

Phase Transitions in Nonlinear Filtering*

Patrick Rebeschini[†]

Ramon van Handel[‡]

Abstract

It has been established under very general conditions that the ergodic properties of Markov processes are inherited by their conditional distributions given partial information. While the existing theory provides a rather complete picture of classical filtering models, many infinite-dimensional problems are outside its scope. Far from being a technical issue, the infinite-dimensional setting gives rise to surprising phenomena and new questions in filtering theory. The aim of this paper is to discuss some elementary examples, conjectures, and general theory that arise in this setting, and to highlight connections with problems in statistical mechanics and ergodic theory. In particular, we exhibit a simple example of a uniformly ergodic model in which ergodicity of the filter undergoes a phase transition, and we develop some qualitative understanding as to when such phenomena can and cannot occur. We also discuss closely related problems in the setting of conditional Markov random fields.

Keywords: filtering in infinite dimension; conditional ergodicity & mixing; phase transitions.

AMS MSC 2010: 37A50; 60G35; 60K35; 82B26; 82B44.

1 Introduction

Let $(X_k, Y_k)_{k \geq 0}$ be a bivariate Markov chain. Such a model represents the setting of partial information: it is presumed that only $(Y_k)_{k \geq 0}$ can be observed, while $(X_k)_{k \geq 0}$ defines the unobserved dynamics. In order to understand the behavior of the unobserved process given the observations, it is natural to ‘lift’ the unobserved dynamics to the level of conditional distributions, that is, to investigate the nonlinear filter

$$\pi_k := \mathbf{P}[X_k \in \cdot | Y_1, \dots, Y_k].$$

Under standard assumptions on the observation structure (cf. section 2), the process $(\pi_k)_{k \geq 0}$ is itself a measure-valued Markov chain. The fundamental question that arises in this setting is to understand in what manner the probabilistic structure of the model $(X_k, Y_k)_{k \geq 0}$ ‘lifts’ to the conditional distributions $(\pi_k)_{k \geq 0}$.

Of particular interest in this context is the behavior of ergodic properties under conditioning. It is natural to suppose that the ergodic properties of $(X_k, Y_k)_{k \geq 0}$ will be inherited by the filter $(\pi_k)_{k \geq 0}$: for example, if X_k forgets its initial condition as $k \rightarrow \infty$, then the optimal mean-square estimate of X_k (and therefore the filter π_k) should intuitively possess the same property. Such a conclusion was already conjectured by Blackwell as early as

*Supported in part by NSF grant CAREER-DMS-1148711 and by the ARO through a PECASE award.

[†]Yale University, 17 Hillhouse Avenue, New Haven, CT 06520. E-mail: patrick.rebeschini@yale.edu

[‡]Sherrerd Hall, Princeton University, Princeton, NJ 08544. E-mail: rvan@princeton.edu

1957 [1], and a proof was provided by Kunita in 1971 [29]. Unfortunately, both the proof and the conclusion are erroneous: it is elementary to construct a finite-state Markov chain $(X_k, Y_k)_{k \geq 0}$ that is 1-dependent (as strong an ergodic property as one could hope for) with observations of the form $Y_k = h(X_{k-1}, X_k)$ such that the corresponding filtering process $(\pi_k)_{k \geq 0}$ is nonergodic, see Example 2.1 below.¹

Despite the appearance of counterexamples already in the most elementary setting, recent advances have provided a surprisingly complete picture of such problems in a general setting. On the one hand, it has been shown under very general assumptions [50, 46] that ergodicity of the underlying model is inherited by the filter when the observations are *nondegenerate*, that is, when the conditional law of each observation $\mathbf{P}[Y_k \in \cdot | X]$ has a positive density with respect to some fixed reference measure. This is a mild condition in classical filtering models that serves mainly to rule out the singular case of noiseless observations: for example, the addition of any observation noise to the above counterexample would render the filter ergodic. On the other hand, even in the noiseless case, ergodicity is inherited in the absence of certain symmetries that are closely related to systems-theoretic notions of *observability* [48, 49, 51, 9]. One can therefore conclude that while there exist elementary examples where the ergodicity of the model fails to be inherited by the filter, such examples must be very fragile as they require both a singular observation structure and the presence of unusual symmetries, either of which is readily broken by a small perturbation of the model.

The theory outlined above provides a satisfactory understanding of conditional ergodicity in classical filtering models. Some care must be taken, however, in interpreting this conclusion. The ubiquitous applicability of the theory hinges on the notion that most filtering models possess observation densities, an assumption made almost universally in the filtering literature (cf. [13] and the references therein). This assumption is largely innocuous in finite-dimensional systems. The situation is entirely different in infinite dimension, where singularity of probability measures is the norm. There exists almost no mathematical literature on filtering in infinite dimension, despite the substantial practical importance of infinite-dimensional filtering models in data assimilation problems that arise in areas such as weather forecasting or geophysics [44]. The aim of this paper is to draw attention to the fact that, far from being a technical issue, the infinite-dimensional setting gives rise to new probabilistic phenomena and questions in filtering theory that are fundamentally different than those that have been studied in the literature to date, and whose understanding remains limited.

Remark 1.1. The singularity of measures in infinite dimension is already a significant hurdle in the ergodic theory of infinite-dimensional Markov processes, as it precludes application of standard methods based on Harris recurrence (e.g., [34]). An infinite-dimensional unobserved process is not, however, the primary source of the problems considered in the sequel: as long as the observations are nondegenerate, it was shown in [46] that ergodicity is still inherited by the filter. The problem is that nondegeneracy is a restrictive assumption on the observations in infinite dimension: it implies that the observations must be effectively finite-dimensional [46, Remark 5.20]. It is in models where the observations are genuinely infinite-dimensional that new phenomena appear.

To model a filtering problem in infinite dimension, we suppose that $(X_k, Y_k)_{k \geq 0}$ is a Markov chain in the product state space $E^V \times F^V$, where E, F are local state spaces and

¹Surprisingly, the counterexample (intended for a different purpose) appears in Blackwell's own paper [1].

V is a countably infinite set of sites (for concreteness, we fix $V = \mathbb{Z}^d$ throughout). Each element of V should be viewed as a single dimension of the model. A more practical interpretation is that V defines a spatial degree of freedom and that $(X_k, Y_k)_{k \geq 0}$ describes the dynamics of a time-varying random field, as is the case in data assimilation applications. In accordance with this interpretation, we will assume that the dynamics of the state X_k and the observations Y_k are local in nature: that is, the conditional distributions of the local state X_k^v given the previous state X_{k-1} , and of the local observation Y_k^v given the underlying process X , depend only on X_{k-1}^w and X_k^w for sites $w \in V$ that are neighbors of v . In essence, our basic model therefore consists of an infinite family of local filtering models $(X_k^v, Y_k^v)_{k \geq 0}$ whose dynamics are locally coupled according to the graph structure of $V = \mathbb{Z}^d$. The general model is described in detail in section 2.

In section 3 we investigate the natural infinite-dimensional version of Blackwell's Example 2.1. Recall that it was crucial in the finite-dimensional setting that the observations $Y_k = h(X_{k-1}, X_k)$ are noiseless: the addition of any noise renders the observations nondegenerate and then ergodicity is preserved. This is no longer the case in infinite dimension: even if the local observations Y_k^v are nondegenerate, the failure of the filter to inherit ergodicity can persist. In fact, we observe a phase transition: the filter fails to be ergodic when the noise is small, but becomes ergodic when the noise strength exceeds a strictly positive threshold. The remarkable feature of this phenomenon is that no qualitative change of any kind occurs in the ergodic properties of the underlying model: $(X_k^v, Y_k^v)_{k \geq 0, v \in V}$ is a 1-dependent random field for every value of the noise parameter. We are therefore in the surprising situation that complex ergodic behavior emerges in an otherwise trivial model when we consider its conditional distributions. Such conditional phase transitions cannot arise in finite dimension.

The above example indicates that our intuition about inheritance of ergodicity, which fails in classical filtering models only in pathological cases, cannot be taken for granted in infinite dimension even under local nondegeneracy assumptions. This raises the question as to whether there are situations in which the inheritance of ergodicity is guaranteed. In view of the finite-dimensional theory, it is natural to conjecture that this might be the case under a symmetry breaking assumption. As will be discussed in section 4, the existing observability theory is not satisfactory in infinite dimension, and even in its simplest form the verification of this conjecture remains an open problem. However, we will show in section 4.2 that the problem can be resolved in translation-invariant models for a (not entirely natural) modification of the usual filtering process using tools from multidimensional ergodic theory. In section 4.3 the analogous problem will be considered in continuous time, where the method of section 4.2 can be extended to provide a full proof of the conjecture in the translation-invariant case.

In section 5 we turn our attention to the counterpart of the filter stability problem in the setting of Markov random fields. Such problems provide a simple setting for the investigation of decay of correlations in filtering problems, and are of interest in their own right as models that arise, for example, in image analysis. Here the natural question of interest is whether the spatial mixing properties of random fields are inherited by conditioning on local observations. A direct adaptation of the example of section 3 shows that the answer is negative in general. However, we will show in section 5.3 that the conditional random field inherits mixing properties of the underlying model when the latter possesses certain monotonicity properties. This provides an entirely different mechanism for conditional ergodicity than in the observability theory of section 4.

As will rapidly become evident, the above results repeatedly exploit connections between filtering in infinite dimension, the statistical mechanics of disordered systems [5], and multidimensional ergodic theory [10, 22]. Some further connections with other probabilistic problems will be discussed in the final section 6 of this paper.

The consideration of filtering problems in infinite dimension is far from esoteric. It is a long-standing problem in applied probability to develop efficient algorithms to compute nonlinear filters in high-dimensional models (current algorithms require a computational effort exponential in the dimension, a severe problem in data assimilation applications). Improved understanding of the ergodic and spatial mixing properties of the filter may be essential for progress on such practical problems. In recent work [40, 41], we have shown that local filtering algorithms can attain dimension-free approximation errors in models that exhibit conditional decay of correlations. The machinery developed in [40, 41] to investigate high-dimensional filtering problems is complementary to the present paper: the former provides strong quantitative results under strong (‘high-temperature’) assumptions, while our aim here is to address fundamental issues that arise in infinite dimension. Regardless of any practical implications, however, the investigation of emergent phenomena that arise from conditioning is arguably of fundamental interest to the understanding of conditional distributions, and provides a compelling motivation for the investigation of conditional phenomena in probability theory.

2 Filtering in infinite dimension

The goal of this section is to set up the basic filtering problem that will be studied in the sequel. We begin by defining a general setting for nonlinear filtering and introduce and discuss the basic ergodicity question in section 2.1. We subsequently introduce our canonical infinite-dimensional filtering model in section 2.2.

2.1 Nonlinear filtering and ergodicity

Throughout this paper, we model dynamics with partial information as a Markov chain $(X_k, Y_k)_{k \geq 0}$ that has the additional property that its transition kernel factorizes as

$$\mathbf{P}[(X_k, Y_k) \in A | X_{k-1}, Y_{k-1}] = \int \mathbf{1}_A(x, y) P(X_{k-1}, dx) \Phi(X_{k-1}, x, dy)$$

for given transition kernels P and Φ : the factorization corresponds to the assumption that $(X_k)_{k \geq 0}$ is a Markov chain in its own right, and that the observations $(Y_k)_{k \geq 0}$ are conditionally independent given $(X_k)_{k \geq 0}$. Such models are frequently called *hidden Markov models*. While not essential for the development of our theory (see, e.g., [45] for a more general setting), the hidden Markov model setting is convenient mathematically and is ubiquitous in practice as a model of noisy observations of random dynamics.

For the time being, we assume that X_k and Y_k take values in an arbitrary Polish space (we will define a more concrete infinite-dimensional setting in section 2.2 below). The *nonlinear filter* is defined as the regular conditional probability

$$\pi_k := \mathbf{P}[X_k \in \cdot | Y_1, \dots, Y_k].$$

We are interested in the question of whether $(\pi_k)_{k \geq 0}$ inherits the ergodic properties of the underlying dynamics $(X_k)_{k \geq 0}$. There are several different but closely connected ways to make this question precise (cf. Remark 2.3 below). For concreteness, we will focus attention

on one particularly elementary formulation of this question that will serve as the guiding problem to be investigated throughout this paper.

We will assume in the sequel that the Markov chain $(X_k)_{k \geq 0}$ admits a unique invariant measure λ . As $\mathbf{P}[X_k, Y_k \in \cdot | X_{k-1}, Y_{k-1}]$ does not depend on Y_{k-1} due to the hidden Markov structure, the invariant measure λ extends uniquely to an invariant measure for the chain $(X_k, Y_k)_{k \geq 0}$, and we denote the unique stationary law of this process as \mathbf{P} . By stationarity, we can assume in the sequel that $(X_k, Y_k)_{k \in \mathbb{Z}}$ is defined also for $k < 0$.

The ergodic property of $(X_k)_{k \geq 0}$ that we will consider is *stability* in the sense that

$$|\mathbf{P}[X_k \in A | X_0] - \lambda(A)| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1$$

for every measurable set A : that is, the law of X_k ‘forgets’ the initial condition X_0 as $k \rightarrow \infty$. The analogous conditional property is *filter stability* in the sense that

$$|\mathbf{P}[X_k \in A | X_0, Y_1, \dots, Y_k] - \mathbf{P}[X_k \in A | Y_1, \dots, Y_k]| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1$$

for every measurable set A : that is, the conditional distribution of X_k given the observed data ‘forgets’ the initial condition X_0 as $k \rightarrow \infty$. It is natural to suppose that stability of the underlying dynamics will imply stability of the filter. This conclusion is incorrect, however, as is illustrated by the following classical example [1].

Example 2.1. Let $(X_k)_{k \geq 0}$ be an i.i.d. sequence of random variables with $\mathbf{P}[X_k = 1] = \mathbf{P}[X_k = -1] = 1/2$, and let $Y_k = X_k X_{k-1}$ for $k \geq 1$. This evidently defines a stationary hidden Markov model with $P(x, \cdot) = (\delta_1 + \delta_{-1})/2$ and $\Phi(x', x, \cdot) = \delta_{xx'}$. Note that

$$X_k = X_0 Y_1 Y_2 \cdots Y_k.$$

We can therefore easily compute for every $k \geq 0$

$$\begin{aligned} \mathbf{P}[X_k = 1 | X_0, Y_1, \dots, Y_k] &= \mathbf{1}_{X_k=1}, \\ \mathbf{P}[X_k = 1 | Y_1, \dots, Y_k] &= 1/2. \end{aligned}$$

Thus the filter is certainly not stable. On the other hand, underlying dynamics $(X_k)_{k \geq 0}$ is an i.i.d. sequence, and is therefore stable in the strongest possible sense:

$$\mathbf{P}[X_k \in A | X_0] = \lambda(A) \quad \text{for all } k \geq 1.$$

Moreover, even the process $(X_k, Y_k)_{k \geq 0}$ is stable in the strongest possible sense: it is a 1-dependent sequence, so that $\mathbf{P}[(X_k, Y_k) \in A | X_0, Y_0] = \mathbf{P}[(X_k, Y_k) \in A]$ for all $k \geq 2$.

Example 2.1 shows that the inheritance of ergodicity under conditioning cannot be taken for granted. Nonetheless, the phenomenon exhibited here is very fragile: if the observations are perturbed by any noise (for example, if we set $Y_k = X_k X_{k-1} \xi_k$ with $\mathbf{P}[\xi_k = -1] = 1 - \mathbf{P}[\xi_k = 1] = p$ and any $0 < p < 1$), the filter will become stable. The inheritance of ergodicity is therefore apparently obstructed by the singularity of the observation kernel Φ . To rule out such singular behavior, it is natural to require that the observation kernel Φ possesses a positive density with respect to some reference measure φ . A model with this property is said to possess *nondegenerate observations*. One might now expect that nondegeneracy of the observations removes the obstruction to inheritance of ergodicity observed in Example 2.1. Unfortunately, this is still not the case in complete generality, as is demonstrated by an esoteric counterexample in [52]. However, the conclusion does hold if we use a stronger *uniform* notion of stability.

Theorem 2.2 ([50]). *Suppose that the following hold.*

1. *The underlying dynamics is uniformly stable in the sense*

$$\sup_A |\mathbf{P}[X_k \in A|X_0] - \lambda(A)| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1.$$

2. *The observations are nondegenerate in the sense*

$$\Phi(x', x, dy) = g(x', x, y) \varphi(dy), \quad g(x', x, y) > 0 \text{ for all } x, x', y.$$

Then the filter is uniformly stable in the sense

$$\sup_A |\mathbf{P}[X_k \in A|X_0, Y_1, \dots, Y_k] - \mathbf{P}[X_k \in A|Y_1, \dots, Y_k]| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1.$$

This result, together with the mathematical theory behind its proof (cf. section 6.1), provides a very general qualitative understanding of the inheritance of ergodicity in classical filtering models. However, as will be explained below, this theory breaks down completely in infinite-dimensional models. In the remainder of this paper, we will see that new phenomena arise in the infinite-dimensional setting.

Remark 2.3. The question of inheritance of ergodic properties under conditioning can be formulated in a number of different ways. For concreteness, we focus our attention in this paper on the elementary formulation introduced above. As the choice of problem is somewhat arbitrary, let us briefly describe a number of alternative formulations.

In the setting of stability of the filter, we have considered ‘forgetting’ of the initial condition X_0 under the stationary measure. Similar problems can be formulated, however, in a more general setting. Denote by \mathbf{P}^μ the law of the process $(X_k, Y_k)_{k \geq 0}$ with the initial distribution $X_0 \sim \mu$. A natural notion of stability is to require that

$$\mathbf{P}^\mu[X_k \in \cdot] \xrightarrow{k \rightarrow \infty} \lambda \quad \text{for every } \mu$$

in a suitable topology on probability measures. If we define the filter started at μ as $\pi_k^\mu := \mathbf{P}^\mu[X_k \in \cdot | Y_1, \dots, Y_k]$, we can now investigate the general filter stability problem

$$|\pi_k^\mu(f) - \pi_k^\nu(f)| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1(\mathbf{P}^\gamma)$$

for a suitable class of measures μ, ν, γ and functions f . The formulation that we consider in this paper corresponds to the special case $\nu = \lambda$ and $\mu = \gamma = \delta_x$ for x outside a λ -null set. Nonetheless, our formulation proves to be equivalent in a rather general setting to stability for general initial measures μ, ν, γ , cf. [13, Chapter 12] and [50, 46].

A different and perhaps more natural formulation dates back to Blackwell [1] and Kunita [29]. Using the Markov property of the underlying model, it is not difficult to show that the measure-valued stochastic process $(\pi_k)_{k \geq 0}$ is itself a Markov chain, cf. [52, Appendix A]. One can now ask whether the ergodic properties of the Markov chain $(X_k)_{k \geq 0}$ ‘lift’ to ergodic properties of the Markov chain $(\pi_k)_{k \geq 0}$. For example, if $(X_k)_{k \geq 0}$ admits a unique stationary measure, does $(\pi_k)_{k \geq 0}$ admit a unique stationary measure also? Similarly, if $(X_k)_{k \geq 0}$ converges to its stationary measure starting from any initial condition, does the same property hold for $(\pi_k)_{k \geq 0}$? Remarkably, while these questions appear in first instance

to be quite distinct from the question of filter stability, such properties again prove to be equivalent in a very general setting to the notion of filter stability that we consider in this paper, cf. [29, 43, 9, 13, 52].

A third formulation of inheritance of ergodicity under conditioning is obtained when we consider, rather than the filter, the conditional distribution of the entire process $X = (X_k)_{k \in \mathbb{Z}}$ given the infinite observation sequence $Y = (Y_k)_{k \in \mathbb{Z}}$. Using the Markov property of the underlying model, it is not difficult to establish that X is still a Markov process under the conditional distribution $\mathbf{P}[\cdot | Y]$, albeit time-inhomogeneous and with transition probabilities that depend on the realized observation sequence Y : that is, the conditional process is a Markov chain in a random environment. One can now ask whether the process X inherits its ergodic properties under \mathbf{P} when it is considered under the conditional distribution $\mathbf{P}[\cdot | Y]$. Once again, this apparently distinct formulation proves to be equivalent in a general setting the formulation considered in this paper, a fact that is exploited heavily in the theory of [50, 46].

It is now well understood that the properties described above are equivalent in classical filtering models. While some of these arguments extend directly to the infinite-dimensional setting, others do not, and it remains to be investigated to what extent these equivalences remain valid in infinite dimension. Nonetheless, the problem formulation considered here is arguably the most elementary one, and provides a natural starting point for the investigation of conditional phenomena in infinite dimension.

Remark 2.4. Even when the underlying dynamics $(X_k)_{k \geq 0}$ is not stable, it may be the case that the filter is stable. For example, using the trivial observation model $Y_k = X_k$, the filter is stable regardless of any properties of the underlying model. More generally, the filter is expected to be stable when the observations are ‘sufficiently informative,’ which is made precise in [48, 49, 51] in terms of nonlinear notions of *observability*. Such results are in some sense the opposite of Theorem 2.2: the latter shows that ergodicity is inherited by the filter, while the former show that the filter can be ergodic regardless of ergodicity of the underlying model (even without nondegeneracy). None of these results prove to be satisfactory in infinite dimension: it appears that a general theory for ergodicity of the filter will require both ergodicity of the underlying model and some form of observability, as will become evident in the following sections.

2.2 The infinite-dimensional model

The aim of this paper is to investigate conditional phenomena that arise in infinite dimension. So far, no assumptions have been made on the model dimension: we have set up our theory in any Polish state space. Nonetheless, while no explicit dimensionality requirements appear, for example, in Theorem 2.2, the assumptions of previous results can typically hold only in finite-dimensional situations. To understand the problems that arise in infinite dimension, and to provide a concrete setting for the investigation of conditional phenomena in infinite dimension, we presently introduce a canonical infinite-dimensional filtering model that will be used in the sequel.

The practical interest in infinite-dimensional filtering models stems from problems that have spatial in addition to dynamical structure. To model this situation, let us assume for concreteness that the spatial degrees of freedom are indexed by the infinite lattice \mathbb{Z}^d . We also define Polish spaces E and F that describe the state of the model at each spatial location. We now assume that X_k and Y_k are random fields that are indexed by \mathbb{Z}^d and

take values locally in E and F , respectively, for every time k : that is,

$$X_k = (X_k^v)_{v \in \mathbb{Z}^d} \in E^{\mathbb{Z}^d} \quad \text{and} \quad Y_k = (Y_k^v)_{v \in \mathbb{Z}^d} \in F^{\mathbb{Z}^d}.$$

Each $v \in \mathbb{Z}^d$ should be viewed as a single ‘dimension’ of the model.² We now define a hidden Markov model that respects the spatial structure of the problem by assuming that both the underlying dynamics and the observations are *local*: that is, we assume that the transition and observation kernels P and Φ factorize as

$$P(x, dz) = \prod_{v \in \mathbb{Z}^d} P^v(x, dz^v), \quad \Phi(x, z, dy) = \prod_{v \in \mathbb{Z}^d} \Phi^v(x, z, dy^v),$$

where

$$P^v(x, A) \quad \text{and} \quad \Phi^v(x, z, B) \quad \text{depend only on } x^w, z^w \text{ for } \|w - v\| \leq 1.$$

Such a model should be viewed as a hidden Markov model counterpart of probabilistic cellular automata [30] or interacting particle systems [31] that have been widely investigated in the literature as natural models of space-time dynamics. Alternatively, one might view such a model as an infinite collection $(X_k^v, Y_k^v)_{k \geq 0}$ of hidden Markov models whose dynamics and observations are locally coupled to their neighbors in \mathbb{Z}^d .

While problems of this type have been rarely considered in filtering theory, the infinite-dimensional model that we have formulated is in principle a special case of the general model described in the previous section. However, its structure is such that the assumptions of a result such as Theorem 2.2 typically cannot hold. Let us consider, for example, the setting where each local observation Y^v has a positive density of the form $\Phi^v(x, z, dy^v) = g(z^v, y^v) \varphi(dy^v)$, so that the observations are *locally nondegenerate*. Choose two values $e, e' \in E$ such that $g(e, \cdot) \neq g(e', \cdot)$, and define the constant configurations z, z' as $z^v = e$ and $z'^v = e'$ for all $v \in \mathbb{Z}^d$. Then the measures $\Phi(x, z, \cdot)$ and $\Phi(x, z', \cdot)$ are two distinct laws of an infinite number of i.i.d. random variables, and are therefore mutually singular. This immediately rules out the possibility that the observations are nondegenerate in the sense of Theorem 2.2. It is precisely this problem that lies at the heart of the difficulties in infinite-dimensional models: probability measures in infinite dimension are typically mutually singular, even when they admit densities locally (that is, for any finite-dimensional marginal). In the absence of densities, classical results in filtering theory cannot be taken for granted, and the study of filtering in infinite dimension gives rise to fundamentally different problems than have been studied in the literature to date. We initiate the investigation of such problems in the sequel.

Remark 2.5. The singularity of measures in infinite dimension is problematic not only for the nondegeneracy of observations, but also for the ergodic theory of Markov chains. For example, the uniform stability property in Theorem 2.2 will rarely hold in infinite dimension: it is often the case that the law of X_k is singular with respect to λ for all

²The present setting is easily extended to the setting of more general locally finite graphs and to the setting where each location v may possess a different local state space E^v . Such an extension does not illuminate significantly the phenomena that will be investigated in the sequel. On the other hand, a nontrivial extension of substantial interest in applications is to continuous infinite-dimensional models such as stochastic partial differential equations, cf. [44]. While we do not investigate such models here, the present setting may be viewed as prototypical for more complicated models of this type, see [40] for further discussion.

$k < \infty$, which rules out total variation convergence (see [46, Example 2.3] for a simple illustration). However, this issue is surmounted in [46] using a form of localization: by performing the analysis of Theorem 2.2 locally (that is, to finite-dimensional projections of the original model), we can avoid the singularity of the full infinite-dimensional problem. This allows to extend the conclusion of Theorem 2.2 to a wide range of infinite-dimensional models with nondegenerate observations. In practice, this implies that much of the classical filtering theory extends, at least in spirit, to models where X_k is infinite-dimensional but Y_k is (effectively) finite-dimensional. It is only when the observations Y_k are also infinite-dimensional that new phenomena arise.

Remark 2.6. Let us note that we have used the term ‘infinite-dimensional’ to denote the situation where there are infinitely many independent degrees of freedom, which is the key issue in our setting. The problem of dimension is unrelated to the linear algebraic or metric dimension of the state space: indeed, even each of the local state spaces E and F in our model can itself be an arbitrary Polish space. Conversely, it is possible to have infinite-dimensional systems that are ‘effectively finite-dimensional’ in the sense that only finitely many degrees of freedom carry significant information. This is common, for example, in stochastic partial differential equations (see, e.g., [46]).

At the same time, it should be noted that even in finite-dimensional systems where results such as Theorem 2.2 technically apply, the qualitative information contained in such statements may be misleading from the practical point of view: in finite but high-dimensional systems, phenomena that arise qualitatively in infinite dimension are still manifested in a quantitative fashion (see [40] for quantitative results and discussion on filtering in high dimension). For example, if the filter is not stable for the infinite-dimensional model, it will often still be the case that the filter is stable for every finite-dimensional truncation of the model; however, the quantitative rate of stability will vanish rapidly as the dimension is increased. Conversely, if the filter is stable for the infinite-dimensional model, then the rate of stability of the filter for the finite-dimensional models will be dimension-free. As it is ultimately the quantitative behavior of filtering algorithms that is of importance in practice, the qualitative phenomena investigated here in infinite dimension can still provide more insight into the behavior of practical filtering problems in high dimension than classical results in filtering theory.

3 A conditional phase transition

3.1 Model and results

The goal of this section is to develop a simple example of the general infinite-dimensional setting of section 2.2 where we observe nontrivial behavior of the inheritance of ergodicity. This model, to be described presently, is a natural infinite-dimensional variation on Blackwell’s counterexample (Example 2.1 above).

Throughout this section,

$$X_k = (X_k^v)_{v \in \mathbb{Z}} \in \{-1, 1\}^{\mathbb{Z}} \quad \text{and} \quad Y_k = (\bar{Y}_k^v, \hat{Y}_k^v)_{v \in \mathbb{Z}} \in (\{-1, 1\} \times \{-1, 1\})^{\mathbb{Z}}$$

are binary random fields in one spatial dimension. We let

$$(X_k^v)_{k, v \in \mathbb{Z}} \text{ are i.i.d. with } \mathbf{P}[X_k^v = 1] = 1/2,$$

and we let

$$\bar{Y}_k^v = X_k^v X_{k-1}^v \bar{\xi}_k^v, \quad \hat{Y}_k^v = X_k^v X_k^{v+1} \hat{\xi}_k^v,$$

where

$$(\bar{\xi}_k^v)_{k,v \in \mathbb{Z}}, (\hat{\xi}_k^v)_{k,v \in \mathbb{Z}} \text{ are i.i.d. with } \mathbf{P}[\bar{\xi}_k^v = -1] = p$$

and $(\bar{\xi}_k^v)_{k,v \in \mathbb{Z}}, (\hat{\xi}_k^v)_{k,v \in \mathbb{Z}}$ are independent of $(X_k^v)_{k,v \in \mathbb{Z}}$.

This evidently corresponds to a model of the form discussed in section 2.2. In words, the underlying dynamics is of the simplest possible type: each time and each spatial location is an independent random variable. When $p = 0$, the observations reveal for each site whether its current state differs from its state at the previous time and from the states of its two neighbors at the present time. When $p > 0$, each observation is subject to additional noise that inverts the outcome with probability p . By symmetry, it will suffice to consider the case $p \leq 1/2$, which we will do from now on.

The model that we have constructed is evidently a direct extension of Example 2.1 to infinite dimension. As in Example 2.1, the process $(X_k, Y_k)_{k \in \mathbb{Z}}$ is ergodic in the strongest sense, so that even the uniform stability assumption of Theorem 2.2 is satisfied. When $p = 0$, it is easily seen by the same reasoning as in Example 2.1 that the filter is not stable. However, in Example 2.1 the addition of observation noise with error probability $p > 0$ would yield nondegenerate observations, and thus filter stability by Theorem 2.2. In the present setting, on the other hand, nondegeneracy fails for any p . Nonetheless, the observations are *locally nondegenerate* when $p > 0$, and one might conjecture that this suffices to ensure inheritance of ergodicity. This is not the case.

Theorem 3.1. *For the model of this section, there exist constants $0 < p_\star \leq p^\star < 1/2$ such that the filter is stable for $p^\star < p \leq 1/2$ and is not stable for $0 \leq p < p_\star$.*

Remark 3.2. We naturally believe that one can choose $p_\star = p^\star$ in Theorem 3.1, but we did not succeed in proving that. The proof yields some explicit bounds on p_\star and p^\star .

Theorem 3.1 shows that local nondegeneracy does not suffice to ensure inheritance of ergodicity in infinite dimension: ergodicity of the filter undergoes a *phase transition* at a strictly positive signal to noise ratio of the observations. Remarkably, the underlying model does not seem to exhibit any qualitative change in behavior: $(X_k^v, Y_k^v)_{k,v \in \mathbb{Z}}$ is a one-dependent random field for every value of the error probability p . Thus it is evidently possible in infinite dimension that complex ergodic behavior emerges in an otherwise trivial model when we consider its conditional distributions.

The remainder of this section is devoted to the proof of Theorem 3.1. The proof relies on standard tools from statistical mechanics [5, 22]: a Peierls argument for the low noise regime and a Dobrushin contraction method for the high noise regime.

3.2 Proof of Theorem 3.1: low noise

We begin by noting that as $(X_k^v, Y_k^v)_{k,v \in \mathbb{Z}}$ and $(-X_k^v, Y_k^v)_{k,v \in \mathbb{Z}}$ have the same law, it follows that $\mathbf{E}[X_k^0 | Y_1, \dots, Y_k] = \mathbf{E}[-X_k^0 | Y_1, \dots, Y_k]$, and we therefore have

$$\mathbf{E}[X_k^0 | Y_1, \dots, Y_k] = 0 \quad \text{for all } k \geq 1.$$

To prove that the filter is not stable, it therefore suffices to show that

$$\inf_{k \geq 1} \mathbf{E}[\mathbf{E}[X_k^0 | X_0, Y_1, \dots, Y_k]] > 0.$$

To show this, we begin by reducing the problem to finite dimension.

Lemma 3.3. *Suppose that $0 < p \leq 1/2$. Then*

$$\mathbf{E}[X_k^0 | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}] \xrightarrow{m \rightarrow \infty} \mathbf{E}[X_k^0 | X_0, Y_1, \dots, Y_k] \quad a.s.$$

Proof. Let $\beta := \log \sqrt{(1-p)/p} > 0$. We begin by noting that

$$\mathbf{P}[\hat{Y}_\ell^v = y | X_0, \dots, X_k] = \sqrt{p(1-p)} e^{\beta y X_\ell^v X_\ell^{v+1}}$$

for $1 \leq \ell \leq k$ and $y \in \{-1, 1\}$. Define the probability measure \mathbf{Q} such that

$$\mathbf{P}(A) = \mathbf{E}_{\mathbf{Q}} \left[\mathbf{1}_A \prod_{\ell=1}^k 4p(1-p) e^{\beta \hat{Y}_\ell^m X_\ell^m X_\ell^{m+1}} e^{\beta \hat{Y}_\ell^{-m-1} X_\ell^{-m-1} X_\ell^{-m}} \right].$$

Then under \mathbf{Q} , the observations \hat{Y}_ℓ^m and \hat{Y}_ℓ^{-m-1} , $1 \leq \ell \leq k$ are symmetric Bernoulli and independent from all the remaining variables in the model, while the remainder of the model is the same as defined above. In particular, this implies that

$$\{X_0^v, X_\ell^v, Y_\ell^v : 1 \leq \ell \leq k, |v| > m\} \perp \{X_0^v, X_\ell^v, Y_\ell^v : 1 \leq \ell \leq k, |v| \leq m\} \quad \text{under } \mathbf{Q}.$$

We therefore obtain using the Bayes formula

$$\begin{aligned} \mathbf{P}[A | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}] &= \\ \frac{\mathbf{E}_{\mathbf{Q}}[\mathbf{1}_A \frac{d\mathbf{P}}{d\mathbf{Q}} | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}]}{\mathbf{E}_{\mathbf{Q}}[\frac{d\mathbf{P}}{d\mathbf{Q}} | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}]} & \\ \geq e^{-4\beta k} \mathbf{Q}[A | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}] & \\ = e^{-4\beta k} \mathbf{Q}[A | X_0, Y_1, \dots, Y_k] & \end{aligned}$$

for any $A \in \sigma\{X_0, Y_1, \dots, Y_k, X_1^v, \dots, X_k^v : |v| \leq m\}$.

Define $Z^0 := (X_1^0, \dots, X_k^0)$ and $Z^{-m} := (X_1^m, \dots, X_k^m, X_1^{-m}, \dots, X_k^{-m})$ for $m \geq 1$. Due to the conditional independence structure of the infinite-dimensional filtering model,

$$\mathbf{E}[f(Z^{-m}) | X_0, Y_1, \dots, Y_k, Z^{-m-1}, Z^{-m-2}, \dots] = \mathbf{E}[f(Z^{-m}) | X_0, Y_1, \dots, Y_k, Z^{-m-1}]$$

for every $m \geq 0$. Thus $(Z^m)_{m \leq 0}$ is a Markov chain under any regular version of the conditional distribution $\mathbf{P}[\cdot | X_0, Y_1, \dots, Y_k]$ (almost surely with respect to the realization of X_0, Y_1, \dots, Y_k). Moreover, the above estimate shows that the (random) transition kernels of this Markov chain satisfy the Doeblin condition [35, Theorem 16.2.4], so

$$|\mathbf{E}[X_k^0 | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}] - \mathbf{E}[X_k^0 | X_0, Y_1, \dots, Y_k]| \leq 2(1 - e^{-4\beta k})^{m+1}$$

for all $m \geq 0$. This completes the proof. \square

Lemma 3.3 reduces our problem to a finite-dimensional one. Indeed, it is clear that the filter is not stable for $p = 0$ (for precisely the same reason as in Example 2.1), so we will assume without loss of generality in the sequel that $0 < p \leq 1/2$. Applying Lemma 3.3, it follows that in order to prove that the filter is not stable, it suffices to show that

$$\inf_{k, m \geq 1} \mathbf{E}[\mathbf{E}[X_k^0 | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}]] > 0.$$

But the conditional independence structure of the infinite-dimensional filtering model implies that the conditional expectation inside this expression depends only on X_ℓ^v and Y_ℓ^v for $0 \leq \ell \leq k$ and $|v| \leq m + 1$. We are thus faced with the problem of obtaining a lower bound on this finite-dimensional quantity that is uniform in k, m .

To lighten the notation, it will be convenient to view $(X_k^v)_{k,v \in \mathbb{Z}}$ not as a sequence of spatial random fields on \mathbb{Z} , but rather as a single space-time random field on \mathbb{Z}^2 . To this end, we will write $X^q := X_k^v$ for $q = (k, v) \in \mathbb{Z}^2$. We will similarly write $Y^{qr} := \bar{Y}_k^v$ and $\xi^{qr} := \bar{\xi}_k^v$ if $q = (k - 1, v)$ and $r = (k, v)$, and $Y^{qr} := \hat{Y}_k^v$ and $\xi^{qr} := \hat{\xi}_k^v$ if $q = (k, v)$ and $r = (k, v + 1)$ (the order of the indices q, r is irrelevant, that is, $Y^{qr} := Y^{rq}$ etc.) In this manner, we can view $X = (X^q)_{q \in \mathbb{Z}^2}$ as a random field on the lattice \mathbb{Z}^2 , with observations Y^{qr} attached to each edge $\{q, r\} \subset \mathbb{Z}^2$ with $\|q - r\| = 1$.

Lemma 3.4. *Suppose that $0 < p \leq 1/2$, and let $k, m \geq 1$. Define $\beta := \log \sqrt{(1-p)/p}$, $J := [1, k] \times [-m, m]$, and $\partial J := \{0\} \times [-m, m] \cup [1, k] \times \{-m - 1, m + 1\}$. For any given configuration $x \in \{-1, 1\}^{\mathbb{Z}^2}$, we define the random measure Σ on $\{-1, 1\}^J$ as*

$$\Sigma^x(\{z\}) := \frac{1}{Z} \exp \left(\beta \left\{ \sum_{\{q,r\} \subseteq J: \|q-r\|=1} \xi^{qr} x^q x^r z^q z^r + \sum_{q \in J, r \in \partial J: \|q-r\|=1} \xi^{qr} x^q z^q \right\} \right),$$

where Z is the normalization such that $\Sigma^x(J) = 1$. Then

$$\mathbf{E}[(X^q)_{q \in J} \in A | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}] = \Sigma^X(A).$$

Proof. By the conditional independence structure of the filtering model, we have

$$\begin{aligned} \mathbf{E}[(X^q)_{q \in J} \in A | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}] = \\ \mathbf{E}[(X^q)_{q \in J} \in A | (X^q)_{q \in \partial J}, (Y^{qr})_{q \in J, r \in J \cup \partial J, \|q-r\|=1}]. \end{aligned}$$

The joint distribution of the random variables that appear in this expression is

$$\begin{aligned} \mathbf{P}[(X^q)_{q \in J \cup \partial J} = z, (Y^{qr})_{q \in J, r \in J \cup \partial J, \|q-r\|=1} = y] = 2^{-|J \cup \partial J|} \times \\ \prod_{\{q,r\} \subseteq J: \|q-r\|=1} \sqrt{p(1-p)} e^{\beta y^{qr} z^q z^r} \prod_{q \in J, r \in \partial J: \|q-r\|=1} \sqrt{p(1-p)} e^{\beta y^{qr} z^q z^r}, \end{aligned}$$

where $|A|$ denotes the cardinality of a set A . The result now follows readily from the Bayes formula and the fact that $Y^{qr} = X^q X^r \xi^{qr}$ by construction. \square

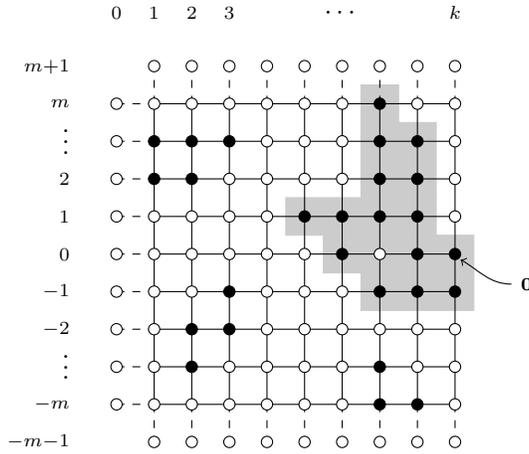
Lemma 3.4 shows that the conditional distribution $\mathbf{P}[\cdot | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}]$ has a familiar form in statistical mechanics: it is (up to the change of variables or *gauge transformation* $\sigma^q = x^q z^q$) an Ising model with random interactions, also known as a *random bond Ising model* or an *Ising spin glass*, with inverse temperature $\beta = \log \sqrt{(1-p)/p}$. The failure of stability of the filter for large β can now be addressed using a standard method in statistical mechanics [5, section 6.4]. For concreteness, we include the requisite arguments in the present setting, which completes the proof.

Remark 3.5. For the benefit of readers who are not familiar with ideas from statistical mechanics, we provide a brief description of the Peierls argument that is used below.

In Lemma 3.4, one should interpret x as the true configuration of the underlying system, and z as the configuration in the set J that is reconstructed from the noisy observations.

The random variable $\Sigma^x(\{z\})$ (which depends only on the realization of the observation noise once x, z are fixed) determines the probability that we reconstruct z when the true configuration is x . As we have conditioned on the true value of the signal at X_0, X_1^v, \dots, X_k^v for $|v| > m$, we can assume that the true configuration is correctly reconstructed on the boundary $z^r = x^r$ for $r \in \partial J$. We want to show that the probability of correctly reconstructing the true configuration at a given site inside J (e.g., X_k^0 which corresponds to the point $\mathbf{0} := (k, 0) \in J$) remains bounded away from $\frac{1}{2}$ with high probability, no matter how far this site is from the boundary ∂J .

Now suppose that $z^{\mathbf{0}} \neq x^{\mathbf{0}}$, that is, that the configuration at $\mathbf{0}$ is incorrectly decoded. Then there must exist an interface between the set of incorrectly decoded vertices (that contains $\mathbf{0}$) and the set of correctly decoded vertices (that contains the boundary ∂J). This is illustrated in the following figure, where correctly decoded vertices are indicated by white circles and incorrectly decoded vertices are indicated by black circles:



We call the shaded region a *contour*: it has the key property that every edge that crosses its boundary connects an incorrectly decoded site inside with a correctly decoded site outside (it also proves to be convenient to assume contours are simply connected, i.e., that they have no holes; this is why the white vertex is included in the shaded region).

The main idea of the proof is that when the error probability p is close to zero, it is very unlikely that a decoded configuration z contains many edges that connect a correctly decoded site r with an incorrectly decoded site q . Intuitively, if the error probability p is small, then the observations reveal with a high degree of confidence whether x^q and x^r coincide or differ; therefore, if we reconstruct one (in)correctly, then it is very likely that the other will be reconstructed (in)correctly as well. More precisely, the Σ^x -probability of reconstructing a configuration that contains a given contour $J' \subseteq J$ is exponentially small in the length of the boundary of J' (Lemma 3.7 below). It then follows from a simple union bound that the probability that the decoded configuration contains *any* contour that includes $\mathbf{0}$ is bounded away from $\frac{1}{2}$. But this completes the proof, as at least one such contour must be present if $\mathbf{0}$ is incorrectly decoded. (Of course, the crucial insight behind this approach is that the probability of containing a given contour and the number of contours of a given length do not depend on the size of the domain, so that all the estimates that appear in this argument are dimension-free.)

We now proceed to make these ideas precise.

Proposition 3.6. *There exists an absolute constant $0 < p_\star < 1/2$ such that*

$$\mathbf{E}|\mathbf{E}[X_k^0 | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}]| \geq \frac{1}{4}$$

for every $k, m \geq 1$ whenever $0 < p < p_\star$.

Proof. Let us fix $k, m \geq 1$ throughout the proof, and define $\mathbf{0} := (k, 0) \in J$. We will prove below the following claim: there exists an absolute constant $0 < p_\star < 1/2$ such that

$$\mathbf{P} \left[\Sigma^x(\{z : z^{\mathbf{0}} = x^{\mathbf{0}}\}) \geq \frac{3}{4} \right] \geq \frac{1}{2}$$

whenever $0 < p < p_\star$: that is, when the noise is sufficiently small, the conditional distribution $\mathbf{P}[X_k^0 \in \cdot | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}]$ assigns a large probability to the actually realized value of X_k^0 at least half of the time (recall Lemma 3.4). Let us complete the proof assuming this claim. Note that $\Sigma^x(\{z : z^{\mathbf{0}} = x^{\mathbf{0}}\}) \geq 3/4$ implies $|\Sigma^x(\{z : z^{\mathbf{0}} = 1\}) - \Sigma^x(\{z : z^{\mathbf{0}} = -1\})| \geq 1/2$. Thus the above claim implies that

$$\mathbf{P} \left[|\mathbf{E}[X_k^0 | X_0, Y_1, \dots, Y_k, \{X_1^v, \dots, X_k^v : |v| > m\}]| \geq \frac{1}{2} \middle| X_0, \dots, X_k \right] \geq \frac{1}{2},$$

where we have used Lemma 3.4 and the fact that $\{X^q\}$ and $\{\xi^{qr}\}$ are independent. The proof is now completed by a straightforward estimate.

It remains to prove the claim. To this end, we use a Peierls argument. Fix for the time being configurations $x, z \in \{-1, 1\}^J$. For any $J' \subseteq J$, define the boundary edges

$$\mathfrak{E}J' := \{\{q, r\} : q \in J', r \in (J \setminus J') \cup \partial J, \|q - r\| = 1\}.$$

A subset $J' \subseteq J$ is called a *contour* if it is simply connected, $z^q = -x^q$ for all $\{q, r\} \in \mathfrak{E}J'$ with $q \in J'$, and $z^r = x^r$ if in addition $r \in J \setminus J'$. We will denote the set of contours as $\mathfrak{C}_{z,x}$ (note that the definition of a contour depends on the given configurations z and x). If $z^{\mathbf{0}} = -x^{\mathbf{0}}$, then there must exist a contour $J' \in \mathfrak{C}_{z,x}$ such that $\mathbf{0} \in J'$: indeed, construct J' by choosing the maximal connected subset of J such that $\mathbf{0} \in J'$ and $z^q = -x^q$ for all $q \in J'$, and then ‘fill in the holes’ to make J' simply connected. Thus

$$\Sigma^x(\{z : z^{\mathbf{0}} = -x^{\mathbf{0}}\}) \leq \Sigma^x(\{z : \exists J' \in \mathfrak{C}_{z,x}, \mathbf{0} \in J'\}) \leq \sum_{J' \ni \mathbf{0}} \Sigma^x(\{z : J' \in \mathfrak{C}_{z,x}\}).$$

Now note that, by the definition of a contour, $x^q z^q = -1$ whenever $\{q, r\} \in \mathfrak{E}J'$ with $q \in J'$, and $x^q x^r z^q z^r = -1$ if in addition $r \in J \setminus J'$. Thus the existence of a contour implies the presence of many such edges. The basic idea of the proof is that the probability that this occurs is small under Σ^x due to Lemma 3.4. Let us make this precise.

Lemma 3.7. *For any $J' \subseteq J$, we have*

$$\Sigma^x(\{z : J' \in \mathfrak{C}_{z,x}\}) \leq \exp \left(-2\beta \sum_{\{q,r\} \in \mathfrak{E}J'} \xi^{qr} \right).$$

Proof. Assume without loss of generality that J' is simply connected. Let us use for simplicity the convention that $z^r = x^r$ for $r \in \partial J$. Define the events

$$\begin{aligned} A &= \{z : z^q = -x^q \text{ and } z^r = x^r \text{ for } \{q, r\} \in \mathfrak{E}J', q \in J'\}, \\ B &= \{z : z^q = x^q \text{ and } z^r = x^r \text{ for } \{q, r\} \in \mathfrak{E}J', q \in J'\}. \end{aligned}$$

Then we evidently have by Lemma 3.4

$$\Sigma^x(\{z : J' \in \mathfrak{C}_{z,x}\}) = \Sigma^x(A) \leq \frac{\Sigma^x(A)}{\Sigma^x(B)}.$$

An elementary computation shows that

$$\frac{\Sigma^x(A)}{\Sigma^x(B)} = \exp\left(-2\beta \sum_{\{q,r\} \in \mathfrak{E}J'} \xi^{qr}\right) \frac{\sum_z \mathbf{1}_A(z) \exp(\beta \sum_{\{q,r\} \subseteq J': \|q-r\|=1} \xi^{qr} x^q x^r z^q z^r)}{\sum_z \mathbf{1}_B(z) \exp(\beta \sum_{\{q,r\} \subseteq J': \|q-r\|=1} \xi^{qr} x^q x^r z^q z^r)}.$$

But the ratio in this expression is unity, as the exponential term inside the sums is invariant under the transformation $z^q \mapsto -z^q$ for all $q \in J'$. The proof is complete. \square

Lemma 3.7 allows us to estimate

$$\begin{aligned} & \mathbf{P}\left[\Sigma^x(\{z : z^{\mathbf{0}} = -x^{\mathbf{0}}\})\right] \geq \sum_{J' \ni \mathbf{0} \text{ simply connected}} e^{-\beta|\mathfrak{E}J'|} \\ & \leq \mathbf{P}\left[\sum_{J' \ni \mathbf{0} \text{ simp. conn.}} \exp\left(-2\beta \sum_{\{q,r\} \in \mathfrak{E}J'} \xi^{qr}\right) \geq \sum_{J' \ni \mathbf{0} \text{ simp. conn.}} e^{-\beta|\mathfrak{E}J'|}\right] \\ & \leq \mathbf{P}\left[\exists J' \ni \mathbf{0} \text{ simply connected with } \sum_{\{q,r\} \in \mathfrak{E}J'} \xi^{qr} \leq \frac{|\mathfrak{E}J'|}{2}\right] \\ & \leq \sum_{J' \ni \mathbf{0} \text{ simply connected}} \mathbf{P}\left[\sum_{\{q,r\} \in \mathfrak{E}J'} \xi^{qr} \leq \frac{|\mathfrak{E}J'|}{2}\right]. \end{aligned}$$

Using a standard combinatorial result [22, Lemma 6.13]

$$|\{J' \subseteq J \text{ simply connected} : \mathbf{0} \in J', |\mathfrak{E}J'| = l\}| \leq l3^{l-1},$$

as well as the simple bound

$$\mathbf{P}\left[\sum_{\{q,r\} \in \mathfrak{E}J'} \xi^{qr} \leq \frac{|\mathfrak{E}J'|}{2}\right] = \mathbf{P}\left[\text{Bin}(|\mathfrak{E}J'|, 1-p) \leq \frac{3}{4}|\mathfrak{E}J'|\right] \leq 2^{|\mathfrak{E}J'|} p^{|\mathfrak{E}J'|/4},$$

we can conclude that

$$\mathbf{P}\left[\Sigma^x(\{z : z^{\mathbf{0}} = -x^{\mathbf{0}}\}) \geq c_1\right] \leq c_2, \quad c_1 = \sum_{l=3}^{\infty} l3^{l-1} \left(\frac{p}{1-p}\right)^{l/2}, \quad c_2 = \sum_{l=3}^{\infty} l3^{l-1} 2^l p^{l/4}.$$

But we can now evidently choose $p_* > 0$ sufficiently small such that $c_1 \leq 1/4$ and $c_2 \leq 1/2$ whenever $p \leq p_*$, which readily yields the desired estimate. \square

3.3 Proof of Theorem 3.1: high noise

We now turn to proving that the filter is stable when the noise is strong. We begin by noting that it suffices to prove stability of finite-dimensional marginals of the filter.

Lemma 3.8. *Suppose that*

$$\mathbf{E}[\mathbf{E}[f(X_k^{-m}, \dots, X_k^m)|X_0, Y_1, \dots, Y_k] - \mathbf{E}[f(X_k^{-m}, \dots, X_k^m)|Y_1, \dots, Y_k]] \xrightarrow{k \rightarrow \infty} 0$$

for every function f and every $m \geq 1$. Then the filter is stable.

Proof. Fix any measurable subset A of $\{-1, 1\}^{\mathbb{Z}}$ and define

$$F_m = f_m(X_0^{-m}, \dots, X_0^m) := \mathbf{P}[X_0 \in A | X_0^{-m}, \dots, X_0^m].$$

We can estimate

$$\begin{aligned} \mathbf{E}|\mathbf{P}[X_k \in A | X_0, Y_1, \dots, Y_k] - \mathbf{P}[X_k \in A | Y_1, \dots, Y_k]| &\leq 2\mathbf{E}|f_m(X_k^{-m}, \dots, X_k^m) - \mathbf{1}_A(X_k)| \\ &+ \mathbf{E}|\mathbf{E}[f_m(X_k^{-m}, \dots, X_k^m)|X_0, Y_1, \dots, Y_k] - \mathbf{E}[f_m(X_k^{-m}, \dots, X_k^m)|Y_1, \dots, Y_k]|. \end{aligned}$$

By stationarity the first term does not depend on k , and the assumption gives

$$\limsup_{k \rightarrow \infty} \mathbf{E}|\mathbf{P}[X_k \in A | X_0, Y_1, \dots, Y_k] - \mathbf{P}[X_k \in A | Y_1, \dots, Y_k]| \leq 2\mathbf{E}|F_m - \mathbf{1}_A(X_0)|.$$

Letting $m \rightarrow \infty$ and using the martingale convergence theorem concludes the proof. \square

We will in fact prove a much stronger *pathwise* bound than is required by the above lemma. The basic tool we will use for this purpose is the Dobrushin comparison theorem [22, Theorem 8.20], which we state here in a convenient form.

Theorem 3.9 (Dobrushin comparison theorem). *Let μ and ν be probability measures on $\{-1, 1\}^I$ for some countable set I , and choose measurable functions m_i, n_i such that*

$$m_i(X) = \mu(X^i = 1 | \{X^j : j \neq i\}), \quad n_i(X) = \nu(X^i = 1 | \{X^j : j \neq i\}).$$

Define

$$b_i := \sup_x |m_i(x) - n_i(x)|, \quad C_{ji} := \sup_{x, z: x^v = z^v \text{ for } v \neq i} |m_j(x) - m_j(z)|,$$

and assume that

$$\sup_{j \in I} \sum_{i \in I} C_{ji} < 1.$$

Then $D := \sum_{n=0}^{\infty} C^n$ exists (in the sense of matrix algebra), and

$$|\mu(f) - \nu(f)| \leq \sum_{j \in J} \sum_{i \in I} D_{ji} b_i$$

whenever J is a finite set, $f(x)$ depends only on $\{x^j : j \in J\}$, and $0 \leq f \leq 1$.

We will apply this result pathwise to compare the filters with and without conditioning on the initial condition. To this end, we must compute the quantities that arise in the Dobrushin comparison theorem for suitably chosen regular conditional probabilities.

Lemma 3.10. *Fix any version of the regular conditional probabilities*

$$\mu_{X,Y} := \mathbf{P}[X_0, \dots, X_k \in \cdot | X_0, Y_1, \dots, Y_k], \quad \nu_Y := \mathbf{P}[X_0, \dots, X_k \in \cdot | Y_1, \dots, Y_k].$$

Then there is a set A with $\mathbf{P}[(X, Y) \in A] = 1$ such that for every $(x, y) \in A$

$$\mu_{x,y}(X_\ell^v = 1 | \{X_r^w : (r, w) \neq (\ell, v)\}) = \nu_y(X_\ell^v = 1 | \{X_r^w : (r, w) \neq (\ell, v)\}) = \frac{e^{\beta\{\bar{y}_\ell^v X_{\ell-1}^v + \hat{y}_\ell^v X_\ell^{v+1} + \bar{y}_{\ell+1}^v X_{\ell+1}^v + \hat{y}_\ell^{v-1} X_\ell^{v-1}\}}}{e^{\beta\{\bar{y}_\ell^v X_{\ell-1}^v + \hat{y}_\ell^v X_\ell^{v+1} + \bar{y}_{\ell+1}^v X_{\ell+1}^v + \hat{y}_\ell^{v-1} X_\ell^{v-1}\}} + e^{-\beta\{\bar{y}_\ell^v X_{\ell-1}^v + \hat{y}_\ell^v X_\ell^{v+1} + \bar{y}_{\ell+1}^v X_{\ell+1}^v + \hat{y}_\ell^{v-1} X_\ell^{v-1}\}}}$$

for $1 \leq \ell < k$ and $v \in \mathbb{Z}$,

$$\mu_{x,y}(X_k^v = 1 | \{X_r^w : (r, w) \neq (k, v)\}) = \nu_y(X_k^v = 1 | \{X_r^w : (r, w) \neq (k, v)\}) = \frac{e^{\beta\{\bar{y}_k^v X_{k-1}^v + \hat{y}_k^v X_k^{v+1} + \bar{y}_k^{v-1} X_k^{v-1}\}}}{e^{\beta\{\bar{y}_k^v X_{k-1}^v + \hat{y}_k^v X_k^{v+1} + \bar{y}_k^{v-1} X_k^{v-1}\}} + e^{-\beta\{\bar{y}_k^v X_{k-1}^v + \hat{y}_k^v X_k^{v+1} + \bar{y}_k^{v-1} X_k^{v-1}\}}}$$

for $v \in \mathbb{Z}$, and $\mu_{x,y}(X_0^v = 1) = \mathbf{1}_{x_0^v=1}$ for $v \in \mathbb{Z}$, where $\beta := \log \sqrt{(1-p)/p}$.

Proof. It is an elementary fact that

$$\begin{aligned} \mu_{X,Y}(X_\ell^v = 1 | \{X_r^w : (r, w) \neq (\ell, v)\}) &= \mathbf{P}[X_\ell^v = 1 | X_0, Y_1, \dots, Y_k, \{X_r^w : (r, w) \neq (\ell, v)\}], \\ \nu_Y(X_\ell^v = 1 | \{X_r^w : (r, w) \neq (\ell, v)\}) &= \mathbf{P}[X_\ell^v = 1 | Y_1, \dots, Y_k, \{X_r^w : (r, w) \neq (\ell, v)\}], \end{aligned}$$

see [53, p. 95–96] or [46, Lemma 3.4]. That each statement in the Lemma holds for \mathbf{P} -a.e. (x, y) can therefore be read off from Lemma 3.4. As there are countably many statements, they can be assumed to hold simultaneously on a set A of unit measure. \square

We can now complete the proof of filter stability for $p > p^*$.

Proposition 3.11. *There exists an absolute constant $0 < p^* < 1/2$ such that*

$$|\mathbf{E}[f(X_k^{-m}, \dots, X_k^m) | X_0, Y_1, \dots, Y_k] - \mathbf{E}[f(X_k^{-m}, \dots, X_k^m) | Y_1, \dots, Y_k]| \leq (8m+4)\|f\|_\infty e^{-k}$$

a.s. for every $k, m \geq 1$ and function f whenever $p^* < p \leq 1/2$.

Proof. We apply Theorem 3.9 with $I = \{0, \dots, k\} \times \mathbb{Z}$ and $\mu = \mu_{x,y}$, $\nu = \nu_y$ as defined in Lemma 3.10. Evidently $b_{(0,v)} \leq 1$ and $b_{(\ell,v)} = 0$ for $1 \leq \ell \leq k$ and $v \in \mathbb{Z}$, so we have

$$|\mu_{x,y}(f(X_k^{-m}, \dots, X_k^m)) - \nu_y(f(X_k^{-m}, \dots, X_k^m))| \leq 2\|f\|_\infty \sum_{w=-m}^m \sum_{v \in \mathbb{Z}} D_{(k,w)(0,v)}$$

by Theorem 3.9 provided that the condition on the matrix C is satisfied.

We proceed to estimate the matrix C using Lemma 3.10. Evidently

$$C_{(\ell',v')(\ell,v)} = 0 \quad \text{if } \ell' = 0 \quad \text{or} \quad |\ell' - \ell| + |v' - v| > 1 \quad \text{or} \quad \ell = \ell', v = v'.$$

On the other hand, note that by Lemma 3.10

$$\frac{e^{-4\beta}}{e^{4\beta} + e^{-4\beta}} \leq \mu_{x,y}(X_\ell^v = 1 | \{X_r^w : (r, w) \neq (\ell, v)\}) \leq \frac{e^{4\beta}}{e^{4\beta} + e^{-4\beta}},$$

so we can estimate

$$C_{ji} \leq \tanh(4\beta) < 1 \quad \text{for all } i, j \in I.$$

It follows readily that

$$\|C\|_* := \sup_{j \in I} \sum_{i \in I} e^{\|j-i\|} C_{ji} \leq 4e \tanh(4\beta).$$

We can now evidently choose $0 < p^* < 1/2$ such that $4e \tanh(4\beta) < 1/2$ for $p^* < p \leq 1/2$. Then the condition of Theorem 3.9 is satisfied. Moreover, as $\|\cdot\|_*$ is a matrix norm

$$\|D\|_* \leq \sum_{n=0}^{\infty} \|C\|_*^n \leq 2.$$

Thus we obtain

$$\begin{aligned} & |\mu_{x,y}(f(X_k^{-m}, \dots, X_k^m)) - \nu_y(f(X_k^{-m}, \dots, X_k^m))| \\ & \leq (4m+2) \|f\|_{\infty} e^{-k} \max_{w=-m, \dots, m} \sum_{v \in \mathbb{Z}} e^{\|(k,w)-(0,v)\|} D_{(k,w)(0,v)} \\ & \leq (4m+2) \|D\|_* \|f\|_{\infty} e^{-k} \leq (8m+4) \|f\|_{\infty} e^{-k}. \end{aligned}$$

As our estimates are valid for \mathbf{P} -a.e. (x, y) , the proof is complete. \square

Remark 3.12. It is natural to conjecture that one can choose $p_* = p^*$ in Theorem 3.1. While we certainly believe this to be true, we were not able to prove this fact using standard methods. The difficulty can be seen in Lemma 3.4, as we presently explain.

Lemma 3.4 shows that the conditional distribution of X_1, \dots, X_n given Y_1, \dots, Y_n can be viewed as an Ising model in the spin variables $\sigma^q := x^q z^q$ with independent random interactions ξ^{qr} . An Ising model is called ferromagnetic if all the interactions are positive. In the ferromagnetic case, it is standard to establish the existence of a unique phase transition point by monotonicity arguments [22, p. 100]. Unfortunately, while our model is ‘ferromagnetic on average’ as $\mathbf{P}[\xi^{qr} = 1] > \mathbf{P}[\xi^{qr} = -1]$, there are always infinitely many interactions of either sign. Thus correlation inequalities cannot be used, and in their absence it is not clear how to prove the existence of a simple phase boundary. Monotonicity arguments will play a central role in section 5.3 below to rule out conditional phase transitions in a somewhat different context.

While we have made no attempt to optimize the estimates for p_* and p^* that can be extracted from the proof of Theorem 3.1, the methods used here are not expected to yield realistic values of these constants. On the other hand, the strong control provided by the Dobrushin comparison theorem has proved to be a powerful tool for the quantitative analysis of filtering algorithms in high dimension, cf. [40, 41].

4 Symmetry breaking and observability

Theorem 3.1 shows that inheritance of ergodicity under conditioning cannot be taken for granted in infinite dimension even when the model is locally nondegenerate. Are such phenomena prevalent in infinite dimension, or are they restricted to some carefully constructed examples? We would like to understand in what situations such phenomena can be ruled out, both from the mathematical perspective and in view of the importance of filter

stability (as well as spatial decay of correlations in infinite dimension) for the performance of practical filtering algorithms [40].

It is not difficult to understand the mechanism that causes the filter to be unstable in Theorem 3.1. In this model, the observations possess a global symmetry: the conditional law of Y is unchanged under the transformation $X \mapsto -X$. This symmetry renders the filter trivially unstable in the absence of observation noise, in precise analogy with Example 2.1. In the finite-dimensional case, however, Theorem 2.2 shows that the addition of any observation noise suffices to ensure that ergodicity of the underlying model is not broken by the additional symmetry introduced by conditioning. The surprise in infinite dimension is that the qualitative effect of the added symmetry still persists in the presence of observation noise. Thus local nondegeneracy in itself does not suffice to ensure the inheritance of ergodicity under conditioning.

On the other hand, the phenomenon exhibited in Theorem 3.1 evidently cannot arise in models that do not possess observation symmetries. It seems natural to conjecture that the presence of such symmetries is the *only* possible obstruction to inheritance of ergodicity under conditioning: that is, inheritance of ergodicity is ensured once observation symmetries are ruled out. It is not entirely obvious, however, how such a principle can be rigorously formulated. On the other hand, even in the absence of a general definition, this intuitive notion should certainly be satisfied in many elementary observation models. For example, let us state the following simple conjecture, which encapsulates the essence of the above intuition in the simplest possible setting.

Conjecture 4.1. *Let $(X_k, Y_k)_{k \in \mathbb{Z}}$ be a stationary infinite-dimensional hidden Markov model as in section 2.2 with $X_k \in \{-1, 1\}^{\mathbb{Z}}$ and with $Y_k \in \{-1, 1\}^{\mathbb{Z}}$ of the form*

$$Y_k^v = X_k^v \xi_k^v, \quad (\xi_k^v)_{k,v \in \mathbb{Z}} \text{ are i.i.d. } \perp X \text{ with } \mathbf{P}[\xi_k^v = -1] = p.$$

If the underlying process $(X_k)_{k \in \mathbb{Z}}$ is stable, then the filter is stable.

The idea behind this conjecture is that the direct observation structure $Y_k^v = X_k^v \xi_k^v$ is evidently devoid of symmetries for any $p \neq \frac{1}{2}$: every configuration $x \in \{-1, 1\}^{\mathbb{Z}}$ gives rise to a distinct observation law $\mathbf{P}[Y_k \in \cdot | X_k = x]$ (the case $p = \frac{1}{2}$ is trivial as then $Y \perp X$; we will therefore assume $p \neq \frac{1}{2}$ in the sequel). Thus any mechanism of the type exhibited by Theorem 3.1 is ruled out, and it seems hard to imagine another mechanism by which ergodicity of the underlying process could be obstructed due to conditioning on such informative observations. Despite the seemingly obvious nature of this conjecture, we were not able to prove such a result in a general setting.

The idea that stability of the filter is related to the absence of symmetries is not new in the infinite-dimensional setting. It arises already in classical filtering models for a somewhat different reason: it may happen that the filter is stable even when the underlying model is *not* ergodic. In such situations, stability properties can emerge under the conditional distribution due to the informative nature of the observations; in essence, the filter will ‘forget’ its initial distribution as the information contained therein is superseded by the information in the observations.³ This phenomenon was made precise in the papers [48, 49, 51]. While the theory developed in these papers is closely related to the symmetry breaking

³This is not the only mechanism that gives rise to stability of nonlinear filters in classical nonergodic models; it is also possible to exploit the local ergodicity of the model, cf. [15] and the references therein. However, this approach is unrelated to the symmetric breaking mechanism that is developed in this section.

properties that we aim to exploit here, these results are not satisfactory in infinite dimension as will be explained below.

In this section, we will aim to extend such observability arguments to *translation-invariant* systems in infinite dimension by exploiting a technique from multidimensional ergodic theory [10]. Somewhat surprisingly, the problem proves to be more tractable in the continuous-time setting, for which we will establish validity of the natural analogue of Conjecture 4.1. In its original discrete time formulation, however, our ultimate result falls short of establishing Conjecture 4.1 even for translation-invariant models. Nonetheless, the theory developed here provides one possible mechanism for symmetry breaking in conditional ergodic theory. An entirely different mechanism will be discussed in the context of conditional random fields in section 5.3 below.

The remainder of this section is organized as follows. In section 4.1 we will recall some basic ideas from the general observability theory developed in [48, 49, 51], and explain why these results are not satisfactory in infinite dimension. We also outline a slightly stronger entropic formulation that will provide the basis for a partial extension to translation-invariant systems in infinite dimension, which is developed in section 4.2. In section 4.3 we consider the continuous-time analogue of Conjecture 4.1, and provide a complete proof in this setting for the translation-invariant case.

4.1 General observability theory

Let $(X_k, Y_k)_{k \geq 0}$ be a general hidden Markov model as in section 2.1, and denote by \mathbf{P}^μ the law of the model with initial distribution $X_0 \sim \mu$. Define the *prediction filter*

$$\tilde{\pi}_k^\mu := \mathbf{P}^\mu[X_k \in \cdot | Y_1, \dots, Y_{k-1}]$$

(note that we are conditioning the state at time k only on observations prior to time k). The basic observation behind the observability theory of [48, 49, 51] is as follows.

Theorem 4.2 ([49, 51]). *Suppose that X_k takes values in a compact metric space and that $(X_k, Y_k)_{k \geq 0}$ is Feller. Suppose that the following observability assumption holds:*

$$\mathbf{P}^\mu[(Y_k)_{k \geq 0} \in \cdot] = \mathbf{P}^\nu[(Y_k)_{k \geq 0} \in \cdot] \quad \text{if and only if} \quad \mu = \nu.$$

Then we have for every bounded and continuous function f

$$|\tilde{\pi}_k^\mu(f) - \tilde{\pi}_k^\nu(f)| \xrightarrow{k \rightarrow \infty} 0 \quad \mathbf{P}^\mu\text{-a.s.}$$

whenever $\mathbf{P}^\mu[(Y_k)_{k \geq 0} \in \cdot] \ll \mathbf{P}^\nu[(Y_k)_{k \geq 0} \in \cdot]$.

Note that no ergodicity or nondegeneracy assumptions are imposed in this result: the only assumption is that of observability, that is, that distinct initial laws give rise to distinct observation laws. The latter can evidently be viewed as a very mild type of symmetry breaking assumption. A more general form of the result that does not require compactness or Feller assumptions can be found in [51].

There is nothing in Theorem 4.2 that prohibits its application in infinite-dimensional systems. In fact, it is readily verified that the observability assumption of Theorem 4.2 is automatically satisfied in the setting of Conjecture 4.1 (provided $p \neq \frac{1}{2}$). Nonetheless, unlike in most finite-dimensional problems, the conclusion of Theorem 4.2 is not satisfactory in the infinite-dimensional setting, as we will presently explain.

There are two ways in which the conclusion of Theorem 4.2 falls short of the filter stability property as it was introduced in section 2.1. On the one hand, the conclusion applies to the prediction filter $\tilde{\pi}_k$ rather than to the filter π_k . This is, however, not a major issue: one could argue that the prediction filter is of equal interest in practice as the filter itself, and thus we would be quite content in first instance to resolve a variant of Conjecture 4.1 for the prediction filter.⁴ Much more serious, however, is the fact that stability is only obtained for initial measures μ, ν that give rise to *absolutely continuous* observation laws. If the state space of X_k is countable, this is not a restriction: then $\delta_x \ll \lambda$ whenever $\lambda(\{x\}) > 0$, so choosing $\mu = \delta_{X_0}$ and $\nu = \lambda$ in Theorem 4.2 yields

$$\|\mathbf{E}[f(X_k)|X_0, Y_1, \dots, Y_{k-1}] - \mathbf{E}[f(X_k)|Y_1, \dots, Y_{k-1}]\| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1,$$

the natural counterpart of the stability property of section 2.1 for the prediction filter. In a continuous state space, this argument does not work, as a point mass is singular with respect to any nonatomic measure. Nonetheless, this proves to be a minor problem in most finite-dimensional models: when the hidden Markov model possesses transition and observation densities, one can generally deduce absolute continuity of the observation laws $\mathbf{P}^\mu[(Y_k)_{k \geq 0} \in \cdot] \ll \mathbf{P}^\nu[(Y_k)_{k \geq 0} \in \cdot]$ even when μ and ν are mutually singular. This is discussed in detail in [13, section 11.5]. In infinite dimension, on the other hand, the existence of transition and observation densities is out of the question, and thus the conclusion of Theorem 4.2 is severely limited. In particular, there is no hope to deduce the very natural filter stability property as introduced in section 2.1, or its prediction filter counterpart, by using the result or the method of proof of Theorem 4.2.

Remark 4.3. While the result of Theorem 4.2 falls short of the natural stability property as introduced in section 2.1, it does ensure a minimal form of stability for absolutely continuous initial conditions in the setting of Conjecture 4.1. Unfortunately, in infinite dimension, absolutely continuous measures are of little practical interest. Indeed, if $\mu \ll \nu$ are absolutely continuous probability measures on an infinite product space $E^{\mathbb{Z}}$, then the density $\frac{d\mu}{d\nu}$ can be approximated arbitrarily well in $L^1(\nu)$ by densities that depend on a finite number of coordinates. This implies that absolutely continuous probability measures in infinite dimension differ effectively only on a finite number of coordinates (indeed, if $\frac{d\mu}{d\nu}$ is a function of $\{x^v\}_{v \in I}$ for a finite set $I \subset \mathbb{Z}$, then $\mu(\{x^v\}_{v \notin I} \in \cdot | \{x^v\}_{v \in I}) = \nu(\{x^v\}_{v \notin I} \in \cdot | \{x^v\}_{v \in I})$ by the Bayes formula, and thus the difference between μ and ν is entirely determined by the marginal on I). Thus the stability property for absolutely continuous initial measures does not capture the dissipation of an initial error in infinitely many coordinates, and is therefore of little relevance to the analysis of any reasonable approximation method that might arise in practice.

The proof of Theorem 4.2 in [49, 51] relies on a classical martingale argument due to Blackwell and Dubins [2]. Instead of explaining this approach, let us present an alternative proof of stability, using the notion of entropy, in a special case: the finite-dimensional counterpart of Conjecture 4.1. The key mechanism behind the proof is equivalent to that of [49, 51]. Nonetheless, the entropic formulation will be crucial to developing a nontrivial infinite-dimensional extension in section 4.2 below.

⁴In finite dimension, stability of the filter can often be deduced from that of the prediction filter and vice versa (e.g., [48, 46]), so that the conditional ergodic theory is not expected to be substantially different in these two formulations. The situation in infinite dimension is not entirely clear, however, cf. section 4.2 below.

Proposition 4.4. Let $(X_k, Y_k)_{k \in \mathbb{Z}}$ be a hidden Markov model as in section 2.1 with $X_k \in \{-1, 1\}^r$ (where $r < \infty$) and with observations $Y_k \in \{-1, 1\}^r$ of the form

$$Y_k^v = X_k^v \xi_k^v, \quad (\xi_k^v)_{k \in \mathbb{Z}, v \in \{1, \dots, r\}} \text{ are i.i.d. } \perp X \text{ with } \mathbf{P}[\xi_k^v = -1] = p.$$

Then the prediction filter is stable in the sense that

$$|\mathbf{P}[X_k \in A | X_0, Y_1, \dots, Y_{k-1}] - \mathbf{P}[X_k \in A | Y_1, \dots, Y_{k-1}]| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1$$

for every set A , provided that $p \neq \frac{1}{2}$.

Proof. We use standard ideas from information theory [12]. Let $H(X)$ denote entropy and $H(X|Y)$ denote conditional entropy of discrete random variables X, Y . Then

$$\begin{aligned} & \sum_{k=1}^n \{H(Y_k | Y_1, \dots, Y_{k-1}) - H(Y_k | X_0, Y_1, \dots, Y_{k-1})\} \\ &= H(Y_1, \dots, Y_n) - H(Y_1, \dots, Y_n | X_0) \\ &= H(X_0) - H(X_0 | Y_1, \dots, Y_n) \leq H(X_0), \end{aligned}$$

where we have used the chain rule for entropy and symmetry of mutual information $H(X) - H(X|Y) = H(Y) - H(Y|X) =: I(X; Y)$. As $H(X_0) \leq r$ is independent of n ,

$$H(Y_k | Y_1, \dots, Y_{k-1}) - H(Y_k | X_0, Y_1, \dots, Y_{k-1}) \xrightarrow{k \rightarrow \infty} 0.$$

But note that

$$\begin{aligned} & H(Y_k | Y_1, \dots, Y_{k-1}) - H(Y_k | X_0, Y_1, \dots, Y_{k-1}) = \\ & \mathbf{E} \left[\sum_{y \in \{-1, 1\}^r} \mathbf{P}[Y_k = y | X_0, Y_1, \dots, Y_{k-1}] \log_2 \frac{\mathbf{P}[Y_k = y | X_0, Y_1, \dots, Y_{k-1}]}{\mathbf{P}[Y_k = y | Y_1, \dots, Y_{k-1}]} \right], \end{aligned}$$

which is precisely the expected relative entropy between the conditional distributions $\mathbf{P}[Y_k \in \cdot | X_0, Y_1, \dots, Y_{k-1}]$ and $\mathbf{P}[Y_k \in \cdot | Y_1, \dots, Y_{k-1}]$. Thus by Pinsker's inequality

$$|\mathbf{E}[g(Y_k) | X_0, Y_1, \dots, Y_{k-1}] - \mathbf{E}[g(Y_k) | Y_1, \dots, Y_{k-1}]| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1$$

for every function g . Now note that, by the conditional independence structure of the observations, we have $\mathbf{E}[g(Y_k) | X_0, Y_1, \dots, Y_{k-1}, X_k] = \mathbf{E}[g(Y_k) | Y_1, \dots, Y_{k-1}, X_k] = \mathbf{E}[g(Y_k) | X_k]$. Thus we can write using the tower property

$$\begin{aligned} \mathbf{E}[g(Y_k) | X_0, Y_1, \dots, Y_{k-1}] &= \mathbf{E}[f(X_k) | X_0, Y_1, \dots, Y_{k-1}], \\ \mathbf{E}[g(Y_k) | Y_1, \dots, Y_{k-1}] &= \mathbf{E}[f(X_k) | Y_1, \dots, Y_{k-1}], \end{aligned}$$

where

$$f(x) = \mathbf{E}[g(Y_k) | X_k = x] = (T_1 \cdots T_r g)(x), \quad (T_i g)(x) := (1-p)g(x) + pg(x^{-i})$$

and $x^{-i} := (x^1, \dots, x^{i-1}, -x^i, x^{i+1}, \dots, x^r)$. But it is readily seen that if $p \neq \frac{1}{2}$, then each operator T_i is invertible. Thus every function f is of the form $f(x) = \mathbf{E}[g(Y_k) | X_k = x]$ for some function g , and the proof is evidently complete. \square

Let us emphasize the two key ideas of the proof. We first establish that the conditional distribution of the *next observation* Y_k given the observation history Y_1, \dots, Y_{k-1} ‘forgets’ the initial condition X_0 as $k \rightarrow \infty$. That is, unlike the prediction filter which aims to predict the *next hidden state* X_k , stability of predicting the next observation holds regardless of any properties of the model: it is a simple consequence of the finite entropy property $H(X_0) < \infty$ of the initial condition (this fails in the continuous case, and for this reason the absolute continuity assumption $\mathbf{P}^\mu[(Y_k)_{k \geq 0} \in \cdot] \ll \mathbf{P}^\nu[(Y_k)_{k \geq 0} \in \cdot]$ is essential in the general setting of Theorem 4.2). However, due to the absence of symmetry in the observations, predicting the next observation Y_k is *equivalent* to predicting the hidden state X_k . It is here that observability enters the picture.

4.2 Translation-invariant systems

The proof of Proposition 4.4 relies crucially on the fact that X_k and Y_k take values in a finite state space $\{-1, 1\}^r$ with $r < \infty$, so that the entropies $H(Y_1, \dots, Y_n)$ and $H(X_0)$ are finite. In the infinite-dimensional setting of Conjecture 4.1 where X_k and Y_k take values in $\{-1, 1\}^{\mathbb{Z}}$, this is no longer the case. Nonetheless, entropy arguments are ubiquitous in the study of infinite-dimensional systems that are *translation-invariant*: in such systems, the role of entropy is replaced by that of the entropy rate (or ‘specific entropy’), which plays an central role in the ergodic theory of measurable dynamical systems, cf. [24, Part 2], and in statistical mechanics where it gives rise to thermodynamic formalism, cf. [22, Chapters 15–16]. The aim of this section is to develop an infinite-dimensional counterpart to Proposition 4.4 using this formalism.

Let $(X_k, Y_k)_{k \in \mathbb{Z}}$ be a stationary infinite-dimensional hidden Markov model as in section 2.2 with $X_k \in \{-1, 1\}^{\mathbb{Z}}$ and with $Y_k \in \{-1, 1\}^{\mathbb{Z}}$ of the form

$$Y_k^v = X_k^v \xi_k^v, \quad (\xi_k^v)_{k, v \in \mathbb{Z}} \text{ are i.i.d. } \perp X \text{ with } \mathbf{P}[\xi_k^v = -1] = p.$$

This model is the same as in Conjecture 4.1. In addition, however, we will assume throughout this section that the model is translation-invariant: that is, that the local transition kernels $P^v(x, dz) = P^w(x, dz)$ for all $v, w \in \mathbb{Z}$ (cf. section 2.2), and that

$$(X_k^v, Y_k^v)_{k, v \in \mathbb{Z}} \stackrel{\text{law}}{=} (X_{k+\ell}^{v+w}, Y_{k+\ell}^{v+w})_{k, v \in \mathbb{Z}} \quad \text{for all } \ell, w \in \mathbb{Z}.$$

We now state the main result of this section.

Theorem 4.5. *For the model of this section, if $p \neq \frac{1}{2}$, we have*

$$|\mathbf{P}[X_k \in A | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}] - \mathbf{P}[X_k \in A | Y_1, \dots, Y_{k-1}, Y_k^{<v}]| \xrightarrow{k \rightarrow \infty} 0 \quad \text{in } L^1$$

for every measurable set A and $v \in \mathbb{Z}$, where $Y_k^{<v} := \{Y_k^w : w < v\}$.

Theorem 4.5 establishes stability of a quantity

$$\mathbf{P}[X_k \in A | Y_1, \dots, Y_{k-1}, Y_k^{<v}]$$

that is intermediate between the prediction filter (section 4.1)

$$\mathbf{P}[X_k \in A | Y_1, \dots, Y_{k-1}]$$

and the filter (section 2.1)

$$\mathbf{P}[X_k \in A | Y_1, \dots, Y_k].$$

It will become evident in the proof that the intermediate filter is the natural infinite-dimensional object that arises in extending the observability theory of section 4.1 to translation-invariant systems. On the other hand, from the practical point of view, the intermediate filter is a somewhat strange object. Unlike the prediction filter, which arises when predicting the next state given the observations to date, and the filter, which arises when tracking the current state given the observation history, the intermediate filter uses an unnatural subset of the observations in the current time step. In particular, Theorem 4.5 falls short of establishing either Conjecture 4.1 or its natural prediction filter counterpart. Nonetheless, Theorem 4.5 could provide a step towards establishing Conjecture 4.1 for translation-invariant models (cf. Remark 4.10 below). Moreover, in section 4.3 below, we will see that this shortcoming of Theorem 4.5 can be resolved for the natural continuous-time analogue of Conjecture 4.1.

Remark 4.6. Note that Theorem 4.5 does not impose any ergodicity assumption on the underlying hidden Markov model: only observability was used to establish the result. In this sense, this result goes beyond the spirit of Conjecture 4.1, which states that ergodicity of the underlying model is inherited by the filter in models with informative observations. Indeed, the result of Theorem 4.5 cannot be interpreted as establishing the inheritance of ergodicity, as ergodicity plays no role in the argument; rather, the intermediate filter is rendered stable here entirely due to the absence of observation symmetries, even when the underlying model is not ergodic.

One might therefore expect that ergodicity should play no role in Conjecture 4.1 either: after all, the mechanism by which we are exploiting the absence of observation symmetries appears to be independent of the ergodic properties of the model. However, it is possible that ergodicity must nonetheless enter the picture in order to extend the conclusion of Theorem 4.5 from the intermediate filter to the filter, so that neither ergodicity nor observability suffices by itself to ensure stability of the filter in infinite dimensional models. Some evidence for this possibility will be discussed in section 6.3. On the other hand, the continuous-time results of section 4.3 below could be viewed as evidence to the contrary. New ideas appear to be needed to resolve these questions.

We now turn to the proof of Theorem 4.5. To this end we begin by proving the following result, which replaces the key step in the proof of Proposition 4.4.

Proposition 4.7. *For the model of this section, we have*

$$H(Y_k^v | Y_1, \dots, Y_{k-1}, Y_k^{<v}) - H(Y_k^v | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}) \xrightarrow{k \rightarrow \infty} 0$$

for every $v \in \mathbb{Z}$.

Remark 4.8. Consider a stationary stochastic process $(Z_k)_{k \in \mathbb{Z}}$ such that Z_k takes values in a finite set. The entropy rate $h(Z)$ of the process can be expressed as

$$h(Z) := \lim_{n \rightarrow \infty} \frac{H(Z_1, \dots, Z_n)}{n} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n H(Z_k | Z_1, \dots, Z_{k-1}) = H(Z_1 | Z_0, Z_{-1}, \dots),$$

where we used the chain rule and stationarity, respectively. This can be used to derive nontrivial identities for conditional entropies. Of particular importance in the present context is that if $Z_k = (X_k, Y_k)$ (still taking values in a finite set), then [24, Lemma 18.2]

$$\lim_{n \rightarrow \infty} H(Y_1 | Y_0, Y_{-1}, \dots; X_{-n}, X_{-n-1}, \dots) = H(Y_1 | Y_0, Y_{-1}, \dots).$$

This result, which could be employed directly in Proposition 4.4 instead of the simpler argument given there, arises in the proof of the Rokhlin-Sinai characterization of Kolmogorov automorphisms in ergodic theory, cf. [24, Chapter 18].

The key to the proof of Proposition 4.7 is that the above ideas admit a multidimensional extension. Let $(Z^q)_{q \in \mathbb{Z}^2}$ be a translation-invariant random field such that Z^q takes values in a finite set. In this setting, the entropy rate (or ‘specific entropy’) $h(Z)$ can be expressed in terms of the lexicographic order \prec on \mathbb{Z}^2 [18, 10]:

$$h(Z) := \lim_{n \rightarrow \infty} \frac{H(Z^{B_n})}{|B_n|} = H(Z^q | Z^{\prec q}),$$

where B_n is the centered box in \mathbb{Z}^2 with radius n and $Z^{B_n} = \{Z^q : q \in B_n\}$, $Z^{\prec q} = \{Z^u : u \prec q\}$. The random field analogue to the above entropy identity was obtained by Conze [10, eq. (20)]: if $Z^q = (X_k^v, Y_k^v)$ for $q = (k, v) \in \mathbb{Z}^2$ take values in a finite set, then

$$\lim_{n \rightarrow \infty} H(Y_1^v | Y_1^{\prec v}, Y_0, Y_{-1}, \dots; X_{-n}^{\prec v}, X_{-n-1}, X_{-n-2}, \dots) = H(Y_1^v | Y_1^{\prec v}, Y_0, Y_{-1}, \dots)$$

for any $v \in \mathbb{Z}$. The proof of Proposition 4.7 follows from this identity.

Remark 4.9. By arguing as in the previous remark, the result of Proposition 4.7 can be rewritten in terms of *conditional entropy rates*. If we define

$$\begin{aligned} h(Y_k | Y_1, \dots, Y_{k-1}) &:= \lim_{n \rightarrow \infty} \frac{H(Y_k^1, \dots, Y_k^n | Y_1, \dots, Y_{k-1})}{n}, \\ h(Y_k | X_0, Y_1, \dots, Y_{k-1}) &:= \lim_{n \rightarrow \infty} \frac{H(Y_k^1, \dots, Y_k^n | X_0, Y_1, \dots, Y_{k-1})}{n}, \end{aligned}$$

then the conclusion of Proposition 4.7 can be expressed as follows:

$$h(Y_k | Y_1, \dots, Y_{k-1}) - h(Y_k | X_0, Y_1, \dots, Y_{k-1}) \xrightarrow{k \rightarrow \infty} 0.$$

This could be interpreted as a direct extension of the key step in the proof of Proposition 4.4 to translation-invariant systems in infinite dimension; the finite-dimensional notion of entropy is simply replaced by its infinite-dimensional counterpart, the entropy rate.

Proof of Proposition 4.7. By stationarity, it suffices to show that

$$H(Y_0^v | Y_{-k+1}, \dots, Y_{-1}, Y_0^{\prec v}) - H(Y_0^v | X_{-k}, Y_{-k+1}, \dots, Y_{-1}, Y_0^{\prec v}) \xrightarrow{k \rightarrow \infty} 0.$$

First, we note that

$$\begin{aligned} \lim_{k \rightarrow \infty} H(Y_0^v | Y_{-k+1}, \dots, Y_{-1}, Y_0^{\prec v}) &= H(Y_0^v | Y_0^{\prec v}, Y_{-1}, Y_{-2}, \dots) \\ &= \lim_{k \rightarrow \infty} H(Y_0^v | Y_0^{\prec v}, Y_{-1}, \dots; X_{-k}^{\prec v}, X_{-k-1}, \dots), \end{aligned}$$

where we have used the identity of Conze [10, eq. (20)] (cf. Remark 4.8 above). As

$$\begin{aligned} H(Y_0^v | Y_0^{\prec v}, Y_{-1}, \dots; X_{-k+1}^{\prec v}, X_{-k}, \dots) &\leq H(Y_0^v | Y_0^{\prec v}, Y_{-1}, \dots; X_{-k}, X_{-k-1}, \dots) \\ &\leq H(Y_0^v | Y_0^{\prec v}, Y_{-1}, \dots; X_{-k}^{\prec v}, X_{-k-1}, \dots), \end{aligned}$$

we obtain

$$\lim_{k \rightarrow \infty} H(Y_0^v | Y_0^{<v}, Y_{-1}, \dots; X_{-k}^{<v}, X_{-k-1}, \dots) = \lim_{k \rightarrow \infty} H(Y_0^v | Y_0^{<v}, Y_{-1}, \dots; X_{-k}, X_{-k-1}, \dots).$$

But by the hidden Markov model structure, $\{Y_{-k}, Y_{-k-1}, \dots; X_{-k-1}, X_{-k-2}, \dots\}$ and $\{Y_0, Y_{-1}, \dots, Y_{-k+1}\}$ are conditionally independent given X_{-k} . Thus

$$H(Y_0^v | Y_0^{<v}, Y_{-1}, \dots; X_{-k}, X_{-k-1}, \dots) = H(Y_0^v | Y_0^{<v}, Y_{-1}, \dots, Y_{-k+1}, X_{-k}),$$

and the proof is complete. \square

Proposition 4.7 concerns the stability of prediction of the next observation. As in the proof of Proposition 4.4, we will transform such properties into stability properties of the filter by using the informative nature of the observations.

Proof of Theorem 4.5. By translation-invariance and Lemma 3.8, it suffices to show that

$$\mathbf{P}[(X_k^1, \dots, X_k^m) \in B | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}] - \mathbf{P}[(X_k^1, \dots, X_k^m) \in B | Y_1, \dots, Y_{k-1}, Y_k^{<v}] \rightarrow 0$$

in L^1 as $k \rightarrow \infty$ for every set B , $m \geq 1$ and $v \in \mathbb{Z}$.

Suppose first that $v \leq 1$. By the chain rule for entropy and Proposition 4.7, we have

$$\begin{aligned} H(Y_k^v, \dots, Y_k^m | Y_1, \dots, Y_{k-1}, Y_k^{<v}) - H(Y_k^v, \dots, Y_k^m | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}) = \\ \sum_{\ell=v}^m \{H(Y_k^\ell | Y_1, \dots, Y_{k-1}, Y_k^{<v}) - H(Y_k^\ell | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v})\} \xrightarrow{k \rightarrow \infty} 0. \end{aligned}$$

Following verbatim the second part of the proof of Proposition 4.4 yields

$$\mathbf{P}[(X_k^v, \dots, X_k^m) \in C | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}] - \mathbf{P}[(X_k^v, \dots, X_k^m) \in C | Y_1, \dots, Y_{k-1}, Y_k^{<v}] \rightarrow 0$$

in L^1 as $k \rightarrow \infty$ for every set C , and thus the result follows.

Now suppose that $v > 1$. We can assume without loss of generality that $v \leq m$ (otherwise the conclusion follows from the result for $m = v$). We may also assume that $0 < p < 1$ (otherwise the conclusion is trivial). By the Bayes formula,

$$\begin{aligned} \mathbf{P}[(X_k^1, \dots, X_k^m) = x | Y_1, \dots, Y_{k-1}, Y_k^{<v}] = \\ \frac{\mathbf{P}[(X_k^1, \dots, X_k^m) = x | Y_1, \dots, Y_{k-1}, Y_k^{<1}] \prod_{i=1}^v g(x^i, Y_k^i)}{\sum_{z \in \{-1, 1\}^m} \mathbf{P}[(X_k^1, \dots, X_k^m) = z | Y_1, \dots, Y_{k-1}, Y_k^{<1}] \prod_{i=1}^v g(z^i, Y_k^i)} \end{aligned}$$

where $g(u, y) = \mathbf{P}[Y_k^i = y | X_k^i = u]$, and $\mathbf{P}[(X_k^1, \dots, X_k^m) = x | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}]$ satisfies the analogous expression. As $0 < \inf g \leq \sup g < \infty$, it follows readily that

$$\mathbf{P}[(X_k^1, \dots, X_k^m) = x | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<1}] - \mathbf{P}[(X_k^1, \dots, X_k^m) = x | Y_1, \dots, Y_{k-1}, Y_k^{<1}] \rightarrow 0$$

for all x implies

$$\mathbf{P}[(X_k^1, \dots, X_k^m) = x | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}] - \mathbf{P}[(X_k^1, \dots, X_k^m) = x | Y_1, \dots, Y_{k-1}, Y_k^{<v}] \rightarrow 0$$

for all x , and thus the proof is complete. \square

Remark 4.10. It is evident that the intermediate filter reduces to the filter if we let $v \rightarrow \infty$; thus Conjecture 4.1 would be established for translation-invariant models if the limits as $k \rightarrow \infty$ and $v \rightarrow \infty$ could be exchanged in Theorem 4.5. Similarly, we could aim to obtain the conclusion of Conjecture 4.1 for the prediction filter by letting $v \rightarrow -\infty$. However, we do not know how to establish the validity of Theorem 4.5 in either limit.

To obtain some insight into this idea, let us rewrite Theorem 4.5 in a measure-theoretic manner. Using the Markov property and translation invariance, we obtain

$$\begin{aligned} & \mathbf{E}|\mathbf{P}[X_k \in A | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}] - \mathbf{P}[X_k \in A | Y_1, \dots, Y_{k-1}, Y_k^{<v}]| = \\ & \mathbf{E}|\mathbf{P}[X_0 \in A | Y_0^{<v}, Y_{-1}, Y_{-2}, \dots; X_{-k}, X_{-k-1}, \dots] - \mathbf{P}[X_0 \in A | Y_0^{<v}, Y_{-1}, \dots, Y_{-k+1}]| \end{aligned}$$

as in the proof of Proposition 4.7. Letting $k \rightarrow \infty$ and using Theorem 4.5 yields

$$\begin{aligned} 0 &= \lim_{k \rightarrow \infty} \mathbf{E}|\mathbf{P}[X_k \in A | X_0, Y_1, \dots, Y_{k-1}, Y_k^{<v}] - \mathbf{P}[X_k \in A | Y_1, \dots, Y_{k-1}, Y_k^{<v}]| \\ &= \mathbf{E}|\mathbf{P}[X_0 \in A | \bigcap_k (\mathcal{Y}_-^v \vee \mathcal{X}_{-k})] - \mathbf{P}[X_0 \in A | \mathcal{Y}_-^v]|, \end{aligned}$$

where we defined the σ -fields $\mathcal{X}_k = \sigma\{X_k, X_{k-1}, \dots\}$ and $\mathcal{Y}_-^v = \sigma\{Y_0^{<v}, Y_{-1}, Y_{-2}, \dots\}$.

Now let us attempt to take the limit as $v \rightarrow -\infty$. This yields

$$\mathbf{E}|\mathbf{P}[X_0 \in A | \bigcap_{k,v} (\mathcal{Y}_-^v \vee \mathcal{X}_{-k})] - \mathbf{P}[X_0 \in A | \bigcap_v \mathcal{Y}_-^v]| = 0.$$

This does not suffice to establish Conjecture 4.1 for the prediction filter. In order to deduce the latter, we would need to establish the identity

$$\bigcap_v \mathcal{Y}_-^v = \mathcal{Y}_- := \sigma\{Y_{-1}, Y_{-2}, \dots\} \quad \text{mod } \mathbf{P}$$

(the notation $\mathcal{F} = \mathcal{G} \text{ mod } \mathbf{P}$ indicates that the \mathbf{P} -completions of the σ -fields \mathcal{F} and \mathcal{G} coincide). Indeed, if this is the case, then we obtain by Jensen's inequality

$$\begin{aligned} & \lim_{k \rightarrow \infty} \mathbf{E}|\mathbf{P}[X_k \in A | X_0, Y_1, \dots, Y_{k-1}] - \mathbf{P}[X_k \in A | Y_1, \dots, Y_{k-1}]| \\ &= \mathbf{E}|\mathbf{P}[X_0 \in A | \bigcap_k (\mathcal{Y}_- \vee \mathcal{X}_{-k})] - \mathbf{P}[X_0 \in A | \mathcal{Y}_-]| \\ &\leq \mathbf{E}|\mathbf{P}[X_0 \in A | \bigcap_{k,v} (\mathcal{Y}_-^v \vee \mathcal{X}_{-k})] - \mathbf{P}[X_0 \in A | \mathcal{Y}_-]| = 0, \end{aligned}$$

where we used $\bigcap_{k,v} (\mathcal{Y}_-^v \vee \mathcal{X}_{-k}) \supseteq \bigcap_k (\mathcal{Y}_- \vee \mathcal{X}_{-k}) \supseteq \mathcal{Y}_-$.

Similarly, let us take $v \rightarrow \infty$. This yields

$$\mathbf{E}|\mathbf{P}[X_0 \in A | \bigvee_v \bigcap_k (\mathcal{Y}_-^v \vee \mathcal{X}_{-k})] - \mathbf{P}[X_0 \in A | \bigvee_v \mathcal{Y}_-^v]| = 0.$$

In order to establish Conjecture 4.1, we would now need the identity

$$\bigvee_v \bigcap_k (\mathcal{Y}_-^v \vee \mathcal{X}_{-k}) = \bigcap_k \bigvee_v (\mathcal{Y}_-^v \vee \mathcal{X}_{-k}) \quad \text{mod } \mathbf{P},$$

the remainder of the argument proceeding in the same manner as for $v \rightarrow -\infty$.

Neither of the above measure-theoretic identities appears to be obvious; indeed, the problem of establishing such identities is closely related to the filter stability problem itself (cf. section 6.1). Nonetheless, the conclusion of Theorem 4.5 appears to be tantalizingly close to establishing Conjecture 4.1 for translation-invariant systems, and the fact that the latter does not appear to follow directly from the former provides one more indication of the delicacy of the filter stability problem in infinite dimension. A very similar argument will be used in the following section to resolve the continuous-time analogue of the problem; the key distinction in this case is that an appropriate measure-theoretic identity can in fact be established (Lemma 4.15 below).

4.3 Continuous time

In the previous section we have developed a partial result on filter stability in the case of translation-invariant models with direct observation structure. Unfortunately, that result concerns a quantity intermediate between the filter and prediction filter, and therefore falls short of resolving Conjecture 4.1 even in the translation-invariant setting. Surprisingly, however, it turns out that this problem can be resolved if we consider the natural *continuous time* analogue of Conjecture 4.1, providing a complete proof of filter stability in the translation-invariant setting for continuous-time models with direct observations. This idea will be developed in the remainder of this section.

To define the continuous-time counterpart of the filtering model of Conjecture 4.1, we begin by considering a stationary Markov process $X = (X_t)_{t \in \mathbb{R}}$ with càdlàg paths with values in $\{-1, 1\}^{\mathbb{Z}}$. We will assume that X satisfies the Feller property, that is, that $x \mapsto \mathbf{E}^x[f(X_t)]$ is a quasilocal function for every t and bounded quasilocal function f (a function is called quasilocal if it is the uniform limit of functions that depend on a finite number of coordinates).⁵ This mild condition ensures that the dynamics are local in a very weak sense (as compared to the much stronger local structure of the model in section 2.2, which was however not used in the previous section). Markov processes of this type arise broadly in the literature on interacting particle systems, cf. [31]. To define the local observations $(Y_t)_{t \in \mathbb{R}}$, we introduce a ‘white noise’ model of the form

$$dY_t^v = X_t^v dt + \sigma dW_t^v, \quad Y_0^v = 0, \quad v \in \mathbb{Z},$$

where $(W_t^v)_{t \in \mathbb{R}}$ are i.i.d. two-sided Brownian motions independent of X , and $\sigma > 0$ denotes the noise strength. The process $(X_t, Y_t)_{t \in \mathbb{R}}$ is a natural continuous-time analogue of the model of Conjecture 4.1, and will be used in the remainder of this section.

Remark 4.11. The details of the present model are not essential for our results. The proof is easily extended to random fields on \mathbb{Z}^d with values in a finite state space, and to observation models other than the usual white noise model (cf. [49]). For concreteness and to avoid additional notation, we will work here in the simplest setting defined above.

As in section 4.2, we will further assume that the model is translation-invariant

$$(X_t^v, Y_t^v)_{t \in \mathbb{R}, v \in \mathbb{Z}} \stackrel{\text{law}}{=} (X_{t+s}^{v+w}, Y_{t+s}^{v+w} - Y_s^{v+w})_{t \in \mathbb{R}, v \in \mathbb{Z}} \quad \text{for all } s \in \mathbb{R}, w \in \mathbb{Z}$$

(note that, due to the additive nature of the observations, it is the increments of Y that are translation invariant and not Y itself). In this setting, we obtain the following result.

Theorem 4.12. *For the model of this section, we have*

$$|\mathbf{P}[X_t \in A | X_0, \{Y_s\}_{0 \leq s \leq t}] - \mathbf{P}[X_t \in A | \{Y_s\}_{0 \leq s \leq t}]| \xrightarrow{t \rightarrow \infty} 0 \quad \text{in } L^1$$

for every measurable set A .

This result evidently resolves, in the translation-invariant case, the continuous-time analogue of Conjecture 4.1. We remark once more that stability of X is not assumed.

To prove Theorem 4.12 we require a sharper version of the entropy identity of Conze that was used in the proof of Theorem 4.5 (cf. Remark 4.8). The proof of the requisite identity, which we state presently, can be found in [10, p. 17, §8].

⁵We recall that a function on $\{-1, 1\}^{\mathbb{Z}}$ is quasilocal if and only if it is continuous [22, Remark 2.21]. Thus the definition given here is equivalent to the usual definition of the Feller property of a Markov process.

Lemma 4.13 ([10]). *Let $(Z^q)_{q \in \mathbb{Z}^2}$ be a translation-invariant random field where $Z^q = (\tilde{X}_k^v, \tilde{Y}_k^v)$ for $q = (k, v) \in \mathbb{Z}^2$ takes values in a finite set. Then for any $n \geq 0$*

$$\begin{aligned} H(\tilde{Y}_0^v | \tilde{Y}_0^{<v}, \tilde{Y}_{<0}) - H(\tilde{Y}_0^v | \tilde{Y}_0^{<v}, \tilde{Y}_{<0}; \tilde{X}_{-n}^{<v}, \tilde{X}_{<-n}) = \\ H(\tilde{X}_0^v | \tilde{X}_0^{<v}, \tilde{X}_{<0}; \tilde{Y}_n^v, \tilde{Y}_n^{<v}, \tilde{Y}_{<n}) - H(\tilde{X}_0^v | \tilde{X}_0^{<v}, \tilde{X}_{<0}; \tilde{Y}), \end{aligned}$$

where $\tilde{Y}_n^{<v} := (\tilde{Y}_n^w)_{w < v}$, $\tilde{Y}_{<k} := (\tilde{Y}_n)_{n < k}$, and $\tilde{Y} := (\tilde{Y}_n)_{n \in \mathbb{Z}}$.

Let us note that the identity in Remark 4.8 follows immediately as $n \rightarrow \infty$.

The problem in continuous time is that we no longer have a discrete random field as in Lemma 4.13. We address this by an appropriate discretization method, which yields the following continuous-time counterpart of Proposition 4.7.

Proposition 4.14. *Let $\delta > 0$ and $v \in \mathbb{Z}$, $m \in \mathbb{N}$ be arbitrary, and define*

$$\begin{aligned} \rho_s &:= \mathbf{P}[\{Y_t^v, \dots, Y_t^{v+m}\}_{t \in [0, \delta]} \in \cdot | Y_{\leq 0}, \{Y_t^{<v}\}_{t \in [0, \delta]}, X_{\leq s}], \\ \rho &:= \mathbf{P}[\{Y_t^v, \dots, Y_t^{v+m}\}_{t \in [0, \delta]} \in \cdot | Y_{\leq 0}, \{Y_t^{<v}\}_{t \in [0, \delta]}], \end{aligned}$$

where $X_{\leq s} := (X_t)_{t \leq s}$ and $Y_{\leq 0} := (Y_t)_{t \leq 0}$. Then

$$\mathbf{E}[D(\rho_s | \rho)] \xrightarrow{s \rightarrow -\infty} 0,$$

where $D(\mu | \nu)$ denotes relative entropy.

Proof. Let us choose $v = 0$ for simplicity. The result for arbitrary v follows immediately by translation invariance. In the following, we fix $\delta > 0$ and $m \in \mathbb{N}$.

Define the random field $Z = (Z^q)_{q \in \mathbb{Z}^2}$ with $Z^q = (\tilde{X}_k^r, \tilde{Y}_k^r)$ for $q = (k, r) \in \mathbb{Z}^2$ as

$$\begin{aligned} \tilde{X}_k^r &:= (X_{k\delta}^{r(m+1)}, \dots, X_{k\delta}^{r(m+1)+m}), \\ \tilde{Y}_k^r &:= (Y_{k\delta+t}^{r(m+1)} - Y_{k\delta}^{r(m+1)}, \dots, Y_{k\delta+t}^{r(m+1)+m} - Y_{k\delta}^{r(m+1)+m})_{t \in [0, \delta]}. \end{aligned}$$

Then evidently Z is translation-invariant and \tilde{X}_k^r is finite-valued, but \tilde{Y}_k^r takes values in the space $C_0([0, \delta]; \mathbb{R}^{m+1})$ of continuous paths $\omega : [0, \delta] \rightarrow \mathbb{R}^{m+1}$ with $\omega(0) = 0$ (which is Polish when endowed with the topology of uniform convergence and the associated Borel σ -field). Thus we cannot directly apply Lemma 4.13.

To surmount this problem, we employ a straightforward discretization procedure. Let $\{A_j\}_{j \geq 1}$ be a countable generating class for the Borel σ -field \mathcal{F} of $C_0([0, \delta]; \mathbb{R}^{m+1})$, and define the functions $\kappa_j : C_0([0, \delta]; \mathbb{R}^{m+1}) \rightarrow \{0, 1\}^j$ as $\kappa_j := (\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_j})$. Then $\mathcal{F}_j := \sigma\{\kappa_j\}$ is an increasing family of σ -fields such that $\bigvee_j \mathcal{F}_j = \mathcal{F}$. Now define $\tilde{Y}_k^r(j) := \kappa_j(\tilde{Y}_k^r)$. Then the random field $(\tilde{X}_k^r, \tilde{Y}_k^r(j))_{k, r \in \mathbb{Z}}$ is translation-invariant and finite-valued for every $j \geq 1$. Thus we can apply Lemma 4.13 to obtain

$$\begin{aligned} H(\tilde{Y}_0^0(j) | \tilde{Y}_0^{<0}(j), \tilde{Y}_{<0}(j)) - H(\tilde{Y}_0^0(j) | \tilde{Y}_0^{<0}(j), \tilde{Y}_{<0}(j); \tilde{X}_{-n}^{<0}, \tilde{X}_{<-n}) = \\ H(\tilde{X}_0^0 | \tilde{X}_0^{<0}, \tilde{X}_{<0}; \tilde{Y}_n^0(j), \tilde{Y}_n^{<0}(j), \tilde{Y}_{<n}(j)) - H(\tilde{X}_0^0 | \tilde{X}_0^{<0}, \tilde{X}_{<0}; \tilde{Y}(j)). \end{aligned}$$

In particular, as the left-hand side of this expression is an expected relative entropy (see the proof of Proposition 4.4), we can estimate

$$\begin{aligned} H(\tilde{Y}_0^0(i) | \tilde{Y}_0^{<0}(j), \tilde{Y}_{<0}(j)) - H(\tilde{Y}_0^0(i) | \tilde{Y}_0^{<0}(j), \tilde{Y}_{<0}(j); \tilde{X}_{-n}^{<0}, \tilde{X}_{<-n}) \\ \leq H(\tilde{X}_0^0 | \tilde{X}_0^{<0}, \tilde{X}_{<0}; \tilde{Y}_n^0(j), \tilde{Y}_n^{<0}(j), \tilde{Y}_{<n}(j)) - H(\tilde{X}_0^0 | \tilde{X}_0^{<0}, \tilde{X}_{<0}; \tilde{Y}(j)) \end{aligned}$$

for every $i \leq j$. Letting $j \rightarrow \infty$ and using that conditioning reduces entropy gives

$$\begin{aligned} & \sup_{i \geq 1} \{H(\tilde{Y}_0^0(i)|\tilde{Y}_0^{<0}, \tilde{Y}_{<0}) - H(\tilde{Y}_0^0(i)|\tilde{Y}_0^{<0}, \tilde{Y}_{<0}; \tilde{X}_{<-n})\} \\ & \leq H(\tilde{X}_0^0|\tilde{X}_0^{<0}, \tilde{X}_{<0}; \tilde{Y}_n^0, \tilde{Y}_n^{<0}, \tilde{Y}_{<n}) - H(\tilde{X}_0^0|\tilde{X}_0^{<0}, \tilde{X}_{<0}; \tilde{Y}). \end{aligned}$$

To proceed, we write the left hand side as an expected relative entropy as in the proof of Proposition 4.4. Using the continuity of the relative entropy in information (e.g., [14, Lemma 4.4.15]) and monotone convergence, this yields

$$\begin{aligned} & \mathbf{E}[D(\mathbf{P}[\tilde{Y}_0^0 \in \cdot | \tilde{Y}_0^{<0}, \tilde{Y}_{<0}, \tilde{X}_{<-n}] || \mathbf{P}[\tilde{Y}_0^0 \in \cdot | \tilde{Y}_0^{<0}, \tilde{Y}_{<0}])] \\ & \leq H(\tilde{X}_0^0|\tilde{X}_0^{<0}, \tilde{X}_{<0}; \tilde{Y}_n^0, \tilde{Y}_n^{<0}, \tilde{Y}_{<n}) - H(\tilde{X}_0^0|\tilde{X}_0^{<0}, \tilde{X}_{<0}; \tilde{Y}). \end{aligned}$$

It therefore follows immediately that

$$\mathbf{E}[D(\mathbf{P}[\tilde{Y}_0^0 \in \cdot | \tilde{Y}_0^{<0}, \tilde{Y}_{<0}, \tilde{X}_{<-n}] || \mathbf{P}[\tilde{Y}_0^0 \in \cdot | \tilde{Y}_0^{<0}, \tilde{Y}_{<0}])] \xrightarrow{n \rightarrow \infty} 0.$$

It remains to note that

$$\mathbf{P}[\tilde{Y}_0^0 \in \cdot | \tilde{Y}_0^{<0}, \tilde{Y}_{<0}] = \rho, \quad \mathbf{P}[\tilde{Y}_0^0 \in \cdot | \tilde{Y}_0^{<0}, \tilde{Y}_{<0}, \tilde{X}_{<-n}] = \rho_{-(n+1)\delta},$$

where we have used the Markov property of $(X_t, Y_t)_{t \geq 0}$ to obtain the latter equality. This establishes that $\mathbf{E}[D(\rho_s || \rho)] \rightarrow 0$ along the subsequence $s = -(n+1)\delta$, and therefore as $s \rightarrow -\infty$ as $\mathbf{E}[D(\rho_s || \rho)]$ is decreasing in s (as is easily verified by Jensen's inequality). \square

We can now complete the proof of Theorem 4.12.

Proof of Theorem 4.12. By translation-invariance and Lemma 3.8, it suffices to show

$$|\mathbf{P}[(X_0^0, \dots, X_0^m) \in A | X_{-t}, \{Y_s\}_{-t \leq s \leq 0}] - \mathbf{P}[(X_0^0, \dots, X_0^m) \in A | \{Y_s\}_{-t \leq s \leq 0}]| \xrightarrow{t \rightarrow \infty} 0 \text{ in } L^1$$

for every set A and $m \geq 1$. But by the martingale convergence theorem, and as $\{X_s, Y_s\}_{s \leq -t}$ is conditionally independent of $\{X_s, Y_s\}_{s \geq -t}$ given X_{-t}, Y_{-t} by the Markov property, it suffices to show for every set A and $m \geq 1$ that

$$|\mathbf{P}[(X_0^0, \dots, X_0^m) \in A | Y_{\leq 0}, X_{\leq -t}] - \mathbf{P}[(X_0^0, \dots, X_0^m) \in A | Y_{\leq 0}]| \xrightarrow{t \rightarrow \infty} 0 \text{ in } L^1.$$

By the martingale convergence theorem, this can be formulated equivalently as

$$\mathbf{P}[(X_0^0, \dots, X_0^m) \in A | \bigcap_t \sigma\{Y_{\leq 0}, X_{\leq -t}\}] = \mathbf{P}[(X_0^0, \dots, X_0^m) \in A | Y_{\leq 0}].$$

The key distinction between the continuous- and discrete-time settings is the following.

Lemma 4.15. *We have*

$$\bigcap_{\delta > 0} \sigma\{Y_{\leq 0}, \{Y_s^{<0}\}_{s \in [0, \delta]}, X_{\leq -t}\} = \sigma\{Y_{\leq 0}, X_{\leq -t}\} \quad \text{mod } \mathbf{P}$$

and

$$\bigcap_{\delta > 0} \sigma\{Y_{\leq 0}, \{Y_s^{<0}\}_{s \in [0, \delta]}\} = \sigma\{Y_{\leq 0}\} \quad \text{mod } \mathbf{P}.$$

Proof. Let us prove the first statement (the second statement follows readily in the same manner). It evidently suffices to prove the stronger identity

$$\bigcap_{\delta > 0} \sigma\{Y_{\leq \delta}, X_{\leq -t}\} = \sigma\{Y_{\leq 0}, X_{\leq -t}\} \quad \text{mod } \mathbf{P}.$$

To this end, it suffices to show that

$$\mathbf{E}[Z|Y_{\leq \delta}, X_{\leq -t}] \xrightarrow{\delta \downarrow 0} \mathbf{E}[Z|Y_{\leq 0}, X_{\leq -t}] \quad \text{in } L^2$$

for every bounded random variable Z . By a standard approximation argument, it suffices to consider Z of the form $Z_s = f(X_{s+t_1}, Y_{s+t_1}, \dots, X_{s+t_n}, Y_{s+t_n})$ for $n \in \mathbb{N}$, $t_1, \dots, t_n \in \mathbb{R}$, and f bounded and local. Now note that by stationarity,

$$\mathbf{E}[\mathbf{E}[Z_s|Y_{\leq \delta}, X_{\leq -t}]^2] = \mathbf{E}[\mathbf{E}[Z_{s-\delta}|Y_{\leq 0}, X_{\leq -t-\delta}]^2].$$

As X is Feller, it is quasi-left continuous [42, p. 101], and therefore $X_{t-\delta} \rightarrow X_t$ a.s. as $\delta \downarrow 0$. In particular, as f is local and Y is continuous by construction, we have $Z_{s-\delta} \rightarrow Z_s$ a.s. as $\delta \downarrow 0$. On the other hand, as f is bounded, we obtain

$$\mathbf{E}[\mathbf{E}[Z_s|Y_{\leq \delta}, X_{\leq -t}]^2] \xrightarrow{\delta \downarrow 0} \mathbf{E}[\mathbf{E}[Z_s|Y_{\leq 0}, X_{\leq -t}]^2] = \mathbf{E}[\mathbf{E}[Z_s|Y_{\leq 0}, X_{\leq -t}]^2],$$

using Hunt's lemma [42, Corollary II.2.4], where the last equality follows again by quasi-left continuity. Thus $\mathbf{E}[Z_s|Y_{\leq \delta}, X_{\leq -t}] \rightarrow \mathbf{E}[Z_s|Y_{\leq 0}, X_{\leq -t}]$ in L^2 , completing the proof. \square

By virtue of Lemma 4.15, to complete the proof of Theorem 4.12, it suffices to show

$$\mathbf{P}[(X_0^0, \dots, X_0^m) \in A | \bigcap_t \sigma\{Y_{\leq 0}, \{Y_s^{<0}\}_{s \in [0, \delta]}, X_{\leq -t}\}] = \mathbf{P}[(X_0^0, \dots, X_0^m) \in A | Y_{\leq 0}, \{Y_s^{<0}\}_{s \in [0, \delta]}]$$

for every set A , $m \geq 1$, and $\delta > 0$: indeed, letting $\delta \downarrow 0$ the yields the expression before the statement of Lemma 4.15. We will deduce this fact from Proposition 4.14. To this end we require a lemma that replaces the analogous argument in Proposition 4.4.

Lemma 4.16. *Let $f : \mathbb{R}^{m+1} \rightarrow \mathbb{R}$ be a bounded continuous function. Then there exists a sequence of bounded continuous functions $g_n : \mathbb{R}^{m+1} \rightarrow \mathbb{R}$ such that*

$$|\mathbf{E}[g_n(nY_{1/n}^0, \dots, nY_{1/n}^m) | X] - f(X_0^0, \dots, X_0^m)| \xrightarrow{n \rightarrow \infty} 0 \quad \text{in } L^1.$$

Proof. By the definition of the observations,

$$\mathbf{E}[g(nY_{1/n}^0, \dots, nY_{1/n}^m) | X] = (g * \xi_n)(n \int_0^{1/n} X_s^0 ds, \dots, n \int_0^{1/n} X_s^m ds),$$

where ξ_n denotes the centered Gaussian measure on \mathbb{R}^{m+1} with covariance $n \text{Id}$ and $*$ denotes convolution. Note the trivial estimate $|\int_0^t X_s^v ds| \leq t$, so the argument of the function $g * \xi_n$ above takes values in the compact set $C = [-1, 1]^{m+1}$.

We now recall that as C is compact, every continuous function on C is contained in the closure of $\{(g * \xi_n)|_C : g \in \mathcal{C}_b(\mathbb{R}^{m+1})\}$ with respect to the uniform convergence topology on C (here $\mathcal{C}_b(\mathbb{R}^{n+1})$ is the family of bounded continuous functions on \mathbb{R}^{n+1}). This follows

from an elementary Hahn-Banach argument, cf. [49, Remark 5]. We can therefore choose, for each n , a bounded continuous function g_n such that

$$|(g_n * \xi_n)(x) - f(x)| \leq 1/n \quad \text{for all } x \in C.$$

Then we evidently have

$$\begin{aligned} |\mathbf{E}[g_n(nY_{1/n}^0, \dots, nY_{1/n}^m)|X] - f(X_0^0, \dots, X_0^m)| \\ \leq 1/n + |f(n \int_0^{1/n} X_s^0 ds, \dots, n \int_0^{1/n} X_s^m ds) - f(X_0^0, \dots, X_0^m)|. \end{aligned}$$

As f is bounded and continuous, and as the paths of X are right-continuous, this expression converges to zero as $n \rightarrow \infty$ a.s. and in L^1 . \square

Note that, by the definition of our model, $Y_{\leq 0}, \{Y_s^{<0}\}_{s \in [0, \delta]}$ is conditionally independent of $\{Y_s^0, \dots, Y_s^m\}_{s \in [0, \delta]}$ given X , so that for every bounded continuous function h

$$\mathbf{E}[h(Y_{1/n}^0, \dots, Y_{1/n}^m)|Y_{\leq 0}, \{Y_s^{<0}\}_{s \in [0, \delta]}, X] = \mathbf{E}[h(Y_{1/n}^0, \dots, Y_{1/n}^m)|X].$$

In particular, in view of Lemma 4.16, it now suffices to show that

$$\mathbf{E}[h(Y_s^0, \dots, Y_s^m)|\bigcap_t \sigma\{Y_{\leq 0}, \{Y_s^{<0}\}_{s \in [0, \delta]}, X_{\leq -t}\}] = \mathbf{E}[h(Y_s^0, \dots, Y_s^m)|Y_{\leq 0}, \{Y_s^{<0}\}_{s \in [0, \delta]}]$$

for every bounded continuous function h and $s \in [0, \delta]$. But this follows readily from Proposition 4.14 using Pinsker's inequality and martingale convergence. \square

5 Conditional random fields

Thus far we have considered infinite-dimensional counterparts of classical stability problems in nonlinear filtering. However, new questions arise in infinite dimension beyond stability that are of interest in their own right. In particular, it is of significant interest (cf. [40]) to understand the spatial mixing and decay of correlations properties of conditional distributions in infinite dimension, which could be viewed as spatial counterparts to the filter stability property. Such questions already arise in the absence of dynamics, and thus we proceed in this section to introduce such problems in the most basic setting of conditional random fields (that is, in models with only spatial degrees of freedom). Our motivations for such questions are threefold:

1. Random fields provide the simplest possible setting to investigate the spatial mixing properties of conditional distributions.
2. Conditional random fields are of practical interest in their own right, for example, in Bayesian image analysis applications [54, 21].
3. Even in the more classical setting of the previous sections, the random field viewpoint proves to be fundamental to the understanding of filter stability in infinite dimension: indeed, the proofs in both sections 3 and 4 above and in [40, 41] exploit the idea that $(X_k^v, Y_k^v)_{k \in \mathbb{Z}, v \in \mathbb{Z}^d}$ can be viewed as a space-time random field.

The remainder of this section is organized as follows. In section 5.1, we recall some basic notions from the theory of Markov random fields. In section 5.2, we develop basic properties of conditional random fields and introduce some of the relevant questions. Finally, in

section 5.3, we develop a general result that ensures the inheritance of ergodicity under conditioning in random fields that possess certain monotonicity properties. The latter provides a mechanism for the resolution of the random field counterpart of Conjecture 4.1 that is quite distinct from the observability theory of section 4.

5.1 Markov random fields

A random field is a collection of random variables X_v that are indexed by the spatial degree of freedom v . For simplicity, we will assume in the sequel that $v \in \mathbb{Z}^d$ (but see section 6.3 below) and that each X_v takes values in a finite set E .

In the following, we define for any $V \subseteq \mathbb{Z}^d$

$$V^c := \mathbb{Z}^d \setminus V, \quad \partial V := \{w \in V^c : \|v - w\| = 1 \text{ for some } v \in V\}, \quad X_V := (X_v)_{v \in V}.$$

If V is a finite subset of \mathbb{Z}^d , we will write $V \subset\subset \mathbb{Z}^d$. We now recall a basic definition.

Definition 5.1. $X = (X_v)_{v \in \mathbb{Z}^d}$ is called a *Markov random field* if it possesses the (local) Markov property, that is, $\mathbf{P}[X_V \in \cdot | X_{V^c}]$ depends only on $X_{\partial V}$ for every $V \subset\subset \mathbb{Z}^d$.

Just as Markov chains are defined by transition probabilities, Markov random fields are defined by a family of local transition kernels called a *specification* [22, Chapter 1].

Definition 5.2. A family $\gamma = (\gamma_V)_{V \subset\subset \mathbb{Z}^d}$ of transition kernels on $E^{\mathbb{Z}^d}$ such that

1. $\gamma_V(x, A)$ is a function of $x_{\partial V}$ for every $A \in \sigma\{X_V\}$ and $V \subset\subset \mathbb{Z}^d$,
2. $\gamma_V(x, A) = \mathbf{1}_A(x)$ for every $A \in \sigma\{X_{V^c}\}$ and $V \subset\subset \mathbb{Z}^d$,
3. $\gamma_V \gamma_W = \gamma_V$ for every $W \subset V \subset\subset \mathbb{Z}^d$,

is called a *specification*. A Markov random field X is said to be *specified* by γ if we have $\mathbf{P}(X \in A | X_{V^c}) = \gamma_V(X, A)$ for every measurable set A and $V \subset\subset \mathbb{Z}^d$. The family of all laws of Markov random fields specified by γ is denoted $\mathcal{G}(\gamma)$.

Example 5.3. Standard constructions of Markov random fields arise in statistical mechanics in the following manner. Let $\psi_v : E \rightarrow \mathbb{R}$ and $\varphi_{\{v,w\}} : E \times E \rightarrow \mathbb{R}$ for $v, w \in \mathbb{Z}^d$ with $\|v - w\| = 1$ be given potential functions, and let

$$\gamma_V(x, A) = \frac{1}{Z} \sum_{x_V \in E^V} \mathbf{1}_A(x) \exp \left(\sum_{\{v,w\} \subset V \cup \partial V : \|v-w\|=1} \varphi_{\{v,w\}}(x_v, x_w) + \sum_{v \in V} \psi_v(x_v) \right)$$

where Z is the appropriate normalization factor. It is easily verified that $\gamma = (\gamma_V)_{V \subset\subset \mathbb{Z}^d}$ defines a specification. The potentials ψ_v and $\varphi_{\{v,w\}}$ describe the local external and interaction forces between different sites, and are defined directly in terms of the physical parameters of the problem. For example, if $E = \{-1, 1\}$, $\varphi_{\{v,w\}}(\sigma, \sigma') = \beta J \sigma \sigma'$, and $\psi_v(\sigma) = \beta \mu \sigma$ with $\beta, J > 0$ and $\mu \in \mathbb{R}$, this is the well known ferromagnetic Ising model with inverse temperature β , interaction strength J and magnetic field strength μ . The construction in terms of potentials will be inessential in the sequel, however.

Given a specification γ , there always exists a random field in $\mathcal{G}(\gamma)$ under our assumptions. However, just as a Markov chain with given transition probabilities may admit more than one stationary distribution, the random field associated to a given specification need

not be unique. In fact, the structure of the set $\mathcal{G}(\gamma)$ is closely related to the spatial mixing properties of the associated random fields, as is shown by the following result [22, section 4.4, Proposition 7.11, Theorem 7.7]. To interpret the notion of extremality that arises here, note that if \mathbf{P} and \mathbf{Q} are the laws of two random fields in $\mathcal{G}(\gamma)$, then $\lambda\mathbf{P} + (1 - \lambda)\mathbf{Q}$ is also in $\mathcal{G}(\gamma)$ for $0 \leq \lambda \leq 1$ [22, Chapter 7]; thus $\mathcal{G}(\gamma)$ is a convex set, and a random field is called *extremal* if it is an extreme point of this set.

Theorem 5.4. *For a given specification γ , the following hold.*

1. *Existence of a random field: $\mathcal{G}(\gamma) \neq \emptyset$.*
2. *Uniqueness \Leftrightarrow uniform mixing: $|\mathcal{G}(\gamma)| = 1$ iff a random field in $\mathcal{G}(\gamma)$ satisfies^{6,7}*

$$\lim_{W \subset \subset \mathbb{Z}^d} \sup_x |\mathbf{P}[X_V \in A | X_{W^c} = x_{W^c}] - \mathbf{P}[X_V \in A]| = 0$$

for every set A and $V \subset \subset \mathbb{Z}^d$.

3. *Extremality \Leftrightarrow mixing: the random field X is an extreme point of $\mathcal{G}(\gamma)$ iff*

$$\lim_{W \subset \subset \mathbb{Z}^d} \mathbf{E} |\mathbf{P}[X_V \in A | X_{W^c}] - \mathbf{P}[X_V \in A]| = 0$$

for every set A and $V \subset \subset \mathbb{Z}^d$.

The mixing property in Theorem 5.4 is a direct spatial analogue of the stability property of a Markov chain introduced in section 2.1. Indeed, a Markov chain is stable if it forgets its initial condition after a long time: that is, the Markov chain has a ‘finite memory.’ Similarly, a random field is mixing if the distribution of any finite set of sites V is insensitive to knowledge of the configuration of the field outside a larger set W when the distance between V and W^c is large. This implies in particular that distant sites are nearly independent, that is, the field has ‘finite correlation length.’ The *uniform* mixing property is a strictly stronger notion, where the forgetting property holds uniformly in the boundary configuration $x_{\partial W}$ (recall that by the Markov property of the random field, $\mathbf{P}[X \in C | X_{W^c} = x_{W^c}]$ depends on $x_{\partial W}$ only).

5.2 Conditional random fields and conditional mixing

In the following, let us fix a specification γ and a Markov random field $X = (X_v)_{v \in \mathbb{Z}^d}$ that is specified by γ . In order to investigate the conditional distributions of random fields, we must introduce a suitable observation structure. To this end, in analogy with section 2.2, let us fix for each $v \in \mathbb{Z}^d$ a transition kernel Φ_v from the state space E of the random field to a measurable space F in which the observations take their values. We now construct the observations $Y = (Y_v)_{v \in \mathbb{Z}^d}$ such that

$$\mathbf{P}[Y \in dy | X] = \prod_{v \in \mathbb{Z}^d} \Phi_v(X_v, dy_v);$$

that is, each site of the underlying field is observed independently with $\mathbf{P}[Y_v \in A | X_v] = \Phi_v(X_v, A)$. The resulting model $(X_v, Y_v)_{v \in \mathbb{Z}^d}$ is called a *hidden Markov random field*.

⁶Here we used the suggestive notation $\mathbf{P}[X \in C | X_{W^c} = x_{W^c}] := \gamma_W(x, C)$ to emphasize the significance of the mixing property. Note that $\mathbf{P}[X \in C | X_{W^c}] = \gamma_W(X, C)$ holds a.s. by the definition of $\mathcal{G}(\gamma)$, but the equivalence between uniqueness and uniform mixing is false if a null set is omitted in the supremum over x .

⁷The notation $\lim_W a_W$ denotes the limit of the net $\{a_W\}$, where $\{W \subset \subset \mathbb{Z}^d\}$ is directed by inclusion.

Remark 5.5. For notational simplicity, we have formulated our model such that the observations are attached to individual sites $v \in \mathbb{Z}^d$. One could also consider more general models, for example, where an observation $Y_{\{v,w\}}$ is attached to every edge $\{v,w\} \subset \mathbb{Z}^d$, $\|v-w\| = 1$ with $\mathbf{P}[Y_{\{v,w\}} \in A | X] = \Phi_{\{v,w\}}(X_v, X_w, A)$ (cf. Example 5.8). The results of this section will continue to hold in this setting with minor modifications.

We can now formulate the natural counterpart of the filter stability property in hidden Markov random fields: the model is said to be conditionally mixing if the conditional distribution of the underlying process in a finite set of sites given the observations is insensitive to knowledge of the configuration of the field at distant sites.

Definition 5.6. The hidden Markov random field $(X_v, Y_v)_{v \in \mathbb{Z}^d}$ is *conditionally mixing* if

$$\lim_{W \subset \subset \mathbb{Z}^d} \mathbf{E} |\mathbf{P}[X_V \in A | X_{W^c}, Y] - \mathbf{P}[X_V \in A | Y]| = 0$$

for every set A and $V \subset \subset \mathbb{Z}^d$.

The basic question to be addressed in this setting is therefore: *when is the mixing property inherited by conditioning*, that is, when does the mixing property of the random field X imply the conditional mixing property of (X, Y) ?

It will be insightful to reformulate the problem in different terms. For simplicity, we will assume in the sequel that the observations are locally nondegenerate, that is, that $\Phi_v(x_v, dy_v) = g_v(x_v, y_v) \varphi(dy_v)$ for some positive density $g_v(x_v, y_v) > 0$ for all x_v, y_v (the reference measure $\varphi(dy_v)$ on F may be any σ -finite measure.)

Proposition 5.7. Define for every $y \in F^{\mathbb{Z}^d}$ and $V \subset \subset \mathbb{Z}^d$ the transition kernel on $E^{\mathbb{Z}^d}$

$$\gamma_V^y(x, A) = \frac{\int \mathbf{1}_A(z) \prod_{v \in V} g_v(z_v, y_v) \gamma_V(x, dz)}{\int \prod_{v \in V} g_v(z_v, y_v) \gamma_V(x, dz)}.$$

Then the following hold.

1. $\gamma^y = (\gamma_V^y)_{V \subset \subset \mathbb{Z}^d}$ is a specification for every $y \in \mathbb{Z}^d$.
2. $\mathbf{P}[X \in \cdot | Y]$ is in $\mathcal{G}(\gamma^Y)$ a.s.
3. (X, Y) is conditionally mixing iff $\mathbf{P}[X \in \cdot | Y]$ is extremal in $\mathcal{G}(\gamma^Y)$ a.s.

Proof. We begin by verifying that γ^y is a specification. To this end, let $W \subset V \subset \subset \mathbb{Z}^d$. As $\gamma_V \gamma_W = \gamma_V$ and $\gamma_W(fg) = g \gamma_W f$ if $g(x)$ depends only on x_{W^c} , we can write

$$\begin{aligned} & \int \mathbf{1}_A(z) \prod_{v \in V} g_v(z_v, y_v) \gamma_V(x, dz) \\ &= \int \gamma_W^y(z', A) \int \prod_{w \in W} g_w(z_w, y_w) \gamma_W(z', dz) \prod_{v \in V \setminus W} g_v(z'_v, y_v) \gamma_V(x, dz') \\ &= \int \gamma_W^y(z, A) \prod_{v \in V} g_v(z_v, y_v) \gamma_V(x, dz). \end{aligned}$$

Thus $\gamma_V^y \gamma_W^y = \gamma_V^y$, and the remaining properties of a specification hold trivially.

Next, we show that $\mathbf{P}[X \in \cdot | Y]$ is in $\mathcal{G}(\gamma^Y)$ a.s. To this end, let us fix any regular version \mathbf{P}^Y of the conditional distribution $\mathbf{P}[\cdot | Y]$. We must show that for a.e. observation record y , we have $\mathbf{P}^y[X \in A | X_{V^c}] = \gamma_V^y(X, A)$ for all A , that is, we must show that

$$\mathbf{E}^y[\gamma_V^y(X, A)\mathbf{1}_B] = \mathbf{P}^y[\{X \in A\} \cap B] \quad \text{for every measurable } A \text{ and } B \in \sigma\{X_{V^c}\}$$

holds for \mathbf{P} -a.e. y . It is easily seen by the definition of a hidden Markov random field that

$$\gamma_V^Y(X, A) = \mathbf{P}[X \in A | X_{V^c}, Y].$$

We therefore have

$$\mathbf{E}[\gamma_V^Y(X, A)\mathbf{1}_B\mathbf{1}_C] = \mathbf{P}[\{X \in A\} \cap B \cap C]$$

for every A and $B \in \sigma\{X_{V^c}\}$, $C \in \sigma\{Y\}$. It follows by disintegration that

$$\mathbf{E}^Y[\gamma_V^Y(X, A)\mathbf{1}_B] = \mathbf{P}^Y[\{X \in A\} \cap B]$$

holds \mathbf{P} -a.s. for a fixed choice of $A, B \in \sigma\{X_{V^c}\}$, and thus simultaneously for a countable family of sets A and $B \in \sigma\{X_{V^c}\}$. By choosing the countable family to be a generating class (note that all our σ -fields are countably generated), the above identity holds simultaneously for every A and $B \in \sigma\{X_{V^c}\}$ by a monotone class argument. As there are only countably many $V \subset \subset \mathbb{Z}^d$, we have proved that $\mathbf{P}[X \in \cdot | Y]$ is in $\mathcal{G}(\gamma^Y)$ a.s.

Finally, we consider the conditional mixing property. As the limit in the definition of (conditional) mixing is over a decreasing net (by Jensen's inequality), it suffices to consider the limit along any fixed cofinal increasing sequence $W_n \subset \subset \mathbb{Z}^d$. Thus by the martingale convergence theorem, the conditional mixing property holds if and only if

$$\lim_{n \rightarrow \infty} \mathbf{E}[|\mathbf{P}[X \in A | X_{W_n^c}, Y] - \mathbf{P}[X \in A | Y]| | Y] = 0 \quad \text{a.s.}$$

for every $V \subset \subset \mathbb{Z}^d$ and $A \in \sigma\{X_V\}$. As we have shown that $\mathbf{P}[X \in A | X_{W_n^c}, Y] = \gamma_{W_n}^Y(X, A) = \mathbf{P}^Y[X \in A | X_{W_n^c}]$, the conditional mixing property is equivalent to

$$\lim_{n \rightarrow \infty} \mathbf{E}^y|\mathbf{P}^y[X \in A | X_{W_n^c}] - \mathbf{P}^y[X \in A]| = 0 \quad \text{for } \mathbf{P}\text{-a.e. } y$$

for every $V \subset \subset \mathbb{Z}^d$ and $A \in \sigma\{X_V\}$. But by the martingale convergence theorem

$$\lim_{n \rightarrow \infty} \mathbf{E}^y|\mathbf{P}^y[X \in A | X_{W_n^c}] - \mathbf{P}^y[X \in A]| = \mathbf{E}^y|\mathbf{P}^y[X \in A | \bigcap_n \sigma\{X_{W_n^c}\}] - \mathbf{P}^y[X \in A]|.$$

Thus we can again use a monotone class argument as above to remove the dependence of the \mathbf{P} -null set on V and A . Thus $(X_v, Y_v)_{v \in \mathbb{Z}^d}$ is conditionally mixing if and only if

$$\lim_{W \subset \subset \mathbb{Z}^d} \mathbf{E}^y|\mathbf{P}^y[X \in A | X_{W^c}] - \mathbf{P}^y[X \in A]| = 0 \quad \text{for every } V \subset \subset \mathbb{Z}^d, A \in \sigma\{X_V\}$$

holds for \mathbf{P} -a.e. y , which is precisely the mixing property of $\mathbf{P}[X \in \cdot | Y]$. \square

Proposition 5.7 shows that the conditional distribution $\mathbf{P}[X \in \cdot | Y]$ defines again a (random) Markov random field, and gives an explicit expression for its specification γ^Y . The inheritance of ergodicity can now be formulated in terms of the ergodic properties of the conditional field. In particular, we can pose two natural questions:

1. If $\mathbf{P}[X \in \cdot]$ is extremal in $\mathcal{G}(\gamma)$, when is $\mathbf{P}[X \in \cdot | Y]$ extremal in $\mathcal{G}(\gamma^Y)$ a.s.?

2. If $|\mathcal{G}(\gamma)| = 1$, when is $|\mathcal{G}(\gamma^Y)| = 1$ a.s.?

The first question is evidently the direct spatial analogue of the filter stability problem: when is the mixing property inherited by the conditional distribution? The second question is analogous, but for the uniform mixing property. It is evident from Theorem 5.4 that $|\mathcal{G}(\gamma^Y)| = 1$ a.s. implies the conditional mixing property. The stronger conclusion $|\mathcal{G}(\gamma^Y)| = 1$ a.s. is perhaps less natural from the point of view of conditional distributions, but is of practical relevance in its own right as it is closely connected with the computational complexity of MCMC methods for Bayesian image analysis [21].

As in the filter stability problem, local nondegeneracy of the observations does not suffice to obtain an affirmative answer to either of the above questions. In fact, we have a direct analogue of the example given in section 3.

Example 5.8. Let $E = F = \{-1, 1\}$, and define the random field $(X_v)_{v \in \mathbb{Z}^2}$ such that X_v are i.i.d. symmetric Bernoulli random variables. It is evident that this model is uniformly mixing in the most trivial sense (thus uniqueness and extremality both hold).

We now attach an observation $Y_{\{v,w\}}$ to each edge $\{v,w\} \subset \mathbb{Z}^d$, $\|v-w\| = 1$ by setting $Y_{\{v,w\}} = X_v X_w \xi_{\{v,w\}}$ with $\xi_{\{v,w\}}$ i.i.d. and independent of X with $\mathbf{P}[\xi_{\{v,w\}} = -1] = p$. In this manner, we evidently obtain a direct counterpart of the model of section 3. While the observations in this model are defined on the edges rather than on the vertices as we have done in this section, a result that is entirely analogous to Proposition 5.7 holds in this setting (see also Remark 5.5 above and Remark 5.9 below).

We can now proceed identically as in the proof of Theorem 3.1 to show that there exists $0 < p_\star < 1/2$ such that the hidden Markov random field (X, Y) fails to be conditionally mixing for $p < p_\star$. In fact, this is precisely the idea behind the proof of Theorem 3.1 in the first place: the model $(X_k^v, Y_k^v)_{k,v \in \mathbb{Z}}$ is considered as a space-time random field, and the problem is addressed using classical methods from statistical mechanics.

The present example could be considered as a toy model in image analysis. The underlying field X represents a grid of black or white pixels of an image, and the observations Y correspond to noisy measurements of the gradient of the image at each point. Thus we see that the ability to reconstruct the image based on the noisy gradient information undergoes a phase transition at a positive signal-to-noise ratio.

Remark 5.9. The use of edge observations in Example 5.8 is merely cosmetic: the same example can be reformulated in terms of vertex observations. Indeed, let us define the random field $(\tilde{X}_v, \tilde{Y}_v)_{v \in \mathbb{Z}^d}$ with $\tilde{X}_v \in \{-1, 1\}^3$ and $\tilde{Y}_v \in \{-1, 1\}^2$ by setting $\tilde{X}_v = (X_v, X_{v+(0,1)}, X_{v+(1,0)})$ and $\tilde{Y}_v = (X_v X_{v+(0,1)} \xi_{\{v,v+(0,1)\}}, X_v X_{v+(1,0)} \xi_{\{v,v+(1,0)\}})$, where X_v and $\xi_{\{v,w\}}$ are as in Example 5.8. Then \tilde{X} is still a uniformly mixing Markov random field, the observations \tilde{Y} are locally nondegenerate, and $\mathbf{P}[\tilde{X}^1 \in \cdot | \tilde{Y}] = \mathbf{P}[X \in \cdot | Y]$. In particular, the above conditional phase transition arises identically in this formulation.

In view of the above, the inheritance of mixing properties of random fields under conditioning cannot be taken for granted. Just as in the filter stability problem, however, it is natural to expect that conditional mixing will hold in the absence of observation symmetries. Such a conjecture is often implicit in work on Bayesian image analysis (cf. [21, p. 6]). For example, we can formulate the natural analogue of Conjecture 4.1.

Conjecture 5.10. Let $(X_v, Y_v)_{v \in \mathbb{Z}^2}$ be a hidden Markov field with $E = F = \{-1, 1\}$ and

$$Y_v = X_v \xi_v, \quad (\xi_v)_{v \in \mathbb{Z}^2} \text{ are i.i.d. } \perp X \text{ with } \mathbf{P}[\xi_v = -1] = p.$$

If the underlying random field X is mixing, then the model is conditionally mixing.

We do not know how to prove such a conjecture in a general setting. However, we will presently establish the validity of such a result under monotonicity assumptions on the underlying field. This provides an entirely different mechanism for the inheritance of ergodicity than the observability theory that was developed in section 4 above.

5.3 Monotonicity

The goal of this section is to prove a variant of Conjecture 5.10 under monotonicity assumptions on the underlying field. In such models, the direct observation structure $Y_v = X_v \xi_v$ ensures that monotonicity properties are preserved under conditioning on the observations, which greatly facilitates the analysis of the conditional random field. The arguments used here are directly inspired by the methods used in [26, 19, 17] to investigate the *global* Markov property of random fields.

In this section, we will assume that $E = F = \{-1, 1\}$. Fix a specification γ and a Markov random field $X = (X_v)_{v \in \mathbb{Z}^d}$ specified by γ , and introduce the observations

$$Y_v = X_v \xi_v, \quad (\xi_v)_{v \in \mathbb{Z}^d} \text{ are independent } \perp\!\!\!\perp X \text{ with } \mathbf{P}[\xi_v = -1] = p_v.$$

For simplicity, we will assume throughout this section that $0 < p_v \leq 1/2$ for all v .⁸

The following monotonicity property is the key assumption of this section. This property has various useful characterizations, cf. [28, Theorem 2.27]. Here a function $f : \{-1, 1\}^{\mathbb{Z}^d} \rightarrow \mathbb{R}$ is called increasing if $f(x) \geq f(z)$ whenever $x_v \geq z_v$ for all $v \in \mathbb{Z}^d$.

Definition 5.11. A specification γ for a $\{-1, 1\}$ -valued random field is called *monotone* if for every bounded increasing function f and $V \subset \subset \mathbb{Z}^d$, the function $\gamma_V f$ is increasing.

We can now formulate the main result of this section.

Theorem 5.12. *For the model of this section, suppose that the specification γ is monotone. Then $|\mathcal{G}(\gamma)| = 1$ implies that $|\mathcal{G}(\gamma^Y)| = 1$ a.s. In particular, under the monotonicity assumption, uniqueness of the underlying random field implies conditional mixing.*

Under the monotonicity assumption, Theorem 5.12 essentially resolves the analogue of Conjecture 5.10 for uniqueness rather than extremality (this is the special case $p_v = p \neq 1/2$ for all v). A partial result on the inheritance of extremality when the field is not unique can be deduced from the proof as well, see Remark 5.17 below.

Nonetheless, the result of Theorem 5.12 is arguably quite different in spirit from Conjecture 5.10, as it does not require the absence of observation symmetries. For example, we could choose the noise parameters p_v such that $p_v = \frac{1}{2}$ for alternating sites in the lattice \mathbb{Z}^d : then there is certainly a large class of observation symmetries, as only every other site in the lattice is observed. Thus Theorem 5.12 is not addressing a symmetry-breaking phenomenon of the type that motivated the observability conjecture in section 4. On the other hand, as will become clear in the proof of Theorem 5.12, the present approach directly implements the idea that ergodicity is inherited from the underlying model to the conditional distribution, in contrast with the observability theory of section 4 which does not exploit at all the ergodic properties of the model.

⁸The assumption $p_v > 0$ ensures local nondegeneracy, which is convenient in view of Proposition 5.7 but is not essential for the results in this section. The assumption $p_v \leq 1/2$ does not entail any loss of generality: if $p_v > 1/2$, we can reduce to $p_v < 1/2$ by considering the inverted observation $-Y_v$ instead of Y_v .

Remark 5.13. It would be interesting to obtain a counterpart to Theorem 5.12 for the filter stability problem. Unfortunately, this does not appear to be possible, at least in the setting of section 2.2. The proof of Theorem 5.12 relies on the fact that the conditional distributions of the random field given the observations are monotone, which is essentially equivalent to the validity of Definition 5.11. However, in the filtering model of section 2.2, the associated space-time random field must generally fail to be monotone in the sense of Definition 5.11, cf. [32] for a discussion in the continuous time setting. While the space-time distributions of interacting particle systems with monotone transition probabilities do satisfy some weaker monotonicity properties, cf. [30, section 3.1] or [32], such weaker properties do not suffice for the proof of Theorem 5.12.

We now turn to the proof of Theorem 5.12. We begin by establishing the inheritance of monotonicity by the conditional specification γ^y .

Lemma 5.14. *Suppose that γ is monotone. Then the following hold:*

1. γ^y is monotone for every $y \in \{-1, 1\}^{\mathbb{Z}^d}$.
2. $y \mapsto \gamma_V^y f$ is increasing for every $V \subset \subset \mathbb{Z}^d$ and bounded increasing f .

Proof. It will be convenient to write

$$g_v(x_v, y_v) = \mathbf{P}[Y_v = y_v | X_v = x_v] = \sqrt{p_v(1-p_v)} e^{\beta_v y_v x_v}$$

with $\beta_v = \log \sqrt{(1-p_v)/p_v} \geq 0$ in Proposition 5.7.

Define for every $y \in F^{\mathbb{Z}^d}$ and $W, V \subset \subset \mathbb{Z}^d$ the transition kernel on $E^{\mathbb{Z}^d}$

$$\gamma_{V,W}^y(x, A) = \frac{\int \mathbf{1}_A(z) \prod_{w \in W \cap V} e^{\beta_w y_w z_w} \gamma_V(x, dz)}{\int \prod_{w \in W \cap V} e^{\beta_w y_w z_w} \gamma_V(x, dz)}.$$

Evidently $\gamma_V^y = \gamma_{V,V}^y$ and $\gamma_V = \gamma_{V,\emptyset}^y$, and for simplicity we will write $\gamma_{\emptyset,W}^y f = f$. We now prove that $(\gamma_{V,W}^y f)(x)$ is increasing in x, y for every bounded increasing function f by induction on $|V| + |W|$. The claim is obviously true for $|V| + |W| = 1$ by the monotonicity of γ . In the remainder of the proof, we suppose that the claim is true for all V, W such that $|V| + |W| = m$, and we proceed with the induction step.

Fix V, W such that $|V| + |W| = m + 1$. The claim is trivially true if $W \cap V = \emptyset$ by the monotonicity of γ . Otherwise, fix $v \in W \cap V$ and a bounded increasing function f . Then

$$\gamma_{V,W}^y f = \gamma_{V,W}^y \gamma_{V \setminus \{v\}, W}^y f$$

holds by the same argument as in the proof of Proposition 5.7. Moreover, by the induction hypothesis, $(\gamma_{V \setminus \{v\}, W}^y f)(x)$ is increasing in x, y , and by construction $(\gamma_{V \setminus \{v\}, W}^y f)(x)$ depends only on x_v and x_{V^c}, y . To show that $(\gamma_{V,W}^y f)(x)$ is increasing in x, y , it thus suffices to show that $\gamma_{V,W}^y(x, A)$ is increasing in x, y for $\mathbf{1}_A(x) = \mathbf{1}_{x_v=1}$. But note that

$$\frac{\gamma_{V,W}^y(x, A)}{1 - \gamma_{V,W}^y(x, A)} = e^{2\beta_v y_v} \frac{\gamma_{V,W \setminus \{v\}}^y(x, A)}{1 - \gamma_{V,W \setminus \{v\}}^y(x, A)}.$$

As $\gamma_{V,W \setminus \{v\}}^y(x, A)$ is increasing in x, y by the induction hypothesis, $\gamma_{V,W}^y(x, A)$ is also increasing in x, y . Thus the induction step is established, and the proof is complete. \square

A classical consequence of the monotonicity of γ is that it implies that $\mathcal{G}(\gamma)$ has a maximal and a minimal element. This substantially simplifies the characterization of uniqueness. The following standard result (see, for example, [5, section 4.3.3]) makes this idea precise. In the sequel, we define the maximal and minimal configurations $+, - \in \{-1, 1\}^{\mathbb{Z}^d}$ as $+_v = 1$ and $-_v = -1$ for all $v \in \mathbb{Z}^d$; moreover, a function $f : \{-1, 1\}^{\mathbb{Z}^d} \rightarrow \mathbb{R}$ is said to be local if it depends only on a finite number of coordinates.

Lemma 5.15. *Suppose that γ is monotone. There exist laws $\mathbf{P}_+, \mathbf{P}_-$ in $\mathcal{G}(\gamma)$ such that*

$$(\gamma_V f)(-) \uparrow \mathbf{E}_-[f(X)] \quad \text{and} \quad (\gamma_V f)(+) \downarrow \mathbf{E}_+[f(X)] \quad \text{as} \quad V \uparrow \mathbb{Z}^d$$

for every local increasing function f . Moreover,

$$\mathbf{E}_-[f(X)] \leq \mathbf{E}[f(X)] \leq \mathbf{E}_+[f(X)]$$

for every \mathbf{P} in $\mathcal{G}(\gamma)$ and local increasing function f . In particular, $|\mathcal{G}(\gamma)| = 1$ iff $\mathbf{P}_+ = \mathbf{P}_-$.

Proof. Let f be a bounded increasing function and let $W \subset V \subset \subset \mathbb{Z}^d$. As $\gamma_W f$ is increasing and $\gamma_V = \gamma_V \gamma_W$, we readily obtain $\gamma_W f(-) \leq \gamma_V f \leq \gamma_W f(+)$. Thus the net $(\gamma_V f(-))_{V \subset \subset \mathbb{Z}^d}$ is increasing and the net $(\gamma_V f(+))_{V \subset \subset \mathbb{Z}^d}$ is decreasing.

Now note that $\{-1, 1\}^{\mathbb{Z}^d}$ is compact and metrizable (for the product topology). Thus $(\gamma_V(-, \cdot))_{V \subset \subset \mathbb{Z}^d}$ and $(\gamma_V(+, \cdot))_{V \subset \subset \mathbb{Z}^d}$ are precompact for the weak convergence topology. In particular, we can choose a cofinal increasing sequence $V_n \subset \subset \mathbb{Z}^d$ such that $\gamma_{V_n}(-, \cdot) \rightarrow \mathbf{P}_-$ and $\gamma_{V_n}(+, \cdot) \rightarrow \mathbf{P}_+$ weakly for some probability measures \mathbf{P}_- and \mathbf{P}_+ , respectively. By the above monotonicity, it follows readily that

$$\lim_{V \subset \subset \mathbb{Z}^d} (\gamma_V f)(-) = \mathbf{E}_-[f(X)] \quad \text{and} \quad \lim_{V \subset \subset \mathbb{Z}^d} (\gamma_V f)(+) = \mathbf{E}_+[f(X)]$$

for every local increasing function f . Moreover, for any measure \mathbf{P} in $\mathcal{G}(\gamma)$, we have

$$\mathbf{E}_-[f(X)] = \lim_{V \subset \subset \mathbb{Z}^d} (\gamma_V f)(-) \leq \lim_{V \subset \subset \mathbb{Z}^d} \mathbf{E}[(\gamma_V f)(X)] = \mathbf{E}[f(X)]$$

for every local increasing function f , and similarly for the upper bound.

We now argue that \mathbf{P}_- and \mathbf{P}_+ are in fact in $\mathcal{G}(\gamma)$. To this end, fix any $W \subset \subset \mathbb{Z}^d$. As $W \subset V_n$ for all n sufficiently large (as $\{V_n\}$ is cofinal), and as $\gamma_W f$ is local if f is local (by the Markov property of the random field), we can write

$$\mathbf{E}_-[g(X_{W^c}) (\gamma_W f)(X)] = \lim_{n \rightarrow \infty} (\gamma_{V_n} \gamma_W f g)(-) = \lim_{n \rightarrow \infty} (\gamma_{V_n} f g)(-) = \mathbf{E}_-[g(X_{W^c}) f(X)]$$

for all local functions f, g . Thus $\mathbf{E}_-[f(X)|X_{W^c}] = (\gamma_W f)(X)$ for all $W \subset \subset \mathbb{Z}^d$, which implies that \mathbf{P}_- is in $\mathcal{G}(\gamma)$. The conclusion for \mathbf{P}_+ follows identically.

Finally, we argue that $|\mathcal{G}(\gamma)| = 1$ iff $\mathbf{P}_+ = \mathbf{P}_-$. If $\mathbf{P}_+ \neq \mathbf{P}_-$, then evidently $|\mathcal{G}(\gamma)| \geq 2$. On the other hand, if $\mathbf{P}_+ = \mathbf{P}_-$, then the expectation of every local increasing function must coincide for all γ -specified random fields. But then all γ -specified random fields coincide, as the local increasing functions are measure-determining (the moment generating function $\mathbf{E}[e^{\lambda_1 X_{v_1} + \dots + \lambda_m X_{v_m}}]$ determines uniquely the joint law of X_{v_1}, \dots, X_{v_m} , and $f(x_{v_1}, \dots, x_{v_m}) = e^{\lambda_1 x_{v_1} + \dots + \lambda_m x_{v_m}}$ is local and increasing for every $\lambda_1, \dots, \lambda_m \geq 0$). \square

We call \mathbf{P}_+ and \mathbf{P}_- the maximal and minimal elements of $\mathcal{G}(\gamma)$. Now recall that if γ is monotone, then so is γ^y . Thus $\mathcal{G}(\gamma^y)$ also has a maximal and a minimal element. The key step in the proof of Theorem 5.12 is the following observation due to Föllmer [19].

Lemma 5.16. *Suppose that γ is monotone. Let \mathbf{P}_+ and \mathbf{P}_- be the maximal and minimal element of $\mathcal{G}(\gamma)$, and let \mathbf{P}_+^y and \mathbf{P}_-^y be the maximal and minimal element of $\mathcal{G}(\gamma^y)$. Then \mathbf{P}_+^Y is a version of $\mathbf{P}_+[X \in \cdot | Y]$ and \mathbf{P}_-^Y is a version of $\mathbf{P}_-[X \in \cdot | Y]$.*

Proof. Let us prove the result for \mathbf{P}_+ ; the conclusion for \mathbf{P}_- follows in the same manner. Let f, g be local increasing functions. It suffices to show that

$$\mathbf{E}_+[\mathbf{E}_+^Y[f(X)]g(Y)] = \mathbf{E}_+[f(X)g(Y)]$$

for all local increasing functions f, g . By Lemma 5.15

$$\mathbf{E}_+[\mathbf{E}_+^Y[f(X)]g(Y)] = \inf_{V \subset \mathbb{C}\mathbb{Z}^d} \mathbf{E}_+[(\gamma_V^Y f)(+)g(Y)].$$

Now note that $(\gamma_V^Y f)(+)$ is a local increasing function of y by Lemma 5.14. It is easily verified (as $p_v \leq \frac{1}{2}$ for all v) that $\mathbf{E}_+[h(Y_V)|X] = \mathbf{E}_+[h(Y_V)|X_V]$ is an increasing function of X_V for every increasing function h . Thus we obtain using Lemma 5.15

$$\begin{aligned} \mathbf{E}_+[\mathbf{E}_+^Y[f(X)]g(Y)] &= \inf_{V \subset \mathbb{C}\mathbb{Z}^d} \mathbf{E}_+[\mathbf{E}_+[(\gamma_V^Y f)(+)g(Y)|X]] \\ &= \inf_{V \subset \mathbb{C}\mathbb{Z}^d} \inf_{W \subset \mathbb{C}\mathbb{Z}^d} \int \mathbf{E}_+[(\gamma_V^Y f)(+)g(Y)|X=x] \gamma_W(+, dx) \\ &= \inf_{V \subset \mathbb{C}\mathbb{Z}^d} \int \mathbf{E}_+[(\gamma_V^Y f)(+)g(Y)|X=x] \gamma_V(+, dx), \end{aligned}$$

where the last equality follows as the quantity inside the infimum is decreasing in both V and W . But note that we have for every V sufficiently large that $g(y) = g(y_V)$

$$\begin{aligned} &\int \mathbf{E}_+[(\gamma_V^Y f)(+)g(Y)|X=x] \gamma_V(+, dx) \\ &= \sum_{y \in \{-1, +1\}^V} (\gamma_V^y f)(+)g(y) \int \prod_{v \in V} g(x_v, y_v) \gamma_V(+, dx) \\ &= \int \sum_{y \in \{-1, +1\}^V} f(x)g(y) \prod_{v \in V} g(x_v, y_v) \gamma_V(+, dx) \\ &= \int f(x) \mathbf{E}_+[g(Y)|X=x] \gamma_V(+, dx). \end{aligned}$$

Taking the infimum over V , it follows that

$$\mathbf{E}_+[\mathbf{E}_+^Y[f(X)]g(Y)] = \mathbf{E}_+[f(X) \mathbf{E}_+[g(Y)|X]] = \mathbf{E}_+[f(X)g(Y)],$$

and the proof is complete. □

We can now easily complete the proof of Theorem 5.12.

Proof of Theorem 5.12. By Lemma 5.15, $|\mathcal{G}(\gamma)| = 1$ implies $\mathbf{P}_+ = \mathbf{P}_-$. But then

$$\mathbf{P}_+^Y = \mathbf{P}_+[X \in \cdot | Y] = \mathbf{P}_-[X \in \cdot | Y] = \mathbf{P}_-^Y \quad \text{a.s.}$$

by Lemma 5.16. Thus $|\mathcal{G}(\gamma^Y)| = 1$ a.s. by Lemma 5.15. □

Remark 5.17. Even when the underlying field is not unique, that is, when $\mathbf{P}_+ \neq \mathbf{P}_-$, Lemma 5.16 immediately shows that the maximal and minimal models \mathbf{P}_+ and \mathbf{P}_- are both conditionally mixing. Indeed, Lemma 5.16 implies that the conditional distributions $\mathbf{P}_+[X \in \cdot | Y]$ and $\mathbf{P}_-[X \in \cdot | Y]$ are maximal and minimal in $\mathcal{G}(\gamma^Y)$ a.s., and must therefore be extremal a.s. (so that conditional mixing follows by Proposition 5.7).

Lemma 5.15 might lead one to expect that if γ is monotone, then \mathbf{P}_+ and \mathbf{P}_- are the only extremal random fields in $\mathcal{G}(\gamma)$. If a monotone specification γ satisfies this property, then we have shown that mixing implies conditional mixing, that is, we have resolved Conjecture 5.10 for monotone random fields. However, it turns out that there may exist extremal random fields defined by monotone specifications that differ from $\mathbf{P}_+, \mathbf{P}_-$. For example, for the ferromagnetic Ising model, the situation is as follows. On the two-dimensional lattice \mathbb{Z}^2 , the maximal and minimal fields $\mathbf{P}_+, \mathbf{P}_-$ are the only extremal fields. However, in higher dimensions \mathbb{Z}^d , $d \geq 3$ there exist extremal fields other than $\mathbf{P}_+, \mathbf{P}_-$. On the other hand, in any dimension, $\mathbf{P}_+, \mathbf{P}_-$ are the only *translation-invariant* extremal fields. These are highly nontrivial results, and we refer to [11] and the references therein for further details. By combining these facts, we can resolve several variants of Conjecture 5.10 for the ferromagnetic Ising model. However, the situation for general monotone fields remains unclear.

6 Discussion

Our main aim in this paper has been to draw attention to the existence of surprising probabilistic phenomena in nonlinear filtering that arise in infinite dimension, and that are fundamentally different in nature from the problems that have been investigated in filtering theory to date. In particular, we have exhibited the existence of conditional phase transitions, and we made some first steps toward a general theory. It is only fair to emphasize, however, that none of our general results are entirely satisfactory, and it appears in particular that we are far from understanding the issues surrounding Conjectures 4.1 and 5.10 at any level of generality. It is likely that new ideas are needed to develop a deeper understanding of these questions.

Throughout this paper, we have repeatedly exploited connections between filtering in infinite dimension and problems in statistical mechanics and in multidimensional ergodic theory. In this final section, we will briefly discuss some further connections with other probabilistic problems that did not arise in the previous sections.

6.1 Measure-theoretic identities and conditional ergodicity

In this paper we have formulated the filter stability problem as one of inheritance of ergodicity. It is insightful, however, to consider the problem from a different perspective: following Kunita [29], the filter stability problem can be equivalently rephrased in terms of the validity of a simple *measure-theoretic* identity (see [52] for further discussion and references). In view of this connection, the counterexamples given above provide some new insight into the failure of such measure-theoretic identities.

We discuss this issue briefly here. Let us begin by rephrasing the filter stability property.

Let $(X_k, Y_k)_{k \in \mathbb{Z}}$ be a stationary hidden Markov model as in section 2.1. Then

$$\begin{aligned} \mathbf{E}|\mathbf{P}[X_k \in A|X_0, Y_1, \dots, Y_k] - \mathbf{P}[X_k \in A|Y_1, \dots, Y_k]| &= && \text{(stationarity)} \\ \mathbf{E}|\mathbf{P}[X_0 \in A|X_{-k}, Y_{-k+1}, \dots, Y_0] - \mathbf{P}[X_0 \in A|Y_{-k+1}, \dots, Y_0]| &= && \text{(Markov property)} \\ \mathbf{E}|\mathbf{P}[X_0 \in A|\mathcal{X}_{-\infty}^{-k} \vee \mathcal{Y}_{-\infty}^0] - \mathbf{P}[X_0 \in A|\mathcal{Y}_{-k+1}^0]| &\xrightarrow{k \rightarrow \infty} && \text{(martingale cvg.)} \\ \mathbf{E}|\mathbf{P}[X_0 \in A|\bigcap_k(\mathcal{X}_{-\infty}^{-k} \vee \mathcal{Y}_{-\infty}^0)] - \mathbf{P}[X_0 \in A|\mathcal{Y}_{-\infty}^0]|, & & & \end{aligned}$$

where we define here and below the σ -fields

$$\mathcal{X}_m^n := \sigma\{X_m, \dots, X_n\}, \quad \mathcal{Y}_m^n := \sigma\{Y_m, \dots, Y_n\}.$$

Thus the filter is stable if and only if

$$\mathbf{P}[X_0 \in \cdot | \bigcap_k(\mathcal{X}_{-\infty}^{-k} \vee \mathcal{Y}_{-\infty}^0)] = \mathbf{P}[X_0 \in \cdot | \mathcal{Y}_{-\infty}^0].$$

In particular, validity of the measure-theoretic identity

$$\bigcap_k(\mathcal{X}_{-\infty}^{-k} \vee \mathcal{Y}_{-\infty}^0) = \mathcal{Y}_{-\infty}^0 \quad \text{mod } \mathbf{P}$$

is sufficient for filter stability. In many cases, this identity can also be shown to be necessary [52]. In precisely the same manner, it is not difficult to show that the unobserved Markov chain $(X_k)_{k \in \mathbb{Z}}$ is stable if and only if the tail σ -field

$$\bigcap_k \mathcal{X}_{-\infty}^{-k} \quad \text{is } \mathbf{P}\text{-trivial.}$$

Thus we can formulate a measure-theoretic version of the filter stability problem:

$$\text{When does } \mathbf{P}\text{-triviality of } \bigcap_k \mathcal{X}_{-\infty}^{-k} \text{ imply } \bigcap_k(\mathcal{X}_{-\infty}^{-k} \vee \mathcal{Y}_{-\infty}^0) = \mathcal{Y}_{-\infty}^0 \text{ mod } \mathbf{P}?$$

It is tempting to conclude that this is always the case: as $\mathcal{Y}_{-\infty}^0$ does not depend on k , one would expect that $\bigcap_k(\mathcal{X}_{-\infty}^{-k} \vee \mathcal{Y}_{-\infty}^0) \stackrel{?}{=} (\bigcap_k \mathcal{X}_{-\infty}^{-k}) \vee \mathcal{Y}_{-\infty}^0$ and thus the filter stability property would automatically follow from stability of the underlying model. This (incorrect) reasoning was used by Kunita [29] in the proof of his main result. The conclusion is already contradicted, however, by Example 2.1! The fundamental issue is that *the exchange of intersection \cap and supremum \vee of σ -fields is not permitted* in general.

An entirely analogous measure-theoretic formulation appears in the random field setting of section 5. Indeed, let $(X_v, Y_v)_{v \in \mathbb{Z}^d}$ be a partially observed random field model as in section 5.2, and define $\mathcal{X}_V := \sigma\{X_v : v \in V\}$ and $\mathcal{Y}_V := \sigma\{Y_v : v \in V\}$ for $V \subseteq \mathbb{Z}^d$. Then the question whether mixing implies conditional mixing can be phrased as:

When does

$$\bigcap_{V \subset \subset \mathbb{Z}^d} \mathcal{X}_{\mathbb{Z}^d \setminus V} \text{ is } \mathbf{P}\text{-trivial}$$

imply

$$\bigcap_{V \subset \subset \mathbb{Z}^d} (\mathcal{X}_{\mathbb{Z}^d \setminus V} \vee \mathcal{Y}_{\mathbb{Z}^d}) = \mathcal{Y}_{\mathbb{Z}^d} \text{ mod } \mathbf{P}?$$

Once again, Example 2.1 already shows this is not always the case (even for $d = 1$).

Establishing the validity of the exchange of intersection and supremum of σ -fields is a measure-theoretic conundrum that poses a tantalizing problem in a diverse range of probabilistic questions [53, 52]. As this problem evidently lies at the heart of the ergodic theory of nonlinear filters, it is interesting to view results in this area from a measure-theoretic perspective. For example, as the inheritance of ergodicity has been established under nondegeneracy of the observations (Theorem 2.2 and generalizations developed in [46]), one might ask whether this result is a manifestation of a more general measure-theoretic principle that enables the exchange of intersection and supremum of σ -fields. Such a hope could be justified as follows. If \mathcal{G}_n is a decreasing filtration and \mathcal{F} is a σ -field, it is easily established that $\bigcap_n(\mathcal{G}_n \vee \mathcal{F}) = (\bigcap_n \mathcal{G}_n) \vee \mathcal{F} \bmod \mathbf{P}$ when \mathcal{G}_0 and \mathcal{F} are independent. Using the Bayes formula, one can readily deduce that the conclusion still follows if $\mathbf{P} \ll \mathbf{Q}$ such that \mathcal{G}_0 and \mathcal{F} are independent under \mathbf{Q} . The nondegeneracy assumption, however, implies a weaker property: it implies that the laws of $(X_k)_{|k| \leq K}$ and $(Y_k)_{|k| \leq K}$ can be made independent by an equivalent change of measure for every *finite* K (but not for infinite K as would be required to apply the above principle). Thus Theorem 2.2 might suggest the existence of a more general *local* form of the above measure-theoretic principle in the setting of stationary processes.

Unfortunately, this hope proves to be unfounded: a counterexample in [52] shows that the exchange of intersection and supremum may fail even when the observations are nondegenerate (note that this does not contradict Theorem 2.2, as its uniform ergodicity assumption is strictly stronger than the stability property: thus the stronger hypothesis is essential to utilize nondegeneracy). However, the counterexample of [52] is decidedly esoteric. In contrast, the results of the present paper show that the validity of the exchange of intersection and supremum of σ -fields can fail in a very robust manner in infinite dimension. For example, Example 5.8 defines an almost trivial model of a stationary and ergodic random field with values in a finite set for which all finite-dimensional marginals admit a density that is bounded away from zero, while the validity of exchange of intersection and supremum of σ -fields exhibits a phase transition. These new measure-theoretic examples illustrate the highly nontrivial nature of the exchange of intersection and supremum problem even in very regular models.

6.2 Spin glasses and conditional Gibbs measures

We have seen in section 5.2 that when a Markov random field is conditioned on partial observations, the conditional distribution again defines a Markov random field, albeit with a *random* specification (through dependence on the observations whose realization is random). This idea has played a central role throughout this paper, both in section 5 where the objects of interest were the conditional random fields themselves, and in sections 3 and 4 where the space-time random field was a basic tool in the investigation of the corresponding filtering problems in infinite dimension.

Random fields whose specification depends on external random variables have been widely studied in the statistical mechanics of disordered systems. The randomness, which is typically assumed to be i.i.d. across sites, is used here to model the irregular spatial distribution of impurities in materials such as magnetic alloys. Such disordered materials, known as *spin glasses*, exhibit many remarkable properties and complex behavior; we refer to [5] for an introduction to the mathematical study of spin glasses. One of the chief difficulties in spin glasses is that the random interactions in the system do not favor a

particular alignment of the configurations at different sites, a phenomenon known as ‘frustration’. This creates a very complicated energy landscape that gives rise to many unusual properties of spin glasses, with a large number of mathematical mysteries remaining to be understood (see, e.g., [5, section 6.5]).

In a sense, our conditional random fields could be interpreted as a form of spin glass models. For example, if the underlying field X is an Ising model (cf. Example 5.3), then the conditional random field in Example 5.8 can be viewed as a form of *random bond Ising model* or *Ising spin glass*, while the setting of Conjecture 5.10 leads to a form of *random field Ising model*, cf. [5, 23]. There is an essential difference, however, with the usual definition of a spin glass: the random variables that enter the conditional specification are not defined by an exogenous i.i.d. disorder, but rather by observations whose law is generated by the underlying random field. Thus the statistics of the randomness is perfectly matched to the underlying model. In particular, the conditional distribution should favor the same configurations as the underlying model, so that one might expect that the effect of frustration is mitigated. It is not clear, however, how such ideas could be made precise. More generally, this highlights the fundamental challenge in investigating conditional random field problems from the point of view of spin glasses: while conditional random fields are expected to exhibit much ‘nicer’ behavior than general spin glasses due to the precise matching of the randomness to the model, it is entirely unclear how the latter can be taken into account in the analysis of such problems.

Remark 6.1. It should be noted that the above intuition for the special nature of the problems considered in this paper fails to hold if we were to consider model misspecification. For example, if the conditional random field in Example 5.8 is computed assuming an error probability p that is misspecified with respect to the true error probability of the underlying observations, then there is no reason to expect improved behavior with respect to general spin glasses. The misspecified setting is completely unnatural from the point of view of conditional distributions: if \mathbf{P}_p denotes the law of the model with error probability p , then $\mathbf{P}_p[X \in \cdot | Y]$ is not even defined under \mathbf{P}_q for $p \neq q$ as then $\mathbf{P}_p \perp \mathbf{P}_q$. Such problems are therefore of a fundamentally different nature than those considered in this paper. On the other hand, one can never expect to have perfect knowledge of the underlying model in real-world applications, so that the present example highlights the need for caution in the practical interpretation of our results.

Remark 6.2. Unlike in most spin glass models, the observations that enter the specification of the conditional random field fail to be independent in almost any model of interest. However, in models that possess certain symmetries, such as Example 5.8, it is possible to transform the conditional random field to a spin glass model with i.i.d. randomness by means of a gauge transformation. This fact was exploited implicitly in the proofs in section 3 to analyze the phase transition behavior in this model. The converse observation has been noted in the physics literature: in spin glasses that possess gauge symmetries, it is possible to make a special choice of the randomness so that various quantities become explicitly computable. This restricts consideration to a distinguished subset of the parameter space of the model called the *Nishimori line*, which corresponds precisely to the situation where the model describes a conditional distribution (after gauge transformation). We refer to [39] and the references therein.

An entirely different manner in which conditional random fields arise in statistical mechanics is in the study of renormalization group transformations. In section 5 we have

investigated Markov random fields where $\mathbf{P}[X_V \in \cdot | X_{V^c}]$ depends on $X_{\partial V}$ only. In renormalization group analyses, one aims to perform computations at the level of a transformation Y of the random field X obtained, for example, by working on a sublattice. In general, such a transformation is no longer Markov, that is, $\mathbf{P}[Y_W \in \cdot | Y_{W^c}]$ depends on all sites outside $W \subset \subset \mathbb{Z}^d$. However, one still expects that the resulting model is *local* in the sense that $\mathbf{P}[Y_W \in \cdot | Y_{W^c}]$ only depends significantly on Y_w for w in some finite neighborhood of W . This idea is formalized by the requirement that the functions $y \mapsto \mathbf{P}[Y_W \in A | Y_{W^c} = y_{W^c}]$ are quasilocal, that is, uniform limits of local functions, for every $W \subset \subset \mathbb{Z}^d$. Such random fields are called *Gibbsian*. In order for the renormalization procedure to be well defined, one therefore faces the question whether a transformation Y of a Markov random field X is guaranteed to be Gibbsian.

Surprisingly, this natural property proves to be false in general: it may be that Y fails to be Gibbsian even when X is Markov. Such *renormalization group pathologies* are discussed at great length in the paper [47]. This issue proves to be closely connected, at least in monotone fields, with the uniqueness property of the conditional random field $\mathbf{P}[X \in \cdot | Y]$, and thus with the type of phenomena investigated in section 5. We refer for further details to the paper [17], which also develops the connection with global Markov property [26, 19] (already mentioned in the proof of Theorem 5.12).

6.3 Trees and other graphs

Throughout this paper, we have considered spatial variables indexed by the lattice \mathbb{Z}^d . However, both infinite-dimensional filtering models and partially observed Markov random fields can be meaningfully formulated on any infinite graph of finite degree $G = (V, \mathcal{E})$. While the lattice setting is a natural prototype for the investigation of infinite-dimensional systems, the consideration of other graphical structures could provide complementary insights on the phenomena considered in this paper.

Of particular interest is the case where the spatial graph G is a tree. Such models are motivated by applications in information theory, statistical physics, and mathematical biology, cf. [16] and the references therein. More importantly, trees play a special role in the theory of Markov random fields, as many problems that are exceedingly difficult in general graphs are amenable to exact computations on a tree by means of recursive formulas (see, for example, [22, Chapter 12]). It is therefore possible that significant insight could be obtained on conjectures that arise from this paper by investigating these on trees using the special tools that are available in this setting. We have not systematically pursued this avenue of investigation. Nonetheless, the statistical mechanics literature on spin glass models on Bethe lattices (infinite regular trees in which every vertex has exactly $q \geq 3$ neighbors) already provides some suggestive evidence towards questions that arise from this paper, as we will now briefly discuss.

1. The main conjecture of this paper is that stability properties are preserved by conditioning in the absence of symmetries. However, the theory of section 4 raises the question whether stability of the underlying model even plays a role in this setting: it is possible that the stability of conditional distributions follows already from the symmetry-breaking assumption alone even when the underlying model is not stable. On the other hand, the theory of section 5.3 does require stability of the model, suggesting that the latter should play a role after all.

Some insight into this question can be obtained from results in [4] on the random field Ising model on the Bethe lattice, which can be interpreted as the conditional random

field for the variant of the model of section 5.3 where the underlying field X is a ferromagnetic Ising model on the Bethe lattice. It is shown in [4] that if the underlying field X is not unique and the error probability p is sufficiently close to $\frac{1}{2}$, then the conditional random field is also not unique for any realization of the observations. Thus the conclusion of Theorem 5.12 fails in this case, even though the symmetry-breaking condition holds here and the model is monotone and translation invariant (in the sense appropriate to the Bethe lattice). Thus, at least in the setting of trees, the symmetry-breaking property alone is not sufficient to ensure the inheritance of ergodicity in the sense of uniform mixing. This provides some evidence towards the necessity of the stability assumption in our conjectures.

2. Section 5 introduced two different notions of conditional ergodicity for Markov random fields: uniform mixing (uniqueness) and mixing (extremality) of the conditional random field. It is not clear whether inheritance of the mixing property is connected to inheritance of the uniform mixing property: for example, is it possible for the conditional distribution of a uniformly mixing random field to be mixing but not uniformly mixing? While we do not have any concrete evidence in the present setting, related results on the ferromagnetic Ising model on the Bethe lattice [3] show that it is indeed possible, albeit in a different setting, that the uniqueness and extremality properties appear at different phase transition points.
3. The conditional random field for the analogue of Example 5.8 on the Bethe lattice can be viewed as a random bond Ising model on the Bethe lattice. This model has been systematically investigated in the statistical mechanics literature [8, 6, 7], which leads to a detailed description of the phase diagram and, in particular, the exact location of the phase transition point on the Nishimori line (cf. Remark 6.2).

Another widely studied class of models in statistical mechanics are mean-field models, where the spatial graph is chosen to be the complete graph on a finite number n of spatial locations (that is, every pair of sites interacts with each other). The large-scale properties of such models do not arise at the level of a fixed graph, but rather in the limit as $n \rightarrow \infty$ where the strength of the interactions are scaled down appropriately to obtain nontrivial limiting behavior. Mean-field analogues of Example 5.8 arise naturally in applications in statistics and computer science [38] and in information theory [36, 37]. However, it is arguably less surprising that nontrivial phase transition phenomena arise in this setting, as these problems are concerned with the limiting behavior of a sequence of conditional distributions where the underlying model is scaled in a nontrivial manner. In contrast, in the problems we have investigated in this paper, the emergent phase transition phenomena arise from conditioning alone.

6.4 Stability in time vs. spatial mixing

We have discussed in this paper two different questions of inheritance of ergodicity: the stability property in time of dynamical filtering models, and the mixing property in space in the setting of static random fields. While we have utilized the notion of a space-time random field even in the dynamical setting, it is not at all clear to what extent the problems of inheritance of spatial and temporal ergodicity are related to one another. In particular, we do not know whether it is possible that there exist filtering models for which the filter is stable, but does not also exhibit a spatial mixing property. From the practical perspective,

the design of filtering algorithms in high dimension relies crucially both on the temporal stability and spatial mixing of the filter [40].

Connections between spatial and temporal mixing properties arise frequently in the ergodic theory of interacting Markov chains. For example, [27] shows a correspondence in the translation-invariant setting between invariant measures of an infinite-dimensional Markov chain and random fields corresponding to the specification of the space-time system (related ideas in the context of the global Markov property can be found in [25, 20]). Similarly, connections between exponential mixing in space and time are well known, see for example [33] and the references therein. We do not know, however, whether similar connections are likely to arise for conditional distributions. Let us emphasize in this context that the arguments that were used in this paper to obtain filter stability (section 4) and conditional mixing (section 5) are of a very different nature, but this is likely to be a limitation of our analysis rather than a genuine phenomenon.

References

- [1] David Blackwell, *The entropy of functions of finite-state Markov chains*, Transactions of the first Prague conference on information theory, Statistical decision functions, random processes held at Liblice near Prague from November 28 to 30, 1956, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1957, pp. 13–20. MR 0100297 (20 #6730)
- [2] David Blackwell and Lester Dubins, *Merging of opinions with increasing information*, Ann. Math. Statist. **33** (1962), 882–886. MR 0149577 (26 #7062)
- [3] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov, *On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice*, J. Statist. Phys. **79** (1995), no. 1-2, 473–482. MR 1325591 (96d:82009)
- [4] ———, *On the phase diagram of the random field Ising model on the Bethe lattice*, J. Statist. Phys. **93** (1998), no. 1-2, 33–78. MR 1656364 (99h:82043)
- [5] Anton Bovier, *Statistical mechanics of disordered systems*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2006, A mathematical perspective. MR 2252929 (2007g:82028)
- [6] J. M. Carlson, J. T. Chayes, L. Chayes, J. P. Sethna, and D. J. Thouless, *Bethe lattice spin glass: the effects of a ferromagnetic bias and external fields. I. Bifurcation analysis*, J. Statist. Phys. **61** (1990), no. 5-6, 987–1067. MR 1083895 (92d:82050)
- [7] J. M. Carlson, J. T. Chayes, J. P. Sethna, and D. J. Thouless, *Bethe lattice spin glass: the effects of a ferromagnetic bias and external fields. II. Magnetized spin-glass phase and the de Almeida-Thouless line*, J. Statist. Phys. **61** (1990), no. 5-6, 1069–1084. MR 1083896 (92d:82051)
- [8] J. T. Chayes, L. Chayes, James P. Sethna, and D. J. Thouless, *A mean field spin glass with short-range interactions*, Comm. Math. Phys. **106** (1986), no. 1, 41–89. MR 853978 (88i:82054)
- [9] Pavel Chigansky and Ramon van Handel, *A complete solution to Blackwell’s unique ergodicity problem for hidden Markov chains*, Ann. Appl. Probab. **20** (2010), no. 6, 2318–2345. MR 2759736 (2011j:93118)

- [10] J. P. Conze, *Entropie d'un groupe abélien de transformations*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **25** (1972/73), 11–30. MR 0335754 (49 #534)
- [11] Loren Coquille and Yvan Velenik, *A finite-volume version of Aizenman-Higuchi theorem for the 2d Ising model*, Probab. Theory Related Fields **153** (2012), no. 1-2, 25–44. MR 2925569
- [12] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, second ed., Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006. MR 2239987 (2007h:00002)
- [13] Dan Crisan and Boris Rozovskiĭ (eds.), *The Oxford handbook of nonlinear filtering*, Oxford University Press, Oxford, 2011. MR 2882749 (2012j:60003)
- [14] Jean-Dominique Deuschel and Daniel W. Stroock, *Large deviations*, Pure and Applied Mathematics, vol. 137, Academic Press Inc., Boston, MA, 1989. MR 997938 (90h:60026)
- [15] Randal Douc, Elisabeth Gassiat, Benoit Landelle, and Eric Moulines, *Forgetting of the initial distribution for nonergodic hidden Markov chains*, Ann. Appl. Probab. **20** (2010), no. 5, 1638–1662. MR 2724416 (2011k:93140)
- [16] William Evans, Claire Kenyon, Yuval Peres, and Leonard J. Schulman, *Broadcasting on trees and the Ising model*, Ann. Appl. Probab. **10** (2000), no. 2, 410–433. MR 1768240 (2001g:60243)
- [17] R. Fernández and C.-E. Pfister, *Global specifications and nonquasilocality of projections of Gibbs measures*, Ann. Probab. **25** (1997), no. 3, 1284–1315. MR 1457620 (98h:60066)
- [18] Hans Föllmer, *On entropy and information gain in random fields*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **26** (1973), 207–217. MR 0348829 (50 #1324)
- [19] ———, *On the global Markov property*, Quantum fields—algebras, processes (Proc. Sympos., Univ. Bielefeld, Bielefeld, 1978), Springer, Vienna, 1980, pp. 293–302. MR 601817 (82d:60191)
- [20] ———, *Random fields and diffusion processes*, École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87, Lecture Notes in Math., vol. 1362, Springer, Berlin, 1988, pp. 101–203. MR 983373 (90f:60099)
- [21] Arnaldo Frigessi, Fabio Martinelli, and Julian Stander, *Computational complexity of Markov chain Monte Carlo methods for finite Markov random fields*, Biometrika **84** (1997), no. 1, 1–18. MR 1450188 (98d:60093)
- [22] Hans-Otto Georgii, *Gibbs measures and phase transitions*, second ed., de Gruyter Studies in Mathematics, vol. 9, Walter de Gruyter & Co., Berlin, 2011. MR 2807681 (2012d:82015)
- [23] Hans-Otto Georgii, Olle Häggström, and Christian Maes, *The random geometry of equilibrium phases*, Phase transitions and critical phenomena, Vol. 18, Phase Transit. Crit. Phenom., vol. 18, Academic Press, San Diego, CA, 2001, pp. 1–142. MR 2014387 (2004h:82022)

- [24] Eli Glasner, *Ergodic theory via joinings*, Mathematical Surveys and Monographs, vol. 101, American Mathematical Society, Providence, RI, 2003. MR 1958753 (2004c:37011)
- [25] S. Goldstein, R. Kuik, and A. G. Schlijper, *Entropy and global Markov properties*, Comm. Math. Phys. **126** (1990), no. 3, 469–482. MR 1032869 (90m:82002)
- [26] Sheldon Goldstein, *Remarks on the global Markov property*, Comm. Math. Phys. **74** (1980), no. 3, 223–234. MR 578041 (83c:82006)
- [27] Sheldon Goldstein, Roelof Kuik, Joel L. Lebowitz, and Christian Maes, *From PCAs to equilibrium systems and back*, Comm. Math. Phys. **125** (1989), no. 1, 71–79. MR 1017739 (91b:82031)
- [28] Geoffrey Grimmett, *The random-cluster model*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 333, Springer-Verlag, Berlin, 2006. MR 2243761 (2007m:60295)
- [29] Hiroshi Kunita, *Asymptotic behavior of the nonlinear filtering errors of Markov processes*, J. Multivariate Anal. **1** (1971), 365–393. MR 0301812 (46 #967)
- [30] Joel L. Lebowitz, Christian Maes, and Eugene R. Speer, *Statistical mechanics of probabilistic cellular automata*, J. Statist. Phys. **59** (1990), no. 1-2, 117–170. MR 1049965 (91i:82012)
- [31] Thomas M. Liggett, *Interacting particle systems*, Classics in Mathematics, Springer-Verlag, Berlin, 2005, Reprint of the 1985 original. MR 2108619 (2006b:60003)
- [32] ———, *Conditional association and spin systems*, ALEA Lat. Am. J. Probab. Math. Stat. **1** (2006), 1–19. MR 2235171 (2007b:60241)
- [33] Pierre-Yves Louis, *Ergodicity of PCA: equivalence between spatial and temporal mixing conditions*, Electron. Comm. Probab. **9** (2004), 119–131 (electronic). MR 2108858 (2005k:60158)
- [34] J. C. Mattingly, *Ergodicity of dissipative SPDEs*, Lecture notes, École d’été de Probabilités de Saint-Flour, July 8–21, 2007 (unpublished), 2007.
- [35] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Communications and Control Engineering Series, Springer-Verlag London Ltd., London, 1993. MR 1287609 (95j:60103)
- [36] Marc Mézard and Andrea Montanari, *Information, physics, and computation*, Oxford Graduate Texts, Oxford University Press, Oxford, 2009. MR 2518205 (2010k:94019)
- [37] Andrea Montanari and Emmanuel Abbe, *Conditional random fields, planted constraint satisfaction and entropy concentration*, Approximation, Randomization, and Combinatorial Optimization (P. Raghavendra, S. Raskhodnikova, K. Jansen, and J. Rolim, eds.), Lecture Notes in Computer Science, vol. 8096, Springer, 2013, pp. 332–346.
- [38] Elchanan Mossel, Joe Neeman, and Allan Sly, *Reconstruction and estimation in the planted partition model*, Probab. Th. Rel. Fields (2014), to appear.

- [39] Hidetoshi Nishimori, *Duality in finite-dimensional spin glasses*, J. Stat. Phys. **126** (2007), no. 4-5, 977–986. MR 2311893 (2008f:82041)
- [40] Patrick Rebeschini and Ramon van Handel, *Can local particle filters beat the curse of dimensionality?*, Ann. Appl. Probab. (2014), to appear.
- [41] Patrick Rebeschini and Ramon van Handel, *Comparison theorems for Gibbs measures*, J. Stat. Phys. **157** (2014), 234–281.
- [42] Daniel Revuz and Marc Yor, *Continuous martingales and Brownian motion*, third ed., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 293, Springer-Verlag, Berlin, 1999. MR 1725357 (2000h:60050)
- [43] Lukasz Stettner, *On invariant measures of filtering processes*, Stochastic differential systems (Bad Honnef, 1988), Lecture Notes in Control and Inform. Sci., vol. 126, Springer, Berlin, 1989, pp. 279–292. MR 1236074
- [44] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer. **19** (2010), 451–559. MR 2652785 (2011i:65093)
- [45] Xin Thomson Tong and Ramon van Handel, *Ergodicity and stability of the conditional distributions of nondegenerate Markov chains*, Ann. Appl. Probab. **22** (2012), no. 4, 1495–1540. MR 2985168
- [46] ———, *Conditional ergodicity in infinite dimension*, Ann. Probab. **42** (2014), 2243–2313.
- [47] Aernout C. D. van Enter, Roberto Fernández, and Alan D. Sokal, *Regularity properties and pathologies of position-space renormalization-group transformations: scope and limitations of Gibbsian theory*, J. Statist. Phys. **72** (1993), no. 5-6, 879–1167. MR 1241537 (94m:82012)
- [48] Ramon van Handel, *Discrete time nonlinear filters with informative observations are stable*, Electron. Commun. Probab. **13** (2008), 562–575. MR 2461532 (2010a:60237)
- [49] ———, *Observability and nonlinear filtering*, Probab. Theory Related Fields **145** (2009), no. 1-2, 35–74. MR 2520121 (2010h:93123)
- [50] ———, *The stability of conditional Markov processes and Markov chains in random environments*, Ann. Probab. **37** (2009), no. 5, 1876–1925. MR 2561436 (2011a:60256)
- [51] ———, *Uniform observability of hidden Markov models and filter stability for unstable signals*, Ann. Appl. Probab. **19** (2009), no. 3, 1172–1199. MR 2537203 (2010j:60185)
- [52] ———, *On the exchange of intersection and supremum of σ -fields in filtering theory*, Israel J. Math. **192** (2012), no. 2, 763–784. MR 3009741
- [53] Heinrich von Weizsäcker, *Exchanging the order of taking suprema and countable intersections of σ -algebras*, Ann. Inst. H. Poincaré Sect. B (N.S.) **19** (1983), no. 1, 91–100. MR 699981 (85c:28001)
- [54] Gerhard Winkler, *Image analysis, random fields and Markov chain Monte Carlo methods*, second ed., Applications of Mathematics (New York), vol. 27, Springer-Verlag, Berlin, 2003. MR 1950762 (2004c:94028)