MEMORISATION, CONVERGENCE AND GENERALISATION IN GENERATIVE MODELS FOR TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

How to define generalisation for a generative model? According to an influential recent proposal, an image diffusion model is said to generalise if two models trained independently on disjoint subsets of a dataset generate nearly identical images given the same latent noise. Extending this idea to autoregressive transformers is challenging: unlike diffusion models, transformers lack a shared latent variable that naturally couples generation, and discrete tokens lack an obvious notion of overlap. Here, we introduce the survival probability – the probability that two transformers trained on disjoint subsets of the training data generate the same sequence under maximal coupling – and use it to empirically demonstrate a transition from verbatim memorisation to generalisation in transformers as training set size increases. We further show that independently trained transformers develop a shared predictive subspace, defined via the leading principal components of next-token probabilities, when trained on data sets that are substantially smaller than those required for convergence. We then introduce a simple auto-regressive generative model, the spiked Markov chain, and develop an asymptotically sharp characterisation of convergence and subspace recovery for this model. Our framework provides new tools for assessing generalisation in auto-regressive generative models and motivate investigating how transformer inductive biases shape both the emergence of shared predictive structure and the agreement between generated sequences.

1 INTRODUCTION

How much training data does a generative neural network need to learn how to generate diverse, high-quality samples, rather than just memorise its training set? This question is a key challenge for the theory of neural networks, and is currently attracting a lot of interest, mostly in the context of diffusion models (Achilli et al., 2024; 2025; Favero et al., 2025; George et al., 2026; Bonnaire et al., 2025; Nicoletti et al., 2026; Marion & Wu, 2026; Maillard & Goldt, 2026; Latourelle-Vigeant et al., 2026; Garnier-Brun et al., 2026; Wang et al., 2026; Mendes et al., 2026; Merger & Goldt, 2026). Some even argue that it requires rethinking generalisation (again) (Marion & Wu, 2026), after the success of deep, over-parametrised neural networks in supervised learning challenged common wisdom in statistical learning theory (Neyshabur et al., 2014; Zhang et al., 2017; Neyshabur et al., 2017).

In a remarkable paper, Kadkhodaie, Guth, Simoncelli and Mallat (Kadkhodaie et al., 2024) (KGSM) showed that diffusion models trained independently on disjoint data subsets with n samples produced near-duplicates of training images at small n, but converged to nearly identical outputs from the same latent noise at large n. KGSM therefore defined generalisation as convergence of independently trained denoisers near-zero variance of the learnt denoisers with respect to variations of the training data. Maillard & Goldt (2026) showed theoretically that convergence does indeed capture learning the bulk of the data distribution, like the spectrum of the covariance, but highlighted that convergence itself can be insensitive to learning the leading latent variables that organise the data.

Given the success of auto-regressive generative models like transformers for language and multimodal modelling, an important question is whether an analogous transition from memorisation to generalisation can be identified for transformers trained on next-token prediction. Transformers are known to memorise (a lot of) facts (Petroni et al., 2019; Geva et al., 2021; Meng et al., 2022) and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the transition of memorisation to generalisation has been studied for the narrow case of in-context learning (Chan et al., 2022; Raventós et al., 2023; Nguyen & Nallamala, 2025; Lu et al., 2025) or "grokking" on carefully designed tasks like modular addition (Power et al., 2022). However, to the best of our knowledge, there is no study of the transition from memorisation to generalisation for what is arguably the fundamental task of a transformer: next-token prediction.

Extending the methodology of Kadkhodaie et al. (2024) to transformers is not straightforward. KGSM rely on a natural coupling (generating images from the same latent noise) and a simple notion of overlap between outputs (the cross-correlation between two images). Transformers instead generate discrete sequences auto-regressively rather than all-at-once, with no obvious latent variable that makes generation deterministic and no natural measure of overlap. These differences leave open how to define and measure convergence, and hence the transition from memorisation to generalisation, in transformers.

Here, we introduce a new method, based on the idea of maximal coupling between stochastic processes, to show that transformers do undergo a similar transition from verbatim memorisation to generalisation. We also introduce a simple structured Markov chain model, the spiked Markov chain, and show that our approach lends itself to a sharp asymptotic analysis in this model.

Our main results are as follows:

- **A convergence criterion for transformers** based on the expected number of matching generated tokens before the first disagreement under maximal coupling. Using this criterion, we **empirically demonstrate a transition from verbatim memorisation to generalisation** in transformers, see section 2;
- **Predictive subspace recovery.** We show that a shared predictive subspace emerge in transformers trained on disjoints subsets of the same data set at substantially smaller sample sizes than convergence, see section 3;
- **Sharp asymptotics in Markov chains.** We finally give asymptotically sharp characterisations of convergence and subspace recovery in the simplest auto-regressive generative model: a Markov chain, see section 4.

Further related work

Memorisation and generalisation in diffusion models The transition from memorisation to generalisation has attracted a lot of recent interest in the context of diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021). A number of recent studies have highlighted that the train and test losses of generative models are by themselves poor proxies for the quality of the generative models (Garnier-Brun et al., 2026; Mendes et al., 2026). Several studies analysed the transition from memorisation to generalisation in associative memories (Pham et al., 2025; Kalaj et al., 2025) and in energy-based models (Catania et al., 2025). In diffusion models, the transition was analysed *during training* by comparing the training and test loss both empirically (Favero et al., 2025) and in random feature models (Bonnaire et al., 2025; George et al., 2026; Latourelle-Vigeant et al., 2026), and from a geometric perspective (Achilli et al., 2024; 2025). Zhang et al. (2024) showed convergence of models trained independently on the *same* dataset. Inspired by Kadkhodaie et al. (2024), Wang et al. (2026) analysed the variance of linear denoisers trained over different realisations of a finite dataset.

2 Memorisation and generalisation in transformers

We now introduce our maximal-coupling framework for quantifying memorisation and convergence in autoregressive transformers through the survival probability that we define in eq. (4). Applying this framework across three text data sets, we show how increasing training set size shifts models from verbatimly reproducing training sequences towards greater agreement between models trained independently on disjoint data.

Method: sampling under maximal coupling To compare two autoregressive models, we must specify how to couple their sampling. Since transformers generate each token by sampling from a

next-token distribution conditioned on the preceding sequence, we treat the *randomness used in sampling tokens* as the shared source of randomness between models. A simple choice is *synchronous coupling*: at each generation step, we draw a single $u_t \sim \mathrm{Unif}([0,1))$ and sample from the two next-token distributions by evaluating their inverse CDFs using the same u_t. While intuitive, this procedure does not yield the strongest possible correlation between the two samples. We therefore sample under *maximal coupling* which maximises the probability that the two models emit the same token at the next step (Levin & Peres, 2017, sec 4.2.). If $p_t^{(1)}$ and $p_t^{(2)}$ denote their next-token distributions given a common history h_t, and $X_t^{(1)}, X_t^{(2)}$ are sampled from these distributions, maximal coupling achieves

$$\mathrm{pr}\left(X_t^{(1)} = X_t^{(2)} \mid h_t\right) = 1 - \mathrm{d_{TV}}\left(p_t^{(1)}, p_t^{(2)}\right). \tag{1}$$

We derive the sampling procedure in section A.3, see algorithm 1.

The survival probability The maximal coupling naturally leads to the *survival probability* $S(t)$, which we define as the probability that two transformers generate the same sequence under maximal coupling up to the t-th token. More precisely, for a given initial history h_0, the survival probability $S(t \mid h_0)$ that the two sequences agree up to time t is

$$S(t \mid h_0) \coloneqq \mathbb{E}_{\max}\left[\mathbb{1}\left\{X_s^{(1)} = X_s^{(2)} \;\; \forall\, s \in \{1,\ldots,t\}\right\} \,\middle|\, h_0^{(1)} = h_0^{(2)} = h_0\right] \tag{2}$$

$$= \mathbb{E}_{\max}\left[\prod_{s=1}^{t}\left(1 - d_{\mathrm{TV}}\left(p_s^{(1)}(\cdot \mid h_s), p_s^{(2)}(\cdot \mid h_s)\right)\right) \,\middle|\, h_0\right], \tag{3}$$

where $\mathbb{E}_{\max}$ denotes the expectation over sequences sampled under maximal coupling and we have used eq. (1) to derive a convenient form of the survival probability that we can evaluate directly.

To quantify **memorisation**, we measure $S_{\mathrm{mem}}(t)$ by maximally coupling a transformer to the empirical distribution of its training set, treated as an autoregressive model: given a prefix h, this distribution is the normalised histogram of next tokens among training sequences matching h. As the prefix grows, it concentrates on individual training sequences, so long survival indicates exact memorisation. We initialise from a randomly chosen training sequence and measure how long the transformer remains identical to its empirical training distribution.

To quantify **convergence**, we measure $S_{\mathrm{con}}(t)$ by maximally coupling two transformers trained independently on disjoint subsets of the same data set. We initialise both models with the same prompt (a short sequence from the held-out test set) and measure their survival length. In our experiments, we average over 2^{12} prompts of 16 tokens and follow the coupled continuations for up to 32 generated tokens. We verified that our results did change substantially when using 8 or 32 tokens as prompts.

Results We trained pairs of transformers to predict the next token on disjoint subsets of three data sets: OpenWebText, TinyStories, and WikiText, varying the number n of training sequences of 128 token each (see section B for details). We plot the survival probabilities $S_{\mathrm{mem}}(t)$ and $S_{\mathrm{con}}(t)$ for memorisation and convergence, respectively, as a function of training set size in fig. 1(a,b). For memorisation, the survival length for small data sets with $n \lesssim 2^{10}$ training sequences remains near zero, indicating a near-zero chance of the transformer deviating from the training set sequence: a clear sign of memorisation. For convergence, we find that $S_{\mathrm{con}}(t)$ always decays as the number of generated tokens increases – in other words, the two transformer never learn *exactly* the same distribution. However, the rate of decay of $S_{\mathrm{con}}(t)$ decreases with n, suggesting convergence of the models to the same distribution. We summarise this data using the **survival length**, i.e. the expected number of matching tokens before the first disagreement,

$$L \coloneqq \sum_{t=1}^{T} S(t), \tag{4}$$

where T is the maximum number of generated tokens considered. In fig. 1(c), we plot L for three different values of $T = 8, 16, 32$ for memorisation (blue). A value of $L = T$ indicates that the transformer exactly reproduces its training sequences; we see in fig. 1(c) that this is indeed the case for data sets with up to $n \approx 2^{10}$ training sequences with 2^{17} tokens. For what concerns

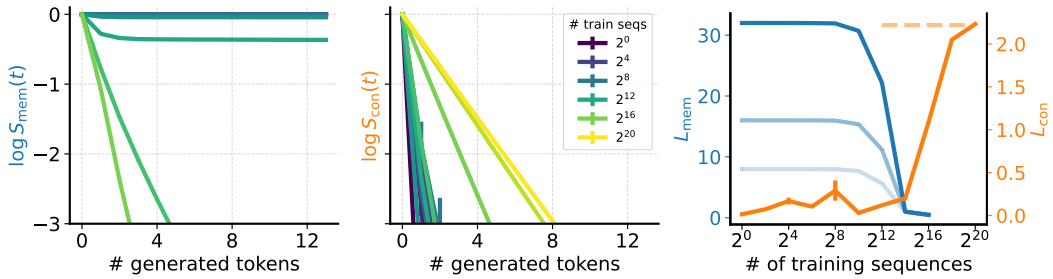


Figure 1: **Memorisation and convergence in transformers** We consider the survival probability $S(t)$ that two sequences sampled under maximal coupling remain identical through the first t generated tokens, eq. (2); the corresponding survival length is $L := \sum_{t=1}^{T} S(t)$. Transformers are trained on OpenWebText with n training sequences. **(a) Memorisation:** Survival probability between a transformer and its training-set distribution. **(b) Convergence:** Survival probability between two transformers trained independently on disjoint subsets of the data. **(c) Survival length** for memorisation and convergence. Larger values indicate stronger memorisation or convergence, and hence generalisation, respectively. The dashed orange line is the survival length of two transformers trained on the *same* data set. Error bars denote the error in the mean over three runs. Experimental details in section B.

convergence, the survival length of independently trained models remains close to zero for small n, before increasing sharply around $n \approx 2^{16}$ sequences to reach a final value of $L = 2.3$. We note that the transition from memorisation to convergence occurs precisely at the training set size at which the validation loss approaches the training loss, see fig. B.1.

Discussion To put the survival length into perspective, consider two models of a fair coin that predict probabilities (p, q) and (q, p) for heads and tails, with $p \geq q$. Under maximal coupling, the models agree with probability $2q$ at each toss, yielding $S(t) = (2q)^t$ and a mean survival length $L \simeq 2q/(1-2q)$, where the finite T introduces a negligible correction. A survival length of $L = 2.3$ as we find on OpenWebText corresponds to $q \simeq 0.348$, almost 70% agreement per toss, while a survival length of $L = 4$ that we find on tinystories, the simplest data set we considered, corresponds to $q = 0.4$ or about 80% agreement. Survival length of only a few tokens can therefore represent substantial agreement between independently trained models. Furthermore, we find that for the largest n, the survival length of two independent transformers matches the survival length of two transformers trained on the *same* data set of the same size, indicated by the dashed orange line. This suggests that resampling the training data adds little disagreement between models beyond the variation between training runs on the same data, as measured by survival length, akin to the convergence observed in diffusion models by Kadkhodaie et al. (2024).

3 Convergence versus predictive subspace recovery in auto-regressive models

Kadkhodaie et al. (2024) defined generalisation in terms of convergence: independently trained denoisers are said to generalise if they become (close to) independent of the realisation of their training data. The survival length introduced in the previous section demonstrates a similar behaviour in transformers, too. However, the theoretical analysis of Maillard & Goldt (2026) showed that, in linear generative models, convergence is driven by learning the bulk spectrum of the data covariance and is insensitive to consistent recovery of the principal latent variables that organise the data. Their observation raises a question for transformers: when do they consistently learn a shared latent structure that organises the data, and how does this relate to the convergence measured in section 2? To answer these questions, let us introduce a simple Markov chain model that will let us distinguish the two precisely.

The simplest language model: a spiked Markov chain We consider a particularly simple model of a language where there are two types of words (think nouns and verbs) that alternate to yield

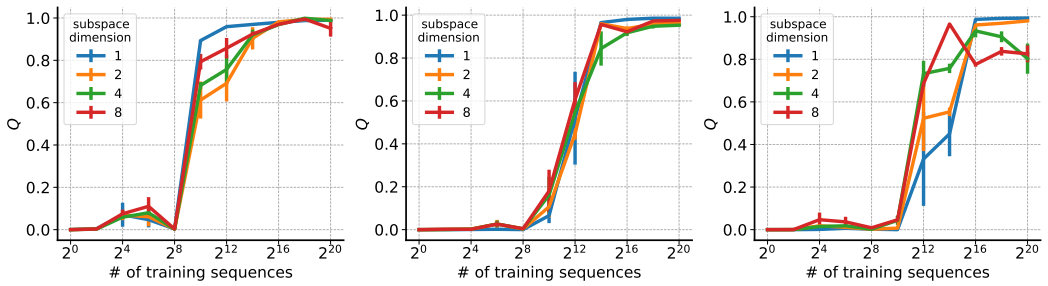


Figure 2: **Emergence of a shared predictive subspace in transformers.** Overlap $Q(k)$, eq. (7), between the leading k frequency-rescaled principal directions of next-token probability covariances, as a function of the number n of training sequences. We compare transformers trained independently on disjoint subsets of TinyStories, English Wikipedia, and OpenWebText (left to right), using held-out contexts to prompt both models. The overlap increases sharply between 2^8 and 2^{10} training sequences in all three data sets. Experimental details in section B.

simple sentences of the form *subject - verb - object*. We model text from this language as a sequence of tokens from a vocabulary of V tokens $\mathcal{V} = [V]$. We encode the group of each token in a vector $u = (u_i)_{i\in[V]}$, where $u_i = \pm 1$. We will assume that the tokens in our training set are sampled from a ground-truth Markov chain, the teacher, that has stationary state $\pi^\star \in \Delta^{V-1}$ and a transition matrix

$$P^\star_{ij} = \mathrm{pr}(w_{t+1} = j | w_t = i) = \pi^*_j \left(1 - \beta u_i u_j\right), \tag{5}$$

where w_t are the tokens in a sequence indexed by $t \geq 0$. We assume that the two groups have equal stationary probability mass, $\sum_i \pi^\star_i u_i = 0$, which ensures that $P^\star$ is a valid transition matrix with stationary distribution $\pi^\star$. At $\beta = 0$, successive tokens are independent and the group assignments are unidentifiable. For $\beta > 0$, transitions between groups are favoured, with strict alternation at $\beta = 1$.

Given a generative model trained on data set sampled from $P^\star$, we can ask whether the trained model has recovered the group structure u. If we fit a Markov chain with transition matrix $\hat{P}$ to a data set sampled from $P^\star$, we can estimate the principal predictive subspace using a spectral estimator. In particular, we can compute the covariance of the next-token probabilities as the current token w varies according to $\pi^\star$. The stationary mean is $\mathbb{E}_{w\sim\pi^\star} P^\star_{wi} = \pi^\star_i$, and we have under population expectation that

$$C^\star_{ij} = \mathbb{E}_{w\sim\pi^\star}\left[(P^\star_{wi} - \pi^\star_i)(P^\star_{wj} - \pi^\star_j)\right] = \beta^2 \pi^\star_i \pi^\star_j u_i u_j, \tag{6}$$

see section C for details. So computing the leading eigenvector of the covariance under the trained model $\hat{P}$ yields an estimate $\hat{u}$ of the group assignments (up to an overall sign).

Predictive subspace recovery in transformers The same spectral method can be applied to the covariance of next-token probabilities obtained from transformers, where we use sequences from the validation set as prompts (see section B for details). By evaluating the two transformers, indexed by $b \in \{1, 2\}$ and trained independently on disjoint data sets, on the same M held-out contexts h_a, we compute the leading eigenvectors $v^{(b)}_r$ of the empirical covaraince matrices $C^{(b)}$. Since the leading eigenvectors contain a contribution due to the unigram frequencies, see eq. (C.4), we divide the eigenvectors by the empirical unigram distribution $\hat{\pi}$. Denoting the resulting directions by $\hat{v}^{(b)}_r$, we define the overlap $Q(k)$ as the overlap of the subspaces spanned by the first k eigenvectors,

$$Q(k) := \frac{1}{k} \sum_{r,s=1}^{k} \left| \left\langle \hat{v}^{(1)}_r, \hat{v}^{(2)}_s \right\rangle \right|^2. \tag{7}$$

For transformers trained on natural language, $Q(k)$ measures reproducibility of predictive structure across training sets, without establishing a semantic interpretation of the recovered directions.

Results We show $Q(k)$ in fig. 2 for transformers trained on tinystories, wikipedia.en, and OpenWebText. Across all three data sets, the overlap increases sharply between 2^8 and 2^{10} training

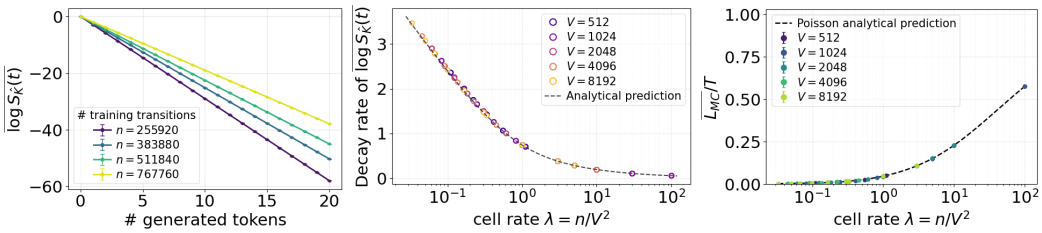


Figure 3: **Convergence in Markov chains. (a)** Survival probability for independently trained Markov chains, eq. (9), for different training set sizes. **(b)** Decay rate of the survival probability as a function of training set size. **(c)** Survival length, i.e. the mean number of tokens before sequences diverge under maximal coupling. In (b) and (c), symbols indicate simulations, while the dashed line reports the analytical prediction of theorem 4.1. Results are averaged over 500 independent pairs of training trajectories for each (V, n); error bars in (a) smaller than the plotting resolution.

sequences. This increase precedes the onset of convergence seen in the survival length: on OpenWebText, for example, the convergence survival length grows sharply only around $n \approx 2^{16}$, see fig. 1. Independently trained transformers therefore learn shared leading structure in their context-dependent predictions well before substantial agreement emerges between their maximally coupled continuations. These results extend the distinction between convergence and latent recovery of Maillard & Goldt (2026) to auto-regressive models, and reveal a separation between the data scales at which shared predictive structure and agreement of generated sequences emerge.

4 Convergence and subspace recovery in a Markov chain

We now turn to studying convergence and predictive subspace recovery in Markov chains as a tractable setting for understanding how autoregressive models learn from finite data. We derive an analytical prediction for survival under maximal coupling in a homogeneous chain and develop a heuristic theory of the transition to subspace recovery in a structured chain.

4.1 Convergence for a Markov Chain

Survival rate of Markov chains The simplest possible setting in which we can study the convergence of independently trained auto-regressive models is that of a first-order Markov chain over V states with transition matrix $P^{\star}_{ij} = \mathrm{pr}(y_{t+1} = j | y_t = i)$. To replicate our transformer experiment, we sample two data sets containing n transitions from a ground-truth Markov chain at steady state and use each data set to estimate an empirical transition matrix $\hat{P}_{ij}$ using the maximum-likelihood estimator (MLE) for observed transitions. This choice consists into defining

$$\hat{P}_{ij} := \#\{t \in [0, n-1] : \ (y_t, y_{t+1}) = (i, j)\} \ / \ \#\{t \in [0, n-1] : \ y_t = i\}, \tag{8}$$

where we set $\hat{P}_{ij} = 0 \ \forall j$ if the system never visits state i. In section D.1, we show that the survival probability for joint sampling under maximal coupling from two Markov chains with transition matrices $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ has the particularly simple form

$$S_{\hat{K}}(t) = \nu_0^{\top} \hat{K}^t \mathbf{1}, \tag{9}$$

where ν_0 is the distribution over the initial token, which could be the steady state distribution or an atomic distribution, $\mathbf{1}$ is the all-ones vector, while the matrix $\hat{K}_{ij} := \min\{\hat{P}^{(1)}_{ij}, \hat{P}^{(2)}_{ij}\}$ is a (substochastic) kernel. The rows of $\hat{K}$ satisfy $\sum_j \hat{K}_{ij} = 1 - d_{\mathrm{TV}}(\hat{P}^{(1)}_{i\cdot}, \hat{P}^{(2)}_{i\cdot})$; hence $\nu_0^{\top} \hat{K}^t \mathbf{1}$ is the total probability mass of all length-t trajectories on which the two maximally coupled chains have not yet diverged. Applying Perron-Frobenius theorem (Horn & Johnson, 2012, sec. 8.2), for a strictly positive initial distribution ν_0 and positive spectral radius $\rho(\hat{K})$ of the transition kernel, we obtain that the decay rate of the survival probability is given by the spectral radius of $\hat{K}$,

$$\lim_{t \to \infty} \frac{1}{t} \log S_{\hat{K}}(t) = \log \rho(\hat{K}). \tag{10}$$

Case study: homogeneous Markov chain We can obtain an analytical prediction for the decay of the survival probability for the simplest case of a homogeneous teacher $P^\star_{ij} = 1/V$ and uniform stationary distribution corresponding to the distribution over the initial token. The key idea of the analysis is to express the empirical transition matrix $\hat{P}^{(b)}$ through a count matrix $M^{(b)}_{ij}$ where transition counts are drawn i.i.d. from a suitable Poisson distribution with rate λ. We then define $(K_0)_{ij} = \min\{M^{(1)}_{ij}, M^{(2)}_{ij}\}/(\lambda V)$ where λV corresponds to the average number of counts per row. We then obtain the following theorem.

Theorem 4.1 (Survival rate for the homogeneous teacher)**.** *Consider the homogeneous teacher $P^\star_{ij} = 1/V$ and the Poisson surrogate K_0 defined above, with $\lambda := \frac{n}{V^2}$ and $\nu_{0i} = \pi_i = 1/V$. For $n \gg V^{3/2}\sqrt{\log V}$, we have*

$$\left[\lim_{t\to\infty} \frac{1}{t} \log S_{K_0}(t)\right] - \log h(\lambda) \xrightarrow[V\to\infty]{\mathbb{P}} 0, \tag{11}$$

where

$$h(\lambda) = 1 - e^{-2\lambda}\left[I_0(2\lambda) + I_1(2\lambda)\right], \tag{12}$$

and I_0, I_1 are modified Bessel functions of the first kind.

We relegate the proof to section D.5. Theorem 4.1 shows that the asymptotic survival rate depends on n only through $\lambda = n/V^2$; below the scale $n \asymp V^{3/2}$, the surrogate instead becomes too sparse and we show in section D.7 that $\rho(K_0) = 0$ with probability tending to one.

Experimental verification We verified theorem 4.1 numerically by estimating two Markov chains estimated from independent trajectories of a homogeneous teacher with $P^\star_{ij}$. Figure 3(a) shows the mean log survival probability, its fitted decay rate, and the corresponding survival length as functions of the rescaled training size $\lambda = n/V^2$. The predicted decay rate, showing with a dashed line in fig. 3(b), agrees closely with simulations across vocabulary sizes V: increasing λ increases the overlap between the learned transition matrices and slows the decay of survival. For $n \gg V^{3/2}\sqrt{\log V}$, the survival length takes the simple form

$$\frac{L_{\mathrm{MC}}}{T} = \frac{1}{T}\sum_{t=1}^{T} S_{K_0}(t) \approx \frac{h(\lambda)}{T}\frac{1-h(\lambda)^T}{1-h(\lambda)}, \tag{13}$$

if we approximate the behaviour of $S_K(t)$ at small t using the asymptotic behaviour theorem 4.1. We plot the resulting prediction with the dashed line in fig. 3(c) and again find close agreement with simulations. This confirms that the Poisson surrogate captures the behaviour of the empirical Markov chain. We note that the simple Markov chain model does not seem to show a sharp transition in convergence as we increase λ, similarly to the continuous increase of convergence in the linear model analysed by Maillard & Goldt (2026).

4.2 Subspace recovery for a Markov chain

We now return to the spiked Markov chain introduced in section 3, and ask how much training data is needed for the leading principal components of two empirical next-token probability matrices, estimated from independent training sequences, to align with each other – akin the transformers analysed in section 3. We assume a uniform stationary distribution and two balanced groups labelled by $u \in \{-1, 1\}^V$ with $\sum_i \pi_i u_i = 0$, so that eq. (5) gives $P^\star_{ij} = (1 - \beta u_i u_j)/V$, where $\beta \in [0, 1)$ controls the group structure strength. Each empirical transition matrix $\hat{P}$ is estimated from a trajectory containing n observed transitions, in the limit $n, V \to \infty$ with $n/V \to \alpha \in (0, \infty)$. After projecting out the uniform mode, independence of the trajectories implies that the leading right singular vectors of the transition matrices have asymptotically orthogonal residuals outside the signal direction $u/\sqrt{V}$, so the onset of agreement between independently trained estimators coincides with the onset of signal recovery by each Markov chain.

This onset of signal recovery is analogous to the classical Baik–Ben Arous–Péché (BBP) transition in spiked sample covariance matrices (Baik et al., 2005), first described by Edwards & Jones (1976). In their setting, a population spike produces an isolated empirical eigenvalue only above a critical signal-to-noise ratio; the corresponding eigenvector then acquires a nonzero asymptotic overlap

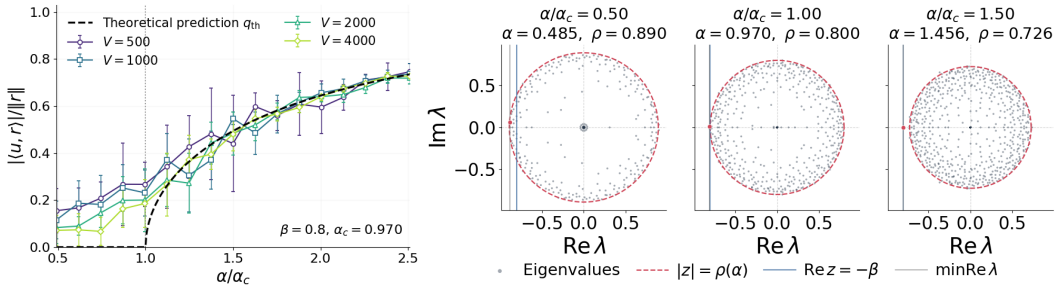


Figure 4: **Recovery of predictive subspace in Markov chains. (a)** Mean right-eigenvector overlap curve as presented in claim 4.1 as a function of α/α_c, where $\alpha = n/V$ and $\beta^2 = (1 - e^{-\alpha_c})/\alpha_c$ for different vocabulary sizes. **(b)** Spectrum of the Markov chain transition matrix for different numbers of transitions; The dashed circle has radius $\rho(\alpha)$We observe in (a) that the more the vocabulary size increases, the closer the experimental result is to the analytical dashed line. Error bars correspond to one standard deviation over independent trials. In (b) we can observe how increasing the length of the chain, the noise bulk shrinks since the estimator gets progressively closer to the ground truth, and the outlier emerges. Here the vocabulary size considered is 1000 and we vary n. The trivial leading eigenvalue is not presented.

with the population direction. Here, the signal is the group mode at eigenvalue $-\beta$, while finite sampling generates a bulk of noise eigenvalues. Since the empirical transition matrix is generally non-Hermitian, its noise spectrum occupies a region of the complex plane. Closely related phenomena involving outliers and eigenvector transitions for finite-rank deformations of non-Hermitian random matrices were investigated by Tao (2013); Bordenave & Capitaine (2016); Benaych-Georges & Rochet (2016). The sparse setting with $O(1)$ nonzero entries per row, was considered by Neri & Metz (2016), who analysed outliers and eigenvectors in locally tree-like models; in addition, Bordenave et al. (2023); Coste & Stephan (2021) provided results on isolated eigenvalues and eigenvectors for models with i.i.d. entries. The sparse setting also arises in community detection for the stochastic block model with bounded mean degree. In this case, high-degree vertices can mask the adjacency matrix outliers, whereas the outliers of the non-backtracking operator become observable once above the Kesten–Stigum threshold. This question was first considered in Krzakala et al. (2013) and was later established on a rigorous basis in Bordenave et al. (2015).

In our setting, the dependence between transition counts along a single trajectory and their empirical row normalisation introduces additional structure that is not covered by the previous results. To deal with this structure, we introduce again a surrogate random matrix model. We fix $\alpha > 0$, $\beta \in [0, 1)$ and $u \in \{-1, 1\}^V$. We then draw $m_1, \dots, m_V$ i.i.d. $\mathrm{Pois}(\alpha)$, and set $n = \sum_i m_i$, $\mathcal{S} = \{i : m_i > 0\}$. Given m, define the *size-biased jump law* out of state i,

$$q_{ij} := \frac{m_j\,(1 - \beta u_i u_j)}{\sum_l m_l\,(1 - \beta u_i u_l)}, \qquad j \in \mathcal{V}, \tag{14}$$

and draw the rows of $\widetilde{M}$ independently, row i being $\mathrm{Mult}(m_i, q_{i\cdot})$: the m_i jumps out of i are i.i.d. with law $q_{i\cdot}$, and $\widetilde{M}_{ij}$ counts those landing in j. We finally set $\widetilde{P} = D^{-1}\widetilde{M}$ with $D = \mathrm{Diag}(m)$, rows with $m_i = 0$ being zero. As we discuss in detail in section E.1.2, we claim that this matrix model has the same threshold and eigenvector overlap as the count matrix for three reasons. First, for any fixed collection of states, the visit counts converge jointly to independent $\mathrm{Pois}(\alpha)$ variables as $V \to \infty$, supporting the surrogate's Poisson count assumption; second, transition probabilities proportional to $m_j(1 - \beta u_i u_j)$ capture the preference for frequently visited states; and third, if we condition on the visit counts, any fixed number of transitions are asymptotically independent, motivating the surrogate's replacement of weak residual dependence by exactly independent rows.

The following claim gives our asymptotically sharp prediction for the recovery threshold and the resulting overlap:

Claim 4.1 ("BBP" transition)**.** *Consider the size-biased jump model above (a surrogate for the empirical count matrix of the Markov chain). Let $\mathcal{S} := \{i \in \mathcal{V} : m_i > 0\}$ the visited set, and let $\rho_\alpha := (1 - e^{-\alpha})/\alpha \in (0, 1)$. Then, as $n, V \to \infty$ with $n/V \to \alpha \in (0, \infty)$ the (complex) spectrum of $\widetilde{P}$ consists of:*

- *A non-degenerate eigenvalue* $z_{\mathrm{triv.}} = 1 + o(1)$.
- *If and only if* $\beta > \rho_\alpha$*, a non-degenerate eigenvalue at* $z_\star = -\beta + o(1)$*. We denote* $u_\star$ *a corresponding eigenvector with* $\|u_\star\|^2 = V$.
- *The rest of the eigenvalues converge to a well-defined limiting density, supported in the circle* $\{z \in \mathbb{C} : |z| \le \rho_\alpha\}$.

Moreover, we denote

$$h_{\mathrm{BBP}}(\alpha) \coloneqq \mathbb{E}\left[\frac{1}{m}\,\middle|\, m \ge 1\right] = \int_0^1 \frac{e^{\alpha u} - 1}{e^\alpha - 1}\,\frac{\mathrm{d}u}{u}, \qquad m \sim \mathrm{Poisson}(\alpha). \tag{15}$$

Then we have, in probability as $V \to \infty$*:*

$$\lim_{V\to\infty} \frac{\langle u_\star, u\rangle^2}{V} = \begin{cases} \dfrac{(1-e^{-\alpha})(\beta^2 - \rho_\alpha^2)}{\beta^2 - \rho_\alpha^2 + (1-\beta^2)h_{\mathrm{BBP}}(\alpha)} & \text{if } \beta > \rho_\alpha, \\ 0 & \text{if } \beta \le \rho_\alpha. \end{cases} \tag{16}$$

In section E, we provide a heuristic justification of claim 4.1. Figure 4 shows that eq. (16) matches simulations of a finite Markov chain well. We emphasize that our justification of claim 4.1 is based on several simplifications and approximations: a fully rigorous derivation is an interesting problem of random matrix theory, which we leave to future work. We note that the outlier eigenvalue of $\widetilde{P}$ corresponds to the position of an eigenvalue of $P^\star$. This is especially noteworthy because, for a real Hermitian random matrix, the two typically do not coincide, and the outlier appears at a location different from the ground truth matrix eigenvalue location.

5 Concluding perspectives

We introduced a method for comparing the generative behaviour of autoregressive transformers through maximal coupling, providing a direct way to probe their transition from memorisation to generalisation. In the future, it will be important to scale our experiments to larger transformers, longer contexts, and larger training sets. It will also be interesting to study how post-training affects generalisation: will fine-tuning or preference optimisation increase convergence and hence improve generalisation, e.g. by imposing additional shared objectives and biases, or decrease convergence by amplifying differences from pre-training. On the theoretical side, analysing memorisation for Markov chains and considering richer data-generating models is an important challenge. It would be interesting to extend the idea of maximal coupling to entire sequences (Völlering, 2016).

We have defined memorisation in terms of literal regurgitation of training set sequences. While this form of memorisation captures the failure mode of generative models, memorisation in the sense of factual recall – the ability to retrieve specific facts, such as the capital city of a country – is desirable, and large language models excel at it (Petroni et al., 2019). Analysing the limits of factual recall in simple associative memories has attracted a lot of interest recently from the experimental (Roberts et al., 2020; Allen-Zhu & Li, 2025) and the theoretical point of view (Cabannes et al., 2024a;b; Nichani et al., 2025; Vural et al., 2026; Kim et al., 2026; Giorlandino et al., 2026; Barnfield et al., 2026). Developing a finer understanding of how storing facts and good generation can co-exist, and what constraints one puts on the other, is an interesting direction for future research; see Farné et al. (2026) for a recent analysis in the context of supervised classification.

Going back to diffusion models and real images, Kadkhodaie et al. (2024); Kamb & Ganguli (2025) argued that the inductive biases of convolutional denoisers are crucial in achieving the transition from memorisation to generalisation. More precisely, their analysis shows that convolutional denoisers are biased toward smooth, geometry-adaptive harmonic representations which match the statistical structure of natural images (Mallat, 1998). How to formulate and then test an analogous statement for transformers is far less clear. Do autoregressive transformers possess an implicit bias that restricts them to a low-complexity class of text distributions? If so, is this bias aligned with the structure of natural language, or does successful generalisation emerge for different reasons? In images, the relevant bases are geometric and multiscale; in text, the corresponding objects might involve hierarchies of syntax, semantics, or topical correlations over long contexts. Identifying the right "basis" or representation class for text and testing whether transformers are biased towards it is, in our view, the central problem raised by the present work.

AI USE STATEMENT

In this work, we used generative AI tools (Fable 5.1 and Claude Opus 5.5) for help with the derivation of claim 4.1 and suggestions of approaches to prove it and for writing (suggestions of approaches and simplifications, corrections, drafts) of section E. We also used generative AI tools (ChatGPT 5.5) for literature review and to verify the soundness of statements of section D. We also used ChatGPT 5.5 for polishing of main text and the appendices. We used ChatGPT 5.5 to review code for training and evaluating transformer models. Beyond what we described above, we have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, or to interpret results. Additionally, we used generative AI tools to create or modify scientific figures or images and to summarize or analyse existing literature. We have reviewed all AI-assisted work: we carefully performed all the calculations presented ourselves. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

REPRODUCIBILITY STATEMENT

To ensure reproducibility of our results, we:

- provide full experimental details of our experiments with transformers in section B
- provide code for dataset preparation, training our models, and for evaluating memorisation, convergence, and subspace recovery with the supplementary material
- For our theoretical results, we give full details of our analysis of convergence in Markov chains in section D and for the BPP transition in section E.

REFERENCES

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion. *arXiv:2410.08727*, 2024.

Beatrice Achilli, Luca Ambrogioni, Carlo Lucibello, Marc Mézard, and Enrico Ventura. Memorization and generalization in generative diffusion under the manifold hypothesis. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073401, 2025.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, 2025.

Johannes Alt, László Erdős, and Torben Krüger. Spectral radius of random matrices with independent entries. *Probability and Mathematical Physics*, 2(2):221–280, 2021.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of probability*, 33(5):1643–1697, 2005.

Nicholas Barnfield, Juno Kim, Eshaan Nichani, Jason D Lee, and Yue M Lu. Sharp capacity thresholds in linear associative memory: From winner-take-all to listwise retrieval. *arXiv preprint arXiv:2605.05189*, 2026.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

Florent Benaych-Georges and Jean Rochet. Outliers in the Single Ring Theorem. *Probability Theory and Related Fields*, 165(1):313–363, June 2016.

Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Charles Bordenave and Mireille Capitaine. Outlier eigenvalues for deformed iid random matrices. *Communications on Pure and Applied Mathematics*, 69(11):2131–2194, 2016.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 1347–1357. IEEE, 2015.

Charles Bordenave, Simon Coste, and Raj Rao Nadakuditi. Detection thresholds in very sparse matrix completion. *Foundations of Computational Mathematics*, 23(5):1619–1743, 2023.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, February 2013.

Pierre Bousseyroux and Marc Potters. The eigenvalues and eigenvectors of finite-rank normal perturbations of large rotationally invariant non-hermitian matrices. *arXiv preprint arXiv:2601.10427*, 2026.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations*, 2024a.

Vivien Cabannes, Berfin Simsek, and Alberto Bietti. Learning associative memories with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Giovanni Catania, Aurélien Decelle, Cyril Furtlehner, and Beatriz Seoane. A theoretical framework for overfitting in energy-based modeling. In *ICML 2025-Forty-Second International Conference on Machine Learning*. PLMR, 2025.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 2022.

Nicholas Cook, Walid Hachem, Jamal Najim, and David Renfrew. Non-Hermitian random matrices with a variance profile (I): deterministic equivalents and limiting ESDs. *Electronic Journal of Probability*, 23:1 – 61, 2018.

Simon Coste and Ludovic Stephan. A simpler spectral approach for clustering in directed networks. *arXiv preprint arXiv:2102.03188*, 2021.

Samuel F Edwards and Raymund C Jones. The eigenvalue spectrum of a large symmetric random matrix. *Journal of Physics A: Mathematical and General*, 9(10):1595–1603, 1976.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv:2305.07759*, 2023.

Gabriele Farné, Fabrizio Boncoraglio, and Lenka Zdeborová. The rules-and-facts model for simultaneous generalization and memorization in neural networks. *arXiv:2603.25579*, 2026.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models. *arXiv:2505.16959*, 2025.

Robert G. Gallager. *Stochastic processes: theory for applications*. Cambridge University Press, Cambridge, United Kingdom ; New York, 2013. ISBN 978-1-107-03975-9.

Jerome Garnier-Brun, Luca Biggio, Davide Beltrame, Marc Mézard, and Luca Saglietti. Biased generalization in diffusion models. *arXiv:2603.03469*, 2026.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Alessio Giorlandino, Sebastian Goldt, and Antoine Maillard. Factual recall in linear associative memories: sharp asymptotics and mechanistic insights. *arXiv preprint arXiv:2605.10795*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Silvio Kalaj, Clarissa Lauditi, Gabriele Perugini, Carlo Lucibello, Enrico M Malatesta, and Matteo Negri. Random features hopfield networks generalize retrieval to previously unseen examples. *Physica A: Statistical Mechanics and its Applications*, pp. 130946, 2025.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

Andrej Karpathy. nanoGPT: The simplest, fastest repository for training/finetuning medium-sized gpts. `https://github.com/karpathy/nanoGPT`, 2022. GitHub repository, accessed 2026-06-11.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D. Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. *arXiv:2603.26554*, 2026.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

Hugo Latourelle-Vigeant, Sinho Chewi, Aram-Alexandre Pooladian, John Sous, and Theodor Misiakiewicz. Generalization, memorization, and overfitting for diffusion models trained in the lazy high-dimensional regime. *arXiv preprint arXiv:2608.23938*, 2026.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Yue M Lu, Mary Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025.

Antoine Maillard and Sebastian Goldt. Memorisation, convergence and generalisation in generative models. *arXiv:2605.21402*, 2026.

Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1998.

Pierre Marion and Yu-Han Wu. Understanding diffusion models requires rethinking (again) generalization. *arXiv:2605.06077*, 2026.

Vicente Conde Mendes, Lorenzo Bardone, Cédric Koller, Jorge Medina Moreira, Vittorio Erba, Emanuele Troiani, and Lenka Zdeborová. A solvable high-dimensional model where nonlinear autoencoders learn structure invisible to pca while test loss misaligns with generalization. *arXiv:2602.10680*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Claudia Merger and Sebastian Goldt. Local coverage governs memorization in diffusion models. *arXiv:2606.14390*, 2026.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Izaak Neri and Fernando Lucas Metz. Eigenvalue Outliers of Non-Hermitian Random Matrices with a Local Tree Structure. *Physical Review Letters*, 117(22):224101, November 2016.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv:1412.6614*, 2014.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Alex Nguyen and Gautam Reddy Nallamala. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 8677–8697, 2025.

Eshaan Nichani, Jason Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations*, 2025.

Flavio Nicoletti, Chenxiao Ma, Enrico Ventura, Luca Saglietti, and Stefano Sarao Mannelli. The interplay of data structure and imbalance in the learning dynamics of diffusion models. *arXiv:2605.06367*, 2026.

OpenAI. tiktoken, 2025. URL `https://github.com/openai/tiktoken`.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv:2505.21777*, 2025.

Marc Potters and Jean-Philippe Bouchaud. *A first course in random matrix theory: for physicists, engineers and data scientists*. Cambridge University Press, 2020.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in neural information processing systems*, 36:14228–14246, 2023.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, November 2020.

Jaron Sanders and Albert Senen-Cerda. Spectral norm bounds for block markov chain random matrices. *Stochastic Processes and their Applications*, 158:134–169, 2023.

Jaron Sanders, Alexandre Proutière, and Se-Young Yun. Clustering in block markov chains. *The Annals of Statistics*, 48(6):3488–3512, 2020.

E. Seneta. *Non-negative Matrices and Markov Chains*. Springer Series in Statistics. Springer, New York, NY, 1981.

John G Skellam. The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 109(3):296–296, 1946.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Terence Tao. Outliers in the spectrum of iid matrices with bounded rank perturbations. *Probability Theory and Related Fields*, 155(1):231–263, 2013.

Florian Völlering. On maximal agreement couplings. *arXiv:1608.01511*, 2016.

Nuri Mert Vural, Alberto Bietti, Mahdi Soltanolkotabi, and Denny Wu. Learning to recall with transformers beyond orthogonal embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026.

Binxu Wang, Jacob Zavatone-Veth, and Cengiz Pehlevan. A random matrix theory perspective on the consistency of diffusion models. *arXiv:2602.02908*, 2026.

TLH Watkin and J-P Nadal. Optimal unsupervised learning. *Journal of Physics A: Mathematical and General*, 27(6):1899–1915, 1994.

Peter Whittle. Some distribution and moment formulae for the markov chain. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 17(2):235–242, 1955.

Anru Zhang and Mengdi Wang. Spectral state compression of markov processes. *IEEE transactions on information theory*, 66(5):3202–3231, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Huijie Zhang, Jinfan Zhou, Yifu Lu, Minzhe Guo, Peng Wang, Liyue Shen, and Qing Qu. The emergence of reproducibility and consistency in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60558–60590. PMLR, 2024.

Ziwei Zhu, Xudong Li, Mengdi Wang, and Anru Zhang. Learning markov models via low-rank optimization. *Operations Research*, 70(4):2384–2398, 2022.

A SAMPLING FROM A PAIR OF TRANSFORMERS USING MAXIMAL COUPLING

In this section, we describe how we generate pairs of sequences from two autoregressive transformers using some shared randomness to couple the sampling process across the models. For diffusion models, the coupling considered by Kadkhodaie et al. (2024) consists in initialising two independently trained denoising diffusion models with the same Gaussian noise to start the generation. For autoregressive models, there is no analogous continuous latent variable. We therefore construct a joint sampling procedure by coupling the categorical next-token distributions of the two transformers at every generation step. Several choices are possible; we discuss an intuitive choice that we call "synchronous coupling", and the maximal coupling that is borrowed from the theory of Markov chains; see sections 4.1–4.3 of (Levin & Peres, 2017) for an introduction.

A.1 SETTING

To make this section self-contained, we briefly recall the setting. Let $\mathcal{V}$ denote a vocabulary of $V \coloneqq |\mathcal{V}|$ tokens. We consider autoregressive transformers which, given a sequence or "history" of tokens $h_t = (x_1, \dots, x_{t-1})$, compute a probability distribution over the next token,

$$p_t(v \mid h_t) \coloneqq \mathrm{pr}(x_t = v \mid h_t), \qquad v \in \mathcal{V}. \tag{A.1}$$

The probability assigned by the transformer to a sequence $x_{1:T}$ then factorises as

$$p(x_{1:T}) = \prod_{t=1}^{T} p_t(x_t \mid x_{<t}). \tag{A.2}$$

We will sample jointly from two transformers indexed by $b \in \{0, 1\}$. At generation step t, let $h_t^{(b)}$ denote the current history of transformer b, including the common prompt when one is provided. For compactness, we write

$$p_t(v) \coloneqq p_t^{(0)}(v \mid h_t^{(0)}), \qquad q_t(v) \coloneqq p_t^{(1)}(v \mid h_t^{(1)}) \tag{A.3}$$

for the two next-token distributions. A **coupling** of p_t and q_t is a joint distribution for a pair of tokens

$$\left(X_t^{(0)}, X_t^{(1)}\right) \in \mathcal{V} \times \mathcal{V} \tag{A.4}$$

whose marginals are p_t and q_t, respectively. Sampling such a coupling at every step, and appending the resulting tokens to the two histories, defines a joint sampling procedure for the two autoregressive models.

A.2 SYNCHRONOUS COUPLING

A simple way of coupling the two next-token distributions that is analogous to the construction of Kadkhodaie et al. (2024) is to sample from p_t, q_t via the inverse of the cumulative density function (CDF) using the *same* the same scalar random variable. Denoting the CDFs of p_t and q_t by

$$F_{p_t}(k) \coloneqq \sum_{j=1}^{k} p_t(v_j), \qquad F_{q_t}(k) \coloneqq \sum_{j=1}^{k} q_t(v_j), \tag{A.5}$$

for each new token we draw a "latent" $u_t \sim \mathrm{Unif}([0, 1))$ and set

$$X_t^{(0)} = F_{p_t}^{-1}(u_t), \qquad X_t^{(1)} = F_{q_t}^{-1}(u_t), \tag{A.6}$$

where we (slightly abusively) identified a token $v_k \in \mathcal{V}$ with its index $k \in [V]$. This sampling procedure constitutes what we will call the *synchronous coupling*. Each token separately has the correct marginal distribution, while the use of the "shared" latent u_t ensures the next tokens $X_t^{(b)}$ are correlated. Moreover, given the latent vector $u = (u_t)$, the output of the transformers is deterministic, akin to how a diffusion model that is sampled using the ODE flow is deterministic given the initial latent noise variable.

However, it turns out that synchronous coupling does not in general maximise the probability that the two models generate the same token. We therefore use a different coupling that maximises token agreement at every generation step. This maximal coupling has the additional advantage that it lends itself to an appealing theoretical description in the case of Markov chains.

Algorithm 1 Maximal coupling of two next-token distributions

Require: Probability distributions p, q on the vocabulary $\mathcal{V}$

$r(v) \leftarrow \min\{p(v), q(v)\}$ for all $v \in \mathcal{V}$

$\alpha \leftarrow \sum_{v \in \mathcal{V}} r(v)$

Sample $B \sim \text{Bernoulli}(\alpha)$

if $B = 1$ **then**

  Sample

$$\bar{r}(v) \leftarrow \frac{r(v)}{\alpha}, \qquad Z \sim \bar{r}$$

  $X^{(0)} \leftarrow Z, \quad X^{(1)} \leftarrow Z$

else

  Define the residual distributions

$$\bar{p}(v) \leftarrow \frac{p(v) - r(v)}{1 - \alpha}, \qquad \bar{q}(v) \leftarrow \frac{q(v) - r(v)}{1 - \alpha}$$

  Sample $X^{(0)} \sim \bar{p}$

  Sample $X^{(1)} \sim \bar{q}$ independently

end if

return $(X^{(0)}, X^{(1)})$

A.3 Maximal coupling

For two probability distributions p_t and q_t on $\mathcal{V}$, their total variation distance is given by

$$\mathrm{d_{TV}}(p_t, q_t) = \frac{1}{2} \sum_{v \in \mathcal{V}} |p_t(v) - q_t(v)| \,. \tag{A.7}$$

A standard result from coupling theory – see for example Levin & Peres (2017, sec. 4.2) – states that for any coupling $(X_t^{(0)}, X_t^{(1)})$ of p_t and q_t,

$$\mathbb{P}\left(X_t^{(0)} \neq X_t^{(1)} \,\middle|\, h_t^{(0)}, h_t^{(1)}\right) \geq d_{\mathrm{TV}}(p_t, q_t). \tag{A.8}$$

The *maximal coupling* (also referred to as an optimal coupling in Levin & Peres (2017)) saturates this inequality; in other words, under maximal coupling, the probability that the two transformers produce the same next token is given by

$$\mathrm{pr}\left(X_t^{(0)} = X_t^{(1)} \,\middle|\, h_t^{(0)}, h_t^{(1)}\right) = 1 - d_{\mathrm{TV}}(p_t, q_t). \tag{A.9}$$

For completeness, we give the sampling algorithm for maximal coupling. Define the common part of the two distributions,

$$r_t(v) \coloneqq \min\{p_t(v), q_t(v)\}, \qquad \alpha_t \coloneqq \sum_{v \in \mathcal{V}} r_t(v) = 1 - d_{\mathrm{TV}}(p_t, q_t). \tag{A.10}$$

With probability α_t, we draw a token from the normalised common distribution

$$\bar{r}_t(v) \coloneqq \frac{r_t(v)}{\alpha_t}, \tag{A.11}$$

and assign it to both transformers,

$$X_t^{(0)} = X_t^{(1)} \sim \bar{r}_t. \tag{A.12}$$

With the remaining probability $1 - \alpha_t$, we sample independently from the two residual distributions

$$\bar{p}_t(v) \coloneqq \frac{p_t(v) - r_t(v)}{1 - \alpha_t}, \qquad \bar{q}_t(v) \coloneqq \frac{q_t(v) - r_t(v)}{1 - \alpha_t}, \tag{A.13}$$

and set

$$X_t^{(0)} \sim \bar{p}_t, \qquad X_t^{(1)} \sim \bar{q}_t. \tag{A.14}$$

The supports of $\bar{p}_t$ and $\bar{q}_t$ are disjoint, so that the two tokens are necessarily different in the second branch. Furthermore,

$$\alpha_t \bar{r}_t + (1 - \alpha_t)\bar{p}_t = p_t, \qquad \alpha_t \bar{r}_t + (1 - \alpha_t)\bar{q}_t = q_t, \tag{A.15}$$

showing explicitly that the two marginal distributions are unchanged. We summarise the sampling algorithm at a single generation step t in algorithm 1.

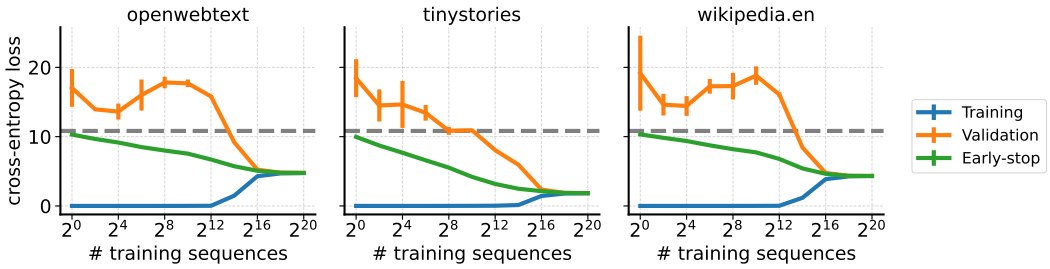


Figure B.1: **Train, test and early-stopping error of the transformers.** We show the training, testing and early-stopping error for the transformers trained on three different data sets as a function of the number of sequences used for training ($L = 128$ tokens).

B Details on transformer training and sampling

Data sets We performed our experiments on three different datasets: OpenWebText, which aims to reconstruct the training corpus of GPT2 in an open way, and contains about 40GB of text across 8,013,769 documents; tinystories, a smaller data set for research purposes (Eldan & Li, 2023), and `wikipedia.en`, a collection of Wikipedia articles in English. We tokenised all three data sets using the the OpenAI GPT2 tokenizer provided by the `tiktoken` library (OpenAI, 2025), which has a vocabulary size of 50257.

Architecture We used Karpathy's nanoGPT implementation (Karpathy, 2022) of a GPT-style decoder-only Transformer. The model maps token indices to learned token embeddings and adds learned positional embeddings for sequences of length at most 128, followed by dropout. The network consists of 8 pre-normalisation Transformer blocks with embedding dimension 256 and 8 attention heads. Each block applies causal multi-head self-attention, followed by an output projection and residual connection. This is followed by a second layer normalisation, a position-wise MLP with hidden dimension four times the embedding dimension, GELU nonlinearity, output projection, dropout, and another residual connection. We tied the weights of token embeddings and unembeddings.

Training We trained all transformer models on next-token prediction with the cross-entropy loss. We used the AdamW optimizer with weight decay 0.1, $\beta_1 = 0.9, \beta_2 = 0.95$ for all the weight tensors and the embeddings, but not for the biases and LayerNorm parameters. We used a maximum learning rate of 6×10^{-4}, with a linear warmup for 2,000 SGD steps, and cosine decay to 6×10^{-5} over the prescribed number of optimization steps. We stopped stops early if the running training loss fell below 10^{-3}, otherwise we trained all models for the same number of SGD steps that assured good performance for models trained on the largest data set.

Sampling For details on our sampling procedure, which relies on the idea of maximum coupling, see section A and algorithm 1.

Subspace overlap To compute the subspace overlap Q that we discuss in section 3, we evaluated the two transformers trained on different splits or branches of the training set on the same held-out test set. For each branch $b \in \{0, 1\}$, we used the first 64 tokens of each test sequence as a context h_a and computed the conditional next-token probabilities $p_j^{(b)}(h_a)$ by applying the softmax to the output logits. We then applied a spectral estimator to the next-token probabilities to estimate shared group structure. The form of the spectral estimator is motivated by the Markov chain model defined in section C and in particular eq. (C.4). In particular, we remove the context-independent component and extract the leading spectral directions. Here the relevant statistics are averaged over the M held-out contexts:

$$\bar{p}^{(b)} = \frac{1}{M} \sum_{a=1}^{M} p^{(b)}(h_a), \qquad C^{(b)} = \frac{1}{M} \sum_{a=1}^{M} \left(p^{(b)}(h_a) - \bar{p}^{(b)} \right) \left(p^{(b)}(h_a) - \bar{p}^{(b)} \right)^{\top}. \tag{B.1}$$

We estimate the leading eigenvectors $v_r^{(b)}$ of the probability covariance $C^{(b)}$; to account for unequal token frequencies, we rescale each direction by the empirical unigram frequency $\hat{\pi}$ that we estimate from all test tokens; we set the frequency of tokens that do not appear in the test set to $1/V$, where V is the vocabulary size. We then measure the agreement between the two branches by computing all pairwise inner products between the first k rescaled principal directions of the two branches, for all k. We sum the squared inner products and divide by k to obtain $Q(k)$, which is invariant to the arbitrary signs of the principal directions. This quantity compares the leading patterns of context-dependent variation in the predicted probabilities and is invariant to the arbitrary signs of the spectral directions, and is shown in fig. 2 for three different data sets.

C Detecting group structure in Markov chains and transformers

C.1 The simplest language model

We consider a particularly simple model of language where there are two types of words that can be though of as nouns and verbs. In a typical string of text, words of the two groups alternate, to yield simple sentences of the form *subject - verb - object*. We model text from this language as a sequence of tokens from a vocabulary of V tokens $\mathcal{V} = [V]$. We encode the group of each token in a vector $u = (u_i)_{i<V}$, where $u_i = \pm 1$. We will assume that the tokens in our training set are sampled from a ground-truth Markov chain, the teacher, that has stationary state $\pi^* \in \Delta^{V-1}$ and a transition matrix

$$P^*_{ij} = \mathrm{pr}(w_{t+1} = j | w_t = i) = \pi^*_j \left(1 - \beta u_i u_j\right), \tag{C.1}$$

where w_t are the tokens in a sequence indexed by $t \geq 0$. The constant $\beta \in [0, 1]$ can be as a signal-to-noise ratio for the group structure of the language. If $\beta = 0$, subsequent tokens are drawn independently of each other, and the group structure is lost from the generated sequence. Instead, if $\beta > 0$, transitions from token i to j are more likely if i and j are in *different* groups, and hence $u_i u_j = -1$, while transitions between tokens in the same group are suppressed.

It is easy to verify that the transition matrix P^*_{ij} is a valid row-stochastic matrix with non-negative entries and $\sum_j P^*_{ij} = 1$ for all $i, j \in [V]$, and that its stationary state is given by π^*, $\sum_i \pi^*_i P^*_{ij} = \pi^*_j$, provided that the tokens in each group have equal probability mass in the steady state,

$$\sum_i \pi^*_i u_i = 0. \tag{C.2}$$

The transition matrix P^* has thus the mathematical structure of a degree-correted stochastic block model (Karrer & Newman, 2011) that is disassociative, i.e. nodes across groups are more likely to "connect" then nodes within groups. For reviews of the enormous literature on the stochastic block model and related matrix models in our context, see our discussion in sections 4.2 and E.4.

C.2 Spectral estimators of the group structure

A robust approach to recovering the group structure of a Markov chain is via spectral methods applied ot the transition matrix. To see this, we define $D^\star = \mathrm{Diag}(\pi^\star)$, $s^\star_i = \sqrt{\pi^\star_i}$, and $z^\star_i = s^\star_i u_i$. From eq. (5), we obtain

$$B^\star \equiv (D^\star)^{1/2} P^\star (D^\star)^{-1/2} - s^\star (s^\star)^\top = -\beta z^\star (z^\star)^\top. \tag{C.3}$$

From the balance condition eq. (C.2), we have $(s^\star)^\top z^\star = 0$, so subtracting the stationary component leaves a rank-one matrix with eigenvalue $-\beta$. For $\beta > 0$, its leading right singular vector is $\pm z^\star$, whose signs recover u up to an overall sign. For $\beta = 0$, this component vanishes. The positive factors $\sqrt{\pi^\star_i}$ do not affect the population signs, so the same construction applies to both stationary-state models.

Given a sequence of length L, we can thus estimate a transition matrix $\hat{P}$, for example using the maximum-likelihood or a Dirichlet-regularized estimator, compute its stationary distribution $\hat{\pi}$, and compute the leading right singular vector of $\hat{B}$ to yield a spectral estimate $\hat{u}$ of the group structure.

For the analysis of next-token probabilities, as they are computed by transformers, it is also convenient to define a matrix in terms of next-token probabilities. Given a set of test sequences (or tokens for the Markov chain), we can compute the empirical covariance of the centered next-token probabilities. At the population level, we can replace the average over a finite test set via the average over the steady-state distribution and find that

$$C_{ij} = \sum_n \pi_n \left(P_{ni} - \sum_\ell \pi_\ell P_{\ell i} \right) \left(P_{nj} - \sum_\ell \pi_\ell P_{\ell j} \right) \tag{C.4}$$

$$= \beta^2 u_i u_j \pi_i \pi_j \tag{C.5}$$

so performing PCA on C and taking $\hat{u} = \mathrm{sign}(v_1)$ gives a good estimate of u, since $\mathrm{sign}(u_i) = \mathrm{sign}(u_i \pi_i)$.

D Analysis of convergence in the Markov chain model

We want to analyze the dependence of the survival probability quantity on the spectral radius of the transition kernel; we denote V the vocabulary size, which corresponds here to the number of states and probability matrix dimension, and n the number of observations.

D.1 Survival probability for a pair of Markov chains

We consider two generated sequences, indexed by $a \in \{1, 2\}$, over a finite vocabulary $\mathcal{V}$, with histories $h_s^{(a)} = (X_0^{(a)}, \dots, X_s^{(a)})$ and conditional next-token distributions $p_s^{(a)}(\cdot \mid h_s^{(a)})$. Both start from the same token, $X_0^{(1)} = X_0^{(2)} \sim \nu_0$, and are run under the step-wise maximal coupling (Levin & Peres, 2017, sec 4.2.), whose law and expectation we denote $\mathrm{pr}_{\max}$ and $\mathbb{E}_{\max}$. When the two histories coincide we write X_s and $h_s = (X_0, \dots, X_s)$ for their common values. For a history $\eta = (\eta_0, \dots, \eta_s) \in \mathcal{V}^{s+1}$ we write $\eta_{0:r} = (\eta_0, \dots, \eta_r)$ for its truncation and (η, j) for its extension by a token j. We want to quantify the probability that the two sequences match up to a given time t.

We introduce two quantities: the distribution of the token on which the chains agree at the next step, and the probability that they agree at all. Under the maximal coupling,

$$k_s(\eta, j) \coloneqq \mathrm{pr}_{\max}\Big(X_{s+1}^{(1)} = X_{s+1}^{(2)} = j \,\Big|\, h_s^{(1)} = h_s^{(2)} = \eta \Big)$$

$$= \min\Big\{ p_s^{(1)}(j \mid \eta),\, p_s^{(2)}(j \mid \eta) \Big\}, \tag{D.1}$$

$$g_s(\eta) \coloneqq \sum_{j \in \mathcal{V}} k_s(\eta, j) = 1 - d_{\mathrm{TV}}\Big(p_s^{(1)}(\cdot \mid \eta),\, p_s^{(2)}(\cdot \mid \eta) \Big). \tag{D.2}$$

Thus $g_s(\eta)$ is the survival weight of the step, i.e. the probability of agreeing at step $s+1$ given the common history η, while $k_s(\eta, \cdot)/g_s(\eta)$ is the law of the next token given that the chains agreed on it. If $g_s(\eta) = 0$ the two next-token distributions have disjoint supports and the chains separate at step $s+1$ with probability one.

Let $\tau \coloneqq \inf\{s \geq 1 : X_s^{(1)} \neq X_s^{(2)}\}$ be the splitting time and

$$S(t \mid \nu_0) \coloneqq \mathrm{pr}_{\max}(\tau > t) = \mathbb{E}_{\max}\left[\prod_{s=1}^{t} \mathbb{1}\{X_s^{(1)} = X_s^{(2)}\} \right]. \tag{D.3}$$

Let us define the weight of a matched trajectory, $\pi_s(\eta) \coloneqq \mathrm{pr}_{\max}(\tau > s,\ h_s = \eta)$, with $\pi_0 = \nu_0$. Two histories are equal exactly when the chains have agreed at every step, $\{h_s^{(1)} = h_s^{(2)} = \eta\} = \{\tau > s,\ h_s = \eta\}$, so eq. (D.1) gives the exact recursion

$$\pi_{s+1}(\eta, j) = \pi_s(\eta)\, k_s(\eta, j), \qquad S(t \mid \nu_0) = \sum_{\eta \in \mathcal{V}^{\,t+1}} \pi_t(\eta). \tag{D.4}$$

Unrolling eq. (D.4):

$$\pi_t(\eta) = \nu_{0,\eta_0} \prod_{s=0}^{t-1} k_s(\eta_{0:s}, \eta_{s+1}) = \underbrace{\left[\nu_{0,\eta_0} \prod_{s=0}^{t-1} \frac{k_s(\eta_{0:s}, \eta_{s+1})}{g_s(\eta_{0:s})} \right]}_{\text{path probability under } \mathrm{pr}_K} \prod_{s=0}^{t-1} g_s(\eta_{0:s}). \tag{D.5}$$

The bracket is a normalised probability over histories of length $t+1$: it defines the matched-path process pr_K, with expectation $\mathbb{E}_K$, in which a *single* trajectory is generated by drawing $X_0 \sim \nu_0$ and then each token from $k_s(h_s,\cdot)/g_s(h_s)$, i.e. by resampling at every step conditionally on the two chains having agreed. Histories on which some g_s vanishes carry no weight in either factor and may be discarded. Summing eq. (D.5) over η,

$$\begin{aligned} S(t \mid \nu_0) &= \mathbb{E}_K\left[\prod_{s=0}^{t-1} g_s(h_s)\right] \\ &= \mathbb{E}_K\left[\prod_{s=0}^{t-1}\Big(1 - d_{\mathrm{TV}}\Big(p_s^{(1)}(\cdot \mid h_s),\, p_s^{(2)}(\cdot \mid h_s)\Big)\Big)\right]. \end{aligned} \tag{D.6}$$

Notice that the expectation is taken under the law pr_K described above.

For our first-order, time-homogeneous model the chains have transition matrices $P^{(a)}$, so eq. (D.1)–eq. (D.2) depend on the history only through its last token $i = \eta_s$, and not on s:

$$k_s(\eta, j) = K_{ij} \coloneqq \min\{P_{ij}^{(1)}, P_{ij}^{(2)}\}, \qquad g_s(\eta) = g(i) = (K\mathbf{1})_i, \tag{D.7}$$

with $\mathbf{1}$ the all-ones vector, so that $K_{ij}/g(i) = \mathrm{pr}_K(j \mid i)$. The matrix K is sub-stochastic, its row deficit $1 - g(i) = d_{\mathrm{TV}}(P_{i\cdot}^{(1)}, P_{i\cdot}^{(2)})$ being the splitting probability on token i. The recursion of eq. (D.4) then closes on the token marginal $\varrho_s(i) \coloneqq \mathrm{pr}_{\max}(\tau > s,\, X_s = i)$, giving $\varrho_{s+1}^\top = \varrho_s^\top K$ with $\varrho_0 = \nu_0$, so that

$$S(t \mid \nu_0) = \sum_{i \in \mathcal{V}} \varrho_t(i) = \nu_0^\top K^t \mathbf{1}, \qquad S(t \mid h_0 = i) = (K^t\mathbf{1})_i. \tag{D.8}$$

D.2 Setup and model

Our strategy for estimating the asymptotic logarithmic decay rate of $S_{\hat{K}}(t)$ is to

1. build a tractable proxy model K to describe a Markov chain and model the coupling of two Markov chains introduced in the main text and described through the kernel $\hat{K}$,
2. show that, under appropriate sampling conditions, the spectral radius $\rho(K)$ of this proxy model concentrates on $\rho(A)$ (with $A = \mathbb{E}K$) as $V \to \infty$.

While far from a rigorous analysis of the decay rate of $S_{\hat{K}}(t)$, this represents a first step, and shows very good agreement with numerical simulations (see fig. 3).

We first model the transition matrix associated to a trajectory a by assuming that the transitions between two states $X_{ij}^{(a)}$ within each row i are described by a multinomial distribution. Denoting by $N_i^{(a)}$ the variable describing the number of transitions in row i, this is expressed by the follwoing:

$$\big(X_{i1}^{(a)}, \ldots, X_{iV}^{(a)}\big) \mid \{N_i^{(a)} = k\} \sim \text{Multinomial}\big(k; P_{i1}^*, \ldots, P_{iV}^*\big). \tag{D.9}$$

We then Poissonize the number of observations per row $N_i^{(a)}$ for each row i by consider it a Poisson random variable of positive mean m_i, hence $N_i^{(a)} \sim \mathrm{Pois}(m_i)$. Through Poisson splitting (Gallager, 2013, Chapter 2), we then obtain that the transitions between two states are independent Poisson random variables, each with rate λ_{ij}:

$$X_{ij}^{(a)} \sim \mathrm{Pois}(\lambda_{ij}), \qquad \lambda_{ij} = m_i P_{ij}^* \tag{D.10}$$

We observe that we can then write $m_i = \sum_j \lambda_{ij}$, and, by construction, $\sum_i m_i = n$. To obtain the proxy probability matrix, we observe that normalizing each row by its total would introduce a dependence within the same row entries: in a further simplification step, we thus normalize every row by the corresponding deterministic expectation m_i.

We now want to model the transition kernel. This operator takes element wise the minimum between the probability matrices associated to two different independent trajectories, normalized by the corresponding counts per row. We define the transition kernel K for this model as

$$M_{ij} = \min\{X_{ij}^{(1)}, X_{ij}^{(2)}\}, \qquad K_{ij} = \frac{M_{ij}}{m_i}, \qquad A_{ij} = \mathbb{E}K_{ij} = \frac{\mathbb{E}M_{ij}}{m_i} = \frac{\mu(\lambda_{ij})}{m_i} \tag{D.11}$$

The stochasticity in our transition kernel comes only from the counts matrix, since we consider the counts per row fixed to its expected value. In this setting, to characterize the behavior of the transition kernel we utilize the random variable corresponding to the minimum of two $\mathrm{Pois}(x)$ independent variables X_x, Y_x. This has mean and variance that can be computed exploiting the Skellam distribution Skellam (1946) and correspond respectively to

$$\begin{aligned} \mu(x) &= \mathbb{E}(\min\{X_x, Y_x\}) = x\big[1 - e^{-2x}(I_0(2x) + I_1(2x))\big] = xh(x), \\ v(x) &= \mathrm{Var}(\min\{X_x, Y_x\}) = x - xe^{-2x}I_0(2x) - x^2e^{-4x}(I_0(2x) + I_1(2x))^2, \end{aligned} \tag{D.12}$$

where I_0 and I_1 are modified Bessel functions of the first kind. In the limit of x going to zero we can approximate $h(x) \sim x$ and $\mu(x) \sim v(x) \sim x^2$. The kernel can be now decomposed into its mean and centered fluctuation writing

$$K = Z + A, \qquad Z_{ij} = \frac{(M_{ij} - \mu(\lambda_{ij}))}{m_i}. \tag{D.13}$$

Thus A gives the deterministic spectral prediction, while Z contains the random fluctuations that must be controlled.

Given an initial strictly positive probability vector ν_0, we have seen in section D.1 that the survival probability $S_K(t)$ can be expressed as

$$S_K(t) = \nu_0^\top K^t \mathbf{1}, \qquad S_{\mathrm{ana}}(t) = \nu_0^\top A^t \mathbf{1}, \qquad t \in \mathbb{N}_0. \tag{D.14}$$

The first object corresponds to the survival probability obtained from the proposed proxy Poisson model, and the second is the survival probability obtained from the mean matrix A. We stress that we are using a proxy model for the transition kernel constructed using independent Poisson distributed matrix entries and deterministic denominators, so that $S_K(t)$ is itself a proxy for the survival probability of the original transition kernel.

In the analysis below, we use o, O to represent deterministic asymptotic bounds, and $o_{\mathbb{P}}, O_{\mathbb{P}}$ for the corresponding probabilistic bounds.

D.3 Spectral concentration: the main result

Our goal is to compare the random realization of the spectral radius $\rho(K)$ with the analytically computable one $\rho(A)$.

We refer to section D.2 for the construction of A and K. We recall that $A = \mathbb{E}K$ is deterministic and $K = A - Z$, with A and K both non-negative matrices.

Let us consider for further generality any matrix A such that $r > 0$ is a positive right eigenvector of A, i.e. $Ar = \rho(A)r$. By the Perron-Frobenius theory (Seneta, 1981, Chapter 1) this condition is given for example when A is irreducible or primitive. Lemma D.2 below shows that the deviation of $\rho(K)$ from the mean matrix radius $\rho(A)$ can be controlled by the following quantity:

$$\eta_r(Z) \coloneqq \max_i \frac{|(Zr)_i|}{r_i} \tag{D.15}$$

We note that $\eta_r(Z)$ is not affected by the normalization of the eigenvector r so in the setting in which the mean row sums are constant and $A \propto \mathbf{1}\mathbf{1}^\top$ we will set $r = \mathbf{1}$. We further notice that in the model of section D.2, we have

$$\eta_r(Z) = \max_i \frac{\left|\sum_j (M_{ij} - \mu(\lambda_{ij}))r_j\right|}{m_i r_i},$$

and we define

$$r_{\max} \coloneqq \max_j r_j.$$

Theorem D.1 (Concentration with fixed denominators)**.** *Consider the construction in section D.2 with K the transition kernel using Poissonization and fixed denominators. Suppose that A is a non-negative matrix such that $Ar = \rho(A)r$ with $r > 0$ a positive right eigenvector. Suppose also that there exists a constant λ_0, independent of V, such that $0 \le \lambda_{ij} \le \lambda_0 < \infty$ for every V and i, j. Then we have*

$$|\rho(K) - \rho(A)| \le \eta_r(Z), \qquad \eta_r(Z) = O_{\mathbb{P}}\left(\max_i \frac{\sqrt{\log V \sum_j v(\lambda_{ij}) r_j^2} + r_{\max} \log V}{m_i r_i} \right). \tag{D.16}$$

If the argument of $O_{\mathbb{P}}(\,\cdot\,)$ in eq. (D.16) *is $o(\rho(A))$, thenwe have further*

$$\mathbb{P}(\rho(K) > 0) \longrightarrow 1, \qquad \frac{\rho(K)}{\rho(A)} \xrightarrow{\mathbb{P}} 1, \qquad \log \rho(K) - \log \rho(A) \xrightarrow{\mathbb{P}} 0. \tag{D.17}$$

The criterion given by theorem D.1 compares the largest weighted row fluctuation with the spectral radius of the mean matrix. We prove theorem D.1 in section D.5.

D.4 Connection between concentration and survival probability

For every nonnegative matrix K with positive spectral radius $\rho(K)$ and every initial strictly positive probability vector, we have

$$\lim_{t \to \infty} \frac{1}{t} \log S_K(t) = \log \rho(K). \tag{D.18}$$

To show eq. (D.18), notice that when $\rho(K) > 0$, Perron-Frobenius theory (Seneta, 1981, Chapter 1) guarantees the existence of a nonnegative eigenvector $x \neq 0$ satisfying $Kx = \rho(K)x$, whose normalization we set as $\|x\|_\infty = 1$. Since $0 \le x \le \mathbf{1}$ and $\nu_0^\top x > 0$, positivity then implies

$$(\nu_0^\top x)\rho(K)^t \le \nu_0^\top K^t \mathbf{1} \le \|K^t\|_\infty. \tag{D.19}$$

We can now take the tth root, and use Gelfand's formula (Horn & Johnson, 2012, Corollary 5.6.14) ; this gives eq. (D.18). In an analogous way we obtain $\lim_{t\to\infty} \log S_{\mathrm{ana}}(t)/t = \log \rho(A)$. We observe that if $\rho(K) = 0$, then K is nilpotent and $S_K(t)$ is eventually zero, giving rate $-\infty$ with $\log 0 = -\infty$. Under the assumptions of theorem D.1, $\mathbb{P}(\rho(K) > 0) \to 1$ and under the same assumptions we thus get,

$$\lim_{t \to \infty} \frac{1}{t} \log S_K(t) - \log \rho(A) \xrightarrow{\mathbb{P}} 0. \tag{D.20}$$

D.5 Proof of Theorem D.1

D.5.1 Spectral comparison

Lemma D.2 (Collatz-Wielandt spectral comparison)**.** *Let $A \ge 0$ a non-negative matrix, $Ar = \rho(A)r$ with $r > 0$, and $K = A + Z \ge 0$. Recall the definition of η_r in eq.* (D.15)*. Then $|\rho(K) - \rho(A)| \le \eta_r(Z)$. Moreover, if $\eta_r(Z) < \rho(A)$, then*

$$|\log \rho(K) - \log \rho(A)| \le \frac{\eta_r(Z)}{\rho(A) - \eta_r(Z)}. \tag{D.21}$$

Proof. Since $Kr = \rho(A)r + Zr$ by construction and $|(Zr)_i| \le \eta_r(Z) r_i$, we can write

$$\rho(A) - \eta_r(Z) \le \frac{(Kr)_i}{r_i} \le \rho(A) + \eta_r(Z) \quad \text{for every } i. \tag{D.22}$$

For any non-negative matrix K and any strictly positive vector r the Collatz–Wielandt inequalities (Seneta, 1981, Chapter 1) imply that $\min_i (Kr)_i/r_i \le \rho(K) \le \max_i (Kr)_i/r_i$. This proves the spectral bound. If $\eta_r(Z) < \rho(A)$, then $\rho(K) \ge \rho(A) - \eta_r(Z) > 0$; applying $|\log x - \log y| \le |x - y|/\min\{x, y\}$ proves eq. (D.21). □

D.5.2 BERNSTEIN'S INEQUALITY

Exploiting lemma D.2 the objective is now to obtain a probabilistic upper bound on $\eta_r(Z)$. Since $\eta_r(Z)$ is defined as the maximum of weighted deviations across rows, we apply Bernstein's inequality to each row and then take a union bound.

For a fixed row i, define

$$\xi_{ij} = \big(M_{ij} - \mu(\lambda_{ij})\big) r_j, \qquad Y_i = \sum_j \xi_{ij} \quad \text{with} \quad \mathrm{Var}(Y_i) = \sum_j \mathrm{Var}(\xi_{ij}) = \sum_j v(\lambda_{ij}) r_j^2 =: \sigma_i^2. \tag{D.23}$$

This last equality follows as ξ_{ij} are independent and centred. Moreover,

$$(Zr)_i = \frac{Y_i}{m_i}, \qquad \eta_r(Z) = \max_i \frac{|Y_i|}{m_i r_i}. \tag{D.24}$$

To use the moment form of Bernstein's inequality (Boucheron et al., 2013, Theorem 2.10), it is enough to find a constant b, depending only on λ_0, such that

$$\mathbb{E}|M_s - \mu(s)|^k \le \frac{k!}{2} v(s) b^{k-2}, \qquad k \ge 3, \quad 0 \le s \le \lambda_0, \tag{D.25}$$

where $s \in [0, \lambda_0]$, $M_s \coloneqq \min\{X_s^{(1)}, X_s^{(2)}\}$, and $X_s^{(1)}, X_s^{(2)} \overset{\text{ind}}{\sim} \mathrm{Pois}(s)$ such that $\mathbb{E}M_s = \mu(s)$ and $\mathrm{Var}(M_s) = v(s)$.

We derive this estimate by controlling both its upper and lower tails. Write $z_+ = \max\{z, 0\}$ and $z_- = \max\{-z, 0\}$. For the negative tail, since $M_s \ge 0$ and $(M_s - \mu(s))_- \le \mu(s) \le s \le \lambda_0$, we have that for every $k \ge 3$

$$\mathbb{E}(M_s - \mu(s))_-^k \le \lambda_0^{k-2} \mathbb{E}(M_s - \mu(s))_-^2 \le \lambda_0^{k-2} v(s). \tag{D.26}$$

For the positive one instead,

$$(M_s - \mu(s))_+^k \le M_s^k \le X_s^k \mathbf{1}_{\{Y_s \ge 1\}}. \tag{D.27}$$

Thus, using independence, the fact that $\mathbb{P}(Y_s \ge 1) \le s$, $y^k \le k!(e^y - 1)$, and $\mathbb{E}e^{X_s} = e^{s(e-1)}$ we obtain

$$\mathbb{E}(M_s - \mu(s))_+^k \le s k! \big(e^{s(e-1)} - 1\big) \le c_1 k! s^2, \qquad \text{where} \qquad c_1 = \frac{e^{\lambda_0(e-1)} - 1}{\lambda_0}. \tag{D.28}$$

The last inequality follows from the convexity of the function $s \mapsto e^{s(e-1)} - 1$. Since $v(s) \sim s^2$ as $s \to 0$ and their ratio is positive and continuous for $s > 0$, we define $c_0 \coloneqq \inf_{0 < s \le \lambda_0} \frac{v(s)}{s^2} > 0$. Hence

$$\mathbb{E}(M_s - \mu(s))_+^k \le \frac{c_1}{c_0} k! v(s). \tag{D.29}$$

Combining the two tails and setting $b = \max\left\{1, \lambda_0, \frac{3c_1}{c_0}\right\}$, we obtain, for every $k \ge 3$, eq. (D.25).

It follows that

$$\sum_j \mathbb{E}|\xi_{ij}|^k = \sum_j r_j^k \mathbb{E}|M_{\lambda_{ij}} - \mu(\lambda_{ij})|^k \le \frac{k!}{2} \sum_j v(\lambda_{ij}) r_j^2 (b r_{\max})^{k-2} = \frac{k!}{2} \sigma_i^2 (b r_{\max})^{k-2}. \tag{D.30}$$

For a fixed row i, equation eq. (D.30) and Bernstein's inequality give,

$$\mathbb{P}\left(|Y_i| \ge \sqrt{2\sigma_i^2 x} + b r_{\max} x\right) \le 2e^{-x}, \qquad x \ge 0. \tag{D.31}$$

Applying the union bound over the V rows gives

$$\mathbb{P}\left(\exists i : |Y_i| \ge \sqrt{2\sigma_i^2 x} + b r_{\max} x\right) \le 2V e^{-x}. \tag{D.32}$$

Setting $x = 2\log V$, we obtain that, with probability at least $1 - 2/V$,

$$|Y_i| \le 2\sqrt{\sigma_i^2 \log V} + 2br_{\max}\log V \qquad \text{for every } i. \tag{D.33}$$

Since $\eta_r(Z)$ is defined as eq. (D.24) it follows that with probability at least $1 - 2/V$

$$\eta_r(Z) \le C \max_i \frac{\sqrt{\log V \sum_j v(\lambda_{ij}) r_j^2} + r_{\max}\log V}{m_i r_i}, \tag{D.34}$$

where C depends only on λ_0.

This translates in

$$\eta_r(Z) = O_{\mathbb{P}}\left(\max_i \frac{\sqrt{\log V \sum_j v(\lambda_{ij}) r_j^2} + r_{\max}\log V}{m_i r_i} \right), \tag{D.35}$$

where the constants depend only on λ_0. This proves eq. (D.16).

D.5.3 Completion of the proof of theorem D.1

Suppose that the argument of $O_{\mathbb{P}}(\,\cdot\,)$ in eq. (D.16) is $o(\rho(A))$: then $\frac{\eta_r(Z)}{\rho(A)} = o_{\mathbb{P}}(1)$. Lemma D.2 therefore yields $|\rho(K)/\rho(A) - 1| \le \eta_r(Z)/\rho(A) = o_{\mathbb{P}}(1)$. Moreover, we obtained that eq. (D.34) holds with probability at least$1 - 2/V$. Since its deterministic right-hand side is $o(\rho(A))$, it follows that $\mathbb{P}\big(\eta_r(Z) < \rho(A)\big) \longrightarrow 1$. On this event, lemma D.2 gives

$$\rho(K) \ge \rho(A) - \eta_r(Z) > 0, \qquad |\log\rho(K) - \log\rho(A)| \le \frac{\eta_r(Z)/\rho(A)}{1 - \eta_r(Z)/\rho(A)} = o_{\mathbb{P}}(1). \tag{D.36}$$

This completes the proof.

D.6 Homogeneous transition probability model

We now focus on the homogeneous model presented in the main text in section 4.1. For this model, we have that the teacher transition matrix is $P^*_{ij} = 1/V$, and expected number of observation for each row equal to $m_i = m = n/V$. The counts from state i to state j associated to the trajectory a (denoted $X_{ij}^{(a)}$) and the transition kernel are then the following:

$$X_{ij}^{(a)} \sim \mathrm{Pois}(mP^*_{ij}), \qquad a \in \{1,2\}, \qquad (K_0)_{ij} = \frac{\min\{X_{ij}^{(1)}, X_{ij}^{(2)}\}}{m}. \tag{D.37}$$

For this model notice that the common Poisson parameter is $\lambda = mP^*_{ij} = n/V^2$.

The corresponding deterministic mean matrix is

$$A_0 = \mathbb{E}K_0 = \frac{h(\lambda)}{V}\mathbf{1}\mathbf{1}^T, \qquad \rho(A_0) = h(\lambda) \tag{D.38}$$

Since $\mathbf{1}$ is an eigenvector of A_0, we can also rewrite $S_{ana}(t) = h(\lambda)^t$; moreover, decomposing $K_0 = A_0 + Z_0$ eq. (D.16) becomes

$$\eta_{\mathbf{1}}(Z_0) = O_{\mathbb{P}}\left(\frac{\sqrt{\log V v(\lambda) V} + \log V}{m} \right) \tag{D.39}$$

Corollary D.3 (Homogeneous spectral prediction)**.** *Suppose that* $0 < \lambda = n/V^2 \le \lambda_0 \le \infty$ *and that* $n \gg V^{3/2}\sqrt{\log V}$*. Then for any* $\nu_0 > 0$*,*

$$\frac{\rho(K_0)}{\rho(A_0)} \xrightarrow{\mathbb{P}} 1, \qquad \lim_{t\to\infty} \frac{1}{t}\log S_{K_0}(t) - \log h(\lambda) \xrightarrow{\mathbb{P}} 0. \tag{D.40}$$

Proof. Since the ratios $v(\lambda)/\lambda^2$ and $\mu(\lambda)/\lambda^2$ converge to 1 as $\lambda \to 0$, they admit continuous extensions to $\lambda = 0$. Consequently, on the interval $0 < \lambda \leq \lambda_0$ these ratios are bounded both above and away from zero. Using this, we can divide eq. (D.39) by $\rho(A_0)$ to obtain

$$\frac{\eta_{\mathbf{1}}(Z_0)}{\rho(A_0)} = O_{\mathbb{P}}\left(\sqrt{\frac{\log V}{V\lambda^2}} + \frac{\log V}{V\lambda^2}\right) = o_{\mathbb{P}}(1). \tag{D.41}$$

where the last equality is given by the fact that we are considering the regime $n \gg V^{3/2}\sqrt{\log V}$. We can then apply theorem D.1 and eq. (D.18). □

D.7 Sampling argument using the homogeneous model support graph

We want here to determine for which parameter regimes the kernel K_0 can have a nonzero radius at all. Our investigation identifies a sparser regime in which the spectral radius equals zero with probability one; therefore, by construction, concentration around the mean matrix cannot occur.

In the homogeneous setting, let G_{K_0} denote the directed support graph associated with the transition kernel K_0, where an edge $i \to j$ is present whenever $M_{ij} > 0$. Since M_{ij} is the minimum of two independent $\mathrm{Pois}(\lambda)$ variables, the edges are independent each with probability $q_\lambda := \mathbb{P}(M_{ij} > 0) = (1 - e^{-\lambda})^2$. Calling D_i^+ denotes the out-degree of vertex i, then

$$\mathbb{E}D_i^+ = \sum_{j=1}^{V} \mathbb{P}(M_{ij} > 0) = Vq_\lambda =: s_{V,n}. \tag{D.42}$$

Here then $s_{V,n}$ can be interpreted as the expected number of positive transitions in a given row of M.

A directed cycle exists if and only if $\rho(K_0) > 0$: if there is no cycle, relabeling makes K_0 strictly upper triangular and hence nilpotent, whereas if a cycle exists then for some i, k we have $(K_0^k)_{ii} > 0$, implying $\rho(K_0) > 0$.

We want then to estimate the probability of the existence of a directed cycle. Call N_k the number of simple directed cycles of length k, where $k = 1$ denotes self-loops. There is a number $(V)_k/k = V(V-1)\dots(V-k+1)/k$ of possible directed k-cycles where the weight $1/k$ enters because the cycle can start at any vertex. Each exists with probability q_λ^k since we consider the edges distinct and they occur independently. Hence

$$\mathbb{E}N_k = \frac{(V)_k}{k} q_\lambda^k \leq \frac{(Vq_\lambda)^k}{k} = \frac{s_{V,n}^k}{k}. \tag{D.43}$$

Rewriting the total number of cycles as $N_{\mathrm{cyc}} = \sum_{k=1}^{V} N_k$ and exploiting Markov's inequality, we obtain for $s_{V,n} < 1$

$$\mathbb{P}\big(\rho(K_0) > 0\big) = \mathbb{P}(N_{\mathrm{cyc}} \geq 1) \leq \mathbb{E}N_{\mathrm{cyc}} \leq \sum_{k=1}^{V} \frac{s_{V,n}^k}{k} \leq -\log(1 - s_{V,n}). \tag{D.44}$$

For $\lambda \to 0$, $s_{V,n} \sim V\mu(\lambda) = mh(\lambda) \sim n^2/V^3$. Hence, when $V \ll n \ll V^{3/2}$, $\rho(K_0) = 0$ with probability tending to one, although $\rho(A_0) > 0$. We observe then that the mean spectral radius fails to predict the typical logarithmic survival rate in this regime.

This argument identifies $n \asymp V^{3/2}$ as the natural scale where the expected out-degree D_i^+ in the support graph becomes of order one. In comparison, the Bernstein-based argument above provides a stronger sufficient requirement, namely $n/V^{3/2} \gg \sqrt{\log V}$, to ensure simultaneous concentration of fluctuations over all rows.

E Derivation of Claim 4.1

Disclaimer – Except for the statements labelled as propositions, lemmas or theorems, which are proved or cited with their hypotheses, this section is written at a heuristic level of rigour, with some assumptions detailed throughout, and the derivation of section E.4 is a detailed sketch. The main objective of this section is to provide a heuristic explanation for Claim 4.1, and to sketch the ideas towards a rigorous proof, which is left to a future and extended version of this work.

E.1 THE SIZE-BIASED MODEL, AND WHY IT DESCRIBES THE MARKOV CHAIN

E.1.1 THE MODELS

The chain and the question. Let V be even, $u \in \{-1,1\}^V$ with $\sum_i u_i = 0$, and $0 \le \beta < 1$. Recall that we consider the Markov chain on $\mathcal{V} = \{1,\dots,V\}$ with transition matrix

$$P^\star_{ij} \coloneqq \frac{1-\beta u_i u_j}{V}. \tag{E.1}$$

$P^\star$ is doubly stochastic (its rows and its columns all sum to one), so the uniform law is stationary. We observe a trajectory $y_0,\dots,y_n$ started from the uniform law on $\mathcal{V}$, with $n = \alpha V$ transitions, $\alpha > 0$ fixed, and form

$$M_{ij} \coloneqq \#\{t < n :\ (y_t, y_{t+1}) = (i,j)\}, \qquad m_i \coloneqq \sum_j M_{ij}, \qquad K \coloneqq D^{-1}M, \quad D \coloneqq \mathrm{Diag}(m), \tag{E.2}$$

rows with $m_i = 0$ being set to zero: M_{ij} counts the transitions $i \to j$, m_i the visits to i, and K is the empirical transition matrix. In the sparse regime $n = \alpha V$, each state is typically visited $O(1)$ times. We note that typically K has a trivial eigenvalue at 1 which carries no information on u. Let $z_\star$ be the eigenvalue of K closest to $-\beta$ and $u_\star$ an associated right eigenvector, normalised so that $\|u_\star\| = \sqrt{V} = \|u\|$. We ask whether $z_\star$ is an outlier, isolated from the bulk of the spectrum and close to $-\beta$, and how much of u its eigenvector retains as $V \to \infty$:

$$q_{\mathrm{R}} \coloneqq \frac{|\langle u_\star, u\rangle|}{V} \in [0,1]. \tag{E.3}$$

Both change as β crosses a threshold depending on α, a so-called "BBP" transition, see e.g. Baik et al. (2005); Benaych-Georges & Nadakuditi (2011) for references in the context of Hermitian matrices. Throughout, α and β are fixed and $V \to \infty$.

The size-biased model. In the Markov chain, the entries of K are built from a single path and are all correlated. In this part, we study instead the following model of K (that we introduced in section 4.2 and recall below for completeness), which keeps the local statistics of the path once the visit counts are known. Its justification follows the definition.

Model E.1 (size-biased model)**.** Fix $\alpha > 0$, $\beta \in [0,1)$ and u as above.

(1) Draw $m_1,\dots,m_V$ i.i.d. $\mathrm{Pois}(\alpha)$, and set $n \coloneqq \sum_i m_i$ and $\mathcal{S} \coloneqq \{i : m_i > 0\}$.

(2) Given m, define the *size-biased* law of a transition out of i,

$$q_{ij} \coloneqq \frac{m_j\,(1-\beta u_i u_j)}{\sum_l m_l\,(1-\beta u_i u_l)}, \qquad j \in \mathcal{V}, \tag{E.4}$$

and draw the rows of $\widetilde{M}$ independently, row i being $\mathrm{Mult}(m_i, q_{i\cdot})$: the m_i transitions out of i land on i.i.d. destinations $J_{i1},\dots,J_{im_i}$ of law $q_{i\cdot}$, and $\widetilde{M}_{ij} \coloneqq \#\{k : J_{ik} = j\}$.

(3) Set $\widetilde{K} \coloneqq D^{-1}\widetilde{M}$, rows with $m_i = 0$ being zero.

In Model E.1, a transition thus lands on a state in proportion to its number of visits, tilted by $P^\star$, and never on an unvisited state. We write $a \coloneqq \mathbb{1}_{\mathcal{S}}$, $b \coloneqq u \circ a$, $s \coloneqq a/m$ entrywise ($s_i = 1/m_i$ on $\mathcal{S}$, 0 elsewhere), $y^{\circ 2} \coloneqq y \circ y$, and $\mathbf{1}$ for the all-ones vector. In Model E.1, $n = \sum_i m_i$ is random, with $n/V \to \alpha$. By the law of large numbers, with $m \sim \mathrm{Pois}(\alpha)$:

$$\begin{gathered}\frac{n}{V} \to \alpha, \qquad \frac{|\mathcal{S}|}{V} \to 1 - e^{-\alpha}, \qquad \rho_V^2 \coloneqq \frac{|\mathcal{S}|}{n} \to \rho_\alpha^2 \coloneqq \frac{1-e^{-\alpha}}{\alpha}, \\ \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \frac{1}{m_i} \to h(\alpha) \coloneqq \mathbb{E}\Big[\frac{1}{m}\ \Big|\ m \ge 1\Big].\end{gathered} \tag{E.5}$$

E.1.2 WHY THIS IS A GOOD MODEL FOR THE CHAIN.

Three facts about the chain of eq. (E.1) motivate Model E.1.

(i) The visit counts are Poisson. Since $\mathbf{1}\mathbf{1}^\top$ and $uu^\top$ commute, are orthogonal ($\mathbf{1}^\top u = 0$) and satisfy $(\mathbf{1}\mathbf{1}^\top)^2 = V\mathbf{1}\mathbf{1}^\top$, $(uu^\top)^2 = Vuu^\top$, one has $(P^\star)^s = \frac{1}{V}(\mathbf{1}\mathbf{1}^\top + (-\beta)^s uu^\top)$ for $s \geq 1$. By stationarity and the Markov property, for times $t_1 < \cdots < t_k$ and states $w_1, \dots, w_k$,

$$\mathbb{P}\big(y_{t_1} = w_1, \dots, y_{t_k} = w_k\big) = \frac{1}{V^k} \prod_{l=1}^{k-1} \Big(1 + (-\beta)^{t_{l+1}-t_l}\, u_{w_l} u_{w_{l+1}}\Big). \tag{E.6}$$

From this, we can deduce the following.

Proposition E.2 (Poisson counts)**.** *Let $(m)_r \coloneqq m(m-1)\cdots(m-r+1)$. Fix $k \geq 1$, integers $r_1, \dots, r_k \geq 1$, and set $R \coloneqq \sum_l r_l$. There is a constant C_R, depending on R, α, β only, such that for all V and all distinct states $i_1, \dots, i_k$,*

$$\Big| \mathbb{E}\Big[\prod_{l=1}^{k} (m_{i_l})_{r_l}\Big] - \alpha^R \Big| \leq \frac{C_R}{V}. \tag{E.7}$$

Consequently: (a) $(m_{i_1}, \dots, m_{i_k})$ converges in law to k independent $\mathrm{Pois}(\alpha)$ variables; (b) $\mathrm{Var}(m_i) = \alpha + O(V^{-1})$ and $\mathrm{Cov}(m_i, m_j) = O(V^{-1})$ for $i \neq j$; (c) for every $f : \mathbb{N} \to \mathbb{R}$ of polynomial growth, $\frac{1}{V}\sum_{i=1}^{V} f(m_i) \to \mathbb{E}[f(X)]$, $X \sim \mathrm{Pois}(\alpha)$, in L^2 and in probability.

Proof of Proposition E.2. Since m_i is the number of times $t \in \{0, \dots, n-1\}$ with $y_t = i$, the falling factorial $(m_i)_r$ is the number of r-tuples $(t^{(1)}, \dots, t^{(r)})$ of pairwise distinct times with $y_{t^{(1)}} = \cdots = y_{t^{(r)}} = i$, that is

$$(m_i)_r = \sum_{t^{(1)}, \dots, t^{(r)} \text{ distinct}} \prod_{q=1}^{r} \mathbb{1}\{y_{t^{(q)}} = i\}.$$

Multiplying these identities over $l = 1, \dots, k$ and taking expectations,

$$\mathbb{E}\Big[\prod_{l=1}^{k} (m_{i_l})_{r_l}\Big] = \sum_{\mathcal{T}} \mathbb{P}\big(y_t = i_l \text{ for all } l \text{ and all } t \in T_l\big), \tag{E.8}$$

where the sum runs over the families $\mathcal{T} = (T_1, \dots, T_k)$ in which T_l is an r_l-tuple of pairwise distinct times. If two tuples T_l and $T_{l'}$, $l \neq l'$, share a time t, the event requires $y_t = i_l$ and $y_t = i_{l'}$ with $i_l \neq i_{l'}$, and its probability is 0. Only the families whose R times are pairwise distinct remain, and such a family is described without loss by three things: the set of its times, written in increasing order as $t_1 < \cdots < t_R$; the labelling $\sigma : \{1, \dots, R\} \to \{1, \dots, k\}$, where $\sigma(j) = l$ if t_j belongs to T_l, which takes each value l exactly r_l times; and, for each l, the order in which the r_l times with label l appear in the tuple T_l. The probability in eq. (E.8) depends on the set of times and on σ only, and there are $\prod_l r_l!$ choices of the orders. Finally, eq. (E.6) with $w_j = i_{\sigma(j)}$ gives the probability that the chain is at $i_{\sigma(1)}, \dots, i_{\sigma(R)}$ at times $t_1 < \cdots < t_R$. Altogether,

$$\begin{gathered}\mathbb{E}\Big[\prod_{l=1}^{k} (m_{i_l})_{r_l}\Big] = \prod_{l=1}^{k} r_l! \sum_{\sigma} \sum_{t_1 < \cdots < t_R} \frac{1}{V^R} \prod_{j=1}^{R-1} \Big(1 + \chi_j (-\beta)^{t_{j+1}-t_j}\Big), \\ \chi_j \coloneqq u_{i_{\sigma(j)}} u_{i_{\sigma(j+1)}} \in \{-1, 1\}.\end{gathered} \tag{E.9}$$

Expand the product over j into 2^{R-1} terms indexed by the subsets $A \subset \{1, \dots, R-1\}$ of retained factors $\chi_j(-\beta)^{t_{j+1}-t_j}$. The term $A = \emptyset$ equals V^{-R} whatever the times and the labelling. There are $\binom{n}{R}$ sets of times and $R!/\prod_l r_l!$ labellings, so the terms $A = \emptyset$ sum to

$$\prod_l r_l! \cdot \frac{R!}{\prod_l r_l!} \cdot \binom{n}{R} \cdot \frac{1}{V^R} = \frac{(n)_R}{V^R} = \alpha^R + O_R\big(V^{-1}\big).$$

We now tackle the other terms. Fix σ and A with $|A| = p \geq 1$. The corresponding term is bounded in absolute value by $V^{-R} \prod_{j \in A} \beta^{t_{j+1}-t_j}$. Parametrise the set of times by t_1 and the gaps

$\tau_j \coloneqq t_{j+1} - t_j \geq 1$, all of which lie in $\{0, \dots, n\}$, and sum over them: each $j \in A$ contributes $\sum_{\tau \geq 1} \beta^\tau \leq \beta/(1-\beta)$, and t_1 and each of the $R-1-p$ other gaps contribute at most n. The term is therefore at most $V^{-R} n^{R-p} (\beta/(1-\beta))^p = \alpha^{R-p} (\beta/(1-\beta))^p V^{-p} \leq C\, V^{-1}$. All these terms together contribute $O_R(V^{-1})$ to eq. (E.9), uniformly in the states. This proves eq. (E.7).

(b) is then the case $R = 2$: $\mathbb{E}[m_i(m_i - 1)] = \alpha^2 + O(V^{-1})$ and $\mathbb{E}[m_i m_j] = \alpha^2 + O(V^{-1})$ for $i \neq j$, while $\mathbb{E} m_i = n/V = \alpha$ exactly since the chain is stationary.

(a) Let $X_1, \dots, X_k$ be independent $\mathrm{Pois}(\alpha)$ variables, whose mixed factorial moments are $\mathbb{E} \prod_l (X_l)_{r_l} = \alpha^R$. Ordinary moments are finite linear combinations of factorial moments. Hence, by eq. (E.7), every mixed moment of $(m_{i_1}, \dots, m_{i_k})$ converges to the corresponding moment of $(X_1, \dots, X_k)$, and so does every moment of the linear combination $\sum_l \lambda_l m_{i_l}$ to that of $Y_\lambda \coloneqq \sum_l \lambda_l X_l$, for each $\lambda \in \mathbb{R}^k$. The variable Y_λ has a finite moment generating function on all of $\mathbb{R}$, so its law is determined by its moments, and a classical use of the method of moments gives $\sum_l \lambda_l m_{i_l} \to Y_\lambda$ in law, see e.g (Billingsley, 1995, Section 30). The Cramér–Wold device (see e.g. (Billingsley, 1995, Theorem 29.4)) then gives (a).

(c) Let $|f(x)| \leq C(1 + x^p)$, $\bar{f} \coloneqq \mathbb{E} f(X)$, and $g(x, y) \coloneqq (f(x) - \bar{f})(f(y) - \bar{f})$. Then

$$\mathbb{E}\Big[\Big(\frac{1}{V} \sum_{i=1}^{V} f(m_i) - \bar{f}\Big)^2\Big] = \frac{1}{V^2} \sum_{i,j=1}^{V} \mathbb{E}\big[g(m_i, m_j)\big]. \tag{E.10}$$

We use two facts. First, a moment bound: by eq. (E.7) and the previous paragraph, for each fixed q the moments $\mathbb{E}[m_i^q]$ and $\mathbb{E}[m_i^q m_j^q]$ are bounded uniformly in V, i and j, hence so is $\mathbb{E}[g(m_i, m_j)^2]$, by a constant C_1 (this includes $i = j$). Second, a symmetry: since $P^\star_{ij}$ depends on (i, j) only through $u_i u_j$ and the initial law is uniform, the law of the trajectory is invariant under any permutation π of the states with $u_{\pi(i)} u_{\pi(j)} = u_i u_j$ for all i, j, that is any permutation preserving u or exchanging the two classes. Given two pairs $i \neq j$ and $i' \neq j'$ with $u_i u_j = u_{i'} u_{j'}$, such a π with $\pi(i) = i'$ and $\pi(j) = j'$ exists, so (m_i, m_j) and $(m_{i'}, m_{j'})$ have the same law. Hence $\mathbb{E}[g(m_i, m_j)]$ takes at most two values on the pairs $i \neq j$, one on pairs in the same class, c_V^+, and one on pairs in different classes, c_V^-. Bounding the V diagonal terms of eq. (E.10) by $\mathbb{E}[g(m_i, m_i)] \leq C_1^{1/2}$,

$$\mathbb{E}\Big[\Big(\frac{1}{V} \sum_{i=1}^{V} f(m_i) - \bar{f}\Big)^2\Big] \leq \frac{C_1^{1/2}}{V} + \max\big(|c_V^+|, |c_V^-|\big),$$

and it remains to show that $c_V^\pm \to 0$, that is, for a fixed pair $i \neq j$, $\mathbb{E}[g(m_i, m_j)] \to \mathbb{E}[g(X_1, X_2)] = 0$, where X_1, X_2 are independent $\mathrm{Pois}(\alpha)$ and the last equality holds because $\mathbb{E}[f(X_1) - \bar{f}] = 0$. Fix $L \geq 1$ and split $g = g \mathbb{1}_{B_L} + g \mathbb{1}_{B_L^c}$ with $B_L = \{(x, y) : x \leq L,\ y \leq L\}$. On B_L, $\mathbb{E}[g(m_i, m_j) \mathbb{1}_{B_L}(m_i, m_j)] = \sum_{x,y \leq L} g(x, y) \mathbb{P}(m_i = x, m_j = y)$ is a finite sum, and each probability converges to $\mathbb{P}(X_1 = x, X_2 = y)$ by (a) with $k = 2$ (for integer-valued vectors, convergence in law is the convergence of these point masses). So $\mathbb{E}[g(m_i, m_j) \mathbb{1}_{B_L}(m_i, m_j)]$ converges to $\mathbb{E}[g(X_1, X_2) \mathbb{1}_{B_L}(X_1, X_2)]$. Moreover, by the Cauchy–Schwarz and Markov inequalities and $\mathbb{E} m_i = \alpha$,

$$\big|\mathbb{E}[g(m_i, m_j) \mathbb{1}_{B_L^c}(m_i, m_j)]\big| \leq C_1^{1/2}\, \mathbb{P}(m_i > L \text{ or } m_j > L)^{1/2} \leq C_1^{1/2} \Big(\frac{2\alpha}{L}\Big)^{1/2},$$

and the same bound holds with (X_1, X_2) in place of (m_i, m_j). Therefore $\limsup_{V \to \infty} |\mathbb{E}[g(m_i, m_j)] - \mathbb{E}[g(X_1, X_2)]| \leq 2 C_1^{1/2} (2\alpha/L)^{1/2}$ for every L, which gives

$$\limsup_{V \to \infty} |\mathbb{E}[g(m_i, m_j)] - \mathbb{E}[g(X_1, X_2)]| = 0.$$

This gives the L^2 convergence, hence the convergence in probability. □

(ii) Size-biasing is the law of a transition given the counts. The law of the count matrix M is known exactly: given the number n of transitions and the endpoints y_0, y_n (Whittle, 1955, eq. (8)),

$$\mathbb{P}(M \mid y_0, y_n) = T(M) \prod_i \frac{m_i!}{\prod_j M_{ij}!} \prod_{i,j} \big(P^\star_{ij}\big)^{M_{ij}} \qquad \text{if} \quad m_j - \sum_i M_{ij} = \mathbb{1}_{j = y_0} - \mathbb{1}_{j = y_n} \ \text{ for all } j, \tag{E.11}$$

and 0 otherwise, where $T(M)$ is the (y_n, y_0) cofactor of $I - D^{-1}M$. The factor $T(M)$ is global: it expresses that the transitions of M can be ordered into a single path from y_0 to y_n (similar to a connectivity constraint in a sparse random graph), and we drop it in building our surrogate model. What remains is the law of independent rows $\mathrm{Mult}(m_i, P^\star_{i\cdot})$, restricted to the matrices whose column sums equal their row sums up to the endpoints: every visit to j is an arrival at j followed by a departure from j, so j receives exactly m_j transitions. This restriction is what makes a transition land preferentially on much-visited states. Given the counts m, a transition out of i must land on one of the n arrival slots, m_j of them at state j, and $P^\star$ weights a slot at j by $1 - \beta u_i u_j$. The probability of landing at j is therefore proportional to $m_j(1 - \beta u_i u_j)$, which is eq. (E.4). We emphasise that $P^\star_{i\cdot}$ is the law of the next state before the trajectory is seen, while $q_{i\cdot}$ is its law *conditional on the counts* $(m_i)_{i\in\mathcal{V}}$. Step (2) of Model E.1 keeps this conditional law of every transition.

(iii) Independent rows. In the Markov chain, two transitions are not exactly independent given m, since they cannot land on the same arrival slot. However, this dependence is weak. To explicit it, notice that at $\beta = 0$, two departures from i and i' land on (j, j') with probability $m_j(m_{j'} - \mathbb{1}_{j=j'})/(n(n-1))$, which is $q_{ij}\, q_{i'j'}\, n/(n-1)$ for $j \neq j'$, the first departure having removed one slot among n, while the event $j = j'$ has probability $O(1/V)$. Adding the "tilt" caused by the signal u does not change this picture: in general, any fixed number of transitions are asymptotically independent. Stated otherwise, the path followed by the chain is locally a tree: with probability $1 - O(1/V)$, no state is reached twice in a bounded number of steps. Step (2) of Model E.1 replaces this near-independence by exact independence, which makes the rows of $\widetilde{M}$ independent, allowing for some simplifications in the analytical treatment.

E.2 First two moments of $\widetilde{K}$ given m

Leading order. The denominator of eq. (E.4) is $n(1 - \beta u_i \bar{u}_m)$ with $\bar{u}_m \coloneqq \frac{1}{n}\sum_l m_l u_l$. Since the counts are independent of u and the classes are balanced, $\bar{u}_m$ and $\frac{1}{n}\sum_{j\in\mathcal{S}} u_j$ are typically of order $\mathcal{O}(V^{-1/2})$. As we will consider statements as $V \to \infty$, from now on, we set them to zero and keep the notation q_{ij} for the leading-order

$$q_{ij} \coloneqq \frac{m_j}{n}\,(1 - \beta u_i u_j), \tag{E.12}$$

Recall that $\widetilde{K}_{ij} = \frac{1}{m_i}\sum_{k=1}^{m_i} \mathbb{1}_{J_{ik}=j}$ for $i \in \mathcal{S}$, with J_{ik} i.i.d. of law $q_{i\cdot}$ and independent across i.

Lemma E.3 (mean)**.** *$\overline{K} \coloneqq \mathbb{E}[\widetilde{K} \mid m] = (a_i q_{ij})_{ij}$, that is*

$$\overline{K} = \frac{1}{n}\Big(a\, m^\top - \beta\, b\,(u \circ m)^\top\Big) = AB^\top, \qquad A \coloneqq \frac{1}{n}\,[\,a,\ -\beta b\,], \qquad B \coloneqq [\,m,\ u \circ m\,], \tag{E.13}$$

a matrix of rank two with $\overline{K}a = a$ and $\overline{K}b = -\beta b$.

Proof of Lemma E.3. $\mathbb{E}\mathbb{1}_{J_{ik}=j} = q_{ij}$ and eq. (E.12) give eq. (E.13), since $a_i u_i = b_i$. Then $(\overline{K}a)_i = a_i \sum_j q_{ij} = a_i$ and $(\overline{K}b)_i = a_i \sum_j q_{ij} u_j a_j = a_i\big(\bar{u}_m - \beta u_i \frac{1}{n}\sum_j m_j\big) = -\beta b_i$. □

Lemma E.4 (covariance)**.** *Let $\widetilde{W} \coloneqq \widetilde{K} - \overline{K}$. The rows of $\widetilde{W}$ are independent and centred, and for $i \in \mathcal{S}$*

$$\mathrm{Cov}\big(\widetilde{K}_{ij}, \widetilde{K}_{ik} \,\big|\, m\big) = \frac{(C_i)_{jk}}{m_i}, \qquad C_i \coloneqq \mathrm{Diag}(q_{i\cdot}) - q_{i\cdot} q_{i\cdot}^\top, \tag{E.14}$$

i.e. $\mathbb{E}[\widetilde{W}_{ij}^2 \mid m] = q_{ij}(1 - q_{ij})/m_i$ and $\mathbb{E}[\widetilde{W}_{ij}\widetilde{W}_{ik} \mid m] = -q_{ij}q_{ik}/m_i$ for $j \neq k$. Equivalently, for every vector y not depending on the transitions,

$$(\widetilde{W}y)_i = \frac{1}{m_i}\sum_{k=1}^{m_i}\big(y_{J_{ik}} - q_{i\cdot}^\top y\big), \qquad \mathbb{E}\big[(\widetilde{W}y)_i^2 \,\big|\, m\big] = \frac{y^\top C_i y}{m_i} = \frac{\mathrm{Var}_{q_i}(y)}{m_i}, \tag{E.15}$$

where $\mathrm{Var}_{q_i}(y) \coloneqq \sum_j q_{ij} y_j^2 - (\sum_j q_{ij} y_j)^2$ is the variance of y at a random destination of a transition out of i.

Proof of Lemma E.4. Row i of $\widetilde{K}$ is a function of the destinations $J_{i1}, \dots, J_{im_i}$ alone, and these families are independent across i, so the rows of $\widetilde{W}$ are independent. They are centred by Lemma E.3. For $i \in \mathcal{S}$, write $\widetilde{K}_{ij} = \frac{1}{m_i} \sum_{l=1}^{m_i} \mathbb{1}_{J_{il}=j}$ and expand the covariance,

$$\mathrm{Cov}\big(\widetilde{K}_{ij}, \widetilde{K}_{ik} \mid m\big) = \frac{1}{m_i^2} \sum_{l,l'=1}^{m_i} \mathrm{Cov}\big(\mathbb{1}_{J_{il}=j}, \mathbb{1}_{J_{il'}=k} \mid m\big).$$

For $l \neq l'$ the two indicators are independent and the covariance vanishes. For $l = l'$, since $\mathbb{1}_{J_{il}=j} \mathbb{1}_{J_{il}=k} = \mathbb{1}_{j=k} \mathbb{1}_{J_{il}=j}$,

$$\mathrm{Cov}\big(\mathbb{1}_{J_{il}=j}, \mathbb{1}_{J_{il}=k} \mid m\big) = \mathbb{E}\big[\mathbb{1}_{J_{il}=j} \mathbb{1}_{J_{il}=k} \mid m\big] - q_{ij} q_{ik} = q_{ij} \mathbb{1}_{j=k} - q_{ij} q_{ik} = (C_i)_{jk}.$$

The m_i diagonal terms $l = l'$ are equal, which gives eq. (E.14). For eq. (E.15), use $\overline{K}_{ij} = q_{ij}$ for $i \in \mathcal{S}$ and $\sum_j \mathbb{1}_{J_{il}=j} y_j = y_{J_{il}}$:

$$(\widetilde{W} y)_i = \sum_j \big(\widetilde{K}_{ij} - q_{ij}\big) y_j = \frac{1}{m_i} \sum_{l=1}^{m_i} \Big(\sum_j \mathbb{1}_{J_{il}=j} y_j - \sum_j q_{ij} y_j \Big) = \frac{1}{m_i} \sum_{l=1}^{m_i} \big(y_{J_{il}} - q_{i\cdot}^\top y \big).$$

The m_i summands are i.i.d. and centred, since $\mathbb{E}[y_{J_{il}} \mid m] = \sum_j q_{ij} y_j = q_{i\cdot}^\top y$, with variance $\mathbb{E}[y_{J_{il}}^2 \mid m] - (q_{i\cdot}^\top y)^2 = \sum_j q_{ij} y_j^2 - (\sum_j q_{ij} y_j)^2 = \mathrm{Var}_{q_i}(y)$. The variance of their average is therefore $\mathrm{Var}_{q_i}(y)/m_i$. Finally $y^\top C_i y = \sum_j q_{ij} y_j^2 - (q_{i\cdot}^\top y)^2 = \mathrm{Var}_{q_i}(y)$ by the definition of C_i, which gives the two expressions in eq. (E.15). □

Define the *variance profile*

$$\Sigma_{ij} \coloneqq \frac{a_i q_{ij}}{m_i} = \frac{1}{n} \Big(s\, m^\top - \beta\, (u \circ s)\, (u \circ m)^\top \Big)_{ij}, \tag{E.16}$$

Since $\sum_j q_{ij} y_j = (\overline{K} y)_i$, eq. (E.15) reads, entrywise,

$$\mathbb{E}\big[(\widetilde{W} y)^{\circ 2} \mid m\big] = \Sigma\, y^{\circ 2} - \frac{(\overline{K} y)^{\circ 2}}{m}. \tag{E.17}$$

The first term is what independent entries of variance Σ_{ij} would give. The second, due to the correlations within a row, matters only when $\overline{K} y$ is of order one: for instance, for the signal $y = b$, $\overline{K} b = -\beta b$, $(\overline{K} b)^{\circ 2}/m = \beta^2 s$ and $\Sigma b^{\circ 2} = \Sigma a = s$ (the rows of Σ sum to s_i), so

$$\mathbb{E}\big[(\widetilde{W} b)^{\circ 2} \mid m\big] = (1 - \beta^2)\, s. \tag{E.18}$$

Informally, eq. (E.18) expresses that the tilt in the direction of u makes the class of the next state partially predictable, which reduces the variance of a transition along the signal from 1 to $1 - \beta^2$. We now characterise the spectrum of Σ. Throughout, $\rho(A)$ denotes the spectral radius of a square matrix A, the largest modulus of its eigenvalues. We also use the leading-order convention $\sum_{j \in \mathcal{S}} u_j = 0$ (otherwise one would need to correct the following statements by $o(1)$ quantities as $V \to \infty$).

Lemma E.5. *Σ has rank two, with $\Sigma a = s$, $\Sigma s = \rho_V^2\, s$, $m^\top \Sigma = \rho_V^2\, m^\top$, and second eigenvalue $-\beta \rho_V^2$ (right eigenvector $u \circ s$). Hence the spectral radius of Σ is $\rho(\Sigma) = \rho_V^2 = |\mathcal{S}|/n$.*

Proof of Lemma E.5. One computes easily:

$$(\Sigma a)_i = \frac{1}{m_i} \sum_j q_{ij} = s_i, \qquad (\Sigma s)_i = \frac{s_i}{n} \sum_{j \in \mathcal{S}} (1 - \beta u_i u_j) = s_i\, \frac{|\mathcal{S}|}{n} = \rho_V^2 s_i,$$

$$(m^\top \Sigma)_j = \frac{m_j}{n} \sum_{i \in \mathcal{S}} (1 - \beta u_i u_j) = \rho_V^2 m_j, \quad \Sigma(u \circ s) = \frac{1}{n} \Big[s \sum_{j \in \mathcal{S}} u_j - \beta\, (u \circ s)\, |\mathcal{S}| \Big] = -\beta \rho_V^2\, (u \circ s).$$

□

E.3 THE GAUSSIAN MODEL

We now introduce our surrogate Gaussian model for $\widetilde{K}$.

Model E.6 (Gaussian model)**.** Given m, let $X \coloneqq \overline{K} + W$ with $\overline{K} = AB^\top$ as in eq. (E.13) and W a centred Gaussian matrix with independent rows, zero rows off $\mathcal{S}$, and for $i \in \mathcal{S}$

$$\mathbb{E}\big[W_{ij}W_{ik}\big] = \frac{q_{ij}\mathbb{1}_{j=k} - q_{ij}q_{ik}}{m_i} = \frac{(C_i)_{jk}}{m_i}, \qquad q_{ij} = \frac{m_j\,(1-\beta u_i u_j)}{n}. \tag{E.19}$$

These are the second moments of $\widetilde{W}$ in Lemma E.4: the Gaussian model has the same first two conditional moments as the size-biased model, and the Gaussian model is the one we will use as a heuristic to characterise the spectrum of $\widetilde{W}$ and the "BBP" transition in $\widetilde{K}$. We will leverage some properties of the Gaussian distribution, as described in the following lemma.

Lemma E.7 (three descriptions of W)**.** *The following define the same law for W, with $g_i \sim \mathcal{N}(0, I_V)$ i.i.d. over i and W^0 a matrix with independent entries $W^0_{ij} \sim \mathcal{N}(0, \Sigma_{ij})$.*

(i) $W_{i\cdot} \sim \mathcal{N}(0, C_i/m_i)$ *independently over* $i \in \mathcal{S}$.

(ii) $W_{i\cdot} = m_i^{-1/2}\,\mathrm{Diag}(\sqrt{q_{i\cdot}})\big(I - \sqrt{q_{i\cdot}}\sqrt{q_{i\cdot}}^{\,\top}\big)g_i$*: a standard Gaussian vector projected orthogonally to* $\sqrt{q_{i\cdot}}$*, then multiplied entrywise by* $\sqrt{q_{i\cdot}/m_i}$.

(iii) $W = W^0 - \mathrm{Diag}(W^0\mathbf{1})\,\overline{K}$*, i.e.* $W_{ij} = W^0_{ij} - (W^0\mathbf{1})_i q_{ij}$*; equivalently, W is W^0 conditioned on* $W^0\mathbf{1} = 0$.

In particular $W\mathbf{1} = 0$*: the rows of X sum to a, as those of $\widetilde{K}$ do.*

Proof of Lemma E.7. In (ii) and (iii), row i of W is a linear function of g_i, respectively of row i of W^0, so the rows are independent centred Gaussian vectors. The law of a centred Gaussian vector is determined by its covariance, so it suffices to check that the covariance of row i is C_i/m_i in both cases.

(ii). If $g \sim \mathcal{N}(0, I_V)$ and A is a matrix, Ag is centred Gaussian with covariance $AA^\top$. Here $A = m_i^{-1/2}\mathrm{Diag}(p)(I - pp^\top)$ with $p \coloneqq \sqrt{q_{i\cdot}}$ entrywise. Since $\|p\|^2 = \sum_j q_{ij} = 1$:

$$AA^\top = \frac{1}{m_i}\mathrm{Diag}(p)(I-pp^\top)^2\mathrm{Diag}(p) = \frac{1}{m_i}\mathrm{Diag}(p)(I-pp^\top)\mathrm{Diag}(p) = \frac{\mathrm{Diag}(q_{i\cdot}) - q_{i\cdot}q_{i\cdot}^\top}{m_i} = \frac{C_i}{m_i}.$$

(iii). Write $\xi_i \coloneqq (W^0\mathbf{1})_i = \sum_l W^0_{il}$, so that $W_{ij} = W^0_{ij} - q_{ij}\xi_i$. For $i \in \mathcal{S}$, $\Sigma_{il} = q_{il}/m_i$, hence $\sum_l \Sigma_{il} = 1/m_i$. Using $\mathbb{E}[W^0_{ij}W^0_{ik}] = \Sigma_{ij}\mathbb{1}_{j=k}$ and expanding the product,

$$\begin{aligned}\mathbb{E}[W_{ij}W_{ik}] &= \mathbb{E}[W^0_{ij}W^0_{ik}] - q_{ik}\sum_l \mathbb{E}[W^0_{ij}W^0_{il}] - q_{ij}\sum_l \mathbb{E}[W^0_{il}W^0_{ik}] + q_{ij}q_{ik}\sum_{l,l'}\mathbb{E}[W^0_{il}W^0_{il'}] \\ &= \Sigma_{ij}\mathbb{1}_{j=k} - q_{ik}\Sigma_{ij} - q_{ij}\Sigma_{ik} + q_{ij}q_{ik}\sum_l \Sigma_{il} = \frac{(C_i)_{jk}}{m_i}.\end{aligned}$$

The conditioning statement follows from classical Gaussian conditioning. □

In order to characterise its spectrum, it will be useful to show that W^0 can be realised as an independent-elements Gaussian matrix with a simple variance profile.

Lemma E.8 (similarity to a flat profile)**.** *Let $D^{\pm 1/2} \coloneqq \mathrm{Diag}(m^{\pm 1/2})$ on $\mathcal{S}$, and let Y be the Gaussian matrix supported on $\mathcal{S}\times\mathcal{S}$ with independent entries*

$$Y_{ij} \coloneqq \frac{1}{\sqrt{n}}\sqrt{1-\beta u_i u_j}\;g_{ij}, \qquad g_{ij}\ \textit{i.i.d.}\ \mathcal{N}(0,1), \qquad i,j \in \mathcal{S}. \tag{E.20}$$

Let also $\mathcal{X} \coloneqq \frac{1}{n}\big[\sqrt{m}\sqrt{m}^\top - \beta(u\circ\sqrt{m})(u\circ\sqrt{m})^\top\big]$, a symmetric matrix of rank two. Then:

(i) $W^0 \overset{\mathrm{d}}{=} D^{-1/2} Y D^{1/2}$*, where recall that* $\overset{\mathrm{d}}{=}$ *means equality in law;*

(ii) $\overline{K} = D^{-1/2} \mathcal{X} D^{1/2}$.

(iii) if W^0 *is realised as* $W^0 = D^{-1/2} Y D^{1/2}$ *and* $W = W^0 - \mathrm{Diag}(W^0 \mathbf{1}) \overline{K}$ *as in Lemma E.7(iii), then* $X = \overline{K} + W$ *satisfies the identity*

$$X = D^{-1/2} \Big[\big(I - \mathrm{Diag}(\xi) \big) \mathcal{X} + Y \Big] D^{1/2}, \qquad \xi := W^0 \mathbf{1} = D^{-1/2} Y D^{1/2} \mathbf{1}. \quad \text{(E.21)}$$

In particular, the right-hand side of eq. (E.21) *has the law of* X *in Model E.6.*

Proof of Lemma E.8. (i). Since $D^{\pm 1/2}$ are diagonal, for $i, j \in \mathcal{S}$,

$$(D^{-1/2} Y D^{1/2})_{ij} = m_i^{-1/2} \, Y_{ij} \, m_j^{1/2} = \sqrt{\frac{m_j}{n \, m_i}} \sqrt{1 - \beta u_i u_j} \, g_{ij},$$

and the entry vanishes if $i \notin \mathcal{S}$ or $j \notin \mathcal{S}$. Entry (i, j) is a function of g_{ij} alone, so the entries of $D^{-1/2} Y D^{1/2}$ are independent centred Gaussian variables, like those of W^0. Two such matrices have the same law if and only if their entries have the same variances, and here the variance of entry (i, j) is $\frac{m_j}{n m_i}(1 - \beta u_i u_j) = q_{ij}/m_i = \Sigma_{ij}$ by eq. (E.16). This proves (i). (ii) is trivial to establish by a direct computation. (iii) is also immediate by a direct computation (using that diagonal matrices commute). □

In words, the variance profile of Y is flat: $(1 - \beta)/n$ within a class and $(1 + \beta)/n$ across, with all row and column sums equal to $|\mathcal{S}|/n = \rho_V^2$. Moreover, Y and X share the same asymptotic spectrum by eq. (E.21) up to at most 2 outliers.

E.4 The BBP transition in the Gaussian model

E.4.1 Statement

We write here the main claim describing the "BBP" transition in the spectrum of the Gaussian model E.6.

Claim E.9 (derived in section E.4.3)**.** *In the Gaussian model, conditionally on* m*, as* $V \to \infty$*:*

(i) The spectrum of X *consists of the eigenvalue* 1 *(eigenvector* a*), a bulk filling the disc* $\{|z| \le \rho_\alpha\}$ *(plus* $|\mathcal{S}^c|$ *zeros), and, if and only if* $\beta > \rho_\alpha$*, an outlier* $z_\star = -\beta + o(1)$*.*

(ii) If $\beta > \rho_\alpha$*, the right eigenvector* $u_\star$ *of* $z_\star$ *satisfies, as* $V \to \infty$*,*

$$q_{\mathrm{R}}^2 := \frac{\langle u_\star, u \rangle^2}{V^2} \longrightarrow \frac{(1 - e^{-\alpha})(\beta^2 - \rho_\alpha^2)}{(\beta^2 - \rho_\alpha^2) + h(\alpha)(1 - \beta^2)} \quad \text{(E.22)}$$

If $\beta \le \rho_\alpha$ *there is no outlier near* $-\beta$*, and* $q_{\mathrm{R}} \to 0$ *for the eigenvector of the eigenvalue closest to* $-\beta$*.*

Generalisation to the size-biased model. The sketch of derivation presented below is done for the Gaussian model for simplicity of exposition. It uses the mean $\overline{K}$, the covariances C_i/m_i of the rows, the independence of the rows, and one approximation, which neglects the paths of the noise that pass twice through the same state (i.e. assuming that the path of the chain is locally treelike). The size-biased model has the same mean, the same row covariances (Lemma E.4) and independent rows, and consequently we expect Claim E.9 to hold for $\widetilde{K}$ as well, and even for the empirical counts of the Markov chain, as illustrated by our numerics, see fig. 4.

Related literature. The BBP transition for the outliers of finite-rank perturbations of large random matrices was first studied in physics Edwards & Jones (1976), then in the context of unsupervised learning Watkin & Nadal (1994), and takes its name from the study Baik et al. (2005) in probability and statistics. A description of its mechanism, with general theorems for Hermitian matrices, is in Benaych-Georges & Nadakuditi (2011), and we refer to the book Potters & Bouchaud (2020) for further references. The following account does not claim to be exhaustive: the literature on spiked

random matrices is vast, and we only point to some of the works closest to the present setting. For i.i.d. non-Hermitian noise, the outliers of a finite-rank perturbation converge to the eigenvalues of the perturbation lying outside the disc of the noise, without shift Tao (2013); Bordenave & Capitaine (2016), and the case of non-Hermitian noise invariant under unitary conjugation was studied recently in Bousseyroux & Potters (2026) (see also further references by the same authors). The sparse regime, with a bounded number of non-zero entries per row, is the one of community detection in the stochastic block model with bounded average degree. There the high-degree vertices hide the informative outliers of the adjacency matrix, while those of the non-backtracking matrix persist down to the Kesten–Stigum threshold, as conjectured in Krzakala et al. (2013) and proved in Bordenave et al. (2015). For sparse non-Hermitian matrices with independent entries and a finite-rank mean, the outliers and their eigenvector overlaps are known rigorously Bordenave et al. (2023); Coste & Stephan (2021). Most recent works on block Markov chains observed along one trajectory (see e.g. Sanders et al. (2020); Sanders & Senen-Cerda (2023) and references therein) assume $n/V \to \infty$. They give error rates, information-theoretic lower bounds and bulk singular value laws of the count matrix, but no detectability threshold at fixed α. For low-rank chains, Zhang & Wang (2019); Zhu et al. (2022) show that n of order V, up to logarithmic factors, is necessary and sufficient for consistent estimation. The exact law of the count matrix M goes back to Whittle (1955). To the best of our knowledge, the asymptotic outlier and its eigenvector overlap have however not been treated in the setting of Claim E.9: the sparse regime $n = \alpha V$, and a non-Hermitian noise whose rows have dependent entries.

E.4.2 THE BULK OF THE NOISE

Let W^0 be the matrix with independent entries $W^0_{ij} \sim \mathcal{N}(0, \Sigma_{ij})$ of Lemma E.7. Its spectrum is described by existing results, which we take from Cook et al. (2018); Alt et al. (2021), and which, together with Lemma E.8, justifies Claim E.9-(i).

Theorem E.10. *Conditionally on m, as $V \to \infty$, with $N \coloneqq |\mathcal{S}|$ and $\delta_N \coloneqq \frac{1}{N}\sum_{j\in\mathcal{S}} u_j$:*

(i) under the leading-order convention $\delta_N = 0$ of section E.2, the empirical spectral distribution of W^0 restricted to $\mathcal{S}\times\mathcal{S}$ converges in probability to the uniform law on the disc $\{|z| \le \rho_\alpha\}$;

(ii) without any convention, for every $\varepsilon, p > 0$ there is $C_{\varepsilon,p}$ such that

$$\mathbb{P}\Big(\big|\rho(W^0) - \sqrt{\rho(\Sigma^Y)}\big| \ge N^{-1/2+\varepsilon}\Big) \le C_{\varepsilon,p}\, N^{-p}, \qquad \rho(\Sigma^Y) = \rho_V^2\, \frac{(1-\beta) + \sqrt{(1+\beta)^2 - 4\beta\delta_N^2}}{2}, \tag{E.23}$$

where $\Sigma^Y_{ij} \coloneqq (1-\beta u_i u_j)/n$, $i, j \in \mathcal{S}$, is the variance profile of the matrix Y of Lemma E.8. In particular, since $\delta_N = O(N^{-1/2})$ in probability, $\rho(W^0) \to \rho_\alpha$ in probability.

Proof of Theorem E.10. By Lemma E.8(i), W^0 restricted to $\mathcal{S}\times\mathcal{S}$ has the law of $D^{-1/2} Y D^{1/2}$, hence the same spectrum as the $N \times N$ matrix Y, whose entries are independent centred Gaussians of variance Σ^Y_{ij}. Both statements are results on matrices with independent entries and a variance profile, applied to Y.

(i) is (Cook et al., 2018, Theorem 2.4). That theorem considers $Y_N \coloneqq \frac{1}{\sqrt{N}}(\sigma_{ij} Z_{ij})_{ij}$, where the Z_{ij} are i.i.d. with $\mathbb{E}Z_{11} = 0$, $\mathbb{E}|Z_{11}|^2 = 1$ and $\mathbb{E}|Z_{11}|^{4+\varepsilon} \le M_0$ for some fixed $\varepsilon > 0$, and (σ_{ij}) is deterministic with $0 \le \sigma_{ij} \le \sigma_{\max}$ and doubly stochastic variance profile, $\frac{1}{N}\sum_i \sigma_{ij}^2 = \frac{1}{N}\sum_j \sigma_{ij}^2 = 1$ for all i, j. It states that the empirical spectral distribution of Y_N converges in probability to the uniform law on the unit disc. To match this normalisation, write

$$Y_{ij} = \frac{1}{\sqrt{n}}\sqrt{1-\beta u_i u_j}\, g_{ij} = \sqrt{\frac{N}{n}} \cdot \frac{1}{\sqrt{N}}\, \sigma_{ij}\, g_{ij}, \qquad \sigma_{ij}^2 \coloneqq 1 - \beta u_i u_j,$$

$$\text{that is} \qquad Y = \rho_V\, \widehat{Y}, \qquad \widehat{Y} \coloneqq \frac{1}{\sqrt{N}}(\sigma_{ij} g_{ij})_{ij},$$

since $\rho_V^2 = N/n$ by eq. (E.5). The entries g_{ij} are standard Gaussian and $\sigma_{ij}^2 \leq 2$. The row sums of the profile are $\frac{1}{N}\sum_{j\in\mathcal{S}}(1-\beta u_i u_j) = 1-\beta u_i \delta_N$, and the column sums are the same by symmetry of the profile. Under the convention $\delta_N = 0$ the profile is therefore exactly doubly stochastic, the theorem applies to $\widehat{Y}$, the eigenvalues of $\widehat{Y}$ fill the unit disc, and those of $Y = \rho_V \widehat{Y}$ fill the disc of radius $\rho_V \to \rho_\alpha$.

(ii) is (Alt et al., 2021, Theorem 2.1). That theorem considers an $N \times N$ matrix Z with independent centred entries and variance matrix $S := (\mathbb{E}|Z_{ij}|^2)$ such that (A1) S is flat, $\kappa_*/N \leq S_{ij} \leq \kappa^*/N$ for constants $\kappa^* > \kappa_* > 0$; (A2) for each $p \in \mathbb{N}$ there is μ_p with $\mathbb{E}|Z_{ij}|^p \leq \mu_p N^{-p/2}$; (A3) $\sqrt{N}\, Z_{ij}$ has a density ν_{ij} with $\|\nu_{ij}\|_{L^{1+\delta}} \leq N^\omega$ for some constants $\delta, \omega > 0$. It states that for every $\varepsilon, p > 0$ there is $C_{\varepsilon,p}$ with $\mathbb{P}\big(|\rho(Z) - \sqrt{\rho(S)}| \geq N^{-1/2+\varepsilon}\big) \leq C_{\varepsilon,p} N^{-p}$. For $Z = Y$, the entries $(1 \pm \beta)/n$ of $S = \Sigma^Y$ are flat because $\beta < 1$ and $N/n = \rho_V^2 \to \rho_\alpha^2 > 0$, and Gaussian entries satisfy (A2) and (A3). This gives the bound in eq. (E.23). For the value of $\rho(\Sigma^Y)$, the profile is constant on the four blocks defined by the classes of i and j, so its Perron eigenvector is constant on each class and $\rho(\Sigma^Y)$ is the Perron eigenvalue of the reduced 2×2 matrix

$$\frac{1}{n}\begin{pmatrix} N_+(1-\beta) & N_-(1+\beta) \\ N_+(1+\beta) & N_-(1-\beta) \end{pmatrix}, \qquad N_\pm := \#\{j \in \mathcal{S} : u_j = \pm 1\} = \frac{N}{2}(1 \pm \delta_N),$$

whose trace is $\frac{N}{n}(1-\beta)$ and determinant $-\frac{N^2}{n^2}\beta(1-\delta_N^2)$. This yields the largest eigenvalue given in eq. (E.23). □

On the balance convention in (i). In Models E.1 and E.6 the balance $\delta_N = 0$ is not exact: the variance profile of $\widehat{Y}$ is doubly stochastic only up to the imbalance $\delta_N = O(N^{-1/2})$ of the visited classes, and (Cook et al., 2018, Theorem 2.4) does not apply verbatim. We expect the general case to follow from the more general result (Cook et al., 2018, Theorem 2.3), which applies to the unbalanced profile and gives a deterministic equivalent μ_N of the empirical spectral distribution, combined with a perturbation argument showing that μ_N converges to the uniform law on the unit disc as $\delta_N \to 0$. Since the main point of the derivation is to justify the (non-rigorous) Claim E.9, we leave this technical point to a future rigorous work.

From W^0 to W. Recall that $W = W^0 - \mathrm{Diag}(W^0\mathbf{1})\overline{K}$ of Lemma E.7(iii) differs from W^0 by a matrix of rank two. We assume that this perturbation of finite rank does not change the limiting spectral distribution nor the spectral radius: i.e. that the eigenvalues of W fill the disc of radius $\rho_V \to \rho_\alpha$ with the density of Theorem E.10, and that the spectral radius of W is ρ_V (up to $o(1)$ corrections).

E.4.3 A HIGH-LEVEL SKETCH OF THE DERIVATION

We derive the outlier and the overlap of Claim E.9 in three steps, using the resolvent $G(z) := (z - W)^{-1}$. The derivation leverages the following two consequences of Model E.6, and uses classical ideas described in the literature on BBP transitions, see e.g. Benaych-Georges & Nadakuditi (2011).

(a) If v is a vector that does not depend on the row i of W (e.g. if v is fixed, or built from the other rows), then $(Wv)_i = \sum_j W_{ij} v_j$ is a centred Gaussian variable of variance

$$\frac{v^\top C_i v}{m_i} = \frac{1}{m_i}\Big[\sum_j q_{ij} v_j^2 - \big(q_{i\cdot}^\top v\big)^2\Big] = \big(\Sigma\, v^{\circ 2}\big)_i - \frac{(\overline{K}v)_i^2}{m_i}, \tag{E.24}$$

using $q_{i\cdot}^\top v = (\overline{K}v)_i$. Since $\sum_j q_{ij} = 1$, this variance is of order one for a vector with entries of size one.

(b) *Removing the dependence on row i.* We shall apply (a) to $v = W^{k-1}y$, which depends on W. Expanding the product, v_j is the sum over the paths $j \to i_1 \to \cdots \to i_{k-1}$ of the products $W_{ji_1}W_{i_1i_2}\cdots W_{i_{k-2}i_{k-1}}y_{i_{k-1}}$, and row i of W enters v_j only through the paths that pass through i. In $(Wv)_i = \sum_j W_{ij}v_j$, these are the paths of length k started at i that come back to i. For fixed k, such a path has $k-1$ intermediate states, each ranging over V values, and it comes back to i for a fraction $O(k/V)$ of the choices: the neighbourhood of i explored in k steps (for k finite as $V \to \infty$) is a tree, up to this small fraction of the paths

that we neglect. Under this approximation, v no longer depends on row i, and (a) applies to $(Wv)_i$ conditionally on the other rows. This approximation can be seen as an example of a cavity approximation from statistical physics, see e.g. Mezard & Montanari (2009).

Step 1: exact outlier equation. Recall $\overline{K} = AB^\top$ with $A = \frac{1}{n}[a, -\beta b]$ and $B = [m, u \circ m]$. If z is not an eigenvalue of W, the eigenvalue equation $(\overline{K} + W)x = zx$ can be rewritten as $(z - W)x = A(B^\top x)$, that is

$$x = G(z)\, A\, \eta, \qquad \eta := B^\top x \in \mathbb{R}^2, \qquad \text{hence} \qquad \eta = B^\top G(z) A\, \eta. \tag{E.25}$$

So z is an eigenvalue of X outside the spectrum of W if and only if the 2×2 matrix $I_2 - B^\top G(z)A$ is singular, and the corresponding eigenvector is then $G(z)A\eta$ for η in its null space. This reduces the problem of characterising the outlier and its eigenvector to computing the four numbers $\frac{1}{n} w^\top G(z) y$, for $w \in \{m, u \circ m\}$, $y \in \{a, -\beta b\}$, and the vector $G(z)A\eta$.

Step 2: the threshold. For $|z|$ larger than the spectral radius of W, we can expand

$$\frac{1}{n}\, w^\top G(z)\, y = \frac{w^\top y}{n\, z} + \sum_{k \geq 1} \frac{w^\top W^k y}{n\, z^{k+1}}.$$

We claim that every term $k \geq 1$ in this sum is negligible as $V \to \infty$ for $w \in \{m,\, u \circ m,\, u\}$ and $y \in \{a,\, b\}$. For $k = 1$ this is (a) alone: $w^\top W y = \sum_i w_i (Wy)_i$ is a sum of V independent centred Gaussian variables of variance $w_i^2\, y^\top C_i y / m_i = O(1)$, hence a centred Gaussian of variance $O(V)$, and thus negligible with respect to $n \asymp V$. For $k \geq 2$, write $W^k y = Wv$ with $v = W^{k-1} y$. By (b), conditionally on the other rows, $(Wv)_i$ is a centred Gaussian of variance $v^\top C_i v / m_i \leq \frac{1}{m_i} \sum_j q_{ij} v_j^2 = (\Sigma\, v^{\circ 2})_i$. Taking expectations,

$$\mathbb{E}\big[(W^k y)^{\circ 2}\big] \leq \Sigma\, \mathbb{E}\big[(W^{k-1} y)^{\circ 2}\big] \leq \cdots \leq \Sigma^k\, y^{\circ 2} \qquad \text{entrywise}, \tag{E.26}$$

where the iteration is legitimate because the entries of Σ are nonnegative, so that multiplication by Σ preserves entrywise inequalities. Since the rows of Σ sum to $s_i = 1/m_i \leq 1$, the entries of $W^k y$ are centred with mean square at most $\max_j y_j^2$, of order one, for every fixed k. Hence $w^\top W^k y = \sum_i w_i (W^k y)_i$ is a sum of V centred terms of order one, nearly independent across i by (b): it is of size $\sqrt{V}$, negligible with respect to n. Therefore, for $w \in \{m,\, u \circ m,\, u\}$ and $y \in \{a,\, b\}$,

$$\frac{1}{n}\, w^\top G(z)\, y \simeq \frac{w^\top y}{n\, z} \qquad (|z| > \rho_V). \tag{E.27}$$

For the reader accustomed to the BBP transition in symmetric or Hermitian settings, let us remark that this simplification fails for a symmetric noise, and the failure explains the shift of the outlier in the Hermitian case Baik et al. (2005); Benaych-Georges & Nadakuditi (2011). Take W symmetric Gaussian with entries of variance σ^2/V and compute the term $k = 2$:

$$(W^2 y)_i = \sum_{j,l} W_{ij} W_{jl}\, y_l = \underbrace{y_i \sum_j W_{ij}^2}_{\simeq \sigma^2 y_i} + \sum_j \sum_{l \neq i} W_{ij} W_{jl}\, y_l.$$

The first part comes from the paths $i \to j \to i$ that return to i along the same entry, $W_{ji} = W_{ij}$: it is not centred, it is aligned with y, and it gives $\frac{1}{n} w^\top W^2 y \simeq \sigma^2 w^\top y / n$, of the same order as the $k = 0$ term. Summing the series then produces the Stieltjes transform $g(z) = 1/(z - \sigma^2 g(z))$ of the semicircle law in place of $1/z$ as in the present setting.

Eq. (E.27) has the following consequences.

- *The outliers.* Since $\frac{1}{n} m^\top a = \frac{1}{n}(u \circ m)^\top b = 1$ and $\frac{1}{n} m^\top b = \frac{1}{n}(u \circ m)^\top a = 0$, eq. (E.27) gives $B^\top G(z) A \simeq \frac{1}{z}\mathrm{Diag}(1, -\beta)$, and the outlier equation reads $(1 - \frac{1}{z})(1 + \frac{\beta}{z}) = 0$: $z = 1$ and $z = -\beta$ are the possible asymptotic outliers.
- *The threshold.* The expansion of Step 2 converges exactly when $|z|$ exceeds the spectral radius of W (which is ρ_V as $V \to \infty$, see section E.4.2). Hence eq. (E.27), and the outlier equation in the form $(1 - \frac{1}{z})(1 + \frac{\beta}{z}) = 0$, hold for $|z| > \rho_V$ only, and the zeros lying in this region are the asymptotic outliers of X. Since $\rho_V < 1$, $z = 1$ is always an outlier, while the value $z = -\beta$ is an outlier if and only if $\beta > \rho_V$.

Step 3: the asymptotic overlap. Assume $\beta > \rho_V$. By Step 2, the outlier is $z_\star = -\beta$, its right eigenvector is $u_\star = \sqrt{V}\, G(z_\star)b/\|G(z_\star)b\|$ (normalised so that $\|u_\star\| = \sqrt{V}$), and $u^\top G(z_\star)b \simeq |\mathcal{S}|/z_\star$. The overlap of eq. (E.3) is therefore

$$q_{\mathrm{R}}^2 = \frac{\langle u_\star, u\rangle^2}{V^2} = \frac{(u^\top G(z_\star)b)^2}{V\,\|G(z_\star)b\|^2} \simeq \frac{|\mathcal{S}|^2}{\beta^2\, V\,\|G(z_\star)b\|^2}, \tag{E.28}$$

and what is left to compute is the norm $\|G(z)b\|^2$ at $z = z_\star$. Let $v \coloneqq G(z)b$ and $\mathcal{E}(z) \coloneqq \mathbb{E}[v^{\circ 2}] \in \mathbb{R}^V$, so that $\|G(z)b\|^2 \simeq \mathbf{1}^\top \mathcal{E}(z)$ as $V \to \infty$ (we assume it concentrates on its average as $V \to \infty$). We will characterise $\mathcal{E}(z)$ as $V \to \infty$. The identity $z\,G(z) = I + W\,G(z)$ multiplied by b gives

$$v_i = \frac{b_i}{z} + \frac{(Wv)_i}{z}, \qquad \text{hence} \qquad \mathcal{E}_i(z) = \frac{1}{z^2}\Big[\, b_i^2 + 2\,b_i\,\mathbb{E}[(Wv)_i] + \mathbb{E}[(Wv)_i^2]\,\Big]. \tag{E.29}$$

Recall that the vector v depends on W, but by the cavity argument (b) it may be replaced by its version without row i, and (a) then applies conditionally on the other rows. This yields $\mathbb{E}[(Wv)_i] \simeq 0$ and, by eq. (E.24),

$$\mathbb{E}[(Wv)_i^2] \simeq \frac{\mathbb{E}[v^\top C_i v]}{m_i} = \big(\Sigma\,\mathcal{E}(z)\big)_i - \frac{\mathbb{E}[(\overline{K}v)_i^2]}{m_i}\,: \tag{E.30}$$

By eq. (E.13), $(\overline{K}v)_i = \frac{1}{n}\big(a_i\, m^\top v - \beta\, b_i\,(u\circ m)^\top v\big)$, and eq. (E.27) applied to $v = G(z)b$ gives $\frac{1}{n}m^\top v \simeq 0$ and $\frac{1}{n}(u\circ m)^\top v \simeq 1/z$. So $(\overline{K}v)_i \simeq -\beta b_i/z$, a deterministic number to leading order, and $\mathbb{E}[(\overline{K}v)_i^2]/m_i \simeq \beta^2 s_i/z^2$. Inserting this and $b_i^2 = a_i$ in eq. (E.29), we reach:

$$\mathcal{E}(z) \simeq \frac{1}{z^2}\Big[\, a + \Sigma\,\mathcal{E}(z) - \frac{\beta^2}{z^2}\, s\,\Big], \tag{E.31}$$

a linear self-consistent equation for the vector $\mathcal{E}(z) \in \mathbb{R}^V$. To leading order, it can be rewritten as

$$\Big(I - \frac{\Sigma}{z^2}\Big)\mathcal{E}(z) = \frac{1}{z^2}\Big(a - \frac{\beta^2}{z^2}\, s\Big), \tag{E.32}$$

which has a unique solution when $z^2 > \rho(\Sigma) = \rho_V^2$ (see Theorem E.10), given by the convergent series

$$\mathcal{E}(z) = \sum_{k\geq 0} \frac{\Sigma^k}{z^{2k}} \cdot \frac{1}{z^2}\Big(a - \frac{\beta^2}{z^2}\, s\Big). \tag{E.33}$$

By Lemma E.5, $\Sigma a = s$ and $\Sigma s = \rho_V^2 s$, so that $\Sigma^k s = \rho_V^{2k} s$ for $k \geq 0$ and $\Sigma^k a = \rho_V^{2(k-1)} s$ for $k \geq 1$. The series in eq. (E.33) can therefore be summed explicitly for $z^2 > \rho_V^2$:

$$\sum_{k\geq 0} \frac{\Sigma^k a}{z^{2k}} = a + \frac{s}{z^2}\sum_{k\geq 1}\Big(\frac{\rho_V^2}{z^2}\Big)^{k-1} = a + \frac{s}{z^2 - \rho_V^2}, \qquad \sum_{k\geq 0}\frac{\Sigma^k s}{z^{2k}} = \frac{z^2}{z^2-\rho_V^2}\, s. \tag{E.34}$$

Inserting eq. (E.34) in eq. (E.33), and using $\mathbf{1}^\top a = |\mathcal{S}|$ and $\mathbf{1}^\top s = \sum_{i\in\mathcal{S}} 1/m_i \simeq |\mathcal{S}| h(\alpha)$ by eq. (E.5), we get

$$\mathcal{E}(z) \simeq \frac{1}{z^2}\Big[\, a + \frac{1-\beta^2}{z^2-\rho_V^2}\, s\,\Big], \qquad \|G(z)b\|^2 \simeq \frac{|\mathcal{S}|}{z^2}\Big[1 + \frac{h(\alpha)\,(1-\beta^2)}{z^2-\rho_V^2}\Big]. \tag{E.35}$$

Conclusion. Inserting the value of $\|G(z_\star)b\|^2$ at $z_\star = -\beta$ into eq. (E.28),

$$q_{\mathrm{R}}^2 \simeq \frac{|\mathcal{S}|^2/\beta^2}{V\,\frac{|\mathcal{S}|}{\beta^2}\Big[1 + \frac{h(\alpha)(1-\beta^2)}{\beta^2-\rho_V^2}\Big]} = \frac{|\mathcal{S}|}{V}\cdot\frac{\beta^2-\rho_V^2}{(\beta^2-\rho_V^2) + h(\alpha)(1-\beta^2)},$$

which is eq. (E.22) since $|S|/V \to 1 - e^{-\alpha}$ as $V \to \infty$.

F Additional experimental results

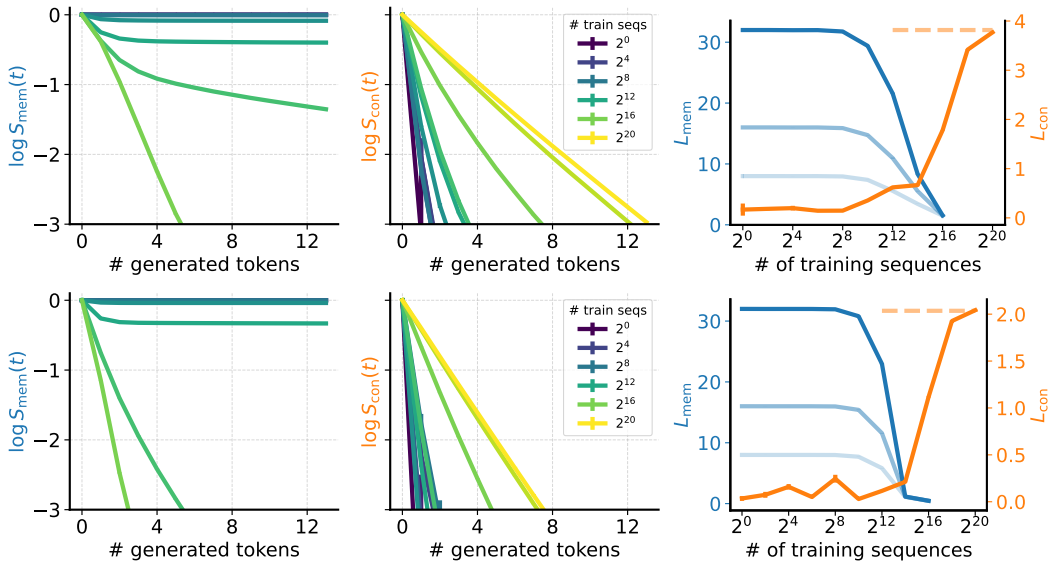


Figure F.1: **Memorisation and convergence in transformers** Same figure as fig. 1, but for the `tinystories` (top) and `wikipedia.en` data sets. Experimental details in section B.