2026) and Hellinger localization for parameter recovery (Shekhtman et al., 2025) studies different metrics or stronger parameter-identification goals.

Density estimation and real-parameter dimension. Rho-estimators provide robust Hellinger oracle inequalities for density and conditional-density models (Baraud & Birgé, 2018). Minimum-distance estimators based on likelihood-ratio VC dimension give a related route to Hellinger density estimation (Compton & Li, 2026). We use rho-estimation because conditional likelihood ratios cancel the unknown prompt marginal directly. The algebraic dimension step uses the sign-pattern method of Warren (1968) and Goldberg & Jerrum (1995). To our knowledge, prior work has not applied observed-path VC-subgraph dimension to unrestricted shared-parameter stochastic autoregression.

8 Conclusion

Full trajectories turn a potentially complicated marginal autoregressive map into a low-degree observed likelihood. For unrestricted logistic memory-d generators, this yields a proper nearly parametric rate and closes the statistical chain-of-thought side of the d versus d^2 question. The proof separates into a structural path-dimension theorem and a conditional Hellinger estimator, which makes the result robust to misspecification and portable to other rational generator families. The main remaining questions are computational: whether the nearly linear rate is efficiently attainable, and statistical: whether latent end-to-end trajectories admit an equally sharp analysis.

Reproducibility Statement

All assumptions and learner guarantees are stated in Sections 2–3. Complete proofs are included below. Appendix C gives the exact path-likelihood identities, Appendix B specializes the cited rho-estimation result to unknown prompt marginals, and Appendix D gives the lower-bound reduction. The algebraic identities were also checked by finite symbolic and numerical tests, but no empirical claim depends on those tests.

References

Awni Altabaa, Omar Montasser, and John D. Lafferty. CoT information: Improved sample complexity under chain-of-thought supervision. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yannick Baraud and Lucien Birgé. Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B):3767–3804, 2018. doi: 10.1214/17-AOS1675.

Spencer Compton and Jerry Li. Density estimation for hellinger via minimum-distance estimators: Mixtures of gaussians, log-concave, and more. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1436–1475, 2026.

Ilan Doron-Arad, Idan Mehalel, and Elchanan Mossel. Stochastic autoregressive learning. *arXiv preprint arXiv:2608.07224*, 2026.

Dylan J. Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120602–120666, 2024. doi: 10.52202/079017-3834.

Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik–Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2–3):131–148, 1995. doi: 10.1007/BF00993408.

Steve Hanneke, Idan Mehalel, and Shay Moran. Sample complexity of autoregressive reasoning: Chain-of-thought vs. end-to-end. *arXiv preprint arXiv:2604.12013*, 2026.

Nirmit Joshi, Gal Vardi, Adam Block, Surbhi Goel, Zhiyuan Li, Theodor Misiakiewicz, and Nathan Srebro. A theory of learning with autoregressive chain of thought. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3161–3212, 2025.

Zhiyuan Li. The optimal sample complexity of learning autoregressive chain-of-thought. *arXiv preprint arXiv:2607.07423*, 2026.

Atsuyoshi Nakamura, Michael Schmitt, Niels Schmitt, and Hans Ulrich Simon. Inner product spaces for Bayesian networks. *Journal of Machine Learning Research*, 6:1383–1403, 2005.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J. Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4831–4837, 2025. Extended abstract. Full version: arXiv:2502.12465.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Eliot Shekhtman, Yichen Zhou, Ingvar Ziemann, Nikolai Matni, and Stephen Tu. Nearly instance-optimal parameter recovery from many trajectories via hellinger localization. *arXiv preprint arXiv:2510.06434*, 2025.

Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

Yunbei Xu, Yuzhe Yuan, and Ruohan Zhan. Autoregressive learning in joint KL: Sharp oracle bounds and lower bounds. *arXiv preprint arXiv:2605.12316*, 2026.

A SIGN-PATTERN PROOF

We provide the numerical constant in Theorem 2. Let $L = \max\{1, MD\}$. Suppose N subgraph inputs are shattered. If $N \le p$, then

$$N \le p < 2p \log_2(8eL).$$

Now suppose $N > p$. By (8), the 2^N dichotomies would all occur among the sign patterns of N polynomials in p variables, each of degree at most L. Warren's sign-pattern bound in the all-signs form used by Goldberg & Jerrum (1995, Corollary 2.1) gives at most $(8eLN/p)^p$ patterns. Hence

$$2^N \le \left(\frac{8eLN}{p}\right)^p.$$

Put $r = N/p > 1$ and $a = 8eL$. Taking logarithms gives

$$r \le \log_2 a + \log_2 r.$$

If $a > r$, then $r < 2\log_2 a$. Otherwise $a \le r$, so $r \le 2\log_2 r$, which forces $r \le 4$. Since $a \ge 8e$, the latter case also gives $r < 2\log_2 a$. Hence

$$N < 2p \log_2(8eL),$$

as claimed.

The argument does not require one joint polynomial formula in the string input. VC dimension fixes the N inputs first. Each fixed input produces one polynomial in the shared concept parameters, and Warren's theorem applies to those N polynomials.

B Conditional rho-estimation details

B.1 Unknown prompt marginal

Let the true joint law be $R_\star(dx, dz) = P_X(dx)S_x(dz)$. Counting measure ν dominates every law on the finite set $\mathcal{A}^M$. The measure $\mu = P_X \otimes \nu$ dominates both the truth and every candidate. Their densities are

$$r_\star(x, z) = S_x(z), \qquad r_\theta(x, z) = q_\theta(z \mid x).$$

Section 8 of Baraud & Birgé (2018) treats precisely this conditional-density construction. The design marginal is allowed to be unknown because the rho-test between r_v and r_w depends on the ratio $r_v/r_w = q_v/q_w$. In particular, the statistic can be computed from the observed (X_i, Z_i) and conditional path likelihoods only.

For completeness, a rho-test has the schematic form

$$T_n(w, v) = \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q_v(Z_i \mid X_i)}{q_w(Z_i \mid X_i)}}\right),$$

with the bounded monotone transform ψ specified by Baraud & Birgé (2018). The estimator approximately minimizes the worst pairwise test. The general theory controls this criterion without requiring log-likelihoods to be bounded. This is why unrestricted logistic weights cause no tail or margin condition.

B.2 Risk specialization

Proposition 7 of Baraud & Birgé (2018) bounds the rho dimension of a VC-subgraph density model of index V by

$$C(V \wedge n)[1 + \log_+(n/V)].$$

Their Corollary 3 yields, with probability at least $1 - e^{-\xi}$,

$$h^2(R_\star, \widehat{R}) \le C\left[\inf_\theta h^2(R_\star, R_\theta) + \frac{(V \wedge n)[1 + \log_+(n/V)] + \xi}{n}\right].$$

Set $\xi = \log(1/\delta)$. Direct calculation using the common first marginal gives

$$\begin{aligned} h^2(R_\star, R_\theta) &= 1 - \int P_X(dx) \sum_{z \in \mathcal{A}^M} \sqrt{S_x(z) q_\theta(z \mid x)} \\ &= \mathbb{E}_X h^2(S_X, Q_{\theta,X}). \end{aligned}$$

The same identity holds for $\widehat{R}$. Substitution proves (4).

B.3 Properness and separability

Let $\mathcal{Q}_M^0$ be a countable pointwise-dense subclass. Run the rho construction on $\mathcal{Q}_M^0$. Enumerate its elements and select the first candidate whose criterion is within the prescribed strict tolerance of the countable infimum. This selection is measurable and belongs to $\mathcal{Q}_M^0$. The universal-separability extension in Baraud & Birgé (2018) identifies the approximation error of the countable model with that of its pointwise closure.

For logistic generators, take $\mathbb{Q}^d$. If $w_k \to w$, then $q_{w_k}(z \mid x) \to q_w(z \mid x)$ for every (x, z). Because these are probability mass functions on a finite path space, pointwise convergence also gives Hellinger convergence for each x. The selected rational vector is a valid unrestricted logistic generator, so the estimator is proper.

C Logistic path algebra and sample bound

Let $a_t = \mathrm{tail}_d(xz_{<t})$. For $u_j = e^{w_j}$ and $u^a = \prod_j u_j^{a_j}$,

$$p_w(1 \mid xz_{<t}) = \frac{u^{a_t}}{1 + u^{a_t}}, \qquad p_w(0 \mid xz_{<t}) = \frac{1}{1 + u^{a_t}}.$$

Therefore

$$q_w(z \mid x) = \prod_{t=1}^{M} \frac{(u^{a_t})^{z_t}}{1+u^{a_t}} = \frac{u^{\sum_{t:z_t=1} a_t}}{\prod_{t=1}^{M}(1+u^{a_t})}.$$

The numerator degree is

$$\left\| \sum_{t:z_t=1} a_t \right\|_1 \le Md.$$

The denominator is a product of M polynomials of degree at most d, so its degree is at most Md. Positivity on $u \in \mathbb{R}^d_{>0}$ permits denominator clearing. Applying Theorem 2 yields

$$V_{d,M} \le 1 + \lceil 2d \log_2(8e \max\{1, Md\}) \rceil .$$

In the realizable case, Theorem 3 and (3) give error at most

$$C \frac{V_{d,M}[1 + \log_+(n/V_{d,M})] + \log(1/\delta)}{n}.$$

To invert this bound, put $A = \log(C_0/\epsilon) \ge 1$, $L = \log(1/\delta)$, and

$$n_0 = \left\lceil \frac{C_0}{\epsilon}(V_{d,M} A + L) \right\rceil .$$

If more samples are available, the learner may retain only n_0. With $r = L/V_{d,M}$, the inequalities $\log(A + r) \le \log A + r/A \le A + r$ and $A \ge 1$ show that

$$1 + \log(n_0/V_{d,M}) \le C_1 A + r.$$

Hence the numerator in the risk bound is at most $C_2(V_{d,M} A + L)$. Taking C_0 large enough makes the risk at most ϵ. Substituting $V_{d,M} = O(d \log(CMd))$ proves Theorem 5.

D LOWER-BOUND DETAILS

Fix $B \ge \log 7$ and $\gamma = \sigma(-B) \le 1/8$. Let the prompt set contain strings $x_1, \dots, x_d$ with $\mathrm{tail}_d(x_j) = e_j$. For $s \in \{0,1\}^d$, define $w_s = B(2s - 1)$. Then

$$q^1_{w_s}(x_j) = \begin{cases} 1-\gamma, & s_j = 1, \\ \gamma, & s_j = 0. \end{cases}$$

Consider the realizable coordinate-label class $\{j \mapsto s_j : s \in \{0,1\}^d\}$. Given an i.i.d. realizable classification sample (J_i, s_{J_i}), independently replace each label by $Z_i = s_{J_i} \oplus N_i$ with $N_i \sim \mathrm{Ber}(\gamma)$. Map J_i to prompt x_{J_i} and give (x_{J_i}, Z_i) to the stochastic learner. Conditional on $J_i = j$, this is exactly a one-step sample from w_s.

Let the learner output a probability predictor $\widehat{q}$. Define $\widehat{s}(j) = \mathbf{1}\{\widehat{q}(x_j) \ge 1/2\}$. Whenever $\widehat{s}(j) \ne s_j$,

$$(\widehat{q}(x_j) - q^1_{w_s}(x_j))^2 \ge (1/2 - \gamma)^2 \ge 9/64.$$

Consequently,

$$\mathbb{P}_J\{\widehat{s}(J) \ne s_J\} \le \frac{64}{9} \mathbb{E}_J(\widehat{q}(x_J) - q^1_{w_s}(x_J))^2.$$

An ϵ-accurate stochastic learner therefore gives a $C\epsilon$-accurate realizable PAC learner with the same sample size and confidence. The coordinate-label class has VC dimension d. The distribution-free realizable lower bound, including the rare-point construction for confidence $1 - \delta$, is $\Omega((d + \log(1/\delta))/\epsilon)$ (Shalev-Shwartz & Ben-David, 2014, Chapter 6). This proves Theorem 6. Fixing one coordinate if desired supplies a common high-mass label while preserving dimension $d - 1$.

E MULTICLASS SOFTMAX DETAILS

For $c < K$, set $u_{c,j} = e^{w_{c,j}}$ and $u_c^a = \prod_j u_{c,j}^{a_j}$. For the reference class, set $u_K^a = 1$. Then

$$p_W(c \mid s) = \frac{u_c^{\varphi(s)}}{\sum_{b=1}^{K} u_b^{\varphi(s)}}.$$

Each numerator and the common denominator are polynomials of degree at most r in $p = (K-1)r$ positive real parameters. A length-M observed path has numerator and denominator degree at most Mr. Theorem 2 gives

$$V \le 1 + \lceil 2(K-1)r \log_2(8e \max\{1, Mr\}) \rceil .$$

Rational weight matrices are pointwise dense in the unrestricted real-weight class. Theorem 3 and Corollary 4 complete the proof of Corollary 7.

Boolean features are used so exponentiating coordinates produces monomials. The same argument permits fixed nonnegative integer features bounded in ℓ_1 by D, with r replaced by D in the degree term.

F ADAPTATION TO UNKNOWN MEMORY

Let $\mathcal{Q}_d$ be the binary logistic path model of memory d and put

$$V_{d,M} = 1 + \lceil 2d \log_2(8e \max\{1, Md\}) \rceil .$$

Choose $\Delta(d) = 2\log(d+1) + c_\Delta$, where c_Δ is large enough that $\sum_{d \ge 1} e^{-\Delta(d)} \le 1$. Apply the penalized rho model-selection theorem of Baraud & Birgé (2018) to the countable rational subclass in each $\mathcal{Q}_d$.

Theorem 8 (Adaptive memory and misspecification)**.** *There is a proper estimator $(\widehat{d}, \widehat{w})$ such that, with probability at least $1 - \delta$,*

$$\mathbb{E}_X h^2(S_X, Q_{\widehat{d},\widehat{w},X}) \le C \inf_{d \ge 1} \left\{ \inf_{w \in \mathbb{R}^d} \mathbb{E}_X h^2(S_X, Q_{d,w,X}) + \frac{(V_{d,M} \wedge n)[1 + \log_+(n/V_{d,M})] + \log(d+1) + \log(1/\delta)}{n} \right\}.$$

The Kraft inequality for Δ permits a simultaneous high-probability union over the countable family of models. The per-model rho-dimension bound is supplied by Proposition 7 of Baraud & Birgé (2018). Universal separability transfers the oracle from rational dense subclasses to the unrestricted logistic models. The selected pair itself is rational and finite, so it defines a proper generator.

G FURTHER SCOPE AND LIMITATIONS

Theorem 2 is an observed-data statement. It applies when the complete path used in the product likelihood is revealed. If some intermediate tokens are latent, summing over their possible values can increase algebraic degree or the number of polynomial predicates. The theorem cannot be applied to such a marginal likelihood without a separate complexity calculation.

The statistical oracle is also different from parameter recovery. When prompt distributions fail to excite some directions, no distribution-free method can recover the corresponding coordinates of w. Hellinger trajectory risk asks only for the induced conditional distributions on the prompt law. Results that localize parameters from many trajectories under Fisher-information or identifiability conditions address a complementary goal (Shekhtman et al., 2025).

Finally, the logarithmic dependence on M in our upper bound is not known to be necessary. The lower bound uses $M = 1$ and certifies the dependence on d, ϵ, and δ. A sharper estimator might remove $\log M$ for this particular logistic class. The present result proves that no polynomial dependence on the horizon, and no second factor of d, is statistically required under full-trace supervision.