How Deep Are Deep GPs, Really? A Sharp Threshold and a Non-Gaussian Limit for Compositional GPs

Mark Kozdoba [1] **Shie Mannor** [1 2]

Abstract

Compositional priors describe the generic properties of layered functions in deep Bayesian models, where deep neural networks with random weights are a canonical example. In the wide-network limit, the prior is a Gaussian process (GP) with a depth-dependent kernel, and its behaviour as depth grows has been extensively studied through this kernel. Here, we study another case, where each layer itself is a vector valued Gaussian process, and our aim is similarly to understand the structure induced by these priors as depth grows.

Previous GP work has established that for the radial basis function (RBF) kernel and a certain range of bandwidths r, the prior degenerates in the limit, converging to the set of constant functions — which is not useful as a probabilistic or a generative model. In this paper we establish several new results. First, we identify a sharp bandwidth threshold $r_c(d) = \Theta(\sqrt{d})$ above which the limit is degenerate, strengthening the earlier bounds. Second, and more importantly, we show that for r below the threshold $r_c(d)$ the prior converges to a limit distribution $\pi_{\bar{Z}}$. We also prove that these distributions are non-degenerate and non-Gaussian, with non-vanishing dependence between coordinates. In contrast to the previously known degenerate regime, deep Gaussian process priors can therefore admit non-trivial limits.

Empirically, we verify the threshold across a range of dimensions d, and demonstrate that the limit distributions $\pi_{\bar{Z}}$ are capable of generating complex multimodal samples — a regime that becomes increasingly narrow with d and would be hard to identify without knowing the threshold.

[1]Technion – Israel Institute of Technology, Haifa, Israel [2]NVIDIA. Correspondence to: Mark Kozdoba <markk@technion.ac.il>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

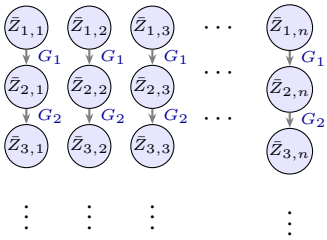

Figure 1. The compositional chain (1) on n input points. At depth i the *same* random Gaussian function G_i is applied to every one of the n points, coupling all trajectories through the shared randomness.

1. Introduction

A central approach to modelling uncertainty is Bayesian inference, in which epistemic uncertainty is carried by the posterior. The posterior, however, is shaped by the prior: the prior encodes the class of models one believes a priori plausible, and any uncertainty estimate extracted from the posterior inherits the prior's biases. Understanding the prior — where its mass lives, what structure it prefers — is thus a prerequisite to understanding Bayesian methods built on top of it. The properties of the prior are critical for the sample complexity of learning, and hence for generalization and memorization (Fortuin, 2022; Tran et al., 2022; Wilson & Izmailov, 2020). Some classical priors, such as shallow Gaussian processes (GPs) (Rasmussen & Williams, 2006) or composition of wide neural network layers (Neal, 1996; Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019), are by now well understood. In particular, wide neural network priors are themselves a single Gaussian process, with a depth-dependent kernel (Lee et al., 2018; Matthews et al., 2018) (see also Section 2), and the properties of the prior can be read off from the properties of that kernel. However, less is known about compositions in which each layer is itself a Gaussian process.

Why compose Gaussian processes. Draw $G_1, \dots, G_l$ i.i.d. from a centred vector valued Gaussian process prior and consider the composition $G_l \circ \cdots \circ G_1$. These *compositional*, or *deep*, Gaussian process priors (Damianou & Lawrence, 2013) have been shown to better fit empirical data than shallow GPs in some cases (Damianou & Lawrence,

2013; Lu et al., 2019), with intermediate layers playing the role of learned features in a way that mirrors the argument for depth in neural networks. The price of this flexibility however is complexity of analysis: past a single layer, the joint law of the composed function at any finite set of inputs is no longer Gaussian, and its qualitative dependence on depth remains poorly understood.

Generation, composition, and GPs. Compositional priors, and Gaussian-process compositions in particular, have been studied in the literature as generative models. The Gaussian process latent variable model (Lawrence, 2005; Titsias & Lawrence, 2010; Damianou et al., 2016) treats a single GP as a generator from a low-dimensional latent space to data, and deep Gaussian processes (Damianou & Lawrence, 2013) extend this to a layered generative pipeline in which a latent input is pushed through a stack of i.i.d. random Gaussian functions. Beyond the GP family, deep latent-variable models more generally (Rezende et al., 2014) and Bayesian treatments of deep generators (Saatci & Wilson, 2017) place priors over compositions of random maps with the same generative use in mind.

The coupled walk and synchronisation. The central object in this paper is the Markov chain that the compositional prior induces on any collection of n input points simultaneously. Let $\bar{Z}_1 = (\bar{Z}_{1,1}, \dots, \bar{Z}_{1,n}) \in (\mathbb{R}^d)^n$ be n starting points and iterate

$$\bar{Z}_{i+1} \;=\; G_i \bar{Z}_i \;=\; (G_i \bar{Z}_{i,1}, \dots, G_i \bar{Z}_{i,n}), \tag{1}$$

with G_i i.i.d. samples from a centred vector GP with rotation-invariant kernel $k(x, x') = \exp(-\|x - x'\|^2 / (2r^2))$. That is, $G_i : \mathbb{R}^d \to \mathbb{R}^d$ are random Gaussian functions (see Section 4 for a definition); see Figure 1 for the structure. Because the *same* G_i is applied to every coordinate, the n trajectories $\{\bar{Z}_{i,t}\}_{i \geq 1}$, where $t \leq n$, are coupled through the shared randomness. Dunlop et al. (2018) showed that for r large enough the process has a *degenerate* limit: the pairwise distances $\|\bar{Z}_{i,s} - \bar{Z}_{i,t}\|$ collapse to zero with depth (i.e. $i \to \infty$), and consequently, the prior concentrates on constant functions. Note that when the bandwidth r is large, the values of the process G_i are strongly correlated, and therefore the process itself is somewhat close to the constant functions. The above result then shows that composition further amplifies this, yielding a limit that coincides with the constant functions. However, their argument leaves open both whether their particular threshold for r is sharp, and, more importantly, what happens on the other side of it.

We note that the general framework of coupled walks (1) is also studied in the Theory of Iterated Random Functions (see Section 2); there, the phenomenon of a degenerate limit is known as *synchronisation*, with the interpretation that all particles eventually follow the same trajectory. We will use the terms synchronisation and degenerate limit interchangeably.

Convergence and sharp thresholds. We give a sharp almost-sure threshold separating a supercritical regime, in which the chain synchronises, from a subcritical regime, in which it admits a nontrivial stationary law. The critical radius is

$$r_c(d) \;=\; \sqrt{2}\, e^{\psi(d/2)/2},$$

where ψ is the digamma function; asymptotically $r_c(d) \sim \sqrt{d}$ as $d \to \infty$ (see Remark 4.6). For $r > r_c(d)$ the coupled walk synchronises almost surely at an explicit exponential rate (Theorem 4.1), improving on Dunlop et al. (2018), whose threshold lies strictly above $r_c(d)$ for every d. More importantly, for $r < r_c(d)$ — a regime not addressed by Dunlop et al. (2018) — we show that the scalar pairwise-distance chain converges in total variation (TV) to a unique nontrivial stationary law on $(0, \infty)$ (Theorem 4.1), and the full position chain $\bar{Z}_i$ converges in total variation to a unique stationary law $\pi_{\bar{Z}}$ on $(\mathbb{R}^d)^n$ for every $n \geq 2$ (Theorem 4.3).

We note that Dunlop et al. (2018) also proved convergence to stationarity, but for a different class of processes. Their direct approach, of constructing Lyapunov functions directly for $\bar{Z}$, does not seem to apply to the composition class, for which they only prove synchronization as discussed above. We avoid the issue by taking an indirect approach: observing that $\bar{Z}$ is determined by the pairwise distances $\|\bar{Z}_{i,t} - \bar{Z}_{i,s}\|$, we work with the induced chain on those distances. Methodologically, Theorem 4.3 is an extension of Theorem 4.1.

The limit is non-Gaussian with non-trivial dependence. A priori one might expect the stationary distribution $\pi_{\bar{Z}}$ to be Gaussian — it is built entirely from Gaussian ingredients, and each marginal $\bar{Z}_{\infty,t}$ is standard Gaussian. Remarkably, however, the joint law of any two coordinates $(\bar{Z}_{\infty,s}, \bar{Z}_{\infty,t})$ is not jointly Gaussian (Theorem 4.5). Moreover, we show that there is dependence between coordinates of $\pi_{\bar{Z}}$, and that it can be estimated in terms of r and d (Corollary 4.9). Setting $r = \lambda\, r_c(d)$, the dependence is weak at small λ and grows strong as λ approaches 1. Crucially, the range of λ in which it is visible on a finite sample of $\pi_{\bar{Z}}$ shrinks rapidly with d — at $d = 100$ the entire non-trivial regime sits inside roughly $\lambda \in (0.99, 1)$ (Section 5). Without the bounds of Proposition 4.7 and Corollary 4.9 and the precise value of $r_c(d)$, this narrow strip is hard to locate — a moderate-λ experiment in high d looks indistinguishable from the independent-coordinate Gaussian baseline. However, when the relevant scale of $1 - \lambda$ is identified, λ can be used to tune the strength of dependence.

Experiments We numerically verify the threshold $r_c(d)$ across $d \in \{1, 10, 100\}$, and sample from $\pi_{\bar{Z}}$ at λ values selected per dimension using the bounds discussed above. The samples display complex chain-specific multimodal structure, consistent with the non-Gaussianity of Theorem 4.5.

Summary of contributions. Our main contributions are: (i) convergence in total variation to a unique stationary law $\pi_{\bar{Z}}$ at every $r < r_c(d)$. Until now, the only depth-asymptotic behaviour established for the composition class was synchronisation onto constants; we show that below the threshold the prior in fact admits a non-degenerate limit, qualitatively changing the picture; (ii) the structural finding that $\pi_{\bar{Z}}$ is non-Gaussian, with coordinate dependence that does not wash out as $d \to \infty$; and (iii) a sharp almost-sure threshold $r_c(d)$ separating the synchronised regime from this non-trivial limit, with an explicit decay rate above the threshold. We verify (iii) across $d \in \{1, 10, 100\}$, and exhibit (i)–(ii) by visualising samples from $\pi_{\bar{Z}}$ at λ values selected per dimension using the bounds of (iii); the samples display complex chain-specific multimodal structure.

Organization. The rest of this paper is organized as follows. Section 2 reviews related work on deep and compositional Gaussian processes, infinite-width neural-network priors, iterated random functions, and previous work on GP composition. Section 3 fixes notation and recalls the objects we use. Section 4 contains the formal statements of our main results together with proof sketches. Section 5 presents numerical experiments verifying the threshold across dimensions and Section 6 discusses open directions. The proofs are collected in Supplementary A.

2. Previous Work

Deep and compositional Gaussian processes. Deep Gaussian processes were introduced by Damianou & Lawrence (2013) as a hierarchical prior formed by stacking GP layers. Their expressivity — non-stationary effective kernels, input-dependent length scales, richer induced push-forward distributions — has been the subject of a substantial body of follow-up work, including approximate inference methods (Bui et al., 2016; Salimbeni & Deisenroth, 2017; Cutajar et al., 2017; Havasi et al., 2018), conditional moment calculations (Lu et al., 2019), and depth-asymptotic studies of several specific constructions (Duvenaud et al., 2014; Dunlop et al., 2018). Among the constructions in this literature, the *composition* class — in which a single GP sample is applied layer-wise to the coupled input vector — is the object of our analysis.

Infinite-width neural-network priors. A parallel mean-field literature analyses depth asymptotics of Bayesian neural-network priors with i.i.d. Gaussian weights and biases of variances σ_w^2 and σ_b^2. Poole et al. (2016); Schoenholz et al. (2017); Hayou et al. (2019) identify, at infinite width and for tanh activations, a synchronisation / non-synchronisation dichotomy in (σ_w^2, σ_b^2) analogous to ours: in some parameter regimes the network synchronises, i.e. the covariance of two input points approaches 1, while in others the covariance approaches a strictly smaller fixed point. For ReLU activations the dichotomy collapses and synchronisation is unavoidable for every choice of these parameters (Hayou et al., 2019).

A structural difference from our work is that in this setting the entire infinite-width composition collapses to a single Gaussian process, with all depth dependence absorbed into its kernel, termed the *neural network Gaussian process* (NNGP) kernel, which satisfies a deterministic depth recursion (Matthews et al., 2018; Lee et al., 2018; Yang, 2019). The synchronisation behaviour is therefore a property of the depth-evolution of this kernel. Thus, the setup here can be considered as a single Gaussian process whose kernel encodes the entire depth dependence. In contrast, the composition we analyse is a genuine composition of many GPs, and is non-Gaussian at every finite depth. In that setting Theorem 4.1 establishes the existence of a non-trivial limiting law, and we further show that this law is also non-Gaussian.

Non-Gaussian limits of deep priors. Two recent lines of work in the deep-neural-network literature establish non-Gaussian limits in different parameter regimes. Bordino et al. (2023), building on Peluchetti et al. (2020), replace Gaussian weights with heavy-tailed Stable weights and obtain Stable-process limits in the infinite-width regime. Hayou (2023) fix the width and let depth grow in a residual architecture, obtaining continuous-time SDE limits whose form depends on the activation function (see also Li et al., 2022). Both routes deliver non-Gaussianity by changing the mechanism behind the standard Gaussian-limit picture: Bordino et al. (2023) change the noise mechanism, replacing Gaussian weights with Stable ones to begin with; Hayou (2023) change the iteration mechanism, since the residual block is additive rather than compositional and is rescaled with the total depth. These architectures are different from our approach, which is purely compositional and uses only Gaussian ingredients. The non-Gaussianity in (1) arises directly from the GP composition mechanism.

Iterated random functions. Another area in which the chain (1) appears is the theory of iterated random functions (Diaconis & Freedman, 1999; Wu & Shao, 2004; Stenflo, 2012), which studies Markov chains obtained by iterating maps G_i drawn i.i.d. from a distribution on function space. The classical theory focuses on contraction-on-average criteria and backward-iteration limits, primarily for finite-dimensional parametrised families such as random affine

maps or Kesten-type perpetuities. The specific setting we study — G_i a sample from a Gaussian process with spatially stationary kernel — gives a genuinely infinite-dimensional random function with non-trivial spatial correlations, and the sharp-threshold / non-Gaussian-limit phenomena established here are not addressed in the classical literature.

On the composition class. The closest technical precedent to our work is Dunlop et al. (2018), working with the scalar pairwise-distance chain (analogous to v in Theorem 4.1). They prove synchronization in L^2 at threshold $r > \sqrt{d}$, which upgrades to almost-sure convergence but only at this cruder L^2-driven threshold. Theorem 4.1 sharpens the a.s. threshold to $r_c(d) = \sqrt{2}\, e^{\psi(d/2)/2}$, opening the window $r \in (r_c(d), \sqrt{d})$ of almost-sure synchronization that is invisible to any L^2 argument. As noted in the introduction, $r_c(d) = \Theta(\sqrt{d})$, with $r_c(d) < \sqrt{d}$ strictly for every $d \geq 1$ but $r_c(d)/\sqrt{d} \to 1$ as $d \to \infty$ (Remark 4.6); the gap with Dunlop's threshold therefore closes with dimension. More importantly, Dunlop et al. (2018) do not address the subcritical regime $r < r_c(d)$, which is covered by Theorem 4.1 and, along with Theorems 4.3 and 4.5, is our principal contribution beyond their work.

As discussed earlier, Dunlop et al. (2018) show ergodicity results (their Theorem 8), but for a *different* construction (Paciorek-style hierarchical kernels), and these do not transfer to the composition class.

Lu et al. (2019) compute, in closed form, the second and fourth moments of a single two-layer composition for several specific kernel families. These are one-step identities and cannot be iterated to larger depth: the moments at layer $L+1$ depend on the full distribution of the layer-L output, not on its moments alone, since the chain is non-Gaussian at every finite depth and no finite collection of moments closes the recursion.

3. Notation and preliminaries

In this section we collect a few standard objects used throughout. The *digamma function* ψ is the logarithmic derivative of the gamma function, $\psi(z) := \Gamma'(z)/\Gamma(z)$ ((Abramowitz & Stegun, 1964)). The *Euler–Mascheroni constant* is $\gamma := -\psi(1) \approx 0.5772$, and we will use the value $\psi(1/2) = -\gamma - 2\log 2$.

For $d \geq 1$, the *chi-squared distribution with d degrees of freedom*, denoted χ^2_d, is the law of $\sum_{k=1}^d g_k^2$ for $g_1, \ldots, g_d \sim \mathcal{N}(0,1)$ i.i.d. standard Gaussians ((Abramowitz & Stegun, 1964)). Its Lebesgue density on $(0,\infty)$ is $f_d(x) = \frac{1}{2^{d/2}\,\Gamma(d/2)}\, x^{d/2-1}\, e^{-x/2}$, and the logarithmic moment is $\mathbb{E}\log X = \psi(d/2) + \log 2$ for $X \sim \chi^2_d$. In particular χ^2_1 is the law of g^2 for a single $g \sim \mathcal{N}(0,1)$, and $\mathbb{E}\log g^2 = \psi(1/2) + \log 2 = -\gamma - \log 2$.

For matrices $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{p\times p}$, the *Kronecker product* $A \otimes B \in \mathbb{R}^{mp\times mp}$ is the block matrix $(A\otimes B)_{(i,k),(j,l)} = A_{ij}B_{kl}$ for $1 \leq i,j \leq m$ and $1 \leq k,l \leq p$. We use it to describe joint Gaussian laws on $(\mathbb{R}^d)^n$: if $R \in \mathbb{R}^{n\times n}$ is positive semidefinite and we identify $\mathbb{R}^{nd}$ with $(\mathbb{R}^d)^n$ by stacking, then $\mathcal{N}(0,\, R\otimes I_d)$ is the law of a random vector $\bar{z} = (z_1,\ldots,z_n) \in (\mathbb{R}^d)^n$ whose blocks are jointly centred Gaussian with cross-covariance $\mathrm{Cov}(z_s, z_t) = R_{st}\, I_d$; equivalently, the d output coordinates are i.i.d. across $k = 1,\ldots,d$, each an $\mathbb{R}^n$-valued centred Gaussian with covariance R.

4. Results

We consider the chain (1) driven by Gaussian process innovations. Specifically, each G_i is sampled i.i.d. from a centred vector-valued Gaussian process on $\mathbb{R}^d$ whose d output coordinates $G_i^{(1)},\ldots,G_i^{(d)}$ are themselves i.i.d. scalar centred GPs with the rotation-invariant radial basis function (RBF) kernel

$$k(x,x') \;=\; \phi(\|x-x'\|), \qquad \phi(t) \;:=\; e^{-t^2/(2r^2)}.$$

This setting is similar to the one considered in (Dunlop et al., 2018), although they consider a slightly more general family of kernels k. At any fixed $x \in \mathbb{R}^d$ the evaluation has law $G_i(x) \sim \mathcal{N}(0, \phi(0)\, I_d) = \mathcal{N}(0, I_d)$, independent of $\bar{Z}_i$, so every marginal trajectory satisfies $\bar{Z}_{i,t} \sim \mathcal{N}(0, I_d)$ at every depth $i \geq 1$. The depth-dependence lies entirely in the *joint* law across the n trajectories.

The first object we study is the pairwise distance for a fixed pair of coordinates $s \neq t$, held fixed throughout this and the next subsection. Let $V_i := \bar{Z}_{i,s} - \bar{Z}_{i,t} \in \mathbb{R}^d$ and $v_i := \|V_i\|$. Conditional on $\bar{Z}_i$, the pair $(\bar{Z}_{i+1,s}, \bar{Z}_{i+1,t})$ is jointly centred Gaussian with cross-covariance $\phi(v_i)\, I_d$, so

$$v_{i+1}^2 \;=\; 2\,(1-\phi(v_i))\; X_i, \qquad X_i \sim \chi^2_d, \tag{2}$$

with X_i independent of v_i by rotation invariance. Because the G_i are i.i.d. across i, the innovations $\{X_i\}_{i\geq 1}$ at this fixed pair are i.i.d., and the scalar chain $\{v_i^2\}$ is Markov on $[0,\infty)$.

The behaviour of (2) is governed by the local contraction rate near 0: for v_i small, $1-\phi(v_i) \approx v_i^2/(2r^2)$, so

$$v_{i+1}^2 \;\approx\; \frac{v_i^2}{r^2}\, X_i \qquad \text{for small } v_i, \tag{3}$$

and on the log-chain $L_i := \log v_i^2$ the increment $L_{i+1} - L_i \approx \log X_i - 2\log r$ is i.i.d. with mean

$$\rho_d \;:=\; \psi(d/2) + \log 2 - 2\log r \tag{4}$$

(using $\mathbb{E}\log X = \psi(d/2) + \log 2$; see Section 3). The sign of ρ_d separates two regimes, made precise in the following theorem.

Theorem 4.1 (Sharp dichotomy for the scalar pairwise-distance chain)**.** *Let*

$$r_c(d) \;:=\; \sqrt{2}\, e^{\psi(d/2)/2}$$

(where ψ is the digamma function), and let $\{v_i^2\}$ be the Markov chain on $[0,\infty)$ given by (2).

(i) If $r > r_c(d)$, then for every $v_1 \in \mathbb{R}$, $v_i \to 0$ almost surely, exponentially fast: with ρ_d as in (4),

$$\begin{aligned} \limsup_{i\to\infty} \frac{1}{i} \log v_i^2 &\le \rho_d \\ &= \psi(d/2) + \log 2 - 2\log r \\ &< 0 \quad \text{a.s.} \end{aligned}$$

(ii) If $r < r_c(d)$ and $v_1 \neq 0$, then $v_i \not\to 0$ almost surely, $\{v_i^2\}$ admits a unique nontrivial stationary distribution $\pi^{\mathbf{v}}$ on $(0,\infty)$, and the law of v_i^2 converges to $\pi^{\mathbf{v}}$ in total variation.

Remark 4.2. *The asymptotic behaviour is independent of the initial condition v_1: in (i) the exponential decay holds for every $v_1 \in \mathbb{R}$ with a v_1-independent rate bound.*

Sketch: supercritical regime ($\rho_d < 0$). The pointwise inequality $1 - e^{-x} \le x$ upgrades the near-origin linearization (3) to a *global* bound $v_{i+1}^2 \le v_i^2\, X_i/r^2$, valid for every v_i. Iterating and taking logs gives $L_i \le L_1 + \sum_{j<i}(\log X_j - 2\log r)$; the Kolmogorov strong law of large numbers (SLLN) on the i.i.d. sequence $\{\log X_j - 2\log r\}$ then forces $\limsup_i L_i/i \le \rho_d < 0$ a.s., so $v_i \to 0$ exponentially fast at rate at least $|\rho_d|$.

Sketch: subcritical regime ($\rho_d > 0$). Near $L_i = -\infty$ the linearization gives positive mean increment ρ_d, so the SLLN forces the log-chain to leave any far-left region in finite time; thus $v_i \not\to 0$. At the other end, the global bound $v_{i+1}^2 \le 2X_i$ (from $1 - e^{-x} \le 1$) caps the chain by an i.i.d. envelope independent of v_i, providing downward drift from far right. Using these facts we construct a Foster–Lyapunov function decreasing at both ends of L; together with ψ-irreducibility from the strictly positive transition density on $(0,\infty)$, standard arguments then yield positive Harris recurrence and a unique invariant probability on $(0,\infty)$ (see (Meyn & Tweedie, 1993)).

Convergence of $\bar{Z}_i$ for general n. Theorem 4.1 describes the scalar dynamics at any single pair (s,t). Controlling the full position chain $\bar{Z}_i \in (\mathbb{R}^d)^n$ requires more: each pairwise distance $u_{st}^{(i)} = \left\|\bar{Z}_{i,s} - \bar{Z}_{i,t}\right\|^2$ follows the scalar kernel marginally, but the pairs are coupled through the shared GP innovation G_i, so a unique joint stationary law needs a Foster–Lyapunov argument controlling all pairs simultaneously. We carry this out on the position chain itself, via a sum-of-squared-log-distances Lyapunov function whose drift decomposes pair-by-pair, while the Gaussian transition kernel supplies irreducibility and a small set on $(\mathbb{R}^d)^n$. This yields a unique stationary law $\pi_{\bar{Z}}$ and total-variation convergence of $\bar{Z}_i$, with the distance vector $\mathbf{u}^{(i)}$ inheriting a unique stationary law $\pi^{\mathbf{u}}$ (all pairwise marginals equal to $\pi^{\mathbf{v}}$) as its image. We work on $\bar{Z}_i$ rather than directly on $\mathbf{u}^{(i)}$ because for $n \ge d+2$ the distance vector is confined to a lower-dimensional set and carries no Lebesgue density (Remark 4.4).

Theorem 4.3 (TV convergence of $\bar{Z}_i$, general n)**.** *Let $\mathcal{Z} = \mathbb{R}^d$, $n \ge 2$, and G_i as in Theorem 4.1; assume $r < r_c(d)$. The pairwise-distance chain*

$$\mathbf{u}^{(i)} \;:=\; \left(u_{st}^{(i)}\right)_{s<t}, \qquad u_{st}^{(i)} := \left\|\bar{Z}_{i,s} - \bar{Z}_{i,t}\right\|^2,$$

is Markov on $(0,\infty)^{\binom{n}{2}}$ and admits a unique stationary distribution $\pi^{\mathbf{u}}$ whose pairwise coordinate marginals all equal $\pi^{\mathbf{v}}$ (the scalar stationary law from Theorem 4.1 (ii)). The full position chain $\{\bar{Z}_i\}$ on $(\mathbb{R}^d)^n$ admits a unique stationary law

$$\pi_{\bar{Z}} \;:=\; \int_{(0,\infty)^{\binom{n}{2}}} \mathcal{N}\big(0,\, R(\mathbf{u}) \otimes I_d\big)\, d\pi^{\mathbf{u}}(\mathbf{u}), \tag{5}$$

with $R(\mathbf{u})_{st} = \phi(\sqrt{u_{st}})$, and for every initial condition $\bar{Z}_1 \in (\mathbb{R}^d)^n$ with $\bar{Z}_{1,s} \neq \bar{Z}_{1,t}$ for all $s \neq t$,

$$\begin{aligned} \left\|\mathrm{Law}(\bar{Z}_i) - \pi_{\bar{Z}}\right\|_{\mathrm{TV}} &\le \left\|\mathrm{Law}(\mathbf{u}^{(i-1)}) - \pi^{\mathbf{u}}\right\|_{\mathrm{TV}} \\ &\xrightarrow{i\to\infty} 0. \end{aligned} \tag{6}$$

Remark 4.4 (State space of the distance chain)**.** *For $n \le d+1$ the squared-distance vector $\mathbf{u}^{(i)}$ ranges over a full-dimensional subset of $(0,\infty)^{\binom{n}{2}}$, but for $n \ge d+2$ it is constrained: squared Euclidean distance matrices of n points in $\mathbb{R}^d$ obey rank (Cayley–Menger) conditions and fill out only a submanifold of dimension $nd - \binom{d+1}{2} < \binom{n}{2}$. The law of $\mathbf{u}^{(i)}$ is then singular with respect to Lebesgue measure on $(0,\infty)^{\binom{n}{2}}$, so $\{\mathbf{u}^{(i)}\}$ is not irreducible there. We therefore establish ergodicity on the position chain $\{\bar{Z}_i\}$ on $(\mathbb{R}^d)^n$, whose Gaussian kernel $\mathcal{N}(0, R(\mathbf{u}(\bar{z})) \otimes I_d)$ has a strictly positive density, and obtain $\pi^{\mathbf{u}}$ as its image under $\bar{z} \mapsto \mathbf{u}(\bar{z})$; see Supplementary A. The experiments all operate in this regime (n in the hundreds or more, $d \le 100$).*

Theorem 4.5 ($\pi_{\bar{Z}}$ is marginally Gaussian but not jointly Gaussian)**.** *Assume $r < r_c(d)$, and let $(\bar{Z}_{\infty,1}, \ldots, \bar{Z}_{\infty,n}) \sim \pi_{\bar{Z}}$ be a sample from the stationary law of Theorem 4.3. Each coordinate is marginally standard Gaussian,*

$$\bar{Z}_{\infty,t} \;\sim\; \mathcal{N}(0, I_d) \qquad \text{for every } t \in \{1,\ldots,n\},$$

but for every pair $s \neq t$ the joint distribution of $(\bar{Z}_{\infty,s}, \bar{Z}_{\infty,t})$ is not *Gaussian. Consequently $\pi_{\bar{Z}}$ itself is not jointly Gaussian: if it were, every pair would be too, contradicting the previous claim.*

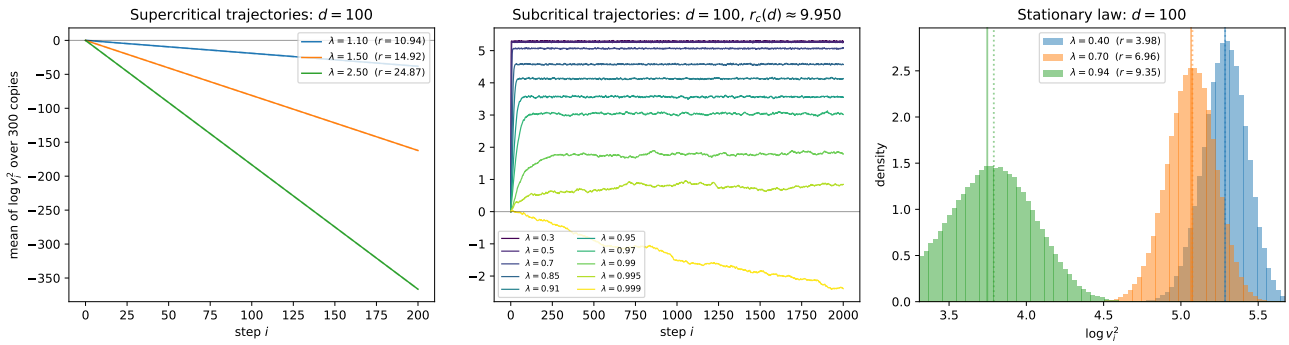

Figure 2. The dichotomy of Theorem 4.1 at $d = 100$. *Left (supercritical):* trajectories of mean $\log v_i^2$ over 300 i.i.d. copies starting at $v_1 = 1$, for $\lambda \in \{1.1, 1.5, 2.5\}$; the grey dashed reference line has slope $-2\log\lambda$ predicted by Theorem 4.1 (i) and is hidden behind the simulated curve in every case. *Middle (subcritical):* the same kind of trajectories for ten λ values spanning 0.30 to 0.999 (purple to yellow). All but the near-critical $\lambda = 0.999$ have equilibrated within the first few hundred steps; $\lambda = 0.999$ continues to drift, consistent with the slowdown of Theorem 4.1 (ii) as $\lambda \uparrow 1$. *Right:* stationary law of $\log v_i^2$ at $\lambda \in \{0.40, 0.70, 0.94\}$, from a chain of $2 \cdot 10^5$ samples after $2 \cdot 10^4$-step burn-in. Solid verticals: empirical stationary mean. Dotted verticals: Jensen upper bound $L_*(r, d)$ from Proposition 4.7 (b). The bound is nearly tight at every λ. The $d = 1, 10$ companions are in Supplementary B.

The pairwise non-Gaussianity holds because, as we show, no distribution of v_i arising from a Gaussian difference of two coordinates can be stationary for (2).

Argument for the proof of Theorem 4.5. *(1, marginal.)* As discussed above, the marginal $\bar{Z}_{i,t}$ is Gaussian, $\bar{Z}_{i,t} \sim \mathcal{N}(0, I_d)$, at every finite step, and the TV convergence $\mathrm{Law}(\bar{Z}_i) \to \pi_{\bar{Z}}$ from Theorem 4.3 forces every coordinate marginal under $\pi_{\bar{Z}}$ to be standard Gaussian. *(2, isotropy.)* Fix any $s \neq t$. The one-step dynamics make $V_i := \bar{Z}_{i,s} - \bar{Z}_{i,t}$ rotationally invariant in $\mathbb{R}^d$ at every finite i (its conditional covariance given $\bar{Z}_{i-1}$ is a scalar multiple of I_d), and rotational invariance passes to the weak limit V_∞. If now $(\bar{Z}_{\infty,s}, \bar{Z}_{\infty,t})$ were Gaussian, V_∞ would be a centred isotropic Gaussian, hence $V_\infty \sim \mathcal{N}(0, \sigma_v^2 I_d)$ for some $\sigma_v^2 \in (0, \infty)$, and $v_\infty^2 := \|V_\infty\|^2$ would be a scaled χ_d^2. *(3, punchline.)* No scaled χ_d^2 is stationary for (2) — proven via characteristic functions in Supplementary A — contradicting the stationarity of $\pi^{\mathbf{v}}$ in Theorem 4.1 (ii). The pair (s, t) was arbitrary, so this rules out joint Gaussianity for every pair under $\pi_{\bar{Z}}$, and hence for $\pi_{\bar{Z}}$ itself.

Remark 4.6. *The map $d \mapsto r_c(d) = \sqrt{2}\, e^{\psi(d/2)/2}$ is increasing, and asymptotically $\psi(d/2) = \log(d/2) - 1/d + O(1/d^2)$, so $r_c(d) \sim \sqrt{d}$ as $d \to \infty$.*

Proposition 4.7 (Moments of $\log v^2$ in the subcritical regime)**.** *Assume $r < r_c(d)$. Under $\pi^{\mathbf{v}}$ (the stationary law of $\{v_i^2\}$ on $(0, \infty)$ from Theorem 4.1 (ii)), write $L := \log v^2$; then $\mathbb{E}_{\pi^{\mathbf{v}}}|L| < \infty$, and*

(a) Stationarity identity. $\mathbb{E}_{\pi^{\mathbf{v}}} H(L) = -\psi(d/2) - \log 2$, *where $H(L) := \log\left(F(e^L)/e^L\right)$ and $F(u) = 2(1 - e^{-u/(2r^2)})$.*

(b) Jensen upper bound. $\mathbb{E}_{\pi^{\mathbf{v}}} \log v^2 \leq L_*(r, d)$, *where L_* is the unique root of $H(L_*) = -\psi(d/2) - \log 2$; explicitly*

$$L_*(r, d) = \log\left(2r^2\alpha\right), \tag{7}$$

with $\alpha \in (0, \infty)$ the unique solution of

$$\frac{1 - e^{-\alpha}}{\alpha} = (r/r_c(d))^2 .$$

$L_*(r, d) \to -\infty$ as $r \uparrow r_c(d)$; in particular $\mathbb{E}_{\pi^{\mathbf{v}}} \log v^2 \to -\infty$ at criticality.

Corollary 4.8 (Dimension factorization of L_* at matched subcriticality)**.** *Fix $\lambda \in (0, 1)$ and set $r = \lambda r_c(d)$. The Jensen bound of Proposition 4.7 (b) factorizes as*

$$L_*(\lambda r_c(d), d) = \psi(d/2) + 2\log 2 + \log\left(\lambda^2 \alpha(\lambda)\right),$$

where $\alpha(\lambda) \in (0, \infty)$ is the unique solution of $(1 - e^{-\alpha(\lambda)})/\alpha(\lambda) = \lambda^2$ (universal in d).

Corollary 4.9 (Persistence of coupling: comparison with independent Gaussians in $\mathbb{R}^d$)**.** *Consider a null model in which a pair of coordinates has the same marginals as our process $\bar{Z}_{i,t}$ but is independent: $\tilde{Z}_1, \tilde{Z}_2 \sim \mathcal{N}(0, I_d)$ and $\tilde{Z}_1 \perp \tilde{Z}_2$. Then $\tilde{v}^2 = \left\|\tilde{Z}_1 - \tilde{Z}_2\right\|^2 \sim 2\chi_d^2$ and*

$$\mathbb{E}\log\tilde{v}^2 = \psi(d/2) + 2\log 2 = \log\left(2\, r_c(d)^2\right). \tag{8}$$

In the notation of Corollary 4.8, at $r = \lambda r_c(d)$ with $\lambda \in (0, 1)$, the Jensen bound sits strictly below *this benchmark:*

$$L_*(\lambda r_c(d), d) - \mathbb{E}\log\tilde{v}^2 = \log\left(\lambda^2 \alpha(\lambda)\right) < 0. \tag{9}$$

The deficit $\log[\lambda^2\alpha(\lambda)]$ is independent of d *and diverges to $-\infty$ as $\lambda \uparrow 1$ (critical threshold).*

A non-zero deficit witnesses dependence directly: the marginals $\bar{Z}_{\infty,1}, \bar{Z}_{\infty,2} \sim \mathcal{N}(0, I_d)$ are fixed by the chain,

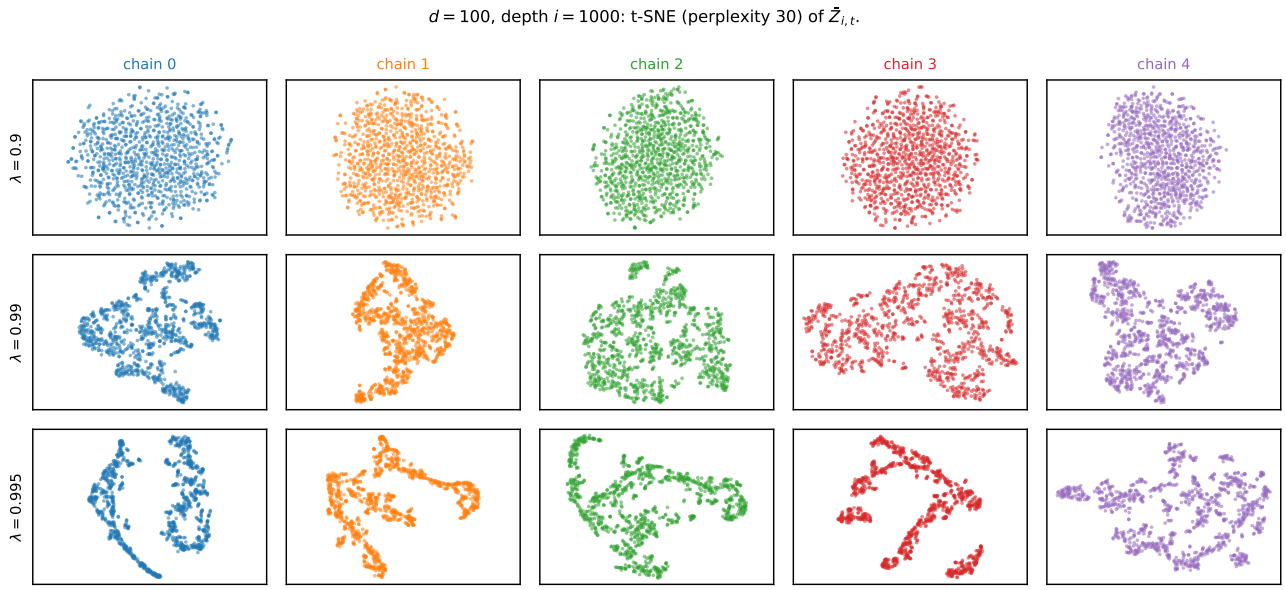

Figure 3. Per-chain t-SNE embeddings of $\bar{Z}_{i,t} \in \mathbb{R}^{100}$ at depth $i = 1000$, for three subcritical λ (rows) and five i.i.d. chains (columns); $n = 1000$ points per panel, perplexity 30, PCA initialisation. The visible-structure window in $d = 100$ is narrow: $\lambda = 0.90$ (top row) embeds as a featureless isotropic blob in every chain, indistinguishable across chains; $\lambda = 0.99$ and $\lambda = 0.995$ (lower rows) produce chain-specific multimodal structure — the non-Gaussianity of $\pi_{\bar{Z}}$ from Theorem 4.5 made visible. The same chain-specific structure is recovered by linear PCA (Figure 12 in Supplementary B), so the patterns here are not an artefact of the t-SNE embedding.

so under independence one would have $v^2 \sim 2\chi^2_d$ exactly and $\mathbb{E}\log v^2 = \log[2\,r_c(d)^2]$. The strict deficit $\log[\lambda^2\,\alpha(\lambda)] < 0$ therefore rules out independence between $\bar{Z}_{\infty,1}$ and $\bar{Z}_{\infty,2}$. In particular, for any d, the dependence can be made arbitrarily strong by taking $\lambda \in (0,1)$ large enough.

We observe that, as expected, by taking $\lambda \uparrow 1$ in (9) we approach the supercritical full-dependence regime, where $\bar{Z}_{\infty,1} = \bar{Z}_{\infty,2}$ and $\mathbb{E}\log v^2 = -\infty$.

The strict negativity in (9) is a consequence of the defining equation for $\alpha(\lambda)$: multiplying $(1 - e^{-\alpha(\lambda)})/\alpha(\lambda) = \lambda^2$ by $\alpha(\lambda)$ gives $\lambda^2\,\alpha(\lambda) = 1 - e^{-\alpha(\lambda)} < 1$.

5. Experiments

In this section we verify the threshold $r_c(d)$ empirically across $d \in \{1, 10, 100\}$, and then use the bounds of Section 4 to choose λ per dimension and explore the unstructured (Gaussian-baseline) and structured (chain-specific multimodal) regimes of $\pi_{\bar{Z}}$.

For the threshold check we simulate the scalar pairwise-distance chain (2) in log-space ($L_i = \log v_i^2$) to avoid spurious absorption at 0 when v_i becomes moderately small. Throughout we parametrize $r = \lambda\,r_c(d)$, so the asymptotic rate of decay of L_i, $-\rho_d = 2\log r - \psi(d/2) - \log 2$ (see Theorem 4.1 (i)), equals $-2\log\lambda$ *independently of* d.

Figure 2 confirms the dichotomy of Theorem 4.1 at $d = 100$; the analogous figures at $d = 1, 10$ are provided in supplementary Supplementary B, Figures 4 to 6, and exhibit a generally similar behaviour. Above $r_c(d)$ (left pane of Figure 2), the mean of $\log v_i^2$ decays linearly at the predicted rate $-2\log\lambda$, with the dashed reference slope hidden behind the simulated curve. Below $r_c(d)$ (middle pane of Figure 2), the mean of $\log v_i^2$ stabilises at a fixed value at every λ we tested except the most near-critical, consistent with convergence; the stabilisation slows as $\lambda \uparrow 1$, and at $\lambda = 0.999$ the trajectory is still drifting at depth 2000. The stationary law (right pane of Figure 2) acquires a heavier left tail as $\lambda \uparrow 1$, with the Jensen bound L_* nearly tight at $d = 100$. At low d the bound becomes loose: at $d = 1, \lambda = 0.94$ the empirical mean sits markedly below L_* (Supplementary Figure 6).

From bounds to a choice of λ. The theory of Section 4 gives more than a sharp threshold: it tells us how to choose λ per dimension to see informative structure under $\pi_{\bar{Z}}$. The relevant tool is the i.i.d.-Gaussian comparison of Corollary 4.9: at $r = \lambda\,r_c(d)$, the stationary mean of $\log v^2$ sits below its i.i.d.-Gaussian benchmark by exactly $\log[\lambda^2\,\alpha(\lambda)] < 0$ (see eq. (9)), which is d-independent at fixed λ. The benchmark itself, $\log[2\,r_c(d)^2] \sim \log d$, grows with dimension, so for $\bar{Z}$ to look distinguishable from the independent-Gaussian null we would like the deficit to be of similar order. This is what fixes the appropriate λ at each dimension: a moderate λ suffices at small d, but at

large d, $\alpha(\lambda)$ has to shrink to make the deficit $\log d$-large, forcing λ close to 1. Concretely, the defining equation $(1 - e^{-\alpha(\lambda)})/\alpha(\lambda) = \lambda^2$ gives $\alpha(\lambda) \approx 2(1 - \lambda^2)$ as $\lambda \uparrow 1$, and hence

$$\log[\lambda^2 \alpha(\lambda)] \;\approx\; \log[2(1 - \lambda^2)]. \tag{10}$$

The deficit therefore reaches order $-\log d$ at $1 - \lambda^2 \sim 1/d$, i.e. $\lambda \approx 1 - 1/(2d)$ at large d. This estimate agrees with the experiments below: the visible-structure window we observe at $d = 100$ sits at $\lambda \in (0.99, 1)$ and at $d = 10$ at $\lambda \in (0.91, 0.97)$, both consistent with $1 - \lambda \asymp 1/d$.

Sampling from $\pi_{\bar{Z}}$. We simulate the position chain $\bar{Z}_i \in (\mathbb{R}^d)^n$ for $n = 1000$ and visualise the resulting cloud one chain at a time, with λ values chosen per dimension to span the regime transition from a near-independent Gaussian baseline to chain-specific multimodal structure: $d = 1$ at $\lambda \in \{0.10, 0.30, 0.60, 0.85\}$, $d = 10$ at $\lambda \in \{0.66, 0.91, 0.95, 0.97\}$, $d = 100$ at $\lambda \in \{0.90, 0.99, 0.995\}$. Five i.i.d. chains per λ. Figure 3 shows the $d = 100$ case as a per-chain t-SNE embedding at depth $i = 1000$. Note that according to Figure 2, for the λ in the above range, at that depth the chains have already stabilised. At the smallest λ ($\lambda = 0.90$, top row) the cloud is an approximately isotropic blob and the chains are visually indistinguishable from one another; as λ approaches 1 the embeddings fragment into chain-specific multimodal structure — clusters, strings, loops — in the narrow strip $\lambda \in (0.99, 1)$, with the structure becoming more pronounced for higher λ. The structure and the run-to-run variation of the chains visualise the non-Gaussianity of $\pi_{\bar{Z}}$ established in Theorem 4.5: a Gaussian limit would force every chain to look like a single common ellipsoidal cloud. Note that the structure visible at large λ is not an artefact of the embedding: all the figures use the same (default) t-SNE parameters, and at small λ the same t-SNE pipeline returns an essentially featureless blob. Moreover, the corresponding linear PCA view in Supplementary B (Figure 12) shows similar chain-specific bananas, U-shapes, and arcs without any non-linear embedding step. The analogous figures at $d = 1$ (per-chain histograms) and $d = 10$ (PCA and t-SNE, where the visible-structure window is wider and the transition smoother) are reported in Supplementary B; as with the $d = 100$ case, at $d = 1, 10$ the chain-specific structure becomes more pronounced as λ enters its d-dependent near-critical strip.

6. Conclusions and Future Work

We have established the existence and several basic properties of the depth-infinite limit $\bar{Z}_\infty$ of the compositional-GP prior chain $\bar{Z}_i$. For the RBF kernel we identified a sharp threshold $r_c(d) = \sqrt{2}\, e^{\psi(d/2)/2}$ separating degeneration into a constant function from survival of a nontrivial limit; below the threshold we showed that $\bar{Z}_\infty$ exists, is not Gaussian, and its components exhibit nontrivial dependence, so that the prior retains genuine structure at depth.

Several directions remain open. First, while we have established the existence and uniqueness of the limit $\pi_{\bar{Z}}$, the *rate* of convergence $\mathrm{Law}(\bar{Z}_i) \to \pi_{\bar{Z}}$ remains open; establishing quantitative, and ideally uniform, rates — with prefactors independent of the initial state $\bar{Z}_1$ — would be a natural next step and is a prerequisite for any finite-depth application. The question connects to a substantial literature on quantitative ergodicity for iterated random functions (Diaconis & Freedman, 1999; Wu & Shao, 2004; Stenflo, 2012), though the infinite-dimensional GP setting we consider falls outside the parametric families that line of work has typically addressed.

Second, the structure of $\pi_{\bar{Z}}$ itself is largely unexplored beyond the basic properties we establish here, and analysing its dependence structure in more detail is a natural next step. A concrete and likely tractable question is whether $\pi_{\bar{Z}}$ belongs to any known family of probability distributions, or constitutes a genuinely new family — the empirical multimodality observed in Section 5 suggests it is not a simple modification of a familiar parametric class.

Third, the arguments we use here appear not to apply at the critical bandwidth value $r = r_c(d)$ itself, and it will be interesting to understand the behaviour there as well. Our results are also specific to the RBF kernel; whether the same dichotomy — a sharp threshold separating synchronisation from a non-degenerate limit — holds for other kernel families (e.g. Matérn) is an obvious open problem. The arguments here use the RBF kernel structure both for the global one-step bound $1 - \phi(v) \le v^2/(2r^2)$ underlying the supercritical SLLN and for the boundary linearisation used in the Foster–Lyapunov drift; replacing or generalising these with kernel-agnostic arguments would extend the picture beyond the Gaussian-form case.

References

Abramowitz, M. and Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, 1964.

Bordino, A., Favaro, S., and Fortini, S. Infinitely wide limits for deep Stable neural networks: sub-linear, linear and super-linear activation functions. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=A5tIluhDW6.

Bui, T. D., Hernández-Lobato, D., Hernández-Lobato, J. M., Li, Y., and Turner, R. E. Deep Gaussian processes for regression using approximate expectation propagation.

In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Cutajar, K., Bonilla, E. V., Michiardi, P., and Filippone, M. Random feature expansions for deep Gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Damianou, A. and Lawrence, N. D. Deep Gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 31 of *Proceedings of Machine Learning Research*, pp. 207–215. PMLR, 2013.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research*, 17(42):1–62, 2016.

Diaconis, P. and Freedman, D. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999.

Dunlop, M. M., Girolami, M. A., Stuart, A. M., and Teckentrup, A. L. How deep are deep Gaussian processes? *Journal of Machine Learning Research*, 19(54):1–46, 2018.

Durrett, R. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.

Duvenaud, D., Rippel, O., Adams, R. P., and Ghahramani, Z. Avoiding pathologies in very deep networks. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

Fortuin, V. Priors in Bayesian deep learning: A review. *International Statistical Review*, 90(3):563–591, 2022. arXiv:2105.06868.

Havasi, M., Hernández-Lobato, J. M., and Murillo-Fuentes, J. J. Inference in deep Gaussian processes using stochastic gradient Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Hayou, S. On the infinite-depth limit of finite-width neural networks. *Transactions on Machine Learning Research*, 2023. arXiv:2210.00688.

Hayou, S., Doucet, A., and Rousseau, J. On the impact of the activation function on deep neural networks training. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Lawrence, N. D. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2018.

Li, M., Nica, M., and Roy, D. M. The neural covariance SDE: Shaped infinite depth-and-width networks at initialization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Lu, C.-K., Yang, S. C.-H., Hao, X., and Shafto, P. Interpretable deep Gaussian processes with moments. *arXiv preprint arXiv:1905.10963*, 2019.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Meyn, S. P. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag London, 1993.

Neal, R. M. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1996. Lecture Notes in Statistics, vol. 118, Springer.

Peluchetti, S., Favaro, S., and Fortini, S. Stable behaviour of infinitely wide deep neural networks. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286. PMLR, 2014.

Saatci, Y. and Wilson, A. G. Bayesian GAN. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Salimbeni, H. and Deisenroth, M. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. In *International Conference on Learning Representations (ICLR)*, 2017.

Stenflo, Ö. A survey of average contractive iterated function systems. *Journal of Difference Equations and Applications*, 18(8):1355–1380, 2012.

Titsias, M. K. and Lawrence, N. D. Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Tran, B.-H., Rossi, S., Milios, D., and Filippone, M. All you need is a good functional prior for Bayesian deep learning. *Journal of Machine Learning Research*, 23(74): 1–56, 2022.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Wu, W. B. and Shao, X. Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436, 2004.

Yang, G. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. In *arXiv preprint arXiv:1902.04760*, 2019.

A. Proofs

The dichotomy of Theorem 4.1 is proved in two stages. We first treat the scalar case $d = 1$ (Proposition A.1 below), where the innovation is g_i^2 with $g_i \sim \mathcal{N}(0,1)$; all of the analytic work is here. The $\mathbb{R}^d$ case is then recovered by substituting $X_i \sim \chi_d^2$ for g_i^2 throughout: the functional form of the log-chain, the Foster–Lyapunov function, and the structural ingredients (irreducibility, aperiodicity, small sets) are unchanged, and only the noise mean $\mathbb{E}\log g^2 = -\gamma - \log 2$ is replaced by $\mathbb{E}\log X = \psi(d/2) + \log 2$.

For $d = 1$, the recursion (2) reduces to

$$v_{i+1} \;=\; \sqrt{2(1-\phi(v_i))}\, g_i, \qquad g_i \sim \mathcal{N}(0,1) \text{ i.i.d.}, \tag{11}$$

with $v_i := Z_{i,1} - Z_{i,2} \in \mathbb{R}$.

Proposition A.1 (Scalar dichotomy, $d = 1$)**.** *Let $\{v_i\}_{i\geq 1}$ be the chain defined by (11) with $v_1 \in \mathbb{R}$ arbitrary, and let*

$$r_c \;:=\; \frac{1}{\sqrt{2}\, e^{\gamma/2}} \;=\; \sqrt{\frac{1}{2e^{\gamma}}} \;\approx\; 0.5298.$$

(i) If $r > r_c$, then for every $v_1 \in \mathbb{R}$, $v_i \to 0$ almost surely, and

$$\limsup_{i\to\infty} \frac{1}{i}\, \log v_i^2 \;\leq\; -\gamma - \log 2 - 2\log r \;<\; 0 \quad \textit{a.s.}$$

(ii) If $r < r_c$ and $v_1 \neq 0$, then $v_i \not\to 0$ almost surely, and $\{v_i^2\}$ admits a unique nontrivial stationary distribution on $(0,\infty)$, independent of v_1.

Proof of Proposition A.1. Set $u_i := v_i^2 \geq 0$, $L_i := \log u_i$, and $F(u) := 2(1 - e^{-u/(2r^2)})$. Then (11) reads

$$u_{i+1} \;=\; F(u_i)\, g_i^2. \tag{12}$$

If $u_1 = 0$ then $u_i \equiv 0$; otherwise $u_i > 0$ a.s. for all i (since $g_i \neq 0$ a.s.), and (12) gives

$$L_{i+1} \;=\; L_i + H(L_i) + \log g_i^2, \tag{13}$$

where $H(L) := \log\big(F(e^L)/e^L\big)$. The function H is continuous on $\mathbb{R}$, satisfies $H(L) \to -2\log r$ as $L \to -\infty$ (since $F(u)/u \to 1/r^2$ as $u \to 0$), and obeys the uniform upper bound

$$H(L) \;\leq\; -2\log r \qquad \text{for all } L \in \mathbb{R}, \tag{14}$$

which follows from $1 - e^{-x} \leq x$. For $g \sim \mathcal{N}(0,1)$, $\mathbb{E}\log g^2 = \psi(1/2) + \log 2 = -\gamma - \log 2$; set $\rho := -\gamma - \log 2 - 2\log r$, so that $r > r_c$ iff $\rho < 0$, and $r < r_c$ iff $\rho > 0$.

Proof of (i). Assume $r > r_c$, so $\rho < 0$. From (13) and (14),

$$L_i \;\leq\; L_1 + \sum_{j=1}^{i-1} \left(-2\log r + \log g_j^2\right).$$

We invoke Kolmogorov's strong law of large numbers for i.i.d. sequences (Theorem 2.4.1 of (Durrett, 2019)): if $\{Y_j\}_{j\geq 1}$ are i.i.d. with $\mathbb{E}|Y_1| < \infty$, then $n^{-1}\sum_{j=1}^n Y_j \to \mathbb{E}Y_1$ almost surely. With $Y_j := -2\log r + \log g_j^2$ the sequence is i.i.d. (since the g_j are), and the integrability hypothesis $\mathbb{E}|\log g^2| < \infty$ holds by Remark A.2. The SLLN therefore yields $(i-1)^{-1}\sum_{j=1}^{i-1}(-2\log r + \log g_j^2) \to \rho$ a.s., whence $\limsup_{i\to\infty} L_i/i \leq \rho < 0$ a.s. So $u_i \to 0$ a.s. at exponential rate at least $|\rho|$, i.e. $v_i \to 0$ a.s.

Proof of (ii), non-convergence. Assume $r < r_c$, so $\rho > 0$. Fix $\varepsilon \in (0, \rho/2)$. Since $H(L) \to -2\log r$ as $L \to -\infty$, there exists $L_* < 0$ such that

$$H(L) \;\geq\; -2\log r - \varepsilon \qquad \text{for all } L \leq L_*; \tag{15}$$

write $\delta := e^{L_*} > 0$. Fix $u_1 > 0$ and define the hitting time

$$\tau \;:=\; \inf\{i \geq 1 : u_i > \delta\}\,.$$

We claim $\tau < \infty$ a.s. On the event $\{\tau = \infty\}$, $L_i \leq L_*$ for every i, so (13) and (15) give the pathwise lower bound

$$\begin{aligned} L_{n+1} &\geq L_1 + \sum_{j=1}^{n} \left(-2\log r - \varepsilon + \log g_j^2\right) \\ &= L_1 + n\,(\rho - \varepsilon) + S_n, \end{aligned} \tag{16}$$

where $S_n := \sum_{j=1}^{n}(\log g_j^2 - \mathbb{E}\log g^2)$ is a zero-mean random walk with i.i.d. increments. By the same SLLN as in part (i), applied *unconditionally* to the i.i.d. sequence $\left\{\log g_j^2 - \mathbb{E}\log g^2\right\}_{j\geq 1}$ on the underlying probability space, $\mathbb{P}(\Omega_0) = 1$ for $\Omega_0 := \{S_n/n \to 0\}$; we do *not* condition on $\{\tau = \infty\}$ before invoking the SLLN, since under that conditioning the g_j would no longer be i.i.d. Now consider any sample point $\omega \in \{\tau = \infty\} \cap \Omega_0$. The pathwise bound (16) together with $S_n(\omega)/n \to 0$ and $\rho - \varepsilon > \rho/2 > 0$ force $L_{n+1}(\omega) \to +\infty$, contradicting $L_{n+1}(\omega) \leq L_*$ for all n (which holds since $\omega \in \{\tau = \infty\}$). Hence $\{\tau = \infty\} \cap \Omega_0 = \emptyset$, and combined with $\mathbb{P}(\Omega_0) = 1$ this gives $\mathbb{P}(\tau = \infty) = 0$.

Iterate: set $\tau_0 := 0$ and $\tau_{k+1} := \inf\{i > \tau_k : u_i > \delta\}$. By the strong Markov property applied at $\tau_k + 1$ (from which point the chain restarts from the positive state u_{τ_k+1}), the same argument gives $\tau_{k+1} < \infty$ a.s. on $\{\tau_k < \infty\}$. By induction every τ_k is a.s. finite, so $u_i > \delta$ for infinitely many i a.s.; in particular

$$\limsup_{i\to\infty} u_i \;\geq\; \delta \;>\; 0 \qquad \text{a.s.},$$

so $v_i = \pm\sqrt{u_i}$ does *not* converge to 0 a.s.

Proof of (ii), stationary distribution. We establish positive Harris recurrence via a Foster–Lyapunov drift argument; we briefly recall the framework before applying it.

Let $\{X_i\}$ be a Markov chain on a state space $\mathcal{X}$ with transition kernel $P(x, \cdot)$. The chain is *ψ-irreducible* if there exists a non-trivial measure ψ on $\mathcal{X}$ such that for every $x \in \mathcal{X}$ and every measurable A with $\psi(A) > 0$, the chain reaches A from x with positive probability in some finite number of steps; intuitively, no part of the state space is invisible from any starting point. A set $C \subset \mathcal{X}$ is *small* if there exist $m \geq 1$, $\eta > 0$, and a probability measure ν such that $P^m(x, \cdot) \geq \eta\,\nu(\cdot)$ for every $x \in C$, i.e. C is a uniform regeneration region: from anywhere in C, the m-step law dominates a fixed probability measure. For a ψ-irreducible chain on a general state space there is a notion of *period* $d \geq 1$, defined via the d-cycle of small sets associated with the chain (see Section 5.4 of (Meyn & Tweedie, 1993), where the period is defined for general state-space ψ-irreducible chains), and the chain is *aperiodic* precisely when $d = 1$. A convenient sufficient condition — which we will use — is that some small set C admit a one-step minorisation $P(x, \cdot) \geq \eta\,\nu$ for every $x \in C$ (i.e. the minorising step m above can be taken equal to 1); a chain satisfying this is called *strongly aperiodic*, and strong aperiodicity implies aperiodicity.

Two results from (Meyn & Tweedie, 1993) combine to give positive recurrence and convergence to equilibrium under a Lyapunov drift condition. First, *Foster's drift criterion* (Theorem 11.3.4 of (Meyn & Tweedie, 1993)): if the chain is ψ-irreducible and there exist a measurable function $V : \mathcal{X} \to [0, \infty)$, a small set C, and constants $c > 0$, $b < \infty$, such that the drift condition

$$\mathbb{E}V(X_{i+1}) \mid X_i = x \;\leq\; V(x) - c + b\,\mathbf{1}_C(x) \tag{17}$$

holds for every $x \in \mathcal{X}$, then the chain is positive Harris recurrent and admits a unique invariant probability measure π. Second, the *Aperiodic Ergodic Theorem* (Theorem 13.0.1 of (Meyn & Tweedie, 1993)): once the chain is positive Harris recurrent, aperiodic, and has invariant probability π, the law of X_i converges to π in total variation from every starting point $x \in \mathcal{X}$,

$$\left\|P^i(x, \cdot) - \pi\right\|_{TV} \to 0 \qquad \text{as } i \to \infty.$$

The function V plays the role of a Lyapunov function: it measures a generalised "distance" to the small set C, and (17) asserts that this distance shrinks on average by at least c outside C, while inside C it may grow but only by a bounded

amount b. Together these prevent both escape to infinity and trapping at the boundary of $\mathcal{X}$, forcing the chain to return to C infinitely often with controlled hitting times — which is exactly what positive recurrence requires.

We apply this framework to the chain $\{u_i\}$ on $\mathcal{X} = (0, \infty)$ with transition $u \mapsto F(u)\, g^2$. Since g^2 has a smooth strictly positive Lebesgue density on $(0, \infty)$ and $F(u) > 0$ for $u > 0$, the one-step transition law of u_{i+1} given $u_i = u$ has a strictly positive Lebesgue density on $(0, \infty)$, smoothly depending on u. This has three consequences. *(a)* The chain is ψ-irreducible with respect to Lebesgue measure on $(0, \infty)$: any set of positive Lebesgue measure has positive one-step transition probability from every $u \in (0, \infty)$. *(b)* Every compact $K \subset (0, \infty)$ is a small set with $m = 1$: the one-step density is bounded below by some $\eta_K > 0$ uniformly on $K \times K$ by continuity, hence $P(u, \cdot) \geq \eta_K\, |K| \cdot \mathrm{Unif}(K)$ for every $u \in K$. *(c)* The chain is aperiodic: $m = 1$ in (b) gives the strong aperiodicity condition. Since $L = \log u$ is a homeomorphism between $(0, \infty)$ and $\mathbb{R}$, all of these properties transfer verbatim to the log-chain $\{L_i\}$ on $\mathbb{R}$ (with Lebesgue measure pulled back to Lebesgue on $\mathbb{R}$).

For the Lyapunov function we work in log-coordinates, where the multiplicative dynamics become additive (cf. (13)). Define

$$V(L) \;:=\; (L_0 - L)^+ + (L - L^0)^+,$$

where $L_0 > L_*$ and $L^0 > L_0$ are to be fixed. This is a "tent" function: $V \equiv 0$ on the strip $L \in [L_0, L^0]$, and V grows linearly outside it (with slope -1 to the left of L_0 and slope $+1$ to the right of L^0). Geometrically, $V(L)$ is the distance from L to the strip $[L_0, L^0]$, which in u-coordinates is the compact interval $[e^{L_0}, e^{L^0}] \subset (0, \infty)$ — a small set by the previous paragraph. Verifying (17) thus amounts to showing that $\mathbb{E}V(L_{i+1}) \mid L_i = L < V(L) - c$ for L in the two tails $\{L \leq L_*\}$ and $\{L \geq L^0 + 2\kappa\}$, with the in-between region $C := [L_*, L^0 + 2\kappa]$ acting as the small set on which V is allowed to grow by the constant b. The mechanisms in the two tails are different: on the *left* ($L \leq L_*$) the linearisation-based positive drift $\rho - \varepsilon$ on L from (15) pushes L_{i+1} rightward, decreasing V; on the *right* ($L \geq L^0$) the global cap $u_{i+1} \leq 2g_i^2$ resets L_{i+1} to a state-independent random level of bounded mean, which sits far below L^0 once L^0 is large, again decreasing V. We verify these now.

Region 1: $L \leq L_*$. Here $V(L) = L_0 - L$. Using $(a)^+ = a + (-a)^+$,

$$\begin{aligned} &\mathbb{E}(L_0 - L_{i+1})^+ \mid L_i = L \\ &\quad = \quad L_0 - \mathbb{E}L_{i+1} \mid L_i = L \\ &\qquad\quad + \mathbb{E}(L_{i+1} - L_0)^+ \mid L_i = L. \end{aligned}$$

By (13) and (15), $\mathbb{E}L_{i+1} \mid L_i = L = L + H(L) - \gamma - \log 2 \geq L + (\rho - \varepsilon)$. By (14), $L_{i+1} \leq L - 2\log r + \log g_i^2$, so for $L \leq L_*$,

$$\mathbb{E}(L_{i+1} - L_0)^+ \mid L_i = L \quad \leq \quad \mathbb{E}(\log g^2 - M)^+,$$

where $M := L_0 - L_* + 2\log r$. Since $\log g^2$ has finite mean, $\mathbb{E}(\log g^2 - M)^+ \to 0$ as $M \to \infty$; choose L_0 large enough that this is $\leq (\rho - \varepsilon)/4$, and by the same bound applied with L^0 in place of L_0, also $\mathbb{E}(L_{i+1} - L^0)^+ \mid L_i = L \leq (\rho - \varepsilon)/4$ for $L \leq L_*$. Using the identity $(L_0 - L_{i+1})^+ = (L_0 - L_{i+1}) + (L_{i+1} - L_0)^+$ together with $L_0 - \mathbb{E}L_{i+1} \mid L_i = L \leq L_0 - L - (\rho - \varepsilon) = V(L) - (\rho - \varepsilon)$ (valid since $V(L) = L_0 - L$ for $L \leq L_*$), we obtain

$$\begin{aligned} &\mathbb{E}(L_0 - L_{i+1})^+ \mid L_i = L \\ &\quad \leq \quad V(L) - (\rho - \varepsilon) + (\rho - \varepsilon)/4 \\ &\quad = \quad V(L) - 3(\rho - \varepsilon)/4. \end{aligned}$$

Adding $\mathbb{E}(L_{i+1} - L^0)^+ \mid L_i = L \leq (\rho - \varepsilon)/4$ gives the drift bound

$$\mathbb{E}V(L_{i+1}) \mid L_i = L \quad \leq \quad V(L) - (\rho - \varepsilon)/2 \tag{18}$$

for all $L \leq L_*$, i.e. $c_1 := (\rho - \varepsilon)/2 > 0$ on $\{L \leq L_*\}$.

Region 2: $L \geq L^0$. Here $V(L) = L - L^0$. From (12), $u_{i+1} \leq 2g_i^2$, so $L_{i+1} \leq \log 2 + \log g_i^2$ regardless of L_i. On Region 2 the lower bound $u_{i+1} \geq F(e^{L^0})\, g_i^2$ also holds (since F is increasing and $u_i \geq e^{L^0}$), giving $L_{i+1} \geq \log F(e^{L^0}) + \log g_i^2$. Combining, for every $L_i \geq L^0$ the new state satisfies

$$|L_{i+1}| \;\leq\; \max\Big(\log 2,\, -\log F(e^{L^0})\Big) + \big|\log g_i^2\big| \;=:\; W_i,$$

an envelope independent of L_i with finite mean (since $\mathbb{E}|\log g^2| < \infty$ by Remark A.2). Hence

$$V(L_{i+1}) \;\le\; |L_0 - L_{i+1}| + \left|L_{i+1} - L^0\right| \;\le\; 2W_i + L_0 + L^0,$$

so that

$$\kappa \;:=\; \mathbb{E}2W_i + L_0 + L^0 \;<\; \infty,$$

and $\mathbb{E}V(L_{i+1}) \mid L_i = L \le \kappa$ for every $L \ge L^0$. The constant κ is fixed by the chain and the choices of L_0, L^0.

Now we use this κ to delimit the small set. We want a strictly negative drift $\le -\kappa$ in Region 2, i.e. $\mathbb{E}V(L_{i+1}) \mid L_i = L \le V(L) - \kappa$, which (using the uniform bound above) holds whenever $V(L) \ge 2\kappa$. Since $V(L) = L - L^0$ in Region 2, this requires $L \ge L^0 + 2\kappa$ — and this is precisely why the small set is taken to extend up to $L^0 + 2\kappa$ rather than just to L^0: the strip $[L^0, L^0 + 2\kappa]$ is a buffer in which $V(L)$ is too small (between 0 and 2κ) for the uniform-envelope bound to dominate it, so we cannot yet claim a useful negative drift there. Outside this buffer, on $\{L \ge L^0 + 2\kappa\}$, we have the drift $\mathbb{E}V(L_{i+1}) \mid L_i = L \le \kappa \le V(L) - \kappa$.

Region 3: the compact set $C := [L_*, L^0 + 2\kappa]$. Here $V(L) \le \max(L_0 - L_*, 2\kappa)$. The map $L \mapsto \mathbb{E}V(L_{i+1}) \mid L_i = L$ is continuous in L (the transition density is continuous and has finite first moment of $|L_{i+1}|$ locally in L, since $L_{i+1} = \log F(e^L) + \log g^2$ with $\log F(e^L)$ continuous and $\mathbb{E}|\log g^2| < \infty$), hence bounded on the compact C by some $b < \infty$. On C we make no claim of a negative drift; this is the small set on which V is allowed to grow, contributing the term $b\,\mathbf{1}_C(L)$ in (17).

Combining the three regions yields the drift condition

$$\mathbb{E}V(L_{i+1}) \mid L_i = L \;\le\; V(L) - c + b\,\mathbf{1}_C(L),$$

for some $c > 0$ and $b < \infty$, exactly the form (17), with small set $C = [L_*, L^0 + 2\kappa]$. Together with the ψ-irreducibility verified above, Foster's drift criterion (Theorem 11.3.4 of (Meyn & Tweedie, 1993)) yields that $\{u_i\}$ is positive Harris recurrent and admits a unique invariant probability measure π on $(0, \infty)$. Aperiodicity was also verified above ($m = 1$ minorisation on every compact). The Aperiodic Ergodic Theorem (Theorem 13.0.1 of (Meyn & Tweedie, 1993)) then applies, giving that the law of u_i converges to π in total variation from every $u_1 > 0$; in particular π is independent of v_1. Non-degeneracy $\pi((\delta, \infty)) > 0$ follows from the non-convergence step (the chain spends a positive fraction of time above δ). □

Remark A.2. *For $g \sim \mathcal{N}(0,1)$, $\mathbb{E}|\log g^2| < \infty$ (used in the proof of (i) and in the non-convergence step of (ii)). Indeed, $g^2 \sim \chi^2_1$ has density $f(x) = (2\pi x)^{-1/2} e^{-x/2}$ on $(0, \infty)$, and*

$$\mathbb{E}|\log g^2| \;=\; \int_0^\infty |\log x|\,(2\pi x)^{-1/2} e^{-x/2}\,dx.$$

Split at $x = 1$. The upper tail $\int_1^\infty (\log x)(2\pi x)^{-1/2} e^{-x/2}\,dx$ is finite because $(\log x)\,x^{-1/2} e^{-x/2}$ decays exponentially. For the lower tail, $|\log x| = -\log x$ on $(0,1)$ and $e^{-x/2} \le 1$, so

$$\int_0^1 (-\log x)\,(2\pi x)^{-1/2} e^{-x/2}\,dx \;\le\; (2\pi)^{-1/2} \int_0^1 \frac{-\log x}{\sqrt{x}}\,dx.$$

Substitute $u := -\log x$, so $x = e^{-u}$, $dx = -e^{-u}\,du$, $\sqrt{x} = e^{-u/2}$, and the limits $x \in (0,1)$ map to $u \in (\infty, 0)$. The integrand transforms as

$$\frac{-\log x}{\sqrt{x}}\,dx \;=\; \frac{u}{e^{-u/2}} \cdot (-e^{-u}\,du) \;=\; -u\,e^{-u/2}\,du,$$

and the sign flip cancels against the reversed limits, yielding

$$\int_0^1 \frac{-\log x}{\sqrt{x}}\,dx \;=\; \int_0^\infty u\,e^{-u/2}\,du \;=\; 4.$$

The substitution thus replaces the logarithmic singularity at $x = 0$ by a linear factor u against an exponentially decaying weight, which is harmless. Combined with finiteness of the upper tail, this gives $\mathbb{E}|\log g^2| < \infty$, and a fortiori $\mathbb{E}\log g^2$ is well-defined and equals $-\gamma - \log 2$.

Proof of Theorem 4.1. Reduction to (2). Conditional on $\bar{Z}_i$, the k-th coordinate $V_{i+1}^{(k)} = G_i^{(k)}(\bar{Z}_{i,1}) - G_i^{(k)}(\bar{Z}_{i,2})$ is centred Gaussian with variance $2(1-\phi(v_i))$, since $G_i^{(k)}(z) \sim \mathcal{N}(0,1)$ pointwise and $\mathrm{Cov}(G_i^{(k)}(\bar{Z}_{i,1}), G_i^{(k)}(\bar{Z}_{i,2})) = \phi(\|\bar{Z}_{i,1} - \bar{Z}_{i,2}\|) = \phi(v_i)$. The coordinates are independent across k because the GPs $G_i^{(1)}, \dots, G_i^{(d)}$ are independent. Hence

$$V_{i+1} \;=\; \sqrt{2(1-\phi(v_i))}\,\mathbf{g}_i, \qquad \mathbf{g}_i \sim \mathcal{N}(0, I_d) \text{ ind. of } v_i,$$

and $v_{i+1}^2 = 2(1-\phi(v_i))\,\|\mathbf{g}_i\|^2$ with $\|\mathbf{g}_i\|^2 \sim \chi_d^2$. By rotation invariance the conditional law of v_{i+1}^2 given $\bar{Z}_i$ depends only on v_i, so $\{v_i^2\}$ is Markov with transition law (2).

Dichotomy. The proof of Proposition A.1 extends to the $\mathbb{R}^d$ chain by substituting $X_i \sim \chi_d^2$ for g_i^2 throughout. The functional ingredients of the argument — the map $F(u) = 2(1-e^{-u/(2r^2)})$, the log-chain function $H(L) = \log(F(e^L)/e^L)$, the uniform upper bound (14), the boundary linearisation $H(L) \to -2\log r$ as $L \to -\infty$, and the Lyapunov function V — depend only on the recursion (2) and not on the specific law of the noise, so they remain unchanged. The only quantity that changes is the noise mean: for $X \sim \chi_d^2 = 2\,\mathrm{Gamma}(d/2, 1)$,

$$\mathbb{E}\log X \;=\; \psi(d/2) + \log 2,$$

finite for every $d \geq 1$. Define $\rho_d := \psi(d/2) + \log 2 - 2\log r$, so $r > r_c(d) \iff \rho_d < 0$ and $r < r_c(d) \iff \rho_d > 0$. The pointwise bound $H(L) \leq -2\log r$ from (14) is unchanged (it uses only $1 - e^{-x} \leq x$); the limit $H(L) \to -2\log r$ as $L \to -\infty$ is unchanged. The SLLN argument of part (i) goes through verbatim with $\log g_j^2$ replaced by $\log X_j$ and $\mathbb{E}\log g^2$ replaced by $\mathbb{E}\log X$, giving $\limsup_i L_i/i \leq \rho_d$ a.s. The non-convergence step in part (ii) is identical (zero-mean random walk $S_n := \sum_{j=1}^n (\log X_j - \mathbb{E}\log X)$ obeys $S_n/n \to 0$ a.s. by the SLLN, and the strong-Markov iteration is unchanged).

For the Foster–Lyapunov step in part (ii): the transition law of $\{u_i = v_i^2\}$ given $u_i = u > 0$ is the law of $2(1-\phi(\sqrt{u}))\,X$ with $X \sim \chi_d^2$. The χ_d^2 density $\frac{1}{2^{d/2}\Gamma(d/2)} x^{d/2-1} e^{-x/2}$ is strictly positive and continuous on $(0,\infty)$ for every $d \geq 1$, so this transition law has a strictly positive Lebesgue density on $(0,\infty)$, smoothly depending on u. The same three-fold consequence as in the proof of Proposition A.1 holds: *(a)* the chain $\{u_i\}$ is ψ-irreducible w.r.t. Lebesgue measure on $(0,\infty)$; *(b)* every compact subset of $(0,\infty)$ is small with $m = 1$; *(c)* the chain is strongly aperiodic. The Lyapunov function $V(L) = (L_0 - L)^+ + (L - L^0)^+$ from the proof of Proposition A.1 works unchanged, with g_i^2 replaced by $X_i \sim \chi_d^2$ in every drift computation; the only input needed is $\mathbb{E}(\log X - M)^+ \to 0$ as $M \to \infty$, which holds because $\mathbb{E}|\log X| < \infty$ (the verification is identical to Remark A.2, using the χ_d^2 density in place of χ_1^2). The drift condition (17) is therefore satisfied with the same small set $C = [L_*, L^0 + 2\kappa]$ (where now κ depends on the χ_d^2 envelope), and Foster's drift criterion (Theorem 11.3.4 of (Meyn & Tweedie, 1993)) yields that $\{u_i\}$ is positive Harris recurrent with a unique invariant probability $\pi^{\mathbf{v}}$ on $(0,\infty)$. The Aperiodic Ergodic Theorem (Theorem 13.0.1 of (Meyn & Tweedie, 1993)), whose hypotheses (positive Harris recurrence, aperiodicity, invariant probability) are now in place, then gives convergence of the law of u_i to $\pi^{\mathbf{v}}$ in total variation from every $u_1 > 0$; in particular $\pi^{\mathbf{v}}$ is independent of v_1. □

Proof of Theorem 4.3. The proof has two parts: (I) ergodicity of the pairwise-distance chain via a sum-of-pairs Lyapunov function, and (II) transfer from $\mathbf{u}^{(i)}$ to $\bar{Z}_i$ via a mixture bound.

Part I: ergodicity of $\{\mathbf{u}^{(i)}\}$.

I.1: Markov property and marginal structure. Conditional on $\bar{Z}_i$, $\bar{Z}_{i+1} \sim \mathcal{N}(0, R(\bar{Z}_i) \otimes I_d)$ with $R(\bar{Z}_i)_{st} = \phi(\sqrt{u_{st}^{(i)}})$, a function of $\mathbf{u}^{(i)}$ alone. Hence $\mathbf{u}^{(i+1)}$ is a function of $\bar{Z}_{i+1}$ whose conditional distribution given $\bar{Z}_i$ depends on $\bar{Z}_i$ only through $\mathbf{u}^{(i)}$, so $\{\mathbf{u}^{(i)}\}$ is Markov on $(0,\infty)^{\binom{n}{2}}$. For each fixed pair (s,t),

$$V_{st}^{(i+1)} := Z_{i+1,s} - Z_{i+1,t} \mid \bar{Z}_i \;\sim\; \mathcal{N}\big(0,\, 2(1-\phi(\sqrt{u_{st}^{(i)}}))\, I_d\big),$$

because the d output coordinates of G_i are i.i.d. and the (s,t)-marginal of $R(\bar{Z}_i) \otimes I_d$ contributes variance $2(1-\phi(\sqrt{u_{st}^{(i)}}))$. Consequently

$$u_{st}^{(i+1)} \;=\; 2\big(1-\phi(\sqrt{u_{st}^{(i)}})\big)\, X_{st}^{(i)}, \qquad X_{st}^{(i)} := \left\|V_{st}^{(i+1)}\right\|^2 / \big(2(1-\phi(\sqrt{u_{st}^{(i)}}))\big) \sim \chi_d^2 \text{ marginally},$$

the same recursion as the scalar chain (2). Although $\left\{X_{st}^{(i)}\right\}_{s<t}$ are jointly correlated (they share the single GP sample G_i), each is marginally χ_d^2, and in particular the *marginal* conditional distribution of $u_{st}^{(i+1)}$ given $\mathbf{u}^{(i)}$ depends only on $u_{st}^{(i)}$ and is identical to the scalar transition kernel.

I.2: Per-pair quadratic drift in log-coordinates. Work in the log-chain $L_{st}^{(i)} := \log u_{st}^{(i)}$. By *I.1*, the marginal recursion is

$$L_{st}^{(i+1)} \;=\; L_{st}^{(i)} + H(L_{st}^{(i)}) + \log X_{st}^{(i)}, \qquad X_{st}^{(i)} \sim \chi_d^2 \text{ marginally,}$$

with $H(L) = \log(F(e^L)/e^L)$ as in the proof of Proposition A.1, $\psi' := \psi(d/2)+\log 2 = \mathbb{E}\log X$, and $\sigma^2 := \mathrm{Var}(\log X) = \psi_1(d/2) < \infty$ (the trigamma function is finite for every $d \geq 1$). Set $\mu(L) := H(L) + \psi' = \mathbb{E}L_{st}^{(i+1)} - L_{st}^{(i)} \mid L_{st}^{(i)} = L$ and define the scalar per-pair quadratic drift

$$\Delta(L) \;:=\; \mathbb{E}(L_{st}^{(i+1)})^2 - (L_{st}^{(i)})^2 \mid L_{st}^{(i)} = L.$$

Expanding $(L+Y)^2 - L^2 = 2LY + Y^2$ with $Y := L_{st}^{(i+1)} - L_{st}^{(i)}$ and completing the square via $\mu + L = \log F(e^L) + \psi'$,

$$\Delta(L) \;=\; \mu(L)^2 + \sigma^2 + 2L\mu(L) \;=\; \left(\log F(e^L) + \psi'\right)^2 - L^2 + \sigma^2. \tag{19}$$

The function Δ is continuous on $\mathbb{R}$ and $\Delta(L) \to -\infty$ as $|L| \to \infty$:

Left tail. As $L \to -\infty$, $F(u) = u/r^2 + O(u^2)$ gives $\log F(e^L) = L - 2\log r + o(1)$, so $\mu(L) \to \psi' - 2\log r = \rho_d > 0$ by subcriticality $r < r_c(d)$. Hence $\mu(L)^2$ and σ^2 stay bounded while $2L\mu(L) \sim 2L\rho_d \to -\infty$.

Right tail. As $L \to +\infty$, $F(e^L) \to 2$, so $\log F(e^L) + \psi'$ is bounded and $\Delta(L) = O(1) - L^2 \to -\infty$.

In particular $B := \sup_{L\in\mathbb{R}} \Delta(L) < \infty$ (depending only on r, d), and we can choose $M > 0$ so that $\Delta(L) \leq -B\binom{n}{2} - 1$ whenever $|L| \geq M$.

The distance vector itself does not give a chain we can analyse through a Lebesgue density: for $n \geq d+2$ the vector $\mathbf{u}^{(i)} = \mathbf{u}(\bar{Z}_i)$ is confined to the cone of squared Euclidean distance matrices of n points in $\mathbb{R}^d$, a submanifold of dimension $nd - \binom{d+1}{2} < \binom{n}{2}$, so its law is singular with respect to Lebesgue measure on $(0,\infty)^{\binom{n}{2}}$ (Remark 4.4). We therefore run the Foster–Lyapunov argument on the position chain $\left\{\bar{Z}_i\right\}$ on $(\mathbb{R}^d)^n$ itself, whose Gaussian transition kernel *does* admit a strictly positive density, and recover $\pi^{\mathbf{u}}$ as the image of $\pi_{\bar{Z}}$ under the distance map.

Define the Lyapunov function on $(\mathbb{R}^d)^n$ through the pairwise distances,

$$V(\bar{z}) \;:=\; \sum_{s<t} \left(\log \|z_s - z_t\|^2\right)^2 \;\geq\; 0,$$

so that $V(\bar{Z}_i) = \sum_{s<t}(L_{st}^{(i)})^2$. By *I.1*, the conditional law of each $u_{st}^{(i+1)}$ given $\bar{Z}_i$ depends on $\bar{Z}_i$ only through $u_{st}^{(i)}$, so the drift of V decomposes as a sum of scalar per-pair drifts:

$$\mathbb{E}V(\bar{Z}_{i+1}) \mid \bar{Z}_i - V(\bar{Z}_i) \;=\; \sum_{s<t} \Delta(L_{st}^{(i)}). \tag{20}$$

I.3: Small set and drift off it. Define $C := \left\{\bar{z} \in (\mathbb{R}^d)^n : \max_{s<t} \left|\log \|z_s - z_t\|^2\right| \leq M\right\}$. For $\bar{z} \notin C$ there exists at least one pair (s_0, t_0) with $|L_{s_0t_0}| > M$; that pair contributes $\Delta(L_{s_0t_0}) \leq -B\binom{n}{2} - 1$ by the choice of M, while every other pair contributes at most B, giving

$$\sum_{s<t} \Delta(L_{st}) \;\leq\; -B\binom{n}{2} - 1 + B\left(\binom{n}{2} - 1\right) \;=\; -B - 1 \;\leq\; -1.$$

For $\bar{z} \in C$, the trivial bound $\sum_{s<t} \Delta(L_{st}) \leq B\binom{n}{2}$ holds. Combining,

$$\mathbb{E}V(\bar{Z}_{i+1}) \mid \bar{Z}_i = \bar{z} \;\leq\; V(\bar{z}) - 1 + \left(B\binom{n}{2} + 1\right) \mathbf{1}_C(\bar{z}),$$

which is the Foster–Lyapunov drift criterion (Theorem 11.3.4 of (Meyn & Tweedie, 1993)) for $\{\bar{Z}_i\}$ on $(\mathbb{R}^d)^n$.

I.4: Irreducibility, small-set minorisation, aperiodicity. Given $\bar{Z}_i = \bar{z}$ with pairwise-distinct coordinates, the law of $\bar{Z}_{i+1}$ is $\mathcal{N}(0, R(\mathbf{u}(\bar{z})) \otimes I_d)$, which has a strictly positive continuous density on all of $(\mathbb{R}^d)^n$: the correlation matrix $R(\mathbf{u}(\bar{z}))$ is positive definite whenever the points are distinct ($R_{ss} = 1$, $|R_{st}| = \phi(\|z_s - z_t\|) < 1$, and positive-definiteness of the RBF correlation kernel on any finite set of distinct points is standard; see Proposition 2 of (Dunlop et al., 2018)), so $R(\mathbf{u}(\bar{z})) \otimes I_d$ is nonsingular. Since this density has full support for every such $\bar{z}$, the chain $\{\bar{Z}_i\}$ is ψ-irreducible with respect to Lebesgue measure on $(\mathbb{R}^d)^n$. On C the squared distances lie in the compact box $[e^{-M}, e^{M}]^{\binom{n}{2}}$, so $R(\mathbf{u}(\bar{z}))$ ranges over a compact set of positive-definite matrices with eigenvalues bounded away from 0 and ∞; the corresponding centred Gaussian densities are therefore uniformly bounded below on any fixed compact ball $B_0 \subset (\mathbb{R}^d)^n$, yielding the minorisation $P(\bar{z}, \cdot) \geq \eta\, \nu(\cdot)$ for all $\bar{z} \in C$, with $\eta > 0$ and ν the normalised restriction of Lebesgue measure to B_0. Hence C is a small set with $m = 1$, and the chain is strongly aperiodic. (Note C is the small set even though it is unbounded under translations: the kernel $\mathcal{N}(0, R(\mathbf{u}(\bar{z})) \otimes I_d)$ is centred at the origin and depends on $\bar{z}$ only through the distances, which are bounded on C.)

I.5: Ergodic theorem. Foster's drift criterion (Theorem 11.3.4 of (Meyn & Tweedie, 1993)) with the drift from *I.3* and the small set from *I.4* yields positive Harris recurrence of $\{\bar{Z}_i\}$ and a unique invariant probability measure $\pi_{\bar{Z}}$ on $(\mathbb{R}^d)^n$. The Aperiodic Ergodic Theorem (Theorem 13.0.1 of (Meyn & Tweedie, 1993)) then gives $\left\|\mathrm{Law}(\bar{Z}_i) - \pi_{\bar{Z}}\right\|_{\mathrm{TV}} \to 0$ from every initial condition $\bar{Z}_1$ with pairwise-distinct coordinates. Let $\pi^{\mathbf{u}}$ be the image of $\pi_{\bar{Z}}$ under the distance map $\bar{z} \mapsto \mathbf{u}(\bar{z})$; it is invariant for the distance chain $\{\mathbf{u}^{(i)}\}$ (well defined and Markov by *I.1*), and $\left\|\mathrm{Law}(\mathbf{u}^{(i)}) - \pi^{\mathbf{u}}\right\|_{\mathrm{TV}} \to 0$ as well, the distance map being a contraction in total variation. The marginal of $\pi^{\mathbf{u}}$ at each pair (s,t) is stationary for the scalar chain (by *I.1*), hence equal to $\pi^{\mathbf{v}}$ by uniqueness in Theorem 4.1 (ii). Uniqueness of $\pi^{\mathbf{u}}$ follows from that of $\pi_{\bar{Z}}$: any invariant law of $\{\mathbf{u}^{(i)}\}$ lifts through $k(\mathbf{u}, \cdot) = \mathcal{N}(0, R(\mathbf{u}) \otimes I_d)$ to an invariant law of $\{\bar{Z}_i\}$, necessarily $\pi_{\bar{Z}}$, whose image is $\pi^{\mathbf{u}}$.

Part II: explicit TV bound and mixture representation.

Denote $P(\bar{z}, A) := \mathbb{P}\left(\bar{Z}_{i+1} \in A \mid \bar{Z}_i = \bar{z}\right) = \mathcal{N}(0, R(\bar{z}) \otimes I_d)(A)$. This depends on $\bar{z}$ only through $\mathbf{u}(\bar{z}) = (\|z_s - z_t\|^2)_{s<t}$, so there is a probability kernel $k(\mathbf{u}, \cdot) := \mathcal{N}(0, R(\mathbf{u}) \otimes I_d)$ on $(\mathbb{R}^d)^n$ with $P(\bar{z}, A) = k(\mathbf{u}(\bar{z}), A)$. By the Markov property,

$$\mathbb{P}\left(\bar{Z}_{i+1} \in A\right) = \mathbb{E}k(\mathbf{u}^{(i)}, A) = \int k(\mathbf{u}, A)\, d\mathrm{Law}(\mathbf{u}^{(i)})(\mathbf{u}),$$

and by (5), $\pi_{\bar{Z}}(A) = \int k(\mathbf{u}, A)\, d\pi^{\mathbf{u}}(\mathbf{u})$. Subtracting and using $k(\mathbf{u}, A) \in [0,1]$,

$$\left|\mathbb{P}\left(\bar{Z}_{i+1} \in A\right) - \pi_{\bar{Z}}(A)\right| = \left|\int k(\mathbf{u}, A)\, d(\mathrm{Law}(\mathbf{u}^{(i)}) - \pi^{\mathbf{u}})(\mathbf{u})\right| \leq \left\|\mathrm{Law}(\mathbf{u}^{(i)}) - \pi^{\mathbf{u}}\right\|_{\mathrm{TV}}.$$

Taking the supremum over A and shifting $i \mapsto i-1$ gives (6). By Part I, the RHS tends to 0 in TV.

Mixture representation. Existence and uniqueness of $\pi_{\bar{Z}}$ were established in Part I; it remains only to identify it with (5). Invariance of $\pi_{\bar{Z}}$ gives $\pi_{\bar{Z}}(A) = \int P(\bar{z}, A)\, d\pi_{\bar{Z}}(\bar{z}) = \int k(\mathbf{u}, A)\, d\pi^{\mathbf{u}}(\mathbf{u})$, with $\pi^{\mathbf{u}}$ the image of $\pi_{\bar{Z}}$ under $\bar{z} \mapsto \mathbf{u}(\bar{z})$; this is exactly (5). □

Proof of Theorem 4.5. Marginals. For every $i \geq 1$ and $t \in \{1,2\}$, conditional on $\bar{Z}_{i-1}$ each coordinate of $\bar{Z}_{i,t} = G_{i-1}(Z_{i-1,t})$ is a marginal evaluation of a centred unit-variance GP, so $\bar{Z}_{i,t} \mid \bar{Z}_{i-1} \sim \mathcal{N}(0, \phi(0)\, I_d) = \mathcal{N}(0, I_d)$. The (t,t)-block is independent of $\bar{Z}_{i-1}$, so unconditionally $\bar{Z}_{i,t} \sim \mathcal{N}(0, I_d)$. Weak convergence then forces $\bar{Z}_{\infty,t} \sim \mathcal{N}(0, I_d)$. It remains to prove non-joint-Gaussianity.

Step 1: any Gaussian joint limit has isotropic V_∞. Let $V_i := \bar{Z}_{i,1} - \bar{Z}_{i,2} \in \mathbb{R}^d$. Conditional on $\bar{Z}_{i-1}$, the d output coordinates of G_{i-1} are i.i.d., so the d coordinates of $V_i = G_{i-1}(\bar{Z}_{i-1,1}) - G_{i-1}(\bar{Z}_{i-1,2})$ are i.i.d. with marginal $\mathcal{N}(0, 2(1 - \phi(\|V_{i-1}\|)))$. Hence

$$V_i \mid \bar{Z}_{i-1} \sim \mathcal{N}\big(0,\ 2(1 - \phi(\|V_{i-1}\|))\, I_d\big),$$

a scalar multiple of I_d depending on V_{i-1} only through $\|V_{i-1}\|$. Therefore the unconditional law of V_i is a mixture of isotropic Gaussians and is itself rotationally invariant in $\mathbb{R}^d$ for every $i \geq 1$; so is its weak limit V_∞. Suppose now, for contradiction, that $(\bar{Z}_{\infty,1}, \bar{Z}_{\infty,2})$ is jointly Gaussian. Then V_∞ is a centred Gaussian in $\mathbb{R}^d$; combined with rotational invariance, its covariance matrix is a scalar multiple of I_d, i.e. $V_\infty \sim \mathcal{N}(0, \sigma_v^2 I_d)$ for some $\sigma_v^2 \in [0, \infty)$. The case $\sigma_v^2 = 0$

gives $V_\infty = 0$ a.s., which contradicts Theorem 4.1 (ii) (nontrivial stationary law on $(0,\infty)$), so $\sigma_v^2 > 0$. Therefore $u_\infty := \|V_\infty\|^2 \overset{d}{=} \sigma_v^2\,\chi_d^2$, equivalently $u_\infty \overset{d}{=} e^\zeta \chi_d^2$ with $\zeta := \log \sigma_v^2 \in \mathbb{R}$.

Step 2: no $e^\zeta\chi_d^2$ is stationary for (2). Suppose $u \overset{d}{=} e^\zeta X$ with $X \sim \chi_d^2$, and let $X_i \sim \chi_d^2$ be the independent innovation in (2), so that the one-step image is $u' = 2(1-\phi(\sqrt{u}))\,X_i = A(X)\,X_i$ with

$$A(X) \;:=\; 2\big(1-\phi(e^{\zeta/2}\sqrt{X})\big) \;=\; 2\big(1-e^{-e^\zeta X/(2r^2)}\big).$$

Stationarity requires $u' \overset{d}{=} e^\zeta X''$ with $X'' \sim \chi_d^2$, i.e. $A(X)\,X_i \overset{d}{=} e^\zeta X''$. Taking logarithms,

$$\log A(X) + \log X_i \;\overset{d}{=}\; \zeta + \log X'', \tag{21}$$

where $\log X_i$ and $\log X''$ are both distributed as $\log\chi_d^2$ and each is independent of the other terms on its side.

Recall that the *characteristic function* of a random variable Y is $\varphi_Y(t) := \mathbb{E}e^{itY}$ for $t \in \mathbb{R}$; φ_Y is continuous with $\varphi_Y(0) = 1$, is multiplicative under convolution (so $\varphi_{Y_1+Y_2} = \varphi_{Y_1}\,\varphi_{Y_2}$ for independent Y_1, Y_2), and uniquely determines the law of Y ((Durrett, 2019)). Applying φ to both sides of (21) and using independence yields, for every $t \in \mathbb{R}$,

$$\varphi_{\log A(X)}(t)\,\varphi_{\log\chi_d^2}(t) \;=\; e^{it\zeta}\,\varphi_{\log\chi_d^2}(t). \tag{22}$$

We claim $\varphi_{\log\chi_d^2}$ is non-vanishing on all of $\mathbb{R}$. By the Mellin transform of the χ_d^2 density,

$$\varphi_{\log\chi_d^2}(t) \;=\; \mathbb{E}X^{it} \;=\; \frac{2^{it}\,\Gamma(d/2+it)}{\Gamma(d/2)},$$

and Γ has no zeros on $\mathbb{C}$ ((Abramowitz & Stegun, 1964)), so $\Gamma(d/2+it) \neq 0$ for all $t \in \mathbb{R}$ and the claim follows. Dividing (22) through by $\varphi_{\log\chi_d^2}(t)$ gives

$$\varphi_{\log A(X)}(t) \;=\; e^{it\zeta} \qquad \text{for every } t \in \mathbb{R}.$$

The right-hand side is the characteristic function of the point mass δ_ζ, so by uniqueness of characteristic functions $\log A(X) \overset{d}{=} \delta_\zeta$, i.e. $A(X)$ is almost surely equal to e^ζ. But $X \mapsto A(X) = 2(1-e^{-e^\zeta X/(2r^2)})$ is strictly increasing and non-constant on $(0,\infty)$, and $X \sim \chi_d^2$ has full support on $(0,\infty)$, so $A(X)$ is non-degenerate. Contradiction.

Step 3: combining. By Step 2, no $e^\zeta\chi_d^2$ is invariant under (2). Since the stationary law of u_i is unique by Theorem 4.1 (ii) and the law of $u_i = \|V_i\|^2$ converges to this stationary law, u_∞ is not of the form $e^\zeta\chi_d^2$ for any $\zeta \in \mathbb{R}$. This contradicts the Gaussianity-plus-isotropy consequence of Step 1. Hence $(\bar{Z}_{\infty,1}, \bar{Z}_{\infty,2})$ is not jointly Gaussian. □

Proof of Proposition 4.7. Finiteness of $\mathbb{E}_{\pi^{\mathbf{v}}}|L|$: the Lyapunov function $V(L) = (L_0 - L)^+ + (L - L^0)^+$ from the proof of Proposition A.1 (ii) satisfies $\mathbb{E}_{\pi^{\mathbf{v}}}V < \infty$ (standard for positive Harris recurrent chains with a Foster–Lyapunov drift function; Theorem 14.3.7 of (Meyn & Tweedie, 1993)), and $|L| \le V(L) + \max(|L_0|, |L^0|)$, so $\mathbb{E}_{\pi^{\mathbf{v}}}|L| < \infty$.

(a) We first verify that the expectations below are finite: as $L \to -\infty$, $H(L) = \log F(e^L) - L \to L - 2\log r - L + o(1) = -2\log r$ stays bounded, and as $L \to +\infty$, $\log F(e^L) \to \log 2$ stays bounded so $H(L) = \log 2 - L + o(1)$, giving the linear envelope $|H(L)| \le |L| + C$ for a constant $C = C(r)$. Hence $\mathbb{E}_{\pi^{\mathbf{v}}}|H(L)| \le \mathbb{E}_{\pi^{\mathbf{v}}}|L| + C < \infty$ since $\mathbb{E}_{\pi^{\mathbf{v}}}|L| < \infty$ from the previous paragraph. Under $\pi^{\mathbf{v}}$, $L_i \overset{d}{=} L_{i+1}$, so taking expectations in (13) (with $\log g_i^2$ replaced by $\log X_i$, $X_i \sim \chi_d^2$) and using $X_i \perp L_i$ gives $0 = \mathbb{E}_{\pi^{\mathbf{v}}}H(L) + \mathbb{E}\log X = \mathbb{E}_{\pi^{\mathbf{v}}}H(L) + \psi(d/2) + \log 2$.

(b) We claim H is strictly concave and strictly decreasing on $\mathbb{R}$. With $x := e^L/(2r^2)$,

$$H'(L) \;=\; \frac{x}{e^x - 1} - 1,$$

and $d/dx[x/(e^x-1)] = [e^x(1-x)-1]/(e^x-1)^2 < 0$ for $x > 0$ (the numerator is 0 at $x = 0$ and has derivative $-xe^x < 0$). Since $dx/dL = x > 0$, H' is strictly decreasing in L and negative (as $H'(L) \to 0^-$ at $L \to -\infty$); hence H is strictly concave and strictly decreasing. Jensen's inequality, combined with (a), gives

$$H(\mathbb{E}_{\pi^{\mathbf{v}}}L) \;\ge\; \mathbb{E}_{\pi^{\mathbf{v}}}H(L) \;=\; -\psi(d/2) - \log 2 \;=\; H(L_*),$$

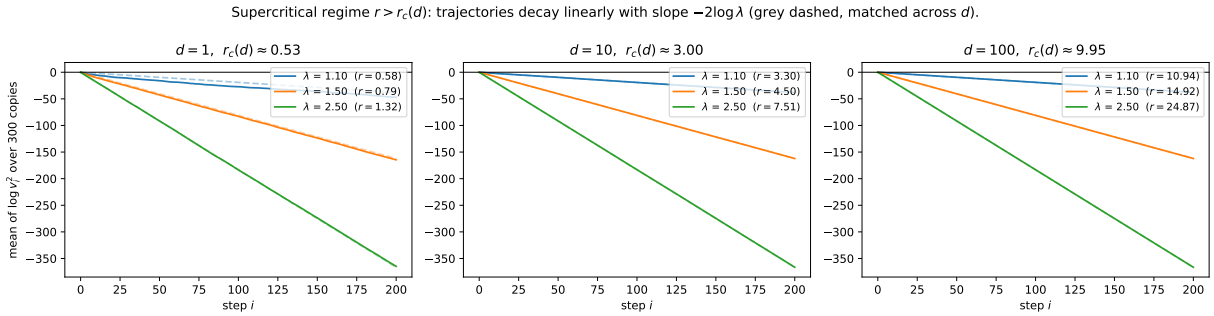

Figure 4. Supercritical regime $r > r_c(d)$: trajectories of $\log v_i^2$ for $\lambda \in \{1.1, 1.5, 2.5\}$ across $d = 1, 10, 100$. Each curve is the mean of $\log v_i^2$ over 300 i.i.d. trajectories starting at $v_1 = 1$; the grey dashed reference line has slope $-2\log\lambda$ predicted by Theorem 4.1 (i), and is hidden behind the simulated curve in every case. At matched λ, the decay rate is the same across dimensions, confirming that the rate bound from Theorem 4.1 is d-independent in units of λ. The $d = 100$ pane is the same as the left pane of Figure 2, repeated here for direct cross-d comparison.

and monotonicity of H inverts this to $\mathbb{E}_{\pi^{\mathrm{v}}} L \le L_*$. The explicit form (7) follows from $H(L_*) = -\psi(d/2) - \log 2$, which upon exponentiating reads $F(e^{L_*})/e^{L_*} = 1/r_c(d)^2$: setting $\alpha := e^{L_*}/(2r^2)$ yields $(1 - e^{-\alpha})/\alpha = (r/r_c(d))^2$. The map $\alpha \mapsto (1 - e^{-\alpha})/\alpha$ is smooth and strictly decreasing from 1 to 0 on $(0, \infty)$, so α is uniquely determined by $r < r_c(d)$. The limits as $r \downarrow 0$ (RHS $\to 0$, $\alpha \to \infty$, $(1 - e^{-\alpha})/\alpha \sim 1/\alpha$, so $\alpha \sim (r_c/r)^2$ and $2r^2\alpha \to 2r_c^2$) and $r \uparrow r_c(d)$ (RHS $\to 1$, $\alpha \to 0$, $L_* \to -\infty$) are direct. □

Proof of Corollary 4.8. $L_* = \log(2r^2\alpha)$ with α determined by $(1 - e^{-\alpha})/\alpha = (r/r_c(d))^2$; at $r = \lambda r_c(d)$ this RHS equals λ^2, so $\alpha = \alpha(\lambda)$ is independent of d. Then $L_* = \log(2\lambda^2 r_c(d)^2\alpha(\lambda)) = \log(\lambda^2\alpha(\lambda)) + \log 2 + \log r_c(d)^2 = \log(\lambda^2\alpha(\lambda)) + \log 2 + (\psi(d/2) + \log 2)$. □

B. Additional empirical figures

This appendix collects the empirical figures supporting Section 5 that are not in the main paper. All simulations are CPU-only Python (NumPy / SciPy / matplotlib / scikit-learn); the full set of experiments finishes well within an hour on a single laptop core.

Figure 4 verifies the supercritical decay rate of Theorem 4.1 (i) across $d \in \{1, 10, 100\}$. Figures 5 and 6 are the $d = 1, 10$ companions of Figure 2: subcritical trajectories of $\log v_i^2$ across all three dimensions, and the empirical stationary law of $\log v_i^2$ across all three dimensions. At $d = 1$ near-critical λ the convergence becomes very slow and the stationary law has an extremely heavy left tail. Figure 7 reports the sample std of $\log v_i^2$ across the 300 i.i.d. replicates used to compute the mean trajectories in Figures 4 and 5; dividing this std by $\sqrt{300}$ yields the standard error of the corresponding mean trajectory estimator.

The remaining figures are per-chain visualisations of $\pi_{\bar{Z}}$, all using $n = 1000$ and five i.i.d. chains per λ, with λ values chosen per dimension to span the regime transition (see "From bounds to a choice of λ" in Section 5): $d = 1$ at $\lambda \in \{0.10, 0.30, 0.60, 0.85\}$ and depth $i = 300$, $d = 10$ at $\lambda \in \{0.66, 0.91, 0.95, 0.97\}$ and depth $i = 600$, $d = 100$ at $\lambda \in \{0.90, 0.99, 0.995\}$ and depth $i = 1000$. The smallest λ in each panel is the near-independent Gaussian baseline against which the structure at the larger λ should be read.

Figure 12 is the linear-PCA companion to the main-paper t-SNE figure at $d = 100$: PC1–PC2 of the same chains as Figure 3, with no non-linear embedding step. The chain-specific bananas, U-shapes, and arcs that appear at $\lambda = 0.99, 0.995$ in the t-SNE figure are also visible in PCA, confirming that the structure is in the data and not invented by t-SNE. See also Figure 13, where we show the *density* rather than just the scatter of the PCA plots, which makes the multimodality even clearer. Figure 8 shows the $d = 1$ case as a histogram of the n scalar entries of $\bar{Z}_{i,t}$ per chain (the top PC of 1-D data is the data itself). Figures 9 and 11 show per-chain PCA and t-SNE at $d = 10$, where the visible-structure window is wider than at $d = 100$ and the transition Gaussian-cloud $\to$ clusters $\to$ ribbons spans four λ rows. Figure 10 shows the empirical density of the PCA projections at $d = 10$.

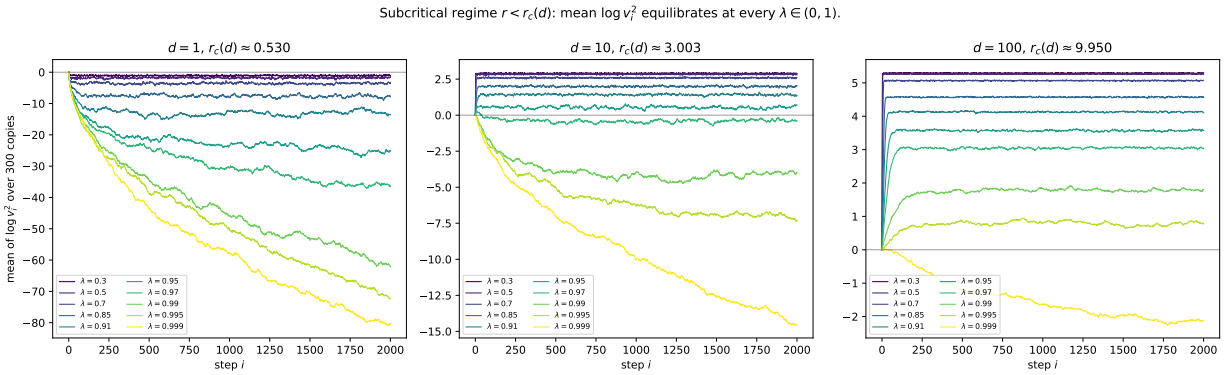

Figure 5. Subcritical regime $r < r_c(d)$: trajectories of $\log v_i^2$ for ten λ values spanning 0.30 to 0.999, across $d = 1, 10, 100$, all panels at uniform depth 2000. Each curve is the mean of $\log v_i^2$ over 300 i.i.d. trajectories starting at $v_1 = 1$, colour-coded from low (purple) to high (yellow). The convergence to equilibrium claimed by Theorem 4.1 (ii) is observed; it slows as $\lambda \uparrow 1$, especially at $d = 1$, where the empirical mean of $\log v^2$ continues drifting beyond the depth shown.

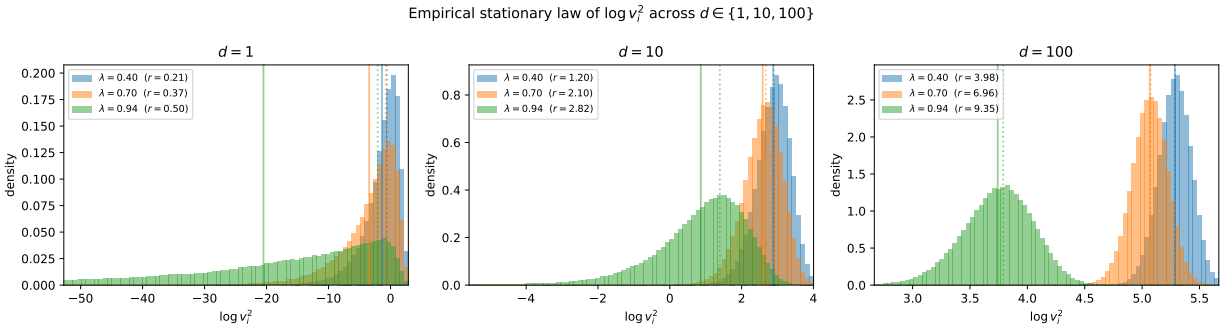

Figure 6. Empirical stationary law of $\log v_i^2$ at $\lambda \in \{0.40, 0.70, 0.94\}$, overlaid per dimension ($d = 1, 10, 100$ left to right). Each histogram is from a chain of $2 \cdot 10^5$ samples after a $2 \cdot 10^4$-step burn-in; the $d = 1$ pane trims the leftmost 8% of each sample to keep the bulk visible. Solid verticals: empirical stationary mean $\mathbb{E}_{\pi^{\mathbf{v}}} \log v^2$ (from the full, unclipped sample). Dotted verticals: Jensen upper bound $L_*(r, d)$ of Proposition 4.7 (b). The bound is nearly tight at $d = 100$ and loose at $d = 1, \lambda = 0.94$ (the heavy left tail pulls the empirical mean far below L_*). The $d = 100$ histograms are markedly narrower than at $d = 10$ or $d = 1$ because $\log X$ for $X \sim \chi^2_d$ concentrates at rate $d^{-1/2}$.

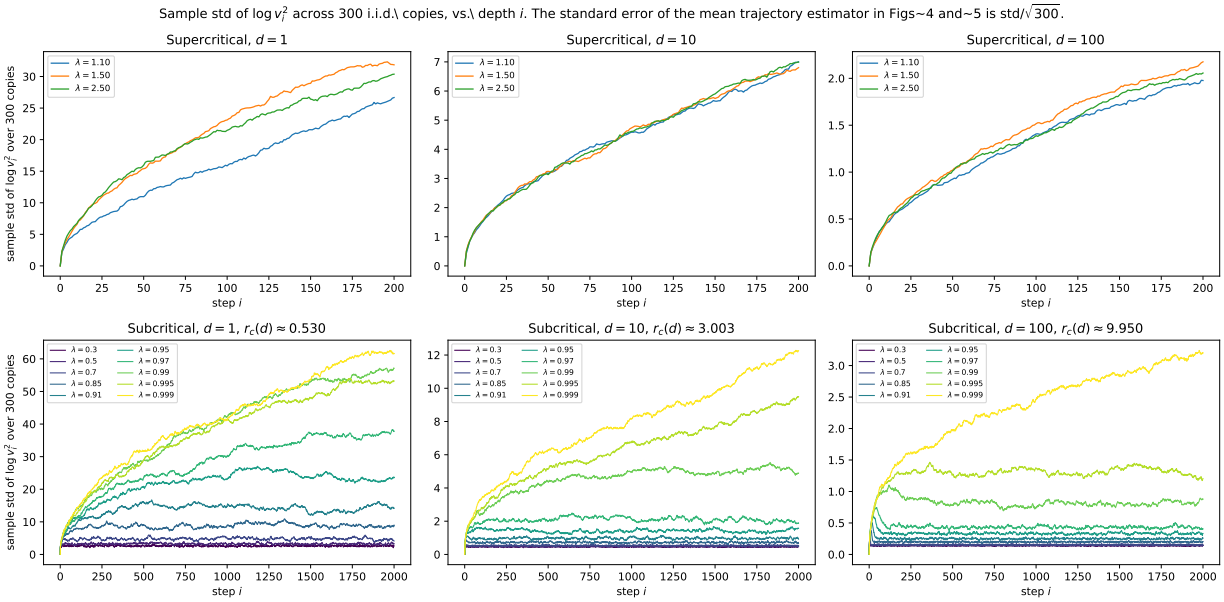

Figure 7. Sample std of $\log v_i^2$ across the 300 i.i.d. copies used in Figures 4 and 5, vs. depth i. *Top row:* supercritical regime, $\lambda \in \{1.1, 1.5, 2.5\}$, depth 200. *Bottom row:* subcritical regime, $\lambda \in \{0.30, \ldots, 0.999\}$, depth 2000; the std saturates as the chain approaches its stationary law, with saturation level increasing in λ. The standard error of the mean trajectory estimator displayed in Figures 4 and 5 is the std shown here divided by $\sqrt{300}$; this is below figure resolution at $d = 10, 100$ for every λ used in the paper, and only the most near-critical $\lambda \in \{0.99, 0.995, 0.999\}$ at $d = 1$ pushes the standard error to a fraction of the mean trajectory's range, so the empirical mean of $\log v^2$ at $d = 1$ near-critical λ is an indicative estimate rather than a tight one.

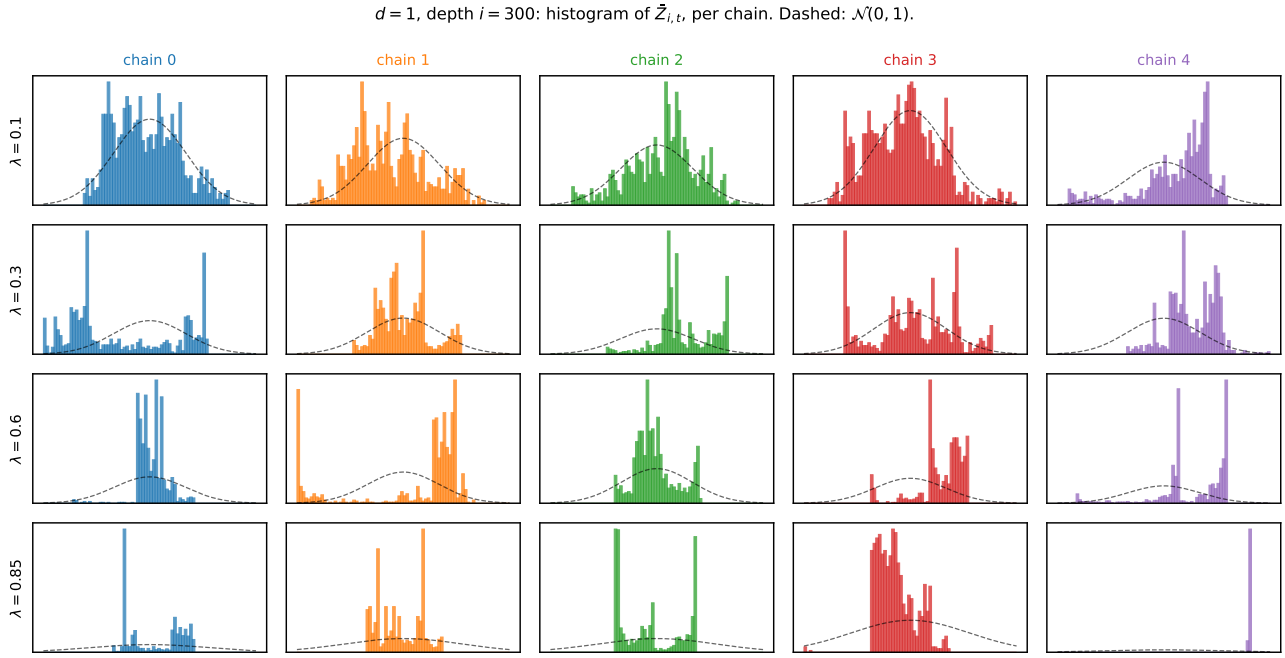

Figure 8. Per-chain histograms of $\bar{Z}_{i,t} \in \mathbb{R}$ at depth $i = 300$, $d = 1$. Rows: $\lambda \in \{0.10, 0.30, 0.60, 0.85\}$. Dashed: $\mathcal{N}(0, 1)$. The marginal stays exactly $\mathcal{N}(0, 1)$ at every depth, so any departure from the dashed curve is a within-chain dependence effect. The top row ($\lambda = 0.10$) is essentially indistinguishable from a $\mathcal{N}(0, 1)$ sample; as $\lambda \uparrow 1$ the histograms become heavily peaked / asymmetric and vary visibly between chains.

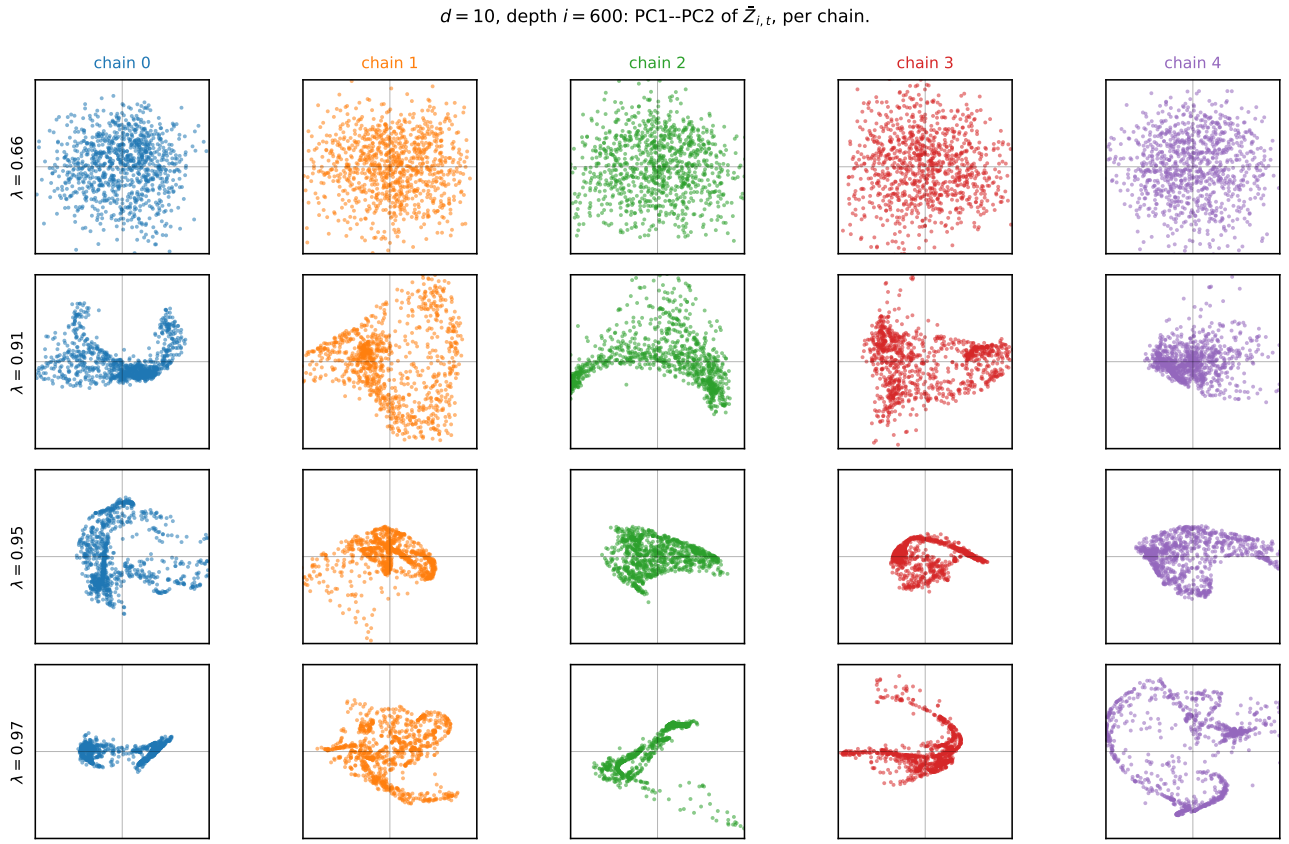

Figure 9. Per-chain PC1–PC2 scatters of $\bar{Z}_{i,t} \in \mathbb{R}^{10}$ at depth $i = 600$, five i.i.d. chains (columns); $n = 1000$ points per panel, centred and projected onto the top two PCs. Rows: $\lambda \in \{0.66, 0.91, 0.95, 0.97\}$. The top two rows are still close to a Gaussian cloud; the lower rows show the cluster-then-ribbon transition characteristic of the wider visible-structure window at moderate d. See also Figure 10, where we plot the *density* of the points above, which makes the structure and multimodality even clearer.

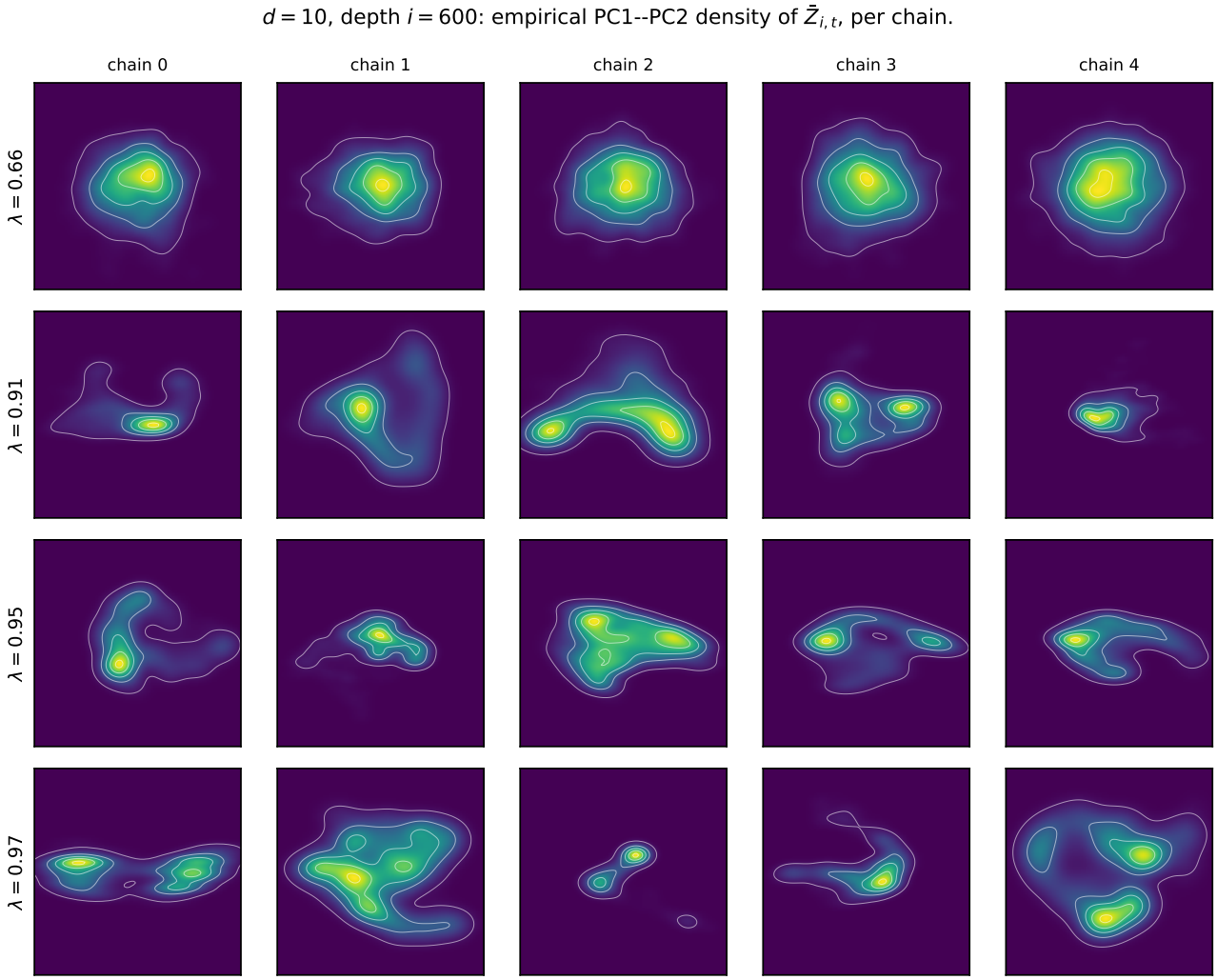

Figure 10. Empirical PC1–PC2 density (Gaussian KDE, Scott bandwidth) of the same per-chain projections shown in Figure 9. The $\lambda = 0.66$ row is the near-independent Gaussian baseline; from $\lambda = 0.91$ upward the densities are clearly multimodal and chain-specific, and by $\lambda = 0.95, 0.97$ each chain shows two or three well-separated modes.

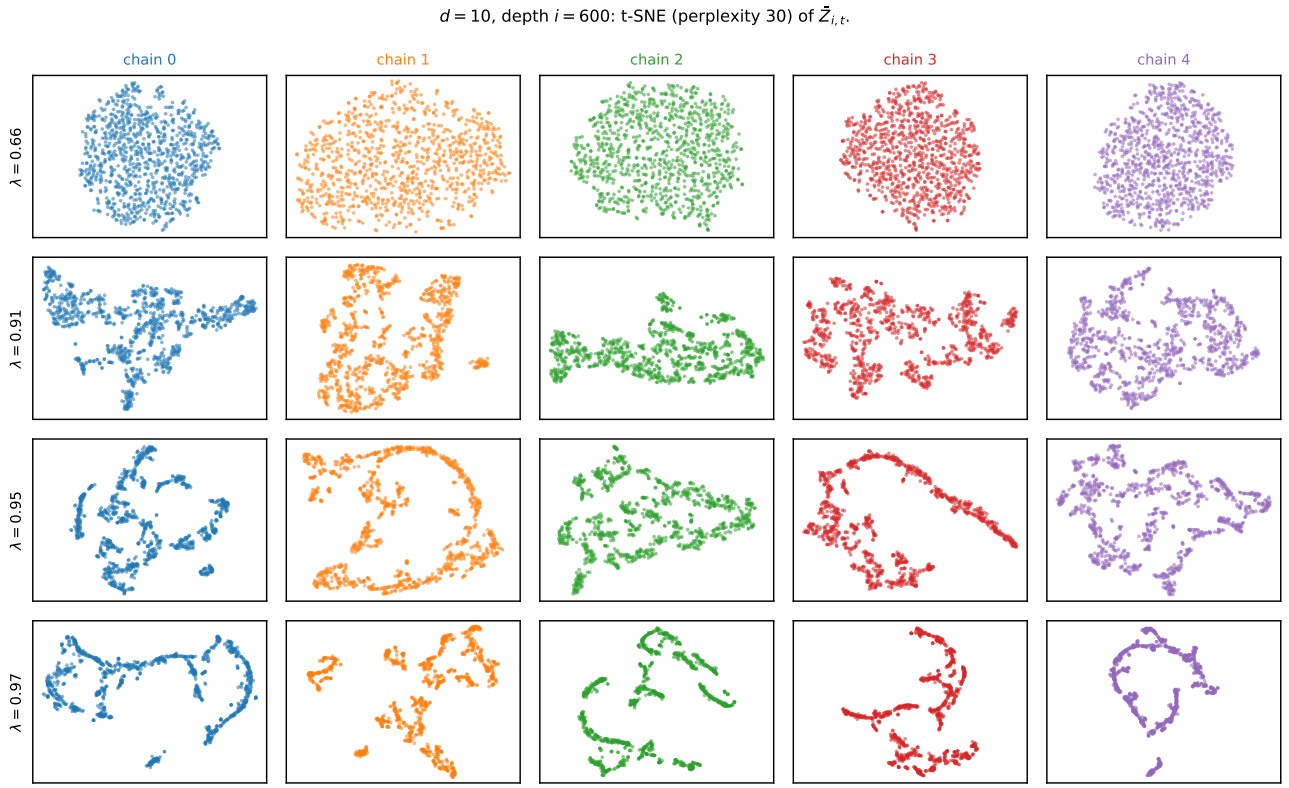

Figure 11. Per-chain t-SNE embeddings of $\bar{Z}_{i,t}$ at $d = 10$, depth $i = 600$, perplexity 30, PCA initialisation. Same λ and chains as Figure 9.

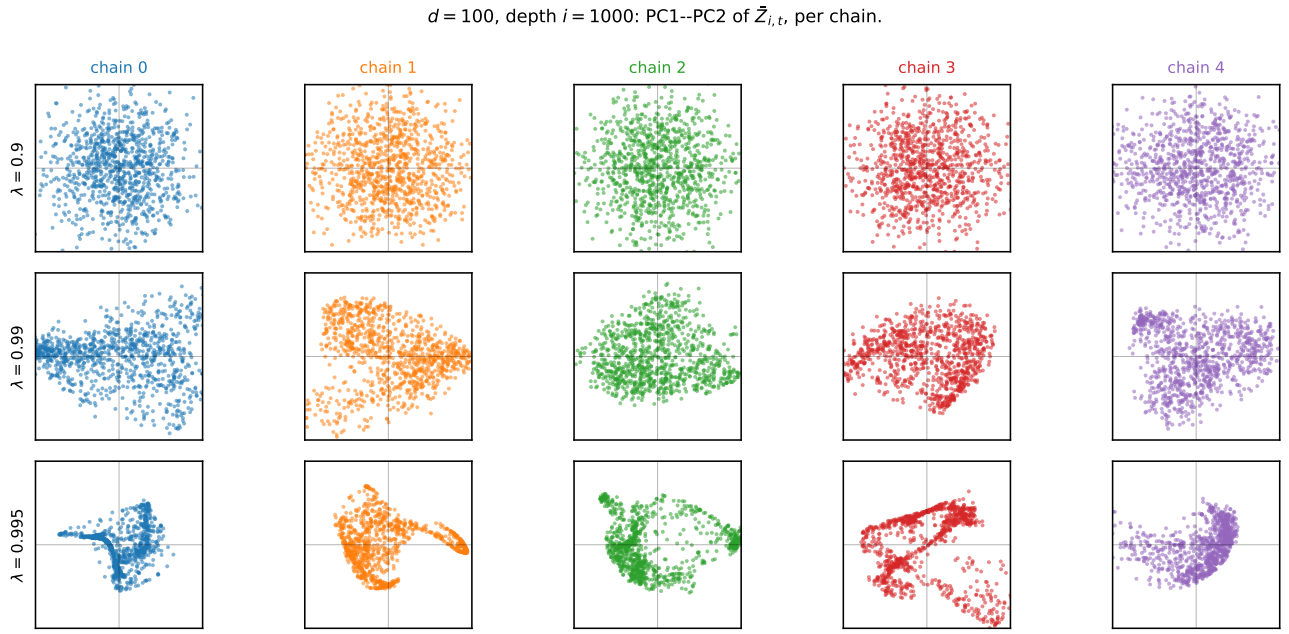

Figure 12. Per-chain PC1–PC2 scatters of $\bar{Z}_{i,t} \in \mathbb{R}^{100}$ at depth $i = 1000$, the linear-PCA companion to Figure 3. Same λ values and same five i.i.d. chains as the main-paper t-SNE figure; $n = 1000$ points per panel, centred and projected onto the top two PCs. The chain-specific bananas, U-shapes, and arcs at $\lambda = 0.99, 0.995$ that appear in t-SNE are also visible here, so they reflect the data geometry rather than the embedding. See also Figure 13, where we plot the *density* of the points above, which makes the structure and multimodality even clearer.

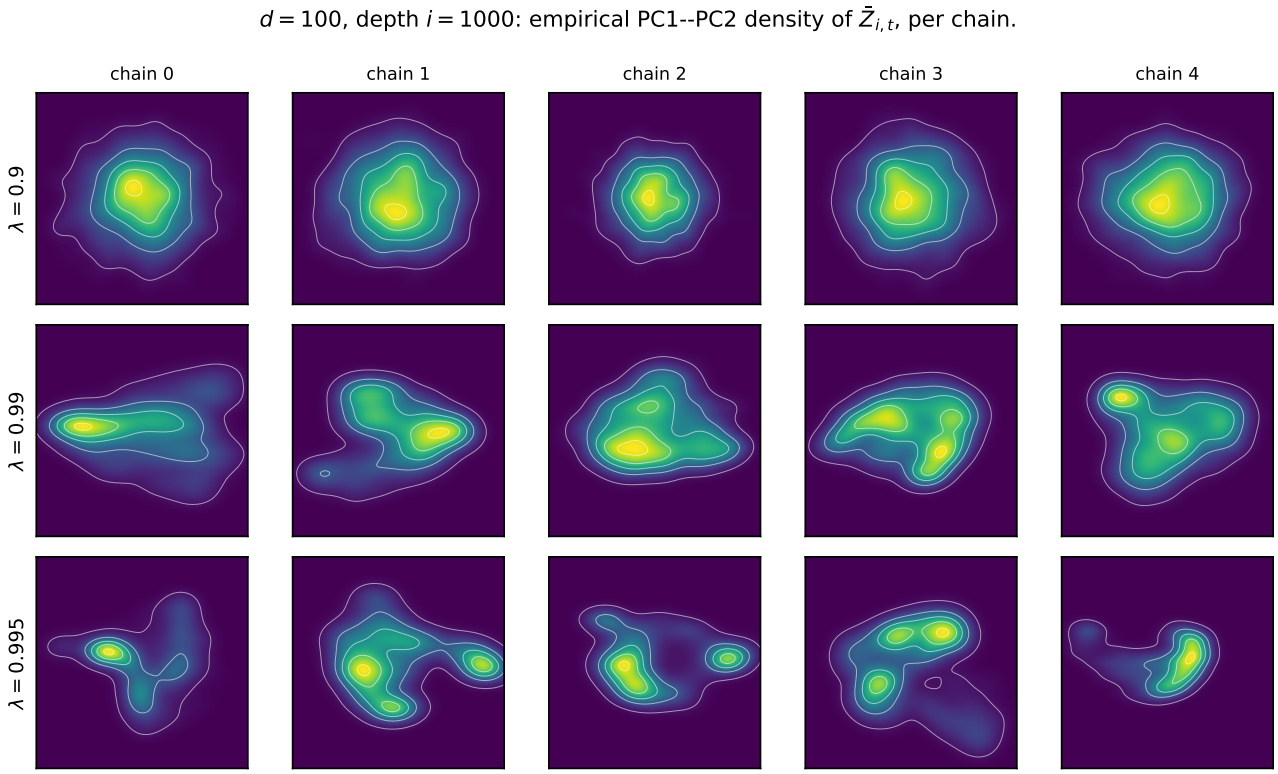

Figure 13. Empirical PC1–PC2 density (Gaussian KDE, Scott bandwidth) of the same per-chain projections shown in Figure 12. The $\lambda = 0.90$ row is the near-independent Gaussian baseline; at $\lambda = 0.99$ and $\lambda = 0.995$ the densities show pronounced chain-specific multimodality with two or three well-separated peaks per chain.