DOI:
10.1103/PhysRevE.95.012122

Document Version
Peer reviewed version

Citation for published version (APA):
Bravi, B., Opper, M., & Sollich, P. (2017). Inferring hidden states in Langevin dynamics on large networks: Average case performance. *PHYSICAL REVIEW E*, *95*(1), 1-20. Article 10.1103/PhysRevE.95.012122. https://doi.org/10.1103/PhysRevE.95.012122

Inferring hidden states in Langevin dynamics on large networks: Average case performance

B. Bravi,[1,*] M. Opper,[2] and P. Sollich[1]

[1]*Department of Mathematics, King's College London, Strand, London, WC2R 2LS UK*
[2]*Department of Artificial Intelligence, Technische Universität Berlin, Marchstraße 23, Berlin 10587, Germany*

We present average performance results for dynamical inference problems in large networks, where a set of nodes is hidden while the time trajectories of the others are observed. Examples of this scenario can occur in signal transduction and gene regulation networks. We focus on the linear stochastic dynamics of continuous variables interacting via random Gaussian couplings of generic symmetry. We analyze the inference error, given by the variance of the posterior distribution over hidden paths, in the thermodynamic limit and as a function of the system parameters and the ratio α between the number of hidden and observed nodes. By applying Kalman filter recursions we find that the posterior dynamics is governed by an "effective" drift that incorporates the effect of the observations. We present two approaches for characterizing the posterior variance that allow us to tackle, respectively, equilibrium and non-equilibrium dynamics. The first appeals to Random Matrix Theory and reveals average spectral properties of the inference error and typical posterior relaxation times, the second is based on dynamical functionals and yields the inference error as the solution of an algebraic equation.

PACS numbers: 87.10.Mn 02.50.Tt 05.10.Gg
Keywords: Inference, Linear Dynamics, Kalman Filter, Random Matrix Theory, Dynamical Functional, Gaussian posterior distributions

I. INTRODUCTION

Inferring the time evolution of a partially observed system of continuous degrees of freedom (d.o.f.) is an important problem in statistical physics. In systems biology these d.o.f. might for example be concentrations of interacting molecular species in biochemical networks. Inference of unobserved or hidden d.o.f. is then often crucial, e.g. for an understanding of molecular mechanisms underlying genetic and metabolic processes. Hidden d.o.f. can occur because the behavior of part of a network is simply not recorded, or because the amount of experimental data available might be limited [1]. If as in our analysis one studies generic continuous d.o.f., a potentially broad and interdisciplinary range of applications can be envisaged beyond biology, e.g. in financial data [2] or weather forecasting [3].

Inference has been studied using statistical mechanics approaches predominantly in scenarios without a temporal dimension, e.g. when learning from examples in neural networks [4, 5]. Several studies have, like ours, focused on performance analysis in the thermodynamic limit of large systems [6, 7]. Especially for linear learning problems, the spectrum of the input correlation matrix (or equivalently the average response function) has turned out to be a key quantity and has been studied by different means, including the replica method [5, 8, 9] based on the pioneering work of [10], diagrammatic techniques [11] and partial differential equations from matrix identities [7]. A key system parameter is the "storage" ratio between the number of training examples and the number of parameters to be learned [8, 11].

Rather less work has been done for inference based on entire temporal trajectories, with most efforts focused on the dynamics of discrete variables, typically Ising spins with random asymmetric couplings: see [12] for a review and [13–16] for examples. We extend these studies significantly by accounting for generic interaction symmetry, thus allowing us to interpolate across a range of non-equilibrium situations all the way to equilibrium dynamics. The results we present are exact in the thermodynamic limit and complement our previous study using an a priori approximate method, the Extended Plefka Expansion [17, 18]. Our emphasis on non-equilibrium dynamics is motivated by the fact that many biological processes are out of equilibrium. Indeed, recent studies [19] and computational models [20] have called for a non-equilibrium approach to gene expression dynamics that would allow one to infer regulatory interactions and transcription factor activity from time-resolved measurements.

We focus on a paradigmatic scenario: stochastic linear dynamics on a network of continuous d.o.f.

[*] barbara.bravi@kcl.ac.uk

that interact via random Gaussian couplings. Such linear dynamics should give a reasonable account also of the behavior of generic nonlinear networks of continuous d.o.f. near stable fixed points. We show that our setting is closely related to (linear Gaussian) state space modelling in statistics [21], where the dynamics of a set of hidden variables can only be observed indirectly. This allows us to deploy inference methods developed for such models [21–23], specifically the Kalman filter (and smoother) [24].

The distribution over network trajectories is Gaussian in our setting, and hence so is the posterior over hidden trajectories given a time trajectory of the observed nodes, as we will make clear. Its mean gives the optimal prediction of the time-dependent hidden state, while the second order statistics give information on the certainty of this prediction. In particular, the normalized trace of the equal-time posterior covariance matrix will be our measure of inference error. Posterior covariances between different times quantify temporal correlations of prediction uncertainties.

The novelty of our approach is that we assess the inference error of the Kalman filter for *random* interactions, which induce a random distribution in the eigenvalues of the posterior covariance. In the thermodynamic limit of large networks that we consider, the spectrum becomes self-averaging: its fluctuations tend to zero, and it becomes equal to the disorder (random interaction) average of the spectrum. We tackle this disorder average by exploiting Random Matrix Theory (RMT) results [25]. For related approaches that connect RMT and Bayesian statistics see [26, 27] and references therein.

We will see that the combination of Kalman filter and RMT gives a wealth of information for inference in systems with equilibrium dynamics, i.e. obeying detailed balance, but cannot be extended in an obvious way to non-equilibrium dynamics. For these scenarios we choose an alternative avenue, using dynamical functionals and defining the normalization factor of the posterior as a partition function. Again we consider the disorder average, for which in our case an annealed approximation is sufficient instead of a replica treatment. The replica approach was used for inference of spins trajectories in [13] generalizing to dynamics an approach that was already used for learning in static networks (see [4–6]).

The aim of this paper is to provide exact results on the average inference error for large size networks, against which other approximation methods or algorithms, can be compared. Exactness in the thermodynamic limit relies crucially on the assumption of weak long-range (mean field) interactions. In addition to the use of Kalman filter recursions combined with RMT, as well as dynamical functionals, we provide a link to variational methods.

The paper is organized as follows. After presenting the governing Kalman filter equations for the posterior variance and the effective posterior drift (section II), we use RMT to study the equilibrium dynamics case in section III, first for the elementary case of hidden variables with only self-interactions (section III B), then for symmetric hidden-hidden couplings (section III C), where we apply free probability methods. Moving on to non-equilibrium dynamics, we describe in section IV the dynamical functional method. We focus on the fully asymmetric case (section IV A) initially, which then generalizes to arbitrary symmetry (section IV B). The result is an algebraic equation for the stationary posterior variance in the Laplace domain which coincides with the one we derived using the Extended Plefka Expansion in [17, 18]. We summarize and discuss the outlook for future work in section V.

II. MODEL AND GENERAL EXPRESSION FOR POSTERIOR COVARIANCE

The setting we study consists of two sets of variables: the subnetwork, which models the *observed* d.o.f. and the bulk, which stays *hidden* and whose values we want to infer from the observations. To allow explicit insight into how the level of accuracy in this inference task depends on the structural parameters of the problem we consider a tractable scenario, where subnetwork and bulk interact *linearly.*

Our model, then, is a linear dynamical system specified by the following equations

$$\partial_t \boldsymbol{x}^{\mathrm{b}}(t) = \boldsymbol{K}^{\mathrm{bs}}\boldsymbol{x}^{\mathrm{s}}(t) + \boldsymbol{K}^{\mathrm{bb}}\boldsymbol{x}^{\mathrm{b}}(t) + \boldsymbol{\xi}^{\mathrm{b}}(t) \qquad (1)$$

$$\partial_t \boldsymbol{x}^{\mathrm{s}}(t) = \boldsymbol{K}^{\mathrm{ss}}\boldsymbol{x}^{\mathrm{s}}(t) + \boldsymbol{K}^{\mathrm{sb}}\boldsymbol{x}^{\mathrm{b}}(t) + \boldsymbol{\xi}^{\mathrm{s}}(t), \qquad (2)$$

where subnetwork and bulk variables are denoted respectively by the superscript s and b; $\boldsymbol{\xi}^{\mathrm{s}}(t)$ and $\boldsymbol{\xi}^{\mathrm{b}}$ are independent white Gaussian noises with zero mean and variance

$$\langle \boldsymbol{\xi}^{\mathrm{s}}(t)\boldsymbol{\xi}^{\mathrm{s}}(t')^T\rangle = \boldsymbol{\Sigma}^{\mathrm{ss}}\delta(t-t') \qquad (3)$$

$$\langle \boldsymbol{\xi}^{\mathrm{b}}(t)\boldsymbol{\xi}^{\mathrm{b}}(t')^T\rangle = \boldsymbol{\Sigma}^{\mathrm{bb}}\delta(t-t'). \qquad (4)$$

In addition the matrix $\boldsymbol{K}^{\mathrm{ss}}$ ($\boldsymbol{K}^{\mathrm{bb}}$) contains the linear couplings between subnetwork (bulk) variables while $\boldsymbol{K}^{\mathrm{bs}}, \boldsymbol{K}^{\mathrm{sb}}$ specify the interactions between subnetwork and bulk.

As pointed out in the introduction, a linear system with Gaussian noise produces a Gaussian distribution over the dynamical trajectories of the entire network. By this we mean that the collection of trajectories of all variables is a Gaussian process: the joint distribution of any finite collection of variables $\{x_i(t_j)\}$ is a multivariate Gaussian. To make this more intuitive it can be helpful to think about a time discretized version of the dynamics (1) and (2), for which the joint distribution of the collection of subnetwork and bulk variables across all time steps is then Gaussian, as also shown in appendix A. Inferring the hidden dynamics then corresponds to Gaussian conditioning. In particular, the aim is to evaluate the posterior probability distribution over hidden trajectories, conditioned on the observed subnetwork trajectory. We denote the latter $\boldsymbol{X}^{\mathrm{s}}$, as a shorthand for the data sequence $\{\boldsymbol{x}^{\mathrm{s}}(t)|t \in [0,T]\}$. The posterior distribution is then fully characterized by the first and second moments

$$\langle \boldsymbol{x}^{\mathrm{b}}(t)\rangle = \boldsymbol{\mu}^{\mathrm{b}}(t) \tag{5}$$

$$\langle \delta\boldsymbol{x}^{\mathrm{b}}(t)\delta\boldsymbol{x}^{\mathrm{b}}(t')^{T}\rangle = \boldsymbol{C}^{\mathrm{bb|s}}(t,t'), \tag{6}$$

where $\delta\boldsymbol{x}^{\mathrm{b}}(t) = \boldsymbol{x}^{\mathrm{b}}(t) - \boldsymbol{\mu}^{\mathrm{b|s}}(t)$ is the deviation from the posterior mean and the T superscript denotes vector or matrix transpose. As defined, $\boldsymbol{C}^{\mathrm{bb|s}}(t,t)$ is then the posterior covariance matrix of $\boldsymbol{x}^{\mathrm{b}}(t)$. We shall drop the superscripts for the sake of brevity so will denote $\boldsymbol{\mu}^{\mathrm{b|s}}(t)$ simply by $\boldsymbol{\mu}(t)$ and $\boldsymbol{C}^{\mathrm{bb|s}}(t,t')$ by $\boldsymbol{C}(t,t')$. The best estimate – in the mean-square sense – of the hidden dynamics based on the observed time series $\boldsymbol{X}^{\mathrm{s}}$ is then just $\boldsymbol{\mu}(t)$, while $\boldsymbol{C}(t,t)$ determines the uncertainty in this prediction: in particular, the trace of $\boldsymbol{C}(t,t)$ is the total mean squared prediction error for the hidden variables. Normalizing by the number of hidden nodes defines what we will call the inference error.

To find the posterior means and variances in linear-Gaussian state models one can use a message passing algorithm known as *Kalman Filter* [24] (see appendix A). For a long time series, the algorithm will converge to stationary values for the covariances when well away from the two ends $t = 0$ and $t = T$; note though that the state prediction $\boldsymbol{\mu}(t)$ remains time dependent as it is driven by the time dependence of the observed $\boldsymbol{x}^{\mathrm{s}}(t)$. The covariances, on the other hand, are entirely independent of the $\boldsymbol{x}^{\mathrm{s}}(t)$, by a general property of conditional Gaussian distributions: they depend only on which variables are observed, but not their values. Note that this contrasts with the case of e.g. binary spins, where mean and variance are directly related so that variances of individual spins would generally also be non-stationary.

The stationary inference error, i.e. the normalized trace of the stationary equal time posterior covariance $\boldsymbol{C}(t,t) = \boldsymbol{C}$, will be the main focus of our attention. As shown in appendix A, $\boldsymbol{C}$ satisfies

$$\boldsymbol{K}^{\mathrm{bb|s}}\boldsymbol{C} + \boldsymbol{C}\boldsymbol{K}^{\mathrm{bb|s}\,T} + \boldsymbol{\Sigma}^{\mathrm{bb}} = 0. \tag{7}$$

This is a Lyapunov equation with an "effective" or "posterior" drift $\boldsymbol{K}^{\mathrm{bb|s}}$, where we use the superscript bb|s to indicate that this is the bulk-bulk coupling matrix conditioned on the observed subnetwork trajectory. By "posterior" we mean then that $\boldsymbol{K}^{\mathrm{bb|s}}$ incorporates the effect of the observations and defines an effective posterior dynamics

$$\partial_t \delta\boldsymbol{x}^{\mathrm{b}}(t) = \boldsymbol{K}^{\mathrm{bb|s}}\delta\boldsymbol{x}^{\mathrm{b}}(t) + \boldsymbol{\xi}^{\mathrm{b}}(t). \tag{8}$$

The effective drift can be written as

$$\boldsymbol{K}^{\mathrm{bb|s}} = \boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{A}, \tag{9}$$

where $\boldsymbol{A} = \boldsymbol{A}^{T}$ is a symmetric matrix that is a solution of the matrix Riccati (i.e. quadratic) equation

$$\boldsymbol{A}\boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{A} - \boldsymbol{A}\boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{K}^{\mathrm{bb}\,T}\boldsymbol{A} = \boldsymbol{W}. \tag{10}$$

Here the *feedback* matrix $\boldsymbol{W} = \boldsymbol{K}^{\mathrm{sb}\,T}(\boldsymbol{\Sigma}^{\mathrm{ss}})^{-1}\boldsymbol{K}^{\mathrm{sb}}$ describes how observations affect the inferred statistics. This matrix is determined by the interplay between the strength of hidden-observed interactions $\boldsymbol{K}^{\mathrm{sb}}$ and the dynamical noise on the observed variables, namely $\boldsymbol{\Sigma}^{\mathrm{ss}}$. (We stress here that this is noise acting on the time evolution of $\boldsymbol{x}^{\mathrm{s}}$, *not* noise affecting our measurement of the observed trajectory.)

The matrix $\boldsymbol{A}$ in (9) is directly related to the backwards messages sent in the Kalman filter method. Specifically, the distribution of $\delta\boldsymbol{x}^{\mathrm{b}}(t)$ conditioned only on observations *from time t onwards* is Gaussian, and $\boldsymbol{A}$ is its inverse covariance in the stationary regime.

Accordingly, equation (10) can be derived as the stationary limit of what is known as a Riccati recursion, for the backward pass in the Kalman Filter (see appendix A). Without observations the distribution of $\boldsymbol{x}^{\mathrm{b}}(t)$ conditional only on data beyond t is flat, hence $\boldsymbol{A}$ vanishes. Then $\boldsymbol{K}^{\mathrm{bb|s}}$ reduces to $\boldsymbol{K}^{\mathrm{bb}}$ as expected and the posterior covariance to the unconditional covariance because (7) becomes simply $\boldsymbol{K}^{\mathrm{bb}}\boldsymbol{C} + \boldsymbol{C}\boldsymbol{K}^{\mathrm{bb}\,T} + \boldsymbol{\Sigma}^{\mathrm{bb}} = 0$. One sees therefore that $\boldsymbol{A}$ is the key quantity that captures the effects of the observations on the (second order) posterior statistics. This insight is supported by an alternative variational derivation of (7), (9) and (10), outlined in Appendix B, where $\boldsymbol{A}$ appears as a Lagrange multiplier implementing the constraints resulting from the observed data.

Once the stationary equal-time covariance $\boldsymbol{C}$ has been found, it is clear from (8) that the two-time covariance must be given by

$$\boldsymbol{C}(t-t') = e^{\boldsymbol{K}^{\mathrm{bb|s}}(t-t')}\boldsymbol{C} \tag{11}$$

for $t > t'$. This exponential decay with the effective drift matrix $\boldsymbol{K}^{\mathrm{bb|s}}$ can be derived explicitly by generalizing the filtering-smoothing procedure (see appendix A and references there). We have emphasized in the notation the fact that $\boldsymbol{C}(t-t')$ depends only on the time difference because the stationary regime obeys time-translation invariance. Stability of the conditional hidden dynamics, where (11) decays to zero as $t-t'$ grows, requires $\boldsymbol{K}^{\mathrm{bb|s}}$ to be negative definite. Assuming that the dynamical matrix $\boldsymbol{K}^{\mathrm{bb}}$ of the isolated hidden dynamics has this property, then also $\boldsymbol{K}^{\mathrm{bb|s}}$ does because $\boldsymbol{A}$, as the inverse covariance matrix in the stationary backwards messages, is non-negative definite.

So far in this section we have derived expressions for $\boldsymbol{C}$ and $\boldsymbol{C}(t-t')$ that specify the second order posterior statistics in our setting of inferring hidden state trajectories. These results are valid for given values of the interaction matrices $\boldsymbol{K}^{\mathrm{bb}}$ etc. In the remainder of the paper we consider these interactions to be drawn from some probability distribution, acting as *quenched disorder*. In an appropriately defined infinite size or thermodynamic limit we then expect key results such as the eigenvalue spectrum of $\boldsymbol{C}$ to be self-averaging, i.e. independent of the specific realization. In particular we look at a fully connected system interacting via Gaussian couplings. This is a standard scenario used to analyze the mean-field regime of e.g. spin glass models [28]. It can also be thought of as the large connectivity limit of an Erdős-Rényi graph [29] with Gaussian weights [30]; studying dynamical processes on such random graphs to predict the evolution of each node from partial observations is of interest in e.g. epidemic forecasting [31, 32]. A precedent for the use of RMT techniques, such as Stieltjes transforms and free probability, in the study of asymptotic eigenvalue distributions for random Lyapunov and Riccati recursions – like those occuring in filtering – can be found in [26]. Ref. [26] takes a control and systems theory perspective, however, while we focus on inference for dynamics. It is worth stressing that this makes our approach more general, as we look at a time dependent problem with quenched, "frozen" randomness rather than a sequence of signals where the randomness in the interactions is re-sampled at each step. From the spectrum $\boldsymbol{C}$ we will obtain the inference error; we will also study the properties of the posterior drift $\boldsymbol{K}^{\mathrm{bb|s}}$, whose inverse defines the spectrum of relaxation times of the posterior dynamics.

III. THERMODYNAMIC LIMIT BY RANDOM MATRIX THEORY

To investigate the thermodynamic limit, we first apply tools from random matrix theory (RMT) to *equilibrium* dynamics, where detailed balance holds. We study two such scenarios. In the first, the hidden variables only have self-interactions (Sec. III B); in the second we add random symmmetric hidden-to-hidden interactions (Sec. III C). The main results are explicit mathematical expressions which establish a link between the inference error and the parameters describing the dynamics. In both cases we make the same assumptions regarding the hidden-to-observed interactions $\boldsymbol{K}^{\mathrm{sb}}$, and therefore discuss first the resulting statistics of the feedback matrix $\boldsymbol{W}$.

A. Feedback matrix: Wishart ensemble

The feedback matrix $\boldsymbol{W} = \boldsymbol{K}^{\mathrm{sb}\,T}(\boldsymbol{\Sigma}^{\mathrm{ss}})^{-1}\boldsymbol{K}^{\mathrm{sb}}$ is a positive definite symmetric matrix of size $N^{\mathrm{b}} \times N^{\mathrm{b}}$, where N^{b} is the number of hidden variables, i.e. the number of components of the vector $\boldsymbol{x}^{\mathrm{b}}$. We assume throughout in the following that the elements of the $N^{\mathrm{s}} \times N^{\mathrm{b}}$ matrix $\boldsymbol{K}^{\mathrm{sb}}$ are independent zero mean Gaussian random variables of fixed variance k^2/N^{b}. If $\boldsymbol{\Sigma}^{\mathrm{ss}} = \sigma_{\mathrm{s}}^2\mathbb{1}$ is isotropic, $\boldsymbol{W}$ is then a sample from a *Wishart* random matrix ensemble, whose spectral properties are well understood [25]. In the thermodynamic limit of infinitely large matrices, $N^{\mathrm{b}} \to \infty$, and up to an overall scale of the eigenvalues, the eigenvalue density of $\boldsymbol{W}$ is thus given by the *Marčenko-Pastur* law (MP) [33]

$$\rho_\alpha(\hat{w}) = (1-\alpha)\Theta(1-\alpha)\delta(\hat{w}) + f_\alpha(\hat{w}), \qquad (12)$$

where

$$f_\alpha(\hat{w}) = \frac{1}{2\pi\hat{w}}\sqrt{(\hat{w}-\hat{w}_-)(\hat{w}_+-\hat{w})} \qquad (13)$$

and is to be read as nonzero only when $\hat{w}$ lies in the interval $[\hat{w}_-, \hat{w}_+]$ with $\hat{w}_\pm = \left(\sqrt{\alpha} \pm 1\right)^2$. The delta peak at $\hat{w} = 0$ in (12) contributes only when $\alpha < 1$, as indicated by the Heaviside step function $\Theta(\cdot)$. Here we have defined $\alpha = N^{\mathrm{s}}/N^{\mathrm{b}} = N^{\mathrm{observed}}/N^{\mathrm{hidden}}$ as the fundamental parameter of our analysis, giving the ratio and thus the relative importance of the sizes of the observed and unknown "sectors" of our network. This parameter resembles the storage ratio [4, 6], or number of training examples per parameter to be learned, in neural network learning. Indeed, in the context of learning linear relationships from examples, the distribution (12) also gives the spectrum of the input correlation matrix governing the learning dynamics [7–9, 11].

In the spectrum (12) the δ peak at $\hat{w} = 0$ arises from the $N^{\mathrm{b}} - N^{\mathrm{s}} = N^{\mathrm{b}}(1-\alpha)$ directions in the hidden state space that are not directly constrained by observations when $\alpha < 1$. The remaining $f_\alpha(\hat{w})$ piece is a semi-circle in the interval $[\hat{w}_-, \hat{w}_+]$, distorted by a factor $1/\hat{w}$. For $\alpha > 1$ this is the only contribution; in the limit $\alpha \gg 1$ the relative variance of the eigenvalues around their mean $\langle \hat{w} \rangle = \alpha$ goes to zero.

B. Self-interacting hidden variables

1. Inference error and relaxation times

We assume below that the noise acting on bulk variables is isotropic, $\boldsymbol{\Sigma}^{\mathrm{bb}} = \sigma_{\mathrm{b}}^2 \mathbb{1}$, as already assumed for the subnetwork noise. This is equivalent to assuming that the amplitude of fluctuations is homogeneous within the hidden system, as it would be if it was given by a physical temperature. Anisotropies would add non-trivial correlations between d.o.f. that would obscure the effect of interactions, which is our main focus here. In this section we further restrict ourselves to interactions between bulk and subnetwork, by taking $\boldsymbol{K}^{\mathrm{bb}} = -\lambda \mathbb{1}$ where the self-interaction λ is the only interaction among hidden variables. Given this, any interesting behavior has to come from observations.

By simultaneously diagonalizing $\boldsymbol{W}$ and $\boldsymbol{A}$, (10) reduces to a scalar equation relating the eigenvalues of these matrices, respectively w and a, as

$$\sigma_{\mathrm{b}}^2 a^2 + 2\lambda\, a = \frac{k^2}{\sigma_{\mathrm{s}}^2} \hat{w}, \tag{14}$$

where we have extracted from w an amplitude factor by writing $w = k^2 \hat{w}/\sigma_{\mathrm{s}}^2$, k being the amplitude for the $\boldsymbol{K}^{\mathrm{sb}}$ entries and $\hat{w}$ a dimensionless Wishart random variable. The physical solution for a is

$$a = \frac{-\lambda + \sqrt{\lambda^2 + \sigma^2 \hat{w}}}{\sigma_{\mathrm{b}}^2}, \tag{15}$$

with the shorthand $\sigma = \sigma_{\mathrm{b}} k/\sigma_{\mathrm{s}}$. By diagonalizing (9) one then gets for the eigenvalues of $\boldsymbol{K}^{\mathrm{bb|s}}$, which we denote by r

$$r = -\lambda - a\, \sigma_{\mathrm{b}}^2 = -\sqrt{\lambda^2 + \sigma^2 \hat{w}}. \tag{16}$$

From (8) and (11), the distribution of $-r$ gives the relaxation rate spectrum of the posterior dynamics, and (16) shows that these rates are increased by observations, i.e. correlations get shorter in time. As expected this effect gets stronger as the hidden-observed interaction amplitude k increases, at fixed ratio $\sigma_{\mathrm{b}}/\sigma_{\mathrm{s}}$.

From (16) we can now find the spectrum of r as the appropriate transformation of the MP law

$$\rho(r) = (1-\alpha)\Theta(1-\alpha)\delta(r+\lambda) + f(\hat{w}(r))|\hat{w}'(r)|, \tag{17}$$

where $f(\hat{w}(r))$ is defined only between $r_\pm = \sqrt{\sigma^2(\sqrt{\alpha} \pm 1)^2 + \lambda^2}$ and $\hat{w}(r) = -(r^2 + \lambda^2)/\sigma^2$ is the inverse function of (16). The first piece, a δ-function at $r = -\lambda$, describes the behavior for hidden state space directions unconstrained by observations. The above result for the spectrum can also be expressed as a spectrum $\rho(\tau) = \rho(r)/\tau^2$ of relaxation times $\tau = -1/r$ for the posterior dynamics. We sometimes plot $\rho(\ln \tau) = \tau \rho(\tau)$ to show the full range of τ; this $\ln \tau$-spectrum is the same as the one of $\ln r$ up to a sign change, with spectral edges at $\tau_\pm = -1/r_\mp$ (see figure 1(a)).

The long-time ($t - t' \gg 1$) behavior of the posterior covariance is an exponential decay whose characteristic time can be defined in different ways. The slowest relaxation time is $\tau_{\max} = 1/r_{\min}$, where $r_{\min}$ is the minimum eigenvalue of $-\boldsymbol{K}^{\mathrm{bb|s}}$

$$r_{\min} = \sqrt{\lambda^2 + \sigma^2 \hat{w}_{\min}} = \\ = \begin{cases} \lambda & \alpha \le 1 \\ \sqrt{\lambda^2 + \sigma^2(\sqrt{\alpha} - 1)^2} & \alpha > 1. \end{cases} \tag{18}$$

One can also look at a relaxation time defined as the average over the spectrum $\rho(\tau)$, i.e. $\langle \tau \rangle = \int d\tau \rho(\tau)\, \tau$. Or finally one can consider a root mean square correlation decay time

$$\tau^{*\,2} = \frac{\int_{-\infty}^{+\infty} t^2 C(t) dt}{2\tilde{C}(0)} = -\frac{1}{2\tilde{C}(0)} \frac{d^2 \tilde{C}(\mathrm{i}\omega)}{d^2 \omega}\bigg|_{\omega=0}, \tag{19}$$

where the power spectrum $\tilde{C}(\mathrm{i}\omega)$ is obtained by setting $z = \mathrm{i}\omega$ in the Laplace transform (see equation (24) below) of the correlator $C(t-t') = \mathrm{Tr}\, \boldsymbol{C}(t-t')$ (trace normalized by N^{b}). It is easy to verify that all three relaxation times exhibit the same asymptotic decay $\sim 1/(\sigma\sqrt{\alpha})$ for large α. In figure 2(a) we show a comparison at smaller α. With only few observations, all measures of posterior correlation time are close to the $\alpha = 0$ value $1/\lambda$ while for $\alpha > 1$ they start decreasing, crossing over to the $1/\sqrt{\alpha}$ large α tail; $\tau_{\max}$ shows the least smooth transition between these two regimes. We can summarize the behavior by saying that with more observations the posterior fluctuations (or error bars on the inferred means) become less correlated in time as predictions become more "tied" to the data observed at any specific moment. This effect is seen in more detail in figure 1(b) where with increasing α the relaxation time spectrum becomes more peaked and shifts towards shorter times. The posterior covariance matrix $\boldsymbol{C}$ has the same set of eigenmodes as $\boldsymbol{K}^{\mathrm{bb|s}}$ in the current scenario because in (7) all matrices can be simultaneously diagonalized. The eigenvalues C of $\boldsymbol{C}$ give the posterior variance for each mode, which from (7) is related to r or τ by

$$C = -\frac{\sigma_{\mathrm{b}}^2}{2r} = \frac{\sigma_{\mathrm{b}}^2}{2}\tau = \frac{\sigma_{\mathrm{b}}^2}{2\sqrt{\lambda^2+\sigma^2\hat{w}}}. \tag{20}$$

This shows that C decreases with increasing feedback values $\hat{w}$: observations increase prediction accuracy as they should. Because $C \propto \tau$, the above results for the spectrum of τ also apply to that of C; see figures 1 and 2(a). For large α in particular the spectrum of C becomes a narrow peak around the asymptotic inference error $C \approx \sigma_{\mathrm{b}}^2/(\sigma\sqrt{\alpha})$.

We note as an aside that from the proportionality $C \propto \tau$ one can show that the relaxation time τ^* defined in (19) can be written in terms of spectral averages as

$$\tau^* = \sqrt{\frac{\langle\tau^4\rangle}{\langle\tau^2\rangle}}. \tag{21}$$

Because $\langle\tau\rangle^2\langle\tau^2\rangle \le \langle\tau^4\rangle$, this implies generally $\langle\tau\rangle \le \tau^*$ in agreement with the results in figure 2(a).

2. *Posterior covariance in Laplace space*

We next turn to the temporal dependence of the posterior covariance (11). Its trace, normalized by N^{b}, is an average of the contributions from the different eigenmodes of $\boldsymbol{K}^{\mathrm{bb|s}}$. In terms of the relevant eigenvalues $\hat{w}$ and using (20) these are

$$C_{\hat{w}}(t-t') = e^{r|t-t'|}C = -\frac{\sigma_{\mathrm{b}}^2}{2r}e^{r|t-t'|}, \tag{22}$$

with an added subscript $\hat{w}$ to indicate this is the contribution from a single eigenmode, characterized by a specific value of $\hat{w}$. We take the double-sided Laplace transform

$$\begin{aligned}\tilde{C}_{\hat{w}}(z) &= \frac{\sigma_{\mathrm{b}}^2}{2r}\int_{-\infty}^{+\infty} e^{-(z+r)|t'-t|}dt' \\ &= \frac{\sigma_{\mathrm{s}}^2}{k^2}\frac{1}{\frac{\lambda^2-z^2}{\sigma^2}+\hat{w}},\end{aligned} \tag{23}$$

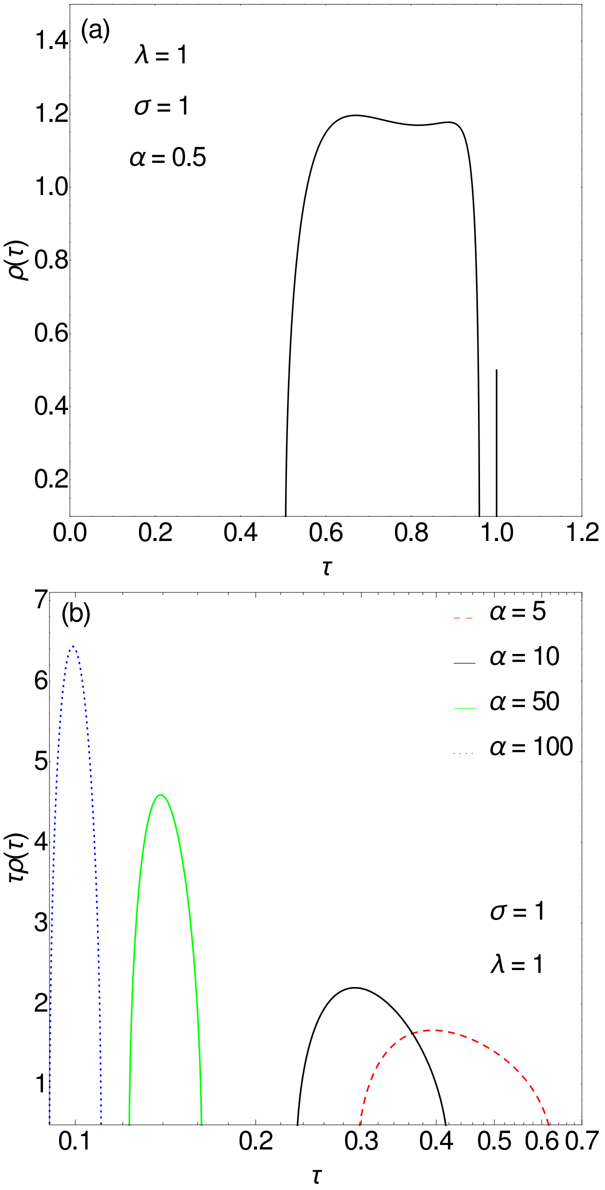

Figure 1. (a) Spectral density $\rho(\tau)$ for $\alpha = 0.5$: the vertical line indicates the δ-peak of height $1-\alpha$ at $\tau = 1/\lambda$, the relaxation time in the absence of observations. (b) Spectral density $\rho(\ln\tau) = \tau\rho(\tau)$ of $\ln\tau$: this shifts to smaller $\ln\tau$ as α increases, indicating shorter posterior correlation times. The spectrum also narrows and becomes concentrated around $\tau = 1/\sigma\sqrt{\alpha}$ for large α. As the posterior variance $C \propto \tau$ for each hidden space mode, the distributions of $\ln C$ differ only from those of $\ln\tau$ by a horizontal shift.

where we have substituted (16) for r in terms of the self-interaction λ and the feedback matrix eigenvalues $k^2\hat{w}/\sigma_{\rm s}^2$.

In the thermodynamic limit, we can then get the Laplace transform of the overall covariance normalized trace $C(t-t') = \mathrm{Tr}\,\boldsymbol{C}(t-t')$ by averaging over the Marčenko-Pastur spectrum $\rho(\hat{w})$, yielding

$$\tilde{C}(z) = \frac{\sigma_{\rm s}^2}{2k^2}\frac{\sigma^2}{(\lambda^2-z^2)}\left\{1-\alpha-\left(\frac{\lambda^2-z^2}{\sigma^2}\right)+\sqrt{\left[1-\alpha-\left(\frac{\lambda^2-z^2}{\sigma^2}\right)\right]^2+4\left(\frac{\lambda^2-z^2}{\sigma^2}\right)}\right\}. \tag{24}$$

One can verify that $\tilde{C}(0)$ has a divergence for $\lambda/\sigma \to 0$ and $\alpha \leq 1$; the small α-curves in figure 2(b) illustrate this effect. See also [18] for a systematic study of the approach to such divergences.

C. Symmetric hidden-hidden couplings

In this section we generalize the above scenario by assuming that $\boldsymbol{K}^{\rm bb} = -\lambda\mathbb{1} + \boldsymbol{J}$. Here the matrix $\boldsymbol{J}$ provides explicit hidden-to-hidden interactions beyond the self-interaction term $-\lambda\mathbb{1}$ we have had so far. To ensure stability of the hidden system, one requires $\lambda > \lambda_{\rm c}$ where $\lambda_{\rm c}$ is the largest eigenvalue of $\boldsymbol{J}$.

We assume that $\boldsymbol{J}$ is symmetric, which is required for any steady state of the whole system to be at equilibrium, i.e. to obey detailed balance. The posterior drift $\boldsymbol{K}^{\rm bb|s}$ from (9) is then also a symmetric matrix. This is crucial as it allows one to solve (7) and (10) in closed form. Eq. (7) gives

$$\boldsymbol{C} = -\frac{\sigma_{\rm b}^2}{2}\left(\boldsymbol{K}^{\rm bb|s}\right)^{-1}, \tag{25}$$

which is positive definite because $\boldsymbol{K}^{\rm bb|s} = (-\lambda+\boldsymbol{J}) - \sigma_{\rm b}^2\boldsymbol{A}$ is negative definite. To eliminate the unknown $\boldsymbol{A}$, note from (10) that

$$\begin{aligned}\Big((-\lambda+\boldsymbol{J})-\sigma_{\rm b}^2\boldsymbol{A}\Big)^2 &= (-\lambda+\boldsymbol{J})^2+\sigma_{\rm b}^4\boldsymbol{A}^2\\ -\sigma_{\rm b}^2(-\lambda+\boldsymbol{J})\boldsymbol{A}-\boldsymbol{A}(-\lambda+\boldsymbol{J})\sigma_{\rm b}^2 &=\\ =(-\lambda+\boldsymbol{J})^2+\sigma_{\rm b}^2\boldsymbol{W} \doteq \boldsymbol{M},\end{aligned} \tag{26}$$

where the last equality defines $\boldsymbol{M}$. Hence

$$\boldsymbol{C} = \frac{\sigma_{\rm b}^2}{2}\boldsymbol{M}^{-1/2}, \qquad \boldsymbol{K}^{\rm bb|s} = -\boldsymbol{M}^{1/2}, \tag{27}$$

where $\boldsymbol{M}^{1/2}$ is the positive definite square root of $\boldsymbol{M}$ and $\boldsymbol{M}^{-1/2}$ its inverse.

1. *Free probability*

From (27), the spectrum of $\boldsymbol{M}$ directly determines those of $\boldsymbol{C}$ and $\boldsymbol{K}^{\rm bb|s}$. As a paradigmatic example where this spectrum can be obtained in the thermodynamic limit we consider the case where the elements of $\boldsymbol{J}$ are independently drawn from a Gaussian distribution, i.e. we set $\boldsymbol{J} = j\hat{\boldsymbol{J}}$ with $\hat{\boldsymbol{J}}$ a random matrix from the *Wigner* ensemble [25]. From the Wigner semi-circular law this has largest eigenvalue 2, thus $\lambda_{\rm c} = 2j$. We will write the feedback matrix as in section III B 2: $\boldsymbol{W} = \frac{k^2}{\sigma_{\rm s}^2}\hat{\boldsymbol{W}}$ with $\hat{\boldsymbol{W}}$ from the *Wishart* ensemble.

With the above assumptions, $\boldsymbol{M} = (-\lambda+\boldsymbol{J})^2 + \sigma_{\rm b}^2\boldsymbol{W}$ is a sum of two independently drawn, symmetric random matrices with known spectrum. Its spectrum can then be found using *free probability* theory. Reviews can be found in [34] for the theory and [35, 36] for applications to RMT. Briefly, the sum defining $\boldsymbol{M}$ is effectively a *free* addition [34] in the sense that because of independent sampling, the eigenvector bases of the two matrices in the sum are randomly rotated against each other. It then turns out that the spectrum of the sum depends only on the eigenvalues and not the eigenvectors of the individual matrices. The intuition beyond this is that, in the limit of infinite matrix size, the detailed statistics of eigenvalues, e.g. whether they are correlated or not, can be neglected [36]. While in an ordinary sum of independent random variables it is the cumulants that add, in a free sum of two random matrices it is the R-transforms that are additive [34], and this allows the spectrum of the sum to be determined.

The R transform of a random matrix is related to its Green's function by

$$G(z) = \frac{1}{z - R(G(z))}. \tag{28}$$

The Green's function or resolvent, in turn, is defined for a generic random matrix $\boldsymbol{M}$ as the normalized trace $G_M(z) = \mathrm{Tr}(z-\boldsymbol{M})^{-1}$. It can be written in

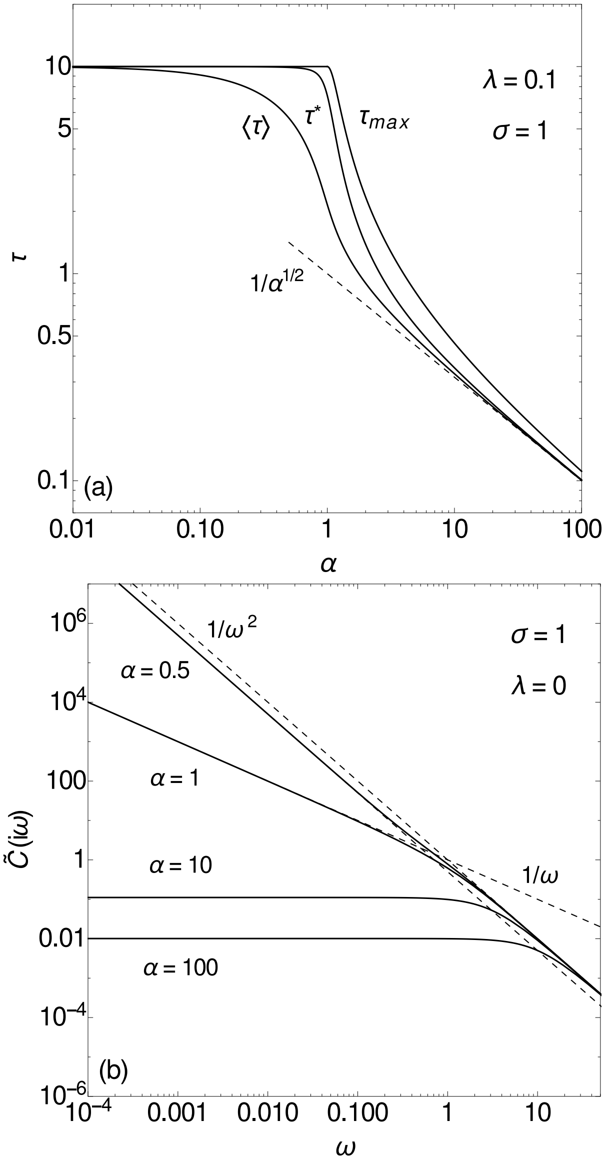

Figure 2. (a) Characteristic posterior relaxation time τ as a function of α, for $\lambda = 0.1$ and $\sigma = 1$, defined in three different ways (see text). For $\alpha \to 0$ all three curves approach $\tau = 1/\lambda = 10$; asymptotically they decay as $1/\sqrt{\alpha}$. (b) Posterior power spectrum (obtained by setting $z = \mathrm{i}\omega$ in (24)) for various α, at $\lambda = 0$. The power spectrum diverges as $\omega \to 0$ when $\alpha \leq 1$. For small α the divergence is $\propto 1/\omega^2$, crossing over to $\propto 1/\omega$ as $\alpha \to 1$. Beyond $\omega \sim O(1)$ the curves for all α exhibit a standard Lorentzian tail $1/\omega^2$. See [18] for a derivation of these power laws.

terms of the eigenvalue density $\rho(m)$ as

$$G_M(z) = \int \frac{\rho(m)}{z - m} dm, \tag{29}$$

which is also known as a Stieltjes transform. Conversely, $\rho(m)$ can be retrieved from the Green's function via

$$\rho(m) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \operatorname{Im} G_M(m + \mathrm{i}\epsilon). \tag{30}$$

The route to finding the spectrum of $\boldsymbol{M}$ in our case is then clear: we need to write the Green's functions and associated R-transforms of $(-\lambda + \boldsymbol{J})^2$ and $\sigma_\mathrm{b}^2 \boldsymbol{W}$, respectively, add these two R-transforms to obtain the R-transform of $\boldsymbol{M}$, and then work backwards to $G_M(z)$ and finally $\rho(m)$.

We denote by $G_1(z)$ the Green's function of $(-\lambda + \boldsymbol{J})^2$, which is given by the integral

$$\begin{aligned} G_1(z) &= \int \frac{\rho(\hat{j})}{z - (-\lambda + j\hat{j})^2} d\hat{j} \\ &= \int_{-2}^{2} \frac{\sqrt{4 - \hat{j}^2}}{2\pi} \frac{1}{z - (-\lambda + j\hat{j})^2} d\hat{j}, \end{aligned} \tag{31}$$

where the Wigner semicircular law has been used. The integral can be performed in closed form

$$\begin{aligned} G_1(z) = \frac{1}{2j^2} &- \frac{1}{4j^2} \sqrt{\frac{\left(\lambda - \sqrt{z}\right)^2 - 4j^2}{z}} \\ &- \frac{1}{4j^2} \sqrt{\frac{\left(\lambda + \sqrt{z}\right)^2 - 4j^2}{z}} \end{aligned} \tag{32}$$

and (28) then gives the R-transform

$$R_1(z) = \frac{j^2}{1 - zj^2} + \frac{\lambda^2}{\left(1 - 2zj^2\right)^2}. \tag{33}$$

The Green's function for a Wishart matrix is well known [11] and the related R transform reads

$$R_2(z) = \frac{\alpha v}{1 - vz}, \tag{34}$$

where we recall that $\alpha = N^\mathrm{s}/N^\mathrm{b}$ and v, the variance, in our case is $v = k^2\sigma_\mathrm{b}^2/\sigma_\mathrm{s}^2$. The two above R-transforms now simply add to give the one for $\boldsymbol{M}$, $R_M(z) = R_1(z) + R_2(z)$. The result can be written as an implicit expression for the Green's function $G_M(z)$, given that from (28) one has generally $z(G) = 1/G + R(G)$

$$z = \frac{1}{G} + \frac{\alpha \frac{k^2\sigma_\mathrm{b}^2}{\sigma_\mathrm{s}^2}}{1 - \frac{k^2\sigma_\mathrm{b}^2}{\sigma_\mathrm{s}^2} G} + \frac{j^2}{1 - j^2 G} + \frac{\lambda^2}{\left(1 - 2j^2 G\right)^2}. \tag{35}$$

We have abbreviated $G \equiv G_M$ on the r.h.s. here. Rearranging the above equation one sees that $G(z)$ is the solution of a fifth order polynomial equation. This can be found numerically, with the correct solution branch being determined from the asymptotic behavior $G \approx 1/z$ for large z. Once $G(z)$ is in hand, $\rho(m)$ can be found using (30).

By a transformation of the spectrum of $\boldsymbol{M}$ we can characterize the spectrum of the posterior covariance matrix $\boldsymbol{C} = \sigma_{\mathrm{b}}^2 \boldsymbol{M}^{-1/2}/2$ as well as the spectrum of relaxation rates as determined by the effective drift $\boldsymbol{K}^{\mathrm{bb|s}} = -\boldsymbol{M}^{1/2}$. The spectrum of $(-\boldsymbol{K}^{\mathrm{bb|s}})^{-1} = \boldsymbol{M}^{-1/2}$ then gives the distribution of relaxation times. As this matrix is proportional to $\boldsymbol{C}$, plots of $\rho(\tau)$ (figure 3) provide information also about the inference error as a function of α. The overall picture is that predictions become increasingly precise when the pool of observed data is expanded, i.e. α increases, while correlation times between posterior fluctuations decrease in proportion.

For qualitative analysis one can rewrite (35) in dimensionless variables $\tilde{z} = \sigma_{\mathrm{s}}^2 z/(k^2\sigma_{\mathrm{b}}^2)$ and $\tilde{G} = k^2\sigma_{\mathrm{b}}^2 G/\sigma_{\mathrm{s}}^2$ as

$$\tilde{z} = \frac{1}{\tilde{G}} + \frac{\alpha}{1-\tilde{G}} + \frac{(\gamma p)^2}{1-(\gamma p)^2\tilde{G}} + \frac{p^2}{\left(1-2(\gamma p)^2\tilde{G}\right)^2}, \tag{36}$$

where $\gamma = j/\lambda$ and $p = \lambda/\sigma$. This reduces the number of parameters and variables, from seven (α, j, k, λ, σ_{s}, σ_{b}, z) to four (p, γ, α, $\tilde{z}$). Here γ and $1/p$ measure the strength of hidden-hidden and hidden-observed couplings relative to the decay weight λ.

We have seen in figure 1(a) that for $\gamma = 0$, i.e. in the absence of hidden-hidden interactions (see section III B 1) the spectrum consists of two separate pieces for $\alpha < 1$, while with such interactions present ($\gamma > 0$) the spectrum can be supported on a single interval. There must be a transition between these two cases at some value of γ that will depend on p and α - see figure 4 (a). Locating this transition numerically gives the results shown in figure 4(b). The spectrum consists of a single piece *above* the line drawn in the (p, γ) plane. One sees that for large $p = \lambda/\sigma = \lambda\sigma_{\mathrm{s}}/(\sigma_{\mathrm{b}}k)$, i.e. weaker hidden-observed couplings, small values of $\gamma = j/\lambda$ and hence weak hidden-hidden interactions are sufficient to merge the two pieces of the spectrum.

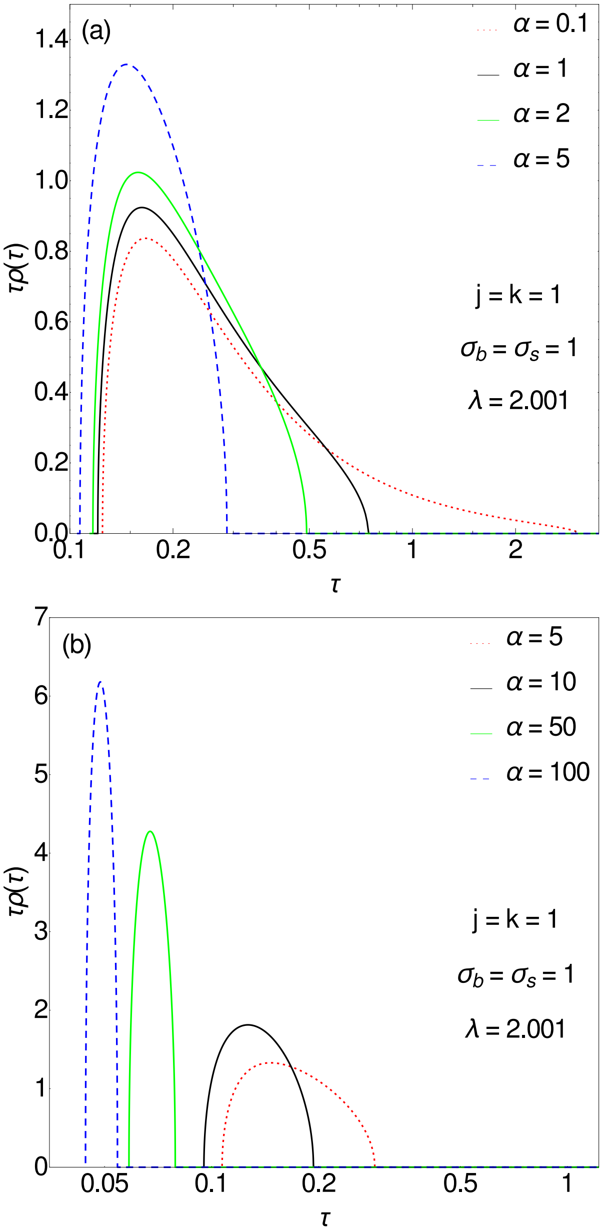

Figure 3. Spectral density $\rho(\ln\tau) = \tau\rho(\tau)$, of relaxation times τ, for different values of α. We plot $\rho(\ln\tau)$ to make the normalization of the densities more obvious. The spectra of posterior variances C, which define the inference error, are identical up to a horizontal shift as $C \propto \tau$. (a) At small α the spectrum is broad, indicating that there is much variation in how different hidden state space directions are constrained by observations. For increasing α the spectrum becomes more peaked, and centred around decreasing τ or C: different directions become determined more strongly, and more evenly, by observations, a trend more clearly visible in (b).

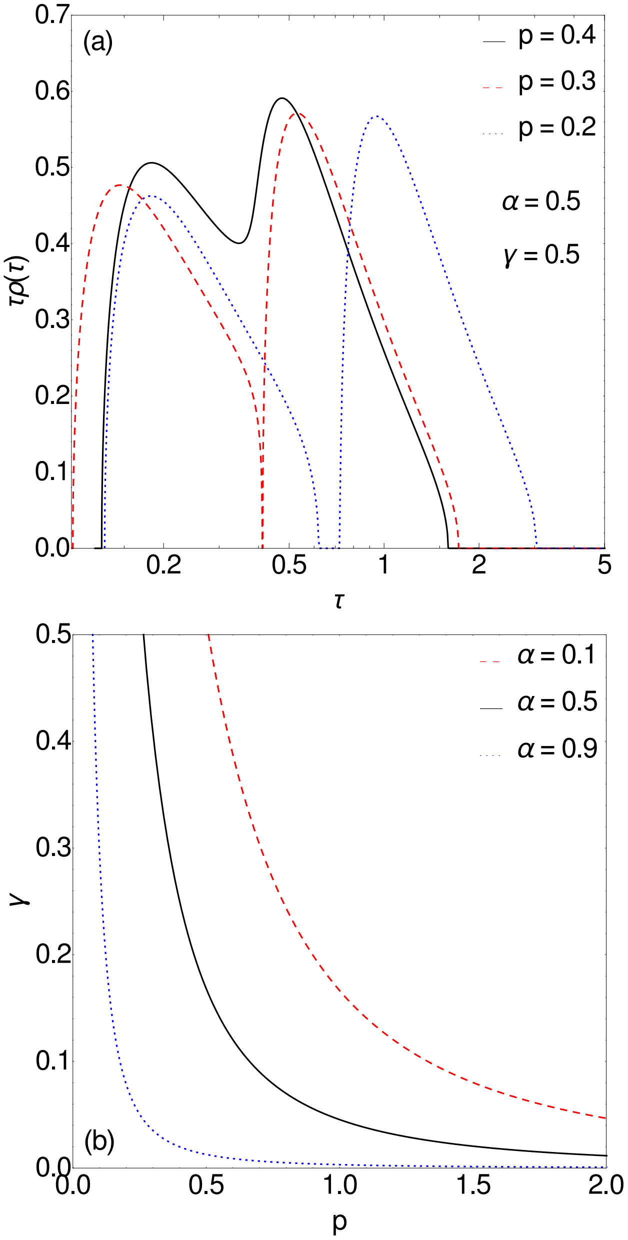

Figure 4. (a) Spectral density $\rho(\ln\tau) = \tau\rho(\tau)$, at $\gamma = j/\lambda = 0.5$ (critical value for internal stability, with $j = 0.2$ and $\lambda = 0.4$) and $\alpha = 0.5$ for different values of p: the two pieces of the spectrum at $p = 0.2$ merge at $p = 0.3$, giving a spectrum supported on a single interval for $p > 0.3$. (b) Curve in the (p, γ) plane for which the two pieces of the spectrum merge when coming from low γ: the black line refers to $\alpha = 0.5$, the case shown in (a). The two-piece region near the origin shrinks (see curve for $\alpha = 0.9$, blue dotted line) and vanishes for $\alpha \to 1$.

2. *Posterior correlations in Laplace space*

From (11) and (27) we can obtain explicitly the posterior correlations in time: for $t > t'$,

$$\boldsymbol{C}(t - t') = \frac{\sigma_{\rm b}^2}{2} e^{-\boldsymbol{M}^{1/2}(t-t')} \boldsymbol{M}^{-1/2}. \qquad (37)$$

We consider the trace, which at $t = t'$ gives the total posterior variance. The double-sided Laplace transform can then be shown to have the simple form

$$\tilde{C}(z) = \sigma_{\rm b}^2 \, \mathrm{Tr} \left(-z^2 + \boldsymbol{M}\right)^{-1} = -\sigma_{\rm b}^2 G_M(z^2). \qquad (38)$$

This relation to the Green's function is in fact a statement of the Fluctuation-Dissipation Theorem [37] (see [38] for details) and holds true because of the symmetry of $\boldsymbol{J}$.

From (38), the Laplace transformed posterior correlation function has to satisfy the equation for $-\sigma_{\rm b}^2 G_M(z^2)$, giving

$$z^2 = -\frac{\sigma_{\rm b}^2}{\tilde{C}} + \frac{\alpha \frac{k^2 \sigma_{\rm b}^2}{\sigma_{\rm s}^2}}{1 + \frac{k^2}{\sigma_{\rm s}^2}\tilde{C}} + \frac{j^2}{1 + \frac{j^2}{\sigma_{\rm b}^2}\tilde{C}} + \frac{\lambda^2}{\left(1 + 2\frac{j^2}{\sigma_{\rm b}^2}\tilde{C}\right)^2}, \qquad (39)$$

where we have set $\tilde{C}(z) = \tilde{C}$. Interestingly, and similarly to (35) which determines the spectrum of $\boldsymbol{M}$, this equation does not become singular at $\lambda = 0$. This fact can be understood in the following way. If directions exist along which the hidden dynamics would grow exponentially without observations, then these always have a non-zero overlap with directions constrained by observed data. This is clear from the independent sampling of the two terms in $\boldsymbol{M}$, and explains how the posterior variance, the uncertainty on the hidden dynamics, can stay finite even when the hidden dynamics without observations would diverge. Nevertheless, such a diverging hidden dynamics is an unphysical situation. We therefore continue to consider only parameter sets with $\lambda > \lambda_{\rm c}$, the internal dynamical condition for a finite and well-defined marginal dynamics of the bulk.

Finally, by setting $z = \mathrm{i}\omega$ one can evaluate the posterior power spectrum $\tilde{C}(\mathrm{i}\omega)$. It can be written in terms of a dimensionless function $\mathcal{C}_{\alpha,p,\gamma}(\Omega)$

$$\tilde{C}(\mathrm{i}\omega) = \frac{\sigma_{\rm s}^2}{k^2} \mathcal{C}_{\alpha,p,\gamma}(\Omega), \qquad (40)$$

with $\Omega = \omega/\sigma$ a rescaled frequency. The prefactor shows that the entire power spectrum of the posterior variance or prediction uncertainty is directly proportional to the dynamical noise acting on the observed subnetwork $\sigma_{\rm s}^2$ and inversely proportional to k^2, the strength with which it interacts with the bulk. As before one can find from (39) an equation for the dimensionless part $\mathcal{C}$

$$-\Omega^2 = -\frac{1}{\mathcal{C}} + \frac{\alpha}{1+\mathcal{C}} + \frac{(\gamma p)^2}{1+(\gamma p)^2\mathcal{C}} + \frac{p^2}{\left(1+2(\gamma p)^2\mathcal{C}\right)^2}, \tag{41}$$

where γ and p are defined as before. One can verify that for $p = 0$ and $0 \le \alpha \le 1$, $\mathcal{C}(0)$ has a divergence, implying also that the time integral of $\operatorname{Tr} \boldsymbol{C}(t-t')$ diverges. This comes physically from the fact that while a fraction α of hidden space directions have variances (and co-variances) of the expected order $\propto 1/k^2$, the others have variances that are independent of k and therefore much larger for large k.

A second region in the α, p, γ parameter space where $\mathcal{C}(0)$ diverges is $\alpha \to 0$ and $\gamma \to \gamma_c = 1/2$. This is as expected: without observations, the hidden dynamics starts to diverge at $\lambda \to \lambda_{\rm c} = 2j$, hence at $\gamma_c = 1/2$. We refer to [18] for further discussion of the behavior in the vicinity of such critical points.

The results of this section are of conceptual and practical significance. First, equation (35) for the Green's function provides a tool to study in a controlled way how spectra change with the number of observations and the interaction strength: this is what we show in figures 1, 3 and 4. Second, as more thoroughly analyzed in [18], from equations (39) and (41) one can calculate posterior equal time variances (by Fourier Transform) and relaxation times (by the second derivative at zero frequency, see (19)), which are exact in the thermodynamic limit and thus expected to be good approximations for large size datasets. Importantly, exact values such these can serve as a reference point around which one could systematically investigate finite size effects.

IV. THERMODYNAMIC LIMIT BY DYNAMICAL FUNCTIONALS

So far we have studied the posterior variance and time-dependent covariance in settings where the dynamics of the entire network obeys detailed balance, and where the relevant Green's functions can be derived using RMT tools.

In the absence of detailed balance, dynamical functionals can be used as an alternative, within a statistical mechanics approach to inference (for a systematic discussion see [4, 39]). The main result here is a generalization of (39) to any degree of symmetry, which therefore provides important insights into the strength of non-equilibrium effects on the inference error. We recall that the aim is to characterize a posterior path distribution, $P(\boldsymbol{X}^{\rm b}|\boldsymbol{X}^{\rm s})$, known to be Gaussian. The likelihood of the observed trajectory $P(\boldsymbol{X}^{\rm s})$ can be seen as a "partition function" Z that is obtained by summing $P(\boldsymbol{X}^{\rm b}, \boldsymbol{X}^{\rm s})$ over all possible hidden paths $\boldsymbol{X}^{\rm b}$. From Z, one can define a free energy (density) to study macroscopic quantities such as mean and covariance of $P(\boldsymbol{X}^{\rm b}|\boldsymbol{X}^{\rm s})$. If the interactions are chosen randomly, they act as quenched disorder and the physically relevant quantity is the quenched average of the free energy,

$$f = -\lim_{N\to\infty} N^{-1} \langle \ln Z(\boldsymbol{J}, \boldsymbol{K}^{\rm sb})\rangle_{\boldsymbol{J},\boldsymbol{K}^{\rm sb}}, \tag{42}$$

where we have abbreviated $N^{\rm b} \equiv N$. The free energy $-N^{-1}\ln Z$ is self-averaging, i.e. its fluctuations around f for different realizations of the disorder vanish for $N \to \infty$. The same is true for the order parameters that arise in the calculation, which include the posterior variance, i.e. inference error.

Dynamical functionals appear in the above approach once we write the joint path probability $P(\boldsymbol{X}^{\rm b}, \boldsymbol{X}^{\rm s})$ defined by the dynamics (1) and (2) in Onsager-Machlup form as proportional to

$$\begin{aligned} &P(\boldsymbol{X}^{\rm b}, \boldsymbol{X}^{\rm s}) \propto \qquad\qquad (43) \\ &\exp\left[-\frac{1}{2\sigma_{\rm b}^2}\int_0^T \left|\left|\partial_t \boldsymbol{x}^{\rm b} - \boldsymbol{K}^{\rm bs}\boldsymbol{x}^{\rm s}(t) - \boldsymbol{K}^{\rm bb}\boldsymbol{x}^{\rm b}(t)\right|\right|^2 dt\right] \\ &\cdot \exp\left[-\frac{1}{2\sigma_{\rm s}^2}\int_0^T \left|\left|\partial_t \boldsymbol{x}^{\rm s} - \boldsymbol{K}^{\rm ss}\boldsymbol{x}^{\rm s}(t) - \boldsymbol{K}^{\rm sb}\boldsymbol{x}^{\rm b}(t)\right|\right|^2 dt\right], \end{aligned}$$

with $\boldsymbol{K}^{\rm bb} = -\lambda\mathbb{1} + \boldsymbol{J}$. From the Gaussian form of this, the second order statistics of the posterior $P(\boldsymbol{X}^{\rm b}|\boldsymbol{X}^{\rm s})$ are independent of the value of the observed $\boldsymbol{X}^{\rm s}$. Hence to obtain the posterior variance it is sufficient to consider *zero observations*, i.e. $x_a(t) = 0$ for all a and t. All $\boldsymbol{x}^{\rm b}$ are then effectively deviations $\delta\boldsymbol{x}^{\rm b}$ from the posterior mean, though we will not write the δ explicitly to save space. The only remaining contribution from observations in (43) is in the couplings K_{aj} and the relevant partition function becomes

$$Z = \left\langle \exp\left[-\frac{1}{2\sigma_{\rm s}^2}\sum_{a=1}^{N^{\rm s}}\int_0^T \left(\sum_{j=1}^{N} K_{aj}x_j(t)\right)^2 dt\right]\right\rangle_{\boldsymbol{x}}, \tag{44}$$

where $\boldsymbol{x} \equiv \boldsymbol{x}^{\rm b} = \{x_i\}_{i=1}^N$. The average is the marginalization over the hidden dynamics with the weight given by the second term in (43). This weight corresponds to the dynamics of the isolated hidden network, viz.

$$\partial_t x_i(t) = -\lambda x_i(t) + \sum_j J_{ij} x_j(t) + \xi_i(t), \qquad (45)$$

with white noise $\langle \xi_i(t)\xi_j(t')\rangle = \sigma_{\rm b}^2 \delta_{ij}\delta(t-t')$ as before.

A. Asymmetric hidden-hidden couplings

1. Annealed average

The average of $\ln Z$ over the quenched couplings $\boldsymbol{J}$ and $\boldsymbol{K}^{\rm sb}$ would conventionally be performed by the replica method. However, for fully connected systems with quadratic interaction terms such as the one here, similar calculations [9, 10] indicate that the annealed calculation, which replaces $\langle \ln Z\rangle$ by $\ln\langle Z\rangle$, will give the exact result. We therefore calculate

$$f = -{\rm lim}_{N\to\infty} N^{-1} \ln \langle Z(\boldsymbol{J}, \boldsymbol{K}^{\rm sb})\rangle_{\boldsymbol{J},\boldsymbol{K}^{\rm sb}}. \qquad (46)$$

We shall again assume $\boldsymbol{J}$ and $\boldsymbol{K}^{\rm sb}$ to have Gaussian-distributed elements with zero mean, but now consider the case where $\boldsymbol{J}$ is *asymmetric*, i.e. $\langle J_{ij}J_{ji}\rangle = 0$, thus breaking detailed balance. (We comment on the case of general symmetry of $\boldsymbol{J}$ below.) For the calculation we introduce

$$\chi_i(t) = \sum_{j=1}^{N} J_{ij} x_j(t) + \xi_i(t), \qquad (47)$$

$$\phi_a(t) = \sum_{j=1}^{N} K_{aj} x_j(t). \qquad (48)$$

With regards to the quenched disorder average these are two Gaussian fields, which become independent when conditioned on the x_i. Defining as before amplitudes j and k so that $\langle J_{ij}^2\rangle = j^2/N$ and $\langle K_{aj}^2\rangle = k^2/N$, we have

$$\langle \chi_i(t)\chi_i(t')\rangle_{\boldsymbol{J}} = \sigma_{\rm b}^2\delta(t-t') + j^2 C(t,t'), \qquad (49)$$

$$\langle \phi_a(t)\phi_b(t')\rangle_{\boldsymbol{J}} = k^2 C(t,t')\delta_{ab}, \qquad (50)$$

where we have introduced the order parameter

$$C(t,t') \doteq \frac{1}{N}\sum_{j=1}^{N} x_j(t)x_j(t'). \qquad (51)$$

Hence, we will calculate

$$Z_{\rm ann} = \left\langle \exp\left[\frac{1}{2\sigma_{\rm s}^2}\sum_{a=1}^{N^{\rm s}}\int_0^T \phi_a^2(t)dt\right]\right\rangle_{\boldsymbol{\phi},\boldsymbol{x}}, \qquad (52)$$

where now the process has an effective prior dynamics given by

$$\partial_t x_i(t) = -\lambda x_i(t) + \chi_i(t). \qquad (53)$$

Here $\boldsymbol{\phi} = \{\phi_a\}_{a=1}^{N^{\rm s}}$ and $\boldsymbol{\chi} = \{\chi_i\}_{i=1}^{N}$ are still coupled to $\boldsymbol{x}$ because of the covariances $C(t,t')$.

2. Decoupling the degrees of freedom

To decouple the degrees of freedom we constrain the value of the order parameter function $C(t,t')$. Formally this means writing $Z_{\rm ann}$ as an integral of $\exp(N\Xi[C])$ over all possible values of $C(t,t')$, where

$$\Xi[C] = \frac{1}{N}\ln\left\langle \exp\left\{-\frac{1}{2\sigma_{\rm s}^2}\sum_{a=1}^{N^{\rm s}}\int_0^T \phi_a^2(t)dt\right\}\prod_{t,t'}\delta\left(NC(t,t') - \sum_{i=1}^{N} x_i(t)x_i(t')\right)\right\rangle_{\boldsymbol{\phi},\boldsymbol{x}} \equiv \Xi_1[C] + \Xi_2[C] (54)$$

with

$$\Xi_1[C] = \frac{1}{N}\ln\left\langle \prod_{t,t'}\delta\left(NC(t,t') - \sum_{i=1}^{N} x_i(t)x_i(t')\right)\right\rangle_{\boldsymbol{x}}, \qquad (55)$$

$$\Xi_2[C] = \frac{N^{\rm s}}{N}\ln\left\langle \exp\left\{-\frac{1}{2\sigma_{\rm s}^2}\int_0^T \phi^2(t)dt\right\}\right\rangle_{\phi}. \qquad (56)$$

In equation (56) the decoupling has allowed us to drop the index a and consider a representative ϕ.

The first equation (55) is dealt with by introducing an order parameter to $C(t,t')$. This means that for $N \to \infty$, we replace the "hard" δ constraints by an extra Gaussian term yielding a new effective measure over independent $x_i(t)$, which is adjusted such that $\langle x_i(t)x_i(t')\rangle_e = C(t,t')$ (here e denotes the effective "posterior" average). Equivalently one can write δ-function constraints in Fourier representation and evaluate $\exp(N\Xi[C])$ using a saddle point method. Either way one has

$$\Xi_1 = \frac{1}{2}\int_0^T dt \int_0^T dt'\, D(t,t')C(t,t') + \ln\left\langle \exp\left\{-\frac{1}{2}\int_0^T dt \int_0^T dt'\, D(t,t')x(t)x(t')\right\}\right\rangle_x . \tag{57}$$

This path integral is now also for a single representative coordinate x. Extremization over $D(t,t')$ is understood in (57), and similarly one needs to extremize over $C(t,t')$ in evaluating the resulting Z_{ann}.

3. Evaluating the order parameters

As before we focus on the steady state of the system for $t \to \infty$. The order parameters then depend on time differences only and the path integrals can be evaluated using Fourier or Laplace modes $\tilde{x}(z)$. These decouple into independent Gaussians and we get from (49), (50) and (53) that

$$\tilde{C}_0(z) \doteq \left\langle |\tilde{x}(z)|^2\right\rangle_{\tilde{x}} = \frac{j^2\tilde{C}(z) + \sigma_{\mathrm{b}}^2}{-z^2+\lambda^2}, \tag{58}$$

$$\left\langle |\tilde{\phi}(z)|^2\right\rangle_{\tilde{\phi}} = k^2\tilde{C}(z). \tag{59}$$

$\tilde{C}_0(z)$ is the covariance of the prior effective dynamics while $\tilde{C}(z)$ relates to the posterior dynamics that includes the conditioning on observations. Carrying out the prior average, the second term in (57) becomes

$$\ln\left\langle \exp\left\{-\frac{1}{2}\int_0^T dt \int_0^T dt'\, D(t,t')x(t)x(t')\right\}\right\rangle_x = -\frac{1}{2}\int dz \ln\left(1+\tilde{C}_0(z)\tilde{D}(z)\right). \tag{60}$$

In a similar way, we have for Ξ_2, from (56)

$$\Xi_2 = \left\langle \exp\left\{-\frac{1}{2\sigma_{\mathrm{s}}^2}\int_0^T \phi^2(t)dt\right\}\right\rangle_\phi = -\frac{1}{2}\int dz \ln\left(1+\frac{k^2}{\sigma_{\mathrm{s}}^2}\tilde{C}(z)\right). \tag{61}$$

Hence, finally, by substituting (60) into (57) and from (61) we get

$$\Xi = \frac{1}{2}\int dz\left[\tilde{D}(z)\tilde{C}(z) - \ln\left(1+\tilde{C}_0(z)\tilde{D}(z)\right)\right] - \frac{\alpha}{2}\int dz \ln\left(1+\frac{k^2}{\sigma_{\mathrm{s}}^2}\tilde{C}(z)\right), \tag{62}$$

where $\alpha = N^{\mathrm{s}}/N$ as before. The order parameter equations $\partial\Xi/\partial\tilde{C}(z) = 0$ and $\partial\Xi/\partial\tilde{D}(z) = 0$ result as

$$\tilde{D}(z) = \frac{\alpha k^2}{\sigma_{\mathrm{s}}^2 + k^2\tilde{C}(z)} + \frac{\tilde{D}(z)}{1+\tilde{C}_0(z)\tilde{D}(z)}\frac{j^2}{-z^2+\lambda^2}, \tag{63}$$

$$\frac{\tilde{C}(z)}{\tilde{C}_0(z)} + \tilde{D}(z)\tilde{C}(z) = 1. \tag{64}$$

Combining these and using (58) gives a closed algebraic equation for $\tilde{C}(z)$

$$z^2 = \left[-\frac{\sigma_{\rm b}^2}{\tilde{C}} + \frac{\alpha\frac{k^2\sigma_{\rm b}^2}{\sigma_{\rm s}^2}}{1+\frac{k^2}{\sigma_{\rm s}^2}\tilde{C}}\right]\left(1+\frac{j^2}{\sigma_{\rm b}^2}\tilde{C}\right)^2 + j^2\left(1+\frac{j^2}{\sigma_{\rm b}^2}\tilde{C}\right) + \lambda^2 \tag{65}$$

with the abbreviation $\tilde{C}(z) = \tilde{C}$. This is the analog of (39) for the non-equilibrium case of asymmetric couplings $\boldsymbol{J}$, and our final result for this section.

B. Generalization to arbitrary interaction symmetry

The above approach based on dynamical functionals can be extended to the case of hidden-hidden interactions of arbitrary degree of symmetry, defined by $\langle J_{ij}J_{ji}\rangle = \eta j^2/N$. Asymmetric couplings (section IV A) correspond to $\eta = 0$ while $\eta = 1$ gives symmetric $\boldsymbol{J}$ (section III C). We do not detail the calculations for the case of general η here. The main change is that the nonzero correlation $\langle J_{ij}J_{ji}\rangle$ causes the effective prior dynamics to contain a response term where each $x_i(t)$ reacts to its values $x_i(t')$ in the past (see e.g. [28]).

The final result is again a closed algebraic equation for $\tilde{C}(z)$

$$z^2 = \left[-\frac{\sigma_{\rm b}^2}{\tilde{C}} + \frac{\alpha\frac{k^2\sigma_{\rm b}^2}{\sigma_{\rm s}^2}}{1+\frac{k^2}{\sigma_{\rm s}^2}\tilde{C}} + \frac{j^2}{1+\frac{j^2}{\sigma_{\rm b}^2}\tilde{C}} + \frac{\lambda^2}{\left(1+(1+\eta)\frac{j^2}{\sigma_{\rm b}^2}\tilde{C}\right)^2}\right]\left(1+(1-\eta)\frac{j^2}{\sigma_{\rm b}^2}\tilde{C}\right)^2. \tag{66}$$

For $\eta = 1$ and $\eta = 0$ this leads back to (39) and (65), respectively, as it should.

The result (66) characterizes the average case posterior variance – and hence inference error – for our partially observed network dynamics. Remarkably, it does so across an entire range of non-equilibrium settings parameterized by η. Equation (66) is derived within the annealed approximation but as discussed above this should be exact here so that our result acts as a baseline for the assessment of other approximations. One such approximation, the Extended Plefka Expansion [17, 18], can be shown to give exactly (66), demonstrating that this approximate scheme is also exact (in the large system limit studied here).

The dependence on various parameters, especially the level of symmetry η, of inference errors and posterior relaxation times as they result from (66) is sufficiently rich that we devote a separate paper to it [18]. It turns out that the behavior can be organized around critical regions in the parameter space of α, γ and p. There are two such regions. Generalizing from section III C 2, these are defined by $p \to 0$ for $0 \leq \alpha \leq 1$ for the first region, and for the second $\alpha \to 0$ and $\gamma \to \gamma_{\rm c} = 1/(1+\eta)$. One key finding is that across the entire range of eta from 0 to just below 1, i.e. the regime where interaction symmetry is broken, there are no qualitative changes in behavior. On the other hand, interesting crossovers then occur in the vicinity of $\eta = 1$, i.e. as interaction symmetry is approached. We refer the interested reader to [18] for further details.

V. DISCUSSION AND CONCLUSIONS

We have considered in this paper linear stochastic dynamics in a large network of continuous degrees of freedom, where given a time trajectory of the nodes in some observable part of the network the task is to infer the trajectory of the hidden nodes. By varying interaction symmetry we were able to study both equilibrium and non-equilibrium settings, thus creating a paradigmatic example of inference from temporal data. Given the increasing availability of large scale temporal data sets such problems are becoming prevalent in e.g. biology, where interpretation of data and prediction are highly challenging when observations only partially characterize a system.

Our main goal was to explore the average case inference error. To ensure analytical tractability we focused on stationary dynamics on large networks. More precisely it is the variance of hidden state estimates that becomes stationary in time; mean predictions for the hidden states have to depend on time in our dynamical context. The large network assumption is realistic in many situations, e.g. for metabolic or neural networks that can be composed of thousands of interacting elements (chemical species, neurons etc).

We deployed two different methods of analysis. For the first, the starting point (section II) is a Lyapunov-type equation for the posterior variance matrix $\boldsymbol{C}$, where an effective drift matrix $\boldsymbol{K}^{\mathrm{bb|s}}$ captures the effect of the observations. In section III we derived average case performance results by appeal to RMT. This is possible because the Lyapunov equation can be solved in the case of self-interacting hidden variables (section III B) or more generally, symmetric hidden-hidden couplings (section III C), corresponding to equilibrium dynamics. With suitable assumptions of couplings being Gaussian and long-range, and taking the thermodynamic limit of large networks, we then used free probability methods to derive the Green's functions and then the spectra of $\boldsymbol{C}$ and $\boldsymbol{K}^{\mathrm{bb|s}}$, which are closely linked.

For the opposite case of *asymmetric* hidden-hidden couplings, where the dynamics is non-equilibrium, we presented in section IV A a calculation based on dynamical functionals. This leads to an algebraic equation for the stationary posterior variance (in Laplace space). We sketched how the approach can be extended to the analysis of non-equilibrium stationary regimes arising from couplings of *generic* symmetry (section IV B).

We focused on the inference error as an average macroscopic quantity. For large networks this is independent of the specific realization of the microscopic (Gaussian) interactions, but does depend on structural parameters such as overall interaction strengths as well as α, the ratio between the number of hidden and observed nodes. Predictions on such structural dependences of macroscopic properties should be testable in practice and may give information on microscopic features such as the degree of interaction symmetry. The emerging picture, consisting of algebraic expressions that link inference errors and parameters, suggests possible connections to experiment design, as we discuss further in [18]. There we quantify these dependences in terms of scaling laws; of particular importance is the dependence on α, as it tells us how many observed nodes are needed to attain a specified precision for the hidden node inference.

The RMT approach to our problem has the benefit that it gives information on spectral densities - our main focus here - including the spectrum of relaxation times in the posterior dynamics. This then allowed us to compare different definitions of a characteristic posterior relaxation time, such as slowest mode and average time (section III B 1). The spectral shapes proved revealing: when there are few observations (small α), the spectrum can be split into two parts corresponding to constrained and unconstrained directions (section III C), but this distinction is then lost as hidden nodes interact more strongly.

One open question for the inference setting we have considered is to answer the question of the spectral density of relaxation times and its support in the *non-equilibrium* case $\eta < 1$. For example, does our result (66) for generic η still have a free probability interpretation? Generalizing the derivation of the equilibrium ($\eta = 1$) result (39) to $\eta < 1$ appears non-trivial. One might consider assuming that the equilibrium relation $\widetilde{C}(z) = -\sigma_{\mathrm{b}}^2 \widetilde{G}(z^2)$ continues to hold and analyze the spectrum corresponding to the Green's function $\widetilde{G}(z)$.

There are a number of avenues for further work, as the setting we have begun to study is still rather new in the statistical physics community [12–14, 16]. An obvious extension would be to sparse networks, where for static analyses statistical mechanics has been successfully deployed [30, 40]. The sparse case would be worth developing because of its relevance to applications such as gene expression networks [1]. As a starting point one could investigate progressive degrees of dilution. Consider for example an average degree of connectivity c, which corresponds to the J_{ij} being drawn as Gaussian random variables with probability c/N, and zero with probability $1 - c/N$; one would set then the amplitude of the nonzero J_{ij} such that $\langle J_{ij}^2 \rangle = j^2/c$ in order to obtain a sensible thermodynamic limit. In this paper, we have effectively considered $c = N$, but from previous studies [41, 42] it is clear that one can take $c \ll N$ (in fact as low as $c \sim \ln N$) without changing the results derived in this paper. This already goes a long way towards making our work applicable to real networks. The strong dilution regime, where $c = O(1)$, would require a separate analysis that goes beyond the scope of the present paper. Cavity and population dynamics methods developed for sparse network spectra (e.g. [30, 40]) would probably need to be deployed there.

A second important consideration for applications to real networks is their finite size N. We have begun to investigate the resulting finite size effects numerically. Encouragingly, we find [17] that even for moderate network sizes ($N \approx 100$) there is good agreement between numerically exact calculations of the inference error on the one hand and our large-N theory on the other.

Variants of the dynamics could also be considered, for example, by adding non-linearities that can be treated perturbatively. One could also extend to measurements of the trajectory of the observable nodes that would be available at a regular or irregular grid of time points only rather than along the entire time interval considered; or to measurements which are noisy rather than just incomplete as in our case [43, 44].

Finally, we have concentrated on the *forward* problem of predicting hidden states given known interactions. This is relevant also for inverse problems such as learning the couplings from dynamical data, where typically a forward problem has to be solved at every iteration (e.g. in Expectation Propagation [45]). Learning which couplings are non-zero is effectively a network reconstruction problem, with potential applications to signaling pathways and gene expression data. In either case, modelling data as explicitly dynamical rather than as uncorrelated snapshots is expected to lead to performance improvements in inference and learning. Such algorithmic advances have already been achieved by adapting equilibrium statistical physics tools [1, 46] to learning of regulatory networks from steady state data.

ACKNOWLEDGEMENTS

This work was supported by the Marie Curie Training Network NETADIS (FP7, grant 290038). We are grateful to Pierpaolo Vivo, Ludovica Bachschmid-Romano and Reimer Kühn for helpful discussions.

Appendix A: Kalman filter and smoother

In this appendix we derive the results (7)-(11) in the main text, using a reduction of our inference problem to a linear Gaussian state space model, to which standard Kalman filter techniques [21] can then be applied.

Let us consider a time discretized version of our dynamics (1) and (2), with elementary time step Δ,

$$\boldsymbol{x}^{\rm b}(t) - \boldsymbol{x}^{\rm b}(t-\Delta) = \tag{A1}$$
$$\Delta\boldsymbol{K}^{\rm bs}\boldsymbol{x}^{\rm s}(t-\Delta) + \Delta\boldsymbol{K}^{\rm bb}\boldsymbol{x}^{\rm b}(t-\Delta) + \Delta\bar{\boldsymbol{\xi}}^{\rm b}(t-\Delta),$$

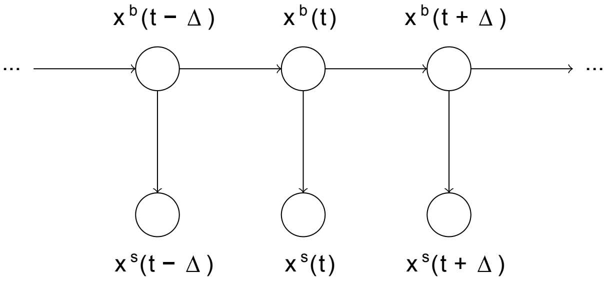

Figure 5. Illustration of a linear-Gaussian state space model.

$$\boldsymbol{x}^{\rm s}(t) - \boldsymbol{x}^{\rm s}(t-\Delta) = \tag{A2}$$
$$\Delta\boldsymbol{K}^{\rm ss}\boldsymbol{x}^{\rm s}(t-\Delta) + \Delta\boldsymbol{K}^{\rm sb}\boldsymbol{x}^{\rm b}(t-\Delta) + \Delta\bar{\boldsymbol{\xi}}^{\rm s}(t-\Delta),$$

where the white noises $\bar{\boldsymbol{\xi}}^{\rm s}$ and $\bar{\boldsymbol{\xi}}^{\rm b}$ are averages of the continuous time noise over the time interval Δ with covariance

$$\langle \bar{\boldsymbol{\xi}}^{\rm s}(t)\bar{\boldsymbol{\xi}}^{{\rm s}\,T}(t')\rangle = \Delta^{-1}\boldsymbol{\Sigma}^{\rm ss}\delta_{tt'} \tag{A3}$$

and similarly for $\bar{\boldsymbol{\xi}}^{\rm b}$. The above dynamics is Markovian, with transition probabilities

$$P(\boldsymbol{x}^{\rm b}(t)|\boldsymbol{x}^{\rm b}(t-\Delta),\boldsymbol{x}^{\rm s}(t-\Delta)) = \tag{A4}$$
$$\mathcal{N}(\boldsymbol{x}^{\rm b}(t)|(\mathbb{1}+\Delta\boldsymbol{K}^{\rm bb})\boldsymbol{x}^{\rm b}(t-\Delta) + \Delta\boldsymbol{K}^{\rm sb}\boldsymbol{x}^{\rm s}(t-\Delta), \Delta\boldsymbol{\Sigma}^{\rm bb}),$$
$$P(\boldsymbol{x}^{\rm s}(t+\Delta)|\boldsymbol{x}^{\rm b}(t),\boldsymbol{x}^{\rm s}(t)) = \tag{A5}$$
$$\mathcal{N}(\boldsymbol{x}^{\rm s}(t+\Delta)|(\mathbb{1}+\Delta\boldsymbol{K}^{\rm ss})\boldsymbol{x}^{\rm s}(t) + \Delta\boldsymbol{K}^{\rm sb}\boldsymbol{x}^{\rm b}(t), \Delta\boldsymbol{\Sigma}^{\rm ss})$$

and we are interested in the posterior probability $P(\boldsymbol{X}^{\rm b}|\boldsymbol{X}^{\rm s})$ of a time trajectory $\boldsymbol{X}^{\rm b}$ of hidden variables given a trajectory $\boldsymbol{X}^{\rm s}$ of observed variables.

To bring this inference problem into a standard form, we exploit the fact that the joint distribution $P(\boldsymbol{X}^{\rm b},\boldsymbol{X}^{\rm s})$ is Gaussian, and hence so is the posterior $P(\boldsymbol{X}^{\rm b}|\boldsymbol{X}^{\rm s})$. From general properties of Gaussian conditioning, the second order statistics of the posterior are then *independent* of the specific observed trajectory $\boldsymbol{X}^{\rm s}$. We can therefore choose the most convenient $\boldsymbol{X}^{\rm s}$ to find the second order statistics, which is the identically zero trajectory. The second order statistics we find then determine the inference error, which is the trace of the covariance matrix of $\boldsymbol{x}^{\rm b}(t)$. For zero observations, the transition probabilities (A5), (A6) simplify to

$$P(\boldsymbol{x}^{\rm b}(t)|\boldsymbol{x}^{\rm b}(t-\Delta)) = \tag{A6}$$
$$\mathcal{N}(\boldsymbol{x}^{\rm b}(t)|(\mathbb{1}+\Delta\boldsymbol{K}^{\rm bb})\boldsymbol{x}^{\rm b}(t-\Delta), \Delta\boldsymbol{\Sigma}^{\rm bb}),$$
$$P(\boldsymbol{x}^{\rm s}(t+\Delta)=0|\boldsymbol{x}^{\rm b}(t)) = \tag{A7}$$
$$\mathcal{N}(\boldsymbol{x}^{\rm s}(t+\Delta)=0|\Delta\boldsymbol{K}^{\rm sb}\boldsymbol{x}^{\rm b}(t), \Delta\boldsymbol{\Sigma}^{\rm ss}).$$

These now have the conventional form of a linear-Gaussian state space model [21], where (A6) specifies the dynamics of the hidden state $\boldsymbol{x}^{\rm b}$ while (A7) defines the "emission probability" at time t, with $\boldsymbol{x}^{\mathrm{s}}(t+\Delta)$ taking the role of the emitted signal or observation. To conform with standard notation, we will shift the time index on $\boldsymbol{x}^{\mathrm{s}}(t+\Delta)$ to $\boldsymbol{x}^{\mathrm{s}}(t)$ for the rest of this discussion; see figure 5. Note that while we are dealing with real-valued states and emissions here, the probabilistic "graphical model" [21] of figure 5 could also capture cases, e.g. *Hidden Markov Models* (HMMs) where the hidden states are discrete.

The chain structure of figure 5 means that posterior probabilities can be computed efficiently by message passing methods, denoted *Forward-Backward* algorithm in the context of HMMs [47] and *Kalman Filter* [24] [48] here.

The forward propagation computes forward messages $\hat{\alpha}_t$ that absorb the effect of previous observations (the past), while the backward propagation accounts for observations from the future. Formally the messages can be defined as

$$\hat{\alpha}(\boldsymbol{x}^{\mathrm{b}}(t)) = P(\boldsymbol{x}^{\mathrm{b}}(t)|\boldsymbol{x}^{\mathrm{s}}(\Delta), ..., \boldsymbol{x}^{\mathrm{s}}(t)) = \hat{\alpha}_t, \quad \text{(A8)}$$

$$\begin{aligned}\hat{\beta}(\boldsymbol{x}^{\mathrm{b}}(t)) &= \frac{P(\boldsymbol{x}^{\mathrm{s}}(t+\Delta), ..., \boldsymbol{x}^{\mathrm{s}}(T)|\boldsymbol{x}^{\mathrm{b}}(t))}{P(\boldsymbol{x}^{\mathrm{s}}(t+\Delta), ..., \boldsymbol{x}^{\mathrm{s}}(T)|\boldsymbol{x}^{\mathrm{s}}(\Delta), ..., \boldsymbol{x}^{\mathrm{s}}(t))} \\ &= \hat{\beta}_t. \qquad \text{(A9)}\end{aligned}$$

Once $\hat{\alpha}_t$ and $\hat{\beta}_t$ have been computed, the desired posterior probability is simply

$$\gamma_t = \hat{\alpha}_t\hat{\beta}_t = \frac{P(\boldsymbol{x}^{\mathrm{b}}(t), \boldsymbol{X}^{\mathrm{s}})}{P(\boldsymbol{X}^{\mathrm{s}})} = P(\boldsymbol{x}^{\mathrm{b}}(t)|\boldsymbol{X}^{\mathrm{s}}). \quad \text{(A10)}$$

The forward propagation for continuous variables reads

$$\begin{aligned}\hat{\alpha}_t \propto\ & P(\boldsymbol{x}^{\mathrm{s}}(t)|\boldsymbol{x}^{\mathrm{b}}(t)) \cdot \qquad \text{(A11)} \\ & \int d\boldsymbol{x}^{\mathrm{b}}(t-\Delta) P(\boldsymbol{x}^{\mathrm{b}}(t)|\boldsymbol{x}^{\mathrm{b}}(t-\Delta))\hat{\alpha}_{t-\Delta}.\end{aligned}$$

In our case, all distributions involved are Gaussian and we denote in particular

$$\hat{\alpha}_t = \mathcal{N}(\boldsymbol{x}^{\mathrm{b}}(t)|0, \boldsymbol{C}_{\mathrm{f}}(t)). \quad \text{(A12)}$$

$\boldsymbol{C}_{\mathrm{f}}(t) = \langle \boldsymbol{x}^{\mathrm{b}}(t)\boldsymbol{x}^{\mathrm{b}}(t)^T\rangle$ is the equal time forward (or "filtered") posterior covariance. By substituting (A6), (A7) and (A12) into (A11) and identifying the quadratic terms in $\boldsymbol{x}^{\mathrm{b}}(t)$ in the exponents one obtains the recursive Kalman filter expression for $\boldsymbol{C}_{\mathrm{f}}^{-1}(t)$

$$\begin{aligned}\boldsymbol{C}_{\mathrm{f}}^{-1}(t) = \big[&\big(\mathbb{1} + \Delta\boldsymbol{K}^{\mathrm{bb}}\big)\boldsymbol{C}_{\mathrm{f}}(t-\Delta)\big(\mathbb{1} + \Delta\boldsymbol{K}^{\mathrm{bb}}\big)^T \\ &+ \Delta\boldsymbol{\Sigma}^{\mathrm{bb}}\big]^{-1} + \Delta\boldsymbol{W}, \qquad \text{(A13)}\end{aligned}$$

where $\boldsymbol{W} = \boldsymbol{K}^{\mathrm{sb}\,T}(\boldsymbol{\Sigma}^{\mathrm{ss}})^{-1}\boldsymbol{K}^{\mathrm{sb}}$ is the *feedback* matrix. Equation (A13) is a discrete time Riccati (i.e. second order matrix) recursion. We are interested in the continuous time limit $\Delta \to 0$, where it becomes

$$\begin{aligned}&\frac{d}{dt}\boldsymbol{C}_{\mathrm{f}}^{-1}(t) = \qquad \text{(A14)} \\ &\boldsymbol{C}_{\mathrm{f}}^{-1}(t)\boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{C}_{\mathrm{f}}^{-1}(t) + \boldsymbol{C}_{\mathrm{f}}^{-1}(t)\boldsymbol{K}^{\mathrm{bb}} + \boldsymbol{K}^{\mathrm{bb}\,T}\boldsymbol{C}_{\mathrm{f}}^{-1}(t) + \boldsymbol{W}.\end{aligned}$$

The backward propagation incorporates in the algorithm the observations from all later time steps

$$\begin{aligned}\hat{\beta}_t \propto \int & d\boldsymbol{x}^{\mathrm{b}}(t+\Delta)\hat{\beta}_{t+\Delta}P(\boldsymbol{x}^{\mathrm{s}}(t+\Delta)|\boldsymbol{x}^{\mathrm{b}}(t+\Delta)) \\ & \cdot P(\boldsymbol{x}^{\mathrm{b}}(t+\Delta)|\boldsymbol{x}^{\mathrm{b}}(t)) \qquad \text{(A15)}\end{aligned}$$

and we set

$$\hat{\beta}_t \propto \mathcal{N}(\boldsymbol{x}^{\mathrm{b}}(t)|0, \boldsymbol{C}_{\mathrm{b}}(t)) \quad \text{(A16)}$$

with $\boldsymbol{C}_{\mathrm{b}}(t) = \langle \boldsymbol{x}^{\mathrm{b}}(t)\boldsymbol{x}^{\mathrm{b}}(t)^T\rangle$ defined as the equal time posterior variance in the backward propagation. Inserting (A16) into (A15) one finds the backward recursion for $\boldsymbol{C}_{\mathrm{b}}^{-1}(t)$

$$\begin{aligned}\boldsymbol{C}_{\mathrm{b}}^{-1}(t) = &\big(\mathbb{1} + \Delta\boldsymbol{K}^{\mathrm{bb}}\big)^T(\Delta\boldsymbol{\Sigma}^{\mathrm{bb}})^{-1} \cdot \qquad \text{(A17)} \\ &\big[\mathbb{1} - \big(\mathbb{1} + \Delta\boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{C}_{\mathrm{b}}^{-1}(t+\Delta) + \Delta^2\boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{W}\big)^{-1}\big] \\ &\cdot(\mathbb{1} + \Delta\boldsymbol{K}^{\mathrm{bb}}).\end{aligned}$$

Taking $\Delta \to 0$, which requires keeping all terms up to $O(\Delta)$ on the r.h.s., gives the continuous time limit

$$\begin{aligned}&\frac{d}{dt}\boldsymbol{C}_{\mathrm{b}}^{-1}(t) = \qquad \text{(A18)} \\ &-\boldsymbol{K}^{\mathrm{bb}\,T}\boldsymbol{C}_{\mathrm{b}}^{-1}(t) - \boldsymbol{C}_{\mathrm{b}}^{-1}(t)\boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{W} + \boldsymbol{C}_{\mathrm{b}}^{-1}(t)\boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{C}_{\mathrm{b}}^{-1}(t).\end{aligned}$$

The changes of sign compared to (A14) come from the backward direction.

Finally the posterior γ_t also has a Gaussian form,

$$\gamma_t = \mathcal{N}(\boldsymbol{x}^{\mathrm{b}}(t)|0, \boldsymbol{C}^{\mathrm{bb|s}}(t)). \quad \text{(A19)}$$

We drop the superscripts on $\boldsymbol{C}^{\mathrm{bb|s}}(t)$ as in the main text and write this overall ("smoothed") covariance as $\boldsymbol{C}(t)$. From (A10) one has $\boldsymbol{C}^{-1}(t) = \boldsymbol{C}_{\mathrm{f}}^{-1}(t) + \boldsymbol{C}_{\mathrm{b}}^{-1}(t)$, so from the sum of (A14) and (A18)

$$\begin{aligned}&\frac{d}{dt}\boldsymbol{C}^{-1}(t) = \qquad \text{(A20)} \\ &\boldsymbol{C}^{-1}(t)\boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{C}^{-1}(t) + \boldsymbol{C}^{-1}(t)\boldsymbol{K}^{\mathrm{bb|s}} + \boldsymbol{K}^{\mathrm{bb|s}\,T}\boldsymbol{C}^{-1}(t),\end{aligned}$$

where we have set

$$\boldsymbol{K}^{\mathrm{bb|s}} = \boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{C}_{\mathrm{b}}^{-1} \quad \text{(A21)}$$

and we have taken $\boldsymbol{C}_{\mathrm{b}}^{-1}$ as the stationary limit of $\boldsymbol{C}_{\mathrm{b}}^{-1}(t)$.

To interpret $\boldsymbol{K}^{\mathrm{bb|s}}$ one can look at $P(\boldsymbol{x}^{\mathrm{b}}(t+\Delta),\boldsymbol{x}^{\mathrm{b}}(t)|\boldsymbol{X}^{\mathrm{s}})$, given by the integrand of (A15). Conditioning on $\boldsymbol{x}^{\mathrm{b}}(t)$ and using (A6), (A7) and (A16) one finds easily that the mean of $\boldsymbol{x}^{\mathrm{b}}(t+\Delta)$ conditioned on $\boldsymbol{x}^{\mathrm{b}}(t)$ is

$$\big(\mathbb{1}+\Delta\boldsymbol{K}^{\mathrm{bb|s}}(t)+O(\Delta^2)\big)\boldsymbol{x}^{\mathrm{b}}(t). \tag{A22}$$

Hence $\boldsymbol{K}^{\mathrm{bb|s}}(t)$ has the meaning of a posterior drift, i.e. it determines the time evolution for the posterior dynamics.

Focusing on the stationary state now, we can drop all dependences on t. From (A20), the posterior covariance $\boldsymbol{C}$ then satisfies the Lyapunov equation (7)

$$\boldsymbol{K}^{\mathrm{bb|s}}\boldsymbol{C}+\boldsymbol{C}\boldsymbol{K}^{\mathrm{bb|s}\,T}+\boldsymbol{\Sigma}^{\mathrm{bb}}=0 \tag{A23}$$

with the stationary posterior drift $\boldsymbol{K}^{\mathrm{bb|s}}$ given by

$$\boldsymbol{K}^{\mathrm{bb|s}}=\boldsymbol{K}^{\mathrm{bb}}-\boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{C}_{\mathrm{b}}^{-1} \tag{A24}$$

and the stationary backward covariance satisfying, from (A18)

$$\boldsymbol{C}_{\mathrm{b}}^{-1}\boldsymbol{\Sigma}^{\mathrm{bb}}\boldsymbol{C}_{\mathrm{b}}^{-1}-\boldsymbol{K}^{\mathrm{bb}\,T}\boldsymbol{C}_{\mathrm{b}}^{-1}-\boldsymbol{C}_{\mathrm{b}}^{-1}\boldsymbol{K}^{\mathrm{bb}}=\boldsymbol{W}. \tag{A25}$$

Apart from the relabelling of $\boldsymbol{C}_{\mathrm{b}}^{-1}$ as $\boldsymbol{A}$, we have therefore derived (7), (9) and (10) in the main text. Note that $\boldsymbol{C}_{\mathrm{b}}^{-1}$ is symmetric by definition; it is also positive semi-definite. As it enters the effective drift with a minus sign, we see that the presence of observations drives the hidden dynamics back towards its mean (zero) more quickly.

To find the evolution of the two-time posterior variance $\boldsymbol{C}(t,t')$, we first look at the case $\boldsymbol{C}(t'+\Delta,t')$ of adjacent time steps. Here (A22) gives directly

$$\boldsymbol{C}(t'+\Delta,t')=\big(\mathbb{1}+\Delta\boldsymbol{K}^{\mathrm{bb|s}}(t')+O(\Delta^2)\big)\boldsymbol{C}(t',t'). \tag{A26}$$

This easily generalizes to the correlations τ steps apart as

$$\boldsymbol{C}(t'+\tau\Delta,t')=\big(\mathbb{1}+\Delta\boldsymbol{K}^{\mathrm{bb|s}}+O(\Delta^2)\big)^{\tau}\boldsymbol{C}, \tag{A27}$$

where we have directly written the stationary version. Setting $t=t'+\tau\Delta$ and taking $\Delta\to 0$ then gives equation (11) in the main text, i.e.

$$\boldsymbol{C}(t-t')=e^{\boldsymbol{K}^{\mathrm{bb|s}}(t-t')}\boldsymbol{C}. \tag{A28}$$

Appendix B: Variational method

As is often the case, the fixed point of a recursion (such as the Forward-Backward algorithm) can also be retrieved variationally, i.e. as the solution of a constrained optimization problem. We show this connection in this appendix.

Let us start from $P(\boldsymbol{X}^{\mathrm{b}},\boldsymbol{X}^{\mathrm{s}})$, the joint probability of subnetwork and bulk trajectories obeying (1) and (2), and denote $Q(\boldsymbol{X}^{\mathrm{b}})$ a variational approximation to the posterior $P(\boldsymbol{X}^{\mathrm{b}}|\boldsymbol{X}^{\mathrm{s}})$ of the effective dynamics (8). As before if we are interested only in the posterior second order statistics, we can remove the means by assuming $\boldsymbol{x}^{\mathrm{s}}(t)=0\ \forall t$ and can then drop the δ in (8). One aim is to determine the effective drift $\boldsymbol{K}^{\mathrm{bb|s}}$ by variational methods. Note that parameterizing Q in terms of $\boldsymbol{K}^{\mathrm{bb|s}}$ gives us enough flexibility to retrieve the *exact* posterior because of the Gaussian nature of our problem.

We can write the joint trajectory probability and the variational posterior, directly in continuous time form, as

$$P(\boldsymbol{X}^{\mathrm{b}},\boldsymbol{X}^{\mathrm{s}})\propto \tag{B1}$$
$$\exp\left[-\frac{1}{2}\int_0^T dt\big(\boldsymbol{\xi}^{\mathrm{b}\,T}(t)\boldsymbol{\Sigma}^{\mathrm{bb}\,-1}\boldsymbol{\xi}^{\mathrm{b}}(t)+\boldsymbol{\xi}^{\mathrm{s}\,T}(t)\boldsymbol{\Sigma}^{\mathrm{ss}\,-1}\boldsymbol{\xi}^{\mathrm{s}}(t)\big)\right]$$

$$Q(\boldsymbol{X}^{\mathrm{b}})\propto\exp\left[-\frac{1}{2}\int_0^T dt\,\boldsymbol{\xi}^{\mathrm{b}\,T}(t)\boldsymbol{\Sigma}^{\mathrm{bb}\,-1}\boldsymbol{\xi}^{\mathrm{b}}(t)\right], \tag{B2}$$

where the noises $\boldsymbol{\xi}^{\mathrm{b}}$ and $\boldsymbol{\xi}^{\mathrm{s}}$ should be expressed as a function of $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{x}^{\mathrm{s}}$ using respectively equations (1) and (2) for $P(\boldsymbol{X}^{\mathrm{b}},\boldsymbol{X}^{\mathrm{s}})$ and (8) for $Q(\boldsymbol{X}^{\mathrm{b}})$.
We find Q in the standard variational way by finding the stationary point of the Kullback-Leibler divergence [49] between P and Q

$$\mathrm{KL}(P||Q)=-\left\langle\log\frac{Q}{P}\right\rangle_Q=F, \tag{B3}$$

which is analogous to a thermodynamic free energy. Inserting (B1) and (B2) and simplifying gives

$$F=\int_0^T dt\,\frac{1}{2}\left\langle\boldsymbol{x}^{\mathrm{b}\,T}(t)(\boldsymbol{K}^{\mathrm{bb}}-\boldsymbol{K}^{\mathrm{bb|s}})^T\boldsymbol{\Sigma}^{\mathrm{bb}\,-1}(\boldsymbol{K}^{\mathrm{bb}}-\boldsymbol{K}^{\mathrm{bb|s}})\boldsymbol{x}^{\mathrm{b}}(t)\right\rangle_Q+\int_0^T dt\,\frac{1}{2}\left\langle\boldsymbol{x}^{\mathrm{b}\,T}(t)\boldsymbol{W}\boldsymbol{x}^{\mathrm{b}}(t)\right\rangle_Q \tag{B4}$$

with $\boldsymbol{W} \doteq (\boldsymbol{K}^{\mathrm{sb}})^T \boldsymbol{\Sigma}^{\mathrm{ss}\,-1} \boldsymbol{K}^{\mathrm{sb}}$ the feedback matrix as before. Here we have performed an integration by parts and assumed that $\boldsymbol{x}^{\mathrm{b}}$ vanishes at the boundaries of the time domain.

In the stationary limit, we can drop the time integrals, drop the resulting factor T and use the definition $\boldsymbol{C} = \langle \boldsymbol{x}^{\mathrm{b}} \boldsymbol{x}^{\mathrm{b}\,T} \rangle_Q$ to write

$$F = \frac{1}{2}\mathrm{Tr}\left[(\boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{K}^{\mathrm{bb|s}})^T \boldsymbol{\Sigma}^{\mathrm{bb}\,-1} (\boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{K}^{\mathrm{bb|s}}) \boldsymbol{C}\right] + \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{W}\boldsymbol{C}\right]. \tag{B5}$$

We now want to optimize over $\boldsymbol{K}^{\mathrm{bb|s}}$, bearing in mind that the stationary posterior variance $\boldsymbol{C}$ is linked to the effective drift by the Lyapunov equation

$$\boldsymbol{K}^{\mathrm{bb|s}} \boldsymbol{C} + \boldsymbol{C} \boldsymbol{K}^{\mathrm{bb|s}\,T} + \boldsymbol{\Sigma}^{\mathrm{bb}} = 0 \tag{B6}$$

(see (7) in the main text). Introducing a Lagrange multiplier matrix $\boldsymbol{A}/2$ to implement this constraint, we optimize

$$\mathcal{L}[\boldsymbol{C}, \boldsymbol{K}^{\mathrm{bb|s}}, \boldsymbol{A}] = F + \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{A}^T (\boldsymbol{K}^{\mathrm{bb|s}} \boldsymbol{C} + \boldsymbol{C} \boldsymbol{K}^{\mathrm{bb|s}\,T} + \boldsymbol{\Sigma}^{\mathrm{bb}})\right]. \tag{B7}$$

Optimization w.r.t. $\boldsymbol{K}^{\mathrm{bb|s}}$ gives

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{K}^{\mathrm{bb|s}}} = \boldsymbol{\Sigma}^{\mathrm{bb}\,-1} (\boldsymbol{K}^{\mathrm{bb|s}} - \boldsymbol{K}^{\mathrm{bb}}) \boldsymbol{C} + \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^T) \boldsymbol{C} = 0, \tag{B8}$$

from which one has the expression (9) for the posterior drift matrix

$$\boldsymbol{K}^{\mathrm{bb|s}} = \boldsymbol{K}^{\mathrm{bb}} - \frac{\boldsymbol{\Sigma}^{\mathrm{bb}}}{2}(\boldsymbol{A} + \boldsymbol{A}^T) = \boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{\Sigma}^{\mathrm{bb}} \boldsymbol{A}_{\mathrm{s}}, \tag{B9}$$

where we have denoted the symmetric part of $\boldsymbol{A}$ by $\boldsymbol{A}_{\mathrm{s}} = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^T)$. We will then write $\boldsymbol{A} = \boldsymbol{A}_{\mathrm{s}} + \boldsymbol{A}_{\mathrm{a}}$ with $\boldsymbol{A}_{\mathrm{a}} = \frac{1}{2}(\boldsymbol{A} - \boldsymbol{A}^T)$ the antisymmetric part. The second optimization condition reads

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{C}} = \frac{1}{2}(\boldsymbol{K}^{\mathrm{bb|s}} - \boldsymbol{K}^{\mathrm{bb}})^T \boldsymbol{\Sigma}^{\mathrm{bb}\,-1} (\boldsymbol{K}^{\mathrm{bb|s}} - \boldsymbol{K}^{\mathrm{bb}}) + \frac{1}{2}\boldsymbol{W} + \frac{1}{2}(\boldsymbol{A}\boldsymbol{K}^{\mathrm{bb|s}} + \boldsymbol{K}^{\mathrm{bb|s}\,T} \boldsymbol{A}) = 0. \tag{B10}$$

By substitution of (B9) into (B10) one obtains

$$\boldsymbol{A}_{\mathrm{s}} \boldsymbol{\Sigma}^{\mathrm{bb}} \boldsymbol{A}_{\mathrm{s}} - \boldsymbol{K}^{\mathrm{bb}\,T} \boldsymbol{A}_{\mathrm{s}} - \boldsymbol{A}_{\mathrm{s}} \boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{A}_{\mathrm{a}} \big(\boldsymbol{K}^{\mathrm{bb}} - \boldsymbol{\Sigma}^{\mathrm{bb}} \boldsymbol{A}_{\mathrm{s}}\big) - \big(\boldsymbol{K}^{\mathrm{bb}\,T} - \boldsymbol{\Sigma}^{\mathrm{bb}} \boldsymbol{A}_{\mathrm{s}}\big) \boldsymbol{A}_{\mathrm{a}} - \boldsymbol{W} = 0. \tag{B11}$$

The symmetric part of this determines $\boldsymbol{A}_{\mathrm{s}}$, which is all we need for (B9), as

$$\boldsymbol{A}_{\mathrm{s}} \boldsymbol{\Sigma}^{\mathrm{bb}} \boldsymbol{A}_{\mathrm{s}} - \boldsymbol{K}^{\mathrm{bb}\,T} \boldsymbol{A}_{\mathrm{s}} - \boldsymbol{A}_{\mathrm{s}} \boldsymbol{K}^{\mathrm{bb}} = \boldsymbol{W}. \tag{B12}$$

This is equation (10) in the main text – we dropped the subscript “s” there – and shows that the Lagrange multiplier $\boldsymbol{A}$ is identical to the (stationary) inverse backward covariance matrix, $\boldsymbol{C}_{\mathrm{b}}^{-1}$.

[1] A. Braunstein, A. Pagnani, M. Weigt, and R. Zecchina. Inference algorithms for gene networks: A statistical-mechanics analysis. *J. Stat. Mech.*, P12001, 2008.

[2] R. S. Tsay. *Analysis of Financial Time Series, 3rd edition.* Wiley, 2010.

[3] A. S. Cofiño, J. M. Gutiérrez, B. Jakubiak, and M. Melonek. Implementation of data mining techniques for meteorological applications. *Realizing Teracomputing*, W. Zwieflhofer and N. Kreitz (Eds.), World Scientific:215–140, 2013.

[4] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning.* Cambridge University Press, 2004.

[5] W. Kinzel and M. Opper. *Models of Neural Networks III*, chapter Statistical Mechanics of Generalization. Springer, 1996.

[6] H. Sompolinsky, N. Tishby, and H. S. Seung. Learning from examples in large neural networks. *Phys. Rev. Lett.*, 65(13):1683–1687, 1990.

[7] P. Sollich. Finite-size effects in learning and generalization in linear perceptrons. *J. Phys. A. Math. Gen.*, 27:7771–7784, 1994.

[8] Y. Le Cun, I. Kanter, and S. A. Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Phys. Rev. Lett.*, 66(18):2396–2399, 1991.

[9] M. Opper. Learning in neural networks: Solvable dynamics. *Europhys. Lett.*, 8(4):389–392, 1989.

[10] S. F. Edwards and R. C. Jones. The eigenvalue spectrum of a large symmetric random matrix. *J. Phys. A: Math. Gen.*, 9(10):1595–1603, 1976.

[11] J. A. Hertz, A. Krogh, and G. I. Thorbergsson. Phase transitions in simple learning. *J. Phys. A: Math. Gen.*, 22(12):2133, 1989.

[12] L. Bachschmid-Romano, C. Battistin, M. Opper, and Y. Roudi. Variational perturbation and extended Plefka approaches to dynamics on random networks: the case of the kinetic Ising model. *J. Phys. A: Math. Gen.*, 49(43):434003, 2016.

[13] L. Bachschmid-Romano and M. Opper. Inferring

hidden states in a random kinetic Ising model: replica analysis. *J. Stat. Mech.*, P06013, 2014.

[14] C. Battistin, J. Hertz, J. Tyrcha, and Y. Roudi. Belief-propagation and replicas for inference and learning in a kinetic Ising model with hidden spins. *J. Stat. Mech.*, P05021, 2015.

[15] Y. Roudi and J. Hertz. Mean field theory for nonequilibrium network reconstruction. *Phys. Rev. Lett.*, 106(048702), 2011.

[16] B. Dunn and Y. Roudi. Learning and inference in a nonequilibrium Ising model with hidden spins. *Phys. Rev. E*, 87(022127), 2013.

[17] B. Bravi and P. Sollich. Inference for dynamics of continuous variables: the Extended Plefka Expansion with hidden nodes. *Arxiv preprint 1603.05538*, 2016.

[18] B. Bravi and P. Sollich. Critical scaling in hidden state inference for linear Langevin dynamics. *Arxiv preprint 1612.01976*, 2016.

[19] J. Berg. Out-of-equilibrium dynamics of gene expression and the Jarzynski equality. *Phys. Rev. Lett.*, 18(100):188101–188105, 2008.

[20] M. Opper and G. Sanguinetti. Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, 26(13):1623–1629, 2010.

[21] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[22] D. F. Anderson and T. G. Kurtz. Continuous time Markov chain models for chemical reaction networks. In H. Koeppl et al., editor, *Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology*, chapter 1, pages 1–44. Springer, 2011.

[23] M. H. A. Davis and R. B. Vinter. *Stochastic Modelling and Control.* Chapman and Hall, 1985.

[24] R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82(1):35–45, 1960.

[25] M.L. Mehta. *Random Matrices.* Elsevier-Academic Press, Amsterdam, 3rd edition, 2004.

[26] A. Vakili and B. Hassibi. On the asymptotic eigenvalue distribution of certain random Lyapunov and Riccati recursions. *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS)*, pages 453–458, 2010.

[27] J. Bun, R. Allez, J. P. Bouchaud, and M. Potters. Rotational invariant estimator for general noisy matrices. *IEEE Transactions on Information Theory*, 62(12), 2015.

[28] A. Crisanti and H. Sompolinsky. Dynamics of spin systems with random asymmetric bonds: Langevin dynamics and a spherical model. *Phys. Rev. A*, 36 (10):4922–4939, 1987.

[29] P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae*, 6(290-297), 1959.

[30] T. Rogers, I. Pérez Castillo, R. Kühn, and K. Takeda. Cavity approach to the spectral density of sparse symmetric random matrices. *Phys. Rev. E*, 78(031116), 2008.

[31] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina. Bayesian inference of epidemics on networks via Belief Propagation. *Phys. Rev. Lett.*, 112(11):118701, 2014.

[32] J. Bindi, A. Braunstein, and L. Dall'Asta. Predicting epidemic evolution on contact networks from partial observations. *Arxiv pre-print 1608.06516*, 2016.

[33] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-sb*, 1(457), 1967.

[34] D. V. Voiculescu, K. J. Dykema, and A. Nica. *Free random variables*, volume I of *CRM Monograph Series.* AMS, 1996.

[35] R. Speicher. Free probability and random matrices. *Proceedings of the ICM*, III:477–501, 2014.

[36] Z. Burda. Free products of large random matrices-a short review of recent developments. *J. Phys.: Conf. Ser.*, 473(012002), 2013.

[37] H. B. Callen and T. A. Welton. Irreversibility and generalized noise. *Phys. Rev.*, 83:34–40, 1951.

[38] B. Bravi. *Path integral approaches to subnetwork dynamics and inference.* PhD thesis, King's College London, 2016.

[39] E. Domany, J. L. van Hemmen, and K. Schulten, editors. *Models of Neural Networks III.* Springer, 1996.

[40] R. Kühn. Spectra of sparse random matrices. *J. Phys. A: Math. Theor.*, 41(295002), 2008.

[41] L. Erdős, A. Knowles, H. T. Yau, and J. Yin. Spectral Statistics of Erdős-Renyi Graphs II: Eigenvalue Spacing and the Extreme Eigenvalues. *Comm. Math. Phys.*, 314:587–640, 2012.

[42] L. Erdős, A. Knowles, H. T. Yau, and J. Yin. Spectral Statistics of Erdős-Renyi Graphs I: Local Semicircle Law. *Ann. Prob.*, 41:2279–2375, 2013.

[43] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *JMLR: Workshop and Conference Proceedings*, 1:1–16, 2007.

[44] B. Cseke, M. Opper, and G. Sanguinetti. Approximate inference in latent Gaussian-Markov models from continuous time observations. *Adv. Neural Inf. Process. Syst.*, 26:971–979, 2013.

[45] M. Opper and O. Winther. Expectation consistent approximate inference. *JMLR*, 6:2177–2204, 2005.

[46] Molinelli E. J. et al. Perturbation biology: Inferring signaling networks in cellular systems. *PLoS Comput. Biol.*, 2013.

[47] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[48] Note1. Rigorously only the recursive computation of forward messages should be referred to as Kalman filter [24], while equations of backward messages are known as Kalman smoothers.

[49] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1):7986, 1951.